Statistical Signal Processing: Introduction to Classical Estimation Theory & Classical and Quantum Detection Theory

Ranjith Nair

Quantum Measurement Group
Department of Electrical & Computer Engineering
National University of Singapore, Singapore

July 14, 2018

- Statistical Signal Processing (SSP) deals with optimal methods for extracting unknown/random signals from noisy observations.
- "statistical" – noise is treated statistically.
- "signal" – feature(s) of interest.
- "unknown/random" – is the signal suitably described by a statistical model or not?
- "optimal" – in terms of a performance criterion suitable to the problem.
- **Detection theory** studies problems in which the set of signals is finite or discrete: $\mathcal{M} = \{1, \ldots, M\}$
    - **Applications:** Digital communication; radar/sonar target detection; medical diagnosis; astronomical observations.
- **Estimation theory** studies problems in which the set of signals $\Theta \subset \mathbb{R}^N$.
    - **Applications:** Analog communication; measurements of physical quantities; radar ranging; siesmology; imaging.
- Why study **classical** SSP?
    - In quantum SSP, signals are carried by quantum systems and affected by quantum noise, but problem objectives are similar.
    - Theoretical concepts and techniques are analogous.

Probability Theory Concepts

- Random variables X, Y, Z, R etc. (upper case).
- Instances $x \in \mathcal{X}, y \in \mathcal{Y}$, etc. (lower case).
- Calligraphic case for the ranges $x \in \mathcal{X}$, $y \in \mathcal{Y}$, etc.
- Probability distributions $P(x), P(x,y), P(x|y)$, etc. (sometimes $P_X(x), P_{XY}(x,y)$ for clarity).
- When the distribution function P is known, we can also consider $P_(X)$ as a random variable.
- **Statistical expectation**:

$$\mathbb{E}\left[f(X,Y)\right] := \int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{Y}} \mathrm{d}y\, P(x,y)\, f(x,y) \tag{1}$$

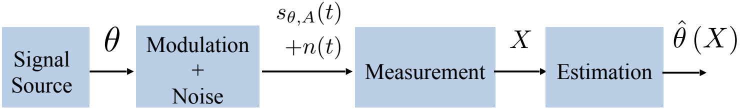
$$= \int_{\mathcal{X}} \mathrm{d}x P(x) \int_{\mathcal{Y}} \mathrm{d}y\, P(y|x)\, f(x,y) \tag{2}$$

$$= \mathbb{E}_X \left[\mathbb{E}_{Y|X=x}\left[f(X,Y)\right]\right]. \text{ (iterated expectation)} \tag{3}$$

- Expectation is additive: $\mathbb{E}\left[f(X,Y) + g(X,Y)\right] = \mathbb{E}\left[f(X,Y)\right] + \mathbb{E}\left[g(X,Y)\right]$.
- **Statistical independence**: Random variables $X_1, \ldots, X_M$ are said to be statistically independent if

$$P_{X_1 \cdots X_M}(x_1, \ldots, x_M) = P_{X_1}(x_1) P_{X_2}(x_2) \cdots P_{X_M}(x_M)\ \forall\, x_1, \ldots, x_M. \tag{4}$$

- **Classical (traditional)** estimation theory vs. **Bayesian** estimation theory: the parameter $\theta \in \Theta \subset \mathbb{R}^N$ is regarded as **unknown** vs. **random** with a prior probability distribution $P_\Theta(\theta)$. We will confine ourselves to classical estimation theory.



- $\theta \in \Theta$, the set of possible parameter values.
- $\mathcal{X}$, the space of observations.
- The probability distribution $P(x|\theta)$ on $\mathcal{X}$ induced by each value $\theta \in \Theta$ (**likelihood function**).
- **Estimator** $\hat{\theta}(X)$: rule that gives the estimate of θ from the observation (cannot depend on θ!).
- **Note**: $\hat{\theta} \in \Theta$ will refer to both the random variable and its instances.

- Performance criterion is the **mean square error** (MSE)

$$\mathrm{MSE}_\theta[\hat{\theta}] := \mathbb{E}\left[\hat{\theta} - \theta\right]^2 \tag{5}$$

$$= \int_{\mathcal{X}} \mathrm{d}x P(x|\theta)\left[\hat{\theta}(x) - \theta\right]^2 . \tag{6}$$

- Ideally, we would like to find an estimator $\hat{\theta}^{\text{ideal}}$ whose MSE satisfies

$$\mathrm{MSE}_\theta\left[\hat{\theta}^{\text{ideal}}\right] \leq \mathrm{MSE}_\theta\left[\hat{\theta}\right] \text{ for all } \theta \in \Theta. \tag{7}$$

- Very difficult to solve this problem (how to optimize over *all* estimators?). We will have to restrict the class of allowed estimators.
- **Bias** of an estimator: $b_\theta\left[\hat{\theta}\right] = \mathbb{E}\left[\hat{\theta}\right] - \theta$.
- An estimator $\hat{\theta}$ is **unbiased** if

$$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta \ \textbf{ for all } \theta \in \Theta. \tag{8}$$

- For unbiased estimators, $\mathrm{MSE}_\theta\left[\hat{\theta}\right] = \mathrm{Var}\left[\hat{\theta}\right]$.

