Convex Functions
Part 1: Linear Algebra Review

Korea University

Spring Semester

Linear Algebra review

- Vector space
- Basis/Dimension
- Nullspace
- Range
- Rank
- Determinant
- ... and more to cover as we move on

Vector Space

Vector space

- a vector space $\mathcal{V}$ consists of
  - A set of vectors
  - Addition operator
  - multiplication with scalar
  - special element 0 vector

Vector space

- a vector space $\mathcal{V}$ consists of
  - A set of vectors
  - Addition operator
  - multiplication with scalar
  - special element 0 vector
- Example:
  - $\mathcal{V}_1 = \mathbb{R}^n$
  - $\mathcal{V}_2 = \{0\}$
  - $\mathcal{V}_3 = \mathrm{span}(v_1, \ldots, v_k)$ with $v_1, \ldots, v_k \in \mathbb{R}^n$ where

$$\mathrm{span}(v_1, \ldots, v_k) = \{c_1 v_1 + \cdots + c_k v_k | c_1, \ldots, c_k \in \mathbb{R}\}$$

Subspace

- Subspace of a vector space is i) subset of a vector space and ii) itself is a vector space
- $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3$ are subspaces

independent set of vectors

- we say vectors $v_1, \ldots, v_k$ are linearly independent when

$$c_1 v_1 + \cdots + c_k v_k = 0 \Rightarrow c_1 = \cdots = c_k = 0$$

- The only way to make the linear combinations of linearly independent vectors is to make all the coefficients zero

independent set of vectors

- we say vectors $v_1, \ldots, v_k$ are linearly independent when

$$c_1 v_1 + \cdots + c_k v_k = 0 \Rightarrow c_1 = \cdots = c_k = 0$$

- The only way to make the linear combinations of linearly independent vectors is to make all the coefficients zero
- No vector v_i, $1 \leq i \leq k$, can be expressed as linear combination of other vectors

independent set of vectors

- we say vectors $v_1, \ldots, v_k$ are linearly independent when

$$c_1 v_1 + \cdots + c_k v_k = 0 \Rightarrow c_1 = \cdots = c_k = 0$$

- The only way to make the linear combinations of linearly independent vectors is to make all the coefficients zero
- No vector v_i, $1 \leq i \leq k$, can be expressed as linear combination of other vectors
- Not to be confused with orthogonality of vectors
  - If $v_1, \ldots, v_k$ are mutually orthogonal, they are linearly independent
  - converse is not necessarily true

Basis and Dimension

- set of vectors $\{v_1, \dots, v_k\}$ is a basis of vector space $\mathcal{V}$ if
  1. $\{v_1, \dots, v_k\}$ spans $\mathcal{V}$, or
  $$\mathcal{V} = \mathrm{span}(v_1, \dots, v_k)$$
  2. $v_1, \dots, v_k$ are linearly independent
- Any point $x \in \mathcal{V}$ can be uniquely expressed as
  $$c_1 v_1 + \cdots + c_k v_k$$
  for some $c_1, \dots, c_k$

Basis and Dimension

- set of vectors $\{v_1, \ldots, v_k\}$ is a basis of vector space $\mathcal{V}$ if
  1. $\{v_1, \ldots, v_k\}$ spans $\mathcal{V}$, or
  $$\mathcal{V} = \operatorname{span}(v_1, \ldots, v_k)$$
  2. $v_1, \ldots, v_k$ are linearly independent
- Any point $x \in \mathcal{V}$ can be uniquely expressed as
  $$c_1 v_1 + \cdots + c_k v_k$$
  for some $c_1, \ldots, c_k$
- For given vector space $\mathcal{V}$ and any of its basis, the number of vectors in the basis is fixed
- The number of basis vectors is called **dimension** of $\mathcal{V}$, denoted by $\mathbf{dim}(\mathcal{V})$

Basis and Dimension

- By default, we let **dim**$(\{0\}) = 0$ (in other mathematical definition of dimensions, a single point other than 0 is also defined to have 0 dimension)

Basis and Dimension

- By default, we let $\textbf{dim}(\{0\}) = 0$ (in other mathematical definition of dimensions, a single point other than 0 is also defined to have 0 dimension)
- Examples: consider $V_1 = \{\alpha v | \alpha \in \mathbb{R}\}$ for some $v \in \mathbb{R}^n$
  - V_1 represents a line going through origin, and is parallel to v
  - V_1 is a subspace of $\mathbb{R}^n$: it is a subset of $\mathbb{R}^n$, and is vector space, and contains $\{0\}$
  - Dimension of V_1 is 1, although it contains a point from $\mathbb{R}^n$!

Basis and Dimension

- By default, we let $\mathbf{dim}(\{0\}) = 0$ (in other mathematical definition of dimensions, a single point other than 0 is also defined to have 0 dimension)
- Examples: consider $V_1 = \{\alpha v | \alpha \in \mathbb{R}\}$ for some $v \in \mathbb{R}^n$
  - V_1 represents a line going through origin, and is parallel to v
  - V_1 is a subspace of $\mathbb{R}^n$: it is a subset of $\mathbb{R}^n$, and is vector space, and contains $\{0\}$
  - Dimension of V_1 is 1, although it contains a point from $\mathbb{R}^n$!
- Consider $v_1, v_2 \in \mathbb{R}^3$ where v_1, v_2 are linearly independent. Plane $V_2 = \{\alpha_1 v_1 + \alpha_2 v_2 | \alpha_1, \alpha_2 \in \mathbb{R}\}$ goes through the origin and is a subspace with dimension 2

Basis and Dimension

- By default, we let $\mathbf{dim}(\{0\}) = 0$ (in other mathematical definition of dimensions, a single point other than 0 is also defined to have 0 dimension)
- Examples: consider $V_1 = \{\alpha v | \alpha \in \mathbb{R}\}$ for some $v \in \mathbb{R}^n$
  - V_1 represents a line going through origin, and is parallel to v
  - V_1 is a subspace of $\mathbb{R}^n$: it is a subset of $\mathbb{R}^n$, and is vector space, and contains $\{0\}$
  - Dimension of V_1 is 1, although it contains a point from $\mathbb{R}^n$!
- Consider $v_1, v_2 \in \mathbb{R}^3$ where v_1, v_2 are linearly independent. Plane $V_2 = \{\alpha_1 v_1 + \alpha_2 v_2 | \alpha_1, \alpha_2 \in \mathbb{R}\}$ goes through the origin and is a subspace with dimension 2
- But note the same plane can be expressed as $\{v \in \mathbb{R}^3 | c^T v = 0\}$ using some vector $c \in \mathbb{R}^n$ **orthogonal** to vectors on the plane!

Matrix vector multiplication

- Useful things to know
- Let $A \in \mathbb{R}^{m \times n}$ and

$$A = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

where a_i is the ith column of A and $x = (x_1, \dots, x_n)$ then

$$Ax = x_1 a_1 + \cdots + x_n a_n = \sum_{i=1}^{n} x_i a_i$$

That is, it is linear combination of columns

Matrix vector multiplication

- Let $A \in \mathbb{R}^{m \times n}$ and

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \cdots \\ a_m^T \end{bmatrix}$$

where a_i^T is the ith row of A and

$$x^T A = x_1 a_1^T + \cdots + x_m a_m^T = \sum_{i=1}^{m} x_i a_i^T$$

That is, it is linear combination of rows

Matrix matrix multiplication

- Let $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{r \times n}$

$$AB = A \begin{bmatrix} b_1 & b_2 & \dots & b_n \end{bmatrix} = \begin{bmatrix} Ab_1 & Ab_2 & \dots & Ab_n \end{bmatrix}$$

or

$$AB = \begin{bmatrix} a_1^T \\ a_2^T \\ \dots \\ a_m^T \end{bmatrix} B = \begin{bmatrix} a_1^T B \\ a_2^T B \\ \dots \\ a_m^T B \end{bmatrix}$$

Range

- Range of a matrix $A \in \mathbb{R}^{m \times n}$, denoted by $\mathcal{R}(A)$ is defined as

$$\mathcal{R}(A) = \left\{Ax | x \in \mathbb{R}^n\right\}$$

  - $\mathcal{R}(A)$ is equivalent to $\operatorname{span}(a_1, \ldots, a_n)$ where $a_i \in \mathbb{R}^m$ are columns of A
  - That is, $\mathcal{R}(A)$ is the subspace (subset of $\mathbb{R}^M$) spanned by columns of A
- set of vectors 'hit' by linear mapping $y = Ax$
- set of vectors such that, for given y in $\mathcal{R}(A)$, equation $Ax - y = 0$ w.r.t. x has solution

