Random Variables and Distribution Functions

Probabilistic Decision Making — Lecture 3

15th October 2025

Robert Peharz
Institute of Machine Learning and Neural Computation
Graz University of Technology

Probability Spaces

- a **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ containing
  - **sample space** Ω (**any** non-empty set)
  - **sigma-algebra** $\mathcal{F}$ (logically closed system of subsets of Ω)
  - **probability measure** $\mathbb{P}$ (mapping $\mathcal{F}$ to $[0, 1]$)
- **rules of probability:**
  - $\mathbb{P}(\Omega) \geq 0$
  - $\mathbb{P}(\Omega) = 1$
  - For any **disjoint** $A_1, A_2, A_3, \ldots$ from $\mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \qquad \text{(sigma-additivity)}$$

- **how can we interpret $\mathbb{P}(A)$?**
  - **fair prices**, "anti-Dutch-book"
  - **frequentist interpretation**: normalized frequency
  - **Bayesian interpretation**: relaxed logical value
  - **Shannon information**: information content, surprise

Random Variables

Omnipresent Random Variables

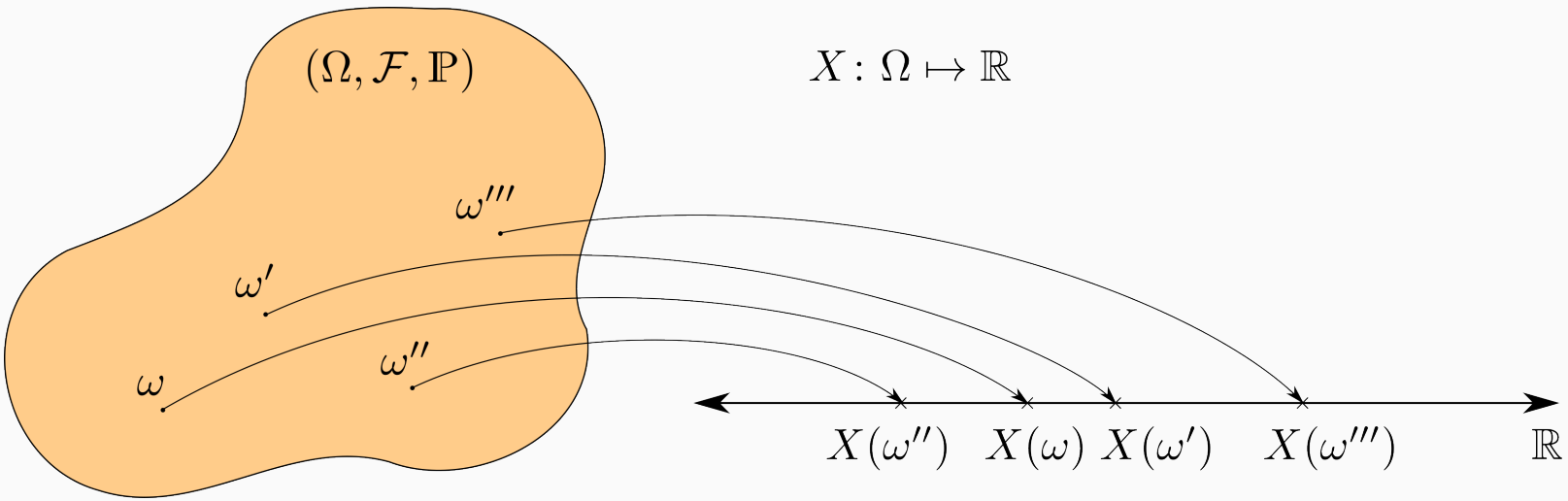
- **all data are random variables**
- images, speech, text, class labels, regression targets, molecules, social network connections, etc., are all "too complicated" to be modelled deterministically
- to treat **uncertainty**, any data are treated as **random variables**
- these random variables are of course **correlated**
- **in a nutshell, machine learning is about capturing and exploiting these correlations**
- **probabilistic machine learning** puts the fact at the center of attention that data are actually random variables

Random Variables

- the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is our mathematical model of randomness
- **random variables** are functions mapping the probability space to "numeric values" (usually the real numbers)

Probability Space
"Random Generator"

Random Variable
"Function of Randomness"



Random Variable

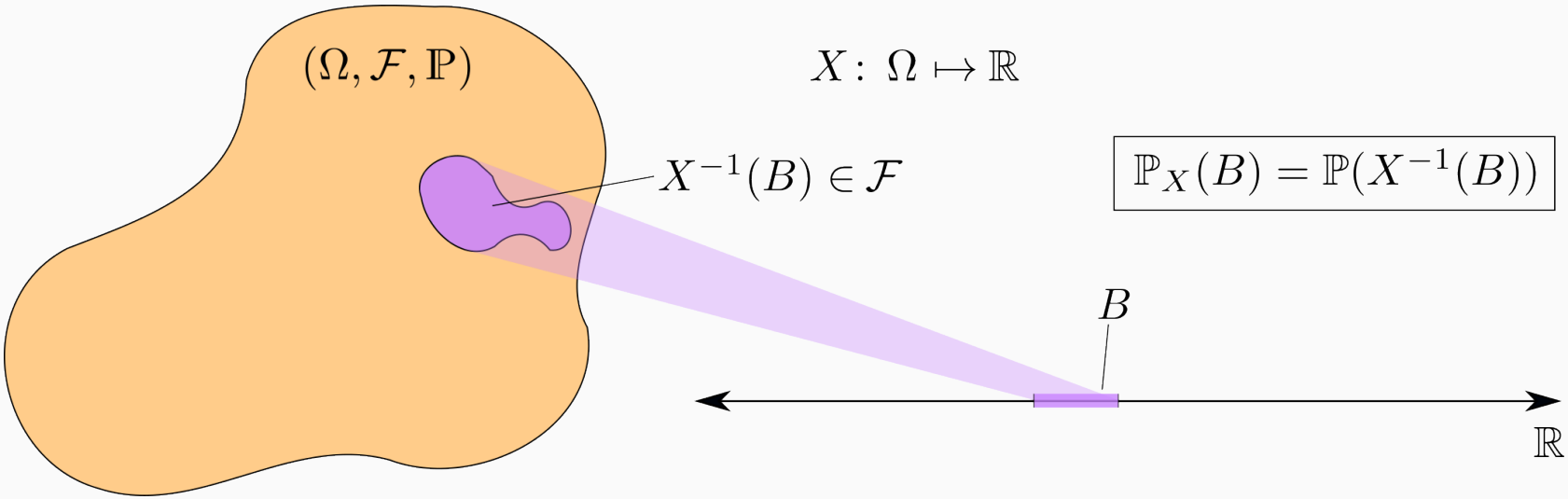
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{X}$ be some space, very often the real numbers $\mathcal{X} = \mathbb{R}$. A function $X \colon \Omega \mapsto \mathcal{X}$ is called a **random variable** (RV).

Technically, X must be a so-called **measurable** function. This means, that we need to equip the output space $\mathcal{X}$ with another sigma-algebra $\mathcal{F}_X$, and make sure that for any $B \in \mathcal{F}_X$ the **pre-image** of B under function X must be in $\mathcal{F}$, the sigma-algebra of the probability space. This is a technical assumption which "make the math work." It holds for all RVs we will consider in this course.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \mapsto \mathcal{X}$ a random variable. For any (measurable) $B \subseteq \mathcal{X}$ let

$$\mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B))$$

where $X^{-1}(B)$ is the **pre-image** of B under X (i.e., the set of elements in Ω which map to some element in B). $\mathbb{P}_X$ is called the **distribution** or **push-forward measure** of X.



- **random variables (RV)**: omnipresent in ML
- an RV is really just a (measurable) function defined on a probability space
- transforms the probability measure to a "numerical" space (**push-forward measure**)
- **we will usually work with RVs, while the underlying probability space (the "Linux kernel") remains implicit**

Distribution Functions

Distribution Functions

- probability measures, whether they are constructed directly or are induced by RVs, are the general description of probability
- however, they are **painful to work with**
- **distribution functions** make our life easier
  - **probability mass functions** (discrete RVs)
  - **probability density functions** (continuous RVs)
  - **cumulative distribution functions**

Discrete Random Variable

Let X be an RV with state space $\mathcal{X}$, RV X is **discrete** if its state space $\mathcal{X}$ is **countable**.

Probability Mass Function (PMF)

Let X be a **discrete** RV with state space $\mathcal{X}$. The **probability mass function** (PMF) $p_X : \mathcal{X} \mapsto \mathbb{R}$ of X is defined as

$$p_X(x) \coloneqq \mathbb{P}_X(\{x\}).$$

Note: $\mathbb{P}_X$ takes **sets** of states as arguments while p_X takes states

Let X be a RV with state space $\mathcal{X} = \{0, 1\}$ and PMF

$$p_X(x \mid \theta) = \begin{cases} 1 - \theta & \text{if } x = 0 \\ \theta & \text{if } x = 1 \end{cases}$$

The distribution of such a binary RV X is called **Bernoulli distribution** with **success probability** θ, where $0 \leq \theta \leq 1$.

Jakob Bernoulli, 1654–1705

A typical example of a Bernoulli RV is the outcome of a coin toss. For a fair coin, $\theta = 0.5$.

Let X be a RV with finite state space $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$ and let its PMF p_X be:

$$p_X(x \mid \boldsymbol{\theta}) = \begin{cases} \theta_1 & \text{if } x = x_1 \\ \theta_2 & \text{if } x = x_2 \\ \vdots & \\ \theta_K & \text{if } x = x_K \end{cases}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ with $\theta_i \geq 0$ and $\sum_{i=1}^{K} \theta_i = 1$. Any such p_X is a **categorical distribution**.

For example, let $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ be the state space of some RV X representing the outcome of a die. For a fair die, $\theta_1 = \theta_2 = \cdots = \theta_6 = \frac{1}{6}$.