Cramér-Rao lower bound

Theorem (Cramér-Rao bound)

Suppose the PDF $P(X|\theta)$ satisfies the "regularity condition"

$$\mathbb{E}\left[\frac{\partial \ln P(X|\theta)}{\partial\theta}\right] = \int_{\mathcal{X}} \mathrm{d}x \frac{\partial P(x|\theta)}{\partial\theta} = 0 \;\forall\; \theta \in \Theta. \tag{9}$$

Then the variance of any <u>unbiased</u> estimator $\hat{\theta}(X)$ of θ satisfies

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \left(\mathbb{E}\left[\left(\frac{\partial \ln P(X|\theta)}{\partial\theta}\right)^2\right]\right)^{-1} \tag{10}$$

$$= -\left(\mathbb{E}\left[\frac{\partial^2 \ln P(X|\theta)}{\partial\theta^2}\right]\right)^{-1}. \tag{11}$$

$\frac{\partial \ln P(X|\theta)}{\partial\theta} \equiv L_\theta(X) =$ **logarithmic derivative**

Proof.

$$\frac{\partial \ln P(x|\theta)}{\partial \theta} = \frac{1}{P(x|\theta)} \frac{\partial P(x|\theta)}{\partial \theta}$$

$$-\frac{\partial^2 \ln P(x|\theta)}{\partial \theta^2} = \frac{1}{P^2(x|\theta)} \left[\frac{\partial P(x|\theta)}{\partial \theta}\right]^2 - \frac{1}{P(x|\theta)} \frac{\partial^2 P(x|\theta)}{\partial \theta^2}.$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln P(x|\theta)}{\partial \theta^2}\right] = \mathbb{E}\left[L_\theta^2(X)\right] - \int_{\mathcal{X}} \mathrm{d}x\, \frac{\partial^2 P(x|\theta)}{\partial \theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln P(x|\theta)}{\partial \theta^2}\right] = \mathbb{E}\left[L_\theta^2(X)\right] - \frac{\partial}{\partial \theta} \int_{\mathcal{X}} \mathrm{d}x\, \frac{\partial P(x|\theta)}{\partial \theta}$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln P(x|\theta)}{\partial \theta^2}\right] = \mathbb{E}\left[L_\theta^2(X)\right] \quad \text{(regularity)}$$

□

$\mathbb{E}\left[L_\theta^2(X)\right] = -\mathbb{E}\left[\frac{\partial L_\theta(X)}{\partial \theta}\right] \equiv \mathcal{J}_\theta[X] =$ **Fisher information** (of X on θ).

Examples

Example (DC value in additive white Gaussian noise (AWGN))

Suppose $\theta \in \Theta = \mathbb{R}; \mathbf{X} = (X_1, X_2, \ldots, X_N)^\mathsf{T} \in \mathcal{X} = \mathbb{R}^N$ such that

$$X_n = \theta + W_n, \;\; 1 \leq n \leq N, \tag{12}$$

where $W_n \sim \mathcal{N}(0, \sigma^2)$, and are independent and identically distributed (i.i.d.).

$$P(\mathbf{x}|\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{\sum_{n=1}^{N}(x_n - \theta)^2}{2\sigma^2}\right]. \tag{13}$$

$$\ln P(\mathbf{x}|\theta) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sum_{n=1}^{N}(x_n - \theta)^2}{2\sigma^2} \tag{14}$$

$$\frac{\partial \ln P(\mathbf{x}|\theta)}{\partial\theta} = \frac{\sum_{n=1}^{N} x_n - N\theta}{\sigma^2} \tag{15}$$

$$-\frac{\partial^2 \ln P(\mathbf{x}|\theta)}{\partial\theta^2} = \frac{N}{\sigma^2} = \mathcal{J}_\theta[\mathbf{X}]. \tag{16}$$

$$\text{Sample mean } \hat{\theta}[\mathbf{X}] := \frac{\sum_{n=1}^{N} X_n}{N} \sim \mathcal{N}\left(\theta, \sigma^2/N\right). \tag{17}$$

Example (DC value in additive white noise)

Suppose $\theta \in \Theta = \mathbb{R}; \mathbf{X} = (X_1, X_2, \ldots, X_N)^\mathsf{T} \in \mathcal{X} = \mathbb{R}^N$ such that

$$X_n = \theta + W_n, \quad 1 \leq n \leq N, \tag{18}$$

with i.i.d. $\{W_n\}$ such that $\mathbb{E}[W_n] = 0, \mathbb{E}\left[W_n^2\right] = \sigma^2$, but otherwise *arbitrarily* distributed.

