Introduction to Probability

Anima Anandkumar
Electrical Engineering and Computer Engineering
University of California, Irvine, CA 92697
a.anandkumar@uci.edu

Outline

Concepts

- Probability space
- Conditional probability and statistical independence.
- Random variables, distributions and densities.
- Expectations and conditional expectations.
- Real and complex Gaussian variables and vectors.
- Inequalities
- Convergence, LLN and CLT.

The Probability Space: Definition

Definition:

A probability space is defined by $(\Omega, \mathcal{F}, \Pr)$

1. Ω is the sample space that contains the set of outcomes.
2. $\mathcal{F}$ is a σ-field of subsets of Ω (events):

   (i) $\Omega \in \mathcal{F}$. (ii) If $\mathcal{E} \in \mathcal{F}$, then $\mathcal{E}^c \in \mathcal{F}$.

   (iii) If $\mathcal{E}_i \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} \mathcal{E}_i \in \mathcal{F}$.
3. $\Pr$ is a function on $\mathcal{F}$ satisfying

   (i) $0 \leq \Pr(\mathcal{E}) \leq 1$. (ii) $P(\Omega) = 1$.

   (iii) If $\mathcal{E}_1, \mathcal{E}_2, \cdots$ are disjoint, then
   $\Pr(\bigcup_{i=1}^{\infty} \mathcal{E}_i) = \sum \Pr(\mathcal{E}_i)$

Why Do We Need Restrictions on Events?

Let $\Omega \triangleq \{(x, y) | x^2 + y^2 = 1\}$. There exists[†] a set $\mathcal{E} \in \Omega$ such that

1. for any rational $\phi, \theta \in [0, 2\pi)$ and $\phi \neq \theta$, the rotation of $\mathcal{E}$ by θ and ϕ are disjoint, *i.e.*, $\mathcal{E}(\theta) \bigcap \mathcal{E}(\phi) = \emptyset$.
2. The union of all $\mathcal{E}$ rotated by rational θ is Ω.

If $\Pr(\mathcal{E}) = x$, then

$$1 = \Pr(\Omega) = \Pr(\bigcup \mathcal{E}(\theta)) = \sum \Pr(\mathcal{E}(\theta)) = \sum x$$

[†] M. Capiński and P. Knopp, *Measure, Integral and Probability*, Springer, 1999.

The Probability Space: Examples

Sample Space Ω

- Picking the "lucky" person out of a class of 30 to receive an A: $\Omega_1 = \{1, 2, \cdots, 29, 30\}$.
- Taking the qualify exam until pass: $\Omega_2 = \{P, FP, FFP, FFFP, \cdots, \}$.
- The time you wake up: $\Omega_3 = \{(00:00, 24:00]\}$
- Throwing a dart to a unit disk: $\Omega_4 = \{(x, y) | x^2 + y^2 \leq 1\}$.

Events

Consider Ω_1

$\mathcal{E}_0$: Someone is lucky: $\mathcal{E}_0 = \Omega_1$.

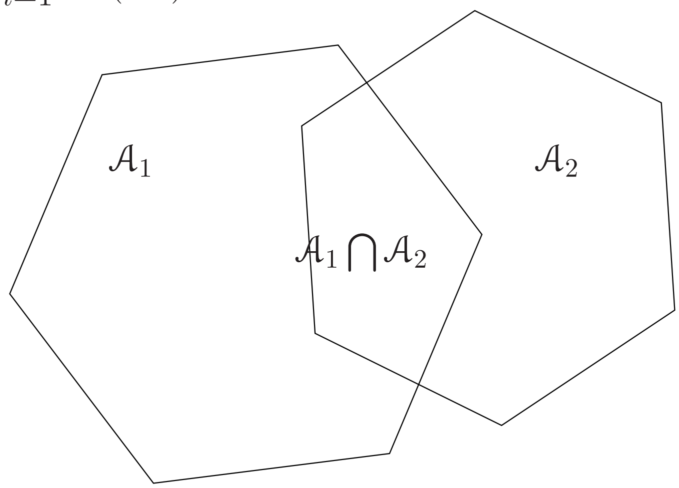
$\mathcal{E}_1$: the "lucky" person has an even ID: $\mathcal{E}_1 = \{2, 4, 6, \cdots, 30\}$.

$\mathcal{E}_2$ The "lucky" person has an even number or a number between 10 and 20. $\mathcal{E}_2 = \mathcal{E}_1 \bigcup \{11, 13, \cdots, 19\}$.

$\mathcal{E}_4$ The "lucky" person has an odd number less than 10. $\mathcal{E}_4 = \mathcal{E}_1^c \bigcap \{1, \cdots, 10\}$.

Elementary Properties

- $\Pr(\mathcal{A}^c) = 1 - \Pr(\mathcal{A}), \quad \Pr(\emptyset) = 0$.
- If $\mathcal{A} \subset \mathcal{B}$, then $\Pr(\mathcal{B}) = \Pr(\mathcal{A}) + \Pr(\mathcal{B} - \mathcal{A}) \geq \Pr(\mathcal{A})$.
- Union bound (Boole's inequality):
  $\Pr(\bigcup_{i=1}^{\infty} \mathcal{A}_i) \leq \sum_{i=1}^{\infty} \Pr(\mathcal{A}_i)$



- Inclusion-exclusion:

$$\begin{aligned}
\Pr(\mathcal{A}_1 \bigcup \mathcal{A}_2) &= \Pr(\mathcal{A}_1) + \Pr(\mathcal{A}_2) - \Pr(\mathcal{A}_1 \bigcap \mathcal{A}_2) \\
\Pr(\bigcup_{i=1}^{n} \mathcal{A}_i) &= \sum_{i=1}^{n} \Pr(\mathcal{A}_i) - \sum_{i<j} \Pr(\mathcal{A}_i \bigcap \mathcal{A}_j) \\
&\quad + \sum_{i<j<k} \Pr(\mathcal{A}_i \bigcap \mathcal{A}_j \bigcap \mathcal{A}_k) - \cdots \\
&\quad + (-1)^{k+1} \sum_{i_1<i_2<\cdots<i_k} \Pr(\bigcap_{r=1}^{k} \mathcal{A}_{i_r}) + \cdots
\end{aligned}$$

- Bonferroni's inequality: $\Pr(\bigcap_{i=1}^{n} \mathcal{A}_i) \geq 1 - \sum_{i=1}^{n} \Pr(\mathcal{A}_i^c)$