Probability Mass Function cont'd

- for any PMF p_X, we have
  - $p_X(x) \geq 0$
  - $\sum_{x \in \mathcal{X}} p_X(x) = 1$
- conversely, any function p with this property is the PMF **for some RV** (defined on some probability space)
- specifies the probability of all **singleton events** $\{x\}$:

$$p_X(x) = \mathbb{P}_X(\{x\})$$

- it also completely specifies $\mathbb{P}_X$, as for any event $A \subseteq \mathcal{X}$

$$\mathbb{P}_X(A) = \sum_{x \in A} p_X(x)$$

- **this works for discrete (countable) spaces; for continuous (uncountable) spaces, we need something else**

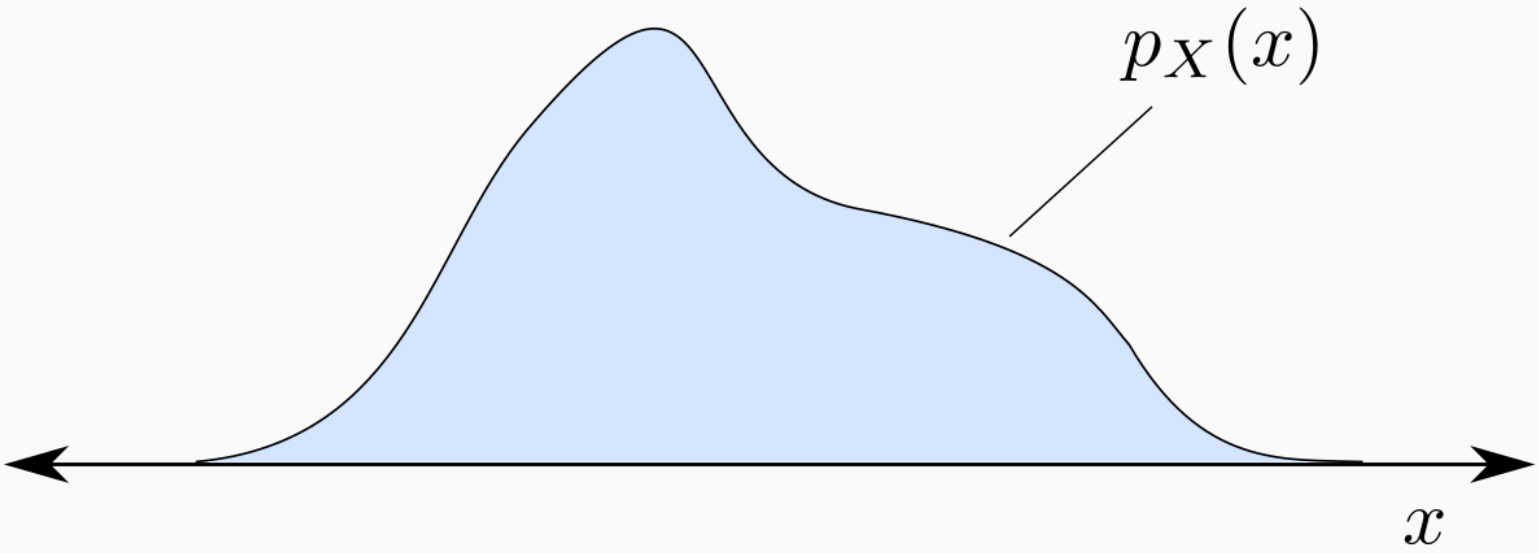
Probability Density Function (PDF)

Let X be a RV with state space $\mathcal{X} = \mathbb{R}$. Let $p_X : \mathbb{R} \mapsto [0, \infty]$, where $\int_A p_X(x)\,\mathrm{d}x = \mathbb{P}_X(A)$ for any event A. Function p_X, if it exists, is called **probability density function** (PDF) or also simply **density** of X.

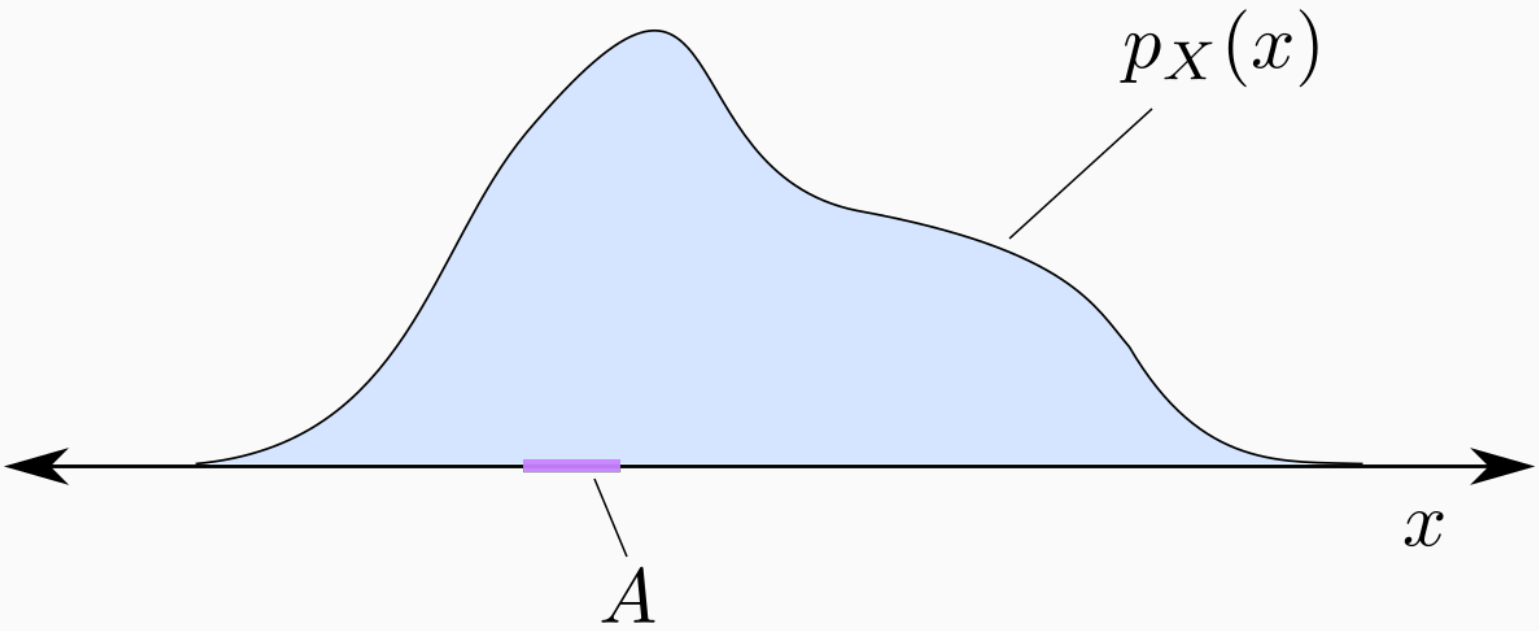
Continuous Random Variable

An RV X which has a density p_X is called **continuous**.

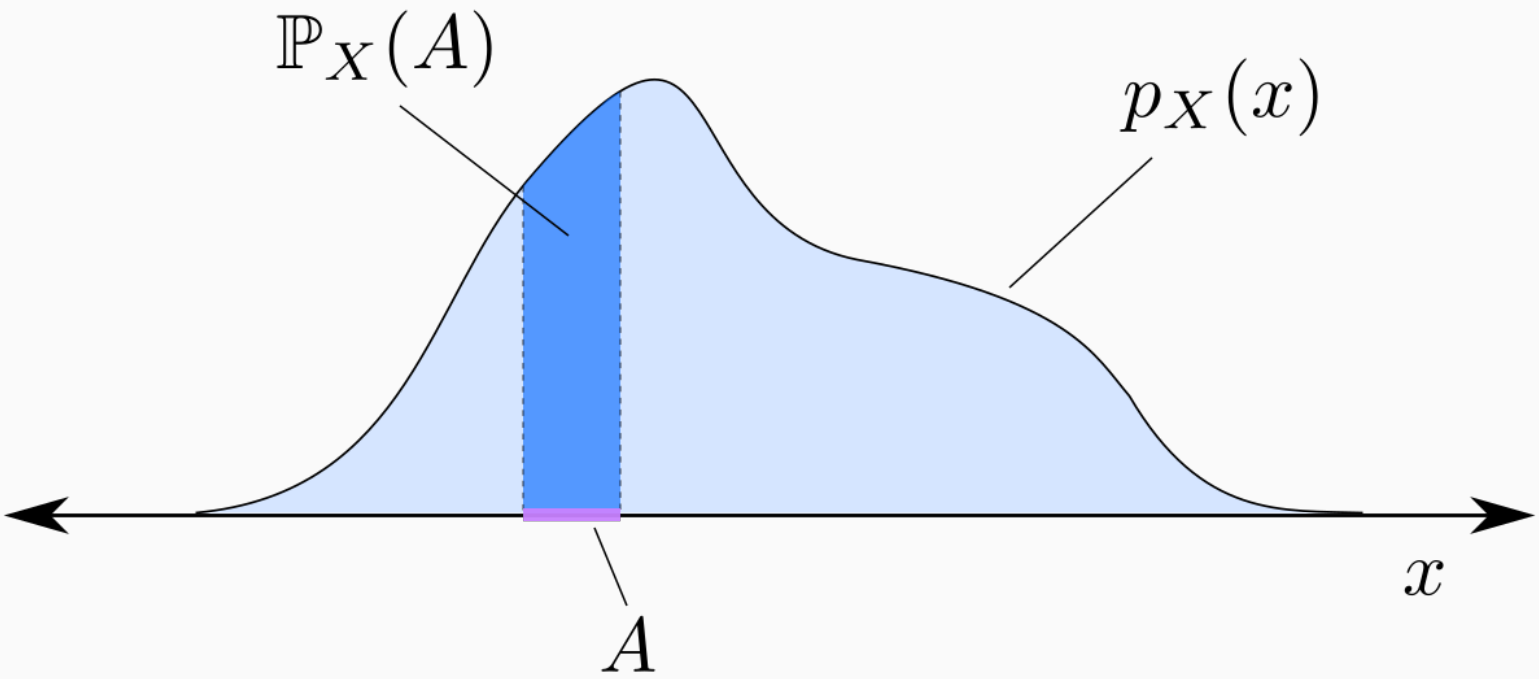
$\int_A p_X(x)\,\mathrm{d}x = \mathbb{P}_X(A)$.



