Review of Basic Probability Theory: Part 1

Yuxin Chen

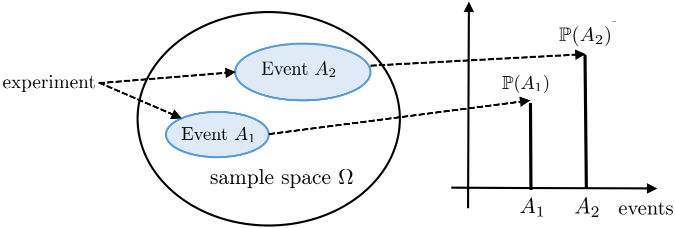
Princeton University, Fall 2018

Outline

- Sample space and probability function
- Conditioning
- Independence
- Random variables
- Functions of random variables
  - Application: generating random variables
- Joint distributions

Elements of probability models

A formal mathematical description of an uncertain situation



- **Sample space** Ω: set of all possible outcomes of an experiment
- **Set of events** $\mathcal{F}$: each event is a subset of Ω, i.e. a collection of possible outcomes
- **Probability function** $\mathbb{P}$

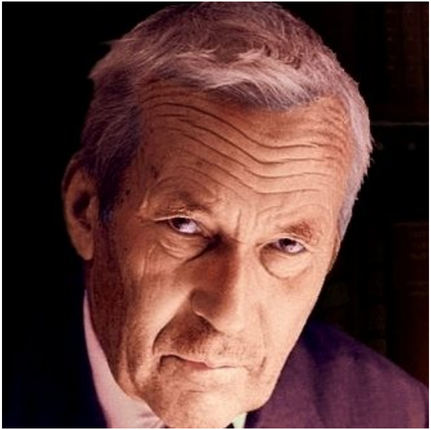
Examples

- Flip a coin 2 times: $\Omega = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}, \mathsf{TT}\}$
  - A possible event: get exactly one H $\quad (A = \{\mathsf{HT}, \mathsf{TH}\})$

- Generate a random number between 0 and 1: $\Omega = [0, 1]$
  - A possible event: get a number between 0.3 and 0.5 $(A = [0.3, 0.5])$

Axioms of probability

Axiomatize the theory to enable a firm mathematical footing



Andrey Kolmogorov introduced axiom system for probability in 1933

Axioms of probability

A **probability function** $\mathbb{P}$ is an assignment of "likelihood" to every event such that:

- (Non-negativity) $\mathbb{P}(A) \geq 0$ for every event A
- (Additivity) If $A_1, A_2, \cdots$ are disjoint events, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots$$

- (Normalization) $\mathbb{P}(\Omega) = 1$, i.e. probability of entire space is 1

Analogy: $\mathbb{P}$ is a measure much like mass, length, volume, ...

Remark: these axioms make no attempt to tell what particular function $\mathbb{P}$ to choose; it merely requires $\mathbb{P}$ to satisfy these properties

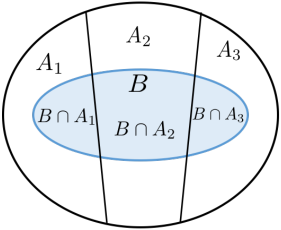
Conditional probability

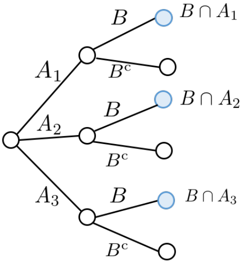
For any events A and B such that $\mathbb{P}(B) \neq 0$, the **conditional probability of A given B** is

$$\mathbb{P}(A \mid B) \stackrel{\text{def}}{=} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- provides a way to reason about outcome of an experiment based on partial information
- think of B as partial observation (e.g. a medical test on a patient), and A as complete outcome (e.g. whether a patient is sick); then $\mathbb{P}(A \mid B)$ is the prob. of A from observer's viewpoint
- **chain rule:** $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$

Law of total probability





Let $A_1, \cdots, A_n$ be *disjoint* events that partition sample space (i.e. $A_1 \cup \cdots \cup A_n = \Omega$, and $A_i \cap A_j = \emptyset$). Then for any event B,

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A_1 \cap B) + \cdots + \mathbb{P}(A_n \cap B) \\ &= \underbrace{\mathbb{P}(B \mid A_1)\, \mathbb{P}(A_1) + \cdots + \mathbb{P}(B \mid A_n)\, \mathbb{P}(A_n)}_{\text{a ``divide-and-conquer'' approach}} \end{aligned}$$

