Course Topics

- **Basic probability and random variables:**
  - Lecture 1: Probability space and Axioms, basic laws, conditional probability, Bayes rule, and independence
  - Lecture 2: Random variables, probability mass function (PMF), cumulative distribution function (CDF), probability density function (PDF), functions of random variables
  - Lecture 3: Joint, marginal and conditional distributions, mixed random variables, signal detection, functions of two random variables
  - Lecture 4: Expectation: mean and variance, covariance and correlation, Markov and Chebyshev inequalities, conditional expectation.

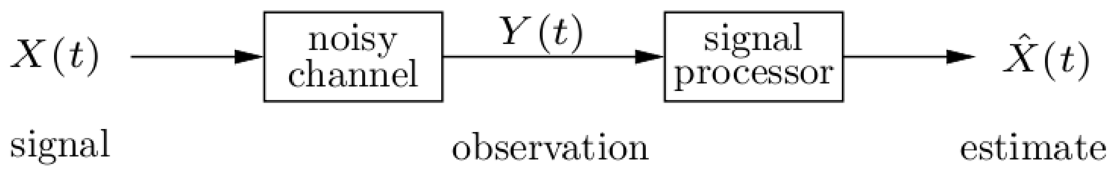
- **Random vectors**
  - Lecture 5: Minimum MSE estimation, linear estimation, jointly Gaussian random variables, independence and conditional independence
  - Lecture 6: Mean and covariance matrix, Gaussian random vectors, Gaussian random vectors

- **Convergence**
  - Lecture 7: Modes of convergence and convergence laws

- **Random processes** Definition and examples of discrete and continous random processes: IID, random walk, independent increment processes, Gaussian random processes. Stationary, autocorrelation function and power spectral density. White noise, bandlimited processes. Applications: Response of linear systems to random inputs.

Statistical Signal Processing

- Focus is on extracting *information* (signals) from *noisy observations*
- Applications are all around us — cell phones, digital cameras, digital TV, DSL modem, . . .
- Generic signal processing problem:



- Signal: audio, image, video, geophysical, medical, . . .
- Channel: twisted pair, optical, wireless, satellite, electronic circuit, biological, . . .

- Channel is modeled as a *statistical system* — linear vs. nonlinear, time invariant vs. time varying
- Noise (physically generated or due to interference) and often signals are modelled as *random processes*, i.e., collections of random variables indexed by time
- Signal processor: attempts to recover the signal from observation via
  - estimation: find an estimate that is close to the signal $X(t)$, for example, that minimizes the *mean square error* (MSE)
  - detection: decide which signal out of a finite number of possible signals (e.g., 0 and 1) was sent — goal: minimize the *error probability*
  - Statistical signal processing deals with both *modeling* of signals and channels and *design* of signal processors
  - Example: many real channels (twisted pair/wireless model, . . .) are modeled as linear time invariant (LTI) system with additive noise

Basic Probability Theory

- Probability theory provides the mathematical rules for assigning probabilities to outcomes of random experiments, e.g., coin flips, packet arrivals, noisy voltage

- Basic elements of probability theory:
  - *Sample space* Ω: set of all possible "elementary" or "finest grain" outcomes of the random experiment
  - *Set of events* $\mathcal{F}$: set of (all?) subsets of Ω — an event $A \subset \Omega$ occurs if the outcome $\omega \in A$
  - *Probability measure* P: function of $\mathcal{F}$ that assigns probabilities to events according to the axioms of probability

- Formally, a *probability* space is the triple $(\Omega, \mathcal{F}, \mathrm{P})$

Probability Measure

- A probability measure $\mathrm{P} : \mathcal{F} \mapsto [0,1]$ satisfies the following axioms:
  (i) $\mathrm{P}(\Omega) = 1$.
  (ii) If $A_1, A_2, \ldots \in \mathcal{F}$ with $A_i \cap A_j = \emptyset$, $\forall\, i \neq j$ (disjoint events), then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

  (countably additive).
- P is a measure in the same sense as *mass*, *length*, *area* and *volume*. E.g., all these are countably additive.
- Unlike those measures, P is bounded by 1.
- New concepts in probability theory not in measure theory: independence (and conditioning).

Discrete Probability Spaces

- Sample space Ω is said to be *discrete* if it is countable

- Examples:

  - Flipping a coin: $\Omega = \{H, T\}$
  - Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Flipping a coin n times: $\Omega = \{H, T\}^n$, sequences of heads/tails of length n
  - Flipping a coin until the first heads appears: $\Omega = \{H, TH, TTH, TTTH, \ldots\}$
  - Number of packets arriving at a node in a communication network in time interval $(0, T]$: $\Omega = \{0, 1, 2, 3, \ldots\}$

  The first three examples have *finite* Ω, whereas the last two have *countably infinite* Ω. Both types are considered *discrete*

- For discrete sample spaces, the set of events $\mathcal{F}$ can be taken to be the set of all subsets of Ω, sometimes called the *power set* of Ω

- Example: For the coin flipping experiment,

$$\mathcal{F} = \{\emptyset,\, \{H\},\, \{T\},\, \Omega\}$$

- $\mathcal{F}$ does not have to be the entire power set (more on this later)

- The probability measure P can be defined by assigning probabilities to individual outcomes — single outcome events $\{\omega\}$ — so that:

$$\mathrm{P}(\{\omega\}) \geq 0 \text{ for every } \omega \in \Omega$$

$$\sum_{\omega \in \Omega} \mathrm{P}(\{\omega\}) = 1$$

- The probability of any other event A is simply

$$\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\{\omega\})$$

- Examples:
  - For the coin flipping experiment, for some p with $0 \le p \le 1$ assign

    $$\mathrm{P}(\{H\}) = p \quad \text{and} \quad \mathrm{P}(\{T\}) = 1 - p$$

    Note: p is called the *bias*. A coin is *fair* if $p = \frac{1}{2}$
  - For the die rolling experiment, assign

    $$\mathrm{P}(\{i\}) = \tfrac{1}{6} \quad \text{for } i = 1, 2, \ldots, 6$$

    The probability of the event "the outcome is even," $A = \{2, 4, 6\}$, is

    $$\mathrm{P}(A) = \mathrm{P}(\{2\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{6\}) = \tfrac{3}{6} = \tfrac{1}{2}$$
  - For the number of packets arriving in $(0, T\,]$, for parameter $\lambda > 0$ assign

    $$\mathrm{P}(\{k\}) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \quad \text{for } k = 0, 1, 2, \ldots$$

    This is the *Poisson* probability distribution with parameter λ, which is the average number of packets per unit time

Continuous Probability Spaces

- Sample space is said to be *continuous* if Ω is *uncountably infinite*
- Examples:
  - Random number between 0 and 1: $\Omega = (0, 1\,]$
  - Packet arrival time: $\Omega = (0, \infty)$
  - Arrival times of n packets: $\Omega = (0, \infty)^n$
- For continuous Ω, we cannot in general define the probability measure P by first assigning probabilities to outcomes
- To see why, consider assigning a unifrom probability measure over $(0, 1\,]$
  - In this case the probability of each single outcome event is zero
  - How do we find the probability of an event such as $A = [0.25, 0.75]$?
- Another difference for continuous Ω: we cannot take the set of events $\mathcal{F}$ as the power set of Ω. (To learn why you need to study measure theory, which is beyond the scope of this course)

- The set of events $\mathcal{F}$ cannot be an arbitrary collection of subsets of Ω. It must make sense, e.g., if A is an event, then its complement A^c must also be an event, the union of two events must be an event, and so on

- Formally, $\mathcal{F}$ must be a *sigma algebra* (σ*-algebra*, σ*-field*), which satisfies the following axioms:

  1. $\emptyset \in \mathcal{F}$
  2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
  3. If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

- Of course, the power set is a sigma algebra. But we can define smaller σ-algebras. For example, for rolling a die, we could define the set of events as

$$\mathcal{F} = \{\phi, \text{odd}, \text{even}, \Omega\}$$

- For $\Omega = R = (-\infty, \infty)$ (or $(0, \infty)$, $(0, 1)$, etc.) $\mathcal{F}$ is typically defined as the family of sets obtained by countable unions, intersections, and complements of intervals in R (including the intervals themselves)

- The resulting $\mathcal{F}$ is called the *Borel field*

- Note: Amazingly there are subsets in R that cannot be generated in this way! (Not ones that you are likely to encounter in your life as an engineer or even as a mathematician)

- To define a probability measure over a Borel field, we first assign probabilities to the intervals in a consistent way, i.e., in a way that satisfies the axioms of probability

  For example to define uniform probability measure over $(0, 1)$, we first assign $\mathrm{P}((a, b)) = b - a$ to all intervals

Basic Probability Properties

- $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$
- If $A \subset B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$
- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
- $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$
- More generally, the *Union of Events Bound*:

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) \;\leq\; \sum_{i=1}^{n} \mathrm{P}(A_i)$$

- *Law of Total Probability*: Let $A_1, A_2, A_3, \ldots$ be events that partition Ω, i.e., disjoint $\big(A_i \cap A_j = \emptyset$ for $i \neq j\big)$ and $\bigcup_{i=1} A_i = \Omega$. Then for any event B

$$\mathrm{P}(B) \;=\; \textstyle\sum_{i=1} \mathrm{P}(A_i \cap B)$$

The Law of Total Probability is very useful for finding probabilities of sets

Conditional Probability

- Let B be an event such that $\mathrm{P}(B) \neq 0$. The *conditional probability* of event A given B is defined to be

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A, B)}{\mathrm{P}(B)}$$

- The function $\mathrm{P}(\cdot \mid B)$ is a probability measure over $\mathcal{F}$, i.e., it satisfies the axioms of probability

- Chain rule: $\mathrm{P}(A, B) = \mathrm{P}(A)\mathrm{P}(B|A) = \mathrm{P}(B)\mathrm{P}(A|B)$ (can be generalized to $n$events

- The probability of event A given B, a nonzero probability event — the *a posteriori* probability of A — is related to the unconditional probability of A — the *a priori* probability — by

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)}{\mathrm{P}(B)} \mathrm{P}(A)$$

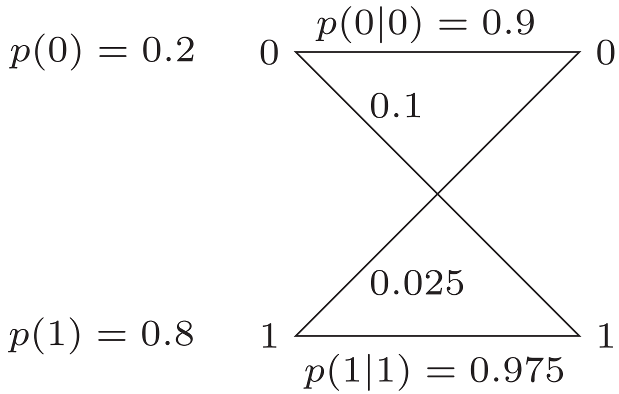
  This follows directly from the definition of conditional probablility

Bayes Rule

- Let $A_1, A_2, \ldots, A_n$ be nonzero probability events that partition Ω, and let B be a nonzero probability event
- We know $\mathrm{P}(A_i)$ and $\mathrm{P}(B \mid A_i)$ and want a posteriori probabilities $\mathrm{P}(A_j \mid B)$
- We know that
$$\mathrm{P}(A_j \mid B) = \frac{\mathrm{P}(B \mid A_j)}{\mathrm{P}(B)} \mathrm{P}(A_j)$$
- By the law of total probability
$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(A_i, B) = \sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B \mid A_i)$$
- Substituting, we obtain *Bayes rule*
$$\mathrm{P}(A_j \mid B) = \frac{\mathrm{P}(B \mid A_j)}{\sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B \mid A_i)} \mathrm{P}(A_j)$$
- Bayes rule also applies to a (countably) infinite number of events

Example: Binary Communication Channel

- Consider the following *probability transition diagram* for a noisy binary channel



