LECTURE 2
Vectors and Functions

Mathematicians are like Frenchmen: whatever you say to them, they translate into their own language, and turn it into something entirely different.

Goethe

Outline

- Introduction
  - Basics
  - Examples
  - Vector spaces
- Inner product, angle, orthogonality
- Projections
- Functions and maps
  - Hyperplanes and halfspaces
  - Gradients

Introduction

- A vector is a collection of numbers, arranged in a column or a row, which can be thought of as the coordinates of a point in n-dimensional space.
- Equipping vectors with sum and scalar multiplication allows to define notions such as independence, span, subspaces, and dimension. Further, the scalar product introduces a notion of angle between two vectors, and induces the concept of length, or norm.
- Via the scalar product, we can also view a vector as a linear function. We can compute the projection of a vector onto a line defined by another vector, onto a plane, or more generally onto a subspace.
- Projections can be viewed as a first elementary optimization problem (finding the point in a given set at minimum distance from a given point), and they constitute a basic ingredient in many processing and visualization techniques for high-dimensional data.
We usually write vectors in column format:

\[ x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \]

Element \( x_i \) is said to be the \( i \)-th component (or the \( i \)-th element, or entry) of vector \( x \), and the number \( n \) of components is usually referred to as the dimension of \( x \).

- When the components of \( x \) are real numbers, i.e. \( x_i \in \mathbb{R} \), then \( x \) is a real vector of dimension \( n \), which we indicate with the notation \( x \in \mathbb{R}^n \).
- We shall seldom need complex vectors, which are collections of complex numbers \( x_i \in \mathbb{C} \), \( i = 1, \ldots, n \). We denote the set of such vectors by \( \mathbb{C}^n \).
- To transform a column-vector \( x \) in row format and vice versa, we define an operation called transpose, denoted with a superscript \( ^\top \):

\[ x^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}; \quad x^{\top\top} = x. \]

### Examples

#### Example 1 (Bag-of-words representations of text)

Consider the following text:

"A (real) vector is just a collection of real numbers, referred to as the components (or, elements) of the vector; \( \mathbb{R}^n \) denotes the set of all vectors with \( n \) elements. If \( x \in \mathbb{R}^n \) denotes a vector, we use subscripts to denote elements, so that \( x_i \) is the \( i \)-th component of \( x \). Vectors are arranged in a column, or a row. If \( x \) is a column vector, \( x^\top \) denotes the corresponding row vector, and vice-versa."

- Row vector \( c = [5, 3, 4] \) contains the number of times each word in the list \( V = \{ \text{vector}, \text{elements}, \text{of}, \text{the} \} \) appears in the above paragraph.
- Dividing each entry in \( c \) by the total number of occurrences of words in the list (15, in this example), we obtain a vector \( x = \begin{bmatrix} 1/3 & 1/5 & 1/5 & 4/15 \end{bmatrix} \) of relative word frequencies.
- Frequency-based representation of text documents (bag-of-words).

#### Example 2 (Time series)

A time series represents the evolution in (discrete) time of a physical or economical quantity.

- If \( x(k), k = 1, \ldots, T \), describes the numerical value of the quantity of interest at time \( k \), then the whole time series, over the time horizon from 1 to \( T \), can be represented as a \( T \)-dimensional vector \( x \) containing all the values of \( x(k) \), for \( k = 1 \) to \( k = T \), that is

\[ x = [x(1) \ x(2) \ \cdots \ x(T)]^\top \in \mathbb{R}^T. \]

Adjusted close price of the Dow Jones Industrial Average Index, over a 66 days period from April 19, 2012 to July 20, 2012.

#### Example 3 (Images)

We are given a gray-scale image where each pixel has a certain value representing the level of darkness. We can arrange the image as a vector of pixels.

