Probability Theory

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

My goal is to give you theory foundations and practical tools for your research

I'll give lots of definitions, but the underlying concepts are typically simple

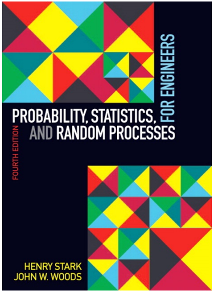
Do the exercises to check your understanding

All referenced Python code is in the `probability_theory` folder

I'm only giving you a small taste of this rich field - take further courses and study on your own!

I will cover material from

- **Stark & Wood's textbook "Probability, Statistics, and Random Processes for Engineers" [1]**
- Assorted other textbooks
- My own experience



Outline

1. What is probability?
2. Boolean and set algebra
3. Axiomatic definition of probability
4. Basic rules of probability

What is probability?

"Probability is a mathematical model to help us study physical systems in an average sense. We have to be able to repeat the experiment many times under the same conditions. Probability then tells us how often to expect the various outcomes." [1]

Why study and use probabilistic models?

"We are forced to use probabilistic models in the real world because we do not know, cannot calculate, or cannot measure all the causes contributing to an effect. The causes may be too complicated, too numerous, or too faint." [1]

Generic
"Probability" means the chance of something

Frequentist
"Probability" means the relative frequency of events

Bayesian
"Probability" means the degree to which we believe something to be true

Axiomatic
"Probability" is a mathematical construct that follows a set of rules

- No interpretation needed - conclusions follow logically from premises
- Be prepared for **counter-intuitive** conclusions

Preliminaries

Set

A **set** is a collection of individual **elements**.

Sets are denoted by braces, with the elements e_i contained inside

$$S = \{e_1, e_2, e_3, \ldots\} \tag{1}$$

Often constructed via set-builder notation

$$S = \{e_i \mid \text{predicate}(e_i)\} \tag{2}$$

"the set of all elements ee-eye such that the predicate holds for ee-eye"

An element e is "in" a set S if S contains e, denoted as $e \in S$.

The **cardinality** of a set is the number of elements in the set.

Examples of Sets

- The set of people reading this slide right now
- The set of hairs on your head
- The **empty set**, denoted $\emptyset$, the set containing nothing at all
  - $\emptyset$ is the only set with cardinality zero
- The set containing the empty set $\{\emptyset\}$
  - This set is not itself empty - it has cardinality one
- The **universal set**, denoted $\mathbb{U}$, the set containing every possible element
- The set of **whole numbers**, denoted $\mathbb{W} = \{0, 1, 2, 3, \ldots\}$
  - It has cardinality $\aleph_0$, a countable infinity
- The set of **real numbers**, denoted $\mathbb{R}$
  - It has cardinality $\mathfrak{c} = 2^{\aleph_0} > \aleph_0$, an uncountable infinity
    - See Cantor's diagonal argument from 1891

Boolean algebra

Basic mathematical operations that apply to **truth/false statements**

- Just like "standard" math operations that apply to numbers like addition, multiplication, etc.

Let x and y be two truth values

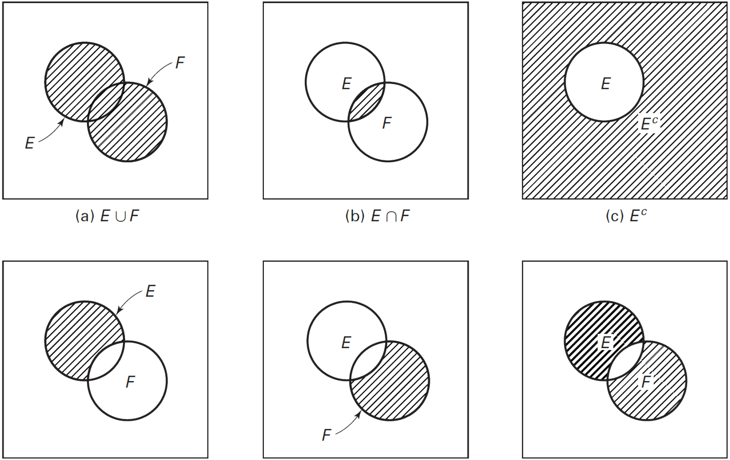
Operation	Notation	Definition
Disjunction	$x \vee y$	x is true **or** y is true
Conjunction	$x \wedge y$	x is true **and** y is true
Negation	$\neg x$	x is **not** true
Equivalence	$x \leftrightarrow y$	x is true **if and only if** y is true

Basic mathematical operations that apply to **sets**

- Defined with Boolean algebra applied to set membership

Let E and F be two sets

Operation	Notation	Definition
Union	$E \cup F$	Set of all elements in E **or** in F
Intersection	$E \cap F$	Set of all elements in E **and** F
Complement	E^c	Set of all elements **not** in E
Difference	$E - F$	Set of all elements in E **and not** in F
Exclusive Union	$E \oplus F$	Set of all elements in E **or** F **and not in both**
Subset	$E \subset F$	Every element in E is also in F
Superset	$E \supset F$	Every element in F is also in E
Equality	$E = C$	Every element in E is also in F and vice versa.



(a) $E \cup F$ (b) $E \cap F$ (c) E^c

(d) $E - F$ (e) $F - E$ (g) $E \oplus F$

Figure 1: Set operations: (a) Union (b) Intersection (c) Complement (d) Difference (e) Difference (f) Exclusive Union

Set terminology

Let $\{E_i\}$ be a collection of sets

Let A be another set (if unspecified, the universal set $A = \mathbb{U}$ is implied)

- $\{E_i\}$ is **disjoint** or **mutually exclusive** if no elements are shared between any two different sets
- $\{E_i\}$ **collectively exhausts** A if the union of $\{E_i\}$ is A
- $\{E_i\}$ **partitions** A if $\{E_i\}$ is disjoint and collectively exhausts A

Set algebra laws

Set operations are related by simple laws, can be proved using Boolean logic (e.g. truth tables) and definitions

Examples:

- $E = F \quad \leftrightarrow \quad (E \subset F) \wedge (E \supset F)$
- $E \cap E^c = \emptyset$
- $E \cup E^c = \mathbb{U}$
- $E - F = E \cap F^c$
- $E \oplus F = (E - F) \cup (F - E) = (E \cup F) \cap (E \cap F)^c$

More set algebra laws

De Morgan's laws

- $\left[\bigcup_{i=1}^{n} E_i\right]^c = \bigcap_{i=1}^{n} E_i^c$
- $\left[\bigcap_{i=1}^{n} E_i\right]^c = \bigcup_{i=1}^{n} E_i^c$

Associative laws

- $A \cup (B \cup C) = (A \cup B) \cup C$
- $A \cap (B \cap C) = (A \cap B) \cap C$

Distributive laws

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Events

Outcome

A **random experiment** results in **individual outcomes**, denoted as ζ.

Sample space

The **sample space** of a random experiment is the set of all possible outcomes of the experiment, denoted as Ω.

Event

An **event** is a subset of the sample space i.e. a set of outcomes.

In probability

- The sample space plays Ω the role of the universal set $\mathbb{U}$, and is called the **certain event**.
- The empty set $\emptyset$ is called the **null event**.
- Any individual outcome ζ is an element of Ω.

Sigma fields (sigma algebras)

Field

The collection of events $\mathcal{F} = \{E_i\}$ is a **field** if

1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$
2. If $E_i \in \mathcal{F}$ for all $i = 1, \ldots, n$, then $\bigcup_{i=1}^{n} E_i \in \mathcal{F}$ and $\bigcap_{i=1}^{n} E_i \in \mathcal{F}$
   - "Closed under **finite** union and intersection"
3. If $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$
   - "Closed under complement"

If condition 2 further holds with n countably infinite i.e. "closed under **countably infinite** union and intersection", then $\mathcal{F}$ is a **sigma (σ) field**.

Ensures any union, intersection, and complement of any set of events is well-defined (by construction).

Sigma fields (sigma algebras)

If Ω is continuous and thus uncountable, e.g. $\Omega = \mathbb{R}$, we can generate a sigma field from the set of all open and closed intervals in Ω.

- In this case the sigma field is called the **Borel field**.

We can compute sigma fields of finite and discrete Ω using combinatorics

- See `sigma_field.py`

Probability measure

Axiomatic definition of probability

Probability is a function that maps events to real numbers $P[\cdot] : \mathcal{F} \to [0, 1]$ that satisfies three axioms

1. $P[E] \geq 0$
2. $P[\Omega] = 1$
3. $P[E \cup F] = P[E] + P[F]$ if $P[EF] = \emptyset$

From the axioms we can establish the additional properties

4. $P[\emptyset] = 0$
5. $P[E - F] = P[E] - P[E \cap F]$
6. $P[E^c] = 1 - P[E]$
7. $P[E \cup F] = P[E] + P[F] - P[EF]$

Probability formulation example

Example: Single coin flip

- Sample space is $\Omega = \{H, T\}$ where $H =$ heads, $T =$ tails
- There are 2^2 possible events, $\emptyset$, H, T, Ω
  - Consider events H and T with equal probability
- σ-field is $\mathcal{F} = \{\emptyset, H, T, \Omega\}$

Example: Die roll

- Sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
- There are 2^6 possible events, each one containing, or not, each of the 6 possible outcomes
  - Consider events $\{1, 3\}$ and $\{2, 3, 4\}$
  - Consider each singleton event equally probable i.e. $P[\{i\}] = 1/6$
- σ-field is...tedious - see Example 1.4-9 [1]

Probability of a union

Probability of a union of disjoint events

Let $\{E_i\}_{i=1}^n$ be a set of mutually disjoint events, i.e.
$E_i \cap E_j = \phi$ for all $i \neq j$.
Then

$$P\left[\bigcup_{i=1}^n E_i\right] = \sum_{i=1}^n P\left[E_i\right]. \tag{3}$$

Proof: Use mathematical induction with Axiom 3.

Probability of a union

Union bound (Boole's inequality)

Let $\{E_i\}_{i=1}^{n}$ be a set of events.
Then

$$P\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} P\left[E_i\right]. \tag{4}$$

Proof: Use mathematical induction with Axiom 7.

Note: The only difference vs the previous result is that the events E_i are not assumed disjoint - the union bound always applies!

Bonferroni inequality

Let $\{E_i\}_{i=1}^n$ be a set of events. Define the sums

$$S_m = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} P\left[\bigcap_{j=1}^{m} E_{i_j}\right] \tag{5}$$

Then for any $k \in \{1, \ldots, n\}$

$$P\left[\bigcup_{i=1}^{n} E_i\right] \left\{\begin{array}{ll} \leq & \text{if } k \text{ odd} \\ \geq & \text{if } k \text{ even} \\ = & \text{if } k = n \end{array}\right\} \sum_{j=1}^{k} (-1)^{j-1} S_j \tag{6}$$

Proof: Use mathematical induction, see Theorem 1.5-1 in [1].

Note: Bonferroni is more tedious, but gives tighter bounds than Boole

Joint and conditional probability

Let A and B be two events with nonzero probability.

Joint probability

The **joint probability** of events A and B is the probability of their intersection $P[A \cap B]$.

