CSED700H: Convex Optimization

Mathematical background[1]

Namhoon Lee

POSTECH

Fall 2023

[1] These slides are created based on Appendix A in Convex Optimization by Boyd and Vandenberghe.

Table of Contents

Table of Contents

Inner product, Euclidean norms, and angle

- *standard inner product* for $x, y \in \mathbb{R}^n$

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i$$

- *Euclidean norm* or l_2-norm of a vector $x \in \mathbb{R}^n$

$$\|x\|_2 = (x^\top x)^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}$$

- *Cauchy-Schwartz inequality* for any $x, y \in \mathbb{R}^n$

$$|x^\top y| \leq \|x\|_2 \|y\|_2$$

- (unsigned) *angle* between nonzero vectors $x, y \in \mathbb{R}^n$

$$\angle(x, y) = \cos^{-1}\left(\frac{x^\top y}{\|x\|_2 \|y\|_2}\right)$$

  where $\cos^{-1}(u) \in [0, \pi]$; x and y are *orthogonal* if $x^\top y = 0$

- standard inner product on $\mathbb{R}^{m\times n}$ for $X, Y \in \mathbb{R}^{m\times n}$

$$\langle X, Y\rangle = \mathrm{tr}(X^\top Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij}$$

- *Frobenius norm* of a matrix $X \in \mathbb{R}^{m\times n}$

$$\|X\|_\mathsf{F} = \left(\mathrm{tr}(X^\top X)\right)^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)^{1/2}$$

- standard inner product on $\mathbb{S}^n$

$$\langle X, Y\rangle = \mathrm{tr}(XY) = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}Y_{ij} = \sum_{i=1}^{n} X_{ii}Y_{ii} + 2\sum_{i<j} X_{ij}Y_{ij}$$

Norms

- A function $f : \mathbb{R}^n \to \mathbb{R}$ with $\operatorname{dom} f = \mathbb{R}^n$ is called a *norm* if
  - f is nonnegative: $f(x) \geq 0$ for all $x \in \mathbb{R}^n$
  - f is definite: $f(x) = 0$ only if $x = 0$
  - f is homogeneous: $f(tx) = |t| f(x)$, for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$
  - f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$, for all $x, y \in \mathbb{R}^n$
- $f(x) = \|x\|$ suggests that a norm is a generalization of the absolute value on $\mathbb{R}$.
- When we specify a particular norm, we use the notation $\|x\|_{\mathsf{symb}}$.

Distance

- A norm is a measure of the *length* of a vector x.
- We can measure the *distance* between two vectors x and y as the length of their difference, *i.e.*,
$$\operatorname{dist}(x, y) = \|x - y\|.$$

Unit ball

- *unit ball* of the norm $\|\cdot\|$

$$\mathcal{B} = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$$

- The unit ball satisfies the following properties:
  - $\mathcal{B}$ is symmetric about the origin, *i.e.*, $x \in \mathcal{B}$ if and only if $-x \in \mathcal{B}$
  - $\mathcal{B}$ is convex
  - $\mathcal{B}$ is closed, bounded, and has nonempty interior
- Conversely, if $C \subseteq \mathbb{R}^n$ is any set satisfying these three conditions, then it is the unit ball of a norm

$$\|x\| = (\sup\{t \geq 0 \mid tx \in \mathbb{C}\})^{-1}.$$

Examples

- absolute value on $\mathbb{R}$
- l_p-*norm* on $\mathbb{R}^n$

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$$

  - $p = 2$: $\|x\|_2 = (x^\top x)^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}$
  - $p = 1$: $\|x\|_1 = |x_1| + \cdots + |x_n|$
  - $p = \infty$: $\|x\|_\infty = \max\{|x_1|, \cdots, |x_n|\}$
- P-*quadratic norms* for $P \in \mathbb{S}^n_{++}$

$$\|x\|_P = (x^\top Px)^{1/2} = \|P^{1/2}x\|_2$$

  - The unit ball of a quadratic norm is an ellipsoid (and conversely, if the unit ball of a norm is an ellipsoid, the norm is a quadratic norm).

