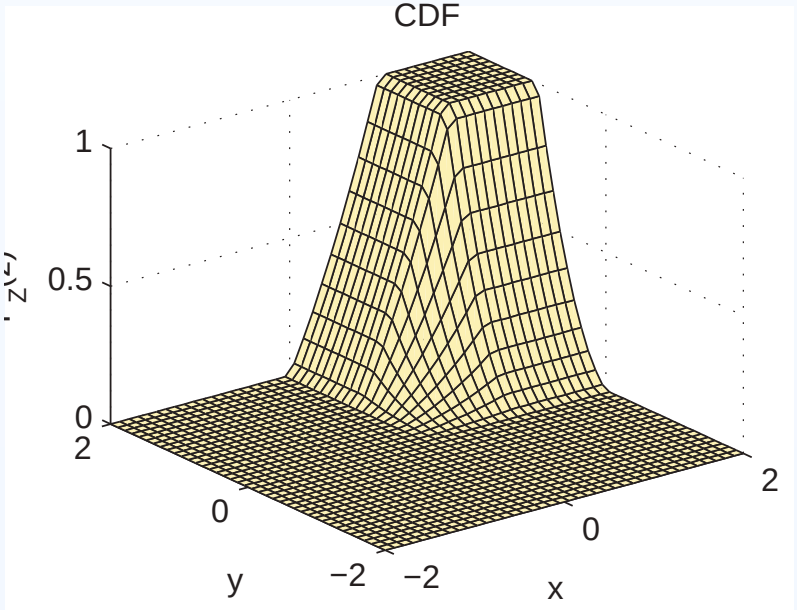
Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. The cdf is plotted here:



A plot of the cumulative distribution function.

Marginal Density Function

The joint pdf characterises the random vector; the so-called **marginal pdf** describes a subset of RVs from the random vector.

Marginal Density Function

The joint pdf characterises the random vector; the so-called **marginal pdf** describes a subset of RVs from the random vector.

Let $\mathbf{k}$ be an M-dimensional vector containing unique indices to elements in the N-dimensional random vector $\mathbf{X}(\zeta)$,

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_M \end{bmatrix}$$

Marginal Density Function

The joint pdf characterises the random vector; the so-called **marginal pdf** describes a subset of RVs from the random vector.

Let $\mathbf{k}$ be an M-dimensional vector containing unique indices to elements in the N-dimensional random vector $\mathbf{X}(\zeta)$,

Now define a M-dimensional random vector, $\mathbf{X_k}(\zeta)$, that contains the M random variables which are components of $\mathbf{X}(\zeta)$ and indexed by the elements of $\mathbf{k}$. In other-words, if

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_M \end{bmatrix} \quad \text{then} \quad \mathbf{X_k}(\zeta) = \begin{bmatrix} X_{k_1}(\zeta) \\ X_{k_2}(\zeta) \\ \vdots \\ X_{k_M}(\zeta) \end{bmatrix}$$

Marginal Density Function

The **marginal pdf** is then given by:

$$f_{\mathbf{X_k}}\left(\mathbf{x_k}\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x_{-k}}$$

Marginal Density Function

The **marginal pdf** is then given by:

$$f_{\mathbf{X_k}}\left(\mathbf{x_k}\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x_{-k}}$$

A special case is the **marginal pdf** describing the individual RV X_j:

$$f_{X_j}\left(x_j\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-1 \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_N$$

Marginal Density Function

The **marginal pdf** is then given by:

$$f_{\mathbf{X_k}}\left(\mathbf{x_k}\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x_{-k}}$$

A special case is the **marginal pdf** describing the individual RV X_j:

$$f_{X_j}\left(x_j\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-1 \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_N$$

Marginal pdfs will become particular useful when dealing with Bayesian parameter estimation later in the course.

Marginal Density Function

Example (Marginalisation). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

⋈

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

Marginal Density Function

Example (Marginalisation). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. By definition:

$$f_X(x) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z})\, dy$$

$$f_Y(y) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z})\, dx$$

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. Taking $f_X(x)$, then:

$$f_X(x) = \begin{cases} \frac{1}{2}\int_0^1 (x+3y)\,dy & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

□

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. Taking $f_X(x)$, then:

$$f_X(x) = \begin{cases} \frac{1}{2}\int_0^1 (x+3y)\,dy & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

which after a simple integration gives:

$$f_X(x) = \begin{cases} \frac{1}{2}\left(x+\frac{3}{2}\right) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. The cdf, $F_X(x)$, is thus given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2}\int_0^x \left(u+\frac{3}{2}\right) du & 0 \leq x \leq 1 \\ \frac{1}{2}\int_0^1 \left(u+\frac{3}{2}\right) du & x > 1 \end{cases}$$

□

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. The cdf, $F_X(x)$, is thus given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2}\int_0^x \left(u + \frac{3}{2}\right) du & 0 \leq x \leq 1 \\ \frac{1}{2}\int_0^1 \left(u + \frac{3}{2}\right) du & x > 1 \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4}(x+3) & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. Similarly, it can be shown that:

$$f_Y(y) = \begin{cases} \frac{1}{2}\left(\frac{1}{2}+3y\right) & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

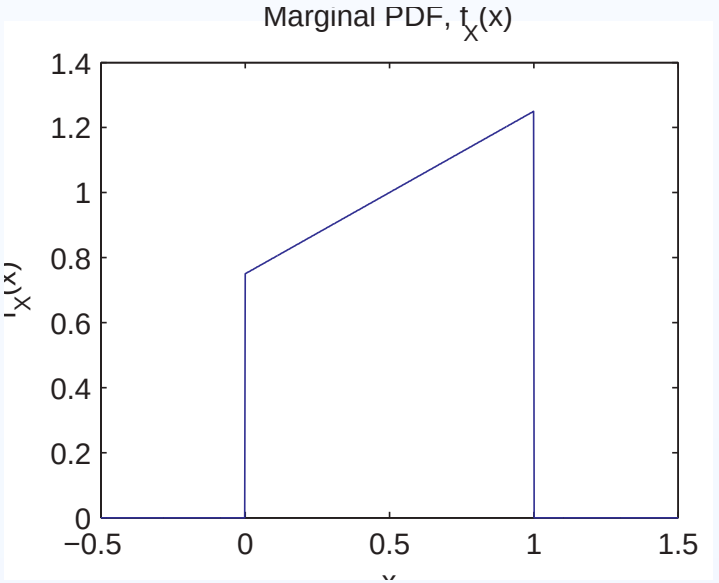
$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ \frac{y}{4}(1+3y) & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

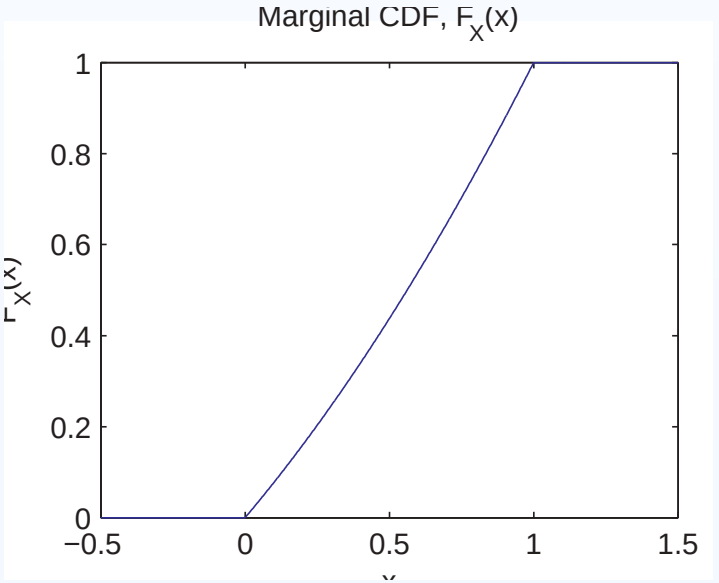
Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

SOLUTION. The marginal-pdfs and cdfs are shown below.





