ECE 250: Stochastic Processes: Week #8

Instructor: Behrouz Touri

Outline:

- Some Estimation Theory, MSE, and MMSE
- Geometry of Random Variables, Conditional Expectation, and Mean Square Estimation
- Orthogonality Principle and Applications to LMSE and MMSE

Estimation Theory

- Main Question: Given an observation Y of a random variable X, how to estimate X?
- In other words, what is the best function g such that $\hat{X} = g(Y)$ is the best estimator?
- More generally: given a sequence of observation of $X_1, \ldots, X_k$, how to estimate X?
- Example: Radar detection: Suppose that X is the radial distance of an aircraft from a radar station and $Y = X + Z$ is the radar's observed location where Z is independent of X and $Z \sim \mathcal{N}(0, \sigma^2)$. What is the best estimator $\hat{X} = g(Y)$ of the location X?
- What if $X_k = X + Z_k$ where Z_k are i.i.d. with $Z_k \sim \mathcal{N}(0, \sigma^2)$? Then what function $g(\cdot)$ of the observations is the best estimator?
- The best is always subjective until we set a criteria. One popular criteria is Mean Square Error (MSE).
- For measurements $X_1, \ldots, X_k$ of a random variable X, we define the MSE of (a measurable) an estimator (function) $g : \mathbb{R}^k \to \mathbb{R}$ to be $\mathbb{E}[|g(X_1, \ldots, X_k) - X|^2]$. In this setting, we view $\mathbb{E}[|U - X|^2]$ as the squared *distance* of random variables U and X.
- Once we fix the MSE criteria for the best estimator, then the problem of finding the best MSE estimator for X based on the measurements $X_1, \ldots, X_k$ is:

$$\arg \min_{g:\mathbb{R}^k \to \mathbb{R}} \mathbb{E}[|g(X_1, \ldots, X_k) - X|^2].$$

- Any g that minimizes the above criteria is called a Minimum Mean Square Error (MMSE) estimator.
- When solving for MMSE, we always assume that X has finite mean and variance.

MMSE

- In practice finding the MMSE *might be* hard.
- We can restrict our attention to special classes of functions g.
- Let $k = 0$, and suppose that we want to find the best *constant* c that estimates X. Note that in this case, we view c as a constant random variable.

$$\text{objective: finding } c \in \operatorname{argmin}_c \mathbb{E}[|X - c|^2]. \tag{1}$$

- In this case,

$$\begin{aligned}
\mathbb{E}[|X - c|^2] &= \mathbb{E}[|X - \bar{X} + \bar{X} - c|^2] \\
&= \mathbb{E}[|X - \bar{X}|^2 + 2(\bar{X} - c)\mathbb{E}[(X - \bar{X})] + (\bar{X} - c)^2 \\
&= \mathbb{E}[(X - \bar{X})^2] + \mathbb{E}[(\bar{X} - c)^2].
\end{aligned}$$

- **Estimation theory interpretation of mean and variance**: The best constant MMSE estimator of X is $\mathbb{E}[X]$ and the corresponding MMSE value is Var(X).

Probability, Linear Algebra, Orthogonality Principle: Geometric View

- In the case of constant estimator, we note that:
  - The space V of constant (r.v.s) is a linear subspace.
  - We have: $\mathbb{E}[(X-\bar{X})c] = \mathbb{E}[c(X-\bar{X})] = 0$.
  - This can be viewed as the orthogonality principle.
- Generally: Let $(\Omega, \mathcal{F}_o, \Pr(\cdot))$ be the underlying probability space.
- The set V of random variables on this space is a vector space (over reals) with $+$ being the regular summation of two functions (r.v.s) and the scalar product as a regular product of a number and a function:
  - Summation of random variables satisfies Abelian group property:
    (i) *Existence of identity:* $\mathbf{1}_{\emptyset} = 0_\Omega$ is a random variable with $X + = 0_\Omega = X$.
    (ii) *Commutative*: $X + Y = Y + X$ for all $X, Y \in V$
    (iii) *Existence of inverse:* $X + (-X) = 0_\Omega$ for all X.
    (iv) *Associativity*: $X + (Y + Z) = (X + Y) + Z$.
  - Scalar product satisfies:
    (i) $1X = X$.
    (ii) $(ab)X = a(bX)$ for all $a, b \in \mathbb{R}$ and $X \in V$.
  - Summation $+$ and scalar product are connected through distributive properties:
    (i) $(\alpha + \beta)X = \alpha X + \beta X$ for all $\alpha, \beta \in \mathbb{R}$ and $X \in V$.
    (ii) $\alpha(X + Y) = \alpha X + \alpha Y$ for all $\alpha \in \mathbb{R}$ and $X, Y \in V$.