Intuitively, it is the probability that both A and B will occur.

Conditional probability

The **conditional probability** of event A given B is the ratio

$$P[A|B] = \frac{P[A \cap B]}{P[B]}. \tag{7}$$

Intuitively, it is the probability that event A will occur, given the knowledge that event B already occurred.

Product Rule for events

Product Rule for events

The joint probability of events A and B can be computed as

$$P[A \cap B] = P[B|A]P[A] \tag{8}$$

When the events are independent we recover the

Proof: Follows by rearranging the definition of conditional probability.

Sum Rule for events

Suppose the events $\{A_i\}_{i=1}^n$ are disjoint and collectively exhaustive, i.e.

- $A_i \cap A_j = \emptyset$ for any $i \neq j$
- $\bigcup_{i=1}^n A_i = \Omega$

Then the **total probability** of event B can be computed as

$$P[B] = \sum_{i=1}^{n} P[B|A_i]P[A_i] = \sum_{i=1}^{n} P[B \cap A_i] \tag{9}$$

Proof: Follows by the product rule and the assumptions on the A_i's.

The sum rule is useful when the conditional probabilities or intersection probabilities are readily available but the total probability is not.

The sum rule is also known as the **law of total probability**.

The total probability is also known as the **marginal probability**, since we are *marginalizing out* the other events A_i.

Total probability - example

Microchip factories

Given information:

1. Factory A makes 4000 chips/day with defect rate of 5%
2. Factory B makes 2000 chips/day with defect rate of 2%
3. Chips from both factories are mixed together at the end of each day then sent to a lab for testing

Question:
What is the probability of getting a defective chip at the lab?

Solution:
Denote the following events:

- D: Chip is defective
- A: Chip is from factory A
- B: Chip is from factory B

First compute base probabilities from frequency of occurrence:

$$P[A] = \frac{4000}{4000 + 2000} = 66.7\% \tag{10}$$

$$P[B] = \frac{2000}{4000 + 2000} = 33.3\% \tag{11}$$

Now use the law of total probability:

$$\begin{aligned} P[D] &= P[D|A]P[A] + P[D|B]P[B] && (12) \\ &= (5\%)(66.7\%) + (2\%)(33.3\%) && (13) \\ &= \boxed{4\%} && (14) \end{aligned}$$

Statistical independence

Two events A and B are **statistically independent** if and only if

$$P[A \cap B] = P[A]P[B]. \tag{15}$$

Equivalently, the conditional and unconditional probabilities of A and B are equal:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A] \tag{16}$$

$$P[B|A] = \frac{P[B \cap A]}{P[A]} = \frac{P[B]P[A]}{P[A]} = P[B] \tag{17}$$

Intuitively, the outcome B has no effect on the chance of A occurring, and vice versa.

Statistical independence

What if there are more than 2 events?

Joint statistical independence

The events $\{A_i\}_{i=1}^n$ are **jointly statistically independent** if and only if for all $k \in \{1, 2, \dots, n\}$

$$P\left[\bigcap_{1 \le i_1 < i_2 < \cdots \le i_k} A_{i_k}\right] = \prod_{1 \le i_1 < i_2 < \cdots \le i_k} P\left[A_{i_k}\right] \tag{18}$$

Note: pairwise independence does not suffice!

- See e.g. this note `http://faculty.washington.edu/fm1/394/Materials/2-3indep.pdf`

Statistical independence

Pit-stop to build your intuition

Question: Can two disjoint events A and B with $P[A] > 0$, $P[B] > 0$ be statistically independent?

Think about it for a moment

Claim: No, A and B **must be dependent**

Explanation:

1. A, B disjoint means $A \cap B = \emptyset$ which implies $P[A \cap B] = 0$
2. $P[A] > 0$, $P[B] > 0$ implies $P[A]P[B] > 0$
3. Therefore $P[A \cap B] \neq P[A]P[B]$ and the claim follows

Intuition: If we know we flipped heads on a coin, that tells us we did not flip tails.

Bayes' theorem

Derivation from definition of conditional probabilities:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}, \tag{19}$$

$$P[B|A] = \frac{P[A \cap B]}{P[A]} \tag{20}$$

Notice the numerators of the right sides are the same!

Rearrange first line into

$$P[A \cap B] = P[A|B]P[B] \tag{21}$$

and put it into the second line to get Bayes' theorem

$$\boxed{P[B|A] = \frac{P[A|B]P[B]}{P[A]}} \tag{22}$$

Intuition: Lets us reason about conditional probability of "flipped" events

Bayes' theorem

Cancer test

Denote the events

- A: test says patient has cancer
- B: patient actually has cancer

Given information:

- Test has an accuracy of 95%
  - 95% of the time when the test says the patient has cancer, they actually do
  - 95% of the time when the test says the patient does not have cancer, they actually do not
- The cancer rate in the population is 0.5%

Question: The patient being tested for cancer cares about the chance they actually have cancer given the test says they do.
What is this probability?

Solution:
Translate given information into math:

$$P[A|B] = P[A^c|B^c] = 95\%, \qquad P[B] = 0.5\% \tag{23}$$

Use the law of total probability to find $P[A]$, the probability of the test saying a patient has cancer:

$$\begin{aligned} P[A] &= P[A|B]P[B] + P[A|B^c]P[B^c] && (24) \\ &= (95\%)(0.5\%) + (100\% - 95\%)(100\% - 0.5\%) && (25) \\ &= 5.45\% && (26) \end{aligned}$$

Now use Bayes' theorem:

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]} = \frac{(95\%)(0.5\%)}{5.45\%} \approx \boxed{8.72\%} \tag{27}$$

Bayes' theorem

How do we resolve this counter-intuitive result?

Even though the test is highly accurate (95%), the chance of actually having cancer is low (8.72%), despite a positive test result. This is because the base rate of cancer is very small, only 0.5%.

On the other hand, conditioning on a positive test result makes the chance of cancer increase dramatically in a relative sense from 0.5% to 8.72%.

From the standpoint of the designer of the cancer test, the smaller the base rate of cancer, the more accurate the test has to be to yield the same probability of a patient actually having cancer.

Homework P1-1:
Consider the previous example. Compute the probability that a patient has cancer, given a negative test result.

Homework P1-2: (1.33 in [1])
A large class in probability theory is taking a multiple-choice test. For a particular question on the test, the fraction of examinees who know the answer is p; $1-p$ is the fraction that will guess. The probability of answering a question correctly is unity for an examinee who knows the answer and $1/m$ for a guessee; m is the number of multiple-choice alternatives.

1. Compute the probability that an examinee knew the answer to a question given that he or she has correctly answered it in terms of m and p.
2. Then evaluate this probability for the specific choice $m = 4$ and $p = 50\%$.

Homework P1-3: (1.35 in [1])
Assume there are three machines A, B, and C in a semiconductor manufacturing facility that make chips. They manufacture, respectively, 25, 35, and 40 percent of the total semiconductor chips there. Of their outputs, respectively, 5, 4, and 2 percent of the chips are defective. A chip is drawn randomly from the combined output of the three machines and is found defective. What is the probability that this defective chip was manufactured by machine A? by machine B? by machine C?

Homework P1-4: (1.55 in [1])
An automatic breathing apparatus (B) used in anesthesia fails with probability P_B. A failure means death to the patient unless a monitor system (M) detects the failure and alerts the physician. The monitor system fails with probability P_M. The failures of the system components are independent events. Professor X, an M.D. at Hevardi Medical School, argues that if $P_M > P_B$ installation of M is useless. Compute the probability of a patient dying with and without the monitor system in place. Take $P_M = 0.1 = 2P_B$. Is Professor X correct in his assessment?

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

Random Variables

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Random variables

Random variable

A **random variable (RV)** X is a function that maps the sample space Ω to real numbers $\mathbb{R}$ i.e. $X : \Omega \to \mathbb{R}$ that satisfies the following properties:

1. For every Borel set of numbers B, the set $E_B = \{\zeta \in \Omega, X(\zeta) \in B\}$ is an event.
2. $P[X = \infty] = P[X = -\infty] = 0$

Realizations

Realizations

Upon outcome ζ, a random variable produces a **realization** / **observation** $X(\zeta)$, which is simply a number.

- Think of a realization "popping into being" upon some trigger.
- As shorthand we often refer to the realizations by the same name/variable as the RV.
- We can only observe realizations of the random variable, but not the random variable itself.
- Qualities of the random variable must either be
  1. Assumed before-hand (model)
  2. Inferred from realizations (data)

Random variables - examples

Flip a coin:
X is one or zero for heads or tails respectively

Roll a die:
X is 1, 2, 3, 4, 5, 6, corresponding to the number of dots on the die face

Spin a wheel:
X is the angle at which it lands between 0 and 360 degrees

Cumulative distribution function (cdf)

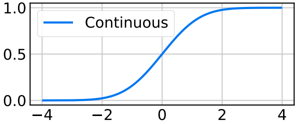
The **cumulative distribution function (cdf)** is defined as

$$F_X(x) = P[\{\zeta | X(\zeta) \leq x\}] \tag{1}$$

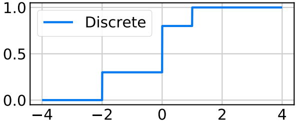
Notation: From here we will usually drop the notation of ζ related to the underlying probability space, so $P[\{\zeta | X(\zeta) \leq x\}]$ becomes $P[X \leq x]$.

Continuous and discrete random variables

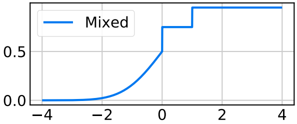
If the cdf $F_X(x)$ is everywhere continuous and differentiable, then X is a **continuous random variable**.



If the cdf $F_X(x)$ is piecewise constant (stairstep shape), then X is a **discrete random variable**.



If neither holds, then X is a **mixed random variable**.



See `mixed.py`

Mass and density functions

Probability mass function (pmf)

The **probability mass function (pmf)** of a discrete random variable is defined as

$$P_X(x) = P[X = x] \tag{2}$$
$$= P[X \leq x] - P[X < x] \tag{3}$$

Probability density function (pdf)

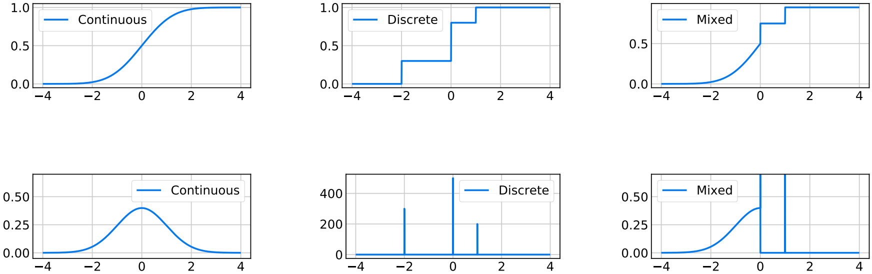
The **probability density function (pdf)** of a continuous random variable* is defined as

$$f_X(x) = \frac{d}{dx} F_X(x) \tag{4}$$

* By introducing Dirac delta functions, the pdf can be defined for discrete and mixed random variables.

