System Identification

Control Engineering EN, 3rd year B.Sc.
Technical University of Cluj-Napoca
Romania

Lecturer: Lucian Buşoniu

Part III

Mathematical Background: Linear Regression and Statistics

Motivation

So far, we have been dealing with transient analysis of step-response models. This mostly involved familiar concepts related to linear systems and their time-domain responses.

Many upcoming methods for system identification require additional tools: linear regression and some concepts from probability theory and statistics. We will discuss these tools here.

In this part we redefine some notation (e.g. x, A) to have a different meaning than in the rest of the course.

Table of contents

Table of contents

Regression problem

Problem elements:

- A collection of known samples $y(k) \in \mathbb{R}$, indexed by $k = 1, \ldots, N$: y is the *regressed variable*.
- For each k, a known vector $\varphi(k) \in \mathbb{R}^n$: contains the *regressors* $\varphi_i(k)$, $i = 1, \ldots, n$, $\varphi(k) = [\varphi_1(k), \varphi_2(k) \ldots, \varphi_n(k)]^\top$.
- An unknown *parameter vector* $\theta \in \mathbb{R}^n$.

Objective: identify the behavior of the regressed variable from the data, using the linear model:

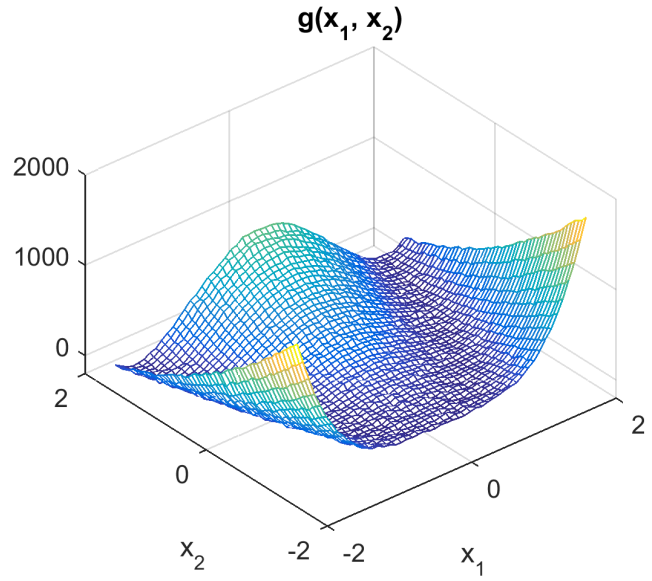
$$\boxed{y(k) = \varphi^\top(k)\theta}$$

Linear regression is classical and very common, e.g. it was used by Gauss to compute the orbits of the planets in 1809.

Regression problem: Two major uses

1. k is a time variable, and we wish to model the time series $y(k)$.
2. k is just a data index, and $\varphi(k) = \phi(x(k))$ where x is an input of some unknown function g. Then $y(k)$ is the corresponding output (possibly corrupted by noise), and the goal is to identify a model of g from the data.
   This problem is also called *function fitting*, *function approximation*, or *supervised learning*.

unknown $g(x)$ $\widehat{g}(x)$



?

Function approximation: Basis functions

For function approximation, the regressors $\phi_i(x)$ in:

$$\phi(x(k)) = [\phi_1(x(k)), \phi_2(x(k)), \ldots, \phi_n(x(k))]^\top$$

are also called *basis functions*.

Function approximation: Motivating example 1

We study the yearly income y (in EUR) of a person based on their education level x_1 and job experience x_2 (both measured in years).

We are given a set of tuples $(x_1(k), x_2(k), y(k))$ from a representative set of persons. The goal is to predict the income of any other person by knowing how educated (x_1) and experienced (x_2) they are.

- Take basis functions $\phi(x) = [x_1, x_2, 1]^\top$. So we expect the income to behave like $\theta_1 x_1 + \theta_2 x_2 + \theta_3 = \phi^\top(x)\theta$, growing linearly with education and experience (from some minimum level). Regression involves finding the parameters θ in order to best fit the given data.
- Reality is of course more complicated... so we would likely need more input variables, better basis functions, etc.

Function approximation: Motivating example 2

We study the reaction time y (in ms) of a driver based on their age x_1 (in years) and fatigue x_2 (e.g. on a scale from 0 to 1).

We are given a set of tuples $(x_1(k), x_2(k), y(k))$ from a representative set of persons of various ages and stages of fatigue. The goal is to predict the rection time of any other person by knowing how old (x_1) and tired (x_2) they are.

Regressors example 1: Polynomial of k

Suitable for time series modeling.

$$
\begin{aligned}
y(k) &= \theta_1 + \theta_2 k + \theta_3 k^2 + \ldots + \theta_n k^{n-1} \\
&= \begin{bmatrix} 1 & k & k^2 & \ldots & k^{n-1} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \ldots \\ \theta_n \end{bmatrix} \\
&= \varphi^\top(k)\theta
\end{aligned}
$$

Regressors example 2: Polynomial of x

Suitable for function approximation. For instance, polynomial of degree 2 with two input variables $x = [x_1, x_2]^\top$:

$$y(k) = \theta_1 + \theta_2 x_1(k) + \theta_3 x_2(k) + \theta_4 x_1^2(k) + \theta_5 x_2^2(k) + \theta_6 x_1(k) x_2(k)$$

$$= \begin{bmatrix} 1 & x_1(k) & x_2(k) & x_1^2(k) & x_2^2(k) & x_1(k)x_2(k) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{bmatrix}$$

$$= \phi^\top(x(k))\theta = \varphi^\top(k)\theta$$

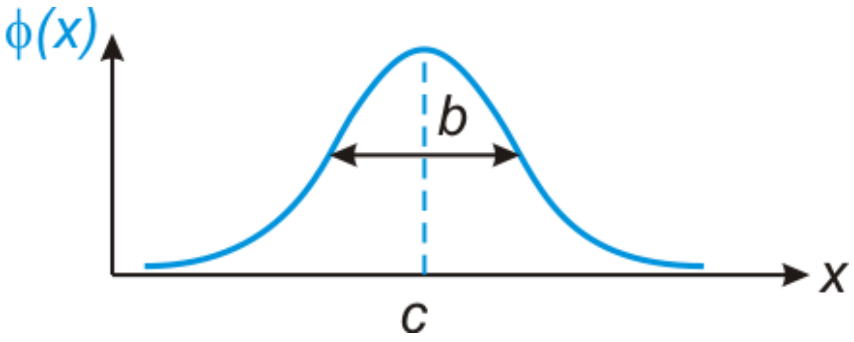
Connection: Project part 1

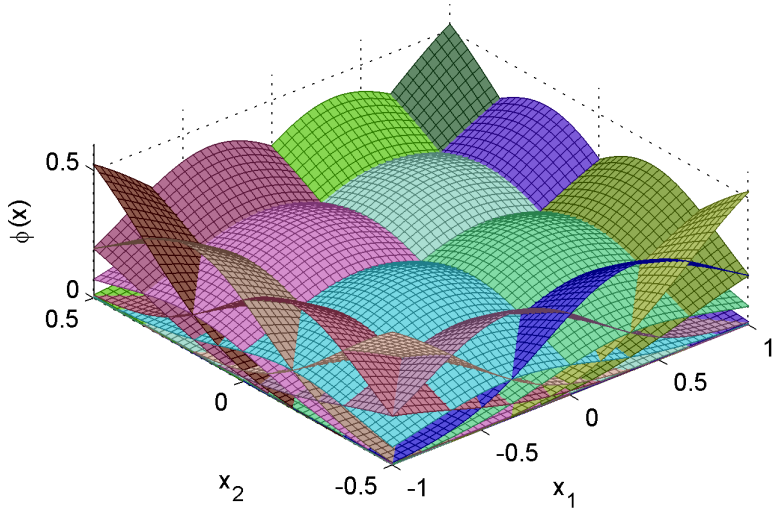
Regressors example 3: Gaussian basis functions

