Basic math for machine learning

Departament de Teoria
del Senyal i Comunicacions
UPC

CONTENTS

Definition of a random variable

Definition: A random variable X is an application from the result of a random experiment to a real value:

$$X : \Omega \quad \rightarrow \quad \mathbb{R}$$

Characterizacion of a random variable

The density function $f(x)$ charactereizes X:

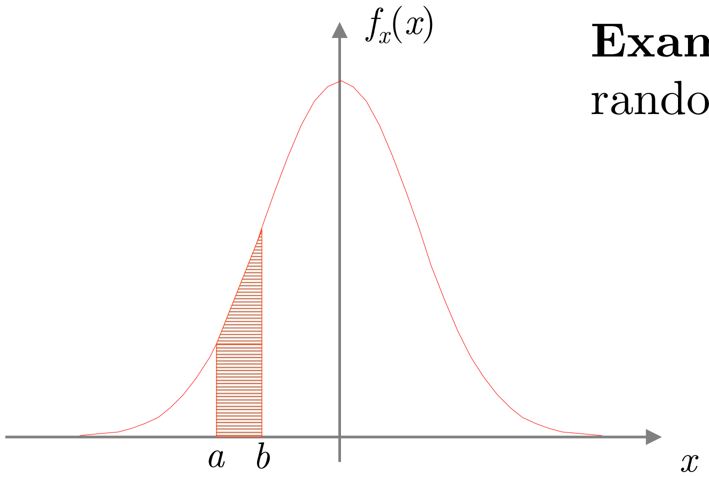
Property 1. It is non-negative.

Property 2. The area under an interval is the probability that X takes values within that interval.

$$\Pr\left(a \le x \le b\right) = \int_a^b f_x(x)dx$$

Property 3. The joint density of several independent variables is the product of density functions.

$$f_{xy}\left(x, y\right) = f_x\left(x\right) f_y\left(y\right)$$



Example 1. $f(x)$ for a real-valued Gaussian random variable

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{\left|x - m_x\right|^2}{2\sigma_x^2}\right) \qquad \text{if } x \in \mathbb{R}$$

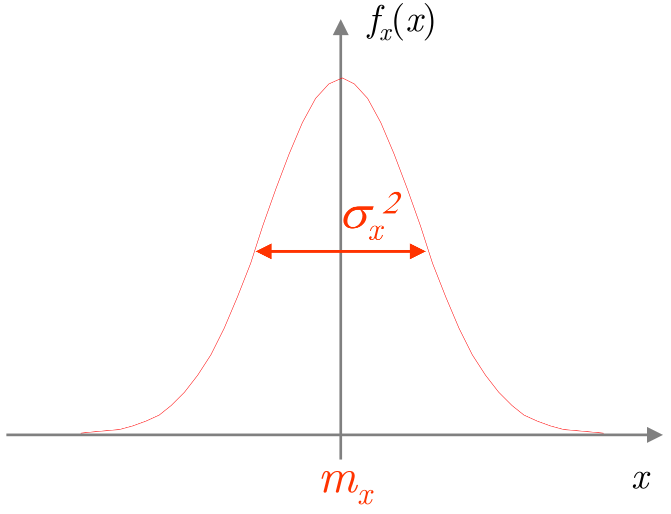
The characterization is simpler if it is built from the statistical moments:

Mean $$m_x = E\{x\} = \int_{-\infty}^{\infty} x f_x(x)\,dx$$ Average value

Power $$P_x = E\left\{|x|^2\right\} = \int_{-\infty}^{\infty} |x|^2 f_x(x)\,dx$$

Cross-correlation $$r_{xy} = E\{xy\} = \int_{-\infty}^{\infty} xy\, f_{xy}(x,y)\,dxdy$$

Variance $$\sigma_x^2 = E\left\{|x-m_x|^2\right\} = \int_{-\infty}^{\infty} |x-m_x|^2 f_x(x)\,dx$$ Dispersion around the mean



Cross-covariance

$$c_{xy}=E\left\{\left(x-E\left\{x\right\}\right)\left(y-E\left\{y\right\}\right)\right\}=\int_{-\infty}^{\infty}\left(x-E\left\{x\right\}\right)\left(y-E\left\{y\right\}\right)f_{xy}(x,y)dxdy$$

If random variables are independent:

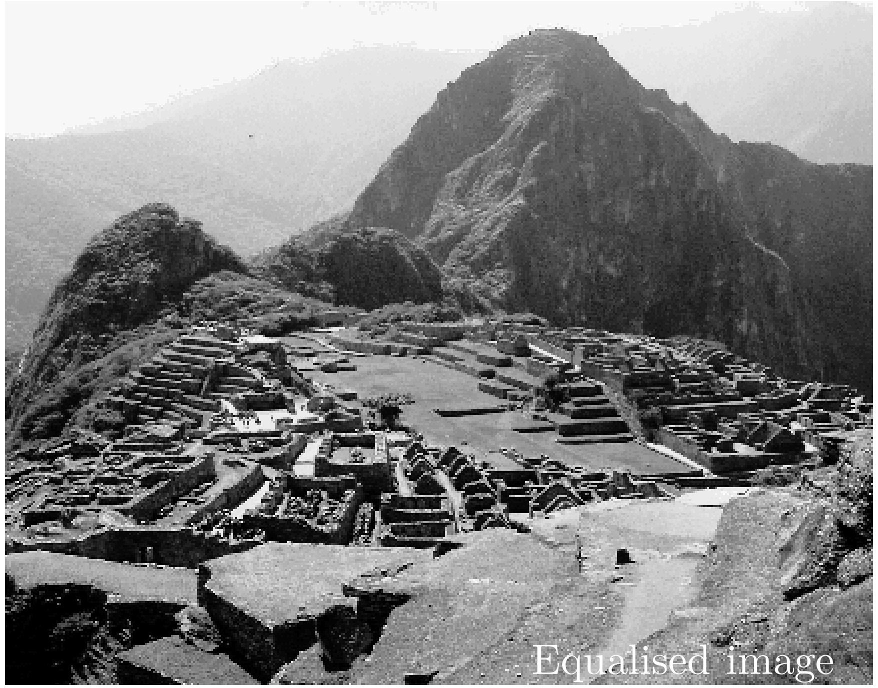
$$c_{xy}=\int_{-\infty}^{\infty}\left(x-E\left\{x\right\}\right)\left(y-E\left\{y\right\}\right)f_{x}(x)f_{y}(y)dxdy=$$

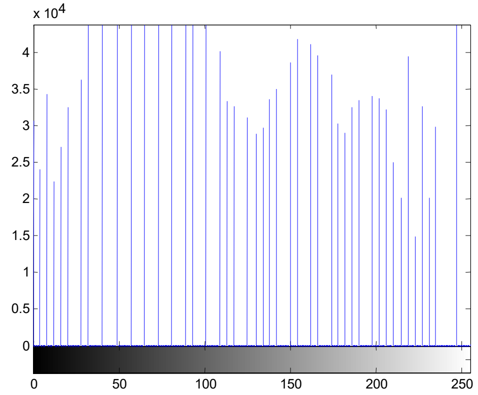
$$=\int_{-\infty}^{\infty}\left(x-E\left\{x\right\}\right)f_{x}(x)dx\int_{-\infty}^{\infty}\left(y-E\left\{y\right\}\right)f_{y}(y)dy=0$$

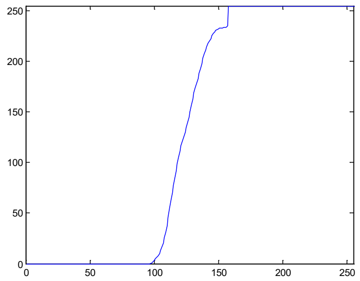
then, they are uncorrelated. The reverse implication is only valid for Gaussian random variables.

Histogram equalisation

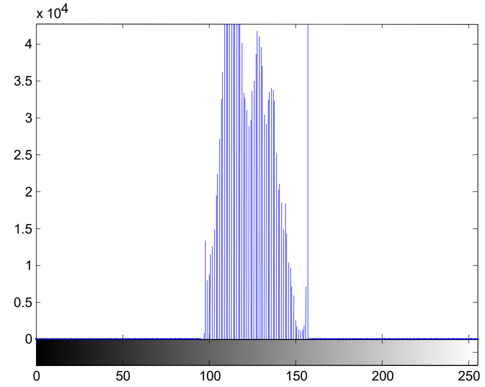
Example 2: Each pixel is interpreted as an observation of a random variable