Sequence of Events

Monotone Convergence

If $\mathcal{E}_i$ increases, *i.e.*, $\mathcal{E}_1 \subseteq \mathcal{E}_2 \subseteq \cdots$, and let $\mathcal{E} \triangleq \bigcup_{i=1}^{\infty} \mathcal{E}_i$. Then

$$\Pr(\mathcal{E}) = \lim_{i \to \infty} \Pr(\mathcal{E}_i)$$

If $\mathcal{E}_i$ decreases, *i.e.*, $\mathcal{E}_1 \supseteq \mathcal{E}_2 \supseteq \cdots$, and let $\mathcal{E} = \bigcap_{i=1}^{\infty} \mathcal{E}_i$. Then

$$\Pr(\mathcal{E}) = \lim_{i \to \infty} \Pr(\mathcal{E}_i)$$

Limits of Sequences

Let $\{\mathcal{E}_n\}$ be an arbitrary sequence of events. Define limits

$$\mathcal{E}^* = \limsup_{i \to \infty} \mathcal{E}_i \triangleq \bigcap_{i=1}^{\infty} \bigcup_{n=i}^{\infty} \mathcal{E}_n, \quad \mathcal{E}_* = \liminf_{i \to \infty} \mathcal{E}_i \triangleq \bigcup_{i=1}^{\infty} \bigcap_{n=i}^{\infty} \mathcal{E}_n$$

Then $\mathcal{E}^*$ is the event that infinitely many of $\{\mathcal{E}_n\}$ occur and $\mathcal{E}^*$ is the event that all except a finite number of $\mathcal{E}_i$ occur, *i.e.*,

$$\begin{aligned} \mathcal{E}^* &= \{\omega \in \Omega : \omega \in \mathcal{E}_i, \text{for infinitely many values of } i\}, \\ \mathcal{E}_* &= \{\omega \in \Omega : \omega \in \mathcal{E}_i, \text{for all but finite many of } i\}, \end{aligned}$$

Now if we know $\Pr(\mathcal{E}_n)$, what can we say about $\Pr(\mathcal{E}^*)$?

Borel-Cantelli Lemmas

1. If $\sum \Pr(\mathcal{E}_i) < \infty$, then $\Pr(\mathcal{E}^*) = 0$.
2. If $\sum \Pr(\mathcal{E}_i)$ diverges, and $\{\mathcal{E}_n\}$ are independent, then $\Pr(\mathcal{E}^*) = 1$.

Example: Passing the Qualify

Consider the random experiment: taking the Qualify exam. The probability model is given by $(\Omega, \mathcal{F}, P)$ where

- the sample space $\Omega_2 = \{P, FP, FFP, FFFP, \cdots,\}$;
- the σ-field $\mathcal{F}$ includes all subsets of Ω_2, *i.e.*, $\mathcal{F} = 2^{\Omega}$.
- If the probability of passing is p, and assume that you learned nothing from the last time, then

$$\Pr(\underbrace{FF\cdots F}_{k}P) = (1-p)^k p$$

Q: What is the probability that you will pass in no more than three tries?

$$\mathcal{E} = \{P, FP, FFP\}, \quad \Pr(\mathcal{E}) = p + (1-p)p + (1-p)p^2$$

Q: What is the probability that you pass eventually?

Let $\mathcal{E}_i$ be the event that you pass in no more than i tries. Then $\mathcal{E}_i^c$ is the event that you have not succeeded after i tries.

$$\Pr(\mathcal{E}_i) = 1 - \Pr(\mathcal{E}_i^c) = 1 - (1-p)^i$$

The event of pass eventually is given by

$$\mathcal{E} = \bigcup_{i=1}^{\infty} \mathcal{E}_i, \quad \Pr(\mathcal{E}) = \lim_{i\to\infty} \Pr(\mathcal{E}_i) = 1$$

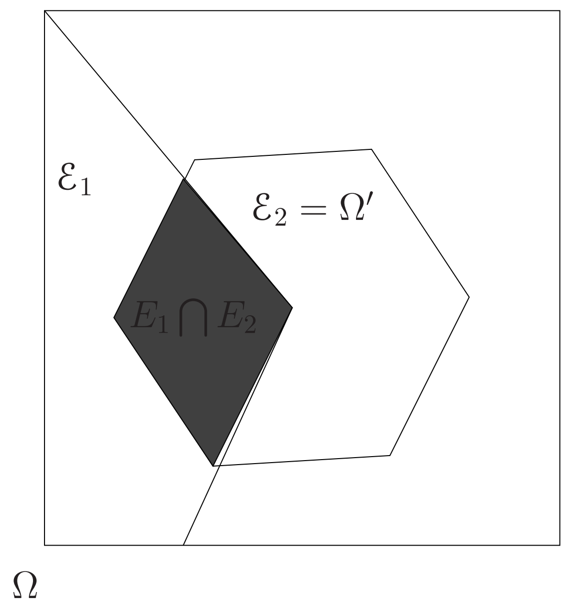
Q: What if your chance of passing increases with the number of tries, you would expect to do better, and $\Pr(\mathcal{E}) = 1$. How about your chance actually decreases with the number of tries?

Conditional Probability

Definition

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two events. Assuming that $\Pr(\mathcal{E}_2) \neq 0$, the conditional probability of the event $\mathcal{E}_1$ given that $\mathcal{E}_2$ has already occurred is given by

$$\Pr(\mathcal{E}_1|\mathcal{E}_2) = \frac{\Pr(\mathcal{E}_1 \bigcap \mathcal{E}_2)}{\Pr(\mathcal{E}_2)}$$



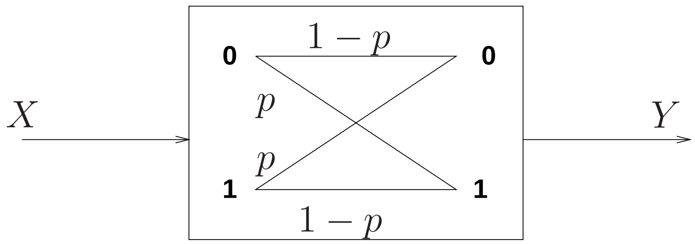
We can think "conditioning" as generating a new probability model (based on the observation of event $\mathcal{E}_2$) from the old by treating $\mathcal{E}_2$ as the new sample space Ω'

Example: Binary Symmetrical Channel

The Channel

The binary symmetric channel (BSC) is defined by the conditional probability

$$\begin{aligned} \Pr(Y=0|X=0) &= \Pr(Y=1|X=1) = 1-p, \\ \Pr(Y=1|X=0) &= \Pr(Y=0|X=1) = p \end{aligned}$$



The Sample Space

$$\Omega = \{(X=x, Y=y), x, y, \in \{0,1\}\} = \{(0,0),(0,1),(1,0),(1,1)\}.$$

The σ-field

$$\mathcal{F} = \{\emptyset, \Omega, \{(0,0)\}, \cdots, \{(1,1)\}, \{(0,0)\} \bigcup \{(0,1)\} \cdots\}$$

The Probability Measure

Suppose that $\{X=0\}$ and $\{X=1\}$ are equally likely.

$$\begin{aligned} \Pr[\{(0,0)\}] &= \Pr(X=0)\Pr(Y=0|X=0) = \frac{1-p}{2}, \\ \Pr[\{(1,1)\}] &= \Pr(X=1)\Pr(Y=1|X=1) = \frac{1-p}{2} \\ \Pr[\{(1,0)\}] &= \Pr(X=0)\Pr(Y=1|X=0) = \frac{p}{2}, \\ \Pr[\{(0,1)\}] &= \Pr(X=1)\Pr(Y=0|X=1) = \frac{p}{2} \end{aligned}$$

