Probability and Random Processes

Lecture 12

- Detection
- Estimation
- Capacity
- Information

Detection

- A random process on $(E^T, \mathcal{E}^T)$ with two possible distributions μ_0 and μ_1
- Assume that $\mu_0 \gg \mu_1$ and $\mu_1 \gg \mu_0$ (the distributions are equivalent)
- Observe $f \in E^T$ and based on the observation

$$\text{decide } H_0 : \mu_0 \text{ or } H_1 : \mu_1$$

$\Longleftrightarrow$ design a measurable mapping $g : E^T \to \{0, 1\}$

Criteria

- Classical: minimize

$$P(g(f) = 0|H_1)$$

subject to $P(g(f) = 1|H_0) \leq \alpha$

- Bayesian: minimize

$$P_e = P(g(f) = 0|H_1)P(H_1) + P(g(f) = 1|H_0)P(H_0)$$

Bayesian detection

- Let $G_1 = g^{-1}(\{1\})$, $G_0 = g^{-1}(\{0\})$, assuming $G_1 \cup G_0 = E^T$ and $G_1 \cap G_0 = \emptyset$, then

$$\begin{aligned} P_e &= P(H_1) \int_{G_0} d\mu_1 + P(H_0) \int_{G_1} d\mu_0 \\ &= P(H_1) \int_{G_0} \left(\frac{d\mu_1}{d\mu_0} - \frac{P(H_0)}{P(H_1)} \right) d\mu_0 + P(H_0) \end{aligned}$$

- Hence, we should set

$$G_0 = \left\{ f : \frac{d\mu_1}{d\mu_0}(f) < \frac{P(H_0)}{P(H_1)} \right\}$$

$$G_1 = \left\{ f : \frac{d\mu_1}{d\mu_0}(f) > \frac{P(H_0)}{P(H_1)} \right\}$$

- Compare the likelihood ratio

$$\lambda(f) = \frac{d\mu_1}{d\mu_0}(f)$$

  to a threshold
- Classical $\Rightarrow$ Neyman–Pearson: also based on comparing λ to a threshold
- Given $f \in E^T$, how do we compute $\lambda(f)$?

Grenander's theorem

- Look at $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ and $T = \mathbb{R}^+$
- $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^T, \mathcal{B}^T, \mu)$ is separable if there is a set $N \subset \Omega$ for which $P(N) = 0$ and a sequence $S = \{t_k\} \subset T$ such that for any open interval I and closed set C

$$\{\omega : \pi_t(X(\omega)) \in C,\ t \in I \cap T\} \setminus \{\omega : \pi_t(X(\omega)) \in C,\ t \in I \cap S\} \subset N$$

  - i.e., $f(t) \in \mathbb{R}^T$ can be sampled without loss at the t_k's
- For any $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^T, \mathcal{B}^T, \mu)$ there is a $\tilde{X} : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^T, \mathcal{B}^T, \tilde{\mu})$ such that $\tilde{X}$ is separable and

$$P(\{\omega : \pi_t(X) = \pi_t(\tilde{X}), t \in T\}) = 1$$

- Consider the detection problem, and assume that μ_1 and μ_2 when restricted to $\{t_k\}_1^n$, for any finite n, are both absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$
- Observe $f(t)$, sample as $f_k = f(t_k)$ (with $\{t_k\}$ as in the definition of separability), let $f^n = (f_1, \dots, f_n)$ and denote the densities $g_1(f^n)$ and $g_2(f^n)$ (corresponding to μ_1 and μ_2)
- Then the entity

$$g_n = \frac{g_1(f^n)}{g_2(f^n)}$$

converges with probability one to $\lambda(f)$ under both H_0 and H_1

Gaussian waveforms

- Consider the continuous-time Gaussian example with two possible mean-value functions, that is, $f(t)$ is Gaussian with $E[f(t)] = m_i(t)$ under H_i, and has a positive-definite covariance kernel $V(s,u)$ (under both H_0 and H_1)
- Without loss we can assume $m_0(t) = 0$ and $m_1(t) = m(t)$
- Assume that $m(t)$ can be expressed as

$$m(t) = \int V(t,s)h(s)ds$$

for some $h(t)$, then

$$\ln \lambda(f) = \int f(t)h(t)dt - \frac{1}{2}\int m(t)h(t)dt$$

(with probability one under H_0 and H_1)