![Row vector representation of an image.](image-url)
Vector spaces

- The operations of sum, difference and scalar multiplication are defined in an obvious way for vectors: for any two vectors \( v^{(1)}, v^{(2)} \) having equal number of elements, we have that the sum \( v^{(1)} + v^{(2)} \) is simply a vector having as components the sum of the corresponding components of the addends, and the same holds for the difference.
- If \( v \) is a vector and \( \alpha \) is a scalar (i.e., a real or complex number), then \( \alpha v \) is obtained multiplying each component of \( v \) by \( \alpha \). If \( \alpha = 0 \), then \( \alpha v \) is the zero vector, or origin.
- A vector space \( X \), is obtained by equipping vectors with the operations of addition and multiplication by a scalar.

A simple example of a vector space is \( X = \mathbb{R}^n \), the space of \( n \)-tuples of real numbers. A less obvious example is the set of single-variable polynomials of a given degree.

Subspaces and span

- A nonempty subset \( V \) of a vector space \( X \) is called a subspace of \( X \) if, for any scalars \( \alpha, \beta \),
  \[ x, y \in V \implies \alpha x + \beta y \in V. \]
  In other words, \( V \) is “closed” under addition and scalar multiplication.
- A linear combination of a set of vectors \( S = \{x^{(1)}, \ldots, x^{(m)}\} \) in a vector space \( X \) is a vector of the form \( \alpha_1 x^{(1)} + \cdots + \alpha_m x^{(m)} \), where \( \alpha_1, \ldots, \alpha_m \) are given scalars.
- The set of all possible linear combinations of the vectors in \( S = \{x^{(1)}, \ldots, x^{(m)}\} \) forms a subspace, which is called the subspace generated by \( S \), or the span of \( S \), denoted with \( \text{span}(S) \).
- Given two subspaces \( X, Y \) in \( \mathbb{R}^n \), the direct sum of \( X, Y \), which we denote by \( X \oplus Y \), is the set of vectors of the form \( x + y \), with \( x \in X, y \in Y \). It is readily checked that \( X \oplus Y \) is itself a subspace.

Bases and dimensions

- A collection \( x^{(1)}, \ldots, x^{(m)} \) of vectors in a vector space \( X \) is said to be linearly independent if no vector in the collection can be expressed as a linear combination of the others. This is the same as the condition
  \[ \sum_{i=1}^m \alpha_i x^{(i)} = 0 \iff \alpha = 0. \]
- Given a subspace \( S \) of a vector space \( X \), a basis of \( S \) is a set \( B \) of vectors of minimal cardinality, such that \( \text{span}(B) = S \). The cardinality of a basis is called the dimension of \( S \).
- If we have a basis \( \{x^{(1)}, \ldots, x^{(d)}\} \) for a subspace \( S \), then we can write any element in the subspace as a linear combination of elements in the basis. That is, any \( x \in S \) can be written as
  \[ x = \sum_{i=1}^d \alpha_i x^{(i)}, \]
  for appropriate scalars \( \alpha_i \).

Affine sets

- An affine set is a set of the form \( A = \{x \in X : x = t + v^{(0)}, \ v \in Y\} \), where \( v^{(0)} \) is a given point and \( Y \) is a given subspace of \( X \). Subspaces are just affine spaces containing the origin.
- Geometrically, an affine set is a flat passing through \( v^{(0)} \). The dimension of an affine set \( A \) is defined as the dimension of its generating subspace \( Y \).
- A line is a one-dimensional affine set. The line through \( x_0 \) along direction \( \omega \) is the set
  \[ L = \{x \in X : x = x_0 + \lambda v, \ v \in \text{span}(\omega)\}, \]
  where in this case \( \text{span}(\omega) = \{\lambda v : \lambda \in \mathbb{R}\} \).
Euclidean length

The Euclidean length of a vector $x \in \mathbb{R}^n$ is the square-root of the sum of squares of the components of $x$, that is

$$\text{Euclidean length of } x = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$  

This formula is an obvious extension to the multidimensional case of the Pythagoras theorem in $\mathbb{R}^2$.

The Euclidean length represents the actual distance to be "travelled" for reaching point $x$ from the origin 0, along the most direct way (the straight line passing through 0 and $x$).

Basics

Norms and $\ell_p$ norms

A norm on a vector space $\mathcal{X}$ is a real-valued function with special properties that maps any element $x \in \mathcal{X}$ into a real number $\|x\|$.