Cumulative function



$$F_X(x) = \int_{-\infty}^{x} f_X(\lambda)\, d\lambda$$

Characterizacion of a vector random variable

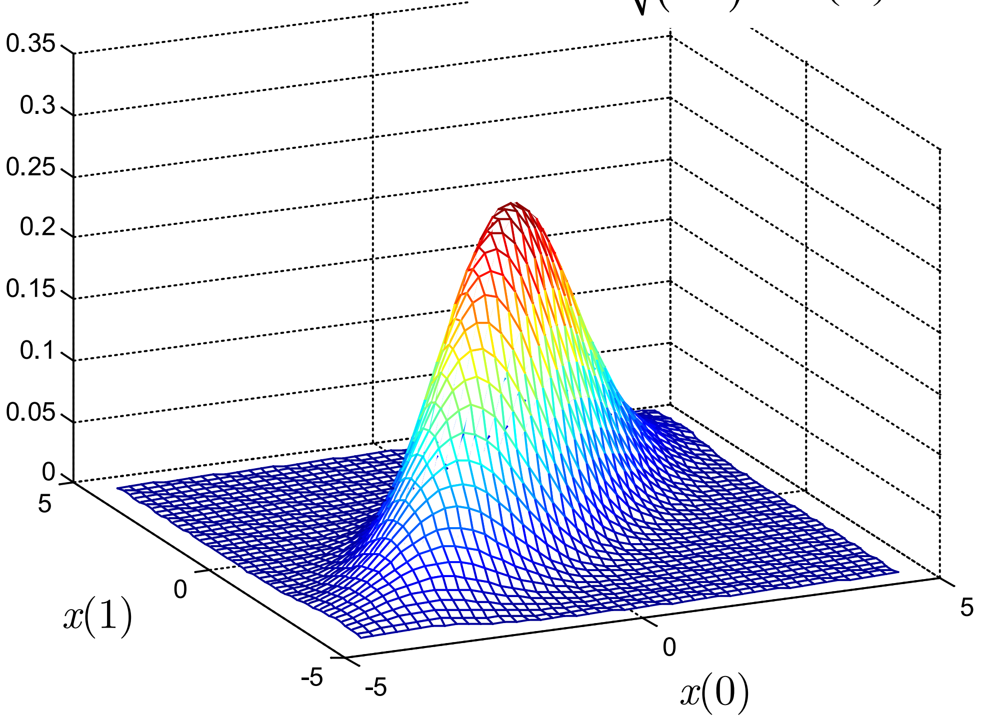
Arrange the random variables in a vector and define a joint density function.

$$\mathbf{x} = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(d) \end{bmatrix}$$

Example 3: Joint Gaussian probability density function for the elements of $\mathbf{x}$

$d=2$

$$f_x(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{C})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right) \quad \text{if } \mathbf{x} \in \mathbb{R}^{d\times 1}$$



The first and second order moments are now defined as

Mean $$\mathbf{m}_x = E\{\mathbf{x}\} = \begin{bmatrix} E\{x(1)\} \\ E\{x(2)\} \\ \vdots \\ E\{x(d)\} \end{bmatrix}$$ Average value

Correlation matrix

$$\mathbf{R}_x = E\{\mathbf{x}\mathbf{x}^T\} = \begin{bmatrix} E\{|x(1)|^2\} & E\{x(1)x(2)\} & \cdots & E\{x(1)x(N)\} \\ E\{x(2)x(1)\} & E\{|x(2)|^2\} & \cdots & E\{x(2)x(N)\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x(N)x(1)\} & E\{x(N)x(2)\} & \cdots & E\{|x(N)|^2\} \end{bmatrix}$$

Contains all possible correlations and cross-correlations between the elements of vector $\mathbf{x}$

Covariance matrix

$$\mathbf{C}_x = E\left\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\right\} =$$

$$= \begin{bmatrix} E\left\{|x(1)-m(1)|^2\right\} & E\left\{(x(1)-m(1))(x(2)-m(2))\right\} & \cdots & \\ E\left\{(x(2)-m(2))(x(1)-m(1))\right\} & E\left\{|x(2)-m(2)|^2\right\} & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ E\left\{(x(N)-m(N))(x(1)-m(1))\right\} & & \cdots & E\left\{|x(N)-m(N)|^2\right\} \end{bmatrix} =$$

$$= \mathbf{R}_x - \mathbf{m}\mathbf{m}^T$$

Contains all possible covariances and cross-covariances between the elements of vector $\mathbf{x}$