Range: interpretation

- let $v \in \mathcal{R}(A)$ and $w \notin \mathcal{R}(A)$
- let $y = Ax$ output of a sensor to input x
  - $y = v$ is possible/consistent output
  - $y = w$ is impossible/inconsistent
- $\mathcal{R}(A)$ represents achievable outputs
- $\mathcal{R}(A)$ is subspace
- suppose $\mathcal{R}(A) = \mathbb{R}^m$
  - any output $y \in \mathbb{R}^m$ is possible

Range examples

- $$A_1 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$
- For any $v \in \mathbb{R}^2$, we have

  $$Av = (v_1 + 2v_2, 2v_1 + 4v_2) = c(1, 2)$$

  for some constant c Thus, $\mathcal{R}(A) = \{c(1, 2) | c \in \mathbb{R}\}$ and is a subspace with dimension of 1
- $$A_2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$
- it turns out that whole $\mathbb{R}^2$ can be mapped onto with Ax for $x \in \mathbb{R}^2$
- Thus $\mathcal{R}(A) = \mathbb{R}^2$ and has dimension 2

Nullspace

- Nullspace of a matrix $A \in \mathbb{R}^{m \times n}$, denoted by $\mathcal{N}(A)$ is defined as

$$\mathcal{N}(A) = \left\{x \in \mathbb{R}^n | Ax = 0\right\}$$

  - $\mathcal{N}(A)$ is set of vectors that is mapped to 0, under linear transformation A
  - vectors in $\mathcal{N}(A)$ are orthogonal to the rows of A
- $\mathcal{N}(A)$ gives the ambiguity of system A
  - for any $v \in \mathcal{N}(A)$, we have $A(x+v) = Ax$
  - conversely, if we have $Ax = A\tilde{x}$ then $\tilde{x} = x + v$ for some $v \in \mathcal{N}(A)$
- $\mathcal{N}(A)$ is a subspace

Interpretation of Nullspace

- Suppose A is a system measures (sensor) input signal $x$and outputs y, so that $y = Ax$
- suppose $z \in \mathcal{N}(A)$
  - z is undetectable from sensor A
  - That is, a signal x and a mixture $x + z$ looks same at the output of sensor A

$$Ax = A(x + z)$$

- $\mathcal{N}(A)$ characterizes ambiguity
  - the 'smaller' $\mathcal{N}(A)$, the less ambiguity

Interpretation of Nullspace

- Suppose A is a system measures (sensor) input signal $x$and outputs y, so that $y = Ax$
- suppose $z \in \mathcal{N}(A)$
    - z is undetectable from sensor A
    - That is, a signal x and a mixture $x + z$ looks same at the output of sensor A
$$Ax = A(x + z)$$
- $\mathcal{N}(A)$ characterizes ambiguity
    - the 'smaller' $\mathcal{N}(A)$, the less ambiguity
- suppose A such that there is **no ambiguity** for $y = Ax$, that is, by looking at $y = Ax$ we can uniquely find x!

Interpretation of Nullspace

- Suppose A is a system measures (sensor) input signal $x$and outputs y, so that $y = Ax$
- suppose $z \in \mathcal{N}(A)$
  - z is undetectable from sensor A
  - That is, a signal x and a mixture $x + z$ looks same at the output of sensor A

$$Ax = A(x + z)$$

- $\mathcal{N}(A)$ characterizes ambiguity
  - the 'smaller' $\mathcal{N}(A)$, the less ambiguity
- suppose A such that there is **no ambiguity** for $y = Ax$, that is, by looking at $y = Ax$ we can uniquely find x!
  - In that case $\mathcal{N}(A) = \{0\}$
  - equivalent to state that the mapping A is unique

Nullspace examples

- $$A_1 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$
- let $v = (2, -1)$, then for any c we have $A(cv) = 0$.
- it turns out that $\mathcal{N}(A_1) = \{cv | c \in \mathbb{R}^n\}$
- $\mathcal{N}(A_1)$ is 1-dimensional subspace
- $$A_2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$
- it turns out that $\mathcal{N}(A_2) = \{0\}$; that is, there is no $v \neq 0$ such that $A_2 v = 0$
- for $y = A_2 x$, if we know y, x can be uniquely determined (in this case $(A_2)^{-1} y$)

Rank

- Rank of matrix $A \in \mathbb{R}^{m \times n}$ is defined as

$$\textbf{rank}(A) = \textbf{dim}(\mathcal{R}(A))$$

- $\mathcal{R}(A)$ is a subspace spanned by columns of A, so rank is the number of independent columns of A
- One can show that number of independent columns and rows are same for any matrix $A \in \mathbb{R}^{m \times n}$
- Rank of A is the number of independent rows/columns
- This implies $\textbf{rank}(A) = \textbf{rank}(A^T)$

Rank

- we have
$$\textbf{rank}(A) \leq \min(m, n)$$
- Why? Suppose $m \leq n$.
  - $\mathcal{R}(A)$ is subspace spanned by vectors in $\mathbb{R}^m$, so $\textbf{rank}(A) \leq m \leq n$
- Why? Suppose $m > n$.
  - $\mathcal{R}(A)$ is subspace spanned by n vectors in $\mathbb{R}^m$, so the basis of $\mathcal{R}(A)$ have at most n basis: $\textbf{rank}(A) \leq n < m$
- Rank can be considered as degree of freedom (information) preserved by going through linear system A
- $y = Ax$. x has DoF n: going through A, output y has DoF at most m – but this cannot exceed n!

rank of matrix products

- we have
$$\textbf{rank}(BC) \leq \min(\textbf{rank}(B), \textbf{rank}(C))$$
- So if $A = BC$ with $B \in \mathbb{R}^{m \times r}$, $C \in \mathbb{R}^{r \times n}$ then
- $y = BCx = B(Cx)$ first DoF reduces to no more than $\textbf{rank}(C)$, but going through B again reduces rank no more than $\textbf{rank}(B)$
- conversely if $\textbf{rank}(A) = r$ then $A \in \mathbb{R}^{m \times n}$ can be factorized to $A = BC$ with $B \in \mathbb{R}^{m \times r}$, $C \in \mathbb{R}^{r \times n}$
- Here r can be considered as the width of information bottleneck

Full Rank

- if
$$\textbf{rank}(A) = \min(m, n)$$
we say A has full rank
- for square matrices, full rank means nonsingular (invertible)
- for skinny matrices ($m \geq n$) full rank means all the columns are independent
- for fat matrices ($m \leq n$) full rank means all the rows are independent

Full rank: case $\mathcal{R}(A) = \mathbb{R}^m$

statement $\mathcal{R}(A) = \mathbb{R}^m$ ($\mathbf{rank}(A) = m$) is equivalent to the following:

- columns of A spans $\mathbb{R}^m$
- the rows of A are independent
- $A^T c = 0$ implies that $c = 0$
- $\det(AA^T) \neq 0$
- A has right inverse, that is, there exists B such that

$$AB = I$$

with $B = A^T(AA^T)^{-1}$

Full rank: Case $\mathcal{N}(A) = \{0\}$

0 is the only element of the nullspace of A:

$$Ac = 0 \text{ implies } c = 0$$

A has *zero nullspace* or A is one-to-one ($\mathbf{rank}(A) = n$)

- linear transformation $y = Ax$ has unique x for each output y
- columns of A are independent (they form basis for a span)
- $\det(A^T A) \neq 0$
- A has a left inverse, that is, there exists B such that $BA = I$ with $B = (A^T A)^{-1} A^T$

Full rank: Inverse

- $A \in \mathbb{R}^{n \times n}$ is invertible or nonsinglar if $\det A \neq 0$
- columns of A are independent.
- rows of A are independents
- columns/rows of A are basis of $\mathbb{R}^n$
- $y = Ax$ has unique solution x for any y
- A has inverse A^{-1} where $AA^{-1} = A^{-1}A = I$
- $\mathcal{R}(A) = \mathbb{R}^n$

Determinant

Determinant

- Determinant is a function that maps a square matrix to a real number: $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$
- Def: signed volume of parallelepiped formed by columns of A
- properties:
  1. multilinear

$$\begin{aligned}&\det\begin{bmatrix}a_1 & a_2 & \dots & v_1+v_2 & \dots & a_n\end{bmatrix}\\ &= \det\begin{bmatrix}a_1 & \dots & v_1 & \dots & a_n\end{bmatrix} + \det\begin{bmatrix}a_1 & \dots & v_2 & \dots & a_n\end{bmatrix}\end{aligned}$$