Definition 1

A function from $\mathcal{X}$ to $\mathbb{R}$ is a norm if

$$\|x\| \geq 0 \quad \forall x \in \mathcal{X},$$

and $\|x\| = 0$ if and only if $x = 0$;

$$\|x + y\| \leq \|x\| + \|y\|,$$  

for any $x, y \in \mathcal{X}$ (triangle inequality);

$$\|\alpha x\| = |\alpha| \|x\|,$$  

for any scalar $\alpha$ and any $x \in \mathcal{X}$.

$\ell_p$ norms are defined as

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p}, \quad 1 \leq p < \infty.$$  

For $p = 2$ we obtain the standard Euclidean length

$$\|x\|_2 = \sqrt{\sum_{k=1}^{n} x_k^2},$$

or $p = 1$ we obtain the sum-of-absolute-values length

$$\|x\|_1 = \sum_{k=1}^{n} |x_k|.$$  

The limit case $p = \infty$ defines the $\ell_\infty$ norm (max absolute value norm, or Chebyshev norm)

$$\|x\|_\infty = \max_{k=1,\ldots,n} |x_k|.$$  

The cardinality of a vector $x$ is often called the $\ell_0$ (pseudo) norm and denoted with $\|x\|_0$.

Inner product

An inner product on a (real) vector space $\mathcal{X}$ is a real-valued function which maps any pair of elements $x, y \in \mathcal{X}$ into a scalar denoted as $\langle x, y \rangle$. The inner product satisfies the following axioms: for any $x, y, z \in \mathcal{X}$ and scalar $\alpha$

$$\langle x, x \rangle \geq 0;$$

$$\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle;$$

$$\langle \alpha x, y \rangle = \alpha \langle x, y \rangle;$$

$$\langle x, y \rangle = \langle y, x \rangle.$$  

A vector space equipped with an inner product is called an inner product space.

The standard inner product defined in $\mathbb{R}^n$ is the "row-column" product of two vectors

$$\langle x, y \rangle = x^\top y = \sum_{k=1}^{n} x_k y_k.$$  

The inner product induces a norm: $\|x\| = \sqrt{\langle x, x \rangle}$. 

Notes
Angle between vectors

The angle between \( x \) and \( y \) is defined via the relation
\[
\cos \theta = \frac{x^\top y}{\|x\| \|y\|}
\]

When \( x^\top y = 0 \), the angle between \( x \) and \( y \) is \( \theta = \pm 90^\circ \), i.e., \( x, y \) are orthogonal.

When the angle \( \theta \) is \( 0^\circ \), or \( \pm 180^\circ \), then \( x \) is aligned with \( y, \) that is \( y = \alpha x \), for some scalar \( \alpha \), i.e., \( x \) and \( y \) are parallel. In this situation \( |x^\top y| \) achieves its maximum value \( |\alpha| \|x\|^2 \).

Cauchy-Schwartz and Hölder inequality

Since \( |\cos \theta| \leq 1 \), it follows from the angle equation that
\[
|x^\top y| \leq \|x\| \|y\|
\]
and this inequality is known as the Cauchy-Schwartz inequality.

A generalization of this inequality involves general \( \ell_p \) norms and it is known as the Hölder inequality.

For any vectors \( x, y \in \mathbb{R}^n \) and for any \( p, q \geq 1 \) such that \( 1/p + 1/q = 1 \), it holds that
\[
|x^\top y| \leq \sum_{i=1}^n |x_i y_i| \leq \|x\|_p \|y\|_q.
\]

Maximization of inner product over norm balls

Our first optimization problem:
\[
\max_{\|x\|_p \leq 1} x^\top y.
\]

For \( p = 2 \):
\[
x^*_2 = \frac{y}{\|y\|},
\]
hence \( \max_{\|x\|_2 \leq 1} x^\top y = \|y\| \).

For \( p = \infty \):
\[
x^*_\infty = \text{sgn}(y),
\]
and \( \max_{\|x\|_\infty \leq 1} x^\top y = \sum_{i=1}^n |x_i| = |y|_1 \).

