PROBABILISTIC MODELS – PART 2: FUNDAMENTAL CONCEPTS OF PROBABILITY

Gerhard Widmer

Institute of Computational Perception
Johannes Kepler University
Linz, Austria

gerhard.widmer@jku.at
www.cp.jku.at/people/widmer

October 2, 2025

Presentation partly based on and inspired by [Koller & Friedman, 2009] and [Russell & Norvig, 2021], including the use of some figures from their books and/or lecture slides.

Many thanks to Daphne Koller, Nir Friedman, Stuart Russell, and Peter Norvig for making these available (`pgm.stanford.edu`; `aima.cs.berkeley.edu`).

Goals of this Lecture

- Establish terminology and mathematical notation for this class
- Remind you of fundamental concepts of probability
- Introduce the probabilistic query
- Show that the answer to any probabilistic query can be computed from the full joint distribution via *Inference by Enumeration*
- Introduce the central concept for this entire class: *Independencies*
- Motivate you to internalise these concepts
  (to be able to follow the rest of the class)

Outline

Basics of Probability: Random Variables

Definition

A **Random Variable** X is a variable with a fixed (finite or infinite) **domain** (i.e., set of possible values) $Val(X)$, which represents some aspect of the system's world (or of the system's internal state).

Different types:

- **Boolean variables** take values $\in \{false, true\}$ (or $\{0, 1\}$)
- **Discrete (categorical) variables** have a finite domain of symbolic values. Example: $Season$ with $Val(Season) = \{spring, summer, fall, winter\}$
- **Continuous (real-valued) variables** take a numeric value $\in \mathbb{R}$ Most real-world sensors (e.g,. temperature, sonar, radar, ...) give real-valued readings.

☞ **Basic atomic building blocks of our models and world representation.**

A Simple Example

Consider a system for medical diagnosis
in a very simple world where there are only

- two diseases (not directly observable): influenza (flu) and hayfever;
- two symptoms (possibly observable): congestion and muscle pain; and
- four seasons (observable): spring, summer, fall, and winter.

and where at a given moment, there is exactly one patient to be diagnosed.

World model using 5 random variables:

Flu	$\in \{true, false\}$
$Hayfever$	$\in \{true, false\}$
$Congestion$	$\in \{true, false\}$
$MusclePain$	$\in \{true, false\}$
$Season$	$\in \{spring, summer, fall, winter\}$

Note the use of boolean variables for diseases and symptoms – to allow for the case where a patient may have both symptoms and/or both diseases!

Basics of Probability: Events

Definition

An **Event** is a fixed assignment of values to some or all of the variables in the system's world.

Definition

An **Atomic Event** is an event where **all** random variables in the system's world have a specific value assigned.

Notes:

- An *atomic event* corresponds to one particular possible state of the world
- An *event* corresponds to a *set* of possible states of the world (it comprises all those states that share the specific values fixed in the event)
- Events (possible value assignments to variables) are what the system will try to compute probabilities for.

A Simple Example

Simple medical diagnosis world:

Flu	$\in \{true, false\}$
$Hayfever$	$\in \{true, false\}$
$Congestion$	$\in \{true, false\}$
$MusclePain$	$\in \{true, false\}$
$Season$	$\in \{spring, summer, fall, winter\}$

Some events:

- $Season = spring \wedge MusclePain = true$
  denotes all situations where it is spring and the patient has a muscle pain (regardless of whether or not she has the flu, hayfever, or congestion)
- $Congestion = false$
- $Season = fall \wedge MusclePain = true \wedge Congestion = false \wedge Hayfever = false \wedge Flu = false$ (an atomic event)

☞ **Total number of possible *atomic* events:** $2 \times 2 \times 2 \times 2 \times 4 = 64$

Some Notational Conventions (1)

NOTATION

Examples	Type of object
$\mathcal{X}$	Complete set of variables that a model is defined over
A, X_i, E_j	Single random variable (e.g., $X_i \in \mathcal{X}$)
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{U}, \ldots$	Sets of random variables (e.g., $\boldsymbol{X} = \{X_1, X_2, X_3, X_4\}$)
$Val(X)$	Domain of a variable X (set of values that X can take)
$x^0, x^1, \ldots$	$x^i = i^{th}$ value in the domain $Val(X)$ of variable X
x^0, x^1	Special convention for boolean variables: $x^0 = false, x^1 = true$
$\boldsymbol{x}_i, \mathbf{e}, \ldots$	Set/list of specific values of a variable (e.g., $\boldsymbol{E} = \boldsymbol{e}$)
$Val(\boldsymbol{X})$	Set of possible value assignments to the variables in $\boldsymbol{X}$ (= Cartesian product $Val(X_1) \times Val(X_2) \times \ldots$)
$P(X = x^i)$	Probability (probability mass function)
$P(x^i), P(x^i, y^j)$	Shorthand notation for $P(X = x^i), P(X = x^i \wedge Y = y^j)$
$P(X, Y)$	Probability distribution over discrete variables X, Y

Probabilities and Distributions

Definition

Given: A set $\mathcal{S}$ of (mutually exclusive) atomic events.

A **Probability Distribution** over $\mathcal{S}$ is a function $P : \mathcal{S} \mapsto \mathbb{R}$ that satisfies the following:

1. $P(\alpha) \geq 0$ for all $\alpha \in \mathcal{S}$
2. $P(\Omega) = 1$, where $\Omega = \bigcup_{\alpha_i \in \mathcal{S}} \alpha_i$ (the disjunction of all possible events)[a]
3. If $\alpha, \beta \in \mathcal{S}$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.

A **Probability** $P(\alpha)$ is the value that the probability distribution P assigns to the specific event α.

[a] We will use $\cap$ and $\cup$ to denote logical AND and OR, respectively, between events

☞ **Look at the argument of an expression** $P(\cdot)$ to decide whether the P denotes a *probability* (a single value) or a *distribution* (a list of values):

- If the argument is an event, it is a probability.
- If the argument is a (set of) random variable(s), it is a distribution.

A Simple Example

Even simpler medical diagnosis world – 2 random variables only:

$$\begin{array}{ll} Flu & \in \{false, true\} \\ MusclePain & \in \{false, true\} \end{array}$$

The set $\mathcal{S}$ of possible atomic events:

$$\begin{array}{ll} \alpha_1 = (\ \ flu \wedge musclepain) & \alpha_2 = (\ \ flu \wedge \neg musclepain) \\ \alpha_3 = (\neg flu \wedge musclepain) & \alpha_4 = (\neg flu \wedge \neg musclepain) \end{array}$$

A possible probability distribution over these 4 atomic events:

$P(Flu, MP)$	$MP = false$	$MP = true$
$Flu = false$	0.7	0.15
$Flu = true$	0.05	0.1

(That is, the probability that a randomly chosen person suffers from a muscle pain *and* has the flu is 0.1, etc.)