$$\mathcal{J}_\theta[X_n] = \int_\mathcal{X} \mathrm{d}x \, \overbrace{\frac{1}{P(x_n|\theta)} \left[\frac{\partial P(x_n|\theta)}{\partial \theta}\right]^2}^{f^2(x)}$$
$$\geq \frac{\left|\int_\mathcal{X} \mathrm{d}x \, f(x) g(x)\right|^2}{\int_\mathcal{X} \mathrm{d}x \, g^2(x)}$$

$$\text{Take } g(x) := \sqrt{P(x|\theta)}(x - \theta) \implies \int_\mathcal{X} \mathrm{d}x \, g^2(x) = \mathrm{Var}[W] = \sigma^2.$$

$$\int_\mathcal{X} \mathrm{d}x \, f(x) g(x) = \int_\mathcal{X} \mathrm{d}x \, (x - \theta) \frac{\partial P(x|\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \int_\mathcal{X} \mathrm{d}x \, x \, P(x|\theta) = 1.$$
$$\implies \mathcal{J}_\theta[\mathbf{X}] \geq N/\sigma^2.$$

For additive noise with a given variance, it is "hardest" to estimate a DC value if the noise is Gaussian.

Example (Arbitrarily modulated parameter in AWGN)

Suppose $\theta \in \Theta \subset \mathbb{R}$ is modulated onto a given discrete-time waveform $s_{n;\theta}, 1 \leq n \leq N$, which is observed in AWGN:

$$X_n = s_{n;\theta} + W_n, \quad 1 \leq n \leq N, \tag{19}$$

with $W_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

$$P(\mathbf{X}|\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{\sum_{n=1}^N (X_n - s_{n;\theta})^2}{2\sigma^2}\right]. \tag{20}$$

$$\ln P(\mathbf{X}|\theta) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sum_{n=1}^N (X_n - s_{n;\theta})^2}{2\sigma^2} \tag{21}$$

$$\frac{\partial \ln P(\mathbf{X}|\theta)}{\partial\theta} = \frac{1}{\sigma^2}\sum_{n=1}^N \frac{\partial s_{n;\theta}}{\partial\theta}(X_n - s_{n;\theta}) \tag{22}$$

$$-\frac{\partial^2 \ln P(\mathbf{X}|\theta)}{\partial\theta^2} = \frac{1}{\sigma^2}\left[\sum_{n=1}^N \left(\frac{\partial s_{n;\theta}}{\partial\theta}\right)^2 - \sum_{n=1}^N \frac{\partial^2 s_{n;\theta}}{\partial\theta^2}(X_n - s_{n;\theta})\right] \tag{23}$$

$$\mathcal{J}_\theta[\mathbf{X}] = \frac{\sum_{n=1}^N \left(\frac{\partial s_{n;\theta}}{\partial\theta}\right)^2}{\sigma^2} \equiv \frac{\left\|\dot{\vec{\mathbf{s}}}\right\|^2}{\sigma^2} \implies \mathrm{Var}\left[\hat{\theta}(\mathbf{X})\right] \geq \left\|\dot{\vec{\mathbf{s}}}\right\|^2/\sigma^2. \tag{24}$$

Examples

Example (Estimating the variance of a Gaussian)

Suppose $v \in \mathbb{R}^+; \mathcal{X} = \mathbb{R}^N$ and $X_n \sim \mathcal{N}(0, v)$ and i.i.d. for $1 \leq n \leq N$.

$$P(\mathbf{X}|v) = (2\pi v)^{-N/2} \exp\left[-\frac{\sum_{n=1}^N X_n^2}{2v}\right].$$

$$\ln P(\mathbf{X}|v) = -\frac{N}{2}\ln(2\pi v) - \frac{\sum_{n=1}^N X_n^2}{2v}.$$

$$\frac{\partial \ln P(\mathbf{X}|v)}{\partial v} = -\frac{N}{2v} + \frac{\sum_{n=1}^N X_n^2}{2v^2}.$$

$$-\frac{\partial^2 \ln P(\mathbf{X}|v)}{\partial v^2} = -\frac{N}{2v^2} + \frac{\sum_{n=1}^N X_n^2}{v^3} \implies \mathcal{J}_v[\mathbf{X}] = \frac{N}{2v^2}.$$

$\hat{V}_{\text{unbias}} := \frac{1}{N}\left(\sum_{n=1}^N X_n^2\right)$ achieves the CRB (Exercise).

$\hat{V}_{\text{bias}} := \frac{N}{N+2}\left(\sum_{n=1}^N X_n^2\right)$ has MSE $\frac{2v^2}{N+2} < \frac{2v^2}{N}$ (Exercise*).

* P. Stoica and R. Moses, *Signal Processing*, vol. 21, pp. 349-350 (1990).