For \( p = 1 \):
\[
[x^*_i] = \begin{cases} 
\text{sgn}(y_i) & \text{if } i = m, \\
0 & \text{otherwise},
\end{cases} \quad i = 1, \ldots, n.
\]
We thus have \( \max_{\|x\|_1 \leq 1} x^\top y = \max |x_i| = |y|_\infty \).

Orthogonal vectors

Generalizing the concept of orthogonality to generic inner product spaces, we say that two vectors \( x, y \) in an inner product space \( \mathcal{X} \) are orthogonal if \( \langle x, y \rangle = 0 \).

Orthogonality of two vectors \( x, y \in \mathcal{X} \) is symbolized by \( x \perp y \).

Nonzero vectors \( x^{(1)}, \ldots, x^{(d)} \) are said to be mutually orthogonal if \( \langle x^{(i)}, x^{(j)} \rangle = 0 \) whenever \( i \neq j \). In words, each vector is orthogonal to all other vectors in the collection.

Proposition 1

Mutually orthogonal vectors are linearly independent.

A collection of vectors \( S = \{x^{(1)}, \ldots, x^{(d)}\} \) is said to be orthonormal if, for \( i, j = 1, \ldots, d \),
\[
\langle x^{(i)}, x^{(j)} \rangle = \begin{cases} 
0 & \text{if } i \neq j, \\
1 & \text{if } i = j.
\end{cases}
\]
In words, \( S \) is orthonormal if every element has unit norm, and all elements are orthogonal to each other. A collection of orthonormal vectors \( S \) forms an orthonormal basis for the span of \( S \).
Orthogonal complement

- A vector \( x \in X \) is orthogonal to a subset \( S \) of an inner product space \( X \) if \( x \perp s \) for all \( s \in S \).
- The set of vectors in \( X \) that are orthogonal to \( S \) is called the orthogonal complement of \( S \), and it is denoted with \( S^\perp \).

Theorem 1 (Orthogonal decomposition)

If \( S \) is a subspace of an inner-product space \( X \), then any vector \( x \in X \) can be written in a unique way as the sum of an element in \( S \) and one in the orthogonal complement \( S^\perp \):

\[ X = S \oplus S^\perp \]

for any subspace \( S \subset X \).

Projections

- The idea of projection is central in optimization, and it corresponds to the problem of finding a point on a given set that is closest (in norm) to a given point.
- Given a vector \( x \) in an inner product space \( X \) (say, e.g., \( X = \mathbb{R}^2 \)) and a closed set \( S \subset X \), the projection of \( x \) onto \( S \), denoted as \( \Pi_S(x) \), is defined as the point in \( S \) at minimal distance from \( x \):

\[
\Pi_S(x) = \arg \min_{y \in S} \| y - x \|,
\]

where the norm used here is the norm induced by the inner product, that is \( \| y - x \| = \sqrt{\langle y - x, y - x \rangle} \).
- This simply reduces to the Euclidean norm, when using the standard inner product, in which case the projection is called Euclidean projection.

Theorem 2 (Projection Theorem)

Let \( X \) be an inner product space, let \( x \) be a given element in \( X \), and let \( S \) be a subspace of \( X \). Then, there exists a unique vector \( x^* \in S \) which is solution to the problem

\[
\min_{y \in S} \| y - x \|.
\]

Moreover, a necessary and sufficient condition for \( x^* \) being the optimal solution for this problem is that

\[ x^* \in S, \quad (x - x^*) \perp S. \]

Corollary 1 (Projection on affine set)

Let \( X \) be an inner product space, let \( x \) be a given element in \( X \), and let \( A = x^{(0)} + S \) be the affine set obtained by translating a given subspace \( S \) by a given vector \( x^{(0)} \). Then, there exists a unique vector \( x^* \in A \) which is solution to the problem

\[
\min_{y \in A} \| y - x \|.
\]