Bayes' rule

Let $A_1, \cdots, A_n$ be *disjoint* events that partition sample space, and assume $\mathbb{P}(A_i) > 0$ for all i. Then for any event B s.t. $\mathbb{P}(B) > 0$:

$$\begin{aligned}
\mathbb{P}(A_i \mid B) &= \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\mathbb{P}(B)} && \text{(turning around conditional prob.)} \\
&= \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\sum_{i=1}^n \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)} && \text{(law of total probability)}
\end{aligned}$$

- A way to reverse the order of conditioning
- Useful when one wants to calculate $\mathbb{P}(A_i \mid B)$ but is given $\mathbb{P}(B \mid A_i)$ instead

Inference via Bayes' rule

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)} \;\rightarrow\; \text{normalization}$$

posterior prob. $\propto$ prior prob. $\times$ likelihood

Suppose B is a certain *"effect"*; A_i's are possible *"causes"*

- **prior probability** $\mathbb{P}(A_i)$
- **likelihood** $\mathbb{P}(B \mid A_i)$: the probability that B is observed when cause A_i is present
- **posterior probability** $\mathbb{P}(A_i \mid B)$: given B is observed, the probability that cause A_i is present

Independence

Two events A and B are said to be independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

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

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

- Says that occurrence of B reveals absolutely no information on the probability of A occurring

Conditional independence

Given an event C, two events A and B are said to be conditionally independent if

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$$

When $\mathbb{P}(B \mid C) \neq 0$, this is equivalent to

$$\mathbb{P}(A \mid B \cap C) = \mathbb{P}(A \mid C)$$

- Says that if C is known to have occurred, additional knowledge about B reveals no new information about whether A occurs
- Conditional independence does NOT imply independence; and vice versa (Examples 1.20 & 1.21 of Bertsekas' book)

Independence does NOT imply conditional independence

Example: consider 2 independent fair coin tosses, and let

$$\begin{aligned} H_1 &= \{\text{1st toss is H}\} \\ H_2 &= \{\text{2nd toss is H}\} \\ D &= \{\text{the 2 tosses have different results}\} \end{aligned}$$

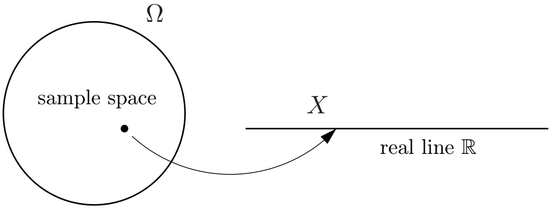
While H_1 and H_2 are independent, one has

$$\mathbb{P}(H_1 \mid D) = \frac{1}{2}, \quad \mathbb{P}(H_2 \mid D) = \frac{1}{2}, \quad \mathbb{P}(H_1 \cap H_2 \mid D) = 0$$

implying that H_1 and H_2 are NOT conditionally independent (as $\mathbb{P}(H_1 \cap H_2 \mid D) \neq \mathbb{P}(H_1 \mid D)\, \mathbb{P}(H_2 \mid D)$)

Random variables

A *random variable* X is a function from a sample space Ω into real numbers, i.e. a real-valued function of experimental outcome



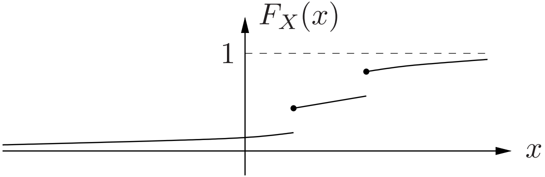
- Example: in an experiment of 2 rolls of dice, X can be sum of two rolls, or the second roll raised to the 5th power
- Often use upper case letters for r.v.s $X, Y, Z, \cdots$
- Often use lower case letters for *values* of r.v.s: $X = x$ means r.v. X takes on value x

Cumulative distribution function (CDF)

The *cumulative distribution function* of a random variable X is

$$F_X(x) \stackrel{\text{def}}{=} \mathbb{P}(X \le x), \qquad \text{for all } x \in \mathbb{R}$$

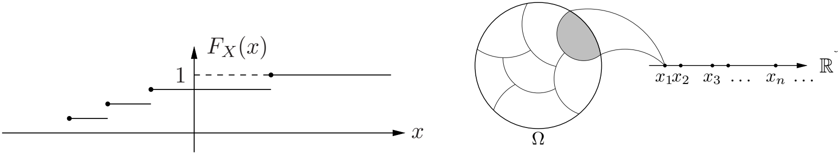
— $F_X(x)$ "accumulates" probability "up to" the value x



