Revision: Probability Theory

(CIV6540 - Probabilistic Machine Learning for Civil Engineers)

Professor: James-A. Goulet

Département des génies civil, géologique et des mines
Polytechnique Montréal

Chapter 3 – Goulet (2020)
Probabilistic Machine Learning for Civil Engineers
MIT Press

Probability theory, why?

"... *théorie des probabilités* n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct..."

— Pierre-Simon, marquis de Laplace [1749 – 1827]

[Wikipedia]

Probability theory, why? 🎥

In order to describe aleatory phenomenons? ⚠ no.

Few phenomenons are intrinsically aleatory

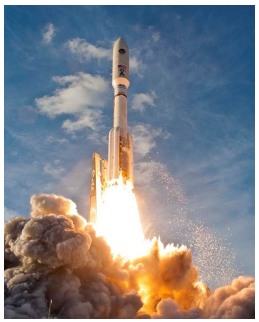
We employ probability to **describe our knowledge (epistemic uncertainties)**

- The Young's modulus E does not vary (locally)
- We do not know its value
- We describe our uncertainty for E using probabilities

The less we know, the more we should use probability theory

[Google images / Nitipak Samsen]

Who/what is using probability theory?




Google

If it works for them, it works for us.

Module #1b Outline

Set theory
Probabilities
X– **R.V.**
X– **M.R.V.**
$\mathbb{E}[X]$
$g(\mathbf{X})$
Linearization

Topics organization

Background
- 1 Revision probability & linear algebra
- 2 Probability distributions

Machine Learning Basics
- 0 Introduction
- 3 Bayesian Estimation $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
- 4 MCMC sampling & Newton

Supervised learning
- 5 Regression
- 6 Classification
- 7 LSTM networks for time series

Unsupervised learning
- 7 State-space model for time-series

Decision Making & RL
- 8 Decision Theory
- 9 AI & Sequential decision problems

Section Outline

Set theory

1.1 Definitions

1.2 Examples

1.3 Elementary rules

Definitions

Set: Ensemble of events or elements

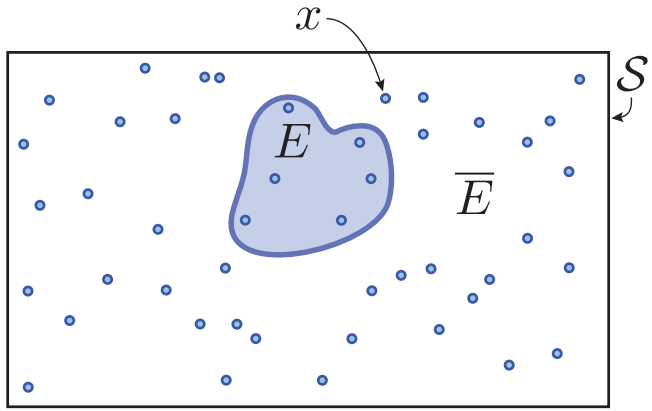
Universe/sampling space ($\mathcal{S}$): Ensemble of possible results (can be discrete or continuous & finite/infinite)

Elementary event (x): a single event, $x \in \mathcal{S}$

Event (E): set of elementary events

$E \subseteq \mathcal{S}$: subset of $\mathcal{S}$
$E = \mathcal{S}$: certain event
$E = \emptyset$: impossible event
$\overline{E}$: complement of E

Venn Diagram



Example #1 – State of a structure after an earthquake

Reality: continuum, continuous sampling space

$$\mathcal{S} = \{\text{no damage} \longleftrightarrow \text{total collapse}\}$$

Abstraction: discrete sampling space

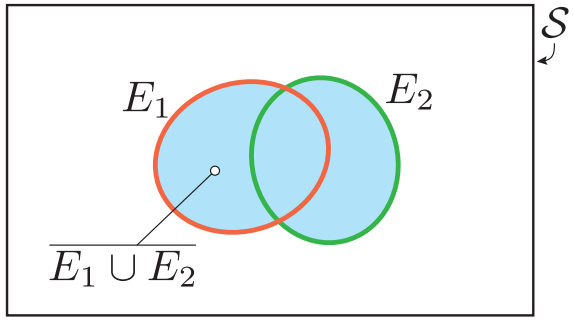
$$\mathcal{S} = \{\text{no damage}, \text{light damage}, \text{important damage}, \text{collapse}\}$$

$$\begin{aligned} \mathcal{S} &= \{ND, LD, ID, CO\} \\ E_1 &= \{ND, LD\} \\ E_2 &= \{CO\} \\ \overline{E_1} &= \{ID, CO\} \\ \overline{E_2} &= \{ND, LD, ID\} \end{aligned}$$

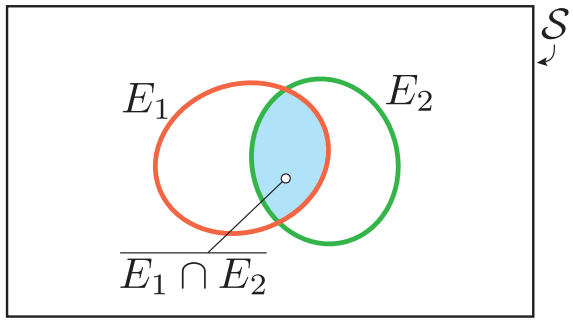
Operations on events

Given the events $\{E_1, E_2, \cdots, E_n\} \in \mathcal{S}$

Union (i.e. "or"): $E_1 \cup E_2$



Intersection (i.e. "and"): $E_1 \cap E_2 \equiv E_1 E_2$

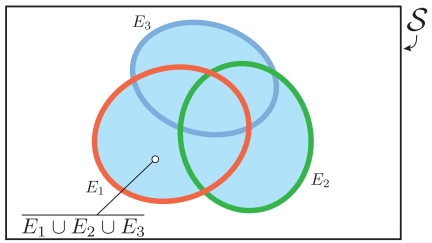


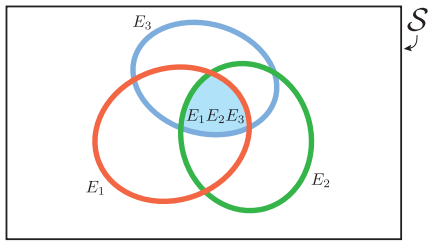
Rules

Commutativity: $E_1 \cup E_2 = E_2 \cup E_1, \quad E_1 E_2 = E_2 E_1$

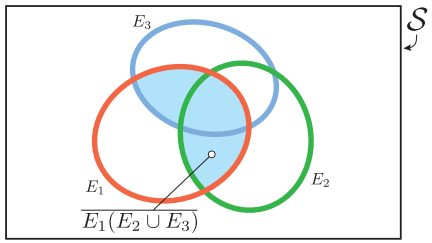
Associativity: $(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3) = E_1 \cup E_2 \cup E_3$

$\cup_{i=1}^{n} E_i = E_1 \cup E_2 \cup \cdots \cup E_n$ $\quad$ $\cap_{i=1}^{n} E_i = E_1 \cap E_2 \cap \cdots \cap E_n$





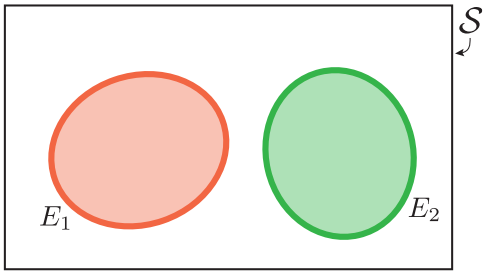
Distributivity: $E_1(E_2 \cup E_3) = (E_1 E_2 \cup E_1 E_3)$



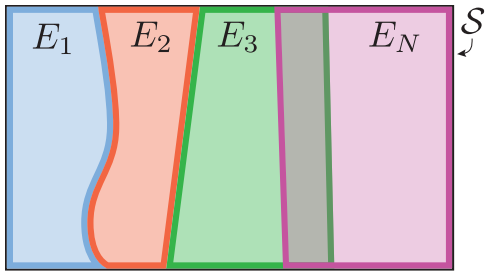
⚠ **Convention:** Intersection has priority over union...otherwise ( )

Types of events: $\{E_1, E_2, \cdots, E_n\} \in \mathcal{S}$

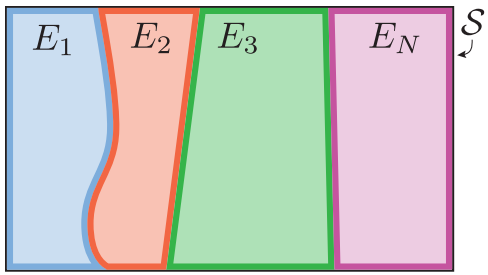
mutually exclusive (ME): $E_1, E_2, \cdots, E_n$ are ME if $E_i E_j = \emptyset, \ \forall i \neq j$

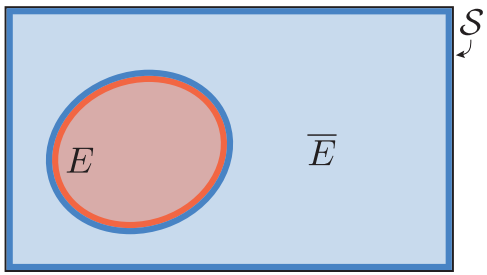


collectively exhaustives (CE):
$E_1, E_2, \cdots, E_n$ are CE if $\cup_{i=1}^{n} E_i = \mathcal{S}$



mutually exclusive & collectively exhaustives (ME & CE) :





Section Outline

Probabilities

2.1 Introduction

2.2 Axioms and elementary rules

2.3 Conditional probabilities & Bayes

Interpretation of probabilities

Probability of an event E_i: $\boxed{\Pr(E_i)}$

⚠ Bayesian interpretation:

$\Pr(E_i)$ quantifies the likelihood with respect to events in $\mathcal{S}$

- Depends on our prior knowledge
- Changes when we obtain new information

⚠ Frequentist interpretation:

$$\Pr(E_i) = \lim_{n\to\infty} \frac{\#\{E_i\}}{n}$$

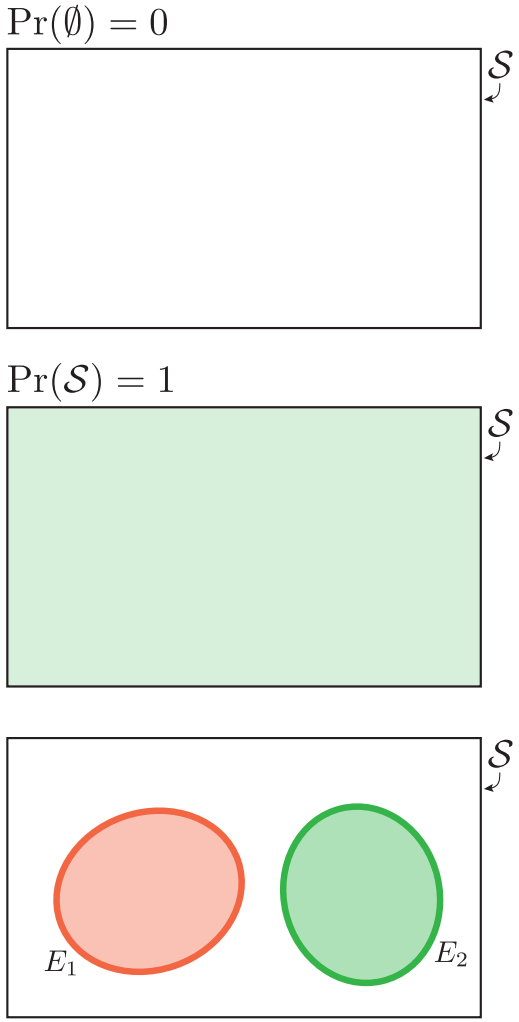
Fundamental Axioms

1. $0 \leq \Pr(E_i) \leq 1$

2. $\Pr(\mathcal{S}) = 1$
   (probability of a certain event $= 1$)