Continuous and discrete random variables

cdfs on top row, pdfs on bottom row



See `mixed.py`

1. $F_X(\infty) = 1$, $F_X(-\infty) = 0$
2. $F_X(x)$ is nondecreasing in x,
   i.e. $X_1 \leq x_2$ implies $F_X(x_1) \leq F_X(x_2)$
3. $F_X(x)$ is continuous from the right,
   i.e. $F_X(x) = \lim_{\epsilon \to 0^+} F_X(x + \epsilon)$

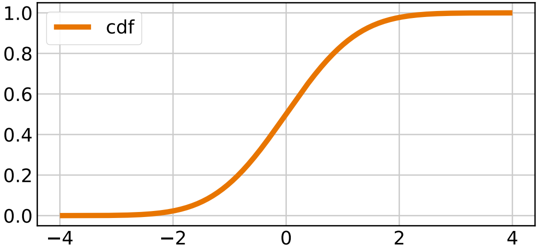


Figure 1: Plot of a typical cdf (std normal)

Properties of the pdf

1. $f_X(x) \geq 0$
2. $\int_{-\infty}^{\infty} f_X(\xi) d\xi = F_X(\infty) - F_X(-\infty) = 1$
3. $F_X(x) = \int_{-\infty}^{x} f_X(\xi) d\xi = P[X \leq x]$
4. $F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(\xi) d\xi = P[x_1 < X \leq x_2]$

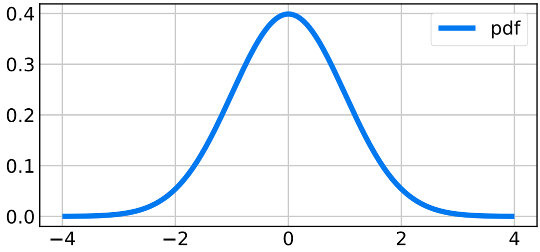


Figure 2: Plot of a typical pdf (std normal)

Knowledge of either the pdf or cdf is sufficient to compute the other, via integration or differentiation.

When we refer to a “distribution,” we mean anything that fully specifies a random variable:

- pdf / pmf
- cdf
- Moment generating function (see Ch. 4.5 of [1])
- Characteristic function (see Ch. 4.7 of [1])

Let’s introduce a couple of quick concepts before we survey various distributions

Support of a distribution

The **support** of a distribution is the set of values that the random variable X can take with nonzero probability density, i.e.

$$\mathrm{supp}(X) = \{x \mid f_X(x) > 0\}. \tag{5}$$

The distinction between the support and the sample space only comes into effect when the sample space is bigger than required by X

- Sometimes convenient when working with different random variables on a shared sample space
- Example: Two dice with faces $\{1, 1, 1, 2, 3, 3\}$ and $\{2, 3, 4, 5, 6, 6\}$ have different supports $\{1, 2, 3\}$ and $\{2, 3, 4, 5, 6\}$, but we might want a sample space $\{1, 2, 3, 4, 5, 6\}$ to accommodate every possible outcome from either of dice

Mixture distributions

Mixture distribution

A **mixture distribution** is the distribution of a **mixture random variable** Y formed as a composite of other component random variables $X_1, X_2, \ldots, X_N$ by selecting among them at random according to weights $w_1, w_2, \ldots, w_N$.

If the component pdfs are $f_{X_1}, f_{X_2}, \ldots f_{X_N}$, then the **mixture pdf** is simply the weighted average

$$f_Y(Y) = \sum_{i=1}^{N} w_i f_{X_i} \tag{6}$$

Discrete distributions

Trivial distribution

What happens if we treat a non-random, fixed, constant number as a random variable? (w.l.o.g. set $X = 0$)

Trivial distribution

A **trivial** random variable has the pmf

$$P[X = x] = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Accordingly, the pdf is the Dirac delta function

$$f_X(x) = \delta(x) \tag{8}$$

and the cdf is the Heaviside step function

$$F_X(x) = H(x) \tag{9}$$

All discrete distributions can be "built" from mixtures of this distribution.

- Follows by definition of pmf

Bernoulli distribution

A **Bernoulli** random variable has the pmf

$$P[X = x] = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (10)$$

If p is not specified, then assume $p = 1/2$.

Example: A coin flip is Bernoulli where heads $= 1$ and tails $= 0$.

Rademacher distribution

A **Rademacher** random variable has the pmf

$$P[X=x] = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = -1, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Basically just the symmetric version of Bernoulli (which is asymmetric)

- Use whichever is most convenient for the task at hand

Example: A coin flip is Rademacher where heads $= 1$ and tails $= -1$.

Homework P2-1: If X is a Bernoulli random variable, write down a function g such that $Y = g(X)$ is a Rademacher random variable. Also, write down an inverse function $h = g^{-1}$ such that $X = h(Y)$ recovers a Bernoulli distribution. Prove that your functions are correct by directly evaluating the pmfs of $g(X)$ and $h(Y)$.

Binomial distribution

Consider the binomial experiment with n independent success/fail trials, each governed by a Bernoulli RV.

The number of ways to choose k elements from a population of size n (irrespective of their ordering) is called the number of **combinations** and is determined by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{12}$$

The probability of an experiment with k successes and $n-k$ failures is

$$p^k (1-p)^{n-k} \tag{13}$$

Since there are $\binom{n}{k}$ ways in which the experiment could end like this, the probability of seeing an experiment with k successes and $n-k$ failures is

$$\binom{n}{k} p^k (1-p)^{n-k} \tag{14}$$

Binomial distribution

A random variable X follows a **binomial distribution** if it represents getting exactly k successes out of the n trials, whose pmf is

$$P[X=k] = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, \ldots, n, \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $p \in [0,1]$ is a parameter representing the success probability of each trial.

Homework P2-2: (1.56 in [1])
In a particular communication network, the server broadcasts a packet of data to N receivers. The server then waits to receive an acknowledgment message from each of the N receivers before proceeding to broadcast the next packet. If the server does not receive all the acknowledgments within a certain time period, it will rebroadcast (retransmit) the same packet. The server is then said to be in the "retransmission mode." It will continue retransmitting the packet until all N acknowledgments are received. Then it will proceed to broadcast the next packet.

Let $p := P$[successful transmission of a single packet to a single receiver along with successful acknowledgment]. Assume that these events are independent for different receivers and separate transmission attempts. Due to random impairments in the transmission media and the variable condition of the receivers, we have that $p < 1$.

(continued on next slide)

Homework P2-2 (cont.):
(a) In a fixed protocol of method of operation, we require that all N of the acknowledgments be received in response to a given transmission attempt for that packet transmission to be declared successful. Let the event $S(m)$ be defined as follows: $S(m) := \{$ a successful transmission of one packet to all N receivers in m or fewer attempts $\}$.

Find the probability

$$P(m) := P[S(m)]$$

Hint: Consider the complement of the event $S(m)$.

(continued on next slide)

Homework P2-2 (cont.):
(b) An improved system operates according to a dynamic protocol as follows. Here we relax the acknowledgment requirement on retransmission attempts, so as to only require acknowledgments from those receivers that have not yet been heard from on previous attempts to transmit the current packet. Let $S_D(m)$ be the same event as in part (a) but using the dynamic protocol. Find the probability

$$P_D(m) := P[S_D(m)]$$

Hint: First consider the probability of the event $S_D(m)$ for an individual receiver, and then generalize to the N receivers.

(continued on next slide)

Homework P2-2 (cont.):
(c) Compare the performance of the two protocols from parts (a) and (b) by comparing $P(m)$ and $P_D(m)$ for $N = 5$ receivers, $m = 2$ transmission attempts, and success probability $p = 0.9$.

Continuous distributions

Uniform distribution

A random variable is **uniform** if the pdf is constant over a finite interval $[a, b]$, i.e. of the form

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Tail behavior: density drops to zero instantly outside $[a, b]$

- Log density decays "infinitely" fast

Homework P2-3: Derive an expression for the cdf of a uniform random variable.

Gaussian distribution

A random variable X is **Gaussian** or **normal** if it has a pdf of the form

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{17}$$

where μ and σ^2 are parameters (we will define and see later they are the mean and variance).

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$ is read as "X is distributed according to a normal distribution with mean mu and variance sigma-squared."

Special case: If $\mu = 0$ and $\sigma^2 = 1$, then the distribution is called the **standard normal**.

Tail behavior: log density decays quadratically

See `gaussian.py`

Exponential distributions

Exponential distribution

A random variable X is **exponential** if it has a pdf of the form

$$f_X(x) = \begin{cases} \lambda \text{exp}(-\lambda x) & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where $\lambda > 0$ is a parameter.

Homework P2-4: Derive an expression for the cdf of an exponential random variable.

Laplace distribution

A random variable X is **Laplace** or **double exponential** if it has a pdf

$$f_X(x) = \frac{1}{2\beta}\exp\left(-\frac{|x-\mu|}{\beta}\right) \tag{19}$$

where μ and β are location and scale parameters.

Notice how similar the Laplace distribution is to a Gaussian

Tail behavior: log density decays linearly - heavier than a Gaussian!

Cauchy distribution

A random variable X is **Cauchy** if it has a pdf of the form

$$f_X(x) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x-x_0)^2+\gamma^2}\right) \tag{20}$$

where x_0 and γ are location and scale parameters.

Example: The ratio of two independent normal variables $X = Z_1/Z_2$ is Cauchy

The Cauchy distribution is very bizarre pathological distribution

- It actually has an undefined mean and variance! (discussed later)
- Makes parameter estimation tricky

Tail behavior: log density decays logarithmically - heavier than a Laplace!

Other distributions

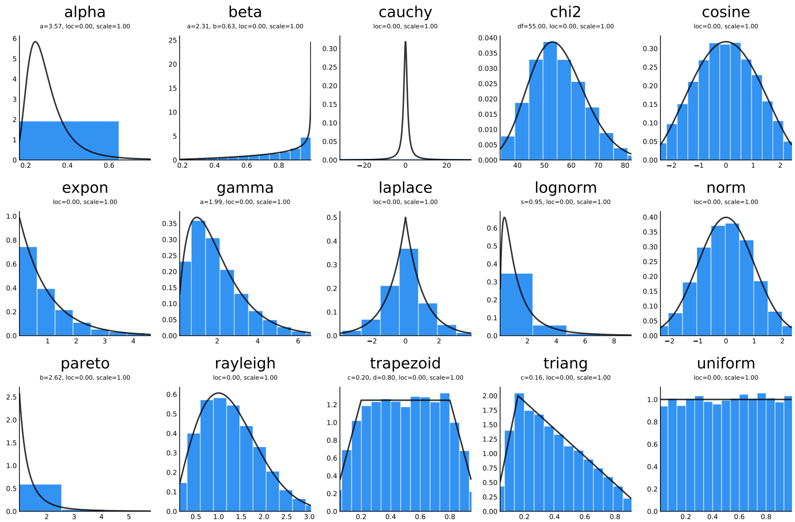


Figure 3: Plot of various pdfs available in SciPy - see `distributions.py`

Conditional distributions

We can condition random variables on random events

Conditional distribution function

The **conditional distribution function of X given event B** is

$$F_X(x|B) = \frac{P[X \leq x \text{ and } B]}{P[B]} \tag{21}$$

Conditional density function

The **conditional density function of X given event B** is

$$f_X(x|B) = \frac{d}{dx} F_X(x|B) \tag{22}$$

Joint distributions

Just as we had the joint probability of two events, we have the joint distribution of two random variables

Joint distribution function

The **joint (cumulative) distribution function** of X and Y is

$$F_{XY}(x, y) = P[X \leq x \text{ and } Y \leq y] \tag{23}$$

Joint probability mass function

The **joint probability mass function** of X and Y is

$$P_{XY}(x, y) = P[X = x, Y = y] \tag{24}$$

Joint density function

The **joint density function** of X and Y is

$$f_{XY}(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{XY}(x, y) \tag{25}$$

Multinomial distribution

Here is an example of a joint distribution

Idea: Generalize the binomial distribution to trials with more than two outcomes

Consider the multinomial experiment with n independent trials with m outcomes, with each trial governed by a discrete RV with success probabilities $\{p_i\}_{i=1}^m$.

