STAT 391
Probability and Statistics for Computer Science
Lecture Notes, Version 3.4

Made available in .pdf form to the STAT 391 students in Spring 2019.

DO NOT DISTRIBUTE

April 6, 2022

Contents

Chapter 1

Introduction

1.1 Why should a computer scientist or engineer learn probability?

• Computers were designed from the beginning to be "thinking machines". They are billions times better than people at Boolean logic, arithmetic, remembering things, communicating with other computers. But they are worse than most people at understanding even simple images, speech. Why this difference?

One reason is that most real-life reasoning is "reasoning in uncertainty". Even if we didn't admit uncertainty exists (in fact it's something relative or even subjective!) we still must recognize this: if we want to reason by rules in real life, we must provide for exceptions. If there are many rules, and every rule has exceptions, any working reasoning system must specify how the exceptions to rule A interact with the exceptions to rule B and so on. This can become very complicated! And it can become too much work even for a computer. This is why there are so few expert systems deployed in practice.

• Computers are used to collect and store data. Computer scientists are required to help analyze these data, draw conclusions from them or make predictions.

Computer scientists are of course not the only ones who work on this. Scientists and statisticians have been analyzing data for much longer. Why should computer scientists be called to help? Because when the data sets are large, specific problems occur that need understanding of data bases, algorithms, etc. Did you ever encounter some examples?

In fact, in recent years, CS has made some really important contributions to

statistics, especially in the areas of machine learning and probabilistic reasoning (belief networks).

• Computer systems are not deterministic. Think of: delays in packet routing, communication through a network in general, load balancing on servers, memory allocation and garbage collection, cache misses.

• Computers interact with people, and people are non-deterministic as well. Can you give some examples? [Graphics, speech synthesis.]

1.2 Probability and statistics in computer science

Probability and statistics have started to be used in practically all areas of computer science:

- **Algorithms and data structures** – randomized algorithms and proofs using probability in deterministic algorithms. For example: randomized sort, some polynomial time primality testing algorithms, randomized rounding in integer programming.
- **Compilers** – modern compilers optimize code at run time, based on collecting data about the running time of different sections of the code
- **Cryptography**
- **Data bases** – to maximize speed of access data bases are indexed and structured taking into account the most frequent queries and their respective probability
- **Networking and communications** – computer networks behave non-deterministically from the point of view of the user. The probabilistic analysis of computer networks is in its beginnings.
- **Circuit design** – both testing the functionality of a circuit and testing that a given chip is working correctly involve probabilistic techniques
- **Computer engineering** – cache hits and misses, bus accesses, jumps in the code, interruptions are all modeled as random events from the point of view of the system designer.
- **Artificial intelligence** – probabilistic methods are present and play a central role in most areas of AI. Here are just a few examples: machine learning, machine vision, robotics, probabilistic reasoning, planning, natural language understanding, information retrieval.

- **Computer graphics** – machine learning techniques and their underlying statistical framework are starting to be used in graphics; also, making rendered scenes look "natural" is often done by injecting a certain amount of randomness (for example rendering of clouds, smoke, fields with grass and flowers, tree foliage)

1.3 Why is probability hard?

• Probability involves math.

• Even if you like math, you may hate probability. A science which deals with gambling, failures and noise. Who would ever like this? We computer scientists like clear, predictable things, like computers. We like the numbers 0 and 1 and these are the only ones we need. Or is it so?

• Probability attempts to describe uncertainty in a general way. This is not an easy task. Probability is both abstract and complex. "Make everything as simple as possible, but not simpler." (Einstein)

1.4 Probability is like a language

When you start learning probability and statistics, it helps remembering the times when you last learned a foreign language. Or of the times when you first learned to program a computer.

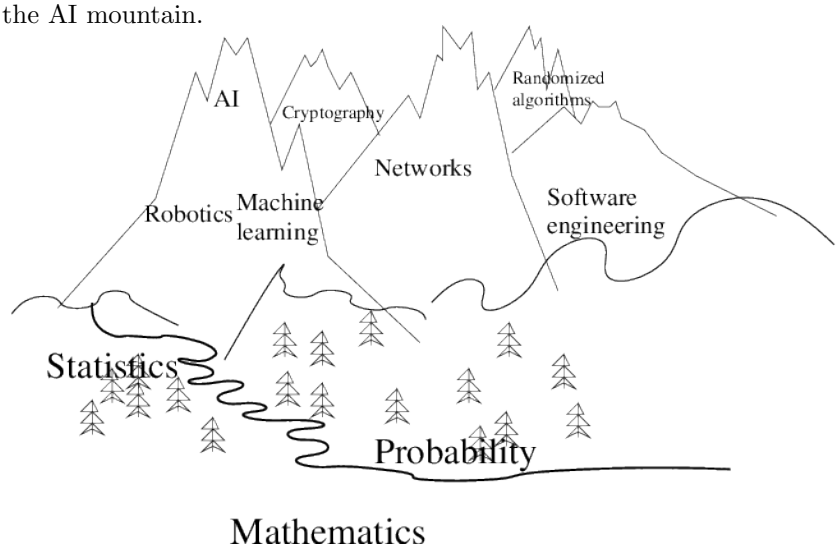
Just like programming languages are languages that describe computation, probability is a language meant to describe uncertainty. It is a very powerful language (proved mathematically). There are many programming languages (why?). [There are many human languages. Why?] Probability as a language for describing uncertainty has practically no competitor.

This is good news: **you only need to learn it once**. Then you can converse with everyone else who understands probability, be they particle physicists or psychologists or casino-managers.

1.5 What we will do in this course

We will learn the fundamental concepts of probability and statistics. We will develop a set of tools and use some of them for practice. We will see examples from CS where our tools apply.

Below is a pictorial view of our road, starting in the plains of Mathematics and winding up the hills of Probability to the heights of Statistics. From there, we will look onto and take short explorations on the nearby mountains, most often on the AI mountain.



1.5.1 Describing randomness

As a first step toward the study of randomness, note that not all random processes are alike.

- How is a coin toss different from a dice roll?
- A fair coin from an unfair one?
- How is the coin toss different from the salt-and-pepper noise in the TV signal?
- How is the salt-and-pepper noise different from the noise you hear in a radio signal?

The concepts of **sample space** and **probability distribution** will help us distinguish between these. As you notice here and in the examples below, some processes are "more random" than others (or should we say "less predictable"?). Concepts like **variance** and **entropy** allow us to measure the predictability of a random variable. Also to measure the quality of our predictions.

We are often interested in the relationship between two (random) phenomena. The concepts of **conditional probability, probabilistic dependence** (and **independence**) will help us handle these.

- The stock market is a random process. However, there is some dependence between one day and the next, and there are clear trends over longer periods of time.
- The weather from one year to the next is highly random (e.g. on some years it rains a lot in winter in Seattle, on others there are many sunny days). However meteorologists have discovered some correlations: about every three years, Seattle gets a dryer than usual fall and winter, while Californians get a rainier than usual season. This is the El Niño effect.
- "Smoking causes lung cancer." This effect has been proved scientifically, but it is a non-deterministic one. You may be a smoker and live a long life without lung cancer, or you may get the disease without being a smoker. But overall, a smoker is *more likely* to have lung cancer than a non-smoker.

1.5.2 Predictions and decisions

Probability helps us make **predictions**. This is usually the ultimate goal of an engineer using probability. Of course, if the phenomenon we are interested in is non-deterministic, we can never predict the future with certainty. We will predict the future in the language of probability as well. In other words, **all predictions are guesses.** But some guesses are better than others (probability and statistics study which and why), and sometimes we will be able to also compute a measure of **confidence**, for example a **confidence interval**, for our guess. This will be a guess too, of course.

What kind of predictions can we make? Here are some examples.

Assume that the probability that a plane's engine fails during the period of 1 hour is $\mathrm{p} = 10^{-6}$. Then we can predict that the probability that the engine fails in 100 hours of flight is no more than 10^{-4}.

If a server has 2 processors, the chance that each of them being busy is 1/3, then the chance that the server can take our job is at least 2/3.

The outcomes of the toss of a fair coin are 0 or 1 with probabilities equal to 1/2. We cannot predict the outcome of the next coin toss, but we can predict that if the coin is flipped n =100 times, we will observe about 50 1's. If the coin is flipped n =1000 times, we can predict that we'll see about 500 1's and the second is a better approximation than the first.

It is interesting to contrast the last example with the properties of computer simulation. Suppose that instead of applying the laws of probability, you write a program that simulates the n coin tosses and counts the number of 1's. The running time of the program will increase if n grows. Intuitively, if you want to simulate a larger system, you need to put more effort into the computation.

Probability eludes this problem in cases like the one above: the computation is just the same for every n (it consists of dividing n by 2) and the result becomes ever more accurate when n is increased! This kind behavior of probabilistic predicitions is called the **law of large numbers**.

Statistical decisions. As the future is never completely known, every decision we make is a "decision in uncertainty". For example:

- **Choosing the optimal cache size.** There is a certain cost (e.g time delay) to a cache miss, but increasing the size of the cache has a cost too. Based on some estimate of the characteristics of the applications the processor will be running, the designer needs to find the optimal cache size, i.e the cache size that give the best trade-off between cost and performance.

- **Stagewise diagnosis.** A doctor sees a new patient who describes her symptoms. Or, a computer specialist is trying to fix a broken computer. The doctor needs to diagnose the disease (and prescribe a treatment). She can prescribe the treatment right away, or can perform more investigations (like MRI scans, blood tests). With each new tests, the doctor accquires more knowlege (and reduces her uncertainty) about the patient's "internal state", so presumably she can make a better treatment recommendation. However, medical investigations carry costs (monetary, in time, discomfort to the patient, risks of secondary effects). The costs themselves are sometimes not known before the test is performed (for example, the discomfort to the patient, or, in case of a treatment like a surgery, the benefits it will bring). Also, later decisions to perform a certain lab test may depend on the results of a previous one. (For example, a test for disease A is negative, so the doctor proceeds to test for the less frequent disease B). The doctor's problem is a stagewise decision problem, because the doctor must make a series of decisions, each of them based on previously accquired information and some probabilistic guesses about the future.

The theory of statistical decision tells us how reason about this act, how to express mathematically our goals, our knowledge and our uncertainty, and how from them to obtain the "optimal" choice.

1.5.3 What is statistics?

A **model** is a probabilistic description of how the data is generated. For example, a fair coin is a model. If we have a model, probability allows us to make predictions about the data that it will generate. For example, we can predict to see roughly 50 1's in 100 tosses of a fair coin.

Statistics does the reverse: we observe the data and try to infer something about the source of the data. For example if I observe 50 1's in 100 tosses what can I say about the coin that produced the data? Or, if I observe that 31 out of 50 patients who had a new treatment got well while in the control group 18 out of 38 got well, what can I say about the new treatment? [Of course, since the data is random, everything I said would still be a guess.]

In engineering, one studies data and constructs models in order to make predictions. For example, suppose a book selling company has collected data on the sales of "Harry Potter" during the first 10 months of 2001. The company's statistician analyses this data and constructs a probabilistic model of the demand for "Harry Potter". The company wants to use this model to predict the sales of "Harry Potter" in the 11th and 12th month of the year. [What factors should the model take into account in order to make good predictions? Remember that the movie "Harry Potter" was released in November and December is Christmas shopping month.] Or, in another scenario, the company uses data on "Harry Potter" from Dec 00 to Nov 01 to construct a model of demand jumps. Then the model is applied to the sales of "Lord of the Rings" in December 01.

Chapter 2

The Sample Space, Events, Probability Distributions

2.1 Summary

S	the sample space (or outcome space)
$x \in S$	an outcome
$E \subseteq S$	an event
$P : \mathcal{P}(S) \longrightarrow [0,1]$	a probability distribution on S
$E \longmapsto P(E)$	maps an event into its probability
$X : S \longrightarrow \mathbb{R}$	a random variable is a function of the outcome

2.2 The sample space

The **sample space** (or **outcome space**) S is the set of outcomes of a **random experiment**.

Example 2.1 *Tossing a coin.* $S = \{0, 1\}$

Example 2.2 *Rolling a die.* $S = \{1, 2, 3, 4, 5, 6\}$

Example 2.3 *Rolling two dice.* $S = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ *i.e the set of all pairs of integers between 1 and 6.*

Example 2.4 *Tossing a coin 10 times.* S *= the set of all binary sequences of length 10.* $|S| = 2^{10}$

Example 2.5 *Measuring the height of a student in this class.* $S = (0, 8\text{ft}]$.

Example 2.5 shows a **continuous** sample space. The previous examples showed **finite** sample spaces. A sample space which is finite or countable is called **discrete**. The last example also shows that the sample space can include outcomes that will never appear (no student will be 0.1 ft high; also, there may be no student 8 ft high, and there may even not be a student 6.02133457 ft high). Including more outcomes than one may necessarily need is not a mistake. However, **to miss any outcomes that are possible** is a serious mistake that typically results in erroneous conclusions further down the line.

Example 2.6 *The position of a robot in a room.* $S = \{(x, y),\ 0 \leq x \leq L,\ 0 \leq y \leq W\}$ *(L and W are the length and width of the room.)*

Example 2.7 *There can be very large discrete sample spaces. For example, the set of all 1024×1024 BW images. If we assume that the value of each pixel is contained in a byte then we have* $|S| = 1024 \times 1024 \times 256$.

Example 2.8 *A random number generated by* `srand48()` *the C random number generator.* $S = [0, 1]$*; or is it so?*

2.3 Events

An **event** is a subset of S [1]

Example 2.9 *For the die roll example,* $E_1 = \{1\}$, $E_2 =$*"the outcome is even"* $= \{2, 4, 6\}$, $E_3 =$*"the outcome is* ≥ 3*"*$= \{3, 4, 5, 6\}$, $E_4 =$*"the outcome is* ≤ 0*"*$= \emptyset$ *are all events.*

Example 2.10 *For the image sample space of example 2.7,* $E_5 =$*"pixel (0,0) is white" is an event. It consists of all the possible images that have a white upper left corner pixel, hence* $|E_5| = (1024 \times 1024 - 1) \times 256$. *The event* $E_6 =$*"the first row is black" is the set of images whose first row is black and it has 1023×1024×256 elements.*

[1] For a continuous outcome space, not all subsets of S are events, but only what is called measurable sets. In practice you will never encounter a set that's not measurable so from now on we shall assume that all subsets that we deal with in any sample space are events.

Example 2.11 *For the robot position experiment of example 2.6, the event* E_7 = *"the robot is 3 ft from the right wall" represents the set* $\{(x,y), L-3\text{ft} \leq x \leq L, 0 \leq y \leq W\} \sim \{(x,y) \in S, x \geq L-3\text{ft}\}$. *The event* E_8 = *"the robot is no more than* ε *away from the middle of the room" is described by the set* $E_8 = \{(x,y) \in S,\ (x-L/2)^2 + (y-W/2)^2 \leq \varepsilon\}$.

One can map events (i.e subsets of S) into propositions about the elements of S: each event is the domain where a certain proposition is true and vice versa. Therefore we can apply propositional logic operations to events.

For example, the event "E_2 and E_3" (i.e "the outcome is even and ≥ 3") represents the set $E_2 \cap E_3 = \{4,6\}$. The set of BW images that "have a white (0,0) pixel or a black first row" is $E_5 \cup E_6$. If "the first row is black" then "pixel (0,0) is not white"; in other words "$E_6 \Rightarrow \bar{E}_5$" ($E_6$ implies non-E_5). Below is a table containing all the relationships between propositional logic and set operations.

Event	**Propositional operation**
$A \cup B$	A OR B
$A \cap B$	A AND B
$\bar{A} = S \setminus A$	NOT A
$A \subseteq B$	A IMPLIES B
$(A \setminus B) \cup (B \setminus A)$	A XOR B
S	TRUE (sure)
$\emptyset$	FALSE (surely NOT)

With this "translation" one can express "in words" other relationships from set algebra, like

$$(A \cap B) \cup (A \cap \bar{B}) \ = \ A$$

"If A is true, then either A and B are both true, or A and $\bar{B}$ are true". For example, let A="the alarm is on" and B="there is a burglar in the house". Lets us decompose the event A (that I observe) into two disjoint events "the alarm is on and there is a burglar in the house" and "the alarm is on and there is no burglar in the house (perhaps an earthquake set it on)" (that I have procedures to deal with: for example call the police in the first case and turn the the alarm off in the second case).

2.4 Probability

2.4.1 The definition

A **probability distribution** P is a function that assignes to each event E a positive number $P(E)$, called its **probability**.

To be a probability distribution, P has to satisfy the following 3 axioms:

1. **Positivity:** $P(E) \geq 0$ for all events $E \subseteq S$.
2. S **is the certain event:** $P(S) = 1$.
3. **Additivity:** If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

Intuition. It will be useful throughout this course to think of probability as behaving like mass (or volume, or area, or number of elements). Then one can interpret the axioms as describing how a function should be in order to "behave like mass". In terms of mass (or volume, etc) the axioms read: "Mass is a property of subsets of the universe". "Atoms", i.e the elements of S, are themselves subsets, so they have "mass" (probability) too.

1. "Mass is always positive".
2. "The total mass in the universe is finite (and by convention equal to 1)."
3. "If we decompose a slice of the universe into two disjoint parts, then the total mass of the slice equals the sum of its parts (conservation of mass in a fashion)."

The truth about Axiom 3 – A mathematical digression

Axiom 3 is in reality somewhat more complicated. It's precise formulation is:

$3'$. If A_1, A_2, ... A_n, ... is a sequence of mutually disjoint sets (i.e $A_n \cap A_m = \emptyset$ for all $m \neq n$) then $P(A_1 \cup A_2 \cup \ldots \cup A_n \cup \ldots) = \sum_{n=1}^{\infty} P(A_n)$.

Note that the above axiom cannot (and should not) be extended to non-countable unions of sets. To understand the difference, the positive integers 1, 2, 3, ... are a countable set, while the points in a square are not countable. There are more points in the unit square then there are positive integers. It is wrong to say that the probability of the unit square is equal to the sum of the probabilities of its points.

2.4.2 Two examples

Let us first describe two examples to illustrate the properties. The first example applies to the dice roll experiment of example 2.2.

Example 2.12 *For a fair die, each outcome has equal probability, so we know that*

$$P(\{i\}) \;=\; \frac{1}{6} \quad \text{for all } i = 1, \ldots 6$$

This means that for a general event E,

$$P(E) \;=\; \frac{|E|}{6}$$

Let us check that it satisfies the axioms of probability. Axiom 1 is obvious, we verify axiom 2:

$$P(S) \;=\; \frac{|S|}{6} \;=\; \frac{6}{6} \;=\; 1$$

To verify axiom 3, note that if two sets A, B are disjoint, then $|A \cup B| \;=\; |A| + |B|$. With this, axiom 3 follows easily.

Example 2.13 *For the robot in a room experiment of example 2.6 let us define*

$$P(E) \;=\; \frac{\text{area}(E)}{\text{area}(S)}$$

i.e the probability of the robot being in a certain region E is proportional to its area.

Again, checking the axioms is straight-forward. Any area is ≥ 0, therefore $P(E) \geq 0$ for all E. $P(S) = \text{area}(S)/\text{area}(S) = 1$. If two regions A, B are disjoint, then $\text{area}(A \cup B) \;=\; \text{area}(A) + \text{area}(B)$. With this, axiom 3 follows easily again.

Both examples illustrate **uniform** distributions, i.e. distributions where "each outcome has the same probability".

2.4.3 Properties of probabilities

The properties that we will derive here give us the opportunity of the first practice with probability calculus. They are also the most general and fundamental properties of any probability distribution and thus worth remembering. Third, they will serve as a "sanity check", showing whether the newly-introduced concept makes sense.

Proposition 2.1 $P(\overline{A}) \;=\; 1 - P(A)$

Proof. $A \cup \overline{A} = S$ and $A \cap \overline{A} = \emptyset$ imply that $P(S) = P(A) + P(\overline{A}) = 1$. From which follows that $P(\overline{A}) \;=\; 1 - P(A)$.

Proposition 2.2 *$P(A) \leq 1$ for all $A \subseteq S$.*

Proof. Every A has a complement $\overline{A}$ whose probability is non-negative. Therefore,

$$P(A) = 1 - P(\overline{A}) \leq 1$$

Proposition 2.3 *$P(\emptyset) = 0$.*

Proof. We have $S \cup \emptyset = S$ and $S \cap \emptyset = \emptyset$ from which $P(\emptyset) + P(S) = P(S)$ or $P(\emptyset) + 1 = 1$.

Proposition 2.4 *If $A \subseteq B$, then $P(A) \leq P(B)$.*

Proof. If $A \subseteq B$, then B can be written as the union of the disjoint subsets A and $B \setminus A$. Therefore,

$$P(B) = P(A) + P(B \setminus A) \geq P(A)$$

Note also that from the above follow that, when $A \subseteq B$, $P(B \setminus A) = P(B) - P(A)$ and $P(A \setminus B) = 0$.

Proposition 2.5 *$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*

Proof. $A \cup B$ can be written as the disjoint union of $A \setminus B$, $B \setminus A$ and $A \cap B$. Therefore,

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &= [P(A) - P(A \cap B)] + [P(B) - P(A \cap B)] + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

2.4.4 Another example – the probability of getting into the CSE major

Example 2.14

At Everdry State University, to get into the CSE major a student needs to get a passing grade in at least two of the following 3 subjects: Computer Programming, Physics and English. Melvin Fooch, freshmen at ESU is calculating his

chances of getting into the CSE major. He has denoted by P the event "passing Physics", by C the event "passing Computer Programming" and by E "passing English". He has spent a sleeples night figuring out the probability for each possible outcome.

CPE	0.20	$\overline{C}PE$	0.12
$CP\overline{E}$	0.13	$\overline{C}P\overline{E}$	0.11
$C\overline{P}E$	0.15	$\bar{C}\bar{P}E$	0.11
$C\bar{P}\bar{E}$	0.11	$\bar{C}\bar{P}\bar{E}$	0.07

Now he is too tired to add them up so let's help him. The event he's most interested in is E_1 entering the CSE major, which is the same as passing at least two of the three courses.

$$\begin{aligned} P(E_1) &= P(CPE \cup \overline{C}PE \cup C\overline{P}E \cup CP\overline{E}) \\ &= P(CPE) + P(\overline{C}PE) + P(C\overline{P}E) + P(CP\overline{E}) \text{(since they're disjoint events)} \\ &= 0.2 + 0.12 + 0.15 + 0.13 \\ &= 0.6 \end{aligned}$$

We can also compute other probabilities:

$$\begin{aligned} P(\text{Melvin passes at most 2 classes}) &= \\ &= 1 - P(CPE) \\ &= 0.8 \end{aligned}$$

$$\begin{aligned} P(\text{Melvin passes English}) &= \\ &= P(E) \\ &= P(CPE \cup \overline{C}PE \cup C\overline{P}E \cup \overline{C}\overline{P}E) \\ &= P(CPE) + P(\overline{C}PE) + P(C\overline{P}E) + P(\overline{C}\overline{P}E) \\ &= 0.2 + 0.12 + 0.15 + 0.11 \\ &= 0.58 \end{aligned}$$

$$\begin{aligned} P(\text{Melvin passes Physics but not English}) &= \\ &= P(P \cap \overline{E}) \\ &= P(CP\overline{E} \cup \overline{C}P\overline{E}) \\ &= P(CP\overline{E}) + P(\overline{C}P\overline{E}) \\ &= 0.13 + 0.11 \\ &= 0.24 \end{aligned}$$

Chapter 3

Finite sample spaces. The multinomial distribution

3.1 Discrete probability distributions

If the outcome space S is finite or countable (i.e discrete), a probability P on it is called *discrete*.

3.1.1 The uniform distribution

If all the outcomes have equal probability, i.e $\theta_0 = \theta_1 = \ldots \theta_{m-1} = \frac{1}{m}$, then the distribution is called a **uniform distribution**. Examples 3.2 and 3.1 above represent uniform distributions.

The probability of any event E under a uniform distribution equals

$$P(E) \ = \ \frac{|E|}{m} \tag{3.1}$$

(where $|E|$ represents the number of elements, or **cardinality** of E). Thus computing probabilities of events under uniform distribution is reduced to counting. Equation (3.1) is at the basis of an older definition of probability that you may have encountered:

$$\text{probability} \ = \ \frac{\text{number of "favorable" cases}}{\text{total number of cases}}$$

Example 3.1 *The die roll. A fair die has six faces,* $S = \{f_1, f_2, \ldots f_6\}$ *having equal probabilities of occurring in a roll. Thus* $\theta_{f_1} = \theta_{f_2} = \ldots = \frac{1}{6}$.

3.1.2 The Bernoulli distribution

This distribution describes a biased coin toss.

Example 3.2 *The (biased) coin toss.* $S = \{0, 1\}$, $\theta_0 = 1 - p$, $\theta_1 = p$

Example 3.3 *Component testing. All integrated circuits produced by a factory are tested. The outcomes of the test are given by* $S = \{\text{pass, fail}\}$ *with probabilities* $\theta_{\text{pass}} = 0.99$, $\theta_{\text{fail}} = 0.01$.

3.1.3 The exponential (geometric) distribution

The sample space is the set of integers $S_m = \{0, 1, \ldots m-1\}$ and the probability distribution is given by

$$P(n) \;=\; \frac{1}{Z}\gamma^n, \quad 0 < \gamma < 1 \tag{3.2}$$

The value γ is called the *parameter* of the distribution. In the above Z is the number that assures that the probabilities sum to 1. It is called the *normalization constant* of P.

$$Z \;=\; \sum_{n=0}^{m-1} \gamma^n \;=\; \frac{1-\gamma^m}{1-\gamma} \tag{3.3}$$

Hence, the exponential distribution is

$$P(n) \;=\; \frac{1-\gamma}{1-\gamma^m}\gamma^n \tag{3.4}$$

This distribution is also known as the *geometric* distribution, because the probabilities $P(n)$ are the terms of a geometric progression.

One can define the exponential distribution over the whole set of integers $S = \{0, 1, 2, \ldots n, \ldots\}$ by formula (3.2). Then

$$Z \;=\; \sum_{n=0}^{\infty} \gamma^n \;=\; \frac{1}{1-\gamma} \tag{3.5}$$

and the geometric distribution becomes

$$P(n) \;=\; (1-\gamma)\gamma^n \tag{3.6}$$

Note that for $\gamma = \frac{1}{2}$ we have $P(n) = \frac{1}{2^{n-1}}$

3.1.4 The Poisson distribution

The Poisson distribution is defined over the range of non-negative integers $\{0, 1, \ldots n, \ldots\}$ by

$$P(n) \;=\; e^{-\lambda}\frac{\lambda^n}{n!} \tag{3.7}$$

The parameter $\lambda > 0$ is called the **rate** of the Poisson distribution for reasons that will become clear soon.

The factor $1/e^{\lambda}$ represents the normalization constant of the Poisson distribution. Remember from calculus the identity

$$e^{\lambda} \;=\; \sum_{n\geq 0}\frac{\lambda^n}{n!} \quad \text{for } \lambda \in (-\infty, \infty) \tag{3.8}$$

In contrast to the exponential distribution $(1-\lambda)\lambda^n$ which always has a maximum at $n = 0$, the Poisson is first increasing to a maximum then decreasing asymptotically towards 0. Figure 3.1 shows this distribution for different values of λ.

Mathematical digression: The Poisson distribution as a sum of Bernoulli's

Assume: We have a unit interval, divided into N equal intervals, where N will tend to infinity. For each subinterval, of length $\Delta t = 1/N$, the probability of observing a 1 is $p = \lambda\Delta t$. We want the probability of n successes in the unit interval. This is the sum of N independent Bernoulli trials (n_1 in the course

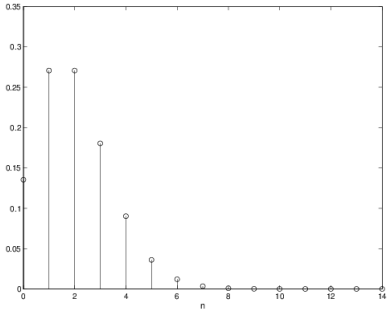


Figure 3.1: The Poisson distribution for $\lambda = 2$.

notes) therefore

$$\begin{aligned} P(n) &= \binom{N}{n} p^n (1-p)^{N-n} && (3.9)\\ &= \frac{N(N-1)\dots(N-n+1)}{n!}\lambda^n \Delta t^n (1-\lambda\Delta t)^{N-n} && (3.10)\\ &= \frac{\lambda^n}{n!}\frac{N(N-1)\dots(N-n+1)}{N^{-n}}(1-\lambda/N)^{N-n} && (3.11) \end{aligned}$$

So we have to prove that $a_N = \frac{N(N-1)\dots(N-n+1)}{N^{-n}}(1-\lambda/N)^{N-n}$ tends to $e^{-\lambda}$.

$$a_N = \prod_{k=0}^{n-1} \frac{1-k/N}{1-\lambda/N} \cdot (1-\lambda/N)^N \tag{3.12}$$

And now it's really easy, because the first product tends to 1 when $N \to \infty$ with n fixed and the second part tends to $e^{-\lambda}$.

Why the last limit? It is known that $\lim_{x\to\infty}(1+1/x)^x = e$. Use $x = -N/\lambda$ and force the exponent to be $1/x$ times $-\lambda$.

3.1.5 Discrete distributions on finite sample spaces – the general case

We can denote the elements of a finite S by $\{x_0, x_2, \dots x_{m-1}\}$ where m is the cardinality of S. A probability P over S is determined by its values $\theta_i = P(x_i)$ on each element of S because

$$P(A) = \sum_{x\in A} P(x) \tag{3.13}$$

In fact P is completely determined by any $m-1$ such values due to the constraints

$$\begin{aligned} \theta_i &\geq 0 \text{ for } i = 0, \dots m-1 && (3.14)\\ \sum_{i=0}^{m-1} \theta_i &= 1 && (3.15) \end{aligned}$$

The numbers θ_i, $i = 0, \dots m-1$ can be given by a rule like for the uniform or exponential distributions, but in general we are free to choose them any way we want, subject to the contraints (3.14). In this case, they are the *parameters* of the distribution. So, in general, a discrete finite distribution is defined by $m-1$ free parameters.

3.2 Sampling from a discrete distribution

How can one generate on a computer samples from an arbitrary distribution with parameters $\theta_0, \dots \theta_{m-1}$?

Here is a method that uses the computer's `rand()` function, that generates a random number uniformly distributed between 0 and 1. We define the numbers $a_0, a_1, \dots a_m$ by:

$$\begin{aligned} a_0 &= 0 && (3.16)\\ a_1 &= a_0 + \theta_0 && (3.17)\\ &\dots && (3.18)\\ a_{k+1} &= a_k + \theta_k && (3.19)\\ &\dots && (3.20)\\ a_m &= 1 && (3.21) \end{aligned}$$

Then we generate a random number r with `rand()`. If $a_k < r \leq a_{k+1}$ the method outputs k.

We will intuitively show why this method is correct. It is because under a uniform distribution over the $[0, 1]$ interval (given by `rand()`) the probability that r falls in the interval $(a_k, a_{k+1}]$ is equal with the length of the interval $a_{k+1} - a_k = \theta_k$. (See also example 5.2.)

3.3 Repeated independent trials

A coin tossed n times, a series of n die rolls are both examples of experiments with repeated trials. (What about rolling n identical dice *simultaneously*?) In a repeated trial, the outcome space S^n is

$$S^n = \underbrace{S \times S \times \dots \times S}_{n \text{ times}}. \quad (3.22)$$

The elements of S^n are length n sequences of elements of S. If S has m elements then $|S^n| = m^n$.

We denote by $x^{(k)}$ the outcome of trial k ($x^{(k)}$ is a random variable); the outcome of the repeated trial is $x^{(1)}, \dots x^{(n)}$.

If in a set of repeated trials, the outcome of a trial $x^{(k)}$ is not influenced in any way by the outcomes of the other trials, either taken together or separately, we say that the trials are **independent**. Independence is a very useful and important property, and will be studied in more detail later. Independent events

have the following property: the probability of two or more independent events is equal to the product of the probabilities of the individual events.

In the case of repeated trials, the above property amounts to:

$$P(x^{(1)}, \ldots x^{(n)}) = P(x^{(1)})P(x^{(2)}) \ldots P(x^{(n)}) \tag{3.23}$$

We also have that $P(x^{(k)}) = \theta_{x^{(k)}}$. Hence, we can compute the probability of every outcome in S^n using the parameters of the original probability over S

$$\begin{aligned} P(x^{(1)}, \ldots x^{(n)}) &= P(x^{(1)})P(x^{(2)}) \ldots P(x^{(n)}) \\ &= \theta_{x^{(1)}}\theta_{x^{(2)}} \ldots \theta_{x^{(n)}} \\ &= \prod_{i=0}^{m-1} \theta_i^{n_i} \end{aligned} \tag{3.24}$$

The exponents n_i represent the number of times value x_i appears in the sequence. Often they are called the **counts** associated with the outcome $(x^{(1)}, \ldots x^{(n)})$. They are integer-valued random variables satisfying the constraints

$$n_i \geq 0, \ i = 0, \ldots m-1 \tag{3.25}$$

$$\sum_{i=0}^{m-1} n_i = n \tag{3.26}$$

Example 3.4 Rolling a die 5 times. *Below are a few outcomes with their counts (all outcomes have the same probability)*

outcome x	n_1	n_2	n_3	n_4	n_5	n_6
11111	5	0	0	0	0	0
23166	1	1	1	0	0	2
63261	1	1	1	0	0	2
16326	1	1	1	0	0	2
42453	0	1	1	2	1	0

Note the difference between $x^{(k)}$ (the outcome of the k-th trial) and x_i (outcome i of S). For example, in the line above $x^{(1)} = 4$ while x_1 is always 1.

Example 3.5 *A fair coin tossed $n = 10$ times. $S^n = \{0, 1\}^n = \{$0000000000, 0000000001, $\ldots$, 1111111111$\}$; $\theta_0 = \theta_1 = \ldots = \theta_{1023} = 0.5^{10}$.*

Both examples above illustrate uniform probabilities over spaces of equal length sequences.

Example 3.6 *The biased coin. A coin is tossed 4 times, and the probability of 1 (Heads) is $p > 0.5$. The outcomes, their probability and their counts are (in order of decreasing probability):*

outcome x	n_0	n_1	$P(x)$	event
1111	0	4	p^4	$E_{0,4}$
1110	1	3	$p^3(1-p)^1$	$E_{1,3}$
1101	1	3	$p^3(1-p)^1$	
1011	1	3	$p^3(1-p)^1$	
0111	1	3	$p^3(1-p)^1$	
1100	2	2	$p^2(1-p)^2$	$E_{2,2}$
1010	2	2	$p^2(1-p)^2$	
1001	2	2	$p^2(1-p)^2$	
0110	2	2	$p^2(1-p)^2$	
0101	2	2	$p^2(1-p)^2$	
0011	2	2	$p^2(1-p)^2$	
0100	3	1	$p^1(1-p)^3$	$E_{3,1}$
1000	3	1	$p^1(1-p)^3$	
0010	3	1	$p^1(1-p)^3$	
0001	3	1	$p^1(1-p)^3$	
0000	4	0	$(1-p)^4$	$E_{4,0}$

3.4 Probabilities of sequences vs. probabilities of events. The multinomial distribution

Note that in the table above, there are several outcomes that have the same probability. In fact, all outcomes that have the same number of zeros n_0 (and correspondingly the same number of ones $n_1 = n - n_0$) have the same probability and this is $(1-p)^{n_0}p^{n_1}$.

We denote by E_{n_0,n_1} the event "the outcome has n_0 zeros and n_1 ones". In other words, the event E_{n_0,n_1} is the set of all sequence with n_0 zeros (and $n_1 = n - n_0$ ones). Events of this kind are so frequently used and so important that their probabilities have the special name of "multinomial distribution". They arise in cases when all we care about the outcome sequence is the number of individual outcomes of each kind but not the order in which they occur.

Example 3.7 *For example, if in the above experiment, we would gain 1\$ for each 1 and nothing for a zero, then from the point of view of the total gain the order of the zeros and ones in the sequence would not matter. The probabilities of gaining 0, 1, 2, 3, 4\$ respectively equals the probabilities of $E_{4,0}, \ldots E_{0,4}$.*

If there are $m > 2$ possible outcomes in a trial, then each outcome sequence is described by m counts $n_0, n_1, \ldots n_{m-1}$. The set of all sequences with the same counts $n_0, n_1, \ldots n_{m-1}$ represents the event $E_{n_0,n_1,\ldots n_{m-1}}$.

Now we will compute the probability of an event $E_{n_0,n_1,\ldots n_{m-1}}$. To simplify notation, we will refer to $E_{n_0,n_1,\ldots n_{m-1}}$ as $(n_0, n_1, \ldots n_{m-1})$ when no confusion is possible.

We shall start with the case of a **binary** experiment (like the coin toss), where there are only 2 possible outcomes ($m = 2$). For a given (n_0, n_1), with $n_0 + n_1 = n$, there are

$$\binom{n}{n_1} \;=\; \binom{n}{n_0} \;=\; \frac{n!}{n_0! n_1!}$$

different outcomes that have counts (n_0, n_1). All the outcomes in the event (n_0, n_1) have the same probability (and are clearly mutually exclusive). Therefore the probability of (n_0, n_1) is given by

$$P(n_0, n_1) \;=\; \theta_0^{n_0} \theta_1^{n_1} \binom{n}{n_1}$$

For an experiment with $m > 2$ outcomes, the number of outcomes that correspond to a set of counts $(n_0,\, n_1, \ldots n_{m-1})$ is

$$\binom{n}{n_0,\, n_1, \ldots n_{m-1}} \;\triangleq\; \frac{n!}{n_0! n_1! \ldots n_{m-1}!} \tag{3.27}$$

read "n choose $n_0,\, n_1, \ldots n_{m-1}$" and called the **multinomial coefficient** indexed by $n_0, \ldots n_{m-1}$. Note the analogy with "n choose k" the well known **binomial** coefficient. Formula (3.27) can be proved by induction over m, starting from $m = 2$.

Then the probability of observing a set of counts $(n_0, \ldots n_{m-1})$ is obtained by multiplying the probability of one sequence, given by (3.24) with the total number of sequences exhibiting those counts:

$$P(n_0, \ldots n_{m-1}) \;=\; \frac{n!}{n_0! n_1! \ldots n_{m-1}!} \prod_{i=0}^{m-1} \theta_i^{n_i} \tag{3.28}$$

$$\;=\; \binom{n}{n_0\, n_1\, \ldots\, n_{m-1}} \prod_{i=0}^{m-1} \theta_i^{n_i} \tag{3.29}$$

Equation (3.28) defines the **multinomial distribution**. For $m = 2$ equation (3.28) is called the **binomial** distribution.

3.5 Examples

We have seen how to define a probability distribution over a discrete space. Let us now practice using it by computing the probabilities of various events.

Example 3.8 *Sending messages through a channel. The probability of a bit being corrupted when it is sent through a channel is $p = 10^{-6}$. We send a message length $n = 10^3$. What is the probability that the message is received correctly? What is the probability that at least one bit is corrupted? It is assumed that the errors on each bit are independent.*

Solution *The probability of receiving one bit correctly is $1 - p$. The probability of receiving all n bits correctly is*

$$(1-p)^n \;=\; (1-10^{-6})^{10^3} \;=\; 0.9990005 \;\approx\; 0.999 \tag{3.30}$$

The event "at least 1 bit is corrupted" is the complement of "all bits are received correctly", hence its probability is

$$1-(1-p)^n \;=\; 1-(1-10^{-6})^{10^3} \;=\; 0.0009995 \;\approx\; 10^{-3} \tag{3.31}$$

Note how these probabilities change if we up the bit error probability p to 10^{-3}:

$$P[\text{all bits correct}] \;=\; (1-10^{-3})^{10^3} \;=\; 0.368 \tag{3.32}$$

$$P[\text{at least one error}] \;=\; 1-0.368 \;=\; 0.632 \tag{3.33}$$

Example 3.9 *Toss a coin until a 1 comes up. The probability of a 1 is θ_1. What is the probability that the experiment takes n trials? Note that the number of trials until 1 comes up is an infinite (but discrete) outcome space.*

$$P(n) \;=\; P(\underbrace{00\ldots0}_{\times n-1}1) \;=\; (1-\theta_1)^{n-1}\theta_1 \tag{3.34}$$

For the fair coin with $\theta_1 = 0.5$, $P(n) = \frac{1}{2^n}$. You can easily verify that $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$.

Let us now compute the probability that the experiment takes at least $n \geq 1$

trials. This is an exercise in geometric series.

$$\begin{aligned} P[\text{at least } n \text{ trials}] &= \sum_{k=n}^{\infty} P(k) && (3.35)\\ &= \sum_{k=n}^{\infty} (1-\theta_1)^{k-1}\theta_1 && (3.36)\\ &= (1-\theta_1)^{n-1}\theta_1 \sum_{k=0}^{\infty} (1-\theta_1)^{k} && (3.37)\\ &= (1-\theta_1)^{n-1}\theta_1 \frac{1}{1-1+\theta_1} && (3.38)\\ &= (1-\theta_1)^{n-1} && (3.39) \end{aligned}$$

Another way to compute the probability of the same event is to notice that the experiment takes at least n trials if and only if the first $n-1$ trials have all outcome 0. The probability of this happening is

$$(1-\theta_1)^{n-1}$$

This problem is encountered in more realistic settings, like: (1) One takes a test as many times as necessary to pass it and one would like to know in advance the probability of having to take the test more than n times. (2) A part of an a car (or computer, or home appliance) will break with probability θ_1 on a given trip (hour of usage). Both the manufacturer and the user want to know what is the probability that the part can be used for n trips (hours) before it has to be replaced.

Example 3.10 *Simple market basket analysis. A grocery store sells Apples, Bread, Cheese, Detergent and Eggs. It is assumed that each customer buys a product randomly with a certain probability, independently of any other products they have already bought and of the products that other customers have bought. (This is a very simple customer model indeed!) The probabilities that a customer buys each of the 5 products are:*

$$\theta_A = 0.2 \;\; \theta_B = 0.4 \;\; \theta_C = 0.3 \;\; \theta_D = 0.1 \;\; \theta_E = 0.1 \qquad (3.40)$$

1. *A customer buys $n = 3$ things. What is the probability that he buys only bread?*

$$P[\text{only bread}] = P(BBB) = \theta_B^3 \qquad (3.41)$$

2. *A customer buys $n = 3$ things. What is the probability that she buys nothing but bread and cheese?*

$$P[\text{nothing but bread and cheese}] = P[(B \vee C), (B \vee C), (B \vee C)] = (\theta_B + \theta_C)^3 \qquad (3.42)$$

3. *A customer buys $n = 5$ things. What is the probability that she buys one of each product?*

$$P[\text{one of each}] \;=\; 5!\theta_A\theta_B\theta_C\theta_D\theta_E \tag{3.43}$$

4. *A customer buys $n = 5$ things. What is the probability that he buys 2 apples, 2 cheeses and one bread?*

$$\begin{aligned} P[\text{2A+2C+1B}] &= \binom{5}{2\,1\,2\,0\,0}\theta_A^2\theta_B\theta_C^2 && (3.44)\\ &= \frac{5!}{2!\,1!\,2!\,0!\,0!}\theta_A^2\theta_B\theta_C^2 && (3.45)\end{aligned}$$

5. *A customer buys $n = 5$ things. What is the probability that he buys at least 3 apples?*

$$\begin{aligned} P[\text{at least 3A}] &= P[\text{3A}] + P[\text{4A}] + P[\text{5A}] && (3.46)\\ &= \binom{5}{3}\theta_A^3(1-\theta_A)^2 + \binom{5}{4}\theta_A^4(1-\theta_A) + \binom{5}{5}\theta_A^5 && (3.47)\end{aligned}$$

Example 3.11 *A coin is tossed n times. The probability of obtaining a 1 on any toss is θ. What is the probability that there are two consecutive 1's in the outcome?*

Solution: Denote $S = \{0,1\}^n$, $A = \{x \in S \,|\, x \text{ contains 2 consecutive 1's }\}$ and $B = \overline{A}$. We will estimate the probability of B, then $P(A) = 1 - P(B)$.

The event B = "the outcome has no 2 consecutive 1's" can be further partitioned into the disjoint sets C_k, D_k defined as follows:

$C_k =$ the set of sequences in B ending in 1 and having exactly k 1's $\quad k = 1, 2, \ldots \lceil \frac{n}{2} \rceil$

$D_k =$ the set of sequences in B ending in 0 and having exactly k 1's $\quad k = 1, 2, \ldots \lfloor \frac{n}{2} \rfloor$

$\mathbf{x \in D_k}$ Then the sequence x can be written as a sequence of k "10" "symbols" and $n - 2k$ "0" symbols. The total number of symbols is $k + (n - 2k) = n - k$. Therefore,

$$P(D_k) \;=\; \binom{n-k}{k}[(1-\theta)\theta]^k(1-\theta)^{n-k} \;=\; \binom{n-k}{k}(1-\theta)^{n-k}\theta^k \tag{3.48}$$

$\mathbf{x \in C_k}$ Then the sequence x can be written as a sequence of $k-1$ "10" "symbols", a final 1 symbol, and $n - 1 - 2(k-1) = n - 2k + 1$ 0 symbols. Since the final 1 has a fixed position, the total number of sequences is

$$\binom{n-2k+1+k-1}{k-1} \;=\; \binom{n-k}{k-1}$$

Therefore,

$$P(C_k) \;=\; \theta \binom{n-k}{k-1} [(1-\theta)\theta]^{k-1}(1-\theta)^{n-2k+1} \;=\; \binom{n-k}{k} (1-\theta)^{n-k}\theta^k \quad (3.49)$$

Note that each two sets $\{C_k, D_k \text{ for all } k\}$ are disjoint and that the union of all the sets is equal to B. So,

$$P(B) \;=\; \sum_k P(C_k) + \sum_k P(D_k) \quad (3.50)$$

Also, recall that

$$\binom{m}{k} + \binom{m}{k-1} \;=\; \binom{m+1}{k} \quad (3.51)$$

Hence, for n even

$$P(B) \;=\; \sum_{k=0}^{n/2} \binom{n-k+1}{k} (1-\theta)^{n-k}\theta^k \quad (3.52)$$

and for n odd

$$P(B) \;=\; \sum_{k=0}^{(n-1)/2} \binom{n-k+1}{k} (1-\theta)^{n-k}\theta^k + (1-\theta)^{\frac{n-1}{2}}\theta^{\frac{n+1}{2}} \quad (3.53)$$

and $P(A) = 1 - P(B)$.