3. if E_1 and E_2 are mutually exclusive

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$$

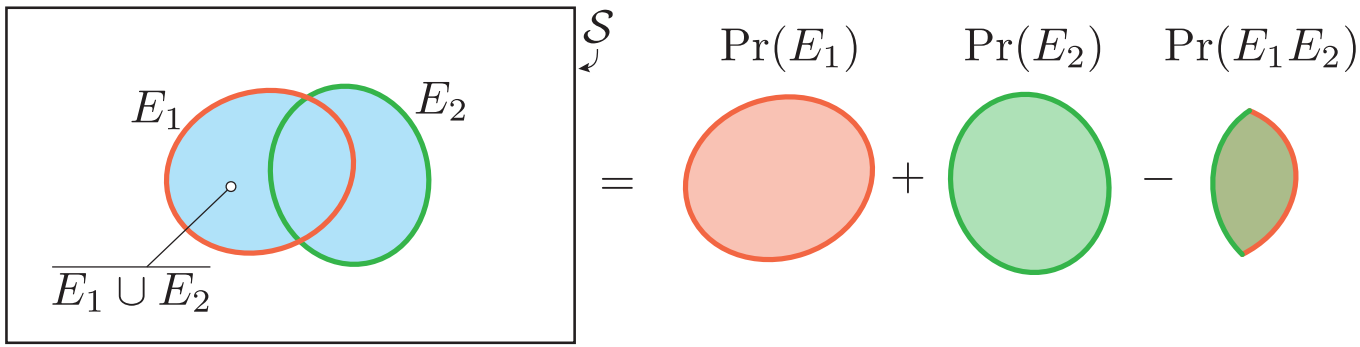


Rules derived from fundamental axioms

Probability of the complement: $\Pr(\overline{E}) = 1 - \Pr(E)$

Probability of an empty set: $\Pr(\emptyset) = 1 - 1 = 0$

Addition rule: $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 E_2)$



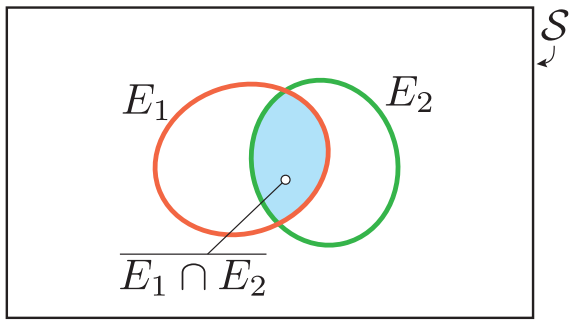
Conditional probabilities

$\Pr(E_1|E_2)$: Probability of the event E_1 conditional on the realization of the event E_2

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 E_2)}{\Pr(E_2)}, \quad (\Pr(E_2) \neq 0)$$

$$\overbrace{\Pr(E_1 E_2)}^{\text{Joint probability}} = \overbrace{\Pr(E_1|E_2)}^{\text{Conditional probability}} \cdot \overbrace{\Pr(E_2)}^{\text{Marginal probability}}$$

$$= \Pr(E_2|E_1) \cdot \Pr(E_1)$$



If E_1 and E_2 are statistically independent ($E_1 \perp\!\!\!\perp E_2$)

$$\Pr(E_1|E_2) = \Pr(E_1) \rightarrow \Pr(E_1 E_2) = \Pr(E_1) \cdot \Pr(E_2)$$

Chain rule

$$\begin{aligned}\Pr(E_1 E_2 \cdots E_n) &= \Pr(E_1|E_2 \cdots E_n)\Pr(E_2 \cdots E_n) \\ &= \Pr(E_1|E_2 \cdots E_n)\Pr(E_2|E_3 \cdots E_n)\Pr(E_3 \cdots E_n) \\ &= \Pr(E_1|E_2 \cdots E_n)\Pr(E_2|E_3 \cdots E_n)\cdots\Pr(E_{n-1}|E_n)\Pr(E_n)\end{aligned}$$

- n terms
- $n-1$ conditional probabilities
- 1 marginal probability

Example:

$$Pr(E_1 E_2 E_3) = \overbrace{\Pr(E_1|E_2 E_3)\underbrace{\Pr(E_2|E_3)\Pr(E_3)}_{=\Pr(E_2 E_3)}}^{=Pr(E_1 E_2 E_3)}$$

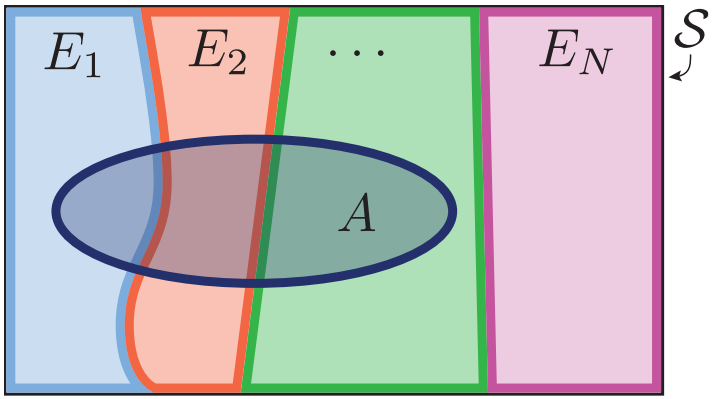
Law of total probability

Given an event A belonging to the sampling space $\mathcal{S}$ (i.e. $A \in \mathcal{S}$)

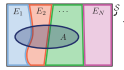
Given $\{E_1, E_2, E_3, \cdots, E_n\} \in \mathcal{S}$ a set of mutually exclusive and collectively exhaustive events (ME&CE)
(i.e. $E_i E_j = \emptyset,\ \forall i \neq j,\ \cup_{i=1}^{n} E_i = \mathcal{S}$)

$$\Pr(A) = \sum_{i=1}^{n} \underbrace{\Pr(A|E_i)\Pr(E_i)}_{=\Pr(AE_i)}$$

(Marginalisation)



Law of total probability (cont.)

REMINDER: (ME&CE, $\cup_{i=1}^{n} E_i = \mathcal{S}$, )

Union rule:

$$\Pr(E_1 \cup E_2|A) = \Pr(E_1|A) + \Pr(E_2|A) - \Pr(E_1 E_2|A)$$

Intersection Rule:

$$\Pr(E_1 E_2|A) = \Pr(E_1|E_2 A)\Pr(E_2|A)$$

$$\Pr(A|B) = \sum_{i=1}^{n} \underbrace{\Pr(A|E_i B)\Pr(E_i|B)}_{=\Pr(AE_i|B)}$$

Bayes's Theorem

Given $E_1, E_2, E_3, \cdots, E_n$ a set of mutually exclusive and collectively exhaustive events (ME&CE)

$$\left.\begin{aligned} \Pr(AE_i) &= \Pr(A|E_i)\Pr(E_i) \\ &= \Pr(E_i|A)\Pr(A) \end{aligned}\right\} \rightarrow \Pr(E_i|A)\Pr(A) = \Pr(A|E_i)\Pr(E_i)$$

Likelihhod function

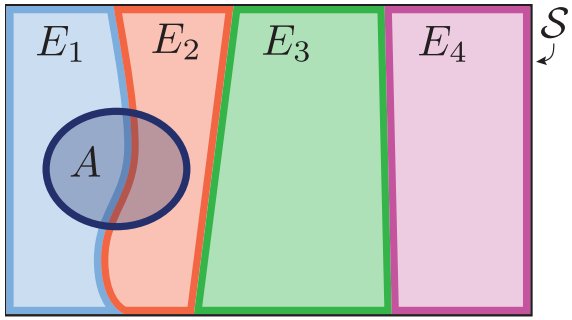
Prior probability

$$\underbrace{\Pr(E_i|A)}_{\text{Posterior probability}} = \frac{\Pr(A|E_i)\Pr(E_i)}{\Pr(A)}$$

Normalization constant

Bayes's Theorem:

$$\Pr(E_1|A) = \frac{\Pr(A|E_1)\Pr(E_1)}{\Pr(A)} = \frac{\Pr(\quad)}{\Pr(\quad)}$$



Normalization constant:

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A|E_i)\Pr(E_i)$$

⚠ typically, $\Pr(A)$ is difficult to estimate

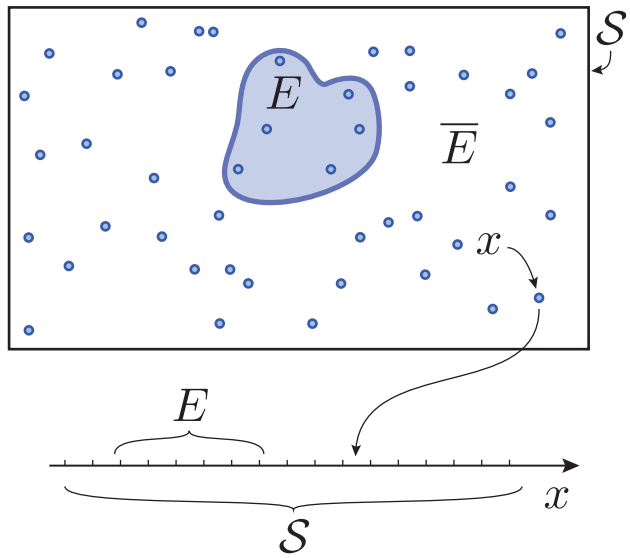
Section Outline

Random variables (R.V.)

Representation of the sampling space $\mathcal{S}$ on a line

Notation:

X : Random variable
x : Realization of a R.V.



