Systems and Signals 344

2025

Lecture 27

Estimation of a Random Variable

Selections from Chapter 12 (12.1-12.3)

Estimation of a Random Variable

- The techniques from the last several lectures use the *outcomes* of experiments to make **inferences** about *probability models*.
- In this chapter we use *observations* to calculate an approximate value (**estimate**) of a *future sample value of a random variable* that has not been observed.
  - We refer to the estimation as a **prediction**.
  - A predictor uses random variables observed in early subexperiments to estimate a random variable produced by a later subexperiment.
- The random variable of interest may be unavailable because it is impractical to measure (for example, the temperature of the sun), or because it is obscured by distortion (a signal corrupted by noise), or because it is not available soon enough.

Estimation of a Random Variable

If X is the random variable to be estimated, we adopt the notation $\hat{X}$ (also a random variable) for the estimate.

We will mostly use the **mean square error** as a measure of the *quality* or *accuracy* of the estimate.

$$e = \mathrm{E}[(X - \hat{X})^2]$$

Estimation of a Random Variable

There are various types of estimates:

- Blind estimation of a random variable
- Estimation of a random variable given an event
- Estimation of a random variable given one other random variable
- Linear estimation of a random variable given a random vector
- Linear estimation of a random vector given another random vector (we won't cover this case in this course)

Minimum Mean Square Error Estimation

- Let's say an experiment produces a random variable X. However, we are unable to observe X directly.
- Instead, we *observe* something (an event or a random variable) that provides *partial information* about the sample value of X.
- If X is a *discrete* random variable, it is possible to use **hypothesis testing** to estimate X.
    - For each $x_i \in S_X$, we could define hypothesis H_i as the probability model $P_X(x_i) = 1, P_X(x) = 0, x \neq x_i$.
    - A hypothesis test would then lead us to choose the most probable x_i given our observations.
    - This however treats all errors in the same manner, regardless of whether they produce answer that are close to or far from the correct value of X.

Minimum Mean Square Error Estimation

- By contrast, the aim of the *estimation procedures* is to find an estimate $\hat{X}$ that, on average, is close to the true value of X, even if the estimate never produces a correct answer.
  - A popular example is an estimate of the number of children in a family. The best estimate, based on available information, might be 2.4 children.
- In an estimation procedure, we aim for a low probability that the estimate is far from the true value of X.
- An *accuracy* measure that helps us achieve this aim is the **mean square error**, which we try to *minimize*.

$$e = \mathrm{E}[(X - \hat{X})^2]$$

The estimator is therefore called the
minimum mean square error (MMSE) estimator.

Blind Estimation of X

- An experiment produces a random variable X.
- Before the experiment is performed, what is the best estimate of X?

Blind Estimation of X

- An experiment produces a random variable X.
- Before the experiment is performed, what is the best estimate of X?
- This is the **blind estimation problem** because it requires us to make an inference about X in the absence of any observations.
- Although it is unlikely that we will guess the correct value of X, we can derive a number that comes as close as possible in the sense that it minimizes the mean square error.
- For this we typically choose:

$$\hat{x}_B = \mathrm{E}[X]$$

Example 12.1

Before a six-sided die is rolled, what is the minimum mean square error estimate of the number of spots X that will appear?

The probability model is $P_X(x) = 1/6$, $x = 1, 2, \ldots, 6$, otherwise $P_X(x) = 0$. For this model, $\mathrm{E}[X] = 3.5$. Even though $\hat{x}_B = 3.5$ is not in the range of X, it is the estimate that minimizes the mean square estimation error.

MMSE Estimation of X Given an Event

- Suppose that in performing an experiment, instead of observing X directly, we learn only that $X \in A$ (which is known).
- Given this information, what is the minimum mean square error estimate of X?

MMSE Estimation of X Given an Event

- Suppose that in performing an experiment, instead of observing X directly, we learn only that $X \in A$.
- Given this information, what is the minimum mean square error estimate of X?
- Given A, X has a conditional PDF $f_{X|A}(x)$ or a conditional PMF $P_{X|A}(x)$.
- Our task is to minimize the conditional mean square error $e_{X|A} = \mathbf{E}[(X - \hat{x})^2|A]$.
- This is essentially the same as the blind estimation problem but now just using the conditional PDF or PMF.
- Therefore:

$$\hat{x}_A = \mathbf{E}[X|A]$$

Example 12.2

The duration T minutes of a phone call is an exponential random variable with expected value $\mathrm{E}[T] = 3$ minutes. If we observe that a call has already lasted 2 minutes, what is the minimum mean square error estimate of the call duration?