  Given that 0 is received, find the probability that 0 was sent

- This is a random experiment with sample space

$$\Omega = \{(0,0), (0,1), (1,0), (1,1)\},$$

  where the first entry is the bit sent and the second is the bit received

- Define the two events

$$A = \{0 \text{ is sent}\} = \{(0,1),(0,0)\}$$
$$B = \{0 \text{ is received}\} = \{(0,0),(1,0)\}$$

- The probability measure on Ω is determined by $\mathrm{P}(A)$, $\mathrm{P}(B \mid A)$, and $\mathrm{P}(B^c \mid A^c)$, which are given on the probability transition diagram of the channel

- To find $\mathrm{P}(A \mid B)$, the *a posteriori* probability that 0 was sent, use Bayes rule:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)}{\mathrm{P}(A)\mathrm{P}(B \mid A) + \mathrm{P}(A^c)\mathrm{P}(B \mid A^c)} \mathrm{P}(A)$$

  to obtain

$$\mathrm{P}(A \mid B) = \frac{0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.025} \cdot 0.2 = \frac{0.9}{0.2} \cdot 0.2 = 0.9$$

  Note that $\mathrm{P}(A \mid B) = 0.9$, the a posteriori probability of A, is much larger than $\mathrm{P}(A) = 0.2$, the a priori probability of A

Independence

- Two events are said to be *statistically independent* if

$$\mathrm{P}(A, B) = \mathrm{P}(A)\mathrm{P}(B)$$

- When $\mathrm{P}(B) \neq 0$, this is equivalent to

$$\mathrm{P}(A \mid B) = \mathrm{P}(A)$$

  In other words, knowing whether B occurs does not change the probability of A

- Example: Assuming that the binary channel of the previous example is used to send two bits independently, what is the probability that both bits are in error?
  - Define the two events

$$E_1 = \{\text{First bit is in error}\}\,, \quad E_2 = \{\text{Second bit is in error}\}$$

  - Since the bits are sent independently, the probability that both are in error is

$$\mathrm{P}(E_1 \cap E_2) \;=\; \mathrm{P}(E_1, E_2) \;=\; \mathrm{P}(E_1)\mathrm{P}(E_2)$$

- To find $\mathrm{P}(E_1)$, we express E_1 in terms of the events A and B as

$$E_1 = (A_1 \cap B_1^c\ )\ \ (A^c{}_1 \cap B_1)$$

- Since E_1 has been expressed as the union of disjoint events,

$$\begin{aligned}\mathrm{P}(E_1) &= \mathrm{P}(A_1, B_1^c) + \mathrm{P}(A_1^c, B_1) \\ &= \mathrm{P}(A_1)\mathrm{P}(B_1^c \,|\, A_1) + \mathrm{P}(A_1^c)\mathrm{P}(B_1 \,|\, A_1^c) \\ &= 0.2 \cdot 0.1 \ + \ 0.8 \cdot 0.025 \ = \ 0.04\end{aligned}$$

- The probability that the two bits are in error is

$$\mathrm{P}(E_1, E_2) \ = \ \mathrm{P}(E_1)\mathrm{P}(E_2) \ = \ (0.04)^2 \ = \ 1.6 \times 10^{-3}$$

- In general $A_1, A_2, \ldots, A_n$ are defined to be independent if for *every* subset $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ of the events,

$$\mathrm{P}(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) \ = \prod_{j=1}^{k} \mathrm{P}(A_{i_j})$$

- Note: $\mathrm{P}(A_1, A_2, \ldots, A_n) = \prod_{j=1}^{n} \mathrm{P}(A_i)$ is *not* sufficient for independence

Example: Roll two fair dice independently. Define the events

$$A = \{\text{First die is 1, 2, or 3}\}$$

$$B = \{\text{First die is 2, 3, or 6}\}$$

$$C = \{\text{Sum of outcomes is 9}\} \ = \ \{(3,6),(4,5),(5,4),(6,3)\}$$

Are A, B, and C independent?

Since the dice are fair and the experiments are done independently, the probability of any pair of outcomes is $\frac{1}{36}$. Therefore

$$\mathrm{P}(A) = \tfrac{1}{2}\,, \quad \mathrm{P}(B) = \tfrac{1}{2}\,, \quad \mathrm{P}(C) = \tfrac{1}{9}$$

Since $A \cap B \cap C = \{(3,6)\}$,

$$\mathrm{P}(A,B,C) = \tfrac{1}{36} = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$$

But A, B, and C are *not* independent because

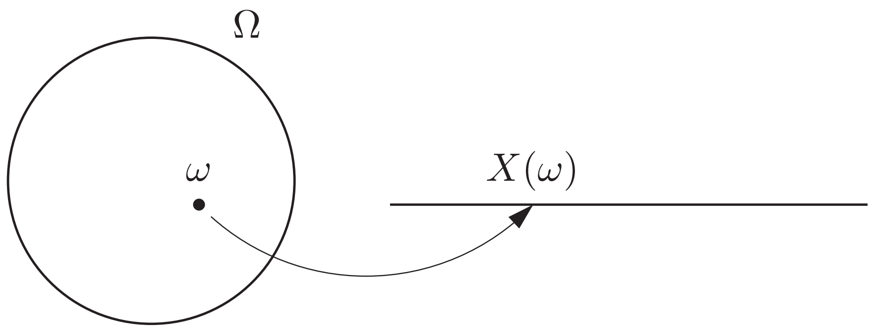
$$\mathrm{P}(A,B) = \tfrac{2}{6} = \tfrac{1}{3} \neq \tfrac{1}{4} = \mathrm{P}(A)\mathrm{P}(B)$$

Lecture 2: Random Variables

- Definition
- Probability Mass Function (PMF)
- Cumulative Distribution Function (CDF)
- Probability Density Function (PDF)
- Functions of a Random Variable
- Application: Generation of Random Variables

Random Variable

- A *random variable* (r.v.) is a real-valued function $X(\omega)$ over a sample space Ω, i.e., $X : \Omega \to \mathbf{R}$



- Notations:
  - We use upper case letters for random variables: $X, Y, Z, \Phi, \Theta, \ldots$
  - We use lower case letters for *values* of random variables: $X = x$ means that random variable X takes on the value x, i.e., $X(\omega) = x$ where ω is the outcome

Examples of Random Variables

1. Let the random variable X be the number of heads in n coin flips. The sample space is $\Omega = \{H, T\}^n$, the possible outcomes of n coin flips, so

$$X \in \{0, 1, 2, \ldots, n\}$$

2. Let $\Omega = \mathbf{R}$, the real numbers. Define random variables X and Y as follows.

   a. $X(\omega) = \omega$

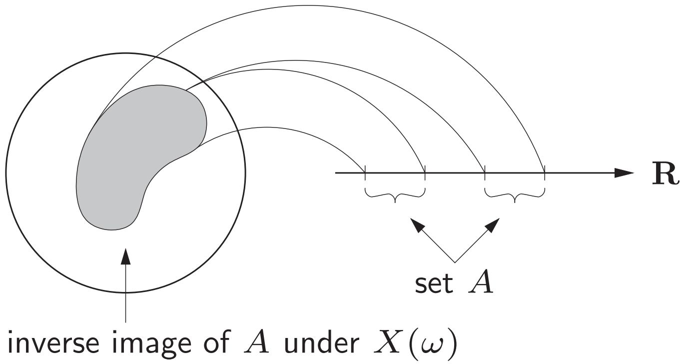
   b. $Y(\omega) = \begin{cases} +1 & \omega \geq 0 \\ -1 & \text{otherwise} \end{cases}$

3. Packet arrival times in the interval $(0, T\,]$. Here Ω is the set of all finite length strings $(t_1, t_2, \ldots, t_n) \in (0, T\,]^*$, where $t_1 \leq t_2 \leq \cdots \leq t_n$. Define the random variable X to be n, the length of the string; $X \in \{0, 1, 2, 3, \ldots\}$

4. Let X be the service time at a router. If the buffer is empty the packet is served immediately, i.e., $X = 0$. If it is not empty, the service time $X \geq 0$

Specifying a Random Variable

- Specifying a random variable means being able to determine the probability that $X \in A$ for any Borel set $A \subset \mathbf{R}$, in particular, for any interval $(a, b]$
- To do so, consider the *inverse image* of A under X, i.e., $\{\omega : X(\omega) \in A\}$



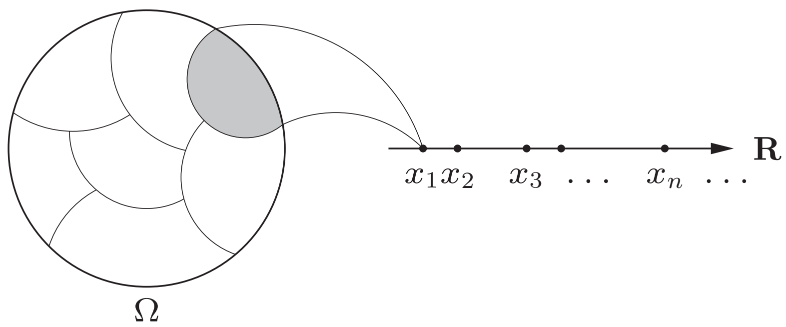
- Since $X \in A$ iff $\omega \in \{\omega : X(\omega) \in A\}$,

$$\mathrm{P}(\{X \in A\}) \;=\; \mathrm{P}(\{\omega : X(\omega) \in A\}) \;=\; \mathrm{P}\{\omega : X(\omega) \in A\}$$

Shorthand: $\mathrm{P}(\{$*set description*$\}) = \mathrm{P}\{$*set description*$\}$

Discrete Random Variables

- A random variable is said to be *discrete* if $\mathrm{P}\{X \in \mathcal{X}\} = 1$ for some *countable* set $\mathcal{X} \subset \mathbf{R}$, i.e., $\mathcal{X} = \{x_1, x_2, \ldots\}$ (finite or infinite)
- Examples 1, 2b, and 3 above are discrete random variables.
- In general, $X(\omega)$ partitions Ω into the sets $\{\omega : X(\omega) = x_i\}$, for $i = 1, 2, \ldots$



In order to specify X, it suffices to know $\mathrm{P}\{X = x_i\}$ for every i

- A discrete random variable is completely specified by its probability mass function

$$p_X(x) \;=\; \mathrm{P}\{X = x\} \text{ for every } x \in \mathcal{X}$$

- Clearly $p_X(x) \geq 0$ and $\sum_{x \in \mathcal{X}} p_X(x) = 1$

- So $p_X(x)$ can be simply viewed as a probability measure over a discrete sample space (even though the original sample space may be continuous as in examples 2b and 3)

- The probability of any (Borel) set $A \subset \mathbf{R}$ is given by

$$\mathrm{P}\{X \in A\} = \sum_{x \in A \cap \mathcal{X}} p_X(x)$$

- Notation: We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to mean that the discrete random variable X has pmf $p_X(x)$ or $p(x)$

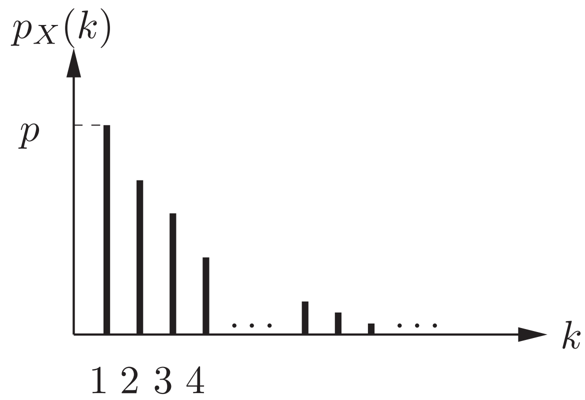
Famous Discrete Random Variables

- *Bernoulli*: $X \sim \mathrm{Bern}(p)$ for $0 \le p \le 1$ has the pmf

$$p_X(1) = p, \text{ and } p_X(0) = 1 - p$$

- *Geometric*: $X \sim \mathrm{Geom}(p)$ for $0 \le p \le 1$ has the pmf

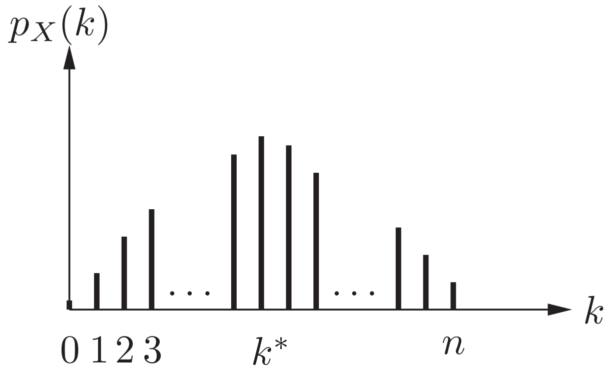
$$p_X(k) = p(1-p)^{k-1}, \text{ for } k = 1, 2, \ldots$$



