
STATISTICAL AND MATHEMATICAL FOUNDATIONS

SETTING THE STAGE



OUTLINE

1. Modeling
2. Distributions
3. Central Limit Theorem
4. Estimation
5. Bayes' Theorem
6. Matrix Algebra
7. Eigenvalues and Eigenvectors
8. Regression
9. Optimization

Real World



Theory

Identification of
details relevant to
description and
translation of real-
world objects into
model variables

Model



MODELS IN GENERAL

First principles modeling

- examine a system
- write down a set of rules/equations that describe the essence of the system
- ignore complicating details that are “less” important

Statistical modeling

- typically a set of equations with parameters
- parameters are learned (model is “trained”) using multiple data observations
- data sample vs. population

MODELING HEURISTICS

Basic steps in building a statistical model:

- **defining the goals**
- **gathering data**
- **deciding on the model structure**
- **preparing the data**
- **selecting and removing features**
- **building candidate models**
- **finalizing the model**
- **implementing and monitoring**

DATA AND DISTRIBUTIONS

If a data feature can be characterized by a distribution, consider asking **four basic questions**:

1. Can the variable only take on **discrete** values? **continuous** values?
2. Is the data distribution **symmetric**?
3. Does the variable have theoretical **upper** and **lower limits**?
4. How likely is it to observe **extreme values** in the distribution?

distribution	pdf $f(x)$	mean	variance	notes
uniform $U(a, b)$	$\frac{1}{b-a}$ for $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	most languages provide rand # generators for $U(a, b)$; used to generate r.v. with other distributions
Gaussian $N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $x \in \mathbb{R}$	μ	σ^2	if $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$ (and vice-versa); very commonly used
Poisson $P(\lambda), \lambda \geq 0$	$\frac{\lambda^x}{x!} e^{-\lambda}$ for $x = 0, 1, 2, \dots$	λ	λ	estimates the # of events that occur in a continuous time interval (# of calls received in 1-hour intervals)
binomial $\mathcal{B}(N, p), N \in \mathbb{N}, p \in [0, 1]$	$\binom{N}{x} p^x (1-p)^{N-x}$ for $x = 0, 1, \dots, N$	Np	$Np(1-p)$	describes the probability of exactly x successes in N independent trials if the probability of a success in a single trial is p (# of heads in N coin tosses)
log-normal $\Lambda(\mu, \sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}$ for $x > 0$	$e^{(\mu + \sigma^2/2)}$	$e^{(2\mu + \sigma^2)} [e^{\sigma^2} - 1]$	if $\ln X \sim N(\mu, \sigma^2)$, then $X \sim \Lambda(\mu, \sigma^2)$ (and vice-versa); positively skewed

JOINT DISTRIBUTIONS

Univariate distributions are useful modeling tools, especially when the variables under consideration are **independent**.

In practice, that is not usually the case. A **joint distribution** $P(X_1, \dots, X_n)$ gives the probability that each of X_1, \dots, X_n falls in a given range. The **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has pdf