Interpretations of Probability

Defining what probabilities *"mean"* is a deep philosophical problem (with a long history ...)

Intuitively:

"The probability $P(\alpha)$ of an event α quantifies the degree of confidence that α will occur."

- $P(\alpha) = 1$: we are certain that α will occur
- $P(\alpha) = 0$: we are certain that α will *not* occur
- But what does $P(\alpha) = x$ mean, if $0 < x < 1$?

Interpretations of Probability

Two main philosophical stances:

'Frequentist interpretation':

The probability of an event is the **proportion of times** that the event α would occur if we repeated the experiment an *infinite number of times*.

Problem: How does this apply to, for example, *"the probability of rain this afternoon is 0.7"*?

'Subjectivist interpretation':

The probability of an event expresses a **subjective degree of belief** that the event α will happen (or the degree to which α is supported by available evidence).

In this class:

- Will adopt the subjectivist (Bayesian) interpretation
- $P(x)$ represents the **system's degree of belief** that x is true in the world.
- (Of course, will have to resort to frequentist view when we want to learn probabilities from observations ...)

Joint and Marginal Distributions

Definition

In a world defined by a set of random variables $\mathcal{X}$, the probability distribution over *all atomic events* possible over $\mathcal{X}$ is called **Full Joint Distribution** over $\mathcal{X}$.
It assigns probabilities to all possible value combinations of the $\mathcal{X}$.

Definition

A probability distribution defined over the events induced by a subset $\boldsymbol{X} \subset \mathcal{X}$ of variables is called **Marginal Distribution** over $\boldsymbol{X}$.
It assigns probabilities to all value combinations of the selected $\boldsymbol{X}$.

Definition

Special Case:
A probability distribution defined over the values of a *single* variable $X \in \mathcal{X}$ is called the **Marginal Distribution** of variable X.

Some More Notational Conventions

NOTATION

- **Probability of conjunction of events:** Write

$$\boxed{P(X = x, Y = y)} \text{ or } \boxed{P(x, y)} \text{ instead of } P((X = x) \cap (Y = y))$$

- **Joint distribution over sets of variables:** Write

$$\boxed{P(\boldsymbol{X}) = P(X_1, X_2, \ldots, X_k)}$$

to denote joint distribution over variables $\boldsymbol{X} = \{X_1, X_2, \ldots X_k\}$

- **Joint distribution over several sets of variables:** Write

$$\boxed{P(\boldsymbol{X}, \boldsymbol{Y}) = P(X_1, \ldots X_k, Y_1, \ldots, Y_l)}$$

to denote joint distribution over variable set $\boldsymbol{X} \cup \boldsymbol{Y}$

- **Conditional distributions:** Write

$$\boxed{P(\boldsymbol{X} \mid \boldsymbol{Y}) = P(X_1, X_2, \ldots, X_k \mid Y_1, Y_2, \ldots, Y_l)}$$

to denote the joint distributions over $\boldsymbol{X}$, conditioned on values of $\boldsymbol{Y}$.

Joint and Marginal Distributions

Example: Diagnosis world with 3 (Boolean) random variables

$$\mathcal{X} = \{Flu, MusclePain(MP), Fever\}$$

Full Joint Distribution:

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Note:

- Sum over all entries is 1.0 (a proper distribution)

Joint and Marginal Distributions

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Marginal distributions

can be computed by summing over the full joint distribution:

$$\begin{aligned} P(flu) &= P(flu, mp, fever) + P(flu, mp, \neg fever) + \\ &\quad P(flu, \neg mp, fever) + P(flu, \neg mp, \neg fever) \\ &= \sum_{x,y} P(flu, MP = x, Fever = y) \\ &= 0.15 + 0.1 + 0.1 + 0.03 = \underline{0.38} \\ P(\neg flu) &= 0.03 + 0.04 + 0.05 + 0.5 = \underline{0.62} \end{aligned}$$

Marginal distribution $P(Flu) = \{0.62, 0.38\}$

Joint and Marginal Distributions

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Likewise:

$$\begin{aligned} P(\neg musclepain) &= 0.1 + 0.03 + 0.05 + 0.5 = \underline{0.68} \\ P(musclepain) &= 0.15 + 0.1 + 0.03 + 0.04 = \underline{0.32} \end{aligned}$$

Marginal distribution $P(MusclePain) = \{0.68, 0.32\}$

$$\begin{aligned} P(\neg fever) &= 0.1 + 0.04 + 0.03 + 0.5 = \underline{0.67} \\ P(fever) &= 0.15 + 0.03 + 0.1 + 0.05 = \underline{0.33} \end{aligned}$$

Marginal distribution $P(Fever) = \{0.67, 0.33\}$

Joint and Marginal Distributions

In Words:

A **marginal probability** $P(\boldsymbol{X} = \boldsymbol{x})$ is computed by summing over all entries in the full joint distribution that have $\boldsymbol{X} = \boldsymbol{x}$:
We are computing the probability of the disjunction ($\cup$) of all atomic events that have $\boldsymbol{X} = \boldsymbol{x}$. (Atomic events are mutually exclusive $\Rightarrow$ sum).

Generally:

The probability of any event α is the sum over all atomic events (= entries in the full joint distribution) that "match" α (i.e., have the same values for those variables whose values are fixed in α).

Leads to a general, simple algorithm for *probabilistic inference*:
☞ **Inference by Enumeration** (see below)

Conditional Probabilities

Conditional probabilities

- will be the main way of modelling questions about a given world:

"Given that I have observed some event β, what is the probability that some other (unobservable) event α of interest to me is true in the world?"

Examples:

- $P(Hayfever = true \mid Congestion = true) = ?$

  Given that my patient suffers from a sinus congestion, what is the probability that she has a hay fever?
- $P(Hayfever = true \mid Congestion = true, MP = true) = ?$

  If I learn, in addition, that she has muscle pain, how does that change my degree of belief in hay fever?
- $P(Location, Velocity \mid GPS =< ., ., . >, Radar =< ., ., . >) = ?$

  Given the readings of a robot's GPS and radar sensors, what is the probability distribution over the possible current locations and velocities of the robot?

Conditional Probabilities

Definition

For two events α, β, the **Conditional Probability** of α, given that we know that β is true, is

$$P(\alpha \mid \beta) = \frac{P(\alpha \cap \beta)}{P(\beta)}$$

Note:

- For a fixed β and some variables $\boldsymbol{X}$, the **conditional distribution** $P(\boldsymbol{X} \mid \beta)$ satisfies all the properties of a proper probability distribution (specifically: sums to 1.0 over all value combinations of $\boldsymbol{X}$)
- $P(\boldsymbol{X} \mid \beta)$ is a legitimate probability distribution in its own right
- ***Conditioning*** is an operation that takes one distribution $P(\boldsymbol{X})$ and returns another distribution $P(\boldsymbol{X} \mid \beta)$.

