II - Probability Basics

Math 728 D - Machine Learning & Data Science - Spring 2019

Contents

Why Probability?

Information given in terms of data (measurements, observations,...) is inherently uncertain.

Probability Theory and Statistics provide proper mathemtical tools to quantify and manipulate uncertainties.

The idea behind quantifying the likelihood of certain events to occur is simple. Suppose you conduct a random experiment, e.g. a fair coin-toss, and record the number $n(N)$ of “heads” among N tosses. The limit P of the fraction $n(N)/N$, as the number N of experiments tends to infinity, is the “probability” - a number between zero and one - of “heads” to happen.

To put this on firm mathematical grounds one treats “events” as subsets $A, B, \ldots,$ of a “sample space” Ω of possible outcomes from which instances/samples are drawn. “Measure” and “Probability Theory” serve as a proper framework. An exhaustive treatment of these topics requires a course by itself. This section introduces some relevant basic notions and a corresponding “way of thinking” to an extent needed in the context of Data Science and Machine Learning.

Probability Spaces Proper framework: Measure Theory ...

Key notions:

- Sample or “outcome” space Ω;
- Set of “events” $\mathcal{B}$ comprised of subsets of Ω, i.e., $\mathcal{B} \subseteq 2^{\Omega}$. Only those subsets of Ω qualify as events that can be measured;
- The probability distribution function or probability measure (function) P returns probabilities of events, i.e., $P : \mathcal{B} \to [0, 1]$.

$P, \mathcal{B}$ are subject to certain structural requirements:

- $P(\Omega) = 1$;
- $\Omega_j \subset \Omega$, $j \in \mathbb{N}$, $\Omega_j \cap \Omega_j = \emptyset$, $j \neq k$, $\Rightarrow$

$$P\Big(\bigcup_{j\in\mathbb{N}} \Omega_j\Big) = \sum_{j\in\mathbb{N}} P(\Omega_j);$$

- $\mathcal{B}$ is σ-algebra: (a) $\Omega \in \mathcal{B}$; (b) $A \in \mathcal{B} \;\Rightarrow\; A^c := \Omega \setminus A \in \mathcal{B}$; (c) $A_i \in \mathcal{B}, i \in \mathbb{N} \Rightarrow \bigcup_{i\in\mathbb{N}} A_i \in \mathcal{B}$

  de Morgan’s law $((A \cup B)^c = A^c \cap B^c)$: (b), (c) $\Rightarrow \bigcap_{j\in\mathbb{N}} A_j \in \mathcal{B}$;

Definition 1

A tripel $(\Omega, \mathcal{B}, P)$ is called probability space

Examples

Flipping a fair coin: (discrete probability space) $\Omega = \{H, T\}$, $\mathcal{B} = \big\{\emptyset, \{H\}, \{T\}, \{H, T\}\big\} \rightsquigarrow$

$$P(\emptyset) = 0, \quad P(\{H\}) = P(\{T\}) = 1/2, \quad P(\{H, T\}) = 1$$

Flipping a fair coin twice: $\Omega = \{HH, HT, TH, TT\}$, $\#(\Omega) = 2^2 = 4$;
Maximal σ-algebra: $\mathcal{B} = 2^\Omega$, $\#(\mathcal{B}) = 16$
events: at most one head: $\{HT, TH, TT\}$; at least one head: $\{HH, HT, TH\}$, etc.

$$P(HH) = P(HT) = P(TH) = P(TT) = 1/4,\ P(\{HH, HT\}) = 1/2,\ P(\{HH, HT, TH\}) = 3/4, \text{...etc.}$$

A sub-σ-algebra: all events with first toss fixed: $A := \{HH, HT\}$, $B := \{TH, TT\}$,
$\mathcal{B} = \{\emptyset, A, B, \Omega\}$

$$P(A) = P(B) = 1/2$$

Borel field: Suppose Ω is a metric space (openness of subsets is defined) $\rightsquigarrow$ the minimal subset of σ-algebra in 2^Ω that contains all open subsets of Ω is a σ-algebra (Borel- σ-algebra)

Properties of Probability Measures

1. $P(A^c) = 1 - P(A)$;
2. $P(B \cap A^c) = P(B) - P(A \cap B)$;
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
4. $P\big(\bigcup_{i \in \mathcal{I}} A_i\big) \leq \sum_{i \in \mathcal{I}} P(A_i)$;
5. $\{C_i\}_{i\imath\mathcal{I}}$ a partition of $\Omega \Rightarrow P(A) = \sum_{i \in \mathcal{I}} P(A \cap C_i)$;
6. $A \subseteq B \Rightarrow P(A) \leq P(B)$.

Exercise: verify these statements

Lebesgue measure: measures subsets of $\mathbb{R}^d$ (volume measure); details on Lebesgue integration will be provided when needed for essential understanding, here only some comments on the basic ideas: for $D \subset \mathbb{R}^d$ define the Lebesgue outer measure as

$$\lambda^*(D) := \inf \left\{ \sum_{k \in \mathbb{N}} \mathrm{vol}_d(R_k) : (R_k)_{k \in \mathbb{N}} \text{ any sequence of } d\text{-hyperrectangles with } D \subseteq \bigcup_{k \in \mathbb{N}} R_k \right\}$$

Define the Lebesgue σ-algebra $\mathcal{B}(\mathbb{R}^d)$ of Lebesgue measurable sets by

$$\mathcal{B}(\mathbb{R}^d) := \{D \subset \mathbb{R}^d : \lambda^*(A) = \lambda^*(A \cap D) + \lambda^*(A \cap D^c), \ \forall\ A \subset \mathbb{R}^d\}$$

One can show that this is indeed a σ-algebra with measure λ ($\lambda(D) = \lambda^*(D)$ when $D \in \mathcal{B}(\mathbb{R}^d)$).
But there exist subsets of $\mathbb{R}^d$ that are not measurable! (see Vitali sets)

Random Variables

$(\Omega, \mathcal{B}, P)$ probability space, $\mathcal{X}$ a measurable space, a function $X : \Omega \to \mathcal{X}$ is called measurable if for any measurable set $S \subseteq \mathcal{X}$

$$\{\omega \in \Omega : X(\omega) \in S\} \in \mathcal{B}, \quad P_X(X \in S) := P(\{\omega \in \Omega : X(\omega) \in S\}).$$

A measurable function $X : \Omega \to \mathcal{X}$ is called a (n $\mathcal{X}$-valued) random variable. When $\mathcal{X} = \mathbb{R}$ we just say “random variable”.

Example: 2-coin toss

$$t_i := \begin{cases} 1 & \text{if} \quad H \text{ occurs in } i\text{th coin toss} \\ 2 & \text{if} \quad T \text{ occurs in } i\text{th coin toss} \end{cases}$$

1) $X := t_1 + t_2$

X	2	3	3	4
Ω	HH	HT	TH	TT

2) $X = \#(\text{heads})$: let us illustrate how a random variable can be used to generate from the original probability space a new one:

Ω	$P(\cdot)$	X	P_X
HH	$1/4$	2	$1/4$
HT	$1/4$	1	$1/4$
TH	$1/4$	1	$1/4$
TT	$1/4$	0	$1/4$

$$\rightsquigarrow \quad X = \begin{cases} 0 & \text{with probability } 1/4 \\ 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \end{cases} \quad \rightsquigarrow \quad (\{0,1,2\}, 2^{\{0,1,2\}}, P_X)$$

Induced Probability Space

In general, consider a probability space $(\Omega, \mathcal{B}, P)$, $X : \Omega \to \mathcal{X}$ a random variable. This induces a probability space

$$(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_X), \quad \mathcal{B}_X(\mathcal{X}) = \{X(A) : A \in \mathcal{B}\}, \quad P_X(G) = P(X^{-1}(G)),\ G \in \mathcal{B}_X(\mathcal{X}).$$

One sometimes writes $X \sim (\mathcal{X}, \mathcal{B}_X(\mathcal{X}), P_X)$.

$\mathcal{X}$ can be quite general, even a function space. To a large extent it suffices, however, to consider $\mathcal{X} = \mathbb{R}$ to introduce the basic facts about random variables. Typical scenario:

- sample space $\mathcal{X} = \mathbb{R}$ (continuous random variable, or $\mathcal{X} = \mathcal{N}, \mathcal{X} = \mathcal{I}$, $\mathcal{I}$ finite or countable (discrete random variable);
- event space $\mathcal{B}(\mathbb{R})$ the Borel σ-algebra of the real line, generated by all open intervals (or half-lines, or closed intervals,...);
- probability measure:

$$P_X(A) := P(\{\omega \in \Omega : X(\omega) \in A\} =: P(X^{-1}(A)) =: P(X \in A), \quad A \in \mathcal{B}(\mathbb{R});$$

Implicit assumption: $\mathcal{B}(\mathbb{R}) = \mathcal{B}_X(\mathbb{R})$, i.e., $X^{-1}(A) \in \mathcal{B}(\Omega)$ for all $A \in \mathcal{B}(\mathbb{R})$.

Example: $X(\omega) = |\omega|$, $(\Omega = [-1, 1], \mathcal{B}([-1, 1]), \lambda/2) \rightsquigarrow$ sample space $X(\Omega) = [0, 1] =: \mathcal{X}$, event space $\mathcal{B}_X([0, 1]) = \mathcal{B}([0, 1])$ (Borel-σ-algebra), probability measure

$$P_X(A) = P(X(\omega) \in A) = \lambda(\{\omega \in [-1, 1] : \omega \in A, -\omega \in A\})/2 = \lambda(A), \quad A \in \mathcal{B}([0, 1]);$$

e.g.: $P_X\left(\left[\frac{1}{3}, \frac{2}{3}\right]\right) = \frac{\lambda\left(\left[\frac{1}{3}, \frac{2}{3}\right]\right)}{2} + \frac{\lambda\left(\left[-\frac{2}{3}, -\frac{1}{3}\right]\right)}{2} = \lambda\left(\left[\frac{1}{3}, \frac{2}{3}\right]\right) = \frac{1}{3}$

Conditional Probability

Probability space $(\Omega, \mathcal{B}, P)$

Definition 2

If $A, B \in \mathcal{B}$, and $P(B) > 0$, then the conditional probability of A given B, denoted $P(A|B)$, is

$$P(A|B) := \frac{P(A \cap B)}{P(B)}. \tag{5.1}$$

Interpretation: $P(A|B)$ is the probability that an outcome of an experiment is in A if one knows it is in B.

Simple facts:

- $A \subset B \Rightarrow P(A|B) = P(A)/P(B) < 1;$
- $B \subset A \Rightarrow P(A|B) = P(B)/P(B) = 1.$

Note: Conditioning on an event B means that B becomes the sample space for a new probability space for which $P(\cdot|B)$ is the probability measure, i.e.,

$$\Omega \to B, \quad \mathcal{B} \to \{A \cap B : A \in \mathcal{B}\}, \quad P \to P(\cdot|B)$$

Example

3 doors, numbered $1, 2, 3$; a prize has been placed randomly behind one door

- Jack bets that the prize is behind door 1;
- Nancy opens door 2 and reveals, no prize behind door 2;
- Nancy asks Jack: "do you want to switch your bet to door 3?"

Jack uses the following probabilistic model to decide whether to switch:

$\Omega = \{1, 2, 3\}^2 = \{(D, N) : D, N \in \{1, 2, 3\}\}$
$D =$ random variable that prize is behind door D,
$N =$ random variable that Nancy opens door N;

Define a probability measure P:
Since Jack bets 1, Nancy never opens 1; $N = D$ leaves no decision to be made, so should not influence decision; since each door is equally probable $(D = 1, N = 2)$, $(D = 1, N = 3)$ have probability $\frac{1}{3} \cdot \frac{1}{2}$; $D = 2, 3$ happens with pobability $\frac{1}{3}$ leaving as relevant events only $N = 3, 2$, respectively

D	1	1	1	2	2	2	3	3	3
N	1	2	3	1	2	3	1	2	3
Prob	0	$\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$	$\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$	0	0	$\frac{1}{3} \cdot 1 = \frac{1}{3}$	0	$\frac{1}{3} \cdot 1 = \frac{1}{3}$	0

Jack should switch if $P(D = 1|N = 2) < P(D = 3|N = 2)$. Since $P(D = 1|N = 2) + P(D = 3|N = 2) = 1$ (why?) Jack should switch if $P(D = 3|N = 2) > 0.5$

$$P(D = 3|N = 2) = \frac{P(D = 3, N = 2)}{P(N = 2)} = \frac{1/3}{\frac{1}{6} + \frac{1}{3}} = \frac{2}{3} \quad \rightsquigarrow \text{ switch}$$

Baye’s Rule

Baye’s Rule:

$A, B \in \mathcal{B}$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}. \tag{5.2}$$

In fact,

$$P(A \cap B) \overset{Def.2}{=} P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \quad \Rightarrow \quad (5.2)$$

More generally: $A_i \in \mathcal{B}, i \in \mathcal{I}$, disjoint events partitioning Ω, then

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i \in \mathcal{I}} P(B|A_i) \cdot P(A_i)}, \tag{5.3}$$

since

$$P(B) = \sum_{i \in \mathcal{I}} P(B \cap A_i) \overset{Def.2}{=} \sum_{i \in \mathcal{I}} P(B|A_i) \cdot P(A_i).$$

Statistical Independence

Two events $A, B \in \mathcal{B}$ are (statistically) independent if and only if

$$P(A \cap B) = P(A) \cdot P(B). \tag{5.4}$$

Independence $\Rightarrow \quad P(A|B) = P(A), \quad P(B|A) = P(B).$

Example: 2-coin-toss: the two events “first toss = head” (HH, HT) and “second toss = head” (HH, TH) are independent.