$$f(x_1, \dots, x_n) := f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

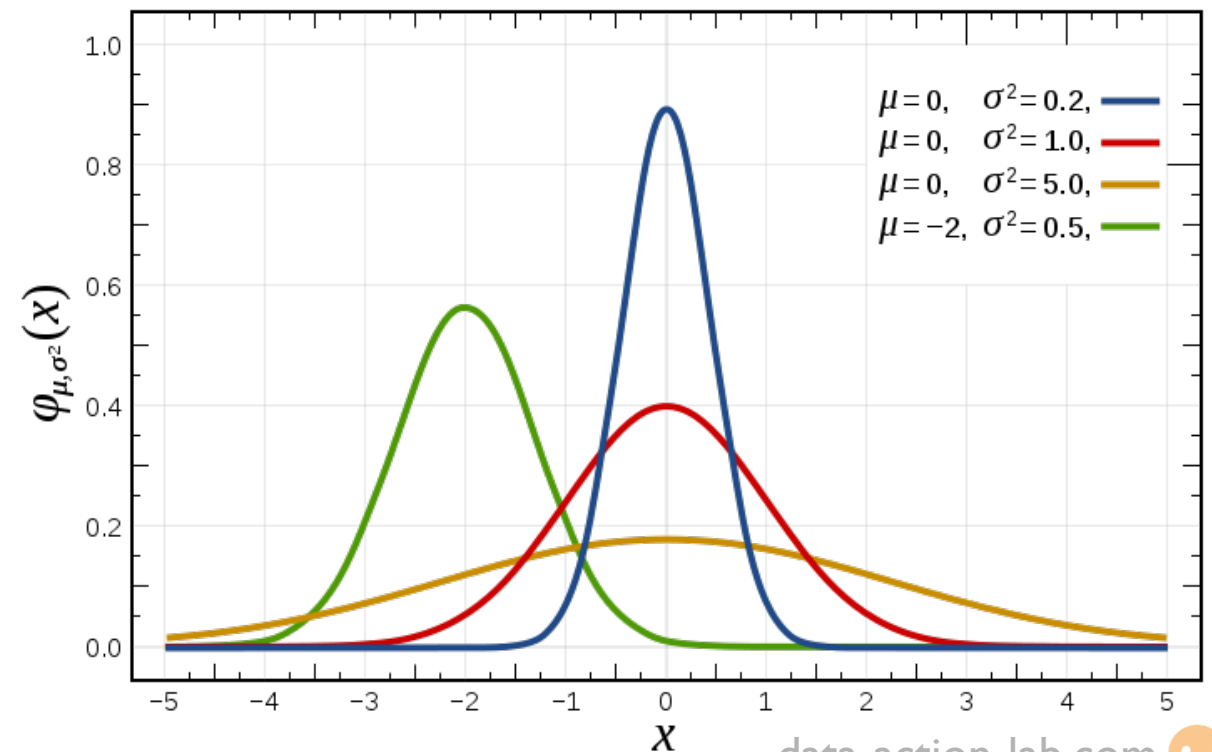
where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ the covariance matrix.

NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$ is **fully characterized** by the mean μ and the standard deviation σ , which reduces estimation requirements.

The probability of a value being drawn can be obtained if we know how many multiples of σ separate it from μ

- within σ from μ : $\approx 68\%$
- within 2σ from μ : $\approx 95\%$
- within 3σ from μ : $\approx 99.7\%$



NORMAL DISTRIBUTION

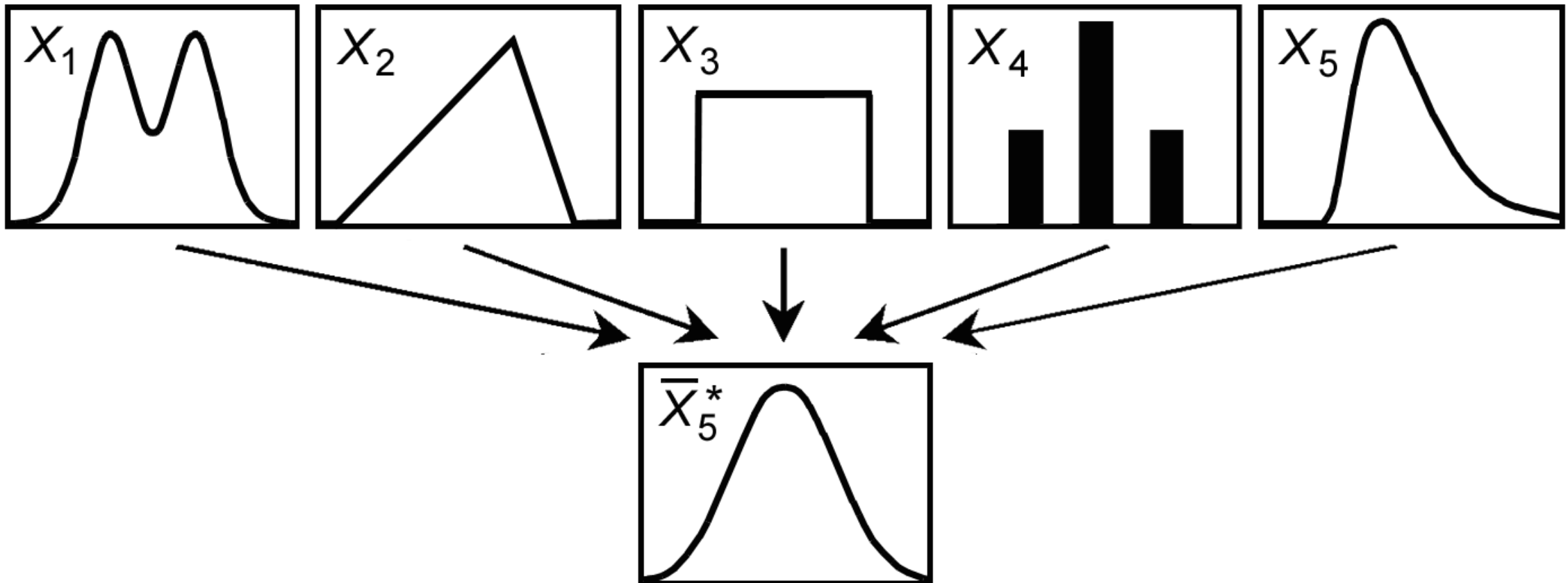
The normal distribution is best suited for data meeting the following minimum requirements:

- strong tendency for the data to take on a central value
- positive, negative deviations from this central value are equally likely
- frequency of the deviations falls off rapidly as we move further away from the central value.

Symmetry of deviations leads to zero **skewness**; low prob. of large deviations from the central value leads to no **kurtosis**.

Its omnipresence in human affairs is linked to the **Central Limit Theorem**.

CENTRAL LIMIT THEOREM IN ACTION



ESTIMATION

One of the goals of statistics is to try to **understand a large population** on the basis of the information available in a small sample.

In particular, we are interested in the population **parameters**, which are estimated using suitable sample statistics.

For example, we may use the **sample mean** $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ as an estimate for the true **population mean** μ .

ESTIMATION

The **estimator** is a random variable; the **estimate** is a number.

As an another example, the **sample standard deviation** S is an estimator of the true **population standard deviation** σ and the computed value

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

of S is an estimate of σ .

An estimator W of ω is **unbiased** if $E(W) = \omega$.

BASIC MATHEMATICAL CONCEPTS

If the estimate $\hat{\beta}$ is unbiased, $E(\hat{\beta} - \beta) = 0$, then an approximate **95% confidence interval** (95% CI) for β is given approximately by

$$\hat{\beta} \pm 2\sqrt{\hat{V}(\hat{\beta})},$$

where $\hat{V}(\hat{\beta})$ is a **sampling design-specific** estimate of $\text{Var}(\hat{\beta})$.

But what is a 95% CI, exactly?

CONDITIONAL PROBABILITIES

A **conditional probability** is the probability of an event taking place given that another event occurred.

The conditional probability of A given B , $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The probability that two events A and B both occur is obtained by applying the multiplication rule:

$$P(A \cap B) = P(B) P(A|B) = P(A) P(B|A)$$

BAYES' THEOREM

The sum rule and the product rules are the **basic rules of probability**.

Bayes' Theorem and the **Marginalization Rule** are simple corollaries of these basic rules.

Bayes' Theorem is sometimes written in a slightly different form

$$P(X|Y, I) = \frac{P(Y|X, I) \times P(X|I)}{P(Y|I)}$$

BAYES' THEOREM

Set-up: assume that an experiment has been conducted to determine the degree of validity of a particular hypothesis, and that experimental data has been collected.

The central data analysis question: given everything that was known *prior* to the experiment, does the collected data support (or invalidate) the hypothesis?

Throughout, let X denote the proposition that the hypothesis in question is true, let Y denote the proposition that the experiment yielded the actual observed data, let I denote (as always) the relevant background information.

BAYES' THEOREM

Central data analysis question (reprise):

What is the value of $P(\text{hypothesis is true} \mid \text{observed data}, I)$?

Problem: this is nearly always impossible to compute directly.

Solution: using Bayes' Theorem,

$$P(\text{hypothesis} \mid \text{data}, I) = \frac{P(\text{data} \mid \text{hypothesis}, I) \times P(\text{hypothesis} \mid I)}{P(\text{data} \mid I)},$$

likelihood prior
posterior evidence

it may be that the terms on the right are easier to compute.

BAYES' THEOREM

Determining the **prior** is a source of considerable controversy

The **evidence** is harder to compute on theoretical grounds – evaluating the probability of observing data requires access to some model as part of I .