Estimation

Bayesian

- Two random objects X and Y on $(\Omega, \mathcal{A}, P)$ with range spaces $(\mathbb{R}, \mathcal{B})$ and $(E, \mathcal{E})$ (standard)
- Estimate $X \in \mathbb{R}$ from observing $Y = y \in E$; MMSE $\Rightarrow$

$$\hat{X}(y) = E[X|Y = y]$$

- That is,

$$\hat{X}(y) = \int x d\mu_y$$

where μ_y is the regular conditional distribution for X given $Y = y$

Classical

- For an absolutely continuous random variable X with pdf $f_\alpha(x)$; given the observation $X = x$ we have the traditional ML estimate

$$\hat{\alpha} = \arg\max_{\alpha} f_\alpha(x)$$

- A pdf $f(x)$ is the Radon–Nikodym derivative of the distribution μ on $(\mathbb{R}, \mathcal{B})$ w.r.t. Lebesgue measure λ; that is, it can be interpreted as the likelihood ratio between the hypothesis $H_0 : \mu = \lambda$ and $H_1 : \mu = \mu_X$ (the correct distribution)
- In the case of a general random object $X : (\Omega, \mathcal{A}, P) \to (E, \mathcal{E}, \mu)$, we can choose a "dummy hypothesis" $H_0 : \mu = \mu_0$ as a reference to $H_1 : \mu = \mu_\alpha$, where μ_α is the correct distribution, with an unknown parameter $\alpha \in \mathbb{R}$

- Then, based on the observation $X = x$, the ML estimate can be computed as

$$\hat{\alpha} = \arg\max_{\alpha} \frac{d\mu_\alpha}{d\mu_0}(x)$$

- The reference distribution can be chosen e.g. such that computing the likelihood ratio is feasible
- Note that in general, the estimate "$\hat{\alpha} = \arg\max \mu_\alpha$" does not make sense; μ_α is a mapping from sets in $\mathcal{E}$

Channels

- Given two measurable spaces $(\Omega, \mathcal{A})$ and $(\Gamma, \mathcal{S})$, a mapping $g : \Omega \times \mathcal{S} \to \mathbb{R}^+$ is called a transition kernel (from Ω to Γ) if
  1. $f(\omega) = g(\omega, S)$ is measurable for any fixed $S \in \mathcal{S}$
  2. $h(S) = g(\omega, S)$ is a measure on $(\Gamma, \mathcal{S})$ for any fixed $\omega \in \Omega$

  If the measure $h(S)$ in 2. is a probability measure, then g is called a stochastic kernel
- If Y is a random object on $(\Omega, \mathcal{A}, P)$ with values in $(E, \mathcal{E})$, then a stochastic kernel g from Ω to E is the regular conditional distribution of Y given $\mathcal{G} \subset \mathcal{A}$ if

$$g(\omega, F) = P(\{Y \in F\}|\mathcal{G})(\omega)$$

  with probability one w.r.t. P and ω, and for all $F \in \mathcal{E}$
- A regular conditional distribution for Y exists if $(E, \mathcal{E})$ is standard

- *The factorization lemma*: Assume two measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ and a measurable mapping $u : \Omega_1 \to \Omega_2$ are given. A function $v : (\Omega_1, \mathcal{A}_1) \to (\mathbb{R}, \mathcal{B})$ is measurable w.r.t $\sigma(u) \subset \mathcal{A}_1$ iff there is a measurable mapping $\phi : (\Omega_2, \mathcal{A}_2) \to (\mathbb{R}, \mathcal{B})$ such that $v = \phi \circ u$
- Let X be an arbitrary random object on $(\Omega, \mathcal{A}, P)$, and let Y be as before, with $(E, \mathcal{E})$ standard. Let

$$g(\omega, F) = P(\{Y \in F\}|\sigma(X))(\omega)$$

and let ϕ_F be the mapping in the factorization lemma $g(\omega, F) = \phi_F(X(\omega))$ (w.r.t. ω for a fixed F), then

$$P(\{Y \in F\}|X = x) = \phi_F(x)$$

is the conditional distribution of Y given $X = x$

- Assume $(\Omega, \mathcal{A}, P)$, a parameter set T and a process $X : (\Omega, \mathcal{A}, P) \to (E^T, \mathcal{E}^T, \mu_X)$ are given
- Given another function/sequence space $(F^T, \mathcal{F}^T)$, a channel is a regular conditional distribution from Ω to F^T given a specific value $X = x \in E^T$; that is, for any $x \in E^T$ the distribution

$$\mu_x(F) = P(\{Y \in F\}|X = x), \ \ F \in \mathcal{F}^T$$

- *Interpretation*: A random channel input X is generated and is then transmitted over the channel, resulting in the channel output Y; given $X = x$ the distribution for Y is μ_x
- A channel exists if the relevant spaces are standard