- $F_X(x) \ge 0$
- $F_X(x)$ is non-decreasing, i.e. if $a > b$ then $F_X(a) \ge F_X(b)$
- $\lim_{x \to +\infty} F_X(x) = 1$ and $\lim_{x \to -\infty} F_X(x) = 0$
- $X \sim F_X$ means X has CDF F_X

Discrete random variables

A *discrete random variable* is a real-valued function of experimental outcome that can take a finite or countably infinite number of values



- The *probability mass function* (PMF) of a discrete r.v. X is

$$p_X(x) \stackrel{\text{def}}{=} \mathbb{P}(X = x)$$

- $p_X(x) \geq 0$ and $\sum_x p_X(x) = 1$
- $X \sim p_X$ means X has PMF p_X

Common discrete random variables

- *Bernoulli*: $X \sim \text{Bern}(p)$ for $0 \leq p \leq 1$ has PMF

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

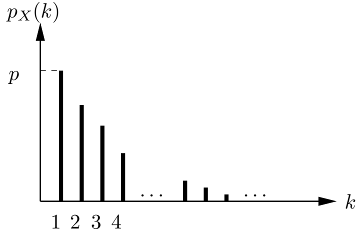
  - e.g. consider the toss of a coin, which comes up a head with probability p, and a tail otherwise

Common discrete random variables

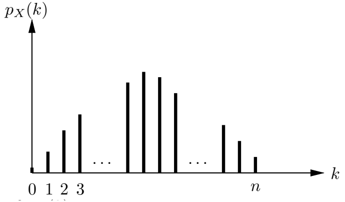
- *Geometric*: $X \sim \text{Geo}(p)$ for $0 \le p \le 1$ has PMF

$$p_X(k) = p(1-p)^{k-1}, \qquad k = 1, 2, \cdots$$

  - e.g., suppose we repeatedly and independently toss a coin with probability of a head equal to p, then X represents # tosses needed for a head to come up for the first time



Common discrete random variables

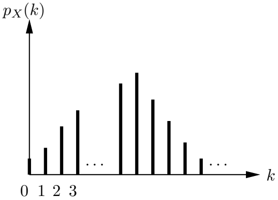


- *Binomial*: $X \sim \text{Bin}(n,p)$ for integer n and $0 \leq p \leq 1$ has PMF

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \cdots, n$$

  - e.g., a coin (with probability of a head equal to p) is tossed independently n times. Then X represents # heads in n tosses

Common discrete random variables



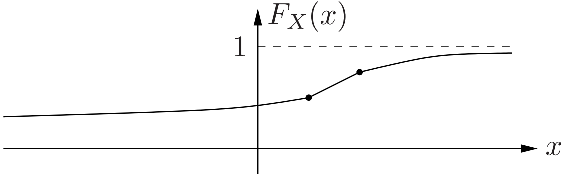
- *Poisson*: $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$ has PMF

$$p_X(k) = \frac{\lambda^k}{k!} \underbrace{e^{-\lambda}}_{\text{normalization}}, \qquad k = 0, 1, \cdots$$

  - Often represents # random events (e.g. arrivals of packets, photons) in some time interval
  - When n is very large and p is small, $\text{Poisson}(np)$ becomes a good approximation for $\text{Bin}(n, p)$ (called *Poisson approximation*)

Continuous random variables

A random variable is said to be *continuous* if its CDF is a continuous function

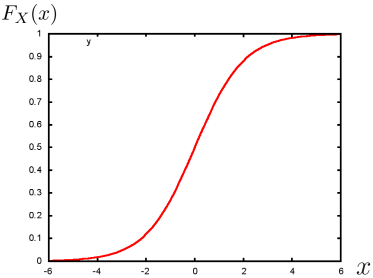


- If $F_X(\cdot)$ is continuous and differentiable, then the *probability density function* (PDF) of X is

$$f_X(x) \stackrel{\text{def}}{=} \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

- $f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$
- $X \sim f_X$ means X has PDF f_X
- $\mathbb{P}\{X \in A\} = \int_{x \in A} f_X(x)\mathrm{d}x$

Continuous random variables



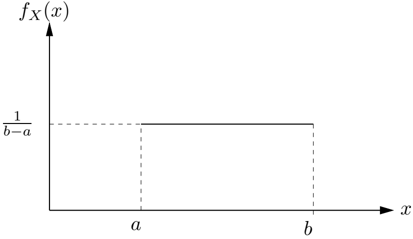
Example (logistic distribution): Suppose

$$F_X(x) = \frac{1}{1+e^{-x}} \qquad \underbrace{\text{(sigmoid function)}}_{\text{widely used in artificial neural networks}}$$