Suitable for function approximation:

$$\phi_i(x) = \exp\left[-\frac{(x - c_i)^2}{b_i^2}\right] \qquad \text{(1-dim);}$$

$$= \exp\left[-\sum_{j=1}^{d} \frac{(x_j - c_j)^2}{b_j^2}\right] \qquad \text{(d-dim)}$$

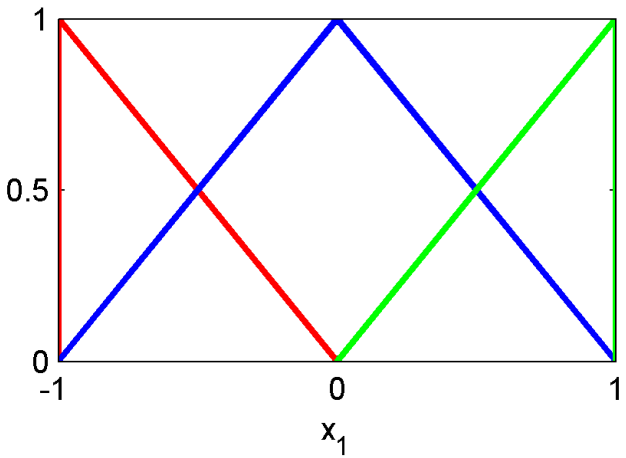


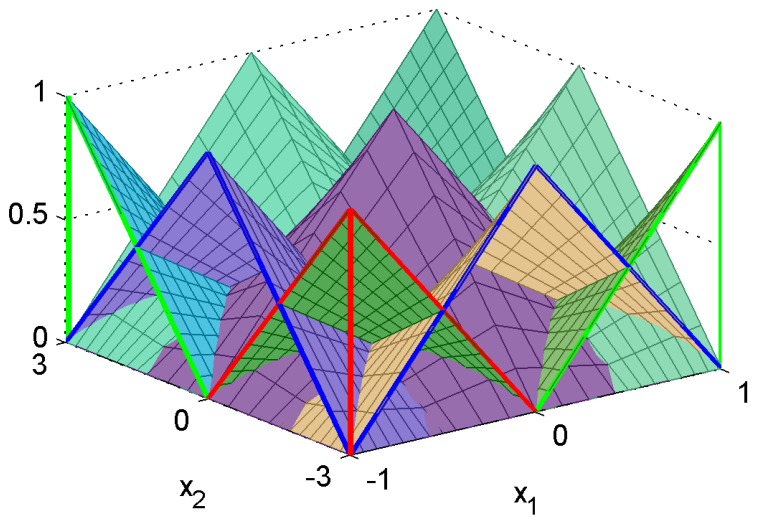


Regressors example 4: Interpolation

Suitable for function approximation.

- *d*-dimensional grid of points in the input space.
- (Multi)-Linear interpolation between the points.
- Equivalent with *pyramidal* basis functions (triangular in 1-dim)





Linear system

Writing the model for each of the N data points, we get a linear system of equations:

$$\begin{aligned} y(1) &= \varphi_1(1)\theta_1 + \varphi_2(1)\theta_2 + \dots \varphi_n(1)\theta_n \\ y(2) &= \varphi_1(2)\theta_1 + \varphi_2(2)\theta_2 + \dots \varphi_n(2)\theta_n \\ &\dots \\ y(N) &= \varphi_1(N)\theta_1 + \varphi_2(N)\theta_2 + \dots \varphi_n(N)\theta_n \end{aligned}$$

Recall that in function approximation, $\varphi_i(k) = \phi_i(x(k))$

This system can be written in a *matrix form*:

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} \varphi_1(1) & \varphi_2(1) & \dots & \varphi_n(1) \\ \varphi_1(2) & \varphi_2(2) & \dots & \varphi_n(2) \\ \dots & \dots & \dots & \dots \\ \varphi_1(N) & \varphi_2(N) & \dots & \varphi_n(N) \end{bmatrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \dots \\ \theta_n \end{bmatrix}$$

$$\boxed{Y = \Phi\theta}$$

with newly introduced variables $Y \in \mathbb{R}^N$ and $\Phi \in \mathbb{R}^{N \times n}$.

Least-squares problem

If $N = n$, the system can be solved with equality.

In practice, it is a good idea to use $N > n$, due e.g. to noise. In this case, the system can no longer be solved with equality, but only in an approximate sense.

- *Error* at k: $\varepsilon(k) = y(k) - \varphi^\top(k)\theta$,
  error vector $\varepsilon = [\varepsilon(1), \varepsilon(2), \ldots, \varepsilon(N)]^\top$.
- Objective function to be minimized:

$$V(\theta) = \boxed{\frac{1}{2}\sum_{k=1}^{N} \varepsilon(k)^2} = \frac{1}{2}\varepsilon^\top\varepsilon$$

Least-squares problem

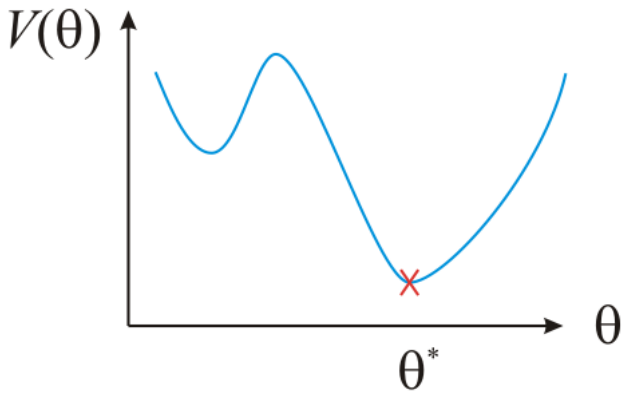
Find the parameter vector $\widehat{\theta}$ that minimizes the objective function:

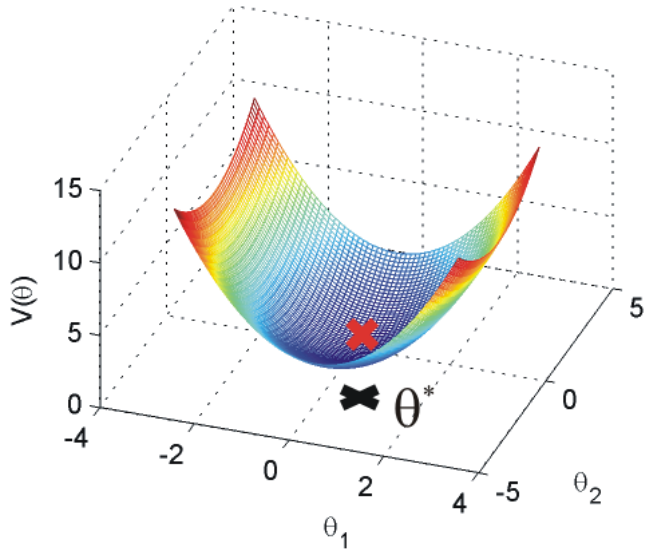
$$\widehat{\theta} = \arg\min_{\theta} V(\theta)$$

Parenthesis: Optimization problem

Given a function V of variables θ, which may be the least-squares objective, or any other function:

find the *optimal function value* $\min_\theta V(\theta)$ and variable values $\theta^* = \arg\min_\theta V(\theta)$ that achieve the minimum.





Note that in the case of linear regression, we use the notation $\widehat{\theta}$; while $\widehat{\theta}$ is still the true solution to the optimization problem given the data, it is still an estimate because the data is noisy

Formal regression solution

After applying some linear algebra:

$$\boxed{\widehat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top Y}$$

Remarks:

- The optimal objective value is $V(\widehat{\theta}) = \frac{1}{2}[Y^\top Y - Y^\top \Phi(\Phi^\top \Phi)^{-1}\Phi^\top Y]$.
- Matrix $\Phi^\top \Phi$ must be invertible. This boils down to a good choice of the model (order n, regressors φ) and having informative data.

Alternative expression

$$\Phi^\top \Phi = \sum_{k=1}^{N} \varphi(k)\varphi^\top(k), \Phi^\top Y = \sum_{k=1}^{N} \varphi(k)y(k)$$

So the solution can be written:

$$\widehat{\theta} = \left[\sum_{k=1}^{N} \varphi(k)\varphi^\top(k)\right]^{-1} \left[\sum_{k=1}^{N} \varphi(k)y(k)\right]$$

Advantage: matrix Φ of size $N \times n$ no longer has to be computed, only smaller matrices and vectors are required, of size $n \times n$ and n, respectively.

Solving the linear system

In practice both these inversion-based techniques perform poorly from a numerical point of view. Better algorithms exist, such as so-called orthogonal triangularization.

In most cases, MATLAB is competent in choosing a good algorithm. If Φ is stored in variable `PHI` and Y in `Y`, then the command to solve the linear system using matrix left division (backslash) is:

```
theta = PHI \ Y;
```

Better control of the algorithm is obtained by using function `linsolve` instead of matrix left division.

Table of contents

Analytical example: Estimating a scalar

Model:

$$y(k) = b = 1 \cdot b = \varphi(k)\theta$$

where $\varphi(k) = 1 \forall k$, $\theta = b$.

For all N data points:

$$\begin{aligned} y(1) &= \varphi(1)\theta = 1 \cdot b \\ &\dots \\ y(N) &= \varphi(N)\theta = 1 \cdot b \end{aligned}$$

In matrix form:

$$\begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \theta$$

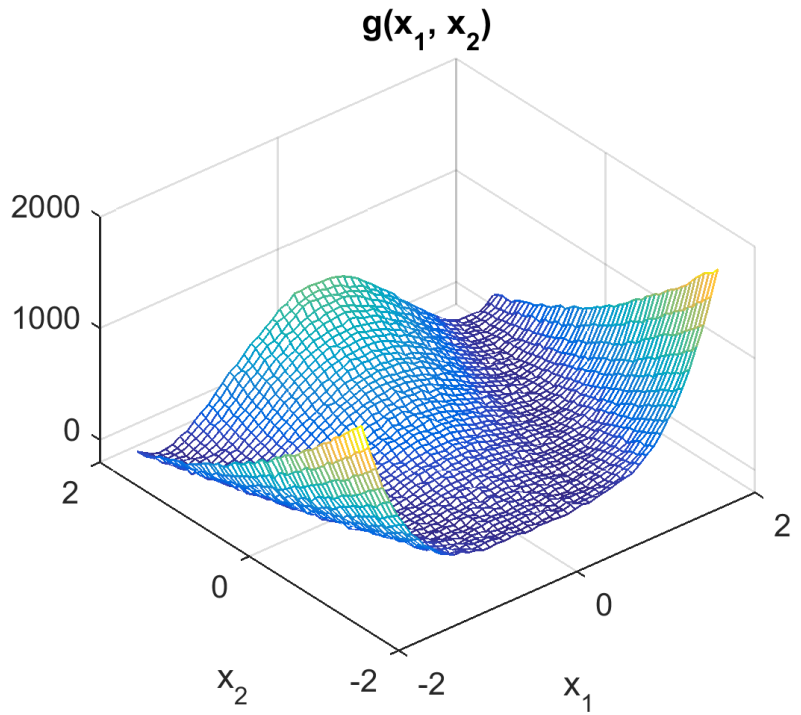
$$Y = \Phi\theta$$

Analytical example: Estimating a scalar (continued)

$$
\begin{aligned}
\widehat{\theta} &= (\Phi^\top \Phi)^{-1} \Phi^\top Y \\
&= \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \\
&= N^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \\
&= \frac{1}{N}(y(1) + \ldots + y(N))
\end{aligned}
$$

Intuition: Estimate is the average of all measurements, filtering out the noise.