The Chain Rule

Consequences of the above definition:

$$P(\alpha \mid \beta) = \frac{P(\alpha \cap \beta)}{P(\beta)}$$

implies:

$$\boxed{P(\alpha \cap \beta) = P(\alpha \mid \beta)P(\beta)}$$

and

$$\boxed{P(\alpha \cap \beta) = P(\beta \cap \alpha) = P(\beta \mid \alpha)P(\alpha) = P(\alpha)P(\beta \mid \alpha)}$$

Definition

The generalisation of the above to n events is known as the **Chain Rule:**

$$\begin{aligned} P(\alpha_1, ..., \alpha_n) &= P(\alpha_1 \mid \alpha_2, ..., \alpha_n)P(\alpha_2, ..., \alpha_n) \\ &= P(\alpha_1 \mid \alpha_2, ..., \alpha_n)P(\alpha_2 \mid \alpha_3, ..., \alpha_n)P(\alpha_3, ..., \alpha_n) \\ &\dots \\ &= P(\alpha_1 \mid \alpha_2, ..., \alpha_n)P(\alpha_2 \mid \alpha_3, ..., \alpha_n) \cdots P(\alpha_{n-1} \mid \alpha_n)P(\alpha_n) \end{aligned}$$

The Chain Rule

Chain Rule

$$P(\alpha_1, ..., \alpha_n) = P(\alpha_1 \mid \alpha_2, ..., \alpha_n)P(\alpha_2 \mid \alpha_3, ..., \alpha_n) \cdots P(\alpha_{n-1} \mid \alpha_n)P(\alpha_n)$$

Notes:

- Because conjunction is commutative and associative, this expression can be expanded using **any order of events**, for example:

$$P(\alpha_1, ..., \alpha_n) = P(\alpha_1)P(\alpha_2 \mid \alpha_1) \cdots P(\alpha_n \mid \alpha_1, ..., \alpha_{n-1})$$

- ... and also **partially**, e.g.,

$$P(\alpha_1, ..., \alpha_n) = P(\alpha_1, ..., \alpha_k)P(\alpha_{k+1}, ..., \alpha_n \mid \alpha_1, ..., \alpha_k)$$

- The chain rule can also be written for whole **distributions**, e.g.:

$$P(X_1, X_2, ..., X_N) = P(X_1, ..., X_k)P(X_{k+1}, ..., X_N \mid X_1, ..., X_k)$$

The Law of Total Probability

Definition

The **Law of Total Probability**:
In a world defined over random variables $\mathcal{X}$, for any event $(\boldsymbol{X} = \boldsymbol{x})$, $\boldsymbol{X} \subset \mathcal{X}$, and any subset of variables $\boldsymbol{Y} \subset \mathcal{X}$, the following holds:

$$P(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in Val(\boldsymbol{Y})} P(\boldsymbol{x}, \boldsymbol{y})$$

or, alternatively, following from the Chain Rule:

$$P(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in Val(\boldsymbol{Y})} P(\boldsymbol{x} \mid \boldsymbol{y}) P(\boldsymbol{y}) = \sum_{\boldsymbol{y} \in Val(\boldsymbol{Y})} P(\boldsymbol{x}) P(\boldsymbol{y} \mid \boldsymbol{x})$$

Proof:

- Obvious. □

The Law of Total Probability

Generalises to:

- **Marginal distributions:**

$$P(\boldsymbol{X}) = \sum_{\boldsymbol{y}} P(\boldsymbol{X}, \boldsymbol{y}) = \sum_{\boldsymbol{y}} P(\boldsymbol{X} \mid \boldsymbol{y}) P(\boldsymbol{y})$$

- **Conditional distributions:**

$$P(\boldsymbol{X} \mid \boldsymbol{Z}) = \sum_{\boldsymbol{y}} P(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{Z}) = \sum_{\boldsymbol{y}} P(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{Z}) P(\boldsymbol{y} \mid \boldsymbol{Z})$$

☞ **The Law of Total Probability justifies our marginalisation and "Inference by Enumeration" algorithms.**

Bayes' Rule

Another view on conditional probability:

$$P(\alpha \mid \beta) = \frac{P(\alpha \cap \beta)}{P(\beta)} \qquad \text{and} \qquad P(\alpha \cap \beta) = P(\beta \mid \alpha)P(\alpha)$$

imply:

Definition

BAYES' RULE:

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}$$

Notes:

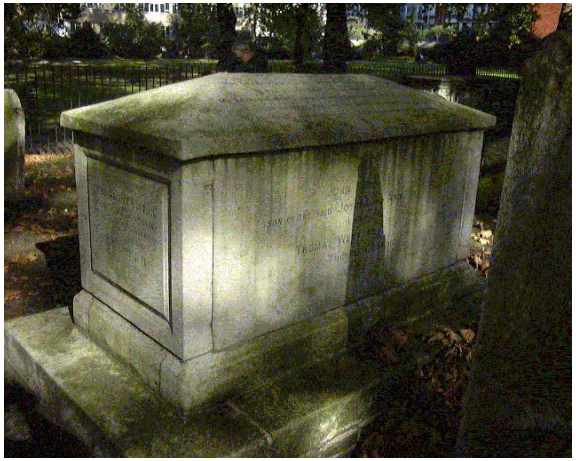
- Bayes' rule allows us to compute a conditional probability $P(\alpha \mid \beta)$ from the "inverse" conditional probability $P(\beta \mid \alpha)$
- Bayes' rule is one of the most fundamental tools in scientific data analysis and probabilistic reasoning.

The Reverend Thomas Bayes

Quote from the Encyclopedia Britannica (quote and image taken from the *International Society for Bayesian Analysis* [`http://bayesian.org/bayes`]):

Bayes, Thomas (b. 1702, London - d. 1761, Tunbridge Wells, Kent), mathematician who first used probability inductively and established a mathematical basis for probability inference (a means of calculating, from the number of times an event has not occured, the probability that it will occur in future trials).
He set down his findings on probability in *"Essay Towards Solving a Problem in the Doctrine of Chances"* (1763), published posthumously in the Philosophical Transactions of the Royal Society of London.
The only works he is known to have published in his lifetime are *"Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures"* (1731) and *"An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst"* (1736) which countered attacks by Bishop Berkeley on the logical foundations of Newton's calculus.