Then the PDF of logistic distribution is

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \frac{e^{-x}}{(1+e^{-x})^2}$$

Common continuous random variables

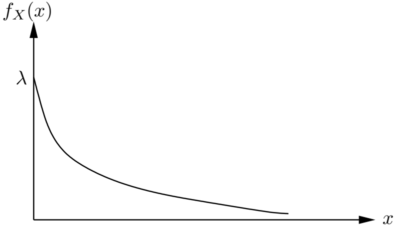


- *Uniform*: $X \sim \text{Unif}(a, b)$ for $a < b$ has PDF and CDF

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

$$F_X(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases}$$

Common continuous random variables



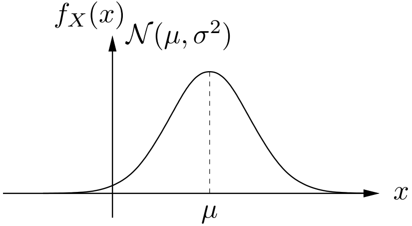
- *Exponential*: $X \sim \text{Exp}(\lambda)$ for $\lambda > 0$ has PDF and CDF

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

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{else} \end{cases}$$

commonly used to model inter-arrival time in a queue

Common continuous random variables



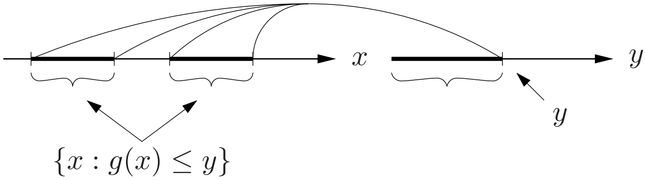
- *Gaussian*: $X \sim \mathcal{N}(\mu, \sigma^2)$ with parameter μ (mean) and σ^2 (variance) has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}$$

Functions of a random variable

Suppose $X \sim F_X$ is a r.v., and let $Y = g(X)$ be a function of X. Then the CDF of Y is

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y)$$



Application: probability integral transform

Given any continuous r.v. with known CDF, can we generate a Unif$(0,1)$ *r.v.?*

Theorem 1.1

Suppose $X \sim F_X$*, and define the r.v.* Y *as* $Y = F_X(X)$*. Then* Y *is uniformly distributed on* $(0,1)$*.*

- As we will discuss later, this fact is used in statistical analysis to test whether acquired data can reasonably be modeled as arising from a specified distribution

Application: probability integral transform

Proof of Theorem 1.1: Assume F_X has an inverse $F_X^{-1}(\cdot)$ such that

$$F_X^{-1}(F_X(y)) = y \quad \text{and} \quad F_X(F_X^{-1}(y)) = y \tag{1.1}$$

(e.g. if $F_X(x) = \frac{1}{1+e^{-x}}$, then $F_X^{-1}(x) = \log \frac{x}{1-x}$) For any $0 < y < 1$,

$$\begin{aligned}
\mathbb{P}\{Y \leq y\} &= \mathbb{P}\{F_X(X) \leq y\} && \text{(definition of } Y\text{)} \\
&= \mathbb{P}\left\{F_X^{-1}\left(F_X(X)\right) \leq F_X^{-1}(y)\right\} \\
&= \mathbb{P}\left\{X \leq F_X^{-1}(y)\right\} && \text{(see (1.1))} \\
&= F_X(F_X^{-1}(y)) && \text{(definition of } F_X\text{)} \\
&= y && \text{(see (1.1))}
\end{aligned}$$

This matches the CDF of Unif$(0, 1)$.

Application: probability integral transform

Example: suppose that $X \sim \mathsf{Exp}(1)$, so that

$$F_X(x) = 1 - \exp(-x), \qquad x \geq 0$$

Then Theorem 1.1 says that

$$Y = F_X(X) = 1 - \exp(-X) \;\sim\; \mathsf{Unif}(0, 1)$$

Application: generating random variables

Another useful scenario: *given a r.v.* $X \sim \mathsf{Unif}(0,1)$, *how to generate a r.v.* Y *with prescribed CDF* F_Y *(e.g. Gaussian)?*

- We learn from Theorem 1.1 that

$$\begin{array}{ccc} Y & \longrightarrow & \overbrace{F_Y(Y)}^{X} \\ P_Y & \longrightarrow & \mathsf{Unif}(0,1) \end{array}$$

- A natural candidate: an inverse transform

$$\begin{array}{rcl} F_Y^{-1}(X) & \longleftarrow & X \\ P_Y & \longleftarrow & \mathsf{Unif}(0,1) \end{array}$$

Application: generating random variables

Theorem 1.2

Suppose $X \sim \mathsf{Unif}(0,1)$, *and define the r.v.* Y *as* $Y = F_Y^{-1}(X)$. *Then* $Y \sim F_Y$.