Total Probability Theorem

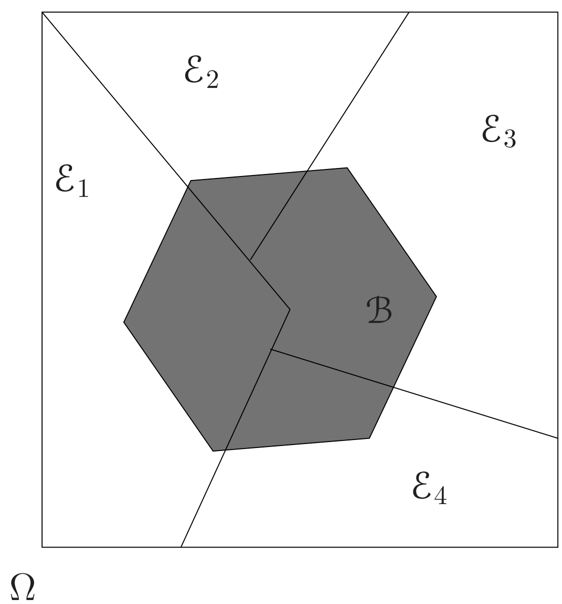
Total Probability Theorem

If $\{\mathcal{E}_i\}$ partition Ω, *i.e.*,

$$\bigcup \mathcal{E}_i = \Omega, \quad \mathcal{E}_i \bigcap \mathcal{E}_j = \emptyset,$$

then

$$\Pr(\mathcal{B}) = \sum \Pr(\mathcal{E}_i) \Pr(\mathcal{B}|\mathcal{E}_i)$$



The Bayes Formula

$$\Pr(\mathcal{E}_i|\mathcal{B}) = \frac{\Pr(\mathcal{B}|\mathcal{E}_i) \Pr(\mathcal{E}_i)}{\sum \Pr(\mathcal{E}_i) \Pr(\mathcal{B}|\mathcal{E}_i)}$$

Statistical Independence

Definition

Two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are statistically independent if

$$\Pr(\mathcal{E}_1 \bigcap \mathcal{E}_2) = \Pr(\mathcal{E}_1)\Pr(\mathcal{E}_2),$$

which implies that

$$\Pr(\mathcal{E}_1|\mathcal{E}_2) = \Pr(\mathcal{E}_1), \qquad \Pr(\mathcal{E}_2|\mathcal{E}_1) = \Pr(\mathcal{E}_2)$$

Events $\{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\}$ are statistically independent if

$$\begin{aligned}
\Pr(\mathcal{E}_1 \bigcap \mathcal{E}_2) &= \Pr(\mathcal{E}_1)\Pr(\mathcal{E}_2) \\
\Pr(\mathcal{E}_1 \bigcap \mathcal{E}_3) &= \Pr(\mathcal{E}_1)\Pr(\mathcal{E}_3) \\
\Pr(\mathcal{E}_2 \bigcap \mathcal{E}_3) &= \Pr(\mathcal{E}_2)\Pr(\mathcal{E}_3) \\
\Pr(\mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_3) &= \Pr(\mathcal{E}_1)\Pr(\mathcal{E}_2)\Pr(\mathcal{E}_3)
\end{aligned}$$

In general, events $\{\mathcal{E}_1, \cdots, \mathcal{E}_n\}$ are statistically independent if

$$\Pr(\mathcal{E}_{i_1} \bigcap \mathcal{E}_{i_2} \bigcap \cdots \bigcap \mathcal{E}_{i_k}) = \Pr(\mathcal{E}_{i_1})\Pr(\mathcal{E}_{i_2})\cdots\Pr(\mathcal{E}_{i_k})$$

for all $\{i_1, \cdots, i_k\} \subset \{1, \cdots, n\}$.

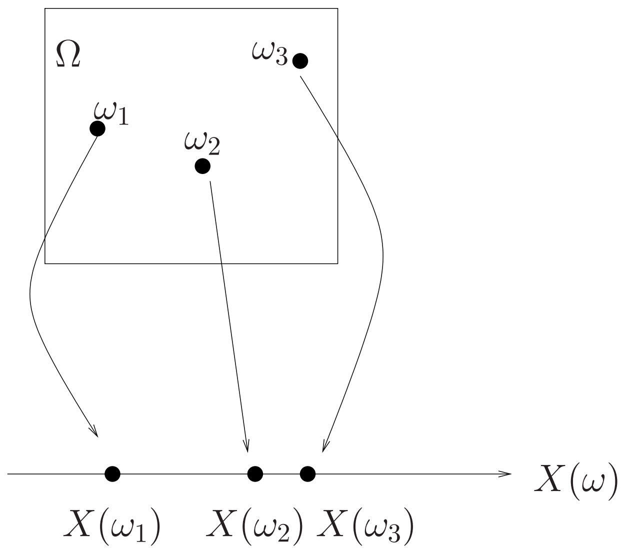
Random Variables

Definition

Given any probability space $(\Omega, \mathcal{F}, \Pr)$, a random variable is a function

$$X : \Omega \to R$$

such that, for all x, $\{\omega \in \Omega : X(\omega) \le x\} \in \mathcal{F}$.



Notations

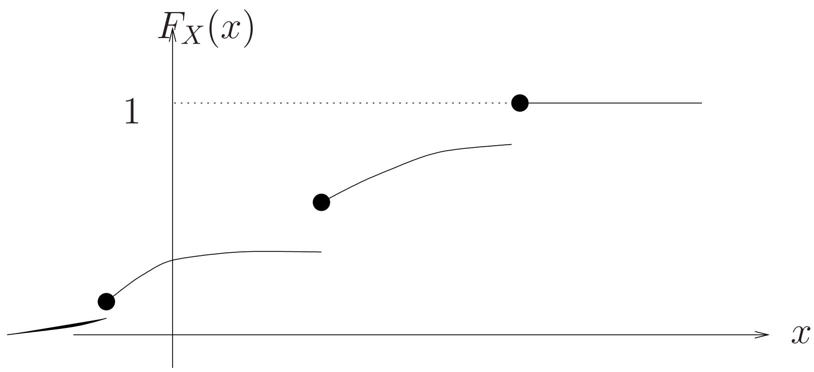
We use capital letters to indicate random variables and their corresponding small letters to indicate their "realizations" ‡. For example, in $X = x$, X is the random variable (a function) and x is the value that X takes (with some probability).