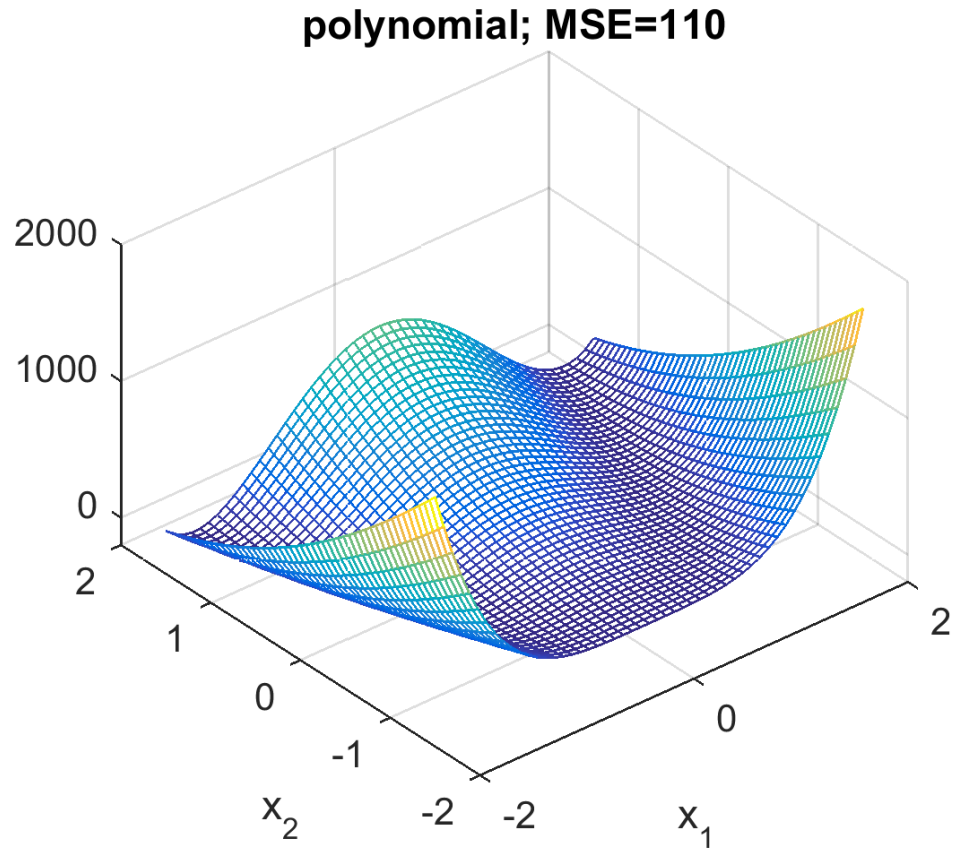
Example: Approximating the "banana" function



- Function $g(x_1, x_2) = (1 - x_1)^2 + 100[(x_2 + 1.5) - x_1^2]^2$, called Rosenbrock's banana function (unknown to the algorithm).
- Approximation data: 200 input points (x_1, x_2), randomly distributed over the space $[-2, 2] \times [-2, 2]$; and corresponding outputs $y = g(x_1, x_2)$, affected by noise.
- Validation data: a uniform grid of 31×31 points over $[-2, 2] \times [-2, 2]$ with their corresponding (noisy) outputs.

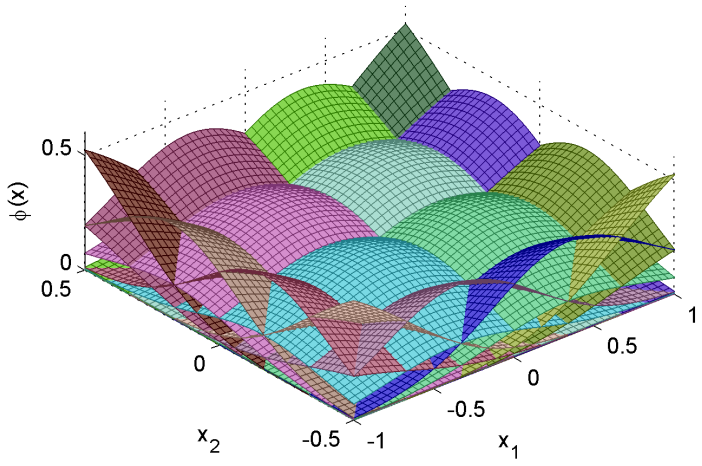
Banana function: Results with polynomial

Polynomial of degree 4 in the two input variables (15 parameters):

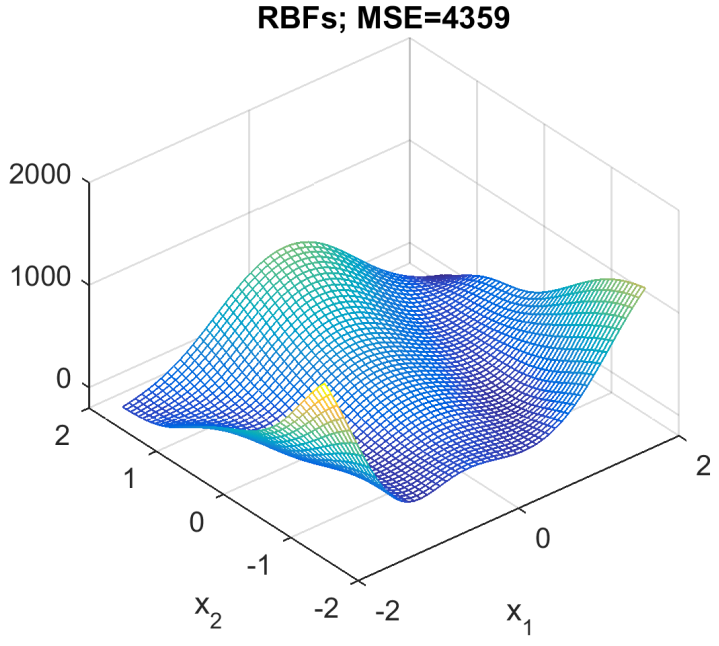


Banana function: Results with radial basis functions

Recall radial basis functions:

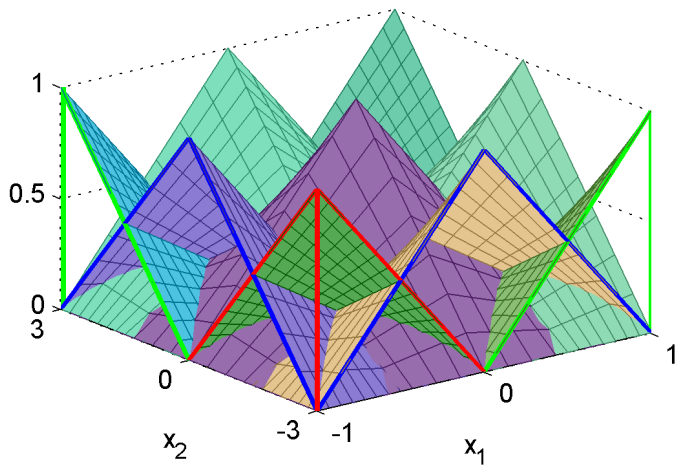


Results with 6×6 RBFs, with centers on an equidistant grid and width equal to the distance between two centers:



Banana function: Results with interpolation

Recall pyramidal basis functions from interpolation:



Results with a 6×6 interpolation grid (corresponding to 6×6 basis functions):