The Reverend Thomas Bayes



Quote and image taken from the *International Society for Bayesian Analysis* [`http://bayesian.org/bayes`]:

> Bayes is buried in Bunhill Fields in the heart of the City of London. The cemetery was used for the burial of nonconformists in the 18th century, but is now a public park maintained by the Corporation of London. Also buried in Bunhill Fields is Bayes's friend Richard Price, a pioneer of insurance, who presented Bayes's famous paper on probability to the Royal Society in 1763, two years after Bayes's death

A Simple Example

Assume we know (or believe, or estimate) from experience that

- $P(mp \mid flu) = 0.5$ (50% of the flu patients tend to have a muscle pain)
- $P(flu) = 0.004$ (0.4% of the population at any time tend to have the flu)

  Or: the probability of a random person having the flu (if we know nothing else about the person's condition) is 0.004
- $P(mp) = 0.1$ (10% of the population tend to suffer from a muscle pain).

Question:

- Given that you meet a person who suffers from a muscle pain, what is the probability that the person has the flu?

  By Bayes' rule:

$$P(flu \mid mp) = \frac{P(mp \mid flu)P(flu)}{P(mp)} = (0.5 \times 0.004)/0.1 = \underline{0.02}$$

Bayes' Rule in Diagnostic Settings

Know about:

- Possible observations (e.g., symptoms, lab tests, ...)
- Problems or diseases that would explain the symptoms
- General probabilities $P(Problem)$ of diseases/problems to occur
- Probabilities with which problems produce specific symptoms: $P(Symptoms \mid Problem)$.

Observe:

- A patient/situation with specific $Symptoms$

Want to determine:

- Probability of the underlying problem, given the symptoms:

$$P(Problem \mid Symptoms) = \frac{P(Symptoms \mid Problem)P(Problem)}{P(Symptoms)}$$

☞ **Diagnosis is a very common reasoning pattern (not only in medicine)!**

Bayes' Rule in Diagnostic Settings

Bayes' Rule in Diagnosis Settings:

$$P(Problem \mid Symptoms) = \frac{P(Symptoms \mid Problem)P(Problem)}{P(Symptoms)}$$

Names of the components:

$P(Problem)$	**Prior Probability** (degree of belief in $Problem$ **before** we have observed anything else about the current case)
$P(Problem \mid Symptoms)$	**Posterior Probability** (degree of belief in $Problem$ **after** we have observed $Symptoms$)
$P(Symptoms \mid Problem)$	**Likelihood** of $Problem$ in the presence of $Symptoms$ = probability with which $Problem$ produces $Symptoms$
$P(Symptoms)$	**"Evidence"** (probability of these $Symptoms$ occurring at all).

Note: By the Law of Total Probability, $P(Symptoms)$ can be calculated as

$$P(Symptoms) = \sum_{p \in Problems} P(Symptoms \mid p)P(p)$$

Probabilistic Inference and Reasoning

Scenario:

- The **full joint distribution** over some set of variables $\mathcal{X}$ will be the system's **Knowledge Base (KB)**, its **model of the world**.
- Gives a probability to each possible state of the world (= atomic event): how likely is it to encounter this specific combination of values?
- The system's task will be to compute, from this knowledge base and the known (observed) values of some variables, the probability of any hypothesis of interest.

Definition

The process of deriving the truth or probability of a hypothesis from a knowledge base and from specific observations is called **INFERENCE**, or **REASONING**.

☞ **In our context, inference will mainly mean *computing (conditional) probability distributions over variables of interest.***

Probabilistic Inference and Reasoning

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Possible queries we might like to answer:

- $P(fever \wedge \neg musclepain) = ?$

  What is the probability that someone has a fever, but no muscle pain?

  (☞ an *event probability*)

- $P(Flu) = ?$

  What is the probability that a random person has the flu (or not)?

  (☞ the *marginal distribution* over variable Flu)

- $P(Flu \mid fever, \neg musclepain) = ?$

  What is the probability that someone does or does not have the flu, if they have a fever, but no muscle pain?

  (☞ a *conditional distribution*)

- and many more ...

Inference by Enumeration

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Remember how we computed marginal probabilities from the full joint:

$$\begin{aligned} P(flu) &= P(flu, mp, fever) + P(flu, mp, \neg fever) + \\ &\quad P(flu, \neg mp, fever) + P(flu, \neg mp, \neg fever) \\ &= \sum_{x,y} P(flu, MP = x, Fever = y) \\ &= 0.15 + 0.1 + 0.1 + 0.03 = \underline{0.38} \end{aligned}$$

- Can compute probability of any event α as the sum over all atomic events that 'match' α
- This is a direct application of the *Law of Total Probability*
- Will permit us to compute the probability of any query we may want to ask of a distribution.

☞ **Very general algorithm!**

Example 1

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Computing Event Probabilities:

$$\begin{aligned} P(\neg fever) &= \sum_{x,y} P(\neg fever, MP = x, Flu = y) \\ &= 0.1 + 0.03 + 0.04 + 0.5 = 0.67 \end{aligned}$$

$$\begin{aligned} P(fever \wedge \neg musclepain) &= \sum_{x \in Val(Flu)} P(fever, \neg musclepain, Flu = x) \\ &= 0.1 + 0.05 = 0.15 \end{aligned}$$

Example 2

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Computing Conditional Probabilities:

$$
\begin{aligned}
P(flu \mid \neg fever) &= \frac{P(flu \wedge \neg fever)}{P(\neg fever)} \quad \text{(by definition of cond. prob.)} \\
&= \frac{\sum_x P(flu, \neg fever, MP = x)}{\sum_{x,y} P(Flu = x, \neg fever, MP = y)} \\
&= \frac{0.1 + 0.03}{0.1 + 0.03 + 0.04 + 0.5} \\
&= \frac{0.13}{0.67} \approx 0.194
\end{aligned}
$$

Example 3

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

Computing Conditional Distributions:

$$P(flu \mid \neg fever) = \frac{P(flu \wedge \neg fever)}{P(\neg fever)}$$
$$= \frac{0.1 + 0.03}{0.1 + 0.03 + 0.04 + 0.5} = \frac{0.13}{0.67} \approx 0.194$$

$$P(\neg flu \mid \neg fever) = \frac{P(\neg flu \wedge \neg fever)}{P(\neg fever)}$$
$$= \frac{0.04 + 0.5}{0.1 + 0.03 + 0.04 + 0.5} = \frac{0.54}{0.67} \approx 0.806$$

$$\Rightarrow \underline{P(Flu \mid \neg fever)} = \{0.806, 0.194\}$$

Example 3 (revisited)

$$
\begin{aligned}
P(\mathit{flu} \mid \neg \mathit{fever}) &= \frac{P(\mathit{flu} \wedge \neg \mathit{fever})}{P(\neg \mathit{fever})} = \frac{0.13}{0.67} \approx 0.194 \\
P(\neg \mathit{flu} \mid \neg \mathit{fever}) &= \frac{P(\neg \mathit{flu} \wedge \neg \mathit{fever})}{P(\neg \mathit{fever})} = \frac{0.54}{0.67} \approx 0.806 \\
\Rightarrow P(\mathit{Flu} \mid \neg \mathit{fever}) &= \{0.194, 0.806\}
\end{aligned}
$$

Notes:

- The denominator $P(\neg \mathit{fever}) = 0.67$ is the same in both cases (because $P(\neg \mathit{fever})$ does not depend on the value of Flu)
- In general: In computing a distribution $P(\boldsymbol{X} \mid \boldsymbol{y})$, the term $P(\boldsymbol{y})$ will be computed $|\mathit{Val}(\boldsymbol{X})|$ times!
- This constant term $P(\boldsymbol{y})$ serves as a **normalisation factor** Z that makes the resulting conditional distribution $P(\boldsymbol{X} \mid \boldsymbol{y})$ sum to 1.0
- $Z = P(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}, \boldsymbol{y})$ can be computed afterwards, in a simple way (without having to sum out over the full joint distribution) ...