Geometry of Random Variables: $L_2(\Omega, \mathcal{F}_o, \Pr(\cdot))$ is an inner-product space

- Define $L_2(\Omega, \mathcal{F}_o, \Pr(\cdot))$ (or simply L_2) to be the set of random variables with finite variance (or second moment), i.e., $L_2 = \{X \mid \mathbb{E}[X^2] < \infty\}$.
- Properties of L_2:
  - L_2 is a linear subspace of random variables:
    (i) $aX \in L_2$ for all $X \in L_2$ and $a \in \mathbb{R}$ as $\mathbb{E}[(aX)^2] = a^2\mathbb{E}[X^2] < \infty$, and
    (ii) $X + Y \in L_2$ for all $X, Y \in L_2$ as:
$$\mathbb{E}[(X+Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] + \mathbb{E}[2XY] \leq 2(\mathbb{E}[X^2] + \mathbb{E}[Y^2]) < \infty.$$
  - **The most important property**: L_2 is an inner-product space[1]. For any two random variables $X, Y \in L_2$, let us define their inner product
$$X \cdot Y := \mathbb{E}[XY].$$
  - Then this operation satisfies the axioms of an inner product:
    (i) $X \cdot X = \mathbb{E}[X^2] \geq 0$ with the equality iff $X = 0$ almost surely.
    (ii) *linearity*: $(\alpha X + Y) \cdot Z = X \cdot Z + \alpha Y \cdot Z$.
- Therefore, L_2 is a normed vector space, with the norm $\| \cdot \|$ defined by
$$\|X\| := \sqrt{X \cdot X} = \sqrt{\mathbb{E}[X^2]}.$$

[1] More importantly it is a Hilbert space.

$L_2(\Omega, \mathcal{F}_o, \Pr(\cdot))$ is an inner-product space: implications

- L_2 is a normed space with the norm $\|X - Y\|^2 := (X - Y) \cdot (X - Y) = E[(X - Y)^2]$.
- Cauchy-Schwartz inequality: for any $X, Y \in L_2$

$$\mathbb{E}^2[XY] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

  Important: The equality holds if and only if $X = cY$ for some constant c.
  Implication: For $X, Y \in L^2$, we define $Cov(X, Y) := \mathbb{E}[(X - \bar{X})(Y - \bar{Y})]$. By Cauchy-Schwartz inequality

$$Cov^2(X, Y) = \mathbb{E}^2[(X - \bar{X})(Y - \bar{Y})] \leq \mathbb{E}[(X - \bar{X})^2]\mathbb{E}[(Y - \bar{Y})^2].$$

  In other words, if we define the correlation coefficient $\rho_{X,Y}$ by

$$\rho_{X,Y} := \frac{Cov(X, Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}},$$

  then $\rho_{X,Y} \in [-1, 1]$.
- In fact, $\rho_{X,Y} = 1$ iff $(Y - \bar{Y}) = c(X - \bar{X})$ for a $c > 0$ and $\rho_{X,Y} = -1$ iff $(Y - \bar{Y}) = c(X - \bar{X})$ for a $c < 0$.

L_2-norm and L_2 convergence

- Since L_2 is a normed space, we can define a new limit of random variables:

  Definition 1. *We say that a sequence $\{X_k\}$ converges in L_2 (or in MSE sense) to X if $\lim_{k\to\infty} \|X - X_k\| = 0$.*

- Note that $\lim_{k\to\infty} \|X - X_k\| = 0$ iff $\lim_{k\to\infty} \mathbb{E}[|X - X_k|^2] = 0$.

- Definition: We say that $H \subseteq L_2$ is a linear subspace if

  (i) for any $X, Y \in H$, we have $X + Y \in H$, and

  (ii) for any $X \in H$ and $a \in \mathbb{R}$, $aX \in H$.

- Definition: We say that $H \subseteq L_2$ is closed if for any process $\{X_k\}$ with

$$\lim_{m,n\to\infty} \|X_m - X_n\|^2 = \lim_{m,n\to\infty} \mathbb{E}[|X_m - X_n|^2] = 0,$$

  we have $\lim_{k\to\infty} X_k \xrightarrow{L_2} X$ for some random variable $X \in L_2$.

- Showing linear subspace is easy, but closedness might be hard.

Orthogonality Principle

- Important Cases:
  1. For random variables $X_1, \ldots, X_k \in L_2$, the set $H = \{\alpha_1 X_1 + \ldots + \alpha_k X_k \mid \alpha_i \in \mathbb{R}\}$ is a closed linear subspace.
  2. For any random variables $X_1, \ldots, X_k \in L_2$, the set $H = \{\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k \mid \alpha_i \in \mathbb{R}\}$ is a closed linear subspace.
  3. For any sub σ-algebra $\mathcal{F}$ of $\mathcal{F}_o$, the set $L_2(\mathcal{F})$ of measurable random variables with respect to $\mathcal{F}$ that are in L_2, is a linear closed subspace of L_2.

Theorem 1. *Let H be a closed linear subspace of L_2 and let $X \in L_2$. Then,*

a. There exists a unique (up to almost sure equivalence) random variable $Y \in H$ such that

$$\|Y - X\|^2 \leq \|Z - X\|^2$$

for all $Z \in H$.

b. Furthermore, $Y \in H$ is the unique random variable with

$$(Y - X) \perp Z$$

for all $Z \in H$.