The number of times each outcome happens throughout the entire experiment is a discrete RV X_i for $i = 1, \ldots, m$.

We are interested in the probability that the ith outcome appears exactly k_i times i.e. the joint distribution of the X_i.

Multinomial distribution

The **multinomial coefficient** is the number of ways that the ith outcome appears exactly k_i times (irrespective of their ordering):

$$\frac{n!}{k_1!k_2!\cdots k_m!} \tag{26}$$

The probability of an experiment with the ith outcome appearing exactly k_i times (irrespective of their ordering) is

$$\prod_{i=1}^{m} p_i^{k_i} \tag{27}$$

Since there are $\frac{n!}{k_1!k_2!\cdots k_m!}$ ways in which the experiment could end with the ith outcome appearing exactly k_i times, the probability of seeing such an experiment is

$$\frac{n!}{k_1!k_2!\cdots k_m!} \prod_{i=1}^{m} p_i^{k_i} \tag{28}$$

Multinomial distribution

A collection of RVs $\{X_i\}_{i=1}^m$ follows a **multinomial distribution** if it represents the multinomial experiment, whose joint pmf is

$$P[X_1 = k_1, X_2 = k_2, \dots, X_m = k_m] \tag{29}$$

$$= \begin{cases} \frac{n!}{k_1! k_2! \cdots k_m!} \prod_{i=1}^m p_i^{k_i} & \text{if } \sum_{i=1}^m k_i = n, \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

where $\{p_i\}_{i=1}^m$ is a set of parameters representing the success probabilities, and must satisfy $\sum_{i=1}^m p_i = 1$.

Exercise: As a special case, how can we recover the binomial distribution from the multinomial distribution?

Marginal distributions

If we have a joint distribution in hand, we can get the distribution of each of the components by integrating ("marginalizing")

Marginal density function

The **marginal density functions** are

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \tag{31}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx \tag{32}$$

Marginal distribution function

The **marginal distribution functions** are

$$F_X(x) = F_{XY}(x,\infty) = \int_{-\infty}^{x} f_X(\xi)d\xi \tag{33}$$

$$F_Y(y) = F_{XY}(\infty,y) = \int_{-\infty}^{y} f_Y(\eta)d\eta \tag{34}$$

Conditional distributions

We can also condition random variables on other random variables

Conditional density function

The **conditional density function of X given Y** is

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{35}$$

Conditional distribution function

The **conditional distribution function of X given Y** is

$$F_{X|Y}(x|y) = P[X \leq x | Y \leq y] = \int_{-\infty}^{x} f_{X|Y}(\xi|y) d\xi \tag{36}$$

Notice that $F_{X|Y}(x|y) \neq \frac{F_{XY}(x,y)}{F_Y(y)}$ (unlike the conditional pdf)

Sum Rule for random variables

Let X and Y be two discrete random variables.

The probability that X takes the value x_i, irrespective of the value of Y, is the **total probability** of $X = x_i$, written as $P[X = x_i]$.

Sum Rule for random variables

The total probability of X can be computed as

$$P[X = x_i] = \sum_j P[X = x_i | Y = y_j] P[Y = y_j] \tag{37}$$

$$= \sum_j P[X = x_i, Y = y_j]. \tag{38}$$

This follows from the law of total probability for the event $X = x_i$ and the fact that all the events $Y = y_j$ partition the sample space of Y.

The **total probability** is also referred to as the **marginal probability**, *since we are marginalizing out* the other variable, Y.

Product Rule for random variables

Let X and Y be two discrete random variables.

Conditional probability (Again!)

For only the instances for which $A = a_i$, the fraction of such instances for which $B = b_j$ is $P[B = b_j | A = a_i]$ and are called the **conditional probability of $B = b_j$ given $A = a_i$**.

Product Rule for random variables

The joint pmf of X and Y can be computed as

$$P[X = x_i, Y = y_j] = P[Y = y_j | X = x_i] P[X = x_i] \tag{39}$$

Bayes' theorem for random variables

RV conditioned on RV

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} \tag{40}$$

Event conditioned on RV

$$P[A|X = x] = \frac{f_{X|A}(x)P[A]}{f_X(x)} \tag{41}$$

RV conditioned on event

$$f_{Y|A}(y) = \frac{P[A|Y = y]f_Y(y)}{P[A]} \tag{42}$$

Independent random variables

Two random variables X and Y are **statistically independent** if the two events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for any pair (x, y).

Equivalently,

$$F_{XY}(x, y) = F_X(x)F_Y(y) \tag{43}$$

or

$$f_{XY}(x, y) = f_X(x)f_Y(y) \tag{44}$$

You can imagine the generalization to more than two RV's - joint distribution is equal to product of the marginals

It is nice when RV's are independent because it makes computing their joint distribution trivial - just multiply the marginals!

Functions of random variables

Core problem:
What is the distribution of a function of a random variable?

Math:
Given $f_X(x)$ and $Y = g(X)$, what is $f_Y(y)$?

"Indirect" procedure:

1. Find the point set C_y such that $\{Y \leq y\} = \{X \in C_y\}$
2. Find the cdf of Y as

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq y] = P[X \in C_y] \tag{45}$$

3. Find the pdf of Y as

$$f_Y(y) = \frac{d}{dy} F_Y(y) \tag{46}$$

Example: affine function of a random variable

Suppose g is affine, i.e. $Y = g(X) = aX + b$.

Case 1: $a > 0$

Step 1: Find the point set

$$\{Y \leq y\} = \{aX + b \leq y\} \tag{47}$$

$$= \left\{X \leq \frac{y-b}{a}\right\} = \{X \in C_y\} \tag{48}$$

Step 2: Find the cdf

$$F_Y(y) = P[Y \leq y] = P[aX + b \leq y] \tag{49}$$

$$= P\left[X \leq \frac{y-b}{a}\right] = F_X\left(\frac{y-b}{a}\right) \tag{50}$$

Step 3: Differentiate cdf to get pdf
Use the change of variables $z = \frac{y-b}{a}$ so

$$
\begin{aligned}
f_Y(y) = \frac{d}{dy} F_Y(y) &= \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) && (51) \\
&= \frac{dF_X(z)}{dz} \cdot \frac{dz}{dy} && \text{(chain rule)} \\
&= f_X(z) \cdot \frac{1}{a} && (52)
\end{aligned}
$$

Optional Exercise: Work out Case 2: $a < 0$

After doing that, you will find the solution is

$$\boxed{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \qquad \text{if } a \neq 0} \tag{53}$$

Optional Exercise: Work out Case 3: $a = 0$ (degenerate case)
Hint: The solution is trivial: $f_Y(y) = \delta(y - b)$*, a Dirac delta at* b*.*

Example 3.2-8 in [1]

Consider the vertical coordinate of a spinner with uniform random angle

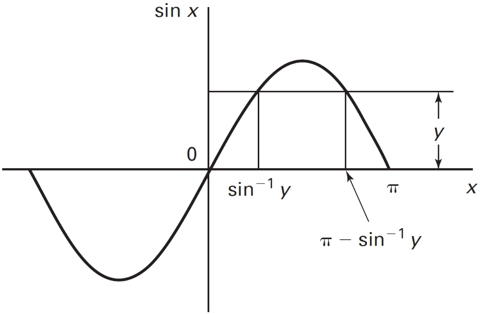
$$g(X) = \sin(X) \quad \text{(sine map)} \tag{54}$$

$$f_X(x) = \begin{cases} \dfrac{1}{2\pi} & \text{if } -\pi \leq X \leq \pi \\ 0 & \text{else} \end{cases} \qquad \text{(uniform distribution)} \tag{55}$$

Case 1: $0 \leq y < 1$

Step 1: Find the point set (this time it's trickier)

$$\{Y \leq y\} = \{\sin(X) \leq y\} \tag{56}$$
$$= \left\{-\pi < X \leq \sin^{-1}(y)\right\} \cup \left\{\pi - \sin^{-1}(y) < X \leq \pi\right\} \tag{57}$$
$$= \{X \in C_y\} \tag{58}$$



Step 2: Find the cdf

$$
\begin{aligned}
F_Y(y) &= P[Y \leq y] && (59)\\
&= P\left[\left\{-\pi < X \leq \sin^{-1}(y)\right\} \cup \left\{\pi - \sin^{-1}(y) < X \leq \pi\right\}\right] && (60)\\
&= P\left[-\pi < X \leq \sin^{-1}(y)\right] + P[\pi - \sin^{-1}(y) < X \leq \pi] && (61)\\
&= \left[F_X(\sin^{-1}(y)) - F_X(-\pi)\right] + \left[F_X(\pi) - F_X(\pi - \sin^{-1}(y))\right] && (62)
\end{aligned}
$$

Step 3: Differentiate cdf to get pdf

$$
\begin{aligned}
f_Y(y) &= \frac{d}{dy} F_Y(y) && (63)\\
&= f_X\left(\pi - \sin^{-1} y\right) \frac{1}{\sqrt{1-y^2}} + f_X\left(\sin^{-1} y\right) \frac{1}{\sqrt{1-y^2}} && (64)\\
&= \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-y^2}} \quad \text{for } 0 \leq y < 1 && (65)
\end{aligned}
$$

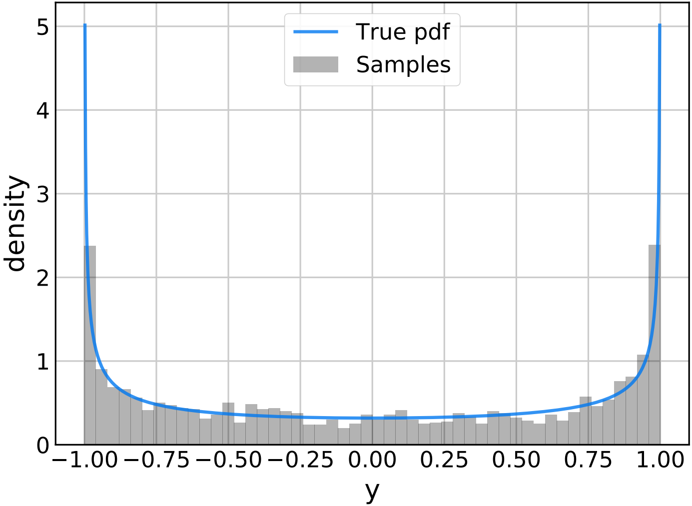
Optional Exercise: Work out Case 2: $-1 < y \leq 0$
Hint: You should find the pdf is the same as for $0 \leq y < 1$

Optional Exercise: Work out Case 3: $|y| >= 1$
Hint: You should find the cdf is constant with respect to y
(either $F_Y(y) = 0$ *or* $F_Y(y) = 1$*) and therefore the pdf is zero.*

Therefore, the complete solution is

$$\boxed{f_Y(y) = \begin{cases} \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-y^2}} & \text{if } |y| < 1 \\ 0 & \text{else} \end{cases}} \tag{66}$$

We can check our solution against a histogram of empirical samples
- see `function_of_rv.py`



Can we go directly from pdf of X to pdf of $Y = g(X)$
(without finding intermediate cdf)?