MMSE estimate of X Given Y

- If an experiment produces two random variables, X and Y, where we can observe Y but we really want to know X.
  - Therefore, the estimation task is to assign to every $y \in S_Y$ a number, $\hat{x}$, that is near X. Again we use the MSE.

$$e_M = \mathrm{E}[(X - \hat{x}_M(y))^2 | Y = y]$$

- Because each $y \in S_Y$ produces a specific $\hat{x}_M(y)$, $\hat{x}_M(y)$ is a *sample value* of a *new random variable* $X_M(Y)$ (since it is a function of RV Y). This is in contrast to the blind estimates of before (where $\hat{x}_B$ and $\hat{x}_A$ were in reality *parameters* of X).

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

Example 12.3

Suppose X and Y are independent random variables with PDFs $f_X(x)$ and $f_Y(y)$. What is the minimum mean square error estimate of X given Y?

In this case, $f_{X|Y}(x|y) = f_X(x)$ and the minimum mean square error estimate is

$$\hat{x}_M(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx = \int_{-\infty}^{\infty} x f_X(x)\, dx = \mathrm{E}[X] = \hat{x}_B. \tag{12.6}$$

That is, when X and Y are independent, the observation Y provides no information about X, and the best estimate of X is the blind estimate.

Linear Estimation of X given Y

- We again use an observation, y, of random variable Y to produce an estimate, $\hat{x}$, of random variable X.
- Again, our *accuracy measure* is the **mean square error**.
- Now the estimate is a single *function* that applies for all Y:

$$\hat{X}_L(y) = g(Y) = aY + b$$

where a and b are the constants we try to *estimate* for all $y \in S_Y$.

- Because $\hat{X}_L(y)$ is a linear function of Y, the procedure is referred to as **linear estimation** (hence subscript L).
- Linear estimation appears in many electrical engineering applications of statistical inference. E.g. Analog Filters (using resistors, capacitors, and inductors).

Linear Estimation of X given Y

$$e_L = \mathrm{E}[(X - \hat{X}_L(Y))^2]$$

In this formula, we use $\hat{X}_L(Y)$ and not $\hat{x}_L(y)$ because the expected value in the formula is an *unconditional* expected value in contrast to the conditional expected value from before (which is a fixed value for a given y).

Linear Estimation of X given Y

Random variables X and Y have expected values μ_X and μ_Y, standard deviations σ_X and σ_Y, and correlation coefficient $\rho_{X,Y}$. The optimum **linear mean square error (LMSE) estimator** of X given Y is

$$\hat{X}_L(Y) = \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) + \mu_X$$

Therefore $a = \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}$ and $b = \mu_X - a\mu_Y$

This linear estimator has the following properties:

1. The *minimum mean square estimation error* for a linear estimate is $e_L^* = \mathrm{E}[(X - \hat{X}_L(Y))^2] = \sigma_X^2(1 - \rho_{X,Y}^2)$
2. The estimation error $\tilde{X} = X - \hat{X}_L(Y)$ is *uncorrelated* with Y

.

Linear Estimation of X given Y

The estimation error $\tilde{X} = X - \hat{X}_L(Y)$ is *uncorrelated* with Y. This is one aspect of the **orthogonality principle** of the LMSE