The probability that a **random variable** X takes a **value** x is described by a **probability mass function (PMF)** or a **probability density function (PDF)**

Discrete random variables

Probability mass function (PMF)

$$\Pr(X = x) \equiv \Pr(x) \equiv p_X(x) \equiv p(x)$$

Properties:

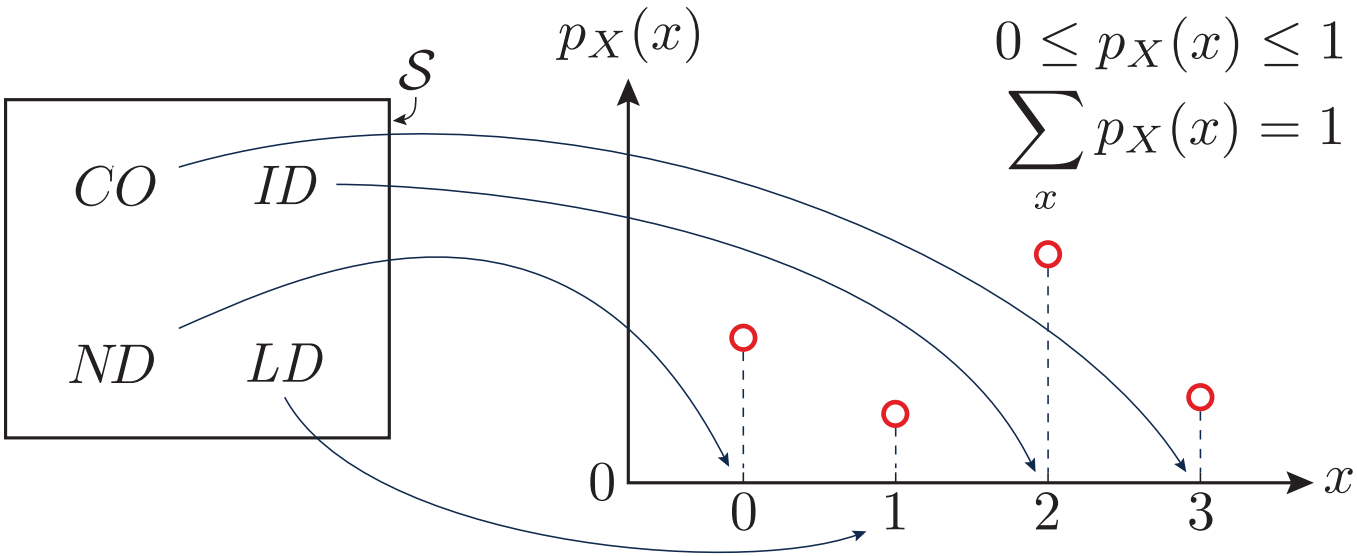
$$0 \leq p_X(x) \leq 1$$

$$\sum_x p_X(x) = 1$$

The set of possible values x is mutually exclusive and collectively exhaustive (ME&CE)

Example – damaged structure

$$\mathcal{S} = \left\{ \begin{array}{r} \text{no damage (ND)} \\ \text{light damage (LD)} \\ \text{important damage (ID)} \\ \text{collapse (CO)} \end{array} \right\}$$



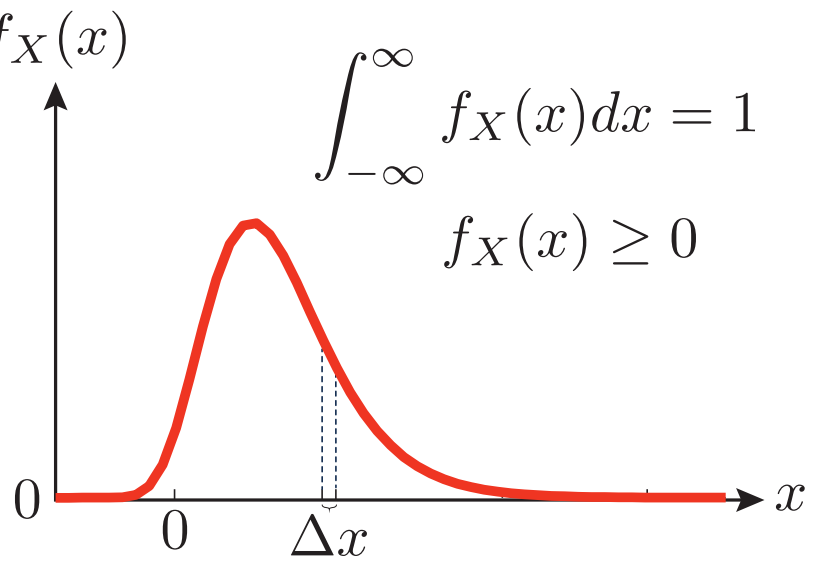
Light or important damage: $\{1 \leq x \leq 2\}$

Continuous random variables

Probability density function (PDF): $f_X(x) \equiv f(x)$

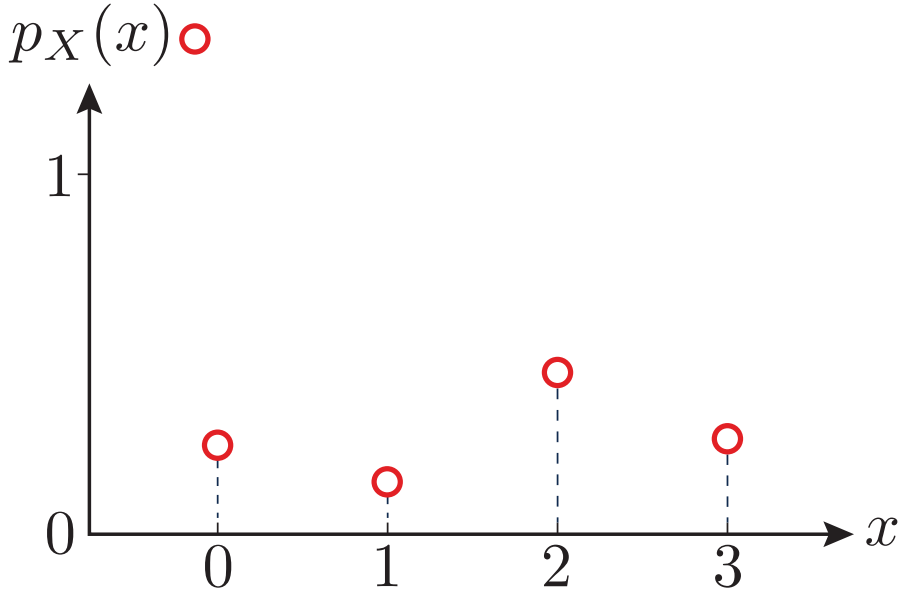
$$\Pr(x < X \leq x + \Delta x) = f_X(x)\Delta x$$

$$\Pr(X = x) = 0$$



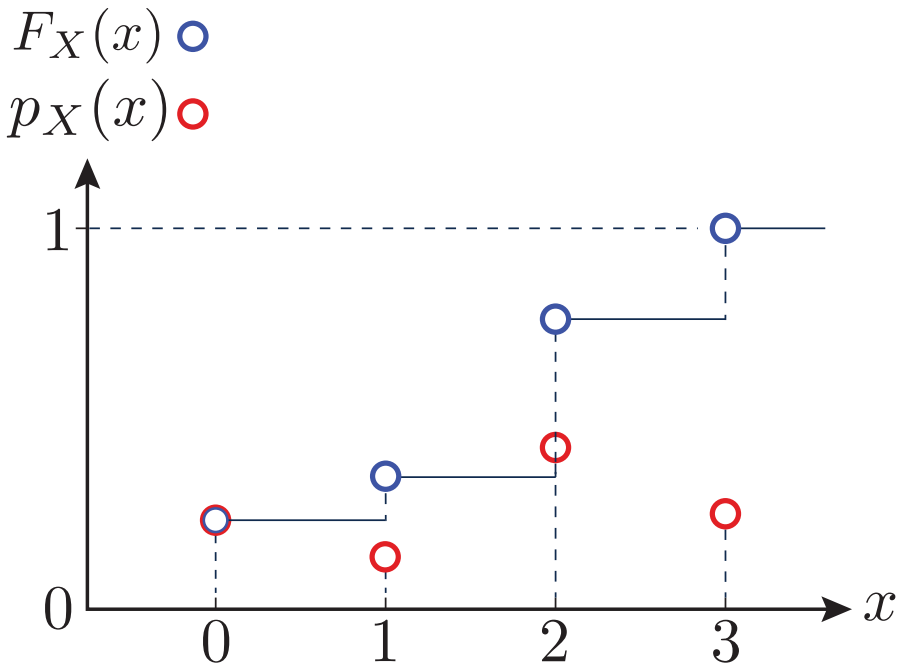
Cumulative distribution function (CDF)

$$F_X(x) = \Pr(X \leq x) = \sum_{x' \leq x} p_X(x'), \text{ (Discrete R.V.)}$$



Cumulative distribution function (CDF)

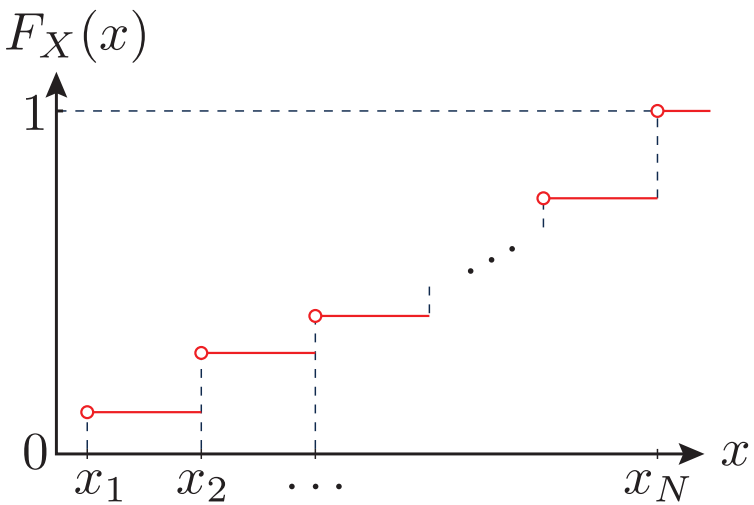
$$F_X(x) = \Pr(X \leq x) = \sum_{x' \leq x} p_X(x'), \text{ (Discrete R.V.)}$$

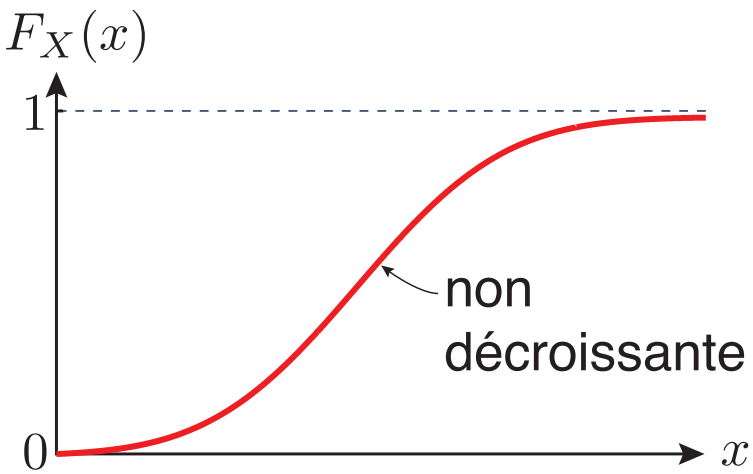


Cumulative distribution function (CDF)

$$F_X(x) = \Pr(X \leq x) = \sum_{x' \leq x} p_X(x'), \text{ (Discrete R.V.)}$$

$$= \int_{-\infty}^{x'} f_X(x')dx, \text{ (Continuous R.V.)}$$





$$F_X(-\infty) = 0, \quad F_X(+\infty) = 1$$

PDF $\leftrightarrow$ CDF

For continuous R.V.

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \leftrightarrow \boxed{f_X(x) = \frac{dF_X(x)}{dx}}$$

For discrete R.V.