- Frobenius norm on $\mathbb{R}^{m\times n}$

$$\|X\|_{\mathsf{F}} = \left(\mathbf{tr}(X^\top X)\right)^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)^{1/2}$$

- sum-absolute-value norm

$$\|X\|_{\mathsf{sav}} = \sum_{i=1}^{m}\sum_{j=1}^{n} |X_{ij}|$$

- maximum-absolute-value norm

$$\|X\|_{\mathsf{mav}} = \max\{|X_{ij}| \mid i=1,\ldots,m,\ j=1,\ldots,n\}$$

Equivalence of norms

- Suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$. A basic result of analysis is that there exist positive constants α and β such that, for all $x \in \mathbb{R}^n$,

$$\alpha\|x\|_a \leq \|x\|_b \leq \beta\|x\|_a.$$

  This means that the norms are *equivalent*, *i.e.*, they define the same set of open subsets, the same set of convergent sequences, and so on.
- Using convex analysis, we can give a more specific result: If $\|\cdot\|$ is any norm on $\mathbb{R}^n$, then there exists a quadratic norm $\|\cdot\|_P$ for which

$$\|x\|_P \leq \|x\| \leq \sqrt{n}\|x\|_P$$

  holds for all x. In other words, any norm on $\mathbb{R}^n$ can be uniformly approximated, within a factor of $\sqrt{n}$, by a quadratic norm.

Operator norms

- Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. We define the *operator norm* of $X \in \mathbb{R}^{m\times n}$, induced by the norms $\|\cdot\|_a$ and $\|\cdot\|_b$, as
$$\|X\|_{a,b} = \sup\{\|Xu\|_a \mid \|u\|_b \leq 1\}.$$
- When $\|\cdot\|_a$ and $\|\cdot\|_b$ are both Euclidean norms, the operator norms of X is its *maximum singular value*, and is denoted $\|X\|_2$:
$$\|X\|_2 = \sigma_{\text{max}}(X) = (\lambda_{\text{max}}(X^\top X))^{1/2}.$$
This norm is also called the *spectral norm* or *l_2-norm* of X.
- *max-row-sum norm*
$$\|X\|_\infty = \sup\{\|Xu\|_\infty \mid \|u\|_\infty \leq 1\} = \max_{i=1,\dots,m} \sum_{j=1}^{n} |X_{ij}|$$
- *max-column-sum norm*
$$\|X\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |X_{ij}|$$

Dual norm

- Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The associated *dual norm*, denoted $\|\cdot\|_*$, is defined as

$$\|z\|_* = \sup\{z^\top x \mid \|x\| \leq 1\}.$$

- The dual norm can be interpreted as the operator norm of $z^\top$, interpreted as a $1 \times n$ matrix, with the norm $\|\cdot\|$ on $\mathbb{R}^n$, and the absolute value on $\mathbb{R}$:

$$\|z\|_* = \sup\{|z^\top x| \mid \|x\| \leq 1\}.$$

- From the definition of dual norm we have the inequality

$$z^\top x \leq \|x\| \|z\|_*,$$

which holds for all x and z.
- The dual norm of the dual norm is the original norm, *i.e.*, $\|x\|_{**} = \|x\|$ for all x.

- The dual of the Euclidean norm is the Euclidean norm, since

$$\sup\{z^\top x \mid \|x\|_2 \le 1\} = \|z\|_2.$$

  (This follows from the Cauchy-Schwarz inequality; for nonzero z, the value of x that maximizes $z^\top x$ over $\|x\|_2 \le 1$ is $z/\|z\|_2$.)

- The dual of the l_∞-norm is the l_1-norm:

$$\sup\{z^\top x \mid \|x\|_\infty \le 1\} = \sum_{i=1}^{n} |z_i| = \|z\|_1,$$

  and the dual of the l_1-norm is the l_∞-norm.

- More generally, the dual of the l_p-norm is the l_q-norm, where q satisfies $1/p + 1/q = 1$, *i.e.*, $q = p/(p-1)$.

- For l_2- or spectral norm on $\mathbb{R}^{m\times n}$, the associated dual norm is

$$\|Z\|_{2*} = \sup\{\operatorname{tr}(Z^\top X) \mid \|X\|_2 \leq 1\},$$

which turns out to be the sum of the singular values,

$$\|Z\|_{2*} = \sigma_1(Z) + \cdots + \sigma_r(Z) = \operatorname{tr}(Z^\top Z)^{1/2},$$

where $r = \operatorname{rank} Z$. This norm is sometimes called the *nuclear* norm.