Computing Conditional Distributions: The General Inference-by-Enumeration Algorithm

Goal:

- Compute $P(\boldsymbol{X} \mid \boldsymbol{y})$

  for some sets of variables $\boldsymbol{X}, \boldsymbol{Y} \subseteq \mathcal{X}$ and some specific value assignment $\boldsymbol{Y} = \boldsymbol{y}$.

Starting Point:

$$\boxed{P(\boldsymbol{X} \mid \boldsymbol{y}) = \frac{P(\boldsymbol{X}, \boldsymbol{y})}{P(\boldsymbol{y})}} \qquad \text{(definition of conditional probability)}$$

The Algorithm in Words:

1. For each possible value combination $\boldsymbol{x} \in Val(\boldsymbol{X})$, compute $P(\boldsymbol{x}, \boldsymbol{y})$ via enumeration
2. Compute denominator = normalisation constant $P(\boldsymbol{y})$ from these $P(\boldsymbol{x}, \boldsymbol{y})$ (simply by summing them up (Law of Total Probability))
3. Normalise $P(\boldsymbol{X}, \boldsymbol{y})$ to obtain conditional distribution $P(\boldsymbol{X} \mid \boldsymbol{y})$.

The General Inference-by-Enumeration Algorithm

Algorithm

1. **Compute** $P(\boldsymbol{X}, \boldsymbol{y})$:

$$P(\boldsymbol{X}, \boldsymbol{y}) = \sum_{\boldsymbol{z} \in Val(\boldsymbol{Z})} P(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z})$$

(*Inference by Enumeration*, $\boldsymbol{Z} = \mathcal{X} - \boldsymbol{X} - \boldsymbol{Y}$)

2. **Compute normalisation constant** Z **as**

$$Z = P(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}, \boldsymbol{y})$$

(*simply sum up all the* $P(\boldsymbol{x}, \boldsymbol{y})$ *computed in previous step*)

3. **Conditioning via renormalisation:**

$$P(\boldsymbol{X} \mid \boldsymbol{y}) = \frac{1}{Z} P(\boldsymbol{X}, \boldsymbol{y})$$

Example

$P(Flu, MP, Fever)$	$musclepain$		$\neg musclepain$	
	$fever$	$\neg fever$	$fever$	$\neg fever$
flu	0.15	0.1	0.1	0.03
$\neg flu$	0.03	0.04	0.05	0.5

$$\boxed{P(Flu, \neg fever)} = \sum_{x \in \{mp, \neg mp\}} P(Flu, \neg fever, MP = x)$$

$$= \begin{bmatrix} 0.1 + 0.03 \\ 0.04 + 0.5 \end{bmatrix} = \begin{bmatrix} 0.13 \\ 0.54 \end{bmatrix} \quad \begin{matrix} \leftarrow P(\ flu, \neg fever) \\ \leftarrow P(\neg flu, \neg fever) \end{matrix}$$

$$\boxed{Z} = \sum_{x \in \{flu, \neg flu\}} P(x, \neg fever) = 0.13 + 0.54 = 0.67$$

$$\boxed{P(Flu \mid \neg fever)} = \frac{1}{Z} \times \begin{bmatrix} 0.13 \\ 0.54 \end{bmatrix} = \begin{bmatrix} \mathbf{0.194} \\ \mathbf{0.806} \end{bmatrix} \quad \begin{matrix} \leftarrow P(\ flu \mid \neg fever) \\ \leftarrow P(\neg flu \mid \neg fever) \end{matrix}$$

An Important Concept: Independence

Starting Point:

- For two events α, β in a given world, we usually expect that $P(\alpha \mid \beta) \neq P(\alpha)$
- In other words, learning that β is true changes our degree of belief in α
- **But** sometimes, this may not be the case: β tells us nothing about α ...

Simple Example: Tossing two independent coins

- $\alpha =$ *"first coin comes up heads"*
- $\beta =$ *"second coin comes up heads"*
- Obvious: $P(\beta \mid \alpha) = P(\beta)$

Independence Between Events

Definition

Two events α and β are said to be **independent** in a joint distribution P, denoted as

$$(\alpha \perp \beta)$$

if[a]

$$P(\alpha \mid \beta) = P(\alpha) \quad \text{or, equivalently,}$$

$$P(\beta \mid \alpha) = P(\beta) \quad \text{or, equivalently,}$$

$$P(\alpha \cap \beta) = P(\alpha)P(\beta)$$

[a] Exercise: Prove that these three conditions are equivalent!

Note that independence is **symmetric:**

$$(\alpha \perp \beta) \Leftrightarrow (\beta \perp \alpha)$$

Independence Between Random Variables

Definition

Two **random variables** X and Y are **independent**, denoted as

$$(X \perp Y)$$

if **for all values** $x \in Val(X), y \in Val(Y)$:

$$P(x \mid y) = P(x) \quad \text{and} \quad P(y \mid x) = P(y)$$

or, in short:

$$P(X \mid Y) = P(X) \quad \text{and} \quad P(Y \mid X) = P(Y)$$

Consequence:

- If X and Y are independent, then

$$\boxed{P(X, Y) = P(X)P(Y)}$$

Independence Between Random Variables

Generalises to *Sets* of Random Variables:

Definition

Two **sets of random variables** $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **independent**, denoted as

$$(\boldsymbol{X} \perp \boldsymbol{Y})$$

if for all value combinations $\boldsymbol{x} \in Val(\boldsymbol{X}), \boldsymbol{y} \in Val(\boldsymbol{Y})$:

$$P(\boldsymbol{x} \mid \boldsymbol{y}) = P(\boldsymbol{x}) \quad \text{and} \quad P(\boldsymbol{y} \mid \boldsymbol{x}) = P(\boldsymbol{y})$$

or, in short:

$$P(\boldsymbol{X} \mid \boldsymbol{Y}) = P(\boldsymbol{X}) \quad \text{and} \quad P(\boldsymbol{Y} \mid \boldsymbol{X}) = P(\boldsymbol{Y})$$