Proof: For any $y \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\{Y \le y\} &= \mathbb{P}\left\{F_Y^{-1}(X) \le y\right\} && \text{(definition of } Y\text{)} \\
&= \mathbb{P}\left\{F_Y\left(F_Y^{-1}(X)\right) \le F_Y(y)\right\} && \\
&= \mathbb{P}\{X \le F_Y(y)\} && \text{(definition of } F_Y^{-1}\text{)} \\
&= F_X(F_Y(y)) && \text{(definition of } F_X\text{)} \\
&= F_Y(y) && (X \sim \mathsf{Unif}(0,1))
\end{aligned}$$

This implies $Y \sim F_Y$.

Application: generating random variables

Example: suppose that $X \sim$ Unif$(0,1)$, and the desired distribution is Exp(1), i.e.

$$F_Y(y) = 1 - \exp(-y), \qquad x \geq 0$$

which has an inverse function

$$F_Y^{-1}(y) = -\log(1-y)$$

Then Theorem 1.2 says that

$$Y = F_Y^{-1}(X) = -\log(1-X) \ \sim \ \text{Exp}(1)$$

Application: generating random variables

Matlab code

- Use a r.v. $Y \sim \mathsf{Exp}(1)$ to generate $X \sim \mathsf{Unif}(0, 1)$

```
Y = exprnd(1);
X = 1 - exp(-Y);
```

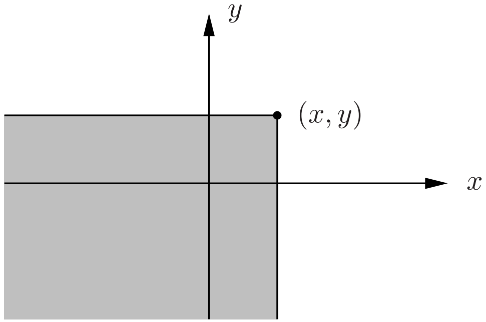
- Use a r.v. $X \sim \mathsf{Unif}(0, 1)$ to generate $Y \sim \mathsf{Exp}(1)$

```
X = rand();
Y = - log(1-X);
```

Joint distributions

Two random variables can be completely described with their *joint CDF*:

$$F_{X,Y}(x,y) = \mathbb{P}\{X \le x, Y \le y\}, \qquad x, y \in \mathbb{R}$$



$F_{X,Y}(x,y)$: probability of the shaded region

Joint distributions

Properties of joint CDF

- $F_{X,Y}(x, y) \geq 0$
- If $x_1 \leq x_2$ and $y_1 \leq y_2$, then $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$
- marginal CDFs:

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y) \quad \text{and} \quad F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x, y)$$

- $\lim_{x,y \to \infty} F_{X,Y}(x, y) = 1$, $\lim_{x,y \to -\infty} F_{X,Y}(x, y) = 0$
- X and Y are independent if for every x and y,

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

Joint, marginal, and conditional PMFs

- Two discrete r.v.s X and Y are specified by their *joint PMF*:

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) \qquad \text{for all } x \text{ and } y$$

- To find *marginal PMF* p_X, use

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

- The *conditional PMF* of X given $Y = y$ is

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \qquad p_Y(y) \neq 0$$

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

Joint, marginal, and conditional PDFs

- Two continuous r.v.s X and Y are specified by their *joint PDF*:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \qquad \text{for all } x, y \in \mathbb{R}$$

- To find *marginal PDF* f_X, use the law of total probability

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

- The *conditional PDF* of X given $Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \qquad f_Y(y) \neq 0$$

- The *conditional CDF* of X given $Y = y$ is

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)}\mathrm{d}u$$

Bayes' rule for random variables

- Bayes' rule for PMFs: given $p_X(x)$ and $p_{Y|X}(y \mid x)$, then

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

- Bayes' rule for PDFs: given $f_X(x)$ and $f_{Y|X}(y \mid x)$, then

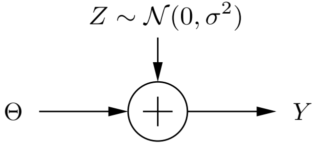
$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)\, f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y \mid u)\, f_X(u) \mathrm{d}u}$$

