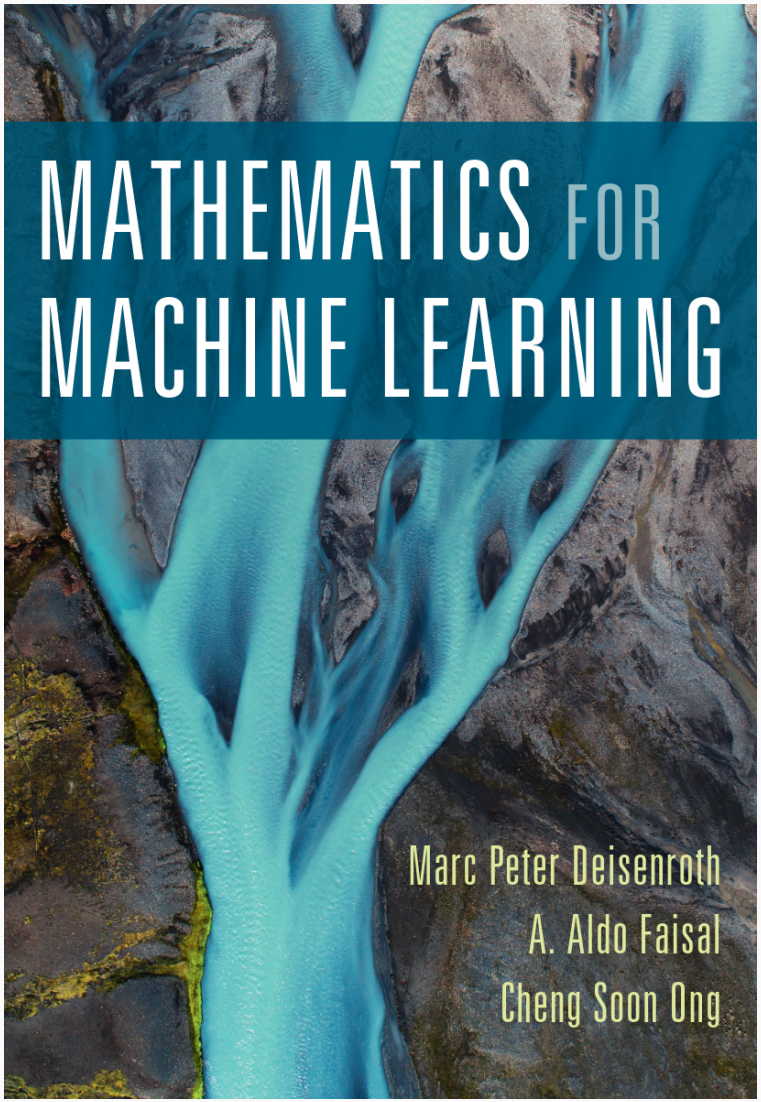
Mathematical Basics

Machine Learning 1 — Lecture 3

21st March 2023

Robert Peharz
Institute of Theoretical Computer Science
Graz University of Technology

Reading Material



I recommend this excellent book, either for further reading or if you need to catch up.

https://mml-book.github.io/

Core Topics

The core mathematical disciplines in Machine Learning are:

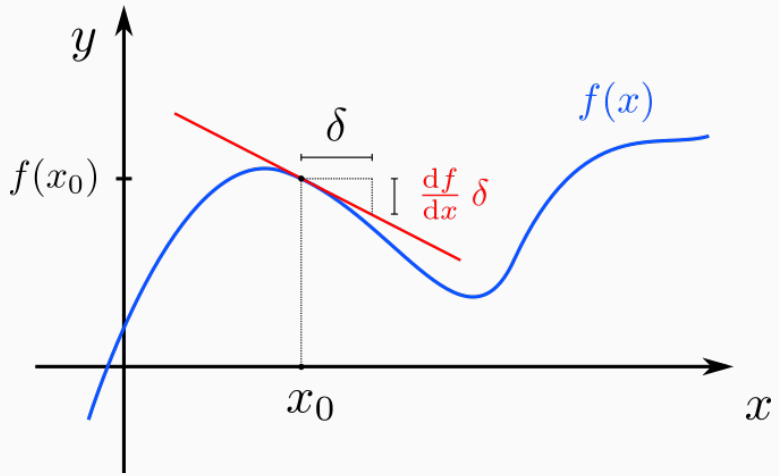
- **Linear algebra**
- **Calculus**
- **Probability**

Differential Calculus

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a univariate function. If it exists, the **derivative** at a point x is defined as

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- rate of change
- slope of the tangent line at some point x_0, i.e. best local linear approximation: $f(x) \approx f_{linear}(x) = \underbrace{f(x_0)}_{b} + \underbrace{f'(x_0)}_{a}\underbrace{(x - x_0)}_{\delta}$



Standard Rules for Derivatives

Let f and g be differentiable univariate functions and a, b arbitrary constants.

- **Derivative is linear:**

$$(af + bg)'(x) = af'(x) + bg'(x)$$

$$\left(\frac{\mathrm{d}(af + bg)}{\mathrm{d}x} = a\frac{\mathrm{d}f}{\mathrm{d}x} + b\frac{\mathrm{d}g}{\mathrm{d}x} \right)$$

- **Product rule:**

$$(fg)'(x) = f'g(x) + fg'(x)$$

$$\left(\frac{\mathrm{d}fg}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}x}g + \frac{\mathrm{d}g}{\mathrm{d}x}f \right)$$

- **Chain rule:**

$$f(g(x))' = f'(g(x))g'(x)$$

$$\left(\frac{\mathrm{d}f \circ g}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}x} \right)$$

Note: $f \circ g$ denotes function composition: $(f \circ g)(x) = f(g(x))$

Derivatives of some well-known functions

Name	$f(x)$	$f'(x)$
Constant	a	0
Affine	$ax + b$	a
Polynomial	x^k	$k\,x^{k-1}$
Exponential	$\exp(x),\ e^x$	$\exp(x),\ e^x$
	b^x	$b^x \log b$
Logarithm	$\log x$	$\frac{1}{x}$
Sine	$\sin(x)$	$\cos(x)$
Cosine	$\cos(x)$	$-sin(x)$

Note: log denotes the **natural logarithm** in this course.

So, far we considered derivatives **uni-variate** functions $y = f(x)$, i.e. with a single input.

Let's generalize to **multi-variate** functions, i.e. functions with multiple inputs — leading to **partial derivatives**.

Let $f : \mathbb{R}^D \mapsto \mathbb{R}$ be a multi-variate scalar function.

If it exists, the **partial derivative** with respect to x_d is the **standard derivative** with respect to x_d, when **keeping all other arguments constant.**

Specifically, let $\boldsymbol{x} = (x_1, \ldots, x_D)^T \in \mathbb{R}^D$ be some vector. Then the partial derivative with respect to x_d at $\boldsymbol{x}$ is

$$\frac{\partial f}{\partial x_d}(\boldsymbol{x}) = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{d-1}, x_d + h, x_{d+1} \ldots, x_D) - f(\boldsymbol{x})}{h}$$

Let $f(x, y) = sin(xy)\, log(x^2)$.

$\frac{\partial f}{\partial x}$ is found by deriving after x, keeping y constant:

$$\frac{\partial f}{\partial x} = \overbrace{cos(xy)\, y}^{\partial sin(xy)/\partial x}\, log(x^2) + sin(xy)\, \overbrace{\frac{1}{x^2}\, 2x}^{\partial \log(x^2)/\partial x}$$

$\frac{\partial f}{\partial y}$ is found by deriving after y, keeping x constant:

$$\frac{\partial f}{\partial y} = \overbrace{cos(xy)\, x}^{\partial sin(xy)/\partial y}\, log(x^2)$$

Let $f : \mathbb{R}^D \mapsto \mathbb{R}$ be a multi-variate scalar function.