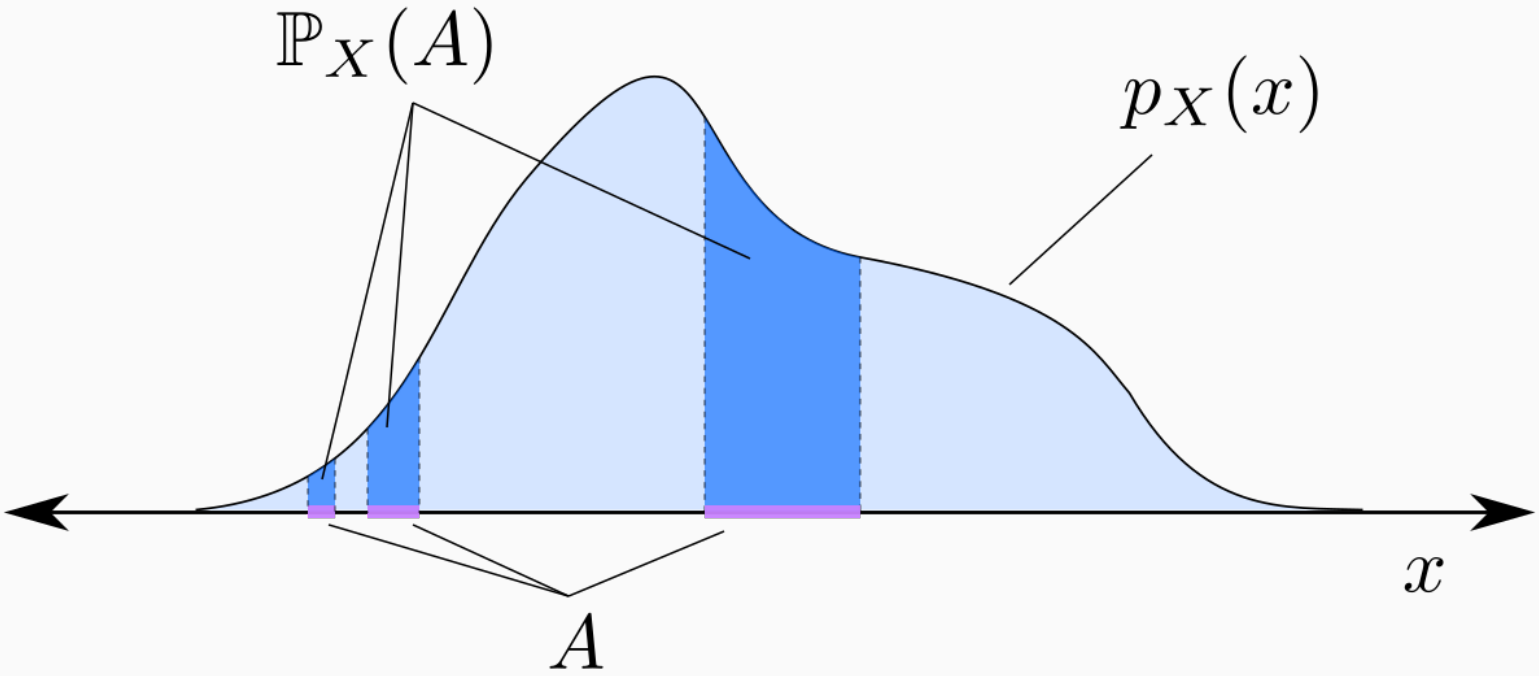
$\int_A p_X(x)\,\mathrm{d}x = \mathbb{P}_X(A)$.



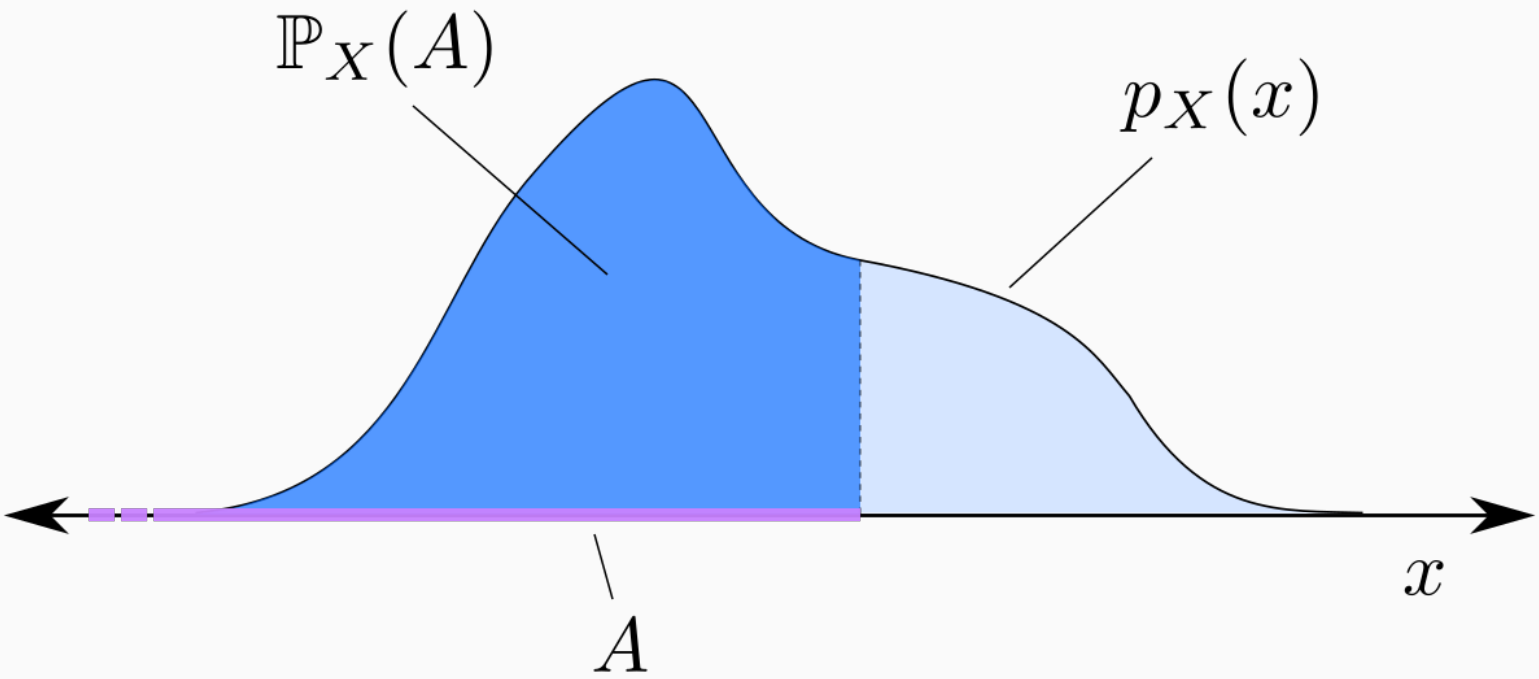
$\int_A p_X(x)\,\mathrm{d}x = \mathbb{P}_X(A)$.



$\int_A p_X(x)\,\mathrm{d}x = \mathbb{P}_X(A)$.



$\int_A p_X(x)\,\mathrm{d}x = \mathbb{P}_X(A)$.



Probability Density cont'd

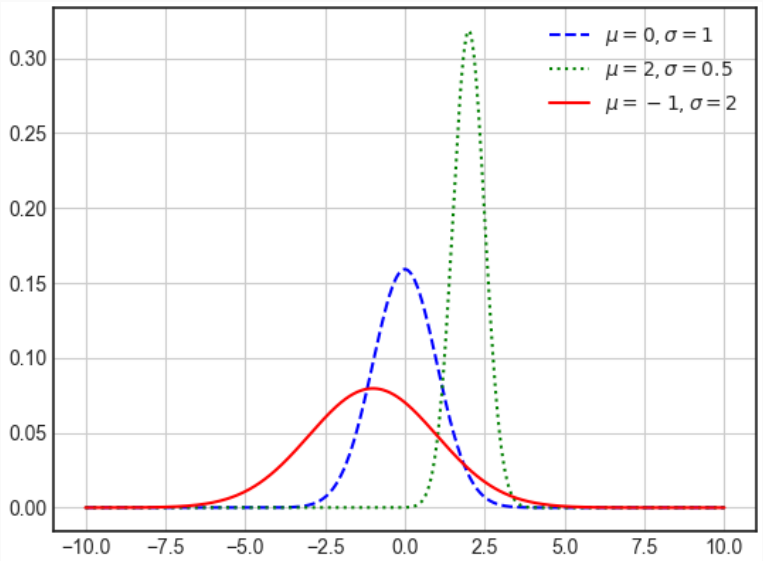
- for any PDF p_X we have
  - $p_X(x) \geq 0$
  - $\int_{\mathcal{X}} p_X(x)\,\mathrm{d}x = 1$
- conversely, any function with these properties is the PDF **for some RV**.
- **warning: probability density $\neq$ probability!**

Gaussian Distribution, Normal Distribution

Let p_X be a probability density defined as

$$p_X(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

with parameters **mean** μ and **standard deviation** $\sigma > 0$. A distribution with density p_X is called **Gaussian distribution** or **normal distribution**.