Theorem (Chain Rule)

Suppose X and Y are two observations with the parameter-dependent joint probability distribution $P(X, Y|\theta)$. Their combined Fisher information on θ equals

$$\begin{aligned}\mathcal{J}_\theta[XY] &= \mathcal{J}_\theta[X] + \mathcal{J}_\theta[Y|X] \\ &\equiv \mathbb{E}\left[\left(\frac{\partial \ln P(X|\theta)}{\partial\theta}\right)^2\right] + \mathbb{E}\left[\left(\frac{\partial \ln P(Y|X,\theta)}{\partial\theta}\right)^2\right]\end{aligned}$$

More generally, for a vector observation $\mathbf{X} = (X_1, X_2, \ldots, X_M)$,

$$\mathcal{J}_\theta\left[X_1 X_2 \cdots X_M\right] = \mathcal{J}_\theta\left[X_1\right] + \mathcal{J}_\theta\left[X_2|X_1\right] + \cdots + \mathcal{J}_\theta\left[X_M|X_1 \cdots X_{M-1}\right].$$

Proof.

We have

$$\left(\frac{\partial \ln P(X,Y|\theta)}{\partial\theta}\right)^2 = \left(\frac{\partial \ln P(X|\theta)}{\partial\theta} + \frac{\partial \ln P(Y|X,\theta)}{\partial\theta}\right)^2$$

$$= \left(\frac{\partial \ln P(X|\theta)}{\partial\theta}\right)^2 + \left(\frac{\partial \ln P(Y|X,\theta)}{\partial\theta}\right)^2 + 2\left(\frac{\partial \ln P(X|\theta)}{\partial\theta}\right)\left(\frac{\partial \ln P(Y|X,\theta)}{\partial\theta}\right).$$

Taking expectations, we get

$$= \mathcal{J}_\theta[X] + \mathcal{J}_\theta[Y|X] + 2\int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{Y}} \mathrm{d}y\, P(x|\theta)P(y|x,\theta)\frac{\partial \ln P(X|\theta)}{\partial\theta}\frac{\partial \ln P(y|x,\theta)}{\partial\theta}$$

$$= \mathcal{J}_\theta[X] + \mathcal{J}_\theta[Y|X] + 2\int_{\mathcal{X}} \mathrm{d}x P(x|\theta)\frac{\partial \ln P(X|\theta)}{\partial\theta}\int_{\mathcal{Y}} \mathrm{d}y \frac{\partial P(y|x,\theta)}{\partial\theta}$$

$$= \mathcal{J}_\theta[X] + \mathcal{J}_\theta[Y|X].$$

□

Corollary (Additivity & Data Processing)

1. *(Additivity for independent observations) If $X_1, \ldots, X_M$ are independent observations, we have*

$$\mathcal{J}_\theta[X_1 X_2 \ldots X_M] = \sum_{m=1}^{M} \mathcal{J}_\theta[X_m]. \tag{25}$$

2. *(Data Processing inequality) For any function f, we have*

$$\mathcal{J}_\theta[X] \geq \mathcal{J}_\theta\left[f(X)\right]. \tag{26}$$

Proof.

1. We have $\ln P(X_1, X_2, \ldots, X_M) = \sum_{m=1}^{M} \ln P(X_m)$.
2. Using the chain rule in two ways, we have

$$\begin{aligned} \mathcal{J}_\theta\left[X\, f(X)\right] &= \mathcal{J}_\theta\left[f(X)\right] + \mathcal{J}_\theta\left[X | f(X)\right] \\ &= \mathcal{J}_\theta\left[X\right] + \mathcal{J}_\theta\left[f(X) | X\right] = \mathcal{J}_\theta\left[X\right]. \\ &\implies \mathcal{J}_\theta[X] \geq \mathcal{J}_\theta\left[f(X)\right]. \end{aligned}$$

□

Is Fisher information subadditive?

If X_1 and X_2 are two observations with joint probability distribution $P(X_1, X_2|\theta)$, do we have

$$\mathcal{J}_\theta[X_1\, X_2] \leq \mathcal{J}_\theta[X_1] + \mathcal{J}_\theta[X_2];$$
$$\mathcal{J}_\theta[X_2|X_1] \leq \mathcal{J}_\theta[X_2]?$$

Is Fisher information subadditive?

If X_1 and X_2 are two observations with joint probability distribution $P(X_1, X_2|\theta)$, do we have

$$\mathcal{J}_\theta[X_1\, X_2] \leq \mathcal{J}_\theta[X_1] + \mathcal{J}_\theta[X_2];$$
$$\mathcal{J}_\theta[X_2|X_1] \leq \mathcal{J}_\theta[X_2]?$$

Example 1: Suppose that, for $\theta \in \mathbb{R}$, we have observations:

$$X_1 = \theta + N_1,$$
$$X_2 = \theta + \alpha N_1 + (1-\alpha)N_2,$$

for some $0 \leq \alpha \leq 1$ and $N_1, N_2 \sim \mathcal{N}(0, \sigma^2)$ are independent noises. Given $X_1 = x_1$ and θ,

$$X_2 = \alpha x_1 + (1-\alpha)\theta + (1-\alpha)N_2, \text{ so that}$$
$$X_2 \sim \mathcal{N}\left(\alpha x_1 + (1-\alpha)\theta, (1-\alpha)^2\sigma^2\right),$$
$$\mathcal{J}_\theta[X_2|X_1] = \begin{cases} \frac{1}{\sigma^2} & \text{if } \alpha < 1, \\ 0 & \text{otherwise,} \end{cases}$$
$$\mathcal{J}_\theta[X_2] = \frac{1}{\left[\alpha^2 + (1-\alpha)^2\right]\sigma^2}.$$
$$\mathcal{J}_\theta[X_2|X_1] < \mathcal{J}_\theta[X_2].$$