The **gradient** of f at some point $\boldsymbol{x} = (x_1, \ldots, x_D)$ is the vector containing all partial derivatives:

$$\nabla_{\boldsymbol{x}} f = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \frac{\partial f}{\partial x_2}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_D}(\boldsymbol{x}) \end{pmatrix} \qquad \text{speak “nabla } \boldsymbol{x} \text{ of f”}$$

- The gradient shows in the **direction of steepest ascend**
- Its magnitude $\|\nabla_{\boldsymbol{x}} f\|_2$ is the **rate of change (slope)**
- At **maxima, minima and saddle points:** $\nabla_{\boldsymbol{x}} f = 0$
- Specifies **best local linear approximation** at any point $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) \approx f_{linear}(\boldsymbol{x}) = \underbrace{f(\boldsymbol{x}_0)}_{\boldsymbol{b}} + \underbrace{\nabla_{\boldsymbol{x}_0} f^T}_{\boldsymbol{w}^T} \underbrace{(\boldsymbol{x} - \boldsymbol{x}_0)}_{\boldsymbol{\delta}}$$

As before, let $f(x, y) = sin(xy)\, log(x^2)$.

We have found that

$$\frac{\partial f}{\partial x} = cos(xy)\, y\, log(x^2) + sin(xy)\, \frac{1}{x^2}\, 2x$$
$$\frac{\partial f}{\partial y} = cos(xy)\, x\, log(x^2)$$

Thus, the gradient at any point $\boldsymbol{x} = (x, y)^T$ is

$$\nabla_{\boldsymbol{x}} f = \begin{pmatrix} cos(xy)\, y\, log(x^2) + sin(xy)\, \frac{1}{x^2}\, 2x \\ cos(xy)\, x\, log(x^2) \end{pmatrix}$$

$$f(x, y) = -2(\cos^2(x) + \cos^2(y))^2$$

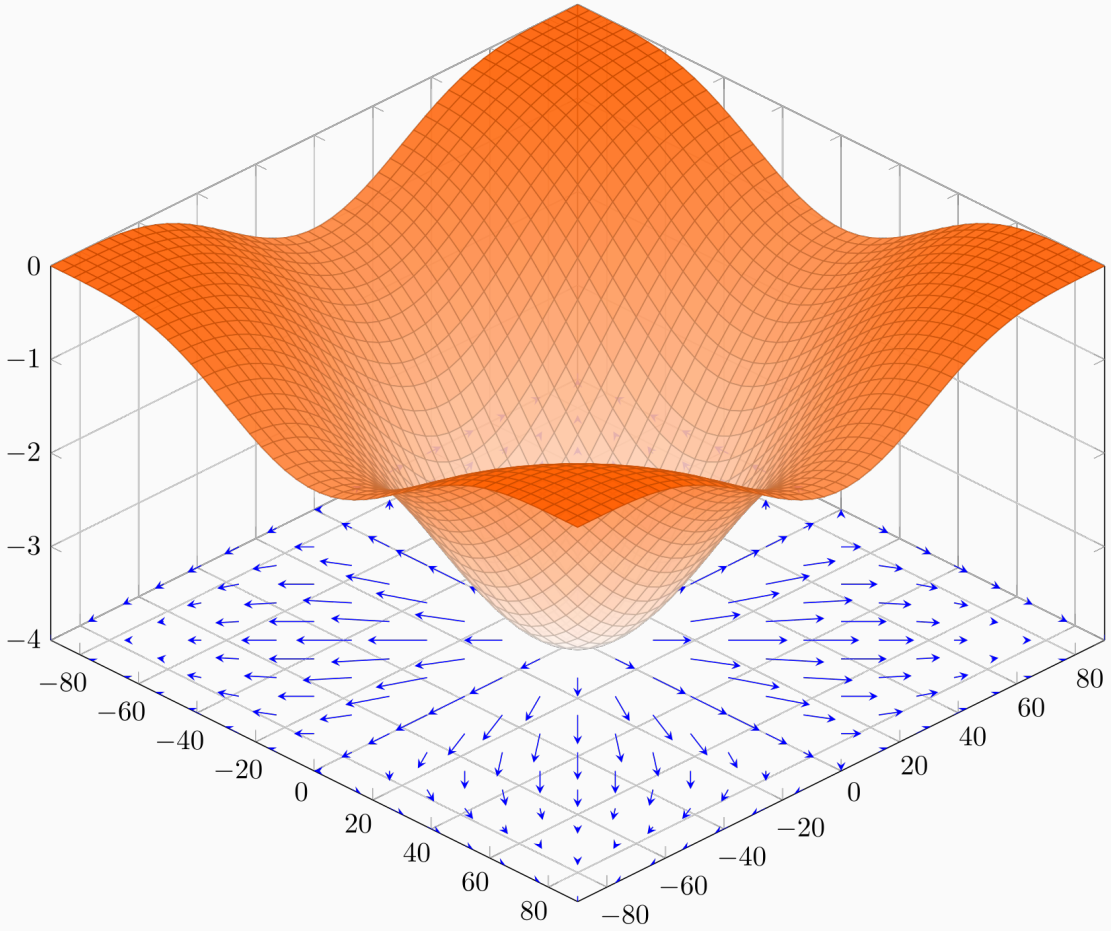
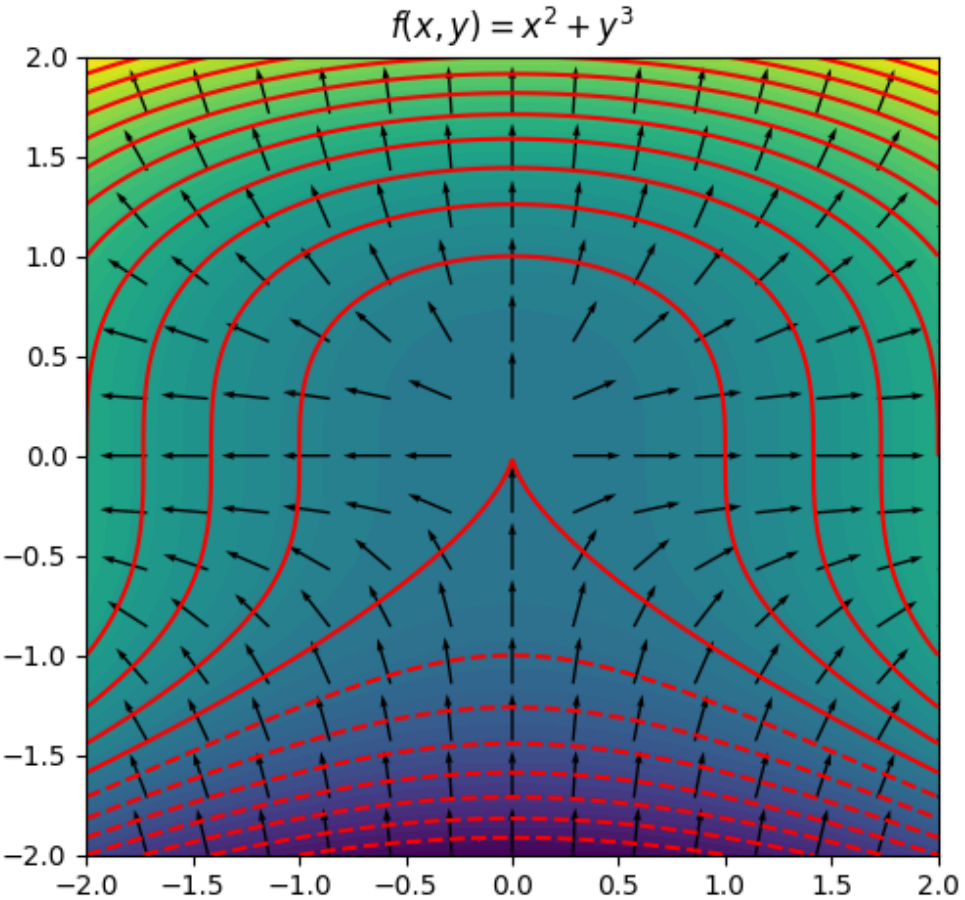


Image: wikipedia

Gradients are orthogonal to the contour sets of the function (contour set: set of points where function has the same value):



Optimization via Gradient Descent

Optimization [see also "Optimization for CS"]

An **optimization problem** in "standard form" is written as

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \textbf{s.t.} \quad & g_i(\boldsymbol{x}) = 0 \qquad i = 1 \dots N \\ & h_j(\boldsymbol{x}) \leq 0 \qquad j = 1 \dots M \end{aligned}$$

"minimize f of $\boldsymbol{x}$, subject to N equality and M inequality constraints."