$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1})$$

Common **probability density functions** will be presented in module #2

Univariate Normal random variable - $X \sim \mathcal{N}(x; \mu, \sigma^2)$ [.m]

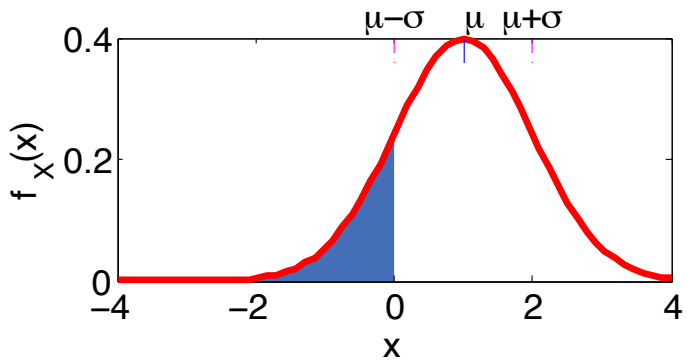
Probability density function (PDF) for a random variable $X : x \in \mathbb{R}$

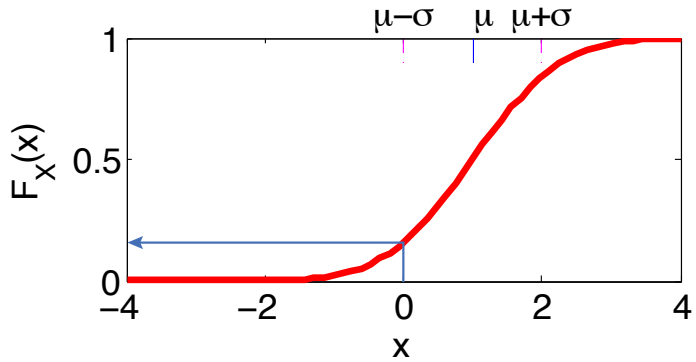
$$f_X(x) = \mathcal{N}(x; \mu, \sigma^2) \begin{cases} \mu & : \text{mean (i.e, C.G.)} \\ \sigma^2 & : \text{variance (i.e, Inertia)} \end{cases}$$

Cumulative density function (CDF)

$$F_X(0) = \int_{-\infty}^{0} f_X(0)dx = 0.16$$

$\mu = 1$
$\sigma = 1$





Conditional probabilities & R.V.

For an event E and a random variable X; the probability of E conditional on the realization of $\{X = x\}$ is:

$$\Pr(E|\{X = x\}) = \frac{\Pr(E \cap \{X = x\})}{\Pr(\{X = x\})}$$

Reminder:

$$\boxed{\Pr(E_1|E_2) = \frac{\Pr(E_1 E_2)}{\Pr(E_2)}}$$

$$\Pr(E|X = x) = \frac{\Pr(E \cap X = x)}{\Pr(X = x)}$$

$$= \frac{\Pr(X = x|E)\Pr(E)}{\Pr(X = x)}$$

$$\Pr(E|x) = \frac{\Pr(x|E)\Pr(E)}{\Pr(x)}$$

Conditional probabilities – events

If X is a **discrete** random variable

$$\boxed{E|x} \quad \rightarrow \quad \Pr(E|x) = \frac{p_{X|E}(x|E)\Pr(E)}{p_X(x)}$$

$$\boxed{X|E} \quad \rightarrow \quad \Pr(x|E) = p_{X|E}(x|E) = \frac{\Pr(E|x)p_X(x)}{\Pr(E)}$$

If X is a **continuous** random variable

$$\boxed{E|x} \quad \rightarrow \quad \Pr(E|x) = \frac{f_{X|E}(x|E)\Pr(E)}{f_X(x)}$$

$$\boxed{X|E} \quad \rightarrow \quad f_{X|E}(x|E) = \frac{\Pr(E|x)f_X(x)}{\Pr(E)}$$

Notation for random variables

Probability

$$\Pr(x) \equiv \Pr(X = x)$$

Random variable

$$\begin{aligned} X &\sim f(x) \equiv f_X(x) \equiv f_X(x;\boldsymbol{\theta}) \rightarrow X \sim \mathcal{N}(\mu, \sigma^2) \equiv \mathcal{N}(x;\mu,\sigma^2) \\ X &\sim p(x) \equiv p_X(x) \equiv p_X(x;\boldsymbol{\theta}) \end{aligned}$$

Conditional random variables

$$X|y \sim f(x|y) \equiv f_{X|Y}(x|y)$$

Section Outline

X– M.R.V.

4.1 Introduction

4.2 Probability density function

4.3 Conditional probabilities

Multivariate random variables

$$\text{Given } \mathbf{X} = [X_1, X_2, \cdots, X_n]^\intercal \begin{cases} \text{vector (column) of} \\ \text{random variables} \end{cases}$$

$$\text{Given } \mathbf{x} = [x_1, x_2, \cdots, x_n]^\intercal \begin{cases} \text{vector describing} \\ \text{the outcomes a} \\ \text{random variable } \mathbf{X} \end{cases}$$

X describes the simultaneous realization of several phenomena

Multivariate probability mass function (PMF)

Definition:

- $p_{\mathbf{X}}(\mathbf{x}) = \Pr(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n)$
- $0 \leq p_{\mathbf{X}}(\mathbf{x}) \leq 1$

Marginalization:

- $\sum_{x_n} p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = p_{X_1,\cdots,X_{n-1}}(x_1, \cdots, x_{n-1})$
- $\sum_{x_1} \cdots \sum_{x_n} p_{\mathbf{X}}(\mathbf{x}) = 1$

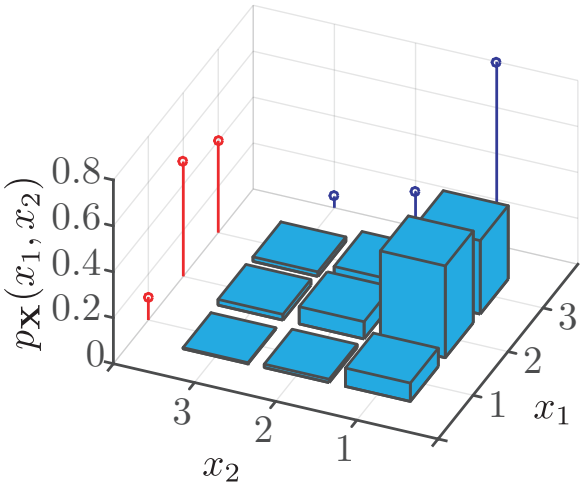
$p_{X_i}(x_i)$: **Marginal PMF**

Example – marginalization [.m]

Given two discrete R.V. $X_1 \perp\!\!\!\perp X_2$

$$p_{X_1}(x_1) \begin{cases} p_{X_1}(1) = 0.1 \\ p_{X_1}(2) = 0.5 \\ p_{X_1}(3) = 0.4 \end{cases}$$

$$p_{X_2}(x_2) \begin{cases} p_{X_2}(1) = 0.8 \\ p_{X_2}(2) = 0.15 \\ p_{X_2}(3) = 0.05 \end{cases}$$



$$p_{X_1 X_2}(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2)$$

$$p_{\mathbf{X}}(x_1, x_2) = \left\{ \begin{array}{c|ccc|c} & x_2 = 1 & x_2 = 2 & x_2 = 3 & \sum_{i=1}^{n=3} p_{\mathbf{X}}(x_1, i) \\ \hline x_1 = 1 & 0.08 & 0.015 & 0.005 & 0.1 \\ x_1 = 2 & 0.4 & 0.075 & 0.025 & 0.5 \\ x_1 = 3 & 0.32 & 0.06 & 0.02 & 0.4 \end{array} \right.$$

[CIVML/marginal.m]

Multivariate probability density function (PDF)

Definition:

- $f_{\mathbf{X}}(\mathbf{x})\Delta\mathbf{x} = \Pr(x_1 < X_1 \leq x_1 + \Delta x_1 \cap \cdots \cap x_n < X_n \leq x_n + \Delta x_n)$
- $0 \leq f_{\mathbf{X}}(\mathbf{x})$ ⚠ can be > 1

Marginalization:

- $\int_{-\infty}^{\infty} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)dx_n = f_{X_1,\cdots,X_{n-1}}(x_1,\cdots,x_{n-1})$
- $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = 1$

$f_{X_i}(x_i)$: **Marginal PDF**

Bivariate Normal PDF [.m]

Probability density function (PDF) for 2 random variables X_1, X_2

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2 + \left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2 - 2\rho\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)\right]\right\}$$

5 parameters: $\mu_{X_1}, \sigma_{X_1}, \mu_{X_2}, \sigma_{X_2}, \rho$

[CIVML/Multivariate_normal_example.m]

Bivariate Normal PDF [.m]

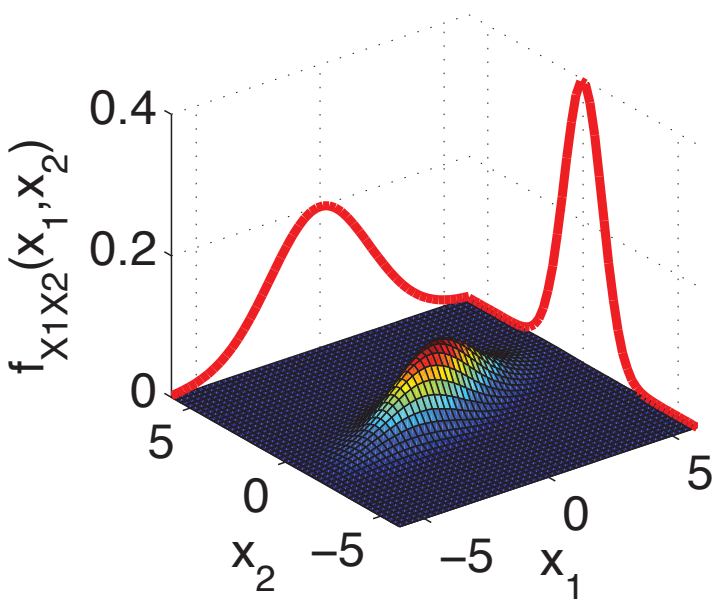
$$\mathbf{X} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

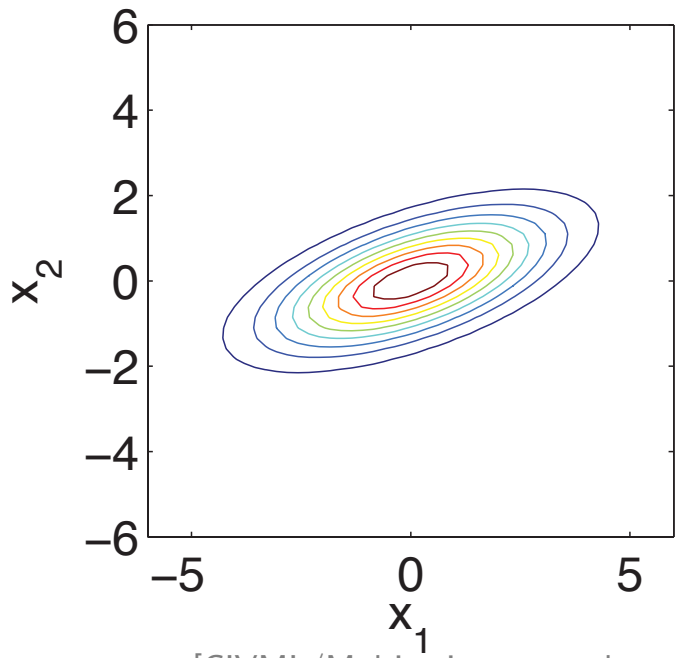
5 parameters:

$\mu_{X_1}, \sigma_{X_1}, \mu_{X_2}, \sigma_{X_2}, \rho$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix}$$

$\mu_{X_1} = 0$
$\sigma_{X_1} = 2$
$\mu_{X_2} = 0$
$\sigma_{X_2} = 1$
$\rho = 0.6$





[CIVML/Multivariate_normal_example.m]

Continuous multivariate CDF

Definition:

- $F_{\mathbf{X}}(\mathbf{x}) = \Pr(X_1 \leq x_1 \cap \cdots \cap X_n \leq x_n)$
- $0 \leq F_{\mathbf{X}}(\mathbf{x}) \leq 1$
- $F_{X_1,\cdots,X_{n-1},X_n}(x_1, \cdots, x_{n-1}, -\infty) = 0$
- $F_{X_1,\cdots,X_{n-1},X_n}(+\infty, \cdots, +\infty, +\infty) = 1$

Marginalization: $(F_{X_1X_2}(x_1, +\infty) = F_{X_1}(x_1))$

- $F_{X_1,\cdots,X_{n-1},X_n}(x_1, \cdots, x_{n-1}, +\infty) = F_{X_1,\cdots,X_{n-1}}(x_1, \cdots, x_{n-1})$

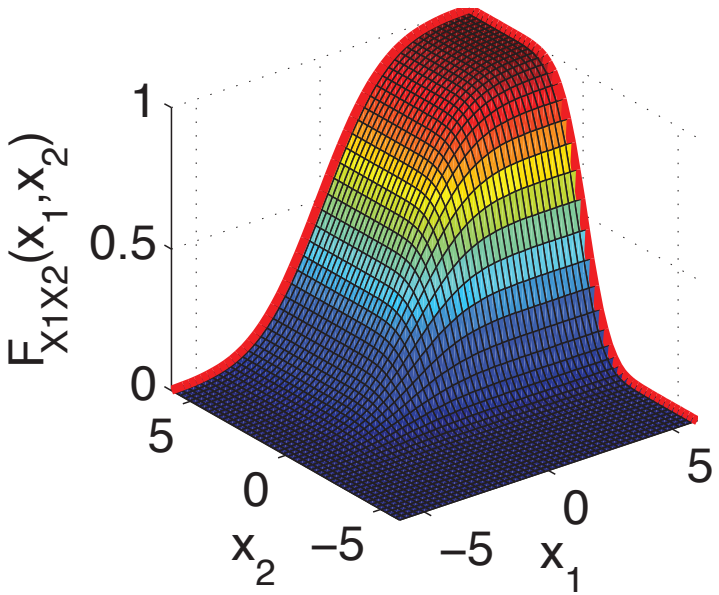
$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \qquad f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \cdots \partial x_n}$$

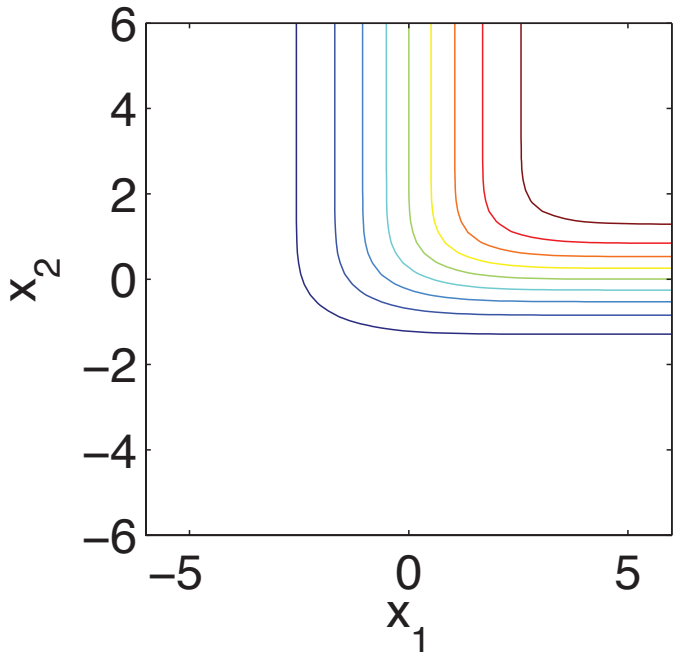
Bivariate Normal CDF [.m]

Cumulative distribution function (CDF) for 2 random variables X_1, X_2

$$F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(x_1, x_2) \partial x \partial y$$

$$\mu_{X_1} = 0$$
$$\sigma_{X_1} = 2$$
$$\mu_{X_2} = 0$$
$$\sigma_{X_2} = 1$$
$$\rho = 0.6$$





[CIVML/Multivariate_normal_example_CDF.m]

Conditional probabilities – Discrete

If X_i is a **discrete** random variable

$$p_{X_1 X_2}(x_1, x_2) = p_{X_1|X_2}(x_1|x_2)p_{X_2}(x_2) = p_{X_2|X_1}(x_2|x_1)p_{X_1}(x_1)$$

$$\boxed{X_1|x_2} \rightarrow p_{X_1|X_2}(x_1|x_2) = \frac{p_{X_2|X_1}(x_2|x_1)p_{X_1}(x_1)}{p_{X_2}(x_2)}$$

$$\boxed{X_2|x_1} \rightarrow p_{X_2|X_1}(x_2|x_1) = \frac{p_{X_1|X_2}(x_1|x_2)p_{X_2}(x_2)}{p_{X_1}(x_1)}$$

Conditional probabilities – Discrete

X_1 and X_2 are statistically independent ($\perp\!\!\!\perp$) if

$$p_{X_1|X_2}(x_1|x_2) = p_{X_1}(x_1)$$

si $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_n$

$$p_{X_1 \cdots X_n}(x_1, \cdots, x_n) = p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n)$$

General case:
X_1 and X_2 **not** statistically independent $\rightarrow$ **chain rule**

$$p_{X_1 \cdots X_n}(x_1, \cdots, x_n) = p_{X_1|X_2 \cdots X_n}(x_1|x_2 \cdots, x_n) \cdots p_{X_{n-1}|X_n}(x_{n-1}|x_n) p_{X_n}(x_n)$$

e.g. $p_{X_1 X_2 X_3}(x_1, x_2, x_3) = p_{X_1|X_2 X_3}(x_1|x_2, x_3) \cdot p_{X_2|X_3}(x_2|x_3) \cdot p_{X_3}(x_3)$

Conditional probabilities – Continuous

Exactly the same thing than for the discrete random variables

$$p_{\mathbf{X}}(\mathbf{x}) \rightarrow f_{\mathbf{X}}(\mathbf{x})$$

Section Outline

$\mathbb{E}[X]$

5.1 Definition

5.2 Moment of order m

5.3 Variance

5.4 Covariance

5.5 Matrix notation

Expected values

Expected value: sum of all possible values weighted by their probability of occurence

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx \quad \text{(Continuous R.V.)}$$

$$\mathbb{E}[X] = \sum x \cdot p_X(x) \quad \text{(Discrete R.V.)}$$

The expectation is a **linear operation**

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Moments

Moment of order m: $\mathbb{E}[X^m]$

$$\mathbb{E}[X^m] = \int_{-\infty}^{\infty} x^m f_X(x) dx$$

For $m = 1$: $\mathbb{E}[X] = \mu_X$ (**Expectation / center of gravity**)

For $m = 2$: $\mathbb{E}[X^2]$ (expectation of the squares)

Centered moments of order m: $\mathbb{E}[(X - \mu_X)^m]$

$$\mathbb{E}[(X - \mu_X)^m] = \int_{-\infty}^{\infty} (X - \mu_X)^m f_X(x) dx$$

For $m = 1$: $\mathbb{E}[(X - \mu_X)^1] = 0$

for $m = 2$: $\mathbb{E}[(X - \mu_X)^2] = \sigma_X^2 = \text{var}[X]$ (**variance / inertia**)

Variance – $\mathbb{E}[(X - \mu_X)^2]$

$$\mathbb{E}[(X - \mu_X)^2] = \sigma_X^2 = \text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

var$[X]$: The variance mesures the dispersion of the probability density function.
This concept is analogue to the inertia of a cross-section.

σ_X: Standard deviation

$\delta_X = \frac{\sigma_X}{\mu_X}$: coefficient of variation (C.O.V.).
adimensional dispersion metric (⚠ $\mu_X \neq 0$)

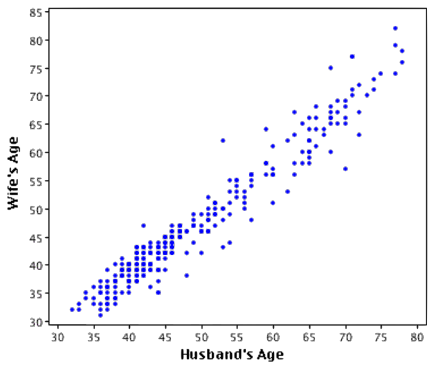
Covariance – $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$

Given two random variables X, Y

$$\text{cov}[XY] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

ρ_{XY}: correlation coefficient.
(Quantifies the **linear dependence** between X and Y)

$$\rho_{ij} = \frac{\text{cov}[XY]}{\sigma_i \sigma_j}, \quad -1 \leq \rho_{XY} \leq +1$$



- $X \perp\!\!\!\perp Y \implies \rho_{XY} = 0$
- $\rho_{XY} = 0 \not\implies X \perp\!\!\!\perp Y$ ⚠
- correlation $\not\iff$ causality ⚠

[onlinestatbook.com, Wikipedia]

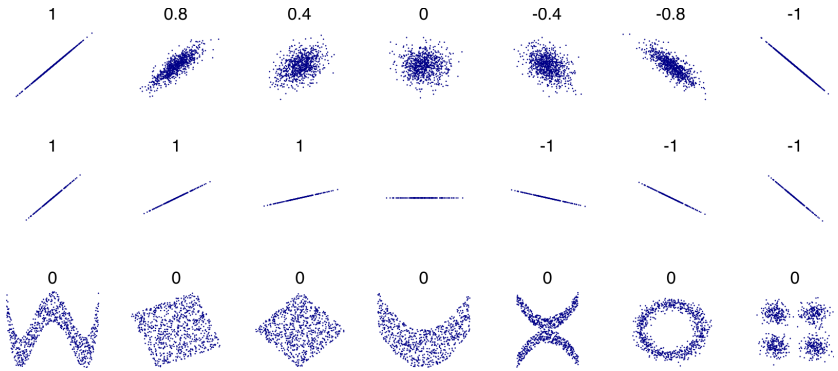
Covariance – $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$

Given two random variables X, Y

$$\text{cov}[XY] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

ρ_{XY}: correlation coefficient.
(Quantifies the **linear dependence** between X and Y)

$$\rho_{ij} = \frac{\text{cov}[XY]}{\sigma_i \sigma_j}, \quad -1 \leq \rho_{XY} \leq +1$$



- $X \perp\!\!\!\perp Y \implies \rho_{XY} = 0$
- $\rho_{XY} = 0 \not\implies X \perp\!\!\!\perp Y$ ⚠
- correlation $\not\Longleftrightarrow$ causality ⚠