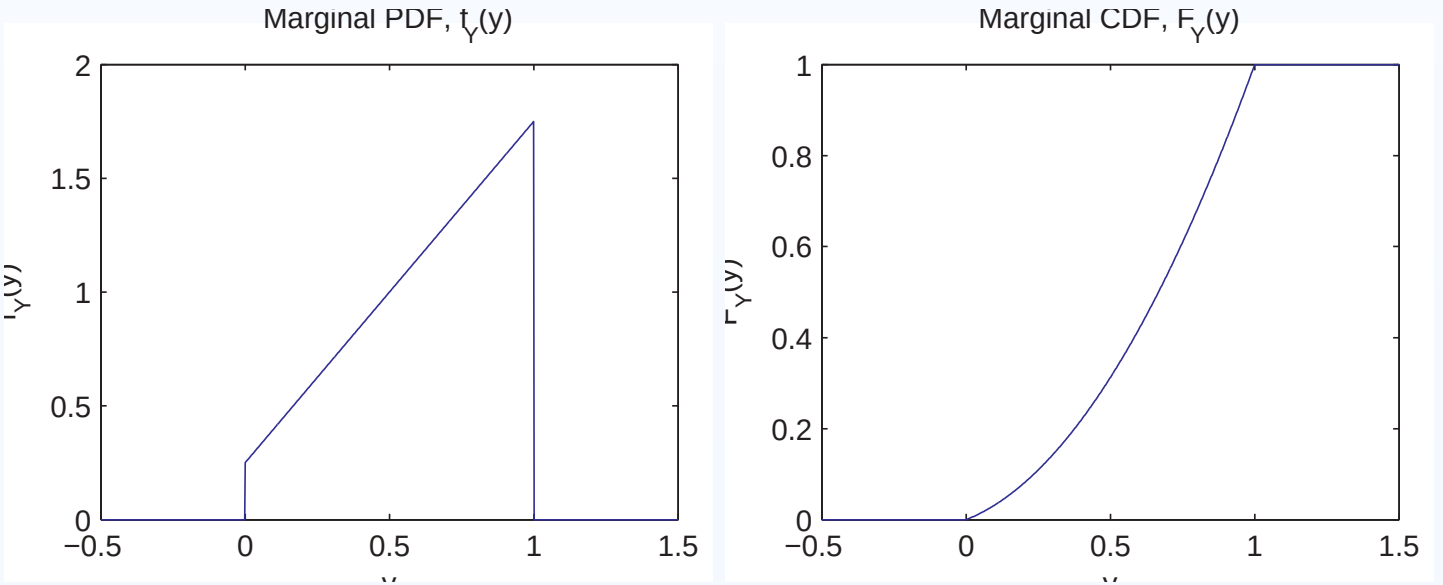
The marginal-pdf, $f_X(x)$, and cdf, $F_X(x)$, for the RV, $X(\zeta)$.

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

SOLUTION. The marginal-pdfs and cdfs are shown below.



The marginal-pdf, $f_Y(y)$, and cdf, $F_Y(y)$, for the RV, $Y(\zeta)$.

Independence

Two random variables, $X_1(\zeta)$ and $X_2(\zeta)$ are **independent** if the events $\{X_1(\zeta) \leq x_1\}$ and $\{X_2(\zeta) \leq x_2\}$ are jointly independent; that is, the events do not influence one another, and

$$\Pr\left(X_1(\zeta) \leq x_1,\, X_2(\zeta) \leq x_2\right) = \Pr\left(X_1(\zeta) \leq x_1\right) \Pr\left(X_2(\zeta) \leq x_2\right)$$

Independence

Two random variables, $X_1(\zeta)$ and $X_2(\zeta)$ are **independent** if the events $\{X_1(\zeta) \leq x_1\}$ and $\{X_2(\zeta) \leq x_2\}$ are jointly independent; that is, the events do not influence one another, and

$$\Pr\left(X_1(\zeta) \leq x_1,\, X_2(\zeta) \leq x_2\right) = \Pr\left(X_1(\zeta) \leq x_1\right) \Pr\left(X_2(\zeta) \leq x_2\right)$$

This then implies that

$$F_{X_1,X_2}\left(x_1,\, x_2\right) = F_{X_1}\left(x_1\right) F_{X_2}\left(x_2\right)$$
$$f_{X_1,X_2}\left(x_1,\, x_2\right) = f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right)$$

Conditionals and Bayes's

The notion of joint probabilities and pdf also leads to the notion of conditional probabilities; what is the probability of a random vector $\mathbf{Y}(\zeta)$, given the random vector $\mathbf{X}(\zeta)$.

Conditionals and Bayes's

The notion of joint probabilities and pdf also leads to the notion of conditional probabilities; what is the probability of a random vector $\mathbf{Y}(\zeta)$, given the random vector $\mathbf{X}(\zeta)$.

The **conditional pdf** of $\mathbf{Y}(\zeta)$ given $\mathbf{X}(\zeta)$ is defined as:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}$$

Conditionals and Bayes's

The notion of joint probabilities and pdf also leads to the notion of conditional probabilities; what is the probability of a random vector $\mathbf{Y}(\zeta)$, given the random vector $\mathbf{X}(\zeta)$.

The **conditional pdf** of $\mathbf{Y}(\zeta)$ given $\mathbf{X}(\zeta)$ is defined as:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}$$

If the random vectors $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ are independent, then the conditional pdf must be identical to the unconditional pdf: $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})$. Hence, it follows that:

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})$$

Conditionals and Bayes's

Since

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{YX}}(\mathbf{y}, \mathbf{x})$$

it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})}$$

Conditionals and Bayes's

Since

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{YX}}(\mathbf{y}, \mathbf{x})$$

it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})}$$

Since $f_{\mathbf{Y}}(\mathbf{y})$ can be expressed as:

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathbb{R}} f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})\, d\mathbf{x} = \int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

then it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}}$$

Statistical Description

Statistical averages are more manageable, but less of a complete description of random vectors.

Statistical Description

Statistical averages are more manageable, but less of a complete description of random vectors.

With care, it is possible to extend many of the statistical descriptors for scalar RVs to random vectors.

Statistical Description

Statistical averages are more manageable, but less of a complete description of random vectors.

With care, it is possible to extend many of the statistical descriptors for scalar RVs to random vectors.

However, it is important to understand that multiple RVs leads to the notion of measuring their interaction or dependence. This concept is useful in abstract, but also when dealing with stochastic processes or time-series.

Statistical Description

Mean vector The **mean vector** is the first-moment of the random vector, and is given by:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)\right] \\ \vdots \\ \mathbb{E}\left[X_N(\zeta)\right] \end{bmatrix} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_N} \end{bmatrix}$$

Statistical Description

Mean vector The **mean vector** is the first-moment of the random vector, and is given by:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)\right] \\ \vdots \\ \mathbb{E}\left[X_N(\zeta)\right] \end{bmatrix} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_N} \end{bmatrix}$$

Correlation Matrix The second-order moments of the random vector describe the spread of the distribution. The **autocorrelation matrix** is defined by:

$$\mathbf{R}_{\mathbf{X}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{X}^{H}\left(\zeta\right)\right] = \begin{bmatrix} r_{X_1X_1} & \cdots & r_{X_1X_N} \\ \vdots & \ddots & \vdots \\ r_{X_NX_1} & \cdots & r_{X_NX_N} \end{bmatrix}$$

Statistical Description

Correlation Matrix The diagonal terms

$$r_{X_i X_i} \triangleq \mathbb{E}\left[|X_i(\zeta)|^2\right], \quad i \in \{1, \ldots, N\}$$

are the second-order moments of each of the RVs, $X_i(\zeta)$.

Statistical Description

Correlation Matrix The diagonal terms

$$r_{X_i X_i} \triangleq \mathbb{E}\left[|X_i(\zeta)|^2\right], \quad i \in \{1, \ldots, N\}$$

are the second-order moments of each of the RVs, $X_i(\zeta)$.

The off-diagonal terms

$$r_{X_i X_j} \triangleq \mathbb{E}\left[X_i(\zeta) X_j^*(\zeta)\right] = r_{X_j X_i}^*, \quad i \neq j$$

measure the **correlation**, or statistical similarity between the RVs $X_i(\zeta)$ and $X_j(\zeta)$.

Statistical Description

Correlation Matrix The diagonal terms

$$r_{X_i X_i} \triangleq \mathbb{E}\left[|X_i(\zeta)|^2\right], \quad i \in \{1, \ldots, N\}$$

are the second-order moments of each of the RVs, $X_i(\zeta)$.

The off-diagonal terms

$$r_{X_i X_j} \triangleq \mathbb{E}\left[X_i(\zeta) X_j^*(\zeta)\right] = r_{X_j X_i}^*, \quad i \neq j$$

measure the **correlation**, or statistical similarity between the RVs $X_i(\zeta)$ and $X_j(\zeta)$.

If the $X_i(\zeta)$ and $X_j(\zeta)$ are **orthogonal** then their **correlation** is zero:

$$r_{X_i X_j} = \mathbb{E}\left[X_i(\zeta) X_j^*(\zeta)\right] = 0, \quad i \neq j$$

Statistical Description

Covariance Matrix The **autocovariance matrix** is defined by:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)^{H}\right] = \begin{bmatrix} \gamma_{X_1X_1} & \cdots & \gamma_{X_1X_N} \\ \vdots & \ddots & \cdots \\ \gamma_{X_NX_1} & \cdots & \gamma_{X_NX_N} \end{bmatrix}$$

Statistical Description

Covariance Matrix The **autocovariance matrix** is defined by:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)^{H}\right] = \begin{bmatrix} \gamma_{X_1X_1} & \cdots & \gamma_{X_1X_N} \\ \vdots & \ddots & \cdots \\ \gamma_{X_NX_1} & \cdots & \gamma_{X_NX_N} \end{bmatrix}$$

The diagonal terms

$$\gamma_{X_iX_i} \triangleq \sigma^2_{X_i} = \mathbb{E}\left[\left|X_i(\zeta)-\mu_{X_i}\right|^2\right], \quad i \in \{1,\ldots,N\}$$

are the **variances** of each of the RVs, $X_i(\zeta)$.