‡We may use small letters to denote random variables when there is no confusion

Cumulative Distribution Function

The cumulative distribution function (CDF) of a random variable X is

$$F_X(x) \triangleq \Pr(X \leq x)$$



Properties

1. $F_X(-\infty) = 0, F_X(\infty) = 1$.
2. If $x < y$, then $F_X(x) \leq F_X(y)$.
3. $F(\cdot)$ is right continuous, *i.e.*, $\lim_{\Delta \to 0^+} F_X(x + \Delta) = F_X(x)$
4. $\Pr(x < X \leq y) = F_X(y) - F_X(x)$.
5. A useful interpretation is

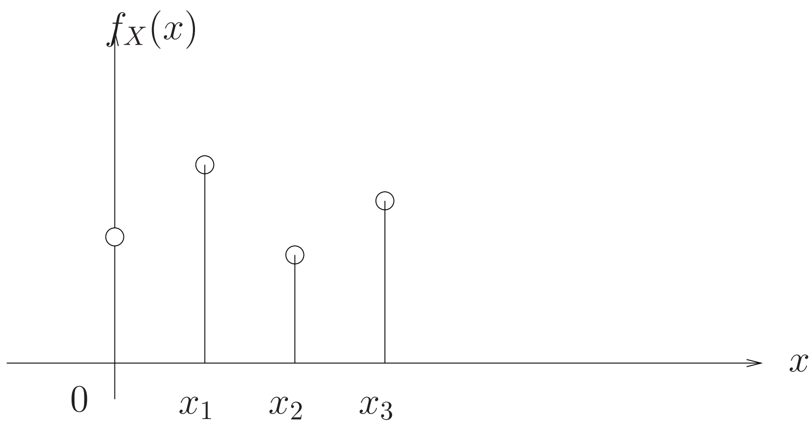
$$\begin{aligned}
\Pr(X \in (x, x + dx)) &= F_X(x + dx) - F_X(x) \triangleq dF_X(x) \\
\Pr(X \in \mathcal{A}) &= \int_{\mathcal{A}} dF_X(x)
\end{aligned}$$

6. $\Pr(X = x_0) = F_X(x_0) - \lim_{y \uparrow x_0} F_X(y)$.

Probability Mass Function

For discrete random variables, *i.e.*, X takes values in a countable set $\{x_i\}$. The probability mass function (PMF) of is given by

$$f_X(x) \triangleq \Pr(X = x)$$



The PMF is related to CDF by

$$F_X(x) = \sum_{u:u \leq x} f_X(u)$$

For any event $\mathcal{E}$, we have

$$\Pr(\mathcal{E}) = \sum_{u \in \mathcal{E}} f_X(u)$$

To unify notations, we also write the above as

$$\Pr(\mathcal{E}) = \int_{\mathcal{E}} f_X(x)dx = \int_{\mathcal{E}} dF_X(x)$$

Probability Density Function

A random variable is continuous if its distribution function can be expressed as

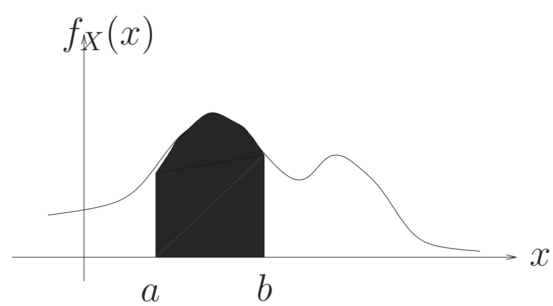
$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{1}$$

for some integrable function $f_X : \mathcal{R} \to [0, \infty)$. Function $f_X(x)$ is the probability density function (pdf) of X:

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Properties:

- $f_X(u) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(u)du = 1$.
- $\int_a^b f_X(u)du = \Pr(a < X \leq b)$.
- $\Pr(\mathcal{E}) = \int_{\mathcal{E}} f_X(u)du$.



Random Vectors

Given a random vector $\mathbf{X} = [X_1, \cdots, X_n]$ defined on the probability space $(\Omega, \mathcal{F}, P)$,

- the joint density distribution function is given by

$$F_{\mathbf{X}}(\mathbf{x}) = \Pr(\mathbf{X} \leq \mathbf{x}) \overset{\Delta}{=} \Pr(X_1 \leq x_1, \cdots, X_n \leq x_n).$$

- The joint density function is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{\mathbf{X}}(\mathbf{x})$$

- The marginal distribution of X_i is given by

$$F_{X_i}(x) \overset{\Delta}{=} \Pr(X_i < x) = F_{\mathbf{X}}(\infty, \cdots, \infty, \underbrace{x}_{ith}, \infty, \cdots, \infty)$$

- The marginal density is given by

$$f_{X_i}(x) = \frac{d}{dx} F_{X_i}(x) = \int f_{\mathbf{X}}(\mathbf{x}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

Independent Random Variables

Recall Independent Events

- $\mathcal{A}$ and $\mathcal{B}$ are statistically independent if

$$\Pr(\mathcal{A} \bigcap \mathcal{B}) = \Pr(A) \Pr(B)$$

- Events $\{A, B, C\}$ are statistically independent if

$$\begin{aligned} \Pr(\mathcal{A} \bigcap \mathcal{B}) &= \Pr(\mathcal{A}) P(\mathcal{B}) \\ \Pr(\mathcal{A} \bigcap \mathcal{C}) &= \Pr(\mathcal{A}) \Pr(\mathcal{C}) \\ \Pr(\mathcal{C} \bigcap \mathcal{B}) &= \Pr(\mathcal{C}) \Pr(\mathcal{B}) \\ \Pr(\mathcal{A} \bigcap \mathcal{B} \bigcap \mathcal{C}) &= \Pr(\mathcal{A}) \Pr(\mathcal{B}) \Pr(\mathcal{C}) \end{aligned}$$

Independent Random Variables

We call n random variables $\mathbf{X} = (X_1, \cdots, X_n)$ statistically independent if

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

or equivalently

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

Conditioning on Random Variables

Conditional Distribution

Consider random variables X and Y with joint distribution (or density) function $F_{X,Y}(x,y)$ ($f_{X,Y}(x,y)$). The conditional distribution of X given $Y = y$ is defined as

$$F_{X|Y}(x|y) \triangleq \Pr(X \leq x | Y = y) = \lim_{\epsilon \downarrow 0} \frac{\Pr(X \leq x, y < Y \leq y + \epsilon)}{\Pr(y < Y \leq y + \epsilon)}$$

The conditional density function of $F_{X|Y}$, written as $f_{X|Y}$, is given by

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & f_Y(y) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $f_Y(y) = \int f_{X,Y}(u,y)du$ is the marginal pdf of Y. Further,

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(u|y)du$$

If X and Y are independent, $f_{X|Y}(x|y) = f_X(x)$.

Example: Consider independent random variables X and N such that

$$Y = X + N,$$

where X is discrete with PMF $f_X(x)$ and N is continuous with PDF $f_N(n)$. Then

$$\begin{aligned} F_{Y|X}(y|x) &= \Pr(Y \leq y | X = x) = \frac{\Pr(N \leq y - x, X = x)}{f_X(x)} = F_N(y - x) \\ F_{X|y}(x|y) &= \Pr(X = x | Y = y) = \lim_{\epsilon \downarrow 0} \frac{\Pr(X = x, y < Y \leq y + \epsilon)}{\Pr(y < Y \leq y + \epsilon)} = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \\ f_{Y|X}(y|x) &= f_N(y - x) \end{aligned}$$

Expectation of Random Variables

Definition

For a random variable X

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x dF_X(x), \quad \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) dF_X(x)$$

Properties

1. The indicator function of an event $\mathcal{E}$ is defined as

   $$1_{\mathcal{E}}(x) = \begin{cases} 1 & x \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

   We then have

   $$\Pr(\mathcal{E}) = \int_{\mathcal{E}} dF_X(x) = \mathbb{E}(1_{\mathcal{E}}(X))$$

2. If X is nonnegative random variable with CDF F,

   $$\mathbb{E}(X) = \int_{0}^{\infty} (1 - F_X(x)) dx$$

3. Linearity: $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y)$.

4. If X and Y are independent, then $\mathbb{E}(h(X)g(Y)) = \mathbb{E}(h(X))\mathbb{E}(g(Y))$.

5. Variance and Covariance

   $$\begin{aligned} \text{Var}(X) &\triangleq \mathbb{E}(X - \mathbb{E}(X))^2, \\ \text{Cov}(X, Y) &\triangleq \mathbb{E}(\mathbb{E}(X - \mathbb{E}(X))\mathbb{E}(Y - \mathbb{E}(Y))). \end{aligned}$$

   The standard deviation of X is $\sqrt{\text{Var}(X)}$.