$A_1, \ldots, A_n \in \mathcal{B}$ are mutually independent iff

$$\text{for any } \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\},\ k \leq n, \quad \Rightarrow \quad P\Big(\bigcap_{j=1}^{k} A_{i_j}\Big) = \prod_{j=1}^{k} P(A_{i_j})$$

CDF and PDF

Consider a probability space $(\Omega, \mathcal{B}, P)$ and a random variable $X \sim (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$:

Cumulative Distribution Function

$$\mathrm{Prob}(X \leq x) := F_X(x) := P_X(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}. \tag{6.1}$$

Recall: $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{B}(\mathbb{R})$.

For a discrete random variable, this is understood as a step function which is continuous from the right.

Proposition 3

F is a CDF iff

1. $\lim_{x\to\infty} F(x) = 1,\ \lim_{x\to-\infty} F(x) = 0$;
2. *F is non-decreasing;*
3. *F is right-continuous, i.e., for every $x_0 \in \mathbb{R}$, $\lim_{x\downarrow x_0} F(x) = F(x_0)$;*
4. *for $\epsilon > 0\ \exists\ M$ such that $\mathrm{Prob}(|X| > M) < \epsilon$ (no complete mass concentration at infinity).*

CDF and PDF

For a discrete random variable $X \sim (\mathcal{I}, \mathcal{B}(\mathcal{I}), P_X)$ the Probability Mass Function (PMF) is defind as

$$f_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}), \quad x \in \mathcal{I}. \tag{6.2}$$

For a continuous random variable $X \sim (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ the Probability Density Function f_X is given by

$$\mathrm{Prob}(X \leq x) := F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \quad \rightsquigarrow \quad P_X(A) = P(X \in A) = \int_A f_X(t)dt =: \int_A dF_X \tag{6.3}$$

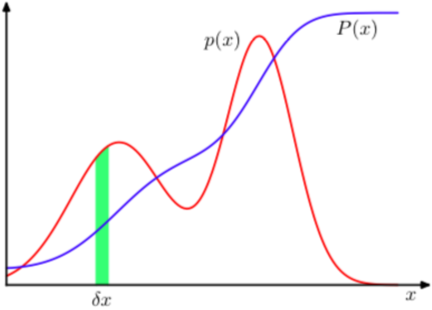
Note: (Fundamental Theorem of Calculus) if f_X is continuous, then $f_X(x) = F'_X(x)$

Proposition 4

A function $f_X(x)$ is a PMF or PDF if and only if

1. $f_X(x) \geq 0$ *for all* x;
2. $$\sum_{x \in \mathcal{I}} f_X(x) = 1 \quad \text{if } X :\to \mathcal{I} \text{ is discrete;} \qquad \int_{\mathbb{R}} f_X(x)dx = 1 \quad \text{if } X \text{ is continuous.}$$

CDF and PDF



Blue curve: CDF $P(x) = F_X(x)$
Red curve: PDF $p(x)$

Interpretation: the probability of X falling into an interval $(x, x + \delta x)$ is given approximately by $p(x)\delta x$. This becomes precise when $\delta x \to 0$ which reflects that the density is the derivative of the cumulative distribution.

Conditional CDF and PDF

Let $X \sim (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$, $A \in \mathcal{B}(\mathbb{R}) \rightsquigarrow$ Conditional CDF

$$P_X(X \leq x | X \in A) := \frac{P_X(\{X(\omega) : \omega \in \Omega,\ X(\omega) \leq x\} \cap A)}{P_X(A)} \qquad \text{briefly:} \quad \frac{P_X(\{X \leq x\} \cap A)}{P_X(A)}$$

Again: differentiation yields Conditional PDF

Example: fix $z \in (0,1)$, $X \sim ([0,1], \mathcal{B}([0,1]), \lambda)$, conditioning on $x > z$:

$$P_X(X \leq x | X \geq z) = \begin{cases} 0, & \text{if} \quad x \leq z, \\ \frac{x-z}{1-z}, & \text{if} \quad x > z. \end{cases}$$

PDF:

$$f_{X|X \geq z}(x) = F'_{X|X \geq z}(x) = \frac{1}{1-z}.$$

Joint Probability Measure

Consider two random variables $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ (on $(\Omega, \mathcal{B}, P)$), $A \in \mathcal{B}(\mathcal{X})$, $B \in \mathcal{B}(\mathcal{Y})$,

$$P_{X,Y}(A, B) := P(X \in A, Y \in B) = P(\{\omega \in \Omega : X(\omega) \in A,\ Y(\omega) \in B\})$$

(in the discrete case $P_{X,Y}(x, y) := P(X = x, Y = y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$)

One has the following equivalent ways of expressing this (as an exercise)

$$\begin{aligned} P_{X,Y}(A, B) &= P(\{\omega \in \Omega : X(\omega) \in A\} \cap \{\omega \in \Omega : Y(\omega) \in B\}) = P(X^{-1}(A) \cap Y^{-1}(B)) \\ &\overset{(5.1)}{=} P(X^{-1}(A)|Y^{-1}(B))P(Y^{-1}(B)) = P(X \in A|Y \in B)P(Y \in B) \\ &= P(Y^{-1}(B)|X^{-1}(A))P(X^{-1}(A)) = P(Y \in B|X \in A)P(X \in A) \qquad (7.1) \end{aligned}$$

Suppose $\{J\}$ is a partition of $\mathcal{Y}$: $\rightsquigarrow$

$$\begin{aligned} \sum_J P_{X,Y}(A, J) &= \sum_J P(X \in A, Y \in J) = \sum_J P(Y \in J|X \in A)P(X \in A) \\ &= P(X \in A) = P_X(A) \quad \text{(since } \textstyle\sum_J P(Y \in J|X \in A) = 1\text{)} \qquad (7.2) \end{aligned}$$

called marginal probability since one variable is “summed out”

Notational Conventions: Marginals

X, Y discrete:

$$P_{X,Y}(X=x, Y=y) =: P_{X,Y}(x,y) \quad \rightsquigarrow \quad P_X(x) = P_{X,Y}(x) = \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) \quad \text{sum rule}$$

X, Y continuous, recall (6.3), think of the partition $\{J\}$ to become finer and finer, summation $\to$ integration, $P(Y\in J) = P_Y(J) = \int_J dF_Y \rightsquigarrow$ (Fubini's Theorem)

$$P_{X,Y}(A) = \int_{\mathcal{Y}} P(X\in A, Y=y)dy, \quad f_{X,Y}(x) = \int_{\mathcal{Y}} f_{X,Y}(x,y)dy, \qquad (7.3)$$
$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x).$$

Sometimes for $p(X,Y) = f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y)$ just briefly:

$$p(X) = \int_{\mathcal{Y}} p(X,Y)dY, \quad p(X,Y) = p(Y|X)p(X)$$

Baye's Rule:

$$p(X|Y) = \frac{p(Y|X)\cdot p(X)}{p(Y)}$$

Interpretation: $p(X) \leftrightarrow$ "prior distribution"; $p(Y) \leftrightarrow$ "data distribution"; $p(X|Y) \leftrightarrow$ "posterior distribution" given the "data/observations" y; $p(Y|X) \leftrightarrow$ "likelihood (of the data, given the prior X) function.

Independent Random Variables

The random variables X, Y are called independent if the joint distribution factors into its marginals (see (5.4)):

$$\begin{aligned} P_{X,Y}(A,B) &= P(X\in A, Y\in B) = P(\{\omega\in\Omega : X(\omega)\in A\}\cap\{\omega\in\Omega : Y(\omega)\in B\}) \\ &= P(X\in A)\cdot P(Y\in B) = P_X(A)\cdot P_Y(B), \end{aligned} \tag{7.4}$$

i.e., the events $\{\omega\in\Omega : X(\omega)\in A\}$, $\{\omega\in\Omega : Y(\omega)\in B\}$ are statistically independent (see (5.4)). In particular this means $P(X\in A|Y\in B) = P(X\in A)$.
In terms of densities: $p = f_X$

$$p(X,Y) = p(X|Y)p(Y) = p(X)\cdot p(Y) \tag{7.5}$$

Integration: Simplified notation $X\sim(\mathcal{X},\mathcal{B}(\mathcal{X}),P_X)\rightsquigarrow X\sim(\mathcal{X},\mathcal{B},P)$ (often $\mathcal{X}=\mathbb{R}$)

$f:\mathcal{X}\to\mathbb{R}$ measurable

$$\int_{\mathcal{X}} f(x)dP := \sup_{\{E_i\}_{i\in\mathcal{I}}}\Big\{\sum_{i\in\mathcal{I}}\big(inf_{x\in E_i}f(x)\big)P(E_i)\Big\}, \tag{7.6}$$

where the supremum is taken over all partitions $\{E_i\}_{i\in\mathcal{I}}$ of $\mathcal{X}$. Often one writes $dP(x) = p(x)dx$.

Identically Distributed Random Variables

Definition 5

Random variables X, Y are called identically distributed (i.d.) if for every set $A \in \mathcal{B}(\mathbb{R})$

$$P_X(X \in A) = P_Y(Y \in A) \qquad \text{i.e., } P(\{\omega \in \mathbb{R} : X(\omega) \in A\}) = P(\{\omega \in \mathbb{R} : Y(\omega) \in A\}). \tag{7.7}$$

Note: two identically distributed (i.d.) random variables X, Y need not be the same

Example: 2-coint toss $X =$ number of Heads, $Y =$ number of Tails, hower

$$X \text{ and } Y \text{ are i.d. } \Leftrightarrow F_X(z) = F_Y(z), \quad \forall z.$$

A “Fruitful” Example a summary of the previous notions

Experiment: two boxes (red and blue), the red one r contains 2 apples, a, and 6 oranges, o, the blue one b contains 3 apples and 1 orange;

randomly pick a box and from that box select randomly a fruit, each fruit being picked equally likely. Having observed the type of fruit (a or o), the fruit is put back.

Suppose we repeat this process many times finding that in 40% of the cases the selected box was r; $\rightsquigarrow$ probabilistic model: the identity (color) of the box is a discrete random variable B(ox) taking $n = 2$ values in $\mathcal{X} = \{r, b\}$, and $P(B = r) = 0.4 \rightsquigarrow P(B = b) = 0.6$. Likewise, the identity of the fruit is also a random variable, denoted by F(ruit), taking $m = 2$ values in $\{a, o\} = \mathcal{Y}$.

Question 1: suppose we have picked randomly a box that turned out to be blue b. Then, the probability of selecting an apple is just the fraction of apples in the blue box which is $3/4$, i.e., $P(F = a|B = b) = 3/4$. Similarly

$$P(F = a|B = r) = \frac{1}{4}, \quad P(F = o|B = r) = \frac{3}{4}, \quad P(F = a|B = b) = \frac{3}{4}, \quad P(F = o|B = b) = \frac{1}{4} \tag{7.8}$$

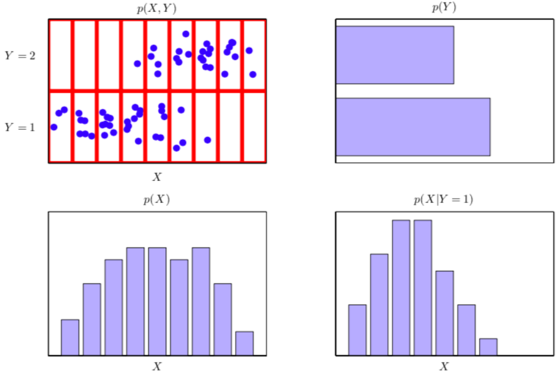
$\rightsquigarrow$ overall probability of selecting an apple:

$$P(F = a) = P(F = a|B = r)P(B = r) + P(F = a|B = b)P(B = b) = \frac{1}{4} \cdot \frac{4}{10} + \frac{3}{4} \cdot \frac{6}{10} = \frac{11}{20}$$

$\rightsquigarrow P(F = o) = 1 - P(F = a) = \frac{9}{20}$

Illustration (see [1, Chapter 1, § 1.2])

Illustration of a distribution over two variables X, taking 9 values, and Y, taking 2 values;



Top left: sample of 60 points drawn from a joint probability distribution over X, Y;

Remaining figures: histogram estimates for the marginals $p(y), p(x)$ and the conditional density $p(X|Y=1)$.