- Given $(E^T, \mathcal{E}^T)$ and $(F^T, \mathcal{F}^T)$, let $\mathcal{E}^T \times \mathcal{F}^T$ be the product σ-algebra on $E^T \times F^T$
- Let $\tilde{\mu}$ be defined by

$$\tilde{\mu}(A, B) = \int_A P(B|X = x) d\mu_X(x)$$

on rectangles, $A \in \mathcal{E}^T, \ B \in \mathcal{F}^T$
- Standard spaces $\Rightarrow$ unique extension of $\tilde{\mu}$ from rectangles to $\mathcal{E}^T \times \mathcal{F}^T$; a joint distribution μ on $(E^T \times F^T, \mathcal{E}^T \times \mathcal{F}^T)$
- Also define the corresponding product distribution π, generated as the extension of $\mu_X(A)\mu_Y(B)$, with

$$\mu_Y(B) = \int_\Omega P(B|X = x) d\mu_X(x)$$

Channel Capacity

- Focus on $T = \mathbb{N}^+$ and $E = F = \mathbb{R}$; a channel $\mu_x(\cdot)$ with input $x \in \mathbb{R}^T$ and output $y \in \mathbb{R}^T$ (sequences), resulting in the joint distribution μ
- A rate R [bits per channel use] is achievable if information can be transmitted at R with error probability below ε for any $\varepsilon > 0$
- The capacity C of the channel $= \sup\{R : \ R \text{ is achievable}\}$

- Let $S_n = \{1, 2, \ldots, n\}$, define the information density

$$i(x, y) = \log \frac{d\mu}{d\pi}$$

(assuming $\pi \gg \mu$), and the corresponding restricted version

$$i_n(x^n, y^n) = \log \frac{d\mu_{|S_n}}{d\pi_{|S_n}}$$

- Let

$$\gamma(\mu_X) = \sup \left\{ \alpha : \lim_{n \to \infty} P\big(n^{-1} i_n \leq \alpha\big) = 0 \right\}$$

- A general formula for channel capacity [Verdú–Han, '94]:

$$C = \sup_{\mu_X} \gamma(\mu_X)$$

- Computing C involves the problem of characterizing the limit γ (for each fixed μ_X) $\Rightarrow$ ergodic theory

Information Measures

- Given $(\Omega, \mathcal{A})$, a measurable partition of Ω is a finite collection $G_1, \ldots, G_n$, $G_i \in \mathcal{A}$, such that $G_k \cap G_l = \emptyset$ for $k \neq l$ and $\cup_i G_i = \Omega$
- Given two probability measures P and Q on $(\Omega, \mathcal{A})$ and a measurable partition $\mathcal{G} = \{G_i\}_{i=1}^n$ of Ω, define

$$D^*(P\|Q)(\mathcal{G}) = \sum_{G \in \mathcal{G}} P(G) \log \frac{P(G)}{Q(G)}$$

- Then, the relative entropy between P and Q is defined as

$$D(P\|Q) = \sup_{\mathcal{G}} D^*(P\|Q)(\mathcal{G})$$

- If $P \ll Q$, then we get

$$D(P\|Q) = \int \log \frac{dP}{dQ}(\omega)\, dP(\omega)$$

- Let X and Y be two random objects on $(\Omega, \mathcal{A}, P)$ with range spaces $(\Gamma, \mathcal{S})$ and $(\Lambda, \mathcal{U})$, and a joint distribution μ_{XY} on $(\Gamma \times \Lambda, \mathcal{S} \times \mathcal{U})$ corresponding to the marginal distributions μ_X and μ_Y
- Let π_{XY} be the corresponding product distribution
- The mutual information between X and Y is then defined as

$$I(X;Y) = \sup_{\mathcal{F}} D^*(\mu_{XY} \| \pi_{XY})(\mathcal{F})$$

  over measurable partitions $\mathcal{F}$ of $\Gamma \times \Lambda$
- The entropy of the single variable X is defined as

$$H(X) = I(X;X)$$

- If $\mu_{XY} \ll \pi_{XY}$, then

$$I(X;Y) = \int \log \frac{d\mu_{XY}}{d\pi_{XY}} \, d\mu_{XY}$$

- Returning to the setup of transmission over a channel (with the previous notation), if $\pi \gg \mu$ we had

$$i_n(x^n, y^n) = \log \frac{d\mu_{|S_n}}{d\pi_{|S_n}}$$

- If for any fixed stationary and ergodic input, with distribution μ_X, the channel is such that the joint input–output process on $(\mathbb{R}^T \times \mathbb{R}^T, \mathcal{B}^T \times \mathcal{B}^T)$ is stationary and ergodic, and in addition satisfies the finite-gap information property (below), then

$$\frac{1}{n} i_n \to \lim_{n\to\infty} \frac{1}{n} I(X^n; Y^n)$$

  with probability one

- finite-gap information: for any $n > 0$ there is a $k \geq n$ such that $I(X_k; X^n | X^k)$ and $I(Y_k; Y^n | Y^k)$ are both finite

- Letting

$$i_\infty(\mu_X) = \lim_{n\to\infty} \frac{1}{n} I(X^n; Y^n)$$

  for any fixed μ_X, we get

$$C = \sup_{\mu_X} i_\infty(\mu_X)$$

- Channels that result in this formula for C have been called information stable
- To prove this, one first needs to see that $\gamma = i_\infty$ for any fixed μ_X such that the input, output and joint input–output are stationary and ergodic. Then one needs to show that the supremum is achieved in this class.