Consequence:

- If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then

$$\boxed{P(\boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{X})P(\boldsymbol{Y})}$$

The Importance of Independence

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then $\boxed{P(\boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{X})P(\boldsymbol{Y})}$

Consequences:

- If two variable sets $\boldsymbol{X}, \boldsymbol{Y}$ are independent, their joint distribution $P(\boldsymbol{X}, \boldsymbol{Y})$ is equal to the product $P(\boldsymbol{X})P(\boldsymbol{Y})$
- Each entry $P(\boldsymbol{x}, \boldsymbol{y})$ in the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y})$ can be reconstructed from the lower-dimensional distributions $P(\boldsymbol{X})$ and $P(\boldsymbol{Y})$:

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{x})P(\boldsymbol{y})$$

- Need to model (store) joint distributions only over smaller subsets $\boldsymbol{X}$ and $\boldsymbol{Y}$ rather than over full set $\boldsymbol{X} \cup \boldsymbol{Y}$
- Remember: Number of entries in joint distribution grows **exponentially** with number of variables!

☞ **Effect: Reduction of Space Complexity.**

An Extreme Example

Consider N independent coins (binary variables $\{X_1, \ldots, X_N\}$)

- Can assume that all X_i, X_j are pairwise independent:

$$P(X_1, \ldots, X_N) = P(X_1)P(X_2) \cdots P(X_N)$$

- Can reconstruct joint distribution over $\{X_1, \ldots, X_N\}$ from one-dimensional distributions $P(X_1), P(X_2) \ldots$
- Joint distribution $P(X_1, \ldots, X_N)$ over all variables has 2^N entries
- Each marginal distribution $P(X_i)$ has only 2 entries $P(heads), P(tails)$[1]
- Need to store only $2N$ probabilities instead of 2^N

☞ **Exponential Reduction of Complexity from $O(2^N)$ to $O(N)$!**

[1] Actually, only one is needed: $P(tails) = 1 - P(heads)$

Conditional Independence

Consider this example:

$\alpha =$ *"patient Smith has a fever"*

$\beta =$ *"patient Smith feels a muscle pain"*

☞ α **and** β **are dependent:** Learning that Smith feels a muscle pain increases our expectation that she will also run a fever (because these often go together):

$$P(\alpha \mid \beta) > P(\alpha)$$

But if we ***already know that*** γ ***is true:***

$\gamma =$*"patient Smith has the flu"*

... then learning about β will probably not change our degree of belief in α (because α is already "explained" by γ):

$$P(\alpha \mid \beta \cap \gamma) = P(\alpha \mid \gamma)$$

☞ β ***tells us nothing new about*** α***, if we already know*** γ ***...***

Conditional Independence

Definition

Two events α and β are said to be **conditionally independent** **given a third event** γ, denoted as

$$(\alpha \perp \beta \mid \gamma)$$

if[a]

$$P(\alpha \mid \beta \cap \gamma) = P(\alpha \mid \gamma) \quad \text{or, equivalently,}$$

$$P(\beta \mid \alpha \cap \gamma) = P(\beta \mid \gamma) \quad \text{or, equivalently,}$$

$$P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma) P(\beta \mid \gamma)$$

[a] Exercise: Prove that these three conditions are equivalent!

Note that conditional independence is **symmetric:**

$$(\alpha \perp \beta \mid \gamma) \Leftrightarrow (\beta \perp \alpha \mid \gamma)$$

Conditional Independence Between Sets of Variables

Definition

Two **sets of random variables** $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **conditionally independent** given a third set $\boldsymbol{Z}$, denoted as

$$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$$

if for all values $\boldsymbol{x} \in Val(\boldsymbol{X}), \boldsymbol{y} \in Val(\boldsymbol{Y}), \boldsymbol{z} \in Val(\boldsymbol{Z})$:

$$P(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z}) \quad \text{and} \quad P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}) = P(\boldsymbol{y} \mid \boldsymbol{z})$$

or, in short:

$$P(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z}) = P(\boldsymbol{X} \mid \boldsymbol{Z}) \quad \text{and} \quad P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}) = P(\boldsymbol{Y} \mid \boldsymbol{Z})$$

Consequence:

- If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{Z}$, then

$$\boxed{P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = P(\boldsymbol{X} \mid \boldsymbol{Z}) P(\boldsymbol{Y} \mid \boldsymbol{Z})}$$

The Importance of Conditional Independence

If X and Y are conditionally independent given Z, then

$$\boxed{P(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z}) = P(\boldsymbol{X} \mid \boldsymbol{Z})}$$

Consequence:

- Joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ over all variables can be decomposed ("factorised") into product of simpler (conditional) distributions:

$$\begin{aligned} \underline{P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})} &= P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z})P(\boldsymbol{Y}, \boldsymbol{Z}) && (1) \\ &= P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z})P(\boldsymbol{Y}|\boldsymbol{Z})P(\boldsymbol{Z}) && (2) \\ &= \underline{P(\boldsymbol{X}|\boldsymbol{Z}) \cdot P(\boldsymbol{Y}|\boldsymbol{Z}) \cdot P(\boldsymbol{Z})} && (3) \end{aligned}$$

(1) *chain rule*
(2) *chain rule*
(3) *conditional independence*

☞ **Most fundamental and important concept in Graphical Models**
☞ **Be sure to understand this!**

Test Yourself[2]

$P(X_1, X_2)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.3	0.2
$X_1 = 2$	0.1	0.4

1. Find the following quantities:

- Marginals: $P(X_1), P(X_2)$
- Conditionals: $P(X_1 \mid X_2), P(X_2 \mid X_1)$
- Joint vs. Posterior: $P(X_1, X_2 = 2)$ vs. $P(X_1 \mid X_2 = 2)$
- Evidence (relative to previous item): $P(X_2 = 2)$
- Normalisation constant $P(\Omega) = P(true)$
- Max: $P(x_1^*) = \max_{x_1} P(x_1 \mid X_2 = 1)$
- Mode: $x_1^* = \arg\max_{x_1} P(x_1 \mid X_2 = 1)$
- Max-marginal: $\max_{x_1} P(x_1, X_2)$

2. Are X_1 and X_2 independent?

[2] Test motivated by Ali Cemgil's ISMIR 2006 Tutorial (`www-sigproc.eng.cam.ac.uk/~atc27/`).