A “Fruitful” Example, Cont’d

Question 2: suppose we have selected an orange and we ask which box did it come from most likely. Now we ask for the probability over the box conditioned on the identity of the fruit. We know the probability over the fruits conditioned on the boxes. Baye’s Rule allows us to reverse the conditional probabilities:

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)} = \frac{3}{4} \cdot \frac{4}{10} \cdot \frac{20}{9} = \frac{2}{3}$$

Since $P(B = r|F = o) + P(B = b|F = o) = 1 \rightsquigarrow P(B = b|F = o) = 1 - 2/3 = 1/3$

Interpretation of Baye’s Theorem: if we had been asked for the identity of the box before being told the selected fruit, the most complete information is provided by $P(B)$. This is called Prior Probability (available before knowing the identity of the fruit). Once we are told $F = o$, Baye’s Theorem allows us to compute $P(B|F)$ which is called the Posterior Probability, the probability obtained after observing F.

More generally, in terms of densities $p = f_{X,Y}$: $F = Y = o$ represents observed data **w** while $B = X = r$ stands for the unknown parameters **x** whose posterior probability is to be estimated. Then Baye’s Theorem reads

$$p(\mathbf{x}|\mathbf{w}) = \frac{p(\mathbf{w}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{w})} \qquad \text{in words:} \quad \text{posterior} \propto \text{likelihood} \times \text{prior} \tag{7.9}$$

where $p(\mathbf{w}|\mathbf{x})$ is called the likelihood function

A “Fruitful” Example, Cont’d

Note: integrating (7.9) on both sides over $\mathbf{x}$ and using that $\int_{\mathcal{X}^d} p(\mathbf{x}|\mathbf{w}) d\mathbf{x} = 1$, yields

$$p(\mathbf{w}) = \int_{\mathcal{X}^d} p(\mathbf{w}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{7.10}$$

Note: the prior probability of selecting the red box was $P(B = r) = \frac{4}{10}$ so we were more likely to select the blue box. However, after observing the selected fruit, we find that the posterior probability of selecting the red box is $P(B = r|F = o) = \frac{2}{3}$, so the red box is actually more likely to be selcted

Intuition: the proportion of oranges in the red box is higher which makes it more likely to come from the red box $\rightsquigarrow$ likelihood function.

If the fraction of oranges and apples were the same in both boxes, the identity of the fruit would not provide any additional information and $P(F|B) = P(F)$ so that $P(B, F) = P(B) \cdot P(F)$ and the probability of selecting a particular fruit is independent of which box has been picked. In this case the random variables B, F are independent.

Remark: When $X \sim (\mathcal{X}, \mathcal{B}, P)$ is a random variable and $f : \mathcal{X} \to \mathbb{R}$ a measurable function, the $f(X)$ is also a random variable.

Baysian Probabilities

So far: probabilities are viewed as frequencies of repeatable events - the “frequentist’s point of view” in statistics.

Baysian Interpretation of Probability: probabilities provide a quantification of uncertaintiy. For instance, will the Arctic cap have disappeared by the end of the century? This cannot be assessed by inference from repeated events numerous times.

Instead: we have an idea of how quickly the ice is melting, e.g. based on an existing physical (mathematical) model.

If we obtain fresh evidence, e.g. by new satellite measurements or novel forms of diagnostic information, we may revise our estimation of uncertainty to be used in subsequent actions or decisions.

Now the issue becomes to quantify “degrees of belief” by numerical values. This can be done in an axiomatic way which eventually leads to “rules” for manipulating “degrees of belief” that are equivalent to the sum- and product rules of probability.

We’ll compare later merits or disadvantages of frequentist versus Baysian approaches.

Statistical Moments Expectation

$X \sim (\mathcal{X}, \mathcal{B}, P)$: the Expectation of a function $f : \mathcal{X} \to \mathbb{R}$ is a “weigted average” of the function values according to the underlying probability density $p(x) = f_X(x)$ (see (7.6)):

$$\mathbb{E}[f] := \sum_{x \in \mathcal{X}} f(x)p(x) \quad \text{(discrete)}, \qquad \mathbb{E}[f] := \int_{\mathcal{X}} f(x)p(x)dx \quad \text{(continuous)} \tag{8.1}$$

Expectations are typically approximated via sampling (see Theorem 11 later below)

$$\mathbb{E}[f] \approx \frac{1}{n} \sum_{j=1}^{N} f(x_j), \tag{8.2}$$

where the x_j are i.i.d. (independent, identically distributed) random samples from $\mathcal{X}$. Note: the samples x_j will cluster in regions where p attains large values.

Marginals, conditional expectations: X, Y random variables

$$\mathbb{E}_X[f(\cdot, y)] = \int_{\mathcal{X}} f(x, y)dx, \quad \mathbb{E}_X[f|y] = \mathbb{E}[f|y] := \int_{\mathcal{X}} f(x)p(x|y)dx. \tag{8.3}$$

Let $\mathbf{1}_A(x)$ denote the (set) characteristic function, i.e., $\mathbf{1}_A(x) = 1$ if $x \in A$ and zero otherwise. Then

$$\mathbb{E}\big[\mathbf{1}_A(X)\big] = \int_A p(x)dx = \mathrm{Prob}(X \in A), \tag{8.4}$$

the expectation of set-characteristic functions completely describes the probability distribution.

Variance $d = 1$

Variance describes the fluctuation of f around its mean (everything is analogous in the discrete case $\sum \leftrightarrow \int$):

$$\begin{aligned}\mathrm{var}[f] &:= \mathbb{E}\left[(f-\mathbb{E}[f])^2\right] = \int_{\mathcal{X}} (f(x)^2 - 2f(x)\mathbb{E}[f] + \mathbb{E}[f]^2)p(x)dx \\ &= \int_{\mathcal{X}} f(x)^2 p(x)dx - 2\mathbb{E}[f]\int_{\mathcal{X}} f(x)p(x)dx + \mathbb{E}[f]^2\int_{\mathcal{X}} p(x)dx \\ &= \mathbb{E}\left[f^2\right] - 2\mathbb{E}[f]^2 + \mathbb{E}[f]^2 = \mathbb{E}\left[f^2\right] - \mathbb{E}[f]^2. \end{aligned} \tag{8.5}$$

Also since $\mathbb{E}[c] = c$ for any constant c one easily verifies

$$\mathrm{var}[X+c] = \mathrm{var}[X], \quad \mathrm{var}[cX] = c^2\mathrm{var}[X], \quad c \text{ constant.} \tag{8.6}$$

Remark 6

Since $\int_{\mathcal{X}} (f(x) - \mathbb{E}[f])p(x)dx = \mathbb{E}[f] - \mathbb{E}[f] = 0$, Lecture I, Theorem 24 $\Rightarrow$ $\mathrm{var}[f]$ is the square of the $L_2(\mathcal{X}, p)$ best approximation error of f by constants:

$$\mathrm{var}[f] = \inf_{c\in\mathbb{R}} \int_{\mathcal{X}} (f(x) - c)^2 p(x)dx. \tag{8.7}$$

$\mathbb{E}[X]$, $\mathrm{var}[X]$ are given by the first and second order moments $\int_{\mathcal{X}} xp(x)dx$, $\int_{\mathcal{X}} x^2p(x)dx$.

Moments for Independent Random Variables

Lemma 7

For random variables $X, Y \sim (\mathbb{R}, \mathcal{B}(R), P)$*, one has*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]. \tag{8.8}$$

if X, Y *are independent, one has*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y], \tag{8.9}$$

and

$$\mathrm{var}[X + Y] = \mathrm{var}[X] + \mathrm{var}[Y]. \tag{8.10}$$

Proof: Let $p = f_{X,Y}$. $\rightsquigarrow \mathbb{E}[\alpha X + \beta Y] = \int\limits_{\mathbb{R}^2} (\alpha x + \beta y) p(x, y) dxdy = \alpha \int\limits_{\mathbb{R}} xp(x)dx + \beta \int\limits_{\mathbb{R}} yp(y)dy$ (marginals) $\Rightarrow$ (8.8).

Concerning (8.9), one has

$$\begin{aligned} \mathbb{E}[XY] &= \int\limits_{\mathbb{R}^2} xyp(x, y)dxdy \overset{(7.5)}{=} \int\limits_{\mathbb{R}^2} xyp(x)p(y)dxdy \\ &= \Big(\int\limits_{\mathbb{R}} xp(x)dx \Big) \Big(\int\limits_{\mathbb{R}} yp(y)dy \Big) = \mathbb{E}[X] \cdot \mathbb{E}[Y] \quad \Rightarrow \quad (8.9) \end{aligned}$$

As for (8.10),

$$\begin{aligned} \mathrm{var}[X + Y] &\overset{(8.5)}{=} \mathbb{E}\left[(X + Y)^2\right] - \mathbb{E}[X + Y]^2 = \mathbb{E}\left[X^2 + 2XY + Y^2\right] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \mathbb{E}\left[X^2\right] + 2\mathbb{E}[XY] + \mathbb{E}\left[Y^2\right] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 + \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 \\ &= \mathrm{var}[X] + \mathrm{var}[Y]. \qquad \square \end{aligned}$$

Variance $d > 1$

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top, \mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ be vectors of random variables (with joint distribution P). The Covariance of $\mathbf{X}$ and $\mathbf{Y}$ is the rank-one matrix

$$\begin{aligned} \mathrm{cov}[\mathbf{X}, \mathbf{Y}] &:= \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top\right] \\ &= \mathbb{E}\left[\mathbf{X}\mathbf{Y}^\top\right] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]^\top \qquad (8.11) \end{aligned}$$

Exercise: Verify the last equality!

Abbreviate: $\mathrm{var}[\mathbf{X}] := \mathrm{cov}[\mathbf{X}, \mathbf{X}]$, so that

$$\mathrm{var}[\mathbf{X}] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top. \qquad (8.12)$$

The Characteristic Function better understanding of PDFs

The characteristic function of a random variable X with density $p(x)$ is defined as

$$\varphi_X(s) := \mathbb{E}\left[e^{-is\cdot}\right] = \int_{\mathbb{R}} p(x) e^{-isx} dx. \tag{8.13}$$

In other words, defining the Fourier transform of a function g as

$$(\mathcal{F}g)(s) := \int_{\mathbb{R}} g(x) e^{-isx} dx,$$

the characteristic function of a random variable X with density p is the Fourier transform of the underlying probability density

$$\varphi_X(s) = (\mathcal{F}p)(s). \tag{8.14}$$

Just as the expectation of the set-charactersitic function completely characterizes a distribution, the expectation of basic Fourier modes e^{-isx} does as well. By properties of the Fourier transform, if the characteristic functions of random variables agree so must the random variables. There are a number of important applications of the characteristic function, among them a convenient proof of the Central Limit Theorem (see later below).

Properties of Characteristic Functions

Remark 8

The characteristic function always exists, is uniformly continuous and $\varphi_X(0) = 1$, $|\varphi_X(s)| \leq 1$. *Moreover, if a random variable has* k*th moments, one has*

$$\mathbb{E}\big[X^k\big] = (-i)^k \varphi_X^{(k)}(0),$$

where i *is the imaginary unit* $i^2 = -1$. *This follows directly from corresponding properties of the Fourier transform.*

Remark 9

For any affine transformation $aX + b$ *of a random variable* X *one has* $\varphi_{aX+b}(s) = e^{-isb}\varphi_X(as)$.

Proposition 10

Suppose that X, Y *are jointly distributed independent random variables. Then*

$$\varphi_{X+Y}(s) = \varphi_X(s) \cdot \varphi_Y(s). \tag{8.15}$$

Proof: $\varphi_{X+Y}(s) = \int\limits_{\mathbb{R}^2} p(x,y)e^{-is(x+y)}dxdy \overset{(7.5)}{=} \int\limits_{\mathbb{R}\times\mathbb{R}} p(x)p(y)e^{-isx}e^{-isy}dxdy = \varphi_X(s) \cdot \varphi_Y(s)\square$

Law of Large Numbers ...

Write briefly: $\mathrm{Prob}(g(\mathbf{x}) \geq \epsilon) := P(\{\mathbf{x} \in \mathcal{X}^N : g(\mathbf{x}) \geq \epsilon\})$.