Statistical Description

Covariance Matrix The off-diagonal terms

$$\begin{aligned}\gamma_{X_i X_j} &\triangleq \mathbb{E}\left[\left(X_i(\zeta)-\mu_{X_i}\right)\left(X_j(\zeta)-\mu_{X_j}\right)^*\right] \\ &= r_{X_i X_j} - \mu_{X_i}\mu^*_{X_j} = \gamma^*_{X_j X_i}, \quad i \neq j\end{aligned}$$

measure the **covariance** $X_i(\zeta)$ and $X_j(\zeta)$.

Statistical Description

Covariance Matrix The off-diagonal terms

$$\begin{aligned}\gamma_{X_i X_j} &\triangleq \mathbb{E}\left[\left(X_i(\zeta)-\mu_{X_i}\right)\left(X_j(\zeta)-\mu_{X_j}\right)^*\right] \\ &= r_{X_i X_j} - \mu_{X_i}\mu^*_{X_j} = \gamma^*_{X_j X_i}, \quad i \neq j\end{aligned}$$

measure the **covariance** $X_i(\zeta)$ and $X_j(\zeta)$.

It should also be noticed that the **covariance** and **correlation** matrices are positive semidefinite; that is, they satisfy the relations:

$$\mathbf{a}^H \mathbf{R_X a} \geq 0$$

$$\mathbf{a}^H \mathbf{\Gamma_X a} \geq 0$$

for any complex vector $\mathbf{a}$.

Statistical Description

The autocorrelation and autocovariance matrices are related, and it can easily be seen that:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]^H\right] = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^H$$

In fact, if $\boldsymbol{\mu}_\mathbf{X} = 0$, then $\mathbf{\Gamma_X} = \mathbf{R_X}$.

Statistical Description

The autocorrelation and autocovariance matrices are related, and it can easily be seen that:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]^H\right] = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^H$$

In fact, if $\boldsymbol{\mu}_\mathbf{X} = 0$, then $\mathbf{\Gamma_X} = \mathbf{R_X}$.

If the random variables $X_i(\zeta)$ and $X_j(\zeta)$ are **independent**, then they are also **uncorrelated** since:

$$\begin{aligned} r_{X_iX_j} &= \mathbb{E}\left[X_i(\zeta)\, X_j(\zeta)^*\right] = \mathbb{E}\left[X_i(\zeta)\right]\mathbb{E}\left[X_j^*(\zeta)\right] \\ &= \mu_{X_i}\mu_{X_j}^* \quad \Rightarrow \quad \gamma_{X_iX_j} = 0 \end{aligned}$$

Statistical Description

The autocorrelation and autocovariance matrices are related, and it can easily be seen that:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]^H\right] = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^H$$

In fact, if $\boldsymbol{\mu}_\mathbf{X} = 0$, then $\mathbf{\Gamma_X} = \mathbf{R_X}$.

If the random variables $X_i(\zeta)$ and $X_j(\zeta)$ are **independent**, then they are also **uncorrelated** since:

$$\begin{aligned} r_{X_iX_j} &= \mathbb{E}\left[X_i(\zeta)\, X_j(\zeta)^*\right] = \mathbb{E}\left[X_i(\zeta)\right]\mathbb{E}\left[X_j^*(\zeta)\right] \\ &= \mu_{X_i}\mu_{X_j}^* \quad \Rightarrow \quad \gamma_{X_iX_j} = 0 \end{aligned}$$

Note, however, that uncorrelatedness does not imply independence, unless the RVs are jointly-Gaussian.

Statistical Description

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{Y}^{H}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix}$$

Statistical Description

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{Y}^{H}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix}$$

Cross-covariance is defined as

$$\begin{aligned} \mathbf{\Gamma_{XY}} &\triangleq \mathbb{E}\left[\left\{\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right\}\left\{\mathbf{Y}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{Y}}\right\}^{H}\right] \\ &= \mathbf{R_{XY}} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^{H} \end{aligned}$$

Statistical Description

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{Y}^{H}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix}$$

Cross-covariance is defined as

$$\begin{aligned} \mathbf{\Gamma_{XY}} &\triangleq \mathbb{E}\left[\left\{\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right\}\left\{\mathbf{Y}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{Y}}\right\}^{H}\right] \\ &= \mathbf{R_{XY}} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^{H} \end{aligned}$$

- Uncorrelated if $\mathbf{\Gamma_{XY}} = 0 \quad \Rightarrow \quad \mathbf{R_{XY}} = \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^{H}.$
- Orthogonal if $\mathbf{R_{XY}} = 0.$

Probability Transformation Rule

Theorem (Probability Transformation Rule). The set of random variables $\mathbf{X}(\zeta) = \{X_n(\zeta),\ n \in \mathcal{N}\}$ are transformed to a new set of RVs, $\mathbf{Y}(\zeta) = \{Y_n(\zeta),\ n \in \mathcal{N}\}$, using the transformations:

$$Y_n(\zeta) = g_n(\mathbf{X}(\zeta)), \quad n \in \mathcal{N}$$

Probability Transformation Rule

Theorem (Probability Transformation Rule). The set of random variables $\mathbf{X}(\zeta) = \{X_n(\zeta),\ n \in \mathcal{N}\}$ are transformed to a new set of RVs, $\mathbf{Y}(\zeta) = \{Y_n(\zeta),\ n \in \mathcal{N}\}$, using the transformations:

$$Y_n(\zeta) = g_n(\mathbf{X}(\zeta)), \quad n \in \mathcal{N}$$

Assuming M-real vector-roots of the equation $\mathbf{y} = \mathbf{g}(\mathbf{x})$ by $\{\mathbf{x}_m,\ m \in \mathcal{M}\}$,

$$\mathbf{y} = \mathbf{g}(\mathbf{x}_1) = \cdots = \mathbf{g}(\mathbf{x}_M)$$

then the joint-pdf of $\mathbf{Y}(\zeta)$ in terms of (i. t. o.) the joint-pdf of $\mathbf{X}(\zeta)$is:

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{m=1}^{M} \frac{f_{\mathbf{X}}(\mathbf{x}_m)}{|J(\mathbf{x}_m)|} \qquad \diamondsuit$$

The Jacobian is defined in the notes, but is the usual definition!

Polar Transformation

Consider the transformation from the random vector $\mathbf{C}(\zeta) = [X(\zeta), Y(\zeta)]^T$ to $\mathbf{P}(\zeta) = [r(\zeta), \theta(\zeta)]^T$, where

$$r(\zeta) = \sqrt{X^2(\zeta) + Y^2(\zeta)}$$

$$\theta(\zeta) = \arctan \frac{Y(\zeta)}{X(\zeta)}$$

Polar Transformation

Consider the transformation from the random vector $\mathbf{C}(\zeta) = [X(\zeta), Y(\zeta)]^T$ to $\mathbf{P}(\zeta) = [r(\zeta), \theta(\zeta)]^T$, where

$$r(\zeta) = \sqrt{X^2(\zeta) + Y^2(\zeta)}$$

$$\theta(\zeta) = \arctan \frac{Y(\zeta)}{X(\zeta)}$$

The Jacobian is given by:

$$J_{\mathbf{g}}(\mathbf{c}) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}^{-1} = \frac{1}{r}$$

Thus, it follows that:

$$f_{R,\Theta}(r,\theta) = r f_{XY}(r\cos\theta, r\sin\theta)$$

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable.**

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable.**

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw$$

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable.**

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw$$

Example (Sum of two RVs). If $X(\zeta)$ and $Y(\zeta)$ have joint-pdf $f_{XY}(x, y)$, find the pdf of the RV $Z(\zeta) = aX(\zeta) + bY(\zeta)$.

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable**.

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw$$

Example (Sum of two RVs). If $X(\zeta)$ and $Y(\zeta)$ have joint-pdf $f_{XY}(x, y)$, find the pdf of the RV $Z(\zeta) = aX(\zeta) + bY(\zeta)$.

SOLUTION. Use as the auxiliary variable the function $W(\zeta) = Y(\zeta)$. The system $z = ax + by$, $w = y$ has a single solution at $x = \frac{z-bw}{a}$, $y = w$.

Thus:

$$f_Z(z) = \frac{1}{|a|} \int_{\mathbb{R}} f_{XY}\left(\frac{z - bw}{a}, w\right) dw \qquad \square$$

Multivariate Gaussian Density Function

Gaussian random vectors play a very important role in the design and analysis of signal processing systems. A Gaussian random vector is characterised by a multivariate Normal or Gaussian density function.

Multivariate Gaussian Density Function

Gaussian random vectors play a very important role in the design and analysis of signal processing systems. A Gaussian random vector is characterised by a multivariate Normal or Gaussian density function.