6. X and Y are uncorrelated if Cov$(X, Y) = 0$.

7. For a real random vector $\mathbf{X} = [X_1, \cdots, X_n]^T$,

$$\text{Mean:} \quad \mathbb{E}(\mathbf{X}) = [\mathbb{E}(X_1), \cdots, \mathbb{E}(X_n)]^T$$
$$\text{Covariance:} \quad \text{Cov}(\mathbf{X}, \mathbf{X}) = \mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T$$

- Cov$(\mathbf{X}, \mathbf{X})$ is always positive (semi) definite.
- If $\mathbf{X}$ is a vector of uncorrelated random variables, then Cov$(\mathbf{X}, \mathbf{X})$ is diagonal with variances as diagonal entries.

Conditional Expectation

The conditional expectation of $g(\mathbf{X})$ given $\mathbf{Y} = \mathbf{y}$ is given by

$$\mathbb{E}(g(\mathbf{X})|\mathbf{Y} = \mathbf{y}) = \int g(\mathbf{x}) f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) d\mathbf{x}$$

Note that $\mathbb{E}(g(\mathbf{X})|\mathbf{Y} = \mathbf{y})$ is a function of $\mathbf{y}$.

Conditional Mean as a Random Variable

- We denote $\mathbb{E}(g(\mathbf{X})|\mathbf{Y})$ as the random variable that takes the value $\mathbb{E}(g(\mathbf{X})|\mathbf{Y} = \mathbf{y})$ when $\mathbf{Y} = \mathbf{y}$.
- Successive conditioning:

$$\mathbb{E}(g(\mathbf{X})) = \mathbb{E}(\mathbb{E}(g(\mathbf{X})|\mathbf{Y}))$$

As an example, suppose that $Y \sim \mathcal{U}(0, 1)$ and $X \sim \mathcal{U}(0, Y)$.

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(\frac{Y}{2}) = \frac{1}{4} \\ \mathbb{E}(X^2) &= \mathbb{E}(\mathbb{E}(X^2|Y)) = \mathbb{E}(\frac{Y^2}{3}) = \frac{1}{9} \end{aligned}$$

Product Expectation Theorem

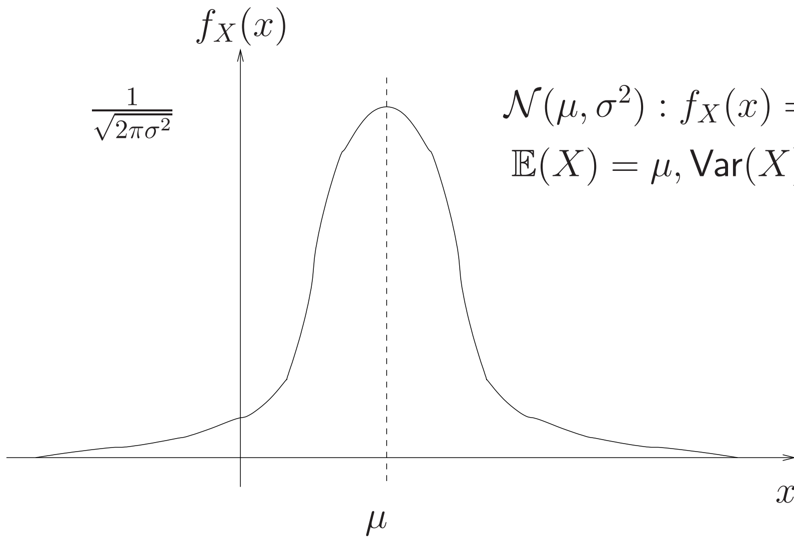
If $g(Y)$ is bounded and $\mathbb{E}(h(X)) \leq \infty$, then

$$\mathbb{E}(h(X)g(Y)) = \mathbb{E}(g(Y)\mathbb{E}(h(X)|Y))$$

A special case is when $g(y) = 1$ and $h(x) = x$

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

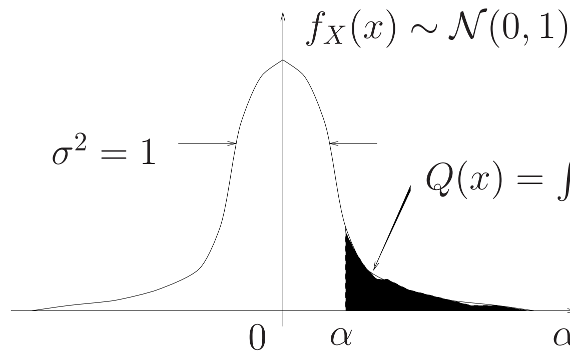
The Gaussian Random Variable



$\mathcal{N}(\mu, \sigma^2) : f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$
$\mathbb{E}(X) = \mu$, $\mathsf{Var}(X) = \sigma^2$.

The $Q(\cdot)$ function

$$Q(\alpha) \triangleq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\infty} e^{-\frac{u^2}{2}} du$$



Properties

1. Probability: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\Pr[X > \alpha] = Q(\frac{\alpha - \mu}{\sigma}), \quad \Pr(X < \alpha) = Q(\frac{\mu - \alpha}{\sigma})$$