$\boldsymbol{x}$: **optimization variables**

f: **objective**, **cost**, **loss function**

g_i: **equality constraint functions**

h_i: **inequality constraint functions**

Note: maximizing a function f is the same as minimizing $-f$.

Gradient Descent

- **Gradient Descent**: a general optimization technique
- start at some point $\boldsymbol{x}$
- take a small step downhill, i.e. in the direction of the **negative gradient** (**steepest descend**)
- repeat, until you are in a valley (**local minimum**)

Gradient Descent

Consider a minimization problem without constraints:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

Gradient descent algorithm:

- initialize $\boldsymbol{x}$ (randomly, heuristically)
- while $\|\nabla_{\boldsymbol{x}} f\| > \epsilon$:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \eta \nabla_{\boldsymbol{x}} f$$

η: **step-size**, **learning rate**

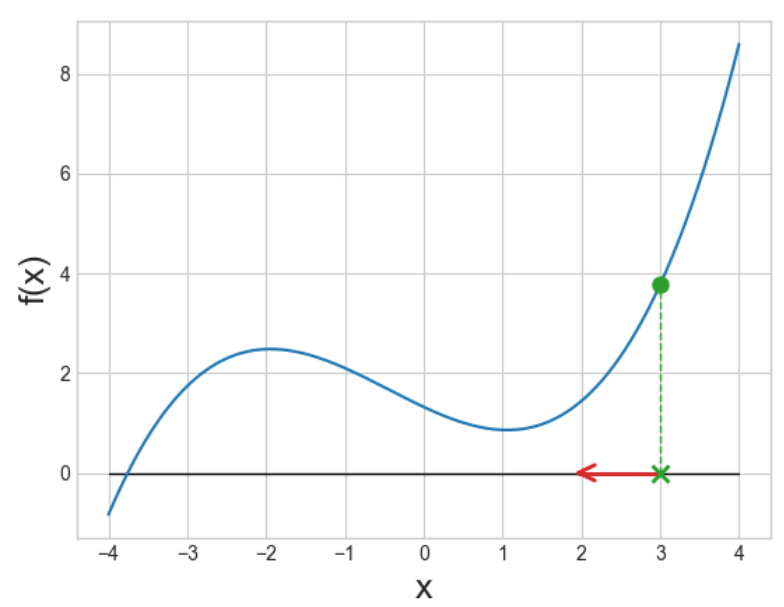
ϵ: small threshold, for deciding whether $\nabla_{\boldsymbol{x}} f \approx 0$

If step-size is sufficiently small, convergence to a **local minimum**.

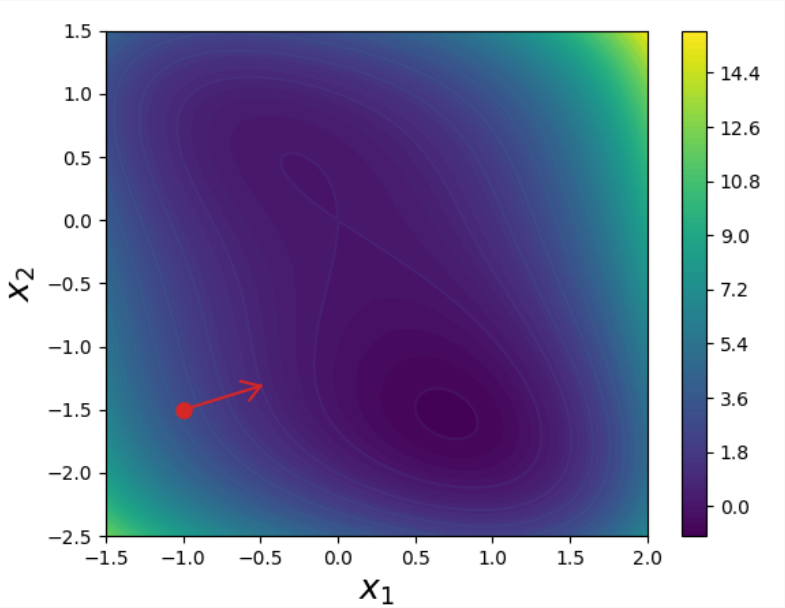
For **convex functions**, any local minimum is also a **global minimum**.

Gradient Descent

Example



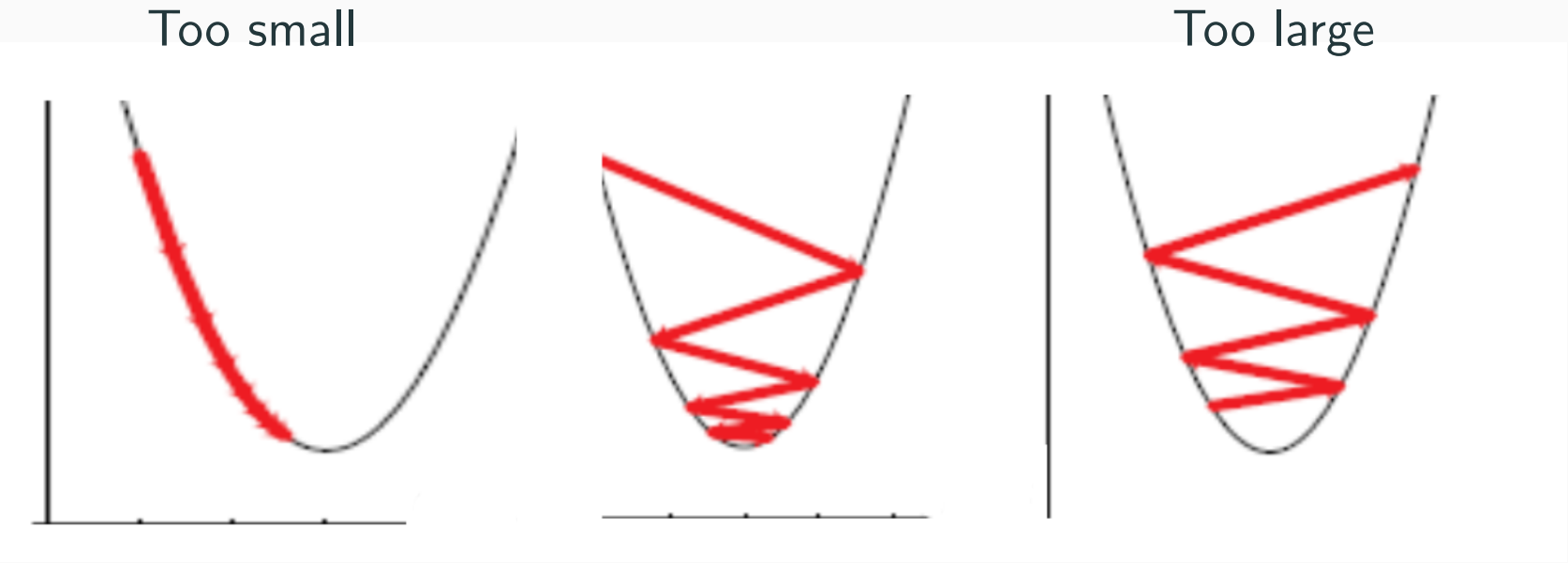
`gradient_descent_demo_1d.py`



`gradient_descent_demo_2d.py`

Role of the Step-Size

- Too small step-size η: slow convergence
- Too large step-size η: divergence, oscillation
- Largest "safe" step-size: depends on Lipschitz constant of f
- In practice, step-size is often set "experimentally"