Moreover, a necessary and sufficient condition for \( x^* \) to be the optimal solution for this problem is that

\[ x^* \in A, \quad (x - x^*) \perp S. \]
Let $p \in \mathbb{R}^n$ be a given point. We want to compute the Euclidean projection $p^*$ of $p$ onto a line $L = \{x_0 + \text{span}(u)\}$, $\|u\| = 1$:

$$p^* = \arg \min_{x \in L} \|x - p\|_2.$$

Since any point $x \in L$ can be written as $x = x_0 + v$, for some $v \in \text{span}(u)$, the above problem is equivalent to finding a value $v^*$ for $v$, such that

$$v^* = \arg \min_{v \in \text{span}(u)} \|v - (p - x_0)\|_2.$$

The solution must satisfy the orthogonality condition $(z - v^*) \perp u$. Recalling that $v^* = \lambda^* u$ and $u \perp \|u\| = 1$, we hence have

$$u^T z - u^T v^* = 0 \iff u^T z - \lambda^* = 0 \iff \lambda^* = u^T z = u^T (p - x_0).$$

The optimal point $p^*$ is thus given by

$$p^* = x_0 + v^* = x_0 + \lambda^* u = x_0 + u^T (p - x_0) u.$$

The squared distance from $p$ to the line is

$$\|p - p^*\|^2 = \|p - x_0\|^2 - \lambda^2 = \|p - x_0\|^2 - (u^T (p - x_0))^2.$$
Projections

Example: Senate voting data

Random direction
Average direction ($u \sim \text{all ones}$)

Clearly some directions are more informative than others . . .

Projections

Euclidean projection of a point onto an hyperplane

- A hyperplane is an affine set defined as
  \[ H = \{ x \in \mathbb{R}^n : a^T x = b \}, \]
  where $a \neq 0$ is called a normal direction of the hyperplane, since for any two vectors $z_1, z_2 \in H$ it holds that $(z_1 - z_2) \perp a$.
- Given $p \in \mathbb{R}^n$ we want to determine the Euclidean projection $p^*$ of $p$ onto $H$.
- The projection theorem requires $p - p^*$ to be orthogonal to $H$. Since $a$ is a direction orthogonal to $H$, the condition $(p - p^*) \perp H$ is equivalent to saying that $p - p^* = \alpha a$, for some $\alpha \in \mathbb{R}$.

To find $\alpha$, consider that $p^* \in H$, thus $a^T p^* = b$, and multiply the previous equation on the left by $a^T$, obtaining

\[ a^T p - b = \alpha |a|^2, \]

whereby

\[ \alpha = \frac{a^T p - b}{|a|^2}, \]

and

\[ p^* = p - \frac{a^T p - b}{|a|^2} a. \]

The distance from $p$ to $H$ is

\[ \| p - p^* \| = |x| \| a \| = \frac{|a^T p - b|}{|a|^2}. \]

Projections

Projection on a vector span

- Suppose we have a basis for a subspace $S \subseteq \mathbb{R}^n$, that is
  \[ S = \text{span}(x^{(1)}, \ldots, x^{(d)}). \]
- Given $x \in \mathbb{R}^n$, the Projection Theorem states that the unique projection $x^*$ of $x$ onto $S$ is characterized by $(x - x^*) \perp S$.
- Since $x^* \in S$, we can write $x^*$ as some (unknown) linear combination of the elements in the basis of $S$, that is
  \[ x^* = \sum_{i=1}^d \alpha_i x^{(i)}. \]
  Then $(x - x^*) \perp S \iff (x - x^*, x^{(k)}) = 0$, $k = 1, \ldots, d$.
  \[ \sum_{i=1}^d \alpha_i (x^{(i)}, x^{(k)}) = (x^{(k)}, x), \quad k = 1, \ldots, d. \]
- Solving this system of linear equations provides the coefficients $\alpha_i$, and hence the desired $x^*$. 

Notes
Projections

Projection onto the span of orthonormal vectors
- If we have an orthonormal basis for a subspace \( S = \text{span}(S) \), then it is immediate to obtain the projection \( x^* \) of \( x \) onto that subspace.
- This is due to the fact that, in this case, the Gram system of equations immediately gives the coefficients 
  \[ a_i = \langle x^{(i)}, x \rangle, \quad i = 1, \ldots, d. \]
- Therefore, we have that 
  \[ x^* = \sum_{i=1}^{d} \langle x^{(i)}, x \rangle x^{(i)}. \]
- Given a basis \( S = \{x^{(1)}, \ldots, x^{(d)}\} \) for a subspace \( S = \text{span}(S) \), there are numerical procedures to construct an orthonormal basis for the same subspace (e.g., the Gram-Schmidt procedure and QR factorization).