For a *real* random vector, this density function has the form:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{\Gamma_X}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T \mathbf{\Gamma}_{\mathbf{X}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})\right]$$

where N is the dimension of $\mathbf{X}(\zeta)$, and $\mathbf{X}(\zeta)$ has mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\mathbf{\Gamma_X}$. It is often denoted as:

$$f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \middle| \boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Gamma_X}\right)$$

Multivariate Gaussian Density Function

The normal distribution is a useful model of a random vector because of its many important properties.

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \middle| \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$ is completely specified by its mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent.

Multivariate Gaussian Density Function

The normal distribution is a useful model of a random vector because of its many important properties.

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$ is completely specified by its mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent.

3. A linear transformation of a normal random vector is also normal.

   This is a particularly useful, since the output of a linear system subject to a Gaussian input is also Gaussian.

Multivariate Gaussian Density Function

The normal distribution is a useful model of a random vector because of its many important properties.

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$ is completely specified by its mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent.

3. A linear transformation of a normal random vector is also normal.

   This is a particularly useful, since the output of a linear system subject to a Gaussian input is also Gaussian.

4. If $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ are *jointly*-Gaussian, then so are their *marginal*-distributions, and their *conditional*-distributions.

Handout 5
Estimation Theory

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.

- In most practical applications, this is the exception rather than the rule.

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.
- In most practical applications, this is the exception rather than the rule.
- The properties and parameters of random events must be obtained by collecting and analysing finite set of measurements.

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.
- In most practical applications, this is the exception rather than the rule.
- The properties and parameters of random events must be obtained by collecting and analysing finite set of measurements.
- This handout will consider the problem of **Parameter Estimation**. This refers to the estimation of a parameter that is fixed, but is unknown.

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

The function $\hat{\theta}\left[\mathcal{X}\right]$ is known as an **estimator** whereas the value taken by the estimator, using a particular set of observations, is called a **point-estimate**.

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

The function $\hat{\theta}\left[\mathcal{X}\right]$ is known as an **estimator** whereas the value taken by the estimator, using a particular set of observations, is called a **point-estimate**.

An aim is to design an estimator, $\hat{\theta}$, that should be as close to the true value of the parameter, θ, as possible.

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

The function $\hat{\theta}\left[\mathcal{X}\right]$ is known as an **estimator** whereas the value taken by the estimator, using a particular set of observations, is called a **point-estimate**.

An aim is to design an estimator, $\hat{\theta}$, that should be as close to the true value of the parameter, θ, as possible.

Since $\hat{\theta}$ is a function of a number of particular realisations of a random outcome (or experiment), then it is itself a RV, and thus has a mean and variance.

Bias of estimator

The **bias** of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

Bias of estimator

The **bias** of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

If θ is large, then a small deviation would give what would appear to be a large bias. Thus, the **normalised bias** is often used instead:

$$\epsilon_b(\hat{\theta}) \triangleq \frac{B(\hat{\theta})}{\theta} = \frac{\mathbb{E}\left[\hat{\theta}\right]}{\theta} - 1, \quad \theta \neq 0$$

Bias of estimator

The **bias** of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

If θ is large, then a small deviation would give what would appear to be a large bias. Thus, the **normalised bias** is often used instead:

$$\epsilon_b(\hat{\theta}) \triangleq \frac{B(\hat{\theta})}{\theta} = \frac{\mathbb{E}\left[\hat{\theta}\right]}{\theta} - 1, \quad \theta \neq 0$$

Example (Biasness of sample mean estimator). Is the sample mean, $\hat{\mu}_x = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$ biased?

SOLUTION. No, since

$$\mathbb{E}\left[\hat{\mu}_x\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[x[n]\right] = \frac{N\mu_X}{N} = \mu_X.$$

Variance of estimator

The **variance** of the estimator $\hat{\theta}$ is defined by:

$$\text{var}\left[\hat{\theta}\right] = \sigma^2_{\hat{\theta}} \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right]$$

However, a minimum variance criterion is not always compatible with the minimum bias requirement; reducing the variance may result in an increase in bias.

Variance of estimator

The **variance** of the estimator $\hat{\theta}$ is defined by:

$$\text{var}\left[\hat{\theta}\right] = \sigma^2_{\hat{\theta}} \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right]$$

However, a minimum variance criterion is not always compatible with the minimum bias requirement; reducing the variance may result in an increase in bias.

Therefore, a compromise or balance between these two conflicting criteria is required, and this is provided by the mean-squared error (MSE) measure described below.

Variance of estimator

The **variance** of the estimator $\hat{\theta}$ is defined by:

$$\text{var}\left[\hat{\theta}\right] = \sigma^2_{\hat{\theta}} \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right]$$

However, a minimum variance criterion is not always compatible with the minimum bias requirement; reducing the variance may result in an increase in bias.

Therefore, a compromise or balance between these two conflicting criteria is required, and this is provided by the mean-squared error (MSE) measure described below.

The **normalised standard deviation** is defined by:

$$\epsilon_r \triangleq \frac{\sigma_{\hat{\theta}}}{\theta}, \quad \theta \neq 0$$

Mean square error

Minimising estimator variance can increase bias. A compromise criterion is the mean-squared error (MSE) of the estimator, which is given by:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2$$

Mean square error

Minimising estimator variance can increase bias. A compromise criterion is the mean-squared error (MSE) of the estimator, which is given by:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2$$

The estimator $\hat{\theta}_{\text{MSE}} = \hat{\theta}_{\text{MSE}}[\mathcal{X}]$ which minimises $\text{MSE}(\hat{\theta})$ is known as the minimum mean-square error:

$$\hat{\theta}_{MSE} = \arg_{\hat{\theta}} \min \text{MSE}(\hat{\theta})$$

Mean square error

Minimising estimator variance can increase bias. A compromise criterion is the mean-squared error (MSE) of the estimator, which is given by:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

The estimator $\hat{\theta}_{\mathrm{MSE}} = \hat{\theta}_{\mathrm{MSE}}[\mathcal{X}]$ which minimises $\mathrm{MSE}(\hat{\theta})$ is known as the minimum mean-square error:

$$\hat{\theta}_{MSE} = \arg_{\hat{\theta}} \min \mathrm{MSE}(\hat{\theta})$$

This measures the average mean squared deviation of the estimator from its true value.

Unfortunately, adoption of this natural criterion leads to unrealisable estimators; ones which cannot be written solely as a function of the data.

Cramer-Rao Lower Bound

If the MSE can be minimised when the bias is zero, then clearly the variance is also minimised. Such estimators are called minimum variance unbiased estimators (MVUEs).

Cramer-Rao Lower Bound

If the MSE can be minimised when the bias is zero, then clearly the variance is also minimised. Such estimators are called MVUEs.

MVUE possess the important property that they attain a minimum bound on the variance of the estimator, called the Cramér-Rao lower-bound (CRLB).

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Alternatively, it may also be expressed as:

$$\text{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]}$$

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Alternatively, it may also be expressed as:

$$\text{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]}$$

$\diamondsuit$

The function $\ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is called the **log-likelihood** of θ.

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Alternatively, it may also be expressed as:

$$\text{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]}$$

Furthermore, an unbiased estimator may be found that attains the bound for all θ if, and only if, (iff)

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} = I(\theta)\left(\hat{\theta} - \theta\right)$$

Consistency of an Estimator

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

approaches zero as the sample size N becomes large, then both the bias and the variance tends toward zero.

Consistency of an Estimator

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

approaches zero as the sample size N becomes large, then both the bias and the variance tends toward zero.

Thus, the sampling distribution tends to concentrate around θ, and as $N \to \infty$, it will become an impulse at θ.

Consistency of an Estimator

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

approaches zero as the sample size N becomes large, then both the bias and the variance tends toward zero.

Thus, the sampling distribution tends to concentrate around θ, and as $N \to \infty$, it will become an impulse at θ.

This is a very important and desirable property, and such an estimator is called a **consistent estimator**.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

This same quantity, viewed as a function of the parameter $\boldsymbol{\theta}$ when a particular set of observations, $\hat{\mathbf{x}}$ is given, is known as the **likelihood function**.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

This same quantity, viewed as a function of the parameter $\boldsymbol{\theta}$ when a particular set of observations, $\hat{\mathbf{x}}$ is given, is known as the **likelihood function**.

The **maximum-likelihood estimate (MLE)** of the parameter $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{ml}$, is defined as that value of $\boldsymbol{\theta}$ that maximises $f_{\mathbf{X}}\left(\hat{\mathbf{x}} \mid \boldsymbol{\theta}\right)$.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

This same quantity, viewed as a function of the parameter $\boldsymbol{\theta}$ when a particular set of observations, $\hat{\mathbf{x}}$ is given, is known as the **likelihood function**.

The **maximum-likelihood estimate (MLE)** of the parameter $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{ml}$, is defined as that value of $\boldsymbol{\theta}$ that maximises $f_{\mathbf{X}}\left(\hat{\mathbf{x}} \mid \boldsymbol{\theta}\right)$.

The MLE for $\boldsymbol{\theta}$ is defined by:

$$\hat{\boldsymbol{\theta}}_{ml}(\mathbf{x}) = \arg_{\boldsymbol{\theta}} \max f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$$

Note that since $\hat{\boldsymbol{\theta}}_{ml}(\mathbf{x})$ depends on the random observation vector $\mathbf{x}$, and so is *itself a RV*.

Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

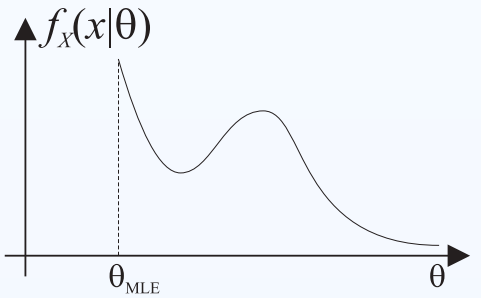
$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$
$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

2. If an MVUE exists and the MLE does not occur at a boundary, then the MLE *is* the MVUE.



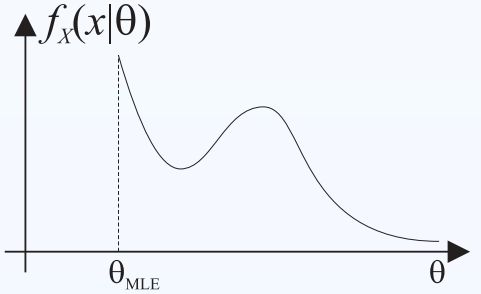
A single parameter MLE that occurs at a boundary

Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$
$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

2. If an MVUE exists and the MLE does not occur at a boundary, then the MLE *is* the MVUE.



A single parameter MLE that occurs at a boundary

3. If the pdf, $f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$, satisfies certain *regularity* conditions, then the MLE is asymptotically distributed according to a Gaussian distribution:

$$\hat{\boldsymbol{\theta}}_{ml} \sim \mathcal{N}\left(\boldsymbol{\theta}, \mathbf{J}^{-1}(\boldsymbol{\theta})\right)$$

DC Level in white Gaussian noise

Example ([Therrien:1991, Example 6.1, Page 282]). A constant but unknown signal is observed in additive white Gaussian noise (WGN). That is,

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}\left(0, \sigma_w^2\right) \qquad \bowtie$$

for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the MLE of the unknown signal A.

DC Level in white Gaussian noise

Example ([Therrien:1991, Example 6.1, Page 282]). A constant but unknown signal is observed in additive WGN. That is,

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}\left(0, \sigma_w^2\right)$$

for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the MLE of the unknown signal A.

SOLUTION. Since this is a memoryless system, and $w(n)$ are independent and identically distributed (i. i. d.), then so is $x[n]$, and

the log-likelihood is given by:

$$\ln f_{\mathbf{x}}\left(\mathbf{x} \mid A\right) = -\frac{N}{2}\ln(2\pi\sigma_w^2) - \frac{\sum_{n\in\mathcal{N}}\left(x[n]-A\right)^2}{2\sigma_w^2}$$

Differentiating this expression w. r. t. A

and setting to zero :

MLE for Transformed Parameter

Theorem (Invariance Property of the MLE). The MLE of the parameter $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g}$ is an r-dimensional function of the $P \times 1$ parameter $\boldsymbol{\theta}$, and the pdf, $f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})$ is parameterised by $\boldsymbol{\theta}$, is given by

$$\hat{\boldsymbol{\alpha}}_{ml} = \mathbf{g}(\hat{\boldsymbol{\theta}}_{ml}) \qquad \diamondsuit$$

where $\hat{\boldsymbol{\theta}}_{ml}$ is the MLE of $\boldsymbol{\theta}$.

MLE for Transformed Parameter

Theorem (Invariance Property of the MLE). The MLE of the parameter $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g}$ is an r-dimensional function of the $P \times 1$ parameter $\boldsymbol{\theta}$, and the pdf, $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ is parameterised by $\boldsymbol{\theta}$, is given by

$$\hat{\boldsymbol{\alpha}}_{ml} = \mathbf{g}(\hat{\boldsymbol{\theta}}_{ml})$$

where $\hat{\boldsymbol{\theta}}_{ml}$ is the MLE of $\boldsymbol{\theta}$.

The MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{ml}$, is obtained by maximising $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$. If the function $\mathbf{g}$ is not an invertible function, then $\hat{\boldsymbol{\alpha}}$ maximises the modified likelihood function $\bar{p}_T\left(\mathbf{x} \mid \boldsymbol{\alpha}\right)$ defined as:

$$\bar{p}_T\left(\mathbf{x} \mid \boldsymbol{\alpha}\right) = \max_{\boldsymbol{\theta}:\boldsymbol{\alpha}=\mathbf{g}(\boldsymbol{\theta})} f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \qquad \diamondsuit$$

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

An alternate philosophy is a class of estimators that in general have no optimality properties associated with them, but make *good sense* for many problems of interest: the **principle of least squares**.

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

An alternate philosophy is a class of estimators that in general have no optimality properties associated with them, but make *good sense* for many problems of interest: the **principle of least squares**.

A salient feature of the method is that *no probabilistic assumptions* are made about the data; only a *signal model* is assumed.

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

An alternate philosophy is a class of estimators that in general have no optimality properties associated with them, but make *good sense* for many problems of interest: the **principle of least squares**.

A salient feature of the method is that *no probabilistic assumptions* are made about the data; only a *signal model* is assumed.

As will be seen, it turns out that the LSE can be calculated when just the first and second moments are known, and through the solution of *linear* equations.

The Least Squares Approach

In the least-squares (LS) approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

The Least Squares Approach

In the LS approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$.

The Least Squares Approach

In the LS approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$.

The LSE of $\boldsymbol{\theta}$ chooses the value that makes $s[n]$ closest to the observed data $x[n]$, and this *closeness* is measured by the LS error criterion:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \left(x[n] - s[n]\right)^2$$

The Least Squares Approach

In the LS approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$.

The LSE of $\boldsymbol{\theta}$ chooses the value that makes $s[n]$ closest to the observed data $x[n]$, and this *closeness* is measured by the LS error criterion:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \left(x[n] - s[n]\right)^2$$

The LSE is given by:

$$\hat{\boldsymbol{\theta}}_{LSE} = \arg_{\boldsymbol{\theta}} \min J(\boldsymbol{\theta})$$

DC Level

Example ([Kay:1993, Example 6.1, Page 221]). It is assumed that an observed signal, $x[n]$, is a perturbed version of an unknown signal, $s[n]$, which is modelled as $s[n] = A$, for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the LSE of the unknown signal A.

DC Level

Example ([Kay:1993, Example 6.1, Page 221]). It is assumed that an observed signal, $x[n]$, is a perturbed version of an unknown signal, $s[n]$, which is modelled as $s[n] = A$, for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the LSE of the unknown signal A.

SOLUTION. According to the LS approach, then:

$$\hat{A}_{LSE} = \arg_A \min J(A) \quad \text{where} \quad J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2$$

Differentiating w. r. t. A and setting the result to zero produces

$$\hat{A}_{LSE} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \qquad \square$$

which is the sample mean estimator.

Linear Least Squares

Thus, the unknown random-vector $\mathbf{s}$ is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

Linear Least Squares

Thus, the unknown random-vector $\mathbf{s}$ is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

Setting the gradient of $J(\boldsymbol{\theta})$ to zero yields the LSE:

$$\hat{\boldsymbol{\theta}}_{LSE} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

Linear Least Squares

Thus, the unknown random-vector $\mathbf{s}$ is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

Setting the gradient of $J(\boldsymbol{\theta})$ to zero yields the LSE:

$$\hat{\boldsymbol{\theta}}_{LSE} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

The equations $\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = \mathbf{H}^T\mathbf{x}$, to be solved for $\hat{\boldsymbol{\theta}}$, are termed the **normal equation**.

Handout 6
MonteCarlo

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Optimisation: involves finding the solution to

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} h(\boldsymbol{\theta})$$

where $h(\cdot)$ is a scalar function of a multi-dimensional vector of parameters, $\boldsymbol{\theta}$.

Typically, $h(\cdot)$ might represent some **cost function**, and it is implicitly assumed that the optimisation cannot be calculated explicitly.

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Integration: involves evaluating an integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

that cannot explicitly be calculated in *closed form*.

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Integration: involves evaluating an integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

that cannot explicitly be calculated in *closed form*.

For example, the Gaussian-error function:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}\, d\theta$$

Again, the integral may be multi-dimensional, and in general $\boldsymbol{\theta}$ is a vector.