`https://blog.skz.dev/gradient-descent`

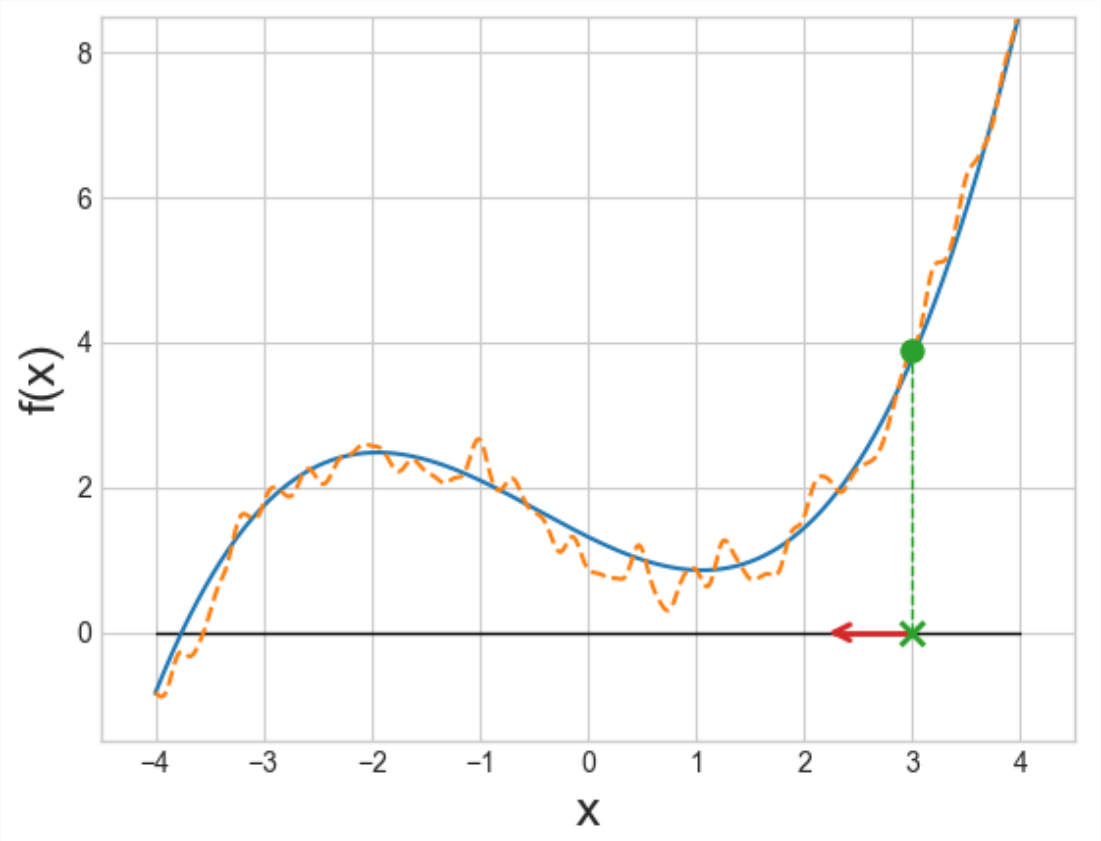
Stochastic Gradient Descent (SGD)

- Sometimes one might not be able to evaluate f exactly, but just a "noisy version" $\tilde{f}$ (formally, this is a stochastic process)
- If $\tilde{f}$ is differentiable and $\mathbb{E}[\tilde{f}] = f$, then one can still use the gradient of $\tilde{f}$ for gradient descent
- This delivers a noisy gradient with $\mathbb{E}[\nabla_{\boldsymbol{x}} \tilde{f}] = \nabla_{\boldsymbol{x}} f$, which "on average shows in the right direction"
- For theoretical convergence, the step-size η_t needs to be reduced over number of iterations t
- When

$$\sum_{t=1}^{\infty} \eta_t = \infty \qquad \sum_{t=1}^{\infty} {\eta_t}^2 < \infty \qquad (e.g.,\ \eta_t = \mathcal{O}(1/t))$$

  SGD converges to a local minimum of f!

Stochastic Gradient Descent (SGD)



`stochastic_gradient_descent_demo.py`

Probability

Probability = consistent reasoning under uncertainty

"Probability is nothing but common sense reduced to calculation."

Pierre-Simon Laplace, 1749–1827

`wikipedia.org`

Random Variables

Intuitively, a **random variable (RV)** X is a "variable assuming random values" — While the precise mathematical definition is a bit more detailed, in essence this intuition is correct.

The set of all possibles values the RV can assume is denoted as $\mathcal{X}$ (calligraphic X) and is called the **state space** of X. Generic values are denoted with lower case letters, x.

Discrete Random Variable

A random variable X whose state space $\mathcal{X}$ is finite or at most countable infinite (e.g. natural numbers $\mathbb{N}$, $\mathbb{Z}$), is called **discrete random variable**.

Probability Mass Function

Let X be a **discrete** RV with state space $\mathcal{X}$. The **probability mass function** (PMF) $p_X : \mathcal{X} \mapsto [0, 1]$ of X is defined as $p_X(x) = \mathbb{P}(X = x)$, i.e. the probability that $X = x$.

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

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

The distribution of such binary RV X is called **Bernoulli distribution** with **success probability** π.

Jakob Bernoulli, 1654–1705

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

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:

$$p_X(x) = \begin{cases} \pi_1 & \text{if } x = x_1 \\ \pi_2 & \text{if } x = x_2 \\ \vdots & \\ \pi_K & \text{if } x = x_K \end{cases}$$

where $\pi_i \geq 0$ and $\sum_{i=1}^{K} \pi_i = 1$. 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, $\pi_1 = \pi_2 = \cdots = \pi_6 = \frac{1}{6}$.