  2. scaling

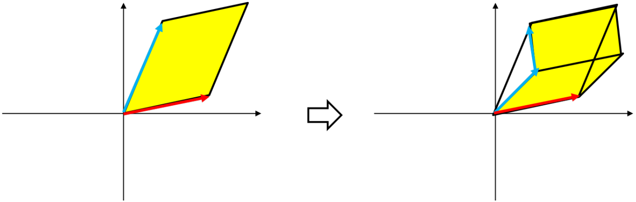
$$\det\begin{bmatrix}a_1 & \dots & ca_i & \dots & a_n\end{bmatrix} = c\det\begin{bmatrix}a_1 & \dots & a_i & \dots & a_n\end{bmatrix}$$

  3. exchange of columns

$$\det\begin{bmatrix}a_1 & \dots & a_i & a_j & \dots & a_n\end{bmatrix} = -\det\begin{bmatrix}a_1 & \dots & a_j & a_i & \dots & a_n\end{bmatrix}$$

Determinant properties

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} a & b_1 \\ c & d_1 \end{vmatrix} + \begin{vmatrix} a & b_2 \\ c & d_2 \end{vmatrix}$$

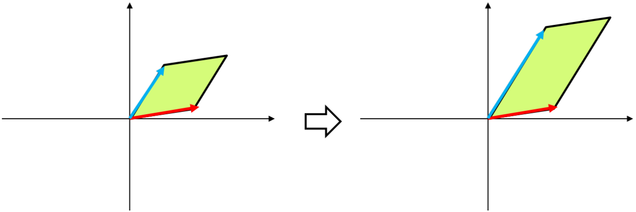


1. multilinear: additive in columns (while fixing other columns)

$$\det \begin{bmatrix} a_1 & a_2 & \dots & v_1 + v_2 & \dots & a_n \end{bmatrix}$$
$$= \det \begin{bmatrix} a_1 & \dots & v_1 & \dots & a_n \end{bmatrix} + \det \begin{bmatrix} a_1 & \dots & v_2 & \dots & a_n \end{bmatrix}$$

Determinant properties

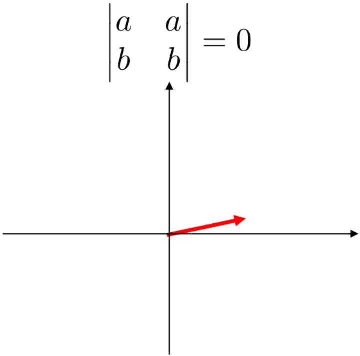
$$\begin{vmatrix} a & s\cdot b \\ c & s\cdot d \end{vmatrix} = s\begin{vmatrix} a & b \\ c & d \end{vmatrix}$$



1. scaling of columns

$$\det\begin{bmatrix} a_1 & \dots & ca_i & \dots & a_n \end{bmatrix} = c\det\begin{bmatrix} a_1 & \dots & a_i & \dots & a_n \end{bmatrix}$$

Determinant properties



1. determinant is zero if two columns are same

$$\det \begin{bmatrix} a_1 & \ldots & a_i & a_i & \ldots & a_n \end{bmatrix} = 0$$

Determinant

- $\det(I) = 1$
- $$\det(cA) = c^n \det(A)$$
- If any row or column of A is 0, then $\det(A) = 0$

Determinant

- $\det(I) = 1$
- $$\det(cA) = c^n \det(A)$$
- If any row or column of A is 0, then $\det(A) = 0$
- If there exist linearly dependent rows/columns, $\det(A) = 0$

Determinant

- determinant of triangular (either upper or lower) matrix A is the product of diagonal elements

$$\prod_{i=1}^{n} a_{ii}$$

Determinant

- determinant of triangular (either upper or lower) matrix A is the product of diagonal elements

$$\prod_{i=1}^{n} a_{ii}$$

- For any matrix B and triangular matrix T, we have $\det(AT) = \det(A)\det(T)$

Determinant

- For square matrices A and B,

$$\det(AB) = \det(A)\det(B)$$

Determinant

- For square matrices A and B,

$$\det(AB) = \det(A)\det(B)$$

- $\det(A) = \det(A^T)$

Determinant

- For square matrices A and B,

$$\det(AB) = \det(A)\det(B)$$

- $\det(A) = \det(A^T)$
- For invertible A,

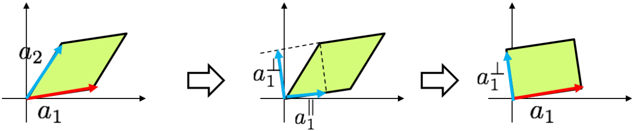
$$\det(A^{-1}) = \det(A)^{-1}$$

Determinant

- $|\det(A)|$ represents the volume of parallelogram formed by columns of A
- can be shown using
  - volume of B where columns $b_1, \ldots, b_n$ are orthogonal, is $|\det(B)|$
  - Gram-Schmidt orthogonalization combined with multilinear property of $\det$
- very useful geometric fact

Determinant

$$\begin{vmatrix} a_1 & a_2 \end{vmatrix} = \begin{vmatrix} a_1 & a_1^{\perp} + a_1^{\parallel} \end{vmatrix} = \begin{vmatrix} a_1 & a_1^{\perp} \end{vmatrix}$$



- Set a_1 as reference vector
- Divide a_2 into sum of components parallel ($a_1^{\parallel}$) and orthogonal ($a_1^{\perp}$) to a_1
- $\det([a_1\ a_1^{\perp}]) = |\det(A)|$: equal to the volume of parallelogram formed by columns of A

Determinant

- Suppose I have set S with volume vol(S). Consider linear transformation

$$T = \{Ax | x \in S\}$$

  Then vol(T) = $|\det(A)|$vol(S)
- Can find the volume of an image of mapping T

Eigenvalues and Symmetric Matrices

Eigenvalues

- Definition: consider square matrix A and if we have for some scalar λ and n-dim vector $v \neq 0$ such that

$$Av = \lambda v$$

we call λ eigenvalue and v eigenvector of A.

Eigenvalues

- Eigenvalues can be found by considering A's characteristic equation
$$\det(A - \lambda I) = 0$$
- This is polynomial equation of order n: n roots exist; they can be real or complex
- roots $\lambda_1, \ldots, \lambda_n$ are eigenvalues of A
- If entries of A are real, the complex eigenvalues come in pairs with conjugate

Eigenvalues: symmetric matrices

- Suppose A is real and symmetric,
  1. eigenvalues are real
  2. eigenvectors are orthogonal to each other

Eigenvalues: symmetric matrices

- symmetric A can be decomposed as

$$A = U \Lambda U^T$$

where Λ is diagonal matrix with λ_i at its i-th diagonal, the columns of U are orthonormal; that is, if

$$U = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix}$$

then $\|u_i\| = 1$ and $u_i^T u_j = 0$ for $i \neq j$

Eigenvalues: symmetric matrices

- Eigenvectors are mutually orthogonal, so

$$U^T U = UU^T = I$$

and

$$U^T = U^{-1}$$

- columns of U form orthonormal basis of $\mathbb{R}^n$
- $A = U \Lambda U^T$ is called spectral decomposition of A

Eigenvalues

- for real symmetric A and its spectral decomposition $U\Lambda U^T$ we have

$$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

  that is, sum of rank-1 matrices

- Now, since u_i are basis, for any $x \in \mathbb{R}^n$ we can write

$$x = \sum_{i=1}^{n} \hat{x}_i u_i$$

  Thus

$$Ax = \sum_{i=1}^{n} \lambda_i \hat{x}_i u_i$$

- λ_i as gains to the direction u_i

Eigenvalues

- A is positive definite iff all of its eigenvalues are positive
  - consider
$$x^T A x = \sum_{i=1}^{n} \lambda_i \hat{x}_i^2$$
- A is positive semi-definite iff all of its eigenvalues are non-negative
- A is invertible iff its eigenvalues are nonzero
- $A^k = U\Lambda U^T U\Lambda U^T \cdots = U\Lambda^k U^T$. So eigenvalues of A^k are λ_i^k's

Eigenvalues

- Consider ellisoid defined as

$$\left\{x | x^T Q x \leq 1\right\}$$

  for some positive definite Q
- The eigenvectors of Q comprise the principal axes of the ellipsoid

$$x^T Q x = \sum_i \lambda_i \hat{x}_i^2 = \sum_i \frac{\hat{x}_i^2}{\left(\frac{1}{\sqrt{\lambda_i}}\right)^2}$$

- In 2-D this is like

$$\frac{x_1^2}{(1/\sqrt{\lambda_1})^2} + \frac{x_2^2}{(1/\sqrt{\lambda_2})^2} \leq 1$$

- if $Q \succeq 0$, $x^T Q x$ can be used as norm (induced norm): 2-norm is special case of $Q = I$

Quadratic Form

Quadratic Form

- Quadratic function: $f : \mathbb{R} \to \mathbb{R}$ is $f(x) = ax^2$.