Orthogonality Principle: Implications 1: LMMSE

- Let $X_1, \ldots, X_k$ be measurements of X.
- LMMSE estimator of X is defined to be the linear estimator $\hat{X} = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k$ that minimizes $\mathbb{E}[|X - \hat{X}|^2] = \|X - \hat{X}\|^2$.
- We derive the LMMSE for the case of $k = 1$. The case $k \geq 2$ can be obtained similarly.
- Let Y be a measurement of X and we want to find optimal $a^*, b^* \in \mathbb{R}$ such that

$$\|X - (a^*Y + b^*)\| \leq \|X - (aY + b)\|,$$

  for any $a, b \in \mathbb{R}$.
- Using orthogonality principle, for all b, we need to have

$$(X - (a^*Y + b^*)) \perp b \Leftrightarrow \mathbb{E}[(X - (a^*Y + b^*))b] = 0.$$

- This holds iff

$$b^* = \bar{X} - a^*\bar{Y}.$$

- Replacing this, we are seeking optimal value for a^* such that

$$[(X - \bar{X}) - a^*(Y - \bar{Y})] \perp aY - b,$$

  for all $a, b \in \mathbb{R}$. Let $a = 1$ and $b = \bar{Y}$. Then, this holds if

$$\mathbb{E}[((X - \bar{X}) - a^*(Y - \bar{Y}))(Y - \bar{Y})] = 0.$$

  This holds iff $a^* = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(Y)}$.
- Therefore, by choosing $a^* = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(Y)}$ and $b^* = \bar{X} - a^*\bar{Y}$, we have $(X - \hat{X}) \perp 1_\Omega$ and $(X - \hat{X}) \perp Y - \bar{Y}$ and hence, $(X - \hat{X}) \perp \alpha Y + \beta$ for all $\alpha, \beta \in \mathbb{R}$ (why?).

Orthogonality Principle: Implications 2: MMSE

Theorem 2. *(MMSE Estimator is the Conditional Expectation) Consider $L_2(\mathcal{F})$ for a sub σ-algebra and $X \in L_2$ and let $\hat{X} = \mathbb{E}[X \mid \mathcal{F}]$. Then,*

$$\mathbb{E}[|X - \hat{X}|^2] \leq \mathbb{E}[|X - Y|^2], \tag{2}$$

for all Y that is measurable with respect to $\mathcal{F}$.

Proof: We show that for any such $Y \in L_2(\mathcal{F})$, we have $Y \perp (X - \hat{X})$. This follows from

$$\begin{aligned}
\mathbb{E}[Y(X - \hat{X})] &= \mathbb{E}[Y(X - \mathbb{E}[X \mid \mathcal{F}])] \\
\text{(by linearity)} &= \mathbb{E}[YX] - \mathbb{E}[Y\mathbb{E}[X \mid \mathcal{F}]] \\
\text{(by product rule)} &= \mathbb{E}[YX] - \mathbb{E}[\mathbb{E}[YX \mid \mathcal{F}]] \\
\text{(by tower rule)} &= \mathbb{E}[YX] - \mathbb{E}[YX] = 0.
\end{aligned}$$

Therefore, by the orthogonality principle, the statement holds.

- Implication: For any measurements $X_1, \ldots, X_k$ of a random variable X, the MMSE estimator for X is $\hat{X} = \mathbb{E}[X \mid X_1, \ldots, X_k]$.

Example: A Gaussian Channel-LMSE Estimator

- Let $Z = X + Y$ be a measurement of X where $X, Y \sim \mathcal{N}(0,1)$ are independent random variables.
- The LMSE estimator for X given Z is

$$\hat{X} = \frac{\mathsf{Cov}(Z,X)}{\mathsf{Var}(Z)} Z = \frac{1}{2} Z.$$

Example: A Gaussian Channel-MMSE Estimator

- What is MMSE of X given Z?
- We need to find out $\tilde{X} = \mathbb{E}[X \mid Z]$. To do so, we proceed the undergrad way:
  - Note that we have:

$$\begin{aligned} F_{Z|X}(z \mid x) &= \Pr(Z \leq z \mid X = x) \\ &= \Pr(X + Y \leq z \mid X = x) \\ &= \Pr(Y \leq z - x \mid X = x) \\ &= F_Y(z - x). \end{aligned}$$

  - Therefore,

$$\begin{aligned} f_{Z|X}(z \mid x) &= \frac{d}{dz} F_Y(z - x) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}}. \end{aligned}$$

  - Therefore,

$$\begin{aligned} f_{X|Z}(x, z) &= \frac{f_{Z|X}(z, x) f_X(x)}{f_Z(z)} \\ &= \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{4\pi}} e^{-\frac{z^2}{4}}} \\ &= \frac{\sqrt{2}}{\sqrt{2\pi}} e^{-(\frac{z^2}{4} - zx + x^2)} \\ &= \frac{\sqrt{2}}{\sqrt{2\pi}} e^{-(x - \frac{z}{2})^2}. \end{aligned}$$

  - Therefore, $X \mid Z = z$ is a Gaussian random variable with mean $\frac{z}{2}$ and variance $\frac{1}{2}$.
  - Finally, $\tilde{X} = \mathbb{E}[X \mid Z = z] = \frac{z}{2}$. Therefore, in this case LMMSE=MMSE.