3.6 Models for text documents

3.6.1 What is information retrieval?

There is a collection of *documents* somewhere (like a library, the web, the archives of a newspaper or newsgroup, an image library) and you want to get information about a certain *topic* (for example "water skiing", "weather in Seattle") from it. The first step, at the library for example, would be to find books on "water skiing", on the web to find the pages mentioning water skiing, and in general to gather the documents that are relevant to your topic. Then you can study the documents and extract the useful information from them. The first step – retrieving all the useful documents from the collection – is called *document retrieval* or *information retrieval*. It is what the search engines (are supposed to) do.

Another related problem that belongs to information retrieval is the following: you have a web page that you like (because it is water skiing for example) and you want your search engine to find other pages that are like it.

What you give to the search engine, be it a whole web page or a few words describing what you want, is called the *query*. The query will be treated as a document, albeit a very short one.

In what follows we will present a simple statistical approach to information retrieval. Although simple, this method is the backbone of most working search engines and document retrieval systems.

3.6.2 Simple models for text

We assume that we deal with documents containing only plain text (i.e no images, links, graphics, etc). This is what some search engines do anyways. Others, like Google, also take into account the link structure of the web.

We shall construct a model for documents. That is we create a distribution P such that when we sample from it the outcomes are documents. So stated, this a *very* ambitious goal. Few have ever come close to attaining it. But something that is more easily done is to create a P such that all documents *could have* come from it. One of the simplest models is the so-called **"multinomial"** model of text. It assumes that documents are produced by sampling the first word from a dictionary, then sampling the second word independently from the first from the same dictionary, and so on until we reach a prescribed word limit n. If we hypothetically sampled from this P, sometimes the outcome would be a document, sometimes something else. This is OK as long as P doesn't assign probability 0 to any document. It means that documents will have in general a lower likelihood than they would under an "ideal" P.

Let us define the outcome space W as the set of all words $W = \{word_0, word_1, \ldots, word_{m-1}\}$. Let P_W be a distribution over W. The probability of sampling word w under P_W is $P_W(w) = \theta_w$. Then, to generate a "document", we will sample its length n from a distribution over lengths P_N, then sample n times independently from P to get the words. To simplify matters even more, we will assume that P_N is uniform for n between 1 and a given maximum length $n = 1000$. This way $P_N(n)$ is a constant and we can ignore it. Hence, the probability of a document d under P is

$$P(d) \;=\; P_N(|d|) \prod_{w \in d} P_W(w) \;=\; P_N(|d|) \prod_{w \in W} \theta_w^{n_w} \tag{3.54}$$

where, as usual, n_w is the number of occurrences of word w in d and $|d|$ is the length of the document.

An alternate model for information retrieval is the so called **bag of words** model. In this model, each word has a probability ϕ_w of appearing in the document. We assume that documents are generated by going through the list of words W, and at each word flipping a coin with a probability ϕ_w of obtaining

one (and probability $1 - \phi_w$ of obtaining 0). The words for which a 1 comes up are included in the document. (This method doesn't really produce a document, just the unordered collection of its words; therefore the name "bag of words" model.)

Under this model, the probability of a document is

$$P_B(d) \;=\; \prod_{w \in d} \phi_w \prod_{w \in W \setminus d} (1 - \phi_w) \tag{3.55}$$

One question is, where do we get the parameters θ or ϕ? We will estimate them from the documents themselves by the Maximum Likelihood method described in chapter 4.

Chapter 4

Maximum likelihood estimation of discrete distributions

4.1 Maximum Likelihood estimation for the discrete distribution

Assume that we have a discrete distribution P over S, with unknown parameters $\{\theta_i\}$. We are given n independent samples from P; they represent the dataset $\mathcal{D} = \{x^{(1)}, \ldots x^{(n)}\}$. The task is to estimate the parameters $\bar{\theta} = \{\theta_i\}$ using the data set.

The Maximum Likelihood (ML) principle tells us to write down the likelihood of the dataset as a function of the parameters and then to choose the vector of parameters $\bar{\theta}$ that maximize the likelihood.

$$\bar{\theta}^{ML} \;=\; \operatorname*{argmax}_{\bar{\theta}} L(\bar{\theta}) \tag{4.1}$$

The *likelihood*, i.e the probability of the data as a function of $\bar{\theta}$ is given by

$$L(\bar{\theta}) \;\equiv\; P(\mathcal{D}|\bar{\theta}) \;=\; \prod_{i=0}^{m-1} \theta_i^{n_i} \tag{4.2}$$

$L(\bar{\theta})$ has to be maximized subject to the constraints

$$\theta_i \geq 0 \text{ for } i = 0, \ldots m-1 \tag{4.3}$$

$$\sum_{i=0}^{m-1} \theta_i = 1 \tag{4.4}$$

The solution to this problem is

$$\theta_i^{ML} = \frac{n_i}{n} \text{ for } i = 0, \ldots m-1 \tag{4.5}$$

This estimate is consistent with the "popular" definition of probability, i.e

$$P(\text{outcome } i) = \frac{\#\text{times outcome } i \text{ occured}}{\text{total } \# \text{ observations}} \tag{4.6}$$

The ML estimate of θ is a function of the outcome, therefore it is a random variable. Note however that the values of $\bar{\theta}$ depend on the outcome only through the counts $(n_0\, n_1\, \ldots\, n_{m-1})$ (i.e the estimate $\bar{\theta}$ is the same for all outcomes that exhibit those counts). For this reason, the counts $(n_0\, n_1\, \ldots\, n_{m-1})$ are called the **sufficient statistics** of the sample. They summarize all the information in the data pertaining to estimating the distribution's parameters.

4.1.1 Proving the ML formula

Here we show two proofs of the result (4.5). A third elegant proof will be given in section 4.4

1. An elementary solution. We shall present here the case $m = 2$. In this case, we have to estimate one parameter θ_0, because $\theta_1 = 1 - \theta_0$. The likelihood is

$$L(\theta_0) = \theta_0^{n_0}(1-\theta_0)^{n-n_0} \tag{4.7}$$

We will work with the logarithm of the likelihood, shortly **log-likelihood** (this is convenient in many other cases).

$$l(\theta_0) \stackrel{\Delta}{=} n_0 \log\theta_0 + n_1 \log(1-\theta_0) \tag{4.8}$$

$$l'(\theta_0) = \frac{n_0}{\theta_0} - \frac{n_1}{1-\theta_0} \tag{4.9}$$

$$= \frac{n_0 - \theta_0(n_0+n_1)}{\theta_0(1-\theta_0)} \tag{4.10}$$

Equating the derivative with 0 we find the maximum of the likelihood at

$$\theta_0 = \frac{n_0}{n_0+n_1} = \frac{n_0}{n} \tag{4.11}$$

2. Advanced calculus solution using Lagrange multipliers. The standard calculus method for solving this problem is by introducing the Lagrange multiplier λ.

$$\bar{\theta}^{ML} = \underset{\bar{\theta}}{\text{argmax}} \underbrace{\left[L(\bar{\theta}) + \lambda(\sum_{i=0}^{m-1} \theta_i - 1) \right]}_{J} \tag{4.12}$$

The solution is found equating the partial derivatives of J w.r.t all the variables with 0.

$$\frac{\partial J}{\partial \theta_i} = \frac{n_i}{\theta_i} L(\bar{\theta}) - \lambda = 0 \tag{4.13}$$

$$\frac{\partial J}{\partial \lambda} = \sum_{j=0}^{m-1} \theta_j - 1 = 0 \tag{4.14}$$

The above partial derivatives exist only if both $n_i > 0$ and $\theta_i > 0$. For $n_i = 0$, it is easy to see that $\theta_i = 0$ maximizes J. If $n_i > 0$, then $\theta_i = 0$ cannot be a maximum. Therefore, the solution can be found by solving the system of equations above. We obtain

$$\theta_i^{ML} = \frac{n_i}{n} \text{ for } i = 0, \ldots m-1$$

4.1.2 Examples

Example 4.1 *The coin toss. A coin is tossed 10 times and the outcome is the sequence* 0010101010. *Hence,* $n = 10$, $n_0 = 6$, $n_1 = 4$. *Figure 4.1 displays the likelihood and the log-likelihood as a function of the parameter* θ_0. *The maximum is attained for* $\theta_0 = 0.6$. *This is verified by computing the ML estimate of the distribution of the outcomes for this coin by formula (4.5):*

$$\theta_0 = \frac{6}{10} = 0.6 \quad \theta_1 = \frac{4}{10} = 0.4$$

Example 4.2 *A grocery store sells Apples, Bread, Cheese, Detergent and Eggs. Last week it made* $n = 1000$ *sales, out of which*

$n_A = 250$ *Apples*
$n_B = 350$ *Breads*
$n_C = 150$ *Cheeses*
$n_D = 150$ *Diapers*
$n_E = 100$ *Eggs*

It is assumed that each customer buys a product randomly with a certain probability, independently of any other products they have already bought and of the products that other custormers have bought. (As we noticed before, this is a very simple customer model indeed!)

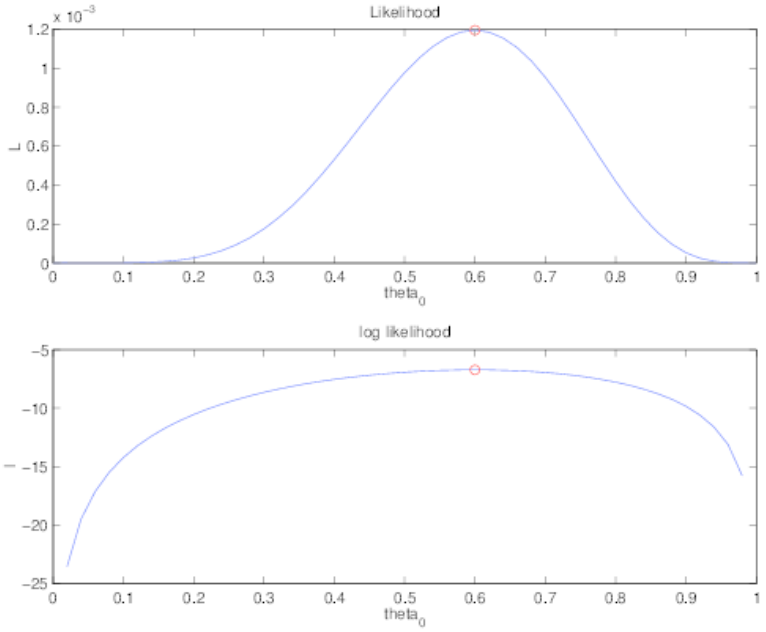


Figure 4.1: The likelihood and log-likelihood of the data in example 4.1 as a function of parameter θ_0.

The grocery store manager wants to estimate the probabilities of a random customer buying each of the 5 products. By the ML method, they are

$$\begin{aligned}
\theta_A &= \frac{250}{1000} = 0.25 \\
\theta_B &= \frac{350}{1000} = 0.35 \\
\theta_C &= \frac{150}{1000} = 0.15 \\
\theta_D &= \frac{150}{1000} = 0.15 \\
\theta_E &= \frac{100}{1000} = 0.1
\end{aligned}$$

Example 4.3 Maximum Likelihood estimation of the λ parameter for the Poisson distribution. *The Poisson distribution is defined in (3.7) as*

$$P(n) = e^{-\lambda}\frac{\lambda^n}{n!} \tag{4.15}$$

Assume that we have a data set $\mathcal{D} = \{n_1, n_2, \ldots n_N\}$ where N represents the number of observations in the data set. The likelihood and log-likelihood of the

parameter are

$$
\begin{aligned}
L(\lambda|\mathcal{D}) &= \prod_{i=1}^{N} P(n_i) && (4.16)\\
&= \prod_{i=1}^{N} e^{-\lambda}\frac{\lambda^{n_i}}{n_i!} && (4.17)\\
&= e^{-N\lambda}\frac{\lambda^{\sum_{i=1}^{N} n_i}}{\prod_{i=1}^{N} n_i!} && (4.18)\\
& && (4.19)\\
l(\lambda|\mathcal{D}) &= \ln L(\lambda|\mathcal{D}) && (4.20)\\
&= -N\lambda + \ln\lambda \sum_{i=1}^{N} n_i - \ln\prod_{i=1}^{N} n_i! && (4.21)
\end{aligned}
$$

Note that the last term above does not depend on λ so in fact has no role in the parameter estimation. To find the maximum of the (log-)likelihood, we differentiate l w.r.t λ, obtaining:

$$l'(\lambda) \;=\; -N + \frac{1}{\lambda}\sum_{i=1}^{N} n_i \;=\; 0 \qquad (4.22)$$

or

$$\lambda^{ML} \;=\; \frac{\sum_{i=1}^{N} n_i}{N} \qquad (4.23)$$

Hence, the rate λ is the arithmetic mean of the observed values of n.

4.2 The ML estimate as a random variable

The estimate of the parameters $\bar{\theta}^{ML}$ is a function of the counts $n_0\, n_1\, \ldots\, n_{m-1}$. The correspondence is one-to-one: for each set of counts, a distinct set of parameter estimates is obtained. Therefore

- the estimate $\bar{\theta}^{ML}$ must take only a finite number of values
- the distribution of the values of $\bar{\theta}^{ML}$ is the same as the distribution of the counts, namely the multinomial distribution.

Example 4.4 *Assume that $m = 2$ and $n = 20$ and let $\theta_1 = \theta$ for simplicity. The true value of θ is* 0.7*. What is the probability that the estimate $\theta^{ML} = 0.8$?*

What is the probability that the estimate $\theta^{ML} = 0.7$*? That it is 0.81? That it lies in the interval* $[0.5, 0.9]$*?*

$$\begin{aligned} P(\theta^{ML} = 0.8) &= P(\frac{n_1}{n} = 0.8) && (4.24)\\ &= P(n_1 = 0.8 \times n = 16) && (4.25)\\ &= \binom{n}{16} \theta^{16}(1-\theta)^4 && (4.26)\\ &= \binom{20}{16} 0.7^{16}(1-0.7)^4 && (4.27) \end{aligned}$$

Note that the true value of θ *(usually unknown) is used to compute this probability.*

Similarly

$$P(\theta^{ML} = 0.7) \;=\; P(n_1 = 20 \times 0.7 = 14) \;=\; \binom{20}{14} 0.7^{14}(1-0.7)^6 \qquad (4.28)$$

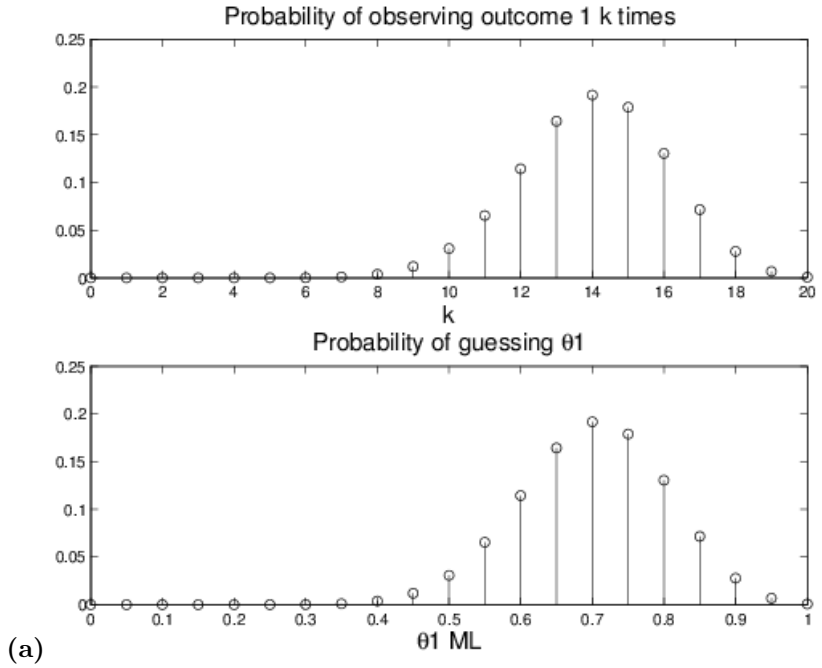
$P(\theta^{ML} = 0.81) = 0$ *because for* $n = 20$ *the estimate can only take values that are multiples of* $1/20$*.*

Figure 4.2,a shows the probability for every value of n_1 *and for the corresponding value of* θ^{ML}*. To compute the probability of an interval, we simply add up the probabilities of all* θ^{ML} *values in that interval. Therefore* $P(\theta^{ML} \in [0.5, 0.9]) = \sum_{\theta \in \{0.5, 0.55, 0.6 \ldots 0.9\}} P(\theta^{ML} = \theta) = 0.975$*.*

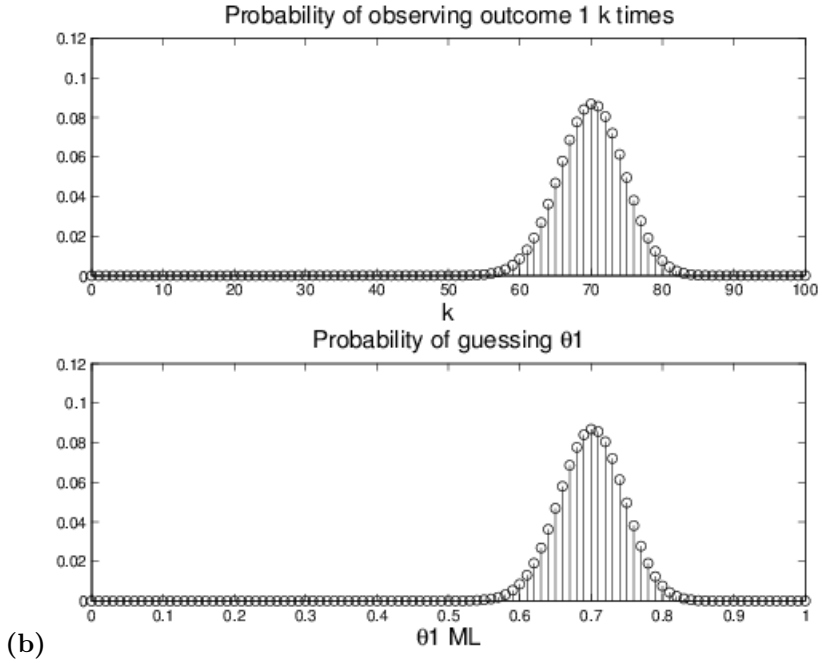
The following experiments show "empirical" distributions for the ML estimates. Unlike the previous figure, these distributions are histograms constructed from a large number $N = 1000$ of different experiments, each of them consisting of drawing n samples from a distribution over $\{0, 1\}$ with fixed parameters and estimating the parameters. Note that the shape of the theoretical and empirical distributions are similar.

Two trends are visible: First, as n grows, the number of possible values grows too. How many possible values can θ take for a given n? As a result, the distribution of the values of θ^{ML} "aproximates better and better" a continuous curve, that has the shape of a bell.

Second, the "bell" is centered on the true value of θ and becomes narrower as n increases. In other words, the distribution becomes more concentrated around the true value. Is this good or bad? Looking at it another way, this fact means that with high probability, θ^{ML} will be a good approximation of the true θ for large n.



(a)



(b)

Figure 4.2: The distribution of the count n_1 and of the ML estimate θ_1^{ML} for $n = 20$ **(a)** and $n = 100$ **(b)** trials from a distribution over $\{0, 1\}$ with $\theta_0 = 0.3$, $\theta_1 = 0.7$.

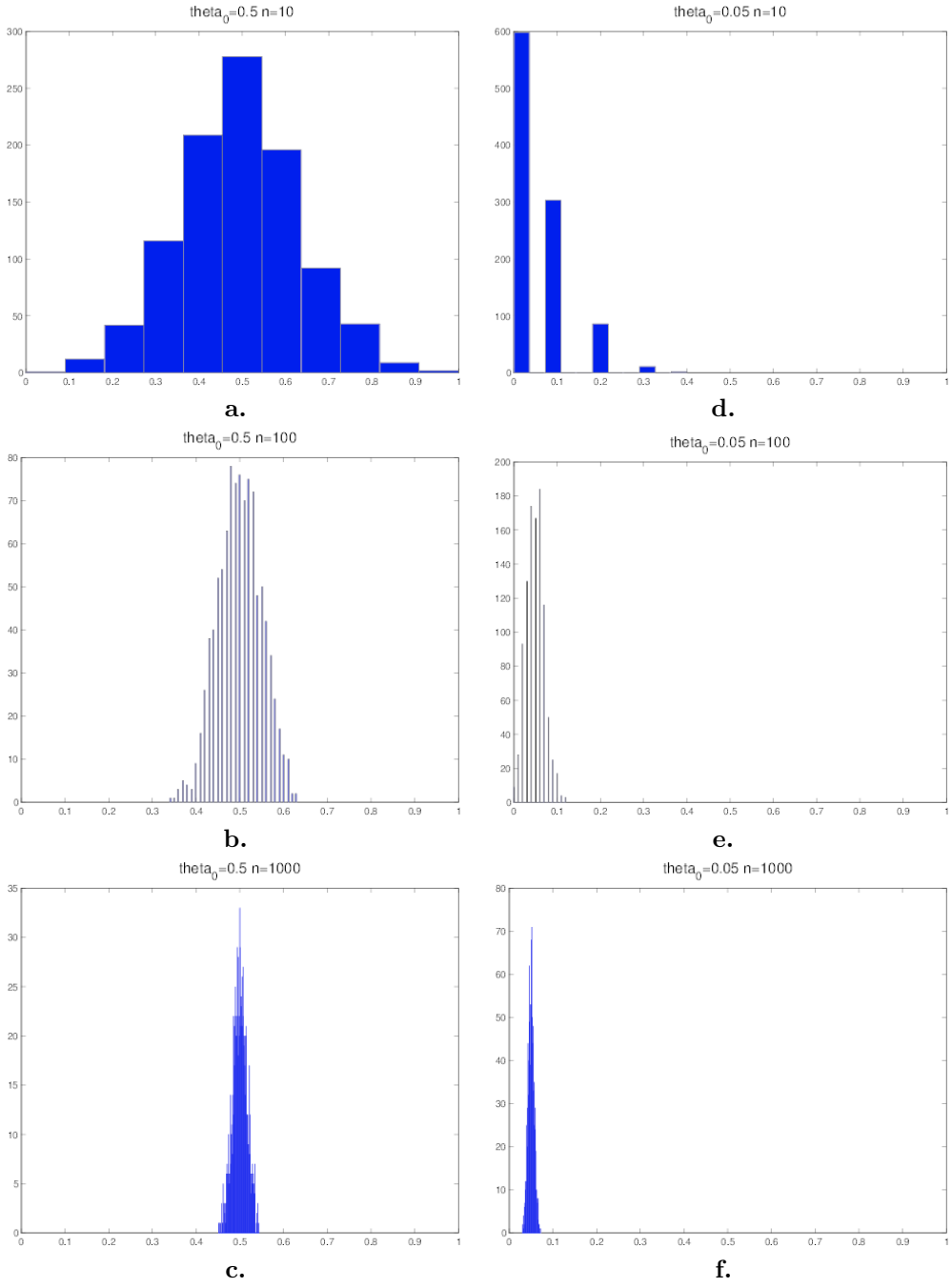


Figure 4.3: Bar graphs of the value of the ML estimate of θ_0 for a binary experiment comprising n identical independent trials. The values of n are 10, 100, 1000 and the values of θ_0 are 0.5 in **a, b, c** and 0.05 in **d, e f**. What are the similarities/differences between the plots for the two values of θ_0?

4.3 Confidence intervals

In the previous section we learned how to compute the probability that the ML estimate θ^{ML} is "near" the true value θ i.e $P[\theta^{ML} \in [\theta - \epsilon, \theta + \epsilon]$ for a given ϵ.

Statisticians often ask the reverse question: *What is ϵ so that $P[\theta^{ML} \in [\theta-\epsilon, \theta+\epsilon]$ is at least, say $p = 0.95$?* If we find such an ϵ, then the interval $[\theta - \epsilon, \theta + \epsilon]$ is called a **confidence interval** for **confidence level** p.

In general, the interval $[a, b]$ is a CI_p for a parameter θ if $P[\,\theta^{ML} \in [a,\, b]\,] \geq p$. The probabililty $P[\,\theta^{ML} \in [a,\, b]\,]$ has the following meaning: If we did many experiments of drawing random data sets of size n from the same distribution, and of estimating θ^{ML} for each of those data sets, at least a fraction p of times θ^{ML} will be contained in $[a, b]$.

It follows that the confidence interval *depends on the true (unknown) distribution* and on the size of the data set n. While n is known, the true distribution is not known (after all, this is why we are estimating its parameters). For now, we will discuss mathematical properties of confidence intervals *assuming that the true distribution is known.* Then, in the next section, we will discuss how get around the fact that the distribution is unknown.

4.3.1 Confidence intervals – the probability viewpoint

In this section we examine properties of confidence intervals, assuming that the true parameters of a distribution are known.

The confidence interval can be obtained numerically using the probability distribution of θ^{ML}. Assume $S = \{0, 1, \ldots m-1\}$, and that P is given by the (true) parameters $\theta_0, \ldots \theta_{m-1}$ as before. The distribution of θ_j^{ML} is

$$P[\theta_j^{ML} = \frac{n_j}{n}] \;=\; P(n_j) \;=\; \binom{n}{n_j} \theta_j^{n_j} (1-\theta_j)^{n-n_j} \tag{4.29}$$

Let $\delta = (1-p)/2$. Then, an algorithm for computing the CI_p is

$i \leftarrow 0$
$q \leftarrow 0$
while $q < \delta$ do
    $q \leftarrow q + P[n_j = i]$
    $i \leftarrow i + 1$
$a \leftarrow (i-1)/n$

$i \leftarrow n$
$q \leftarrow 0$
while $q < \delta$ do
    $q \leftarrow q + P[n_j = i]$
    $i \leftarrow i - 1$
$b \leftarrow (i+1)/n$

output interval $[a, b]$

This algorithm finds an interval $[a, b]$ such that $P[\theta_j^{ML} < a], P[\theta_j^{ML} > b] < \delta$.

The confidence interval is not unique. For once, if $[a, b]$ is CI_p then any interval $[a', b']$ that contains $[a, b]$ is also a CI_p. Indeed, if $a' \leq a \leq b \leq b'$ we have

$$P[\,\theta^{ML} \in [a',\, b']\,] \geq P[\,\theta^{ML} \in [a,\, b]\,] \geq p \tag{4.30}$$

Also, if $p' < p$ and if $[a, b]$ is CI_p then $[a, b]$ is also $CI_{p'}$. Moreover, we can have intervals $[a, b]$, $[a', b']$ overlapping but not included in one another, which are both CI_p.

Computing confidence intervals by using the multinomial distribution of θ^{ML} is computationally intensive. In later chapters we shall learn an approximate but very convenient method to obtain confidence intervals, based on the Gaussian distribution.

4.3.2 Statistics with confidence intervals

Statisticians use confidence interval as an indicator of how "trustworthy" the estimated θ^{ML} is. The smaller a confidence interval (i.e the smaller the difference $b - a$), the more confident we are in the value of our estimate.

However, as we have seen before, the confidence interval depends on the true θ so we cannot compute it if we don't know θ! What is done then is to use the estimate θ^{ML} itself as if it was the true θ. If we replace our estimate θ_j^{ML} in the place of θ_j in the distribution (4.29) we can find an **estimated confidence interval**.

The estimated confidence interval has the following property that makes it useful in practice: Assume that we set p at some value. Then perform a mental experiment: Draw many data sets of size n from the true, unknown distribution. For each data set, estimate θ^{ML} and then calculate a confidence interval of confidence p, pretending that the current θ^{ML} is the true θ. This way we will get a lot of intervals $[a, b]$, one for each data set. One can prove that a fraction p or larger of these intervals contain the true θ.

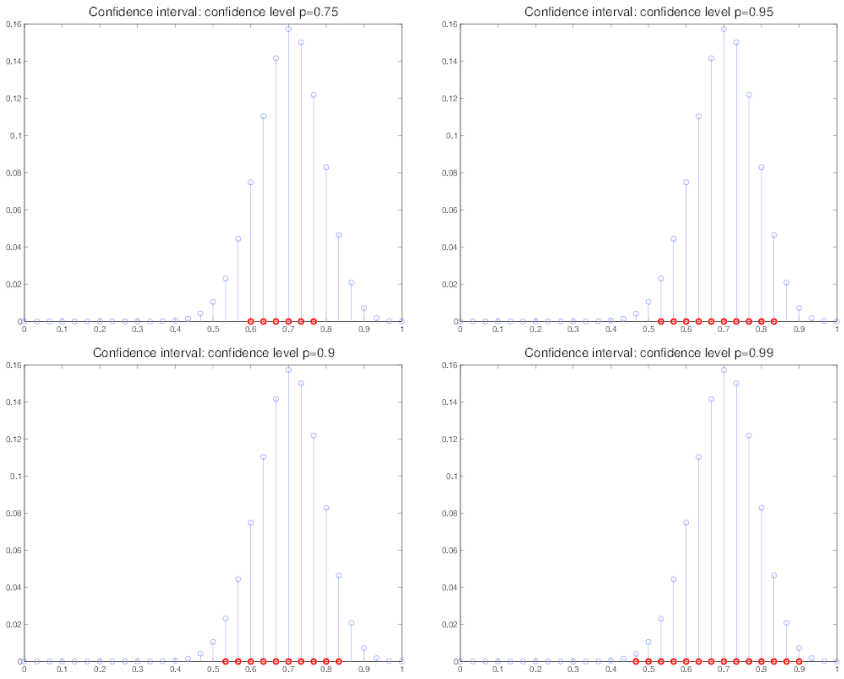


Figure 4.4: Confidence intervals for θ_1 for the distribution in example 4.4 for various confidence levels p

.

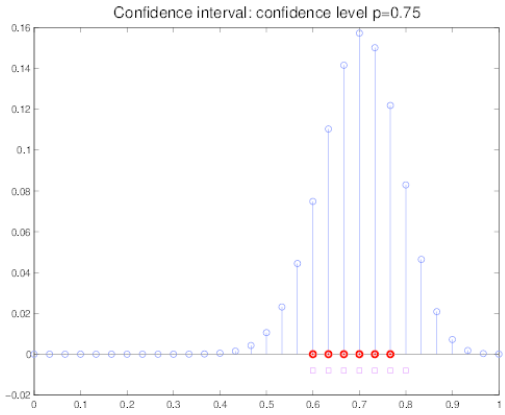


Figure 4.5: The confidence interval for a given p is not unique. Here we show confidence intervals computed by two different methods. The (magenta) squares mark the confidence interval computed by the method presented here, which results in $P[\theta_1^{ML} < a] \leq \delta$ and $P[\theta_1^{ML} \geq b] \leq \delta$. The (red) circles mark a a confidence interval computed by a different method, for which $a - \theta_1 = b - \theta_1$.

.

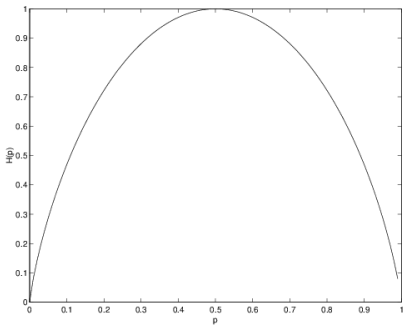


Figure 4.6: The entropy of the distribution representing the coin toss as a function of $p = P(1)$. Note that the entropy is symmetric around 0.5.

In summary, estimating a parameter θ_j from data takes the following steps.

1. Collect data $\mathcal{D} = \{x^{(1)}, \ldots x^{(n)}\}$
2. Compute sufficient statistics $n_0, n_1, \ldots n_{m-1}$
3. Compute the ML estimates $\theta_j^{ML} = n_j/n$
4. Choose a confidence level, say $p = 0.95$. For each $j = 0, \ldots m-1$ use θ_j^{ML} and distribution (4.29) to obtain a CI_p for the parameter θ_j.

4.4 Incursion in information theory

Return to the coin toss experiment. I toss the fair coin and tell you the result. How much information do you receive? By definition, one *bit*. Now I toss the biased coin ($p = 0.9$) and tell you the result again. Do you receive the same amount of information now? Uncertainty is the opposite of information. When I give you information, I remove some of your uncertainty. So when are you more uncertain about the outcome of the toss? At $p = 0.5$, $p = 0.9$ or $p = 0.999$? If $p = 1$ you are certain and the information you'd receive from me is 0. In information theory, uncertainty is measured by the *entropy*. The entropy of a distribution is the amount of randomness of that distribution. If x is an elementary event in S, then the entropy of a distribution P over S is

$$H(P) \;=\; -\sum_{x \in S} P(x) \log P(x) \tag{4.31}$$

Figure 4.6 plots the entropy of the distribution representing the coin toss as a function of p. The maximum entropy is reached at $p = 0.5$, corresponding to the uniform distribution. The entropy is 0 for $p = 0$ and $p = 1$ which are the

deterministic experiments. Note that H is always ≥ 0. The logarithms are in base 2.

Information theory also gives us a way of measuring the "distance" between two probability distributions. It is the **Kullbach-Leibler** (KL) divergence

$$D(P||Q) \;=\; \sum_{x\in S} P(x)\log\frac{P(x)}{Q(x)} \tag{4.32}$$

The KL divergence is 0 if and only if $P \equiv Q$, and positive otherwise. $D(.||.)$ is not a distance, since it is not symmetric and it does not obey the triangle inequality.

Example 4.5 *Let $S = \{0, 1\}$ and let P and Q be two distributions on S defined by $\theta_1^P = 0.5$, $\theta_1^Q = 0.2$.*

$$\begin{aligned} D(P||Q) &= P(0)\log\frac{P(0)}{Q(0)} + P(1)\log\frac{P(1)}{Q(1)} \;=\; \theta_0^P\log\frac{\theta_0^P}{\theta_0^Q} + \theta_1^P\log\frac{\theta_1^P}{\theta_1^Q} \;=\; 0.2231 \\ D(Q||P) &= Q(0)\log\frac{Q(0)}{P(0)} + Q(1)\log\frac{Q(1)}{P(1)} \;=\; 0.1927 \end{aligned}$$

4.4.1 KL divergence and log-likelihood

An interesting connection exists between KL divergence and likelihood. Let us rewrite the logarithm of the likelihood for an experiment with n trials:

$$\begin{aligned} l(\overline{\theta}) &= \log\theta_0^{n_0}\theta_1^{n_1}\ldots\theta_{m-1}^{n_{m-1}} && (4.33)\\ &= n_0\log\theta_0 + n_1\log\theta_1 + \ldots + n_{m-1}\log\theta_{m-1} && (4.34)\\ &= n\sum_{i=0}^{m-1}\frac{n_i}{n}\log\theta_i && (4.35)\\ &= n\sum_{i=0}^{m-1}\frac{n_i}{n}[\log\theta_i + \log\frac{n_i}{n} - \log\frac{n_i}{n}] && (4.36)\\ &= -n\sum_{i=0}^{m-1}\frac{n_i}{n}\log\frac{\frac{n_i}{n}}{\theta_i} - n\sum_{i=0}^{m-1}\frac{n_i}{n}\log\frac{n_i}{n} && (4.37)\\ &= n[-D(\hat{P}||P) + H(\hat{P})] && (4.38) \end{aligned}$$

In the above, we have denoted by P the distribution over S defined by the parameters $\overline{\theta}$ and by $\hat{P}$ the distribution defined by the parameters $\frac{n_i}{n}$, $i = 0, \ldots m-1$. Hence, we can rewrite the log-likelihood as

$$\frac{1}{n}l(\overline{\theta}) \;=\; -D(\hat{P}||P) + H(\hat{P}) \tag{4.39}$$

To maximize the log-likelihood we have to minimize the KL divergence on the r.h.s. because the other term does not depend on $\overline{\theta}$. But this KL divergence has a unique minimum of 0 for $P \equiv \hat{P}$, i.e for

$$\theta_i \;=\; \frac{n_i}{n} \quad \text{for } i \;=\; 0, \ldots\, m-1 \tag{4.40}$$

Chapter 5

Continuous Sample Spaces

Here we study probability distributions whose set of outcomes S is the real line. While a discrete distribution can be directly defined by its values for all elements of S, the case of *continuous* distribution this approach breaks: For example, the probability of an individual point on the real line is zero (almost everywhere) but the probability of an interval containing only zero-probability points is usually non-zero. Hence, distributions on (subsets of) the real line require a different approach.

5.1 The cumulative distribution function and the density

The cumulative distribution function (CDF) corresponding to a distribution P is defined by

$$F(x) \;=\; P(X \leq x) \tag{5.1}$$

In the above, X is a random sample from P, while x, the argument of F is a given point on the real line.

The following properties of F are easy to derive:

1. $F \;\geq\; 0$ positivity.
2. $\lim_{x \to -\infty} F = 0$
3. $\lim_{x \to \infty} F = 1$
4. F is an increasing function

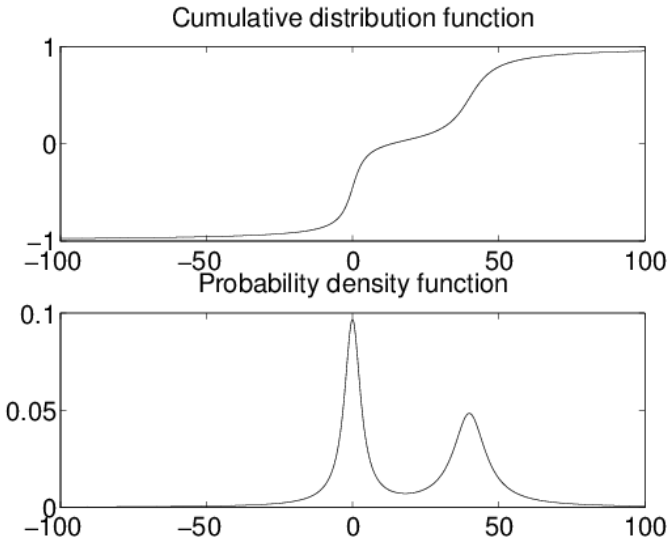


Figure 5.1: A distribution over the real numbers, given by its cumulative distribution function F and by its probability density f. Note that the maxima of the density correspond to the steepest points in F.

5. $P((a,b]) = F(b) - F(a)$; the probability of an interval is the increase of F between its limits.

This last property is useful because having the probability of an interval allows us to compute the probability of any set of interest (read Lebesgue measurable set if you are steeped in measure theory) from F. Hence F is sufficient to determine P.

In the following we shall assume that F is continuous and differentiable. Its derivative w.r.t x is called the *probability density function* or shortly the *density*[1].

$$f = \frac{dF}{dx} \tag{5.2}$$

By Newton's formula we have

$$P(a,b) = P[a,b] = F(b) - F(a) = \int_a^b f(x)dx \tag{5.3}$$

[1]For you measure theorists out there, P is a measure, $F(x)$ is the measure of $(-\infty, x]$ and f is the Radon-Nikodym derivative of P w.r.t to the Lebesgue measure.

and the *normalization condition*

$$\int_{-\infty}^{\infty} f(x)dx \;=\; 1 \tag{5.4}$$

The distribution P can be defined just as well by the density as by the cumulative distribution function. Throughout this course, we shall find it more convenient to use almost exclusively the former.

Any function f defined on the real numbers can be a density, provided it satisfies the following conditions:

1. $f(x) \;\geq\; 0$ for all x (**positivity**)
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (**normalization**)

Note that: **a.** If f is not defined over all real numbers we can extend it by giving it value 0 everywhere where it is not otherwise defined. For example

$$f(x) \;=\; \begin{cases} \sqrt{x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

b. If f integrates to a finite number $K \neq 1$ (and $K > 0$), then we can rescale f to make it integrate to 1.

$$\int_{-\infty}^{\infty} f(x)dx \;=\; K \;>\; 0 \quad\Rightarrow\quad \int_{-\infty}^{\infty} \frac{1}{K} f(x)dx \;=\; 1$$

This operation is called **normalization**.

Example 5.1

$f(x) \;=\; \sin x$ for $x \in (-\infty, \infty)$	**is not** a density because it can take negative values
$f(x) \;=\; 1 + \sin x$ for $x \in (-\infty, \infty)$	**is not** a density because its integral is infinite
$f(x) \;=\; 1 + \sin x$ for $x \in [0, \pi]$ and 0 otherwise	**is not** a density because its integral is $\pi + 2 \neq 1$
$f(x) \;=\; \frac{1+\sin x}{\pi+2}$ for $x \in [0, \pi]$ and 0 otherwise	**is** a density

5.2 Popular examples of continuous distributions

Example 5.2 The uniform distribution *over a given interval $[a, b]$, has a density that's constant inside $[a, b]$ and 0 outside.*

$$f(x) \;=\; \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

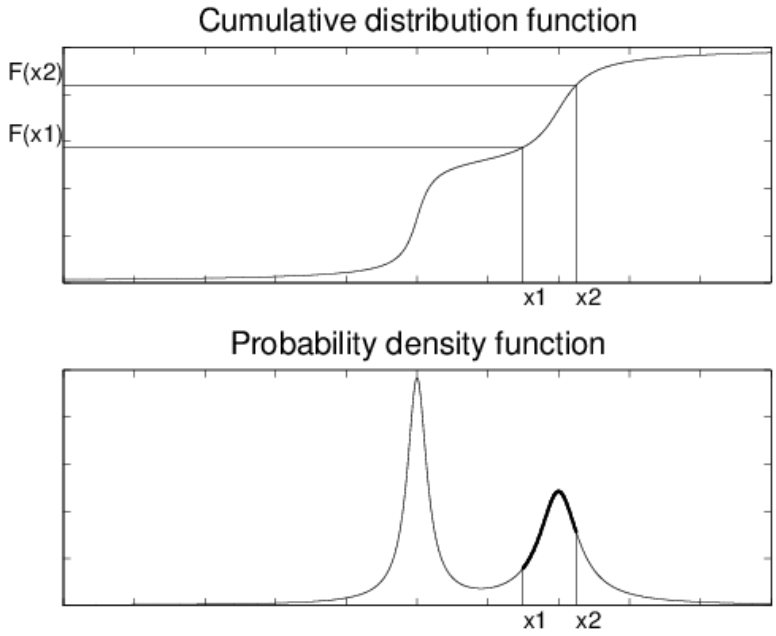


Figure 5.2: Computing the probability of an interval. The probability of $[x1, x2]$ is equal to the difference $F(x2) - F(x1)$ in the plot above, but it is also equal to the integral of f from $x1$ to $x2$ in the plot below.

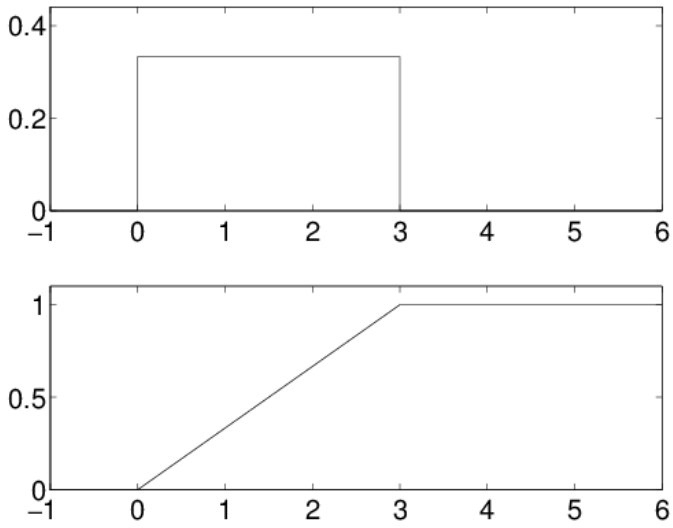


Figure 5.3: The uniform distribution on interval [0,3] represented as density (above) and as cumulative distribution function (below).

Similarly, one can define uniform distributions over $[a, b), (a, b), (a, b]$. The number $b-a$ represents the normalization constant of the uniform distribution. The uniform density and its cumulative distribution function are shown in figure 5.3.

Under a uniform distribution, the probability of an interval contained in $[a, b]$ is proportional to its length. To see this, assume first that $[a, b] = [0, 1]$ the unit interval and that $0 \leq c \leq d \leq 1$, that is, $[c, d] \subseteq [0, 1]$. Then $P[c, d] = F(d) - F(c) = d - c$ and the result is proved. **Exercise** Generalize this to any other interval $[a, b]$.

Example 5.3 *The normal (Gaussian) distribution.*

This is the "bell curve", probably the most important (and famous) of all distributions. There are innumerable real-world phenomena which behave according to the normal distribution (sometime later in this course we shall see one reason why). Can you give an example? Moreover, the Gaussian is popular also with the mathematicians (this is no surprize:) who were charmed by its many nice properties and are still discovering more of them. The normal distribution is given by

$$f(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.6}$$

It has two parameters, μ and σ called respectively the *mean* and *standard deviation* of the distribution. We shall see later what is the significance of these names. For now we can notice that the parameter σ controls the "spread" of the distribution, while μ controls the position of the maximum. Indeed, it is easy to show that f has a unique maximum at $x = \mu$ and that it is symmetric around this value. Figure 5.4 plots normal distributions with different means and standard deviations.