"Direct" procedure:

1. Find the root functions $x_i = x_i(y)$ that satisfy $y - g(x_i) = 0$ for any fixed y
2. Compute derivative $g'(x)$
3. Evaluate $|g'(x_i)|$ check $|g'(x_i)| \neq 0$
4. Compute the pdf directly as

$$f_Y(y) = \sum_i \frac{f_X(x_i)}{|g'(x_i)|} \tag{67}$$

Note: Throughout keep in mind that $x_i = x_i(y)$ are functions!

Example 3.2-9 in [1]

Consider again the problem

$$g(X) = \sin(X) \quad \text{(sine map)} \tag{68}$$

$$f_X(x) = \begin{cases} \frac{1}{2\pi} & \text{if } -\pi \leq X \leq \pi \\ 0 & \text{else} \end{cases} \quad \text{(uniform distribution)} \tag{69}$$

<u>Case 1: $0 \leq y < 1$</u>

Step 1:
For any $0 \leq y < 1$ we have the roots of

$$y - g(x) = y - \sin(x) = 0 \tag{70}$$

are

$$x_1 = \sin^{-1}(y) \quad \text{and} \quad x_2 = \pi - \sin^{-1}(y) \tag{71}$$

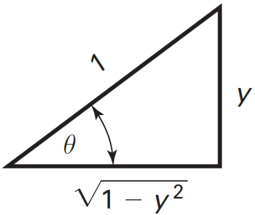
Step 2:
We have the derivative

$$\frac{dg}{dx} = \cos(x) \tag{72}$$

Step 3:
Evaluated at the roots, the derivative is

$$\left.\frac{dg}{dx}\right|_{x_1} = \cos(\sin^{-1}(y)), \qquad \left.\frac{dg}{dx}\right|_{x_2} = -\cos(\sin^{-1}(y)) \tag{73}$$

When you see the **composition of trig and inverse trig**, there is usually a nice simplification to make - use triangle diagram to help



$$\sin(\theta) = \frac{y}{1} \tag{74}$$

$$\theta = \sin^{-1}(y) \tag{75}$$

$$\cos(\theta) = \frac{\sqrt{1-y^2}}{1} \tag{76}$$

$$\cos(\sin^{-1}(y)) = \sqrt{1-y^2} \tag{77}$$

We have the absolute values

$$\left|\frac{dg}{dx}\right|_{x_1} = \left|\frac{dg}{dx}\right|_{x_2} = \sqrt{1-y^2} \neq 0 \text{ for } 0 \leq y < 1 \tag{78}$$

Step 4:
Compute the pdf

$$f_Y(y) = \sum_i \frac{f_X(x_i)}{|g'(x_i)|} \tag{79}$$

$$= \frac{\frac{1}{2\pi}}{\sqrt{1-y^2}} + \frac{\frac{1}{2\pi}}{\sqrt{1-y^2}} \tag{80}$$

$$= \frac{1}{\pi}\sqrt{1-y^2} \quad \text{for } 0 \le y < 1 \tag{81}$$

which is the same result as we got using the "indirect" method.

Optional Exercise: Repeat the procedure for Case 2: $-1 < y \le 0$

Optional Exercise: Repeat the procedure for Case 3: $|y| \ge 1$

Core problem:
What is the distribution of a function of a random variable?

Math:
Given $f_{XY}(x, y)$ and $Z = g(X, Y)$, what is $f_Z(z)$?

"Indirect" procedure:

1. Find the point set C_z such that $\{Z \leq z\} = \{(X, Y) \in C_z\}$
2. Find the cdf of Z as

$$F_Z(z) = \iint_{(x,y)\in C_z} f_{XY}(x, y)dxdy \tag{82}$$

3. Find the pdf of Z as

$$f_Z(z) = \frac{d}{dz} F_Z(z) \tag{83}$$

Product of two RVs

Optional Exercise: Find $f_Z(z)$ where $Z = XY$
Hint: See Example 3.3-1 in [1]
Solution:

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f_{XY}(z/y, y) dy \tag{84}$$

Optional Exercise: Find $f_Z(z)$ where $Z = X + Y$ Eqs. (3.3-13), (3.3-14) in [1]

Solution:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(z-y, y) dy \tag{85}$$

If X and Y are independent

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy \tag{86}$$

which is a **convolution integral**

Evaluate by reversing one function and sliding it

See Examples 3.3-4, 3.3-5, 3.3-6, 3.3-7, 3.3-8 in [1]

Homework P2-4: Find $f_Z(z)$ where $Z = \max(X, Y)$ and X, Y are independent.
Hint: See Example 3.3-2 in [1]

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

Expectation and Moments

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Expectation and moments

Expectation

The **expectation** or **mean** of a random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{1}$$

The expectation of a function of a random variable $g(X)$ is

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{2}$$

If the RV is discrete, these integrals become simple sums:

$$\mathbb{E}[X] = \sum_i x_i P_X(x_i) \tag{3}$$

$$\mathbb{E}[g(X)] = \sum_i g(x_i) P_X(x_i) \tag{4}$$

Linearity of expectation

Expectation is a **linear operator** - follows from linearity of integration

$$
\begin{aligned}
&\mathbb{E}[X+Y] && (5)\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f_{XY}(x,y)dxdy && (6)\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xf_{XY}(x,y)dxdy + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yf_{XY}(x,y)dxdy && (7)\\
&= \int_{-\infty}^{+\infty}x\left(\int_{-\infty}^{+\infty}f_{XY}(x,y)dy\right)dx + \int_{-\infty}^{+\infty}y\left(\int_{-\infty}^{+\infty}f_{XY}(x,y)dx\right)dy && (8)\\
&= \int_{-\infty}^{+\infty}xf_X(x)dx + \int_{-\infty}^{+\infty}yf_Y(x)dy && (9)\\
&= \mathbb{E}[X] + \mathbb{E}[Y] && (10)
\end{aligned}
$$

Use induction to conclude the linearity property

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N}\mathbb{E}\left[X_i\right] \tag{11}$$

Recall the Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$.

Let's show the mean is μ using the change of variable $z = \frac{x-\mu}{\sigma}$

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x) dx && (12)\\
&= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx && (13)\\
&= \int_{-\infty}^{\infty} (\sigma z + \mu) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz && (14)\\
&= \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} z \cdot \exp\left(-\frac{1}{2} z^2\right) dz}_{=0 \text{ because integrand odd}} + \mu \underbrace{\left[\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz\right]}_{=1 \text{ because } P[Z \leq \infty] = 1} && (15)\\
&= \mu && (16)
\end{aligned}$$

Conditional expectation

Conditional expectation

The **conditional expectation** of random variable Y given event B has occurred is

$$\mathbb{E}[Y|B] = \int_{-\infty}^{\infty} y f_{Y|B}(y|B) dy \tag{17}$$

The **conditional expectation** of random variable Y conditioned on random variable X is

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \tag{18}$$

We have a **law of total expectation** (like law of total probability)

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x] f_X(x) dx \tag{19}$$

Moments are expectations of monomials of (shifted and scaled) RVs

Moments

The k^{th} **(raw) moment** of X is

$$m_k = \mathbb{E}[X^k] \tag{20}$$

The k^{th} **central moment** of X is

$$c_k = \mathbb{E}[(X - \mathbb{E}[X])^k] \tag{21}$$

The k^{th} **standardized moment** of X is

$$s_k = \frac{\mathbb{E}[(X - \mathbb{E}[X])^k]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^{k/2}} = \frac{c_k}{c_2^{k/2}} \tag{22}$$

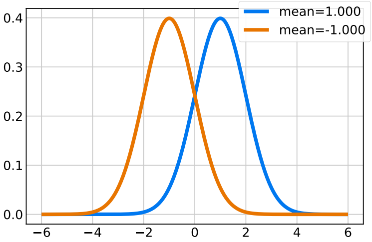
Moments

Moments summarize different aspects of the **shape** of a distribution

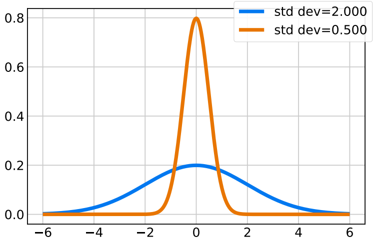
Name	Definition	Intuition
Mean	$\mu = m_1$	Location or center
Variance	$\sigma^2 = c_2$	Dispersion or spread
Std deviation	$\sigma = \sqrt{\sigma^2}$	Dispersion or spread
Skewness	s_3	Asymmetry or tilt
Kurtosis	s_4	Heaviness of tails

See `moments.py`

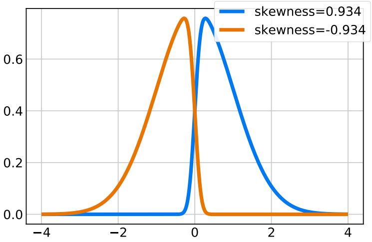
Comparison of pdfs with different moments



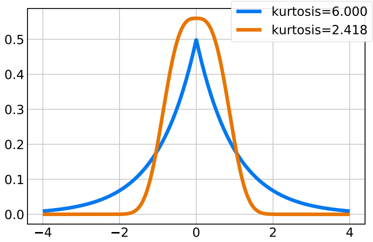
(a) Mean



(b) Standard deviation



(c) Skewness



(d) Kurtosis

Moments

We can convert between raw and central moments

Example: Second moment

$$
\begin{aligned}
c_2 &= \mathbb{E}[(X - \mathbb{E}[X])^2] && (23)\\
&= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] && (24)\\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && (\text{linearity of } \mathbb{E}[\cdot])\\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && (25)\\
&= m_2 - m_1^2 && (26)
\end{aligned}
$$

This relation generalizes to higher-order moments as

$$
c_k = \sum_{i=0}^{k} \binom{k}{i} (-1)^i \mu^i m_{k-i} \qquad (27)
$$

Homework P3-1:
Verify the expression for the variance of a Gaussian.
Hint: See Example 4.1-7 in [1]

Optional Exercise:
Find expressions for all moments of a Gaussian.
Hint: See e.g. https://arxiv.org/abs/1209.4340

Basic probability bounds

Often we want to bound the probability of certain events or random variables without having to specify/compute their distribution

c.f. the first several pages of Wainwright's book [2]

Tail bounds

Markov inequality

Given a non-negative random variable X with finite mean, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all } t > 0 \tag{28}$$

"X is probably small when its mean is small"

The most basic tail bound.