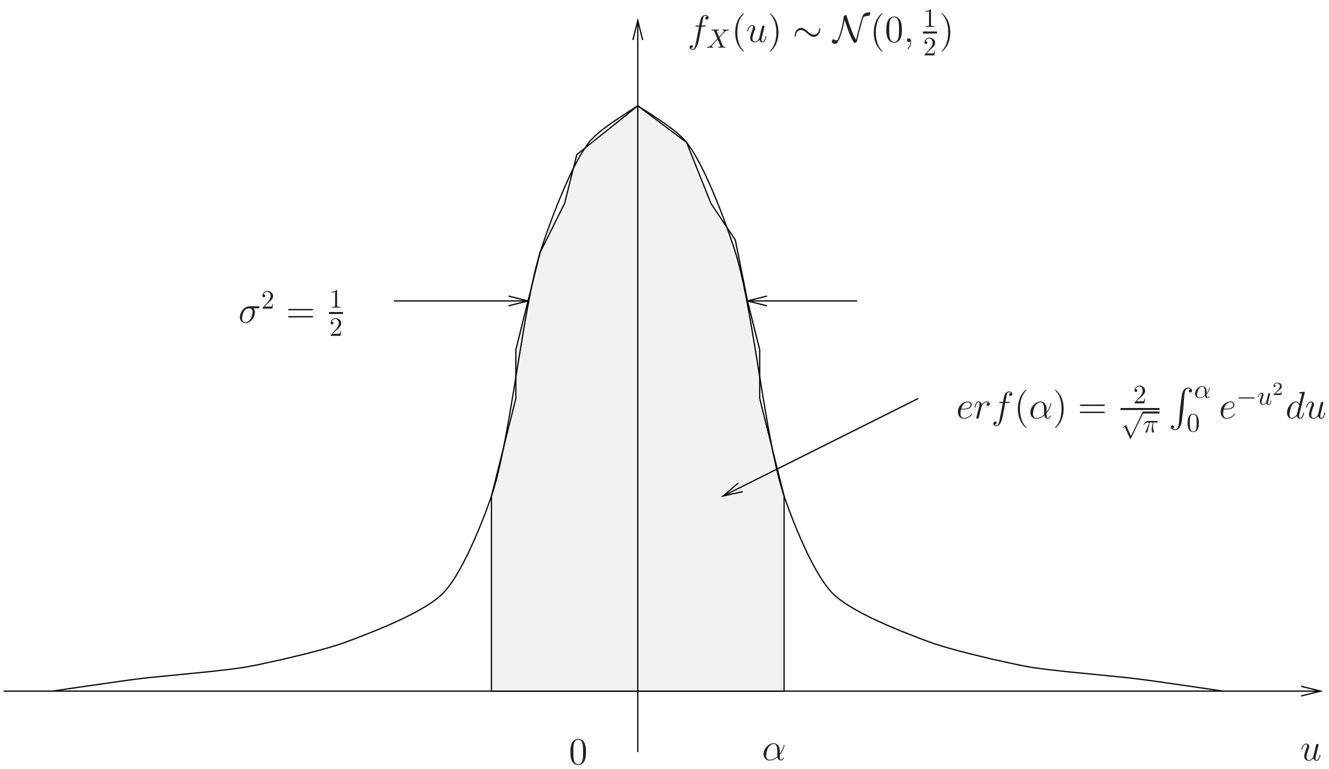
2. Bounds:

$$(1 - \frac{1}{x^2}) \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \leq Q(x) \leq \frac{1}{2} e^{-x^2/2}$$

$Q(\cdot)$, erf$(\cdot)$, and erfc$(\cdot)$

Definitions:

$$
\begin{aligned}
erf(\alpha) &\triangleq \frac{2}{\sqrt{\pi}} \int_0^{\alpha} e^{-u^2} du \\
erfc(\alpha) &\triangleq \frac{2}{\sqrt{\pi}} \int_{\alpha}^{\infty} e^{-u^2} du = 1 - erf(\alpha)
\end{aligned}
$$



Relations

$$
\begin{aligned}
Q(\alpha) &= \frac{1}{2} erfc(\frac{\alpha}{\sqrt{2}}) = \frac{1}{2}(1 - erf(\frac{\alpha}{\sqrt{2}})) \\
erfc(\alpha) &= 2Q(\sqrt{2}\alpha)
\end{aligned}
$$

Gaussian Random Vectors

A random vector $\mathbf{X} = [X_1, \cdots, X_n]^T$ is Gaussian if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma})}} exp\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\}$$

where

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$$

$$\begin{aligned} \mathbf{\Sigma} &= \mathbb{E}\{(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T\} \\ &= \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & & & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{pmatrix} \end{aligned}$$

- Random variables $X_1, \cdots, X_n$ are called jointly Gaussian.
- The Gaussian distribution is completely specified by the mean and the covariance.

Properties of Gaussian Random Vectors

Suppose that $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Jointly Gaussian implies marginally Gaussian. In particular,

$$X_i \sim \mathcal{N}(\mathbb{E}(X_i), \text{Cov}(X_i, X_i)).$$

  Any sub-vector of $\mathbf{X}$ is Gaussian. (The converse is not true in general!)

- For any matrix $\mathbf{A}$ and vector $\mathbf{b}$, $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ is Gaussian and

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t).$$

  Proof:

$$\begin{aligned} \mathbb{E}(\mathbf{Y}) &= \mathbf{A}\mathbb{E}(\mathbf{X}) + \mathbf{b} \\ \text{Cov}(\mathbf{Y}, \mathbf{Y}) &= \mathbb{E}(\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t \mathbf{A}^t) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t \end{aligned}$$

- Uncorrelated Gaussian random variables are independent.

- If

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}\right), \qquad (2)$$

  $f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z})$ is the complex Gaussian density with

$$\begin{aligned} \mathbb{E}(\mathbf{y}|\mathbf{z}) &= \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}(\mathbf{z} - \boldsymbol{\mu}_z) \\ \text{Cov}(\mathbf{y}, \mathbf{y}^H|\mathbf{z}) &= \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yz}\Sigma_{zz}^{-1}\boldsymbol{\Sigma}_{zy} \end{aligned}$$

Complex Random Vectors

Definition

The probability space of a complex random vector $\mathbf{X} = \mathbf{X}_R + j\mathbf{X}_I$ is defined by the joint distribution of $\mathbf{X}_R$ and $\mathbf{X}_I$. A complex random vector $\mathbf{X}$ is proper (or symmetrical) if

$$\text{Cov}(\mathbf{X}\mathbf{X}^T) = \mathbf{0} \;\Rightarrow\; \begin{cases} \text{Cov}(\mathbf{X}_R, \mathbf{X}_R^t) = \text{Cov}(\mathbf{X}_I, \mathbf{X}_I^t) \\ \text{Cov}(\mathbf{X}_R, \mathbf{X}_I^t) = -\text{Cov}(\mathbf{X}_I, \mathbf{X}_R^t) \end{cases}$$

Remarks

- If $\mathbf{X}$ is symmetrical, then all second-order statistics of $\mathbf{X}$ is contained in Cov$(\mathbf{X}, \mathbf{X}^H)$.

$$\begin{aligned} \text{Cov}(\mathbf{X}, \mathbf{X}^H) &= \text{Cov}(\mathbf{X}_R, \mathbf{X}_R^T) + \text{Cov}(\mathbf{X}_I, \mathbf{X}_I^T) \\ &\quad -j(\text{Cov}(\mathbf{X}_R, \mathbf{X}_I^T) - \text{Cov}(\mathbf{X}_I, \mathbf{X}_R^T)) \\ &= 2\text{Cov}(\mathbf{X}_R, \mathbf{X}_R^T) + 2j\text{Cov}(\mathbf{x}_I, \mathbf{x}_R^t) \end{aligned}$$

- If $\mathbf{X}$ is proper, then $\mathbf{A}\mathbf{X} + \mathbf{b}$ is also proper (invariant under affine transforms).
- For proper complex random vectors, we can use complex arithmetics at a lower dimension by changing transpose to Hermitian.