Table of Contents

Interior

- An element $x \in \mathbb{C} \subseteq \mathbb{R}^n$ is called an *interior* point of $\mathbb{C}$ if there exists an $\epsilon > 0$ for which
$$\{y \mid \|y - x\|_2 \leq \epsilon\} \subseteq \mathbb{C},$$
i.e., there exists a ball centered at x that lies entirely in $\mathbb{C}$.
- The set of all points interior to $\mathbb{C}$ is called the *interior* of $\mathbb{C}$ and is denoted $\operatorname{int} \mathbb{C}$.

Closure

- The *closure* of a set $\mathbb{C}$ is defined as

$$\operatorname{cl}\mathbb{C} = \mathbb{R}^n \setminus \operatorname{int}(\mathbb{R}^n \setminus \mathbb{C}),$$

  i.e., the complement of the interior of the complement of $\mathbb{C}$.
- A point x is in the closure of $\mathbb{C}$ if for every $\epsilon > 0$, there is a $y \in \mathbb{C}$ with $\|x - y\|_2 \leq \epsilon$.

Boundary

- The *boundary* of the set $\mathbb{C}$ is defined as

$$\operatorname{bd}\mathbb{C} = \operatorname{cl}\mathbb{C} \setminus \operatorname{int}\mathbb{C}.$$

- A *boundary point* x (*i.e.*, a point $x \in \operatorname{bd}\mathbb{C}$) satisfies the following property: For all $\epsilon > 0$, there exists $y \in \mathbb{C}$ and $z \notin \mathbb{C}$ with

$$\|y - x\|_2 \leq \epsilon, \qquad \|z - x\|_2 \leq \epsilon,$$

  i.e., there exist arbitrarily close points in $\mathbb{C}$, and also arbitrarily close points not in $\mathbb{C}$.

Open and closed sets

- A set $\mathbb{C}$ is *open* if $\operatorname{int} \mathbb{C} = \mathbb{C}$, *i.e.*, every point in $\mathbb{C}$ is an inteior point.
- A set $\mathbb{C} \subseteq \mathbb{R}^n$ is *closed* if its complement $\mathbb{R}^n \setminus \mathbb{C} = \{x \in \mathbb{R}^n \mid x \notin \mathbb{C}\}$ is open.
- A set $\mathbb{C}$ is closed if and only if it contains the limit point of every convergent sequence in it. In other words, if $x_1, x_2, \ldots$ converges to x, and $x_i \in \mathbb{C}$, then $x \in \mathbb{C}$. The closure of $\mathbb{C}$ is the set of all limit points of convergent sequences in $\mathbb{C}$.
- A set $\mathbb{C}$ is *closed* if it contains its boundary, *i.e.*, $\operatorname{bd} \mathbb{C} \subseteq \mathbb{C}$. It is *open* if it contains no boundary points, *i.e.*, $\mathbb{C} \cap \operatorname{bd} \mathbb{C} = \emptyset$

Supremum

- Suppose $\mathbb{C} \subseteq \mathbb{R}$. A number a is an *upper bound* on $\mathbb{C}$ if for each $x \in \mathbb{C}$, $x \leq a$.
- Then the set of upper bounds on a set $\mathbb{C}$ is either
  - empty (in which case we say $\mathbb{C}$ is unbounded above),
  - all of $\mathbb{R}$ (only when $\mathbb{C} = \emptyset$), or
  - a closed infinite interval $[b, \infty)$.
- The number b is called the *least upper bound* or *supremum* of the set $\mathbb{C}$, and is denoted $\sup \mathbb{C}$.
- We take $\sup \emptyset = -\infty$, and $\sup \mathbb{C} = \infty$ if $\mathbb{C}$ is unbounded above.
- When the set $\mathbb{C}$ is finite, $\sup \mathbb{C}$ is the maximum of its elements.