Basis for several "classical" concentration inequalities.

Chebyshev inequality

Given a random variable X with finite mean μ and variance σ^2, we have

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2} \quad \text{for all } t > 0 \tag{29}$$

"X is probably close to its mean whenever its variance is small"

The most basic concentration inequality.

Proof: Follows by applying Markov inequality to the non-negative random variable $(X - \mu)^2$.

Moment bound

Given a non-negative random variable X with finite moments up to order k, we have

$$\mathbb{P}[|X-\mu| \geq t] \leq \frac{\mathbb{E}\left[|X-\mu|^k\right]}{t^k} \quad \text{for all } t > 0 \tag{30}$$

Proof: Follows by applying Markov inequality to the random variable $|X-\mu|^k$

Chernoff bound

Given a non-negative random variable X with a moment generating function in a neighborhood of zero, we have

$$\mathbb{P}[X \geq 0] \leq \inf_{\theta > 0} \mathbb{E}\left[e^{\theta X}\right] \tag{31}$$

Proof: Follows by applying Markov inequality to the random variable $e^{\theta(X-\mu)}$ and optimizing over θ.

The moment bound with an optimal choice of k is never worse than the Chernoff bound.

Nonetheless, the Chernoff bound is most widely used in practice, possibly due to the ease of manipulating moment generating functions.

Homework P3-2:
Compare the Markov inequality bound with the exact tail probability from the exponential cdf with parameter $\lambda = 1$; compute the probability bounds at the level $t = 2$. How bad is the Markov bound compared with the exact tail probability?
Hint: The mean of an exponential random variable is $\mu = 1/\lambda$.

Homework P3-3:
Compare the Chebyshev inequality bound with the exact tail bound from the standard normal cdf; compute the probability bounds at the level $t = 2$. How bad is the Chebyshev bound compared with the exact concentration probability?
Hint: The standard normal cdf does not have a closed-form expression, so either use the `cdf()` method of `scipy.stats.norm` or a table of the standard normal cdf to get the exact value. In case you run into issues, $\Phi(2) = 1 - \Phi(-2) = 0.9772$.

Joint moments summarize different aspects of the shape of a joint distribution

Joint moments

The **ijth (raw) joint moment** of random variables X and Y is

$$m_{ij} = \mathbb{E}[X^i Y^j] \tag{32}$$

The **ijth central joint moment** of random variables X and Y is

$$c_{ij} = \mathbb{E}[(X - \mathbb{E}[X])^i (Y - \mathbb{E}[Y])^j] \tag{33}$$

Joint moments

Some joint moments have special, confusing names

The **correlation** is

$$m_{11} = \mathbb{E}[XY] \tag{34}$$

The **covariance** is

$$c_{11} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{35}$$

The **correlation coefficient** is

$$\rho = \frac{c_{11}}{\sqrt{c_{02}c_{20}}} \tag{36}$$

Homework: Joint moments

Homework P3-4:
Prove the relation

$$m_{11} = c_{11} + \mathbb{E}[X]\mathbb{E}[Y]$$

Hint: It is similar to the earlier second moment relation $m_2 = c_2 + m_1^2$

Homework P3-5:
When are the correlation and covariance equal?
Hint: Use the relation $m_{11} = c_{11} + \mathbb{E}[X]\mathbb{E}[Y]$ *you just proved.*

Homework P3-6:
Prove that $\rho \in [-1, 1]$
Hint: See Ch. 4.3 of [1]

Uncorrelated and orthogonal random variables

Uncorrelated random variables

Two random variables are **uncorrelated** if their **covariance** is zero.

Orthogonal random variables

Two random variables are **orthogonal** if their **correlation** is zero.

- Yes I know the terminology is confusing :/

Homework P3-7:
Prove that if X and Y are uncorrelated, then $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$
i.e. "the variance of the sum is the sum of the variances."
Hint: Use linearity of expectation.

Homework P3-8:
Prove that if X and Y are independent, then they are uncorrelated.
Remark: The converse does not hold unless X and Y are both Gaussian.

Homework P3-9:
Under what condition(s) can a pair of uncorrelated random variables be orthogonal?
Hint: This is a special case of one of the earlier exercises.

Random vectors

Random vector

A **random vector** is a vector of random variables.

The **cdf** of a random vector is defined as

$$F_X(x) = \mathbb{P}[X_1 \leq x_1 \text{ and } X_2 \leq x_2 \text{ and } \ldots X_n \leq x_n] \tag{37}$$

The **pdf** is defined as

$$f_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \partial x_2 \cdots \partial x_n} \tag{38}$$

Similar definitions for joint, marginal, and conditional distributions

- See Ch. 5.1 of [1]

The **expectation** of a random vector X is the vector μ_X with entries

$$[\mu_X]_i = \mathbb{E}[X]_i = \int_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i \tag{39}$$

where $f_{X_i}(x_i)$ is the ith marginal pdf.

Moments are defined similarly as with random variables.

(Auto)-covariance matrix of X

$$K_X = \mathbb{E}[(X - \mu_X)(X - \mu_X)^\intercal] \tag{40}$$

(Cross)-covariance matrix between X and Y

$$C_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\intercal] \tag{41}$$

We can gather these up into the block covariance matrix

$$D_{XY} = \begin{bmatrix} K_X & C_{XY} \\ C_{XY}^\intercal & K_Y \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix} \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix}^\intercal\right] \tag{42}$$

(Auto)-correlation matrix of X

$$R_X = \mathbb{E}[XX^\intercal] \succeq 0 \tag{43}$$

(Cross)-correlation matrix between X and Y

$$S_{XY} = \mathbb{E}[XY^\intercal] \tag{44}$$

We can gather these up into the block correlation matrix

$$B_{XY} = \begin{bmatrix} R_X & S_{XY} \\ S_{XY}^\intercal & R_Y \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} X \\ Y \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix}^\intercal\right] \tag{45}$$

Homework 3-10:
Prove the identity between covariance and correlation matrices

$$R = K + \mu\mu^\intercal \tag{46}$$

Hint: Use linearity of expectation.

Homework 3-11:
Write an expression for D in terms of B, μ_X, μ_Y.
Hint: It follows immediately from $R = K + \mu\mu^\intercal$ by stacking X and Y.

Homework 3-12:
Prove that $R \succeq K \succeq 0$ and $B \succeq D \succeq 0$ where $A \succeq B$ means $A - B$ is symmetric positive semidefinite.
Hint: It follows by the above relations and the property of outer product matrices $AA^\intercal \succeq 0$ for any matrix A, and taking $A = \mu$.

Uncorrelated and orthogonal random vectors

A random vector X is **uncorrelated** with itself if K is diagonal.

A random vector X is **orthogonal** with itself if R is diagonal.

Two random vectors X and Y are **uncorrelated** if $C = 0$.

Two random vectors X and Y are **orthogonal** if $S = 0$.

Optional Exercise:
Think about how these expressions can be summarized in terms of the block matrices C and D.

Optional Exercise:
Under what condition(s) can a pair of uncorrelated random vectors be orthogonal?
Hint: You already solved this in the scalar case.

Whitening transformation

Sometimes we need to get a standardized version of a random variable

In the scalar case we used the standardizing transform

$$Z = \frac{X - \mu}{\sigma} \tag{47}$$

- Subtract out the mean and normalize by the standard deviation, so Z has zero mean and variance one
- Need to assume $\sigma > 0$ for non-degeneracy

The **whitening transformation** is the multivariate generalization of the scalar standardizing transform

- Based on the eigen-decomposition of the covariance matrix

The **whitening transformation** is

$$Z = \Lambda_X^{-1/2} U_X^{\mathsf{T}} (X - \mu) \tag{48}$$

- Subtract the mean out and normalize, so Z has zero mean and identity auto-covariance
- Λ_X is a diagonal matrix whose entries are the n eigenvalues of K_X
  - The eigenvalues λ_i are real numbers since K_X is symmetric
  - Need to assume $\lambda_i > 0$ for $i = 1, \ldots, n$ for non-degeneracy
    - Equivalent to assuming K_X full rank
  - $\Lambda_X^{-1/2}$ is diagonal with entries $\lambda_i^{-1/2}$
- U_X is an orthogonal matrix whose columns are n eigenvectors of K_X

Sometimes we need to get a a random vector Y with nonzero mean μ_Y and non-identity covariance K_Y from a white random vector

- Inverse operation of the whitening transformation

The **coloring transformation** is

$$Y = U_Y \Lambda_Y^{1/2} X + \mu \tag{49}$$

- Λ_Y is a diagonal matrix whose entries are the n eigenvalues of K_Y
- U_Y is an orthogonal matrix whose columns are n eigenvectors of K_Y

The n-**dimensional multivariate Gaussian** pdf is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(K)}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal K^{-1}(x-\mu)\right] \tag{50}$$

- Mean is $\mu \in \mathbb{R}^n$
- Covariance is $K \in \mathbb{R}_+^{n \times n}$

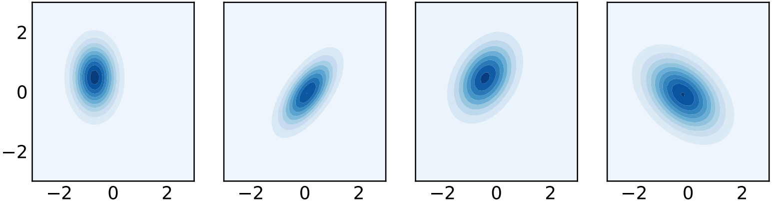


Figure 2: Various multivariate Gaussian pdfs for $n = 2$.
See `multivariate_gaussian.py`

Properties of multivariate Gaussians

Gaussians are extremely special distributions with nice properties

- Marginals of a Gaussian are Gaussian
- Gaussians conditioned on Gaussians are Gaussian
- Any affine transformation of a Gaussian is Gaussian
- All pertinent information about a Gaussian is encoded in the mean and covariance
- Sums of random vectors tend towards a Gaussian (central limit theorem, coming up)

Homework 3-13:
What is the pdf of a white (zero mean and identity covariance) multivariate Gaussian random vector X? Can it be expressed in terms of the marginal densities of each component of X? If so, write the expression. Are the components of X statistically independent?

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

[2] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.
https://people.eecs.berkeley.edu/~wainwrig/BibPapers/Wai19.pdf.

Parameter Estimation

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

1. Parameter estimation
2. Laws of large numbers
3. Central limit theorem

Parameter estimation

Parameter estimation

In many applications:

- Distribution of a random variable X is unknown or too complicated to compute
- Only need some parameter θ that characterizes the distribution

Goal: Obtain a good approximation of parameter θ based only on observations of X.