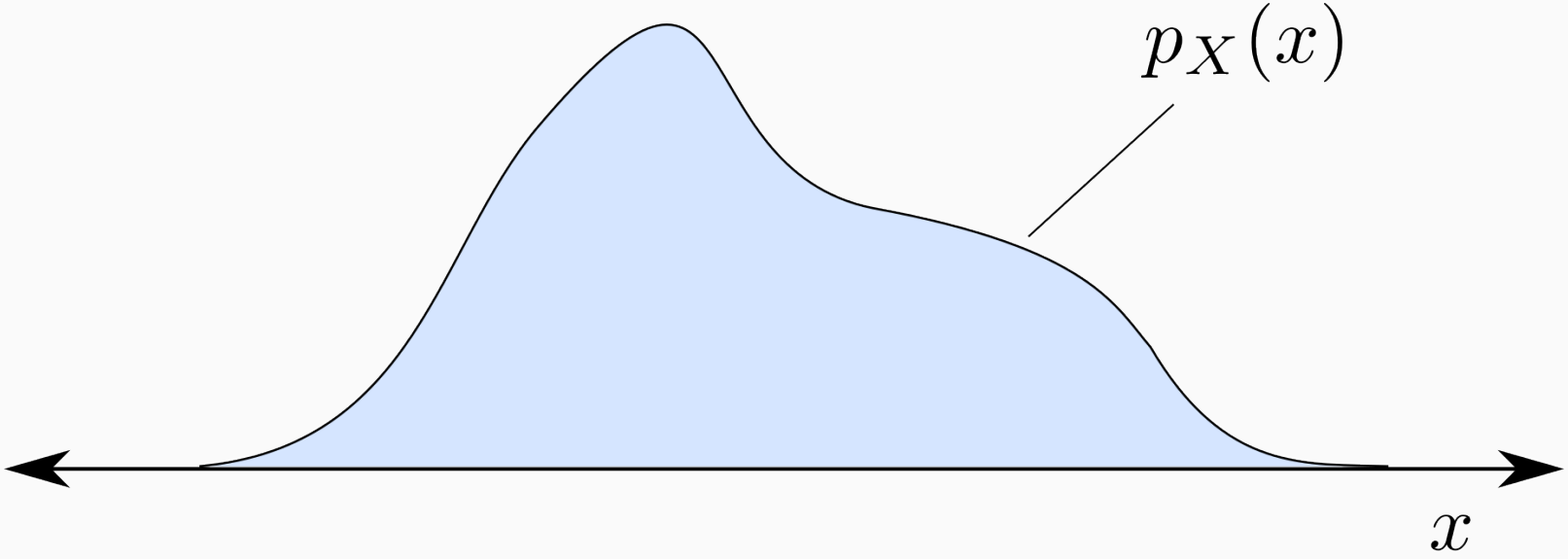
Probability Density Function

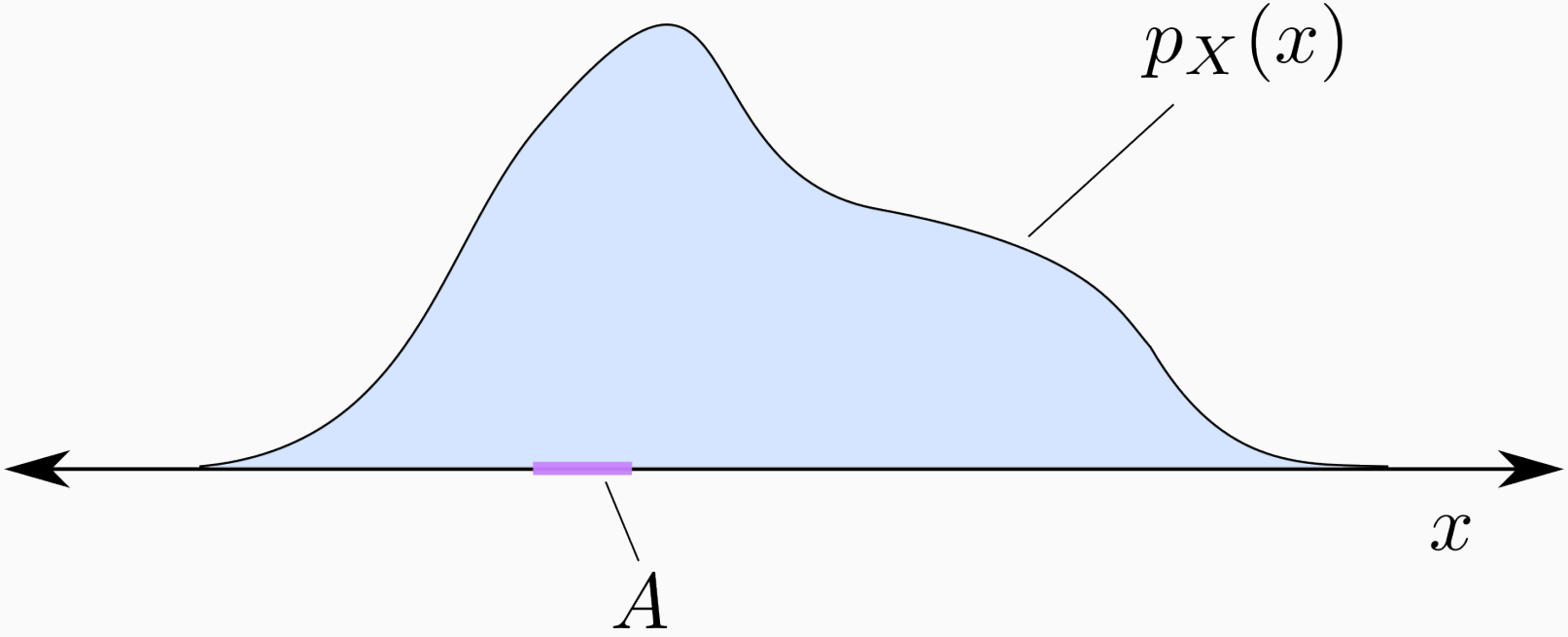
Let X be an RV with state space $\mathcal{X} = \mathbb{R}$. If there exists a function $p_X : \mathcal{X} \mapsto [0, \infty]$ such that

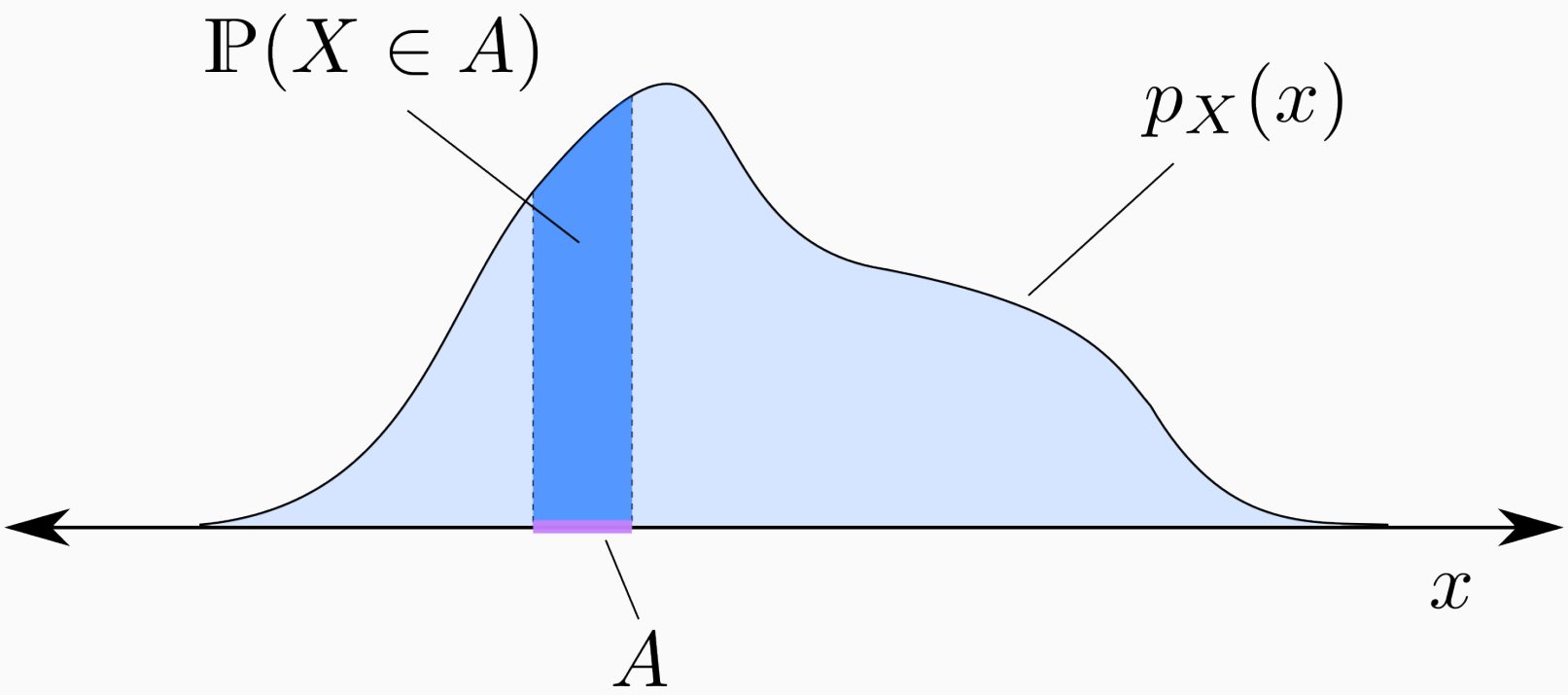
$$\mathbb{P}(X \in [a, b]) = \int_a^b p_X(x)\,\mathrm{d}x$$