The cumulative distribution function of the Gaussian cannot be written in closed form. It is called G, or the error function, and its values for $\mu = 0$ and $\sigma = 1$ are tabulated.

Example 5.4 *The (continuous) exponential distribution* $F(x) = 1 - e^{-\gamma x}$, $f(x) = \gamma e^{-\gamma x}$.

5.3 Another worked out example

Assume that components are produced which have a parameter x with nominal value is $x = a$. However, due to process errors, the distribution of x is described

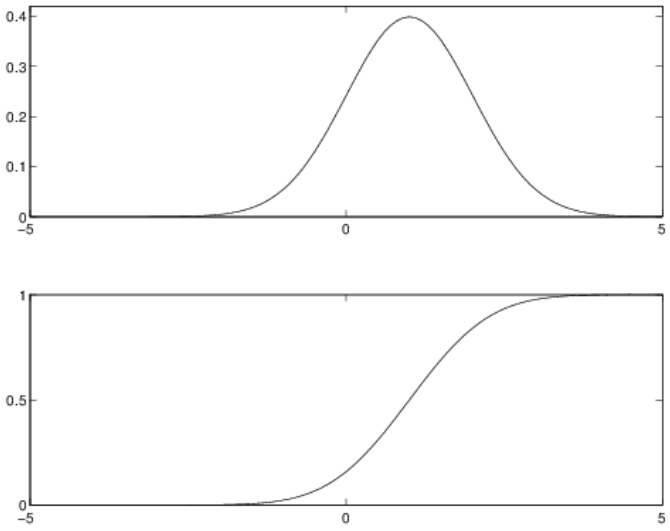


Figure 5.4: The normal distribution with $\mu = 1$ and $\sigma = 1$: the density (above) and the cumulative distribution function (below).

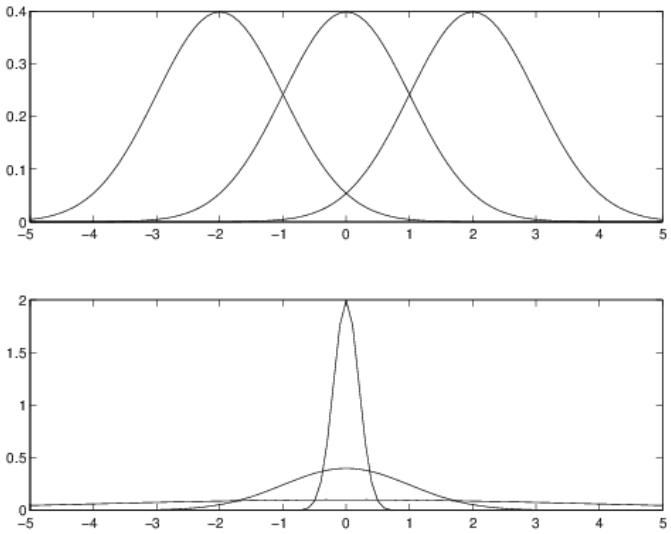


Figure 5.5: Examples of normal distributions. Above: $\sigma = 1,\ \mu = -2, 0, +2$; below: $\mu = 0,\ \sigma = 0.2, 1, 4$. The peak of each distribution is located at μ and the spread increases with σ.

by the density

$$f(x) \;=\; \begin{cases} \frac{hx}{a}, & x \in [0,a] \\ \frac{h(2a-x)}{a}, & x \in (a,2a] \\ 0, & \text{otherwise} \end{cases}$$

1. What is the value of h so that the above function is a proper probability density function?

Solution The function f needs to be ≥ 0, which is easily verified, and to satisfy the normalization condition

$$\int_0^{2a} f(x) \;=\; 1$$

To find h, we compute the area under f and equal it with 1. The area is a triangle, whose area is $2ah/2 = ah$. Therefore

$$1 \;=\; ah \quad \Longrightarrow \quad h \;=\; \frac{1}{a}$$

and

$$f(x) \;=\; \begin{cases} \frac{x}{a^2}, & x \in [0,a] \\ \frac{2a-x}{a^2}, & x \in (a,2a] \\ 0, & \text{otherwise} \end{cases}$$

2. What is the CDF of this distribution?

Solution $F(t) = P(x \leq t)$

For $t \leq a$:

$$F(t) \;=\; \int_0^t \frac{x}{a^2} dx \;=\; \frac{x^2}{2a^2}\Big|_0^t \;=\; \frac{t^2}{2a^2}$$

For $a < t \leq 2a$:

$$\begin{aligned} F(t) &= \int_0^t f(x) \\ &= \frac{1}{2} + \int_a^t \frac{2a-x}{a^2} dx \\ &= \frac{1}{2} + \left[\frac{2ax}{a^2} - \frac{x^2}{2a^2} \right]_a^t \\ &= \frac{1}{2} + \left[\frac{2a(t-a)}{a^2} - \frac{t^2-a^2}{2a^2} \right]_a^t \\ &= \frac{1}{2} + \frac{4at - 4a^2 - t^2 + a^2}{2a^2} \\ &= 1 - \frac{(2a-t)^2}{2a^2} \end{aligned}$$

3. Compute the probabilities: $P(x \le a)$, $P(x \le a/2)$, $p(x \ge 3a/2)$.

Solution By the symmetry of the triangle, $P(x \le a) = P(x \ge a) = 1/2$. Or, we can use the CDF to obtain the same result; $P(x \le a) = F(a) = t^2/2\big|_{t=1} = 1/2$.

$$\begin{aligned} P(x \le a/2) &= \int_0^{a/2} f(x)dx \;=\; \left[\frac{x^2}{2a^2}\right]_0^{a/2} \;=\; \frac{1}{8} \\ &= F(a/2) \;=\; \frac{t^2}{2a^2}\bigg|_{t=a/2} \;=\; \frac{1}{8} \\ &= \left(\frac{1}{2}\right)^2 P(x \le a) \;=\; \frac{1}{8} \quad \text{geometrically: by triangle similarity} \\ P(x \ge 3a/2) &= 1 - P(x < 3a/2) \;=\; 1 - F(3a/2) \;=\; \frac{1}{8} \\ &= P(x \le a/2) \;=\; \frac{1}{8} \quad \text{geometrically: by symmetry} \end{aligned}$$

This also shows that $1/8+1/8$ of the components are $a/2$ away or more from the nominal value. If we call these components "bad" and the others "good", we can conclude that in this process the probability of producing a good component is $p_1 = 1 - 1/4 = 3/4$.

4. Let us compare this distribution with a normal distribution $g(x)$ with $\mu = a, \sigma^2 = 1$. We will choose a so that the two distribution have the same maximum height. We want to establish for which distribution the probability of producing a good component is higher.

Solution We first need to find a. We have

$$g(x) \;=\; \frac{1}{\sqrt{2\pi}} e^{(x-a)^2} 2$$

and therefore $g(a) = 1/\sqrt{2\pi}$. If we require $g(a) = f(a)$ it follows that $h = 1/a = 1/\sqrt{2\pi}$ and $a = \sqrt{2\pi}$.

Now we want to find the probability $P_g(x \le a/2) = P_g(x \le \sqrt{2\pi}/2)$. Recall that g is a normal distribution centered at μ. Tables give us the CDF of a normal distribution centered at 0. Therefore, we change the variable x to $x' = x - a$. Now our question becomes: what is $P(x' \le a/2 - a = -\sqrt{\pi/2})$ under a standard normal distribution? We can find this from the table, looking for the value $\sqrt{\pi/2} = 1.2533$. We find that the corresponding probability is 0.1050.

Finally, we can compute the probability $P_g(|x - a| \le a/2) = 1 - 2 \times 0.1050 = 0.79 = p_2$.

Comparing the $p_1 = 0.75$ with p_2 we see that the normal distribution is more likely to produce a good component.

5. Quantiles Sometimes we are interested in the reverse question: What is t, so that $P(x \leq t)$ equals some value α? Typically, $\alpha = \frac{1}{4}, \frac{3}{4}, 0.1, 0.2, \ldots 0.9$. The corresponding t values, denoted $x_{\frac{1}{4}}, x_{\frac{3}{4}}, x_{0.1}, x_{0.2}, \ldots$ are called *quartiles, deciles,* or more generally *quantiles.*

We will evaluate the quartile $x_{\frac{1}{4}}$.

Solution We have to solve the equation

$$F(x_{1/4}) \;=\; 1/4$$

or equivalently

$$\frac{x^2}{2a^2} \;=\; \frac{1}{4} \quad\Longrightarrow\quad x_{1/4} \;=\; \frac{a}{\sqrt{2}}$$

5.4 Sampling from a continuous distribution

Assume that we have a system-supplied function `rand()` that generates random numbers uniformly distributed in the interval [0, 1]. We want to sample from another continuous density F using `rand()`.

Here is a method:

1. call `rand()` to produce $u \in [0, 1]$
2. output $x = F^{-1}(u)$

Then x's CDF will be F.

Why it works: Assume that the method above produces numbers distributed according to a CDF $\tilde{F}$. We will show that $\tilde{F} = F$. Take an arbitrary $x_0 \in (-\infty, \infty)$.

$$\begin{aligned}
\tilde{F}(x_0) &= P(x \leq x_0) \quad \text{(by definition)} \\
&= P(F(x) \leq F(x_0)) \quad \text{(because } F \text{ is increasing)} \\
&= P(u \leq F(x_0)) \quad \text{(because } u = F(x)) \\
&= F(x_0) \quad \text{(because } u \text{ is uniformly distributed)}
\end{aligned}$$

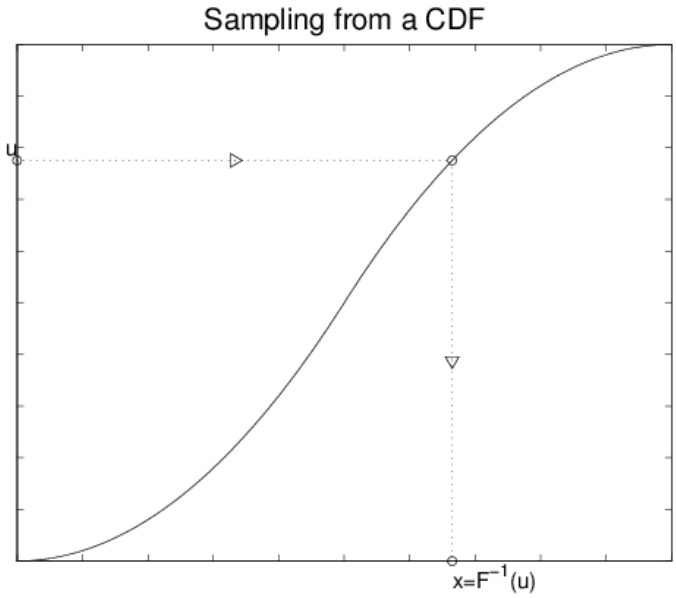


Figure 5.6: Sampling from an arbitrary F.

5.5 Discrete distributions on the real line

We shall now give a unified view of distributions over subsets of the real line, be they discrete (like the Bernoulli) or continuous (like the normal). Let us start with the example of the die roll, whose outcomes set is $S = \{1, 2, 3, 4, 5, 6\}$ a finite subset of $(-\infty, \infty)$.

What is the cumulative distribution function F for this distribution? Applying the definition (5.1) we obtain

$$F(x) \;=\; \begin{cases} 0 & x < 1 \\ \frac{1}{6} & 1 \le x < 2 \\ \frac{2}{6} & 2 \le x < 3 \\ \frac{3}{6} & 3 \le x < 4 \\ \frac{4}{6} & 4 \le x < 5 \\ \frac{5}{6} & 5 \le x < 6 \\ 1 & x \ge 6 \end{cases} \tag{5.7}$$

Hence the cumulative distribution function exists, but is discontinuous. How about the density f? Obviously, f is zero in all points but $1, 2, 3, 4, 5, 6$. In those points, its value must be "infinity". More precisely, we say that f for the die roll is

$$f(x) \;=\; \frac{1}{6}\delta(x-1) + \frac{1}{6}\delta(x-2) + \;\ldots\; \frac{1}{6}\delta(x-6) \tag{5.8}$$

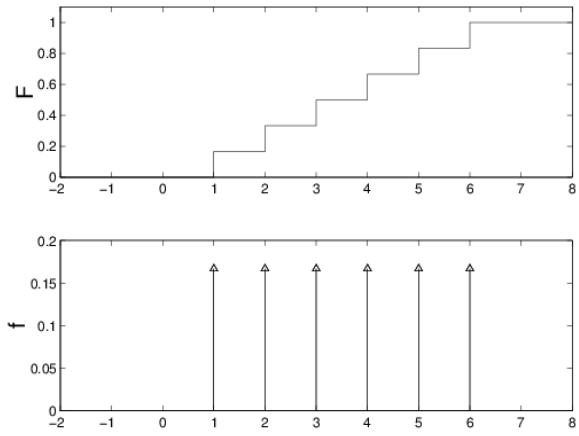


Figure 5.7: The discrete distribution corresponding to the die roll. Its cumulative distribution function F (above) has "steps" of size $\frac{1}{6}$ at the points $1, \ldots 6$; the density f has δ "spikes" at the same points and is 0 otherwise.

The symbol $\delta()$ called Dirac's "function" is defined by

$$\delta(x) \;=\; \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases} \tag{5.9}$$

with

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{5.10}$$

In addition we have the more general relationship holding for any function g defined on the real line

$$\int_{-\infty}^{\infty} \delta(x)g(x)dx = g(0) \tag{5.11}$$

(Parenthesis: Mathematically speaking, Dirac's funtion is not a function but a functional and the last formula above defines it by the values it takes when applied to a real function.)

In general, the density of a discrete distribution over $S = \{x_0, \ldots x_{m-1}\} \subseteq (-\infty, \infty)$ is given by

$$f_x(x) \;=\; \sum_{i=0}^{m-1} P(x_i)\delta(x - x_i) \tag{5.12}$$

Chapter 6

Parametric density estimation

6.1 Parametrized families of functions

In parametric density estimation, we assume that the density to be estimated belongs to a *parametrized family* of densities $\mathcal{F} = \{f_\theta,\ \theta \in \Theta\}$. The following are examples of parametrized families of densities.

- the family of all *uniform* distributions over a closed interval

$$\mathcal{F}_1 \ = \ \{u_{[a,b]},\ a < b\} \tag{6.1}$$

  Under this distribution all outcomes are equally possible. It is sometimes called an **uninformative** distribution, because it gives no outcome a higher preference. $\mathcal{F}_1$ has two parameters a and b. The domain Θ is the half-plane $a < b$.

- the family of all **normal** distributions, parametrized by μ and σ^2; here Θ is the half-plane $(\mu, \sigma^2),\ \sigma^2 > 0$.

$$\mathcal{F}_2 \ = \ \{N(.; \mu, \sigma^2)\} \tag{6.2}$$

  This distribution is the most famous of all. Many natural and social phenomena are well described by this law. Besides, it has compelling mathematical properties which make it a focus point for much of statitics.

- the family of **logistic** cumulative distribution functions (CDF's) given by

$$F(x; a, b) \ = \ \frac{1}{1 + e^{-ax-b}},\ \ a > 0 \tag{6.3}$$

The density that corresponds to this cumulative distribution function is

$$f(x;a,b) \;=\; \frac{ae^{-ax-b}}{(1+e^{-ax-b})^2} \tag{6.4}$$

and the family $\mathcal{F}_3$ is the set of all densities of the form (6.4) with parameters (a,b) belonging to the right half-plane.

The logistic distribution has been used to describe growth phenomena. It is also very useful in classification (chapter 17).

- the family $\mathcal{F}_4$ of exponential distributions parametrized by $\lambda > 0$

$$f(x;\lambda) \;=\; \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{6.5}$$

The **exponential** distribution is used heavily in reliability (the probabilistic theory of how to optimally schedule diagnosis and maintenance for components and systems) to describe occurence of failures, times to fix bugs, component lifetimes. It is also used to describe concentration of pollutants in environmetrics, in physics to model radioactive decay, and so on.

- any subset of one of the above families (e.g all the normal distributions with $\sigma > 1$)
- the union of some of the above families (e.g $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$)

6.2 ML density estimation

Like in non-parametric density estimation, the objective is to estimate f_θ from a given set of data points $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$. Since the form of f is defined up to the parameter (vector) θ, the problem is in fact equivalent to estimating θ. In the framework of ML, we want to find

$$\hat{\theta} \;=\; \operatorname*{argmax}_{\theta \in \Theta} L(\theta) \tag{6.6}$$

where

$$L(\theta) \equiv L(f(.;\theta)) = \prod_{i=1}^{n} f(x_i;\theta). \tag{6.7}$$

Example 6.1 *Estimating a uniform density*

Let $\mathcal{F}$ be the family of all uniform densities over an interval $[a,b]$ of the real line. The likelihood is then

$$L(a,b) \;=\; \begin{cases} \frac{1}{(b-a)^n}, & \text{for } a \leq x_i \leq b \text{ for all } i \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

From this we can easily derive the ML estimates of the parameters:

$$a = \min_i x_i \tag{6.9}$$

$$b = \max_i x_i \tag{6.10}$$

The uniform family is a unusual example, in the sense that the likelihood (and u) are not smooth functions of the data (a small, even infinitezimal change in the data can induce a lage change in the likelihood). I included it more like a curiosity than as a situation you are likely to encounter in practice. The next examples have more to do with real estimation problems.

6.2.1 Estimating the parameters of a normal density

The normal density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.11}$$

Because of the exponential form of the distribution, we will find it more convienient to work with the log-likelihood l.

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \left[-\frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln 2\pi - \frac{(x_i-\mu)^2}{2\sigma^2} \right] \tag{6.12}$$

$$= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2 \right] \tag{6.13}$$

To find the maximum of this expression, we equate its partial derivatives w.r.t μ and σ^2 with 0.

$$\frac{\partial l}{\partial \mu} = \frac{2n\mu}{2\sigma^2} - \frac{2\sum_{i=1}^{n} x_i}{2\sigma^2} = 0 \tag{6.14}$$

From this equation we obtain

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{6.15}$$

Hence, the mean μ is equal to the arithmetic mean of the data. It is also convenient that μ can be estimated from the data independently from σ^2.

Now, let us take the partial derivative w.r.t σ^2:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 0 \tag{6.16}$$

This entails

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{6.17}$$

In other words, the variance σ^2 is the arithmetic mean of the squared deviations of the data from the estimated mean $\hat{\mu}$. Note that an alternative formula for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \hat{\mu}^2 \tag{6.18}$$

Now we also see that for the purpose of parameter estimation, the data are summarized by the **sufficient statistics** $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$.

6.2.2 Estimating the parameters of an exponential density

$$l(\lambda) = \sum_{i=1}^{n}(\log\lambda - \lambda x_i) \tag{6.19}$$

$$= n\log\lambda - \lambda\sum_{i=1}^{n}x_i \tag{6.20}$$

Taking the derivative we obtain:

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}x_i \tag{6.21}$$

Then, solving $\frac{\partial l}{\partial \lambda} = 0$ we obtain

$$\frac{1}{\lambda^{ML}} = \frac{\sum_{i=1}^{n}x_i}{n} \tag{6.22}$$

$$\lambda^{ML} = \frac{n}{\sum_{i=1}^{n}x_i} \tag{6.23}$$

Note that this density, too, has a **sufficient statistic**: it is $\frac{1}{n}\sum_{i=1}^{n}x_i$ the arithmetic mean of the observations. The parameter λ is inversely proportional to the sufficient statistic, which suggests that if x is measured in time units (for example seconds), then λ would be measured in inverse time units (for example s^{-1}). For this reason, λ is called the **rate** of the exponential distribution.

6.2.3 Iterative parameter estimation

Let us now estimate the ML parameters of the third family of functions, the logistic CDF's. The expression of the log-likelihood is:

$$l(a,b) = n\ln a - a\sum_{i}x_i - nb - 2\sum_{i=1}^{n}\ln(1+e^{-ax_i-b}) \tag{6.24}$$

and its partial derivatives w.r.t a, b are

$$\frac{\partial l}{\partial a} = \frac{n}{a} - \sum_{i=1}^{n} x_i \frac{1 - e^{-ax_i - b}}{1 + e^{-ax_i - b}} = 0 \tag{6.25}$$

$$\frac{\partial l}{\partial b} = -\sum_{i=1}^{n} \frac{1 - e^{-ax_i - b}}{1 + e^{-ax_i - b}} = 0 \tag{6.26}$$

The above system of equation cannot be solved analytically. We'll have to settle for a numerical solution, obtained iteratively by *gradient ascent*. We shall start with an initial estimate (aka guess) of a, b and update this estimate iteratively as follows:

$$a \leftarrow a + \eta \frac{\partial l}{\partial a} \tag{6.27}$$

$$b \leftarrow b + \eta \frac{\partial l}{\partial b} \tag{6.28}$$

As the gradient $(\frac{\partial l}{\partial a}, \frac{\partial l}{\partial b})$ approaches zero, this iteration is guaranteed to converge to a (local) maximum of the log-likelihood. This is a serious problem with iterative parameter estimation in general, and one that hasn't been satisfactory solved yet.

Figure 6.1 shows the gradient ascent iteration for a data set of 100 samples, in two situations. Both iterations are run for 100 steps, starting from the initial point $a = 1, b = 0$. The first situation corresponds to a *step size* $\eta = .01$. The iteration converges to $a = 3.5, b = -1.9$, the log-likelihood attains a maximum of -71.6 and the gradient itself converges to 0. The final density is plotted with continuous line vs. the true density. They do not coincide exactly: the true density is a normal density and the best estimate by a logistic CDF has a slightly different shape. In the second situation, the step size $\eta = 0.1$. This is too large a step size, causing the iteration to oscillate; neither the parameters nor the likelihood or the gradients converge. The resulting estimate is disastruous.

Appendix: The likelihood function for the logistic density has no local optima

In section 6.2.3 we computed the gradient of the logistic density and showed how the likelihood $l(a, b)$ can be maximized by gradient ascent. In general, the gradient ascent method, being a **greedy** method, converges to a **local maximum** of the likelihood function. Here we show that this is not a problem for the logistic density, as the likelihood in this case has a unique optimum. More precisely, we will show that the **Hessian matrix** of $l(a, b)$, i.e the matrix of second derivatives of l w.r.t the parameters a, b, is negative definite. When a function has a negative definite Hessian, then the function is concave, and it can have at most one maximum.

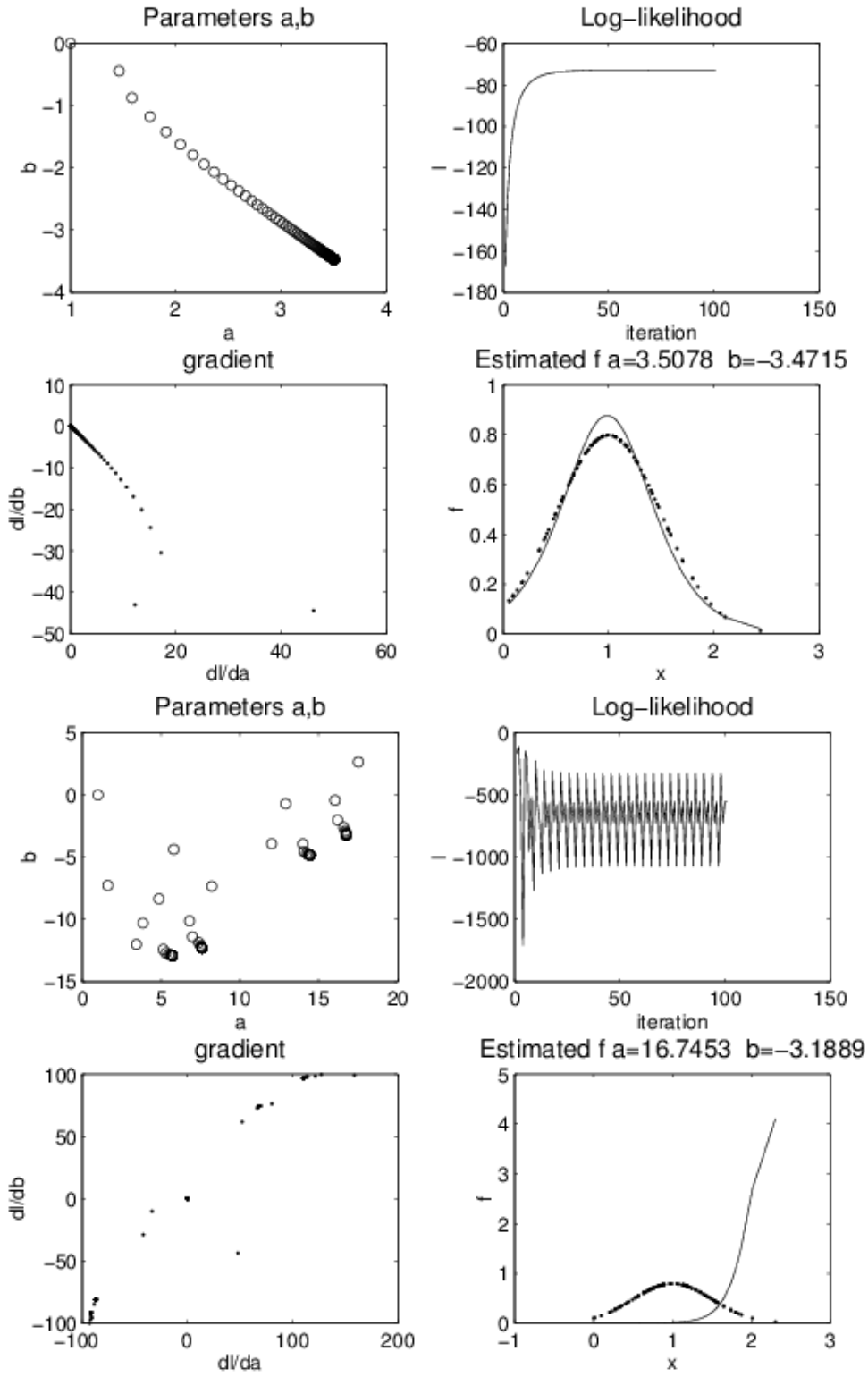


Figure 6.1: Two examples of gradient ascent on the same problem. The data are 100 points generated from a normal distribution. The step size η is 0.01 in the top 4 plots and 0.1 in the bottom set of plots. The estimated f is shown with continuous line, the true f is in dotted line.

We start by computing the second derivatives of l. To simplify notation, let $e_i \leftarrow e^{-ax_i-b}$.

$$\frac{\partial}{\partial a}\left(\frac{1-e_i}{1+e_i}\right) = \frac{2xe_i}{(1+e_i)^2} \tag{6.29}$$

$$\frac{\partial}{\partial b}\left(\frac{1-e_i}{1+e_i}\right) = \frac{2e_i}{(1+e_i)^2} \tag{6.30}$$

$$\frac{\partial^2 l}{\partial a^2} = -\frac{n}{a^2} - \sum_{i=1}^{n} 2x_i^2 \frac{e_i}{(1+e_i)^2} \tag{6.31}$$

$$\frac{\partial^2 l}{\partial b^2} = -\sum_{i=1}^{n} \frac{2e_i}{(1+e_i)^2} \tag{6.32}$$

$$\frac{\partial^2 l}{\partial a \partial b} = -\sum_{i=1}^{n} \frac{2x_i e_i}{(1+e_i)^2} \tag{6.33}$$

The Hessian is the symmetric matrix

$$H = \begin{bmatrix} \frac{\partial^2 l}{\partial a^2} & \frac{\partial^2 l}{\partial a \partial b} \\ \\ \frac{\partial^2 l}{\partial a \partial b} & \frac{\partial^2 l}{\partial b^2} \end{bmatrix} \tag{6.34}$$

A matrix H is negative definite when $-H$ is positive definite. To prove that some 2×2 matrix is A positive definite, we need to show that the determinant $\det A > 0$ and that the two diagonal elements are also positive.

Note that both $\frac{\partial^2 l}{\partial a^2}$, $\frac{\partial^2 l}{\partial b^2} \leq 0$. Therefore, to prove that the matrix H is negative definite, all we need to prove is that the determinant $\det(-H) > 0$ for all $a > 0$ and all b.

To simplify matters again, we denote $y_i = \frac{e_i}{(1+e_i)^2} > 0$ and hence

$$\det(-H) = \frac{\partial^2 l}{\partial a^2}\frac{\partial^2 l}{\partial b^2} - \left(\frac{\partial^2 l}{\partial a \partial b}\right)^2 \tag{6.35}$$

$$= \left(\frac{n}{a^2} + \sum_i 2x_i^2 y_i\right)\left(\sum_i 2y_i\right) - \left(\sum_i 2x_i y_i\right)^2 \tag{6.36}$$

$$= \frac{2n}{a^2}\sum_i y_i + 4\sum_i x_i^2 y_i \sum_i y_i - 4(\sum_i x_i y_i)^2 \tag{6.37}$$

To process the last term, let us first prove the follwing identity:

$$\begin{aligned}
(\sum_i a_i^2)(\sum_j b_j^2) - (\sum_i a_i b_i)^2 \; &= \\
&= \sum_i \sum_j a_i^2 b_j^2 - (\sum_i a_i^2 b_i^2 + 2\sum_{i<j} a_i b_i a_j b_j) \qquad (6.38) \\
&= 2\sum_{i<j} a_i^2 b_j^2 - 2\sum_{i<j} a_i b_i a_j b_j \qquad (6.39) \\
&= \sum_{i<j} (a_i b_j - a_j b_i)^2 \qquad (6.40)
\end{aligned}$$

Set now $a_i^2 = x_i^2 y_i,\ b_i^2 = y_i$ (which implies $x_i y_i = a_i b_i$) and we obtain

$$\det(-H) \;=\; \frac{2n}{a^2}\sum_i y_i + 4\sum_{i<j}(a_i b_j - a_j b_i)^2 \;> 0 \qquad (6.41)$$

This proves that the log likelihood $l(a, b)$ has a negative definite Hessian everywhere and therefore it can have at most one maximum.

6.3 The bootstrap

T.B.D.

Chapter 7

Non-parametric Density Estimation

The objective is to estimate a probability density f_X over the real line (or a subset thereof) from a set of points $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$.

7.1 ML density estimation

The likelihood of $\mathcal{D}$ is

$$L(f_X|\mathcal{D}) \;=\; \prod_{i=1}^{n} f_X(x_i) \tag{7.1}$$

and the log-likelihood

$$l(f_X|\mathcal{D}) \;=\; \sum_{i=1}^{n} \log f_X(x_i) \tag{7.2}$$

Maximizing the above over all functions yields (without proof)

$$\hat{f}_X^{ML} \;=\; \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \tag{7.3}$$

where $\delta_{\bar{x}}$ is the Dirac "function"

$$\delta_{\bar{x}} \;=\; \begin{cases} \infty & \text{for } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

By convention

$$\int_{-\infty}^{\infty} \delta_x(t)dt = 1 \tag{7.5}$$

$$\int_{-\infty}^{\infty} \delta_x(t)g(t)dt = g(x) \tag{7.6}$$

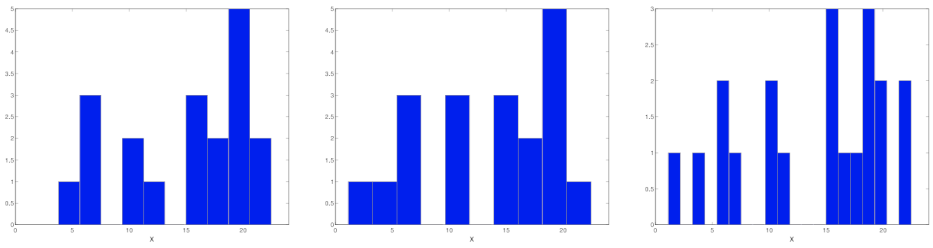
Hence, the ML estimate of f is a weighted sum of δ spikes placed at the sampled points. Such an estimate is counterintuitive - we know that most densities aren't spikes! It is also completely impractical: if we used the model $\hat{f}_X$ for prediction then we would predict that all the future samples from f_X will lie at the locations $x_1, x_2, \ldots x_n$ and nowhere else!

Therefore, instead of maximizing the likelihood over all possible density functions we will impose some restrictions corresponding to our intuition of a "realistic" density. One way to do that is to decide on a model class (e.g uniform, normal) and find the ML estimate in that class. This is called *parametric* density estimation. The alternative is the *non-parametric* way. We will study two non-parametric models: the *histogram* and the *kernel density estimator*.

7.2 Histograms

To construct a histogram, we partition the domain of the distribution into n_b *bins* of equal width h. Then we count the number of points n_i, $i = 1, \ldots n_b$ in each bin and we define f_X to be equal to the $\frac{n_i}{nh}$ over bin i. Note that this way f_X is a piecewise constant function that integrates to 1. The density is zero in all bins that contain no points.

Figure 7.1 shows examples of histograms. The choice of the *bin width* h influences the aspect of the histogram and its *variance* w.r.t to the sample. This is an illustration of the *bias-variance trade-off* that will be discussed further on. Another source of variation in a histogram is the choice of *bin origins*. If all bins are shifted by an amount $\Delta < h$, the numbers n_i may change, because bin boundaries are shifted. The latter variability is entirely due to artefacts - having nothing to do either with the data or with other reasonable assumptions about nature. It is an example of problem to be avoided by a "good" statistical model. The next section will show a class of models which is clearly superior to histograms in all respects. Therefore, histograms are not recommended and should not be trusted except with caution, for a very qualitative look at the data.



a **b** **c**

Figure 7.1: Three histograms (note that they are *unnormalized*, i.e don't sum to 1). The first two are over data sets that differ in only 1 point. The third is from the first data set but has twice as many bins.

7.3 Kernel density estimation

This method constructs the estimate of f_X by placing a "bump" at each data point and then summing them up.

$$f_X(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - x_i}{h}) \tag{7.7}$$

The "bump" function $k(.)$ is called a *kernel* and the parameter h is the *kernel width*. Figure 7.3 shows three typical kernels. A kernel should always be non-negative and satisfy the following conditions

1. $\int_{-\infty}^{\infty} k(x)dx = 1$ integrate to 1
2. $\int_{-\infty}^{\infty} xk(x)dx = 0$ "centered" at 0
3. $\int_{-\infty}^{\infty} x^2k(x)dx < \infty$ "finite variance"

Usual kernels are also symmetric around 0, have a maximum at 0 and decrease monotonically away from the origin. If a kernel is 0 outside a neighborhood of the origin, then we say that it has *compact support*. The uniform and the Epanechnikov kernel have compact support, while the Gaussian kernel doesn't. The Epanechnikov kernel has optimal variance (something we'll discuss next).

Sometimes, the last condtion is replaced with $\int_{-\infty}^{\infty} x^2k(x)dx = 1$. This condition insures that different kernels are comparable w.r.t *width*.

Note that $\hat{f}_X$ defined above is a ML estimate. If we model f_X by summing n bumps of fixed shape and width and maximize the likelihood of the data w.r.t the bumps positions, then, if the width of the bumps is small enough, the optimal placement centers each bump on a data point.

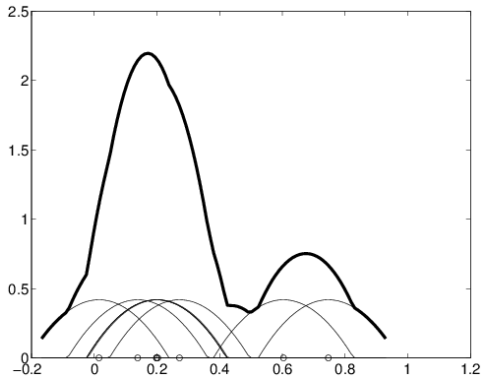


Figure 7.2: Kernel density estimation. A kernel is placed on each data point; the density is (proportional to) the sum of all kernels.

What happens if we also allow the kernel width to vary? Decreasing h will have the effect of increasing the likelihood. It will also make the estimated density look "spikier". The "optimal" h will be zero in which case the original, unconstrained ML solution with its n δ spikes is recovered. This shows that kernel density estimation is ML estimation with a restriction on how "spiky" we allow our solution to be.

Another way of looking at kernel density estimation is as a convolution: the kernel density estimator represents the convolution of the kernel with a set of spikes placed at the data points.

$$\hat{f}_X = \frac{1}{h} \hat{f}_X^{ML} * k \tag{7.8}$$

Choosing a kernel A compactly supported kernel has computational advantages: $k(x)$ being zero outside a finite interval we will only need to compute the non-zero terms in 7.7. If we assume that the original density is defined only on an interval of the real axis (such an f_X is called *compactly supported*), then it also makes sense to choose a kernel with compact support.

On the contrary, the Gaussian kernel assures that $\hat{f}_X$ is non-zero everywhere. To compute such an $\hat{f}_X$ at one point x we have to evaluate k in n points, which can be quite a burden if the data set is large.

The exact shape of the kernel is not critical in practice. Therefore in the next examples we shall only use the Gaussian kernel. Far more important than the kernel shape is the choice kernel width h that controls the *bias-variance trade-off*.

Figure 7.3: Examples of kernel functions: (a) the square kernel; (b) the Epanechnikov kernel; (c) the Gaussian kernel. The area under each kernel equals 1. Note that they have different widths and different maximum heights; therefore we expect different amounts of smoothing for the same h.

7.4 The bias-variance trade-off

Bias. The bias refers to the capacity of a family of functions (in this case the family of kernel density estimators with a given kernel k and a given h) to fit the data. The better the fit to the data, the lower the bias. For example, estimating the density with delta spikes models the data perfectly, hence has 0 bias. On the other hand, if we use a kernel density estimator with h large, then the bumps are wide and their peaks are flat. No matter if the original density was flat or not, the estimator will look flat. Hence densities that have sharp peaks can't be approximated well with a large h. We say that an estimator with large h is *biased* toward slowly varying densities. In the case of the kernel density estimators, the bias increases with h.

Because h controls the smoothness of the resulting density estimate, is also called a *smoothing parameter*. Large bias toward some kind of solution implies potentially large estimation errors, i.e large differences between our solution and the "true" density that generated the data. Therefore we usually want to have low bias.

Variance measures how much the estimated density changes due to the randomness of the data set. The maximum variance is attained for $h = 0$ - the unconstrained ML estimate. Indeed, if the data set contains a data point at a then the density there is ∞; if we draw another sample which doesn't contain a data point exactly at a, then the density in a will be 0. A variation from infinite to 0 due to an infinitesimal change in the data! As h becomes larger, the density becomes less sensitive to small perturbations in the data, therefore the variance of the estimate will decrease with h.

Since we want an estimated density that fits well *all* the possible data sets, a low variance is what we should aim for.

Considering now what we know about bias, we see that minimizing bias (which means reducing h) and minimizing variance (by increasing h) are conflicting goals. This is the *bias-variance* trade-off: finding a value of the kernel width h that is reasonable both for bias and for variance.

The effect of the sample size n. Intuitively, it is harder to fit more data points than less data points. Thus, the bias will in general not decrease when the sample size n increases. For the case of kernel density estimates, the bias doesn't change with n. The variance however will decrease with n, therefore it is at our advantage to obtain as much data as possible. With enough data to compensate for the variance, we can afford using a small h to reduce the bias as well. In conclusion, a larger sample size n has a beneficial effect on the overall quality of the estimate.

How should h be changed with n? Theoretical studies show that the optimal kernel width should be

$$h \propto \frac{1}{n^{\frac{1}{5}}} \tag{7.9}$$

Example 7.1 *Traffic on the I-90 bridge*

Assume that we have placed a sensor on the I-90 bridge that records the moment a car passes in front of it. The data file `fig_h7_traffic.dat` *is a (***fictitious!***) recording of such data over 24 hours. The same data is plotted in figure 7.5 on the time axis (from 0 to 24 hrs). We will visualize the it by constructing a kernel density estimator.*

The figure 7.6 shows the density estimate using 3 different kernels with the same width $h = 1$. The rectangular kernel is easy to recognize by its ruggedness, the other two plots that are very close together are the Gaussian kernel and the Epanechnikov (call it E.!) kernel. Note two things: First, the Gaussian and E. kernels give almost indistinguishable estimates. It doesn't really matter which one we use. The rectangular kernel, at this h, produces a more rugged picture. While for the two other kernels $h = 1$ seems a good kernel width, for the rectangular kernel we may want to use a larger h.

After experimenting with the three kernel types, we decide to use one of the smooth kernels, and the choice falls onto the E. kernel[1]. The next plots show

[1] An alternative formula for the Epanechnikov kernel is $k(x) = \frac{3}{4}(1 - x^2)$ for $|x| \leq 1$. This formula differs from the original one in the scaling of the x-axis (i.e is obtained from the previous one by the change of variable $x \to x\sqrt{5}$). The presence of the $\sqrt{5}$ factor ensures that $\int x^2 k(x)dx = 1$, same as for the Gaussian kernel.

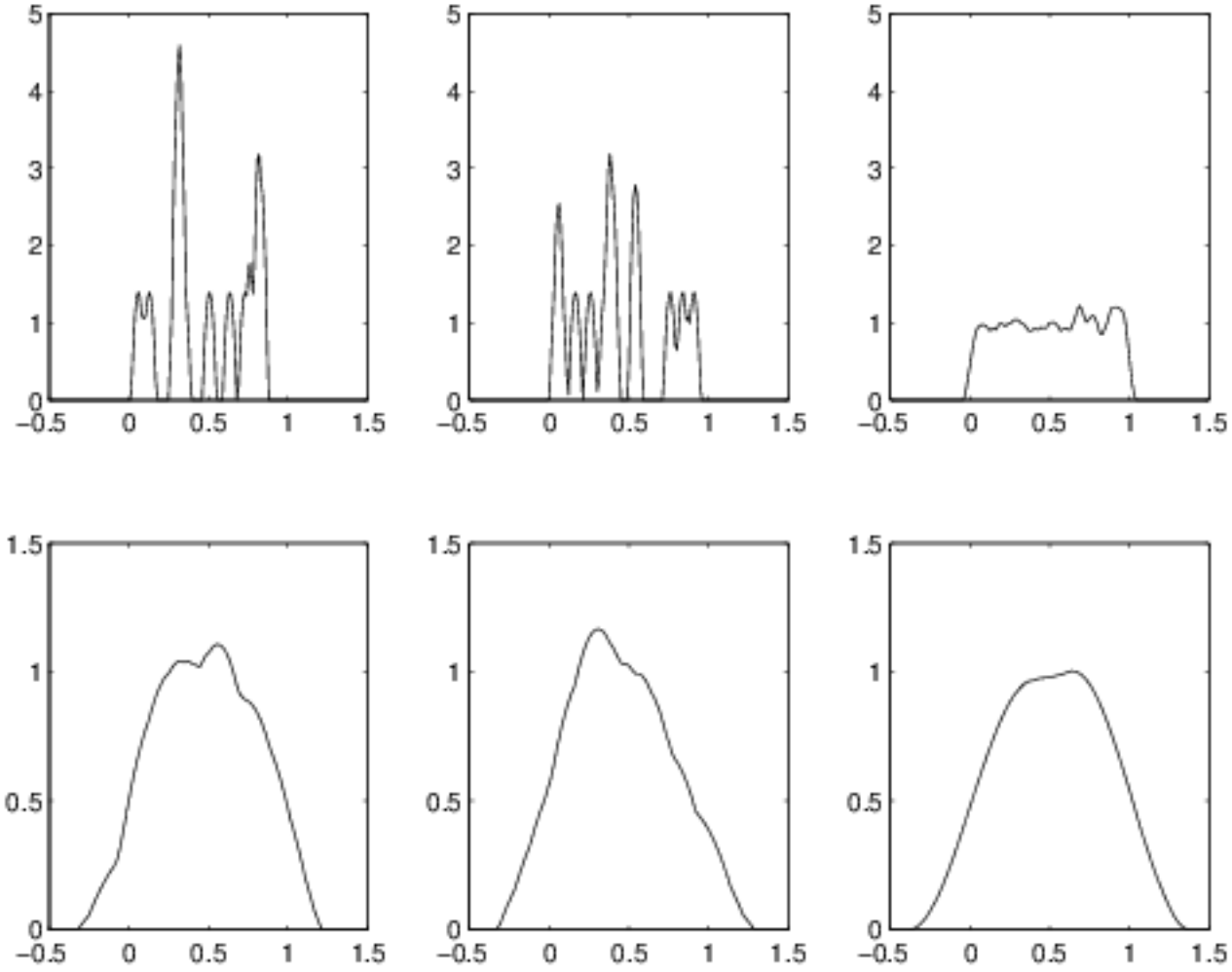


Figure 7.4: The effect of bias, variance and the sample size. The first row plots density estimates with $h = 0.02$ for 3 samples from the same distribution (a uniform over $[0, 1]$). The first two samples have size $n = 12$, the third has $n = 1200$. The density estimate is concentrated at the data points (thus the bias is low); this is beneficial for the large sample, but produces high variance for small samples. The second row shows density estimates from the same three data sets for $h = 0.17$. Now the three curves look very similar – the variance is low. The estimates obtained from the small data sets are much closer to the true distribution now. But this h is too large for the large data set, resulting in a worse estimate then previously. Last, note also that more data points are better than less data points in both cases.

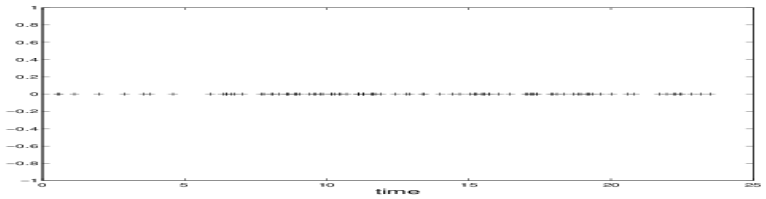


Figure 7.5: The traffic data set.

