Math primer for the Neural Networks course

Roberto Esposito

roberto.esposito@unito.it

Image dreamed by stable diffusion

Prompt: "curved lines with dices in a colored steampunk world, render 3d, dark colors, gray, red"

Probability theory knowledge poll

Probability

Probability can be seen as the extension of logic to deal with uncertainty.

Logic provides a set of formal rules for determining what propositions are implied to be true or false given the assumption that some other set of propositions is true or false.

Probability theory provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions.

Random Variables

A random variable is a variable that can take on different values randomly.

Example: the outcome of a coin flip, the number of heads in a sequence of coin flips, the result of a dice roll, the number of sixes in a sequence of dice rolls.

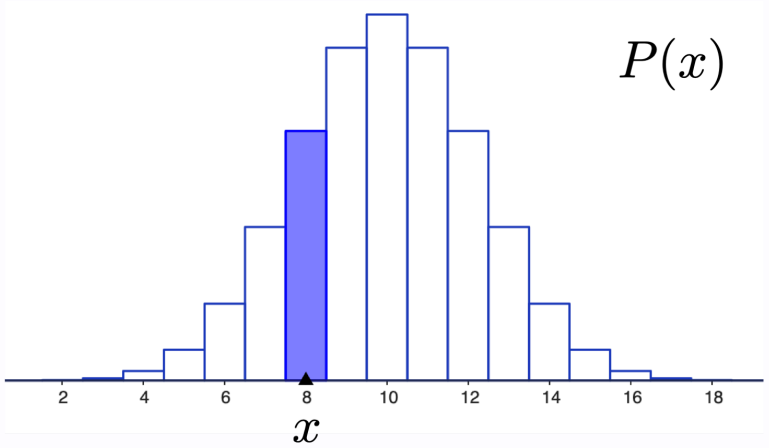
Random variables may be **discrete** or **continuous**.

Notation:
A random variable is denoted in plain typeface, e.g., $\mathrm{x}, \mathrm{y}, \ldots$
Values taken by the variables are typeset in lowercase script letters, e.g., $x, y, \ldots$
The set of all possible values taken by a random variable x is denoted Ω_{x}.
Sometimes we write $x \in \mathrm{x}$ as a shorthand for $x \in \Omega_{\mathrm{x}}$.
For vector-valued variables and their values, boldface characters are used: $\mathbf{x}, \mathbf{y}$ and $\boldsymbol{x}, \boldsymbol{y}, \ldots$.

Probability Distribution (Discrete case)

A probability distribution over *discrete variables* is described using a **probability mass function** (PMF). A PMF is denoted using a capital P as in $P(\mathbf{x})$.

The PMF *maps from a state of a random variable to the probability of that random variable taking on that state*.



Notation

Probability that $P(\mathbf{x} = x)$ is usually denoted as $P(x)$.
The notation $\mathbf{x} \sim P$ is used to state that the random variable $\mathbf{x}$ follows the $P(\mathbf{x})$ distribution.

Properties of a PMF

To be a PMF on a random variablex, a function P must satisfy the following properties:

- The domain of P must be the set of all possible states of $\mathbf{x}$
- $\forall x \in \mathbf{x} : 0 \leq P(x) \leq 1$
- $\sum_{x \in \mathbf{x}} P(x) = 1$

Other properties:

- $P(S) = \sum_{x \in S} P(x)$
- $P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2)$
- $P(\Omega \setminus S) = 1 - P(S)$

with S, S_1, S_2 being sets of possible outcomes; $P(S)$ being a shorthand for $P(\mathbf{x} \in S)$; and Ω being the set of all possible values.

Note: these are not to be confused with the axioms of probabilities. Here we are simplifying the discussion and only report notable properties.

Example

If we consider a random variable $\mathbf{x}$ taking k possible values $x_1, \ldots, x_k$, the PMF $P(\mathbf{x}) = \frac{1}{k}$ is a valid PMF:

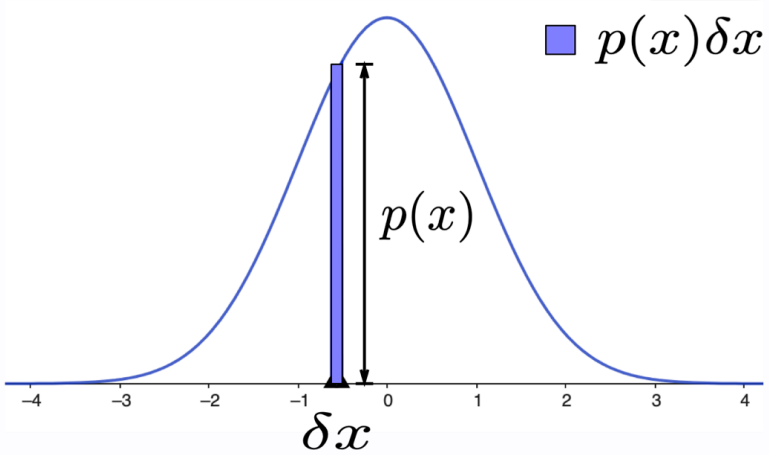
- it is defined over all possible states of $\mathbf{x}$
- $\forall x \in \mathbf{x} : 0 \leq P(x) = \frac{1}{k} \leq 1$
- $P(\Omega) = \sum_{x \in \{x_1, \ldots, x_k\}} P(x) = \sum_{x \in \{x_1, \ldots, x_k\}} \frac{1}{k} = 1$

Thanks to the other properties we can evaluate:

- $P(\{x_1, x_2, x_3\} \cup \{x_4, x_5\}) = P(\{x_1, x_2, x_3\}) + P(\{x_4, x5\}) = \frac{3}{k} + \frac{2}{k} = \frac{5}{k}$
- $P(\Omega \setminus \{x_1, x_2\}) = 1 - \frac{2}{k} = \frac{k-2}{k} = P(\{x_3, \ldots, x_k\})$

Probability distributions (Continuous case)

When working with continuous random variables, we describe probability distributions using a **probability density function** (PDF) rather than a probability mass function.