Correlation function

$$\mathbf{R}_x = E\left\{\mathbf{x}\mathbf{x}^H\right\} = \begin{bmatrix} E\left\{|x(1)|^2\right\} & E\left\{x(1)x(2)\right\} & \cdots & E\left\{x(1)x(N)\right\} \\ E\left\{x(2)x(1)\right\} & E\left\{|x(2)|^2\right\} & \cdots & E\left\{x(2)x(N)\right\} \\ \vdots & \vdots & \ddots & \vdots \\ E\left\{x(N)x(1)\right\} & E\left\{x(N)x(2)\right\} & \cdots & E\left\{|x(N)|^2\right\} \end{bmatrix}$$

$$\mathbf{r}_x = \begin{bmatrix} r_x(0) \\ r_x(1) \\ \vdots \\ r_x(N-1) \end{bmatrix}$$

Property 1. Hermitian symmetry $r_x(k) = r_x(-k)$
Property 2. Its maximum is in $k = 0$, and it is the power of $x(n)$

Cross-correlation

$$\mathbf{R}_{xy} = E\left\{\mathbf{x}\mathbf{y}^T\right\} = \begin{bmatrix} E\{x(1)y(1)\} & \cdots & E\{x(1)y(k)\} & \cdots & E\{x(1)y(N)\} \\ E\{x(2)y(1)\} & \cdots & E\{x(2)y(k)\} & \cdots & E\{x(2)y(N)\} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ E\{x(N)y(1)\} & \cdots & E\{x(N)y(k)\} & \cdots & E\{x(N)y(N)\} \end{bmatrix}$$

$$\mathbf{r}_{xy} = E\left\{\mathbf{x}y(k)\right\} = \begin{bmatrix} E\{x(1)y(k)\} \\ E\{x(2)y(k)\} \\ \vdots \\ E\{x(N)y(k)\} \end{bmatrix}$$

Cross-correlation is a measure of similarity between random variables: the larger it is, the lower is the error.

$$\begin{aligned} MSE &= E\left\{\left|y(n) - x(n+k)\right|^2\right\} = \\ &= r_x(0) + r_y(0) - 2E\left\{x(n+k)y(n)\right\} = \\ &= r_x(0) + r_y(0) - 2r_{xy}(k) \geq 0 \end{aligned}$$

CONTENTS

Notation (I)

x	:	Random variable
$x(n)$	:	Temporal sequence
$X(f)$	:	Fourier transform of a temporal sequence
$\mathbf{x}$, $\underline{x}$	:	Column vector
$\mathbf{x}^T$, $\underline{x}^T$	:	Row vector, transpose of $\mathbf{x}$

$$\Rightarrow \quad \mathbf{x} \in C^N \; ; \mathbf{x} = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(N) \end{bmatrix}$$

$\mathbf{X}$, $\underline{X}$	:	Matrix
$\mathbf{X}^T$, $\underline{X}^T$	:	Transpose matrix $\mathbf{X}$

$\mathbf{x}^T\mathbf{y}$: Scalar product between vectores $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x}^T\mathbf{y} = \sum_{i=0}^{N-1} x(i)y(i)$$

Notation (II)

$\mathbf{x}\,\mathbf{y}^T$: Outer product between vectors $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x}\mathbf{y}^T = \begin{bmatrix} x(1)y(1) & x(1)y(2) & \cdots & x(1)y(N) \\ x(2)y(1) & x(2)y(2) & \cdots & x(2)y(N) \\ \vdots & \vdots & \ddots & \vdots \\ x(N)y(1) & x(N)y(2) & \cdots & x(N)y(N) \end{bmatrix}$$

$\mathbf{C}\,\mathbf{y}$: Matrix-vector product

$$\mathbf{C}\mathbf{y} = \begin{bmatrix} \mathbf{c}_0^T \\ \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_{N-1}^T \end{bmatrix} \mathbf{y} = \begin{bmatrix} \sum_{i=0}^{N-1} c(0,i)y(i) \\ \sum_{i=0}^{N-1} c(1,i)y(i) \\ \vdots \\ \sum_{i=0}^{N-1} c(N-1,i)y(i) \end{bmatrix}$$