The Law of Large Numbers states in great generality that the mean of independent samples of a random variable with bounded variance becomes arbitrarily close to its expectation.

Theorem 11

Let $x_1, \dots, x_N$ be i.i.d. samples of a random variable $X \sim (\mathcal{X}, \mathcal{B}, P)$. Then

$$\mathrm{Prob}\left(\left|\frac{x_1 + \cdots + x_N}{N} - \mathbb{E}[X]\right| \geq \epsilon\right) \leq \frac{\mathrm{var}[X]}{N\epsilon^2}, \quad \epsilon > 0. \tag{9.1}$$

The proof is based on two basic but important inequalities: Markov's inequality and Chebyshev's inequality as well as on Lemma 7.

Markov's Inequality

Theorem 12

Let $X \sim (\mathbb{R}, \mathcal{B}, P)$ be a nonnegative random variable. Then one has for $a > 0$

$$\mathrm{Prob}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{9.2}$$

Proof: Let $p = f_X$ denote the density so that

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty xp(x)dx = \int_0^a xp(x)dx + \int_a^\infty xp(x)dx \geq \int_a^\infty xp(x)dx \\ &\geq a\int_a^\infty p(x)dx \geq a\,\mathrm{Prob}(X \geq a), \end{aligned}$$

which proves the assertion. The same argument works for the discrete case. □

Corollary 13

Under the above assumption one has $\mathrm{Prob}(x \geq b\mathbb{E}[X]) \leq \frac{1}{b}$, $b > 0$.

Chebyshev's Inequality

Markov's inequality bounds the tail of a distribution in terms of the mean. It is used to prove Chebyche's inequality which offers the following (in some sense sharper) tail bound.

Theorem 14

Let $X \sim (\mathbb{R}, \mathcal{B}, P)$ be a random variable. Then for $c > 0$

$$\mathrm{Prob}\left(|X - \mathbb{E}[X]| \geq c\right) \leq \frac{\mathrm{var}[X]}{c^2}. \tag{9.3}$$

Proof: $Y := |X - \mathbb{E}[X]|^2$ is a nonnegative random variable and, by definition, $\mathbb{E}[Y] = \mathrm{var}[X]$. Thus, Markov's inequality gives

$$\mathrm{Prob}(|X - \mathbb{E}[X]| \geq c) = \mathrm{Prob}(|X - \mathbb{E}[X]|^2 \geq c^2) = \mathrm{Prob}(Y \geq c^2) \overset{(9.3)}{\leq} \frac{\mathbb{E}[Y]}{c^2} = \frac{\mathrm{var}[X]}{c^2}.$$

□

Proof of Theorem 11

By Chebyshev’s inequality (9.3)

$$\begin{aligned}\mathrm{Prob}\left(\left|\frac{1}{N}\sum_{j=1}^{N} X_j - \mathbb{E}[X]\right| \geq \epsilon\right) &\leq \frac{\mathrm{var}\left[\frac{1}{N}\sum_{j=1}^{N} X_j\right]}{\epsilon^2} \overset{(8.6)}{=} \frac{\mathrm{var}\left[\sum_{j=1}^{N} X_j\right]}{N^2\epsilon^3} \overset{(8.10)}{=} \frac{\sum_{j=1}^{N}\mathrm{var}[X_j]}{N^2\epsilon^2} \\ &= \frac{\mathrm{var}[X]}{N\epsilon^2},\end{aligned}$$

since the X_j are i.e.d. This proves the assertion. □

Exercises

There are some general ideas behind bounding excess probabilities (tail bounds). The following exercises are to hint at them and will be taken up later again.

Exercise 15

1. *Show that for any non-negative random variable X*

$$\mathbb{E}[X] = \int_0^\infty \mathrm{Prob}(X \geq t)dt \tag{9.4}$$

and re-derive Markov's inequality (assume that for the CDF $F_X'(x) = p(x)$).

2. *Let $\phi(t)$ be any strictly monotonely increasing non-negative function. Show that for any random variable X and any $t \in \mathbb{R}$*

$$\mathrm{Prob}(X \geq t) \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}. \tag{9.5}$$

3. *Re-derive Chebyshev's inequality from (2): for an arbitrary random variable X and $t > 0$ one has*

$$\mathrm{Prob}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{var}[X]}{t^2}. \tag{9.6}$$

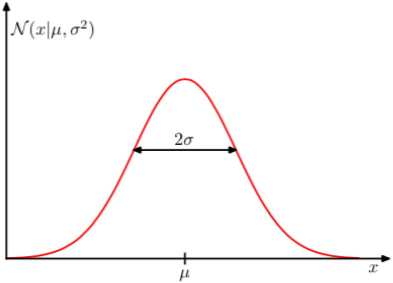
Gaussians

The perhaps most important distribution is the normal or Gaussian density

$$p(x) = \mathcal{N}(x|\mu,\sigma) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \tag{10.1}$$

involving two parameters: the mean μ and the variance σ^2 (σ is called standard deviation). In fact, straightforward calculation $\rightsquigarrow$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\mathcal{N}(x|\mu,\sigma)dx = \mu, \quad \mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2\mathcal{N}(x|\mu,\sigma)dx = \mu^2 + \sigma^2 \tag{10.2}$$



(8.5) $\rightsquigarrow$ $\mathrm{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[x]^2 = \sigma^2$

$\beta = 1/\sigma^2$ is called precision.

Remark 16

$\mathcal{N}(x|\mu,\sigma)$ attains its maximum (called "mode") at its mean μ and $\mathcal{N}(\mu|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}$ (find the zero of the derivative).

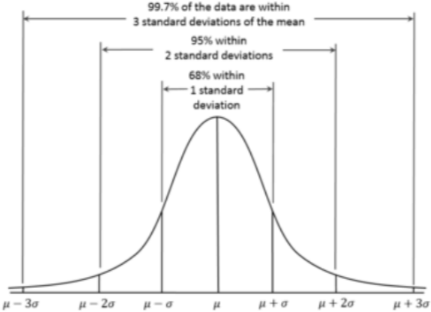
One writes $X \sim \mathcal{N}(\mu,\sigma^2)$

More Comments ...

Note: For $\varphi(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ one has $\varphi'(z) = -z\varphi(z)$, $\varphi''(z) = (z^2 - 1)\varphi(z)$

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sigma}\varphi\Big(\frac{x-\mu}{\sigma}\Big) \quad \Rightarrow \quad \frac{d}{dx}\mathcal{N}(x|\mu,\sigma^2)|_{x=\mu} = 0, \quad \frac{d^2}{dx^2}\mathcal{N}(x|\mu,\sigma^2)|_{x=\mu\pm\sigma} = 0,$$

i.e., $\mathcal{N}(x|\mu,\sigma^2)$ attains its maximum at the mean μ and has inflection points at $\mu \pm \sigma$.



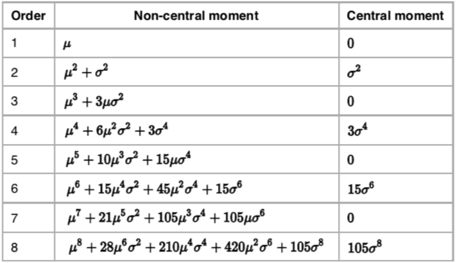
For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

Higher Moments of Gaussians

Mean and expectations are first and second order moments. Plain and absolute moments are the expectations of corresponding powers of the random variable: $\mathbb{E}\left[X^k\right]$, $\mathbb{E}\left[|X|^k\right]$. Bounds on moments of a probability density provide important structural information and lead e.g. to so called tail bounds used in quantifying the performance of estimators in machine learning. For the centered Gaussian $X \sim \mathcal{N}(0, \sigma^2)$ it can be shown that

$$\mathbb{E}\left[X^k\right] = \begin{cases} 0 & \text{if} \quad k \text{ is odd} \\ \sigma^k (k-1)!! & \text{if} \quad k \text{ is even} \end{cases} \qquad \mathbb{E}\left[|X|^k\right] = \sigma^k (k-1)!! \begin{cases} \sqrt{\frac{2}{\pi}} & \text{if} \quad k \text{ is odd} \\ 1 & \text{if} \quad k \text{ is even} \end{cases} \tag{10.3}$$

where $n!! := \prod_{j=1}^{\lceil \frac{n}{2} \rceil - 1} (n-2j)$. For non-centered densities $\mathcal{N}(\cdot|\mu, \sigma^2)$



Order	Non-central moment	Central moment
1	μ	0
2	$\mu^2 + \sigma^2$	σ^2
3	$\mu^3 + 3\mu\sigma^2$	0
4	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$	$3\sigma^4$
5	$\mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4$	0
6	$\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6$	$15\sigma^6$
7	$\mu^7 + 21\mu^5\sigma^2 + 105\mu^3\sigma^4 + 105\mu\sigma^6$	0
8	$\mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8$	$105\sigma^8$

A First Application of the Law of Large Numbers

Draw two points $\mathbf{z}, \mathbf{y} \in \mathbb{R}^d$ where the components z_j, y_j are realizations of independent random variables $Z, Y \sim \mathcal{N}(0,1)$. We are interested in the expected (squared) distance $\|\mathbf{y}-\mathbf{z}\|_2^2 = \sum_{j=1}^d (y_j - z_j)^2$. For each $j \in \{1, \dots, d\}$ one has

$$\mathbb{E}[(Y_j - Z_j)^2] = \mathbb{E}[Y_j^2] + \mathbb{E}[Z_j^2] - 2\mathbb{E}[Y_j Z_j] = \mathrm{var}[Y_j] + \mathrm{var}[Z_j] - 2\mathbb{E}[Y_j]\mathbb{E}[Z_j] \overset{(8.9)}{=} \mathrm{var}[Y_j] + \mathrm{var}[Z_j] = 2$$

since $\mathbb{E}[Y_j] = \mathbb{E}[Z_j] = 0$. Then Theorem 11 says that

$$\mathrm{Prob}\Big(\Big| \frac{1}{d} \sum_{j=1}^d (Y_j - Z_j)^2 - 2 \Big| \geq \epsilon \Big) \leq \frac{\mathrm{var}[(Y-Z)^2]}{d\epsilon^2} \to 0, \quad \text{as} \quad d \to \infty. \tag{10.4}$$

As seen above, all higher moments of Gaussians (expectation of powers of Gaussians) are finite, the right hand side indeed tends to zero as the dimension grows. Thus, with probability at least $1 - \frac{\mathrm{var}[(Y-Z)^2]}{d\epsilon^2}$ the instances satisfy

$$(2-\epsilon)d \leq \|\mathbf{y}-\mathbf{z}\|_2^2 \leq (2+\epsilon)d \tag{10.5}$$

Since by the above argument $\mathbb{E}[(Y_j - Z_j)^2] = \mathbb{E}[Y_j^2] + \mathbb{E}[Z_j^2] = 2$, i.e., $\mathbb{E}[\|\mathbf{Y}\|_2^2] = \mathbb{E}[\|\mathbf{Z}\|_2^2] = d$, one concludes that $\mathbf{y}$ and $\mathbf{z}$ are nearly orthogonal with high probability. Rescaling $\mathbf{Y}, \mathbf{Z}$ to unit length and choosing $\mathbf{y}$ as the north pole, almost all $\mathbf{z}$ concentrate near the equator. A detailed discussion will follow later.