Infimum

- A number a is a lower bound on $\mathbb{C} \subseteq \mathbb{R}$ if for each $x \in \mathbb{C}$, $a \leq x$.
- The *infimum* (or *greatest lower bound*) of a set $\mathbb{C} \subseteq \mathbb{R}$ is defined as $\inf \mathbb{C} = -\sup(-\mathbb{C})$.
- When $\mathbb{C}$ is finite, the infimum is the minimum of its elements.
- We take $\inf \emptyset = \infty$, and $\inf \mathbb{C} = -\infty$ if $\mathbb{C}$ is unbounded below, *i.e.*, has no lower bound.

Table of Contents

Function notation

- f is a function on the set $\mathrm{dom}\, f \subseteq A$ into the set B

$$f : A \to B$$

  the notation indicates *syntax*, not the domain of function
- for example

$$f : \mathbb{R}^n \to \mathbb{R}^m$$

  f maps (some) n-vectors into m-vectors; it does not mean that $f(x)$ is defined for every $x \in \mathbb{R}^n$.
- another example, $f : \mathbb{S}^n \to \mathbb{R}$

$$f(X) = \log\det X$$

  with $\mathrm{dom}\, f = \mathbb{S}^n_{++}$

Continuity

- A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *continuous* at $x \in \mathbf{dom}\, f$ if for all $\epsilon > 0$ there exists a δ such that

$$y \in \mathbf{dom}\, f, \quad \|y - x\|_2 \leq \delta \Rightarrow \|f(y) - f(x)\|_2 \leq \epsilon.$$

- Continuity can be described in terms of limits: whenever the sequence $x_1, x_2, \ldots$ in $\mathbf{dom}\, f$ converges to a point $x \in \mathbf{dom}\, f$, the sequence $f(x_1), f(x_2), \ldots$ converges to $f(x)$, *i.e.*,

$$\lim_{i \to \infty} f(x_i) = f(\lim_{i \to \infty} x_i).$$

- A function f is *continuous* if it is continuous at every point in its domain.

Closed functions

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *closed* if, for each $\alpha \in \mathbb{R}$, the sublevel set

$$\{x \in \mathbf{dom}\, f \mid f(x) \leq \alpha\}$$

  is closed.
- This is equivalent to the condition that the epigraph of f,

$$\mathbf{epi}\, f = \{(x, t) \in \mathbb{R}^{n+1} \mid x \in \mathbf{dom}\, f,\ f(x) \leq t\},$$

  is closed.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, and $\mathbf{dom}\, f$ is closed, then f is closed. If $f : \mathbb{R}^n \to \mathbb{R}$ is continous, with $\mathbf{dom}\, f$ open, then f is closed if and only if f converges to ∞ along every sequence converging to a boundary point of $\mathbf{dom}\, f$. In other words, if $\lim_{i\to\infty} x_i = x \in \mathbf{bd}\,\mathbf{dom}\, f$, with $x_i \in \mathbf{dom}\, f$, we have $\lim_{i\to\infty} f(x_i) = \infty$.

Examples on $\mathbb{R}$

- The function $f : \mathbb{R} \to \mathbb{R}$, with $f(x) = x \log x$, $\operatorname{dom} f = \mathbb{R}_{++}$, is *not* closed.
- The function $f : \mathbb{R} \to \mathbb{R}$, with

$$f(x) = \begin{cases} x \log x & x > 0 \\ 0 & x = 0, \end{cases} \qquad \operatorname{dom} f = \mathbb{R}_{+},$$

  is closed.
- The function $f(x) = -\log x$, $\operatorname{dom} f = \mathbb{R}_{++}$ is closed.

Table of Contents

Derivative

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $x \in \operatorname{int} \operatorname{dom} f$. The function f is differentiable at x if there exists a matrix $Df(x) \in \mathbb{R}^{m \times n}$ that satisfies
$$\lim_{z \in \operatorname{dom} f, z \neq x, z \to x} \frac{\|f(z) - f(x) - Df(x)(z - x)\|_2}{\|z - x\|_2} = 0,$$
in which case we refer to $Df(x)$ as the *derivative* (or *Jacobian*) of f at x.
- The function f is *differentiable* if $\operatorname{dom} f$ is open, and it is differentiable at every point in its domain.
- The affine function of z given by
$$f(x) + Df(x)(z - x)$$
is called the *first-order approximation* of f at (or near) x.
- The derivative can be found by deriving the first-order approximation of the function f at x, or from partial derivatives:
$$Df(x)_{ij} = \frac{\partial f_i(x)}{\partial x_j}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

