Peeter Joot
peeterjoot@protonmail.com

ECE1505H Convex Optimization. Lecture 3: Matrix functions, SVD, and types of Sets. Taught by Prof. Stark Draper

Disclaimer Peeter's lecture notes from class. These may be incoherent and rough.

These are notes for the UofT course ECE1505H, Convex Optimization, taught by Prof. Stark Draper.

1.1 Matrix inner product

Given real matrices $X, Y \in \mathbb{R}^{m \times n}$, one possible matrix inner product definition is

$$\begin{aligned}
\langle X, Y \rangle &= \mathrm{Tr}(X^\mathrm{T} Y) \\
&= \mathrm{Tr}\left(\sum_{k=1}^m X_{ki} Y_{kj} \right) \\
&= \sum_{k=1}^m \sum_{j=1}^n X_{kj} Y_{kj} \\
&= \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij}.
\end{aligned} \tag{1.1}$$

This inner product induces a norm on the (matrix) vector space, called the Frobenius norm

$$\begin{aligned}
\|X\|_F &= \mathrm{Tr}(X^\mathrm{T} X) \\
&= \sqrt{\langle X, X \rangle} \\
&= \sum_{i=1}^m \sum_{j=1}^n X_{ij}^2.
\end{aligned} \tag{1.2}$$

1.2 Range, nullspace.

Definition 1.1: Range.

Given $A \in \mathbb{R}^{m \times n}$, the range of A is the set:

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

Definition 1.2: Nullspace.

Given $A \in \mathbb{R}^{m \times n}$, the nullspace of A is the set:

$$\mathcal{N}(A) = \{\mathbf{x} | A\mathbf{x} = 0\}.$$

1.3 SVD.

To understand operation of $A \in \mathbb{R}^{m \times n}$, a representation of a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, decompose A using the singular value decomposition (SVD).

Definition 1.3: SVD.

Given $A \in \mathbb{R}^{m \times n}$, an operator on $\mathbf{x} \in \mathbb{R}^n$, a decomposition of the following form is always possible

$$\begin{aligned} A &= U \Sigma V^{\mathrm{T}} \\ U &\in \mathbb{R}^{m \times r} \\ V &\in \mathbb{R}^{n \times r}, \end{aligned}$$

where r is the rank of A, and both U and V are orthogonal

$$\begin{aligned} U^{\mathrm{T}} U &= I \in \mathbb{R}^{r \times r} \\ V^{\mathrm{T}} V &= I \in \mathbb{R}^{r \times r}. \end{aligned}$$

Here $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$, is a diagonal matrix of "singular" values, where

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r.$$

For simplicity consider square case $m = n$

$$A\mathbf{x} = \left(U \Sigma V^{\mathrm{T}}\right) \mathbf{x}. \tag{1.3}$$

The first product $V^{\mathrm{T}}\mathbf{x}$ is a rotation, which can be checked by looking at the length

$$\begin{aligned} \left\| V^{\mathrm{T}} \mathbf{x} \right\|_2 &= \sqrt{\mathbf{x}^{\mathrm{T}} V V^{\mathrm{T}} \mathbf{x}} \\ &= \sqrt{\mathbf{x}^{\mathrm{T}} \mathbf{x}} \\ &= \|\mathbf{x}\|_2, \end{aligned} \tag{1.4}$$

which shows that the length of the vector is unchanged after application of the linear transformation represented by V^{T} so that operation must be a rotation.

Similarly the operation of U on $\Sigma V^{\mathrm{T}}\mathbf{x}$ also must be a rotation. The operation $\Sigma = [\sigma_i]_i$ applies a scaling operation to each component of the vector $V^{\mathrm{T}}\mathbf{x}$.

All linear (square) transformations can therefore be thought of as a rotate-scale-rotate operation. Often the A of interest will be symmetric $A = A^{\mathrm{T}}$.

1.4 Set of symmetric matrices

Let S^n be the set of real, symmetric $n \times n$ matrices.

Theorem 1.1: Spectral theorem.

When $A \in S^n$ then it is possible to factor A as

$$A = Q\Lambda Q^{\mathrm{T}},$$

where Q is an orthogonal matrix, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots \lambda_n)$. Here $\lambda_i \in \mathbb{R}\,\forall i$ are the (real) eigenvalues of A.