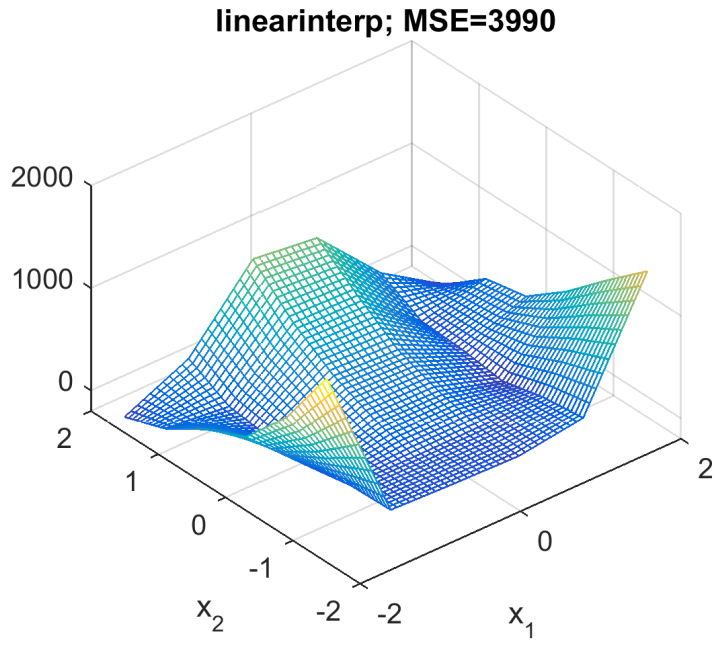


Table of contents

Probability: Formal definition

Preliminary concepts:

- *Outcome* ω, taking possible values in the *sample space* Ω, $\omega \in \Omega$
- *Event* A, defined as a subset of Ω, $A \subseteq \Omega$ (with some extra technical conditions on valid events)

Definition

A probability measure P is a function mapping possible events into probability values in $[0, 1]$, satisfying the conditions:

1. $0 \leq P(A) \leq 1$ (valid probabilities)
2. $P(\Omega) = 1$ (the entire sample space must have probability 1)
3. If events $A_1, \ldots, A_m$ are disjoint, then $P(A_1 \cup A_2 \cup \cdots \cup A_m) = P(A_1) + P(A_2) + \cdots + P(A_m)$. This must hold even as $m \to \infty$.

Much of this section follows Chapter 5 of the SysID lecture notes at Uppsala University, by K. Pelckmans.

Probability: Example

Consider as an example the precipitation on a given day, and to make the definitions precise let h denote precipitation in mm.

- *Sample set*: e.g. $\Omega = \{\text{dry } (h = 0), \text{drizzle } (0 < h \leq 2), \text{rain } (2 < h \leq 10), \text{downpour } (h > 10)\}$, with the *outcomes* ω taking any of these values.
- *Event A*: any one of the outcomes, e.g. $A = \{\text{drizzle}\}$, and in addition any union of outcomes, such as $A = \{\text{drizzle}\} \cup \{\text{rain}\} \cup \{\text{downpour}\}$; call $A = \text{wet}$.

Then, one example of *probability measure* is $P(\{\text{dry}\}) = 0.5$, $P(\{\text{drizzle}\}) = 0.2$, $P(\{\text{rain}\}) = 0.2$, $P(\{\text{downpour}\}) = 0.1$, and we use condition 3 to generate the probabilities of combined events; e.g. $P(\text{wet}) = 0.2 + 0.2 + 0.1 = 0.5$. Note that conditions 1 and 2 ($P(\Omega) = 1$) are satisfied.

Probability: Independence

The *joint probability* of two events A and B is defined as $P(A, B) := P(A \cap B)$.

Definition

Two events A and B are called independent if $P(A, B) = P(A)P(B)$.

Examples:

- The event of rolling a 6 with a dice is independent of the event of getting a 6 at the previous roll (or, indeed, any other value and any previous roll).
- The event of rolling *two consecutive* 6-s, however, is not independent of the previous roll!

(Incidentally, failing to understand the first example leads to the so-called gambler's fallacy. Having just had a long sequence of bad–or good–games at the casino, means nothing for the next game!)

Random variable

Definition

A random variable is a function $X : \Omega \to \mathcal{X}$ defined on the sample set Ω, taking values in some arbitrary space $\mathcal{X}$.

Intuitively, random variables associate interesting quantities to the outcomes. A specific (deterministic) value of X is denoted x. Such a value is called a *realization* of X.

The probability of X taking value x is the probability of all outcomes leading to x:

$$P(X = x) = P(\{\omega \mid X(\omega) = x\})$$

where the first, shorter notation is used for convenience.

Random variable: Example

An urn contains 10 colored balls numbered from 1 to 10; the first 2 balls are white, the others 8 are black. The sample space is $\Omega = \{1, \ldots, 10\}$. Balls are drawn from the urn following a uniform distribution, corresponding to $P(\{i\}) = 1/10, \ \forall i$.

- The *random variable* is the color of the ball, $X : \Omega \to \{\text{white}, \text{black}\}$, defined by $X(1) = X(2) = \text{white}$, $X(3) = \cdots = X(10) = \text{black}$.
- The probability of drawing a white ball is $P(X = \text{white}) = P(\{1, 2\}) = 1/5$.

Discrete random variable

If set $\mathcal{X}$ is discrete, then so is the random variable. Two possibilities:

- $\mathcal{X}$ contains a finite number n of elements
- $\mathcal{X}$ contains infinitely many elements that can be indexed by natural numbers $0, 1, 2, \ldots$ (mathematical term: "countable").

In this case, a sufficient representation of the probability distribution is the PMF:

Definition

The probability mass function (PMF) of X is the list of the probabilities of all individual values $p(x_0), p(x_1), \ldots$.

Example: Ball color is a discrete random variable, with finitely many (two) values, and its PMF is $p(\text{white}) = 1/5$, $p(\text{black}) = 4/5$.

Continuous random variable: Motivation

In the weather example, we wish to characterize the precise quantity of precipitation $h \in [0, h_{\max}]$ where $h_{\max}$ is some reasonable maximum. Assume each value h has equal probability. (For completeness, take sample space $\Omega = [0, h_{\max}]$ so that random variable H is the identity, $H(\omega) = \omega$).

But there are continuously, infinitely many values in the interval $[0, h_{\max}]$, so $P(h)$ must be 0 for any h! (Otherwise, since the probabilities are equal, $P([0, h_{\max}]) \to \infty$ and condition 1 of the probability definition does not hold.) So, we cannot define a meaningful PMF.

Continuous random variable: CDF and PDF

Meaningful probabilities can only be defined for "continuous" subsets.

Definitions

The cumulative distribution function (CDF) of a continuous random variable $X : \Omega \to \mathbb{R}$ is:

$$F(x) := P(X \leq x) = P(\{\omega \mid X(\omega) \leq x\})$$

From the CDF, define the probability density function (PDF):