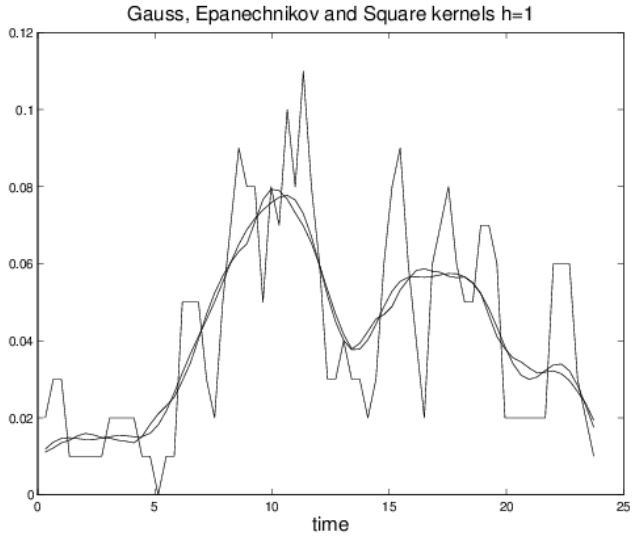


Figure 7.6: Density estimates of the traffic data with three different kernels: square, Epanechnikov and Gaussian. The kernel width is $h = 1$ in all cases.

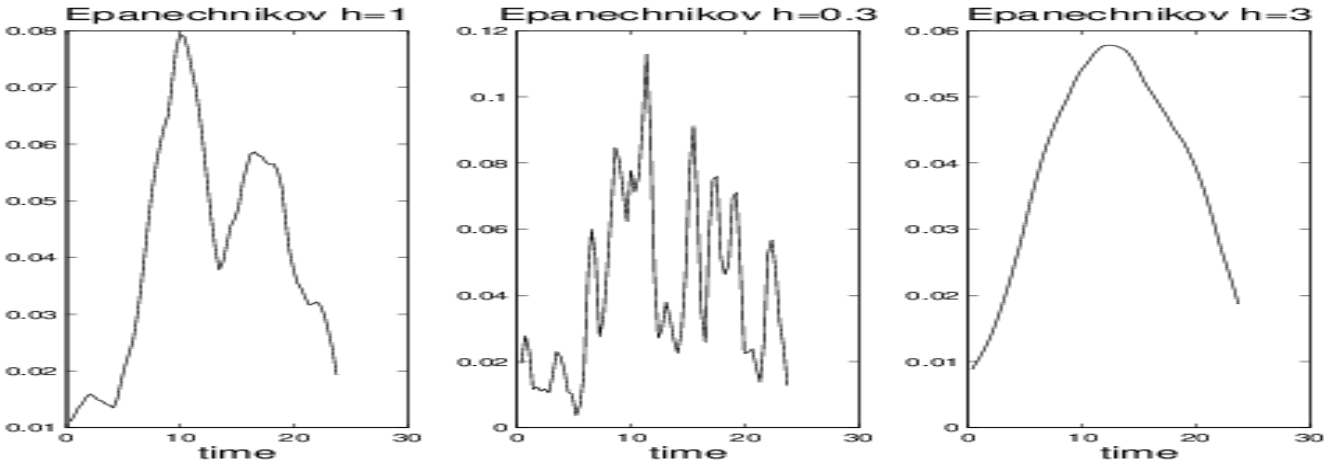


Figure 7.7: Density estimates of the traffic data with the same kernel at three different kernel widths.

the density obtained with this kernel for various kernel widths. At the smallest kernel width, $h = 0.3$ the density has many peaks and valleys. Even without having seen the true f, we may assume that traffic doesn't vary that wildly. The estimate for $h = 1$ is much smoother and on it two peaks - corresponding to the morning and afternoon rush hours - appear clearly. This plot can help anyone trying to learn something about the traffic see the global pattern (in this case two intervals of intense traffic) amids the "sampling noise" that the small h estimate failed to suppress. Thus a density estimator is a tool in data visualization. The last plot, for $h = 3$ shows only one large peak; the kernel width is too large, smoothing out not only the noise but also the structure in the data.

The density estimate can be used also for prediction: How many cars will cross the I-90 bridge tomorrow between noon and 1 pm, if the total number of cars that cross it in a day is 10,000? The answer is

$$10,000 \int_{12.00}^{13.00} f(t)dt \approx 535 \tag{7.10}$$

In the above example, $h = 1$ has been chosen by visually examining the plots. Although "the visual appeal" method is quite popular, one can do something more principled.

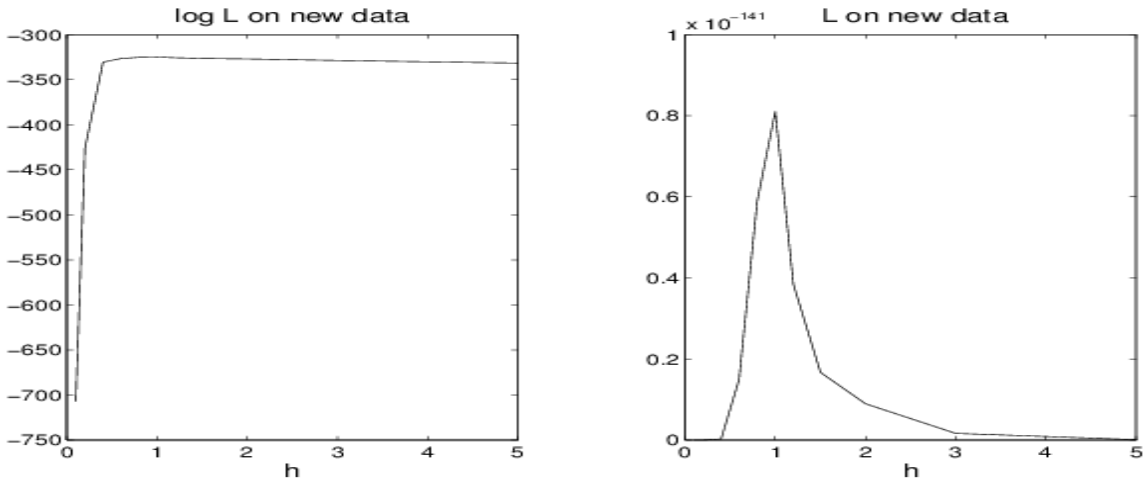


Figure 7.8: Likelihood (right) and log-likelihood of the test set of traffic data for different kernel sizes h. The optimum is at $h = 1$.

7.5 Cross-validation

The idea of cross-validation is to "test" the obtained model on "fresh" data, data that has not been used to construct the model. Of course, we need to have access to such data, or to set aside some data before building the model. In our imaginary example, we are lucky to be given "next day's data", another sample *from the same distribution* (this is the file `fig_h7_traffic_next.dat`). This data set is called *test data* or *hold out* data, in contrast to the data used to build the model which is called *training data*.

We will "test" the model on the holdout data. If the model is accurate, it must be able to predict well unseen data coming from the same distribution. In statistics terms, the unseen data should have high likelihood. Thus, the log-likelihood of the test data

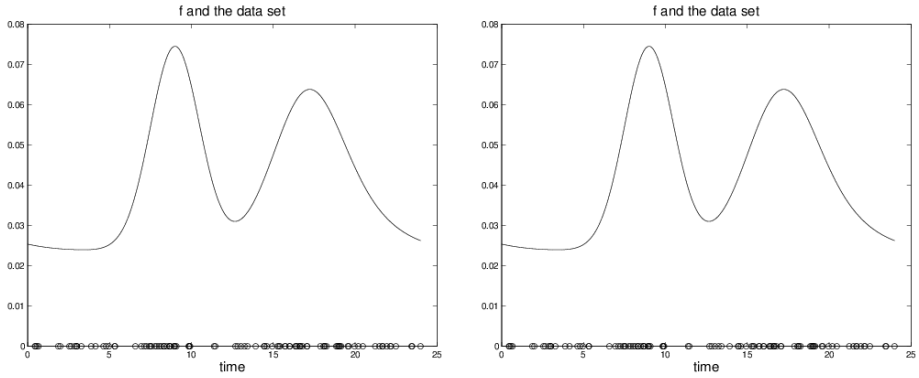
$$l_{test}(h) = \sum_{x \in \mathcal{D}_{test}} \log f_h(x) \tag{7.11}$$

$$= \sum_{x \in \mathcal{D}_{test}} \log \left[\frac{1}{|\mathcal{D}|h} \sum_{y \in \mathcal{D}} k\left(\frac{x-y}{h}\right) \right] \tag{7.12}$$

is a measure of the goodness of our estimator. In the above equation, we have indexed the density by h the kernel width. Now all we need to do is to is to compute f_h and l_{test} for a sufficiently large range of h. This was done and the results, both as likelihood and as log-likelihood are shown in figure 7.8. The

maximum value of the (log-)likelihood is attained for $h = 1$. This is the value that predicts the future data best, confirming our intuition.

Now at least having made all the choices we can allow ourselves to take a look at the "true" density that I used to generate this sample. Figure 7.9 depicts it along with the sample ($n = 100$).



a **b**

Figure 7.9: The data and the true density (left) and cumulative distribution function (right).

7.5.1 Practical issues in cross-validation

1. **The range of h to be tested.** If the kernel is finitely supported, then, once h is smaller than the smallest distance between two data points, each point is under a separate bump and decreasing it further will only create larger 0 density regions between the data. So, this is a good lower limit for h. For the upper limit, a good choice is an kernel width of about the range of the data $x_{max} - x_{min}$, or a fraction of it, e.g $1/2(x_{max} - x_{min})$.
2. **The size of the validation set $\mathcal{D}_{test}$.** If the validation set is too small, then the value of l_{test} will have high variance (i.e will change much if we pick another validation set out of the original data set). So, our decision based on it will be prone to error. But if $n_{test} = |\mathcal{D}_{test}|$ is large, then we may be left with too little data for building the model.

   What is recommended depends on the amount of data available. If data is abundant (several thousands or more data points) then a n_{test} of about 1000 should suffice; the rest of the data should be used for constructing the model. If the available data set is medium (several hundreds), then it is recommende to split it into a ratio of $\frac{n_{test}}{n} \approx \frac{1}{3} \ldots \frac{1}{2}$.

   For smaller data sets, a procedure called **K-fold cross validation** is used.

The whole data is divided at random into equal sized sets $\mathcal{D}_1, \dots \mathcal{D}_K$. Then, for $k = 1, \dots K$, $\mathcal{D}_k$ is used as a validation set, while the rest of the data is used as training set. The log-likelihood $l_k(h)$ of $\mathcal{D}_k$ for the k-th model is calculated. The final score for each h is equal to the arithmetic mean of l_k, $k = 1, \dots K$. In practice, the values of K range from 3–5 to n. If $K = n$ the method is called **leave-one-out cross validation**. You can notice that, the larger the value of K, the more credible are the results of the procedure (why?). The downside is that computational costs also grow with K as follows. The number of kernel computations to evaluate a density estimate from n' points on n_{test} points is $n'n_{test}$. Therefore, to perform K-fold CV we need

$$N \;=\; K\left(\frac{n}{K} \times (K-1)\frac{n}{K}\right) \;=\; n^2(1-1/K) \tag{7.13}$$

kernel evaluations.

3. Other sanity checks include looking at the shape of the density estimate for the chosen h, or even at how this shape changes with h.

Note that in spite of its conceptual elegance, cross-validation is not a completely error-proof method. For example, it can be shown that $h_{CV} \not\to 0$ if the target density f has infinite support and decays exponentially or slower. Also, outliers can cause problems in cross-validation.

Chapter 8

Random variables

8.1 Events associated to random variables

A *random variable* (r.v.) is defined as a function that associates a number to each element of the outcome space. Hence, any r,

$$r : S \longrightarrow (-\infty, \infty) \tag{8.1}$$

is a random variable.

Example 8.1 *Let $S = \{H, T\}^3$ be the outcome space of the experiment consisting of tossing a coin three times. The following 3 numbers that can be associated with each outcome of the experiment are random variables on this space:*

n_H the number of heads in 3 tosses

h the number of the first toss that is heads, or zero if no heads appear

r the length of the longest run of heads only

The following table shows the values of these random variables for each outcome and the respective probabilities, assuming that the coin is fair ($p = .5$).

Outcome	n_H	h	r	
TTT	0	0	0	0.125
TTH	1	3	1	0.125
THH	2	2	2	0.125
THT	1	2	1	0.125
HHT	2	1	2	0.125
HHH	3	1	3	0.125
HTH	2	1	1	0.125
HTT	1	1	1	0.125

We see that n_h takes values in the set $S_{n_h} = \{0, 1, 2, 3\}$, h in S_h and r in S_r which coincidentally are also equal to $\{0, 1, 2, 3\}$.

An important question that will often occur is "What is the probability that a random variable (RV from now on) takes a certain value?". For example what is the probability of the event "$r = 2$"? This event is the set $\{THH, HHT\}$ and its probability is the sum of the probabilities of the individual outcomes

$$P(r = 2) \;=\; P(HHT) + P(THH) \;=\; 0.25 \tag{8.2}$$

The events $r = k$ for $k = 0, 1, 2, 3$ are disjoint events, and their union is equal to the whole sample space S. We say that they form a *partition* of S. If we are interested only in r instead of the experiments outcome itself, then we can ignore the original outcome space S and instead look at the outcome space S_r of r. The probability of an outcome k in S_r is the probability of the event $r = k$ in S. The Karnaugh diagram below shows how the events "$r = k$" partition the sample space S.

x_1x_2	00	01	10	11
$x_3 = 0$	0	1		2
1	1	2		3

In general, for a random variable Y and a general outcome space S, with $Y : S \longrightarrow S_Y \subseteq (-\infty, \infty)$: the range of Y, S_Y is called the **outcome space of the random variable** Y.

If the range (i.e the outcome space) S_Y of a RV Y is discrete, then Y is called a **discrete** random variable. If S_Y is continuos (for example an interval or a union of intervals) then the RV is **continuous**. Since the outcome space of Y cannot have more elements than the original S, on a discrete (finite) S one can have only discrete (finite) RVs. If S is continuous, one can construct both discrete and continuous RV's.

For example, let $S = [0, 10]$. A continuous RV on S is $Y(x) = x^2$ for $x \in S$ and a discrete RV is $Z(x)$ that rounds x to the nearest integer.

8.2 The probability distribution of a random variable

The **probability distribution of a RV** Y denoted by P_Y is a probability over S_Y defined by

$$P_Y(y) \;=\; P(\{x \in S \,|\, Y(x) = y\}) \tag{8.3}$$

(It is a standard notation to designate random variables by capital letters and their values by the same letter, in lower case. We shall use this notation often but not always.)

In this section we show how to derive the distribution of a RV from the original distribution P on the original sample space S.

Conceptually, of course, $P_Y(y)$ is the probability of the event $Y(x) = y$. What we do now is to derive equvalent and (if possible) easily computable formulas that will give us P_Y from P. We break up this task by the sample space type of S and S_Y.

8.2.1 Discrete RV on discrete sample space

This is the case when both S and S_Y are discrete. A discrete distribution is defined by its values on the individual outcomes. Let

$$\theta_x \;=\; P(\{x\}) \quad \text{for } x \in S \tag{8.4}$$
$$\phi_y \;=\; P_Y(\{y\}) \quad \text{for } y \in S_Y \tag{8.5}$$
$$\tag{8.6}$$

The task is now to find the parameters ϕ of P_Y from the parameters θ.

$$\phi_y \;=\; P(Y(x) = y) \tag{8.7}$$
$$\;=\; \sum_{x:Y(x)=y} P(\{x\}) \tag{8.8}$$
$$\;=\; \sum_{x:Y(x)=y} \theta_x \tag{8.9}$$

Hence, the parameters of P_Y are sums of subsets of the parameters of P. Note that as every x belongs to one and only one of the events "$Y(x) = y$", every θ_x participates in one and only one ϕ_y parameter. This ensures that the ϕ parameters sum to 1.

$$\sum_{y \in S_Y} \phi_y \;=\; \sum_{y \in S_Y} \sum_{x:Y(x)=y} \theta_x \;=\; \sum_{x \in S} \theta_x \;=\; 1 \tag{8.10}$$

Example 8.2 *In the above coin toss we have 3 random variables; they define 3 probability distributions:*

value	P_{n_H}	P_r	P_h
0	*0.125*	*0.125*	*0.125*
1	*0.375*	*0.5*	*0.5*
2	*0.375*	*0.25*	*0.25*
3	*0.125*	*0.125*	*0.125*

In the above table, the values of P_r and P_h are the same. However, the two random variables are not identical, because the event "$r = 2$" (equal to $\{HHT, THH\}$) is different from the event "$h = 2$" (which is equal to $\{THH, THT\}$).

8.2.2 Discrete RV on continuous sample space

This is the case when $S = (-\infty, \infty)$ and S_Y is discrete. Hence, the task is to find the parameter ϕ of P_Y as a function of the CDF $F(x)$ or probability density $f(x)$ representation of P on S.

$$\begin{aligned} \phi_y &= P_Y(\{y\}) && (8.11)\\ &= P(Y(x) = y) && (8.12)\\ &= \int_{\{x|Y(x)=y\}} f(x)dx && (8.13)\\ &= \int_{-\infty}^{\infty} f(x)\mathbf{1}_{Y(x)=y}dx && (8.14) \end{aligned}$$

By $\mathbf{1}_{Y(x)=y}$ we denote a function that is 1 when $Y(x) = y$ and 0 otherwise. Note the similarity between the above formula and equation (8.9): they are identical except for replacing the summation with and integral when S is continuous.

If we introduce the notation (see more about this notation in the appendix)

$$Y^{-1}(y) \;=\; \{x \,|\, Y(x) = y\} \tag{8.15}$$

then we can simplify the formula (8.14) to

$$P_Y(y) \;=\; P(Y^{-1}(y)) \tag{8.16}$$

Example 8.3 . *The discrete exponential Assume P is given by*

$$f(x) \;=\; \lambda e^{-\lambda x} \quad \text{for } x \geq\; 0,\;\; \lambda > 0 \tag{8.17}$$

and

$$Y(x) \;=\; \lfloor x \rfloor \tag{8.18}$$

In other words, x has an exponential distribution and Y is its discretization to the integers.

$$\begin{aligned} P_Y(y) &= P(\lfloor x \rfloor = y) && (8.19)\\ &= P(y \;\le\; x \;< y+1) && (8.20)\\ &= \int_y^{y+1} \lambda e^{-\lambda x} dx && (8.21)\\ &= e^{-\lambda y} - e^{-\lambda(y+1)} && (8.22)\\ &= (1-e^{-\lambda})e^{-\lambda y} && (8.23) \end{aligned}$$

Hence, the distribution of the the discrete RV y is proportional to $e^{\lambda y}$. This is the **discrete exponential** distribution.

8.2.3 Continuous RV on continuous sample space

Here $S = (-\infty, \infty)$ as above, but $Y = Y(x) = g(x)$ is a continuous RV. For clarity, in this subsection we denote the function of x represented by Y with the letter g. **Note** also that it is the random variable Y that is continuous-valued (for example S_Y is an interval), but $Y(x)$ aka $g(x)$ as a function may or may not be a continuos function!! For example $g(x) = x - \lfloor x \rfloor$ the fractionary part of x. Then Y is a continous RV taking values in $[0, 1)$ but g is not continuous.

The task is to find the distribution of Y, P_Y being given the distribution of x (by its density f_X for example). A continuous distribution is defined if we know either its CDF (i.e F_Y) or its density f_Y. In the following we will derive F_Y; the density f_Y can then be easily computed as the derivative of F_Y.

$$\begin{aligned} F_Y(y) &= P_Y(Y \le y) && (8.24)\\ &= P(\{x,\; g(x) \le y\}) && (8.25)\\ &= P(g^{-1}(-\infty, y]) && (8.26)\\ &= \int_{g^{-1}(-\infty,y]} f(x)dx && (8.27) \end{aligned}$$

Example 8.4 *f_X is the uniform density on $[-1, 1]$, given by*

$$f(x) \;=\; \begin{cases} \frac{1}{2}, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{8.28}$$

and

$$Y \;=\; g(x) \;=\; x^2 \tag{8.29}$$

Then $S_Y = [0,1]$ *and for every* $y \in S_Y$

$$g^{-1}(y) \;=\; \{x,\; x^2 \le y\} \;=\; [-\sqrt{y}, \sqrt{y}]. \tag{8.30}$$

Applying (8.24) we obtain

$$F_Y(y) \;=\; P([-\sqrt{y}, \sqrt{y}]) \;=\; \begin{cases} \frac{1}{2}\cdot 2\sqrt{y} \;=\; \sqrt{y}, & 0 < y \le 1 \\ 0, & y \le 0 \\ 1, & y > 1 \end{cases} \tag{8.31}$$

$$f_Y(y) \;=\; F'_Y(y) \;=\; \begin{cases} \frac{1}{2\sqrt{y}}, & 0 < y \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{8.32}$$

Note: don't be bothered that the derivative of F_Y doesn't exist at 0 and 1. If this happens in a finite or countable set of points only, we can define the value of f_Y in those points as we wish (I set it to 0) without changing the value of the integral $\int_I f_Y dy$ on any interval I.

Example 8.5 $F(x) = \frac{x^3}{3}$ *for* $x \in [0, 3^{\frac{1}{3}}]$. *Hence*

$$f(x) \;=\; \begin{cases} x^2, & 0 \le x \le 3^{\frac{1}{3}} \\ 0, & \text{otherwise} \end{cases} \tag{8.33}$$

Let $Y = g(x) = x^2$ *as before. Then* $S_Y = [0, 3^{\frac{2}{3}}]$ *and* $g^{-1}((-\infty, y]) = [0, \sqrt{y}]$.

$$F_Y(y) \;=\; P([0, \sqrt{y}]) \tag{8.34}$$

$$\;=\; \int_0^{\sqrt{y}} f(x)dx \tag{8.35}$$

$$\;=\; \int_0^{\sqrt{y}} x^2 dx \tag{8.36}$$

$$\;=\; \frac{x^3}{3}\Big|_0^{\sqrt{y}} \tag{8.37}$$

$$\;=\; \frac{1}{3}y^{\frac{3}{2}} \tag{8.38}$$

$$\tag{8.39}$$

$$f_Y(y) \;=\; F'_Y(y) \;=\; \begin{cases} \frac{1}{2}\sqrt{y}, & 0 \le y \le 3^{\frac{2}{3}} \\ 0, & \text{otherwise} \end{cases} \tag{8.40}$$

Compare with the previous example to see that the same function $g(x) = x^2$ generates two random variables with completely different distributions, due to the different domains and distributions of x.

Example 8.6 . **Linear dependence.** *If* Y *depends linearly on* x, *we can derive a general relationship between their respective CDF's and densities. Let*

$$Y \;=\; g(x) \;=\; ax, \quad a > 0 \tag{8.41}$$

Then

$$\begin{aligned} F_Y(y) &= P(\{x,\ ax \leq y\}) && (8.42)\\ &= P(\{x,\ x \leq \frac{y}{a}\}) && (8.43)\\ &= F(\frac{y}{a}) && (8.44) \end{aligned}$$

and

$$f_Y(y) = \frac{1}{a} f(\frac{y}{a}) \quad (8.45)$$

For example, if x is uniformly distributed between 0 and 1, then $Y = ax,\ a > 0$ will be uniformly distributed between 0 and a and the density will be $f_y(y) = 1/a$ in this interval.

If $a < 0$ and F is continuous then

$$\begin{aligned} F_Y(y) &= P(\{x,\ ax \leq y\}) && (8.46)\\ &= P(\{x,\ x \geq \frac{y}{a}\}) && (8.47)\\ &= 1 - F(\frac{y}{a}) && (8.48)\\ & && (8.49)\\ f_Y(y) &= -\frac{1}{a} f(\frac{y}{a}) && (8.50) \end{aligned}$$

or, summarizing both $a > 0$ and $a < 0$

$$f_Y(ax) = \frac{1}{|a|} f(x) \quad (8.51)$$

where $|a|$ is the magnitude of a.

Example 8.7 *Another special case that can be solved in closed form is the case when $Y = g(x)$ is a strictly monotonic function of x. In this case, if we pick $S_Y = g(S)$ then the mapping g between S and S_Y is one-to-one and its inverse $g^{-1}(y)$ exists. (Do not confuse this inverse with the* inverse image *of y, $g^{-1}(\{y\})$ which can be either the empty set, or a value, or a set of x's and which exists for any g and any y. When the g is one-to-one and the* inverse function *g^{-1} also exists, the two are equal). So, if g is strictly increasing,*

$$\begin{aligned} F_Y(y) &= P_Y(Y \leq y) && (8.52)\\ &= P(g(x) \leq y) && (8.53)\\ &= P(x \leq g^{-1}(y)) && (8.54)\\ &= F(g^{-1}(y)) && (8.55) \end{aligned}$$

If g is also differentiable then we have

$$f_Y(y) \;=\; \frac{d}{dy} g^{-1}(y) f(g^{-1}(y)) \;=\; \frac{1}{g'(x)} f(x)\Big|_{x=g^{-1}(y)} \tag{8.56}$$

If g is strictly decreasing and F is continuous, by a reasoning similar to that of the previous example , we obtain

$$F_Y(y) \;=\; 1 - F(g^{-1}(y)) \tag{8.57}$$

$$f_y(y) \;=\; -\frac{1}{g'(x)} f(x)\Big|_{x=g^{-1}(y)} \tag{8.58}$$

As an example, assume $x \sim N(.;0,1)$, i.e x is normally distributed with mean 0 and variance 1. Set $Y = g(x) = ax + b$ and let us derive the distribution of Y. By the way, the CDF of a normal distribution cannot be computed in closed form. The CDF for $N(;0,1)$ is denoted by G and its values are tabulated. This exercise will allow us to compute the CDF of any other normal distribution using G.

The inverse of g is $g^{-1}(y) = \frac{y-b}{a}$ and hence

$$f_Y(y) \;=\; \frac{1}{|a|} f(\frac{y-b}{a}) \tag{8.59}$$

$$\;=\; \frac{1}{|a|\sqrt{2\pi}} e^{-\frac{(y-b)^2}{2a^2}} \tag{8.60}$$

$$\;=\; N(y; b, a^2) \tag{8.61}$$

Or, the linear transformation of a $N(.;0,1)$ distributed variable is a normally distributed variable whose mean is shifted by b and whose standard deviation is multiplied by a.

If $a > 0$ then

$$F_Y(y) \;=\; F(g^{-1}(y)) \;=\; G(\frac{y-b}{a}) \tag{8.62}$$

8.3 Functions of a random variable

A function of a random variable is itself a random variable. Let Y be a random variable on S with distribution P_Y and outcome space S_Y. Let Z be some function of Y. Then $Z : S_Y \longrightarrow S_Z \subseteq (\infty, \infty)$. The probability distribution associated with Z is

$$P_Z(z) \;=\; P_Y(\{y,\; Z(y) = z\} \;=\; P(x \in S,\; Z(Y(x)) = z\} \tag{8.63}$$

We can also define functions of several random variables; such a function is, of course, a new random variable. For example,

$$n_s = \begin{cases} n_H - r + 1 & r > 0 \\ 0 & r = 0 \end{cases} \tag{8.64}$$

is the number of sequences of heads in the experiment (it can be 0,1 or 2). It's distribution is given by

$$P_{n_s}(0) = P_r(0) = 0.125 \tag{8.65}$$
$$P_{n_s}(1) = P(n_H - r = 0 \text{ and } r \neq 0) = 0.75 \tag{8.66}$$
$$P_{n_s}(2) = P(n_H - r = 1 \text{ and } r \neq 0) = P(n_H - r = 1) = 0.125 \tag{8.67}$$

Note that unlike the previous case, for a function of several RVs, you will usually need to resort to the original outcome space and its probability distribution to compute the probability of the new RV.

To compute the density of a continuous RV Z that is a function of another (continuous) RV Y we need to take two steps:

1. compute the cumulative distribution function (CDF) of Z as

$$F_Z(z_0) = P(z \leq z_0) \tag{8.68}$$

by using the density of Y on S_Y.

2. take the derivative dF_Z/dz to obtain f_Z.

Example 8.8 *Let $X \sim$ uniform$[0,1]$ and $Y = X^2$. We want to find the CDF and density of Y.*

First, we note that $S_Y = [0,1]$. Then,

$$F_Y(a) = P(y \leq a) \tag{8.69}$$
$$= P(x^2 \leq a) \tag{8.70}$$
$$= P(x \leq \sqrt{a}) \tag{8.71}$$
$$= F(\sqrt{a}) \tag{8.72}$$
$$= \sqrt{a} \tag{8.73}$$

Hence $F_Y(y) = \sqrt{y}$ and

$$f_Y(y) = \frac{1}{2\sqrt{y}} \quad \text{for } y \in (0,1) \tag{8.74}$$

8.4 Expectation

The *expectation* of a RV Y is a real number computed by

$$E[Y] \;=\; \sum_{y \in S_Y} y.P_Y(y) \;=\; \sum_{x \in S} Y(x)P(x) \tag{8.75}$$

if Y is discrete and

$$E[Y] \;=\; \int_{y \in S_Y} y.f_Y(y)dy \;=\; \int_{x \in S} Y(x).f(x)dx \tag{8.76}$$

if Y is continuous.

Intuitively, the expectation is the "average value" of the RV, more precisely it is a weighted average of the values of the RV, where the weights are the probabilities of the outcomes. The expectation is also called *mean*, *expected value* and sometimes *average*.

Example 8.9 The fair die roll. *$S_Y = \{1, 2, \ldots\ldots, 5\}$ and $P_Y(y) = \frac{1}{6}$ for all y.*

$$E[Y] \;=\; \frac{1}{6}.1 + \frac{1}{6}.2 + \ldots \frac{1}{6}.6 \;=\; 3.5 \tag{8.77}$$

Example 8.10 The uniform density on $[a, b]$.

$$E[u_{[a,b]}] \;=\; \int_a^b \frac{1}{b-a}.x dx = \frac{1}{b-a}\;\frac{x^2}{2}\bigg|_a^b \;=\; \frac{b^2-a^2}{b-a} \;=\; \frac{b+a}{2} \tag{8.78}$$

Example 8.11 The 0–1 coin toss (Bernoulli distribution) *If X is Bernoulli with $P(1) = p, P(0) = 1 - p$ then*

$$E[X] \;=\; p.1 + (1-p).0 \;=\; p \tag{8.79}$$

Example 8.12 The three coin toss. *Let's compute the expectation of the RVs n_H, h, r from the example above.*

$$E[n_H] \;=\; 0.125.0 + 0.375.1 + 0.375.2 + .125.3 \;=\; 1.5 \tag{8.80}$$

$$E[h] \;=\; 0.125.0 + 0.5.1 + 0.25.2 + .125.3 \;=\; 1.375 \tag{8.81}$$

$$E[r] \;=\; 0.125.0 + 0.5.1 + 0.25.2 + .125.3 \;=\; 1.375 \;=\; E[h] \tag{8.82}$$

Example 8.13 The Poisson distribution

$$P(n) \;=\; e^{-\lambda}\frac{\lambda^n}{n!} \tag{8.83}$$

Therefore,

$$\begin{aligned} E[n] &= \sum_{n\geq 0} n e^{-\lambda}\frac{\lambda^n}{n!} && (8.84)\\ &= \sum_{n\geq 1} e^{-\lambda}\frac{\lambda^n}{(n-1)!} && (8.85)\\ &= \lambda\sum_{n\geq 1} e^{-\lambda}\frac{\lambda^{n-1}}{(n-1)!} && (8.86)\\ &= \lambda\sum_{n\geq 0} e^{-\lambda}\frac{\lambda^n}{n!} && (8.87)\\ &= \lambda && (8.88) \end{aligned}$$

Hence, the average value of n is equal to λ.

Example 8.14 The normal distribution *If $X \sim N(\,;\mu,\sigma^2)$ then*

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx && (8.89)\\ &= \int_{-\infty}^{\infty} (x-\mu+\mu)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx && (8.90)\\ &= \underbrace{\int_{-\infty}^{\infty} (x-\mu)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx}_{0} + \mu\underbrace{\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx}_{1} && (8.91)\\ &= \mu && (8.92) \end{aligned}$$

Hence, the expectation of the normal distribution is μ which is rightly called the mean.

8.4.1 Properties of the expectation

Property 1. The expectation of a constant random variable is the constant itself. If $Y = C$, where C is a constant, then Y is in fact determinintic. The expectation is expressed as

$$E[Y] \;=\; C \times P(C) \;=\; C \times 1 \;=\; C \tag{8.93}$$

Property 2. Multiplication with a constant. Let X be random variable and $Y(X) = CX$. Then

$$E[Y] \;=\; \sum_{x\in S} Cx\,P(x) \;=\; C\sum_{x\in S} x\,P(x) \;=\; CE[X] \tag{8.94}$$

Note: in the above we used the standard statistical notation by which r.v's are denoted by capital letters (e.g X) and their values by lower case letters (e.g x).

Property 3. The expectation of the sum of two r.v's is equal to the sum of their expectations. Let $Y(X), Z(X)$ be two r.v's.

$$\begin{aligned} E[Y+Z] &= \sum_{x\in S}(Y(x)+Z(x))P(x) && (8.95)\\ &= \sum_{x\in S}Y(x)P(x)+\sum_{x\in S}Z(x)P(x) && (8.96)\\ &= E[Y]+E[Z] && (8.97)\end{aligned}$$

Here we have implicitly assumed that Y and Z share a common original sample space. But what happens if they don't? Such a case is the case when Y and Z are the outcomes of two die rolls from different dice. This is not a problem. It is always possible to construct a sample space S and and r.v X so that Y and Z are both functions of the random outcome X. For instance, in the two dice example, we can construct the space S representing the outcomes of the pair of dice. An outcome $X \in S$ is a pair (X_1, X_2). Then, trivially, $Y(X) = X_1$ and $Z(X) = X_2$. Hence, equation (8.97) always holds.

From the last two properties it follows that the expectation is a *linear* operation. This means that for any n random variables $X_1, \ldots X_n$ and real numbers $\lambda_1, \ldots \lambda_n$ we have

$$E[\sum_{i=1}^{n}\lambda_i.X_i] = \sum_{i=1}^{n}\lambda_i E[X_i] \quad (8.98)$$

For example, the expectation of $n_H - r$ is

$$\begin{aligned} E[n_H-r] &= 0.P(n_H-r=0)+1.P(n_H-r=1)+2P(n_H-r=2)+3P(n_H-r=3)\\ &= 0\times 0.875+1\times 0.125+2\times 0+3\times 0 && (8.99)\\ &= 0.125 && (8.100)\\ &= E[n_H]-E[r] && (8.101)\end{aligned}$$

8.5 The median

Like the expectation, the median is another way of summarizing a r.v in a single number. By definition, the **median** of a continuous random variable X with CDF $F(x)$ and density $f(x)$ is the value $m[X]$ for which

$$P(X \le m[X]) = \frac{1}{2} \quad \Leftrightarrow \quad F(m[X]) = \frac{1}{2} \quad \Leftrightarrow \quad \int_{-\infty}^{m[X]} f(x)\,dx = \int_{m[X]}^{-\infty} f(x)\,dx = \frac{1}{2} \quad (8.102)$$

It can be shown (take it as an exercise!) that if the density f_X is symmetric about the value μ then $m[X] = E[X] = \mu$.

The median can be defined as well for a discrete distribution. In this case, it's value is often not unique. For example, take P_X to be the uniform distribution over $\{1, 2, 3\}$; then $m[X] = 2$. But for P_Y uniform over $\{1, 2, 3, 4\}$ any number between 2 and 3 can be the median. By convention, computer scientists take it to be 2.5.

8.6 Variance

A special kind of expectation associated with a RV measures the average amount of deviation from it's mean. This is the *variance*, defined as

$$Var\, X \;=\; E[(X - E[X])^2] \tag{8.103}$$

The variance is always ≥ 0. When the variance is 0, the RV is deterministic (in other words it takes one value only). The square root of the variance is called *standard deviation*.

Let us compute the variances for the examples above.

Example 8.15 The fair die roll

$$Var\, Y \;=\; \frac{1}{6}\left[(1-3.5)^2 + (2-3.5)^2 + \ldots (6-3.5)^2\right] \;=\; 2.9167 \tag{8.104}$$

Example 8.16 The uniform density on $[a, b]$.

$$\begin{aligned} Var\, U &= E\left[(U - \frac{a+b}{2})^2\right] && (8.105)\\ &= \int_a^b \frac{1}{b-a}\left(u - \frac{a+b}{2}\right) du && (8.106)\\ &= \frac{1}{b-a}\int_a^b \left(u^2 - (a+b)u + \frac{(a+b)^2}{4}\right) du && (8.107)\\ &= \frac{1}{b-a}\left[\frac{u^3}{3}\Big|_a^b - (a+b)\frac{u^2}{2}\Big|_a^b + \frac{(a+b)^2}{4}\Big|_a^b\right] && (8.108)\\ &= \frac{(b-a)^2}{12} && (8.109) \end{aligned}$$

Hence, the mean of the uniform distribution is in the middle of the interval, while the variance is proportional to the length of the interval, squared. Thus the standard deviation is proportional to the length of the interval.

Example 8.17 The 0–1 coin toss (Bernoulli distribution)

$$\begin{aligned} Var\, X &= E[(X-p)^2] && (8.110)\\ &= p(1-p)^2 + (1-p)(0-p)^2 && (8.111)\\ &= p(1-p)(1-p+p) && (8.112)\\ &= p(1-p) && (8.113) \end{aligned}$$

Note that the variance of the biased coin toss depends on p. It is largest for $p = 0.5$ and it tends to 0 when $p \to 0$ or 1. Does this make sense?

Example 8.18 The Poisson distribution *To compute this variance, we will use formula (8.133) proved in the next section, plus another trick that will help simplify the calculations. [***Exercise** *After you see the result and all these tricks, you can do the direct calculation of the variance as an exercise.]*

$$\begin{aligned} E[n^2-n] &= \sum_{n\geq 0} n(n-1)e^{-\lambda}\frac{\lambda^n}{n!} && (8.114)\\ &= \lambda^2 \sum_{n\geq 2} e^{-\lambda}\frac{\lambda^{n-2}}{(n-2)!} && (8.115)\\ &= \lambda^2 && (8.116) \end{aligned}$$

Now, by (8.133), the variance of a Poisson random variable is

$$\begin{aligned} Var(n) &= E[n^2-n+n] - (E[n])^2 && (8.117)\\ &= \lambda^2 + \lambda - \lambda^2 && (8.118)\\ &= \lambda && (8.119) \end{aligned}$$

Example 8.19 The normal distribution

$$\begin{aligned} Var\, X &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx && (8.120)\\ &= \int_{-\infty}^{\infty} t^2\sigma^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}} \sigma dt && (8.121)\\ &= \sigma^2 \int_{-\infty}^{\infty} -t.(-t\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}) dt && (8.122)\\ &= -\sigma^2 \left[t.\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \right] && (8.123)\\ &= -\sigma^2[0-1] && (8.124)\\ &= \sigma^2 && (8.125) \end{aligned}$$

Hence, the variance of the normal distribution is σ^2 and its standard deviation is σ.

Properties of the variance

Property 1. The variance is always non-negative. The variance is the expectation of an expression that can never be negative. The variance is zero only if the expression $x - E[X]$ is zero for all x. This happens only if X is a constant.

Property 2. The variance of a constant C is zero. We have that $E[C] = C$ and therefore

$$Var(C) \;=\; E[(C - E[C])^2] \;=\; E[0] \;=\; 0 \tag{8.126}$$

The converse is also true, as shown in property 1. In other words, a variable that has zero variance has no randomness at all, it is deterministic.

Property 3. Multiplication with a constant. If $Y = CX$ then

$$\begin{aligned} Var(Y) &= E[(CX - E[CX])^2] && (8.127)\\ &= E[(CX - CE[X])^2] && (8.128)\\ &= E[C^2(X - E[X])^2] && (8.129)\\ &= C^2E[(X - E[X])^2] \;=\; C^2Var(X) && (8.130) \end{aligned}$$

Hence, if a variable is multiplied by a constant, its variance scales quadratically with the value of the constant. Note that the standard deviation, the square root of the variance, scales linearly:

$$\sqrt{Var(CX)} \;=\; C\sqrt{Var(X)} \tag{8.131}$$

Property 4. The variance of a sum of two random variables does **not** equal the sum of the variances, except in special cases. We will discuss this when we talk about correlation.

8.7 An application: Least squares optimization

8.7.1 Two useful identities

First we prove two useful identitites involving the mean and variance. Let X be a random variable, and denote

$$\mu \;=\; E[X] \qquad \sigma^2 \;=\; Var\, X \tag{8.132}$$

Note that X may be any r.v (not necessarily Gaussian, or continuous).

$$\boxed{Var\,X \;=\; E[X^2] - (E[X])^2} \tag{8.133}$$

Proof

$$\begin{aligned} Var\,X &= E[(X-\mu)^2] && (8.134)\\ &= E[X^2 - 2\mu X + \mu^2] && (8.135)\\ &= E[X^2] - 2\mu \underbrace{E[X]}_{\mu} + \mu^2 && (8.136)\\ &= E[X^2] - \mu^2 \;\; \text{Q.E.D} && (8.137) \end{aligned}$$

If a is a fixed real number

$$\boxed{E[(X-a)^2] \;=\; \sigma^2 + (a-\mu)^2} \tag{8.138}$$

Proof

$$\begin{aligned} E[(X-a)^2] &= E[(X-\mu+\mu-a)^2] && (8.139)\\ &= E[(X-\mu)^2] + 2\underbrace{E[(X-\mu)]}_{0}(\mu-a) + E[(\mu-a)^2] && (8.140)\\ &= \sigma^2 + 2(\mu-a)\times 0 + (a-\mu)^2 && (8.141) \end{aligned}$$

8.7.2 Interpretation of the second identity

Let now $C(a) = (X-a)^2$ be a r.v and $\bar{c}(a)$ be its expectation. By (8.138)

$$\bar{c}(a) \;=\; \sigma^2 + (a-\mu)^2 \tag{8.142}$$

This is a quadratic function of a that has a minimum for $a = \mu$. The minimum value is

$$\bar{c}^* \;=\; \sigma^2 \tag{8.143}$$

The following example illustrates one frequent usage of formula (8.138). Assume X is the temperature at noon on a given day. In the morning you cannot know what will be the value of X, but you know its distribution. You want to dress aprropriately for temperature X but of course you cannot do it before knowing X. All you can do is "guess" that $X = a$ and dress appropriately for a. But if $X \neq a$, you will pay a "cost" for your error. You will suffer of cold of $X < a$ and you will be too hot if $X > a$. Evaluate the **cost of your error** to be $C(X) = (X-a)^2$, thus a **squared** cost. You want to choose a in a way that minimizes your expected cost $E[C]$. The equations above offer the solution to this problem. They say that if you are trying to predict a random outcome

X by the constant a, and that if the cost of your error is a quadratic function of the distance between the true X and a, then the best possible prediction is $a = E[X]$. The cost associated with this prediction, in other words the lowest possible cost to pay, is equal to the variance of X. This last observation confirms that the variance is a measure of predictability of a variable.

8.8 The power law distribution

Assume $S = \{1, 2, \ldots n, \ldots\}$. The "power law" is a distribution over S defined by

$$P(n) \;=\; \frac{1}{Z} n^{-r} \qquad \text{for } n \in S,\; r > 1 \tag{8.144}$$

with Z equal to the normalization constant

$$Z \;=\; \sum_{n=1}^{\infty} \frac{1}{n^r} \tag{8.145}$$

The name "power law" underscores the fact that the probability of the outcome being n is proportional to the inverse r-th power of n. The parameter r can have a fractionary or integer value. Figure 8.1 shows plots of the power law distribution for different values of r.

The power law distribution, and its close relative called Zipf's law, have a pervasive presence in the realm of modelling human activities (but are not restricted to it). Here are a few examples:

- Denote by d the number of links pointing to a given web page. This is called the **in-degree** of the page in the graph representing the world wide web. The probability that a random page has in-degree d follows a power law given by
$$P(d) \;\propto\; d^{-r}$$
FYI, the average degree of a web page, that is $E[d]$ is very close to 2.1.

- If we rank the pages of a web site (Yahoo, CNN) with the most accessed page first and the least accessed page last, then the probability of accessing page i in this list is proportional to $1/i^r$. This kind of distribution is called **Zipf's law**. Thus, Zipf's law is a power law distribution over a domain of ranks.

- Zipf's law is present in information retrieval as well. For example, sort all the words in an English dictionary by their frequency. Usually this frequency is estimated from a large collection of documents, called a **corpus** (plural "corpora"). If this is done, very likely word 1 is "the", word 2 is "and", while rare words like "corpora" have much higher ranks (a natural

language corpus for English may contain between a few thousands and some tens of thousands of words). It was found that the frequency of the i-th word in the list is proportional to $(i + K)^{-r}$.