If X_1 and X_2 are two observations with joint probability distribution $P(X_1, X_2|\theta)$, do we have

$$\mathcal{J}_\theta[X_1\, X_2] \leq \mathcal{J}_\theta[X_1] + \mathcal{J}_\theta[X_2];$$
$$\mathcal{J}_\theta[X_2|X_1] \leq \mathcal{J}_\theta[X_2]?$$

Example 2: Suppose that, for $\theta \in \mathbb{R}$, we have observations:

$$X_1 = \theta + N_1,$$
$$X_2 = \theta - \alpha N_1 + (1-\alpha)N_2,$$

for some $0 \leq \alpha \leq 1$ and $N_1, N_2 \sim \mathcal{N}(0, \sigma^2)$ are independent noise variables. Given $X_1 = x_1$ and θ,

$$X_2 = -\alpha x_1 + (1+\alpha)\theta + (1-\alpha)N_2, \text{ so that}$$
$$X_2 \sim \mathcal{N}\left(-\alpha x_1 + (1+\alpha)\theta, (1-\alpha)^2\sigma^2\right).$$

$$\mathcal{J}_\theta[X_2|X_1] = \frac{(1+\alpha)^2}{(1-\alpha)^2\sigma^2}$$
$$\mathcal{J}_\theta[X_2] = \frac{1}{\left[\alpha^2 + (1-\alpha)^2\right]\sigma^2}.$$
$$\mathcal{J}_\theta[X_2|X_1] > \mathcal{J}_\theta[X_2].$$

- Given a family of probability distributions $P(X|\theta)$ on $\mathcal{X}$, for any $\theta, \theta' \in \Theta$, the **Bhattacharyya coefficient** $B(\theta, \theta')$ is defined as:

$$B(\theta, \theta') = \int_{\mathcal{X}} \mathrm{d}x \sqrt{P(x|\theta)\, P(x|\theta')};$$
$$0 \leq B(\theta, \theta') \leq 1.$$

- We have

$$\frac{\partial B(\theta, \theta')}{\partial \theta'} = \frac{1}{2} \int_{\mathcal{X}} \mathrm{d}x \sqrt{\frac{P(x|\theta)}{P(x|\theta')}} \frac{\partial P(x|\theta')}{\partial \theta'},$$

$$\frac{\partial^2 B(\theta, \theta')}{\partial \theta'^2} = -\frac{1}{4} \int_{\mathcal{X}} \mathrm{d}x \frac{1}{P(x|\theta')} \sqrt{\frac{P(x|\theta)}{P(x|\theta')}} \left[\frac{\partial P(x|\theta')}{\partial \theta'}\right]^2 + \frac{1}{2} \int_{\mathcal{X}} \mathrm{d}x \sqrt{\frac{P(x|\theta)}{P(x|\theta')}} \frac{\partial^2 P(x|\theta')}{\partial \theta'^2}$$

- Setting $\theta' = \theta$, we get

$$\mathcal{J}_\theta[X] = -4 \frac{\partial^2 B(\theta, \theta')}{\partial \theta'^2}\bigg|_{\theta'=\theta}$$

- Given an observation $x \in \mathcal{X}$ drawn from one of a finite family of probability distributions $\{P(x|i)\}_{i=1}^{M}$ on $\mathcal{X}$, we wish to determine the index i.
- Two paradigms: i is a **nonrandom** parameter (Neyman-Pearson) / I is a **random variable** (Bayesian).
- **Applications** (Neyman-Pearson): medical diagnosis, target detection, gravitational wave detection.

  (Bayesian): digital communication, theoretical bounds on other problems.

- Two possible **hypotheses** $i \in \{0, 1\}$:

$$H_0 : P(x|0);\; x \in \mathcal{X} \text{ (null hypothesis).}$$
$$H_1 : P(x|1);\; x \in \mathcal{X} \text{ (alternative hypothesis).}$$

- A decision rule $J(X) = \hat{I}(X)$ representing the "best" decision that can be taken given the observation.
- Equivalent to partitioning the observation space into two regions $\mathcal{X}_0$ and $\mathcal{X}_1$ such that $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1; \mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$.
- Two kinds of errors:

$$\Pr\left[\mathcal{X}_1|H_0\right] = \int_{\mathcal{X}_1} \mathrm{d}x\, P(x|0) \equiv P_F \text{ (false-alarm probability),}$$
$$\Pr\left[\mathcal{X}_0|H_1\right] = \int_{\mathcal{X}_0} \mathrm{d}x\, P(x|1) \equiv P_M \text{ (miss probability).}$$

- In the Bayesian approach, I is a random variable with its two values having the prior probabilities $\{\pi_0, \pi_1\}$.