Properties of Gaussians

(1) The substitution $(x-\mu)/\sigma \to z$ transforms $X \sim \mathcal{N}(\mu, \sigma^2)$ to standard form $Z \sim \mathcal{N}(0,1)$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} = 1$$

(2) The Fourier transform of a Gaussian is a Gaussian. Specifically, one has: (see (8.14))

Proposition 17

$$\varphi_Z(s) = (\mathcal{F}\mathcal{N}(\cdot|0,1))(s) = \sqrt{2\pi}\mathcal{N}(s|0,1) = e^{-\frac{1}{2}s^2}, \quad Z \sim \mathcal{N}(0,1). \tag{10.6}$$

More generally, when $X \sim \mathcal{N}(\mu, \sigma^2)$ one has

$$\varphi_X(s) = (\mathcal{F}\mathcal{N}(\cdot|\mu,\sigma^2))(s) = \sqrt{2\pi}e^{-is\mu}\mathcal{N}(\sigma s|0,1) = e^{-\frac{1}{2}\sigma^2 s^2 - is\mu}. \tag{10.7}$$

Corollary 18

For independent Gaussian random variables $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ the sum $X+Y$ is also Gaussian. More precisely

$$X + Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \tag{10.8}$$

Proof of Proposition 17 and Corollary 18

Proof of Proposition 17: Abbreviate $g(z) := (2\pi)^{-1/2}e^{-\frac{1}{2}z^2} = \mathcal{N}(z|0,1)$ and note that $g'(z) = -zg(z)$. Applying the Fourier transform to both sides, yields on the one hand

$$(\mathcal{F}g')(s) = \int_{\mathbb{R}} g'(z)e^{-isz}dz = -\int_{\mathbb{R}}(-is)g(z)e^{-isz}dz = (is)(\mathcal{F}g)(s),$$

which has to equal

$$-\int_{\mathbb{R}} zg(z)e^{-isz}dz = -\int_{\mathbb{R}} g(z)(-i)^{-1}\frac{d}{ds}e^{-isz}dz = -i\frac{d}{ds}(\mathcal{F}g)(s).$$

Therefore

$$\frac{\frac{d}{ds}(\mathcal{F}g)(s)}{(\mathcal{F}g)(s)} = -s, \quad \Rightarrow \quad \int_0^s \frac{\frac{d}{ds}(\mathcal{F}g)(s')}{(\mathcal{F}g)(s')}ds' = -\frac{s^2}{2}, \quad \Rightarrow \quad \ln((\mathcal{F}g)(s)) - \ln((\mathcal{F}g)(0)) = \ln((\mathcal{F}g)(s)) = -\frac{s^2}{2} \quad \Rightarrow \quad (10.6).$$

Since $X = \sigma Z + \mu$, Remark 9 says that

$$\varphi_X(s) = \varphi_{\sigma Z+\mu}(s) = e^{-is\mu}\varphi_Z(\sigma s) \overset{(10.6)}{=} \sqrt{2\pi}e^{-is\mu}\mathcal{N}(\sigma s|0,1),$$

which is (10.7). □

Proof of Corollary 18: Let p_{X+Y} denote the density of $X + Y$. Then, by (8.14) and Proposition 10,

$$(\mathcal{F}p_{X+Y})(s) = \varphi_{X+Y}(s) \overset{(8.14)}{=} \varphi_X(s)\varphi_Y(s) \overset{(10.7)}{=} e^{-is(\mu_X+\mu_Y)}e^{-\frac{1}{2}s^2(\sigma_X^2+\sigma_Y^2)} \overset{(10.7)}{=} (\mathcal{F}\mathcal{N}(\cdot|\mu_X+\mu_Y,\sigma_X^2+\sigma_Y^2))(s),$$

which confirms (10.8). □

Further Properties of Gaussians

It follows, as in the proof of Corollary 18 from the same factorization of the characteristic function of independent random variables, that for independent $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$

$$\sum_{j=1}^{n} X_j \sim \mathcal{N}\Big(\sum_{j=1}^{n}\mu_j, \sum_{j=1}^{n}\sigma_j^2\Big), \quad \frac{1}{n}\sum_{j=1}^{n} X_j \sim \mathcal{N}\Big(\frac{1}{n}\sum_{j=1}^{n}\mu_j, \frac{1}{n^2}\sum_{j=1}^{n}\sigma_j^2\Big). \tag{10.9}$$

Convolutions: $f, g \in L_2(\mathbb{R})$, $(f \star g)(x) = \int\limits_{\mathbb{R}} f(x-y)g(y)dy \in L_1(\mathbb{R})$. One easily verifies that

$$(\mathcal{F}(f \star g))(s) = (\mathcal{F}f)(s) \cdot (\mathcal{F}g)(s), \tag{10.10}$$

i.e., under the Fourier transform convolution becomes a pointwise multiplication.

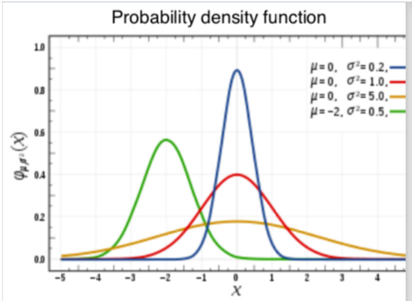
Proposition 19

The convolution of two Gaussians densities is a Gaussian density, i.e.,

$$(\mathcal{N}(\cdot|\mu_X, \sigma_X^2) \star \mathcal{N}(\cdot|\mu_Y, \sigma_Y^2))(x) = \mathcal{N}(x|\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \tag{10.11}$$

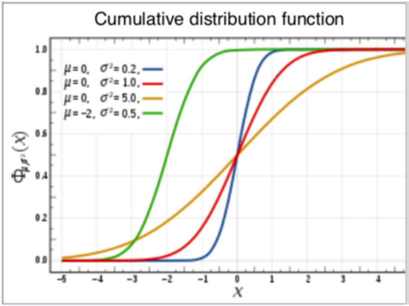
Proof: By (10.10), $(\mathcal{F}(\mathcal{N}(\cdot|\mu_X, \sigma_X^2) \star \mathcal{N}(\cdot|\mu_Y, \sigma_Y^2))(s) = (\mathcal{F}(\mathcal{N}(\cdot|\mu_X, \sigma_X^2))(s) \cdot (\mathcal{F}(\mathcal{N}(\cdot|\mu_Y, \sigma_Y^2))(s)$ which by (10.7), equals $e^{-is(\mu_X+\mu_Y)}e^{-\frac{1}{2}(\sigma_X^2+\sigma_Y^2)s^2}$. Again by (10.7), this is $(\mathcal{F}\mathcal{N}(\cdot|\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2))(s)$. □

An Approximation Aspect



Probability density function

The red curve is the *standard normal distribution*



Cumulative distribution function

One sees from the figures that small variance concentrate the density around the mean. The corresponding PDFs become closer and closer to a a step function. Since the integrals of the densities (by normalization) stay always equal to one the family $\{\mathcal{N}(\cdot|0,\sigma^2)\}_{\sigma>0}$ form a so called approximate identity. This means that for a given function f, the convolutions

$$(f\star\mathcal{N}(\cdot|0,\sigma^2))(x)=\int_{\mathbb{R}} f(y)\mathcal{N}(x-y|0,\sigma^2)dy=\int_{\mathbb{R}} f(x-y)\mathcal{N}(y|0,\sigma^2)dy$$

tend to f in $L_p(\mathbb{R})$, say

$$\lim_{\sigma\to 0}\|f-(f\star\mathcal{N}(\cdot|0,\sigma^2))\|_{L_p(\mathbb{R})}=0. \tag{10.12}$$

This agrees with (10.11) since $\lim_{\sigma\to 0}(\mathcal{N}(\cdot|\mu,\sigma_X^2)\star\mathcal{N}(\cdot|0,\sigma^2))(x)=\lim_{\sigma\to 0}\mathcal{N}(x|\mu,\sigma_X^2+\sigma)=\mathcal{N}(x|\mu,\sigma_X^2)$.

Multivariate Gaussians

For a vector $\mathbf{X} \in \mathbb{R}^d$ of random variables the d-variate (or spherical) Gaussian density is:

$$\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{A}) := \frac{1}{(2\pi)^{d/2}|\det \boldsymbol{A}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\mu)^\top \boldsymbol{A}^{-1}(\mathbf{x}-\mu)\Big\} \qquad (10.13)$$

where: $\mu \in \mathbb{R}^d$ is called the mean and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, symmetric positive definite, the (co)variance of $\mathbf{X}$. We'll verify that this terminology is justified.

A special case: suppose that $\mathbf{X}$ is a vector of independent random variables $X_1, \ldots, X_N$, which are jointly distributed $X_j \sim \mathcal{N}(\cdot|\mu, \sigma^2)$, $j = 1, \ldots, N$. Independence means (see (7.5)) that the joint probability density is the product of the marginals which are the univariate Gaussians. That is

$$p(\mathbf{x}) = \prod_{j=1}^N \mathcal{N}(x_j|\mu, \sigma^2) = \mathcal{N}(\mathbf{x}|(\mu, \ldots, \mu)^\top, \Sigma), \quad \Sigma := \mathrm{diag}\,(\sigma^2, \ldots .\sigma^2) = \sigma^2\mathbf{I}, \qquad (10.14)$$

which is an N-variate Gaussian with mean $(\mu, \ldots, \mu)^\top \in \mathbb{R}^N$ and variance $\sigma^2\mathbf{I}$.

Transformation to Standard Form a useful tool

By Lecture I, Theorem 32: $\boldsymbol{A} = \mathbf{U}\Lambda\mathbf{U}^\top$, $\boldsymbol{A}^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}^\top$ for some orthogonal matrix $\mathbf{U}$ and diagonal matrix Λ

$$(\mathbf{x}-\mu)^\top \boldsymbol{A}^{-1}(\mathbf{x}-\mu) = (\mathbf{U}^\top(\mathbf{x}-\mu))^\top \Lambda^{-1}\mathbf{U}^\top(\mathbf{x}-\mu) = \|\Lambda^{-1/2}\mathbf{z}\|_2^2 \text{ where } \mathbf{z} := \mathbf{U}^\top(\mathbf{x}-\mu).$$

Setting $\sigma_j^2 := \lambda_j$, this yields (since $|\det(\mathbf{U}^\top)| = 1$, see section on unitary matrices in Lecture I)

$$\begin{aligned}
\int_{\mathbb{R}^d} g(\mathbf{x})\mathcal{N}(\mathbf{x}|\mu,\boldsymbol{A})d\mathbf{x} &= \frac{1}{(2\pi)^{d/2}|\det\boldsymbol{A}|^{1/2}}\int_{\mathbb{R}^d} g(\mathbf{x})e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top\boldsymbol{A}^{-1}(\mathbf{x}-\mu)}d\mathbf{x} \\
&= \frac{1}{(2\pi)^{d/2}\prod_{j=1}^d \sigma_j}\int_{\mathbb{R}^d} g(\mathbf{U}\mathbf{z}+\mu)e^{-\sum_{j=1}^d \frac{1}{2\sigma_j^2}z_j^2}d\mathbf{z} \\
&= \int_{\mathbb{R}^d} g(\mathbf{U}\mathbf{z}+\mu)\prod_{j=1}^d \mathcal{N}(z_j|0,\sigma_j^2)d\mathbf{z} \\
&= \int_{\mathbb{R}^d} g(\mathbf{U}\mathbf{z}+\mu)\mathcal{N}(\mathbf{z}|0,\Lambda)d\mathbf{z}. \qquad (10.15)
\end{aligned}$$

Multivariate Gaussian: Expectation, Variance

Proposition 20

Let $\mathbf{X} \sim \mathcal{N}(\mu, \boldsymbol{A})$, $\mu \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, *symmetric positive definite. Then*

$$\mathbb{E}[\mathbf{X}] = \mu, \qquad \mathrm{var}[\mathbf{X}] = \boldsymbol{A}. \tag{10.16}$$

Proof: By (10.15),

$$\mathbb{E}[\mathbf{X}] = \int_{\mathbb{R}^d} \mathbf{x}\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{A})d\mathbf{x} = \int_{\mathbb{R}^d} (\mathbf{Uz} + \mu)\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z}, \quad \boldsymbol{A} = \mathbf{U}\Lambda\mathbf{U}^\top.$$

The subsequent arguments use that $\mathcal{N}(\mathbf{z}|0, \Lambda) = \prod_{j=1}^d \mathcal{N}(z_j|0, \sigma_j^2)$ ($\lambda_j =: \sigma_j^2 > 0$) is a product of 1$d$-centered (normalized) Gaussians, i.e.,

$$\int_{\mathbb{R}} \mathcal{N}(z_j|0, \sigma_j^2) = 1, \quad \int_{\mathbb{R}} z_j\mathcal{N}(z_j|0, \sigma_j^2)dz_j = 0, \quad \int_{\mathbb{R}} z_j^2\mathcal{N}(z_j|0, \sigma_j^2)dz_j = \mu_j^2 + \sigma_j^2. \tag{10.17}$$

By (10.15),

$$\mathbb{E}[\mathbf{X}] = \int_{\mathbb{R}^d} (\mathbf{Uz} + \mu)\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z} = \mathbf{U}\Big(\int_{\mathbb{R}^d} \mathbf{z}\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z}\Big) + \mu\int_{\mathbb{R}^d} \mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z} = 0 + \mu = \mu \;\Rightarrow\; \text{1. rel. in (10.16).}$$

Regarding the variance, we use (8.12) and observe first that, by (10.17),

$$\int_{\mathbb{R}^d} z_k z_j\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z} = \delta_{k,j}\int_{\mathbb{R}} z_k^2\mathcal{N}(z_k|0, \sigma_k^2)dz_k \overset{(10.2)}{=} \delta_{k,j}(\sigma_k^2 + \mu_k^2) = \delta_{k,j}\sigma_k^2.$$

Moreover, $(\mathbf{Uz} + \mu)(\mathbf{Uz} + \mu)^\top = \mathbf{U}(\mathbf{zz}^\top)\mathbf{U}^\top + \mathbf{Uz}\mu^\top + \mu\mathbf{z}^\top\mathbf{U}^\top + \mu\mu^\top =: \mathbf{a}^1 + \mathbf{a}^2 + \mathbf{a}^3 + \mathbf{a}^4$. Again, by (10.17),

$$\int_{\mathbb{R}^d} \mathbf{a}^1\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z} = \mathbf{U}\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\mathbf{U}^\top = \boldsymbol{A}, \quad \int_{\mathbb{R}^d} \mathbf{a}^q\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z} = 0, \;\; q = 2, 3, \quad \int_{\mathbb{R}^d} \mathbf{a}^4\mathcal{N}(\mathbf{z}|0, \Lambda)d\mathbf{z} = \mu\mu^\top.$$

$\rightsquigarrow$ second relation in (10.16). □

The Central Limit Theorem

The Gaussian distribution plays a pivotal role in a number of contexts. We briefly mention some of them here.