Answers

$P(X_1, X_2)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.3	0.2
$X_1 = 2$	0.1	0.4

Marginals:

$P(X_1)$	
$X_1 = 1$	0.5
$X_1 = 2$	0.5

$P(X_2)$	$X_2 = 1$	$X_2 = 2$
	0.4	0.6

Conditionals:

$P(X_1 \mid X_2)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.75	0.33
$X_1 = 2$	0.25	0.67

$P(X_2 \mid X_1)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.6	0.4
$X_1 = 2$	0.2	0.8

Answers

$P(X_1, X_2)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.3	0.2
$X_1 = 2$	0.1	0.4

Joint vs. Posterior:

$P(X_1, X_2 = 2)$	$X_2 = 2$
$X_1 = 1$	0.2
$X_1 = 2$	0.4

$P(X_1 \mid X_2 = 2)$	$X_2 = 2$
$X_1 = 1$	0.33
$X_1 = 2$	0.67

Evidence: (relative to previous item):

$$P(X_2 = 2) = \sum_{x_1} P(x_1, X_2 = 2) = 0.6$$

Normalisation constant:

$$P(\Omega) = P(true) = \sum_{x_1} \sum_{x_2} P(x_1, x_2) = 1$$

Answers

$P(X_1, X_2)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.3	0.2
$X_1 = 2$	0.1	0.4

Max:

$$P(x_1^*) = \max_{x_1} P(x_1 \mid X_2 = 1) = 0.75$$

Mode:

$$x_1^* = \arg\max_{x_1} P(x_1 \mid X2 = 1) = 1$$

Max-marginal (get the "skyline"):

$\max_{x_1} P(x_1, X_2)$	$X_2 = 1$	$X_2 = 2$
	0.3	0.4

Answers

$P(X_1, X_2)$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.3	0.2
$X_1 = 2$	0.1	0.4

2. Are X_1 and X_2 independent?

Remember: Two random variables X, Y are (marginally) independent if

$$P(X \mid Y) = P(X) \quad (\text{or} \quad P(Y \mid X) = P(Y))$$

☞ easy to see that X_1 and X_2 are .....

Probabilities over Continuous Variables

Motivation:

- So far, have considered only *discrete* variables with *finite* domains $Val(X)$
- But: In real-world applications, many (most) of the relevant variables will be **continuous** (real-valued) !
- Examples: Sensors in an autonomous robot – range sensors, temperature and pressure sensors, GPS, camera, ...

Problem:

- Probability of each specific value (from an infinite set) is zero!

Simple Example:

- Assume continuous variable X with $Val(X) = [0, 1] \subset \mathbb{R}$
- Assume we want to assign the same probability to each number in this range (uniform distribution)

$\Rightarrow$ infinite number of possible values x

$\Rightarrow$ $P(X = x)$ must be 0 for all x (otherwise $\sum_x P(x)$ would be infinite)!

(This problem also appears for other, non-uniform distributions)

Density Functions

Definition

A function $p : \mathbb{R} \mapsto \mathbb{R}$ is a **Probability Density Function (PDF)** for a continuous variable X if

- $p(x) \geq 0$ for all $x \in Val(X)$
- $p(x) = 0$ for all $x \notin Val(X)$
- $\int_{-\infty}^{+\infty} p(x)dx = 1$

Important:

- A PDF is **not a probability distribution!**
- $p(x)$ must not be interpreted as a probability (can even be > 1!)
- Probabilities are defined on top of PDFs via *Cumulative Distributions*

Probabilities from Densities

The PDF defines a set of **distributions** over X as follows:

Definition

For any $a \in \mathbb{R}$,

$$P(X \leq a) = \int_{-\infty}^{a} p(x)dx$$

This is called the **Cumulative Distribution Function (CDF)** for X.

Intuitively: PDF $p(x)$ is the incremental amount that x adds to the cumulative distribution in the integration process.

Consequence (follows from rules of probability)

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = \int_{a}^{b} p(x)dx$$

☞ PDFs permit us to compute probabilities for (arbitrarily small) *intervals* of values.

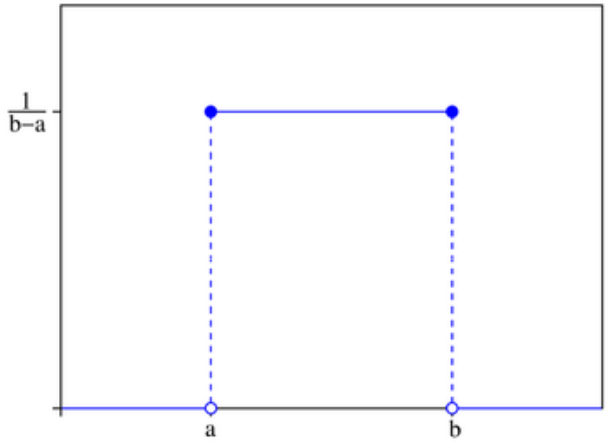
Example 1: The Uniform Distribution

Definition

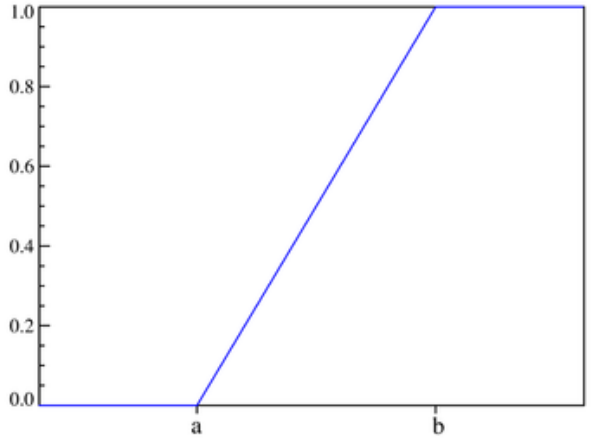
A real variable X has a **Uniform Distribution** over a range $[a, b]$, denoted $X \sim Unif[a, b]$, if it has the PDF

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

PDF:



CDF:



Note:

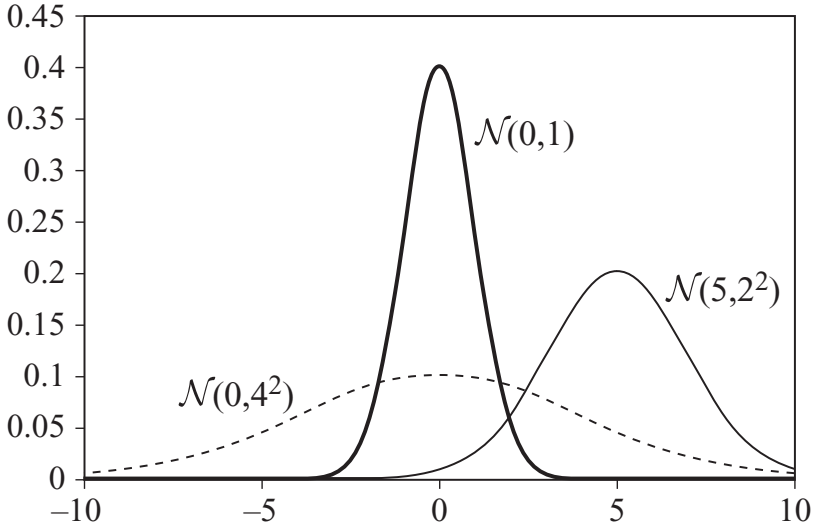
- If $(b - a) < 1$, the PDF $p(x)$ is > 1!
- Can happen in any legal PDF, as long as total area under the PDF is 1.0.