Norms and Schwarz's inequality

The norm is defined using the scalar product $< \cdot \, , \cdot >$

- Given a escalar product $< \cdot \, , \cdot >$, the norm is $\quad \|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$

- Examples… $\quad \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T \mathbf{x}, \quad \|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{N} |x_i|^2}$

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}\left(\mathbf{Y}^T \mathbf{X}\right), \quad \|\mathbf{X}\| = \sqrt{\mathrm{Tr}\left(\mathbf{X}^T \mathbf{X}\right)} = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{M} |x_{i,j}|^2} = \|\mathbf{X}\|_F$$

$$\langle x(t), y(t) \rangle = \int x(t) y(t) dt, \quad \|x(t)\| = \sqrt{\int |x(t)|^2 \, dt} = \sqrt{E_x}$$

$$\langle X, Y \rangle = E\{X \, Y\}, \quad \|X\| = \sqrt{E\left\{|X|^2\right\}} = \sqrt{\sigma_X^2 + m_X^2}$$

Schwarz's inequality (valid in all cases):

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \|\mathbf{x}\|^2 \, \|\mathbf{y}\|^2$$

Equality condition is satisfied when…

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 = \|\mathbf{x}\|^2 \, \|\mathbf{y}\|^2 \qquad \Leftrightarrow \qquad \exists k, \quad \mathbf{x} = k \cdot \mathbf{y}$$

Matrix operators

Trace

$$\mathbf{A} \in \mathbb{R}^{N\times N} \quad \Rightarrow \quad \mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{N} [\mathbf{A}]_{i,i}$$

$$\mathbf{A} \in \mathbb{R}^{N\times M}, \quad \mathbf{B} \in \mathbb{R}^{M\times N} \quad \Rightarrow \quad \mathrm{Tr}(\mathbf{AB}) = \sum_{i=1}^{N} [\mathbf{AB}]_{i,i} = \mathrm{Tr}(\mathbf{BA}) = \sum_{i=1}^{M} [\mathbf{BA}]_{i,i}$$

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{BCA}) = \mathrm{Tr}(\mathbf{CAB})$$

Determinant

$$\mathbf{A} \in \mathbb{R}^{N\times N}, \quad \mathbf{B} \in \mathbb{R}^{N\times N} \quad \Rightarrow \quad \det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$$

$$\det(\mathbf{ABC}) = \det(\mathbf{A})\det(\mathbf{B})\det(\mathbf{C})$$

Frobenius' (or Euclidean) norm

$$\mathbf{x} \in \mathbb{R}^{N\times 1} \quad \Rightarrow \quad \|\mathbf{x}\|_F = \sqrt{\sum_{i=1}^{N} |x_i|^2} = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\mathrm{Tr}\left(\mathbf{x}\mathbf{x}^T\right)}$$

$$\mathbf{X} \in \mathbb{R}^{N\times M} \quad \Rightarrow \quad \|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M} |x_{i,j}|^2} = \sqrt{\mathrm{Tr}\left(\mathbf{X}^T\mathbf{X}\right)} = \sqrt{\mathrm{Tr}\left(\mathbf{X}\mathbf{X}^T\right)}$$

Inverse of a matrix product

$$\mathbf{A} \in \mathbb{R}^{N\times N}, \quad \mathbf{B} \in \mathbb{R}^{N\times N} \quad \Rightarrow \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Inversion lemma (Woodbury's identity)

$$\left(\mathbf{A} + \mathbf{UCV}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

... particular case

$$\left(\mathbf{A} + k\mathbf{uu}^{H}\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{uu}^{H}\mathbf{A}^{-1}}{\frac{1}{k} + \mathbf{u}^{H}\mathbf{A}^{-1}\mathbf{u}}$$

Eigenvalues and eigenvectors of a matrix

The eigenvectors of a matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ are those vectors whose norm is not altered after being transformed:

$$\mathbf{A}\mathbf{q} = \lambda \mathbf{q}$$

$$\Downarrow$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{q} = \mathbf{0}$$

$$\Downarrow$$

$$P(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

This equation is a polynomial in λ (characteristic polynomial of $\mathbf{A}$), that has N roots. Hence, $\mathbf{A}$ has N eigenvalues $\lambda_i \quad i=1,\ldots,N$ (possibly multiple).