Estimator

An **estimator** $\hat{\Theta}$ is a function of the data $\{X_i\}$ that approximates θ, but is not an explicit function of θ.

Judging parameter estimators

How do we judge the quality of an estimator?

Consistency

An estimator $\hat{\Theta}_n$ computed from n samples is **consistent** if

$$\lim_{n \to \infty} P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] = 0 \tag{1}$$

for any positive tolerance $\varepsilon > 0$.

Consistency means "we can guarantee arbitrarily accurate estimates if we use an arbitrarily large amount of data"

Judging parameter estimators

What we really want:

Confidence bound

An estimator $\hat{\Theta}_n$ is **ε-accurate with $1-\delta$ confidence** if

$$P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] \leq \delta \tag{2}$$

- This is like soft consistency w/ finite data
- Consistency allows us to take ε and δ as small as we like (so long as we can pay for it with infinite data $n \to \infty$)
- Quantifying n
  - Can be done exactly in certain special cases
    - e.g. estimating the mean of a Gaussian
  - Can be done conservatively using concentration inequalities in more general cases
    - e.g. estimating the mean of any distribution w/ finite variance

Judging parameter estimators

Confidence interval

Consider an estimator $\hat{\Theta}_n$. Fix the number of samples n and fix a failure probability δ. The $1-\delta$ **confidence interval** is the smallest accuracy tolerance ε such that

$$P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] \leq \delta \tag{3}$$

i.e. the estimator $\hat{\Theta}_n$ is ε-accurate with $1-\delta$ confidence.

Basically the same as the confidence criterion where we fixed ε and sought n, but here we fix n and seek ε

Judging parameter estimators

Many classical results use two proxies for the ε-δ criterion:

- Bias
  - "systematic errors"
  - "location"
- Variance
  - "random errors"
  - "spread"

Bias

The **bias** of an estimator $\hat{\Theta}$ is

$$|\mathbb{E}[\hat{\Theta}] - \theta|. \tag{4}$$

The estimator is **unbiased** if

$$\mathbb{E}[\hat{\Theta}] = \theta. \tag{5}$$

Variance

The **variance** of an estimator $\hat{\Theta}$ is

$$\mathbb{E}[(\hat{\Theta} - \theta)^2]. \tag{6}$$

The estimator is **minimum variance** if

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\ \mathbb{E}[(\Theta - \theta)^2]. \tag{7}$$

Sometimes bias can be eliminated without affecting the variance

- We will see an example of such a correction

Sometimes bias can only be reduced at the expense of higher variance

- In machine learning this is a well-studied phenomenon called the **bias-variance tradeoff**

Sample average estimator

Sample average estimator of a RV

The **sample average estimator** of a random variable X given N observations $\{X_i\}_{i=1}^N$ is

$$\hat{\mu}_X(n) := \frac{1}{N} \sum_{i=1}^{N} X_i$$

Sample average estimator of a function of a RV

The **sample average estimator** of a function g of a random variable X given N observations $\{X_i\}_{i=1}^N$ is

$$\hat{\mu}_{g(X)}(n) := \frac{1}{N} \sum_{i=1}^{N} g(X_i)$$

Properties of the sample average: bias

It's easy to show that the sample average is **unbiased**:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\mu}_X(n)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && \text{(def. of } \hat{\mu}_X(n)\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] && \text{(linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_X && \text{(def. of } \mu_X\text{)} \\
&= \frac{1}{n} \cdot n \cdot \mu_X && (8) \\
&= \mu_X && (9)
\end{aligned}
$$

The **variance** of the sample average is not much harder to find:

$$
\begin{aligned}
\sigma^2_{\hat{\mu}}(n) &:= \mathbb{E}\left[\left(\hat{\mu}_X(n) - \mathbb{E}\left[\hat{\mu}_X(n)\right]\right)^2\right] && \text{(def. of } \sigma^2_{\hat{\mu}}(n)\text{)} \\
&= \mathbb{E}\left[\left(\hat{\mu}_X(n) - \mu_X\right)^2\right] && \text{(since } \hat{\mu} \text{ unbiased)} \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)\right)^2\right] && \text{(def. of } \hat{\mu}\text{)} \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \mu_X)^2\right] + \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(X_i - \mu_X)(X_j - \mu_X)\right] \\
& && \text{(expand squared sum)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu_X)^2\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n}\mathbb{E}\left[(X_i - \mu_X)(X_j - \mu_X)\right] \\
& && \text{(linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_X^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n} 0 && \text{(def. of } \sigma_X^2\text{, uncorrelation of } X_i\text{)} \\
&= \sigma_X^2/n && (10)
\end{aligned}
$$

Properties of the sample average: confidence

We can get a **confidence bound** by using the Chebyshev inequality:

$$P\left[|\hat{\mu}_X(n) - \mu_X| \geq \varepsilon\right] \leq \frac{\sigma^2_{\hat{\mu}}(n)}{\varepsilon^2} = \frac{1}{n} \cdot \frac{\sigma_X^2}{\varepsilon^2} \tag{11}$$

Taking $n \to \infty$ reveals that the **sample average is consistent**:

$$\lim_{n\to\infty} P\left[|\hat{\mu}_X(n) - \mu_X| \geq \varepsilon\right] = \lim_{n\to\infty} \frac{1}{n} \cdot \frac{\sigma_X^2}{\varepsilon^2} = 0 \tag{12}$$

Remark: If we knew the form of the distribution e.g. Gaussian we could get an exact confidence bound using the standard normal CDF.

Remark: This confidence bound involves the true variance σ_X^2, which is typically unknown. If X is Gaussian and σ_X^2 is replaced by a sample variance estimate, an exact confidence bound can still be obtained using the **student T-distribution** CDF - see Ch. 6.3 of [1].

Sample variance

So far we estimated the mean - what about estimating the variance?

If we **knew the true mean μ** we could create the variance estimator

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 \tag{13}$$

But of course we **don't know the true mean μ!**

Natural idea: just use the sample mean in place of the true mean:

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \hat{\mu})^2 \tag{14}$$

But there is an issue with this...

Homework P4-1

Compute the expectation of the sample variance estimator

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \hat{\mu}_X(n))^2 \tag{15}$$

where

$$\hat{\mu}_X(n) = \frac{1}{n}\sum_{i=0}^{n} X_i \tag{16}$$

1. Is this sample variance estimator $\hat{\sigma}_X^2(n)$ biased?
2. If so, how much is the bias?
3. How does the bias change with the number of samples n?
4. What correction needs to be made to $\hat{\sigma}_X^2(n)$ in order to make the estimator unbiased?

Maximum likelihood estimation

Maximum likelihood estimation provides a principled way to design estimators based on optimization.

Likelihood

The **likelihood** function $L(\theta)$ of the random variables $\{X_i\}_{i=1}^n$ for outcome $\{x_i\}_{i=1}^n$ under parameter θ is the parametric joint pdf

$$L(\theta) = f_{\{X_i\}_{i=1}^n}(\{x_i\}_{i=1}^n; \theta). \tag{17}$$

As a special case, if $\{X_i\}_{i=1}^n$ are i.i.d. random variables then

$$L(\theta) = \prod_{i=1}^n f_X(x_i; \theta) \tag{18}$$

Maximum likelihood estimation

Maximum likelihood estimate

The **maximum likelihood estimate** for outcome $\{x_i\}_{i=1}^n$ is the parameter $\theta^*(\{x_i\}_{i=1}^n)$ that maximizes the likelihood, i.e.

$$\theta^*(\{x_i\}_{i=1}^n) = \operatorname*{argmax}_{\theta} L(\theta) \tag{19}$$

The **maximum likelihood estimator** is the random variable

$$\hat{\theta} = \theta^*(\{X_i\}_{i=1}^n) \tag{20}$$

MLE: mean of a Gaussian

We start by assuming the *form* of the distribution is Gaussian with variance σ^2. We are estimating the mean, so the parameter is $\theta = \mu$

The likelihood is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{21}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^{n} -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{22}$$

Since the log function is monotonic increasing, the argmax of $L(\mu)$ is the same as the argmax of $\log L(\mu)$. The log is easier to work with.

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{23}$$

To maximize the log likelihood we find the stationary point

$$0 = \left.\frac{\partial \log L(\mu)}{\partial \mu}\right|_{\mu^*} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu^*) \tag{24}$$

which implies the MLE is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{25}$$

which happens to be the sample mean.

Homework: MLE for variance of a Gaussian

Homework P4-2: Derive the expression for the maximum likelihood estimator of the mean and variance of a Gaussian. Is the MLE variance biased?

Hint: Use the log-likelihood

$$\log L(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (26)$$

Suppose we wish to estimate a vector parameter which is exposed through the **linear observation model**

$$Y = H\theta + N \tag{27}$$

- Y is an **observation vector**
- H is a known constant **observation matrix**
- θ is an unknown constant **parameter vector**
- N is a **random observation noise vector**

The observation Y is directly measured, but the noise N is not.

Least-squares estimation

Define the **residual**

$$E = Y - H\theta \tag{28}$$

which measures the error between the observation and its expected value.

A natural idea is to choose a parameter estimate that minimizes an objective function $v(\theta)$ which increases with the size of the residual.

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \; v(\theta) \tag{29}$$

In particular, choose $v(\theta)$ as the squared norm of the residual:

$$v(\theta) = \|E\|^2 = (Y - H\theta)^\intercal (Y - H\theta) \tag{30}$$

Next we need some basic facts from optimization and matrix calculus.

Fact 1: The minimum of a continuous function $f(\theta)$ can only occur at a **stationary point** where the gradient vanishes

$$0 = \frac{\partial f(\theta)}{\partial \theta} \tag{31}$$

Fact 2: The derivative of an affine form is

$$\frac{d}{dx} a^\intercal x = a \tag{32}$$

and the derivative of a quadratic form is

$$\frac{d}{dx} x^\intercal Q x = 2Qx \tag{33}$$

Since $v(\theta)$ is a quadratic form, we can compute the minimizer in closed-form by finding the **stationary point** where the gradient of the objective vanishes:

$$0 = \left. \frac{\partial v(\theta)}{\partial \theta} \right|_{\hat{\theta}} = 2(H^\intercal H)\hat{\theta} - 2H^\intercal Y \tag{34}$$

Rearranging yields the so-called **normal equation**

$$(H^\intercal H)\hat{\theta} = H^\intercal Y \tag{35}$$

If $H^\intercal H$ is invertible, we obtain the **least-squares estimate (LSE)**

$$\hat{\theta} = (H^\intercal H)^{-1} H^\intercal Y \tag{36}$$

Remark: If N is a white Gaussian noise, i.e. $N \sim \mathcal{N}(0, I)$, then it can be shown that the LSE is an unbiased, minimum variance, and maximum likelihood estimator.

Homework P4-3: We are given the following data:

$$\begin{bmatrix} 6.2 \\ 7.8 \\ 2.2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix} \theta + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{37}$$

where n_i are random variables. Find a least-squares estimate for θ.