Quadratic Form

- Quadratic function: $f : \mathbb{R} \to \mathbb{R}$ is $f(x) = ax^2$.
- **Quadratic form:** a function that maps length n vector to scalar, that is $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\boxed{f(x) = x^T A x}$$

  where $A \in \mathbb{S}^n$
- This is *vector version* of quadratic functions

Quadratic Form

- Quadratic function: $f : \mathbb{R} \to \mathbb{R}$ is $f(x) = ax^2$.
- **Quadratic form:** a function that maps length n vector to scalar, that is $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\boxed{f(x) = x^T A x}$$

  where $A \in \mathbb{S}^n$
- This is *vector version* of quadratic functions
- If x is scalar, $f(x) = x^T a x = ax^2$

Examples

- Let $x = (x_1, x_2)$ and $A = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$. Then

$$\mathbf{x}^T A\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4x_1 \\ 3x_2 \end{bmatrix}$$

  which gives $f(x) = 4x_1^2 + 3x_2^2$

Examples

- $A = \begin{bmatrix} 3 & -2 \\ -2 & 7 \end{bmatrix}$. Then

  $$\mathbf{x}^T A\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 3 & -2 \\ -2 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 3x_1 - 2x_2 \\ -2x_1 + 7x_2 \end{bmatrix}$$

  which gives $f(x) = 3x_1^2 - 4x_1x_2 + 7x_2^2$
- If $A = I$, then $f(x) = x^T x = x_1^2 + x_2^2 = \|x\|^2$

Quadratic Form

- Why is A symmetric? Suppose A is nonsymmetric, then

$$f(x) = x^T A x = f(x)^T = x^T A^T x$$

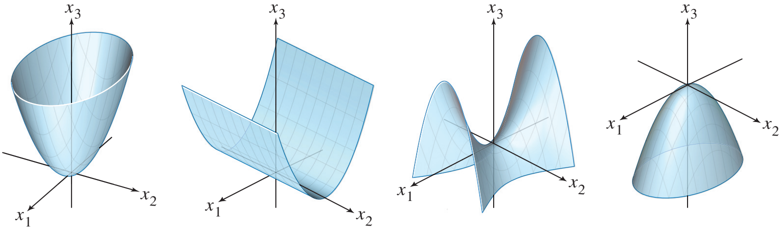
- This means

$$f(x) = \frac{x^T A x}{2} + \frac{x^T A x}{2} = \frac{x^T A x}{2} + \frac{x^T A^T x}{2} = x^T \frac{(A + A^T)}{2} x$$

  Note $Q = \frac{(A+A^T)}{2}$ is symmetric. This means, we can always replace A by Q, and have the same function.

- So it is sufficient to use only symmetric A.

Quadratic Form: examples



(a) $z = 3x_1^2 + 7x_2^2$ (b) $z = 3x_1^2$ (c) $z = 3x_1^2 - 7x_2^2$ (d) $z = -3x_1^2 - 7x_2^2$

Positive Definiteness

- Q is **positive definite** if $x^T Qx > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$
- Q is **positive semidefinite** if $x^T Qx \geq 0$ for all $x \in \mathbb{R}^n$
- Q is **negative definite** if $x^T Qx > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$
- Q is **negative semidefinite** if $x^T Qx \geq 0$ for all $x \in \mathbb{R}^n$
- If the sign of $x^T Qx$ differs by x, Q is **indefinite.**
- postive definiteness is the notion of positiveness of a number extended to matrix
- a is positive $\Rightarrow ax^2 > 0$ for all $x \in \mathbb{R} \setminus \{0\}$

Positive Definiteness

- If $Q \succ 0$, then
$$\left\{x | x^T Q x = c\right\}$$
is an **ellipse**
- special case: if $Q = I$, then
$$\left\{x | x^T x = c\right\}$$
is a sphere

Positive Definiteness

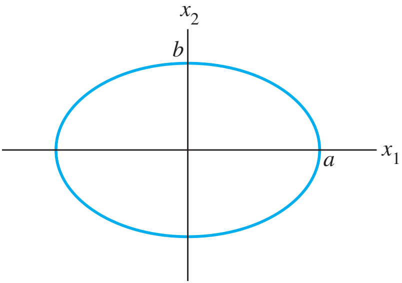
- Let

$$Q = \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{bmatrix}$$

then

$$x^T Q x = \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2}$$

- Plot of $x^T Q x = 1$

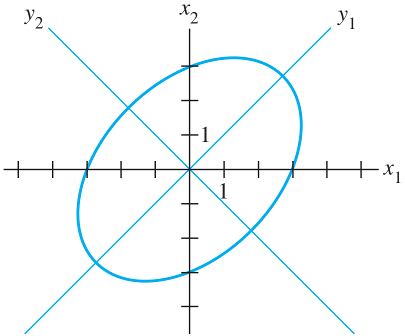


Positive Definiteness

- Let

$$Q = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$$

- $x^T Qx = 5x_1^2 - 4x_1x_2 + 5x_2^2$: Q is pd
- Plot of $x^T Qx = 48$: also an ellipse!



Positive Definiteness

- Why $x^T Q x = 1$ is ellipse? If $Q = U \Lambda U^T$

$$x^T U \Lambda U^T x = y^T \Lambda y$$

  here $y = U^T x$ which is change of coordinates

- In y-coordinate

$$y^T \Lambda y = 1 \Rightarrow \frac{y_1}{\left(\sqrt{\frac{1}{\lambda_1}}\right)^2} + \frac{y_2}{\left(\sqrt{\frac{1}{\lambda_2}}\right)^2} + \cdots + \frac{y_n}{\left(\sqrt{\frac{1}{\lambda_n}}\right)^2} = 1$$

  this is ellipse with axis lengths

$$\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}}$$

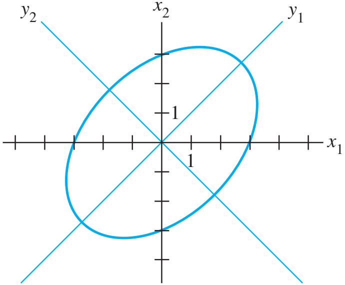
Positive Definiteness

- Thus $x^T Qx = 1$ is simply an ellipse going through rotation and reflections, i.e., change of coordinates $x = Uy$
- This shows that, for ellipse $x^T Qx = 1$
  1. The axis directions are eigenvectors of Q
  2. The lengths of axis are square-root of inverse of eigenvalues of Q

$$\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}}$$

- Because axis $e_1, \dots, e_n$ in y-coordinate are mapped to $u_1, \dots, u_n$ which are eigenvectors of Q
- Eigenvectors with **larger** eigenvalue: **shorter** axis length

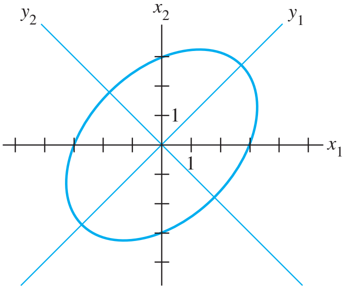
Positive Definiteness



- $$Q = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$$

  so $x^T Qx = 5x_1^2 - 4x_1x_2 + 5x_2^2$

Positive Definiteness



- $Q = U\Lambda U^T$ where

$$u_1 = 1/\sqrt{2}(1,1), u_2 = 1/\sqrt{2}(-1,1)$$

  with $\lambda_1 = 3$, $\lambda_2 = 7$
- $x^T Qx = 1$: axis directions are $(1,1)$ and $(-1,1)$, and the axis lengths are $1/\sqrt{3}$ and $1/\sqrt{7}$
- longer axis length to smaller eigenvalue direction $(1,1)$

Rayleigh Quotient

- Suppose A is symmetric.

$$R(A, x) = \frac{x^T Ax}{x^T x}$$

  is called the Rayleigh quotient
- Suppose $A \succ 0$. What x maximizes $R(A, x)$?
- Let us fix $x^T x = 1$: that is, consider x whose distance from origin is 1
- At what x

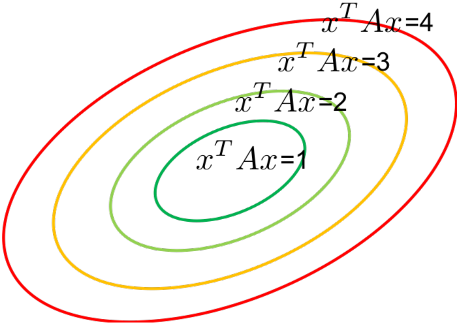
$$x^T Ax$$

  is maximized?