The PDF is usually denoted as p and must satisfy the following properties:

- the domain of p must be the set of all possible states of $\mathbf{x}$;
- $\forall x \in \mathbf{x} : p(x) \geq 0$, note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$.

Example

Consider a uniform distribution on an interval of the real numbers: $\mathbf{x} \sim U(a, b)$, where a and b are the endpoints of the interval, with $b > a$. The PDF $u(x; a, b)$ of the uniform distribution is:

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

Note:

- the domain is the set of all possible values of $\mathbf{x}$;
- $\forall x \in \mathbf{x} : p(x) \geq 0$;
- $\int_{-\infty}^{\infty} u(x; a, b)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b 1dx = \frac{1}{b-a} \cdot \left(x\Big|_a^b\right)$ = 1

Marginal Probability

Sometimes we know the probability distribution over a set of variables and we want to know the probability distribution over just a subset of them. The probability distribution over the subset is known as the **marginal probability** distribution.

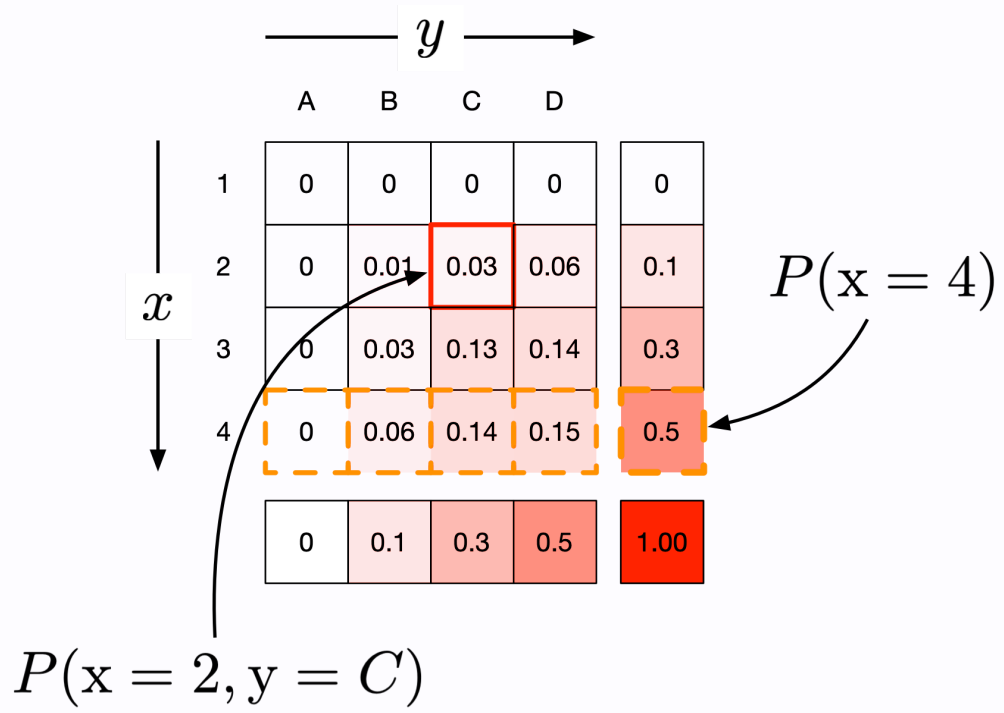
Marginal probabilities are computed by summing over all values *other* variables.

For example, suppose we have discrete random variables x and y, with joint distribution $P(\mathrm{x}, \mathrm{y})$. The marginal distribution $P(\mathrm{x})$ is:

$$\forall x \in \mathrm{x} : P(x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y);$$

for continuous variables:

$$p(x) = \int p(x, y) dy.$$

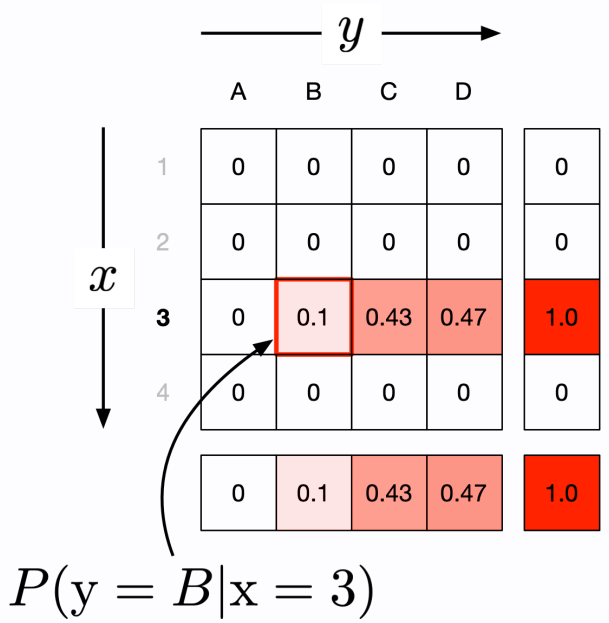


Conditional Probability

In many cases, we are interested in the probability of some event, given that some other event has happened. This is called a conditional probability. We denote the conditional probability that $\mathrm{y} = y$ given $\mathrm{x} = x$ as $P(\mathrm{y} = y | \mathrm{x} = x)$.