BAYES' THEOREM

Thankfully, the evidence is rarely required on problems of parameter estimation (although it is crucial for model selection):

- prior to the experiment, there are numerous competing hypotheses
- the priors and likelihoods will differ, but not the evidence
- the evidence is not needed to differentiate the various hypotheses

Bayes' Theorem is often presented as

$$P(\text{hypothesis} \mid \text{data}, I) \propto P(\text{data} \mid \text{hypothesis}, I) \times P(\text{hypothesis} \mid I)$$

or simply as $\text{posterior} \propto \text{likelihood} \times \text{prior}$, that is to say, **beliefs should be updated in the presence of new information.**

LINEAR ALGEBRA

A **matrix** is an important mathematical tool that allows for easy organization of information, simplifies notation, and facilitates the application of algorithms to data.

Most statistical tools require **rectangular** data:

- each column contains a **variable** (feature, field, attribute)
 - indicator, target, question in a survey, etc.
- each row contains an **observation** (case, unit, item)
 - country, survey respondent, subject in an experiment, etc.
- each cell contains a **value** (measurement) for a particular variable and observation
 - GDP per capita for Canada, answer to a specific question, age, etc.

MATRIX OPERATIONS

A matrix is a rectangular grid of **elements** arranged into **rows** and **columns**.

Matrices are often used in algebra to solve for unknown values in linear equations, and in geometry.

You should know how to do

- **matrix addition, multiplication by a scalar, matrix transposition**
- **matrix multiplication**
- **matrix inversion, matrix determinant, matrix trace**

EIGENVECTORS AND EIGENVALUES

An **eigenvector** of a matrix A is a vector $\mathbf{v} \neq \mathbf{0}$ such that, for some scalar λ , $A\mathbf{v} = \lambda\mathbf{v}$.

The value λ is called an **eigenvalue** of A associated with \mathbf{v} .

The eigenvalues of an $n \times n$ matrix A satisfy $\det(A - \lambda I_n) = 0$. The left-hand side is a polynomial in λ , and is called the **characteristic polynomial** of A , denoted by $p_A(\lambda)$.

To find the eigenvalues of A , we find the roots of $p_A(\lambda)$.

EIGEN-DECOMPOSITION


If an $n \times n$ matrix A has n linearly independent eigenvectors, then A may be **decomposed** in the following manner:

$$A = B\Lambda B^{-1},$$

where Λ is a diagonal matrix whose diagonal entries are the eigenvalues of A and the columns of B are the corresponding eigenvectors of A .

REGRESSION MODELING

The data structure of a general modeling task is represented by



X_1	X_2	\cdots	X_p	Y
x_{11}	x_{12}	\cdots	x_{1p}	y_1
x_{21}	x_{22}	\cdots	x_{2p}	y_2
\cdots	\cdots	\cdots	\cdots	\cdots
x_{n1}	x_{n2}	\cdots	x_{np}	y_n

We consider p independent variables X_i to try to predict the dependent variable Y .

In order to simplify the discussion in the following, we introduce the matrix notation $\mathbf{X}[n \times p]$, $\mathbf{Y}[n \times 1]$, $\boldsymbol{\beta}[p \times 1]$, where n is the # of observations and p is the # of independent variables.

LINEAR REGRESSION

The basic assumption of linear regression is that the dependent variable y can be **approximated** by a linear combination of the independent variables as follows:

$$Y = X\beta + \varepsilon,$$

where $\beta \in \mathbb{R}^p$ is to be determined based on the training set, and for which

$$E(\varepsilon|X) = 0, \quad E(\varepsilon\varepsilon^T|X) = \sigma^2 I.$$

Typically, the errors are also assumed to be normally distributed, that is :

$$\varepsilon|X \sim N(0, \sigma^2 I).$$

LINEAR REGRESSION

If $\hat{\beta}_i$ is the estimate of the true coefficient β_i , the **linear regression** model associated with the data is

$$\hat{Y}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

In matrix form, the regression problem requires a solution $\hat{\boldsymbol{\beta}}$ to the **normal equation** $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}$.

When the symmetric positive definite matrix $\mathbf{X}^T \mathbf{X}$ is invertible, the fitted coefficient is simply $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y})$. Note that $\mathbf{X}^T \mathbf{X}$ is a $p \times p$ matrix, which makes the inversion “easier” to compute, relatively speaking, when n is large.

GENERALIZED LINEAR REGRESSION

Generalized linear models (GLMs) extend linear statistical models by accommodating response variables with **non-normal** conditional distributions.