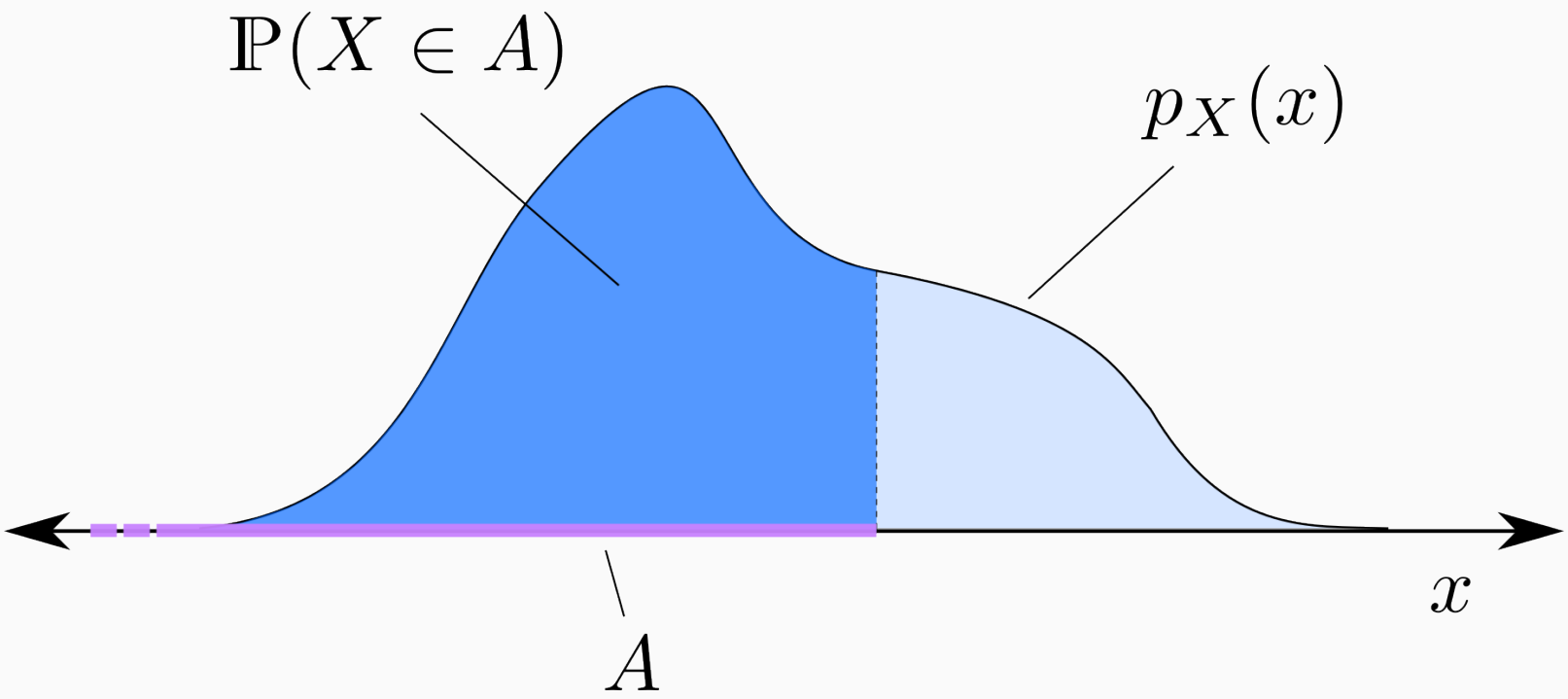
is the probability of $X \in [a, b]$, for all $a \leq b \in \mathbb{R}$, then p_X is called a **probability density function** (PDF) of X (short **probability density** or just **density**).

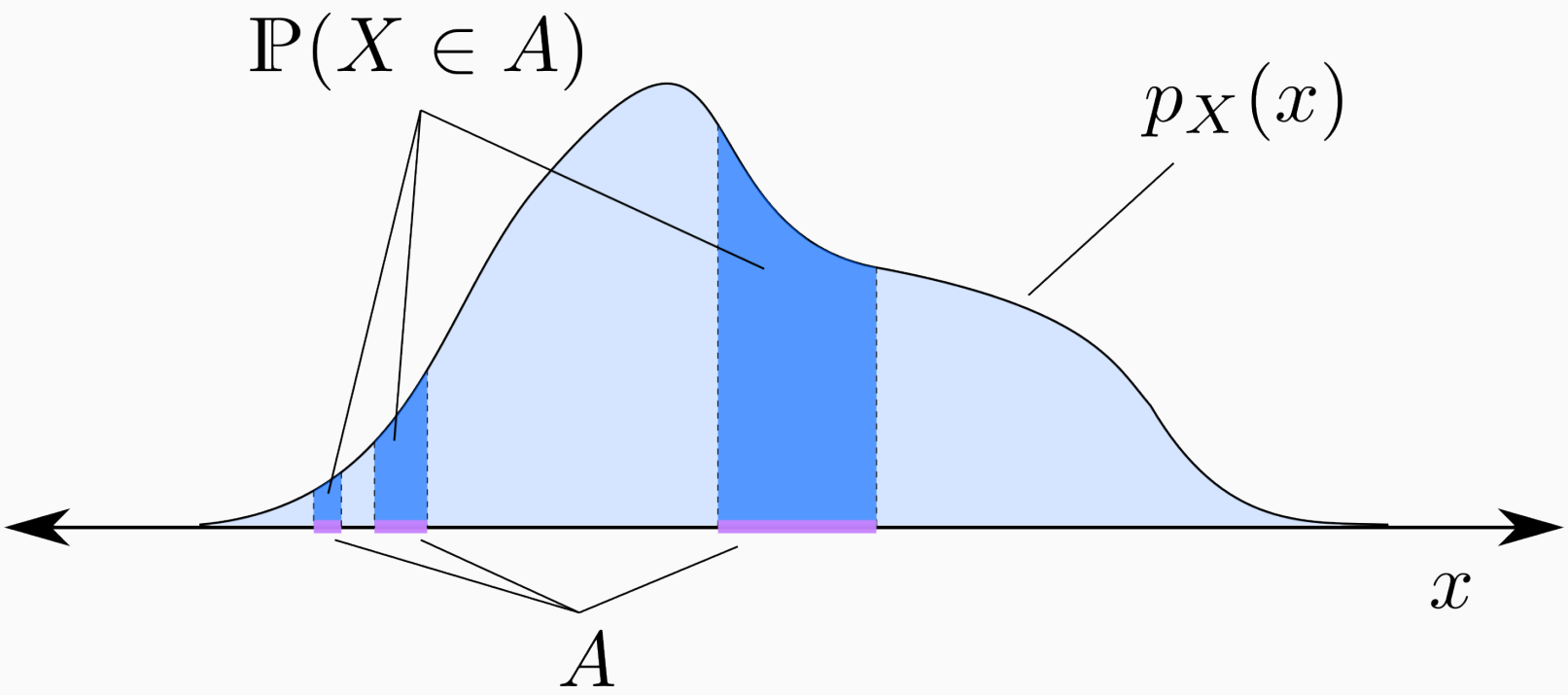
An RV which has a density p_X is called **continuous RV**.


$p_X(x)$
x


$p_X(x)$
x
A


$\mathbb{P}(X \in A)$
$p_X(x)$
x
A


$\mathbb{P}(X \in A)$
$p_X(x)$
x
A

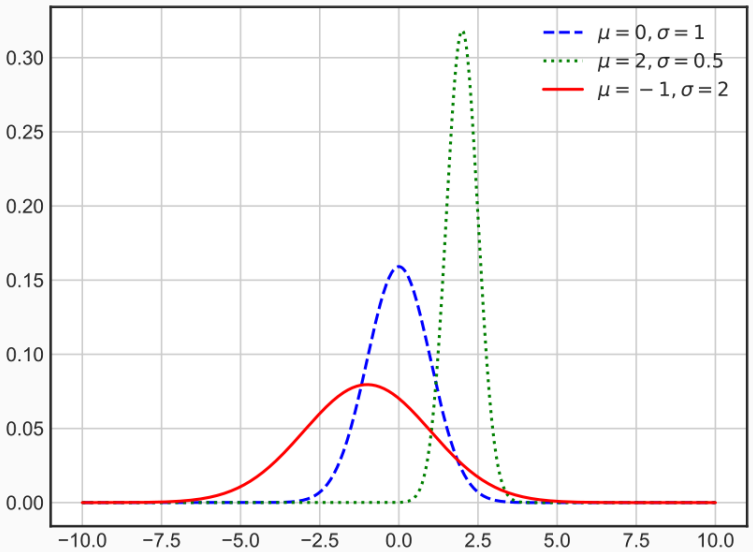

$\mathbb{P}(X \in A)$
$p_X(x)$
x
A

Gaussian Distribution, Normal Distribution

Let p_X be a probability density defined as

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

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

Image: wikipedia

Notes

- We use p_X for **both** probability mass functions and probability density functions $\rightarrow$ **probability distributions functions**
- We will often write $p(x)$ instead of $p_X(x)$
- For any RV X with probability density p_X:

$$\forall x\colon p_X(x) \geq 0 \qquad \int_{\mathbb{R}} p_X(x)\,\mathrm{d}x = 1$$