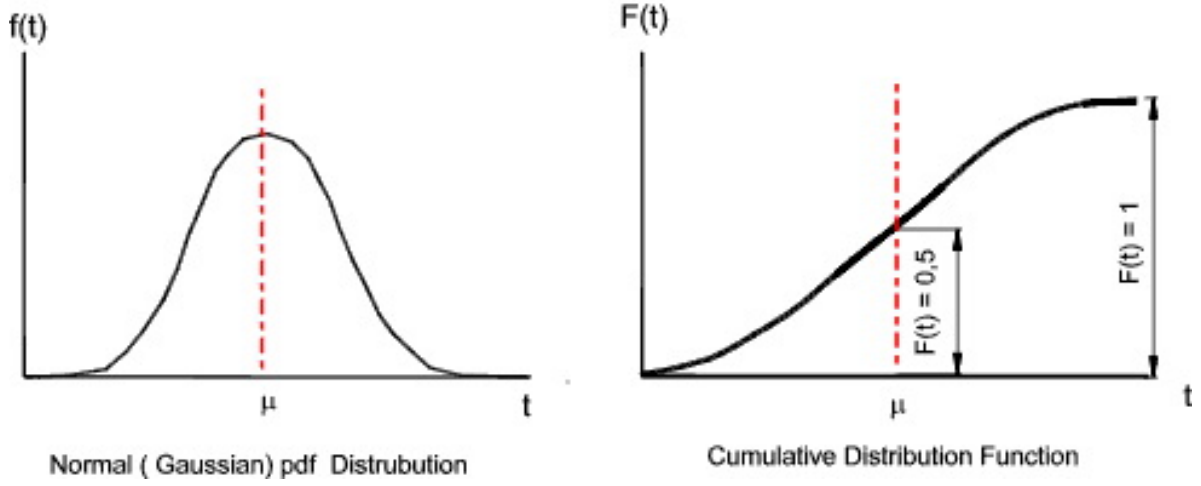
$$f(x) := \frac{dF(x)}{dx}$$

Remarks:

- The PDF is the correspondent to the PMF from the discrete case.
- For any set $Z \subseteq \mathcal{X}$, $P(X \in Z) = \int_{x \in Z} f(x)$ (in the discrete case, $P(X \in Z) = \sum_{x \in Z} P(x)$).

Example: Gaussian

Similar in shape, but not in meaning, with Gaussian basis functions.



Normal (Gaussian) pdf Distrubution

Cumulative Distribution Function

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

Parameters: mean μ, and variance σ^2 (their meaning is clarified later)

The Gaussian distribution arises very often in nature: e.g., distribution of IQs in a human population. It is also called the normal distribution, and denoted $\mathcal{N}(\mu, \sigma^2)$.

Table of contents

Practical probabilities

In engineering, we usually consider *numerical* random variables and often work directly with their PMF $p(x)$ or PDF $f(x)$.

The underlying sample space Ω, outcomes ω, and events A are rarely made explicit.

Expected value

Definition

$$\mathrm{E}\{X\} = \begin{cases} \sum_{x \in \mathcal{X}} p(x)x & \text{for discrete random variables} \\ \int_{x \in \mathcal{X}} f(x)x & \text{for continuous random variables} \end{cases}$$

Intuition: the average of all values, weighted by their probability; the value "expected" beforehand given the probability distribution.

The expected value is also called *mean*, or *expectation*.

Examples:

- For a fair dice with X the value of a face, $\mathrm{E}\{X\} = \frac{1}{6}1 + \frac{1}{6}2 + \ldots + \frac{1}{6}6 = 7/2$.
- If X is distributed with PDF $f(x) = f_{\mathrm{G}}(x)$, Gaussian, then $\mathrm{E}\{X\} = \mu$.

Expected value of a function

Consider a function $g : \mathcal{X} \to \mathbb{R}$, depending on some random variable X. Then $g(X)$ is itself a random variable, and:

$$\mathrm{E}\left\{g(X)\right\} = \begin{cases} \sum_{x \in \mathcal{X}} p(x) g(x) & \text{if discrete} \\ \int_{x \in \mathcal{X}} f(x) g(x) & \text{if continuous} \end{cases}$$

Variance

Definition

$$\operatorname{Var}\{X\} = \boxed{\mathrm{E}\left\{(X - \mathrm{E}\{X\})^2\right\}} = \mathrm{E}\left\{X^2\right\} - (\mathrm{E}\{X\})^2$$

Intuition: the "spread" of the random values around the expectation.

$$\operatorname{Var}\{X\} = \begin{cases} \sum_{x \in \mathcal{X}} p(x)(x - \mathrm{E}\{X\})^2 & \text{if discrete} \\ \int_{x \in \mathcal{X}} f(x)(x - \mathrm{E}\{X\})^2 & \text{if continuous} \end{cases}$$

$$= \begin{cases} \sum_{x \in \mathcal{X}} p(x)x^2 - (\mathrm{E}\{X\})^2 & \text{if discrete} \\ \int_{x \in \mathcal{X}} f(x)x^2 - (\mathrm{E}\{X\})^2 & \text{if continuous} \end{cases}$$

Examples:

- For a fair dice, $\operatorname{Var}\{X\} = \frac{1}{6}1^2 + \frac{1}{6}2^2 + \ldots + \frac{1}{6}6^2 - (7/2)^2 = 35/12$.
- If X is distributed with PDF $f(x) = f_{\mathrm{G}}(x)$, Gaussian, then $\operatorname{Var}\{X\} = \sigma^2$.

Notation

We will generically denote $\mathrm{E}\{X\} = \mu$ and $\mathrm{Var}\{X\} = \sigma^2$.

Quantity $\sigma = \sqrt{\mathrm{Var}\{X\}}$ is called *standard deviation*.

Covariance

Definition

$$\operatorname{Cov}\{X, Y\} = \boxed{\mathrm{E}\{(X - \mathrm{E}\{X\})(Y - \mathrm{E}\{Y\})\}} = \mathrm{E}\{(X - \mu_X)(Y - \mu_Y)\}$$

where μ_X, μ_Y denote the means (expected values) of the two random variables.

Intuition: how much the two variables "change together" (positive if they change in the same way, negative if they change in opposite ways).

Remark: $\operatorname{Var}\{X\} = \operatorname{Cov}\{X, X\}$.

Uncorrelated variables

Definition

Random variables X and Y are uncorrelated if $\mathrm{Cov}\{X, Y\} = 0$. Otherwise, they are correlated.

Examples:

- The education level of a person is correlated with their income.
- Hair color is uncorrelated with income (or at least it should be, ideally).

Remarks:

- If X and Y are independent, they are uncorrelated.
- But the reverse is not necessarily true! Variables can be uncorrelated and still dependent.

Vectors of random variables

Consider a vector $\boldsymbol{X} = [X_1, \ldots, X_N]^\top$ where each X_i is a continuous, real random variable. This vector has a *joint PDF* $f(\boldsymbol{x})$, with $\boldsymbol{x} \in \mathbb{R}^N$.