Example 2a: The Normal (Gaussian) Distribution

Definition

A variable X has a **Normal Distribution** with mean μ and variance σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has the **Gaussian PDF**

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



Example 2b: The Multivariate Gaussian

Definition

N random variables $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ have a **Joint (Multivariate) Normal Distribution** with mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, denoted $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if they have the **Multivariate Gaussian PDF**

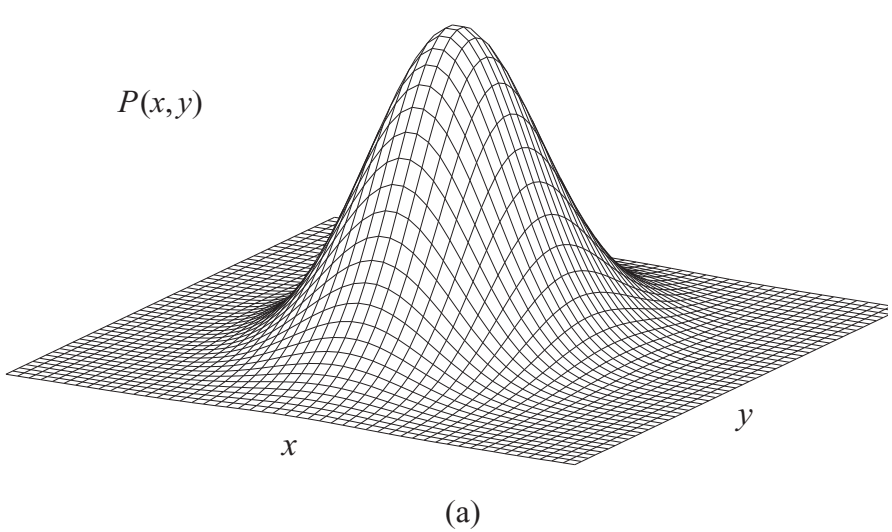
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]}$$

where

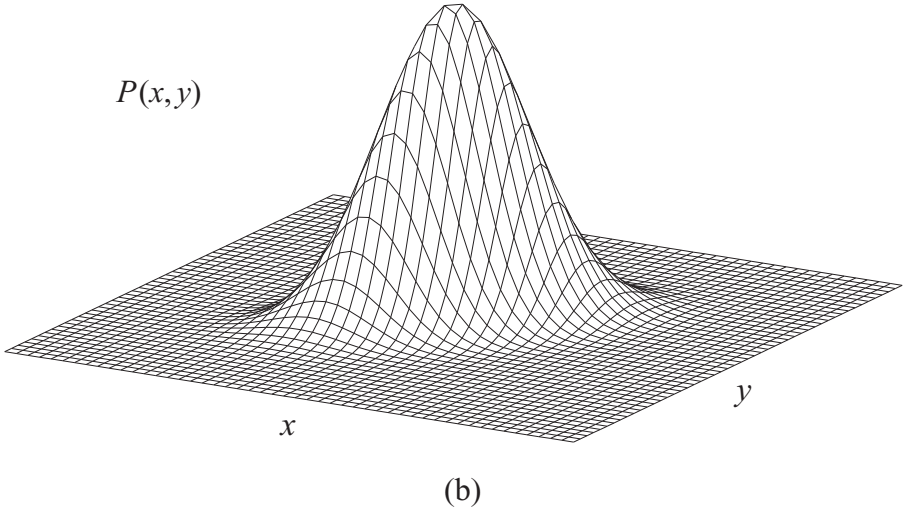
- $\boldsymbol{\mu}$ is a *mean vector* $\boldsymbol{\mu} = [\mu_{X_1}, ..., \mu_{X_N}]$
- $\boldsymbol{\Sigma}$ is the $N \times N$ *covariance matrix* (with $\boldsymbol{\Sigma}_{ij} = \sigma_{ij}$ the *covariance (= pairwise correlation)* between variables X_i and X_j)
- $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$

Example 2b: The Multivariate Gaussian

2-D Gaussians over two variables X and Y:



(a)



(b)

X and Y are uncorrelated ($\sigma_{X,Y} = 0$)

X and Y are (positively) correlated ($\sigma_{X,Y} > 0$)

Notational Conventions: Summary

NOTATION

Examples	Type of object
$\mathcal{X}$	Complete set of variables that a model is defined over
A, X_i, E_j	Single random variable (e.g., $X_i \in \mathcal{X}$)
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{U}, \ldots$	Sets of random variables (e.g., $\boldsymbol{X} = \{X_1, X_2, X_3, X_4\}$)
$Val(X)$	Domain of variable X (set of values that X can take)
$x^0, x^1, \ldots$	$x^i = i^{th}$ value in the domain $Val(X)$ of variable X
x^0, x^1	Special convention for boolean variables: $x^0 = false, x^1 = true$
$\boldsymbol{x}_i, \mathbf{e}, \ldots$	Set/list of specific values of a variable (e.g., $\boldsymbol{E} = \boldsymbol{e}$)
$Val(\boldsymbol{X})$	Set of possible value assignments to the variables in $\boldsymbol{X}$ (= Cartesian product $Val(X_1) \times Val(X_2) \times \ldots$)
$\boldsymbol{x}\langle \boldsymbol{Y} \rangle$	Instance $\boldsymbol{x}$ reduced to its values for variables $\boldsymbol{Y} \subset \boldsymbol{X}$
$X_i^{(0)}, X_i^{(1)}, X_i^{(t)}$	Instantiation of a template variable X_i at time t (see Chapter 6)
$\boldsymbol{\theta}$	Set of parameters of a model
$P(X = x^i)$	Probability (probability mass function)
$P(x^i), P(x^i, y^j)$	Shorthand notation for $P(X = x^i), P(X = x^i, Y = y^j)$
$P(X, Y)$	Probability distribution over discrete variables X, Y
$p(x)$	Probability Density Function (PDF) over continuous variable X

Literature

Cemgil, A. T. (2006).
Bayesian Methods for Music Signal Analysis. Tutorial, 7th International Conference on Music Information Retrieval (ISMIR 2006), Victoria, Canada.
`http://www-sigproc.eng.cam.ac.uk/~atc27/papers.html`.

Koller, Daphne and Friedman, Nir (2009).
Probabilistic Graphical Models: Principles and Techniques.
Cambridge, MA: MIT Press.

Russell, Stuart J. and Norvig, Peter (2021).
Artificial Intelligence: A Modern Approach. Fourth Edition
Pearson Publishers.