[onlinestatbook.com, Wikipedia]

Matrix notation

given n random variables $\mathbf{X} = [X_1, X_2, \cdots, X_n]^\intercal$ where,
$\mathbf{M_X} = [\mu_1, \mu_2, \cdots, \mu_n]^\intercal$ is a vector containing expected values,
$\mathbf{D_X}$ is the matrix containing standard deviations,
$\mathbf{R_X}$ is the correlation matrix, and $\mathbf{\Sigma_X}$ is the covariance matrix.

$$\mathbf{D_X} = \begin{bmatrix} \sigma_{X_1} & 0 & 0 & 0 \\ & \sigma_{X_2} & 0 & 0 \\ & & \ddots & 0 \\ \text{sym.} & & & \sigma_{X_n} \end{bmatrix}, \quad \mathbf{R_X} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ & 1 & \cdots & \rho_{2n} \\ & & \vdots\!\vdots & \rho_{n-1n} \\ \text{sym.} & & & 1 \end{bmatrix}$$

$$\mathbf{\Sigma_X} = \mathbf{D_X R_X D_X} = \begin{bmatrix} \sigma^2_{X_1} & \rho_{12}\sigma_{X_1}\sigma_{X_2} & \cdots & \rho_{1n}\sigma_{X_1}\sigma_{X_n} \\ & \sigma^2_{X_2} & \cdots & \rho_{2n}\sigma_{X_2}\sigma_{X_n} \\ & & \vdots\!\vdots & \vdots \\ \text{sym.} & & & \sigma^2_{X_n} \end{bmatrix}$$

Section Outline

$g(\mathbf{X})$

6.1 Definition

6.2 Transformation rule

Function of random variables – $g(\mathbf{X})$

Given $\mathbf{X} = [X_1, X_2, \cdots, X_n]^\intercal$ defined from their joint PDF $f_{\mathbf{X}}(\mathbf{x})$, and given $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_m]^\intercal$ obtained from a function

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \begin{bmatrix} g_1(\mathbf{X}) \\ g_2(\mathbf{X}) \\ \vdots \\ g_m(\mathbf{X}) \end{bmatrix}$$

Question: what is the PDF $f_{\mathbf{Y}}(\mathbf{y})$,

3 cases:

1. $m = n = 1$
2. $m = n > 1$
3. $m = 1, n > 1$

We are interested in
linear functions
i.e. $Y_k = g_k(X) = aX_k + b$

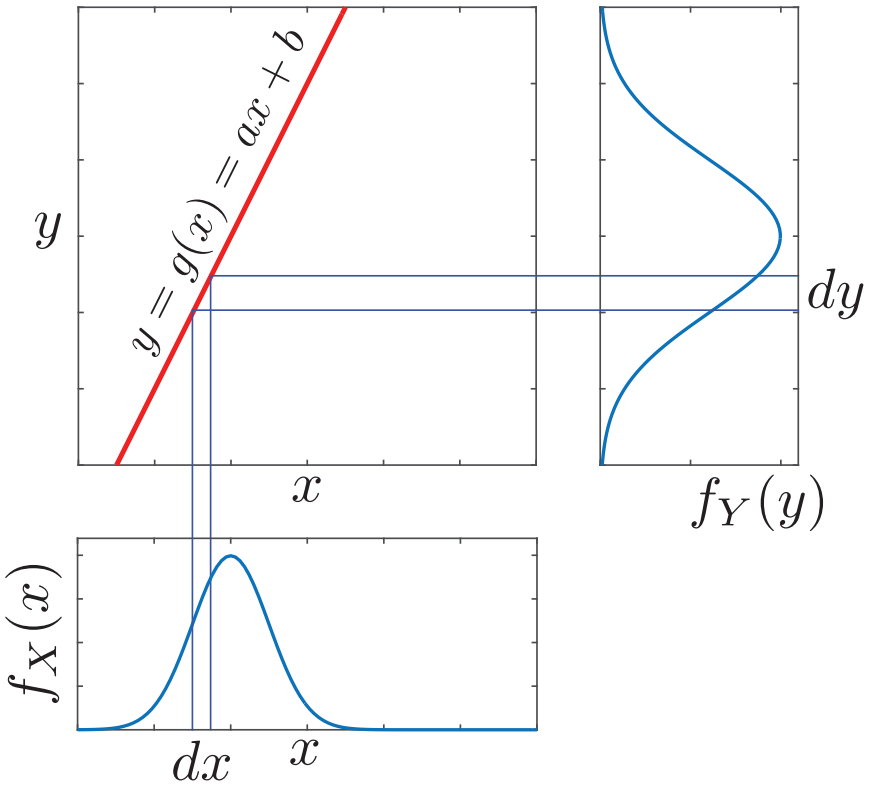
$Y = g(X)$: case 1, $m = n = 1$

$$f_Y(y)dy = f_X(x)dx$$

$$f_Y(y) = \underbrace{f_X(x)\left|\frac{dx}{dy}\right|}_{f_X(x)\geq 0}$$

For $Y = aX + b$

$$\left|\frac{dy}{dx}\right| = |a|$$



$$\mu_Y = g(\mu_X) = a\mu_X + b, \quad \sigma_Y = |a|\sigma_X$$

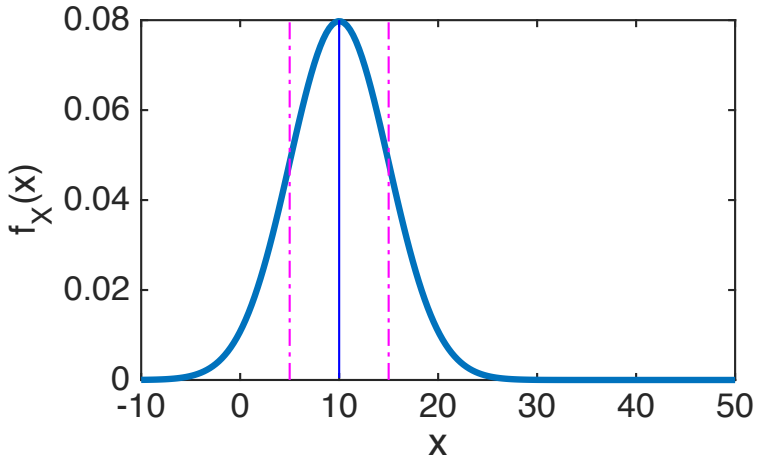
Example: Case 1, $Y = 2X$

Random variable:

$$X \sim \mathcal{N}(x; \underbrace{10}_{\mu_X}, \underbrace{5^2}_{\sigma_X^2})$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$$

$$\frac{dy}{dx} = 2$$



Transformation $X \to Y$:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x)\left|\frac{dy}{dx}\right|^{-1} = f_X(\underbrace{\frac{y}{2}}_{x=y/2}) \cdot \frac{1}{2}$$

$$f_Y(y) = \frac{1}{2\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{\frac{y}{2}-\mu_X}{\sigma_X}\right)^2\right], \quad Y \sim \mathcal{N}(y; \underbrace{2\mu_X}_{\mu_Y}, (\underbrace{2\sigma_X}_{\sigma_Y})^2)$$

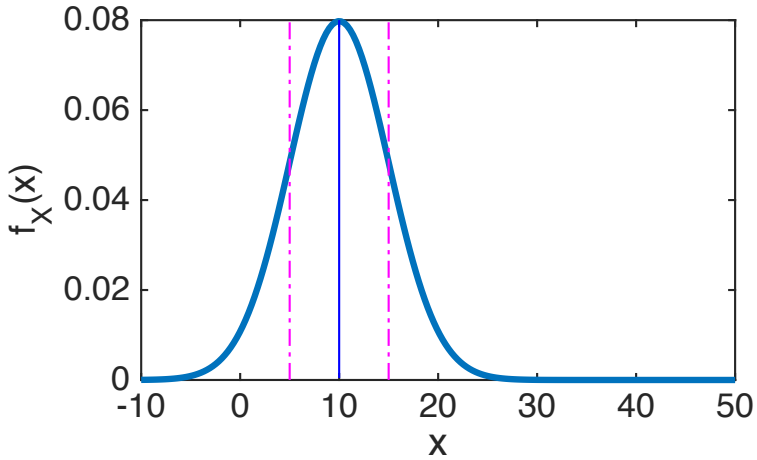
Example: Case 1, $Y = 2X$

Random variable:

$$X \sim \mathcal{N}(x; \underbrace{10}_{\mu_X}, \underbrace{5^2}_{\sigma_X^2})$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$$

$$\frac{dy}{dx} = 2$$



Transformation $X \to Y$:

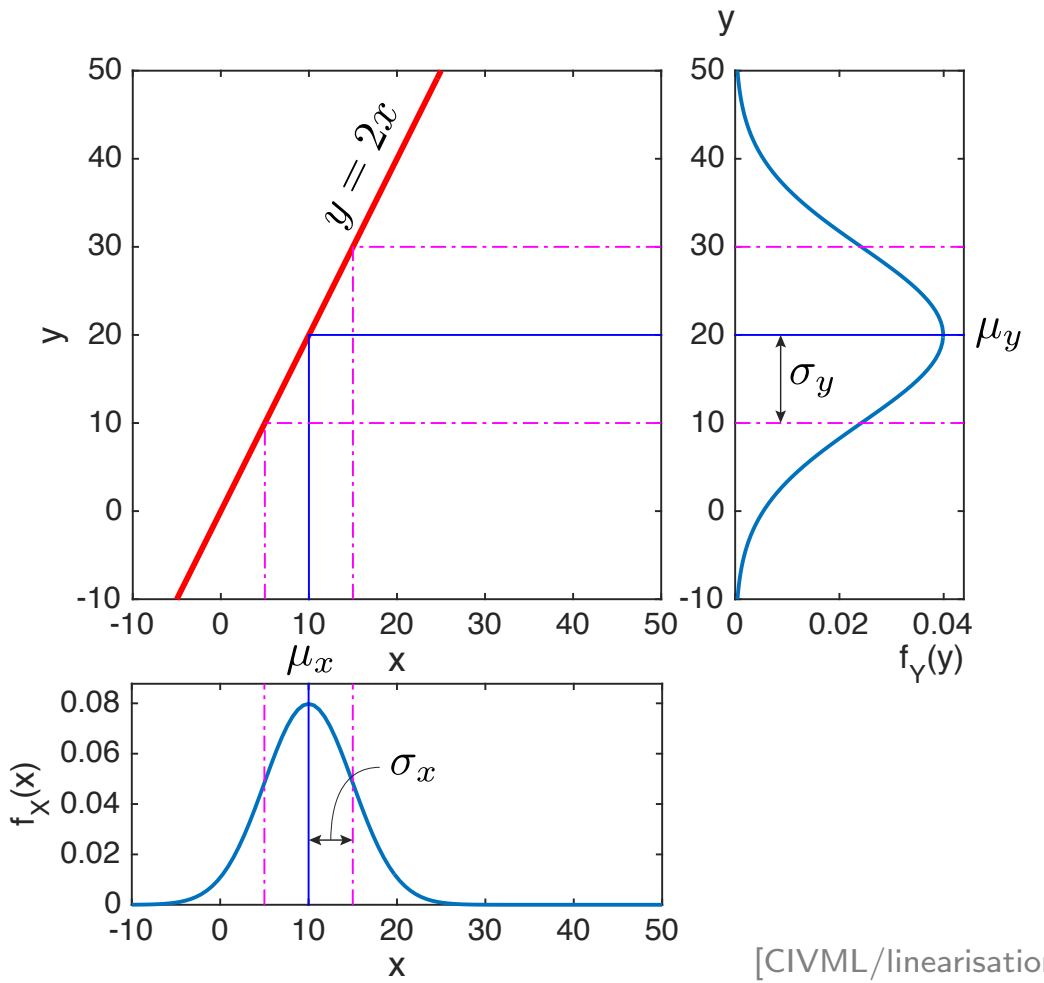
$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x)\left|\frac{dy}{dx}\right|^{-1} = f_X(\underbrace{\frac{y}{2}}_{x=y/2}) \cdot \frac{1}{2}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}2\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{y-2\mu_X}{2\sigma_X}\right)^2\right], \quad Y \sim \mathcal{N}(y; \underbrace{2\mu_X}_{\mu_Y}, (\underbrace{2\sigma_X}_{\sigma_Y})^2)$$