Nonrandom case: Neyman-Pearson lemma

- Minimizing P_F ("expand $\mathcal{X}_0$") and minimizing P_M ("expand $\mathcal{X}_1$") are conflicting objectives.
- For each value of $P_F \in [0, 1]$, find the minimum achievable P_M (and the corresponding $\mathcal{X}_0$).
- The resulting curve of minimum P_M vs. P_F is called the receiver operating characteristic (ROC).

Theorem (Neyman-Pearson lemma)

For a given $P_F = \alpha$, the decision rule that minimizes P_M is to decide H_1 whenever the likelihood ratio

$$L(x) = \frac{P(x|1)}{P(x|0)} > \gamma,$$

where the threshold γ is determined by the condition

$$P_F = \int_{x:L(x)>\gamma} P(x|0)\,\mathrm{d}x = \alpha.$$

- When prior probabilities $\{\pi_0, \pi_1\}$ are assigned, we choose to minimize the *average error probability*

$$P_E = \pi_0 P_F + \pi_1 P_M. \tag{27}$$

- Using the decision function $I_1(x)$ as before, we can write

$$\begin{aligned} P_E &= \int_{\mathcal{X}} \mathrm{d}x \left[I_1(x)\, \pi_0 P(x|0) + I_0(x)\, \pi_1 P(x|1) \right], \\ &\geq \int_{\mathcal{X}} \mathrm{d}x \min \left\{ \pi_0 P(x|0), \pi_1 P(x|1) \right\}, \end{aligned}$$

$$a \leq b \implies a\,(b) = (a+b)/2 \mp |a-b|/2.$$

$$\begin{aligned} P_E &= \frac{1}{2} \int_{\mathcal{X}} \mathrm{d}x \left[\pi_0 P(x|0) + \pi_1 P(x|1) - \left| \pi_0 P(x|0) - \pi_1 P(x|1) \right| \right] \\ &= \frac{1}{2} - \frac{1}{2} \int_{\mathcal{X}} \mathrm{d}x \left| \pi_0 P(x|0) - \pi_1 P(x|1) \right| \\ &\equiv \frac{1}{2} - \frac{1}{2} \left\| \pi_0 P(X|0) - \pi_1 P(X|1) \right\|_1 \text{ (l_1-distance).} \end{aligned}$$

- Optimum decision rule $J(x) = \arg\max_{j \in \{0,1\}} \{\pi_j P(x|j)\}$.

Random case: M-ary hypothesis testing

- $M > 2$ hypotheses $\{H_i \equiv P_X(X|i)\}_{i=1}^M$ with prior probabilities $\{\pi_i\}_{i=1}^M$.
- Define the decision functions $\{M_j(x)\}_{j=1}^M$ such that

$$0 \leq M_j(x) \leq 1, \;\; \forall j = 1, \ldots, M,$$
$$\sum_{j=1}^{M} M_j(x) = 1, \;\; \forall x \in \mathcal{X}.$$

- The *probability of correct decision* is:

$$\begin{aligned} P_C &= \sum_{j=1}^{M} \pi_j \int_{\mathcal{X}} \mathrm{d}x P_X(x|j)\, M_j(x) \\ &= \int_{\mathcal{X}} \mathrm{d}x \, \sum_{j=1}^{M} M_j(x)\, \pi_j \, P_X(x|j) \\ &\leq \int_{\mathcal{X}} \mathrm{d}x \, \max_j \{\pi_j \, P_X(x|j)\} \, . \end{aligned}$$

- **Optimum decision rule**: When $x \in \mathcal{X}$ is observed, set

$$J(x) = \arg\max_j \{\pi_j P_X(x|j)\} \, .$$

Bounds on binary error probability

- In most cases, the exact error probability cannot be calculated (absolute value/minimum).
- In many cases, we have N independent copies of the observation:

$$\mathcal{X} \to \mathcal{X}^N,$$
$$x \to \mathbf{x} = (x_1, \ldots, x_N),$$
$$P_X(x|i) \to P_{\mathbf{X}}(x_1, \ldots, x_N|i) = \prod_{n=1}^{N} P_X(x_n|i) \equiv P_{\mathbf{X}}^{(N)}(\mathbf{x}).$$

- The asymptotic behavior of $P_E^{(N)}$ as $N \to \infty$ is of interest.
- M-ary error probabilities can be bounded in terms of pairwise binary error probabilities.