This conditional probability can be computed with the formula:

$$P(\mathrm{y} = y | \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$



The Chain Rule of Conditional Probabilities

Any joint probability distribution over many random variables may be decomposed into products of conditional distributions over only one variable:

$$P(\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(n)}) = P(\mathbf{x}^{(1)}) \prod_{i=2}^{n} P(\mathbf{x}^{(i)} | \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)})$$

Example

Note:

- $P(a|b,c) = \frac{P(a,b,c)}{P(b,c)} \Rightarrow P(a,b,c) = P(a|b,c)P(b,c)$
- *similarly we have that* $P(b,c) = P(b|c)P(c)$

Implying:

$$\begin{aligned} P(a,b,c) &= P(a|b,c)P(b,c) \\ &= P(a|b,c)P(b|c)P(c) \end{aligned}$$

Independence

Two random variables $\mathbf{x}$ and $\mathbf{y}$ are **independent** (we write $\mathbf{x} \perp \mathbf{y}$) if their probability distribution can be expressed as a product of two factors, one involving only $\mathbf{x}$ and one involving only $\mathbf{y}$:

$$\forall x \in \mathbf{x}, y \in \mathbf{y} : p(\mathbf{x} = x, \mathbf{y} = y) = p(\mathbf{x} = x)p(\mathbf{y} = y)$$

Note:

$$\mathbf{x} \perp \mathbf{y} \iff \forall x \in \mathbf{x}, y \in \mathbf{y} : \quad p(x|y) = p(x) \quad \wedge \quad p(y|x) = p(y).$$

Two random variables $\mathbf{x}$ and $\mathbf{y}$ are **conditionally independent** given a random variable $\mathbf{z}$ (we write $\mathbf{x} \perp \mathbf{y} | \mathbf{z}$) if the conditional probability distribution over $\mathbf{x}$ and $\mathbf{y}$ factorizes in this way for every value of $\mathbf{z}$:

$$\forall x \in \mathbf{x}, y \in \mathbf{y}, z \in \mathbf{z} : p(\mathbf{x} = x, \mathbf{y} = y | \mathbf{z} = z) = p(\mathbf{x} = x | \mathbf{z} = z)p(\mathbf{y} = y | \mathbf{z} = z)$$

Expectation

The expectation or expected value (often denoted μ) of some function $f(x)$ with respect to a probability distribution $P(\mathbf{x})$ is the average or mean value that f takes on when x is drawn from P.

Discrete variables

$$\mathbb{E}_{\mathbf{x}\sim P}[f(x)] = \sum_x P(x)f(x)$$

Continuous variables

$$\mathbb{E}_{\mathbf{x}\sim p}[f(x)] = \int p(x)f(x)dx$$

Note

The expectation is a linear operator: $\mathbb{E}[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}[f(x)] + \beta\mathbb{E}[g(x)]$.

Variance

The variance (often denoted σ^2) gives a measure of how much the values of a function of a random variable $\mathbf{x}$ vary as we sample different values of $\mathbf{x}$ from its probability distribution:

$$\mathbf{Var}[f(\mathbf{x})] = \mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])^2]$$

Example

Variance can be also computed as: $\mathbf{Var}[f(\mathbf{x})] = \mathbb{E}[f(\mathbf{x})^2] - \mathbb{E}[f(\mathbf{x})]^2$

$$\begin{aligned}
\mathbf{Var}[f(\mathbf{x})] &= \mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])^2] \\
&= \mathbb{E}[f(\mathbf{x})^2 - 2f(\mathbf{x})\mathbb{E}[f(\mathbf{x})] + \mathbb{E}[f(\mathbf{x})]^2] \\
&= \mathbb{E}[f(\mathbf{x})^2] - 2\mathbb{E}[f(\mathbf{x})\mathbb{E}[f(\mathbf{x})]] + \mathbb{E}[\mathbb{E}[f(\mathbf{x})]^2] \\
&= \mathbb{E}[f(\mathbf{x})^2] - 2\mathbb{E}[f(\mathbf{x})]\mathbb{E}[f(\mathbf{x})] + \mathbb{E}[f(\mathbf{x})]^2 \\
&= \mathbb{E}[f(\mathbf{x})^2] - \mathbb{E}[f(\mathbf{x})]^2.
\end{aligned}$$

Standard deviation and Covariance

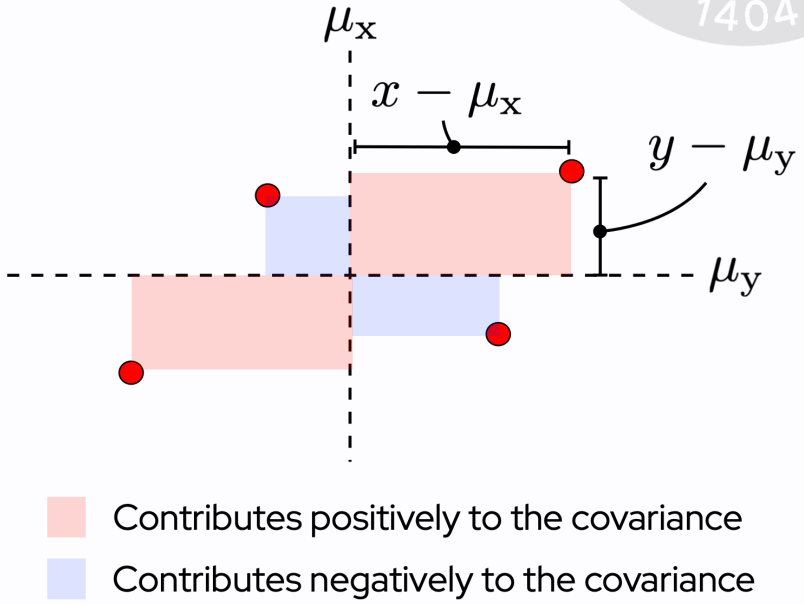
The **standard deviation** (denoted σ) is given by the square root of the variance.