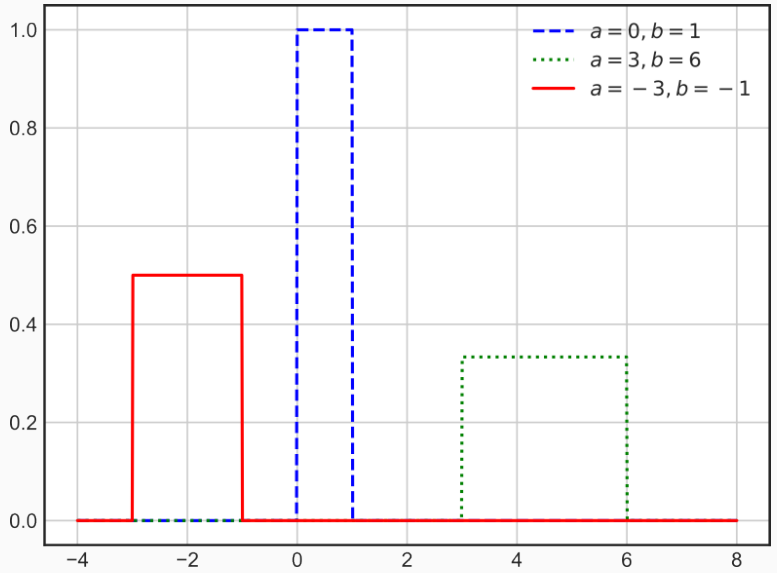


Carl Friedrich Gauss 1777–1855

Uniform Distribution

The **uniform distribution** on the interval $[a, b]$ has the pdf

$$p_X(x \mid a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



Distribution Functions

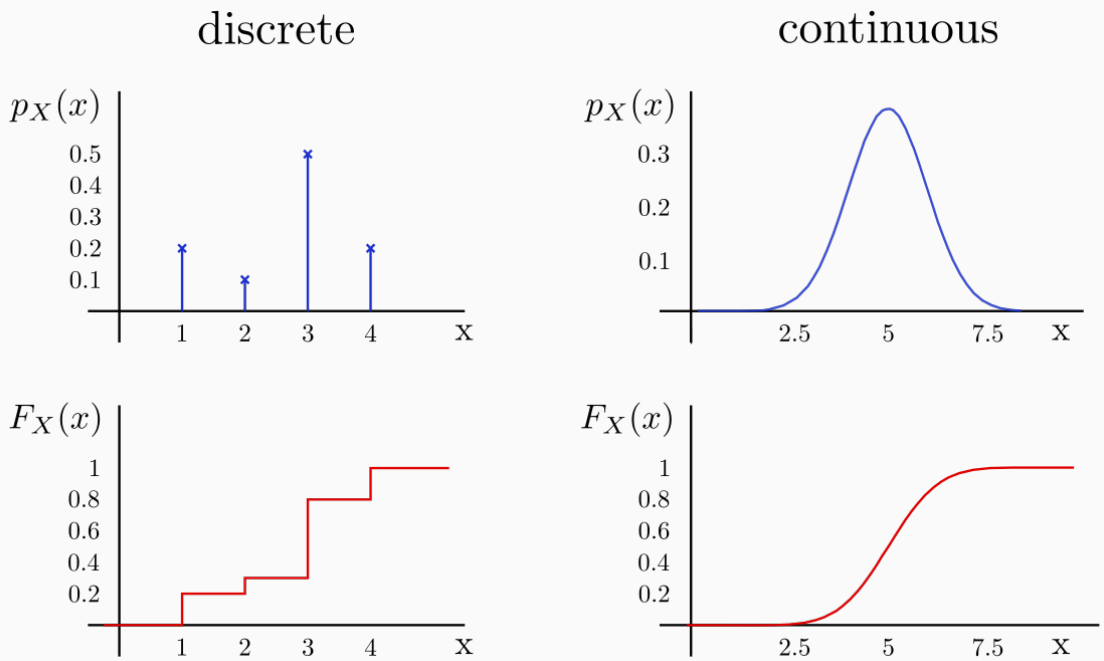
- we have overloaded the symbol p_X to denote both
  - PMF (discrete RV)
  - PDF (continuous RV)
- we refer to either as **distribution function**
- from a measure theoretic perspective, there is no difference between them—both are actually densities w.r.t. to a different **base measure** (counting vs. Lebesgue measure)
- **compact description of probability measure** ("high-level user interface")
- another description are cumulative distribution functions

Cumulative Distribution Function

- Let X be an RV with $\mathcal{X} \subseteq \mathbb{R}$ (discrete **or** continuous)
- The **cumulative distribution function** (CDF) is defined as

$$F_X(x) := \mathbb{P}_X([-\infty, x])$$

- F_X is monotonous increasing from 0 to 1
- for densities, p_X is the derivative of F_X



Notation

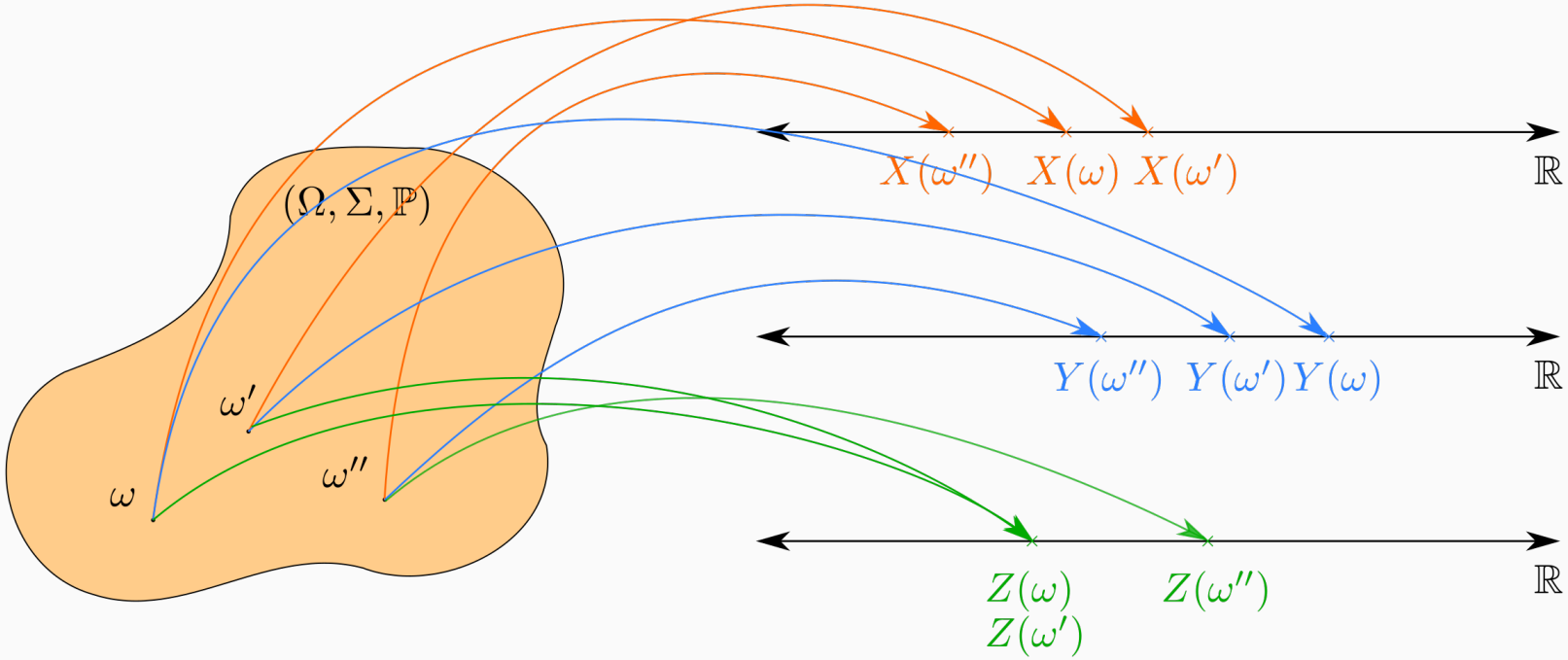
- uppercase letters X, Y, Z for RVs
- RVs take values in some **state space** $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$
- **values** or **states** (elements of state space) are denoted with lowercase letters x, y, z
- distributions (probability measures) of RVs are denoted as $\mathbb{P}_X$, $\mathbb{P}_Y$, $\mathbb{P}_Z$
- distribution function are denoted as p_X, p_Y, p_Z (PMF, PDF)
- cumulative distribution function are denoted F_X, F_Y, F_Z

Multivariate Random Variables

Multivariate Random Variables

- so far, we were talking only about **single** RVs, describing only a single random quantity
- however, real data consists of **many** correlated RVs:
  - pixels in an image are just a bunch of correlated RVs
  - tokens in large language models are correlated RVs
  - class value or regression targets are RVs
- we require **multivariate RVs**, also called **random vectors**
- in contrast, single RVs are called **univariate RV**

While univariate RVs are just a single function mapping from a probability space to the real numbers, a multivariate RV consists of **several** functions defined on the **same** probability space:



Multivariate Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_D$ be some spaces, very often the real numbers $\mathcal{X}_i = \mathbb{R}$. A collection of RVs $\{X_i : \Omega \mapsto \mathcal{X}_i\}_{i=1}^D$ is called a **multivariate RV**.

Equivalently, they represent a vector-valued function

$$\boldsymbol{X} : \Omega \mapsto \boldsymbol{\mathcal{X}}, \quad \boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_D(\omega))$$

denoted as **random vector**, where $\boldsymbol{\mathcal{X}} = \bigtimes_{i=1}^D \mathcal{X}_i$ is the **joint state space** (Cartesian product of $\mathcal{X}_i$'s).

Is there really something new?

- RVs are functions $X : \Omega \mapsto \mathcal{X}$, mapping atomic events $\omega \in \Omega$ to states $x \in \mathcal{X}$
- often we will assume $\mathcal{X} = \mathbb{R}$, **but $\mathcal{X}$ can of course be any set, for example $\mathbb{R}^D$**
- hence, multivariate RVs are still "normal RVs"—they just map to a vectors instead of scalars
- keeping this in mind should help understanding the following material better

Notation

To highlight multivariate RVs, we will use **boldface symbols:**

- uppercase boldface letters $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ for **multivariate random variables**, **random vectors**.
- they take values from some **joint state space** $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}}$, $\boldsymbol{\mathcal{Z}}$
- **joint values** denoted with boldface lowercase letter $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$
- random vectors follow a **distribution** $\mathbb{P}_{\boldsymbol{X}}$, $\mathbb{P}_{\boldsymbol{Y}}$, $\mathbb{P}_{\boldsymbol{Z}}$
- **distribution function** are denoted $p_{\boldsymbol{X}}$, $p_{\boldsymbol{Y}}$, $p_{\boldsymbol{Z}}$

Joint Probability Mass Function (PMF)

Let $X_1, X_2, \ldots, X_D$ be **discrete RVs** forming a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)$. The state space of $\boldsymbol{X}$ is

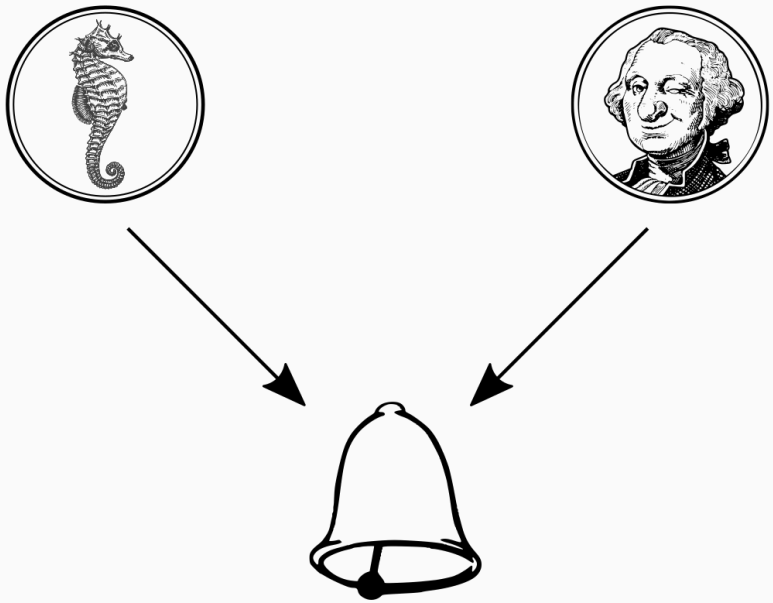
$$\boldsymbol{\mathcal{X}} = \bigtimes_{i=1}^{D} \mathcal{X}_i,$$

where $\mathcal{X}_i$ is the state space of X_i. The **joint probability mass function** (PMF) $p_{\boldsymbol{X}} : \boldsymbol{\mathcal{X}} \mapsto [0, 1]$ is defined as

$$p_{\boldsymbol{X}}(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{X}}(\{\boldsymbol{x}\}).$$

Completely the same definition as in the univariate case—just a "bigger" (combinatorial) state space.

