ELEG 5633 Detection and Estimation
Review of Probability Theory

Jingxian Wu

Department of Electrical Engineering
University of Arkansas

January 19, 2017

Review of Probability Theory

- Basic probabilities
  - Probability measures
  - Conditional probability
  - Independence
  - Bayes rule
- Random variables
- Expectation
- Pairs of random variables
- Random vectors
- Gaussian random variables and random vectors
- Convergence of sums of independent random variables

Probability theory begins with three basic components:

- The set of all possible outcomes, denoted Ω.
- The collection of all sets of outcomes (events), denoted $\mathcal{A}$.
- A probability measure $\mathbb{P}$.

Specification of the triple $(\Omega, \mathcal{A}, \mathbb{P})$ defines the probability space which models a real-world measurement or experimental process.

Example

$$\begin{aligned}
\Omega &= \{\text{all outcomes of the roll of a die}\} \\
&= \{1, 2, 3, 4, 5, 6\} \\
\mathcal{A} &= \{\text{all possible sets of outcomes}\} \\
&= \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, \ldots, 6\}\} \\
\mathbb{P} &= \text{probability of all sets/events}
\end{aligned}$$

Assume all six outcomes are equally probable.
What is the probability of a given $\omega \in \Omega$, say $\omega = 3$?
What is the probability of the event $\omega \in \{1, 2, 3\}$?

Probability Measures

Probability measures must satisfy the following properties:

1. $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{A}$.
2. $\mathbb{P}(\Omega) = 1$
3. if $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Show that the last condition also implies that in general $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ This is a useful inequality sometimes called the union bound.

Conditional Probability

Consider two events $A, B \in \mathcal{A}$. The (conditional) probability that A occurs given B occurs is

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Example

$\Omega = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 2\}$, $B = \{2, 3\}$. Now suppose you are told that B occurs. What is the conditional probability that A occurs?

Independence

Two events A and B are said to be independent if $\mathbb{P}(A|B) = P(A)$.
In other words, B provides no information about whether A has occurred.

Example

Suppose we have two dice. Then
$\Omega = \{$all pairs of outcomes of the roll of two dice$\}$. Let $A = \{$1st die is 1$\}$ and $B = \{$2nd die is 1$\}$. $P(A) =?$, $P(A|B) =?$

Independence

Example

$\Omega = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 2\}$, $B = \{2, 3\}$. A and B are independent or not?

Bayes Rule

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

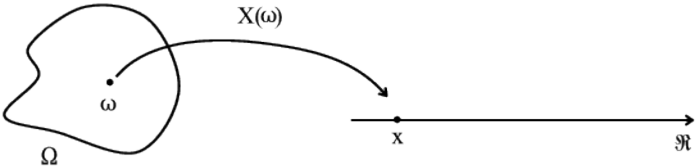
- Bayes rule is a formula for the "inverse" conditional probability, $\mathbb{P}(B|A)$.
- It is easy to verify by recalling the definition of conditional probability.
- This inversion formula will play a central role in signal estimation problems later in the course.

Example

Geneticists have determined that a particular genetic defect is related to a certain disease. Many people with the disease also have this defect, but there are disease-free people with the defect. The geneticists have found that 0.01% of the general population has the disease and that the defect is present in 50% of these cases. They also know that 0.1% of the population has the defect. What is the probability that a person with the defect has the disease?

Random Variables

- A real-valued random variable is a mapping $X : \Omega \to \mathbb{R}$.
- Random variables can also be vectors, e.g., a mapping $X : \Omega \to \mathbb{R}^n$.
- Since $\mathbb{P}$ specifies the probability of every subset of Ω, it also induces probabilities on events expressed in terms of X.
  - E.g., X is a real-valued scalar random variable, then $\{X \geq 0\} \equiv \{X(\omega) \geq 0\}$. Therefore, $\mathbb{P}(X \geq 0) = \mathbb{P}(\{\omega : x(\omega) \geq 0\})$.
  - More generally, for any set A we may consider the event $\{X \in A\}$ and its probability $\mathbb{P}(X \in A)$.



Example

Consider a dice gambling game. You are betting that the next roll will be greater than 3. If the outcome ω is greater than 3 ($\omega = 4, 5$, or 6), you win ω dollars; otherwise, you loose ω dollars. Define a random variable that corresponds to the amount of money that you win (positive) or lose (negative).