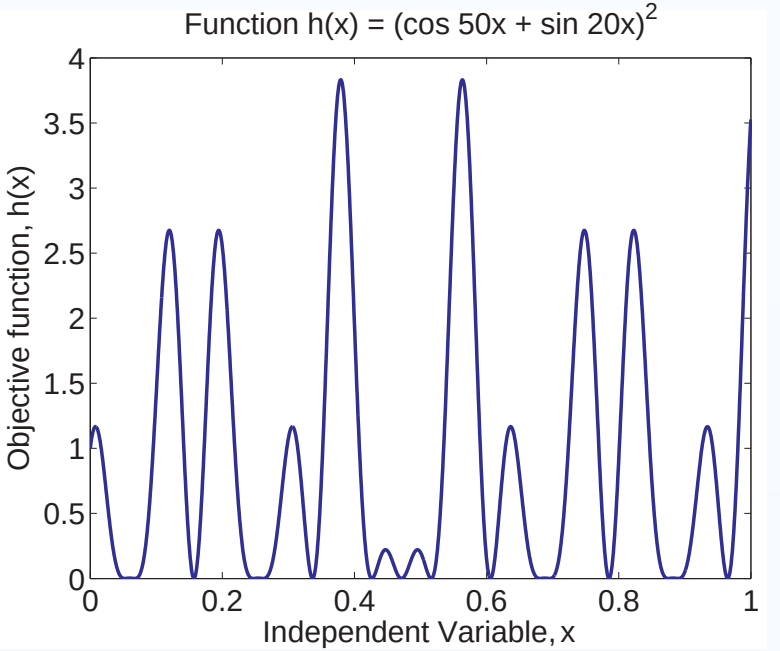
Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Optimisation and Integration Some problems involve both integration and optimisation: a fundamental problem is the maximisation of a marginal distribution:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \int_{\Omega} f(\boldsymbol{\theta}, \boldsymbol{\omega})\, d\boldsymbol{\omega}$$

Deterministic Numerical Methods



Plot of the function $h(x) = (\cos 50x + \sin 20x)^2$, $0 \leq x \leq 1$.

There are various deterministic solutions to the optimisation and integration problems.

Deterministic Numerical Methods

Optimisation: 1. Golden-section search and Brent's Method in one dimension;

2. Nelder and Mead Downhill Simplex method in multi-dimensions;
3. Gradient and Variable-Metric methods in multi-dimensions, typically an extension of Newton-Raphson methods.

Deterministic Numerical Methods

Integration: Most deterministic integration rely on classic formulas for equally spaced abscissas:

1. simple Riemann integration;
2. standard and extended Simpson's and Trapezoidal rules;
3. refinements such as Romberg Integration.

Unfortunately, these methods are not easily extended to multi-dimensions.

Deterministic Numerical Methods

Integration: Most deterministic integration rely on classic formulas for equally spaced abscissas:

1. simple Riemann integration;
2. standard and extended Simpson's and Trapezoidal rules;
3. refinements such as Romberg Integration.

More sophisticated approaches allow non-uniformally spaced abscissas at which the function is evaluated.

These methods tend to use Gaussian quadratures and orthogonal polynomials. Splines are also used.

Unfortunately, these methods are not easily extended to multi-dimensions.

Deterministic Optimisation

The **Nelder-Mead Downhill Simplex method** simply crawls downhill in a straightforward fashion that makes almost no special assumptions about your function.

This can be extremely slow, but it can be robust.

Deterministic Optimisation

Gradient methods are typically based on the Newton-Raphson algorithm which solves $\nabla h(\boldsymbol{\theta}) = \mathbf{0}$.

For a scalar function, $h(\boldsymbol{\theta})$, of a vector of independent variables $\boldsymbol{\theta}$, a sequence $\boldsymbol{\theta}_n$ is produced such that:

Deterministic Optimisation

Gradient methods are typically based on the Newton-Raphson algorithm which solves $\nabla h(\boldsymbol{\theta}) = \mathbf{0}$.

For a scalar function, $h(\boldsymbol{\theta})$, of a vector of independent variables $\boldsymbol{\theta}$, a sequence $\boldsymbol{\theta}_n$ is produced such that:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \left(\nabla \nabla^T h\left(\boldsymbol{\theta}_n\right)\right)^{-1} \nabla h\left(\boldsymbol{\theta}_n\right)$$

Numerous variants of Newton-Raphson-type techniques exist, and include the **steepest descent method**, or the **Levenberg-Marquardt method**.

Deterministic Integration

The integral

$$\mathcal{I} = \int_a^b f(\theta)\, d\theta,$$

where θ is a scalar, and $b > a$, can be solved with the trapezoidal rule using

$$\hat{I} = \frac{1}{2} \sum_{k=0}^{N-1} (\theta_{k+1} - \theta_k)\left(f(\theta_k) + f(\theta_{k+1})\right)$$

where the θ_k's constitute an ordered partition of $[a, b]$.

Deterministic Integration

The integral

$$\mathcal{I} = \int_a^b f(\theta)\, d\theta,$$

where θ is a scalar, and $b > a$, can be solved with the trapezoidal rule using

$$\hat{I} = \frac{1}{2} \sum_{k=0}^{N-1} (\theta_{k+1} - \theta_k)\left(f(\theta_k) + f(\theta_{k+1})\right)$$

where the θ_k's constitute an ordered partition of $[a, b]$.

Another formula is **Simpson's rule**:

$$\hat{I} = \frac{\delta}{3}\left\{ f(a) + 4\sum_{k=1}^{N} f(\theta_{2k-1}) + 2\sum_{k=1}^{N} h(\theta_{2k}) + f(b) \right\}$$

in the case of equally spaced samples with $\delta = \theta_{k+1} - \theta_k$.

Monte Carlo Numerical Methods

Monte Carlo methods are stochastic techniques, in which random numbers are generated and use to examine some problem.

Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Defining a function $\pi(\boldsymbol{\theta})$ which is non-zero and positive for all $\boldsymbol{\theta} \in \Theta$, this integral can be expressed in the alternate form:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where the function $\pi(\boldsymbol{\theta}) > 0$, $\boldsymbol{\theta} \in \Theta$ is a pdf which satisfies

$$\int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$$

Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Defining a function $\pi(\boldsymbol{\theta})$ which is non-zero and positive for all $\boldsymbol{\theta} \in \Theta$, this integral can be expressed in the alternate form:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where the function $\pi(\boldsymbol{\theta}) > 0$, $\boldsymbol{\theta} \in \Theta$ is a pdf which satisfies

$$\int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$$

This may be written as an expectation:

$$\mathcal{I} = \mathbb{E}_{\pi}\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right]$$

Monte Carlo Integration

This expectation can be estimated using the idea of the **sample expectation**, and leads to the idea behind Monte Carlo integration:

1. Sample N random variates from a density function $\pi(\boldsymbol{\theta})$,

$$\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}), \quad k \in \mathcal{N} = \{0, \ldots, N-1\}$$

2. Calculate the sample average of the expectation using

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})} \approx \mathbb{E}_{\pi} \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right]$$

Stochastic Optimisation

There are two distinct approaches to the Monte Carlo optimisation of the objective function $h(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} h(\boldsymbol{\theta})$$

The first method is broadly known as an **exploratory approach**, while the second approach is based on a **probabilistic approximation** of the objective function.

Stochastic Optimisation

Exploratory approach This approach is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

Stochastic Optimisation

Exploratory approach This approach is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

For example, maximisation can be solved by sampling a large number, N, of independent random variables, $\{\boldsymbol{\theta}^{(k)}\}$, from a pdf $\pi(\boldsymbol{\theta})$, and taking the estimate:

$$\hat{\boldsymbol{\theta}} \approx \underset{\{\boldsymbol{\theta}^{(k)}\}}{\arg\max}\, h\left(\boldsymbol{\theta}^{(k)}\right)$$

Typically, when no specific features regarding the function $h(\boldsymbol{\theta})$, are taken into account, $\pi(\boldsymbol{\theta})$ will take on a uniform distribution over Θ.

Stochastic Optimisation

Exploratory approach This approach is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

For example, maximisation can be solved by sampling a large number, N, of independent random variables, $\{\boldsymbol{\theta}^{(k)}\}$, from a pdf $\pi(\boldsymbol{\theta})$, and taking the estimate:

$$\hat{\boldsymbol{\theta}} \approx \underset{\{\boldsymbol{\theta}^{(k)}\}}{\arg\max}\, h\left(\boldsymbol{\theta}^{(k)}\right)$$

Typically, when no specific features regarding the function $h(\boldsymbol{\theta})$, are taken into account, $\pi(\boldsymbol{\theta})$ will take on a uniform distribution over Θ.

Stochastic Approximation The Monte Carlo EM algorithm

Generating Random Variables

This section discusses a variety of techniques for generating random variables from a different distributions.

Uniform Variates

The foundation underpinning all stochastic simulations is the ability to generate a sequence of i. i. d. uniform random variates over the range $(0, 1]$.

Uniform Variates

The foundation underpinning all stochastic simulations is the ability to generate a sequence of i. i. d. uniform random variates over the range $(0, 1]$.

Random variates are *pseudo* or *synthetic* and not truly random since they are usually generated using a recurrence of the form:

$$x_{n+1} = (a\, x_n + b) \mod m$$

Uniform Variates

The foundation underpinning all stochastic simulations is the ability to generate a sequence of i. i. d. uniform random variates over the range $(0, 1]$.

Random variates are *pseudo* or *synthetic* and not truly random since they are usually generated using a recurrence of the form:

$$x_{n+1} = (a\, x_n + b) \mod m$$

This is known as the linear congruential generator.