- two fair coins are tossed, modeled with Bernoulli X_1 and X_2
- if both show heads, a bell (Bernoulli X_3) rings with certainty
- if exactly one shows heads, the Bell rings with 50% probability
- if both show tails, the Bell rings with 1% probability



- RVs X_1, X_2, X_3, with state spaces $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \{0, 1\}$
- random vector $\boldsymbol{X} = (X_1, X_2, X_3)$ with state space $\boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 = \{(0,0,0), (0,0,1), (0,1,0), \ldots, (1,1,1)\}$.
- the PMF is given as:

$\boldsymbol{x}$	$p_{\boldsymbol{X}}(\boldsymbol{x})$
$(0,0,0)$	0.2475
$(0,0,1)$	0.0025
$(0,1,0)$	0.125
$(0,1,1)$	0.125
$(1,0,0)$	0.125
$(1,0,1)$	0.125
$(1,1,0)$	0
$(1,1,1)$	0.25

also written

x_1	x_2	x_3	$p_{\boldsymbol{X}}(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

We will see how to obtain this PDF in a future lecture.

Joint Probability Density Function (PDF)

Let $X_1, X_2, \ldots, X_D$ be RVs, each with state space $\mathbb{R}$. Then $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)$ is a random vector with state space $\mathbb{R}^D$. If there exists a function $p_{\boldsymbol{X}} : \mathbb{R}^D \mapsto [0, \infty]$ with

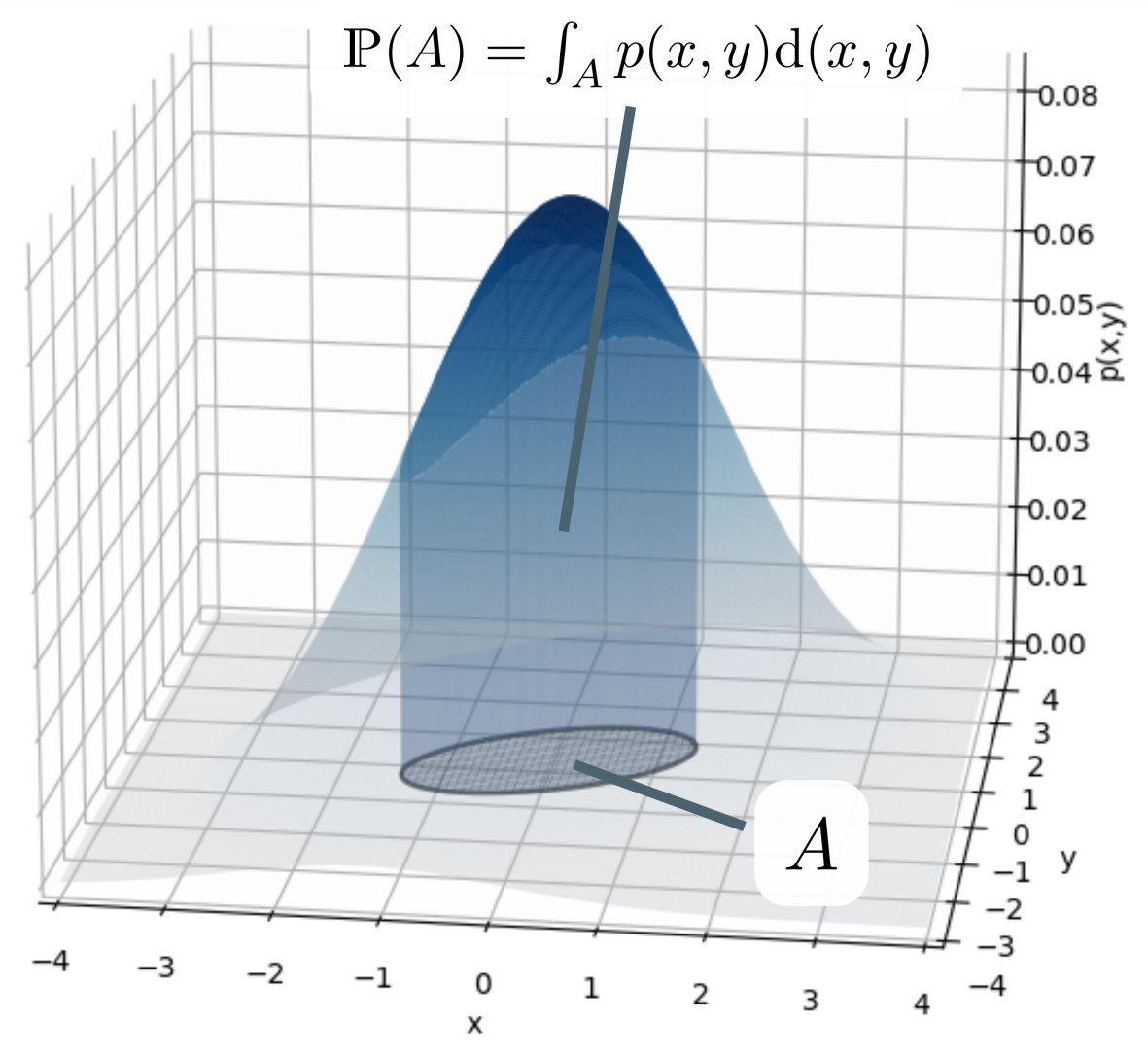
$$\mathbb{P}_{\boldsymbol{X}}(A) = \int_A p_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

then $p_{\boldsymbol{X}}$ is called a **joint probability density function** (PDF) or simply **joint density** of $\boldsymbol{X}$.

Completely the same definition as in the univariate case—just a "bigger" (combinatorial) state space.

Multivariate Probability Density

Example

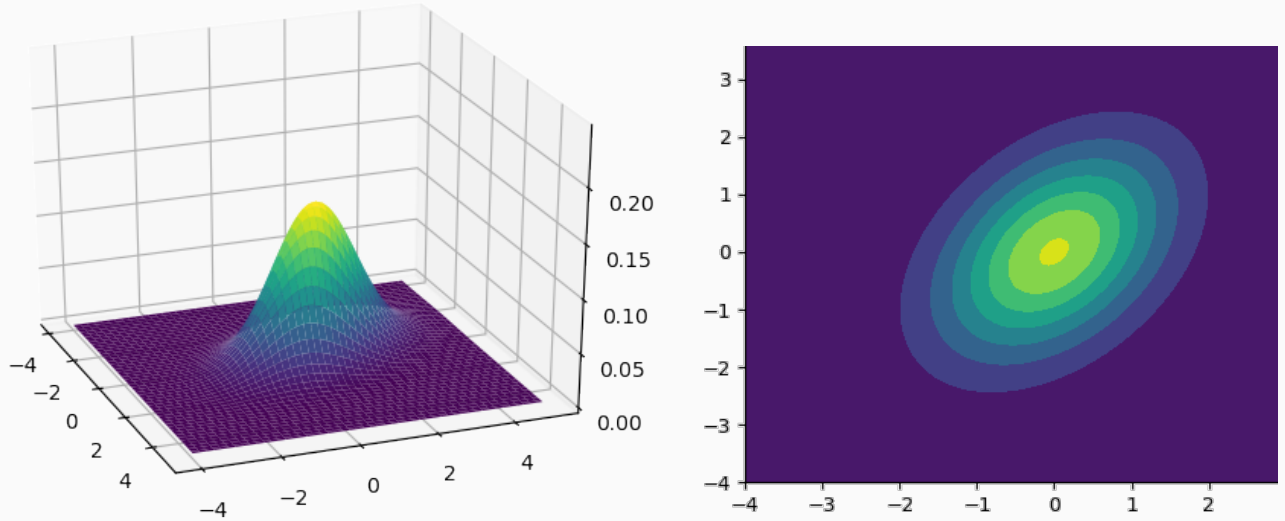


Multivariate Gaussian Distribution

Let $p_{\boldsymbol{X}}$ be a probability density defined as

$$p_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D \, |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

with D-dimensional **mean vector** $\boldsymbol{\mu}$ and $D \times D$ **positive definite covariance matrix** $\boldsymbol{\Sigma}$. A distribution with density $p_{\boldsymbol{X}}$ is called **multivariate Gaussian distribution**.