Rayleigh Quotient

- Answer: the direction of the **shortest** axis length
- Eigenvector with the **largest** eigenvalue

Rayleigh Quotient

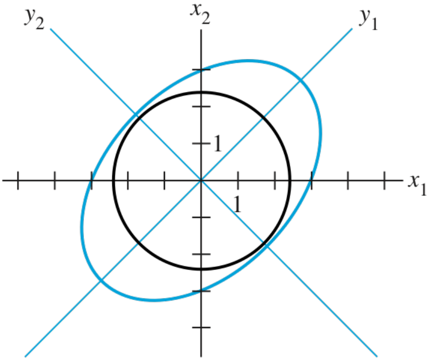


- This can be seen by looking at **contours** of the set

$$x^T Ax = c$$

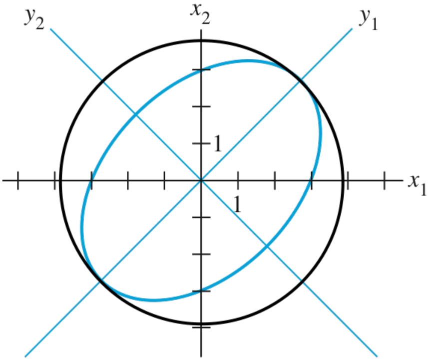
for various c

Rayleigh Quotient



- For fixed $x^T x$, the eigenvector with *maximum* eigenvalue **maximizes** $x^T Ax$

Rayleigh Quotient



- For fixed $x^T x$, the eigenvector with *minimum* eigenvalue **minimizes** $x^T Ax$

Maximum and minimum eigenvalues

- Consider symmetric matrix A. Its maximum eigenvalue, denoted by $\lambda_{\max}(A)$, is given by

$$\max_{x \neq 0} \frac{x^T A x}{x^T x}$$

that is, x which maximizes $R(A, x)$

Maximum and minimum eigenvalues

- Consider symmetric matrix A. Its maximum eigenvalue, denoted by $\lambda_{\max}(A)$, is given by

$$\max_{x\neq 0} \frac{x^T A x}{x^T x}$$

  that is, x which maximizes $R(A, x)$
- Why? Since $A = Q\Lambda Q^T$. Assume $x = \sum_{i=1}^n x_i q_i$. then

$$\max_{x\neq 0} \frac{\sum_{i=1}^n \lambda_i x_i^2}{\sum_{i=1}^n x_i^2} \leq \frac{\lambda_{\max} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} = \lambda_{\max}$$

- This maximum value is achievable by setting x to the eigenvector q_i corresponding to $\lambda_{\max}$.

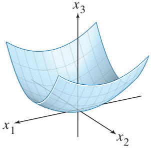
Maximum and minimum eigenvalues

- One can make similar arguments for the minimum eigenvalue of A
- $\lambda_{\min}(A)$, is given by

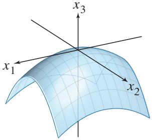
$$\min_{x \neq 0} \frac{x^T A x}{x^T x}$$

  or $\min_{x \neq 0} R(A, x)$
- Try to show it by yourself!

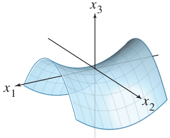
Positive Definiteness



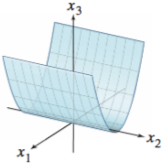
Positive definite



Negative definite



Indefinite



Positive semidefinite

Notation: Positive Definiteness

- Set of positive (semi)definite matrices is denoted by $\mathbb{S}^n_{++}(\mathbb{S}^n_+)$
- We write for positive definite A

$$A \succ 0$$

  Here 0 on the right is $n \times n$ matrix of zeros
- for positive semidefinite A

$$A \succeq 0$$

- similar for $A \prec 0$ and $A \preceq 0$ for negative definite and negative semidefinite matrix

Positive Definiteness: properties

- every positive definite matrix A is nonsingular

$$Ax = 0 \Rightarrow x^T A x = 0 \Rightarrow x = 0$$

- every positive definite matrix A has positive diagonal elements

$$e_i^T A e_i > 0$$

- every positive semidefinite matrix A has nonnegative diagonal elements

$$e_i^T A e_i \geq 0$$

Quadratic form and Eigenvalues

- For $A \in \mathbb{S}^n$, A is
  1. positive definite iff all the eigenvalues are positive
  2. negative definite iff all the eigenvalues are negative
  3. indefinite iff there are both positive and negative eigenvalues

Quadratic form and Eigenvalues

- Since
$$A = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_n u_n u_n^T$$
we have $x^T A x = \sum_{i=1}^n c_i^2 \lambda_i$ where $x_i = \sum_{i=1}^n c_i u_i$.
  - $\lambda_i > 0$, $i = 1, \cdots, n$ implies $x^T A x > 0$.
  - $\lambda_i < 0$, $i = 1, \cdots, n$ implies $x^T A x < 0$.

Quadratic form and Eigenvalues

- Indefiniteness means that, there exist vector x which can make $x^T Ax$ to ∞ or $-\infty$
- for example

$$x = tu_i,\ \lambda_i > 0$$

and let $t \to \infty$, $x^T Ax \to \infty$. But

$$x = tu_j,\ \lambda_j < 0$$

and let $t \to \infty$, $x^T Ax \to -\infty$.

Positive Definiteness

- Examples: Let $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$
- To get eigenvalues, we have

$$\det(A - \lambda I) = 0 \Rightarrow (\lambda - a)(\lambda - c) - b^2 = 0$$

We have quadratic equation

$$\lambda^2 - (a + c)\lambda + ac - b^2 = 0$$

Positive Definiteness

- characteristic equation

$$\lambda^2 - (a + c)\lambda + ac - b^2 = 0$$

- firstly, are roots always real? discriminant

$$(a + c)^2 - 4(ac - b^2) = (a - c)^2 + 4b^2 \geq 0$$

- If $A \succ 0$, we must have two positive solutions:

$$a + c > 0, \; ac - b^2 > 0$$

Positive Definiteness

- condition is:

$$a > 0,\ c > 0,\ ac - b^2 > 0$$

- diagonal elements are positive and determinant is positive
- Make sense, because quadratic form

$$x^T A x = x^T \begin{bmatrix} a & b \\ b & c \end{bmatrix} x = a x_1^2 + 2b x_1 x_2 + c x_2^2$$

to be strictly positive, $a > 0$ and $c > 0$ and *discriminant* of quadratic equation

$$ax^2 + 2bx + c = 0$$

must be negative

Positive definite and inverse

- Suppose $A \succ 0$.
- Does A^{-1} exist?
  - Yes,
    $$A^{-1} = QD^{-1}Q^T$$
    where D^{-1} has $1/\lambda_i$ on its diagonals
- Is A^{-1} pd?
  - Yes, its eigenvalues are positive
- Suppose A, B are pd. Is $A + B \succ 0$?
  - Yes

$A^T A$ is positive (semi)definite

- Consider $m \times n$ matrix A.
- $A^T A$ is positive semidefinite (psd).
- Why?

$$x^T A^T A x = (Ax)^T A x = \|Ax\|^2 \geq 0$$

- If A has independent columns, $A^T A$ is positive definite (pd).
- AA^T is also psd (pd if A has independent rows)

$A^T A$ is positive (semi)definite

- rank-1 matrix of form vv^T is psd
- projection matrix
$$\frac{vv^T}{v^T v}$$
is psd
- Given set of vectors $u_1, \ldots, u_k$
$$A = \sum_{i=1}^{k} c_i u_i u_i^T$$
is psd iff $c_i \geq 0$.