Cumulative Distributions and Densities

- Cumulative distribution function (cdf) of a real-valued RV X:

$$F_X(x) := \mathbb{P}(X \leq x)$$

  - $F_X(x_2) - F_X(x_1) = \mathbb{P}(X \leq x_2) - \mathbb{P}(X \leq x_1) = \mathbb{P}(x_1 < X \leq x_2)$
- Probability density function (pdf) of a continuous random variable X:

$$p_X(x) := \lim_{\Delta \to 0} \frac{F_X(x + \Delta) - F_X(x)}{\Delta}$$

  if the limit exists (F is differentiable at x).
  - $p_X(x) \geq 0$: $F_X(x)$ is a monotonic increasing function.
  - If $F_X(x)$ is differentiable everywhere, then by the Fundamental Theorem of Calculus we have

$$F_X(x) = \int_{-\infty}^{x} p_X(x) dx$$

  - $\lim_{x \to \infty} F_X(x) = \int_{-\infty}^{\infty} p_X(x) dx = 1$
  - $\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x) dx$, which explains the term "density".

Probability Mass Functions (pmf) for Discrete RVs

If X takes values in a discrete set $\{x_1, x_2, \ldots\}$ (which may be finite or infinite), then the pmf of X is given by

$$\mathbb{P}(X = x_i), i = 1, 2, \ldots$$

Example

Suppose you toss a coin n times and count the number of heads. Assume the probability of observing a head in a single toss is p. This number is a random variable X taking a value between 0 and n. What is the pmf of X?

Expectation

If a random variable X has density $p_X(x)$, then the expectation of $f(X)$, where f is any function of X, is

$$E[f(X)] = \int f(x) p_X(x) dx$$

If the random variable is discrete, then the expectation is

$$E[f(X)] = \sum_i f(x_i) P(X = x_i)$$

Some special cases:

- mean: $\mu = \mathbb{E}[X]$
- variance: $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- probability: $\mathbb{P}(X \in A) = \mathbb{E}[\mathbf{1}_{X \in A}]$
- characteristic function: $\phi(\omega) := \mathbb{E}[\exp(-i\omega X)]$.
  The characteristic function of X is the Fourier transform of its density.

Pairs of Random Variables

- Cumulative distribution function (cdf) of a pair of real-valued RVs X, Y:

$$F_{XY}(x, y) := \mathbb{P}(X \leq x, Y \leq y)$$

- Joint probability density function (pdf):

$$p_{XY}(x, y) := \frac{d^2 F_{XY}(x, y)}{dxdy}$$

Pairs of Random Variables

- Conditional pdf:

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

- Marginal pdf:

$$p_X(x) = \int p_{XY}(x,y)dx$$

,

$$p_Y(y) = \int p_{XY}(x,y)dy$$

- Expectation:

$$E[f(X,Y)] = \int f(x,y)p_{XY}(x,y)dxdy$$

Correlation, Covariance, Dependence

- Covariance: $\sigma_{XY} := \mathbb{E}[(X - \mu_x)(Y - \mu_y)] = \mathbb{E}[XY] - \mu_x \mu_y$
- Correlation coefficient: $\rho_{XY} := \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- Conditional expectation: $\mathbb{E}[X|Y = y] = \int x \cdot p_{X|Y}(x|y)dx$
  Note: it is a function of y! $\mathbb{E}[X|Y]$ is a random variable.
- X, Y are statistically independent if $p_{X|Y}(x|y) = p_X(x)$, or $p_{XY}(x, y) = p_X(x)p_Y(y)$.
- Question: X, Y independent $\Leftrightarrow$ X, Y uncorrelated?

Example

Consider two RVs X, Y, where $Y = aX + b$, a, b are constants. What is ρ_{XY}?

Random Vectors

- If X is a d-dimensional random variable, $X = [X_1, ..., X_d]^T$, then
  - mean $\mathbb{E}[X] = [\mu_1, \mu_2, \ldots, \mu_d]^T$.
  - covariance matrix is $\Sigma = \mathbb{E}[(X - \mu)(X - \mu)^T]$.
    Covariance matrices are always symmetric, and positive semi-definite.
- Two random vectors X, Y:
  - cross-covariance matrix $\Sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T]$
  - conditional expectation $\mathbb{E}[f(x)|Y = y] = \int f(x) \cdot p_{X|Y}(x|y)dx$
  - conditional mean $\mathbb{E}[X|Y = y] = \int x \cdot p_{X|Y}(x|y)dx$
  - conditional covariance $\Sigma_{X|Y=y} = \int (x - \mu_{X|Y=y})(x - \mu_{X|Y=y})^T p_{X|Y}(x|y)dx$
- If $Y = \mathbf{A}X$, where $\mathbf{A}$ is an $m \times d$ matrix, then the random vector Y has
  - mean $\mathbb{E}[\mathbf{A}X] = \mathbf{A}\mu$
  - covariance $\mathbb{E}[(\mathbf{A}X - \mathbf{A}\mu)(\mathbf{A}X - \mathbf{A}\mu)^T)] = \mathbf{A}\Sigma\mathbf{A}^T$.