- Consider the case $\pi_0 = \pi_1 = 1/2$ for simplicity.

$$P_E = \frac{1}{2} - \frac{1}{4}\int_{\mathcal{X}} \mathrm{d}x \left|P(x|0) - P(x|1)\right|,$$

$$D\left[P(X|0), P(X|1)\right] = \frac{1}{2}\int_{\mathcal{X}} \mathrm{d}x \left|P(x|0) - P(x|1)\right|.$$

(*Kolmogorov distance*)

$$D^2\left[P(X|0), P(X|1)\right] =$$

$$\frac{1}{4}\left[\int_{\mathcal{X}} \mathrm{d}x \left|\sqrt{P(x|0)} + \sqrt{P(x|1)}\right|\left|\sqrt{P(x|0)} - \sqrt{P(x|1)}\right|\right]^2$$

$$\leq \frac{1}{4}\int_{\mathcal{X}} \mathrm{d}x \left[\sqrt{P(x|0)} + \sqrt{P(x|1)}\right]^2 \times \int_{\mathcal{X}} \mathrm{d}x \left[\sqrt{P(x|0)} - \sqrt{P(x|1)}\right]^2$$

$$= 1 - \left[\int_{\mathcal{X}} \mathrm{d}x \sqrt{P(x|0)\, P(x|1)}\right]^2$$

$$= 1 - B^2\left[P(X|0), P(X|1)\right].$$

Upper bound on binary error probability (Chernoff bound)

- We can upper-bound P_E as follows:

$$P_E = \int_{\mathcal{X}} \mathrm{d}x \min\left\{\pi_0 P(x|0), \pi_1 P(x|1)\right\},$$

$$\min\{a, b\} \leq a^s b^{1-s} \leq sa + (1-s)b, \quad \forall s \in [0,1].$$

$$\begin{aligned} P_E &= \int_{\mathcal{X}} \mathrm{d}x \min\left\{\pi_0 P(x|0), \pi_1 P(x|1)\right\} \\ &\leq \pi_0^s \pi_1^{1-s} \int_{\mathcal{X}} \mathrm{d}x P^s(x|0)\, P^{1-s}(x|1) \\ &\equiv \pi_0^s \pi_1^{1-s} C_s; \\ C &:= \min_{s\in[0,1]} C_s \leq 1 \text{ (Chernoff bound).} \end{aligned}$$

- The Chernoff bound is multiplicative: $P_{\mathbf{X}}^{(N)}(\mathbf{x}) \to C_s^{(N)} = C_s^N$.
- It is exponentially tight in the large N-limit:

$$\alpha C^N \leq P_E^{(N)} \leq \beta C^N \quad \text{for } N \geq N_0.$$

Classical	Quantum
Observation space $\mathcal{X}$	Hilbert space $\mathcal{H}$
Probability distribution $0 \leq P(x\|i);\ 1 \leq i \leq M.$ $\int_{\mathcal{X}} \mathrm{d}x\, P(x\|i) = 1;\ 1 \leq i \leq M.$	Density operator $\rho_i \in \mathcal{S}(\mathcal{H})$ $0 \leq \rho_i;\ 1 \leq i \leq M.$ $\mathrm{Tr}\ \rho_i = 1;\ 1 \leq i \leq M..$
Decision functions $0 \leq M_j(x) \leq 1;\ 1 \leq j \leq M$ $\sum_{j=1}^{M} M_j(x) = 1;\ \ \forall x \in \mathcal{X}.$	POVM (Positive-operator-valued Measure) $0 \leq \hat{M}_j \leq \hat{I};\ 1 \leq j \leq M.$ $\sum_{j=1}^{M} \hat{M}_j = \hat{I}.$

- Quantum Measurements first described by orthogonal projection operators (von Neumann, Lüders):

$$\hat{\Pi}_j = \hat{\Pi}_j^\dagger; \qquad \hat{\Pi}_j = \hat{\Pi}_j^2,$$

$$\hat{\Pi}_j \, \hat{\Pi}_k = \hat{\Pi}_j \, \delta_{jk}; \qquad \sum_{j=1}^{M} \hat{\Pi}_j = \hat{I}.$$

- These Projection-valued measurements (PVMs) insufficient for quantum detection and estimation: Maximum number of outcomes $\leq \dim \mathcal{H}$.
- For an arbitrary state ρ, any set of positive operators $\left\{ \hat{M}_j \right\}_{j=1}^{M}$ satisfying $\sum_{j=1}^{M} \hat{M}_j = \hat{I}$ generates a probability distribution on $j \in \{1, \ldots, M\}$ via $\Pr[j|\rho] = \mathrm{Tr}\ \rho \hat{M}_j$.
- By attaching a measuring system (ancilla) to the system of interest, evolving via a joint unitary, and then measuring the ancilla using a PVM, one can realize *any* given POVM $\left\{ \hat{M}_j \right\}_{j=1}^{M}$ (Naimark's theorem).