$A^T A$ is positive (semi)definite

- difference between A^2 and $A^T A$ (or AA^T):
- A^2 is only defined for square A
- $A^T A$ or AA^T can be defined for arbitrary A
- A^2 is **not** guaranteed to be psd, whereas $A^T A$ and AA^T are always psd
    - If A is symmetric, then A^2 is also symmetric and psd

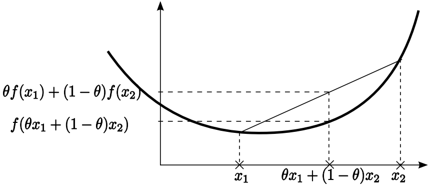
Part 2: Convex Functions

Definition of convex functions

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathcal{D}(f)$ is convex set and

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2)$$

for all $x_1, x_2 \in \mathcal{D}(f)$ and some $0 \leq \theta \leq 1$



- f is concave if $-f$ is convex
- $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if $\mathcal{D}(f)$ is convex set and

$$f(\theta x_1 + (1-\theta)x_2) < \theta f(x_1) + (1-\theta) f(x_2)$$

for all $x_1, x_2 \in \mathcal{D}(f)$ for $x_1 \neq x_2$ for some $0 < \theta < 1$

Examples on $\mathbb{R}$

- convex functions
  - affine: $ax + b$
  - exponential: e^{ax} for any $a \in \mathbb{R}$
  - powers: x^α on $\mathbb{R}_{++}$ for $\alpha \leq 0$ or $\alpha \geq 1$
  - powers of absolute value: $|x|^p$ for $p \geq 1$
  - negative entropy: $x \log x$ on $\mathbb{R}_{++}$

- concave functions
  - affine: $ax + b$
  - powers: x^α on $\mathbb{R}_{++}$ for $0 \leq \alpha \leq 1$
  - logarithm: $\log x$ on $\mathbb{R}_{++}$

Examples on $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$

- Examples for functions $\mathbb{R}^n \to \mathbb{R}$
  - Affine functions: $f(x) = a^T x + b$ for some $a \in \mathbb{R}^n,\ b \in \mathbb{R}$
  - p-norms: $\|x\|_p := (\sum_i |x_i|^p)^{1/p}$, for $p \geq 1$, $\|x\|_\infty = \max_i |x_i|$
  - can be shown using triangular inequality, $\|x+y\| \leq \|x\| + \|y\|$

Examples on $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$

- Examples for functions $\mathbb{R}^{m\times n} \rightarrow \mathbb{R}$
  - affine functions

$$f(X) = \mathbf{tr}(A^T X) + b = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}X_{ij} + b$$

    for $A \in \mathbb{R}^{m\times n}$ $b \in \mathbb{R}$
  - $\mathbf{tr}(B)$ is the sum of the diagonals of a square matrix B
  - $\mathbf{tr}(A^T X)$ is defined as the inner product of A and X
    - Using informal notations: $A(:)^T B(:)$

Examples on $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$

- Matrix norm (spectral norm):

$$\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\max}(A)$$

- triangular inequality holds, because

$$\begin{aligned}\|A+B\| &= \max_x \frac{\|(A+B)x\|}{\|x\|} \le \max_x \frac{\|Ax\|+\|Bx\|}{\|x\|} \\ &\le \max_x \frac{\|Ax\|}{\|x\|} + \max_x \frac{\|Bx\|}{\|x\|} = \|A\| + \|B\|\end{aligned}$$

Restriction to line

- $f(x)$ is convex iff $g : \mathbb{R} \to \mathbb{R}$ defined as

$$g(t) = f(a + bt)$$

  is convex in t where $a, b, a + bt \in \mathcal{D}(f)$
- determine convexity of 1-D function can be simpler

Multivariate functions: gradient and directional derivative

- Consider a function $f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}$
- When $n = 1$, we know the derivative

$$\frac{df(x)}{dx}$$

  or $f'(x)$, is the rate of change in function f
- This can be viewed also as "slope" of line tangent to f
- Or, the first-order approximation of f around $x = x_0$ is given by

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

Partial derivative

- now consider case $n \geq 1$: we consider multivariate function
- Given a multivariate function $f(x, y)$

$$\frac{\partial f}{\partial x} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon, y) - f(x, y)}{\epsilon}$$
$$\frac{\partial f}{\partial y} = \lim_{\epsilon \to 0} \frac{f(x, y+\epsilon) - f(x, y)}{\epsilon}$$

  are the partial derivative of f with respect to x and y
- Rate of change of f in only one direction
- For $x \in \mathbb{R}^n$, $f(x) = f(x_1, x_2, \ldots, x_n)$, $\frac{\partial f}{\partial x_i}$ similarly defined

Partial derivative

- Finding partial derivative: treat other variables as constant
- Example: $f(x, y) = x^2 + xy$

$$\frac{\partial f}{\partial x} = 2x + y, \frac{\partial f}{\partial y} = x$$

Multivariate functions: gradient and directional derivative

- want to define rate of change in f, but in what *direction*?
- consider 1st approximation of $f(x)$ with small change $\Delta x \in \mathbb{R}^n$
- By chain rule,

$$
\begin{aligned}
f(x + \Delta x) &\approx f(x) + \frac{\partial f}{\partial x_1}\Delta x_1 + \cdots + \frac{\partial f}{\partial x_n}\Delta x_n \\
&= f(x) + \nabla f(x)^T \Delta x
\end{aligned}
$$

  Here

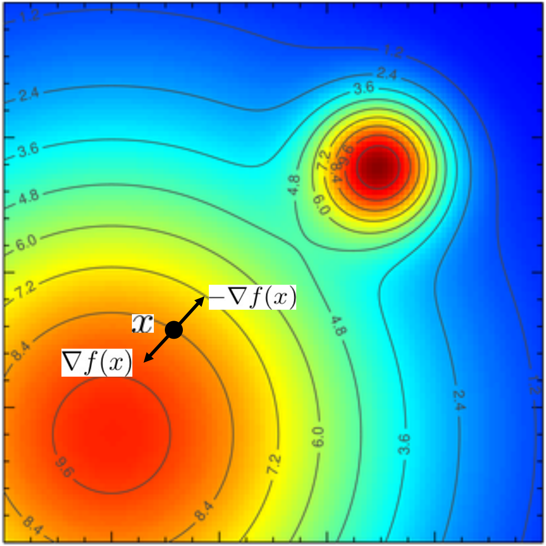
$$
\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)
$$

  is called gradient of f at x
- Gradient is a vector

Multivariate functions: gradient and directional derivative

- For vector v, $\nabla f(x)^T v$ is the directional derivate of f along vector v at point x. This is the rate of change of f to the direction of v at point x
- Equivalent to
$$\frac{d}{dt} f(x+tv)\bigg|_{t=0}$$
- Now consider vectors v with size 1 such that
$$\max_{\|v\|=1} \nabla f^T v$$
- One that maximizes $\nabla f^T v$: parallel to ∇f
- Thus gradient is the direction at which the rate of change is maximum

Multivariate functions: gradient and directional derivative



Multivariate Calculus

- Let $f(x) = c^T x$. Find ∇f
- Since $f(x) = \sum_{i=1}^{n} c_i x_i$, we have $\frac{\partial f}{\partial x_i} = c_i$ thus

$$\nabla f = (c_1, c_2, \dots, c_n) = c$$

Multivariate Calculus

- Let $f(x) = x^T Qx$. Find ∇f
- For any i, we have

$$f(x) = \sum_{i=1}^{n} Q_{i,i} x_i^2 + 2 \sum_{i \neq j} x_i x_j Q_{i,j}$$

So

$$\frac{\partial f}{\partial x_i} = 2Q_{i,i} x_i + 2 \sum_{j \neq i} x_j Q_{i,j} = 2 \sum_{j} Q_{i,j} x_j$$

thus

$$\nabla f = 2Qx$$

Multivariate functions: Hessian

- Consider second order approximation of function f in one dimension

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2$$

- How about in higher dimension?

$$f(x+\Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 f(x) \Delta x$$

- the Hessian $\nabla^2 f(x) \in S^n$ is defined as

$$\left[\nabla^2 f(x)\right]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j},\ i,j = 1,\ldots,n$$

Jacobian

- Consider vector function $f(x) = (f_1(x), f_2(x), \ldots, f_m(x))$. Consider small change Δx.
-

$$f(x+\Delta x) = \begin{bmatrix} f_1(x+\Delta x) \\ \ldots \\ f_m(x+\Delta x) \end{bmatrix} \approx f(x) + \begin{bmatrix} \nabla f_1(x)^T \Delta x \\ \ldots \\ \nabla f_m(x)^T \Delta x \end{bmatrix}$$
$$= f(x) + \begin{bmatrix} \nabla f_1(x)^T \\ \ldots \\ \nabla f_m(x)^T \end{bmatrix} \Delta x$$

  The Jacobian of f at x is

$$J_f = \begin{bmatrix} \nabla f_1(x)^T \\ \ldots \\ \nabla f_m(x)^T \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \ldots & \frac{\partial f_1}{\partial x_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \ldots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