Gaussian Random Variables and Random Vectors

- Gaussian (Normal) distribution:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Shorthand: $X \sim \mathcal{N}(\mu, \sigma^2)$. Standard normal distribution: $\mathcal{N}(0, 1)$.
- CDF for $\mathcal{N}(0, 1)$: $\Phi(x)$. $Q(x) = 1 - \Phi(x)$.
- Multivariate Gaussian (or Normal) distribution

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

- Shorthand: $X \sim \mathcal{N}(\mu, \Sigma)$. $\mu = \mathbb{E}[X]$, $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$.
- If Σ is a diagonal matrix, component random variables are uncorrelated. The multivariate Gaussian density factorizes into univariate component densities, which means that uncorrelated Gaussian random variables are also independent.
- This is a special property for Gaussian!

More Properties about Gaussian Random Vectors

- If X is multivariate Gaussian distributed, $Y = \mathbf{A}X$, $Y \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)$.
- Each individual variate X_i of X is Gaussian distributed, with variance $\sigma_i^2 = \Sigma_{ii}$
- X, Y are joint Gaussian random vectors. Conditional pdf of X given $Y = y$ is still Gaussian, with

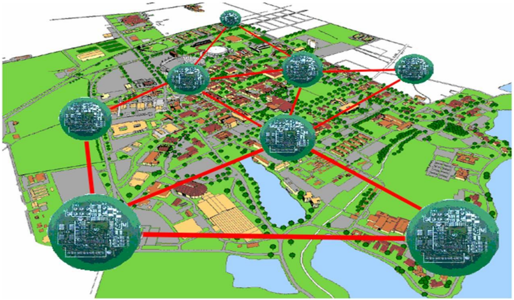
$$\mu_{X|Y}(y) = \mu_X + \Sigma_{XY}\Sigma_Y^{-1}(y - \mu_y)$$
$$\Sigma_{X|Y}(y) = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T$$

Example

Suppose that $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$ What is the distribution of X_1, $2X_1 + X_2$, and X_1 given $X_2 = 2$?

Example

Imagine we have a sensor network monitoring a manufacturing facility. The network consists of n nodes and its function is to monitor for failures in the system. Let $X = [X_1, ..., X_n]^T$ denote the set of scalar measurements produced by the network and suppose that it can modeled as a realization of a $\mathcal{N}(0, \Sigma)$ random vector. Furthermore, assume that if the average of the measurements is greater than a threshold τ, it indicates that the system fails. Then, what is the probability of failure? What if $n = 2$, $\Sigma = I$, and $\tau = 0$?



Convergence of Sums of Independent Random Variables

Example

A data analyst wants to find out the average number of retweets for a tweet at Twitter. Assume tweets are independent with each other. She randomly picks n tweets, and tracks them until no one retweets them. Let $x_1, \ldots, x_n$ denote the recorded number of times that a tweet has been retweeted. She then use the average $\hat{p} := \frac{1}{n}\sum_{i=1}^{n} x_i$ as the estimator.
Question: is the estimator reasonably accurate?

To answer it, we need to understand the behavior of sums of independent RVs!

Let $X_1, \ldots, X_n$ be i.i.d. RVs with mean μ and variance $\sigma^2 < \infty$. Let $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$.

- $\hat{\mu}$ is another RV with mean μ and variance σ^2/n.
- The variance is reduced by a factor of n!
- Lower variance means less uncertainty, i.e., we can reduce the uncertainty by averaging. The more we average, the less the uncertainty.

Can we say more about the distribution of $\hat{\mu}$?

Theorem (Central Limit Theorem)

Let $X_1, X_2, ..., X_n$ be independent random variables with means μ and variances $\sigma^2 < \infty$. Then $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}(\mu, \sigma^2/n)$ in distribution as $n \to \infty$.

- The conditions are sufficient but not necessary.
- The distribution of $\hat{\mu}$ tends to be Gaussian quickly regardless of the form of the distribution of X_i!
- It captures the i.i.d case.