However, suitable values of a, b and m can be chosen such that the random variates pass all statistical tests of randomness.

Transformation Methods

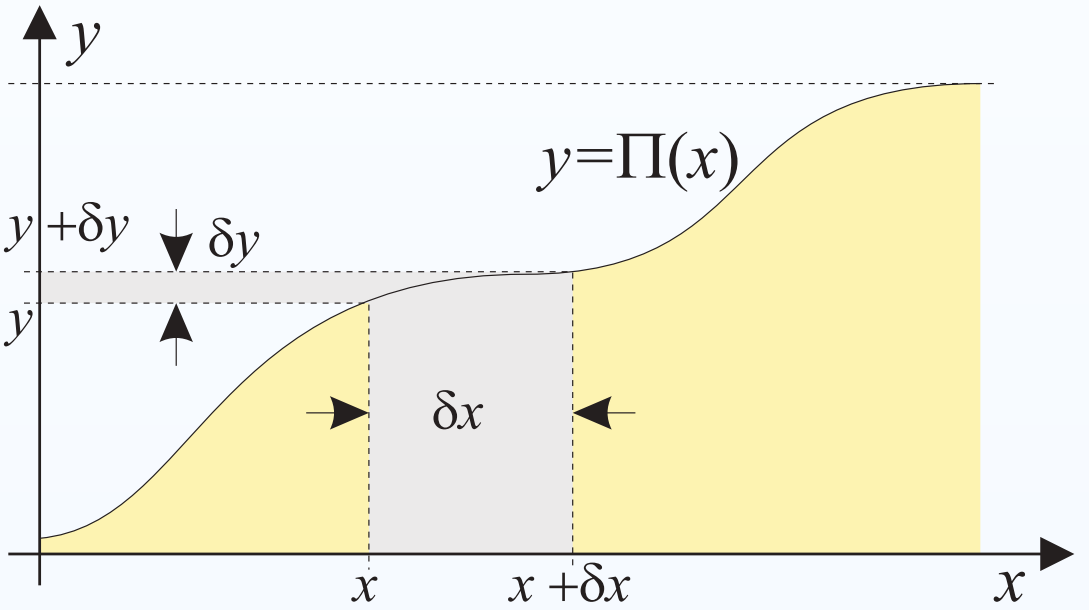
It is possible to sample from a number of extremely important probability distributions by applying various probability transformation methods.

Theorem (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n,\, n \in \mathcal{N}\}$, such that

$$y = g(x_1) = \cdots = g(x_N)$$

PROOF. The proof is given in the handout on scalar random variables.

Inverse Transform Method

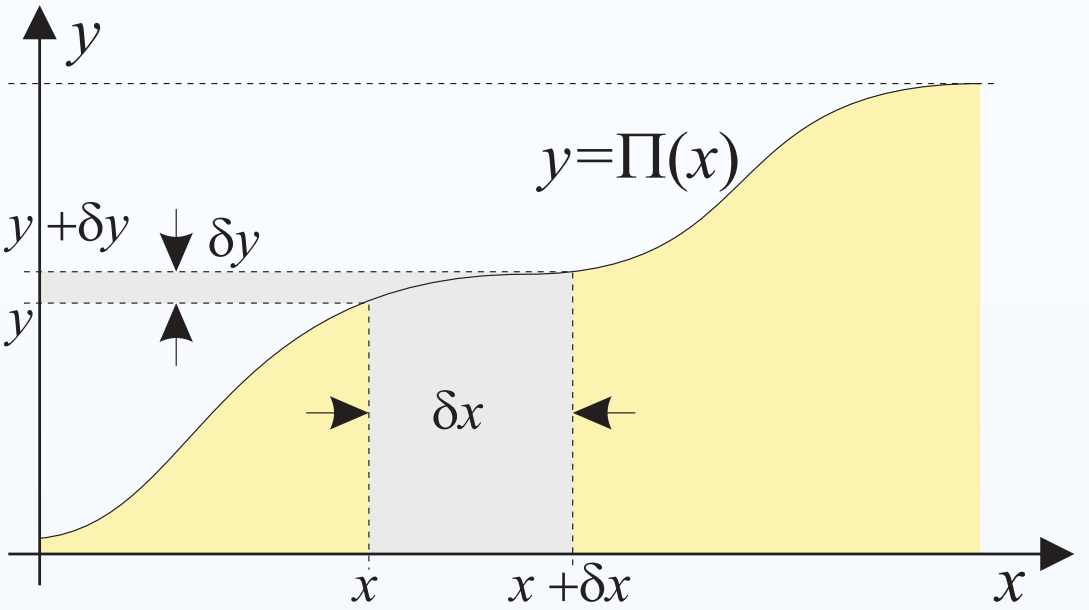


A simple derivation of the inverse transform method

$X(\zeta)$ and $Y(\zeta)$ are RVs related by the function $Y(\zeta) = \Pi(X(\zeta))$.

$\Pi(\zeta)$ is monotonically increasing so that there is only one solution to the equation $y = \Pi(x)$, $x = \Pi^{-1}(y)$.

Inverse Transform Method



A simple derivation of the inverse transform method

$$f_X(x) = \frac{d\Pi(x)}{dx} f_Y(y)$$

Now, suppose $Y(\zeta) \sim \mathcal{U}_{[0,1]}$ is a uniform random variable. If $\Pi(x)$ is the cdf corresponding to a desired pdf $\pi(x)$, then

$$f_X(x) = \pi(x), \quad \text{where} \quad x = \Pi^{-1}(y)$$

Inverse Transform Method

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1} U(\zeta) \sim \pi\left(x\right)$$

Inverse Transform Method

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1} U(\zeta) \sim \pi\left(x\right)$$

Example (Exponential variable generation). If $X(\zeta) \sim \mathcal{E}xp(1)$, such that $\pi(x) = e^{-x}$ and $\Pi(x) = 1 - e^{-x}$, then solving for x in terms of $u = 1 - e^{-x}$ gives $x = -\log(1-u)$.

Inverse Transform Method

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1} U(\zeta) \sim \pi\left(x\right)$$

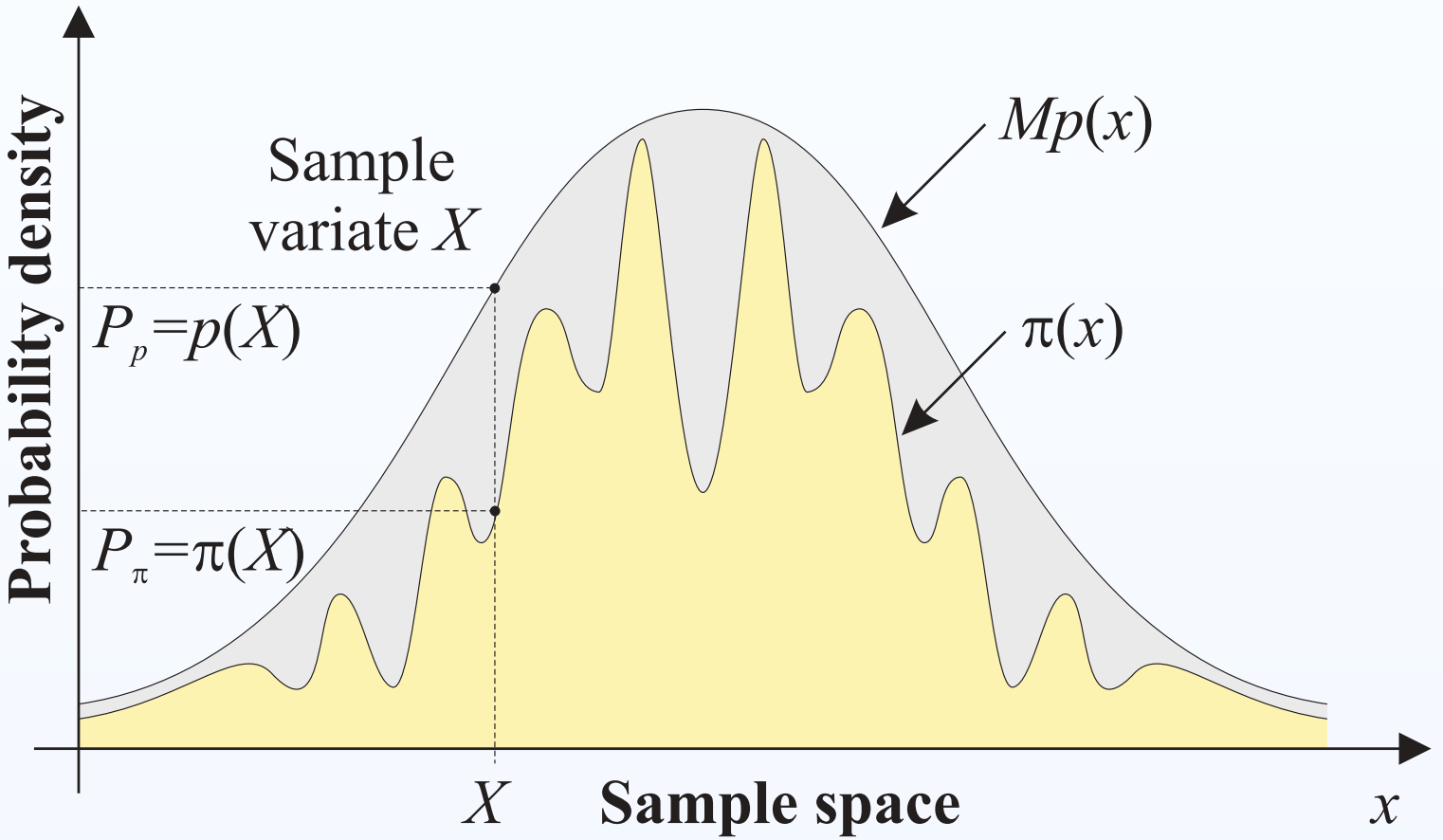
Example (Exponential variable generation). If $X(\zeta) \sim \mathcal{E}xp(1)$, such that $\pi(x) = e^{-x}$ and $\Pi(x) = 1 - e^{-x}$, then solving for x in terms of $u = 1 - e^{-x}$ gives $x = -\log(1-u)$.