Definitions

Expected value and *covariance matrix* of $\boldsymbol{X}$:

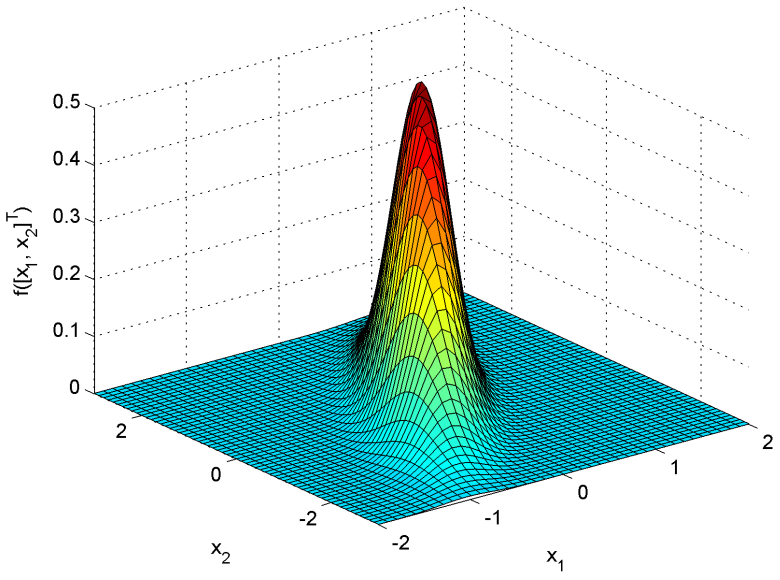
$$\mathrm{E}\{\boldsymbol{X}\} := [\mathrm{E}\{X_1\}, \ldots, \mathrm{E}\{X_N\}]^\top = [\mu_1, \ldots, \mu_N]^\top, \text{ denoted } \boldsymbol{\mu} \in \mathbb{R}^N$$

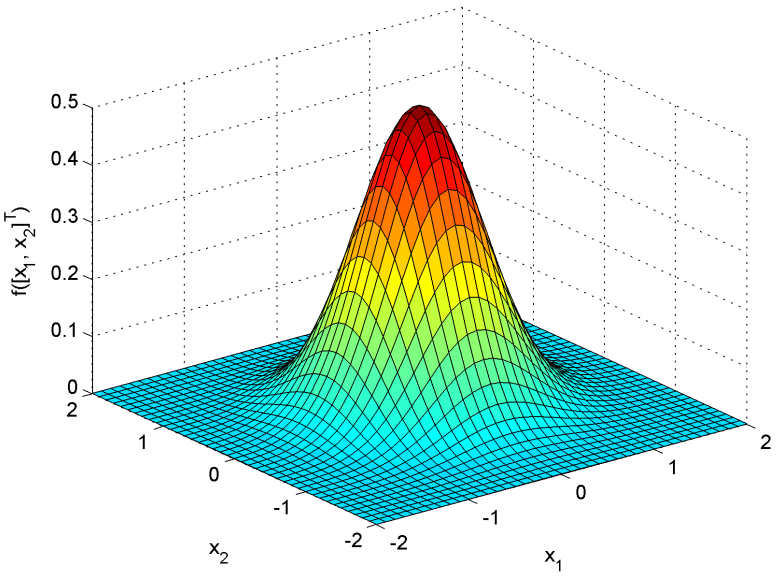
$$\mathrm{Cov}\{\boldsymbol{X}\} := \begin{bmatrix} \mathrm{Cov}\{X_1, X_1\} & \mathrm{Cov}\{X_1, X_2\} & \cdots & \mathrm{Cov}\{X_1, X_N\} \\ \mathrm{Cov}\{X_2, X_1\} & \mathrm{Cov}\{X_2, X_2\} & \cdots & \mathrm{Cov}\{X_2, X_N\} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{Cov}\{X_N, X_1\} & \mathrm{Cov}\{X_N, X_2\} & \cdots & \mathrm{Cov}\{X_N, X_N\} \end{bmatrix}$$

$$= \mathrm{E}\left\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right\}, \text{ denoted } \Sigma \in \mathbb{R}^{N,N}$$

Remarks: $\mathrm{Cov}\{X_i, X_i\} = \mathrm{Var}\{X_i\}$. Also, $\mathrm{Cov}\{X_i, X_j\} = \mathrm{Cov}\{X_j, X_i\}$, so matrix Σ is symmetrical.

Example: Multivariate Gaussian (normal)





The PDF of a vector $\boldsymbol{X}$ with a Gaussian joint distribution can be written:

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

parameterized by the vector mean μ and covariance matrix Σ (assumed positive definite, so $\det(\Sigma) > 0$ and Σ^{-1} exists).

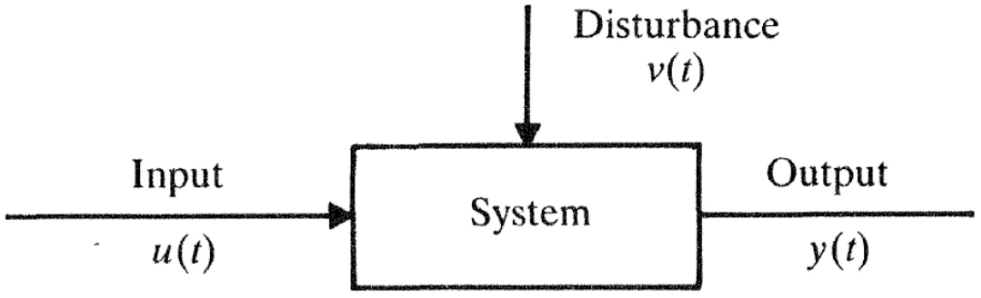
Stochastic process

Definition

A stochastic process $\boldsymbol{X}$ is a sequence of random variables $\boldsymbol{X} = (X_1, \ldots, X_k, \ldots, X_N)$.

It is in fact just a vector of random variables, with the additional structure that the index in the vector has the meaning of time step k.

In system identification, signals (e.g., inputs, outputs) will often be stochastic processes evolving over discrete time steps k.



Zero-mean white noise

Definition

The stochastic process $\boldsymbol{X}$ is zero-mean white noise if: $\forall k$, $\mathrm{E}\{X_k\} = 0$ (zero-mean), and $\forall k, k' \neq k$, $\mathrm{Cov}\{X_k, X_{k'}\} = 0$ (the values at different steps are uncorrelated). In addition, the variance $\mathrm{Var}\{X_k\}$ must be finite $\forall k$.

Stated concisely using vector notation: mean $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{X}\} = \mathbf{0} \in \mathbb{R}^N$ and covariance matrix $\Sigma = \mathrm{Cov}\{\boldsymbol{X}\}$ is diagonal (with positive finite elements on the diagonal).

In system identification, noise processes often affect signal measurements, and we will sometimes assume that the noise is white and zero-mean.