Useful to know

for a normally distributed variable, about 95% of the points fall into the range $\mu \pm 2\sigma$.

The **covariance** gives some sense of how much two random variables are related to each other:

$$\begin{aligned}\mathrm{Cov}(\mathrm{x},\mathrm{y}) &= \mathbb{E}_{x,y\sim P(\mathrm{x},\mathrm{y})}[(x-\mathbb{E}[\mathrm{x}])(y-\mathbb{E}[\mathrm{y}])] \\ &= \mathbb{E}_{x,y\sim P(\mathrm{x},\mathrm{y})}[(x-\mu_{\mathrm{x}})(y-\mu_{\mathrm{y}})]\end{aligned}$$



Note: Covariance is affected by the scale of the variables.

Correlation adjusts for the scale of each variable, ensuring that the relationship between the variables is measured without being influenced by their individual magnitudes.

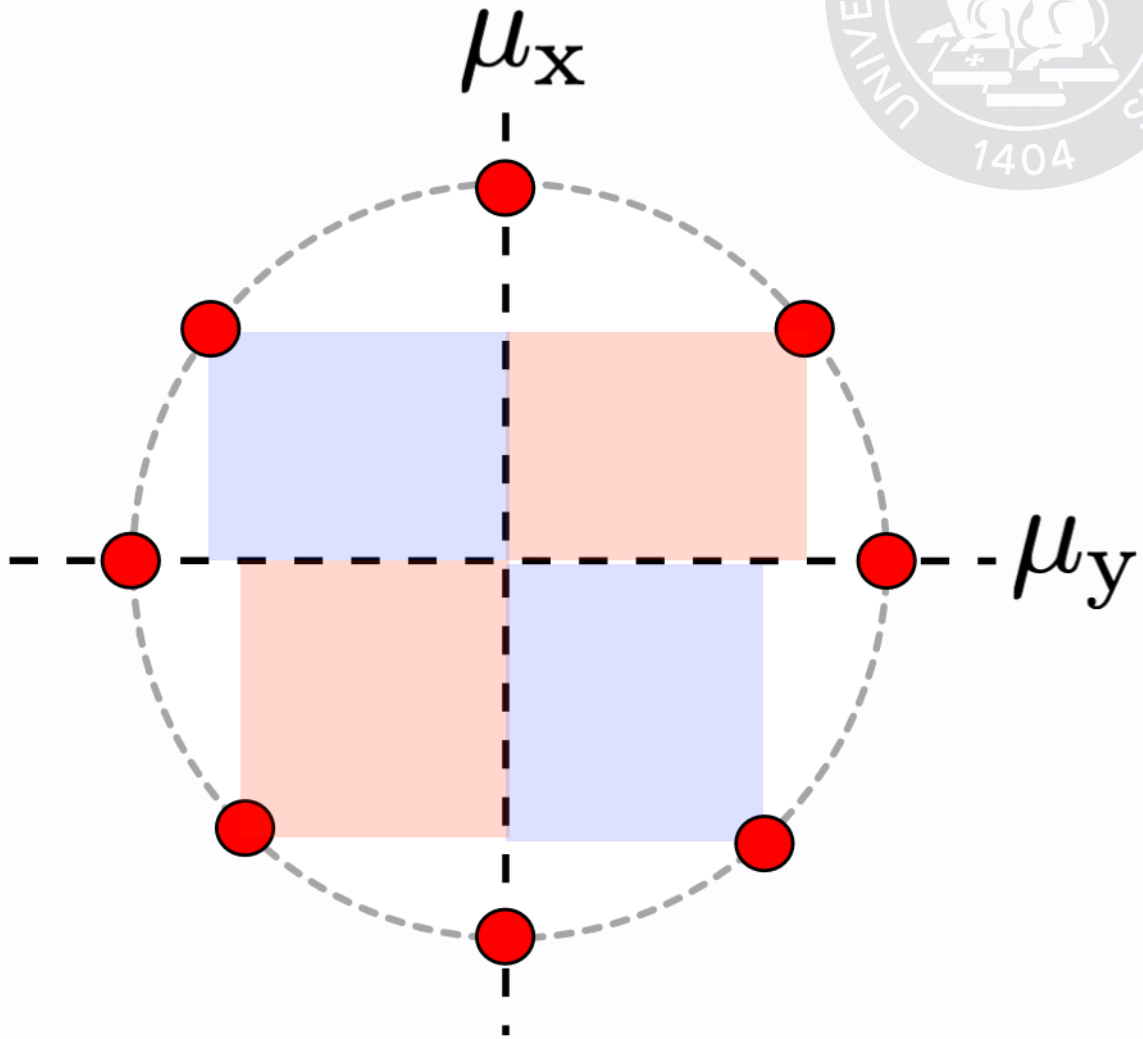
$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

A correlation close to ± 1 indicates a strong relationship between the variables, while a correlation near 0 suggests that the variables may be independent.

Covariance vs Dependence

The notions of **covariance** and **dependence** are related, but are in fact **distinct** concepts.

- two variables that are independent have zero covariance;
- two variables that have non-zero covariance are dependent;
- two variables can have zero covariance and be dependent nonetheless (as shown in the img).



Common Probability Distributions

Bernoulli distribution

The Bernoulli distribution is a distribution over a single **binary** random variable. It is controlled by a single parameter $\phi \in [0, 1]$, which gives the probability of the random variable being equal to 1. It has the following properties:

$$
\begin{aligned}
P(\mathrm{x} = 1) &= \phi \\
P(\mathrm{x} = 0) &= 1 - \phi \\
P(\mathrm{x} = x) &= \phi^{x}(1 - \phi)^{1-x} \\
\mathbb{E}[\mathrm{x}] &= \phi \\
Var(\mathrm{x}) &= \phi(1 - \phi)
\end{aligned}
$$

Multinulli distribution

The **multinoulli** or **categorical** distribution is a distribution over a single discrete variable with k different states, where k is finite.