Covariance Matrix

The covariance matrix Σ of a Gaussian is a **positive definite matrix**, which means that it is **symmetric** and has only **non-negative eigenvalues**

- the diagonal element Σ_{dd} is the **variance** of the d^{th} RV X_d
- the off-diagonal element Σ_{de} is the **co-variance** of the d^{th} and e^{th} RVs X_d and X_e

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1D} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{D1} & \Sigma_{D2} & \cdots & \Sigma_{DD} \end{pmatrix}$$

Notation

- the central object of interest in this course: **joint distribution**
- let's simplify notation a bit
  - when redundant due to the argument, we drop the subscript:

$$\text{instead of } p_{\boldsymbol{X}}(\boldsymbol{x}) \text{ write } p(\boldsymbol{x})$$

  - when referring to whole density we still use $p_{\boldsymbol{X}}$, or being a bit sloppy $p(\boldsymbol{x})$ (even it is technically "p evaluated at $\boldsymbol{x}$")
  - when $\boldsymbol{X} = \{X_1, X_2, \ldots, X_D\}$ we might write

$$p(\boldsymbol{x}) \quad \text{or} \quad p(x_1, \ldots, x_D)$$

  - similarly, for partitions of $\boldsymbol{X}$ ($\boldsymbol{Z} \cup \boldsymbol{Y} = \boldsymbol{X}$ and $\boldsymbol{Y} \cap \boldsymbol{Z} = \emptyset$), we write

$$p(\boldsymbol{x}) \quad \text{or} \quad p(\boldsymbol{y}, \boldsymbol{z})$$

Probabilistic Inference (Part 1)

Logic

$$\forall X : \text{gummy\_bear}(X) \wedge \text{red}(X) \rightarrow \text{eaten}(X)$$

$$\forall X \, \forall Y : \text{gummy\_bear}(X) \wedge \text{stick}(X, Y) \rightarrow \text{gummy\_bear}(Y)$$

Systems of Equations

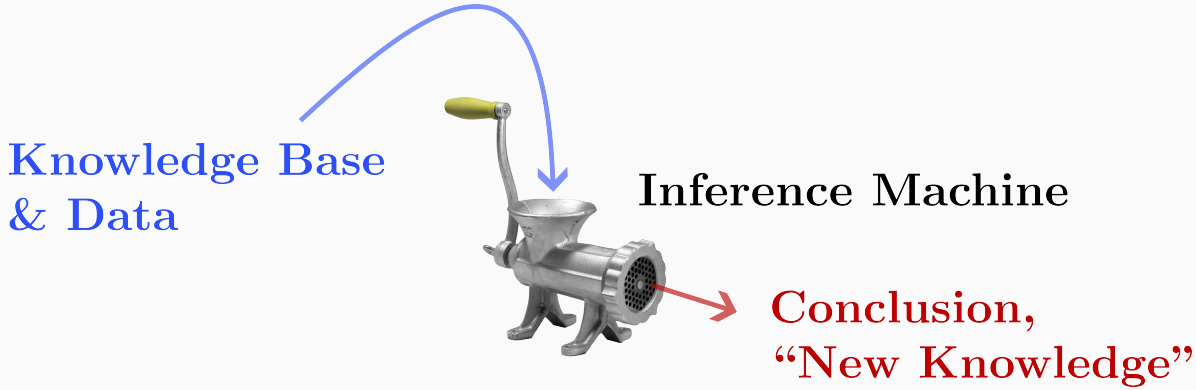
$$a^2 = b^2 + c^2$$

$$a = e + 3f^2$$

Probability

a	b	c	$p(a, b, c)$
0	0	0	0.14
0	0	1	0.05
0	1	0	0.2
0	1	1	0.05
1	0	0	0.01
1	0	1	0.09
1	1	0	0.2
1	1	1	0.3

Reasoning Systems cont'd



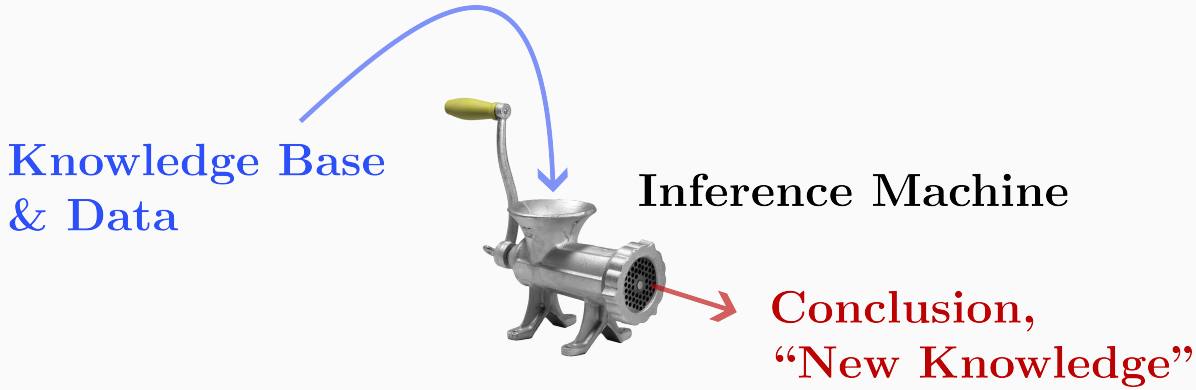
- **automatic reasoning** is a key ability for any AI system
- on a high level, an automatic reasoning system needs some
  - **knowledge base**, **domain model**
    capturing what we know about the world/task/domain
  - **inference machine**
    rules that allow us to manipulate the available knowledge and data in order to derive conclusions

Reasoning Systems cont'd

	Knowledge Base	**Inference Machine**
Logic	axioms, logical rules	rules of deduction
Equational Reasoning	system of equations	algebraic manipulations, calculus
Probability	joint distribution	probabilistic inference

- don't think of probability only in terms of frequencies—that's just one mental picture! (Bayesian, Shannon information, fair prices)
- probabilistic reasoning is **exact and rigorous**—just because it **treats uncertainty** doesn't mean that the reasoning is "fuzzy"
- **caveat:** it is also computationally hard and needs to be approximated—approximation introduces (often random) errors

Probabilistic Inference



- **knowledge base = joint distribution $p_{\boldsymbol{X}}$**
  - represents both **dependencies** between RVs and **uncertainty**
  - many **probabilistic models** for joints exist (to be discussed)
- **inference machine**

  the two key rules:
  - **marginalization** (**sum rule**)
  - **conditioning** (**product rule**)

  but also:
  - **expectations**, **maximization/minimization**, **sampling**

Marginal Distribution

Let p be the distribution of $\boldsymbol{X}$, and let $\boldsymbol{Y}$ and $\boldsymbol{Z} = \{Z_1, \ldots, Z_K\}$ be any partition of $\boldsymbol{X}$ ($\boldsymbol{Y} \cap \boldsymbol{Z} = \emptyset$ and $\boldsymbol{Y} \cup \boldsymbol{Z} = \boldsymbol{X}$). Then

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\mathcal{Z}}} p(\boldsymbol{y}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\mathcal{Z}_1} \cdots \int_{\mathcal{Z}_K} p(\boldsymbol{y}, z_1, \ldots, z_K) \mathrm{d}z_1 \ldots \mathrm{d}z_K$$

is the **marginal distribution** of $\boldsymbol{Y}$.

For discrete RVs, integrals are replaced by sums:

$$p(\boldsymbol{y}) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} p(\boldsymbol{y}, \boldsymbol{z}) = \sum_{z_1 \in \mathcal{Z}_1} \cdots \sum_{z_K \in \mathcal{Z}_K} p(\boldsymbol{y}, z_1, \ldots, z_K)$$

Marginalization is often also called the **sum rule**.

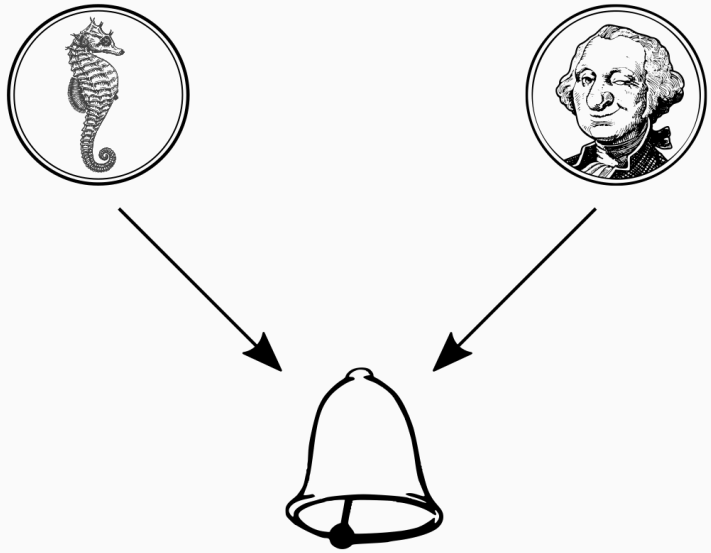
Marginal Distribution cont'd

- the marginal $p(\boldsymbol{y})$ is the (joint) distribution over only $\boldsymbol{Y}$ "as if $\boldsymbol{Z}$ had never existed"
- **this is no approximation!** $p(\boldsymbol{y})$ describes the exact push-forward measure under RVs $\boldsymbol{Y}$
- **semantics of marginalization**:
  - **ignore**
  - **forget**
  - **account for unknowns**
- a joint distribution over D RVs can be thought as containing all 2^D **sub-marginals**. Special cases:
  - marginalizing no RVs yields the original joint
  - marginalizing all RVs yields the normalization constant (1 for normalized distributions)

Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

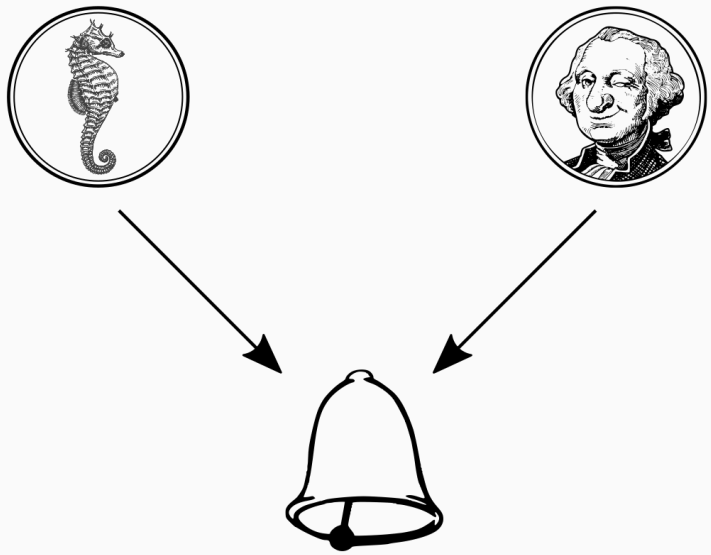


Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

x_1	x_2	$p(x_1, x_2)$
0	0	
0	1	
1	0	
1	1	

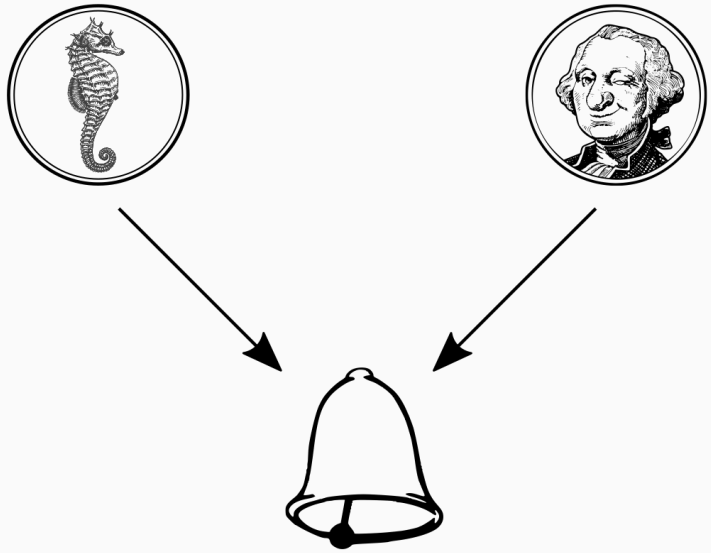


Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

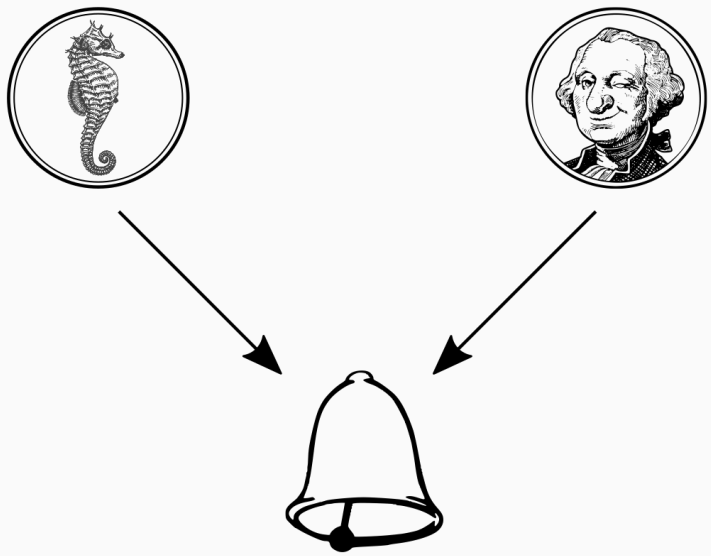
x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25



Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25



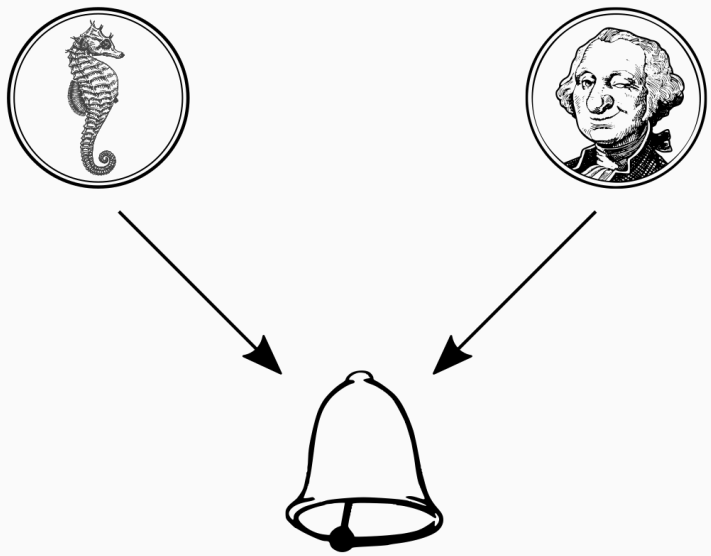
x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	
0	1	
1	0	
1	1	

Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25



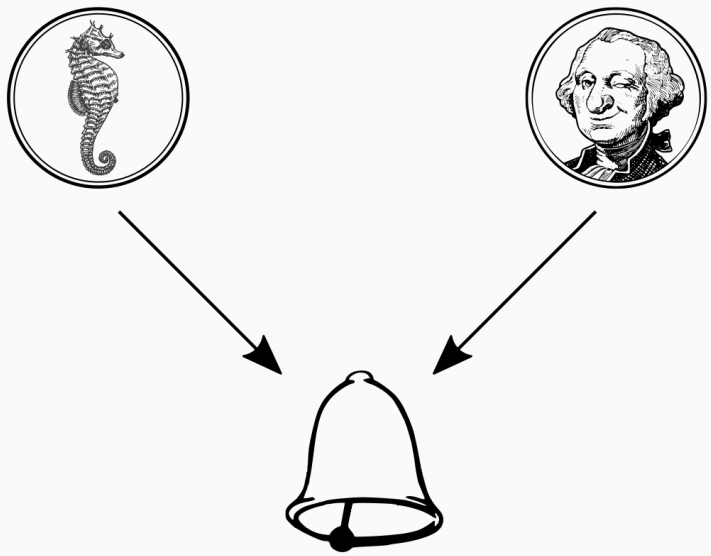
x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25



x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

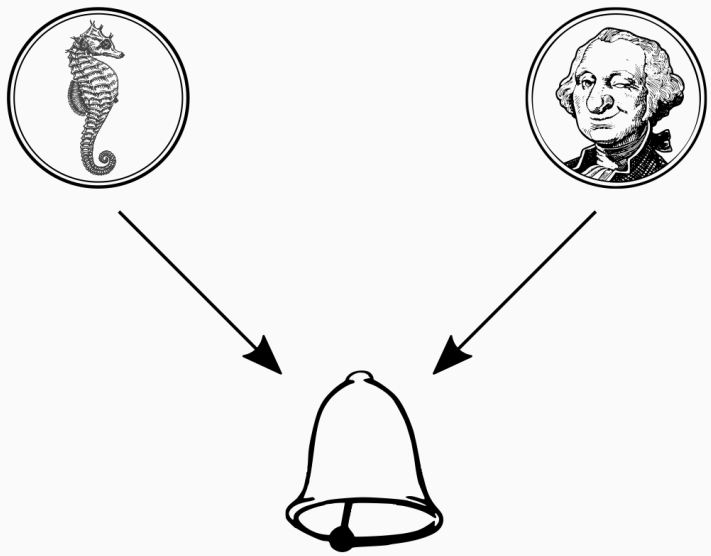
x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	
1	

Marginal Distributions (PMFs)

Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25



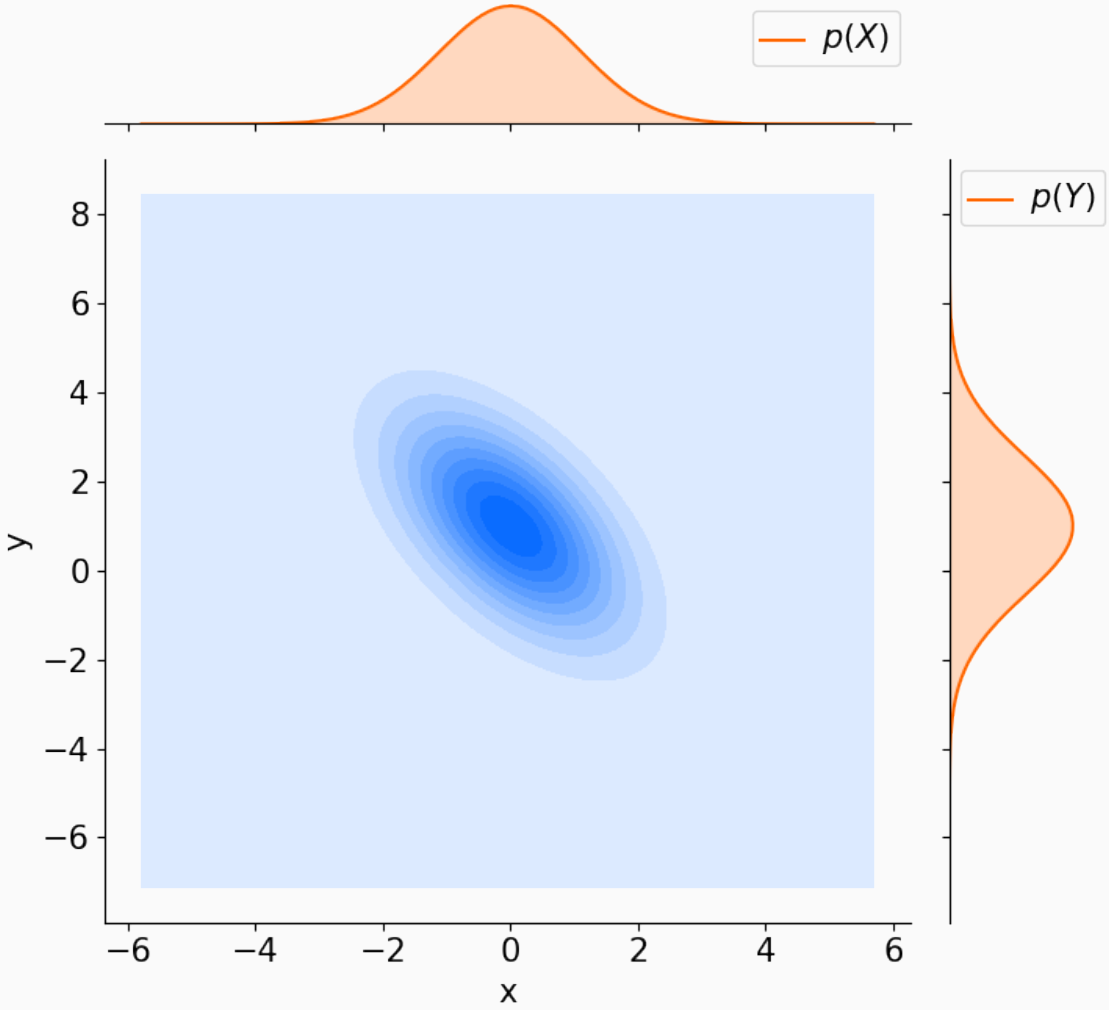
x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	0.4975
1	0.5025

Marginal Distributions (Gaussian)

Example



Gaussian marginals to be discussed formally in an upcoming lecture.