A real symmetric matrix $A \in S^n$ is "positive semi-definite" if

$$\mathbf{v}^{\mathrm{T}} A \mathbf{v} \geq 0 \qquad \forall \mathbf{v} \in \mathbb{R}^n, \mathbf{v} \neq 0,$$

and is "positive definite" if

$$\mathbf{v}^{\mathrm{T}} A \mathbf{v} > 0 \qquad \forall \mathbf{v} \in \mathbb{R}^n, \mathbf{v} \neq 0.$$

The set of such matrices is denoted S^n_+, and S^n_{++} respectively.

Consider $A \in S^n_+$ (or S^n_{++})

$$A = Q\Lambda Q^{\mathrm{T}}, \tag{1.5}$$

possible since the matrix is symmetric. For such a matrix

$$\begin{aligned} \mathbf{v}^{\mathrm{T}} A \mathbf{v} &= \mathbf{v}^{\mathrm{T}} Q \Lambda A^{\mathrm{T}} \mathbf{v} \\ &= \mathbf{w}^{\mathrm{T}} \Lambda \mathbf{w}, \end{aligned} \tag{1.6}$$

where $\mathbf{w} = A^{\mathrm{T}}\mathbf{v}$. Such a product is

$$\mathbf{v}^{\mathrm{T}} A \mathbf{v} = \sum_{i=1}^{n} \lambda_i w_i^2. \tag{1.7}$$

So, if $\lambda_i \geq 0$ ($\lambda_i > 0$) then $\sum_{i=1}^n \lambda_i w_i^2$ is non-negative (positive) $\forall \mathbf{w} \in \mathbb{R}^n, \mathbf{w} \neq 0$. Since $\mathbf{w}$ is just a rotated version of $\mathbf{v}$ this also holds for all $\mathbf{v}$. A necessary and sufficient condition for $A \in S^n_+$ (S^n_{++}) is $\lambda_i \geq 0$ ($\lambda_i > 0$).

1.5 Square root of positive semi-definite matrix

Real symmetric matrix power relationships such as

$$\begin{aligned} A^2 &= Q\Lambda Q^{\mathrm{T}} Q\Lambda Q^{\mathrm{T}} \\ &= Q\Lambda^2 Q^{\mathrm{T}}, \end{aligned} \tag{1.8}$$

or more generally $A^k = Q\Lambda^k Q^{\mathrm{T}}$, $k \in \mathbb{Z}$, can be further generalized to non-integral powers. In particular, the square root (non-unique) of a square matrix can be written

$$A^{1/2} = Q \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_n} \end{bmatrix} Q^{\mathrm{T}}, \tag{1.9}$$

since $A^{1/2}A^{1/2} = A$, regardless of the sign picked for the square roots in question.

1.6 Functions of matrices

Consider $F : S^n \rightarrow \mathbb{R}$, and define

$$F(X) = \log\det X, \tag{1.10}$$

Here $\operatorname{dom} F = S^n_{++}$. The task is to find $\boldsymbol{\nabla} F$, which can be done by looking at the perturbation $\log\det(X + \Delta X)$

$$\begin{aligned} \log\det(X + \Delta X) &= \log\det(X^{1/2}(I + X^{-1/2}\Delta X X^{-1/2})X^{1/2}) \\ &= \log\det(X(I + X^{-1/2}\Delta X X^{-1/2})) \\ &= \log\det X + \log\det(I + X^{-1/2}\Delta X X^{-1/2}). \end{aligned} \tag{1.11}$$

Let $X^{-1/2}\Delta X X^{-1/2} = M$ where λ_i are the eigenvalues of $M : M\mathbf{v} = \lambda_i \mathbf{v}$ when $\mathbf{v}$ is an eigenvector of M. In particular

$$(I + M)\mathbf{v} = (1 + \lambda_i)\mathbf{v}, \tag{1.12}$$

where $1 + \lambda_i$ are the eigenvalues of the $I + M$ matrix. Since the determinant is the product of the eigenvalues, this gives

$$\begin{aligned} \log\det(X + \Delta X) &= \log\det X + \log \prod_{i=1}^{n} (1 + \lambda_i) \\ &= \log\det X + \sum_{i=1}^{n} \log(1 + \lambda_i). \end{aligned} \tag{1.13}$$