- Bayes' rule for 1 discrete r.v. Θ and 1 continuous r.v. Y:

$$p_{\Theta|Y}(\theta \mid y) = \frac{f_{Y|\Theta}(y \mid \theta)\, p_\Theta(\theta)}{\sum_{\theta'} f_{Y|\Theta}(y \mid \theta')\, p_\Theta(\theta')} \tag{1.2}$$

Example: additive Gaussian noise channel

Consider the following communication channel



where the signal Θ is generated such that

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

The received signal is $Y = \Theta + Z$, where Θ and Z are independent

Question: given $Y = y$ is observed, find posterior pmf $p_{\Theta|Y}(\theta \mid y)$

Example: additive Gaussian noise channel

Solution: use Bayes rule (1.2). We already know p_Θ. To compute $f_{Y|\Theta}(y \mid \theta)$, consider

$$\begin{aligned}
\mathbb{P}\{Y \le y \mid \Theta = 1\} &= \mathbb{P}\{\Theta + Z \le y \mid \Theta = 1\} \\
&= \mathbb{P}\{Z \le y - \Theta \mid \Theta = 1\} \\
&= \mathbb{P}\{Z \le y - 1 \mid \Theta = 1\} \\
&= \mathbb{P}\{Z \le y - 1\} \qquad \text{(independent of } \Theta \text{ and } Z\text{)}
\end{aligned}$$

giving that $Y \mid \{\Theta = 1\} \sim \mathcal{N}(1, \sigma^2)$. Similarly, $Y \mid \{\Theta = -1\} \sim \mathcal{N}(-1, \sigma^2)$. Thus,

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

$\square$

Example: additive Gaussian noise channel

If $\sigma = 1$ and $p = 1/2$, $p_{\Theta|Y}(1 \mid y)$ simplifies to

$$p_{\Theta|Y}(1 \mid y) = \frac{pe^y}{pe^y + (1-p)\, e^{-y}}$$

Suppose the receiver decides that the signal transmitted is 1 if $Y > 0$, and -1 otherwise. What is the probability of decision error?

Example: additive Gaussian noise channel

Solution: The decision is incorrect if

- $\Theta = 1$ but $Y \leq 0$ (or equivalently, $Z \leq -1$), or
- $\Theta = -1$ but $Y > 0$ (or equivalently, $Z > 1$).

Therefore, the probability of error is

$$
\begin{aligned}
\mathbb{P}_{\text{error}} &= \mathbb{P}\{\Theta = 1, Y \leq 0\} + \mathbb{P}\{\Theta = -1, Y > 0\} \\
&= \mathbb{P}\{\Theta = 1\}\,\mathbb{P}\{Y \leq 0 \mid \Theta = 1\} + \mathbb{P}\{\Theta = -1\}\,\mathbb{P}\{Y > 0 \mid \Theta = -1\} \\
&= \mathbb{P}\{\Theta = 1\}\,\mathbb{P}\{Z \leq -1\} + \mathbb{P}\{\Theta = -1\}\,\mathbb{P}\{Z > 1\} \\
&= \frac{1}{2}\mathbb{P}\{Z \leq -1\} + \frac{1}{2}\mathbb{P}\{Z > 1\} \\
&= \frac{1}{2}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-1} e^{-x^2/2}\mathrm{d}x + \frac{1}{\sqrt{2\pi}}\int_{1}^{\infty} e^{-x^2/2}\mathrm{d}x\right\} \\
&= \frac{1}{\sqrt{2\pi}}\int_{1}^{\infty} e^{-x^2/2}\mathrm{d}x
\end{aligned}
$$

Reference

[1] "*Introduction to probability (2nd Edition)*," D. Bertsekas, J. Tsitsiklis, *Athena Scientific*, 2008.

[2] "*Lecture notes: Statistical Signal Processing*," A. El Gamal.

[3] "*Probability and Computing (2nd Edition)*," M. Mitzenmacher, E. Upfal, *Cambridge University Press*, 2017.

[4] "*Statistical Inference (2nd Edition)*," G. Casella, R. Berger, *Cengage*, 2002.

[5] "*Stochastic Processes: Theory for Applications*," R. Gallager, *Cambridge University Press*, 2013.