Therefore, if $U(\zeta) \sim \mathcal{U}_{[0,\,1]}$, then the RV from the transformation $X(\zeta) = -\log U(\zeta)$ has the exponential distribution (since $U(\zeta)$ and $1 - U(\zeta)$ are both uniform).

Acceptance-Rejection Sampling

For most distributions, it is often difficult or even impossible to directly simulate using either the inverse transform or probability transformations.

Acceptance-Rejection Sampling



On average, you would expect to have too many variates that take on the value X by a factor of

$$u(X) = \frac{P_p}{P_\pi} = \frac{p(X)}{\pi(X)}$$

Acceptance-Rejection Sampling

Thus, to reduce the number of variates that take on a value of X, simply throw away a number of samples in proportion to the amount of *over sampling*.

Acceptance-Rejection Sampling

Thus, to reduce the number of variates that take on a value of X, simply throw away a number of samples in proportion to the amount of *over sampling*.

1. Generate the random variates $X \sim p(x)$ and $U \sim \mathcal{U}_{[0,\,1]}$;
2. Accept X if $U \leq P_a = \frac{\pi(X)}{Mp(x)}$;
3. Otherwise, reject and return to first step.

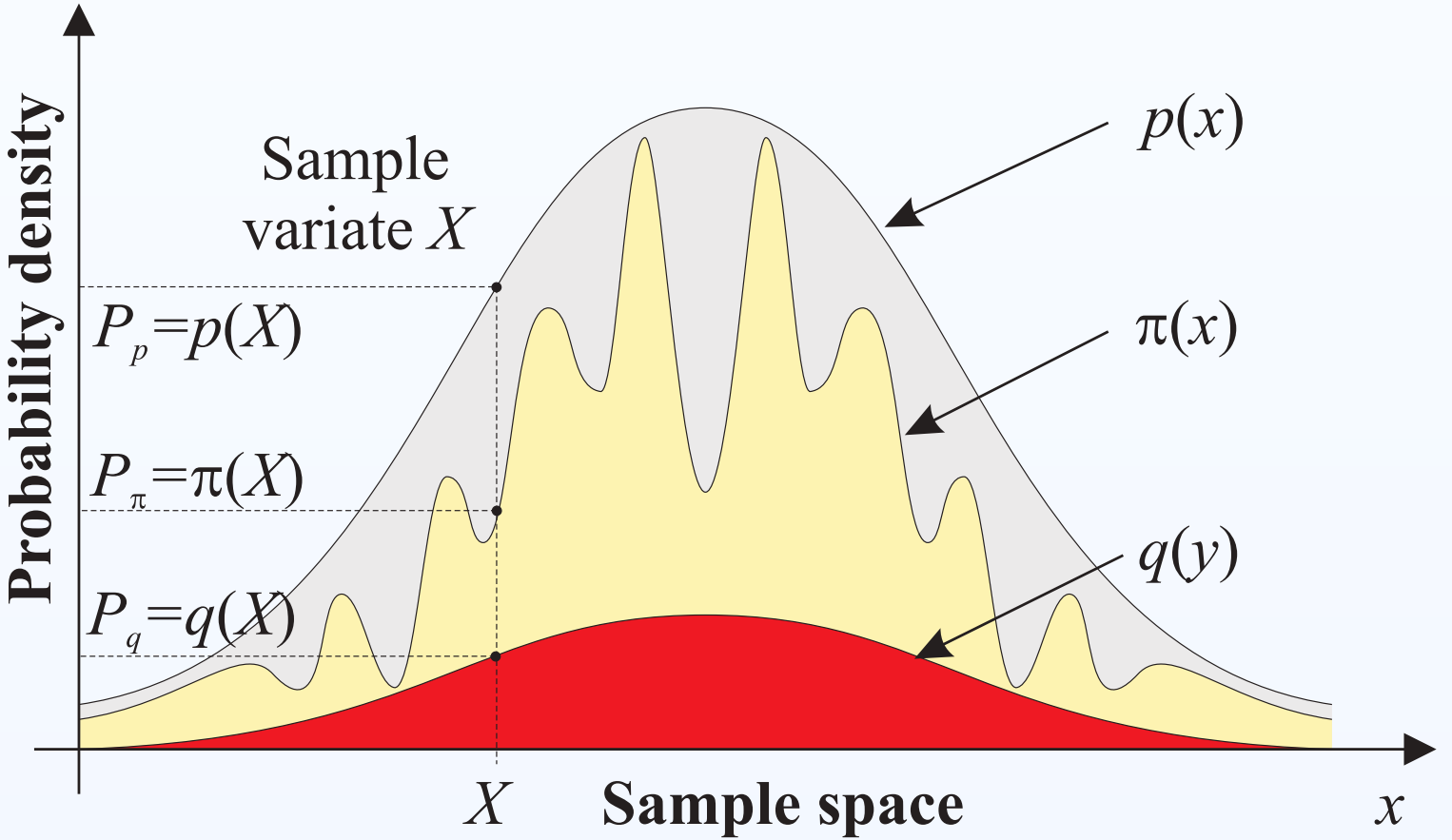
Envelope and Squeeze Methods

A problem with many sampling methods, which can make the density $\pi(x)$ difficult to simulate, is that the function may require substantial computing time at each evaluation.

It is possible to reduce the algorithmic complexity by looking for another computationally simple function, $q(x)$ which *bounds* $\pi(x)$ *from below*.

Envelope and Squeeze Methods

If X satisfies $q(X) \leq \pi(X)$, then it should be accepted when $U \leq \frac{q(X)}{Mp(x)}$, since this also satisfies $U \leq \frac{\pi(X)}{Mp(x)}$.



Envelope and Squeeze Methods

This leads to the **envelope accept-reject algorithm**:

1. Generate the random variates $X \sim p(x)$ and $U \sim \mathcal{U}_{[0,\,1]}$;
2. Accept X if $U \leq \frac{q(X)}{Mp(x)}$;
3. Otherwise, accept X if $U \leq \frac{\pi(X)}{Mp(x)}$;
4. Otherwise, reject and return to first step.

Envelope and Squeeze Methods

This leads to the **envelope accept-reject algorithm**:

1. Generate the random variates $X \sim p(x)$ and $U \sim \mathcal{U}_{[0,\,1]}$;
2. Accept X if $U \leq \frac{q(X)}{Mp(x)}$;
3. Otherwise, accept X if $U \leq \frac{\pi(X)}{Mp(x)}$;
4. Otherwise, reject and return to first step.

By construction of a lower envelope on $\pi\left(x\right)$, the number of function evaluations is potentially decreased by a factor of

$$P_{\bar{\pi}} = \frac{1}{M} \int q\left(x\right)\, dx$$

which is the probability that $\pi\left(x\right)$ is not evaluated.

Importance Sampling

The problem with accept-reject sampling methods is finding the envelope functions and the constant M.

Importance Sampling

The problem with accept-reject sampling methods is finding the envelope functions and the constant M.

The simplest application of **importance sampling** is in Monte Carlo integration. Suppose that is is desired to evaluate the function:

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Importance Sampling

The problem with accept-reject sampling methods is finding the envelope functions and the constant M.

The simplest application of **importance sampling** is in Monte Carlo integration. Suppose that is is desired to evaluate the function:

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Approximate by empirical average:

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{I}_{\Theta}\left(\boldsymbol{\theta}^{(k)}\right), \quad \text{where} \quad \boldsymbol{\theta}^{(k)} \sim f(\boldsymbol{\theta})$$

where $\mathbb{I}_{\mathcal{A}}(a)$ is the indicator function, and is equal to one if $a \in \mathcal{A}$ and zero otherwise.

Importance Sampling

Defining an *easy-to-sample-from* density $\pi(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta} \in \Theta$:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathbb{E}_{\pi}\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right],$$