The Geometric r.v. represents, for example, the number of coin flips until the first heads shows up (assuming independent coin flips)

- Binomial: $X \sim \mathrm{Binom}(n, p)$ for integer $n > 0$ and $0 \leq p \leq 1$ has the pmf

$$p_X(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \text{ for } k = 0, 1, 2, \ldots, n$$



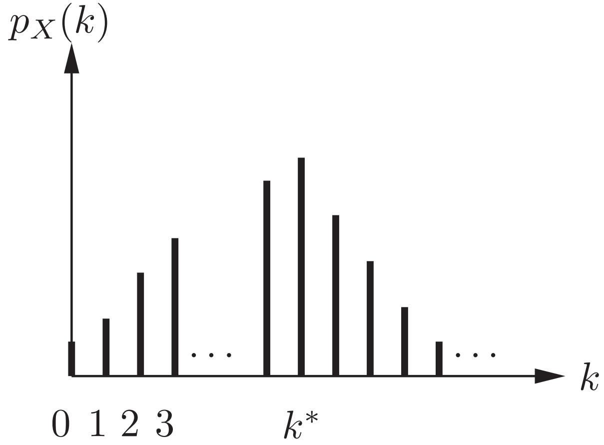
The maximum of $p_X(k)$ is attained at

$$k^* = \begin{cases} (n+1)p,\ (n+1)p - 1, & \text{if } (n+1)p \text{ is an integer} \\ [(n+1)p], & \text{otherwise} \end{cases}$$

The binomial r.v. represents, for example, the number of heads in n independent coin flips

- Poisson: $X \sim \mathrm{Poisson}(\lambda)$ for $\lambda > 0$ has the pmf

$$p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \text{ for } k = 0, 1, 2, \ldots$$



The maximum of $p_X(k)$ attained at

$$k^* = \begin{cases} \lambda,\ \lambda - 1, & \text{if } \lambda \text{ is an integer} \\ [\lambda], & \text{otherwise} \end{cases}$$

The Poisson r.v. represents the number of random events in a unit time, e.g., arrivals of packets, photons, customers — λ is the average arrival rate

- Fact: Poisson is the limit of Binomial when $p \propto \frac{1}{n}$ as $n \to \infty$

To show this let $X_n \sim \mathrm{B}(n, \frac{\lambda}{n})$, for $\lambda > 0$ and any fixed positive integer k

$$
\begin{aligned}
p_{X_n}(k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{(n-k)} \\
&= \frac{n(n-1)\ldots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(\frac{n-\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\ldots(n-k+1)}{(n-\lambda)^k} \frac{\lambda^k}{k!} \left(\frac{n-\lambda}{n}\right)^{n} \\
&= \frac{n(n-1)\ldots(n-k+1)}{(n-\lambda)(n-\lambda)\ldots(n-\lambda)} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n} \\
&\to \frac{\lambda^k}{k!} e^{-\lambda} \text{ as } n \to \infty
\end{aligned}
$$

Cumulative Distribution Function

- To specify a random variable, we need to be able to determine $\mathrm{P}\{X \in A\}$ for any Borel set $A \subset \mathbf{R}$, i.e., any set generated by countable unions, intersections, and complements of intervals

- It suffices to specify $\mathrm{P}\{X \in (a, b]\}$ for all intervals. The probability of any other Borel set can be determined by the axioms of probability

- Equivalently, it suffices to specify its *cumulative distribution function* (cdf):

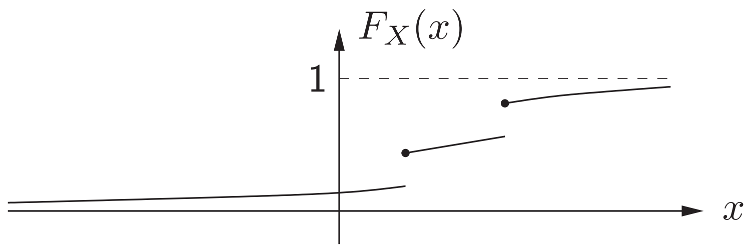
  $$F_X(x) = \mathrm{P}\{X \leq x\} = \mathrm{P}\{X \in (-\infty, x]\}, \quad x \in \mathbf{R}$$

  To see the equivalence:

  $$\mathrm{P}\{X \in (a, b]\} \;=\; \mathrm{P}\{X \leq b\} - \mathrm{P}\{X \leq a\} \;=\; F_X(b) - F_X(a), \quad a < b$$

Properities of CDF

- $F_X(x) \geq 0$ is monotonically nondecreasing, i.e., if $a > b$ then $F_X(a) \geq F_X(b)$



- Limits: $\lim_{x\to+\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$

- $F_X(x)$ is right continuous, i.e., $F_X(a^+) = \lim_{x\to a^+} F_X(x) = F_X(a)$

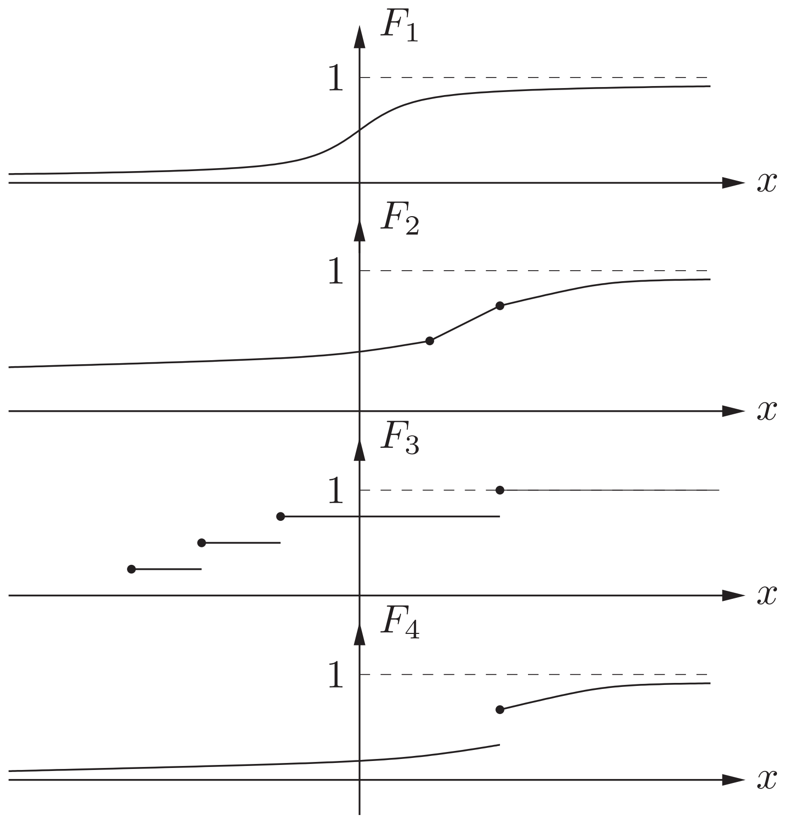
- $\mathrm{P}\{X = a\} = F_X(a) - F_X(a^-)$, where $F_X(a^-) = \lim_{x\to a-} F_X(x)$

- For any Borel set A, $\mathrm{P}\{X \in A\}$ can be determined from $F_X(x)$

- For a discrete random variable, $F_X(x)$ consists of a countable set of steps

- For any Borel set A, $\mathrm{P}\{X \in A\}$ can be determined from $F_X(x)$
- For a discrete random variable, $F_X(x)$ consists only of a countable set of steps
- A random variable is said to be *continuous* if its cdf is a continuous function
- A random variable is said to be *mixed* if its cdf is neither discrete nor continuous

Example CDFs



Probability Density Function

- If $F_X(x)$ is differentiable (except possibly over a countable set), then X has a *probability density function* (pdf) $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(\alpha)\, d\alpha$$

- If $F_X(x)$ is differentiable everywhere, then (by definition of derivative)

$$f_X(x) = \frac{dF_X(x)}{dx} = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\mathrm{P}\{x < X \le x + \Delta x\}}{\Delta x}$$

- Notation: $X \sim F_X(x)$ means that X has cdf $F_X(x)$, and $X \sim f_X(x)$ means that X has pdf $f_X(x)$

Properties of PDF

- $f_X(x) \geq 0$
- $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$
- For any event (Borel set) $A \subset \mathbf{R}$,

$$\mathrm{P}\{X \in A\} \; = \; \int_{x \in A} f_X(x)\,dx$$

  In particular,

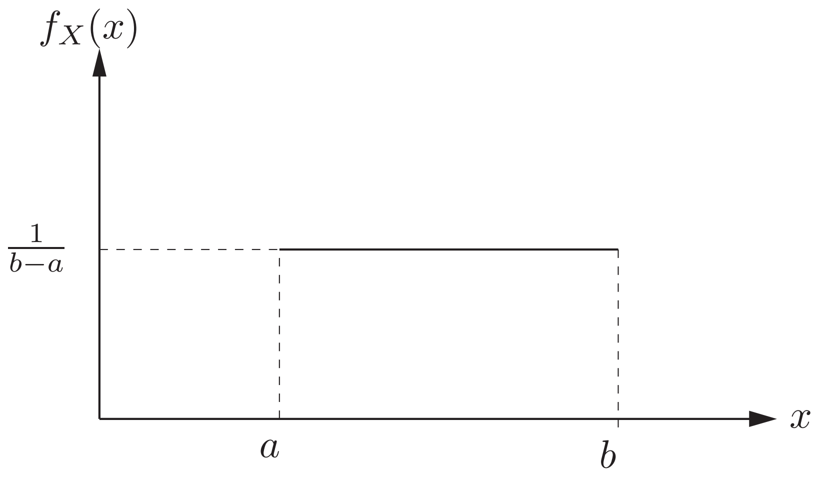
$$\mathrm{P}\{x_1 < X \leq x_2\} = \int_{x_1}^{x_2} f_X(x)\,dx$$

- Important note: $f_X(x)$ should not be interpreted as the probability that $X = x$. In fact, $f_X(x)$ is *not* a probability measure since it can be > 1
- Remark: We can use delta functions to define a pdf for a discrete or a mixed random variable, but this is not commonly done in the field of probability

Famous Continous Random Variables

- *Uniform*: $X \sim \mathrm{U}[a,b]$ where $a < b$ has pdf

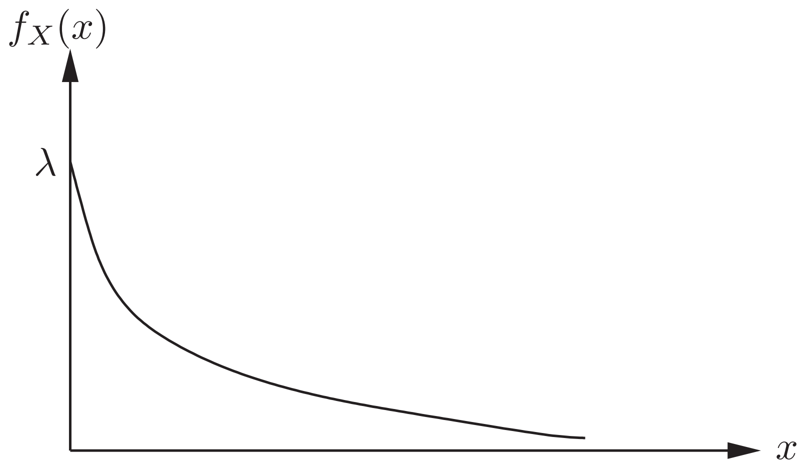
$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



The uniform r.v. is commonly used in modeling quantization noise and finite precision computation error (roundoff error)

• *Exponential:* $X \sim \mathrm{Exp}(\lambda)$ where $\lambda > 0$ has pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



The exponential r.v. represents *interarrival time* in a queue (time between two consecutive packet or customer arrivals) or service time in a queue, or particle lifetime

- Example: Let $X \sim \mathrm{Exp}(0.1)$ be the customer service time in minutes at a bank

  The person ahead of you has been served for 10 minutes. What is the probability that you will wait another 10 minutes or more before getting served?