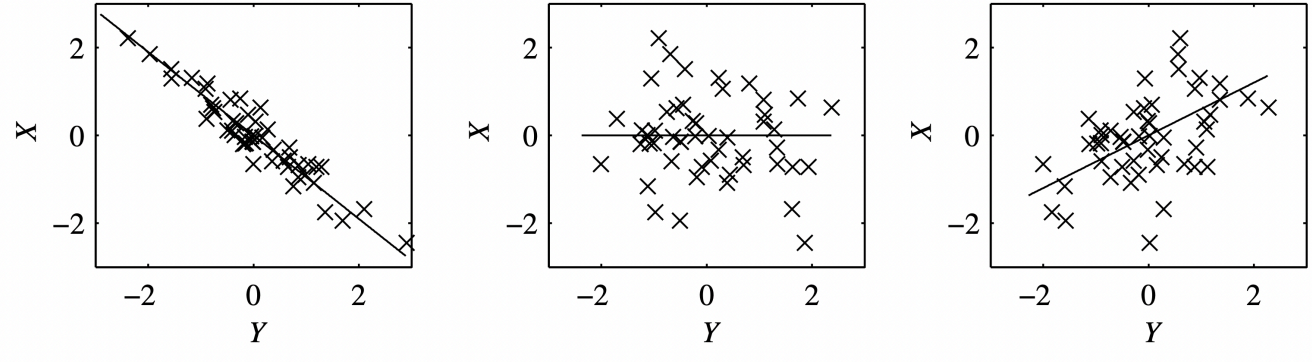
- This means $\mathbf{E}[\tilde{X}] = 0$ and $\mathbf{Cov}[\tilde{X}, Y] = r_{\tilde{X},Y} = 0$
- It states that the estimation error is orthogonal to the data used in the estimate.
- A geometric explanation of linear estimation is that the optimum estimate of X is the *projection* of X into the plane of linear functions of Y.
- Intuitively, this means that once you have the optimal estimate $\hat{X}_L(Y)$, any further "information" about Y cannot help reduce the error any more, since the remaining error $\tilde{X}$ is uncorrelated with Y.

Linear Estimation of X given Y

- The correlation coefficient $\rho_{X,Y}$ plays a key role in the optimum linear estimator.
- Recall that $|\rho_{X,Y}| \leq 1$ and that $\rho_{X,Y} = \pm 1$ corresponds to a deterministic linear relationship between X and Y.
- In this case now, when $\rho_{X,Y} = \pm 1$ and then $e_L^* = 0$
  - the estimate is the **true value**, since you have all the information.
- At the other extreme, when X and Y are *uncorrelated*, $\rho_{X,Y} = 0$ then $\hat{X}_L(Y) = \mathrm{E}[X]$, which is the **blind estimate**.
  - With X and Y uncorrelated, there is no linear function of Y that provides useful information about the value of X.

Linear Estimation of X given Y

- Therefore, the magnitude of the correlation coefficient indicates the extent to which observing Y improves our knowledge of X.
- The optimum linear estimator of X given Y is the line $\hat{X}_L(Y) = \rho_{X,Y} Y$
- For each pair (x, y), the estimation error equals the *vertical distance* to the estimator line.



(a) $\rho_{X,Y} = -0.95$ **(b)** $\rho_{X,Y} = 0$ **(c)** $\rho_{X,Y} = 0.60$

Figure 12.1 Each graph contains 50 sample values of the random variable pair (X, Y), each marked by the symbol $\times$. In each graph, $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, $\mathrm{Var}[X] = \mathrm{Var}[Y] = 1$. The solid line is the optimal linear estimator $\hat{X}_L(Y) = \rho_{X,Y} Y$.

MAP and ML Estimation

- Now we will consider the **maximum a posteriori probability (MAP) estimator** and the **maximum likelihood (ML) estimator**.
- Although neither of these estimates produces the MMSE, they are convenient to obtain in some applications, and they often produce estimates with errors that are not much higher than the minimum mean square error.
- MAP and ML estimators have different objectives to MMSE.
  - MAP seeks to find the most probable value of the parameter given the observed data and a prior probability distribution.
  - ML, on the other hand, aims to find the parameter value that maximizes the likelihood of the observed data.
- As you might expect, MAP and ML estimation are closely related to MAP and ML hypothesis testing.

MAP Estimate

The **maximum a posteriori probability (MAP) estimate** of X given an observation of Y is