If λ_i are sufficiently "small", then $\log(1 + \lambda_i) \approx \lambda_i$, giving

$$
\begin{aligned}
\log\det(X+\Delta X) &= \log\det X + \sum_{i=1}^{n} \lambda_i \\
&\approx \log\det X + \mathrm{Tr}(X^{-1/2}\Delta X X^{-1/2}).
\end{aligned}
\tag{1.14}
$$

Since

$$
\mathrm{Tr}(AB) = \mathrm{Tr}(BA), \tag{1.15}
$$

this trace operation can be written as

$$
\begin{aligned}
\log\det(X+\Delta X) &\approx \log\det X + \mathrm{Tr}(X^{-1}\Delta X) \\
&= \log\det X + \langle X^{-1}, \Delta X\rangle,
\end{aligned}
\tag{1.16}
$$

so

$$
\boldsymbol{\nabla} F(X) = X^{-1}. \tag{1.17}
$$

To check this, consider the simplest example with $X \in \mathbb{R}^{1\times 1}$, where we have

$$
\begin{aligned}
&\frac{d}{dX}\left(\log\det X\right) \\
&= \frac{d}{dX}\left(\log X\right) \\
&= \frac{1}{X} \\
&= X^{-1}.
\end{aligned}
\tag{1.18}
$$

This is a nice example demonstrating how the gradient can be obtained by performing a first order perturbation of the function. The gradient can then be read off from the result.

1.7 Second order perturbations

- To get first order approximation found the part that varied linearly in ΔX.
- To get the second order part, perturb X^{-1} by ΔX and see how that perturbation varies in ΔX.

For $G(X) = X^{-1}$, this is

$$
\begin{aligned}
(X+\Delta X)^{-1} &= \left(X^{1/2}(I + X^{-1/2}\Delta X X^{-1/2})X^{1/2}\right)^{-1} \\
&= X^{-1/2}(I + X^{-1/2}\Delta X X^{-1/2})^{-1}X^{-1/2}
\end{aligned}
\tag{1.19}
$$

To be proven in the homework (for “small” A)

$$
(I+A)^{-1} \approx I - A. \tag{1.20}
$$

This gives

$$\begin{aligned}(X+\Delta X)^{-1} &= X^{-1/2}(I - X^{-1/2}\Delta X X^{-1/2})X^{-1/2} \\ &= X^{-1} - X^{-1}\Delta X X^{-1},\end{aligned} \tag{1.21}$$

or

$$\begin{aligned}G(X+\Delta X) &= G(X) + (DG)\Delta X \\ &= G(X) + (\nabla G)^{\mathrm{T}}\Delta X,\end{aligned} \tag{1.22}$$

so

$$(\nabla G)^{\mathrm{T}}\Delta X = -X^{-1}\Delta X X^{-1}. \tag{1.23}$$

The Taylor expansion of F to second order is

$$F(X+\Delta X) = F(X) + \operatorname{Tr}\left((\nabla F)^{\mathrm{T}}\Delta X\right) + \frac{1}{2}\left((\Delta X)^{\mathrm{T}}(\nabla^2 F)\Delta X\right). \tag{1.24}$$

The first trace can be expressed as an inner product

$$\begin{aligned}\operatorname{Tr}\left((\nabla F)^{\mathrm{T}}\Delta X\right) &= \langle \nabla F, \Delta X\rangle \\ &= \langle X^{-1}, \Delta X\rangle.\end{aligned} \tag{1.25}$$

The second trace also has the structure of an inner product

$$\begin{aligned}(\Delta X)^{\mathrm{T}}(\nabla^2 F)\Delta X &= \operatorname{Tr}\left((\Delta X)^{\mathrm{T}}(\nabla^2 F)\Delta X\right) \\ &= \langle (\nabla^2 F)^{\mathrm{T}}\Delta X, \Delta X\rangle,\end{aligned} \tag{1.26}$$

where a no-op trace could be inserted in the second order term since that quadratic form is already a scalar. This $(\nabla^2 F)^{\mathrm{T}}\Delta X$ term has essentially been found implicitly by performing the linear variation of ∇F in ΔX, showing that we must have