  Solution: We want to find $\mathrm{P}\{X > 20 \,|\, X > 10\}$

  By definition,

$$\begin{aligned}\mathrm{P}\{X > 20 \,|\, X > 10\} &= \frac{\mathrm{P}\{X > 20, X > 10\}}{\mathrm{P}\{X > 10\}} \\ &= \frac{\mathrm{P}\{X > 20\}}{\mathrm{P}\{X > 10\}} \;=\; \frac{e^{-2}}{e^{-1}} \;=\; e^{-1}\end{aligned}$$

  But $\mathrm{P}\{X > 10\} = e^{-1}$. Therefore the conditional probability of waiting more than 10 minutes is the same as the *un*conditional probability of waiting more than 10 minutes!

- In general for any exponential r.v., whenever $0 \le x' < x$,

$$\mathrm{P}\{X > x \,|\, X > x'\} \;=\; \mathrm{P}\{X > x - x'\}$$

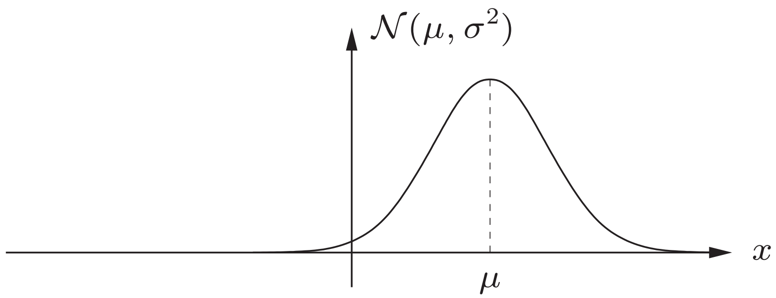
  Because of this property, the exponential r.v. is called *memoryless*

Gaussian Random Variable

- *Gaussian*: $X \sim \mathcal{N}(\mu, \sigma^2)$ has pdf with parameters μ (the *mean*) and σ^2 (the *variance*) (or the *standard deviation* σ):

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The graph of the Gaussian (or *normal*) density is the bell-shaped curve



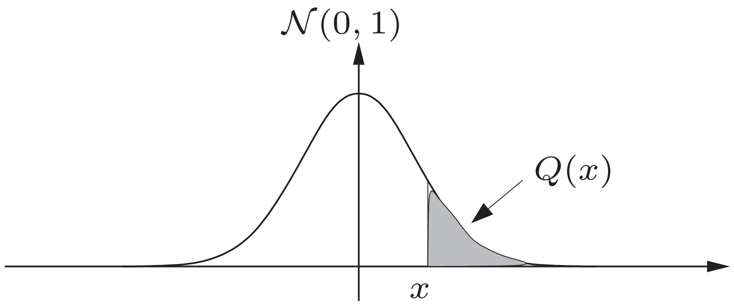
- The Gaussian r.v. is frequently encountered in nature — thermal and shot noise in electronic devices are Gaussian — and very frequently used in modelling various social, biological, and other phenomena. A *lot* more on Gaussian r.v.s later

Φ, Q, and erfc Functions

- The cdf of the standard normal random variable $\mathcal{N}(0,1)$ is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du$$

- Also define the function $Q(x) = 1 - \Phi(x) = \mathrm{P}\{X > x\}$



The $Q(\cdot)$ function can be used to compute $\mathrm{P}\{X > a\}$ for *any* Gaussian r.v. X

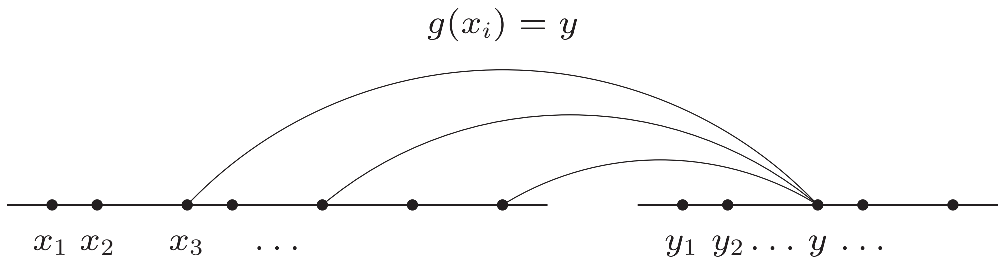
- The *complementary error function* is

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-u^2} du = 2Q(\sqrt{2}x)$$

Functions of a Random Variable

- Suppose we are given a r.v. with known distribution (pmf, cdf, or pdf), a function $y = g(x)$, and want to specify the random variable $Y = g(X)$
- If $X \sim p_X(x)$ is a discrete r.v., then Y is also discrete with pmf

$$p_Y(y) = \sum_{\{x: g(x) = y\}} p_X(x)$$



Derived Densities

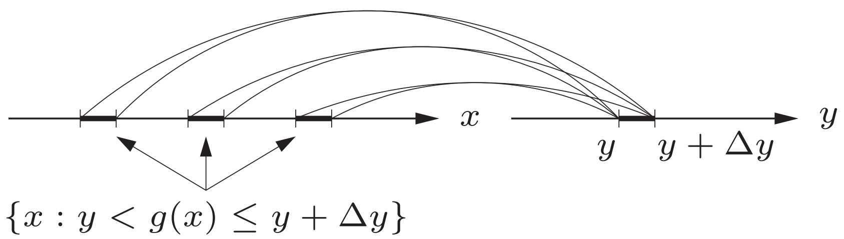
- If $X \sim f_X(x)$ is a continuous r.v., and $g(x)$ is differentiable, then we can find the pdf of Y as follows

  By definition of pdf, $f_Y(y) = \lim_{\Delta y \to 0} \frac{\mathrm{P}\{y < Y \le y + \Delta y\}}{\Delta y}$ To find f_Y, we find

  $$\mathrm{P}\{y < Y \le y + \Delta y\} = \mathrm{P}\{x : y < g(x) \le y + \Delta y\}, \text{ so}$$

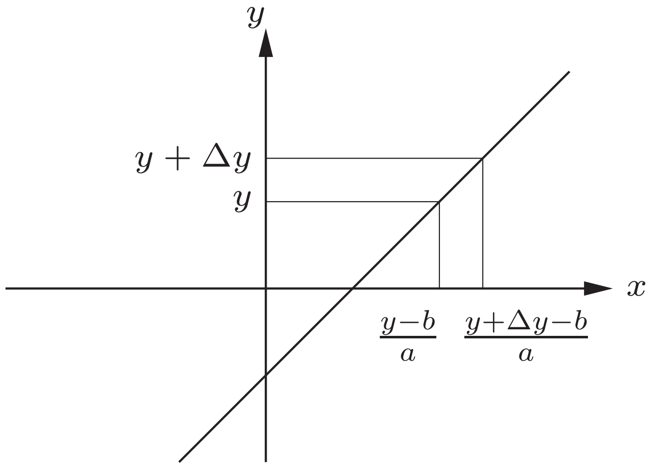
  i.e., the probability of the inverse image under $g(x)$ of $(y, y + \Delta y]$, then take limits

  $$f_Y(y) \;=\; \lim_{\Delta y \to 0} \frac{\mathrm{P}\{x : y < g(x) \le y + \Delta y\}}{\Delta y}$$



  We find $f_Y(y)$ for two examples, then give a general formula for $f_Y(y)$

- Example: *Linear function.* Let $X \sim f_X(x)$ and $Y = aX + b$, $a > 0$. Find $f_Y(y)$



$$
\begin{aligned}
f_Y(y)\Delta y &\approx \mathrm{P}\{x : y < g(x) \leq y + \Delta y\} \\
&= \mathrm{P}\left\{\frac{y-b}{a} < X \leq \frac{y-b}{a} + \frac{\Delta y}{a}\right\} \approx f_X\left(\frac{y-b}{a}\right)\frac{\Delta y}{|a|}
\end{aligned}
$$

As we let $\Delta y \to 0$, we obtain $f_Y(y) = \dfrac{1}{|a|} f_X\left(\dfrac{y-b}{a}\right)$

- Special case: $X \sim \mathcal{N}(\mu, \sigma^2)$, i.e.,

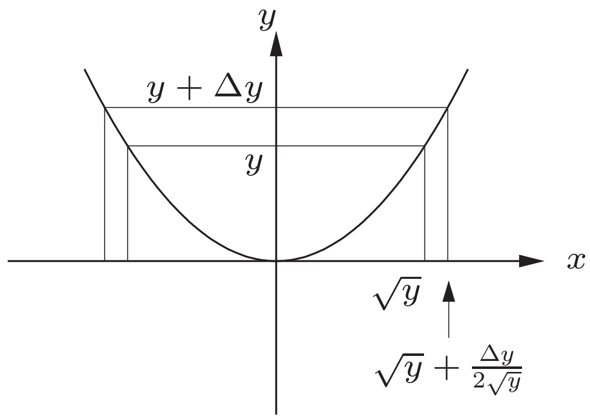
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Again setting $Y = aX + b$,

$$\begin{aligned} f_Y(y) &= \frac{1}{\mid a \mid} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{\mid a \mid} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi(a\sigma)^2}} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}, \text{ for } -\infty < y < \infty \end{aligned}$$

Therefore, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
This result can be used to compute probabilities for an arbitrary Gaussian r.v. from the distribution of the $\mathcal{N}(0,1)$ r.v. (i.e., using the $Q(.)$ function)

- Example: *Quadratic function.* Let $X \sim f_X(x)$ and $Y = X^2$. Find $f_Y(y)$



$$
\begin{aligned}
f_Y(y)\Delta y &\approx \mathrm{P}\{x : y < g(x) \le y + \Delta y\} \\
&= \mathrm{P}\left\{+\sqrt{y} < X \le +\sqrt{y} + \frac{\Delta y}{2\sqrt{y}} \quad \text{or} \; -\sqrt{y} - \frac{\Delta y}{2\sqrt{y}} < X \le -\sqrt{y}\right\} \\
&\approx \left(\frac{1}{2\sqrt{y}} f_X(+\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y})\right) \Delta y
\end{aligned}
$$

Therefore

$$
f_Y(y) = \frac{1}{2\sqrt{y}}\Big(f_X(+\sqrt{y}) + f_X(-\sqrt{y})\Big)
$$

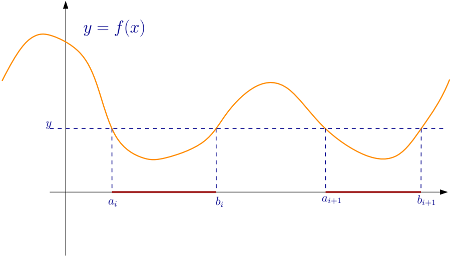
- In general, let $X \sim f_X(x)$ and $Y = g(X)$ be differentiable. Then,

$$f_Y(y) = \sum_{i=1}^{k} \frac{f_X(x_i)}{|g'(x_i)|},$$

where $x_1, x_2, \ldots$ are the solutions to the equation $y = g(x)$, and $g'(x_i)$ is the derivative of g evaluated at x_i.

Transformations (Supplementary)

- Suppose $Y = f(X)$. How do we transform one pdf to the other?



- Change of variables: suppose f is differentiable,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(f(X) \leq y) = \mathbb{P}(X \in \{x : f(x) \leq y\}) \\ &= \mathbb{P}\left(X \in \bigcup_i [a_i, b_i]\right) \\ &= \sum_i (F_X(b_i) - F_X(a_i)). \end{aligned}$$

Note that $f(b_i)$ has positive slope and $f(a_i)$ has negative slope for all i.