- Conversely, any function $p(x)$ with

$$\forall x\colon p(x) \geq 0 \qquad \int_{\mathbb{R}} p(x)\,\mathrm{d}x = 1$$

  is the PDF for **some** RV X.
- **Warning:** probability density $\neq$ probability

Expectations

Expected Value

Let p be a distribution function of an RV X with state space $\mathcal{X}$.
The **expected value** of X is defined as

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} p(x)\, x & \text{if } X \text{ is discrete,} \\ \int_{x \in \mathcal{X}} p(x)\, x \, \mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases}$$

Moreover, let g be a function defined on $\mathcal{X}$, **yielding a new RV** $Y := g(X)$. The expected value of Y is given as

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \begin{cases} \sum_{x \in \mathcal{X}} p(x)\, g(x) & \text{if } X \text{ is discrete,} \\ \int_{x \in \mathcal{X}} p(x)\, g(x) \, \mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases}$$

(known the **Law of the Unconscious Statistician (LOTUS)**)

Summary Statistics: Moments, Mean, Variance

Moments

The k^{th} **moment** of RV X is defined as $\mathbb{E}[X^k]$.

Mean

The **mean** of RV X is its expected value $\mathbb{E}[X]$.

Variance, Standard Deviation

The **variance** of RV X is defined as

$$\text{var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

The **standard deviation** of X is defined as $\text{std}[X] := \sqrt{\text{var}[X]}$.

Random Vectors, Multi Variate RVs

- Uppercase boldface letters $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ for **multivariate random variables**, **random vectors**.
- They take values from some **joint state space** $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}}$, $\boldsymbol{\mathcal{Z}}$, i.e. the **Cartesian product** of the individual state spaces.
- **Joint values** denoted with boldface lowercase letter $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$.

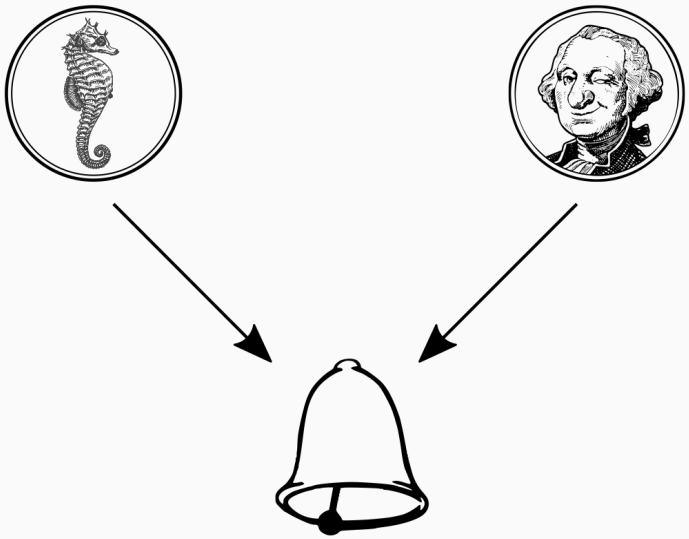
Joint Probability Mass Function (PMF)

Let $X_1, X_2, \ldots, X_D$ be **discrete RVs** and $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)^T$ be the corresponding random vector with state space $\boldsymbol{\mathcal{X}} = \bigtimes_{i=1}^{D} \mathcal{X}_i$, 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

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

- Two fair coins are tossed, modeled with Bernoulli C_1 and C_2
- If both show heads, a Bell (B) rings with 100% probability
- If exactly one shows heads, the Bell rings with 50% probability
- If both show tails, the Bell rings with 1% probability
- The three RVs have the following joint PMF:

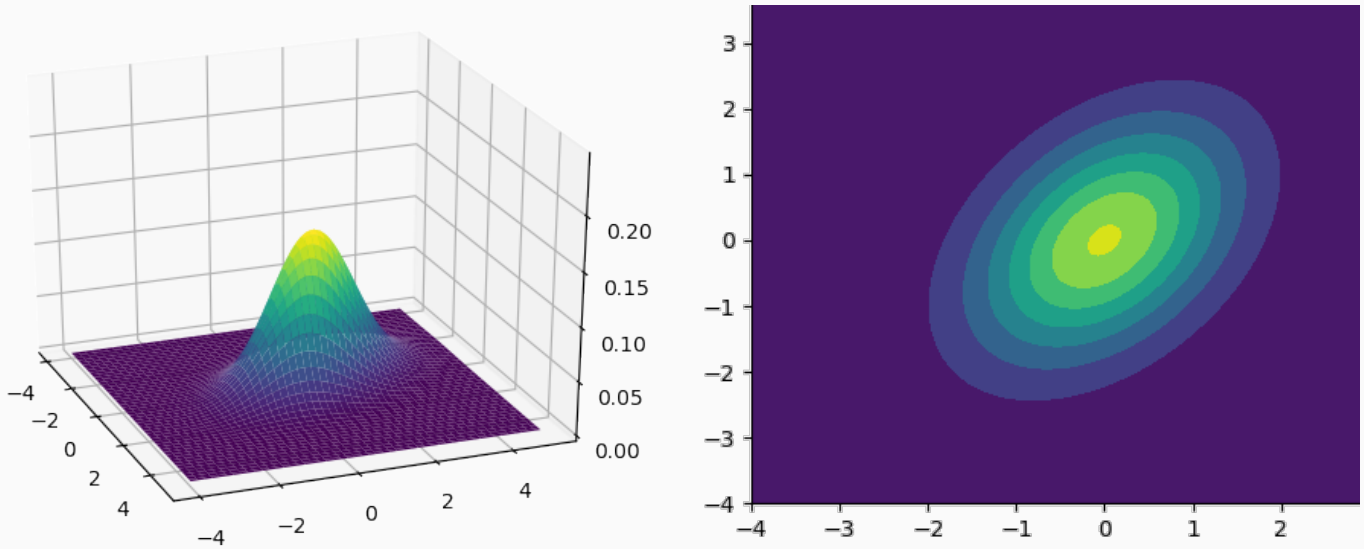


(c_1, c_2, b)	$p(c_1, c_2, b)$
$(0, 0, 0)$	0.2475
$(0, 0, 1)$	0.0025
$(0, 1, 0)$	0.125
$(0, 1, 1)$	0.125
$(1, 0, 0)$	0.125
$(1, 0, 1)$	0.125
$(1, 1, 0)$	0
$(1, 1, 1)$	0.25

Let p be defined as

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

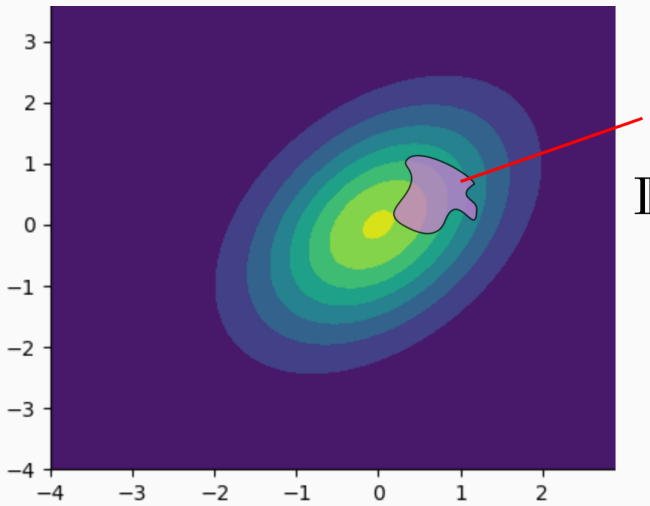
with D-dimensional **mean vector** $\boldsymbol{\mu}$ and $D \times D$ positive definite **covariance matrix** Σ. A distribution with density p is called **multivariate Gaussian distribution**. It generalizes Gaussians to higher dimensions.