Gradient

- When f is real-valued (*i.e.*, $f : \mathbb{R}^n \to \mathbb{R}$) the derivative $Df(x)$ is a $1 \times n$ matrix, *i.e.*, it is a *row* vector. Its transpose is called the *gradient* of the function:

$$\nabla f(x) = Df(x)^\top,$$

  which is a (column) vector, *i.e.*, in $\mathbb{R}^n$.
- Its components are the partial derivatives of f:

$$\nabla f(x)_i = \frac{\partial f(x)}{\partial x_i}, \quad i = 1, \ldots, n.$$

- The first-order approximation of f at a point $x \in \mathbf{int}\,\mathbf{dom}\, f$ can be expressed as (the affine function of z)

$$f(x) + \nabla f(x)^\top (z - x).$$

Chain rule

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbf{int}\,\mathbf{dom}\, f$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(x) \in \mathbf{int}\,\mathbf{dom}\, g$. Define the composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(z) = g(f(z))$. Then h is differentiable at x, with derivative

$$Dh(x) = Dg(f(x))Df(x).$$

- As an example, suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$. Taking the transpose of $Dh(x) = Dg(f(x))Df(x)$ yields

$$\nabla h(x) = g'(f(x))\nabla f(x).$$

Second derivative

- The second derivative or *Hessian matrix* of f at $x \in \operatorname{int} \operatorname{dom} f$, denoted $\nabla^2 f(x)$, is given by

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \qquad i = 1, \dots, n, \quad j = 1, \dots, n,$$

  provided f is twice differentiable at x, where the partial derivatives are evaluated at x.

- The second derivative can be interpreted as the derivative of the first derivative. If f is differentiable, the *gradient mapping* is the function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, with $\operatorname{dom} \nabla f = \operatorname{dom} f$, with value $\nabla f(x)$ at x. The derivative of this mapping is

$$D\nabla f(x) = \nabla^2 f(x).$$

Table of Contents

Range and nullspace

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

- The *range* of A is the set of all vectors in $\mathbb{R}^m$ that can be written as linear combinations of the colums of A, *i.e.*,

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}.$$

  - The range $\mathcal{R}(A)$ is a subspace of $\mathbb{R}^m$. Its dimension is the *rank* of A. The rank of A can never be greater than the minimum of m and n.

- The *nullspace* of A is the set of all vectors x mapped into zero by A:

$$\mathcal{N}(A) = \{x \mid Ax = 0\}.$$

  - The nullspace is a subspace of $\mathbb{R}^n$.

Orthogonal decomposition induced by A

- If $\mathcal{V}$ is a subspace of $\mathbb{R}^n$, its *orthogonal complement* is defined as
$$\mathcal{V}^\perp = \{x \mid z^\top x = 0 \text{ for all } z \in \mathcal{V}\}.$$
- A basic result of linear algebra is that, for any $A \in \mathbb{R}^{m \times n}$, we have
$$\mathcal{N}(A) = \mathcal{R}(A^\top)^\perp.$$
- This result is often stated as
$$\mathcal{N}(A) \overset{\perp}{\oplus} \mathcal{R}(A^\top) = \mathbb{R}^n.$$

  Here the symbol $\overset{\perp}{\oplus}$ refers to *orthogonal direct sum*, *i.e.*, the sum of two subspaces that are orthogonal. The decomposition is called the *orthogonal decomposition induced by* A.

Symmetric eigenvalue decomposition

- Suppose $A \in \mathbb{S}^n$. Then A can be factored as

$$A = Q\Lambda Q^\top,$$

  where $Q \in \mathbb{R}^{n\times n}$ is *orthogonal*, *i.e.*, satisfies $Q^\top Q = I$, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.
- The (real) numbers λ_i are the *eigenvalues* of A, and are the roots of the *characteristic polynomial* $\det(sI - A)$.
- The columns of Q form an orthonormal set of *eigenvectors* of A.
- The factorization is called the *spectral decomposition* or (symmetric) *eigenvalue decomposition* of A.