Conditional Distribution

Let p the **joint distribution** of $\boldsymbol{Y}$ and $\boldsymbol{Z}$. The **conditional distribution** of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is defined as (for $p(\boldsymbol{z}) > 0$):

$$p(\boldsymbol{y} \mid \boldsymbol{z}) = \frac{\overbrace{p(\boldsymbol{y}, \boldsymbol{z})}^{\text{joint}}}{\underbrace{p(\boldsymbol{z})}_{\text{marginal}}} = \frac{p(\boldsymbol{y}, \boldsymbol{z})}{\int_{\mathcal{Y}} p(\boldsymbol{y}, \boldsymbol{z}) \mathrm{d}\boldsymbol{y}}$$

Conversely, if **marginal** $p(\boldsymbol{z})$ and **conditional** $p(\boldsymbol{y} \mid \boldsymbol{z})$ are given, the **joint distribution** is given as

$$p(\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \boldsymbol{z})\, p(\boldsymbol{z})$$

This is often called the **product rule**.

Conditional Distribution cont'd

- the conditional $p(\boldsymbol{y} \mid \boldsymbol{z})$ is the (joint) distribution over $\boldsymbol{Y}$ upon observing that $\boldsymbol{Z} = \boldsymbol{z}$
- hence, it should be seen as a **collection** of distributions, one for each state $\boldsymbol{z} \in \mathcal{Z}$ the RVs $\boldsymbol{Z}$ can assume
- **semantics of conditioning**:
  - **observe**
  - **inject/update information**
- **only stuff on the left of the conditioning bar "|" is random—the stuff on the right is "observed input"**
- a joint distribution over D RVs can be thought as containing all 2^D **sub-conditionals**. Special cases:
  - conditioning on no RVs yields the original joint
  - conditioning on all RVs yields 1

Conditional Distribution (PMFs)

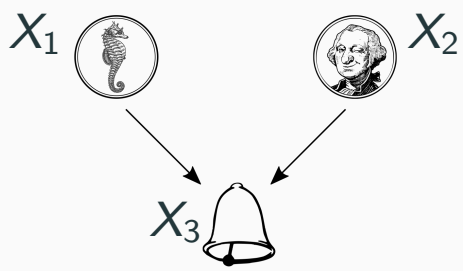
Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	0.4975
1	0.5025



Conditional Distribution (PMFs)

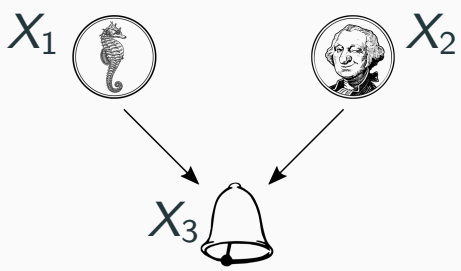
Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	0.4975
1	0.5025



x_3	x_1	x_2	$p(x_3 \mid x_1, x_2)$
0 1	0	0	
0 1	0	1	
0 1	1	0	
0 1	1	1	

Conditional Distribution (PMFs)

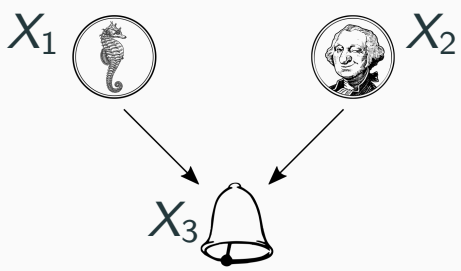
Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	0.4975
1	0.5025



x_3	x_1	x_2	$p(x_3 \mid x_1, x_2)$
0	0	0	0.99
1			0.01
0	0	1	0.5
1			0.5
0	1	0	0.5
1			0.5
0	1	1	0
1			1

Conditional Distribution (PMFs)

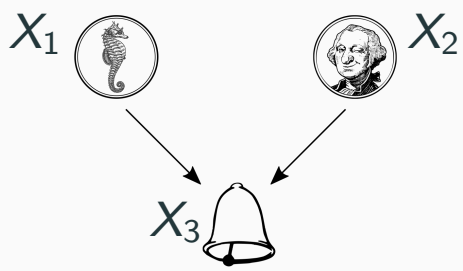
Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	0.4975
1	0.5025



x_3	x_1	x_2	$p(x_3 \mid x_1, x_2)$
0	0	0	0.99
1			0.01
0	0	1	0.5
1			0.5
0	1	0	0.5
1			0.5
0	1	1	0
1			1

x_1	x_3	$p(x_1 \mid x_3)$
0	0	
1		
0	1	
1		

Conditional Distribution (PMFs)

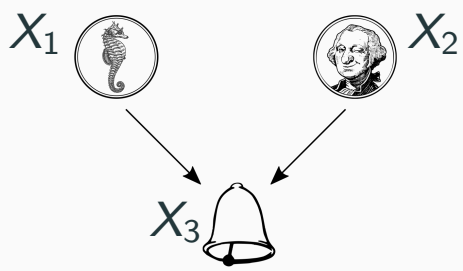
Example

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	0.2475
0	0	1	0.0025
0	1	0	0.125
0	1	1	0.125
1	0	0	0.125
1	0	1	0.125
1	1	0	0
1	1	1	0.25

x_1	x_2	$p(x_1, x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x_1	x_3	$p(x_1, x_3)$
0	0	0.3725
0	1	0.1275
1	0	0.125
1	1	0.375

x_3	$p(x_3)$
0	0.4975
1	0.5025

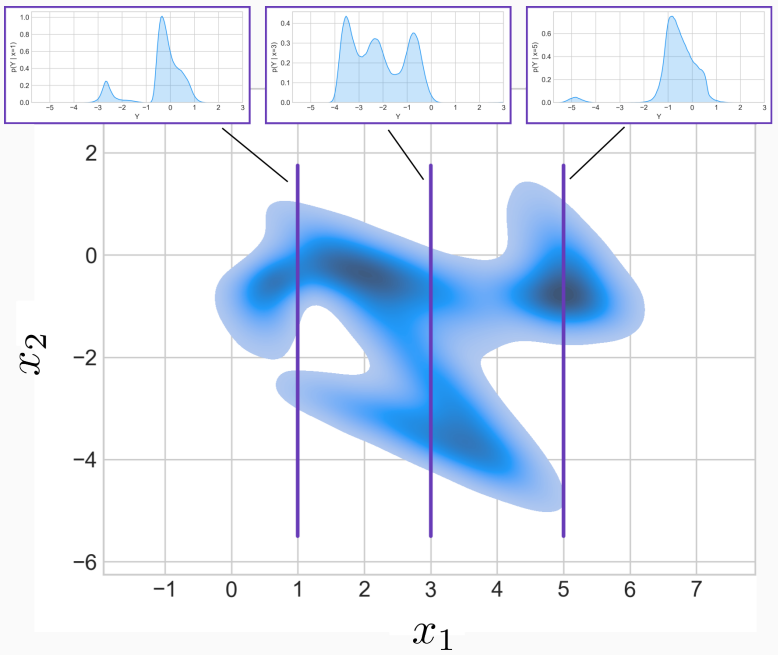


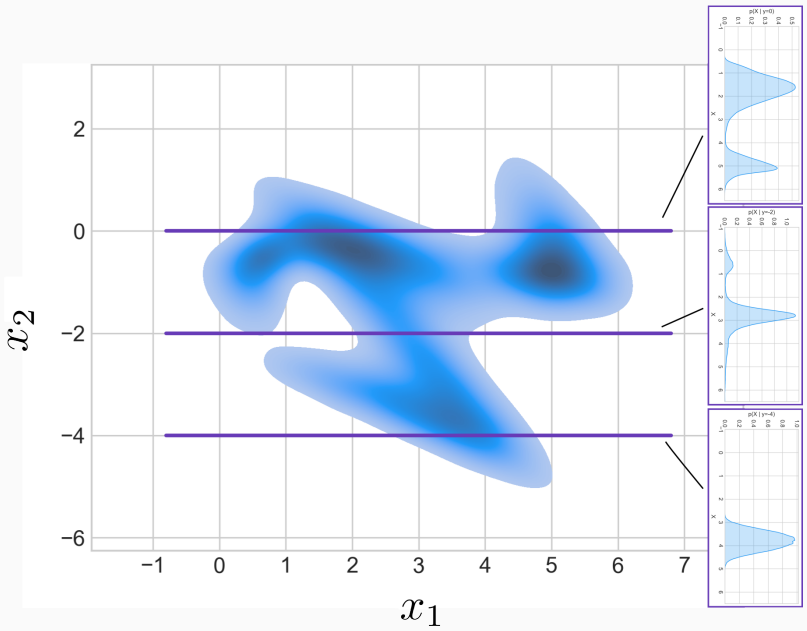
x_3	x_1	x_2	$p(x_3 \mid x_1, x_2)$
0	0	0	0.99
1			0.01
0	0	1	0.5
1			0.5
0	1	0	0.5
1			0.5
0	1	1	0
1			1

x_1	x_3	$p(x_1 \mid x_3)$
0	0	0.749
1		0.251
0	1	0.254
1		0.746

For continuous RVs, conditionals can be thought as axis-aligned slices through the joint, followed by renormalization:

$$p(\boldsymbol{y} \mid \boldsymbol{z}) = \frac{p(\boldsymbol{y}, \boldsymbol{z})}{p(\boldsymbol{z})} = \frac{p(\boldsymbol{y}, \boldsymbol{z})}{\int_{\boldsymbol{y}} p(\boldsymbol{y}, \boldsymbol{z}) \mathrm{d}\boldsymbol{y}}$$





- formally, **random variables (RVs)** are **measurable functions** on some (abstract) probability space
- **distribution functions** describe their distribution compactly
  - **probability mass functions (PMFs)** for **discrete RVs**
  - **probability density functions (PDFs)** for **continuous RVs**
- **central object in probabilistic ML: joint distributions** describing correlations of many RVs
- joints are subject to **probabilistic inference** with two core routines
  - **marginalization**
  - **conditioning**