Transformations (Supplementary)

- Chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}y}F_X(g(y)) = \left.\frac{\mathrm{d}x}{\mathrm{d}y}\right|_{x=g(y)} \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \frac{1}{f'(g(y))}\frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \frac{p_X(x)}{f'(x)},$$

  where $x = g(y)$ or $y = f(x)$ with $f = g^{-1}$.

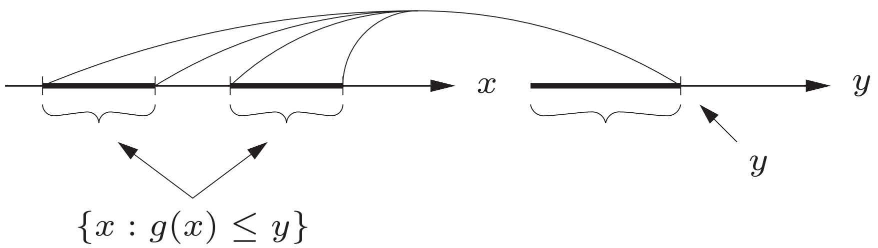
- Taking derivatives, noting that a_i, b_i are functions of y,

$$\begin{aligned} p_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) &= \sum_i \left(\frac{\mathrm{d}F_X(b_i)}{\mathrm{d}y} - \frac{\mathrm{d}F_X(a_i)}{\mathrm{d}y} \right) \\ &= \sum_k \frac{p_X(x_k)}{|f'(x_k)|}, \end{aligned}$$

  where x_k is either a_i or b_i, i.e., the solutions of the equation $y = f(x)$. [Note that f is invertible in a small interval around each x_k.]

Derived CDFs

- If the cdf of X is given and we wish to find the cdf of Y

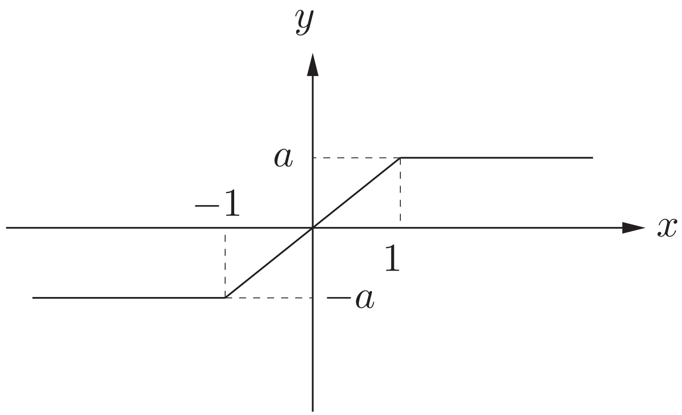


  then we use

$$F_Y(y) = \mathrm{P}\{Y \le y\} = \mathrm{P}\{x : g(x) \le y\}$$

- This method is needed in cases where X does not have a density or the function is not differentiable (or both). In some cases it is also easier to use when X has a density and g is differentiable (in this case we find $F_Y(y)$ first then take derivatives to find $f_Y(y)$)

- Example: *Limiter.* Let X be a r.v. with Laplacian pdf $f_X(x) = \frac{1}{2}e^{-|x|}$, and let Y be defined by the function of X shown in the figure. Find the cdf of Y.



Solution: To find the cdf of Y, we consider the following cases

$y < -a$: Here clearly $F_Y(y) = 0$

$y = -a$: Here

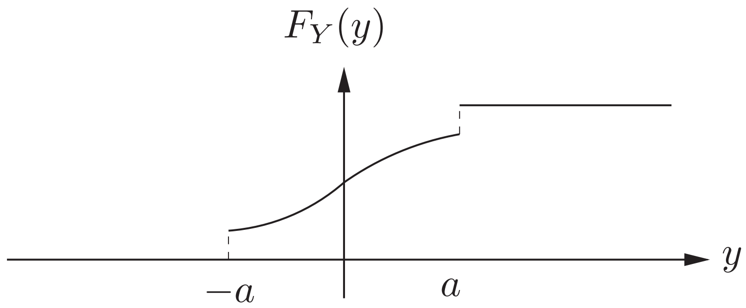
$$F_Y(-a) = F_X(-1)$$
$$= \int_{-\infty}^{-1} \frac{1}{2}e^{x}\,dx = \frac{1}{2}\cdot e^{-1}$$

- $-a < y < b$: Here

$$
\begin{aligned}
F_Y(y) &= \mathrm{P}\{Y \le y\} \\
&= \mathrm{P}\left\{aX \le y\right\} \\
&= \mathrm{P}\left\{X \le \frac{y}{a}\right\} = F_X\left(\frac{y}{a}\right) \\
&= \frac{1}{2}\cdot e^{-1} + \int_{-1}^{y/a} \frac{1}{2} e^{-|x|}\, dx
\end{aligned}
$$

$y \ge a$: Here $F_Y(y) = 1$

Combining the results, the following is a sketch of the cdf of Y



Application: Generation of Random Variables

- Generating a r.v. with a prescribed distribution is often required for performing simulations involving random phenomena, e.g., noise or packet arrivals

- First let $X \sim F(x)$ where the cdf $F(x)$ is continuous and strictly increasing. Define $Y = F(X)$, a real-valued random variable that is a function of X
  What is the cdf of Y?
  Clearly, $F_Y(y) = 0$ for $y < 0$, and $F_Y(y) = 1$ for $y > 1$
  For $0 \leq y \leq 1$, note that by assumption F has an inverse F^{-1},

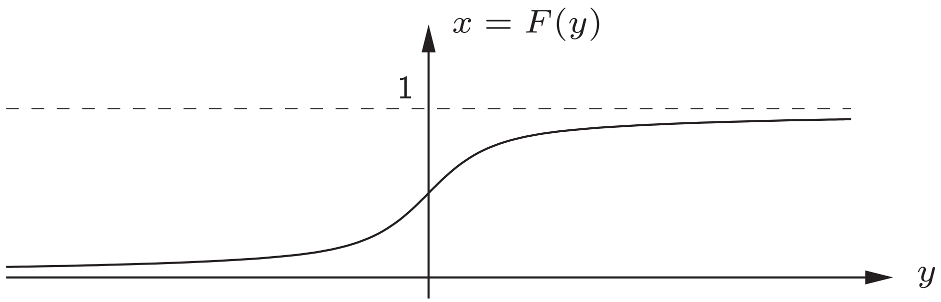
$$F_Y(y) = P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y.$$

Thus, $Y \sim U[0, 1]$, i.e., Y is a uniformly distributed random variable

- Note: $F(x)$ does not need to be invertible. If $F(x) = a$ is a constant over some interval, then the probability that X lies in this interval is zero. Without loss of generality, we can take $F^{-1}(a)$ to be the leftmost point of the interval

- Conclusion: we can generate a $U[0, 1]$ r.v. from any continuous r.v.

- In a more useful scenario where we are given $X \sim U[0,1]$ (a random number generator) and wish to generate a random variable Y with prescribed cdf $F(y)$



- If F is continuous and strictly increasing, set $Y = F^{-1}(X)$. To show $Y \sim F(y)$,

$$\begin{aligned} F_Y(y) &= \mathrm{P}\{Y \le y\} \\ &= \mathrm{P}\{F^{-1}(X) \le y\} \\ &= \mathrm{P}\{X \le F(y)\} \\ &= F(y)\,, \end{aligned}$$

since $X \sim \mathrm{U}[\,0,1\,]$ and $0 \le F(y) \le 1$

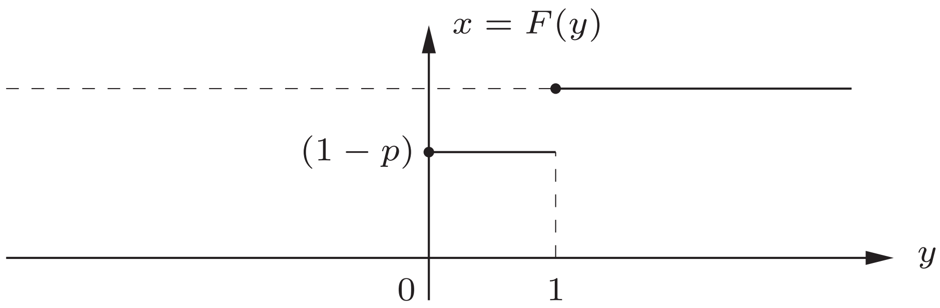
- Example: To generate $Y \sim \mathrm{Exp}(\lambda)$, set

$$Y = -\frac{1}{\lambda} \ln(1 - X)$$

- Note: F does not need to be continuous for the above to work. For example, to generate $Y \sim \mathrm{Bern}(p)$, we set

$$y = \begin{cases} 0 & x \le 1 - p \\ 1 & \text{otherwise} \end{cases}$$

This method is equivalent to defining $y = F^{-1}(x)$ where x is the smallest value such that $F(x) = y$



Lecture 3: Two Random Variables

- Joint, Marginal, and Conditional PMFs
- Joint, Marginal, and Conditional CDFs, PDFs
- One Discrete and One Continous Random Variables
- Signal Detection: MAP Rule
- Functions of Two Random Variables

Joint, Marginal, and Conditional PMFs

- Let X and Y be discrete random variables on the same probability space
- They are completely specified by their *joint pmf*:

$$p_{X,Y}(x,y) = \mathrm{P}\{X = x, Y = y\}\,, \quad x \in \mathcal{X},\, y \in \mathcal{Y}$$

  By axioms of probability, $\displaystyle\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) = 1$

- Example: Consider the pmf $p_{X,Y}(x,y)$ described by the following table

		x		
		0	1	2.5
	-3	0	$\frac{1}{4}$	$\frac{1}{8}$
y	-1	$\frac{1}{8}$	0	$\frac{1}{4}$
	2	$\frac{1}{8}$	$\frac{1}{8}$	0

- To find $p_X(x)$, the *marginal pmf* of X, we use the law of total probability

$$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y)\,, \quad x \in \mathcal{X}$$

- The *conditional pmf* of X given $Y = y$ is defined as

$$p_{X|Y}(x|y) \;=\; \frac{p_{X,Y}(x, y)}{p_Y(y)}\,, \quad p_Y(y) \neq 0,\, x \in \mathcal{X}$$

  Check that if $p_Y(y) \neq 0$ then $p_{X|Y}(x|y)$ is a pmf for X The (elementary) conditional probability of an event $X \in A$ given $Y = y$ is

$$\mathrm{P}(X \in A \mid Y = y) = \sum_{x \in A} p_{X|Y}(x|y)$$

- *Chain rule*: $p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y|x) = p_Y(y) p_{X|Y}(x|y)$

- X and Y are said to be *independent* if for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$p_{X,Y}(x, y) = p_X(x) p_Y(y)\,,$$

  which is equivalent to

$$p_{X|Y}(x|y) \;=\; p_X(x)\,, \quad p_Y(y) \neq 0,\, x \in \mathcal{X}$$

Bayes Rule for PMFs

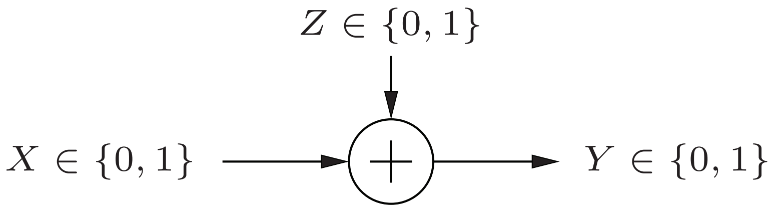
Given $p_X(x)$ and $p_{Y|X}(y|x)$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we can find $p_{X|Y}(x|y)$:

$$
\begin{aligned}
p_{X|Y}(x|y) &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \\
&= \frac{p_X(x) p_{Y|X}(y|x)}{p_Y(y)} \\
&= \frac{p_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} p_{X,Y}(x', y)} p_X(x) \\
&= \frac{p_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} p_{Y|X}(y|x') p_X(x')} p_X(x)
\end{aligned}
$$