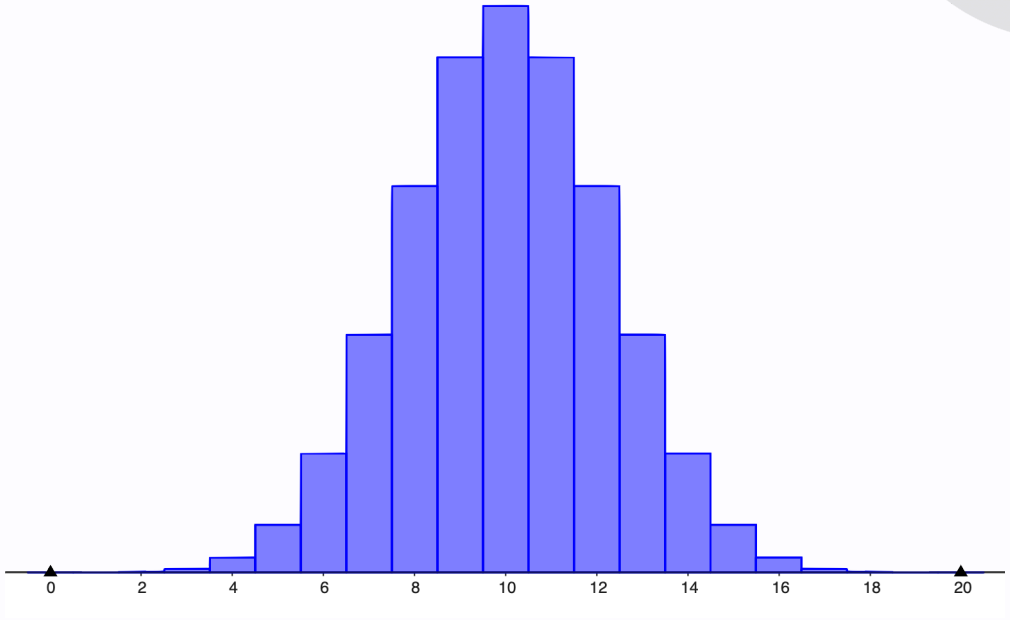
It is parametrized by a vector $\mathbf{p} \in [0, 1]^k$, with $\mathbf{1}^\top \mathbf{p} = 1$ where p_i gives the probability of the i-th state.

Multinoulli distributions are often used to refer to distributions over categories of objects, so we do not usually assume that state 1 has numerical value 1, etc. For this reason, **we do not usually need to compute the expectation or variance** of multinoulli-distributed random variables.

Binomial Distribution

The binomial distribution gives **the probability of observing a given number of successes in a repeated Bernoulli experiment**. It is parametrized by

- p: the probability of successes of the Bernoulli experiment,
- and N: the number of total repetitions of the Bernoulli experiment.



If $\mathrm{x} \sim \mathrm{Bi}(p, N)$, then:

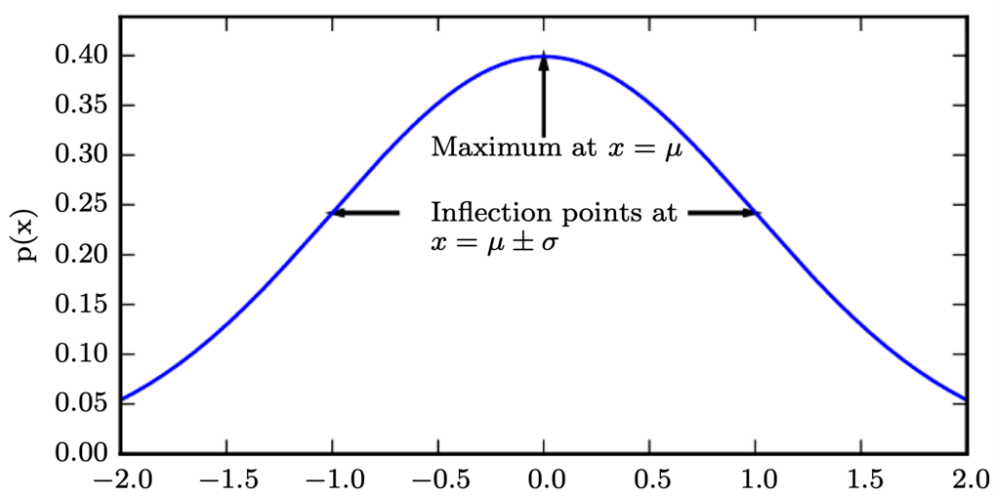
$$P(\mathrm{x} = k) = \binom{N}{k} p^k (1-p)^{N-k} \qquad \mathbb{E}[\mathrm{x}] = Np \qquad \mathrm{Var}[\mathrm{x}] = Np(1-p)$$

Gaussian Distribution

The most commonly used distribution over real numbers is the **normal distribution**, also known as the **Gaussian distribution**.

A Gaussian distribution is parametrized by a mean μ and a variance σ^2. If $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, then the probability density function (PDF) of $\mathbf{x}$ is given by:

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



Gaussian Distribution

Many distributions we wish to model are truly close to being normal distributions.

Central Limit Theorem

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables with finite mean μ and variance σ^2. Then the distribution of $S = \sum_{i=1}^{n} X_i$ approaches a normal distribution $\mathcal{N}(n\mu, n\sigma^2)$ as n approaches infinity. The average $\frac{1}{n}S$ also approaches a normal distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ as n approaches infinity.