Not all matrices can be diagonalised using the eigenvectors. Only in those cases where each eigenvalue is associated to a vector space (generated by the eigenvectors) of dimension equal to the multiplicity of the eigenvalue.

Properties of the correlation and covariance matrices

P1. $\mathbf{R}_x$ is symmetric $\mathbf{R} = \mathbf{R}^T, \qquad \mathbf{R} \in \mathbb{R}^{N \times N}$

P2. $\mathbf{R}_x$ is positive semidefinite: $\mathbf{y}^T \mathbf{R}_x \mathbf{y} \geq 0 \qquad \forall \mathbf{y}$

P3. The eigenvalues of $\mathbf{R}_x$ are real and non-negative.

P4. The trace of a matrix is the sum of its eigevalues.

P5. The eigenvalues of $\mathbf{R}_x$ are bounded by the maximum and the minimum of of the power spectral density of $x(n)$.

P6. The eigenvectors of $\mathbf{R}_x$ are orthogonal, so we can write $\mathbf{R}_x = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$

P7. $\mathbf{R}_x^{-1} = \mathbf{R}_x^{-T}$

Symmetric matrices

A square matrix is symmetric if $\mathbf{R} = \mathbf{R}^T, \qquad \mathbf{R} \in \mathbb{R}^{N\times N}$

Spectral decomposition theorem

A symmetric matrix $\mathbf{R}$ can always diagonalize using a base of orthonormal eigenvectors with real eigenvalues.

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T, \qquad \lambda_i \mathbf{q}_i = \mathbf{R}\mathbf{q}_i, \qquad \lambda_i \in \mathbb{R}, \qquad \|\mathbf{q}_i\| = 1$$

$$\mathbf{Q} = \left[\mathbf{q}_1 \quad \mathbf{q}_2 \quad \cdots \quad \mathbf{q}_N\right] \in \mathbb{R}^{N\times N}, \quad \mathbf{q}_i \in \mathbb{R}^{N\times 1}, \quad \mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \quad \mathbf{Q}^{-1} = \mathbf{Q}^T$$

$$\mathbf{\Lambda} = \mathrm{diag}\left[\lambda_1, \lambda_2, \ldots, \lambda_N\right] \qquad \text{with } \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$$

Positive semidefinite matrices: $\mathbf{v}^T \mathbf{R} \mathbf{v} \geq 0, \quad \forall \mathbf{v} \neq \mathbf{0} \in \mathbb{R}^{N \times 1}$

- ✓ A symmetric matrix is positive semidefinite if all eigenvalues are positive: $\lambda_i \geq 0, \quad \forall i$

CONTENTS

Constrained optimization

Maximization or minimization of $f(\mathbf{x})$ with restrictions on $\mathbf{x} \in \mathbb{R}^{M\times 1}$:

$$\text{optimize } f(\mathbf{x}) \text{ with } g_i(\mathbf{x}) \leq 0 \text{ and/or } g_j(\mathbf{x}) = 0 \qquad i, j = 1, \ldots, K$$

Computation of a solution based on the method of Lagrange multipliers:

1. Build the Lagrangian function:

$$L\left(\mathbf{x}, \{\mu_i\}\right) = f(\mathbf{x}) + \sum_{i=1}^{K} \mu_i g_i(\mathbf{x})$$