Example: $Y = 2X$ (cont.) [.m]

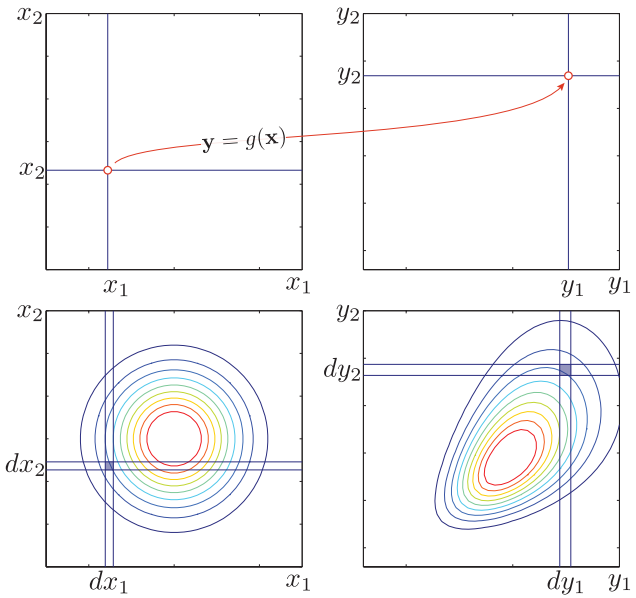
$$X \sim \mathcal{N}(x; \underbrace{10}_{\mu_X}, \underbrace{5^2}_{\sigma_X^2})$$

$$Y \sim \mathcal{N}(y; \underbrace{2\mu_X}_{\mu_Y}, \underbrace{(2\sigma_X)^2}_{\sigma_Y})$$



[CIVML/linearisation.m]

$\mathbf{Y} = \mathbf{g}(\mathbf{X})$: Case 2, $m = n > 1$



Jacobian ($\mathbf{J}_{\mathbf{y},\mathbf{x}}$): size of the neighborhood of $d\mathbf{y}$ with respect to $d\mathbf{x}$

$$\mathbf{J}_{\mathbf{y},\mathbf{x}} = \begin{bmatrix} J_{y_1,x_1} & \cdots & J_{y_1,x_n} \\ & \ddots & \vdots \\ \text{sym.} & & J_{y_m,x_n} \end{bmatrix} = \begin{bmatrix} \nabla g_1(\mathbf{x}) \\ \vdots \\ \nabla g_n(\mathbf{x}) \end{bmatrix}$$

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

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

$$\frac{d\mathbf{x}}{d\mathbf{y}} = |\det \mathbf{J}_{\mathbf{y},\mathbf{x}}|^{-1}$$

$$\frac{d\mathbf{y}}{d\mathbf{x}} = |\det \mathbf{J}_{\mathbf{y},\mathbf{x}}|$$

$$J_{k,i} = J_{y_k,x_i} = \frac{\partial y_k}{\partial x_i}$$

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})\,|\det \mathbf{J}_{\mathbf{y},\mathbf{x}}|^{-1}$$

$\mathbf{Y} = \mathbf{g}(\mathbf{X})$: Case 2, $m = n > 1$ (cont.)

Given a linear function $\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \mathbf{AX} + \mathbf{b}$ where

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \quad \mathbf{M_X} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_n} \end{bmatrix}, \quad \mathbf{\Sigma_X} = \begin{bmatrix} \sigma^2_{X_1} & \cdots & \rho_{1n}\sigma_{X_1}\sigma_{X_n} \\ & \vdots & \vdots \\ \text{sym.} & & \sigma^2_{X_n} \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \\ \\ \end{bmatrix}_{n\times 1}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \\ \\ \end{bmatrix}_{n\times n}}_{\mathbf{A}=\mathbf{J_{y,x}}} \times \underbrace{\begin{bmatrix} \\ \\ \end{bmatrix}_{n\times 1}}_{\mathbf{X}} + \underbrace{\begin{bmatrix} \\ \\ \end{bmatrix}_{n\times 1}}_{\mathbf{b}}$$

$$\begin{aligned} \mathbf{M_Y} &= \mathbf{g}(\mathbf{M_X}) = \mathbf{AM_X} + \mathbf{b} \\ \mathbf{\Sigma_Y} &= \mathbf{A\Sigma_X A}^\intercal = \mathbf{J_{y,x}\Sigma_X J_{y,x}}^\intercal \end{aligned}$$

$Y = g(\mathbf{X})$: Cas 3, $m = 1$, $n > 1$

Given a linear function $Y = g(\mathbf{X}) = \mathbf{a}^\intercal \mathbf{X} + b$ where

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \quad \mathbf{M_X} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_n} \end{bmatrix}, \quad \mathbf{\Sigma_{XX}} = \begin{bmatrix} \sigma^2_{X_1} & \cdots & \rho_{1n}\sigma_{X_1}\sigma_{X_n} \\ & \vdots\vdots\vdots & \vdots \\ \text{sym.} & & \sigma^2_{X_n} \end{bmatrix}$$

$$\underbrace{[\quad]_{1\times 1}}_{Y} = \underbrace{[\qquad]_{1\times n}}_{\mathbf{a}^\intercal = \nabla g(\mathbf{x})} \times \underbrace{\begin{bmatrix} \\ \\ \end{bmatrix}_{n\times 1}}_{\mathbf{X}} + \underbrace{[\quad]_{1\times 1}}_{b}$$

$$\begin{aligned} \mu_Y &= g(\mathbf{M_X}) = \mathbf{a}^\intercal \mathbf{M}_X + b \\ \sigma^2_Y &= \nabla g(\mathbf{X}) \mathbf{\Sigma_X} \nabla g(\mathbf{X})^\intercal \\ &= \mathbf{a}^\intercal \mathbf{\Sigma_X} \mathbf{a} \end{aligned}$$

$$\underbrace{\nabla g(\mathbf{x}) = \left[\frac{\partial g(\mathbf{x})}{\partial x_1}, \cdots, \frac{\partial g(\mathbf{x})}{\partial x_n} \right]_{\mathbf{x}}}_{\text{Gradient}}$$

Non-linear functions: $\mathbf{Y} = \mathbf{g}(\mathbf{X})$

What should we do if $\mathbf{g}(\mathbf{X})$ **is non-linear? Linearize.**

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) \approx \mathbf{A}\mathbf{X} + \mathbf{b}$$

Section Outline

Linearization

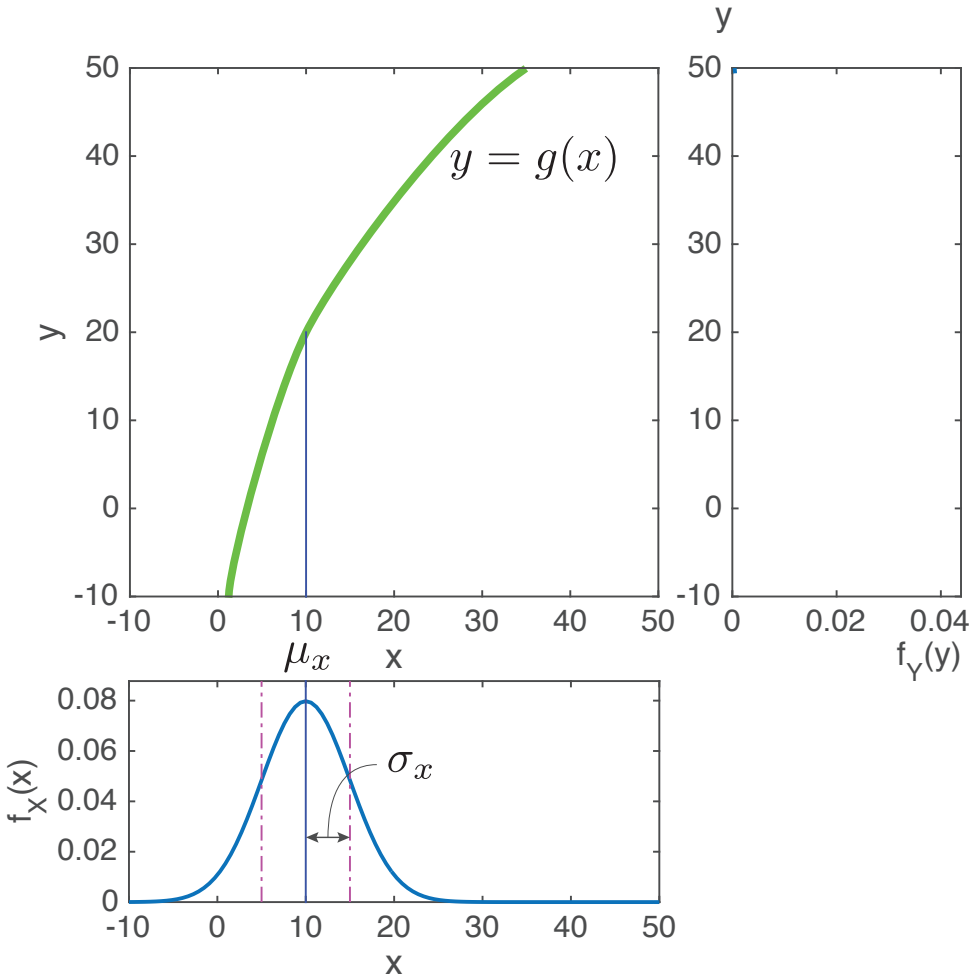
7.1 Definition

7.2 Taylor series

7.3 First order approximation

Linearization

We replace $g(X)$ by a linear function (e.g. a straight line)