Out of all possible probability distributions with the same mean and variance, the normal distribution **encodes the maximum amount of uncertainty** (i.e., it is the distribution having the maximum entropy).

Normal

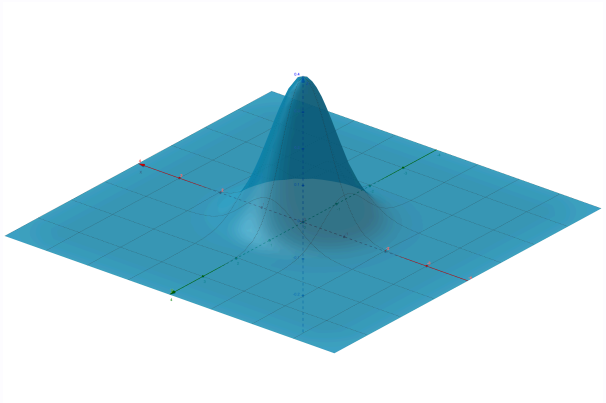
The normal distribution generalizes to $\mathbb{R}^n$:

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- $\boldsymbol{\mu}$: a vector denoting the mean of the distribution;
- $\boldsymbol{\Sigma}$: the covariance matrix of the distribution.

Covariance matrices are **symmetric** and **positive semi-definite** and their main diagonal contains variances.

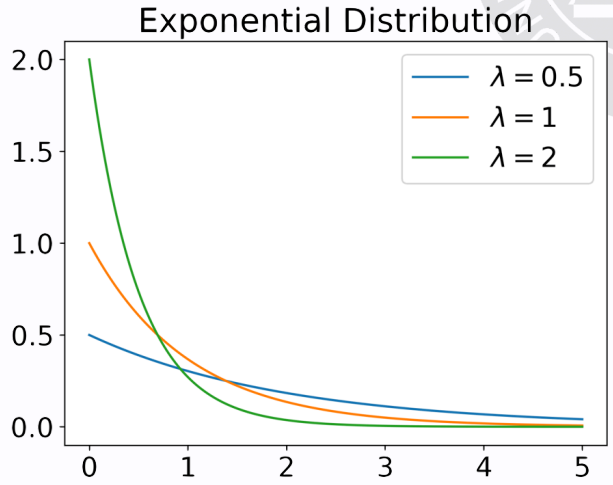
If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, the variance is the same in every direction: the distribution is said to be **isotropic**.



Exponential and Laplace distributions

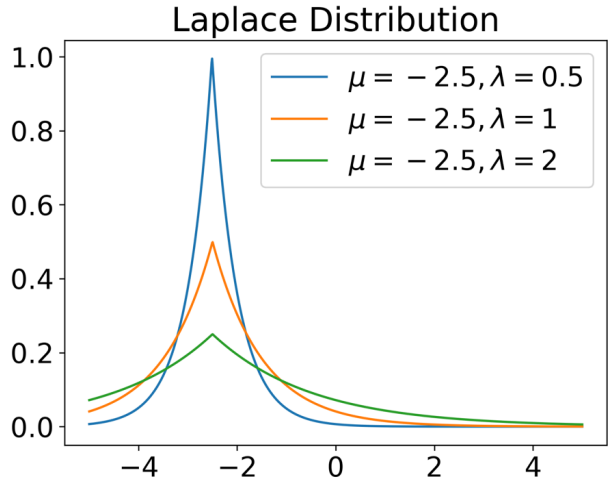
In the context of deep learning, we often want to have a probability distribution with a sharp point at $x = 0$. To accomplish this, we can use the **exponential distribution**:

$$p(x; \lambda) = \lambda \exp(-\lambda x), \quad x \geq 0$$



A closely related probability distribution that allows us to place a sharp peak of probability mass at an arbitrary point μ is the **Laplace distribution**:

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right)$$

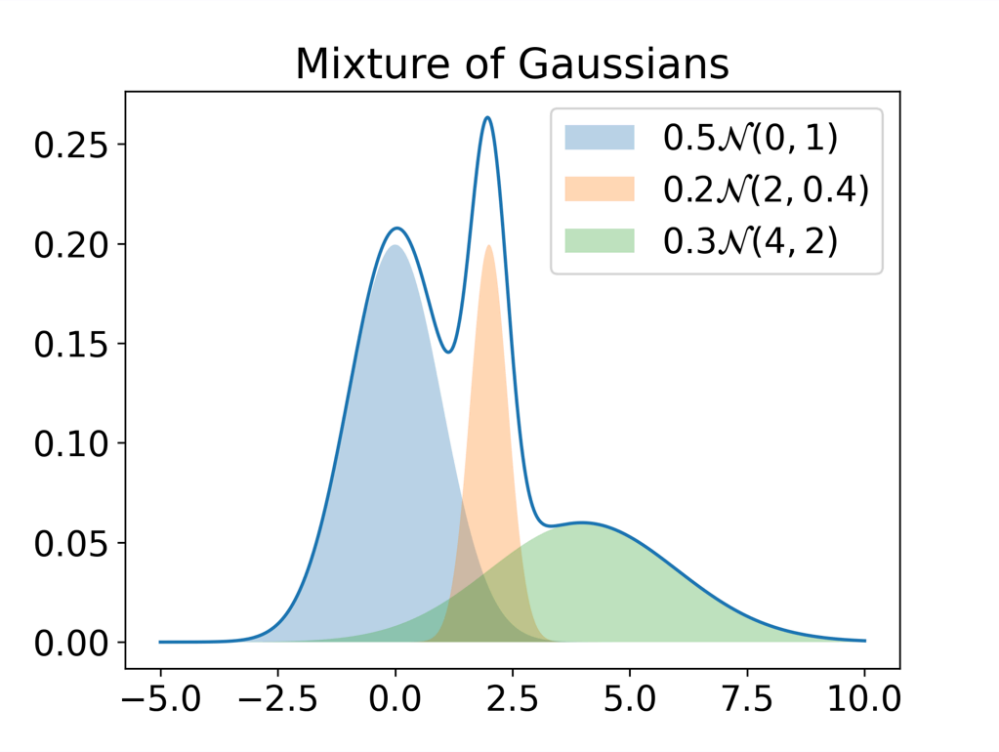


Mixtures of Distributions

It is also common to define probability distributions by combining other simpler probability distributions. One common way of combining distributions is to construct a **mixture distribution**:

$$P(\mathbf{x}) = \sum_i P(\mathbf{c} = i) P(\mathbf{x} | \mathbf{c} = i)$$

where $P(\mathbf{c})$ is a multinulli distribution over the components of the mixture.



Bayes' Rule

We often find ourselves in a situation where we know $P(\mathrm{y}|\mathrm{x})$ and need to know $P(\mathrm{x}|\mathrm{y})$.

Fortunately, if we also know $P(\mathrm{x})$ and $P(\mathrm{y})$, we can compute the desired quantity using Bayes' rule:

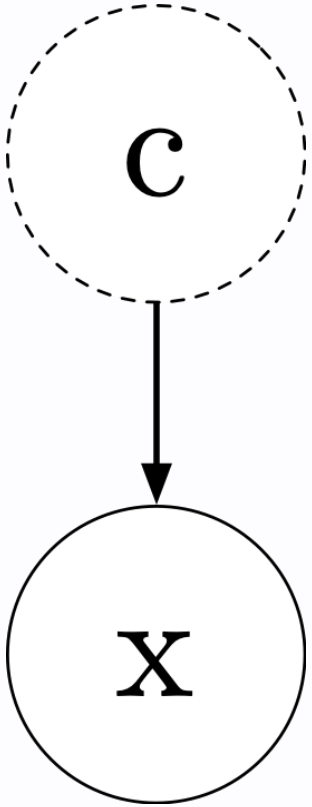
$$P(\mathrm{x}|\mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y}|\mathrm{x})}{P(\mathrm{y})}$$

Latent Variables

The mixture model allows us to briefly glimpse a concept that will be of paramount importance later — the latent variable.

A **latent variable** is a random variable that **we cannot observe directly**. The component identity variable $\mathbf{c}$ of the mixture model provides an example.

Latent variables may be related to $\mathbf{x}$ through the joint distribution, in this case, $P(\mathbf{x}, \mathbf{c}) = P(\mathbf{x}|\mathbf{c})P(\mathbf{c})$.

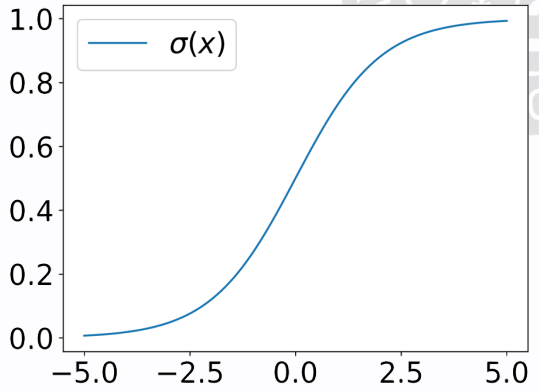


Useful Properties of Common Functions

Sigmoid

Often used to produce the ϕ parameter of a Bernoulli distribution. It is denoted with σ and defined as:

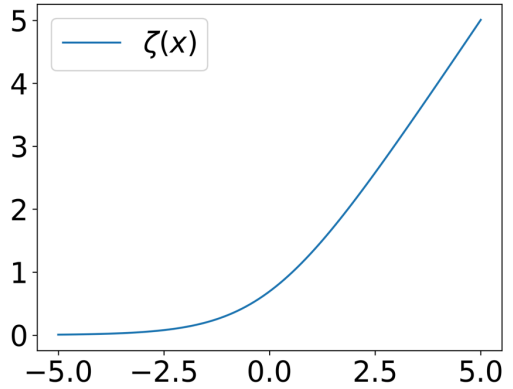
$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$



Softplus

A smoothed version of $x^+ = \max(0, x)$, it is denoted with ζ and defined as:

$$\zeta(x) = \log(1 + \exp(x))$$



Useful Properties of Common Functions

- The sigmoid function **saturates** for large (negative/positive) values of x.
- $\sigma(x) = \frac{\exp(x)}{\exp(x)+\exp(0)}$
- $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$
- $1-\sigma(x) = \sigma(-x)$
- $\log\sigma(x) = -\zeta(-x)$
- $\frac{d}{dx}\zeta(x) = \sigma(x)$
- $\forall x \in (0,1) : \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$
- $\forall x > 0 : \zeta^{-1}(x) = \log(\exp(x)-1)$
- $\zeta(x) = \int_{-\infty}^{x}\sigma(y)dy$
- $\zeta(x) - \zeta(-x) = x$

Technical Details of Continuous Variables

A proper formal understanding of continuous random variables and probability density functions requires developing probability theory through a branch of mathematics known as **measure theory**.

Without measure theory, one might encounter paradoxical situations.

Example

It is possible to construct two sets S_1 and S_2, where $S_1 \cap S_2 = \emptyset$, such that $p(x \in S_1) + p(x \in S_2) > 1$.

These paradoxes usually involve constructing very *exotic* sets, such as fractal-like sets or sets derived from transformations of rational numbers, but the possibility exists.