- So the first order change is $J_f \Delta x$

Jacobian

- Let $f(x) = Ax$. Find the Jacobian J_f.
- We have

$$f(x) = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}$$

$$J_f = \begin{bmatrix} \nabla f_1(x)^T \\ \dots \\ \nabla f_m(x)^T \end{bmatrix} = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{bmatrix} = A$$

Hessian

- Second-order derivative: rate of change of gradient, that is

$$\nabla f(x + \Delta x) \approx \nabla f(x) + J_{\nabla f}(x)\Delta x$$

- $J_{\nabla f}(x)$, Jacobian of gradient of f, is called Hessian

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

  Thus

$$\nabla f(x + \Delta x) \approx \nabla f(x) + \nabla^2 f(x)\Delta x$$

- Hessian is square matrix and is symmetric

Hessian

- Let $f(x) = \frac{1}{2}x^T Qx$. Find the Hessian.
- We have

$$\nabla f = Qx$$

So

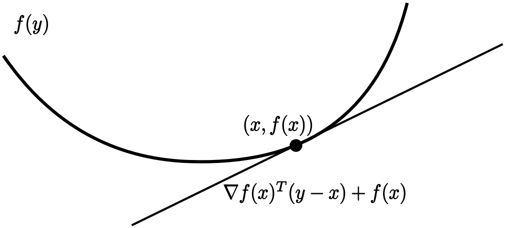
$$J_{\nabla f} = Q$$

First-order condition of convexity

- $\mathcal{D}(f)$ denotes the domain of f: set over which f is defined
- f is differentiable if is open and the gradient ∇f exists at each $x \in \mathcal{D}(f)$
- 1st-order condition: differentiable f with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

for all $x, y \in \mathcal{D}(f)$.



First-order approximation of f is a global underestimator

Second-order condition of convexity

- f is twice differentiable if $\mathcal{D}(f)$ is open and the Hessian $\nabla^2 f(x) \in S^n$ exists at each $x \in \mathcal{D}(f)$

- 2nd-order condition: for twice differentiable f with convex domain
  - f is convex iff
  $$\nabla^2 f(x) \succeq 0$$
  for all $x, y \in \mathcal{D}(f)$.
  - if $\nabla^2 f(x) \succ 0$ for all $x \in \mathcal{D}(f)$, then f is strictly convex.

Examples

- quadratic function:

$$f(x) = (1/2)x^T Px + q^T x + r$$

  with $P \in S^n$, then
  Convex if $P \succeq 0$

- Note

$$\nabla f(x) = Px + q, \ \nabla^2 f(x) = P$$

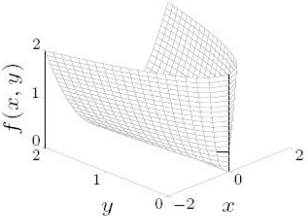
Examples

- least-squares objective: $f(x) = \|Ax - b\|_2^2$

$$\nabla f(x) = 2A^T(Ax - b), \;\; \nabla^2 f(x) = 2A^T A$$

- quadratic-over-linear: $f(x, y) = x^2/y$, convex for $y > 0$

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T$$



log-sum-exponential function

log-sum-exp: $f(x) = \log \sum_{k=1}^{n} \exp x_k$ is convex

$$\nabla^2 f(x) = \frac{1}{\mathbf{1}^T z} \mathbf{diag}(z) - \frac{1}{(\mathbf{1}^T z)^2} z z^T \qquad (z_k = \exp x_k)$$

to show $\nabla^2 f(x) \succeq 0$, we must verify that $v^T \nabla^2 f(x) v \geq 0$ for all v:

$$v^T \nabla^2 f(x) v = \frac{(\sum_k z_k v_k^2)(\sum_k z_k) - (\sum_k v_k z_k)^2}{(\sum_k z_k)^2} \geq 0$$

since $(\sum_k v_k z_k)^2 \leq (\sum_k z_k v_k^2)(\sum_k z_k)$ (from Cauchy-Schwarz inequality)

- Log-sum-exp is “smooth-max” function
- Convexity can be shown using Hölder’s inequality

Epigraph and sublevel set

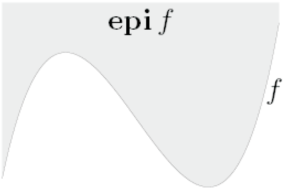
- α-sublevel set of $f : \mathbb{R}^n \to \mathbb{R}$:

$$C_\alpha = \{x \in \mathcal{D}(f) | f(x) \leq \alpha\}$$

  sublevel sets of convex functions are convex sets (converse is false)

- Epigraph of $f : \mathbb{R}^n \to \mathbb{R}$:

$$\mathbf{epi}(f) := \{(x, t) \in \mathbb{R}^{n+1} | x \in \mathcal{D}(f),\ f(x) \leq t\}$$



- f is convex iff **epi** f is convex

Jenson's inequality

- Basic inequality: $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2)$$

  for some $0 \leq \theta \leq 1$.

- Extension: For any random variable z, if f is convex, then

$$f(\mathbb{E}[z]) \leq \mathbb{E}[f(z)]$$

- Basic inequality is a special case for a random variable z such that

$$z = \begin{cases} x_1 & \text{w.p. } \theta \\ x_2 & \text{w.p. } 1-\theta \end{cases}$$

  Here w.p. stands for 'with probability'.

- It is useful to consider simple case to remember the inequality.

Operations that preserve convexity

- practical methods for establishing convexity of a function

  1. verify definition
  2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$.

  3. Show that f is obtained from simple convex functions by operations that preserve convexity:
     - nonnegative weighted sum
     - composition with affine function
     - pointwise maximum and supremum
     - composition
     - minimization

Positive weighted sum & composition with affine function

- nonnegative multiple: af is convex if f is convex, $a \geq 0$
- sum: $f_1 + f_2$ convex if f_1, f_2 convex (extends to infinite sums, integrals)
- composition with affine function: $f(Ax + b)$ is convex if f is convex
- examples
  - log barrier for linear inequalities

$$f(x) = -\sum_{i=1}^{n} \log(b_i - a_i^T x), \ \mathcal{D}(f) = \{x | a_i^T x < b_i, i = 1, \ldots, m\}$$

    - $f(x)$ is defined on the **interior** of the polyhedron, and get to infinity as points move to the polyhedron boundary
  - (any) norm of affine function: $f(x) = \|Ax + b\|_p$

Log determinant

- Let $A \in \mathbb{S}^n_{++}$ be positive definite n by n matrix
- $f(A) = \log \det(A)$ is concave

Pointwise maximum

- if $f_1, \ldots, f_m$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex
- examples
  - piecewise-linear function: $f(x) = \max_{i=1\ldots m}(a_i^T x + b_i)$ is convex
  - Sum of r largest components of $x \in \mathbb{R}^n$:

    $$f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[r]}$$

    is convex ($x_{[i]}$ is ith largest component of x)

    proof: $f(x) = \max\{x_{i_1} + x_{i_2} + \cdots + x_{i_r} | 1 \leq i_1 < i_2 < \ldots < i_r \leq n\}$

Pointwise supremum

- if $f(x,y)$ is convex in x for each $y \in A$, then

$$g(x) = \sup_{y \in A} f(x,y)$$

is convex, proof:

$$\begin{aligned} \theta g(x_1) + (1-\theta) g(x_2) &= \sup_{y \in A} \theta f(x_1, y) + \sup_{y \in A} (1-\theta) f(x_2, y) \\ &\geq \sup_{y \in A} \{\theta f(x_1, y) + (1-\theta) f(x_2, y)\} \\ &\geq \sup_{y \in A} \{f(\theta x_1 + (1-\theta) x_2, y)\} \\ &= g(\theta x_1 + (1-\theta) x_2) \end{aligned}$$

- Note $f(x,y)$ need **NOT** be convex both in x and y (example: $f(x,y) = x^2 \mathbf{log}(\mathbf{1} + \boldsymbol{y})$, for $x, y > 0$ $f(x,y)$ is convex in x for any given y, but is not overall convex.