Stationary process

Signal values at different time steps can be correlated (e.g. when they depend on the output of some dynamic system). Nevertheless, signals are sometimes required to be stationary, in the sense:

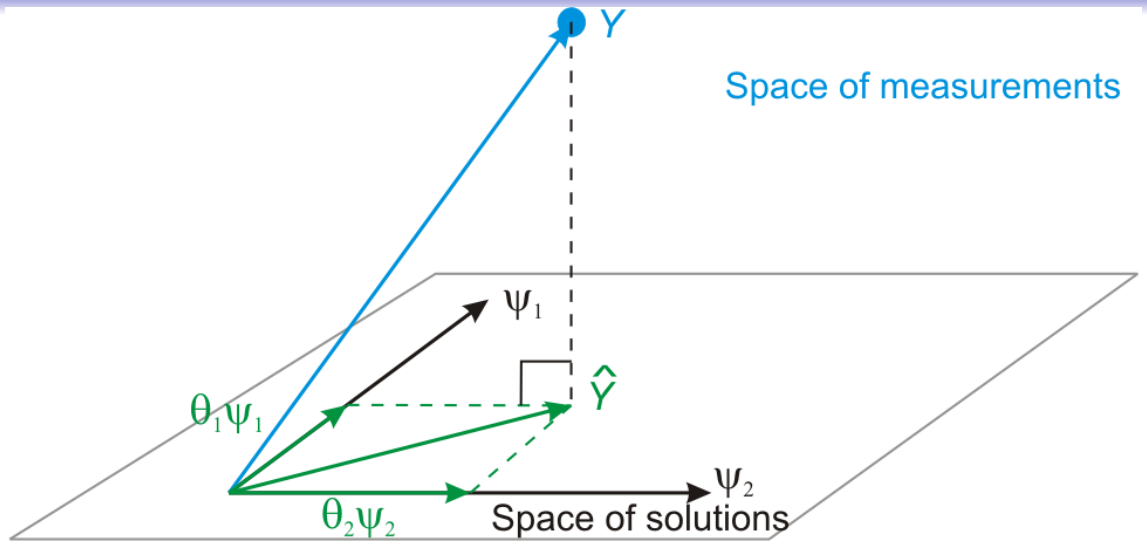
Definition

The stochastic process $\boldsymbol{X}$ is stationary if $\forall k, \mathrm{E}\{X_k\} = \mu$, and $\forall k, k', \tau, \mathrm{Cov}\{X_k, X_{k+\tau}\} = \mathrm{Cov}\{X_{k'}, X_{k'+\tau}\}$.

The mean is the same at every step, whereas the covariance depends on only the relative positions of time steps, not their absolute positions.

Table of contents

Geometrical interpretation



- The space of all measurement vectors Y is an N-dimensional, linear/vector space.
- Denote the ith column of matrix Φ by ψ_i, $i = 1, \ldots, n$. Note $\psi_i = [\varphi_i(1), \ldots, \varphi_i(N)]^\top$.
- Then, the space of solutions representable by the regressors is an n-dimensional subspace spanned by vectors $\psi_1, \ldots, \psi_n$. A solution is achieved by choosing some parameter values $\theta_1, \ldots, \theta_n$ and taking the linear combination $\sum_{i=1}^{n} \theta_i \psi_i$.
- The least-squares solution $\widehat{Y}$ is the projection of the measurement vector Y on this subspace.

Theoretical analysis: Assumptions

1. There exists a true parameter vector θ_0 so that the data satisfy:

$$y(k) = \varphi^\top(k)\theta_0 + e(k)$$

2. The stochastic process $e(k)$ is *zero-mean white noise*, with variance σ^2 at every step.

Intuition: The assumptions say that the true data can be represented by the chosen model, up to some errors that are well-behaved in a statistical sense.

Remark: The new errors $e(k)$ have different meaning from $\varepsilon(k)$ before ($e(k)$ are the ideal errors for the true θ_0, while $\varepsilon(k)$ are the real errors for the parameters θ found in practice).

Theoretical analysis: Guarantees

Theorem

1. The solution $\widehat{\theta}$ of the least-squares problem is *an unbiased estimate* of θ_0. This means: $\mathrm{E}\left\{\widehat{\theta}\right\} = \theta_0$ where the expectation is with respect to the probability distribution of the data.
2. The covariance matrix of the solution is:

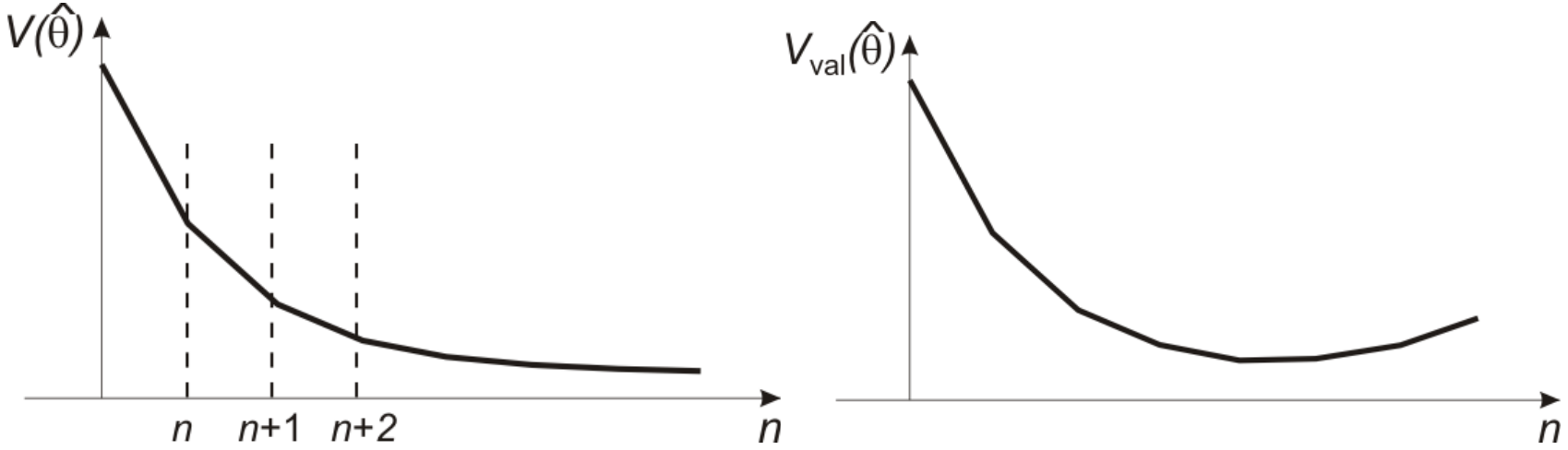
$$\mathrm{Cov}\left\{\widehat{\theta}\right\} = \sigma^2(\Phi^\top\Phi)^{-1}$$

Intuition: Part 1 says that the solution makes (statistical) sense, while Part 2 can be interpreted as measuring the confidence in the solution. E.g., smaller errors $e(k)$ have smaller variance σ^2, which means the covariances are smaller – i.e. better confidence that $\widehat{\theta}$ is close to the true value θ_0.

Remark: σ^2 is unknown, but can be estimated as $\frac{2V(\widehat{\theta})}{N-n}$ (recall that we know $V(\widehat{\theta}) = \frac{1}{2}[Y^\top Y - Y^\top\Phi(\Phi^\top\Phi)^{-1}\Phi^\top Y]$).

Model choice

Assume that given a model size n, we have a way to generate regressors $\varphi(k)$ that make the model more expressive (e.g., basis functions on a finer grid). Then we expect the objective function to behave as follows:



So we can grow n incrementally and stop when there are no significant improvements in V, or the error V_{val} on the validation data starts growing.

Remark: With noisy data, increasing n too much can lead to overfitting: good performance on the training data, but poor performance on other data. Validation on a separate dataset is essential in practice!