Let us now evaluate the mean, standard deviation, and median of the power law distribution.

$$E[n] = \frac{1}{Z}\sum_{n=1}^{\infty} n \times n^{-r} = \frac{1}{Z}\sum_{n=1}^{\infty} n^{-r+1} \tag{8.146}$$

$$E[n^2] = \frac{1}{Z}\sum_{n=1}^{\infty} n^2 \times n^{-r} = \frac{1}{Z}\sum_{n=1}^{\infty} n^{-r+2} \tag{8.147}$$

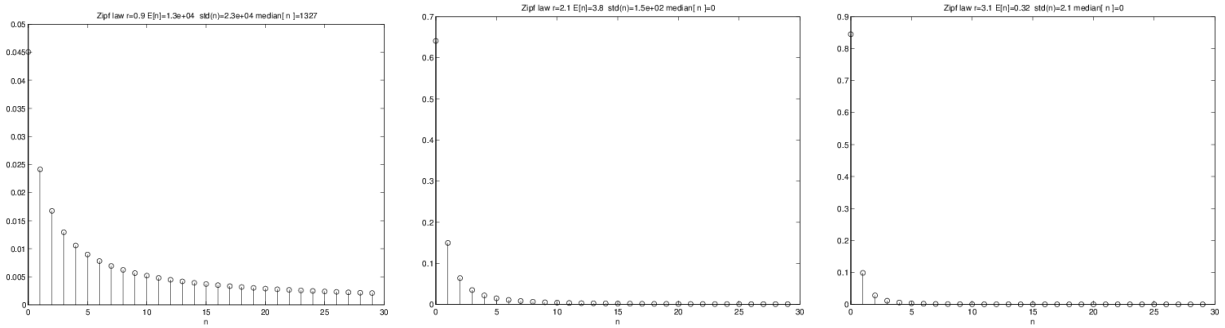
$$Var(n) = E[n^2] - (E[n])^2 \tag{8.148}$$

Figure 8.2 shows the dependence of the mean, standard deviation and median of n as a function of r. It is a well-known fact that

$$\sum_{n=1}^{\infty} n^{-r}$$

converges only for $r > 1$ and is ∞ for $0 \leq r \leq 1$. Therefore, the power law will have infinite mean if $r < leq2$ and infinite variance if $r \leq 3$. In practice, the domain S never extends to infinity (there are a finite number of web pages, a finite number of words in the dictionary, etc.). So in the domain of social networks and social interactions at least, the mean and variance of any process will be finite. However, for r smaller or near 3, the fact that the power law decays slowly makes the "tails" of the distribution have large weight. So when we estimate the distribution, or its mean, or variance, these estimates will be very sensitive to the tails and therefore not robust.

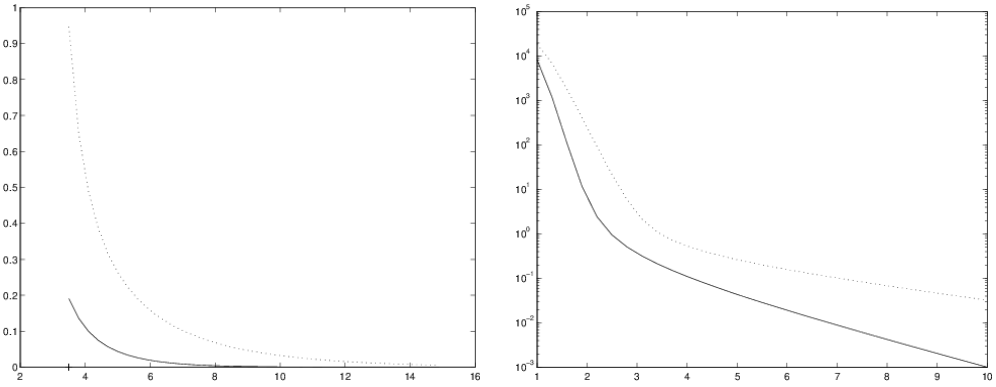
Figure 8.1: The power law distribution $P(n) \propto (n + 1)^{-r}$ for different values of r. The domain S has been truncated to $\{0, 1, \ldots, 100000\}$.



$r = 0.9$ $r = 2.1$ $r = 3.1$

This is not the only reason the power law is a difficult distribution. Another reason is that the mean is not a very informative quantity for this distribution.

Figure 8.2: The the mean (full line) and standard deviation (dotted line) of the power law distribution $P(n) \propto (n+1)^{-r}$ for different values of r. The domain S has been truncated to $\{0, 1, \ldots, 100000\}$. On the left, the mean and standard deviation is plotted for $r > 3$. On the right, the values are plotted for $r > 1$ on a logarithmic scale. Note the fast grouth of the mean to the left of 2 and of the standard deviation to the left of $r = 3$. The median is 0 for all values of r.



Note that in figure 1, the most probability mass is concentrated at 0; for example, for $r = 2.1$, $P(0) = 0.64$. For the same r, the mean equals 3.8. These type of distributions, where most of the mass is on one side of the mean, are called **skewed** distributions. The variance is also rather uninformative; for the case $r = 2.1$, the standard deviation is 156 suggesting that the bulk of the samples is in $[3.8 - 156, 3.8 + 156]$. In fact, the interval that contains the bulk of the distribution is much smaller: $[0, 5]$ contains 92% of the probability mass.

8.9 Appendix: The inverse image of a set and the change of variables formula

In dealing with RVs and their expectations, we shall rely heavily on the change of variable formula, so let's start by recalling what it is.

The setup:

- y is a continuous and differentiable function of x and g' is its derivative

$$y = g(x) \tag{8.149}$$

$$g' = \frac{dg}{dx} \tag{8.150}$$

- For every x in a set A, $y = g(x)$ is in a set B and $g(x) \in B$ only if $x \in A$.

In other words:

$$A = \{x,\ g(x) \in B\} = g^{-1}(B) \tag{8.151}$$
$$B = \{y,\ y = g(x),\ x \in A\} = g(A) \tag{8.152}$$

We denote this equivalently by

$$x \in A \xrightarrow{g} y \in B \tag{8.153}$$

- $f : B \to \mathcal{R}$ is an integrable function of y

Then the change of variable formula is

$$\boxed{\int_B f(y)dy = \int_A f(g(x))g'(x)dx}$$

If you prefer a simplified and somewhat more ambiguous notation, whereby $y = y(x)$ (replacing the letter g by y), then the change of variable formula reads

$$\int_b f(y)dy = \int_A f(y(x))y'(x)dx$$

We shall use the formula in both directions, i.e we will sometimes replace the left hand side by the right hand side of (8.9) and sometimes we'll do the reverse.

Chapter 9

Conditional Probability of Events

9.1 A summary of chapters 8 and 9

$$\text{a random outcome } X \longrightarrow Y \text{ depends on } X \begin{cases} \text{deterministically :} & \text{Random Variable} \\ \text{non} - \text{deterministically :} & \text{Conditional Probability} \end{cases}$$

9.2 Conditional probability

Suppose that we are interested in the probability of an event A occurring during an experiment with outcome space S. Can the occurrence of another event B, or additional knowledge about the conditions of the experiment influence the probability of A?

The answer is YES, the probability of an event A can change when the experimental conditions change. Assume the "experimental conditions" are represented by event B; then to emphasize the dependence of $P(A)$ on B we write $P(A|B)$ and read it as *probability of A conditioned on B* or probability of A *given B*.

Example 9.1 *The probability of rain on a random day in a random city in the US is*

$$P(\text{rain}) = 0.3 \tag{9.1}$$

But, if we know the city, then this probability changes:

$$P(\text{rain}|\text{Seattle}) = 0.5 \tag{9.2}$$
$$P(\text{rain}|\text{Phoenix}) = 0.01 \tag{9.3}$$

Sometimes it's not the experimental conditions that change, but our knowledge about them.

Example 9.2 *In class on Monday your friend Andy tells you:*
A: "I ran into one of your friends this weekend. Guess who it was?"
You have three other friends: Beth, Chris and Dana. Which of them could it have been? You don't have any reason to think one friend was more likely than another to coincidentally meet Andy so

$$P(B) = P(C) = P(D) = \frac{1}{3}$$

But then Andy drops a hint:
A: "I was skiing at Crystal this weekend".
You know that Beth and Dana do not ski and in fact don't like snow at all, but Chris is a fanatic of skiing. Therefore, your probabilities change to

$$P(C|\text{Crystal}) = \frac{98}{100} \qquad P(B|\text{Crystal}) = P(D|\text{Crystal}) = \frac{1}{100}$$

Notation: From now on, we will denote the event "A and B" by A, B or AB or $A \cap B$ interchangeably.

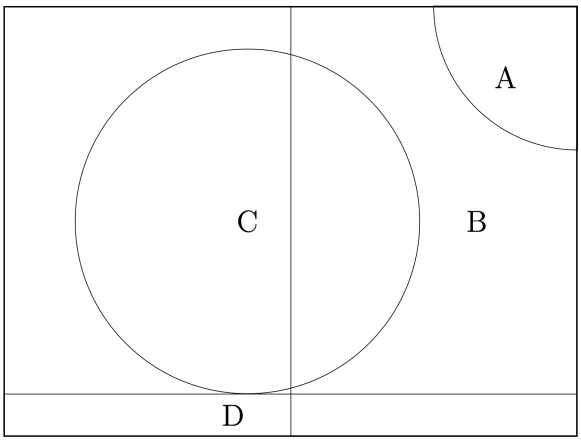
The conditional probability of event A given B is defined as:

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{9.4}$$

In words, $P(A|B)$ equals the proportion of outcomes in which A occurs from the total set of outcomes in which B occurs.

Example 9.3 The dice roll. $P(2) = 1/6$ *but* $P(2 \,|\, \text{"even outcome"}) = 1/3$ *and* $P(2 \,|\, \text{"odd outcome"}) = 0$. $P(5 \,|\, \text{outcome} > 4) = 1/2$, $P(\text{outcome} > 4 \,|\, \text{outcome} > 2) = 1/2$, $P(\text{outcome} > 2 \,|\, \text{outcome} > 4) = 1$.

Example 9.4 The robot in the closet *Below is a picture of a very small room that the Roomba vacuum cleaner (*`www.irobot.com`*) takes care of. The pictures maps some areas of interest:*



A *is the area near the robot's power supply*

B *is the Eastern half of the closet where it's warm and cozy*

C *is the central area that the Janitor cleans weekly*

D *is the strip by the door (there are unhealthy drafts there).*

The robot is programmed to be constantly moving at random, so the probability that it is in a certain area is proportional to the size of the area.

Therefore,

$$\begin{aligned} P(A) &= \frac{\pi}{4 \times 12} \approx 0.065 \\ P(B) &= \frac{1}{2} \\ P(C) &= \frac{\pi(1.5)^2}{12} \approx 0.59 \\ P(D) &= \frac{1.2}{12} = 0.1 \end{aligned}$$

Here are some conditional probabilities:

$$\begin{aligned} P(A|B) &= \frac{\pi}{4 \times 6} \approx 0.129 > P(A) \\ P(A|\overline{B}) &= 0 \\ P(B|A) &= 1 \\ P(C|B) &= \frac{2.39}{6} \approx 0.39 < P(C) \\ P(C|\overline{B}) &= \frac{4.78}{6} \approx 0.78 > P(C) \\ P(D|B) &= \frac{0.6}{6} = 0.1 \end{aligned}$$

9.3 What is conditional probability useful for?

Here are a few situations that can be described using conditional probability:

- Discovering that there is a relationship between two variables or events. For example, $P($ lung cancer $|$ smoking $) > P($ lung cancer $|$ not smoking $)$ suggests that there is a connection between lung cancer and smoking.
- Better predicting or guessing the value of an unobserved variable. The weather man studies the data collected from satellites and weather stations and makes a prediction for the tomorrow's weather. His prediction is $P($sunny$|$ data $)$. You are a trained meteorologist, but you don't have access to other data than looking at the sky today. You can also make a prediction $P($sunny$)$. It is possible that $P($sunny$|$ data $) >, =, < P($sunny$)$. It is possible that the weather tomorrow is closer to your prediction than to the weatherman's (what does "closer" mean in this context?), and both are guesses anyways, but on average the weather man's guess is a more accurate guess.
- Guessing about an unobserved cause. (This is an application of Bayes' rule which will be discussed further in this chapter.) It is known that $P($ fever $|$ infectious disease $)$ is high and $P($ fever $|$ no infectious disease $)$ is low. Therefore, if fever is observed, we conclude that $P($ infectious disease $)$ is higher than if we hadn't measured the temperature.
- Probabilistic reasoning (something that will be discussed later on). The formula of conditional probability is an instrument that allows us to reason and draw conclusions about a variable we don't observe from one that we can observe. With a little mathematical manipulation, one can use this formula to derive the probability of one or several unobserved events from several observed ones. Taking medical diagnosis as an example again: the events of interest are diseases (flu, hepatitis, diabetes). The observed events are temperature, results of blood tests, patient's age, sex, etc. We also know various conditional probabilities relating the observed and unobserved events, as for example $P($fever$|$flu$)$, $P($ high blood sugar$|$ diabetes $)$. What we want is $P($ diabetes $|$ observed blood sugar, temperature, other tests, patient age $)$. This can be computed from the observations and the known conditional probabilities.

9.4 Some properties of the conditional probability

Property 1 $P(\cdot|B)$ is a probability distribution over S, i.e it obeys the axioms

of probability. Let us check them:

First of all, $P(A|B)$ is defined for all events $A \subseteq S$ and clearly, $P(A|B) \geq 0$ for all A. Then,

$$P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1. \tag{9.5}$$

Finally, if A, C are disjoint sets (incompatible events) then $(A \cap B) \cap (C \cap B) = \emptyset$ and thus

$$P(A \cup C|B) \;=\; \frac{P(A \cup C, B)}{P(B)} \;=\; \frac{P(A, B) + P(C, B)}{P(B)} \;=\; P(A|B) + P(C|B) \tag{9.6}$$

Note that under $P(\ |B)$ all the outcomes of S that are not in B and all the events of S that do not intersect B have 0 probability (see Property 4).

Property 2 If $A \subseteq B$ then

$$P(A|B) \;=\; \frac{P(A, B)}{P(B)} \;=\; \frac{P(A)}{P(B)} \;\geq\; P(A) \tag{9.7}$$

Hence, if A implies B then B occurring increases the probability of A.

Property 3 If $B \subseteq A$ then

$$P(A|B) \;=\; \frac{P(A, B)}{P(B)} \;=\; \frac{P(B)}{P(B)} \;=\; 1 \tag{9.8}$$

Hence, if B implies A then B occurring makes A sure.

Property 4 If $B \cap A = \emptyset$ then

$$P(A|B) \;=\; \frac{P(A, B)}{P(B)} \;=\; \frac{0}{P(B)} \;=\; 0 \;=\; P(B|A) \tag{9.9}$$

Intuitively, conditioning on B sets the probability of all the outcomes outside B to 0 and renormalized the probabilities of the remaining outcomes to sum to 1. (Or, in other words, the probability of $S \setminus B$ is set to 0 – it makes sense since we know that B occurred – and the outcome space shrinks to B. The probability of B thus becomes 1 and the probabilities of all events in B are scaled up by $1/P(B)$).

Property 5 Conditioning on several events.

$$P(A|B, C) \;=\; \frac{P(A, B, C)}{P(B, C)} \;=\; \frac{P(A, B|C)P(C)}{P(B|C)P(C)} \;=\; \frac{P(A, B|C)}{P(B|C)} \tag{9.10}$$

In other words, if an event C is a context that is true in all cases, then the formula for the conditional probability between A and B gets C added behind the conditioning bar, but otherwise remains unchanged.

9.5 Marginal probability and the law of total probability

Using conditional probability we can get the following useful rule (sometimes called the *rule of total probability*).

$$\begin{aligned} P(A,B) &= P(B)P(A|B) && (9.11)\\ P(A,\overline{B}) &= P(\overline{B})P(A|\overline{B}) && (9.12)\\ \text{but} & \quad (A,B)\cap(A,\overline{B}) = \emptyset \quad \text{hence} \\ P(A) &= P(A,B)+P(A,\overline{B}) \\ &= P(B)P(A|B)+P(\overline{B})P(A|\overline{B}) && (9.13) \end{aligned}$$

In this context $P(A,B)$ is called the **joint** probability of A and B and $P(A)$ is called the **marginal** probability of A. Using a similar reasoning as above we also have that

$$P(A) = P(A,B)+P(A,\overline{B}) \qquad (9.14)$$

The last equations shows how to use the joint probability of two events to obtain the probability of one of them only.

Example 9.5 *Alice is thinking of going to a party this Saturday, and wants to compute the chances that she has a good time there. Denote by A the event "Alice enjoys herself at the party" and by B the event "Bob, her boyfriend, will also be at the party".*

If Bob is present, Alice is practically sure that she'll have a great time ($P(A|B) = 0.90$) but if he's absent, she may like the party anyways, albeit with lower probability $P(A|\overline{B}) = 0.30$. She knows that Bob has a lot of homeworks to finish, so that $P(B) = 0.6$. Then

$$\begin{aligned} P(A) &= P(B)P(A|B)+P(\overline{B})P(A|\overline{B}) \\ &= 0.6\times 0.9+0.4\times 0.3 \\ &= 0.66 \end{aligned}$$

Note that the total probability of having a good time is somewhere in between 0.9 and 0.3 the probabilities of having a good time in each of the two (mutually exclusive) situations. Is this always the case?

9.6 Bayes' rule

$$P(A,B) = P(B)P(A|B) = P(A)P(B|A) \qquad (9.15)$$

From which we derive the famous *Bayes' rule*

$$P(A|B) \;=\; \frac{P(A)P(B|A)}{P(B)} \tag{9.16}$$

This simple property of conditional probability has generated a whole field of research called *Bayesian* statistics. We will revisit Bayes' formula later in this course.

Example 9.6 Probabilistic medical diagnosis. *A patient tests HIV positive on a test (call this event T) and the doctor wants to know what is the probability that the patient is actually HIV positive (call this event HIV). What the doctor knows is that*

- *The HIV test is not perfect; it will be positive if the patient has HIV with probability $P(T|HIV) = 0.99$ and negative otherwise. The test may also be positive if the patient is not infected with HIV; this happens with probability $P(T|\overline{HIV}) = 0.03$.*
- *The incidence of the HIV virus in the population of the US is $P(HIV) = 0.001$. (These figures are not real figures!)*

How can the doctor compute what she wants to know, namely $P(HIV|T)$ from the information she has?

$$P(HIV|T) \;=\; \frac{P(T|HIV)P(HIV)}{P(T)}$$

We now compute $P(T)$ by the law of total probability

$$P(T) \;=\; P(T|HIV)P(HIV) + P(T|\overline{HIV})P(\overline{HIV})$$

Replacing the numbers we get

$$P(HIV|T) \;=\; \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.03 \times 0.999} \;=\; 0.032$$

This probability is very small, but it is about 30 times larger than the $P(HIV)$ before seeing the positive result of the test. This is due mainly to the fact that the **prior** *probability of HIV in the population $P(HIV)$ is very small.*

Suppose now that the doctor has also examined the patient, and now, based on the patient's symptoms, she thinks that the probability of an HIV infection is $P'(HIV) = 0.3$. Therefore she prescribes an HIV test, which is positive. What is the new value $P'(HIV|T)$ in these conditions? Redoing the above calculations we obtain

$$P'(HIV|T) \;=\; \frac{0.99 \times 0.3}{0.99 \times 0.3 + 0.03 \times 0.7} \;=\; 0.93$$

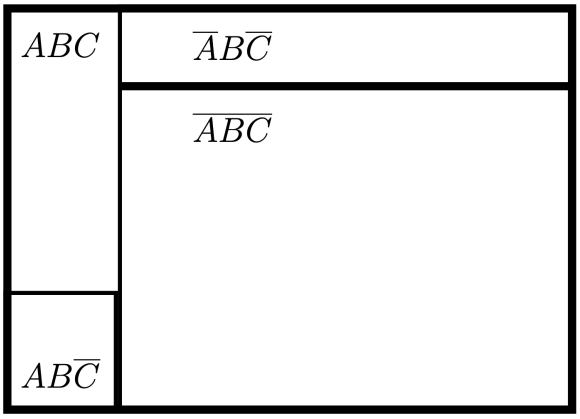
9.7 Examples

Example 9.7 Bayes' rule in court. *(After Al Drake) With probability 0.8 Al is guilty of the crime for which he is about to be tried. Bo and Ci, each of whom knows whether or not Al is guilty, are called to testify.*

Bo is a friend of Al's and will tell the truth if Al is innocent but will lie w.p 0.2 if Al is guilty. Ci hates everybody but the judge and will tell the truth if Al is guilty but will lie w.p 0.3 if Al is innocent.

a. *Draw the outcome space for this problem.*

Solution Denote by A, B, C the events "Al is innocent", "Bo says Al is innocent", "Ci says Al is innocent".



b. What is the probability that Bo testifies that Al is innocent? What is the probability that Ci testifies that Al is innocent?

Solution This means computing the marginal probability of B.

$$\begin{aligned} P(B) &= P(A,B) + P(\overline{A},B) \\ &= P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) \\ &= 1 \times 0.2 + 0.2 \times 0.8 \\ &= 0.36 \end{aligned}$$

Similarly, the marginal of C is

$$\begin{aligned} P(C) &= P(A,C) + P(\overline{A},C) \\ &= P(C|A)P(A) + P(C|\overline{A})P(\overline{A}) \\ &= 0.7 \times 0.2 + 0 \times 0.8 \\ &= 0.14 \end{aligned}$$

c. Which witness is more likely to commit perjury?

Solution

$$P(\text{Bo commits perjury}) = P(B|\overline{A})P(\overline{A}) = 0.2 \times 0.8 = 0.16 \qquad (9.17)$$

$$P(\text{Ci commits perjury}) = P(\overline{C}|A)P(A) = 0.3 \times 0.2 = 0.06 \qquad (9.18)$$

So, Bo is more likely to commit perjury.

d. What is the probability that the witnesses give conflicting testimony?

Solution The witnesses give conflicting testimony if either of them lies but not both (note they will never both lie anyways).

$$\begin{aligned} P(\text{conflicting testimony}) &= P(B,\overline{C}) + P(\overline{B},C) \\ &= P(A,B,\overline{C}) + P(\overline{A},B,\overline{C}) + P(A,\overline{B},C) + P(\overline{A},\overline{B},C) \\ &= P(A)P(B,\overline{C}|A) + P(\overline{A})P(B,\overline{C}|\overline{A}) + 0 + 0 \\ &= 0.2 \times 0.3 + 0.8 \times 0.2 \\ &= 0.22 \end{aligned}$$

e. What is the probability that Al is guilty, given that the witnesses give conflicting testimony?

Solution This is an application of Bayes' rule:

$$\begin{aligned} P(\overline{A}|\text{conflicting testimony}) &= \frac{P(\text{conflicting testimony}|\overline{A})P(\overline{A})}{P(\text{conflicting testimony})} \\ &= \frac{[P(B,\overline{C}|\overline{A}) + P(\overline{B},C|\overline{A})]P(\overline{A})}{P(\text{conflicting testimony})} \\ &= \frac{(0.2+0) \times 0.8}{0.22} \\ &= 0.73 \end{aligned}$$

f. What is the probability that Al is guilty, given that both witnesses say he's innocent? What if both witnesses say he's guilty?

Solution These are also applications of Bayes' rule. The denominator isn't readily computed this time, so we have to apply the rule of total probability to compute it.

$$\begin{aligned} P(B,C) &= P(A,B,C) + P(\overline{A},B,C) \\ &= P(A)P(B,C|A) + P(\overline{A})P(B,C|\overline{A}) \\ &= 0.2 \times 0.7 + 0 \\ &= 0.14 \end{aligned}$$

Since there are only 3 alternatives: either both witnesses say Al is innocent, or both say Al is guilty, or they give conflicting testimony, we have that

$$P(\overline{B},\overline{C}) \;=\; 1 - P(B,c) - P(\text{conflicting testimony}) \;=\; 0.64$$

Now we can apply Bayes' rule:

$$\begin{aligned} P(\overline{A}|B,C) &= \frac{P(B,C|\overline{A})P(\overline{A})}{P(B,C} \;=\; 0 \\ P(\overline{A}|\overline{B},\overline{C}) &= \frac{P(\overline{B},\overline{C}|\overline{A})P(\overline{A})}{P(\overline{B},\overline{C}} \;=\; \frac{0.8 \times 0.8}{0.64} \;=\; 1 \end{aligned}$$

Could we have obtained these results in a more elegant way?

Example 9.8 Communication over a noisy channel *(After Al Drake) Horton and Helen each know that the a-priori probability of Helen's mother being at home (call this event M) on any given evening is 0.6. However, Helen can determine her mother's plan for the evening only at 6 p.m and then, at 6:15 p.m. she has only one chance of sending a signal across the Lake Washington ship canal to Horton. She can either whistle or holler and they decide that she will* **holler if mom is at home** *and whistle otherwise.*

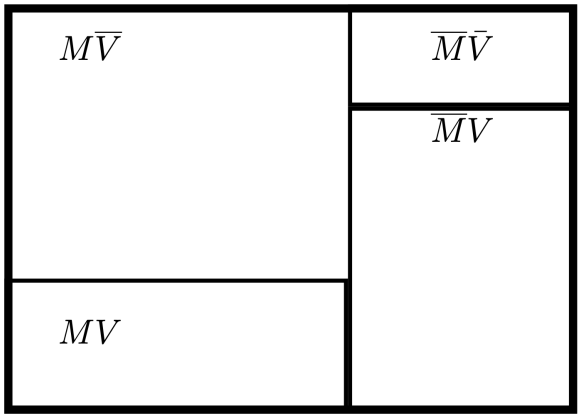
But Helen has a meek voice and the traffic on and across the canal at 6:15 p.m is heavy so that sometimes Horton confuses the signals. Their problem is one of communicating over a noisy channel, where the channel is described by

$$P(\text{Horton hears holler}\,|\,\text{holler}) \;=\; \frac{2}{3} \qquad P(\text{Horton hears holler}\,|\,\text{whistle}) \;=\; \frac{1}{4}$$

Horton will visit Helen if he thinks mom will be away and will play computer games if he thinks mom will be at home. Let's denote the event "Horton believes Helen whistled" which is equivalent with "Horton visits" by V.

a. *What is the probability that Horton visits Helen given that mom will be away? What is this probability given mom will be at home?*

Solution. The sample space is drawn below



$$\begin{aligned} P(V\,|\,\overline{M}) &= P(\text{Horton hears whistle}\,|\,\text{whistle}) = \frac{3}{4} \\ P(V\,|\,M) &= P(\text{Horton hears whistle}\,|\,\text{holler}) = \frac{1}{3} \end{aligned}$$

b. What is the marginal probability that Horton visits Helen on a given evening?

$$P(\text{Horton visits}) = P(\text{Horton visits}\,|\,M)P(M)+P(\text{Horton visits}\,|\,\overline{M})P(\overline{M}) = \frac{1}{3}\times 0.6+\frac{3}{4}\times 0.4 = 0.5$$

c. What is the probability that Horton misunderstands Helen's signal? (This is called the *probability of error* in communications.

$$\begin{aligned} P(\text{error}) &= P(\overline{V},\overline{M}) + P(V,M) \\ &= P(\overline{V}|\overline{M})P(\overline{M}) + P(V|M)P(M) \\ &= \frac{1}{4}\times 0.4 + \frac{1}{3}\times 0.6 = 0.3 \end{aligned}$$

d. Would this probability be lower if they chose to encode "mom will be away" by a holler and "mom at home" by whistle?

Solution. Call *error'* the event "error in the new encoding". The changes are that $P(\text{holler})$ is now equal to $P(\overline{M}) = 0.4$ (and $P(\text{whistle}) = P(M) = 0.6$. So,

$$P(\text{error}') = \frac{1}{3}\times 0.4 + \frac{1}{4}\times 0.6 = 0.28$$

The second encoding method yields a different (and smaller) probability of error!

9.8 Independence

Two events A, B are *independent* if their probabilities satisfy

$$P(A, B) \;=\; P(A)P(B) \tag{9.19}$$

This definition is equivalent to the following one

$$P(A|B) \;=\; P(A) \tag{9.20}$$

The second definition offers more insight into the meaning of independence: A is independent of B if knowing B doesn't affect the probability of A. Note that the relationship is symmetric: if A is independent of B then B is also independent of A by the symmetry of (9.19). We denote independence by the symbol $\perp$; thus, "A independent of B" is expressed by

$$A \;\perp\; B \tag{9.21}$$

Independence, as shown by (9.19), is a symmetric relationship: if A is independent of B, then B is independent of A. Or, in other words, if A provides no information about B, then B cannot provide any information about A. Remember that $P(A|B)$ is not defined if $P(B) = 0$. Therefore, when talking about the independence of two events, we assume that neither event has probability 0.

Example 9.9 *Tossing a coin with $P_H = p$. The events $A =$"the outcome of the 1st toss is 1" and $B =$"the outcome of the 2nd toss is 1" are independent.*

A set of events $A_1, \ldots A_m$ are *mutually independent* if

$$P(A_i|A_j \ldots A_k) \;=\; P(A_i) \tag{9.22}$$

for all subsets $\{j, k, \ldots\} \subseteq \{1, 2, \ldots m\} \setminus \{i\}$. In other words, knowing all and any of the events $A_1, \ldots A_{i-1}, A_{i+1}, \ldots A_m$ does not change the information we have about A_i. The above definition implies that

$$P(A_1 A_2 \ldots A_m) \;=\; P(A_1)P(A_2) \ldots P(A_m). \tag{9.23}$$

If a set of events are mutually independent, they are also pairwise independent, i.e

$$P(A_i A_j) \;=\; P(A_i)P(A_j) \quad \text{for all } i \neq j \tag{9.24}$$

The reverse is not true: pairwise independence does not imply mutual independence.

Example 9.10 *Pairwise independent events that are not mutually independent. Let A be the event "the outcome of a fair coin toss is H" and B be the event*

"the outcome of a second fair coin toss is H". Define C as the event $A \neq B$. Clearly A and B are independent but A, B, C are not mutually independent, since knowing any two of them completely determines the third. However, they are pairwise independent since:

$$\begin{aligned} P(A) &= 0.5 \\ P(B) &= 0.5 \\ P(C) &= P(A,\overline{B}) + P(B,\overline{A}) = P(A)P(\overline{B}) + P(\overline{A})P(B) = 0.25 + 0.25 = 0.5 \\ P(A,B) &= P(A)P(B) \\ P(A,C) &= P(A,\overline{B}) = 0.25 = P(A)P(C) \\ P(B,C) &= P(B,\overline{A}) = 0.25 = P(B)P(C) \end{aligned}$$

9.9 Conditional independence

Two events A, B are *conditionally independent* given event C if

$$P(A,B|C) = P(A|C)P(B|C) \tag{9.25}$$

or, equivalently,

$$P(A|B,C) = P(A|C) \tag{9.26}$$

This means that, if C is true, knowing that B is true does not change the probability of A.

There is no general relationship between the conditional (in)dependence of two events and their unconditional (in)dependence. All combinations are possible.

Example 9.11 *Let the outcome space S be the set of students at UW. Let A be the event "the student plays frisbee", $B =$"the student takes Stat 391" and $C =$"the student is an CS major". Then, we have that A and C are independent, since knowing that someone is a CS major doesn't give any information about her preference for frisbee. C and B are not independent, because knowing that a student takes STAT 391 makes it more likely that she is a CS major than if we didn't have that information. A and C are also independent given B: if I restrict the set of possible students to those who take STAT 391, knowing that one of the students in the class is playing frisbee does not tell me anything new about her being a CS major.*

Example 9.12 *We observe smoke in the classroom if someone smokes or if there is a fire. Assume that smoking is allowed, that fires only start because of defective electrical appliances, and that people choose to light cigarettes independently of the fact that there is a fire or not. Let F denote "there is a fire",*

S = "smoke is observed" and C = "someone is smoking". The events C and F are independent by our definition, if we don't know whether there is smoke in the room or not. But what if we observe smoke (i.e S is true)? If there is smoke and no one is smoking then it is very likely that a fire is on. So knowing something about C gives information about F, if S is also true. Hence, F and C are conditionally dependent given S.

Example 9.13 *A program consists of two modules A and B. The probability of having a bug in module A does not depend of the fact that B has a bug or not. (In other words, bugs are inserted independently in code). Let A be the event "there's a bug in module A", B be the event "there's a bug in module B", and C be the event "the program's output is wrong".*

Before we run the program, A and B are independent: if we test module A, we get information about event A but none about event B.

Suppose we run the program and find that C is true. Then we test A and find that it has no bugs. Knowing that C is true, this means that there must be a bug in B, therefore that B is true. Thus, in the context of C true, observing A gives us strong information about B. Therefore,

$$A \not\perp B \mid C$$

Think what if we run the program and find the output is correct?

Example 9.14 *If $A \subseteq B$ or $A \cap B = \emptyset$ then $A \not\perp B$.*

Example 9.15 *Are shoe size and ability to read dependent in a child? Yes: the shoe size can tell us something about the child's age, and the age in turn gives us information about the ability to read. If you know that a child's shoe size is 3, then you'd guess correctly that she cannot read yet. If the shoe size is 5, then it's more likely she can read. However, once you know the age of the child, shoe size can't give you any extra information about the reading ability (and neither can reading give you on shoe size), so the two are* conditionally *independent given the age.*

$$\text{shoe size} \not\perp \text{reading}$$
$$\text{shoe size} \perp \text{reading} \mid \text{age}$$

Chapter 10

Distributions of two or more random variables

10.1 Discrete random variables. Joint, marginal and conditional probability distributions

Let $X : S \to S_X$ and $Y : S \to S_Y$ be two discrete random variables on the outcome space S. We define the **joint probability** of RVs X, Y to be

$$P_{XY}(x, y) \;=\; P(X = x, Y = y). \tag{10.1}$$

We can think of (X, Y) as a vector-valued RV and of P_{XY} as its distribution. The joint distribution P_{XY} summarizes all the information in P (the distribution on the underlying sample space) that is relevant to X, Y and their interaction. Therefore, after obtaining the joint distribution, we can discard the original sample space altogether as long as we are only concerned with X and Y.

Each of X, Y has also its own distribution $P_X(x)$, $P_Y(y)$. They are related to the joint distribution by

$$P_X(x) \;=\; \sum_{y \in S_Y} P_{XY}(x, y) \tag{10.2}$$

$$P_Y(y) \;=\; \sum_{x \in S_X} P_{XY}(x, y) \tag{10.3}$$

In this context P_X or P_Y are called the **marginal** probabilities of X, respectively Y. The summations above, by which we obtain the marginal distributions P_X, P_Y from the joint distribution P_{XY} is called **marginalization** over Y (respectively X).

The **conditional** probability of X given Y is a function of x and y representing the conditional probability of the event $X = x$ given that $Y = y$ for $x \in S_X,\ y \in S_Y$. The conditional probability can be written as a function of the marginal and joint probabilities of X and Y:

$$P_{X|Y}(x|y) \ = \ \frac{P_{XY}(x,y)}{P_Y(y)} \tag{10.4}$$

Like in the case of events, from (10.4) and (10.2) we get the **law of total probability**

$$P_X(x) \ = \ \sum_{y \in S_Y} P_Y(y) P_{X|Y}(x|y) \tag{10.5}$$

$$P_Y(y) \ = \ \sum_{x \in S_X} P_X(x) P_{Y|X}(y|x) \tag{10.6}$$

10.2 Joint, marginal and conditional densities

Here we will define joint, marginal and conditional distributions for continuous random variables. As you will see, these definitions and other properties are all obtained from their counterparts for discrete random variables by replacing the probabilities with densities and the sums with integrals.

Let $X, Y : S \longrightarrow \mathcal{R}$ be two continuous random variables over a continuous subset S of $\mathcal{R}$. An event in the X, Y space is a set[1] in $\mathcal{R}^2$. We want to define a joint probability distribution for the two variables, i.e a function P_{XY} that associates a positive values to each event in the X, Y space. For continuous one-dimensional distributions, we defined this probability by the density f. We shall do the same here.

The **joint density** for X, Y is an integrable function of x, y that satisfies

$$f_{XY} \ \geq \ 0 \tag{10.7}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) dx dy \ = \ 1 \tag{10.8}$$

If $A \subseteq \mathcal{R}^2$ is an event, then the probability of A is given by

$$P_{XY}(A) \ = \ \int_A f_{XY}(x,y) dx dy \ \leq \ 1 \tag{10.9}$$

[1]Recall that, in theory, for a continuous distribution on $\mathcal{R}$ there exist sets that are not events, but they practically never occur. The situation is similar for two-dimensional or multi-dimensional continuous distributions. One can show by measure theory that there exist sets that are not events, but we shall not be concerned with them since they almost never occur in practice. So from now on we will safely assume that all subsets of a continous S are events.

If we are interested in the distribution of X or Y separately, then these distributions can be obtained from f_{XY} by **marginalization**:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \tag{10.10}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx \tag{10.11}$$

The functions f_X, f_Y are called the **marginal densities** of X and Y respectively (or simply **marginals**). You can easily verify that they integrate to 1 and are positive.

The conditional probability of X given a fixed value $Y = y$ is a continous probability over the range of X. We define it by the **conditional density** $f_{X|Y}(\ |y)$

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{10.12}$$

Note that the denominator in the above expression is the integral of the numerator over all values of x. For every fixed y, $f_{X|Y}$ is a function of x; if y is also allowed to vary, then $f_{X|Y}$ is a function of x and y.

Just like in the discrete case, we also have

$$f_{XY}(x,y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y) \tag{10.13}$$

and the **law of total probability**

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)dy \tag{10.14}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_X(x)f_{Y|X}(y|x)dx \tag{10.15}$$

10.3 Bayes' rule

Bayes' rule is as essential in reasoning with random variables as it is in reasoning with events. It can be derived from (10.13) for continuous RV's

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \tag{10.16}$$

For discrete RV's it follows easily from (10.4)

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} \tag{10.17}$$

Bayes' rule is playing a very important role in estimation and prediction problems. So important, in fact, that a whole subfield of statistics is called **Bayesian statistics**.

Why? Imagine yourself solving an estimation problem, where you observe data $\mathcal{D}$ and want to find the model that generated the data. Let's say for example that the model is a normal distribution with variance 1 and all you need to estimate is its μ. Assume also that from past experience with similar experiments you know a probability distribution for μ, let's call it $P(\mu|\text{past})$. This distribution has the name **prior distribution** or **prior knowledge** or simply **prior**. This is the knowledge about μ that you have prior to seeing the data. After having seen the data you can compute the likelihood, which is $P(\mathcal{D}|\mu)$ – hence a function of μ.

Now let us apply Bayes' rule with $A = \mu$ and $B = \mathcal{D}$. We have

$$P(\mu|\mathcal{D}, \text{past}) \;=\; \frac{P(\mathcal{D}|\mu)P(\mu|\text{past})}{P(\mathcal{D}|\text{past})} \tag{10.18}$$

The left hand of the above equation, $P(\mu|\mathcal{D}, \text{past})$ is what we want: the distribution of the parameter of interest μ given what we know: the dataset and the prior knowledge. On the right hand side are the things we can compute: the probability of the parameter before seeing the data $P(\mu \,|\, \text{past})$, the likelihood of the data. The denominator $P(\mathcal{D} \,|\, \text{past})$ is equal to

$$P(\mathcal{D}|\text{past}) \;=\; \int_{-\infty}^{\infty} P(\mu'|\text{past})P(\mathcal{D}|\mu')d\mu' \tag{10.19}$$

thus being the normalization constant that turns the function in the numerator into a probability density. But what is of importance is that $P(\mathcal{D})$ is not a function of μ and that, at least conceptually, it can be computed from the known functions $P(\mu \,|\, \text{past})$ and $P(\mathcal{D}|\mu)$.

Hence, in contrast to ML estimation that always returns a single number as a best guess, Bayes' rule gives us μ as a distribution over possible values. In many cases, such an answer is more informative and more useful than a single number; for example, when there are several completely different "good guesses".

Another fundamental diference from ML estimation is that Bayesian estimation allows us to fuse two sources of knowledge: previous experience and new information provided by the current experiment.

Often for convenience we drop the reference to the past from Bayes formula, leaving it as

$$P(\mu|\mathcal{D}) \;=\; \frac{P(\mathcal{D}|\mu)P(\mu)}{P(\mathcal{D})} \tag{10.20}$$

10.4 Independence and conditional independence

The RV's X and Y are called **independent** if and only if the events $X = x$ and $Y = y$ are independent for all $x \in S_X$ and $y \in S_Y$.

$$X \perp Y \implies P_{XY}(x,y) = P_X(x)P_Y(y) \;\; x \in S_X,\; y \in S_Y \tag{10.21}$$

Of course, the equivalent definition is also true: X and Y are independent if knowing the value of X does not give any information about the value of Y.

$$X \perp Y \implies P_{X|Y}(x|y) = P_X(x) \;\; x \in S_X,\; y \in S_Y \tag{10.22}$$

In a similar way, we define conditional independence for random variables. Two RVs X, Y are conditionally independent given RV Z if

$$P_{X|YZ}(x|y,z) = P_{X|Z}(x|z) \tag{10.23}$$

If two events or RVs are not (conditionally) independent we say that they are (conditionally) **dependent**.

Two continous RV's X, Y are **independent** if

$$X \perp Y \iff f_{XY}(x,y) = f_X(x)f_Y(y) \;\; x,y \in \mathcal{R} \tag{10.24}$$

Of course, the equivalent definition is also true: X and Y are independent if knowing the value of X does not give any information about the value of Y.

$$X \perp Y \iff f_{X|Y}(x|y) = f_X(x) \;\; x,y \in \mathcal{R} \tag{10.25}$$

We define conditional independence on a third (continuous) RV, Z as

$$X \perp Y \mid Z \iff f_{X|YZ}(x|y,z) = f_{X|Z}(x|z) \quad \text{for all } x,y,z \tag{10.26}$$

The significance is the same as in the discrete case: if Z is known then knowing Y does not add any information about X. If two RVs are not (conditionally) independent we say that they are (conditionally) **dependent**.

10.5 The sum of two random variables

The results below hold for both continuous and discrete RVs. First we define the expectation of a function of one or more RVs. Let $g(X), h(X,Y)$ be integrable functions of one and two RVs respectively. Then, by definition,

$$E[g] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \tag{10.27}$$

$$E[h] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f_{XY}(x,y)dxdy \tag{10.28}$$

Now we show that the expectation is a linear operation, i.e. the expectation of the (weighted) sum of two RVs is the (weighted) sum of the expectations of the individual RVs. The sum is a function of two variables and we can apply the formula above.

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{XY}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\left(x\int_{-\infty}^{\infty}f_{XY}(x,y)dy\right)dx + \int_{-\infty}^{\infty}\left(y\int_{-\infty}^{\infty}f_{XY}(x,y)dx\right)dy \\ &= \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dy \\ &= E[X]+E[Y] \end{aligned} \tag{10.29}$$

We can also show that if a is a real number then

$$E[aX] = \int_{-\infty}^{\infty}axf_X(x)dx \tag{10.30}$$

$$= a\int_{-\infty}^{\infty}xf_X(x)dx \tag{10.31}$$

$$= aE[X] \tag{10.32}$$

Putting the two above results together, we obtain

$$E[aX+bY] = aE[X]+bE[Y] \tag{10.33}$$

for any constants a, b. This result can be generalized by induction to a linear combination of any number of random variables. Note that the RV do not have to be independent. If we replace the integrals with sums we obtain a similar result for discrete distributions.

Example 10.1 The arithmetic mean of n samples from the same distribution. *Let $X_1, \ldots X_n$ be independent samples from the density f. We want to find the expectation of their arithmetic mean*

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}X_i \tag{10.34}$$

Because $X_1, \ldots X_n$ are all drawn from the same distribution, their densities are $f_{X_i} = f$ and their expectations are identical and equal to $E[X_1]$. The expectation

of the arithmetic mean is

$$\begin{aligned} E[\bar{\mu}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] && (10.35)\\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] && (10.36)\\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_1] && (10.37)\\ &= E[X_1] && (10.38) \end{aligned}$$

Hence, the expectation of the arithmetic mean of n identically distributed RVs is equal to the expectation of each of the RVs. We shall see further on that the variance of the arithmetic mean for independent, identically distributed (i.i.d) RVs is different, and much lower than that of the individual RVs.

For two independent RVs, we can compute the distribution of their sum. Let $Z = X + Y$ and F_Z be its CDF. Then

$$\begin{aligned} F_Z(z) &= P(Z \leq z) && (10.39)\\ &= P(X + Y \leq z) && (10.40)\\ &= \int_{x+y\leq z} f_{XY}(x,y)dxdy && (10.41)\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-x} f_{XY}(x,y)dy\right)dx && (10.42)\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-x} f_X(x)f_Y(y)dy\right)dx && (10.43)\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-x} f_X(x)f_Y(y)dy\right)dx && (10.44)\\ &= \int_{-\infty}^{\infty} f_X(x)F_Y(z-x)dx && (10.45)\\ f_Z(z) &= \frac{d}{dz}F_Z(z) && (10.46)\\ &= \frac{d}{dz}\int_{-\infty}^{\infty} f_X(x)F_Y(z-x)dx && (10.47)\\ &= \int_{-\infty}^{\infty} f_X(x)\frac{d}{dz}F_Y(z-x)dx && (10.48)\\ &= \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx && (10.49) \end{aligned}$$

The above operation is called the **convolution** of the two densities. Note that in the proof we made use of the fact that X, Y are independent: the result is

true only for independent RVs. The probability of the sum of two discrete RVs U, V is the **discrete convolution** of their distributions P_U, P_V:

$$P_{U+V}(n) = \sum_{k=-\infty}^{\infty} P_U(k) P_V(n-k) \tag{10.50}$$

10.6 Variance and covariance

Let us now study the variance of a sum of two RVs, trying to write it as a function of the variances of the individual RVs.

$$\begin{aligned}
Var(X+Y) &= E[(X+Y-E[X]-E[Y])^2] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y-E[X]-E[Y])^2 f_{XY}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(x-E[X])^2 + 2(x-E[X])(y-E[Y]) + (y-E[Y])^2\right] f_{XY}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-E[X])^2 f_{XY}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-E[Y])^2 f_{XY}(x,y)dxdy \\
&\quad +2\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-E[X])(y-E[Y]) f_{XY}(x,y)dxdy}_{Cov(X,Y)} \\
&= \int_{-\infty}^{\infty} (x-E[X])^2 \int_{-\infty}^{\infty} f_{XY}(x,y)dydx + \int_{-\infty}^{\infty} (y-E[Y])^2 \int_{-\infty}^{\infty} f_{XY}(x,y)dxdy \\
&\quad + 2Cov(X,Y) \\
&= \int_{-\infty}^{\infty} (x-E[X])^2 f_X(x)dx + \int_{-\infty}^{\infty} (y-E[Y])^2 f_Y(y)dy \; + \; 2Cov(X,Y) \\
&= VarX + VarY + 2Cov(X,Y)
\end{aligned} \tag{10.51}$$

The quantity denoted by $Cov(X,Y)$ is called the **covariance** of the two random variables. The covariance is a measures the "co-variation" of the two RVs around their respective means. If large deviations of X are paired with large deviations of Y in the same direction, then $Cov(X,Y)$ is a large, positive number. If the deviations of X are paired with deviations of Y in opposite direction, then $Cov(X,Y)$ is a negative number of large magnitude. If the deviations of X and Y around their means are unrelated, the covariance is close to 0.