The final formula is entirely in terms of the known quantities $p_X(x)$ and $p_{Y|X}(y|x)$

Example: Binary Symmetric Channel

Consider the following binary communication channel



The bit sent is $X \sim \mathrm{Bern}(p)$, $0 \le p \le 1$, the noise is $Z \sim \mathrm{Bern}(\epsilon)$, $0 \le \epsilon \le 0.5$, the bit received is $Y = (X + Z) \bmod 2 = X \oplus Z$, and X and Z are independent

Find

1. $p_{X|Y}(x|y)$
2. $p_Y(y)$
3. $\mathrm{P}\{X \neq Y\}$, the probability of error

- To find $p_{X|Y}(x|y)$, we use Bayes rule

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} p_{Y|X}(y|x') p_X(x')} p_X(x)$$

We know $p_X(x)$, but we need to find $p_{Y|X}(y|x)$:

$$\begin{aligned} p_{Y|X}(y|x) &= \mathrm{P}\{Y = y \,|\, X = x\} = \mathrm{P}\{X \oplus Z = y \,|\, X = x\} \\ &= \mathrm{P}\{x \oplus Z = y \,|\, X = x\} = \mathrm{P}\{Z = y \oplus x \,|\, X = x\} \\ &= \mathrm{P}\{Z = y \oplus x\} \quad \text{since } Z \text{ and } X \text{ are independent} \\ &= p_Z(y \oplus x) \end{aligned}$$

Therefore

$$\begin{aligned} p_{Y|X}(0 \,|\, 0) &= p_Z(0 \oplus 0) = p_Z(0) = 1 - \epsilon \\ p_{Y|X}(0 \,|\, 1) &= p_Z(0 \oplus 1) = p_Z(1) = \epsilon \\ p_{Y|X}(1 \,|\, 0) &= p_Z(1 \oplus 0) = p_Z(1) = \epsilon \\ p_{Y|X}(1 \,|\, 1) &= p_Z(1 \oplus 1) = p_Z(0) = 1 - \epsilon \end{aligned}$$

Plugging into the Bayes rule equation, we obtain

$$p_{X|Y}(0|0) = \frac{p_{Y|X}(0|0)}{p_{Y|X}(0|0)p_X(0) + p_{Y|X}(0|1)p_X(1)}p_X(0) = \frac{(1-\epsilon)(1-p)}{(1-\epsilon)(1-p)+\epsilon p}$$

$$p_{X|Y}(1|0) = 1 - p_{X|Y}(0|0) = \frac{\epsilon p}{(1-\epsilon)(1-p)+\epsilon p}$$

$$p_{X|Y}(0|1) = \frac{p_{Y|X}(1|0)}{p_{Y|X}(1|0)p_X(0) + p_{Y|X}(1|1)p_X(1)}p_X(0) = \frac{\epsilon(1-p)}{(1-\epsilon)p+\epsilon(1-p)}$$

$$p_{X|Y}(1|1) = 1 - p_{X|Y}(0|1) = \frac{(1-\epsilon)p}{(1-\epsilon)p+\epsilon(1-p)}$$

2. We already found $p_Y(y)$ as

$$\begin{aligned} p_Y(y) &= p_{Y|X}(y|0)p_X(0) + p_{Y|X}(y|1)p_X(1) \\ &= \begin{cases} (1-\epsilon)(1-p)+\epsilon p & \text{for } y = 0 \\ \epsilon(1-p)+(1-\epsilon)p & \text{for } y = 1 \end{cases} \end{aligned}$$

3. Now to find the probability of error $\mathrm{P}\{X \neq Y\}$, consider

$$\begin{aligned}\mathrm{P}\{X \neq Y\} &= p_{X,Y}(0,1) + p_{X,Y}(1,0) \\ &= p_{Y|X}(1|0)p_X(0) + p_{Y|X}(0|1)p_X(1) \\ &= \epsilon(1-p) + \epsilon p \;=\; \epsilon\end{aligned}$$

An interesting special case is $\epsilon = \frac{1}{2}$. Here, $\mathrm{P}\{X \neq Y\} = \frac{1}{2}$, which is the worst possible (no information is sent), and

$$p_Y(0) = \tfrac{1}{2}p + \tfrac{1}{2}(1-p) = \tfrac{1}{2} = p_Y(1)$$

Therefore $Y \sim \mathrm{Bern}(\frac{1}{2})$, independent of the value of p !

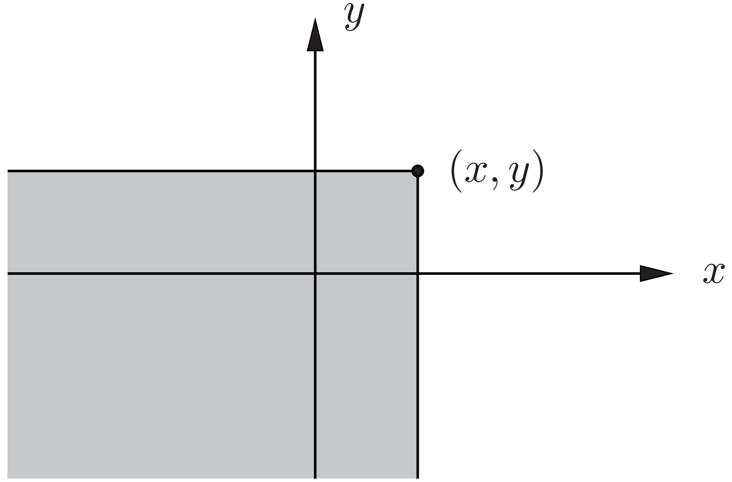
In this case, the bit sent X and the bit received Y are independent (check this)

Joint and Marginal CDF and PDF

- Any two random variables are specified by their *joint cdf*

$$F_{X,Y}(x, y) = \mathrm{P}\{X \le x,\, Y \le y\}\,, \quad x, y \in \mathbf{R}$$

$F_{X,Y}(x, y)$ is the probability of the shaded region of $\mathbf{R}^2$

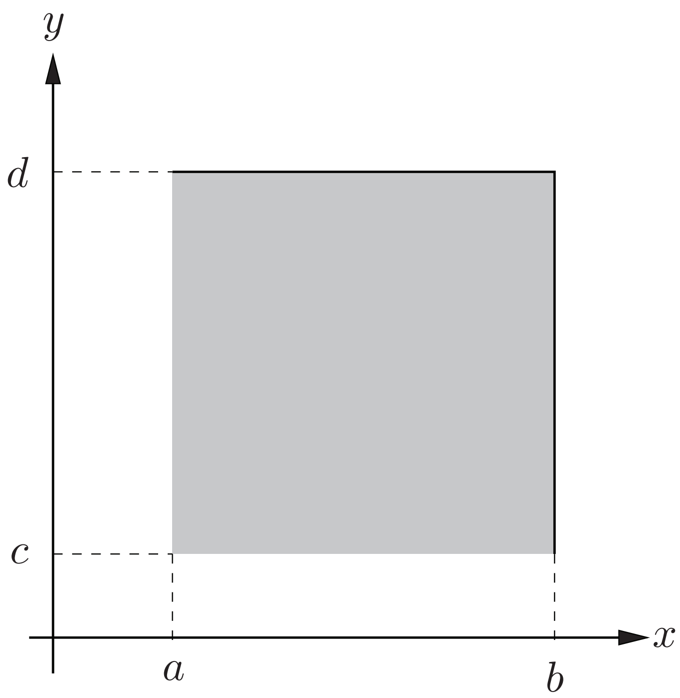


- Properties of the cdf:
  - $F_{X,Y}(x, y) \geq 0$
  - If $x_1 \leq x_2$ and $y_1 \leq y_2$ then $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$
  - $\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1$
  - $\lim_{y\to-\infty} F_{X,Y}(x, y) = 0$ and $\lim_{x\to-\infty} F_{X,Y}(x, y) = 0$
  - $\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x)$ and $\lim_{x\to\infty} F_{X,Y}(x, y) = F_Y(y)$

    This is often abbreviated to $F_{X,Y}(x, \infty) = F_X(x)$, $F_{X,Y}(\infty, y) = F_Y(y)$.

    $F_X(x)$ and $F_Y(y)$ are called the *marginal cdfs* of X and Y

- The probability of any rectangular set can be determined from the joint cdf



For example,

$$\mathrm{P}\{a < X \leq b,\, c < Y \leq d\} = F(b,d) - F(a,d) - F(b,c) + F(a,c)$$

- X and Y are *independent* if for every x and y

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

- X and Y are jointly *continuous* random variables if their joint cdf is continuous in both x and y

  In this case, we can define their *joint pdf*, provided that it exists, as the function $f_{X,Y}(x,y)$ such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,du\,dv\,, \quad x,y \in \mathbf{R}$$

- If $F_{X,Y}(x,y)$ is differentiable in x and y, then

$$f_{X,Y}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = \lim_{\Delta x, \Delta y \to 0} \frac{\mathrm{P}\{x < X \le x + \Delta x,\; y < Y \le y + \Delta y\}}{\Delta x \Delta y}$$

- Properties of $f_{X,Y}(x,y)$:
  - $f_{X,Y}(x,y) \geq 0$
  - $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
  - The probability of any set $A \subset \mathbf{R}$ can be calculated by integrating the joint pdf over A:

$$\mathrm{P}\{(X,Y) \in A\} = \int_{(x,y)\in A} f_{X,Y}(x,y)\,dx\,dy$$

- The *marginal pdf* of X can be obtained from the joint pdf via the law of total probability:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

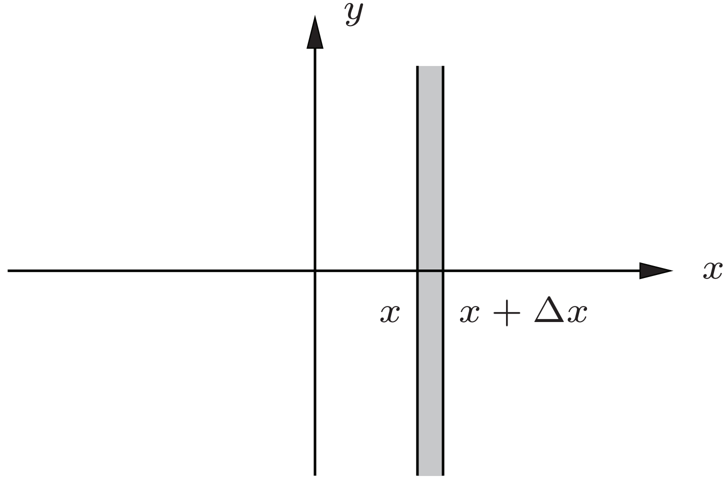
To see this,

$$\begin{aligned}
F_X(x) &= P\{X \leq x\} = P\{X \leq x, Y \leq \infty\} \\
& \quad \infty \\
&= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x',y)dx'dy
\end{aligned}$$

and hence differentiating yields

$$f_X(x) = \frac{d}{dx}\left(\int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x', y)dx'dy\right) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Alternatively, consider the figure on the next page



$$
\begin{aligned}
f_X(x) &= \lim_{\Delta x \to 0} \frac{\mathrm{P}\{x < X \le x + \Delta x\}}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \lim_{\Delta y \to 0} \sum_{n=-\infty}^{\infty} \mathrm{P}\{x < X \le x + \Delta x,\, n\Delta y < Y \le (n+1)\Delta y\} \\
&= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, \Delta x = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy
\end{aligned}
$$

- X and Y are independent iff $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for every x, y

Example

- Let $(X, Y) \sim f(x, y)$, where

$$f(x,y) = \begin{cases} c & x \geq 0,\, y \geq 0,\, x + y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find c
2. Find $f_Y(y)$
3. Are X and Y independent?
4. Find $\mathrm{P}\{X \geq \frac{1}{2}Y\}$

Solution:

1. To find c, note that

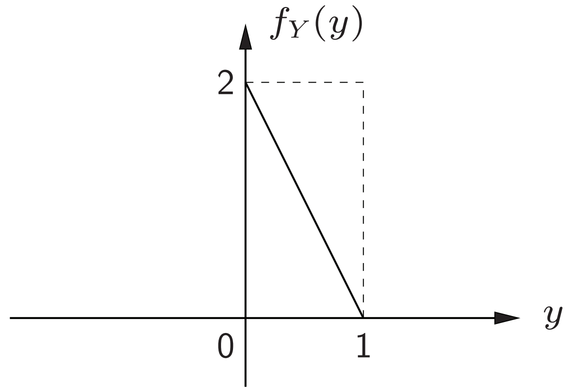
$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy \;=\; \int_0^1\int_0^{1-y} c\,dx\,dy = c\int_0^1 (1-y)\,dy = \tfrac{1}{2}c\,,$$

hence $c = 2$

2. To find $f_Y(y)$, we use the law of total probability

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

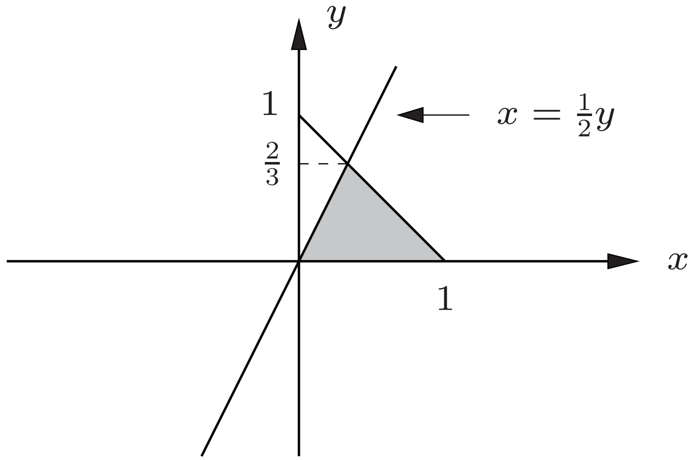
$$= \begin{cases} \int_0^{(1-y)} 2\,dx & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 2(1-y) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$



3. X and Y are independent if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for every x, y

But $f_{X,Y}(0,1) = 2$ and $f_X(0)f_Y(1) = 0$, so X and Y are *not* independent

Another example: $f_{X,Y}(\frac{1}{2},\frac{1}{2}) = 2 \ne 1 \cdot 1 = f_X(\frac{1}{2})f_Y(\frac{1}{2})$



From the figure we find that

$$\mathrm{P}\left\{X \geq \tfrac{1}{2}Y\right\} = \int_{\left\{(x,y):x\geq\frac{1}{2}y\right\}} f_{X,Y}(x,y)\,dx\,dy$$

$$= \int_0^{\frac{2}{3}} \int_{\frac{y}{2}}^{(1-y)} 2\,dx\,dy \;=\; \frac{2}{3}$$

Conditional CDF and PDF

- Let X and Y be continuous random variables with joint pdf $f_{X,Y}(x,y)$. We wish to define $F_{Y|X}(y\,|\,X=x) = \mathrm{P}\{Y \le y\,|\,X=x\}$
- We cannot define the above conditional probability as

$$\frac{\mathrm{P}\{Y \le y,\, X=x\}}{\mathrm{P}\{X=x\}}$$

because both numerator and denominator are equal to zero. Instead, we define conditional probability for continuous random variables as a limit

$$\begin{aligned} F_{Y|X}(y|x) &= \lim_{\Delta x \to 0} \mathrm{P}\{Y \le y\,|\,x < X \le x+\Delta x\} \\ &= \lim_{\Delta x \to 0} \frac{\mathrm{P}\{Y \le y,\, x < X \le x+\Delta x\}}{\mathrm{P}\{x < X \le x+\Delta x\}} \\ &= \lim_{\Delta x \to 0} \frac{\int_{-\infty}^{y} f_{X,Y}(x,u)\,du\,\Delta x}{f_X(x)\Delta x} = \int_{-\infty}^{y} \frac{f_{X,Y}(x,u)}{f_X(x)}\,du \end{aligned}$$

- Define the conditional pdf as $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ if $f_X(x) \neq 0$

- Thus
$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(u|x)\,du$$
which shows that $f^{Y|X}(y|x)$ is a pdf for Y given $X = x$, i.e.,
$$Y \mid \{X = x\} \sim f^{Y|X}(y|x)$$

Note: Conditional distributions are in fact not defined in this way in advanced probability, in particular they are not defined using limits. Like Dirac delta functions, they are defined by their behavior inside integrals — suitable integrals of conditional pdfs yield elementary conditional probabilities.

- Example: Let $f(x, y)$ be defined as
$$f(x,y) = \begin{cases} 2 & x \geq 0,\ y \geq 0,\ x + y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
Find $f_{X|Y}(x|y)$

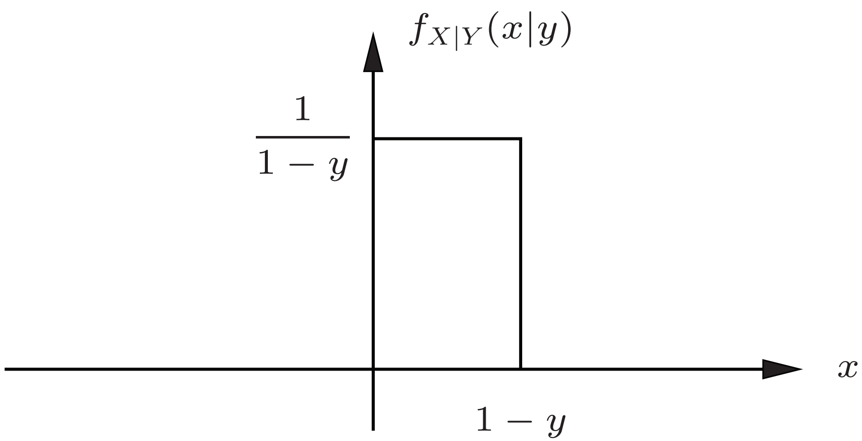
Solution: We already know that

$$f_Y(y) = \begin{cases} 2(1-y) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$f_{X|Y}(x|y) \;=\; \frac{f_{X,Y}(x,y)}{f_Y(y)} \;=\; \begin{cases} \dfrac{1}{1-y} & 0 \le y < 1,\ 0 \le x \le 1-y \\ 0 & \text{otherwise} \end{cases}$$

In other words, $X \mid \{Y = y\} \sim \mathrm{U}[\,0, 1-y\,]$



- Bayes rule for densities: Given $f_X(x)$ and $f_{Y|X}(y|x)$, we can find

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

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

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

- Example: Let $\Lambda \sim \mathrm{U}[0,1]$, and let the conditional pdf of X given $\Lambda = \lambda$ be

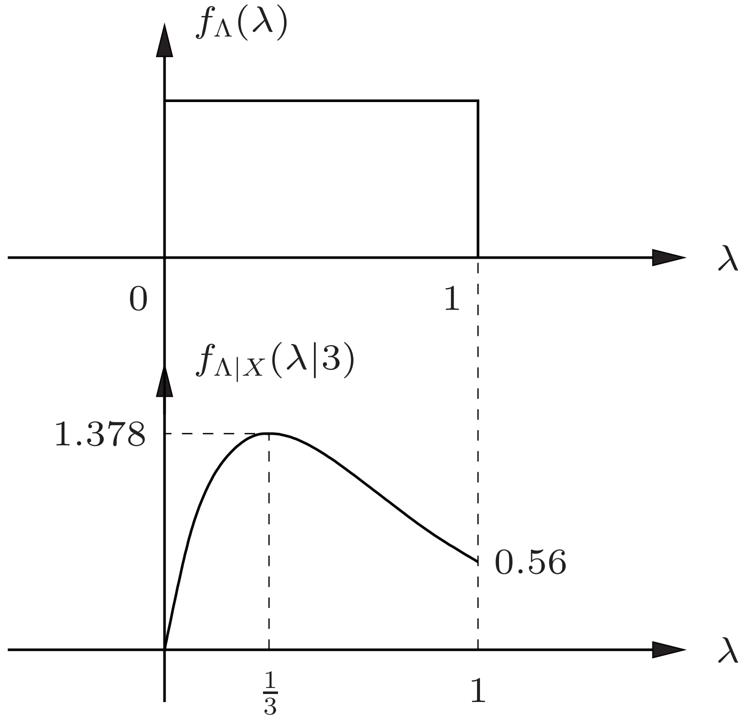
$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}\,, \quad 0 < \lambda \le 1\,,$$

i.e., $X \mid \{\Lambda = \lambda\} \sim \mathrm{Exp}(\lambda)$

Given $X = 3$, find $f_{\Lambda|X}(\lambda|3)$

Solution: Use Bayes rule

$$f_{\Lambda|X}(\lambda|3) = \frac{f_{X|\Lambda}(3|\lambda) f_\Lambda(\lambda)}{\int_0^1 f_\Lambda(u) f_{X|\Lambda}(3|u)\, du} = \begin{cases} \dfrac{\lambda e^{-3\lambda}}{\frac{1}{9}(1-4e^{-3})} & 0 < \lambda \le 1 \\ 0 & \text{otherwise} \end{cases}$$



Mixed Random Variables

- Let Θ be a discrete random variable with pmf $p_\Theta(\theta)$
- For each $\Theta = \theta$ with $p_\Theta(\theta) \neq 0$, let Y be a continuous random variable, i.e., $F_{Y|\Theta}(y|\theta)$ is continuous for all θ. We define $f_{Y|\Theta}(y|\theta)$ in the usual way
- The conditional pmf of Θ given y can be defined as a limit

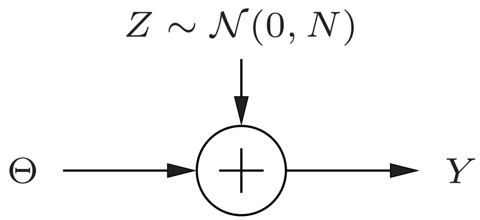
$$
\begin{aligned}
p_{\Theta|Y}(\theta \,|\, y) &= \lim_{\Delta y \to 0} \frac{\mathrm{P}\{\Theta = \theta,\, y < Y \leq y + \Delta y\}}{\mathrm{P}\{y < Y \leq y + \Delta y\}} \\
&= \lim_{\Delta y \to 0} \frac{p_\Theta(\theta) f_{Y|\Theta}(y|\theta)\Delta y}{f_Y(y)\Delta y} = \frac{f_{Y|\Theta}(y|\theta)}{f_Y(y)} p_\Theta(\theta)
\end{aligned}
$$

- So we obtain yet another version of Bayes rule: Given $p_\Theta(\theta)$ and $f_{Y|\Theta}(y|\theta)$, then

$$
p_{\Theta|Y}(\theta \,|\, y) = \frac{f_{Y|\Theta}(y|\theta)}{\sum_{\theta'} p_\Theta(\theta') f_{Y|\Theta}(y|\theta')} p_\Theta(\theta)
$$

- Example: *Additive Gaussian Noise Channel*

  Consider the following communication channel:



  The signal transmitted is a binary random variable Θ:

$$\Theta = \begin{cases} +1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$$

  The received signal, also called the *observation*, is $Y = \Theta + Z$, where Θ and Z are independent

  Given $Y = y$ is received (observed), find $p_{\Theta|Y}(\theta|y)$, the a posteriori pmf of Θ

Solution: We use Bayes rule

$$p_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)}{\sum_{\theta'} p_{\Theta}(\theta') f_{Y|\Theta}(y|\theta')} p_{\Theta}(\theta)$$

We are given $p_{\Theta}(\theta)$:

$$p_{\Theta}(+1) \; = \; p \quad \text{and} \quad p_{\Theta}(-1) \; = \; 1 - p$$

and $f_{Y|\Theta}(y|\theta) = f_Z(y - \theta)$:

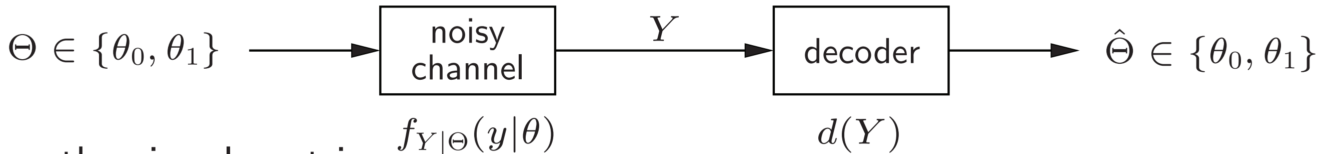
$$Y \,|\, \{\Theta = +1\} \sim \mathcal{N}(+1, N) \quad \text{and} \quad Y \,|\, \{\Theta = -1\} \sim \mathcal{N}(-1, N)$$

Therefore

$$p_{\Theta|Y}(1|y) \; = \; \frac{\frac{p}{\sqrt{2\pi N}} e^{-\frac{(y-1)^2}{2N}}}{\frac{p}{\sqrt{2\pi N}} e^{-\frac{(y-1)^2}{2N}} + \frac{(1-p)}{\sqrt{2\pi N}} e^{-\frac{(y+1)^2}{2N}}} \; = \; \frac{p e^{\frac{y}{N}}}{p e^{\frac{y}{N}} + (1-p) e^{-\frac{y}{N}}}$$

Signal Detection

- Consider the following general digital communication system



  where the signal sent is

$$\Theta = \begin{cases} \theta_0 & \text{with probability } p \\ \theta_1 & \text{with probability } 1-p \end{cases}$$

  and the observation (received signal) is

$$Y \mid \{\Theta = \theta\} \sim f_{Y|\Theta}(y \mid \theta)\,, \quad \theta \in \{\theta_0, \theta_1\}$$

- We wish to find an optimal decoder (or optimal receiver), $d(Y)$, that minimizes the *probability of error*:

$$\begin{aligned} P_e &\stackrel{\triangle}{=} \mathrm{P}\{\hat{\Theta} \neq \Theta\} = \mathrm{P}\{\Theta = \theta_0,\ \hat{\Theta} = \theta_1\} + \mathrm{P}\{\Theta = \theta_1,\ \hat{\Theta} = \theta_0\} \\ &= \mathrm{P}\{\Theta = \theta_0\}\mathrm{P}\{\hat{\Theta} = \theta_1 \mid \Theta = \theta_0\} + \mathrm{P}\{\Theta = \theta_1\}\mathrm{P}\{\hat{\Theta} = \theta_0 \mid \Theta = \theta_1\} \end{aligned}$$

- We define the *maximum a posteriori probability* (MAP) decoder as

$$d(y) = \begin{cases} \theta_0 & \text{if } p_{\Theta|Y}(\theta_0|y) > p_{\Theta|Y}(\theta_1|y) \\ \theta_1 & \text{otherwise} \end{cases}$$

- The MAP decoding rule minimizes P_e, since

$$\begin{aligned} P_e &= 1 - \mathrm{P}\{d(Y) = \Theta\} \\ &= 1 - \int_{-\infty}^{\infty} f_Y(y)\mathrm{P}\{d(y) = \Theta \,|\, Y = y\}\, dy \end{aligned}$$

  and the integral is maximized when we pick the largest $\mathrm{P}\{d(y) = \Theta \,|\, Y = y\}$ for each y, which is precisely the MAP decoder

- If $p = \frac{1}{2}$, i.e., equally likely signals, using Bayes rule, the MAP decoder reduces to the *maximum likelihood* (ML) decoder

$$d(y) = \begin{cases} \theta_0 & \text{if } f_{Y|\Theta}(y|\theta_0) > f_{Y|\Theta}(y|\theta_1) \\ \theta_1 & \text{otherwise} \end{cases}$$

Additive Gaussian Noise Channel

- Consider the additive Gaussian noise channel with signal

$$\Theta = \begin{cases} +\sqrt{P} & \text{with probability } \frac{1}{2} \\ -\sqrt{P} & \text{with probability } \frac{1}{2} \end{cases}$$

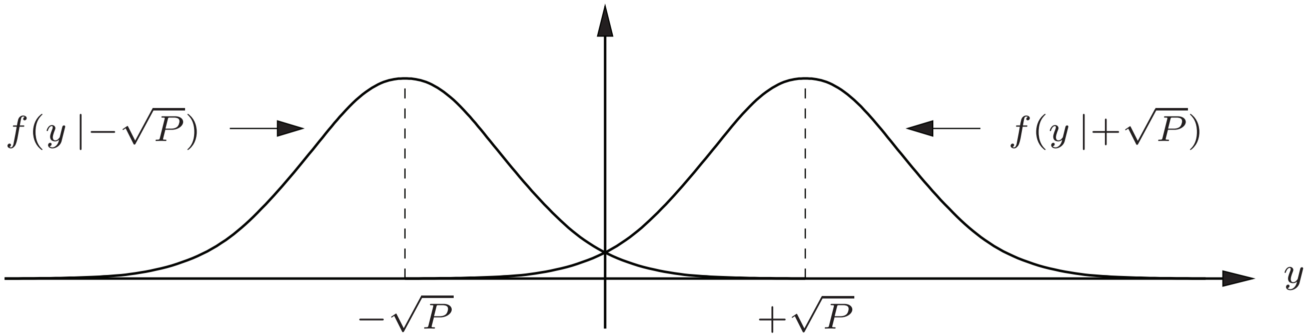
  noise $Z \sim \mathcal{N}(0, N)$ (Θ and Z are independent), and output $Y = \Theta + Z$

- The MAP decoder is

$$d(y) = \begin{cases} +\sqrt{P} & \text{if } \dfrac{\mathrm{P}\{\Theta = +\sqrt{P} \mid Y = y\}}{\mathrm{P}\{\Theta = -\sqrt{P} \mid Y = y\}} > 1 \\ -\sqrt{P} & \text{otherwise} \end{cases}$$

  Since the two signals are equally likely, the MAP decoding rule reduces to the ML decoding rule

$$d(y) = \begin{cases} +\sqrt{P} & \text{if } \dfrac{f_{Y|\Theta}(y \mid +\sqrt{P})}{f_{Y|\Theta}(y \mid -\sqrt{P})} > 1 \\ -\sqrt{P} & \text{otherwise} \end{cases}$$



- From the figure above and using the Gaussian pdf, the MAP decoder reduces to the *minimum distance decoder*

$$d(y) = \begin{cases} +\sqrt{P} & (y-\sqrt{P})^2 < (y-(-\sqrt{P}))^2 \\ -\sqrt{P} & \text{otherwise} \end{cases}$$

which simplifies to

$$d(y) = \begin{cases} +\sqrt{P} & y > 0 \\ -\sqrt{P} & y < 0 \end{cases}$$

Note: The decision when $y = 0$ is arbitrary

- Now to find the *minimum* probability of error, consider

$$\begin{aligned}
P_e &= \mathrm{P}\{d(Y) \neq \Theta\} \\
&= \mathrm{P}\{\Theta = \sqrt{P}\}\mathrm{P}\{d(Y) = -\sqrt{P} \,|\, \Theta = \sqrt{P}\} + \\
&\qquad \mathrm{P}\{\Theta = -\sqrt{P}\}\mathrm{P}\{d(Y) = \sqrt{P} \,|\, \Theta = -\sqrt{P}\} \\
&= \tfrac{1}{2}\mathrm{P}\{Y \leq 0 \,|\, \Theta = \sqrt{P}\} + \tfrac{1}{2}\mathrm{P}\{Y > 0 \,|\, \Theta = -\sqrt{P}\} \\
&= \tfrac{1}{2}\mathrm{P}\{Z \leq -\sqrt{P}\} + \tfrac{1}{2}\mathrm{P}\{Z > \sqrt{P}\} \\
&= Q\left(\sqrt{\frac{P}{N}}\right) = Q\left(\sqrt{\mathrm{SNR}}\right)
\end{aligned}$$

  The probability of error is a decreasing function of P/N, the *signal-to-noise ratio* (SNR)

Functions of Two Random Variables

- Let $(X,Y) \sim f(x,y)$ and let $g(x,y)$ be a differentiable function. To find the pdf of $Z = g(X,Y)$, we first find the inverse image of $\{z < Z \le z + \Delta z\}$ then find its probability expressed as a function of z and Δz

- Example: Let X and Y be independent r.v.s, with $X \sim f_X(x)$, and $Y \sim f_Y(y)$. Find the pdf of $Z = X + Y$

- Solution 1: Using approximation and limit arguments

$$\begin{aligned}
f_Z(z)\Delta z &\approx \mathrm{P}\{z < Z \le z + \Delta z\} \\
&= \int_{-\infty}^{\infty} \mathrm{P}\{z < X + Y \le z + \Delta z \,|\, X = x\} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \mathrm{P}\{z < x + Y < z + \Delta z \,|\, X = x\} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \mathrm{P}\{z - x < Y < z - x + \Delta z\} f_X(x)\, dx \\
&\approx \int_{-\infty}^{\infty} f_Y(z - x)\Delta z f_X(x)\, dx
\end{aligned}$$

Letting $\Delta z \to 0$,

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\, dx\,,$$

which is the *convolution* of $f_X(x)$ and $f_Y(y)$

For example, if $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then it can be shown that

$$Z = X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

- Solution 2: Direct derived distribution:

Independence of X and Y implies that the joint pdf is $f_{X,Y}(x,y) = f_X(x) f_Y(y)$. To find pdf f_Z, first find the cdf $F_Z(z)$ and then differentiate.

$$\begin{aligned}
F_Z(z) &= \Pr(Z \le z) = \Pr(X + Y \le z) \\
&= \int_{x,y:x+y\le z} f_{X,Y}(x,y) dx dy \\
&= \int_{x,y:x+y\le z} f_X(x) f_Y(y) dx dy = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f_X(x) dx \right) f_Y(y) dy
\end{aligned}$$

The pdf is then found by differentiation using the rule for differentiating integrals with respect to variables which appear only in the limits:

$$\begin{aligned}
f_Z(z) &= \frac{dF_Z(z)}{dz} = \frac{d}{dz}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} f_X(x)dx\right) f_Y(y)dy \\
&= \int_{-\infty}^{\infty}\left(\frac{d}{dz}\int_{-\infty}^{z-y} f_X(x)dx\right) f_Y(y)dy \\
&= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)dy
\end{aligned}$$

- The convolution result also holds for the sum of two independent discrete random variables (replacing pdfs with pmfs and integrals with sums)

  For example, if $X \sim \mathrm{Poisson}(\lambda_1)$ and $Y \sim \mathrm{Poisson}(\lambda_2)$ are independent, then $Z = X + Y \sim \mathrm{Poisson}(\lambda_1) * \mathrm{Poisson}(\lambda_2) = \mathrm{Poisson}(\lambda_1 + \lambda_2)$

- The property that the sum of two independent r.v.s with the same distribution has the same distribution, which is obeyed by Gaussian and Poisson r.v.s, is referred to as *infinite divisibility*

  E.g., a Poisson r.v. with parameter λ can be written as the sum of *any* number of independent $\mathrm{Poisson}(\lambda_i)$ r.v.s, so long as $\sum_i \lambda_i = \lambda$

- Example: *Minimum and Maximum of Independent Random Variables*

  Let $X \sim f_X(x)$ and $Y \sim f_Y(y)$ be independent. Define

$$U = \max\{X, Y\} \quad \text{and} \quad V = \min\{X, Y\}$$

  Find the pdfs of U and V

  Solution: To find the pdf of U, we first find its cdf:

$$F_U(u) \,=\, \mathrm{P}\{U \le u\} \,=\, \mathrm{P}\{X \le u,\, Y \le u\} \,=\, F_X(u)F_Y(u)$$

  Using the product rule for derivatives,

$$f_U(u) \,=\, f_X(u)F_Y(u) \,+\, f_Y(u)F_X(u)$$

  Now to find the pdf of V,

$$\mathrm{P}\{V > v\} = \mathrm{P}\{X > v,\, Y > v\} \;\Rightarrow\; 1 - F_V(v) = (1 - F_X(v))(1 - F_Y(v))$$

  Thus

$$f_V(v) \,=\, f_X(v) + f_Y(v) \,-\, f_X(v)F_Y(v) \,-\, f_Y(v)F_X(v)$$