Functions and maps
- A function takes a vector argument in \( \mathbb{R}^n \), and returns a unique value in \( \mathbb{R} \).
- We use the notation \( f: \mathbb{R}^n \to \mathbb{R} \) to refer to a function with "input" space \( \mathbb{R}^n \). The "output" space for functions is \( \mathbb{R} \).
- For example, the function \( f: \mathbb{R}^2 \to \mathbb{R} \) with values 
  \[ f(x) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \]
gives the Euclidean distance from the point \((x_1, x_2)\) to a given point \((y_1, y_2)\).
- We allow functions to take infinity values. The domain of a function \( f \), denoted \( \text{dom} f \), is defined as the set of points where the function is finite.

Sets related to functions
- Consider a function \( f: \mathbb{R}^* \to \mathbb{R} \).
- The graph and the epigraph of a function \( f: \mathbb{R}^n \to \mathbb{R} \) are both subsets of \( \mathbb{R}^{n+1} \).
- The graph of \( f \) is the set of input-output pairs that \( f \) can attain, that is:
  \[ \text{graph} f = \{ (x, f(x)) \in \mathbb{R}^{n+1} : x \in \mathbb{R}^n \}. \]
- The epigraph, denoted \( \text{epi} f \), describes the set of input-output pairs that \( f \) can achieve, as well as "anything above":
  \[ \text{epi} f = \{ (x, t) \in \mathbb{R}^{n+1} : x \in \mathbb{R}^n, \; t \geq f(x) \}. \]
Sets related to functions

- A level set (or contour line) is the set of points that achieve exactly some value for the function $f$. For $t \in \mathbb{R}$, the $t$-level set of the function $f$ is defined as

$$C_t(t) = \{x \in \mathbb{R}^n : f(x) = t\}.$$

- The $t$-sublevel set of $f$ is the set of points that achieve at most a certain value for $f$:

$$L_t(t) = \{x \in \mathbb{R}^n : f(x) \leq t\}.$$  

Linear and affine functions

- Linear functions are functions that preserve scaling and addition of the input argument.
- A function $f : \mathbb{R}^n \to \mathbb{R}$ is linear if and only if

$$\forall x \in \mathbb{R}^n \text{ and } \alpha \in \mathbb{R}, \quad f(\alpha x) = \alpha f(x);$$

$$\forall x_1, x_2 \in \mathbb{R}^n, f(x_1 + x_2) = f(x_1) + f(x_2).$$

- A function $f$ is affine if and only if the function $\tilde{f}(x) = f(x) - f(0)$ is linear (affine = linear + constant).
- Consider the functions $f_1, f_2, f_3 : \mathbb{R}^2 \to \mathbb{R}$ defined below:

$$f_1(x) = 3.2x_1 + 2x_2,$$

$$f_2(x) = 3.2x_1 + 2x_2 + 0.15,$$

$$f_3(x) = 0.001x_2^2 + 2.3x_1 + 0.3x_2.$$

The function $f_1$ is linear; $f_2$ is affine; $f_3$ is neither linear nor affine ($f_3$ is a quadratic function).

Hyperplanes and halfspaces

- A hyperplane in $\mathbb{R}^n$ is a set of the form

$$H_a = \{x \in \mathbb{R}^n : a^\top x = b\},$$

where $a \in \mathbb{R}^n$, $a \neq 0$, and $b \in \mathbb{R}$ are given.

- Equivalently, we can think of hyperplanes as the level sets of linear functions.
- When $b = 0$, the hyperplane is simply the set of points that are orthogonal to $a$ (i.e., $H$ is a $(n-1)$-dimensional subspace).
Hyperplanes and halfspaces

- An hyperplane \( H \) separates the whole space in two regions:
  \[
  H_- = \{x : \alpha^\top x \leq b\}, \quad H_+ = \{x : \alpha^\top x > b\}.
  \]
- These regions are called halfspaces (\( H_- \) is a closed halfspace, \( H_+ \) is an open halfspace).
- The halfspace \( H_- \) is the region delimited by the hyperplane \( \{a \mid x = b\} \) and lying in the direction opposite to vector \( \alpha \). Similarly, the halfspace \( H_+ \) is the region lying above (i.e., in the direction of \( \alpha \)) the hyperplane.