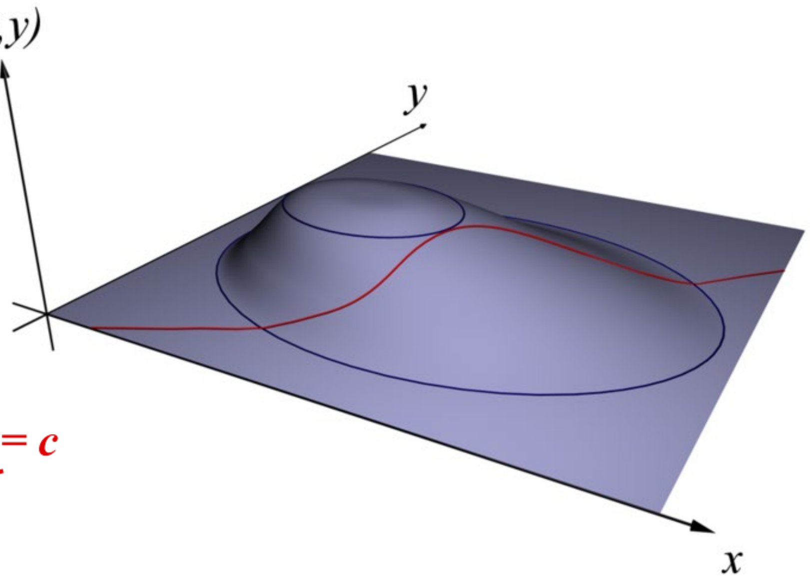
2. Equate the gradient to zero and solve the equation to obtain the possible values of $\mathbf{x}$, that depend on the Lagrange multipliers λ_i

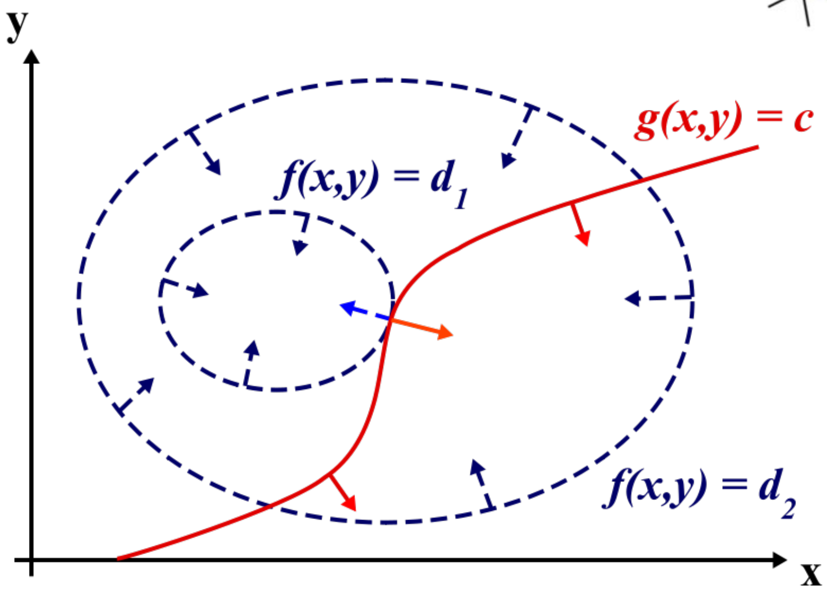
$$\nabla_{\mathbf{x}} L\left(\mathbf{x}, \{\mu_i\}\right) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{K} \mu_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) = \mathbf{0}$$

3. Select the solution among the possible values of $\mathbf{x}$ for which $f(\mathbf{x})$ is maximum/minimum and all restrictions are satisfied simultaneously.

Constrained optimization

The derivation of the Lagrangian provides the sufficient condition if the function is concave/convex. As an illustration: when only one restriction is set, in the optimum point, the gradient of $f(\mathbf{x})$ and $g(\mathbf{z})$ are parallel...





CONTENTS

Derivation of scalar functions of real variables

In many optimization problems we are interested in determining the gradient of a real scalar function with a vector variable:

$$\nabla_{\mathbf{h}} f(\mathbf{h}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{h})}{\partial h(1)} \\ \vdots \\ \dfrac{\partial f(\mathbf{h})}{\partial h(L)} \end{bmatrix}$$

Some useful cases are...

$$\nabla_{\mathbf{h}} \mathbf{a}^T \mathbf{h} = \mathbf{a}$$

$$\nabla_{\mathbf{h}} \mathbf{h}^T \mathbf{a} = \mathbf{a}$$

$$\nabla_{\mathbf{h}} \mathbf{h}^T \mathbf{h} = 2\mathbf{h}$$

$$\nabla_{\mathbf{h}} \mathbf{h}^T \mathbf{R} \mathbf{h} = \begin{cases} 2\mathbf{R}\mathbf{h} & \text{si } \mathbf{R}^T = \mathbf{R} \\ \left(\mathbf{R} + \mathbf{R}^T\right)\mathbf{h} & \text{si } \mathbf{R}^T \neq \mathbf{R} \end{cases}$$