The Central Limit Theorem (CLT): states that properly normalized sums of independent random variables with bounded variance tend to a normal distribution. This result comes in numerous quantified formulations. A classical variant reads as follows.

Theorem 21

Let $X_1, X_2, \dots, X_N$ be a sequence of i.i.d random variables with expectation $\mathbb{E}[X_j] = \mu$ and variance $\mathrm{var}[X_j] = \sigma^2$ for all j. Consider the sample mean

$$S_N := \frac{1}{N}(X_1 + X_2 + \cdots + X_N).$$

Then $\sqrt{N}(S_N - \mu) \overset{d}{\to} \mathcal{N}(0, \sigma^2)$. This convergence in distribution means (see (6.1))

$$\lim_{N\to\infty} \sup_{z\in\mathbb{R}} \left| \mathrm{Prob}_{\mathbb{R}^N}\left(\sqrt{N}(S_N - \mu) \le z\right) - \int_{-\infty}^{z} \mathcal{N}(x|0, \sigma^2)dx \right| = 0. \tag{10.18}$$

This says that the CDF of the sequence of random variables $\sqrt{N}(S_N - \mu)$ tend uniformly to the CDF of the Gaussian density $\mathcal{N}(\cdot|0, \sigma^2)$.

Sketch of a Proof of Theorem 21

$\{X_1, \dots, X_N\}$ i.i.d. samples of a random variable with mean μ and variance σ^2; by Lemma 7, $X_1 + \cdots + X_N$ has mean $N\mu$ and variance $N\sigma^2$. Consider the random variable

$$\frac{1}{\sigma}\sqrt{N}(S_N - \mu) =: Z_N = \frac{X_1 + \cdots + X_N - N\mu}{\sqrt{N\sigma^2}} = \sum_{j=1}^{N} \frac{X_j - \mu}{\sqrt{N\sigma^2}} =: \frac{1}{\sqrt{N}} \sum_{j=1}^{N} Y_j, \quad Y_j := \frac{X_j - \mu}{\sigma}$$

Then, since the Y_j are i.i.d.

$$\varphi_{Z_N}(s) = \varphi_{\sum_{j\le N} \frac{1}{\sqrt{N}} Y_j}(s) \overset{(8.15)}{=} \prod_{j=1}^{N} \varphi_{Y_j}(s/\sqrt{N}) = \varphi_{Y_1}(s/\sqrt{N})^N;$$

Claim:

$$\lim_{N\to\infty} \varphi_{Z_N}(s) = \lim_{N\to\infty} \varphi_{Y_1}(s/\sqrt{N})^N = e^{-\frac{s^2}{2}} \quad \text{pointwise.} \tag{10.19}$$

To that end, observe

$$\varphi_{Y_1}(0) = 1, \quad \varphi'_{Y_1}(s) = -i\int_{\mathbb{R}} yp(y)e^{-isy}dy \quad \rightsquigarrow \quad \varphi'_{Y_1}(0) = -i\,\mathbb{E}[Y_1] = 0,$$

and

$$\varphi''_{Y_1}(s) = (-i)^2 \int_{\mathbb{R}} y^2 p(y)e^{-isy}dy \quad \rightsquigarrow \quad \varphi''_{Y_1}(0) = -\int_{\mathbb{R}} y^2 p(y)dy = -\mathbb{E}[Y_1^2] = -\mathrm{var}[Y_1] = -1.$$

Taylor's Theorem then says

$$\varphi_{Y_1}(s/\sqrt{N}) = \varphi_{Y_1}(0) + \frac{s}{\sqrt{N}}\varphi'_{Y_1}(0) + \frac{s^2}{2N}\varphi''_{Y_1}(0) + o\Big((\frac{s^2}{N}\Big) = 1 - \frac{s^2}{2N} + o\Big(\frac{s^2}{N}\Big), \quad \Big(\frac{s}{\sqrt{N}}\Big) \to 0. \tag{10.20}$$

Sketch of a Proof of Theorem 21

Since $e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ one concludes from (10.20) that

$$\varphi_{Z_N}(s) = \left(1 - \frac{s^2}{2N} + o\Big(\frac{s^2}{N}\Big)\right)^N \to e^{-\frac{1}{2}s^2}, \quad N\to\infty, \quad \text{(pointwise)},$$

because all higher order terms vanish in the limit. This confirms the claim (10.19).

Recall from (10.6) that $e^{-\frac{1}{2}s^2} = \varphi_Z(s)$ when $Z \sim \mathcal{N}(0,1)$. Now we invoke Levy's Continuity Theorem. It states that pointwise convergence of characteristic functions of a sequence of probability measure implies that the limit is the characteristic function of a probability measure to which the measures then converge uniformly.

This implies that the densities of Z_N approach $\mathcal{N}(\cdot|0,1)$ in distribution. Since $Z_N = \frac{\sqrt{N}}{\sigma}(S_N-\mu)$ this yields

$$\sup_{z\in\mathbb{R}}\left|\mathrm{Prob}\Big(\frac{\sqrt{N}}{\sigma}(S_N-\mu)\le z'\Big) - \int_{-\infty}^{z'}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx\right| \to 0, \quad N\to\infty$$

Since $\mathrm{Prob}\Big(\frac{\sqrt{N}}{\sigma}(S_N-\mu)\le z'\Big) = \mathrm{Prob}\Big(\sqrt{N}(S_N-\mu)\le \sigma z'\Big)$, replacing $\sigma z'$ by z and noting that $\int_{-\infty}^{z/\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx = \int_{-\infty}^{z}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx$, (10.18) follows. □

Illustration

Bernoulli distribution: Let $p \in [0,1]$, $X \sim B(1,p)$ is a random variable taking the value 1 with probability p and the value 0 with probability $q = 1 - p$. The PDF (PMF) over $k \in \{0,1\} =: \mathcal{X}$ is

$$f_X(k;p) = \left\{ \begin{array}{lll} p & \text{if} & k = 1, \\ q = 1 - p & \text{if} & k = 0, \end{array} \right\} = p^k(1-p)^{1-k}, \quad k \in \{0,1\}. \tag{10.21}$$

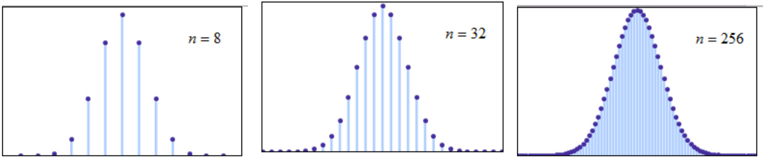
One can check (Exercise)

$$\mathbb{E}[X] = p, \quad \mathrm{var}[X] = p(1-p) = pq. \tag{10.22}$$

This is a special case of the binomial distribution $Y \sim B(N,p)$ obtained by taking the sum of i.i.d Bernoulli trials $X_j \sim B(1,p)$

$$Y = \sum_{j=1}^{N} X_j \sim B(N,p). \tag{10.23}$$

The Central Limit theorem applies. Here are a few binomial instances for various values of N, $p = 1/2$:



Parameter Estimation

A typical estimation problem: suppose we have a data set of N observations/instances $\mathbf{x} = (x_1, \ldots, x_N)^\top \in \mathbb{R}^N$ of the scalar random variable X. Suppose that these observations are drawn independently from a Gaussion $\mathcal{N}(\cdot|\mu, \sigma^2)$ whose μ and variance σ^2 are unknown.

Recall that the samples x_j are called i.d.d. - independent identically distributed. We know from (7.5) and (10.14) that the joint probability density for the random variable $\mathbf{X}$ obtained by the i.i.d. random draws X_j is the product of the marginals which are the univariate Gaussians, i.e.,

$$\mathcal{N}(\mathbf{x}|(\mu, \ldots, \mu)^\top, \Sigma) = \prod_{j=1}^{N} \mathcal{N}(x_j|\mu, \sigma^2) =, \quad \Sigma := \sigma^2 \mathbf{I} \tag{10.24}$$

which is an N-variate Gaussian.

Question: How could one estimate the values of μ and σ^2 from an instance $\mathbf{x}$ of $\mathbf{X}$?

Maximizing the Likelihood Function of Gaussians

Suppose a Gaussian process with unknown mean and variance is observed through independent samples collected in a data vector $\mathbf{x}$.

A possible way of estimating μ, σ from the data $\mathbf{x}$ is to maximize the likelihood function $L(\mu, \sigma^2; \mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \sigma^2)$ over μ, σ^2. That means one seeks parameters μ, σ^2 that are most likely for the given data (rather than fitting the specific observations best in a least squares sense).

Maximizing $\mathcal{N}(\mathbf{x}|\mu, \sigma^2)$ is most conveniently done by maximizing its logarithm

$$\log \mathcal{N}(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{j=1}^{N}(x_j - \mu)^2 - \frac{N}{2}\log(\sigma^2) - \frac{N}{2}\log(2\pi) \quad \text{(log-likelihood function)}. \tag{10.25}$$

Maximizing over μ, yields:

$$\mu_{ML} = \frac{1}{N}\sum_{j=1}^{N} x_j \quad \text{sample mean.} \tag{10.26}$$

Maximizing with respect to σ^2 yields:

$$\sigma^2_{ML} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu_{ML})^2 \quad \text{sample variance w.r.t. sample mean} \tag{10.27}$$

Note: the joint maximization of μ, σ^2 decouples in this case! verify this.

Illustration ([1, Chapter 1, § 1.2])

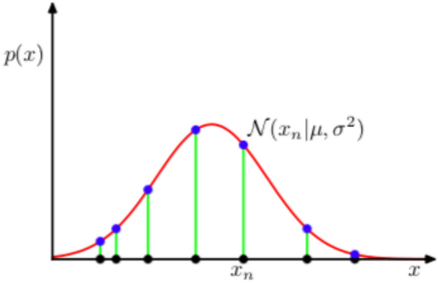


Illustration of the likelihood function for a Gaussian density (red curve); the black points denote a data set of values x_n, and the likelihood function corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.

Maximizing the Likelihood Function of Gaussians

Note: for each draw **x** one obtains estimates $\mu_{ML} = \mu_{ML}(\mathbf{x})$, $\sigma_{ML} = \sigma_{ML}(\mathbf{x})$ which will vary over repeated draws.

Exercise 22

μ_{ML}, σ_{ML} *depend on the random draws* **X** *and are therefore random variables. Hence we can compute the expectation of these quantities: show that*

$$\mathbb{E}\left[\mu_{ML}\right] = \mu, \qquad \mathbb{E}\left[\sigma_{ML}^2\right] = \left(\frac{N-1}{N}\right)\sigma^2. \tag{10.28}$$

Thus, the maximum likelihood estimate systematically underestimates the true variance by the factor $\frac{N-1}{N}$. This results from computing σ_{ML}^2 based on the sample mean not the true mean.
(10.28) $\rightsquigarrow$

$$\tilde{\sigma}_{ML}^2 := \frac{N}{N-1}\sigma_{ML}^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \mu_{ML})^2$$

is an unbiased estimator. These are special effects reflecting a more general feature of maximum likelihood methods.

The Gaussian in Information Theory

Given a (discrete) random variable X, a core question in Information Theory is: how much information is received when observing a specific value x of X. Intuitively, observing a highly unlikely event provides more information than observing a probable event (a certain event would give no additional information). Therefore any "measure of information" must be related to the underlying probability density $p = f_X$. $\rightsquigarrow$

Objective: look for a quantity $h(x)$ which is a monotonic function of $p(x)$ and expresses the "information content" of realizations of X;

Structure of $h(x)$: if X, Y are independent jointly distributed events, commonly observing both events should provide the same information as observing them separately, i.e., $h(x, y) = h(x) + h(y)$. Since by independence (see (7.5)), $p(x, y) = p(x) \cdot p(y)$, the sum has to stem from a product, i.e., h should be the logarithm of p. It is a convention to use the logarithm for base two since this relates directly to binary code length:

$$h(x) : -\log_2 p(x) \geq 0. \qquad \text{(Note: if } p(x) \text{ is small } h(x) \text{ is large).} \tag{10.29}$$

Now suppose a sender wishes to transmit values (samples) x of a random variable X to a receiver. The average amount of information transmitted in the process is its expectation w.r.t. the underlying density (convention: $p(x) \log_2 p(x) = 0$ if $p(x) = 0$)

$$H[X] := -\sum_{x\in\mathcal{X}} p(x) \log_2 p(x), \quad \text{continuous case:} \quad H[X] := -\int_{\mathcal{X}} p(x) \log_2 p(x) dx. \tag{10.30}$$

$H[X]$ is called the entropy of the random variable X.