If the random variables are independent, then $Cov(X,Y) = 0$.

$$\begin{aligned} Cov(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-E[X])(y-E[Y])f_{XY}(x,y)dxdy && (10.52)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-E[X])(y-E[Y])f_X(x)f_Y(y)dxdy && (10.53)\\ &= \underbrace{\left[\int_{-\infty}^{\infty}(x-E[X])f_X(x)dx\right]}_{0}\underbrace{\left[\int_{-\infty}^{\infty}(y-E[Y])f_Y(y)dy\right]}_{0} && (10.54)\\ &= 0 && (10.55)\end{aligned}$$

The **correlation coefficient** of X and Y is

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var\,X\,Var\,Y}} \quad (10.56)$$

One can show that the correlation coefficient is always between -1 and 1. When $|\rho_{XY}|$ is close to 1, the variables are **strongly correlated**; when it is 0, they are **uncorrelated**. Independent variables are always uncorrelated; the converse in in general not true.

It is also useful to note that the variance of a RV scales quadratically with the RV. If a is a real number, then

$$Var[aX] = E[(aX - aE[X])^2] = a^2 Var\,X \quad (10.57)$$

Then it is simple to derive what happens to the variance of the linear combination of two RVs

$$\begin{aligned} Var\,(aX+bY) &= E[(aX - aE[X] + bY - bE[Y])^2]\\ &= E[a^2(X-E[X])^2 + b^2(Y-E[Y])^2 + 2ab(X-E[X])(Y-E[Y])]\\ &= a^2 Var\,X + b^2 Var\,Y + 2ab\,Cov(X,Y) && (10.58)\end{aligned}$$

Example 10.2 The arithmetic mean of n samples from the same distribution. *Let $X_1, \ldots X_n$ be independent samples from the density f. We now want to find the variance of their arithmetic mean $\bar{\mu}$*

$$\begin{aligned} Var\,\bar{\mu} &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) && (10.59)\\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\,X_i && (10.60)\\ &= \frac{nVar\,X_1}{n^2} && (10.61)\\ &= \frac{Var\,X_1}{n} && (10.62)\end{aligned}$$

In the above we used the fact that the RVs are independent, hence uncorrelated, and the fact that they have all the same variance. It results that the variance of the arithmetic mean decreases proportionally to 1/(the number of terms in the sum).

10.7 Some examples

Example 10.3 A discrete distribution

The experiment is tossing a fair coin 4 times. Thus, $S = \{0,1\}^4 = \{(X_1, X_2, X_3, X_4)\}$ *and the probability of every outcome is* $1/2^4 = 1/16$. *We define the random variables:*

- Y = *position of first 1, or 0 if no ones*
- Z = *number of ones* $= X_1 + X_2 + X_3 + X_4$

The values of Y, Z *(in this order) for every outcome are shown in the table below.*

X_1X_2:	00	01	11	10
$X_3X_4 = 00$	0, 0	2, 1	1, 2	1, 1
01	4, 1	2, 2	1, 3	1, 2
11	3, 2	2, 3	1, 4	1, 3
10	3, 1	2, 2	1, 3	1, 2

The joint distribution $P_{YZ}(y,z)$ is represented in the next table. For clarity, the values in the table are multiplied by 16 (so, for example, $P_{YZ}(0,0) = 1/16$, $P_{YZ}(3,1) = 3/16$, etc.).

Y	0	1	2	3	4	
$Z = 0$	1	0	0	0	0	
1	0	1	1	1	1	
2	0	3	2	1	0	
3	0	3	1	0	0	
4	0	1	0	0	0	
						16

By adding up the rows of the table above, we obtain the marginal P_Y. Similarly, the marginal of Z is obtained by adding up the elements in each column of P_{YZ}. Below is the joint table with the marginals added (the values are again multiplied by 16). Note that "adding up the elements in a row/column" corresponds to

implementing the definition of the marginal

$$P_Y(y) \;=\; \sum_{z=0}^{4} P_{YZ}(y,z)$$

Y	0	1	2	3	4	
$Z=0$	1	0	0	0	0	1
1	0	1	1	1	1	4
2	0	3	2	1	0	6
3	0	3	1	0	0	4
4	0	1	0	0	0	1
	1	8	4	2	1	16

The conditional distribution $P_{Z|Y}$ is shown below. Each row contains a distribution over Z, $P_{Z|Y=y}$. They are obtained by normalizing the corresponding row in the P_{YZ} table by the marginal value $P_Y(y)$ in the last column. In other words by implementing the formula

$$P_{Z|Y}(z|y) \;=\; \frac{P_{YZ}(y,z)}{P_Y(y)}$$

Y	0	1	2	3	4
$Z=0$	1	0	0	0	0
1	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1
2	0	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{2}$	0
3	0	$\frac{3}{8}$	$\frac{1}{4}$	0	0
4	0	$\frac{1}{8}$	0	0	0
	1	1	1	1	1

And here is $P_{Y|Z}$:

Y	0	1	2	3	4	
$Z=0$	1	0	0	0	0	1
1	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1
2	0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	0	1
3	0	$\frac{3}{4}$	$\frac{1}{4}$	0	0	1
4	0	1	0	0	0	1

Example 10.4 Rob's fuel consumption

Rob (who is a robot roaming in the basement of Sieg hall) is switching to fossil fuel consumption to save on electric energy. Every morning he is taking in his tank X gallons of fuel, where X is uniformly distributed between 0 and 3.

$$X \;\sim\; \text{Uniform}(0,3]$$

At the end of the day, the amount of fuel remaining in his tank is Y

$$Y \sim \text{Uniform}(0, X)$$

a. *What is the joint density of* X, Y *?*

The joint outcome space is $S = \{0 < Y < X \leq 3\}$. We know that

$$f_X(x) = \frac{1}{3} \text{ for } x \in (0, 3]$$

and

$$f_{Y|X}(y|x) = \frac{1}{x} \text{ for } y \in (0, x]$$

Therefore

$$f_{XY}(x, y) = f_X(x) f_{Y|X}(y|x) = \frac{1}{3x} \text{ for } (x, y) \in S \text{ and 0 otherwise}$$

Note that this is NOT a uniform distribution on S!

b. What is the marginal distribution of Y ?

$$\begin{aligned} f_Y(y) &= \int_y^3 f_{XY}(x, y)\, dx && (10.63) \\ &= \int_y^3 \frac{1}{3x}\, dx && (10.64) \\ &= \frac{1}{3}(\ln 3 - \ln y) && (10.65) \end{aligned}$$

Note that this density is unbounded towards 0.

c. What is the expectation of Y ?

$$\begin{aligned} E[Y] &= \int_0^3 y f_Y(y)\, dy && (10.66) \\ &= \int_0^3 y \frac{1}{3}(\ln 3 - \ln y)\, dy && (10.67) \\ &= \frac{1}{3}\left[\frac{y^2}{2}\ln 3 - (\frac{y^2}{2}\ln y - \frac{y^2}{4})\right]_0^3 && (10.68) \\ &= \frac{3}{4} && (10.69) \end{aligned}$$

d. What is the conditional distribution of X given Y ?

The domain of X is $(y, 3]$ and

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)} && (10.70)\\ &= \frac{\frac{1}{3x}}{\frac{1}{3}(\ln 3 - \ln y)} && (10.71)\\ &= \frac{1}{\ln 3 - \ln y}\frac{1}{x} && (10.72) \end{aligned}$$

e. If $Y = 1$ gallon, what is the probability that on that day Rob started with $X < 2$ gallons in his tank?

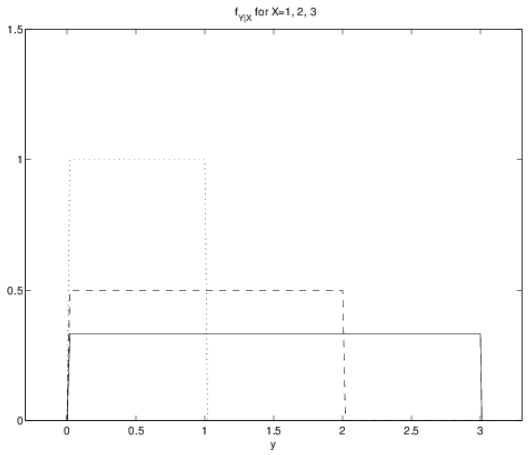
This is

$$\begin{aligned} P(X < 2|Y = 1) &= \int_1^2 f_{X|Y}(x, 1)\, dx && (10.73)\\ &= \int_1^2 \frac{1}{\ln 3 - \ln 1}\frac{1}{x}\, dx && (10.74)\\ &= \frac{\ln 2}{\ln 3} && (10.75) \end{aligned}$$

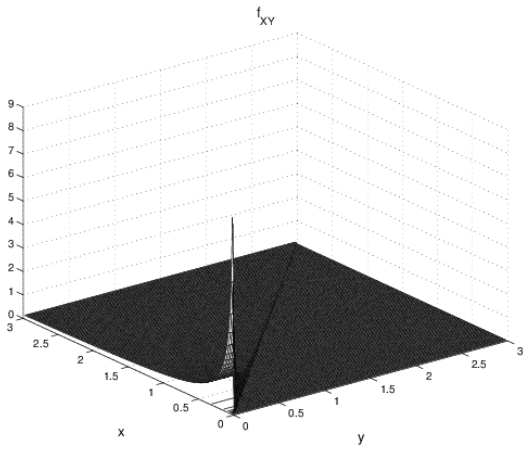
f. What is Rob's daily average fuel consumption?

The fuel consumption is $X - Y$ so the daily average fuel consumption is

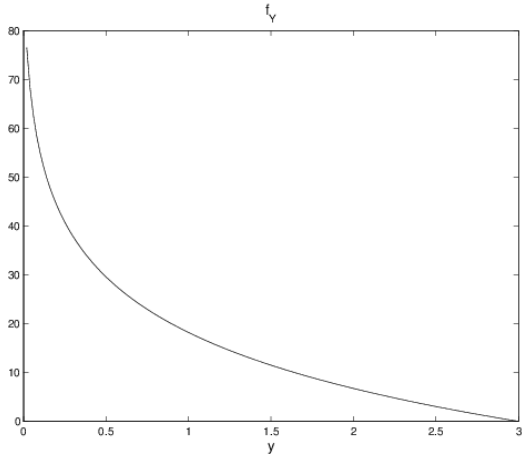
$$E[X - Y] \;=\; E[X] - E[Y] \;=\; \frac{3-0}{2} - \frac{3}{4} \;=\; \frac{3}{4} \qquad (10.76)$$



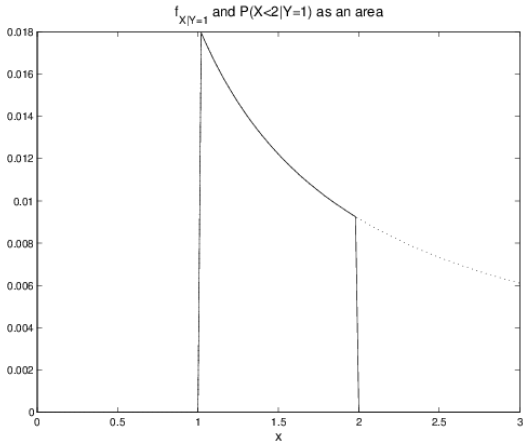
$f_{Y|X}$ for $X = 1,\ 2,\ 3$



The joint density f_{XY}



The marginal of Y, f_Y



$f_{X|Y}$ for $Y = 1$

Example 10.5 *In* **reliability**, *the probability density of failure over time is often modeled as an exponential distribution. In other words, for a certain type of component (e.g. a light bulb), the probability that the component fails in the interval* $[t, t + \Delta)$ *(for a very small* Δ*) equals* $f(t)\Delta$*, with* f *being the p.d.f of the exponential distribution*

$$f(t) \ = \ \gamma e^{-\gamma t} \quad t \geq 0 \tag{10.77}$$

We compute the conditional probability that a component will fail in the next

Δ interval, given that it is working at time t.

$$\begin{aligned} P[\text{fail in } [t,\, t+\Delta) \,|\, \text{working at } t] &= \frac{P[\text{fail in } [t,\, t+\Delta), \text{ working at } t]}{P[\text{working at t}]} && (10.78)\\ &= \frac{f(t)\Delta}{1 - P[\text{failed before } t]} && (10.79)\\ &= \frac{f(t)\Delta}{1 - F(t)} && (10.80)\\ &= \frac{\gamma e^{-\gamma t}\Delta}{1 - (1 - e^{-\gamma t})} && (10.81)\\ &= \gamma\Delta && (10.82) \end{aligned}$$

Note that this probability does NOT depend on t! The parameter γ represents the fraction of failures in a certain short interval divided by the lenght of the interval, and therefore it is known in reliability as the **rate of failure** of the components. The exponential distribution describes those components that have constant rates of failure.

Can you think of other processes that exhibit constant "rates of decay" or "accumulation"?

10.8 The bivariate normal distribution

10.8.1 Definition

Two RVs X, Y are said to be **jointly normal**, or **jointly Gaussian**, or to obey the **bivariate normal** distribution, if their joint density is

$$f_{XY}(x,y) \;=\; \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}[x-\mu_x \;\; y-\mu_y]\Sigma^{-1}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right) \qquad (10.83)$$

The parameters of the bivariate normal distribution are the **mean vector**

$$\mu \;=\; \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

and the **covariance matrix**

$$\Sigma \;=\; \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

The covariance matrix is **positive definite** or, equivalently, its determinant is always positive.

$$|\Sigma| \;=\; \sigma_x^2\sigma_y^2 - \sigma_{xy}^2 \;\geq\; 0 \qquad (10.84)$$

We denote the fact that X, Y have a jointly normal density by

$$(X,Y) \;\sim\; N(\mu,\, \Sigma)$$

10.8.2 Marginals

Proposition 1.
If $X, Y \sim N(\mu, \Sigma)$, then

$$X \sim N(\mu_x, \sigma_x^2) \tag{10.85}$$
$$Y \sim N(\mu_y, \sigma_y^2) \tag{10.86}$$

In other words, the distributions of the individual variables are also normal and their parameters are found by copying the corresponding parameters from the joint distribution.

Proof. For simplicity, let us denote

$$x' = x - \mu_x \qquad y' = y - \mu_y \tag{10.87}$$

and

$$\Sigma^{-1} = \begin{bmatrix} D_x & D_{xy} \\ D_{xy} & D_y \end{bmatrix} = \frac{1}{|\Sigma|} \begin{bmatrix} \sigma_y^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_x^2 \end{bmatrix} \tag{10.88}$$

The marginal of X is defined by

$$f_X(x) = \int f_{XY}(x,y)dy = \int \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}[D_x x'^2 + 2D_{xy}x'y' + D_y y'^2]}\, dy' \tag{10.89}$$

The expression in the exponent is a quadratic in y' and we will separate from it something that looks like a normal distribution in y' alone, leaving out terms depending on x' only. Remember that the ultimate goal is to integrate over y' and in this context x' is a constant. We are guaranteed that the expressions will be normalized so we can ignore tho the constant factors in front of the exponentials.

$$f_X(x) \propto \int e^{-\frac{1}{2}[D_x x'^2 + 2D_{xy}x'y' + D_y y'^2]}\, dy' \tag{10.90}$$
$$\propto \int e^{-\frac{1}{2}D_y[y'^2 + 2\frac{D_{xy}}{D_y}x'y' + \left(\frac{D_{xy}}{D_y}x'\right)^2 - \left(\frac{D_{xy}}{D_y}x'\right)^2 + D_x/D_y x'^2]} dy' \tag{10.91}$$
$$\propto \int \underbrace{e^{-\frac{1}{2}D_y\left(y' + \frac{D_{xy}}{D_y}x'\right)^2}}_{N(-x'D_{xy}/D_y, 1/D_y)} \underbrace{e^{-\frac{1}{2}[D_y\left(\frac{D_{xy}}{D_y}x'\right)^2 + D_x x'^2]}}_{\text{depends on } x' \text{ only}} dy' \tag{10.92}$$
$$\propto e^{-\frac{1}{2}[D_y\left(\frac{D_{xy}}{D_y}x'\right)^2 + D_x x'^2]} \cdot 1 \tag{10.93}$$
$$\propto e^{-\frac{1}{2}[D_x - \frac{D_{xy}}{D_y}]x'^2} \tag{10.94}$$

This is a normal distribution in x'. The coefficient of x'^2 can be written as

$$D_x - \frac{D_{xy}}{D_y} = \frac{D_x D_y - D_{xy}^2}{D_y} = \frac{|D|}{\sigma_x^2/|\Sigma|} = \frac{|D||\Sigma|}{\sigma_x^2} = \frac{1}{\sigma_x^2} \tag{10.95}$$

By replacing x' and (10.95) into (10.94) we obtain

$$f_X(x) \ \propto \ e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \tag{10.96}$$

QED.

Proposition 1 explains the notations and names of the parameters $\mu_x,\ \mu_y,\ \sigma_x,\ \sigma_y$. But what about the parameter σ_{xy} which has no correspondent in the one variable case? This parameter measures the covariance $Cov(X,Y)$ defined in Handout 10.

Proposition 2. If $X, Y \ \sim \ N(\mu,\ \Sigma)$, then

$$Cov(X,Y) \ = \ \sigma_{xy} \tag{10.97}$$

Proposition 3. If $X, Y \ \sim \ N(\mu,\ \Sigma)$ and $\sigma_{xy} = 0$, the variables X, Y are independent.

Proof. First, note that if $\sigma_{xy} = 0$ the inverse covariance Σ^{-1} is diagonal, with elements $1/\sigma_x^2,\ 1/\sigma_y^2$. Then,

$$\begin{aligned} f_{XY} &= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}]} && (10.98)\\ &= \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}} \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu_y)^2}{\sigma_y^2}} && (10.99)\\ &= f_X(x) f_Y(y) && (10.100) \end{aligned}$$

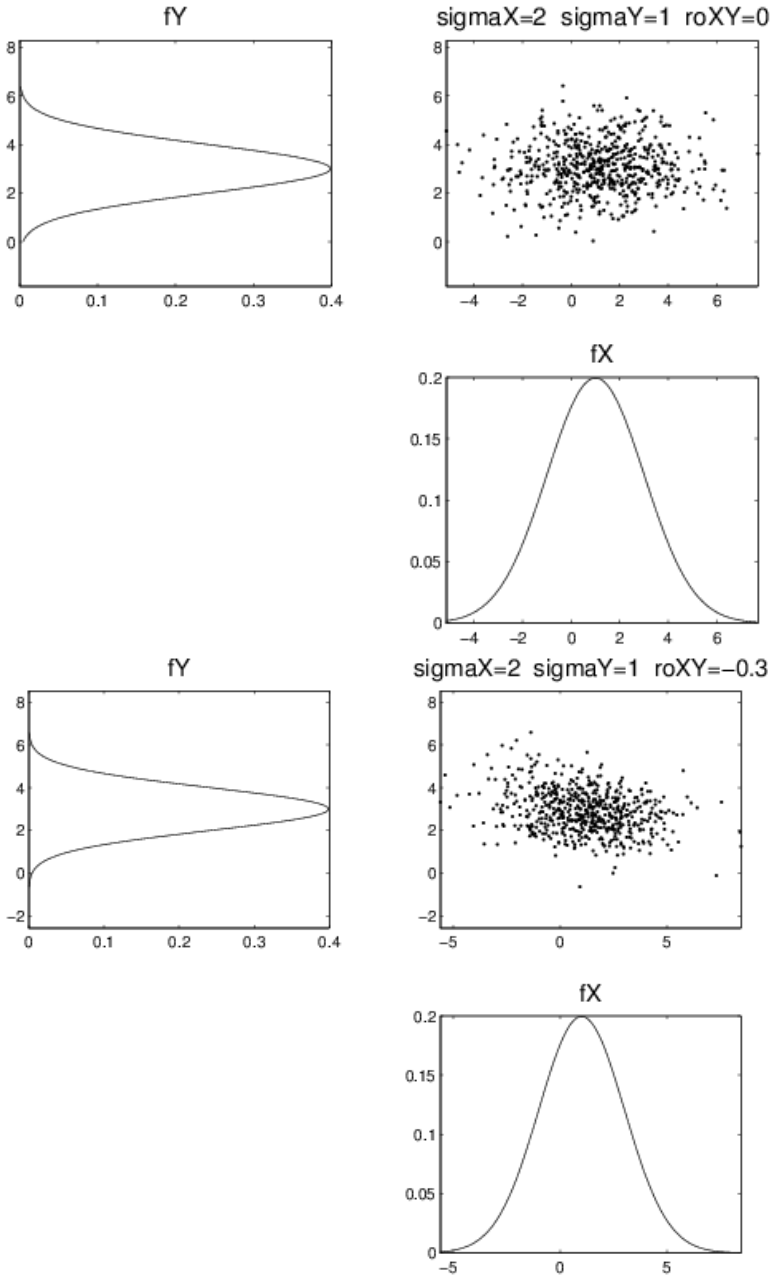
The **correlation coefficient** ρ_{xy} is

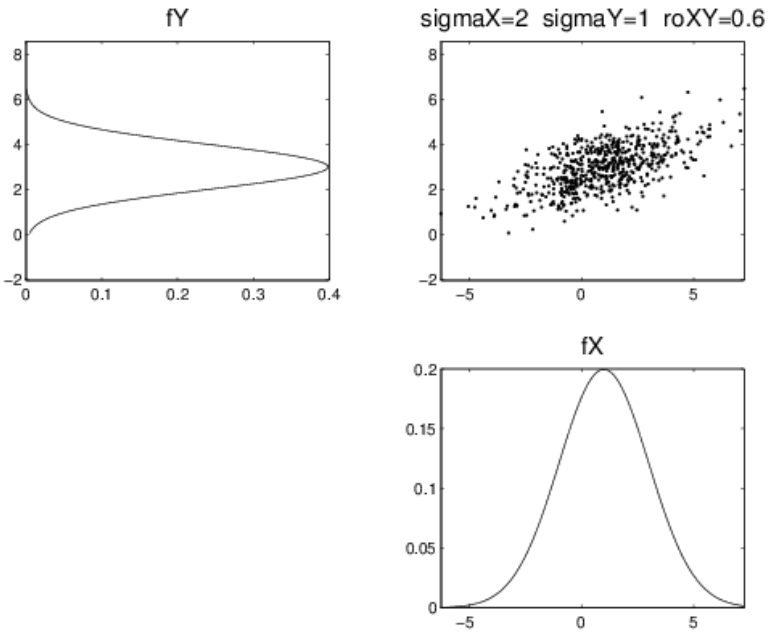
$$\rho_{xy} \ = \ \frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{10.101}$$

Because the determinant of Σ is never negative (10.84), it follows that, as we already knew,

$$-1 \ \leq \ \rho_{xy} \ \leq 1 \tag{10.102}$$

On the next page are depicted 3 jointly normal distributions, together with their marginals f_X, f_Y. The three distributions have the same parameters, except for σ_{xy}.


fY
sigmaX=2 sigmaY=1 roXY=0
fX
fY
sigmaX=2 sigmaY=1 roXY=−0.3
fX



10.8.3 Conditional distributions

Proposition 4 If $(X, Y) \sim N(\mu, \Sigma)$, then the conditional distribution of $f_{X|Y}$ is also normal, with parameters

$$\mu_{x|Y=y} = \mu_x + (y - \mu_y)\frac{\sigma_{xy}}{\sigma_y^2} \tag{10.103}$$

$$\sigma^2_{x|Y=y} = \sigma_x^2(1 - \rho_{xy}^2) \tag{10.104}$$

Intuitively, after we observe Y, the expectation of X deviates from μ_x by an amount proportional to the deviation of the observed y from its own mean μ_y. The proportionality constant is itself proportional to the covariance between the two variables and inversely proportional to the noise in Y as measured by σ_y^2.

Another way of expressing the above equation (prove it as an exercise) is

$$\frac{\mu_{x|Y=y} - \mu_x}{\sigma_x} = \rho_{xy}\frac{y - \mu_y}{\sigma_y} \tag{10.105}$$

The covariance of X after Y is observed is decreased, since we gain information. The decrease is proportional to the square of the correlation coefficient ρ_{xy}.

Proof. With the previous notations we have

$$f_{X|Y}(x,y) \;=\; \frac{f_{XY}(x,y)}{f_Y(y)} \;=\; \frac{\frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}[D_x x'^2 + 2D_{xy} x'y' + D_y y'^2]}}{\frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{y'^2}{2\sigma_y^2}}} \tag{10.106}$$

For any fixed value of y this is obviously a normal distribution. By identifying the above expression with the stardard expression for a (univariate normal) we aim to uncover its parameters. First, note that the coefficient of x^2 in the exponential part of the density must be the inverse variance $1/\sigma^2_{x|Y=y}$. Hence,

$$\sigma_{x|Y=y} \;=\; \frac{1}{D_x} \;=\; \frac{\sigma_x^2\sigma_y^2 - \sigma_{xy}^2}{\sigma_y^2} \;=\; \sigma_x^2(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}) \;=\; \sigma_x^2(1-\rho_{xy}^2) \tag{10.107}$$

Second, the expectation of x' must be the coefficient of x' in the exponent, times $-\frac{1}{2D_x}$. This gives

$$\frac{2D_{xy}}{-2D_x} \;=\; \frac{\sigma_{xy}}{\sigma_y^2} y' \tag{10.108}$$

To obtain the expectation of x we add μ_x to the above expression. QED.

Note that the expected value of X changes with the observed value of Y, but its variance is constant for all values of Y.

10.8.4 Estimating the parameters of a bivariate normal

We now turn to the problem of estimating the parameters of f_{XY} from data by the Maximum Likelihood (ML) method.

The data consists of n independent samples from f_{XY}

$$\mathcal{D} \;=\; \{\,(x_1,y_1),\,(x_2,y_2),\ldots\,(x_n,y_n)\,\} \tag{10.109}$$

The task is to find the parameters μ, Σ that maximize the likelihood of the data

$$(\mu^{ML},\,\Sigma^{ML}) \;=\; \underset{\mu,\Sigma}{\operatorname{argmax}} \prod_{i=1}^{n} f_{XY}(x_i,y_i) \tag{10.110}$$

Because, μ_x, σ_x are at the same time the parameters of the marginal f_X, a

Gaussian distribution in one variable, we can immediately derive that

$$\mu_x^{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10.111}$$

$$(\sigma_x^{ML})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x^{ML})^2 = \frac{1}{n}\sum_{i=1}^{n} x_1^2 - (\mu_x^{ML})^2 \tag{10.112}$$

and

$$\mu_y^{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{10.113}$$

$$(\sigma_y^{ML})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y^{ML})^2 = \frac{1}{n}\sum_{i=1}^{n} y_1^2 - (\mu_y^{ML})^2 \tag{10.114}$$

To obtain the last parameter, σ_{xy}, we have to actually equate the gradient of the log-likelihood with 0, and use the previously obtained estimates of μ_x, μ_y, σ_x, σ_y. Eventually, we get

$$\sigma_{xy}^{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x^{ML})(y_i - \mu_y^{ML}) \tag{10.115}$$

10.8.5 An example

The following $n = 20$ data points were generated from a bivariate normal density with $\mu_x = 1,\ \mu_y = -1,\ \sigma_x = 0.5,\ \sigma_y = 0.3,\ \sigma_{xy} = 0.075$ (which gives $\rho = 0.5$).

	1.383096	-0.712764
	1.365463	-1.190579
	1.127047	-0.911922
	0.433800	-1.308288
	1.518166	-0.932528
	1.199770	-0.937261
	0.634095	-1.220424
	1.125638	-1.226752
	1.438293	-0.762800
	0.571201	-1.154991
	0.204084	-1.740352
	1.433606	-1.051744
	1.183250	-0.901840
	0.913308	-1.282981
	0.499134	-1.198037
	0.316031	-1.414655
	1.727697	-0.841577
	0.398874	-1.129591
	0.315847	-1.492383
	0.964255	-0.754558
	Statistics:	
$\sum x_i$	18.752654	-22.166027
$\sum x_i^2$	21.897914	25.938701
$\sum x_i y_i$	-18.909045	
	ML estimates:	
μ	0.937633	-1.108301
σ	0.464479	0.261922
ρ	0.770421	
	Unbiased ML estimates	
μ	0.937633	-1.108301
σ	0.476545	0.268726
ρ	0.770421	

Chapter 11

Bayesian estimation

Here we study two common examples of Bayesian estimation: the mean of a normal distribution and the parameters of a discrete distribution. Recall that Bayesian estimation assumes that you have a prior distribution on the parameter(s) you want to estimate. You also have samples from the unknown distribution. The task is to combine the information from the samples, in the form of the likelihood, with the prior, and to obtain another, updated, distribution for the parameter, called the posterior.

11.1 Estimating the mean of a normal distribution

The data consists of n independent samples from f_X, where f_X is a normal distribution with (unknown) parameters $\mu_{true}, \sigma^2_{true}$.

$$\mathcal{D} \;=\; \{\, x_1,\, x_2, \ldots\, x_n\} \tag{11.1}$$

The prior distribution of μ is assumed to be also a normal distribution

$$\mu \;\sim\; N(m, s^2) \;=\; f_0 \tag{11.2}$$

The parameters m, s^2 are set by us (according to our presumed knowledge) so they are known. The task is to obtain the posterior distribution of μ.

By Bayes' formula, this is

$$f(\mu \,|\, \mathcal{D}) \;=\; \frac{f_0(\mu) \prod_{i=1}^{n} f_X(x_i|\mu, \sigma)}{f_{\mathcal{D}}} \tag{11.3}$$

If we take logarithms and ignore the factors that don't depend on μ, we obtain

$$\ln f(\mu \,|\, \mathcal{D}) \quad \propto \quad -\frac{1}{2}\left[\frac{(\mu - m)^2}{s^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}\right] \tag{11.4}$$

$$\propto \quad -\frac{1}{2}\left[\underbrace{\left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right)}_{1/s^2_{new}} \mu^2 - 2\underbrace{\left(\frac{m}{s^2} + \frac{\sum_i x_i}{\sigma^2}\right)}_{m_{new}/s^2_{new}} \mu + \ldots\right] \tag{11.5}$$

This shows that the posterior distribution of μ is also a normal distribution. Let us determine its parameters. The coefficient of μ^2 in the above expression represents the inverse variance.

$$\frac{1}{s^2_{new}} \;=\; \frac{1}{s^2} + \frac{1}{\frac{\sigma^2}{n}} \tag{11.6}$$

Therefore

$$s^2_{new} \;=\; \frac{1}{\frac{1}{s^2} + \frac{n}{\sigma^2}} \tag{11.7}$$

Equation (11.6) shows that the inverse variance of the posterior is the sum of the inverse variances given by the data and the prior's inverse variance. Because the sum is larger than either of its terms, it follows that

$$s^2_{new} \;\leq\; \min\left(s^2, \frac{\sigma^2}{n}\right) \tag{11.8}$$

Hence, Bayesian estimation decreases the variance, both w.r.t ML estimation and w.r.t the prior. Note that σ in the above expressions is undetermined. We would like to use σ_{true} but since it is unknown, a reasonable choice in practice is to use σ_{ML}.

Next, let us estimate the mean. This is obtained from the coefficient of μ in (11.5)

$$m_{new} \;=\; \frac{\frac{1}{s^2}m + \frac{n}{\sigma^2}\overbrace{\frac{\sum_i x_i}{n}}^{\mu_{ML}}}{\frac{1}{s^2} + \frac{n}{\sigma^2}} \tag{11.9}$$

Hence, the posterior mean is a weighted average between the prior mean and the ML mean. The weights depend on the respective variances of the prior and ML estimate. In the above, the ideal formula makes use of the unknown σ_{true}; in practice, σ_{true} is replaced with the ML estimate.

The prior variance s^2 is a measure of the strength of the prior (or of the confidence we put into the mean m). The larger s, the weaker the influence of the prior and the closer the m_{new}, s^2_{new} parameters to the ML estimates. Note also that, for sufficiently large n, the posterior will be dominated completely by the data terms. The moral is, in other words, that even if the prior is wrong, with enough data the initial mistakes can be overridden.

11.2 Estimating the parameters of a discrete distribution

A discrete distribution on $S = \{ 0, \ldots, m-1 \}$ is parametrized by $\vec{\theta} = (\theta_0, \ldots, \theta_{m-1}) \in \mathbf{\Theta}$ where

$$\mathbf{\Theta} = \{ \vec{\theta} \mid \sum_{0}^{m-1} \theta_k = 1, \ \theta_k \geq 0, \ k = 0, \ldots m-1, \}. \tag{11.10}$$

The dataset is $\mathcal{D} = \{ x_1, x_2, \ldots x_n \}$. Recall that the likelihood of the data is given by

$$L(\vec{\theta}) \equiv P(\mathcal{D} \mid \vec{\theta}) = \prod_{k=0}^{m-1} \theta_k^{n_k} \tag{11.11}$$

where n_k is the number of times outcome k appears in the dataset. The ML parameter estimates are

$$\theta_k^{ML} = \frac{n_k}{n} \quad \text{for } k = 0, \ldots, m-1 \tag{11.12}$$

The prior is a density over $\mathbf{\Theta}$. We define it (ignoring the normalization constant) as

$$f(\vec{\theta}) \equiv D(\vec{\theta}; \vec{n'}) \propto \prod_{k=0}^{m-1} \theta_k^{n'_k - 1} \tag{11.13}$$

Note that we are free to choose any density over $\mathbf{\Theta}$ as prior (or rather the density that best reflects our knowledge about the parameters before we see the data). We choose this one for reasons that will appear in the forthcoming, one of them being mathematical and algorithmic convenience.

Let us make the notations

$$\vec{n} = (n_0, \ldots n_{m-1}) \quad \vec{n'} = (n'_0, \ldots n'_{m-1}) \tag{11.14}$$

By Bayes' rule, the posterior is proportional to the product of the prior and the likelihood:

$$f(\vec{\theta} \mid \mathcal{D}) \propto \prod_{k=0}^{m-1} \theta_k^{n'_k - 1} \prod_{k=0}^{m-1} \theta_k^{n_k} \tag{11.15}$$

$$\propto \prod_{k=0}^{m-1} \theta_k^{n_k + n'_k - 1} \tag{11.16}$$

$$\propto D(\vec{\theta}; \vec{n} + \vec{n'}) \tag{11.17}$$

Hence if the prior is Dirichlet, the posterior is also in the Dirichlet family of distributions. Moreover, the parameters $\vec{n'}_{new}$ of the posterior are the sum of

the prior parameters and the sufficient statistics of the data. This suggests that the parameters $\vec{n'}$ of a Dirichlet distribution are the sufficient statistics of a "fictitious data set". Thus the prior knowledge that the Dirichlet distribution embodies is equivalent to the knowledge we'd have if we had seen a previous data set with sufficient statistics $\vec{n'}$. Note that the numbers n' are not restricted to integers, and in particular they can be smaller than 1. The sum

$$n' = \sum_{k=0}^{m-1} n'_k \tag{11.18}$$

is the "equivalent sample size" of the fictitious data set and it represents the strength of the prior. The smaller n', the weaker the confidence in our prior knowledge and the weaker the influence the prior has on the parameter distribution.

The mean of the Dirichlet distribution is given by

$$E_{D(\vec{n'})}[\theta_k] = \frac{n'_k}{n'} \tag{11.19}$$

[**Exercise.** Prove (11.19). It's a non-trivial exercise in multivariate integration.] Therefore, the mean values of θ under the posterior distribution is

$$E_{D(n+\vec{n'})}[\theta_k] = \frac{n_k + n'_k}{n + n'} \tag{11.20}$$

Intuitively, if we want to make a "best guess" at the parameters after we compute the posterior, we would obtain an estimate that is like the ML estimate from the data and the fictitious data pooled together.

Chapter 12

Statistical estimators as random variables. The central limit theorem

12.1 The discrete binary distribution (Bernoulli)

Let $S = \{0, 1\}$ and let a distribution P be defined on it by $P(1) = p$. If we have a data set $\mathcal{D} = \{x_1, \ldots x_n\}$ of independent samples from P, the ML estimate of p is the well known

$$\hat{p} \;=\; \frac{\sum_{i=1}^{n} x_i}{n} \tag{12.1}$$

Now, $\hat{p}$ is obviously a function of the data set, hence a function of the outcome of the experiment "sample n points from P", hence a random variable. Let's see what are its mean and variance. In particular, if the formula (12.1) is a good method for estimating p, we would expect that $\hat{p}$ is close to p or even converges to p when n is large.

By the linearity of the mean, we have that

$$E[\hat{p}] \;=\; E[\frac{\sum_{i=1}^{n} X_i}{n}] \;=\; \frac{\sum_{i=1}^{n} E[X_i]}{n} \;=\; p \tag{12.2}$$

We call a random variable (=function of the data) like $\hat{p}$ an *estimator* for p. The expectation of $\hat{p}$ is equal to the true value of the parameter p. Such an estimator is call *unbiased*. By contrast, an estimator for a parameter θ whose expectation is different from the true value of the parameter it estimates is *biased*. The

difference $E[\hat{\theta}] - \theta$ is called *bias*.

$$\text{Bias}(\hat{\theta}) \;=\; E[\hat{\theta}] - \theta \tag{12.3}$$

The bias is in general a function of θ. We'll encounter an example of a biased estimator in one of the following sections.

It is good that the expectation of our estimator is equal to the desired value, but what about the variance?

$$\begin{aligned} Var\ \hat{p} &= E[(\hat{p} - p)^2] && (12.4)\\ &\quad \text{(by example 2 in Handout 9} \\ &= \frac{Var\ X_1}{n} && (12.5)\\ &= \frac{p(1-p)}{n} && (12.6) \end{aligned}$$

Therefore, as $n \to \infty$, the variance of $\hat{p}$ will tend to 0; in other words, $\hat{p}$ converges to the true value p[1]. This means that for large n with very high probability, $\hat{p}$ will be close to p. How close? See the section on the central limit theorem to find out.

12.2 General discrete distribution

A general discrete distribution on an outcome space of size m is defined by m parameters $\theta_0, \ldots \theta_{m-1}$. The ML estimate of $\theta_i,\ i = 0, \ldots m-1$ is

$$\hat{\theta}_i \;=\; \frac{n_i}{n} \tag{12.7}$$

where n is the number of data points and n_i is the number of times the outcome is i. We can estimate the mean and variance of this estimator in a similar way as above, if we note that n_i is the sum of n random variables Y_j that are defined such that

$$Y_j \;=\; \begin{cases} 1 & X_j = i \\ 0 & X_j \neq i \end{cases} \tag{12.8}$$

Hence,

$$\hat{\theta}_i \;=\; \frac{\sum_{j=1}^{n} Y_j}{n} \tag{12.9}$$

and therefore

$$\begin{aligned} E[\hat{\theta}_i] &= P(Y=1) \;=\; \theta_i && (12.10)\\ Var\,\theta_i &= \frac{\theta_i(1-\theta_i)}{n} && (12.11) \end{aligned}$$

[1]This kind of convergence is a weak form of convergence. There are stronger results about the convergence of $\hat{p}$ to p but the one we derived here suffices to illustrate the point

This shows that for a general discrete distribution, the ML estimates converge to the true values of the parameters in the limit of infinite data.

12.3 The normal distribution

The normal density function is given by

$$f(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{12.12}$$

and the ML estimates of its parameters are

$$\hat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \tag{12.13}$$

$$\hat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{12.14}$$

Let us see how these estimates behave. For $\hat{\mu}$ we have (obviously by now)

$$E[\hat{\mu}] \;=\; \frac{1}{n}\sum_{i=1}^{n} E[X_i] \;=\; \mu \tag{12.15}$$

Its variance is (also obviously)

$$Var\,\hat{\mu} \;=\; \frac{1}{n^2}\sum_{i=1}^{n} Var\,X_i \;=\; \frac{\sigma^2}{n} \tag{12.16}$$

So, the estimate of the mean of the normal distribution is also converging to its true values. But with the estimate of the variance there is a surprise in store.

To do the calculations easier, recall that in chapter 8 it was proved that for any random variable Z and any real number a

$$E[(Z-a)^2] \;=\; Var\,Z + (E[Z]-a)^2 \tag{12.17}$$

For $a = 0$ we obtain

$$E[Z^2] \;=\; Var\,Z + E[Z]^2 \tag{12.18}$$

We shall use this relationship below.

$$\begin{aligned}
E[\hat{\sigma^2}] &= E[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2] \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}(X_i^2-2\hat{\mu}X_i+\hat{\mu}^2)\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}E[X_i^2]+E[-2\hat{\mu}\underbrace{\sum_{i=1}^{n}X_i}_{n\hat{\mu}}+n\hat{\mu}^2]\right) \\
&= \frac{1}{n^2}\left[\underbrace{n(\sigma^2+\mu^2)}_{\text{by (12.18)}}-nE[\hat{\mu}^2]\right] \\
&= \frac{1}{n^2}\left[n(\sigma^2+\mu^2)-n\underbrace{(\frac{\sigma^2}{n}+\hat{\mu}^2)}_{\text{by (12.18)}}\right] \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

This is an example of a biased estimator. The bias is

$$\text{Bias}(\hat{\sigma^2}) \;=\; E[\hat{\sigma^2}]-\sigma^2 \;=\; \frac{1}{n}\sigma^2 \tag{12.19}$$

It is easy to fix this problem by using a different estimator for σ^2

$$\bar{\sigma^2} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{12.20}$$

The above is the standard method for estimating the variance of the normal distribution.

12.4 The central limit theorem

We shall end this discussion with a convergence result of another nature. The *central limit theorem* says that if we add up a very large number of independent random variables, the distribution of the sum (which is the convolution of the distributions of the terms) will tend to a normal distribution. Let us make things more simple, by assuming that all the variables are sampled from the

same distribution P. Denote the mean and variance of P respectively by M and V (which are finite).

Let the sum of n samples from P be Z_n. Then, naturally

$$E[Z_n] \;=\; nM \quad \text{and} \quad Var\, Z_n \;=\; nV \tag{12.21}$$

Let us now shift and scale Z_n to obtain a RV with 0 mean and unit variance Y_n.

$$Y_n \;=\; \frac{Z_n - nM}{\sqrt{nV}} \tag{12.22}$$

Now, the Central Limit Theorem says that the CDF of Y_n converges pointwise to the CDF of the normal distribution with 0 mean and unit variance when $n \to \infty$.

$$F_{Y_n}(y) \;\to\; G(y) \tag{12.23}$$

where

$$G'(y) \;=\; \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \tag{12.24}$$

The convergence rate is pretty fast if P is "like a bump", that is having one maximum, being relatively smooth and symmetric.

Example 12.1 The sum of $n = 100$ independent tosses of a biased coin. *Assume that the probability of obtaining a 1 on any given toss is $\theta_1 = 0.6$ and let $X_i \in \{0, 1\}$ denote the outcome of toss i. Let $Z = X_1 + X_2 + \ldots X_n$.*

Find an approximation for the probability that $70 \leq Z \leq 90$.

Solution Because $E[X] = \theta_1$ and $Var(X) = \theta_1(1 - \theta_1)$ we have that $E[Z] = nE[X] = n\theta_1$ and $Var(Z) = nVar(X)$. Define the random variable

$$Y \;=\; \frac{Z - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} \tag{12.25}$$

Y has zero mean and unit variance and for n sufficiently large its CDF is well approximated by the CDF of the standard normal distribution. Therefore

$$\begin{aligned} P(70 \leq Z \leq 90) &= P\left(Y \in \left[\frac{70 - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}}, \frac{90 - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}}\right]\right) && (12.26)\\ &\approx \Phi\left(\frac{90-60}{4.899}\right) - \Phi\left(\frac{70-60}{4.899}\right) && (12.27)\\ &= 1 - 0.9961 \;=\; 0.0039 && (12.28) \end{aligned}$$

Exercise Note that the **exact** value of this probability, as well as the exact distribution of Z are known from chapter 4. Then, why is the approximation described above useful?

Chapter 13

Graphical models of conditional independence

13.1 Distributions of several discrete variables

The "Chest clinic" example - a domain with several discrete variables.

Smoker $\in \{Y, N\}$

Dyspnoea $\in \{Y, N\}$

Lung cancer $\in \{\text{no}, \text{incipient}, \text{advanced}\}$

Bronchitis $\in \{Y, N\}$

The domain has 4 variables, $2 \times 2 \times 3 \times 2 = 24$ possible configurations. The **joint probability distribution** $P_{SDLB}(s, d, l, b)$ is real valued function on $S_{\{S,D,L,B\}} = S_S \times S_D \times S_L \times S_B$. We sometimes call it a **multidimensional probability table**.

The **marginal** distribution of S, L is

$$P_{SL}(s, l) = \sum_{d \in S_D} \sum_{b \in S_B} P_{SDLB}(s, d, l, b)$$

The **conditional** distribution of Bronchitis given Dyspnoea is

$$P_{B|D}(b|d) = \frac{P_{BD}(b, d)}{P_D(d)}$$

Computing the probabilities of some variables (B) when we observe others (D) and we don't know anything about the rest (L, S) is a fundamental operation in probabilistic reasoning. Often it is called **inference** in the model P_{SDLB}.

13.2 How complex are operations with multivariate distributions?

Notations:

$V = \{X_1, X_2, \dots X_n\}$ the domain
$r_i = |S_{X_i}|$
$P_{X_1, X_2, \dots X_n}$ the joint distribution.

Number of configurations $|S_V| = \prod_{i=1}^{n} r_i \geq 2^n$. Required storage depends **exponentially** on n!