Pointwise supremum: examples

1. support function of a set C: $S_C(x) = \sup_{y \in C} y^T x$ is convex
2. distance to farthest point in a set C: $f(x) = \sup_{y \in C} \|x - y\|$
3. maximum eigenvalue: let $f : \mathbb{S}^n \to \mathbb{R}$ such that

$$f(A) = \sup_x \frac{x^T A x}{x^T x}$$

Composition with scalar functions

- composition of $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$:

$$f(x) = h(g(x))$$

f is convex if

$$\begin{array}{c} g \text{ convex, } h \text{ convex, } h \text{ nondecreasing} \\ g \text{ concave, } h \text{ convex, } h \text{ nonincreasing} \end{array}$$

proof: (for $n = 1$, differentiable g, h)

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

- Note the result is general, and holds without differentiability assumption
- examples
  - $\exp g(x)$ is convex if g is convex
  - $1/g(x)$ is convex if g is concave and positive

Vector composition

- composition of $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$

$$f(x) = h(g(x)) = h(g_1(x), g_2(x), ..., g_k(x))$$

- f is convex if
  - g_i convex, h convex, h nondecreasing in each argument
  - g_i concave, h convex, h nonincreasing in each argument
- proof (for $n = 1$, differentiable g, h)

$$f''(x) = g'(x)^T \nabla^2 h(g(x)) g'(x) + \nabla h(g(x))^T g''(x)$$

- examples
  - $\sum_{i=1}^n \log g_i(x)$ is concave if g_i are concave and positive
  - $\log \sum_{i=1}^n \exp g_i(x)$ is convex if g_i are convex

Minimization

- if $f(x, y)$ is convex in (x, y) and for a convex set $\boldsymbol{C}$, then

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex, proof:

$$\begin{aligned}
\theta g(x_1) + (1-\theta) g(x_2) &= \theta f(x_1, y^*(x_1)) + (1-\theta) f(x_2, y^*(x_2)) \\
&\geq f(\theta x_1 + (1-\theta) x_2, \theta y^*(x_1) + (1-\theta) y^*(x_2)) \\
&\geq \inf_{y \in C} f(\theta x_1 + (1-\theta) x_2, y) \\
&= g(\theta x_1 + (1-\theta) x_2)
\end{aligned}$$

- Note how convexity condition on $f(x, y)$ has been **strengthened** compared to supremum case.

Minimization examples

- examples
  1. $f(x, y) = x^T Ax + 2x^T By + y^T Cy$ with

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0, C \succ 0$$

     minimizing over y gives

$$g(x) = \inf_y f(x, y) = x^T (A - BC^{-1} B^T) x$$

     Since g is convex, Schur complement $A - BC^{-1} B^T \succeq 0$.

  2. distance to a set: $\mathrm{dist}(x, S) = \inf_{y \in S} \|x - y\|$ is convex if S is convex

Perspective

- the perspective of $f : \mathbb{R}^n \to \mathbb{R}$ is function $g : \mathbb{R}^{n+1} \to \mathbb{R}$ such that

$$g(x, t) = tf(x/t), \quad t > 0$$

  g is convex if f is convex
- examples:
  - $f(x) = x^T x$ is convex; $g(x, t) = x^T x/t$ is convex
  - $f(x) = -\log x$ is convex; the relative entropy $g(x, t) = t\log t - t\log x$ is convex
  - if f is convex, then

$$g(x) = (c^T x + d)f\left((Ax + b)/(c^T x + d)\right)$$

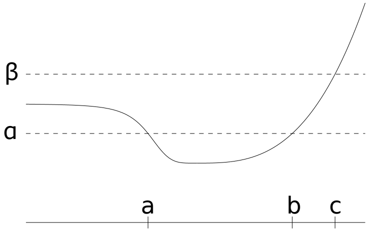
    is convex

Quasiconvex functions

$f : \mathbb{R}^n \to \mathbb{R}$ is *quasiconvex* if $\mathcal{D}(f)$ is convex and the sublevel set

$$S_\alpha = \{x \in \mathcal{D}(f) | f(x) \leq \alpha\}$$

is a convex set for any $\alpha \in \mathbb{R}$



- f is quasiconcave if $-f$ is quasiconvex.
- f is quasilinear if it is quasiconvex and quasiconcave

Examples

- $\sqrt{|x|}$ is quasiconvex on $\mathbb{R}$
- $\mathrm{ceil}(x) = \inf\{z \in \mathbb{Z} | z \geq x\}$ is quasilinear
- $\log(x)$ is quasilinear on $\mathbb{R}_{++}$
- $f(x_1, x_2) = x_1 x_2$ is quasiconcave on $\mathbb{R}^2_{++}$
- linear-fractional function

$$f(x) = \frac{a^T x + b}{c^T x + d}, \quad \mathcal{D}(f) = \left\{ x | c^T x + d > 0 \right\}$$

  is quasilinear
- distance ratio

$$f(x) = \frac{||x - a||_2}{||x - b||_2}, \qquad \mathcal{D}(f) = \{x | ||x - a||_2 \leq ||x - b||_2\}$$

  is quasiconvex

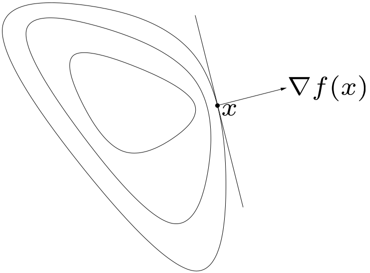
Properties

- **modified Jensen's inequality** for quasiconvex f

$$0 \le \theta \le 1 \Rightarrow f(\theta x + (1-\theta)y) \le \max\{f(x), f(y)\}$$

- **first-order condition** differentiable f with convex domain is quasiconvex iff

$$f(y) \le f(x) \Rightarrow \nabla f(x)^T(y-x) \le 0$$



(Note that this relates to the level sets and the direction of ∇f)

- Sum of quasiconvex functions are not necessarily quasiconvex

Operations that preserve quasi-convexity

- Nonnegative weighted maximum

$$f = \max\{w_1 f_1, \ldots, w_m f_m\}$$

  and

$$f = \sup_{y \in C} w(y) g(x, y)$$

  where $w(y) \geq 0$ and $g(x, y)$ is quasiconvex in x for each y
- $g : \mathbb{R}^n \to \mathbb{R}$ is q.c. and h in nondecreasing, $h(g())$ is q.c.
- $f(x, y)$ is quasiconvex jointly in x and y, then for convex set C,

$$g(x) = \inf_{y \in C} f(x, y)$$

  is q.c.

Log-concave and log-convex

a positive function f is log-concave if $\log f$ is concave.

$$f(\theta x_1 + (1-\theta)x_2) \geq f(x_1)^\theta f(x_2)^{1-\theta}$$

f is log-convex if $\log f$ is convex.

- powers: x^a on $\mathbb{R}_+$ log-convex for $a \leq 0$, log-concave for $a \geq 0$
- many probability densities are log-concave, e.g., normal

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\bar{x})^T \Sigma^{-1}(x-\bar{x})}$$

- cumulative Gaussian distribution Φ is log-concave

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$$

Log-concave and log-convex

- log-convexity $\Rightarrow$ convexity $\Rightarrow$ quasi-convexity
  - If f is convex then $\exp f$ is convex
- concavity $\Rightarrow$ log-concavity $\Rightarrow$ quasi-concavity
  - If f is concave then $\log f$ is concave
- log-convex and log-concave functions can be optimized due to quasi-convex/concavity

Properties of log-concave functions

- twice differentiable f with convex domain is log-concave iff

$$f(x)\nabla^2 f(x) \preceq \nabla f(x)\nabla f(x)^T$$

  for all $x \in \mathbf{dom} f$
- product of log-concave function is log-concave
- sum of log-concave function is not always log-concave
- integration: if $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is log-concave, then

$$g(x) = \int f(x, y) dy$$

  is log-concave.

Consequences of integration property

- convolution $f * g$ is log-concave functions of f, g is log-concave

$$f * g(x) = \int f(x-y)g(y)dy$$

- if $C \subseteq \mathbb{R}^n$ convex and y is a random variable with log-concave pdf then

$$f(x) = \mathbb{P}(x + y \in C)$$

  is log-concave
- proof: write $f(x)$ as integral of product of log-concave functions

$$f(x) = \int g(x+y)p(y)dy, \; g(u) = \begin{cases} 1 & u \in C \\ 0 & u \notin C, \end{cases}$$

  p is pdf of y

Example: yield function

$$Y(x) = \mathbb{P}(x + \omega \in S)$$

- $x \in \mathbb{R}^n$: nominal parameter values for product
- $\omega \in \mathbb{R}^n$: random variations of parameters in manufactured product
- S: set of acceptable values

if S is convex and ω has a log-concave pdf, then

- Y is log-concave
- yield regions $|\{x|Y(x) \geq \alpha\}$ are convex
- log-concave functions can be optimized (maximized)