An Example

Suppose the random variable X has 8 equally likely states, i.e. $p(x) = 1/8, x \in \{1, 2, \ldots, 8\}$. Thus, a message has 3 bits. Its entropy is

$$H[X] = -8 \cdot \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ (bits)}.$$

Now suppose the possible states x_j are unevenly distributed

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$

Then

$$H[X] = -(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{8} \log_2 \frac{1}{8} + \frac{1}{16} \log_2 \frac{1}{16} + \frac{4}{64} \log_2 \frac{1}{64}) = 2 \text{ (bits)}$$

Thus non-uniform distribution has a smaller entropy than a uniform one (entropy can be interpreted as a measure for the amount of “disorder”)

An Example

How to transmit the states? In the example 3 bits suffice to transmit a value. Can one use less by exploiting information content?

Idea: use shorter code for probable events and longer codes for less probable events; one hopes on average the shorter ones will be used more frequently in favor of overall shorter code.

Example:

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	
code	0	10	110	1110	111100	111101	111110	111111	$\leadsto$

average code length $\sum_{x\in\mathcal{X}} p(x)\mathrm{code}(x)$:

$$\frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+\frac{1}{8}\cdot 3+\frac{1}{16}\cdot 4+4\cdot\frac{1}{64}\cdot 6 = 2 \text{ (bits)}$$

i.e., he amount of entropy. (Note: one cannot use shorter code words because one needs to distinguish the states).

There is a general relation between entropy and code length. Shannon's so called Noiseless Coding Theorem (1948) states that the entropy is a lower bound for the code length = number of bits needed to transmit the state of a random variable.

Gaussians Maximize Entropy

Proposition 23

Among all continuous random variables X on $\mathbb{R}$ with mean μ and variance σ^2 the entropy is maximized by $X \sim \mathcal{N}(\mu, \sigma^2)$.

To see this consider the Lagrangian

$$L(p, \lambda_0, \lambda_1) := -\int_{\mathbb{R}} p(x)\ln(p(x))dx + \lambda_0\Big(1 - \int_{\mathbb{R}} p(x)dx\Big) + \lambda_1\Big(\sigma^2 - \int_{\mathbb{R}} p(x)(x-\mu)^2 dx\Big)$$

The Lagrange multiplier λ_0 ensures that the integral of p is one, which is necessary for a density, while λ_2 controls the variance of X. The constrained maximum of the first summand $H[X]$ is obtained by the global maximizer of L over $(p, \lambda_0, \lambda_1)$. The global extremum must be a critical point of the Lagrangian, i.e., the tripel $(p, \lambda_0, \lambda_1)$ for which the variation of L vanishes (zero of the derivative). More precisely, find $(p, \lambda_0, \lambda_1)$ such that

$$\lim_{t\to 0} \frac{1}{t}\left(L(p + t\delta f, \lambda_0, \lambda_1) - L(p, \lambda_0, \lambda_1)\right) = 0 \quad \forall\ \delta f$$

Straightforward computation $\rightsquigarrow \int_{\mathbb{R}} \delta f(x)(\ln p(x) + 1 + \lambda_0 + \lambda_1(x-\mu)^2)dx = 0$ for all $\delta f \Rightarrow$

$$p(x) = e^{-(1+\lambda_0+\lambda_1(x-\mu)^2)}.$$

Solving the constraint equations $\int_{\mathbb{R}} e^{-(1+\lambda_0+\lambda_1(x-\mu)^2)}dx = \int_{\mathbb{R}} e^{-(1+\lambda_0)}e^{-\lambda_1(x-\mu)^2}dx = 1$,

$\int_{\mathbb{R}} e^{-(1+\lambda_0)}e^{-\lambda_1(x-\mu)^2}(x-\mu)^2dx = \sigma^2$ yields $\lambda_1 = 1/(2\sigma^2)$, $e^{(1+\lambda_0)} = \sigma\sqrt{2\pi}$, which shows that

$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \mathcal{N}(x|\mu, \sigma^2).$ □

Preliminary Remarks

Model scenario: suppose we have measurements/observations $(t_1, y_1), \dots, (t_m, y_m)$ taken to understand an underlying physical law. For instance, the y_j could be fuel consumption rates of an engine under certain driving conditions t_j. The searched for law could be described as a function of t taking values y. Finding that function would allow one to predict fuel consumption under any other driving conditions t. This suggests an ansatz

$$y(t) = y(t; x_1, \dots, x_n), \tag{11.1}$$

where $x_1, \dots, x_n$ are (unknown) parameters that are to parametrize (a sufficiently good approximation of) the law. In principle, $y(\cdot; x_1, \dots, x_n)$ could be a non-linear function of the parameters x_j. Deep Neural Networks are currently very prominent examples of highly-nonlinear such parametrizations. In many applications a simpler linear ansatz

$$y(t) = \sum_{j=1}^{n} x_j \phi_j(t), \tag{11.2}$$

suffices though and should be understood first. Here the $\phi_j(t)$ are suitably chosen (background information) ansatz-functions such as polynomials, trigonometric functions, splines, wavelets, radial basis functions, etc.

A Simple Regression Model

In this section we adopt the following model assumptions:

1. The values t_j (which could by imposed by technological constraints, like boreholes for geophysical exploration) are considered here as deterministic quantities.
2. Measurements are never accurate. Therefore, the y_j are considered as instances/samples of some random variables Y_j. Data uncertainty and model-mismatch are then treated by a noise model.
3. Statistical inference with reasonable confidence therefore requires (substantially) more measurements y_i than unknown parameters x_j, i.e., $m \geq n$.

For a linear ansatz (11.2) this leads to the following Linear Regression Model:

$$Y_i = \sum_{k=1}^{n} a_{i,k} x_k + F_i, \quad i = 1, \ldots, m, \quad (a_{i,k} := \phi_k(t_i), \quad i = 1, \ldots, m,\ k = 1, \ldots, n). \qquad (11.3)$$

In brief: $\mathbf{Y} = \boldsymbol{A}\mathbf{x} + \mathbf{F}$. Thus, $\mathbf{Y} = (Y_i, \ldots, Y_m)^\top$ is a vector of random variables whose realizations form the vector $\mathbf{y} = (y_1, \ldots, y_m)^\top$. The vector $\mathbf{F} = (F_1, \ldots, F_m)^\top$ is another random vector representing model- or measurement errors. The x_j are unknown and unobservable parameters “explaining” the measurements.

Goal: find an estimate $\hat{\mathbf{x}}$ for the “true” parameters $\mathbf{x} = (x_1, \ldots, x_n)^\top$ clouded by the noise.

Under such assumptions the Least Squares Estimator turns out to be the best one can do:

$$\hat{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^n} \|\boldsymbol{A}\mathbf{z} - \mathbf{Y}\|_2, \quad \boldsymbol{A} := (a_{i,k})_{i,k=1}^{m,n} \in \mathbb{R}^{m \times n}, \quad \mathbf{Y} = \boldsymbol{A}\mathbf{x} + \mathbf{F}. \qquad (11.4)$$

Linear Estimators

- We know from Lecture I, Theorem 45, (6.18) that the solution of (11.4) is characterized by the normal equations

$$\|\boldsymbol{A}\hat{\mathbf{X}} - \mathbf{Y}\|_2 = \min \quad \Leftrightarrow \quad \boldsymbol{A}^\top \boldsymbol{A}\hat{\mathbf{X}} = \boldsymbol{A}^\top \mathbf{Y} \quad \Leftrightarrow \quad \hat{\mathbf{X}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{Y}, \tag{11.5}$$

  and thus the result of a linear operator.
- Therefore, $\hat{\mathbf{X}}$ is called a linear estimator. Since $\mathbf{Y}$ is a random variable, $\hat{\mathbf{X}}$ is also a random variable. For each realization $\mathbf{y}$ of $\mathbf{Y}$, the realization $\hat{\mathbf{x}}$ of $\hat{\mathbf{X}}$ is the solution of the (deterministic) least squares problem $\min_{\mathbf{x}\in\mathbb{R}^n} \|\boldsymbol{A}\mathbf{x} - \mathbf{y}\|_2$.
- We call

$$\tilde{\mathbf{X}} = \mathbf{C}\mathbf{Y} = \sum_{k=1}^{m} \mathbf{c}^k Y_k \quad \text{for some } \mathbf{C} \in \mathbb{R}^{n\times m}, \quad (\mathbf{c}^k \text{ the columns of } \mathbf{C}) \tag{11.6}$$

  a linear estimator where the matrix $\mathbf{C}$ is independent of (the non-observable) $\mathbf{x}$ but may depend only on the observable $\boldsymbol{A}$. The following theorem says that the least squares estimator is in some sense optimal among all linear estimators.

The Best linear Unbiased Estimator

Model: $\mathbf{Y} = \boldsymbol{A}\mathbf{x} + \mathbf{F}$, $\mathbf{F}$ noise vector, $\mathbb{E}[\mathbf{F}] = 0 \Rightarrow \mathbb{E}[\mathbf{Y}] = \boldsymbol{A}\mathbf{x}$

Theorem 24

Assume that $\mathbf{F} = (F_1, \ldots, F_m)^\top$ *is a vector of random variables* F_i *with zero mean* $\mathbb{E}[F_i] = 0$, $i = 1, \ldots, m$ *(no systematic error), uncorrelated, i.e.,* $\mathbb{E}[F_i F_j] = 0$, $i \neq j$, *with equal variance* $\mathrm{var}[F_i] = \sigma^2$, *i.e., the covariance matrix has the form*

$$\mathrm{var}[\mathbf{F}] = \mathbb{E}[\mathbf{F}\mathbf{F}^\top] = \left(\mathbb{E}[F_i F_j]\right)_{i,j=1}^{m} = \sigma^2 \mathbf{I}. \tag{11.7}$$

Then, if $\boldsymbol{A}$ *has full rank,*

$$\hat{\mathbf{X}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{Y} \tag{11.8}$$

is the Best Linear Unbiased Estimator (BLUE), i.e.,

1. *it is unbiased which means* $\mathbb{E}[\hat{\mathbf{X}}] = \mathbf{x}$, $\mathrm{var}[\hat{\mathbf{X}}] = \sigma^2 (\boldsymbol{A}^\top \boldsymbol{A})^{-1}$;
2. *it minimizes the variance among all linear unbiased estimators, i.e., for any* $\tilde{\mathbf{X}}$ *of the form* (11.6) *one has*

$$\mathrm{var}[\tilde{\mathbf{X}}] - \mathrm{var}[\hat{\mathbf{X}}] \quad \textit{is positive semi-definite.} \tag{11.9}$$

Proof of Theorem 24: We'll use several times that for any verctor of random variables $\mathbf{X}$ and deterministic matrix $\mathbf{B}$ one has $\mathbb{E}[\mathbf{BX}] = \mathbf{B}\mathbb{E}[\mathbf{X}]$, see (8.8).

- The estimator is linear because $\mathbf{Y} \to \hat{\mathbf{X}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{Y}$ is a linear mapping.
- Expectation: $\mathbb{E}[\hat{\mathbf{X}}] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{Y} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} = \mathbf{x}$ because $\mathbb{E}[\mathbf{F}] = \mathbf{0}$.
- Variance: since

$$\hat{\mathbf{X}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{Y}, \quad (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top(\mathbf{Y} - \mathbf{F}) = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} = \mathbf{x},$$

one obtains $\hat{\mathbf{X}} - \mathbf{x} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{Y} - (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top(\mathbf{Y} - \mathbf{F}) = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{F} \rightsquigarrow$

$$(\hat{\mathbf{X}} - \mathbf{x})(\hat{\mathbf{X}} - \mathbf{x})^\top = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{F}\mathbf{F}^\top \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1},$$

and thus ($\mathbb{E}$ is linear, (8.8))

$$\mathbb{E}\left[(\hat{\mathbf{X}} - \mathbf{x})(\hat{\mathbf{X}} - \mathbf{x})^\top\right] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbb{E}\left[\mathbf{F}\mathbf{F}^\top\right]\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} = \sigma^2(\boldsymbol{A}^\top \boldsymbol{A})^{-1}.$$

Regarding (11.9), suppose $\tilde{\mathbf{X}} = \mathbf{CY}$ ($\mathbf{C} \in \mathbb{R}^{n\times m}$) is an arbitrary linear estimator with $\mathbb{E}[\tilde{\mathbf{X}}] = \mathbf{x}$. Then

$$\mathbf{x} = \mathbb{E}[\mathbf{CY}] = \underbrace{\mathbb{E}[\mathbf{C}\boldsymbol{A}\mathbf{x}]}_{=\mathbf{C}\boldsymbol{A}\mathbf{x}} + \underbrace{\mathbb{E}[\mathbf{CF}]}_{=\mathbf{C}\mathbb{E}[\mathbf{F}]=\mathbf{0}} = \mathbf{C}\boldsymbol{A}\mathbf{x}.$$

Since this has to hold for any $\mathbf{x}$ one concludes $\mathbf{C}\boldsymbol{A} = \mathbf{I} \in \mathbb{R}^{n\times n}$. The particular choice

$$\mathbf{C}^\top := \mathbf{B}^\top = \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \tag{11.10}$$

satisfies $\mathbf{C}\boldsymbol{A} = \mathbf{I}$. We show next that this is the only possible choice that minimizes the variance.