Discrete

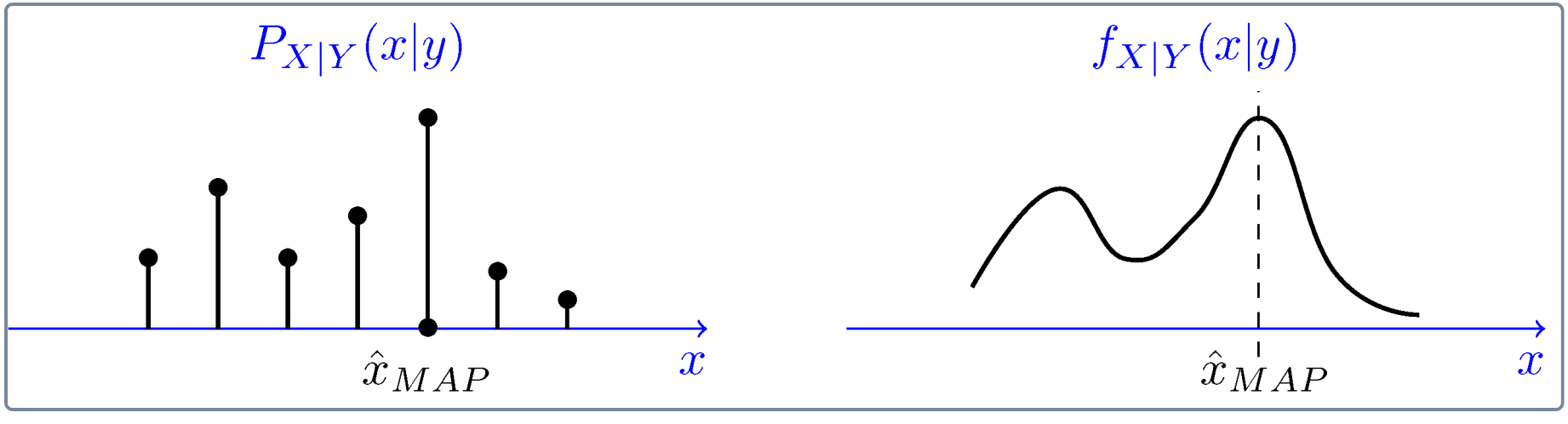
$$\hat{x}_{\text{MAP}}(y_j) = \arg\max_{x \in S_X} P_{X|Y}(x|y_j)$$

Continuous

$$\hat{x}_{\text{MAP}}(y) = \arg\max_{x} f_{X|Y}(x|y)$$

The notation $\arg\max_x g(x)$ denotes a value of x that maximizes $g(x)$, where $g(x)$ is any function of a variable x.

MAP Estimate



Recap: Conditioning by a Random Variable

For *continuous* random variables X and Y with joint PDF $f_{X|Y}(x|y)$, and x and y such that $f_X(x) > 0$ and $f_Y(y) > 0$, respectively,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

For *continuous* random variables X and Y with joint PDF $f_{X,Y}(x,y)$, and x and y such that $f_X(x) > 0$ and $f_Y(y) > 0$,

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$$

MAP Estimate

- Because the denominator $f_Y(y)$ does not depend on x, maximizing $f_{X|Y}(x|y)$ over all x is equivalent to maximizing the numerator $f_{Y|X}(y|x)f_X(x)$.

Discrete

$$\hat{x}_{\text{MAP}}(y_j) = \arg\max_{x \in S_X} P_{Y|X}(y_j|x)P_X(x)$$

Continuous

$$\hat{x}_{\text{MAP}}(y) = \arg\max_{x} f_{Y|X}(y|x)f_X(x)$$

Similar to the hypothesis tests, $P_X(x)$ and $f_X(x)$ indicates the *priori* probability model of X.

ML Estimate

Then in the absence of this *priori* information, we can instead calculate a **maximum likelihood (ML) estimate** of X given the observation $Y = y$:

Discrete

$$\hat{x}_{\mathrm{ML}}(y_j) = \arg\max_{x \in S_X} P_{Y|X}(y_j|x)$$

Continuous

$$\hat{x}_{\mathrm{ML}}(y) = \arg\max_{x} f_{Y|X}(y|x)$$

Example 12.7

Consider a collection of old coins. Each coin has random probability, q of landing with heads up when it is flipped. The probability of heads, q, is a sample value of a beta $(2, 2)$ random variable, Q, with PDF

$$f_Q(q) = \begin{cases} 6q(1-q) & 0 \le q \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{12.27}$$

To estimate q for a coin we flip the coin n times and count the number of heads, k. Because each flip is a Bernoulli trial with probability of success q, k is a sample value of the binomial(n,q) random variable K. Given $K = k$, derive the following estimates of Q:

(a) The blind estimate $\hat{q}_B$,

(b) The maximum likelihood estimate $\hat{q}_{\mathrm{ML}}(k)$,

(c) The maximum a posteriori probability estimate $\hat{q}_{\mathrm{MAP}}(k)$,

(d) The minimum mean square error estimate $\hat{q}_M(k)$,

(e) The optimum linear estimate $\hat{q}_L(k)$.

Plot of PDF