Computing the marginal of $X_1, \dots X_k$ takes $\left(\prod_{i=1}^{k} r_i\right)\left(\prod_{i=k+1}^{n} r_i\right) = |S_V|$ additions. Also exponential.

Computing conditional distributions: they are ratios of two marginals $\Rightarrow$ also exponential.

Sampling: can be done in logarithmic time in the size of S_V, thus is $\mathcal{O}(n)$.

Returning the probability of a configuration is also $\mathcal{O}(n)$.

In conclusion, a multivariate probability distribution becomes intractable when the number of variables is large (practically over 10 – 20). A solution to alleviate this problem (but ONLY in special cases) is offered by **graphical probability models**. They have the potential for compact representation and for efficient computations.

If $A, B \subseteq V$ are disjoint subsets of variables and $C = V \setminus (A \cup B)$ then

$$P_{A|B} = \frac{P_{A \cup B}}{P_B} \tag{13.1}$$

$$P_{A \cup B}(a, b) = \sum_{c \in S_C} P_V(a, b, c) \tag{13.2}$$

$$P_B(b) = \sum_{a \in S_A} \sum_{c \in S_C} P_V(a, b, c) \tag{13.3}$$

$$= \sum_{a \in S_A} P_{A \cup B}(a, b) \tag{13.4}$$

Hence, $P_B(b)$ is the normalization constant that turns $P_{A\cup B}(\cdot, b)$ into $P_{A|B}(\cdot|b)$.

Computing	P_B	directly :	$\|S_V\| = \|S_A\|.\|S_B\|.\|S_C\|$ operations
	P_B	as normalization constant	$\|S_A\|.\|S_B\|$ operations

13.3 Why graphical models?

A graphical model is a joint distribution represented in a certain way. We use the joint distribution to "reason" about the variables of interest (i.e. to compute their conditional probabilities given the evidence). We know that a discrete multivariate distribution represented by its values becomes intractable for high dimensions. Graphical models attempt to alleviate this problem - the model structure controls the computational complexity.

Tasks and domains	- noise - many dimensions - (usually) large data sets - task is not precisely defined (or more than one task)
For example:	Image analysis/segmentation/restauration Medical and technical diagnosis Maximum likelihood decoding, error correcting codes
Related to:	statistics optimization algorithms and computability database management

13.4 What is a graphical model?

Graphical model	= graphical representation of (conditional) independence relationships in a joint distribution = the distribution itself
graphical model	**- structure** (a graph) **- parametrization** (depends on the graph, parameters are "local")

A graph is defined as $G = (V, E)$ where

- V is the set of graph **vertices** (or **nodes**); each node represents a variable
- E is the set of graph **edges**; edges encode the dependencies. More precisely:

A missing edge encodes a conditional independence relationship.

13.5 Representing probabilistic independence in graphs

Idea: Independence in the joint distribution $\longleftrightarrow$ **Separation** in graph

This mapping is **not unique** and **not perfect**.

Reasoning in graphical models (i.e computing $P($ variables of interest $|$ evidence $)$) is performed by propagating beliefs along paths in the graph. We call these mechanisms **local** propagation because they corresponds to operations between variables that are close to each other in terms of graph distance. See the example of Markov chains below.

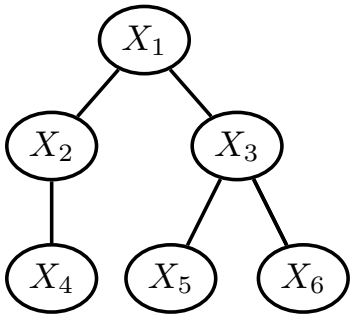
13.5.1 Markov chains

The joint distribution

$$P_{X_1X_2X_3X_4X_5} = P_{X_1}P_{X_2|X_1}P_{X_3|X_2}P_{X_4|X_3}P_{X_5|X_4}$$

is a product of conditional distributions involving X_{t+1}, X_t. X_{t+1}, X_t are neighbors in the chain, hence we say that $P_{X_1X_2X_3X_4X_5}$ can be computed from *local* (conditional) distributions.

13.5.2 Trees



Tree = connected graph with no cycles (we also call it **spanning** tree). If disconnected and no cycles, we call it a **forest**. Sometimes we use the term tree to mean either a spanning tree or a forest.

Property: between every two variables in a spanning tree there is exactly one path (at most one path for forests).

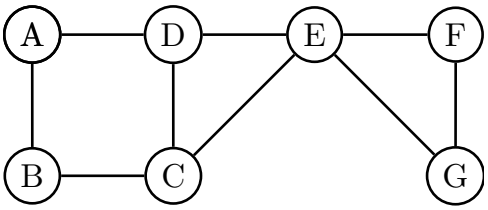
$A \perp B \mid C \iff$ all paths between sets A and B pass through set C

We say that C **blocks** the paths between A and B. Think of it as "blocking the flow of information".

13.5.3 Markov Random Fields (MRF)

Arbitrary undirected graphs.

$U_1 \perp U_2 \mid U_3 \iff$ all paths between sets U_1 and U_2 pass through set U_3



Examples: $F, G \perp A, B, C, D \mid E$
$A \perp C \mid B, D$
$A \perp C \mid B, D, E, F$

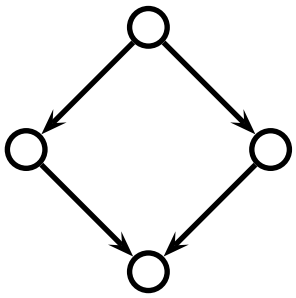
$n(A) =$ the **neighbors** of variable A

Markov property for MRFs:

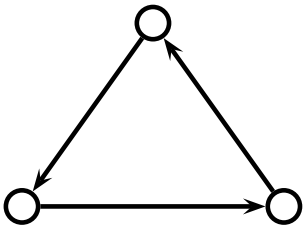
$$\boxed{A \perp \text{everything else} \mid n(A)}$$

13.5.4 Bayesian Networks

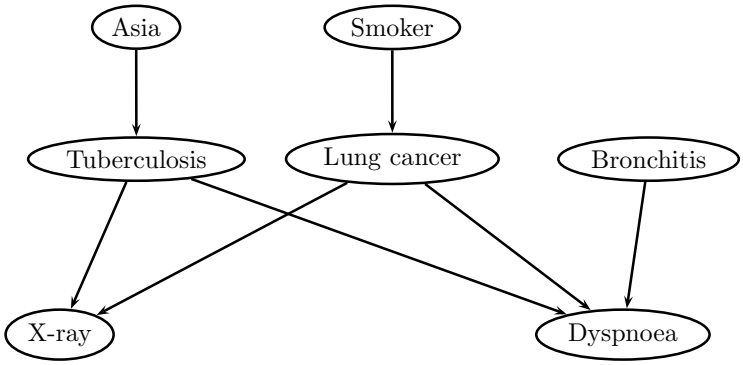
Directed acyclic graphs (DAGs)



this is a DAG



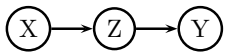
this is not a DAG



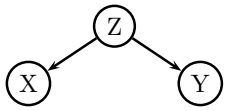
Terminology:	parent	Asia is parent of Tuberculosis
	pa(variable)	the set of parents of a variable
		pa(X-ray) = { Lung cancer, Tuberculosis }
	child	Lung cancer is child of Smoker
	ancestor	Smoker is ancestor of Dyspnoea
	descendent	Dyspnoea is descendent of Smoker
	family	a child and its parents
		Dyspnoea, Tuberculosis, Lung cancer, Bronchitis are a family

$$\boxed{A \perp B \mid C \iff A \textbf{ d-separated} \text{ from } B \text{ by } C}$$

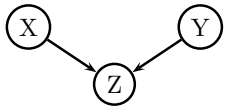
D-separation : A is **d-separated** from B by C if all the paths between sets A and B are blocked by elements of C. The three cases of d-separation:



if $Z \in C$ the path is blocked, otherwise open



if $Z \in C$ the path is blocked, otherwise open



if Z or one of its descendents $\in C$ the path is open, otherwise blocked

The **directed Markov property**: $\boxed{X \perp \text{ its non-descendants } \mid pa(X)}$

13.6 Bayes nets

Here we show how to construct joint probability distributions that have the independencies specified by a given DAG. Assume the set of discrete variables is $V = \{X_1, X_2, \ldots X_n\}$ and that we are given a DAG $G = (V, E)$. The goal is to construct the family of distributions that are represented by the graph. This family is given by

$$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i|\text{pa}(X_i)) \tag{13.5}$$

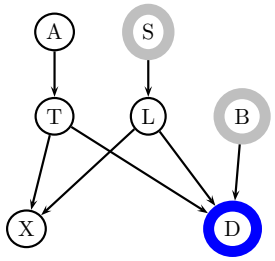
In the above $P(X_i|\text{pa}(X_i))$ represents the conditional distribution of variable X_i given its parents. Because the factors $P(X_i|\text{pa}(X_i))$ involve a variable and

its parents, that is, nodes closely connected in the graph structure, we often call them **local** probability tables (or local distributions).

Note that the parameters of each local table are (functionally) independent of the parameters in the other tables. We can choose them separately, and the set of all parameters for all conditional probability distributions form the family of distributions for which the graph G is an I-map.

If a distribution can be written in the form (13.5) we say that the distribution **factors according to the graph** G. A joint distributions that factors according to some graph G is called a **Bayes net**.

Note that any distribution is a Bayes net in a trivial way: by taking G to be the complete graph, with no missing edges. In general, we want a Bayes net to be as sparse as possible, because representing independences explicitly has many computational advantages.



The Bayes net described by this graph is

$$P(A,S,T,L,B,X,D) = P(A)P(S)P(T|A)P(L|S)P(B)P(X|T,L)P(D|T,L,B)$$

A way of obtaining this decomposition starting from the graph is

1. Construct a topological ordering of the variables. A **topological ordering** is an ordering of the variables where the parents of each variable are always before the variable itself in the ordering.
   A, S, T, L, B, X, D is a topological ordering for the graph above.
2. Apply the chain rule following the topological ordering.
   $$\begin{aligned} P(A,S,T,L,B,X,D) &= P(A)P(S|A)P(T|A,S)P(L|A,S,T)P(B|A,S,T,L) \\ &\quad P(X|A,S,T,L,B)P(D|A,S,T,L,B,X) \end{aligned}$$
3. Use the directed Markov property to simplify the factors
   $$\begin{aligned} P(S|A) &= P(S) \\ P(T|A,S) &= P(T|A) \\ P(L|A,S,T) &= P(L|S) \\ P(B|A,S,T,L) &= P(B), \text{ etc.} \end{aligned}$$

Let us now look at the number of parameters in such a model. Assume that in the example above all variables are binary. Then the number of unconstrained parameters in the model is

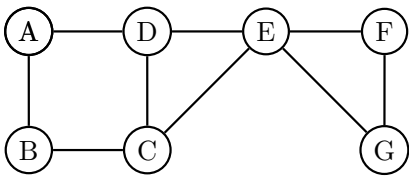
$$1 + 1 + 2 + 2 + 1 + 4 + 8 = 19$$

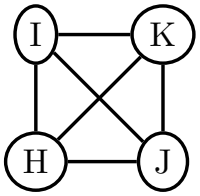
The number of parameters in a 7 way contingency table is $2^7 - 1 = 127$ so we are saving 118 parameters. As we shall see, there are also other computational advantages to joint distribution representations of this form.

13.7 Markov nets

Now we look at joint distributions that factor according to undirected graphs. Such distributions are called **Markov nets** or **Markov random fields**. Just like Bayes nets, Markov nets are a product of local functions, called **clique potentials**.

A **clique** is a completely connected subset of nodes in a graph. For example, in the graph below, $\{A, D\}$, $\{D, E\}$, $\{C, D, E\}$, $\{E, F, G\}$ are cliques. In particular, every node and every edge in a graph is clique. $\{A, B, C, D\}$ is not a clique. A clique of size four is $\{H, I, J, K\}$.





Some cliques are included in other cliques (for example $\{D, E\}$ is included in $\{C, D, E\}$). A clique which is not included in any other clique is called **maximal**. In the example above, $\{B, C\}$ and $\{C, D, E\}$ are some maximal cliques.

A clique potential $\psi(X_{\mathcal{C}})$ is a non-negative function of the variables in the clique $\mathcal{C}$. A joint distribution P **factors according to the undirected graph** G if it can be written as a product of potentials over the maximal cliques of G.

$$P(X_1, X_2, \ldots X_n) = \prod_{\mathcal{C}} \psi(X_{\mathcal{C}}) \tag{13.6}$$

For example, a joint distribution that factors according to the undirected graph above has the form

$$P(A, B, C, D, E, F, G,) = \psi(A, B)\psi(A, D)\psi(B, C)\psi(C, D, E)\psi(E, F, G) \tag{13.7}$$

Note that this factorization is not unique. One can obtain equivalent factorizations by dividing/multiplying with functions of variables that are common between cliques. For example, let $\phi(B)$ be a postive function of variable B. Then the following is also a factorization of $P(A, B, C, D, E, F, G,)$ according to the same graph.

$$P(A,B,C,D,E,F,G,) \;=\; [\underbrace{\psi(A,B)\phi(B)}_{\psi'(A,B)}]\psi(A,D)[\underbrace{\psi(B,D)/\phi(B)}_{\psi'(B,D)}]\,\psi(C,D,E)\psi(E,F,G) \tag{13.8}$$

Unlike in Bayes nets, the potentials ψ do not represent probability distributions.

Again, the savings in terms of number of parameters are significant. Assume that all variables are binary, and all potentials are represented by (unnormalized) tables. Then for the graph above, the total number of parameters is

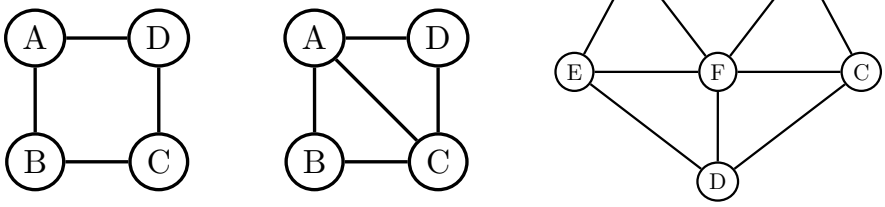
$$3 \times 2^2 + 2 \times 2^3 \;=\; 28$$

The size of a probability table over 7 binary variables is $2^7 - 1 = 127$ thus in this example we save 99 parameters (almost 80%).

13.8 Decomposable models

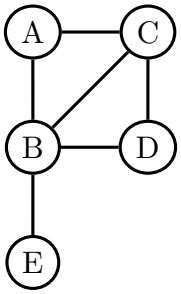
Decomposable models are a category of graphical probability models that factor according to triangulated graphs.

We say that an undirected graph is **triangulated** (or **chordal**) if every cycle of length greater than 3 has a **chord**.

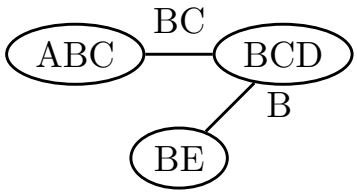


not triangulated triangulated not triangulated

In a triangulated graph, the maximal cliques and their intersections, called **separators**, play an important role. A triangulated graph can be represented as a tree of maximal cliques. Below is an example.



graph



tree of cliques; separators are edges

A joint probability distribution that factors according to a junction tree has the form:

$$P(X_1,\, X_2, \ldots\, X_n) \;=\; \frac{\prod_{\mathcal{C}} P(X_{\mathcal{C}})}{\prod_{\mathcal{S}} P(X_{\mathcal{S}})} \tag{13.9}$$

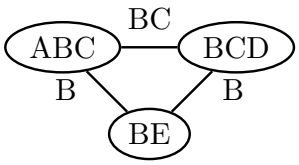
where $\mathcal{C}, \mathcal{S}$ are respectively indices over the cliques and separators of the graph G.

For the graph above, the factorization is

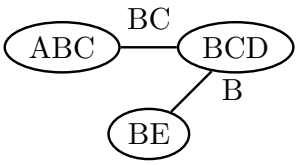
$$P(A, B, C, D, E) \;=\; \frac{P(A, B, C)P(B, C, D)P(B, E)}{P(B, C)P(B)} \tag{13.10}$$

Any junction tree factorization can be easily seen as a Markov net factorization. Obviously, any decomposable model is a Markov net. Therefore, we often refer to $P_{\mathcal{C}}, P_{\mathcal{S}}$ as **clique/separator potentials**. However, in decomposable models the potentials are in a form that exhibits the local probability tables. Note that local, in this context, means within a clique or a separator. In contrast with Bayes nets, the local probability distributions that build a decomposable model are marginal probabilities.

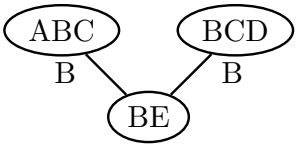
The junction tree structure is not unique. A junction tree is always a **Maximum Spanning tree** w.r.t separator size. (A maximum spanning tree is a tree over V whose sum of edge weights has a maximum value. Here the edge weights are the sizes of the separators.)



graph of cliques and separators

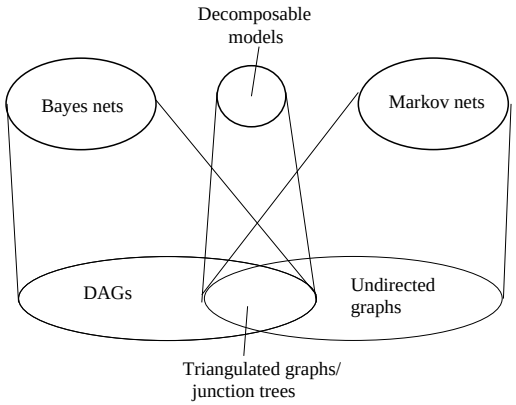


a junction tree



not a junction tree

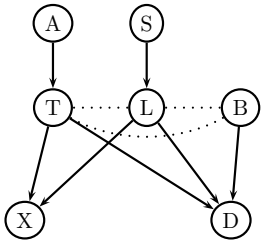
13.9 Relationship between Bayes nets, Markov nets and decomposable models



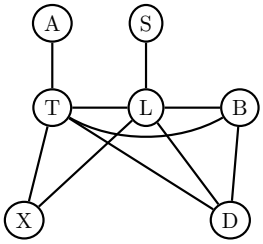
13.10 D-separation as separation in an undirected graph

Here we show that D-separation in a DAG is equivalent to separation in an undirected graph obtained from the DAG and the variables we are interested in. First two definitions, whose meaning will become clear shortly.

Moralization is the graph operation of connecting the parents of a V-structure. A DAG is **moralized** if all nodes that share a child have been connected. After a graph is moralized, all edges, be they original edges or new edges added by moralization, are considered as undirected. If G is a DAG the graph obtained by moralizing G is denoted by G^m and is called the **moral graph** of G.



marrying the parents



dropping the directions

For any variable X the set $\text{an}(X)$ denotes the ancestors of X (including X itself). Similarly, if A is a set of nodes, $\text{an}(A)$ denotes the set of all ancestors of variables in A.

$$\text{an}(A) \;=\; \bigcup_{X \in A} \text{an}(X)$$

The **ancestral graph** of a set of nodes $A \subseteq V$ is the graph $G_A = (\text{an}(A), E_A)$ obtained from G by removing all nodes not in $\text{an}(A)$.

Now we can state the main result.

Theorem Let $A, B, S \subseteq V$ be three disjoint sets of nodes in a DAG G. Then A, B are D-separated by S in G iff they are separated by S in the moral ancestral graph of A, B, S.

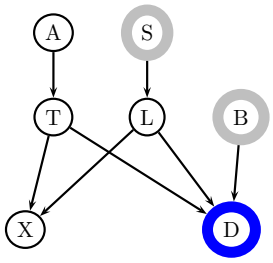
$$A \perp B \,|\, S \text{ in } G \qquad \text{iff} \qquad A \perp B \,|\, S \text{ in } (G_{A \cup B \cup S})^m$$

The intuition is that observing/conditioning on a variable creates a dependence between its parents (if it has any). Moralization represents this link. Now why the ancestral graph? Note that an unobserved descendent cannot produce dependencies between its ancestors (ie cannot open a path in a directed graph). So we can safely remove all descendents of A, B that are not in S. The descendents of S itself that are not in A, B, and all the nodes that are not ancestors of A, B, S can be removed by a similar reasoning. Hence, first the graph is pruned, then dependencies between parents are added by moralization. Now directions on edges can be removed, because DAG's are just like undirected graphs if it weren't for the V-structures, and we have already dealt with those.

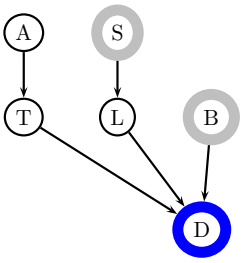
The Theorem immediately suggests an algorithm for testing D-separation using undirected graph separation.

1. remove all nodes in $V \setminus \text{an}(A \cup B \cup S)$ to get $G_{A \cup B \cup S}$
2. moralize the remaining graph to get $(G_{A \cup B \cup S})^m$
3. remove all nodes in S from $(G_{A \cup B \cup S})^m$ to get G'
4. test if there is a path between A and B in G'

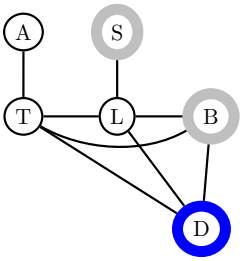
For example, test if $S \perp B \mid D$ in the chest clinic DAG.



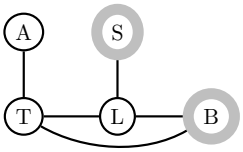
show nodes of interest



ancestral graph



moralize



eliminate conditioning nodes

Chapter 14

Probabilistic reasoning

The concept of conditional independence, together with the formula of conditional probability and the derived formulas (Bayes' rule, total probability), when pieced together, allow us to process complicated systems of (probabilistically) related concepts and to draw conclusions within these systems about events that interest us from observing other events. This is called probabilistic reasoning and is one of the most spectacular and successful applications of probability in Artificial Intelligence.

Below is an example concerning an imaginary problem of medical diagnosis.

Example 14.1 Probabilistic medical diagnosis. *A patient tests HIV positive on a test (call this event T) and the doctor wants to know what is the probability that the patient is actually HIV positive (call this event HIV). What the doctor knows is that*

- *The HIV test is not perfect; it will be positive if the patient has HIV with probability*
$$P(T|HIV) = 0.99 \tag{14.1}$$
and negative otherwise. The test may also be positive if the patient is not infected with HIV; this happens with probability
$$P(T|\overline{HIV}) = 0.03. \tag{14.2}$$
- *The incidence of the HIV virus in the population of the US is $P(HIV) = 0.001$. (These figures are not real figures!)*

How can the doctor compute what she wants to know, namely $P(HIV|T)$ from the information she has?

$$P(HIV|T) \;=\; \frac{P(T|HIV)P(HIV)}{P(T)} \tag{14.3}$$

We now compute $P(T)$ by the law of total probability

$$P(T) \;=\; P(T|HIV)P(HIV) + P(T|\overline{HIV})P(\overline{HIV})$$

Replacing the numbers we get

$$P(HIV|T) \;=\; \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.03 \times 0.999} \;=\; 0.032$$

This probability is very small, but it is about 30 times larger than the $P(HIV)$ before seeing the positive result of the test. This is due mainly to the fact that the **prior** probability of HIV in the population $P(HIV)$ is very small.

Let us now add to the scenario the fact that, when trying to diagnose the HIV infection, the doctor may take into account other **evidence** then the test, namely the patient's symptoms and history. Denote the events "history suggesting HIV infection" and "symptoms suggesting HIV infection" by H and S respectively. The doctor's knowledge gives him the following conditional probabilities relating HIV with S and H:

$$P(HIV|H) \;=\; 0.1 \tag{14.4}$$
$$P(S|HIV) \;=\; 0.8 \tag{14.5}$$
$$P(S|\overline{HIV}) \;=\; 0.1 \tag{14.6}$$

The doctor also knows that the presence of the symptoms depends on nothing else but the HIV infection, the test result T and the symptoms S are independent if we know whether HIV is present or not, and that the test and symptoms are related to the patient history only through the HIV state. This knowledge can be expressed in the probabilistic independencies:

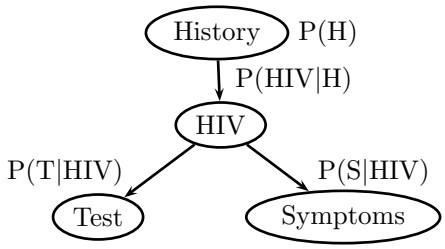
$$T \perp S \;\mid\; HIV \tag{14.7}$$
$$T \perp H \;\mid\; HIV \tag{14.8}$$
$$S \perp H \;\mid\; HIV \tag{14.9}$$

The graph below describes the cause-effect relationship between the 4 events involved and lets us easier remember the conditional independencies (14.7–14.9).

Let us now assume that the doctor learns about the patient's history H. How can the doctor merge the two observations she now posseses T and H to improve his guess about HIV the event she cannot observe directly? In other words, how to compute the conditional probability $P(HIV|T,H)$?

The intuitive solution is to replace $P(HIV)$ by $P(HIV|H)$ in formula (14.3). This gives us after calculations

$$P(HIV|T,H) \;=\; \frac{0.99 \times 0.1}{0.99 \times 0.1 + 0.03 \times 0.9} \;=\; 0.79 \tag{14.10}$$



Let us do the same calculation another way. We will work on $P(HIV|T,H)$ trying to put it in a form that lets us use the probabilites we already know.

$$P(HIV|T,H) = \frac{P(T|HIV,H)P(HIV|H)}{P(T|H)} \quad \text{(Bayes for } HIV,\ T \text{ only)} \tag{14.11}$$

$$= \frac{P(T|HIV)P(HIV|H)}{P(T|H)} \quad (\text{because } T \perp H \,|\, HIV) \tag{14.12}$$

The denominator is computed again by total probability, keeping H behind the conditioning bar.

$$P(T|H) = P(T|HIV,H)P(HIV|H) + P(T|\overline{HIV},H)P(\overline{HIV}|H) \tag{14.13}$$

$$= P(T|HIV)P(HIV|H) + P(T|\overline{HIV})P(\overline{HIV}|H) \tag{14.14}$$

$$= 0.99 \times 0.1 + 0.03 \times 0.9 \ = \ 0.126 \tag{14.15}$$

Putting it all together we get again

$$P(HIV|T,H) \ = \ \frac{0.99 \times 0.1}{0.99 \times 0.1 + 0.03 \times 0.9} \ = \ 0.79$$

Let us now take into account history, test and symptoms. We need to compute $P(HIV|H,T,S)$. Again, we will turn it around to exhibit the probabilities that

we know, using conditional independence whenever we can.

$$P(\underbrace{HIV|T,S}_{\text{Bayes}},H) = \frac{P(T,S|HIV,H)P(HIV|H)}{P(T,S|H)} \tag{14.16}$$

$$= \frac{P(T,S|HIV)P(HIV|H)}{P(T,S|H)} \quad (\text{because } T,S \perp H \,|\, HIV) \tag{14.17}$$

$$= \frac{P(T|HIV)P(S|HIV)P(HIV|H)}{P(T,S|H)} \quad (\text{because } T \perp S \,|\, HIV) \tag{14.18}$$

$$= \frac{P(T|HIV)P(S|HIV)P(HIV|H)}{P(T|HIV)P(S|HIV)P(HIV|H) + P(T|\overline{HIV})P(S|\overline{HIV})P(\overline{HIV}|H)} \quad (\text{total pr}$$

$$= \frac{0.99 \times 0.8 \times 0.1}{0.99 \times 0.8 \times 0.1 + 0.03 \times 0.1 \times 0.9} = 0.967 \tag{14.19}$$

Let us take a qualitative look at the series of probabilities $P(HIV) = 0.001,\ P(HIV|T) = 0.032,\ P(HIV|T,H) = 0.79,\ P(HIV|T,S,H) = 0.967$. The first is the prior, which gives HIV a very low probability. The test is strong evidence for HIV, but in view of the low prior, the probabilityof HIV is still very low. The patient history is weak evidence ($P(HIV|H) = 0.1$) but it replaces the uninformed prior $P(HIV)$ with something 100 times stronger; in corroboration with the positive test, the probability of HIV is now significant. The symptoms represent relatively weak evidence compared to the test, but in the context of the other two observations H,T that they corroborate, they make the probability of HIV almost a certainty.

Exercises. What would be the probability of HIV if the symptoms are negative, i.e what is $P(HIV|T,\overline{S},H)$?

What if the doctor, upon seeing that the test is positive, instead of examining the history and symptoms orders the repetition of the test?

Denote $T1$ = "the first test is positive", $T2$ = "the second test is positive". Assume

$$T1 \perp T2 \,|\, HIV \tag{14.20}$$

$$T1 \perp T2 \,|\, \overline{HIV} \tag{14.21}$$

The results of the tests are independent given the HIV state. The probabilities of the test outcomes are given by (14.1,14.2) for both tests. Compute $P(HIV|T1,T2)$.

Chapter 15

Statistical Decisions

Example 15.1 *Optimizing the number of servers*

An Internet company offers a certain service. The number of requests for service in a second follows a discrete exponential (or **geometric***) distribution with* $\lambda = 2/3$

$$P(n) \;=\; (1-\lambda)\lambda^n \tag{15.1}$$

The distribution is plotted in figure 15.1, **a**. *(***Exercise** *Verify that* $P(n)$ *sum to 1 and that the expectation* $E[n]$ *is equal to* $\frac{\lambda}{1-\lambda}$ *for* $\lambda < 1$*).*

The company wants to choose the number of servers m *so that it's operation costs are as low as possible. Serving one request takes exactly 1 second and a server can serve only one request at a time. The costs associated with the operation are are:*

request refused	$R = 5$
idle server	$I = 1$
request served	$S = -10$ *(a gain)*

To pose the problem in probabilistic terms, we want to minimize the expected cost incurred per second.

First idea. The cost as a function of n the number of requests can be expressed as:

$$cost(n) \;=\; \begin{cases} nS + (m-n)I & \text{if } 0 \le n \le m \\ mS + (n-m)R & \text{if } n > m \end{cases} \tag{15.2}$$

The average cost $C(m)$ will then be

$$C(m) \;=\; E[cost] \;=\; \sum_{n=0}^{\infty} cost(n)P(n) \tag{15.3}$$

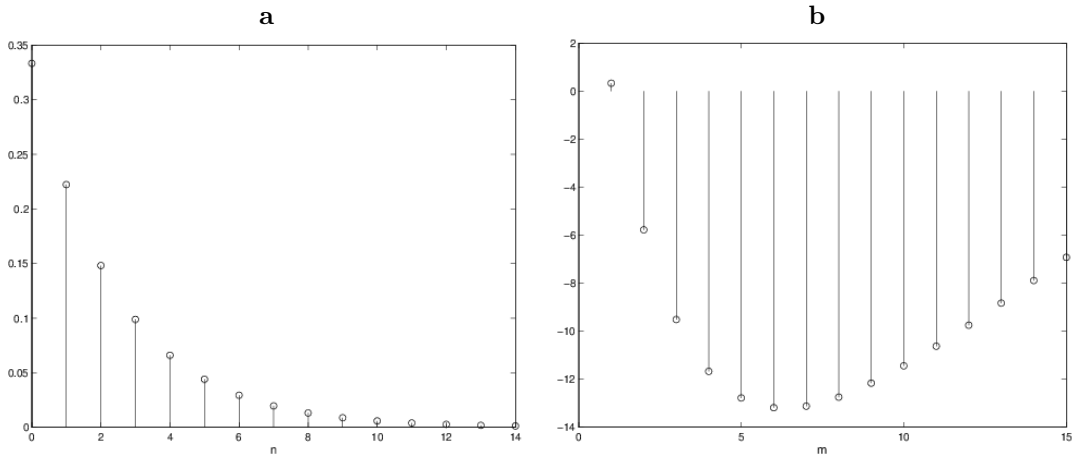


Figure 15.1: The discrete exponential (geometric) distribution for $\lambda = 2/3$ (a); the cost as a function of the number of servers m (b)

Computing this value for every m in a reasonable range we obtain the plot in figure 15.1, **b** and the optimal values $m = 6$, $C = -13.2$. Thus, for the optimal value of m, the company gains 13.2 monetary units per second on average.

Second idea. The cost is equal to $S \times n_s + R \times n_R + I \times n_I$ where n_S, n_R, n_I are respectively the number of requests that are served, the number of requests refused and the number of idle servers. Therefore, we can compute the average cost as

$$E[cost] \;=\; S \times E[n_S] \;+\; R \times E[n_R] \;+\; I \times E[n_I] \tag{15.4}$$

We have

$$E[n_S] \;=\; mP(n \geq m) + \sum_{n<m} nP(n) \tag{15.5}$$

$$E[n_R] \;=\; \sum_{n>m} (n-m)P(n) \;=\; E[n] - E[n_S] \tag{15.6}$$

$$E[n_I] \;=\; \sum_{n<m} (m-n)P(n) \tag{15.7}$$

This is actually the way that the plots in figure 15.1 were obtained. The matlab file is `statistical-decision-server-geometric.m` on the Handouts web page.

Variation: Poisson distribution A more realistic model for the requests distribution is the Poisson model. A Poisson distribution with $\lambda = 2$ is shown in figure 15.2, **a**. Recall that the Poisson distribution is

$$P(n) \;=\; e^{-\lambda} \frac{\lambda^n}{n!} \tag{15.8}$$

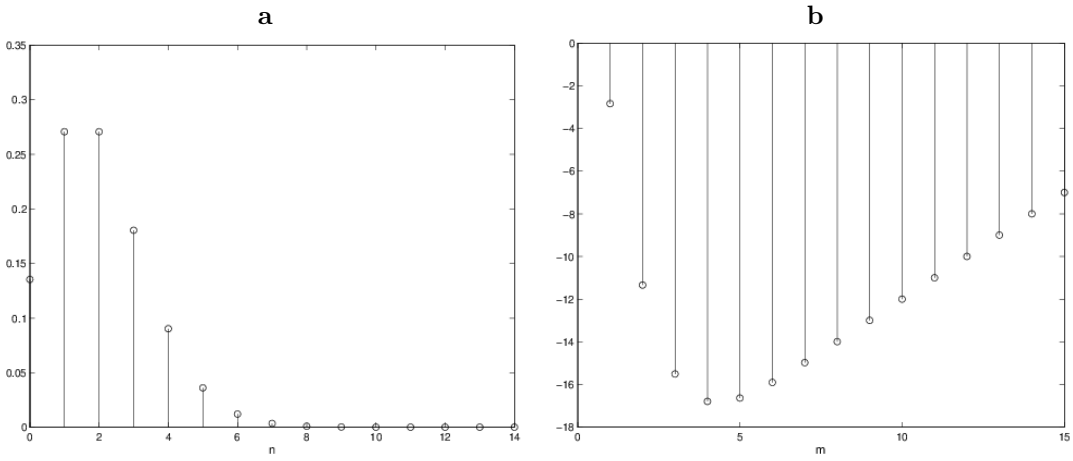


Figure 15.2: The Poisson distribution with $\lambda = 2$ (a); the cost as a function of the number of servers m (b)

and its expectation is λ.

If we redo the calculations above (see the file `statistical-decision-server-poisson.m` on the Handouts web page) assuming a Poisson distribution, we obtain the plot in figure 15.2,**b** and the optimal values $m = 4$, $C = -16.8$.

Example 15.2 *Testing for HIV*

Should a doctor test his patient for HIV?

What the doctor knows is that

- *The HIV test is not perfect; it will be positive if the patient has HIV with probability*

$$P(T|HIV) = 0.99 \tag{15.9}$$

and negative otherwise. The test may also be positive if the patient is not infected with HIV; this happens with probability

$$P(T|\overline{HIV}) = 0.03. \tag{15.10}$$

- *The incidence of the HIV virus in the population of the US is* $P(HIV) = 0.001$. *(These figures are not real figures!)*

Note that for brevity we use $\overline{HIV}$ *to mean* $HIV = 0$ *or equivalently "no HIV"; similarly* $\overline{T}$ *means* $T = 0$ *or "test result negative".*

The doctor also knows that the patient associates the folowing costs with the possible actions and outcomes:

taking the test	*3*
correct diagnosis	*0*
undiagnosed HIV infection (miss)	*100*
false HIV diagnosis (false alarm)	*10*

Like many diagnosis situations, this is a one where the costs of errors are asymmetric: missing a developing infection is much worse than giving the patient a false alarm.

The doctor has the choice of prescribing the test or not, and he will choose the alternative with the lowest expected cost for the patient.

Let us start by evaluating the expected cost in the case the test is not taken. In this case the doctor's diagnosis is "no HIV", therefore the costs table looks like this (C being the cost):

HIV	0	1
C	0	100

The expected cost is

$$E[C \,|\, \text{no test}] \;=\; P(HIV) \times 100 + P(\overline{HIV}) \times 0 \;=\; 0.1 \tag{15.11}$$

In the second case, the outcome is described by the two variables: $HIV \in \{0,1\}$ and $T \in \{0,1\}$.

$$\begin{aligned} E[C \mid \text{test}] &= 3 + P(HIV, \overline{T}) \times 100 + P(\overline{HIV}, T) \times 10 && (15.12) \\ &= 3 + 10^{-3} \times 0.01 \times 100 + 0.999 \times 0.03 \times 10 && (15.13) \\ &\approx 3.3 && (15.14) \end{aligned}$$

Hence, not taking the test is much cheaper for an average person.

Exercise Assume that the test is taken and is positive. Should the doctor repeat the test? If the test is repeated and the result is negative, the doctor will diagnose "no HIV". The results of the tests are independent conditioned on the variable HIV. All the other costs and probabilities remain the same.

Chapter 16

Hypothesis testing

16.1 What is hypothesis testing?

Example 16.1 (From Dekking et. al) How many tanks were produced this month? a WWII true story *Let N be the number of objects (i.e. tanks) produced in a given month. We know that the objects are each labeled with their number, ranging from 1 to N. We observe the numbers $\mathcal{D} = \{61, 19, 56, 24, 16\}$. Obviously, at least 56 objects were produced. It is reported that $N = 350$. Shall we believe this report?*

1. **Framing the problem a hypothesis testing**
   - $H_0 : N = 350$ is the **null hypothesis**. We will believe this hypothesis if the data does not give us reason to **reject** it.
   - $H_1 : N < 350$ is the **alternative hypothesis**. Note that this is not simply the negation of H_0. It is what we would believe if the null was rejected.
2. **Choose a test statistic** T. In our problem, we choose $T = \max \mathcal{D}$. Qualitatively, if T is very small, we favor H_1 over H_0. If T is close to $N = 350$ but not larger than N, then we favor H_0 over H_1, and if T is larger than N, neither hypothesis would be true.
3. **Calculate** $t = T(\mathcal{D})$**.** $t = 61$ in our problem. Is this value "small" (favoring H_1) or "close to 350" (favoring H_0)?
4. **Calculate tail probability for** T **given** H_0. The **tail probability** is the probability that T is *at least as extreme* than the observed $t = 61$, under the null hypothesis. In our case, $Pr[T \leq t = T(\mathcal{D})|H_0]$ is given

Table 16.1: Possible outcomes in hypothesis testing

	H_0 is true	H_1 is true
Not reject H_0	✓	Type II error (FN)
Reject H_0	Type I error (FP)	✓
Total probability	1	1

by the probability of sampling uniformly from $1:N$ without replacement $n=5$ times, and obtaining a sample contained entirely in $1:61$.

$$Pr[T \leq t = T(\mathcal{D}) \,|\, H_0] \;=\; \frac{61}{350}\frac{60}{349}\frac{59}{348}\frac{58}{347}\frac{57}{346} \;=\; 0.00014 \qquad (16.1)$$

Note that "more extreme" depends on the alternative hypothesis H_1. The value $p = Pr[T \leq t = T(\mathcal{D})|H_0]$ is called a **p-value**. It tells us that the probability of getting data that supports H_0 *stronger* than the observed data $\mathcal{D}$ is $1 - p = 0.9986$.

How does the p-value depend on t and n? For this problem, we expect that p will increase if t increases towards N, but that it will decrease if we take a larger sample size n. This is reflected in the following table.

N			350		
t	61	300	300	300	300
n	5	5	10	20	50
p	0.00014	0.447	0.18	0.024	6.10^{-6}

5. **Make decision** based on p. The situation is described by the following table. In the above, FN, FP denote **False Positive** (a "positive" meaning a surprise, a deviation from the null hypothesis) respectively **False Negative** (with "negative" in the sense of no suprise, i.e. null hypothesis holding). The p-value measures the probability of making a Type I error. To make a truly informed decision, we would need to know the probability of a Type II error too, but this probability is not obtainable from the current assumptions (we would need to have a model for N when $N < 350$). In this case, knowing that the probability of Type I error is very close 0, we will probably be comfortable rejecting H_0.

16.2 Advanced concepts of hypothesis testing

In view of the asymmetric information described by Table 16.1, a person performing a hypothesis test by steps 1 to 4 above could include a preliminary decision (Step 0) which is *the maximum value of p for which they would be comfortable rejecting* H_0. This value is called α. In other words, α represents how unusual the observations need to be before we feel compelled to assume the

situation is not normal and take action accordingly. Since Type I error (mis-detection of a potential problem) is very often more costly than Type II error (false alarm), it is fortunate that it is the former probability that is most easily computed.

In statistics, α is called **significance level**. Since p is a function of t, let us denote it p_t in this section. The **critical region** $K_\alpha = \{t \,|\, p_t \leq \alpha\}$. This is the region in the range of the test statistic T where H_0 is rejected "at significance level α". The critical value C_α is the most probable value of t in K_α. In our example it is the maximum t for which H_0 is rejected.

Note that there could be more complicated scenarios where the boundary of K_α is not a single point, in which case the critical value C_α may not be a unique value.

The **power** of a test is $Pr[$ Reject $H_0 \,|\, H_1$ true $]$, and

$$\beta \;=\; 1 - \text{power} \;=\; 1 - Pr[\text{ Reject } H_0 \,|\, H_1 \text{ true }] \;=\; Pr[\text{Type II error}] \quad (16.2)$$

The power of a test depends on

- α: Power increases when α increases [Prove this]
- n number of samples: Power increases with number of samples
- How different is truth (an element of H_1) from H_0. This is called **effect size**.
- The test statistic T (what information from the data we use to make decision). Some tests are more powerful than others.

16.2.1 What do theoretical statisticians work on?

- For a given T (a function of the data), what is the distribution of T given H_0. What are the critical regions? These can then be implemented in computer programs to be readily used in decision problems.
- For given questions: what T's give more powerful tests?

  It turns out that for large classes of questions, there exists a test that is **uniformly most powerful**. This means that there is a function T^*, so that the test based on T^* is the most powerful compared to all tests based other T statistics, *for any data*. This is the Likelihood Ratio test presented next.

16.3 The Likelihood Ratio Test

Originally written by Wenyu Chen `wenyuc@uw.edu`, 2018

16.4 Likelihood Ratio Test Statistics

Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f(x|\theta)$. We define the **likelihood ratio test statistic** for $H_0 : \theta \in \Theta_0$ verses $H_1 : \theta \in \Theta_0^C$ as

$$\lambda(x) = \frac{\sup_{\Theta_0} L(\theta|x)}{\sup_{\Theta} L(\theta|x)}$$

where Θ denotes the space of free parameters. The Likelihood ratio test can be applied in these cases where *either H_0 is true, or H_1 is true.*

If we plug in the MLE, we obtain, equivalently,

$$\lambda = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}.$$

Where $L(\hat{\theta}_0|x)$ is the MLE under H_0, and $L(\hat{\theta}|x)$ is the MLE under H_1.

We call an *unknown* parameter a **free parameter** if its value can not be uniquely determined when knowing all other parameters.

For example, in the coin flip case, let p_0 and p_1 be the unknown probability of getting a tail and a head. Since we always have $p_0 + p_1 = 1$, knowing one will automatically give the value of the other. Hence, even though we can write the likelihood as a function of both p_0 and p_1, there is only one free parameter. If p_0 is free, then p_1 is not, and if p_1 is free, then p_0 is not. However, one of them must be free.

Example 16.2 *Consider a coin flip problem with $p = P(1)$. We test $H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$. We know that in this problem there is only one parameter, p. Under H_1, the parameter p can be any other value in the range $[0, 1]$, making it a* free parameter*. Under H_0, we already know p, so there is no free parameter.*

Example 16.3 *Consider a multinomial problem with 3 possible outcomes, and p_a, p_b, p_c the probability of outcome a, b, c. We test $H_0 : p_a = 0.5$ vs $H_1 : p_a \neq 0.5$. There are three parameters, but since $p_a + p_b + p_c = 1$, knowing two of them will automatically give the value of the third. Thus, under H_1, there are two free parameters. Under H_0, p_a is known, so there is only one free parameters.*

Example 16.4 *Multinomial problem with 3 possible outcomes, and p_a, p_b, p_c the probability of outcome a, b, c, respectively. We test $H_0 : p_a = p_b$ vs $H_1 : p_a \neq p_b$. Under H_1, there are two free parameters. Under H_0, the unknown parametr p_a determines $p_b = p_a$ and $p_c = 1 - 2p_a$, so there is only one free parameters.*

Example 16.5 *Multinomial problem with 3 possible outcomes, and p_a, p_b, p_c the probability of outcome a, b, c, respectively. We test $H_0 : p_a = p_b = 0.2$ vs $H_1 : p_a \neq p_b$. Under H_1, there are two free parameters. Under H_0, $p_a = 0.2$ is known, and so is $p_b = 0.2$ and $p_c = 0.6$, so there is no free parameter.*

16.4.1 Examples: Plug in likelihood to LRT definition

Example 16.6 *Normal distribution with known variance Let $X_1, \ldots, X_n$ be n i.i.d. observations drawn from $Normal(\mu, \sigma^2)$. We want to test $H_0 : \mu = \theta_0$ versus $H_0 : \mu \neq \theta_0$.*

$$\lambda(x) = \exp\left\{\left(-\sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)/2\right\} = \exp[-n(\bar{x} - \theta_0)^2/2] \tag{16.3}$$

H_0 has 0 free parameters, H_1 has 1 free parameter, μ.