Proof of Theorem 24 continued: The condition $\mathbf{C}\boldsymbol{A} = \mathbf{I} = \boldsymbol{A}^\top\mathbf{C}^\top$ means that the columns $\bar{\mathbf{c}}^j$ of $\mathbf{C}^\top$ (rows of $\mathbf{C}$) solve the linear systems

$$\boldsymbol{A}^\top\bar{\mathbf{c}}^j = \mathbf{e}^j, \quad j = 1, \dots, n. \tag{11.11}$$

Since $m \geq n$ this system of linear equations is in general underdetermined in which case the solutions are not unique. Let $\mathbf{b}^j = \mathbf{B}^\top\mathbf{e}^j = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\mathbf{e}^j$ denote the jth column of the above particular choice $\mathbf{B}$ from (11.10). Then, any solution $\bar{\mathbf{c}}^j$ has the form (see Lecture I, Exercise 9, page 16)

$$\bar{\mathbf{c}}^j = \mathbf{b}^j + \mathbf{w}^j, \quad \mathbf{w}^j \in \ker(\boldsymbol{A}^\top).$$

Thus, every solution $\mathbf{C}$ that solves $\boldsymbol{A}^\top\mathbf{C}^\top = \mathbf{I}$ is of the form $\mathbf{C}^\top = \mathbf{B}^\top + \mathbf{W} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1} + \mathbf{W}$ with $\boldsymbol{A}^\top\mathbf{W} = \mathbf{0}$. Now, one verifies (as above)

$$\begin{aligned} \mathrm{var}[\tilde{\mathbf{X}}] &= \mathbb{E}\big[(\tilde{\mathbf{X}} - \mathbf{x})(\tilde{\mathbf{X}} - \mathbf{x})^\top\big] = \sigma^2\mathbf{C}\mathbf{C}^\top = \sigma^2\big((\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top + \mathbf{W}^\top\big)\big(\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1} + \mathbf{W}\big) \\ &= \sigma^2(\boldsymbol{A}^\top\boldsymbol{A})^{-1} + \sigma^2\mathbf{W}^\top\mathbf{W}. \end{aligned}$$

So, whenever $\mathbf{W} \neq \mathbf{0}$, the variance of $\tilde{\mathbf{X}}$ is larger than that of $\hat{\mathbf{X}}$, which finishes the proof. □

Maximum Likelihood Estimator

There is a further important property of the least squares estimator $\hat{\mathbf{X}}$ from (11.8) which holds when the noise model $\mathbf{F}$ satisfies in addition to the above assumptions from Theorem 24

$$F_i \sim \mathcal{N}(0;\sigma^2), \quad f_i(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2}, \quad i = 1,\dots,m, \tag{11.12}$$

i.e., the error components centered Gaussians with equal variance σ^2. By (10.13) and (10.14), the joint density of $\mathbf{Y} = \boldsymbol{A}\mathbf{x} + \mathbf{F}$ is then (because $\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] + \mathbb{E}[\mathbf{F}] = \boldsymbol{A}\mathbf{x}$)

$$\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{m}{2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\boldsymbol{A}\mathbf{x}\|_2^2} = \mathcal{N}(\mathbf{y}|\boldsymbol{A}\mathbf{x}, \sigma^2\mathbf{I}) =: L(\mathbf{x};\mathbf{y})$$

which is called the Likelihood Function. Given the observations $\mathbf{y} \in \mathbb{R}^m$ the vector $\tilde{\mathbf{x}}$ that best explains the observations - “most likely” value - is the one that maximizes the density and hence the Likelihood Function. An estimator $\tilde{\mathbf{X}}$ is called maximum likelihood estimator if

$$L(\tilde{\mathbf{x}};\mathbf{y}) \geq L(\mathbf{x}';\mathbf{y}) \quad \forall\, \mathbf{x}' \in \mathbb{R}^n. \tag{11.13}$$

Clearly, $L(\tilde{\mathbf{x}};\mathbf{y})$ is maximal when $\|\boldsymbol{A}\tilde{\mathbf{x}} - \mathbf{y}\|_2$ is minimal. Therefore:

Proposition 25

Under the above assumptions on $\mathbf{F}$ *the least squares estimator* $\hat{\mathbf{X}}$ *is also the maximum likelihood estimator.*

Comments

The various optimality properties of the least squares estimator should not be misread regarding the quality of the resulting estimation:

1. $\mathrm{var}[\hat{\mathbf{X}}]$ depends not only on **F** (through the value of σ^2) but also on the matrix $\boldsymbol{A}$. First, $\boldsymbol{A}$ is required to have full rank. But even if this is the case, the smallest singular value σ_n of $\boldsymbol{A}$ could be very close to zero which means that $\sigma_n^{-2} = \|(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\|_2$ is very large, and so is the variance. In other words, the estiamted values may fluctuate very much around the expected components in $\mathbf{x}$.
2. Recall the initial example where $\boldsymbol{A} = (\phi_k(t_i))_{i,k=1}^{m,n} \in \mathbb{R}^{m\times n}$ and the $\phi_k(t)$ were some basis functions such as polynomials or splines. While the linear combinations $\sum_{k=1}^n x_k\phi_k(t)$ cannot vanish for all t, unless the coefficients x_k are all zero, since they form a basis, it may very well happen that such a linear combination can vanish at finitely many points. In this case the matrix $\boldsymbol{A}$ does not have full rank. For instance, a polynomial of degree n can have n zeros.
3. If, on the other hand, we use more and more measurements, i.e., sample the basis functions more and more densely, it becomes harder and harder for the linear combination $\sum_{k=1}^n x_k\phi_k(t)$ to attain values close to zero simultaneously at all points t_i. This is expected to increase the smallest singular value and therefore decreases the variance of the estimator. In the limit the smallest singular value of $\boldsymbol{A}$ depends on the condition of the basis, see Lecture I, page 31.

   Upshot: the relation between the number of measurements m and the number n of parameters to be estimated plays a critical role. This issue will be addressed in more depth later in the context of learning algorithms for regression.

Comments

One would, of course, like to understand the behavior of $\|\hat{\mathbf{X}} - \mathbf{x}\|_2$ when m and n vary. Since $\hat{\mathbf{X}}$ is a random variable the quantity $g(\hat{\mathbf{X}}) := \|\hat{\mathbf{X}} - \mathbf{x}\|_2$ is also a random variable, so a meaningful question would be to bound quantities like

$$\mathbb{E}\left[\|\hat{\mathbf{X}} - \mathbf{x}\|_2^2\right] \quad \text{or} \quad \mathrm{Prob}\left(\|\hat{\mathbf{X}} - \mathbf{x}\|_2^2 \geq \eta\right), \tag{11.14}$$

depending on m, n, where expectation and probabilities refer to the distribution underlying $\mathbf{F}$. Note that, by Exercise 15, (1), the second quantity gives sharper information.

Remark 26

- $\mathbb{E}\left[\|\hat{\mathbf{X}} - \mathbf{x}\|_2^2\right] = \mathbb{E}\left[(\hat{\mathbf{X}} - \mathbf{x})^\top(\hat{\mathbf{X}} - \mathbf{x})\right]$ *is different from* $\mathrm{var}[\hat{\mathbf{X}}] = \mathbb{E}\left[(\hat{\mathbf{X}} - \mathbf{x})(\hat{\mathbf{X}} - \mathbf{x})^\top\right]$.
- *Verify:* $\mathbb{E}\left[\|\hat{\mathbf{X}} - \mathbf{x}\|_2^2\right] = \sum_{j=1}^n \mathrm{var}[\hat{X}_j]$.
- *Markov's inequality (Theorem 12) gives a first simple estimate for the second quantity in* (11.14)*:* $\mathrm{Prob}\left(g(\hat{\mathbf{X}}) \geq \eta\right) \leq \mathbb{E}[g(\hat{\mathbf{X}})]/\eta$. *However, that requires knowing* $\mathbb{E}[g(\hat{\mathbf{X}})]$.
- *An idea is to estimate* $\mathbb{E}[g(\hat{\mathbf{X}})]$ *from the empirical mean* $\frac{1}{N}\sum_{j=1}^N g(\hat{\mathbf{x}}^j)$, *where the* $\hat{\mathbf{x}}^j$ *are the least squares solutions with respect to independently drawn* $\mathbf{y}^j$ *of* $\mathbf{Y}$.

Central Issue: (addressed in the next lecture) derive refined tail bounds for (large) sums of random variables on deviations from corresponding expectations.

Comments on Numerical Aspects

While understanding the effect of noise on estimation is a core issue in this lecture, one should be aware that the numerical computation itself, needed to compute the least squares solutions, is subject to numerical errors as well. The theoretical considerations, in particular (11.5), suggest to simply solve the normal equations $\boldsymbol{A}^\top \boldsymbol{A}\hat{\mathbf{x}} = \boldsymbol{A}^\top \mathbf{y}$. There is, however, a wrinkle about this option:

1. $\boldsymbol{A}$ may have a large condition number $\kappa_2(\boldsymbol{A}) = \sigma_1/\sigma_n$;
2. Using SVD it is easy to show that $\kappa_2(\boldsymbol{A}^\top \boldsymbol{A}) = \kappa_2(\boldsymbol{A})^2$ (Exercise), i.e., the actual numerical system has a possibly much larger condition number σ_1^2/σ_n^2, causing significant loss of accuracy in the computations.

Alternate Strategy:

- Calculate the *QR* factorization $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ (see Lecture I, Theorem 37) and compute $\tilde{\mathbf{y}} := \boldsymbol{Q}^\top \mathbf{y}$. We write $\tilde{\mathbf{y}}$ as $((\tilde{\mathbf{y}}^1)^\top, (\tilde{\mathbf{y}}^2)^\top)^\top$ with $\tilde{\mathbf{y}}^1 \in \mathbb{R}^n$, $\tilde{\mathbf{y}}^2 \in \mathbb{R}^{m-n}$. Note, that $\boldsymbol{R}$ has block form $\boldsymbol{R} = \binom{\tilde{\boldsymbol{R}}}{\mathbf{0}}$ where $\tilde{\boldsymbol{R}} \in \mathbb{R}^{n\times n}$ is upper triangular.
- Note: (using Lecture I, Proposition 15, (3))

$$\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 = \|\boldsymbol{Q}\boldsymbol{R}\hat{\mathbf{x}} - \boldsymbol{Q}\boldsymbol{Q}^\top \mathbf{y}\|_2^2 = \|\boldsymbol{Q}(\boldsymbol{R}\hat{\mathbf{x}} - \tilde{\mathbf{y}})\|_2^2 = \|\boldsymbol{R}\hat{\mathbf{x}} - \tilde{\mathbf{y}}\|_2^2 = \|\tilde{\boldsymbol{R}}\hat{\mathbf{x}} - \tilde{\mathbf{y}}^1\|_2^2 + \|\tilde{\mathbf{y}}^2\|_2^2.$$

  This is minimized if and only if $\hat{\mathbf{x}} = \tilde{\boldsymbol{R}}^{-1}\tilde{\mathbf{y}}^1$ which requires solving an $(n \times n)$ upper triangular system. Note! by Lecture I, Proposition 15, (4), one has $\kappa_2(\tilde{\boldsymbol{R}}) = \kappa_2(\boldsymbol{A})$, so squaring of the condition is avoided!

References I

[1] Cristopher M. Bishop, Pattern Recognition and Machine Learning, Springer Science+Business Media, LLC, 2006.

[2] Avrim Blum, John Hopcroft, and Ravindran Kannan, Foundations of Data Science, 2015, https://www.cs.cornell.edu/jeh/book.pdf