- The determinant and trace can be expressed in terms of the eigenvalues,

$$\det A = \prod_{i=1}^{n} \lambda_i, \qquad \operatorname{tr} A = \sum_{i=1}^{n} \lambda_i,$$

as can the spectral and Frobenius norms,

$$\|A\|_2 = \max_{i=1,\ldots,n} |\lambda_i| = \max\{\lambda_1, -\lambda_n\}, \qquad \|A\|_F = \left(\sum_{i=1}^{n} \lambda_i^2 \right)^{1/2}.$$

Definiteness and matrix inequalities

- The largest and smallest eigenvalues satisfy

$$\lambda_{\mathsf{max}}(A) = \sup_{x \neq 0} \frac{x^\top A x}{x^\top x}, \qquad \lambda_{\mathsf{min}}(A) = \inf_{x \neq 0} \frac{x^\top A x}{x^\top x}.$$

- A matrix $A \in \mathbb{S}^n$ is called *positive definite*, denoted as $A \succ 0$, if for all $x \neq 0$, $x^\top A x > 0$. By the inequality above, we see that $A \succ 0$ if and only all its eigenvalues are positive, *i.e.*, $\lambda_{\mathsf{min}}(A) > 0$. If $-A$ is positive definite, we say A is *negative definite*, which we write as $A \prec 0$.
- If A satisfies $x^\top A x \geq 0$ for all x, we say that A is *positive semidefinite* or *nonnegative definite*. If $-A$ is nonnegative definite, *i.e.*, if $x^\top A x \leq 0$ for all x, we say that A is *negative semidefinite* or *nonpositive definite*.
- For $A, B \in \mathbb{S}^n$, we use $A \prec B$ to mean $B - A \succ 0$, and so on. These inequalities are called *matrix inequalities*, or generalized inequalities associated with the positive semidefinite cone.

Symmeric squareroot

- Let $A \in \mathbb{S}^n_+$, with eigenvalue decomposition $A = Q \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Q^\top$. We define the (symmetric) squareroot of A as

$$A^{1/2} = Q \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}) Q^\top.$$

The squareroot $A^{1/2}$ is the unique symmetric positive semidefinite solution of the equation $X^2 = A$.

Generalized eigenvalue decomposition

- The *generalized eigenvalue* of a pair of symmetric matrices $(A, B) \in \mathbb{S}^n \times \mathbb{S}^n$ are defined as the roots of the polynomial $\det(sB - A)$.
- We are usually interested in matrix pairs with $B \in \mathbb{S}^n_{++}$. In this case the generalized eigenvalues are also the eigenvalues of $B^{-1/2}AB^{-1/2}$ (which are real). As with the standard eigenvalue decomposition, we order the generalized eigenvalues in nonincreasing order, as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and denote the maximum generalized eigenvalue by $\lambda_{\mathsf{max}}(A, B)$.
- When $B \in \mathbb{S}^n_{++}$, the pair of matrices can be factored as

$$A = V \Lambda V^\top, \qquad B = VV^\top,$$

  where $V \in \mathbb{R}^{n \times n}$ is nonsingular, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where λ_i are the generalized eigenvalues of the pair (A, B). The decomposition is called the *generalized eigenvalue decomposition*.
- The generalized eigenvalue decomposition is related to the standard eigenvalue decomposition of the matrix $B^{-1/2}AB^{-1/2}$. If $Q\Lambda Q^\top$ is the eigenvalue decomposition of $B^{-1/2}AB^{-1/2}$, then the above holds with $V = B^{1/2}Q$.

Singular value decomposition

- Suppose $A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank} A = r$. Then A can be factored as

$$A = U \Sigma V^\top,$$

  where $U \in \mathbb{R}^{m \times r}$ satisfies $U^\top U = I$, $V \in \mathbb{R}^{n \times r}$ satisfies $V^\top V = I$, and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.
- The columns of U are called *left singular vectors* of A, the columns of V are *right singular vectors*, and the numbers σ_i are the *singular values*.
- The singular value decomposition can be written

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

  where $u_i \in \mathbb{R}^m$ are the left singular vectors, and $v_i \in \mathbb{R}^n$ are the right singular vectors.