Asymptotics

Asymptotics

In this section we see major results from classical statistics

Claims are **asymptotic**; they only hold as the amount of data $\rightarrow \infty$

Claims are all about **convergence** of some kind

Contrast with finite-sample results c.f. [2]

Weak law of large numbers

Let X_i be an infinite sequence of i.i.d. random variables with a finite, common true mean μ and variance σ^2. Consider the sample mean

$$\hat{\mu}(n) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{38}$$

Then for any fixed positive tolerance $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}\left[|\hat{\mu}(n) - \mu| < \varepsilon \right] = 1 \tag{39}$$

i.e. the sample mean **converges in probability** to the true mean.

Proof: We already proved that the sample mean is consistent, which is the same thing as the WLLN.

Strong law of large numbers (SLLN)

Strong law of large numbers

Let X_i be an infinite sequence of i.i.d. random variables with a finite, common true mean μ and variance σ^2. Consider the sample mean

$$\hat{\mu}(n) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{40}$$

Then we have

$$\mathbb{P}\left[\lim_{n \to \infty} \hat{\mu}(n) = \mu\right] = 1 \tag{41}$$

i.e. the sample mean **converges almost surely** to the true mean.

Proof: More involved than the WLLN. Also SLLN implies WLLN.

Notice the difference between weak and strong laws:

1. WLLN: Sequence of success probabilities approaches one
2. SLLN: Sequence of sample means approaches the true mean

Central limit theorem

Let X_i be an infinite sequence of independent random variables with cdf's F_{X_i}, finite means μ_i and finite variances σ_i^2.

Define the variance sum s_n^2 and normalized random variable Z_n

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2, \quad Z_n = \sum_{i=1}^{n} (X_i - \mu_i)/s_n \tag{42}$$

Suppose there exists $\varepsilon > 0$ and for all n sufficiently large that

$$\sigma_i < \varepsilon s_n, \quad i = 1, \dots, n \tag{43}$$

Then

$$\lim_{n \to \infty} F_{Z_n}(z) = \Phi(z) \tag{44}$$

i.e. Z_n **converges in distribution** to a standard normal.

Homework: Central limit theorem

Homework P4-4: Let $\{X_i\}_{i=1}^n$ be a sequence of n i.i.d. random variables. Compute the approximate probability

$$\mathbb{P}[a \leq S \leq b] \tag{45}$$

of the sum

$$S(n) = \sum_{i=1}^{n} X_i \tag{46}$$

using the central limit theorem.
For concreteness, assume the X_i are uniform random variables on the unit interval $[0, 1]$, $n = 100$, $a = 45$, and $b = 52.5$.

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

[2] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.
https://people.eecs.berkeley.edu/~wainwrig/BibPapers/Wai19.pdf.

Information Theory

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Information theory

Information theory

Information theory concerns quantifying the amount of information present in signals

- Originally developed for sending and receiving messages over communication channels
- Deals primarily with discrete random variables

Applications

- Machine learning e.g. classify images
- Reinforcement learning e.g. teach robots how to balance

c.f. Ch. 1-3 of Mackay's "Information Theory, Inference, and Learning Algorithms" [1]

c.f. Ch. 3 of Goodfellow's "Deep Learning" [2]

Information

Intuitively, we want a quantity that measures

- The amount of information communicated by an outcome
- How surprising an outcome is

Our definition of "information" or "surprise" should satisfy three axioms:

1. Certain events yield zero information
   - They always happen, so they are not surprising
2. Less probable events yield more information
   - They happen less, so they are more surprising
3. The total information of independent events is the sum of the information of each individual event
   - Their chances of happening are unrelated, so knowing one outcome has no effect on how surprising the other outcome is

Information

The **(Shannon) information** of measuring random variable X with pmf P_X as outcome x is the quantity

$$I_X(x) = -\log_b(P_X(x)) \tag{1}$$

The log base b is an arbitrary choice which has the effect of fixing the units of information. Common choices:

- $b = 2$, "bits"
- $b = e$, "nats"
- $b = 10$, "dits"

Information is a **description of a distribution** like the pmf or cdf.

Sometimes the random variable $I(X) = I_X(X)$ is also called the information.

Entropy

Entropy

The **entropy** of random variable X is the expected information of X

$$
\begin{aligned}
H(X) &= \mathbb{E}_X[I(X)] && (2)\\
&= \sum_i P_X(x_i) I_X(x_i) && (3)\\
&= -\sum_i P(x_i)\log(P_X(x_i)) && (4)
\end{aligned}
$$

Entropy measures the amount of randomness in X.

Entropy is a **summary statistic** like the mean or variance.

Let X be a Bernoulli random variable with success probability p

Let's compute the entropy of X as a function of the probability p

$$
\begin{aligned}
H(X) &= -\sum_i P(x_i)\log(P_X(x_i)) && (5)\\
&= -p\log(p) - (1-p)\log(1-p) && (6)
\end{aligned}
$$

Example: Entropy of a coin flip

Exercise: Compute p which maximize and minimize entropy.

Solution:

- Max entropy when $p = 1/2$
  - Most random, heads and tails equally likely
- Min entropy when $p = 0$ or $p = 1$
  - Least random, heads or tails is certain

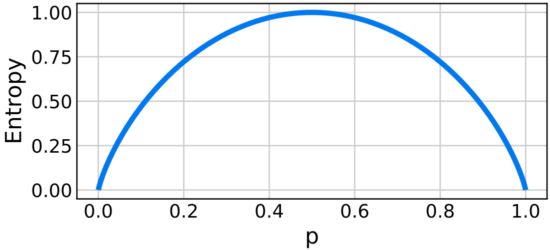


Figure 1: Entropy vs. parameter p for a Bernoulli random variable. See `entropy_bernoulli.py`

Joint entropy

The **joint entropy** between two random variables X and Y with joint pmf P_{XY} is

$$H(X,Y) = -\sum_i \sum_j P_{XY}(x_i, y_i) \log(P_{XY}(x_i, y_i)) \tag{7}$$

Joint entropy measures the amount of randomness in X and Y.

Special case:
X and Y independent if and only if the joint entropy is additive

$$H(X,Y) = H(X) + H(Y) \tag{8}$$

Mutual information

The **mutual information** between two random variables X and Y is

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{9}$$

$$= \sum_i \sum_j P_{XY}(x_i, y_i) \log\left(\frac{P_{XY}(x_i, y_i)}{P_X(x_i) P_Y(y_i)}\right) \tag{10}$$

Mutual information measures the average reduction in uncertainty about X that results from learning the value of Y.

Special case: $I(X,X) = H(X)$, so entropy can be thought of as "self mutual information"

Cross-entropy

The **cross-entropy** from random variable Y to X is the expected information of Y with respect to X

$$
\begin{aligned}
H(X||Y) &= \mathbb{E}_X[I(Y)] && (11)\\
&= \sum_i P_X(x_i) I_Y(x_i) && (12)\\
&= -\sum_i P_X(x_i) \log(P_Y(x_i)) && (13)
\end{aligned}
$$

Cross-entropy measures the amount of randomness in Y, under the fictitious assumption that Y has the distribution of X for the purpose of computing expectation.

Special case: $H(X||X) = H(X)$, so entropy can be thought of as "self cross-entropy"

Relative entropy / Kullback-Leibler divergence

The **relative entropy** or **Kullback–Leibler (KL) divergence** from random variable Y to X is

$$\mathcal{D}_{KL}(X||Y) = H(X||Y) - H(X) \tag{14}$$

$$= \sum_i P_X(x_i) \log \left(\frac{P_X(x_i)}{P_Y(x_i)} \right) \tag{15}$$

KL divergence measures the **difference between two distributions**.

KL divergence is **not a distance metric** because

1. It is not symmetric
2. The triangle inequality fails

See `kl_divergence.py`

Wasserstein metric

Wasserstein metric ("analytic" definition)

The **pth Wasserstein metric** between two pdfs f_X and f_Y is

$$W_p(f_X, f_Y) = \inf_{\pi \in \Pi(f_X, f_Y)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^p d\Pi(x, y) \right)^{1/p} \tag{16}$$

where $\Pi(f_X, f_Y)$ is the space of joint pdfs with marginals f_X and f_Y.

- There are ∞ different joint pdfs with marginals f_X and f_Y!
- The joint pdf π defines a **transport map** between f_X and f_Y.
  - π is a plan for moving the mass from f_X to f_Y (and vice versa)
  - Finding the infimal π is a special case of the general **optimal transport problem** c.f. [3]
  - In many cases, this ∞-dim infimization problem can be solved analytically or by reformulating as a finite-dim optimization program

Wasserstein metric

Wasserstein metric ("probabilistic" definition) [4]

The **pth Wasserstein metric** can be expressed as

$$W_p(f_X, f_Y) = \inf_{X \sim f_X,\ Y \sim f_Y} \left(\mathbb{E}_{XY}[\|X - Y\|^p] \right)^{1/p} \tag{17}$$

More facts:

- The two pdfs f_X and f_Y need not both be continuous or discrete
- $p = 1$ and $p = 2$ are the most common choices

Comparison with KL divergence:

- Like the KL divergence, the Wasserstein metric measures the **difference between two distributions**
- Unlike the KL divergence, the Wasserstein metric **is a valid distance metric**
    - Formal analysis using generic results for distance metrics is easier

Wasserstein metric

Special case: pth Wasserstein metric of two Dirac deltas
$f_X(x) = \delta(x-a)$ and $f_Y(y) = \delta(y-b)$

$$W_p(f_X, f_Y) = \|a-b\| \tag{18}$$

Special case: 2nd Wasserstein metric of two Gaussians
$f_X = \mathcal{N}(\mu_X, \Sigma_X)$ and $f_Y = \mathcal{N}(\mu_Y, \Sigma_Y)$

$$W_2(f_X, f_Y) = \sqrt{\left\|\mu_X - \mu_Y\right\|^2 + \mathbf{Tr}\left[\Sigma_X + \Sigma_Y - 2\left(\Sigma_Y^{\frac{1}{2}}\Sigma_X\Sigma_Y^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \tag{19}$$

Wasserstein metric

For the interested reader:

1. *"Statistical aspects of Wasserstein distances"* [4]
   - https://arxiv.org/abs/1806.05500
   - Contains a nice introduction on the Wasserstein metric.
2. *"Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations"* [5]
   - https://arxiv.org/abs/1505.05116
   - Quickly becoming a classic.
   - Details how to use the Wasserstein metric to solve optimization problems involving random problem data with unknown distribution while being robust to the worst-case distribution.

Bibliography I

[1] David JC MacKay and David JC Mac Kay.
Information theory, inference and learning algorithms.
Cambridge university press, 2003.
https://www.inference.org.uk/itila/.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
http://www.deeplearningbook.org.

[3] Cédric Villani.
Optimal transport: old and new, volume 338.
Springer, 2009.
https://cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf.

[4] Victor M Panaretos and Yoav Zemel.
Statistical aspects of wasserstein distances.
Annual review of statistics and its application, 6:405–431, 2019.
https://arxiv.org/pdf/1806.05500.pdf.

[5] Peyman Mohajerin Esfahani and Daniel Kuhn.
Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.
Mathematical Programming, 171(1):115–166, 2018.
https://arxiv.org/pdf/1505.05116.pdf.