Except for the **error structure**, a GLM is essentially the same as for a linear model:

$$Y_i \sim \text{some distribution with mean } \mu_i, \text{ where } g(\mu_i) = x_i^T \beta$$

A GLM therefore consists of three parts:

- a **systematic** component $x_i^T \beta$
- a **random** component – specified distribution for Y_i
- a **link** function g

OPTIMIZATION

Suppose we have a **cost** (objective) function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ to **optimize** (the maximum likelihood function of linear regression, for instance).

Seeking a maximum for f is equivalent to seeking a minimum for $-f$.

The aim is to find parameter values \mathbf{x} that minimize this function:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} f(\mathbf{x})$$

The cost function could be subjected to a number of constraints

$$c_i(\mathbf{x}) = 0, i = 1, \dots, m; c_j(\mathbf{x}) \geq 0, j = 1, \dots, k; \mathbf{x} \in \Omega \subseteq \mathbb{R}^n.$$

OPTIMIZATION

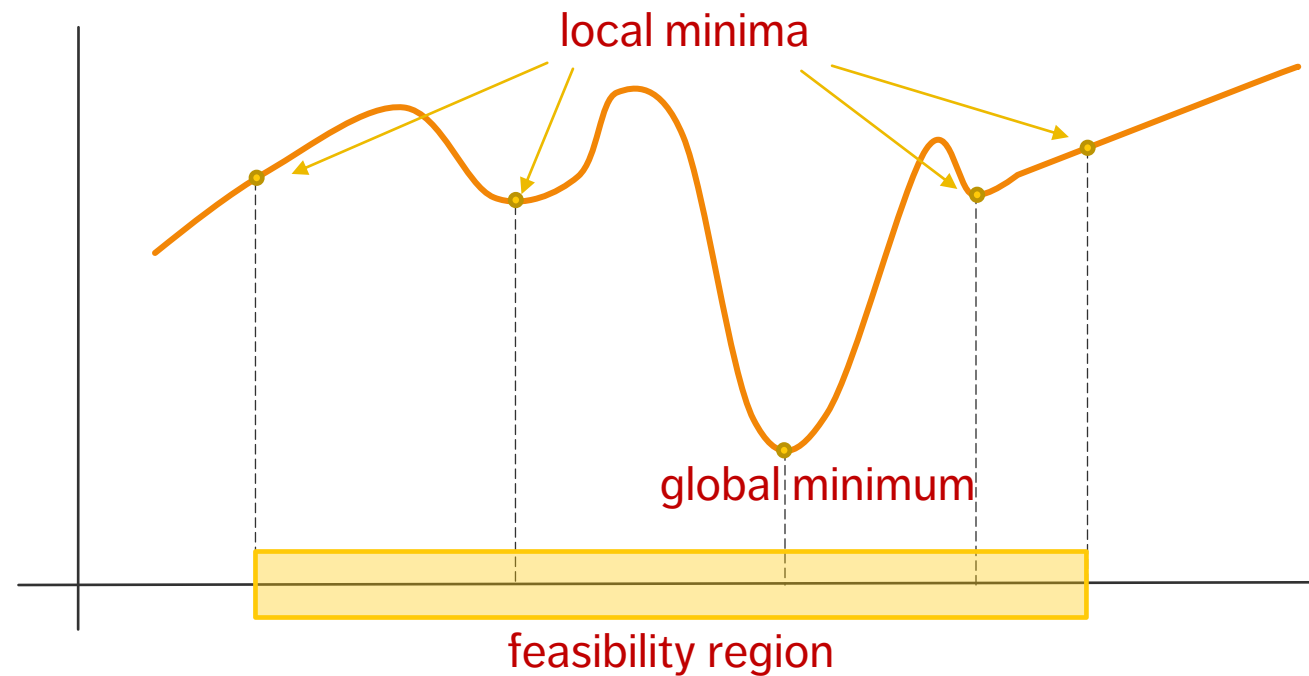
The optimization problem can be viewed as a **decision problem** that involves finding the “best” vector \mathbf{x} over all possible vectors in $\Omega \subseteq \mathbb{R}^n$.

This vector is called the **minimizer** of f over Ω . There may be multiple minimizers, or none.

If $\Omega = \mathbb{R}^n$, then we refer to the problem as an **unconstrained** optimization problem.

In general, this is not a trivial problem (consult the literature).

TYPE OF MINIMA



In many instances, optimization is a **numerical** endeavour. Which of the minima is found depends on the algorithm's **starting point**.