Complex Gaussian Random Vectors

Random vector $\mathbf{x}$ is complex Gaussian if

1. $\mathbf{X}$ is symetrical
2. $\begin{pmatrix} \mathbf{X}_R \\ \mathbf{X}_I \end{pmatrix}$ is Gaussian.

Properties

- Distribution: $\mathbf{X} \sim \mathcal{N}_c(\mu, \Sigma)$ implies

$$\begin{aligned} E(\mathbf{X}) &= \mu, cov(\mathbf{x}, \mathbf{x}^H) = \Sigma, \\ f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{\pi^n |\Sigma|} exp\{-(\mathbf{x} - \mu)^H \Sigma^{-1} (\mathbf{x} - \mu)\}. \end{aligned}$$

- When $\mathbf{X}_R, \mathbf{X}_I \sim \mathcal{N}(0, \frac{N_0}{2}\mathbf{I})$, $\mathbf{X} \sim \mathcal{N}_c(0, N_0\mathbf{I})$,

$$p(\mathbf{x}) = \frac{1}{\pi^n N_0^n} exp\{-\frac{||\mathbf{x}||^2}{N_0}\}.$$

- A userful case: If $\mathbf{X} = \mathbf{S} + \mathbf{N}$ where $\mathbf{S}$ and $\mathbf{N}$ are independent, $\mathbf{N} \sim \mathcal{N}(0, N_0\mathbf{I})$,

$$f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s}) = \frac{1}{\pi^n N_0^n} exp\{-\frac{||\mathbf{x} - \mathbf{s}||^2}{N_0}\}$$

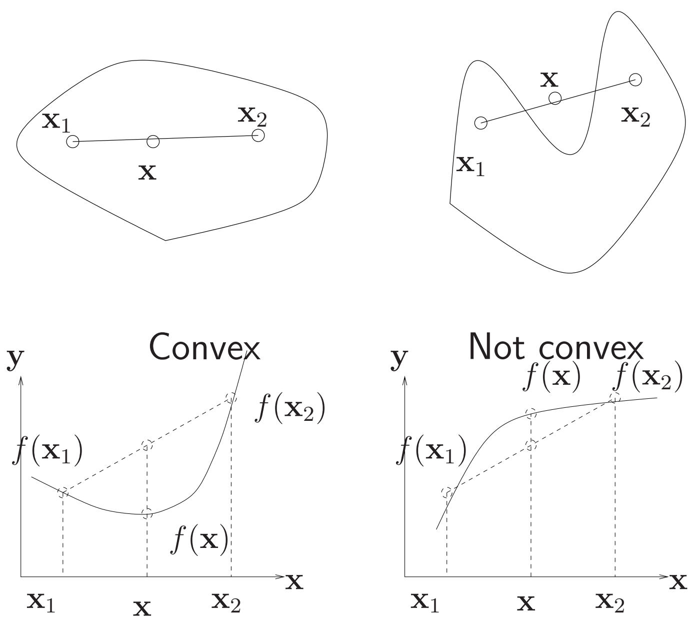
Convexity and Jensen's Inequality

Convex Set and Convex Function

A set $\mathcal{X}$ in $\mathcal{R}^n$ or $\mathcal{C}^n$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\theta \in [0,1]$, $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{X}$. A real valued function $f(\cdot)$ on a convex set $\mathcal{X}$ is convex (convex $\cup$) if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\theta \in [0,1]$,

$$f(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \le \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2)$$

A function is strictly convex if the strict inequality holds. A function $f$is concave (convex $\cap$) if $-f$ is convex.



Convex function   Concave function

Jensen's Inequality

Let f be a real valued convex function. Then

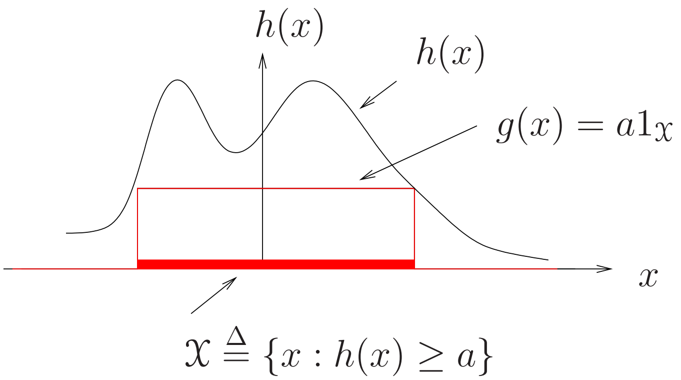
$$f(\mathbb{E}(\mathbf{x})) \le \mathbb{E}(f(\mathbf{x}))$$

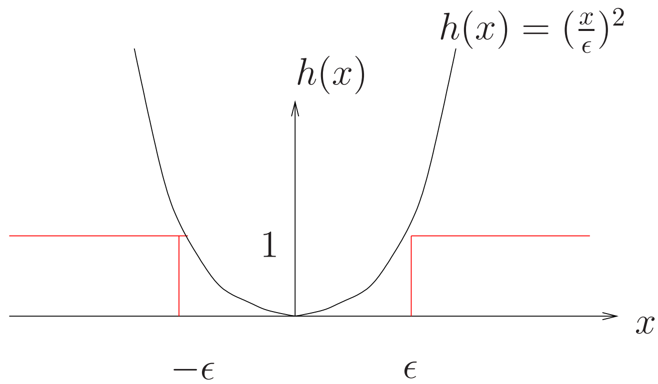
For concave f, the inequality is reversed.

Markov and Chebyshev Inequalities

The Markov Inequality: For any non-negative function $h(\cdot)$,

$$\Pr[h(X) \geq a] \leq \frac{\mathbb{E}(h(X))}{a} \quad \forall a > 0.$$





Chebyshev Inequality: Setting $h(x) = |x - \mathbb{E}(X)|^2$,

$$\Pr[\frac{|X - \mathbb{E}(X)|}{\epsilon} \geq 1] \leq \frac{\text{Var}(X)}{\epsilon^2}$$

As an application, for i.i.d. X_i and $\mathbb{E}(X_i) = p$,

$$Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i \rightarrow \Pr(|Y_N - p| > \epsilon) \leq \frac{\text{Var}(X)}{N\epsilon^2}$$

The probability of Y_N deviates from its mean decreases with $O(\frac{1}{N})$.