- The singular value decomposition of a matrix A is closely related to the eigenvalue decomposition of the (symmetric, nonnegative definite) matrix $A^\top A$.

$$A^\top A = V\Sigma^2 V^\top = \begin{bmatrix} V & \tilde{V} \end{bmatrix} \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V & \tilde{V} \end{bmatrix}^\top,$$

  where $\tilde{V}$ is any matrix for which $\begin{bmatrix} V & \tilde{V} \end{bmatrix}$ is orthogonal.
- The righthand expression is the eigenvalue decomposition of $A^\top A$, so we conclude that its nonzero eigenvalues are the singular values of A squared, and the associated eigenvectors of $A^\top A$ are the right singular vectors of A.
- A similar analysis of $AA^\top$ shows that its nonzero eigenvalues are also the squares of the singular values of A, and the associated eigenvectors are the left singular vectors of A.

- The first or largest singular value is also written as $\sigma_{\text{max}}(A)$. It can be expressed as

$$\sigma_{\text{max}}(A) = \sup_{x,y \neq 0} \frac{x^\top A y}{\|x\|_2 \|y\|_2} = \sup_{y \neq 0} \frac{\|Ay\|_2}{\|y\|_2}.$$

  The righthand expression shows that the maximum singular value is the l_2 operator norm of A.
- The *minimum singular value* of $A \in \mathbb{R}^{m \times n}$ is given by

$$\sigma_{\text{min}}(A) = \begin{cases} \sigma_r(A) & r = \min\{m, n\} \\ 0 & r < \min\{m, n\}, \end{cases}$$

  which is positive if and only if A is full rank.

- The singular values of a symmetric matrix are the absolute values of its nonzero eigenvalues, sorted into descending order. The singular values of a symmetric positive semidefinite matrix are the same as its nonzero eigenvalues.
- The *condition number* of a nonsingular $A \in \mathbb{R}^{n\times n}$, denoted $\mathrm{cond}(A)$ or $\kappa(A)$, is defined as

$$\mathrm{cond}(A) = \|A\|_2 \|A^{-1}\|_2 = \sigma_{\mathsf{max}}(A)/\sigma_{\mathsf{min}}(A).$$

Pseudo-inverse

- Let $A = U\Sigma V^\top$ be the singular value decomposition of $A \in \mathbb{R}^{m\times n}$, with $\operatorname{rank} A = r$. We define the *pseudo-inverse* or *Moore-Penrose inverse* of A as

$$A^\dagger = V\Sigma^{-1}U^\top \in \mathbb{R}^{n\times m}.$$

- Alternative expresions are

$$A^\dagger = \lim_{\epsilon\to 0}(A^\top A + \epsilon I)^{-1}A^\top = \lim_{\epsilon\to 0} A^\top(AA^\top + \epsilon I)^{-1},$$

where the limits are taken with $\epsilon > 0$, which ensures that the inverses in the expressions exist. If $\operatorname{rank} A = m$, then $A^\dagger = A^\top(AA^\top)^{-1}$. If A is square and nonsingular, then $A^\dagger = A^{-1}$.

- The pseudo-inverse comes up in problems involving least-squares, minimum norm, quadratic minimization, and (Euclidean) projection. For example, $A^\dagger b$ is a solution of the least-squares problem

$$\text{minimize} \quad \|Ax - b\|_2^2$$

in general. When the solution is not unique, $A^\dagger b$ gives the solution with minimum (Euclidean) norm. As another example, the matrix $AA^\dagger = UU^\top$ gives (Euclidean) projection on $\mathcal{R}(A)$. The matrix $A^\dagger A = VV^\top$ gives (Euclidean) projection on $\mathcal{R}(A^\top)$.

- The optimal value $p^\star$ of the (general, nonconvex) quadratic optimization problem

$$\text{minimize} \quad (1/2)x^\top P x + q^\top x + r,$$

where $P \in \mathbb{S}^n$, can be expressed as

$$p^\star = \begin{cases} -(1/2)q^\top P^\dagger q + r & P \succeq 0, \quad q \in \mathcal{R}(P) \\ -\infty & \text{otherwise.} \end{cases}$$

(This generalizes the expression $p^\star = -(1/2)q^\top P^{-1} q + r$, valid for $P \succ 0$.)