Gradients

- The gradient of a function \( f : \mathbb{R}^n \to \mathbb{R} \) at a point \( x \) where \( f \) is differentiable, denoted with \( \nabla f(x) \), is a column vector of first derivatives of \( f \) with respect to \( x_1, \ldots, x_n \):
  \[
  \nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^\top.
  \]
- When \( n = 1 \) (there is only one input variable), the gradient is simply the derivative.
- An affine function \( f : \mathbb{R}^n \to \mathbb{R} \), represented as \( f(x) = \alpha^\top x + b \), has a very simple gradient: \( \nabla f(x) = \alpha \).

Example 4

The distance function \( \rho(x) = \|x - p\| = \sqrt{\sum_{i=1}^n (x_i - p_i)^2} \) has gradient
\[
\nabla \rho(x) = \frac{1}{\|x - p\|^2} (x - p).
\]

Affine approximation of nonlinear functions

- A non-linear function \( f : \mathbb{R}^n \to \mathbb{R} \) can be approximated locally via an affine function, using a first-order Taylor series expansion.
- Specifically, if \( f \) is differentiable at point \( x_0 \), then for all points \( x \) in a neighborhood of \( x_0 \), we have that
  \[
  f(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + \epsilon(x),
  \]
  where the error term \( \epsilon(x) \) goes to zero faster than first order, as \( x \to x_0 \), that is
  \[
  \lim_{x \to x_0} \frac{\epsilon(x)}{\|x - x_0\|^2} = 0.
  \]
- In practice, this means that for \( x \) sufficiently close to \( x_0 \), we can write the approximation
  \[
  f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0).
  \]

Geometric interpretation of the gradient

- The gradient of a function can be interpreted in the context of the level sets.
  - Indeed, geometrically, the gradient of \( f \) at a point \( x_0 \) is a vector \( \nabla f(x_0) \) perpendicular to the contour line of \( f \) at level \( \alpha = f(x_0) \), pointing from \( x_0 \) outwards the \( \alpha \)-sublevel set (that is, it points towards higher values of the function).
Geometric interpretation of the gradient

- The gradient $\nabla f(x_0)$ also represents the direction along which the function has the maximum rate of increase (steepest ascent direction).
- Let $v$ be a unit direction vector (i.e., $|v| = 1$), let $\epsilon \geq 0$, and consider moving away at distance $\epsilon$ from $x_0$ along direction $v$, that is, consider a point $x = x_0 + \epsilon v$. We have that
  
  $f(x_0 + \epsilon v) \approx f(x_0) + \epsilon \nabla f(x_0)^\top v$, for $\epsilon \to 0$.

  or, equivalently,
  
  $\lim_{\epsilon \to 0} \frac{f(x_0 + \epsilon v) - f(x_0)}{\epsilon} = \nabla f(x_0)^\top v$.

- Whenever $\epsilon > 0$ and $v$ is such that $\nabla f(x_0)^\top v > 0$, then $f$ is increasing along the direction $v$, for small $\epsilon$.
- The inner product $\nabla f(x_0)^\top v$ measures the rate of variation of $f$ at $x_0$, along direction $v$, and it is usually referred to as the directional derivative of $f$ along $v$. [131x579]f(x_0) = \text{const.} = f(x_0)$

Notes

- The rate of variation is thus zero, if $v$ is orthogonal to $\nabla f(x_0)$: along such a direction the function value remains constant (to first order), that is, this direction is tangent to the contour line of $f$ at $x_0$.
- Contrary, the rate of variation is maximal when $v$ is parallel to $\nabla f(x_0)$, hence along the normal direction to the contour line at $x_0$. [131x579]f(x) = \text{const.} = f(x_0)$

Notes