A Lower Bound

If h is a non-negative uniformly bounded by M, then

$$\Pr(h(X) \geq a) \geq \frac{\mathbb{E}(h(X)) - a}{M - a}, \quad a \in [0, M).$$

Chernoff Bound

If we want to have exponentially decaying probability, we may need the Chernoff bound. Let X be a random variable. For any $\lambda > 0$ and τ,

$$\Pr[X \geq \tau] \leq \exp\{-\lambda\tau + \phi_X(\lambda)\}$$

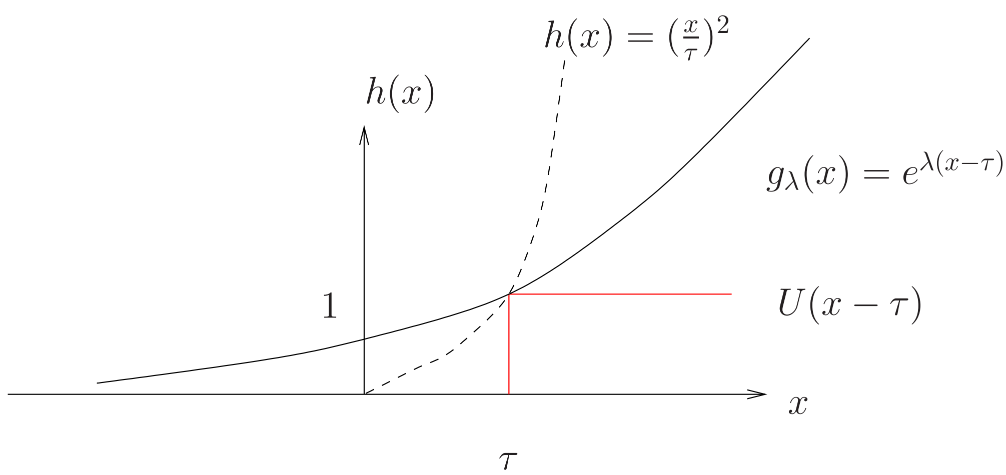
where

$$\phi_X(\lambda) \triangleq \ln \mathbb{E}(e^{\lambda X})$$

is the cumulant generating function. Similarly, we also have

$$\Pr[X \leq \tau] \leq \exp\{\lambda\tau + \phi_X(-\lambda)\}$$

Proof: Use the Markov inequality with $h(X) = e^{\lambda X}$ and $a = e^{\lambda\tau}$



Remark: The Chernoff bound can be tightened by optimizing λ.

An Application of the Chernoff Bound

Consider

$$Y_N \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i, \quad X_i \overset{\text{i.i.d.}}{\sim} \mathcal{B}(p)$$

By the Chernoff bound,

$$\begin{aligned}
\Pr[Y_N \geq a] &= \Pr[\sum X_i \geq Na] \leq e^{-N\lambda a}\mathbb{E}(e^{\lambda \sum X_i}) \\
&= e^{-N\lambda a}[\mathbb{E}(e^{\lambda X_i})]^N \\
&= [\mathbb{E}(e^{\lambda(X_i - a)})]^N
\end{aligned}$$

The best λ is given by solving

$$\frac{d}{d\lambda}\mathbb{E}(e^{\lambda(X_i-a)})|_{\lambda=\lambda_o} = 0 \to \frac{\mathbb{E}(X_i e^{\lambda_o X_i})}{\mathbb{E}(e^{\lambda_o X_i})} = a$$

For Bernoulli r.v. and $a \in (p, 1]$,

$$\frac{pe^{\lambda_o}}{pe^{\lambda_o} + (1-p)} = a \to \lambda_o = \ln\frac{a(1-p)}{p(1-a)} > 0$$

Thus,

$$\Pr[Y_N \geq a] \leq [(\frac{p}{a})^a(\frac{1-p}{1-a})^{1-a}]^N = \exp\{-ND(\mathcal{B}(a)||\mathcal{B}(p)))\}$$

where

$$D(P_1||P_2) \triangleq \mathbb{E}_{P_1}(\log\frac{P_1}{P_2})$$

is the Kullback-Leibler divergence, which is always positive.

Weak Convergence and Weak LLN

Definition

Suppose X and $\{X_n, n = 1, 2, \cdots\}$ are random variables defined on the same probability space. We say that the sequence (X_n) converges in probability, denoted as $X_n \xrightarrow{P} X$ if, for all ϵ,

$$\Pr(|X_n - X| \geq \epsilon) \to 0 \quad \text{as } n \to \infty$$

Example

Let X_n be independent variables with PMF

$$\Pr(X_n = 1) = 1 - \frac{1}{n} \quad \Pr(X_n = n) = \frac{1}{n}$$

For any $\epsilon > 0$,

$$\Pr(|X_n - 1| > \epsilon) = \Pr(X_n = n) = \frac{1}{n} \to 0 \text{ as } n \to \infty$$

Therefore $X_n \xrightarrow{P} 1$.

The Weak Law of Large Numbers

Let X_i be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then,

$$\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

Proof: Use the Chebyshev Inequality for $X = \frac{1}{N} \sum_{i=1}^{N} X_i$.

Strong Convergence and Strong LLN

Definition

The sequence (X_n) converges almost surely (or strongly), denoted by $X_n \xrightarrow{\text{a.s.}} X$, if

$$\Pr(\omega \in \Omega : X_n(\omega) \to X(\omega)) = \Pr(X_n \to X) = 1 \text{ as } n \to \infty$$

Equivalently, $X_n \xrightarrow{\text{a.s.}} X$ if $\forall \epsilon > 0$ and $\delta \in (0,1)$, there exists n_0 such that, for all $n > n_0$,

$$\Pr(\bigcap_{m>n} \{|X_m - X| \leq \epsilon\}) > 1 - \delta$$

Example Revisited Let X_n be independent variables with PMF

$$\Pr(X_n = 1) = 1 - \frac{1}{n} \quad \Pr(X_n = n) = \frac{1}{n}$$

For every $\epsilon > 0$, $\delta \in (0,1)$, and $N > n$,

$$\Pr(\bigcap_{m>n} \{|X_m - 1| \leq \epsilon\}) \leq \Pr(\bigcap_{m=n+1}^{N} \{|X_m - 1| \leq \epsilon\}) = \prod_{m=n+1}^{N} \Pr(|X_m - 1| \leq \epsilon)$$

$$= \prod_{m=n+1}^{N} (1 - \frac{1}{m}) = \frac{n}{N} \leq 1 - \delta$$

Strong Law of Large Numbers

Suppose (X_n) are i.i.d. random variables with mean μ and $\mathbb{E}(|X|^4) < \infty$. Then

$$\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu$$

We can show that

$$\Pr(|\bar{X}_n - \mu| > \epsilon) \leq \frac{A}{n^2},$$

where A is a constant. By the Borel-Cantellis Lemma, $\{|\bar{X}_n - \mu| > \epsilon\}$ happens only finite number of times.

Convergence in Distribution and CLT

Definition

Suppose X and $\{X_n, n = 1, 2, \cdots\}$ are random variables defined on the same probability space. We say that the sequence (X_n) with CDF $F_{X_n}(x)$ converges in distribution to X with CDF $F_X(x)$, denoted as $X_n \xrightarrow{D} X$, if $F_{X_n}(x) \to F_X(x)$ for all x where $F_X(x)$ is continuous.

Central Limit Theorem

Let $\{X_n\}$ be i.i.d. random variables with mean μ and variance σ^2. Denote $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$$

The law of the iterative lograrithm

If $\{X_i\}$ are i.i.d. with mean μ and variance σ^2. Then

$$\Pr(\limsup_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{2n \log\log n}} = 1) = 1$$

This means that the event, with probability 1, the event

$$\{\frac{S_n - n\mu}{\sigma} > \alpha\sqrt{2n \log\log n}\}$$

should happen only finite number of times if $\alpha > 1$ and infinitely many times if $\alpha < 1$.