One of the key contributions of measure theory is that it provides a framework to characterize sets where probabilities can be consistently computed, thus avoiding paradoxes.

Measure theory provides a rigorous way of describing that a set of points is **negligibly small**. In such cases we say that the set has **measure zero**.

Example

A line in $\mathbb{R}^2$ has measure zero.

Any **union of countably many sets** having measure zero has measure zero (so the set of all rational numbers has measure zero).

When a property holds throughout all of space except for points in a set of measure zero, we say that the property holds **almost anywhere**.

Technical details of Continuous Variables

Another technical detail of continuous variables relates to handling continuous random variables that are **deterministic functions of one another**.

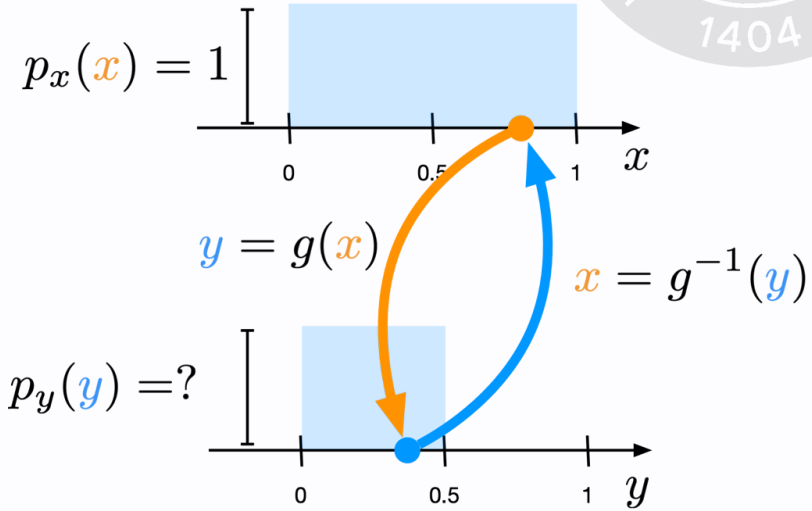
Suppose we have two random variables, $\mathbf{x}$ and $\mathbf{y}$, such that $\boldsymbol{y} = g(\boldsymbol{x})$, where g is an invertible, continuous differentiable transformation.

Unfortunately: $p_{\mathbf{y}}(\boldsymbol{y}) \neq p_{\mathbf{x}}(g^{-1}(\boldsymbol{y}))$

Example

Let $\mathrm{y} = \frac{\mathrm{x}}{2}$ and $\mathrm{x} \sim U(0,1)$. In this case we have:

- $y = g(x) = \frac{x}{2}$,
- $x = g^{-1}(y) = 2y$



One would expect that $p_y(y) = p_x(2y)$, but this is not the case. In fact, if it was so, $p_y(y)$ would be zero everywhere except in the interval $\left[0, \frac{1}{2}\right]$ and it would be 1 in such interval. But then:

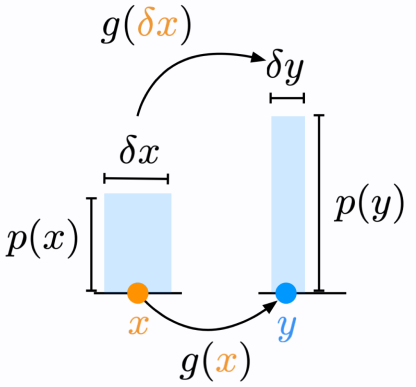
$$\int_{-\infty}^{\infty} p_y(y)dy = \int_{0}^{0.5} p_y(y)dy = \int_{0}^{0.5} 1dy = x\Big|_0^{0.5} = 0.5 - 0 = 0.5$$

Technical details of Continuous Variables

The problem with this approach is that it fails to account for the distortion of space introduced by the function g.

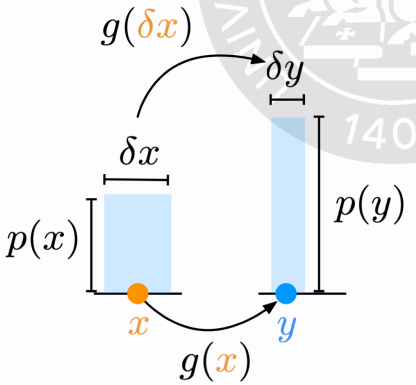
The probability of x lying in an infinitesimally small region with volume δx is given by $p(x)\delta x$.

Since g can expand or contract space, the infinitesimal volume surrounding x in x space may have different volume in y space.



To correct the problem we need to preserve the property:

$$|p_y(g(x))dy| = |p_x(x)dx|$$



which yields

$$p_y(g(x))|dy| = p_x(x)|dx| \quad \Rightarrow \quad p_x(x) = p_y(g(x)) \left| \frac{d}{dx} g(x) \right|$$

or, equivalently:

$$p_y(y)|dy| = p_x(g^{-1}(y))|dx| \quad \Rightarrow \quad p_y(y) = p_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

Technical details of Continuous Variables

In our example:

$$p_y(y) = p_x(g^{-1}(y)) \left| \frac{\partial g^{-1}(y)}{\partial y} \right| = p_x(2y) \left| \frac{d}{dy} 2y \right| = 1 \cdot 2 = 2$$

and $\int_0^{0.5} p_y(y)dy = \int_0^{0.5} 2dy$ = 1.

In **higher dimensions** $g : \mathbb{R}^m \to \mathbb{R}^n$, the derivative generalizes to the Jacobian matrix and the absolute value to the absolute value of the determinant:

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \left| \det\left(J\right) \right|$$

where the Jacobian J is such that $J_{ij} = \frac{\partial g(\boldsymbol{x})_i}{\partial x_j}$

Probability theory knowledge poll