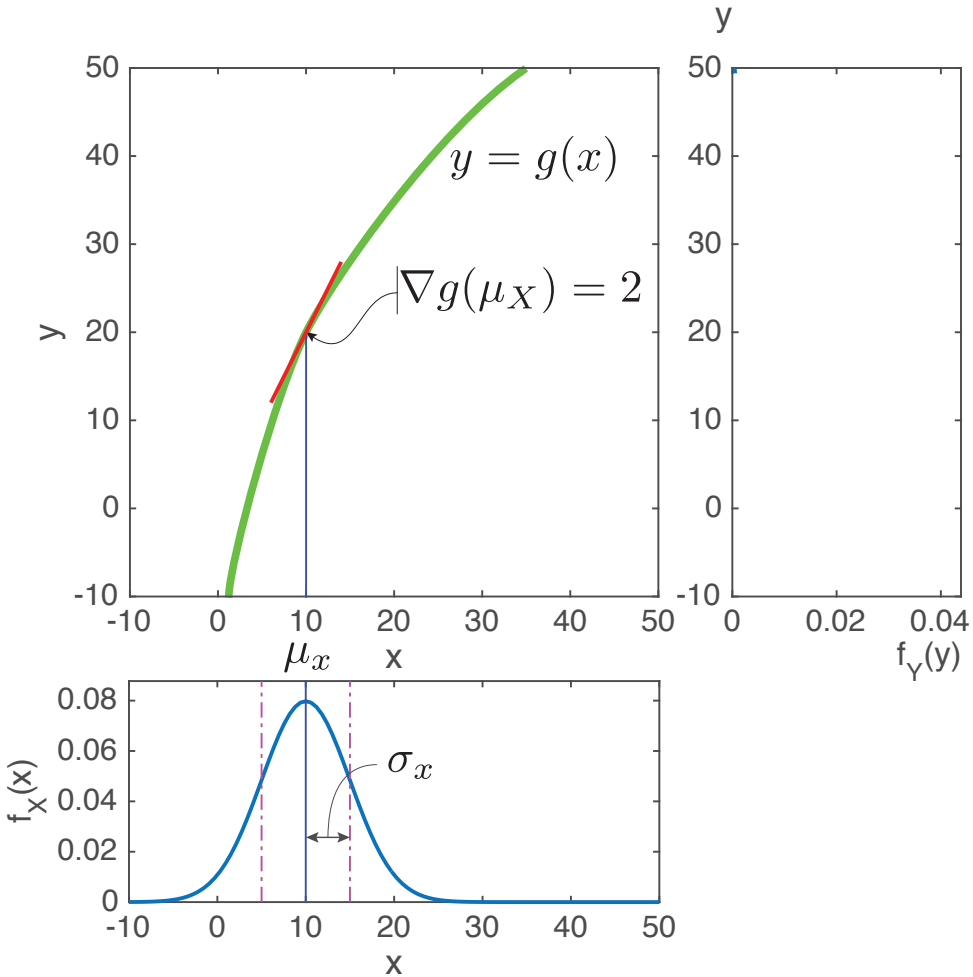


Linearization at the expected value μ_x
⚠ **high probability content**

$$\underbrace{\nabla g(\mu_x) = \left[\frac{dg(x)}{dx}\right]_{x=\mu_x}}_{\text{Gradient}}$$

Linearization

We replace $g(X)$ by a linear function (e.g. a straight line)

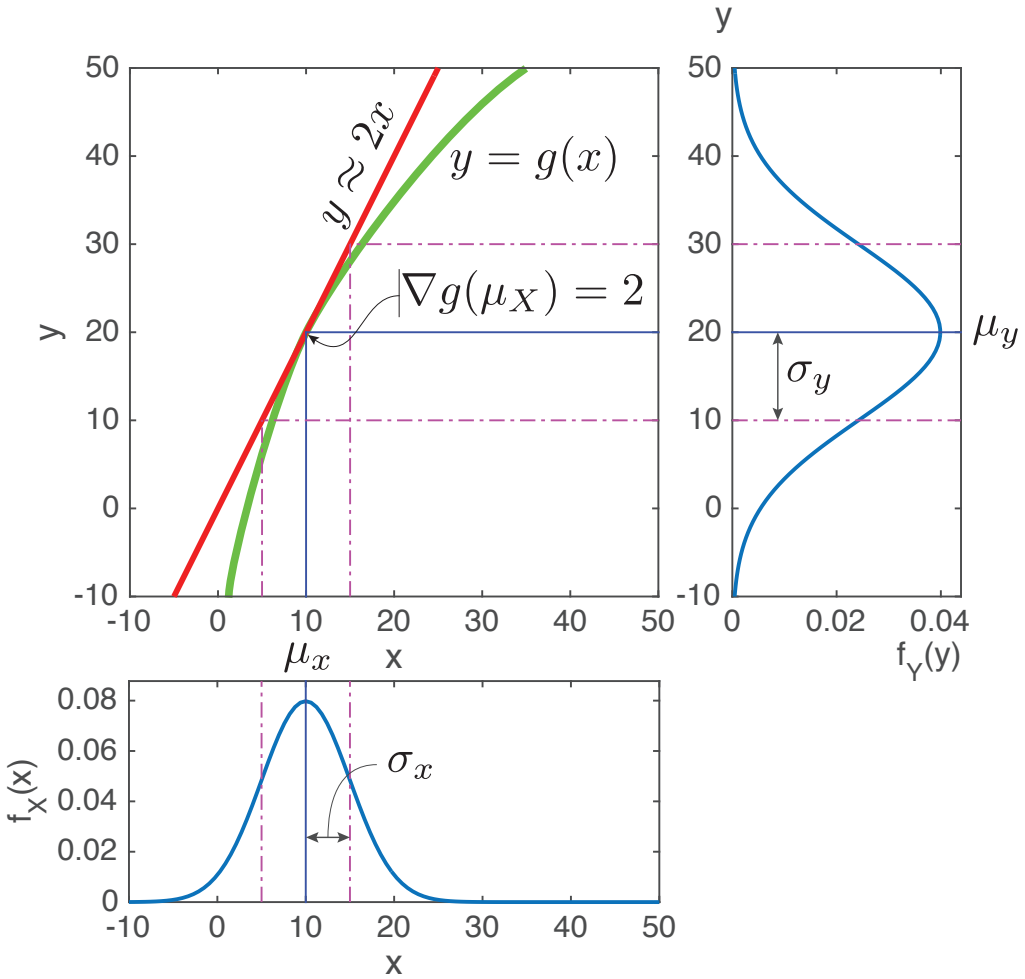


Linearization at the expected value μ_x
⚠ **high probability content**

$$\underbrace{\nabla g(\mu_x) = \left[\frac{dg(x)}{dx}\right]_{x=\mu_x}}_{\text{Gradient}}$$

Linearization

We replace $g(X)$ by a linear function (e.g. a straight line)



Linearization at the expected value μ_x
⚠ **high probability content**

$$\underbrace{\nabla g(\mu_x) = \left[\frac{dg(x)}{dx}\right]_{x=\mu_x}}_{\text{Gradient}}$$

Linearization – Taylor series around expected values $\mathbf{M_X}$

Given a non-linear function $Y = g(\mathbf{X}) \approx \mathbf{a}^\intercal \mathbf{X} + b$

$$Y \approx \overbrace{\underbrace{g(\mathbf{M_X}) + \overbrace{\nabla g(\mathbf{M_X})}^{\text{Gradient}}(\mathbf{X} - \mathbf{M_X})}_{\text{first order approxiamtion}} + \frac{1}{2}(\mathbf{X} - \mathbf{M_X})^\intercal \overbrace{\mathbf{H}(\mathbf{M_X})}^{\text{Hessian}}(\mathbf{X} - \mathbf{M_X}) + \cdots}^{\text{second order approximation}}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{m}^{th} \text{ order approximation}}$$

Note: For a non-linear function, the gradient/Jacobian changes depending where it is evaluated $\rightarrow \nabla g(\mathbf{M_X})$ / $\mathbf{J_{Y,X}}(\mathbf{M_X})$

Linearization – First order approximation

First order approximation centred on the expected value $\equiv$ linearization of $g(\mathbf{X})$ at $\mathbf{M_X}$

$$\begin{aligned} Y = g(\mathbf{X}) &\cong g(\mathbf{M_X}) + \nabla g(\mathbf{M_X})(\mathbf{X} - \mathbf{M_X}) \\ \mu_Y &\cong g(\mathbf{M_X}) \\ \sigma_Y^2 &\cong \nabla g(\mathbf{M_X}) \mathbf{\Sigma_X} \nabla g(\mathbf{M_X})^\intercal \end{aligned}$$

where

$$\underbrace{\nabla g(\mathbf{M_X}) = \left[\frac{\partial g(\mathbf{x})}{\partial x_1}, \cdots, \frac{\partial g(\mathbf{x})}{\partial x_n} \right]_{\mathbf{x}=\mathbf{M_X}}}_{\text{Gradient}}$$

Summary

Probabilities:

- Probabilities **describe our knowledge**
- The less we know, the more we should employ probability theory

Bayesian interpretation: $\Pr(E_i)$ quantifies the likelihood of an event with respect to others in $\mathcal{S}$

Rules/operations events: $\cap, \cup, \subset, \subseteq, \in$

Fundamental Axioms:

1. $0 \leq \Pr(E_i) \leq 1$
2. $\Pr(\mathcal{S}) = 1$
3. Si E_1 et E_2 are mutually exclusives $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$

Inclusion-exclusion rule: $\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \ldots$

Bayes Theorem: $\Pr(E_i|A) = \frac{\Pr(A|E_i)\Pr(E_i)}{\Pr(A)}$

Probability distributions: PDF, CDF, PMF, CMF

Multivariate Normal: $\mu_1, \sigma_1, \mu_2, \sigma_2, \rho_{12}$

Multivariate probability density function:

- $0 \leq f_{\mathbf{X}}(\mathbf{x})$
- $\int \cdots \int f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = 1$

Conditional probabilities:

- si $p_{X_1|X_2}(x_1|x_2) = p_{X_1}(x_1)$, $X_1 \perp\!\!\!\perp X_2$
- si $X_1 \perp\!\!\!\perp X_2$ $p_{X_1X_2}(x_1, x_2) = p_{X_1}(x_1)p_{X_2}(x_2)$

General case: $X_1 \not\perp\!\!\!\perp X_2 \rightarrow$ **Chain rule**

Expectation & Variance:

$$\mathbb{E}[X] = \int x \cdot f_X(x)dx \quad \text{(Continuous R.V.)}$$

$$\mathbb{E}[(X - \mu_X)^2] = \sigma_X^2 = \text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Matrix notation: $\mathbf{\Sigma_X} = \mathbf{D_X R_X D_X}$

Function of random variables: $Y = g(\mathbf{X})$

$$f_Y(y)dy = f_X(x)dx$$

$$\begin{aligned} \mathbf{M}_Y &= \mathbf{g}(\mathbf{M}_X) = \mathbf{A}\mathbf{M}_X + \mathbf{B} \\ \mathbf{\Sigma}_Y &= \mathbf{A}\mathbf{\Sigma_X}\mathbf{A}^\intercal = \mathbf{J_{y,x}}\mathbf{\Sigma}_X\mathbf{J_{y,x}}^\intercal \end{aligned}$$

Linearization – First order approximation

$$\begin{aligned} Y = g(\mathbf{X}) &\cong g(\mathbf{M_X}) + \nabla g(\mathbf{M_X})(\mathbf{X} - \mathbf{M_X}) \\ \mu_Y &\cong g(\mathbf{M_X}) \\ \sigma_Y^2 &\cong \nabla g(\mathbf{M_X})\mathbf{\Sigma_X}\nabla g(\mathbf{M_X})^\intercal \end{aligned}$$