- Bayesian problem: $\rho_0(\pi_0)/\rho_1(\pi_1)$.
- Quantum decision rule : 2-element POVM $0 \leq \hat{M}_0, \hat{M}_1; \hat{M}_0 + \hat{M}_1 = \hat{I}$.
- Error probability

$$P_E\left[\hat{M}_0, \hat{M}_1\right] = \pi_0 \,\mathrm{Tr}\, \rho_0 \hat{M}_1 + \pi_1 \,\mathrm{Tr}\, \rho_1 \hat{M}_0$$
$$= \pi_0 + \mathrm{Tr}\, (\pi_1 \rho_1 - \pi_0 \rho_0)\, \hat{M}_0.$$

- Difference operator:

$$\Delta := \pi_1 \rho_1 - \pi_0 \rho_0 = \Delta^\dagger = \sum_x \lambda_x \,|x\rangle\, \langle x|,$$

$$\Delta_+ := \sum_{x:\lambda_x \geq 0} \lambda_x \,|x\rangle\, \langle x|,$$

$$\Delta_- := \sum_{x:\lambda_x < 0} |\lambda_x|\, |x\rangle\, \langle x|,$$

$\Delta = \Delta_+ - \Delta_-$ such that

$0 \leq \Delta_+, \Delta_-;\ \Delta_+\, \Delta_- = 0$. (Jordan decomposition).

$P_\pm :=$ projection operators onto the range of $\Delta_\pm$.

$P_+\, P_- = 0$.

Binary Quantum Hypothesis Testing: Helstrom bound

- Minimum error probability

$$P_E^{\min} := \min_{0 \leq \hat{M}_0 \leq \hat{I}} P_E\left[\hat{M}_0, \hat{M}_1\right]$$
$$= \pi_0 + \mathrm{Tr}\ \Delta_+ \hat{M}_0 - \mathrm{Tr}\ \Delta_- \hat{M}_0.$$

$$\hat{A}, \hat{B} \geq 0 \Longrightarrow \mathrm{Tr}\ \hat{A}\hat{B} \geq 0.$$
$$\hat{X} \geq \hat{Y}, \hat{A} \geq 0 \Longrightarrow \mathrm{Tr}\ \hat{A}\hat{X} \geq \mathrm{Tr}\ \hat{A}\hat{Y}.$$

- Since

$$\mathrm{Tr}\ \Delta_- \hat{M}_0 = \mathrm{Tr}\ P_- \Delta_- P_- \hat{M}_0$$
$$= \mathrm{Tr}\ \Delta_- P_- \hat{M}_0 P_- \leq \mathrm{Tr}\ \Delta_- P_- \hat{I} P_-$$
$$= \mathrm{Tr}\ \Delta_- P_-.$$

- Let's take $\hat{M}_0 = P_-$:

$$\mathrm{Tr}\ \Delta_+ \hat{M}_0 = 0$$
$$\Longrightarrow \left\{\hat{M}_0 = P_-, \hat{M}_1 = P_+\right\} \text{ is the optimum POVM.}$$

- Minimum error probability

$$P_E^{\min} := \pi_0 - \mathrm{Tr}\ \Delta_-$$
$$\mathrm{Tr}\ \Delta = \mathrm{Tr}\ \Delta_+ - \mathrm{Tr}\ \Delta_- = \pi_1 - \pi_0$$
$$\pi_0 - \mathrm{Tr}\ \Delta_- = \pi_1 - \mathrm{Tr}\ \Delta_+.$$

- So

$$\begin{aligned} P_E^{\min} &:= \frac{\pi_0 - \mathrm{Tr}\ \Delta_- + \pi_1 - \mathrm{Tr}\ \Delta_+}{2} \\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{Tr}\ (\Delta_+ + \Delta_-) \\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{Tr}\ |\Delta|\,, \end{aligned}$$

- where

$$\mathrm{Tr}\ |X| = \mathrm{Tr}\ \sqrt{X^\dagger X} \equiv \|X\|_1 \text{ (trace norm).}$$
$$P_E^{\min} = \frac{1}{2} - \frac{1}{2}\|\pi_1\rho_1 - \pi_0\rho_0\|_1 \text{ } (\textbf{Helstrom limit}).$$

- **Classical Detection & Estimation Theory**
  1. *Fundamentals of Statistical Signal Processing: Vol. I – Estimation Theory*, Steven M. Kay, Prentice-Hall, 1993.
  2. *Fundamentals of Statistical Signal Processing: Vol. II – Detection Theory*, Steven M. Kay, Prentice-Hall, 1998.
  3. *Detection Estimation and Modulation Theory, Part I: Detection, Estimation, and Filtering Theory*, Harry L. Van Trees and Kristine Bell, Wiley, 2013.

- **Quantum Detection & Estimation Theory**
  1. *Quantum Detection and Estimation Theory*, Carl W. Helstrom, Academic Press, 2012.
  2. *Quantum Measurement and Control*, Howard M. Wiseman and Gerard J. Milburn, Cambridge University Press, Cambridge Univ. Press, 2014.

- **Quantum Theory background**
  1. *Quantum Information and Computation*, Michael A. Nielsen and Isaac. L. Chuang, Cambridge Univ. Press, 2000.
  2. *Quantum Systems, Channels, Information: A Mathematical Introduction*, Alexander S. Holevo, De Gruyter, 2012.