Example 16.7 *Binomial Let x be the number of heads in n trials, test $H_0 : \theta = \theta_0$ verses $H_1 : \theta \neq \theta_0$:*

$$\lambda(x) = \frac{\theta_0^x(1 - \theta_0^{n-x})}{[\frac{x}{n}]^x[\frac{n-x}{n}]^{n-x}} = \left[[\frac{n\theta_0}{x}]^x\right][\frac{(1-\theta_0)n}{n-x}]^{n-x} \tag{16.4}$$

H_0 has 0 free parameter, H_1 has 1 free parameter.

Example 16.8 *Multinomial with m possible outcomes Let $X_1, \ldots, X_m$ be the counts of each category. We test $H_0 : \theta_i = \theta_{0,i}$ for all $i = 1, \ldots, m$ versus $H_0 : \theta \neq \theta_0$,*

$$\lambda(x) = \left[\prod_{i=1}^{m-1}(\frac{n\theta_{0,i}}{x_i})^{x_i}\right][\frac{(1 - \sum_{j=1}^{m-1}\theta_{0,j})n}{n - \sum_{j=1}^{m-1}x_j}]^{n-\sum_{j=1}^{m-1}x} \tag{16.5}$$

H_0 has 0 free parameter, H_1 has $m-1$ free parameter. The number is $m-1$ instead of m because $\theta_m = 1 - \sum_{i=1}^{m-1}\theta_i$, so knowing the first $m-1$ parameters will fix the value of the last one. Notice the similarity between this and the previous question.

16.4.2 Wilk's theorem

With MLE being a consistent, asymptotically normal estmator for θ, (which is met in the examples we mentioned above),

$$-2\log\lambda \sim \chi^2_{d-d_0} \text{ as } n \to \infty. \tag{16.6}$$

Here d is the number of free parameters in H_1 and d_0 is number of free parameters in H_0.

Chapter 17

Classification

17.1 What is classification?

In classification, we are given an *input* x which can be an integer, a real number or another type of discrete variable (it can also be a vector of variables) and a set of possible categories, or *classes*. The input x always belongs to one of the classes and only one; we call it the *class of* x. The task is to determine $c(x)$ the class of x for every possible x.

Classification appears in a wide variety of circumstances. It is also known as *pattern recognition*, *concept learning*, *categorization*.

Example 17.1 Handwritten digit recognition *The inputs x are 8-by-8 matrices of black and white pixels (see fig 17.1). There are 10 classes, the digits 0 through 9. The task is to recognize the digits.*

Example 17.2 Document categorization *The inputs are documents (in some representation), for example news articles, books, scientific articles. The classes are categories of documents: sport, politics, international, economic if the documents are news articles.*

Example 17.3 Protein classification *A protein is a sequence of aminoacids. The inputs are proteins, reprented as strings of aminoacids (there are 20 aminoacids) and the output is a class to which the protein belongs (e.g peptides).*

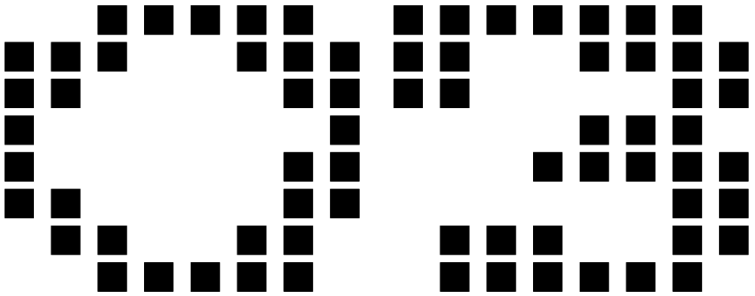


Figure 17.1: Two examples of handwritten digits.

17.2 Likelihood ratio classification

Here, for simplicity we will assume that there are only two classes, labeled 0 and 1. A possible method of classifying data is to assume that for each class there is a distribution $P_{X|c}$ that generates the data belonging to that class. We assign

$$c(x) \;=\; 1 \;\;\Leftrightarrow\;\; P_{X|C}(x|1) \;>\; P_{X|C}(x|0) \tag{17.1}$$

otherwise, we decide that the class of x is 0. The above is equivalent to writing

$$c(x) \;=\; 1 \;\;\Leftrightarrow\;\; \frac{P_{X|C}(x|1)}{P_{X|C}(x|0)} \;>\; 1 \tag{17.2}$$

and because $P_{X|C}(x|c)$ is the likelihood that x belongs to class c the method is known as the *likelihood ratio* method. If x belongs to a continuous set, i.e. x is a real number or a vector of real numbers, then the above becomes:

$$c(x) \;=\; 1 \;\;\Leftrightarrow\;\; \frac{f_{X|C}(x|1)}{f_{X|C}(x|0)} \;>\; 1 \tag{17.3}$$

17.2.1 Classification with different class probabilities

Very often, classes do not appear in the data with the same probability.

Example 17.4 Automatic meteorite classification *The robot NOMAD, constructed by CMU, was deployed in Antarctica to search for meteorites. It would pick up a stone, inspect it visually and chemically and decide whether the stone is one of three types of meteorites or a rock from the nearby mountain. In this situation, there are four classes, but one of them (terrestrial rocks) is thousands of times more frequent then the others.*

In this case, it is good to take into account the class *prior probability* P_C. Then by Bayes' formula, we have

$$P_{C|X}(c|x) \;=\; \frac{P_C(c)P_{X|C}(x|c)}{\sum_{c'} P_C(c')P_{X|C}(x|c')} \tag{17.4}$$

This gives us a probability distribution $P_{C|X}$ over the possible classes. If we want to decide for a class, we choose the class with highest *posterior probability* $P_{C|X}$.

For two classes, we choose class 1 if

$$P_{C|X}(1|x) \;>\; P_{C|X}(0|x) \quad\Longleftrightarrow\quad 1 \;<\; \frac{P_{C|X}(1|x)}{P_{C|X}(0|x)} \;=\; \frac{P_{X|C}(x|1)P_C(1)}{P_{X|C}(x|0)P_C(0)} \tag{17.5}$$

Or, equivalently, we choose

$$c(x) \;=\; 1 \;\Leftrightarrow\; \frac{P_{X|C}(x|1)}{P_{X|C}(x|0)} \;>\; \frac{P_C(0)}{P_C(1)} \tag{17.6}$$

This is again a likelihood ratio method, where the *threshold* $\frac{P_C(0)}{P_C(1)}$ depends on the relative probabilities of the two classes.

17.2.2 Classification with misclassification costs

Example 17.5 Diagnosis as classification *A doctor is faced with a classification problem when she has to decide whether a patient has a certain disease or not. This is a decision in uncertainty, so there is always a non-zero probability of making a diagnosis error. There are two kinds of errors a doctor can make: to diagnose the disease when the patient is healty (false alarm), or, to decide the patient is healty when he is in fact ill. The second error is potentially more damaging than the first.*

One can assign a **loss** L to each kind of possible error and look at classification as a statistical decision problem where the objective is to minimize the expected loss. For a problem with 2 classes, assuming that the loss of a correct guess is 0, that $\hat{c}$ is our guess and c is the truth, we have the loss matrix

$\hat{c}$	0	1
$c=0$	0	L_{01}
1	L_{10}	0

The expected losses for guessing 0, respective 1 are

$$E[L|\hat{c}=0] \;=\; P_{C|X}(1|x)\times L_{10} + P_{C|X}(0|x)\times 0 \tag{17.7}$$

$$E[L|\hat{c}=1] \;=\; P_{C|X}(1|x)\times 0 + P_{C|X}(0|x)\times L_{01} \tag{17.8}$$

Assuming the losses are positive, we choose

$$c(x) = 1 \Leftrightarrow 1 < \frac{E[L|\hat{c}=0]}{E[L|\hat{c}=1]} = \frac{P_{C|X}(1|x) \times L_{10}}{P_{C|X}(0|x) \times L_{01}} \tag{17.9}$$

After a little calculation we obtain

$$c(x) = 1 \Leftrightarrow \frac{P_{X|C}(x|1)}{P_{X|C}(x|0)} > \frac{P_C(0)}{P_C(1)} \frac{L_{01}}{L_{10}} \tag{17.10}$$

This shows that in the case of asymmetric misclassification loss, the classification rule is again a likelihood ratio method, where the threshold depends on the losses incurred by each kind of error.

Example 17.6 *Two Normal Distributions*

If both classes are generated by normal distributions, the resulting likelihood ratio can be brought to a simple form. Let

$$f_{X|C}(x|c) = \frac{1}{\sigma_c \sqrt{2\pi}} e^{-\frac{(x-\mu_c)^2}{2\sigma_c^2}} \tag{17.11}$$

and $P_C(1) = p$. *Then we choose class 1 if*

$$\log \frac{P_{C|X}(1|x)}{P_{C|X}(0|x)} > 0 \tag{17.12}$$

But

$$\log \frac{P_{C|X}(1|x)}{P_{C|X}(0|x)} = -\frac{(x-\mu_1)^2}{2\sigma_1^2} - \log \sigma_1 + \frac{(x-\mu_0)^2}{2\sigma_0^2} + \log \sigma_0 + \log \frac{p}{1-p} \tag{17.13}$$

This is a quadratic function in x. *If the two class variances are equal* $\sigma_0 = \sigma_1$ *then the decision simpifies to*

$$-\frac{(x-\mu_1)^2}{2\sigma^2} - \log \sigma + \frac{(x-\mu_0)^2}{2\sigma^2} - \log \sigma + \log \frac{p}{1-p} > 0 \tag{17.14}$$

$$-(x-\mu_1)^2 + (x-\mu_0)^2 + 2\sigma^2 \log \frac{p}{1-p} > 0 \tag{17.15}$$

$$-2x(\mu_0 - \mu_1) + \mu_0^2 - \mu_1^2 + 2\sigma^2 \log \frac{p}{1-p} > 0 \tag{17.16}$$

$$x > \frac{\mu_1 + \mu_0}{2} - \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{p}{1-p}$$

(assuming $\mu_1 > \mu_0$). Hence in this case classification boils down to a comparison between x and a threshold.

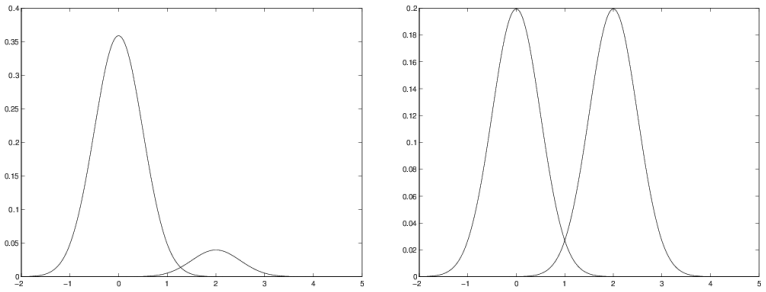


Figure 17.2: Two normal distributions scaled by their prior probabilities; $p = 0.1$ and 0.5, $\mu_0 = 0, \mu_1 = 2$, $\sigma_1 = \sigma_0 = 1$

17.3 The decision boundary

Likelihood ratio classification can be translated into the following "rule":

> If x is in the set $\mathcal{R}_1 = \{x : \frac{P_{C|X}(1|x)}{P_{C|X}(0|x)} > 1\}$ choose $C = 1$
> else choose $C = 0$.

The set $\mathcal{R}_1$ is called the *decision region* for class 1, and its complement, denoted by $\mathcal{R}_0$ is the decision region for 0. The boundary between $\mathcal{R}_0$ and $\mathcal{R}_1$ is the *decision boundary*.

For two classes, the above can usually be summarized in the rule

> If $\phi(x) > 0$ choose class 1, else choose class 0.

In this case the curve $\phi(x) = 0$ is the decision boundary.

A classifier (i.e a program that classifies the x's) is anything that specifies the decision regions for each class, whether it uses a probability model or not. In the following sections we shall see some examples of classifiers that are based on decision regions rather than on probability models for the classes.

If we don't need probability to construct a classifier, why even mention them together? As it turns out, we do need probability in order to analyze classifiers: for example to predict its average error, or how many examples we need in order to learn it accurately. Probability and statistics also provide methods for learning a classifier from examples and for choosing between two classifiers. Last but not least, probabilistic methods often lead us to classification methods that might not have been invented otherwise. Such classifiers are among the best performing classifiers in existence.

17.4 The linear classifier

The linear classifier takes as input a vector of real numbers x. A zero-one valued variable can be also viewed as a real number. The decision rule is

> If
> $$b^T x \;=\; \sum_{i=1}^{m} b_i x_i \;>\; \gamma \tag{17.17}$$
> choose $C = 1$ else choose 0. Here b is an m dimensional vector of real numbers and γ is a real number; b, γ are the *parameters* of the classifier.

In terms of the *decision function* $\phi(x)$ a linear classfier is a classifier for which $\phi(x)$ is linear (plus a constant).

For example, the likelihood ratio classifier for two normal distributions seen in the previous handout is a linear classifier. Indeed, the decision rule for that classifier is

> If
> $$x \;>\; \gamma = \frac{\mu_1 + \mu_0}{2} - \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{p}{1-p}$$
> choose class 1 else class 0.

In this case

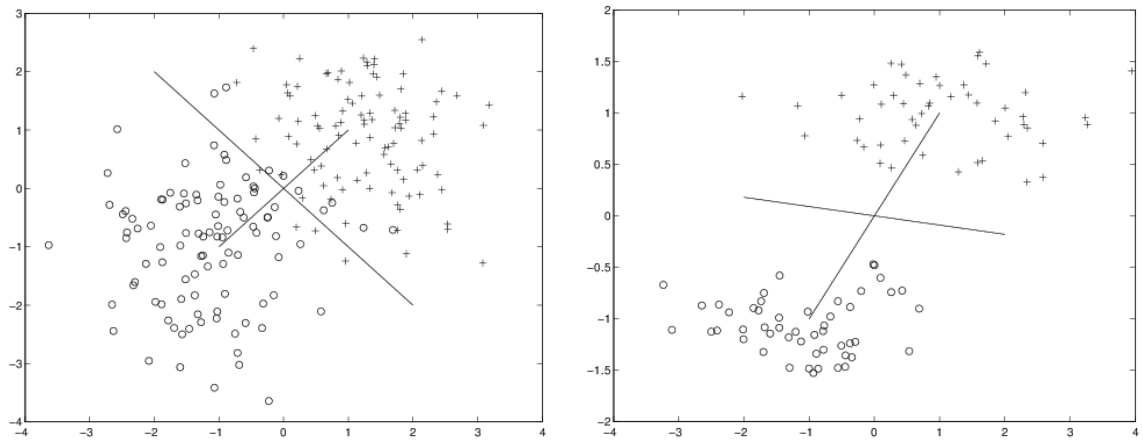
$$\phi(x) = x - \gamma \tag{17.18}$$

The same is true if $x = [x_1\ x_2]^T$ a two-dimensional vector and $P_{X|C}$ are normal densities with means μ_c, $c = 0, 1$ (two-dimensional vectors) and equal covariance matrices $\Sigma_0 = \Sigma_1 = \Sigma$. As before, the prior probability of class 1 is p. Denote

$$D \;=\; \Sigma^{-1} \tag{17.19}$$

Then

$$\begin{aligned}
\phi(x) &= \log P_{C|X}(1|x) - \log P_{C|X}(0|x) && (17.20)\\
&= -\frac{1}{2}(x-\mu_1)^T D(x-\mu_1) - \log p + \frac{1}{2}(x-\mu_0)^T D(x-\mu_0) + \log(1-p)\\
&= \frac{1}{2}\left[-x^T Dx + 2\mu_1^T Dx - \mu_1^T D\mu_1 - \log p\right.\\
&\quad \left. + x^T Dx - 2\mu_0^T Dx + \mu_0^T D\mu_0 + \log(1-p)\right] && (17.21)\\
&= \underbrace{(\mu_1-\mu_0)^T D}_{b^T} x - \underbrace{\frac{1}{2}\left[\mu_1^T D\mu_1 - \mu_0^T D\mu_0 + \log\frac{p}{1-p}\right]}_{\gamma} && (17.22)
\end{aligned}$$



a **b**

Figure 17.3: Linear classification for two normal distributions with equal covariance matrices. In case **a** $\sigma_x = \sigma_y = 1,\ \rho = 0,\ \mu_0 = [-1\ \ -1],\ \mu_1 = [1\ 1]$; in case **b** $\sigma_x = 1,\ \sigma_y = 0.3,\ \rho = 0,\ \mu_0 = [-1\ \ -1],\ \mu_1 = [1\ 1]$ and the decision boundary is not perpendicular to the line connecting the class means.

For example, if Σ is the unit matrix, meaning that $\sigma_{x_1} = \sigma_{x_2} = 1,\ \rho_{x_1 x_2} = 0$ then D is also the unit matrix. Then the above decision function simplifies to

$$\phi(x) \ = \ (\mu_1 - \mu_0)^T x - \frac{1}{2}\left[||\mu_1||^2 - ||\mu_0||^2 + \log\frac{p}{1-p}\right] \tag{17.23}$$

where $||a||$ is the length of vector a. This function is a line perpendicular to the line connecting the two means μ_0, μ_1. The line is closer to μ_1 if p is small, closer to μ_0 otherwise and in the middle of the segment if $p = 0.5$. (Can you prove this?)

In general, for m-dimensional input x, if both classes have normal distributions the decision rule will be linear.

17.5 The classification confidence

If our classifier is obtained from a likelihood ratio, the decision function $\phi(x)$ is

$$\phi(x) \ = \ \log\frac{P_{C|X}(1|x)}{P_{C|X}(0|x)} \tag{17.24}$$

Hence, if $\phi(x)$ is close to 0, then the confidence in our classification is low, since the corresponding $P_{C|X}$ is close to 0.5. If $\phi(x)$ is positive or negative with a large

magnitude, then the confidence in the chosen class is high. We can interpret $\phi(x)$ as a confidence even if it wasn't explicitly obtained from a likelihood ratio.

17.6 Quadratic classifiers

A quadratic classifier is a classifier for which $\phi(x)$ is a polynomial of degree 2. For example, for x in 1 dimension

$$\phi(x) \;=\; x^2 - bx + c \tag{17.25}$$

is a quadratic classifier. Can you show that this classifier is obtained when the two classes have normal distributions with *same* μ and different σ^2?

In two dimensions, a quadratic classifier is

$$\phi(x) \;=\; a_1 x_1^2 + a_2 x_2^2 + a_{12} x_1 x_2 + b_1 x_1 + b_2 x_2 + c \tag{17.26}$$

The curve $\phi(x) = 0$ is a quadratic curve – ellipse, parabola or hyperbola, depending on the values of the parameters.

In m dimensions, the quadratic classifier is

$$\phi(x) \;=\; \sum_{i=1}^{m} a_i x_i^2 + \sum_{i=1}^{m} \sum_{j=i+1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m} b_i x_i + c \tag{17.27}$$

In a similar way, one can obtain classifiers from polynomials of higher degrees. More generally any function $\phi(x)$ depending on some set of parameters θ corresponds to a classifier for x. These are called *parametric* classifiers. Of course the problem is what are the correct parameters for a given classifier and this is usually solved by learning (that is estimating) the best parameters from examples, as we shall see shortly.

17.7 Learning classifiers

Classification has strong ties to learning from data, since for the vast majority of classification tasks, the "classification rule" is not known, or is to long and complicated to implement. Take for example the digit recognition task; a human can do a good job on this task, meaning that (s)he has a fairly accurate decision rule. Implementing this rule, however, is an entirely different story and in practice it is easier (although not easy!) to derive a classification rule from examples than to have an "expert" write one.

Learning a classifier from examples is done in two stages: First, one decides on the type of classifier to be used (e.g linear, quadratic, likelihood ratio, decision

tree, etc); this is calles selecting a *model class*. The model class is chosen based on knowledge about the problem and on the amount of data available. Next, the parameters for the model are estimated using the data. Parameter estimation is often called *learning* or *training*.

17.8 Learning the parameters of a linear classifier

Usually for this problem the inputs are vectors of real numbers (or integers) having dimension m. The data set (or training set) consists of n examples and their classes

$$\mathcal{D} \ = \ \{(x^{(1)}, c^{(1)}), \ \ldots \ (x^{(n)}, c^{(n)})\} \tag{17.28}$$

By this notation, $x_j^{(i)}$ is the coordinate j of the i-th example and $c^{(i)}$ is its class.

When estimating the parameters of a linear binary classifier, it is sometimes practical to consider the class as taking values in $\{+1, -1\}$ instead of $\{0, 1\}$. This will make the forthcoming mathematical expressions more compact and easy to read without changing the basic problem.

A linear classifier is defined by

> If $\phi(x) > 0$ choose class 1, else choose class -1.
>
> $$\phi(x) \ = \ \sum_{j=1}^{m} b_j x_j \ + \ \gamma \tag{17.29}$$
>
> where $(b_1, \ \ldots \ b_m, \gamma)$ are the classifier's parameters.

The problem is to estimate these parameters from the data and for this purpose we use the Maximum Likelihood (ML) approach. To obtain a probability model from $\phi(x)$ we use (17.24) and the fact that $P_{C|X}(1|x) + P_{C|X}(0|x) = 1$. We obtain

$$P_{C|X}(1|x) \ = \ \frac{1}{1 + e^{-\phi(x)}} \ \stackrel{\Delta}{=} \sigma(\phi(x)) \tag{17.30}$$

and

$$\begin{aligned} P_{C|X}(-1|x) &= 1 - \frac{1}{1 + e^{-\phi(x)}} && (17.31)\\ &= \frac{e^{-\phi(x)}}{1 + e^{-\phi(x)}} && (17.32)\\ &= \frac{1}{1 + e^{\phi(x)}} && (17.33)\\ &= \sigma(-\phi(x)) && (17.34) \end{aligned}$$

Figure 17.4: Examples of handwritten digits. The images of the digits were rotated and scaled to fit in a square. (They are also right-left reflected but this is an artefact caused by the way the image was generated.)

We can summarize the last two equations into

$$P_{C|X}(c|x) \;=\; \frac{1}{1+e^{-c\phi(x)}} \;\stackrel{\Delta}{=}\; \sigma(c\phi(x)) \tag{17.35}$$

In the next section we shall use the following property of the sigmoid function σ:

$$\sigma'(u) \;=\; \sigma(u)(1-\sigma(u)) \tag{17.36}$$

This property can be proved directly by taking the derivative of $\frac{1}{1+e^{-u}}$.

17.8.1 Maximizing the likelihood

We define the likelihood of the data as

$$L(b_1, \ldots b_m, \gamma) \;=\; \prod_{i=1}^{n} P_{C|X}(c^{(i)}|x^{(i)}, b_1, \ldots b_m, \gamma) \tag{17.37}$$

and the log-likelihood

$$\begin{aligned} l(b_1, \ldots b_m, \gamma) \;&=\; \sum_{i=1}^{n} \log P_{C|X}(c^{(i)}|x^{(i)}, b_1, \ldots b_m, \gamma) && (17.38)\\ &=\; \sum_{i=1}^{n} \log \sigma(c^{(i)}\phi(x^{(i)})) && (17.39) \end{aligned}$$

This expression cannot be maximized analytically. To find the optimal parameters we use gradient ascent. Let us compute the expression of the gradient of l w.r.t the parameters.

$$\frac{\partial l}{\partial b_j} = \sum_{i=1}^{n} \frac{\partial \log \sigma(c^{(i)}\phi(x^{(i)}))}{\partial b_j} \tag{17.40}$$

$$= \sum_{i=1}^{n} \frac{\partial \sigma(c^{(i)}\phi(x^{(i)}))}{\partial b_j} \Big/ \sigma(c^{(i)}\phi(x^{(i)})) \tag{17.41}$$

$$= \sum_{i=1}^{n} \frac{\sigma(c^{(i)}\phi(x^{(i)}))(1-\sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} x_j^{(i)}}{\sigma(c^{(i)}\phi(x^{(i)}))} \tag{17.42}$$

$$= \sum_{i=1}^{n} (1-\sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} x_j^{(i)} \tag{17.43}$$

Similarly, we obtain

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} (1-\sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} \tag{17.44}$$

A word about the computational complexity of this optimization. Note that all the partial derivatives involve the factor $(1-\sigma(c^{(i)}\phi(x^{(i)})))$. Computing it once requires $m+1$ multiplications (and an exponentiation), rougtly $\mathcal{O}(m)$ operations. We have to compute this term for every data point, which requires $\mathcal{O}(mn)$ operations. The additional multiplications to obtain the gradient once we have the values of the sigmoids are also $\mathcal{O}(mn)$ hence one step of the gradient ascent takes $\mathcal{O}(mn)$ operations. Note that this is the same order of magnitude as classifying all the points in the data set.

Unlike the previous gradient ascent problem that we have encountered, this problem generally has *local optima*. It means that the solution we obtain may depend on the initial point of the iteration. It is wise in such cases to repeat the ascent several times starting from different point to increase the chance of obtaining a good solution.

17.9 ML estimation for quadratic and polynomial classifiers

The quadratic classifier is defined by

$$\phi(x) = \sum_{j=1}^{m} a_j x_j^2 + \sum_{k=1}^{m} \sum_{j=k+1}^{m} a_{kj} x_k x_j + \sum_{j=1}^{m} b_j x_j + \gamma \tag{17.45}$$

and its parameters are (a_j, $j = 1, \ldots m$, a_{kj} $k = 1, \ldots m$ $j = k{+}1, \ldots m$, b_j $j = 1, \ldots m$, γ).

Following the same steps that led to (17.43) we obtain

$$\frac{\partial l}{\partial a_j} = \sum_{i=1}^{n}(1 - \sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} x_j^{(i)2} \tag{17.46}$$

$$\frac{\partial l}{\partial a_{kj}} = \sum_{i=1}^{n}(1 - \sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} x_k^{(i)} x_j^{(i)} \tag{17.47}$$

$$\frac{\partial l}{\partial b_j} = \sum_{i=1}^{n}(1 - \sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} x_j^{(i)} \tag{17.48}$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n}(1 - \sigma(c^{(i)}\phi(x^{(i)}))) \cdot c^{(i)} \tag{17.49}$$

The number of computations is proportional to the number of parameters times the number of data points.

The procedure generalizes readily to decision functions $\phi(x)$ that are polynomial or in general *linear in the parameters*, i.e

$$\phi(x) = \sum_j \theta_j g_j(x) \tag{17.50}$$

17.10 Non-parametric classifiers

17.10.1 The Nearest-Neighbor (NN) classifier

The NN classifier uses the following classification rule:

> To classify point x find the point in the dataset $\mathcal{D}$ that is nearest to x and assign its class to x. Let this point be $x^{(i)}$. Then, $c(x) = c(x^{(i)}) = c^{(i)}$.

Nearest neighbor classification eliminates the decision on the model class. It also eliminates learning altogether. On the other hand, you need to keep the dataset around in order to classify subsequent data. The method has low bias – it can represent any decision surface, given enough data. But the variance is rather high. A popular method to reduce variance is the k-NN method. The k-NN method classifies a point x the following way:

> Find the k points in the data set that are closest to x. Choose the class of x to be the class of the majority of the k neighbors.

In practice k is small (usually 3, 5). Both NN and k-NN can be applied to multi-class problems as well.

The most important draw-back of NN methods in general is not the variance but the dependence of the distance function. If the data are scaled, the results of the classification change as well. Moreover, if there are inputs x_j that have no relevance for the class, then the results of the NN method become very poor. This is often the case with very high dimensional inputs.

Chapter 18

Clustering

18.1 What is clustering?

Clustering is the task of grouping observations into categories. It is the first step in data analysis. Figure 18.1 shows an example of data where 3 groups are visible.

Clustering problems occur very often in the real world and in computer applications. Here are just a few examples:

- In a newspaper, the article topics are grouped into sections like politics, sports, buisness, etc. If we were to do the grouping starting from a collection of mixed articles we would be "clustering" the articles.
- Foods can be clustered into breads, cereal, fruit, vegetables, meats, sweets, etc. Some of the categories, or **clusters** can be further subdivided (for example meats can be grouped into fish, poultry, beef, lamb, etc.)
- An airline groups its passengers into buisness and leisure travellers and offers distinctive fares, discounts and other incentives to the two groups. Of course, the airline does not ask the passengers to identify themselves as "buisness" or "leisure"; it has to realize what cluster the passenger falls into based solely on the characteristics of the passenger that it can observe (for example: that the trip starts on a Friday evening, travel with a child, etc.).
- **Image segmentation** means finding the groups of **pixels** in an image that correspond to the same object. In other words, a computer receives an images represented as a matrix of pixels and it has to group together

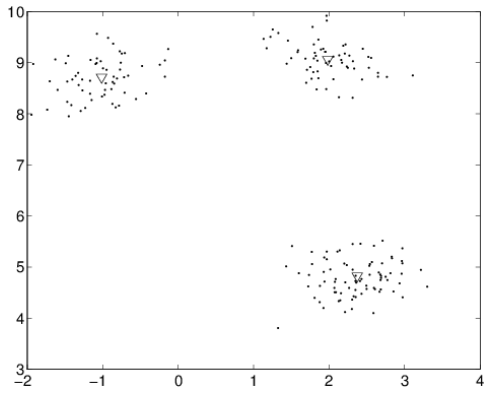


Figure 18.1: A set of points that can be partitioned into 3 clusters. The crosses represent the clusters' centers of mass.

the pixels corresponding to the same object (or object part). Image segmentation is performed naturally by us people, but it is a very difficult task for a computer. In fact, the problem has not been solved satisfactorily to date.

Note that in this task, the computer is not told in advance what the objects would be; it does not even know *how many* objects (= clusters of pixels) there are. This is a fundamental characteric of clustering, which contributes to making it a difficult task.

- **Gene clustering.** The analysis of DNA has identified large numbers of genes in humans and other living species, but for most of them the functionality is yet unknown. One method that scientists use is to cluster the genes (looking at characteristics like: the gene structure, or behaviour in certain experiments). After clustering, if in one cluster fall some genes who function is already known, we may reasonably guess that the other genes in the cluster may be involved in similar processes. In this task, unlike in image segmentation, neigher humans, no computers know the "right" clustering. Again, it is not known in advance how many clusters there are, or what each of them might represent.

18.2 The K-means algorithm

The **K-means algorithm** is one of the simplest algorithms for clustering. It assumes that the clusters are **spherical**; in other words, every cluster has a **center**, and points belonging to the cluster are near that center. The centers are represented as crosses superimposed on the data set of figure 18.1.

To run the algorithm we need to know the number of clusters K. The algorithm starts with assinging the K centers random positions. Then it finds the data points that are nearest to each center. With this as an initial clustering, it can find better positions for the centers, i.e exaclty at the center of mass of each cluster. With the new centers a new clustering is found, and so on. The algorithm is given below in pseudocode: C_k denotes the set of points assigned to cluster k; for a data point i, $clust(i)$ is a number between 1 and K that represents the cluster that point i belongs to.

Algorithm K-MEANS
Input $\{x_1, x_2, \ldots x_n\}$ the data points
K the number of clusters
Output $clust(i)$ for $i = 1, \ldots n$
c_1, c_2, ... c_K the positions of the K centers

Initialize c_1, c_2, ... c_K with random values
Do
for$i = 1, \ldots n$
find k such that $||x_i - c_k|| \leq ||x_i - c_{k'}||$ for all $k' = 1, \ldots K$
$clust(i) \leftarrow k$
for $k = 1, \ldots K$
$C_k = \{x_i,\ clust(i) = k\}$
$c_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$
until $clust(i),\ i = 1, \ldots n$ remain unchanged

A succesful run of the algorithm is depicted in figure 18.2.

Computations per iteration. An algorithm that arrives at the result by gradually improving the current solution in a loop, is called an **iterative** algorithm. The sequence of operations in the loop form an **iteration**. In each iteration, the algorithm has to compute the distance of every data point x_i to all the centers. This takes Kn distance computations, or $\mathcal{O}(Knd)$ elementary operations, if the data lie in d dimensions. Recomputing the centers takes a number of operations equal to the number of data points n. Therefore, the total number of operations required for one iteration of the K-means algorithm is $\mathcal{O}(Knd)$.

Convergence. Note that, if none of the cluster assignments changes between two consecutive steps of the algorithm, then neither the cluster assignments or the cluster centers will change in the future. We say that the algorithm has **converged**. For the K-means algorithm, convergence always occurs after a finite number of iterations, but we cannot know in advance how many iterations it will take.

Local optima. If the initial conditions change, the result of the K-means

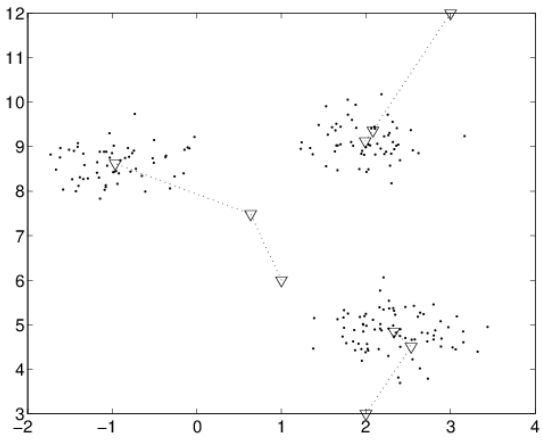


Figure 18.2: The K-means algorithm on the 3 clusters data. The triangles represent the trajectories of the 3 centers from their initial values to the final ones.

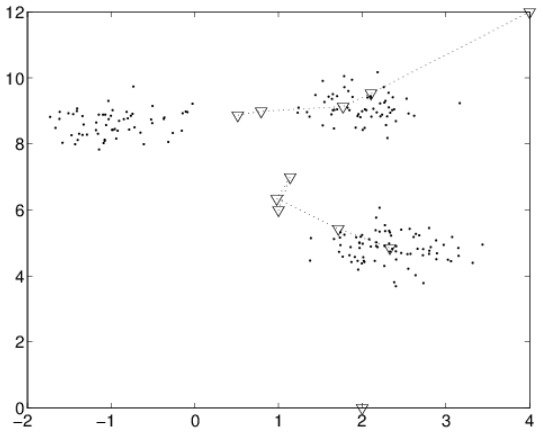


Figure 18.3: An unsuccessful run of the K-means algorithm. The initial values of the centers are such that the algorithm converges to a local optimum where one of the centers is assigned no points.

algorithm may be different. This happens because the algorithm works by improving the current solution; starting in a very defavorable configuration may lead to a bad clustering. An example of this behavior is shown in figure 18.3. We say that the K-means algorithm finds **local optima** of the clustering problem.

18.3 The confusion matrix

Sometimes we want to compare two clusterings of the same data set. For example, when we are given the correct solution, we want to see how well K-means (or another algorithm) has performed on the data, for the purpose of testing the algorithm. To compare two clusterings (with possibly different number of clusters) we use the **confusion matrix**.

Let the two clusterings have K, respectively K' clusters and be described by $clust(i)$, respectively $clust'(i)$ for $i = 1, \ldots n$. Define the set $A_{kk'}$ to contain the points that belong to cluster k in the first clustering and to cluster k' in the second clustering

$$A_{kk'} \;=\; \{x_i,\; clust(i) = k,\; clust'(i) = k'\} \;=\; C_k \cap C'_{k'} \tag{18.1}$$

The confusion matrix A has K rows and K' colums. An element $a_{kk'}$ of A represents the number of points in $A_{kk'}$.

$$a_{kk'} \;=\; |A_{kk'}| \tag{18.2}$$

If the two clusterings are identical, then each row or column of A will contain exactly one non-zero element. The position k, k' of the non-zero element indicates the mapping between the two clusterings: for example if a_{13} is a non-zero element of A, it means that $C_1 \sim C'_3$. If the two clusterings are not identical but are very similar, then A will have some large elements (that indicate the mapping between the two clusterings) and some small elements for the data points on which the two clusterings don't agree. The more different the two clusterings are, the more blurred the difference between the "large" and "small" elements of the confusion matrix.

Example The table below shows two possible clusterings of a set of 10 points.

point i	1	2	3	4	5	6	7	8	9	10
$clust(i)$	1	1	1	2	2	3	3	3	3	3
$clust(i)$	3	3	3	1	2	2	2	2	1	1

The confusion matrix correspoding to the two clusterings is

	C'_1	C'_2	C'_3	
C_1	0	0	3	3
C_2	1	1	0	2
C_3	2	3	0	5
	3	4	3	

In this example, there perfect correspondence between two clusters ($C_1 = C'_3$) which is reflected in row 1 and column 3 of the confusion matrix. The correspondence between the remaining two clusters of each clustering is very weak, which is reflected in the block $A_{2:3,1:2}$ of the confusion matrix.

18.4 Mixtures: A statistical view of clustering

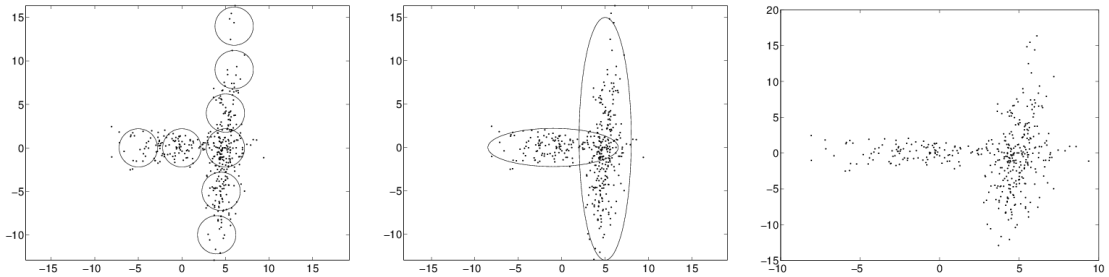
18.4.1 Limitations of the K-means algorithm

The K-means algorithm is simple to understand and implement, but has some serious limitations. For example, the K-means algorithm may fail to find the correct clustering when

- the clusters have different sizes
- the clusters are not spherical, but have elongated (and possibly different) shapes
- the data is rescaled, so that clusters that were previously round become elongated

Another problem is finding the correct number of clusters K, which the algorithm requires as an input. In this context note that clustering data has a subjective component: figure 18.4 shows a three different possibilities of clustering the same data set.

The clustering method that will be presented now, although similar to K-means, is much more flexible, allowing for clusters of arbitrary shape or elongation. It is also rooted in probability, which will provide us eventually with additional insights.



a. **b.** **c.**

Figure 18.4: Clustering data depends on the problem. Does this data set contain many circular clusters (**a**), two elliptical clusters (**b**), or just one cluster?

18.4.2 Mixture models

A **mixture of Gaussians** is a probability density given by

$$f(x) \;=\; \sum_{k=1}^{K} \lambda_k f_k(x) \tag{18.3}$$

where

- $f_k(x)$ are normal (Gaussian) densities with parameters μ_k, σ_k^2 called the **mixture components**
- $\lambda_k \geq 0$ are real numbers satisfying $\sum_{k=1}^{K} \lambda_k \;=\; 1$ called **mixture coefficients**

Figure 18.5 depicts a mixture of Gaussians and its components. Here we assume for simplicity that the data x are one dimensional but the model and algorithm can be generalized for any number of dimensions.

Intuitively, adopting a mixture (of Gaussians) model reflects the assumption that there are K sources that generate data independently (these are the $f_1,\ f_2,\ \ldots f_K$). The probability that an arbitrary data point is generated by f_k is λ_k. Thus, $(\lambda_1, \ldots \lambda_K)$ describe a discrete distribution over the sources. A new data point is generated in two steps: first, the source f_k is picked randomly from $f_1,\ f_2,\ \ldots f_K$ according to the probability given by $(\lambda_1, \ldots \lambda_K)$; second, the data point x is sampled from the chosen f_k. We observe x but we do not observe k, the index of the source that generated x. Because k is unobserved, it is called a **hidden variable**.

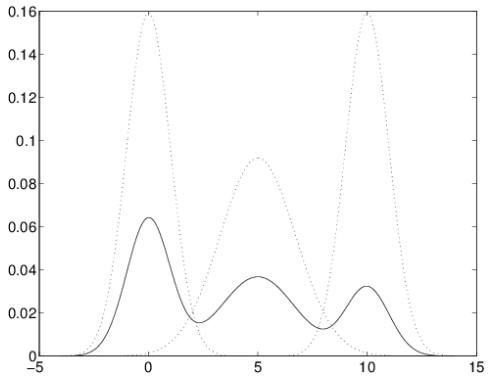


Figure 18.5: A mixture of normal distributions (full line) and its components (dotted line). The mixture components parameters are $\mu_1 = 0,\ \sigma_1^2 = 1,\ \mu_2 = 5,\ \sigma_2^2 = 3,\ \mu_3 = 10,\ \sigma_3^2 = 1$. The mixture coefficients are $\lambda_1 = 0.4,\ \lambda_2 = 0.4,\ \lambda_3 = 0.2$.

One can rewrite $f(x)$ so as to exhibit the two-step data generation model:

$$f(x) \ = \ \sum_{k=1}^{K} P(k) f(x \,|\, k) \tag{18.4}$$

where of course

$$P(k) \ = \ \lambda_k \quad \text{for } k = 1, \ldots K \tag{18.5}$$
$$f(x|k) \ = \ f_k(x) \tag{18.6}$$

In this probabilistic framework, the clustering problem can be translated as follows. Finding the clusters is equivalent to estimating the densities of the K data sources $f_1, \ldots f_K$. Assigning the data to the clusters means recovering the values of the hidden variable k for each data point.

18.5 The EM algorithm

The **Expectation-Maximization (EM)** algorithm solves the clustering problem as a Maximum Likelihood estimation problem. It takes as input the data $\mathcal{D} = \{x_1,\ x_2, \ldots x_n\}$ and the number of clusters K, and it outputs the model parameters $\Theta = \{\lambda_1, \ldots \lambda_K,\ \mu_1, \ldots \mu_K,\ \sigma_1^2, \ldots \sigma_K^2\}$ and the posterior probabilities of the clusters for each data point $\gamma_i(k)$, for $i = 1,\ \ldots n,\ k = 1, \ldots K$.

For any given set of model parameters Θ, we compute the probability $P(k|x_i)$

that observation x_i was generated by the k-th source f_k using Bayes formula

$$P(k|x_i) \;=\; \frac{P(k)f(x_i|k)}{\sum_{k'=1}^{K} P(k')f(x_i|k')} \;=\; \frac{\lambda_k f_k(x_i)}{\sum_{k'=1}^{K} \lambda_{k'} f_{k'}(x_i)} \;=\; \gamma_i(k) \qquad (18.7)$$

The values $\gamma_i(k),\ k = 1, \ldots K$ sum to 1. They are called the **partial assignments** of point x_i to the K clusters.

Algorithm EXPECTATION-MAXIMIZATION
Input $\{x_1, x_2, \ldots x_n\}$ the data points
K the number of clusters
Output$\gamma_i(k)$ for $i = 1, \ldots n$, $k = 1, \ldots K$
μ_k, σ_k^2 for $k = 1, \ldots K$ the parameters of the K mixture components
λ_k for $k = 1, \ldots K$ the mixture coefficients

Initialize μ_k, σ_k^2, λ_k for $k = 1, \ldots K$ with random values

Do
E step
for$i = 1, \ldots n$
$\gamma_i(k) \;=\; \frac{\lambda_k f_k(x_i)}{\sum_{k'=1}^{K} \lambda_{k'} f_{k'}(x_i)}$ for $k = 1, \ldots K$
M step
for $k = 1, \ldots K$
$n_k = \sum_{i=1}^{n} \gamma_i(k)$
$\lambda_k = \frac{n_k}{n}$
$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_i(k) x_i$
$\sigma_k^2 = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_i(k)(x_i - \mu_k)^2$
until convergence

It can be proved that the EM algorithm converges. The parameters Θ obtained at convergence represent a local maximum of the likelihood $L(\Theta)$.

The complexity of each iteration is $\mathcal{O}(Kn)$.

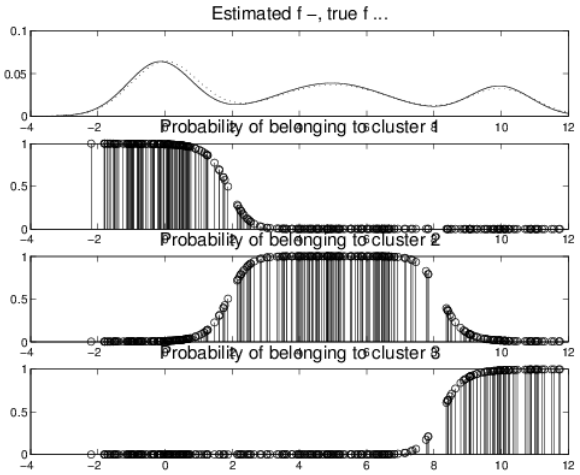


Figure 18.6: The results of the EM algorithm on a data set of size 300 sampled from the mixture distribution in figure 18.5. The top plot shows the estimated density (full) and the true density (dotted). The next three plots show the partial assignments to the clusters for each point in the data set (row 2 is γ_1, row 3 is γ_2, and row 4 is γ_3). Note that the points lieing between two clusters are have non-zero probabilities of belonging to each of the two clusters.

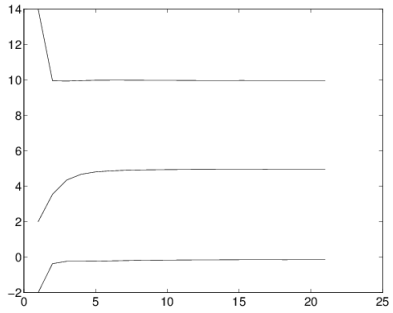


Figure 18.7: The convergence of the 3 cluster centers μ_1, μ_2, μ_3 from the initial values to the final ones for the data and EM algorithm illustrated in figure 18.6.