Probability Represented by Multivariate Densities

- Consider a **random vector** $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)$, consisting of D real-valued RVs
- Like for single RVs, the volume under the density $p(\boldsymbol{x})$ corresponds to the probability of events
- For an set $A \subseteq \mathbb{R}^D$, integrating p over A gives the probability of $\boldsymbol{X} \in A$:

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

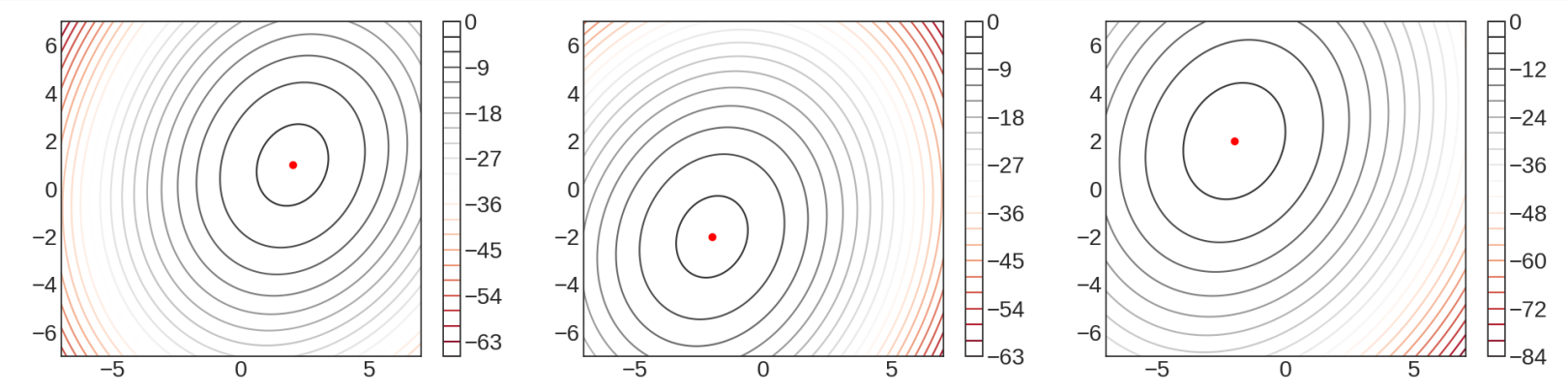


$$\mathbb{P}(\mathbf{X} \in A) = \int_A p(x_1, x_2)\, \mathrm{d}(x_1, x_2)$$

Mean vector μ

The **contour lines** (or **contour sets** in higher dimensions) of Gaussians are ellipsoids.

The mean vector μ determines the **location** and **mode** (maximal point) of the pdf.



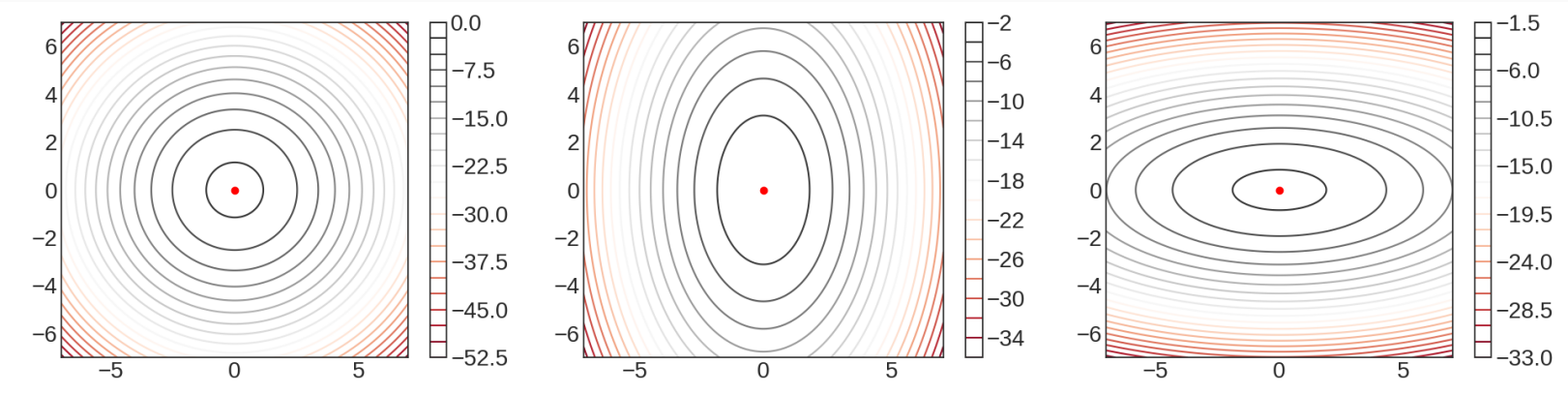
$\mu = (2, 1)^\top$ $\mu = (-2, -2)^\top$ $\mu = (-2, 2)^\top$

Covariance Σ

Σ is a **positive definite matrix**, which means that

- $\boldsymbol{a}^T \Sigma \boldsymbol{a} \geq 0$, for all $\boldsymbol{a} \in \mathbb{R}^D$, or equivalently
- Σ is symmetric and has only non-negative **eigenvalues**

In the special case that Σ is a diagonal matrix, the ellipses are axis-aligned. Elements in diagonal are the variances in each dimension:



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

Diagonal Covariance = Independence

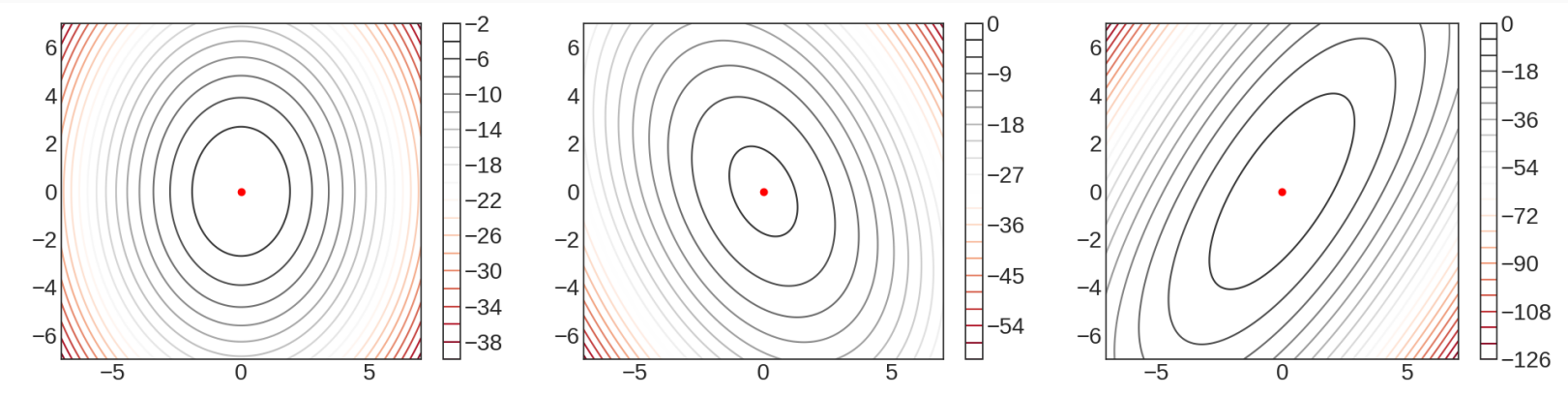
Gaussians with diagonal covariance factorize into indepenent 1d-Gaussians:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \prod_d p(x_d; \mu_d, \Sigma_{dd})$$

Intuitively, this means that none of the dimensions holds any information about any other dimension.

Covariance Σ con't

If Σ is not a diagonal matrix, then some or all off-diagonal elements (**co-variances**) are non-zero. In this case the ellipsoids are "slanted" and the dimensions are not independent:



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$