$$\operatorname{Tr}\left((\Delta X)^{\mathrm{T}}(\nabla^2 F)\Delta X\right) = \langle -X^{-1}\Delta X X^{-1}, \Delta X\rangle, \tag{1.27}$$

so

$$F(X+\Delta X) = F(X) + \langle X^{-1}, \Delta X\rangle + \frac{1}{2}\langle -X^{-1}\Delta X X^{-1}, \Delta X\rangle, \tag{1.28}$$

or

$$\log\det(X+\Delta X) = \log\det X + \operatorname{Tr}(X^{-1}\Delta X) - \frac{1}{2}\operatorname{Tr}(X^{-1}\Delta X X^{-1}\Delta X). \tag{1.29}$$

1.8 Convex Sets

- Types of sets: Affine, convex, cones
- Examples: Hyperplanes, polyhedra, balls, ellipses, norm balls, cone of PSD matrices.

Definition 1.4: Affine set

A set $C \subseteq \mathbb{R}^n$ is affine if $\forall \mathbf{x}_1, \mathbf{x}_2 \in C$ then

$$\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in C, \qquad \forall \theta \in \mathbb{R}.$$

The affine sum above can be rewritten as

$$\mathbf{x}_2 + \theta(\mathbf{x}_1 - \mathbf{x}_2). \tag{1.30}$$

Since θ is a scaling, this is the line containing $\mathbf{x}_2$ in the direction between $\mathbf{x}_1$ and $\mathbf{x}_2$.
Observe that the solution to a set of linear equations

$$C = \{\mathbf{x} | A\mathbf{x} = \mathbf{b}\}, \tag{1.31}$$

is an affine set. To check, note that

$$\begin{aligned} A(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2) &= \theta A \mathbf{x}_1 + (1-\theta) A \mathbf{x}_2 \\ &= \theta \mathbf{b} + (1-\theta)\mathbf{b} \\ &= \mathbf{b}. \end{aligned} \tag{1.32}$$

Definition 1.5: Affine combination.

An affine combination of points $\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n$ is

$$\sum_{i=1}^{n} \theta_i \mathbf{x}_i,$$

such that for $\theta_i \in \mathbb{R}$

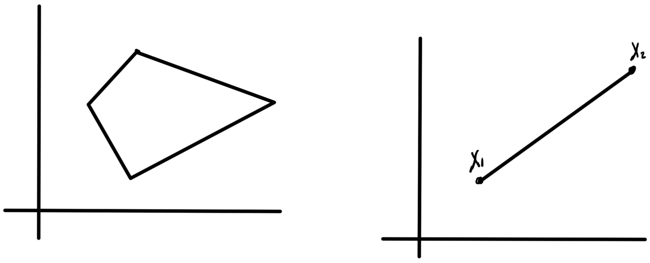
$$\sum_{i=1}^{n} \theta_i = 1.$$

An affine set contains all affine combinations of points in the set. Examples of a couple affine sets are sketched in fig. 1.1.
For comparison, a couple of non-affine sets are sketched in fig. 1.2.

Definition 1.6: Convex set

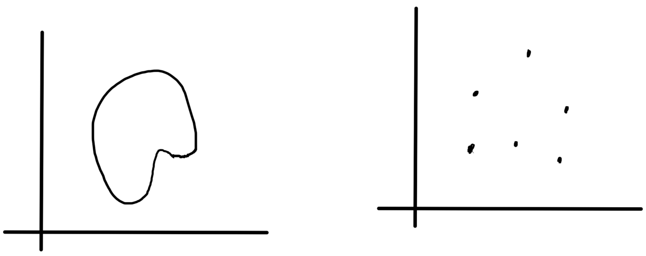
A set $C \subseteq \mathbb{R}^n$ is convex if $\forall \mathbf{x}_1, \mathbf{x}_2 \in C$ and $\forall \theta \in \mathbb{R}, \theta \in [0, 1]$, the combination

$$\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in C. \tag{1.33}$$



(a) **(b)**

Figure 1.1: Affine.



(a) **(b)**

Figure 1.2: Not affine.

Definition 1.7: Convex combination

A convex combination of $\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n$ is

$$\sum_{i=1}^{n} \theta_i \mathbf{x}_i,$$

such that $\forall \theta_i \geq 0$

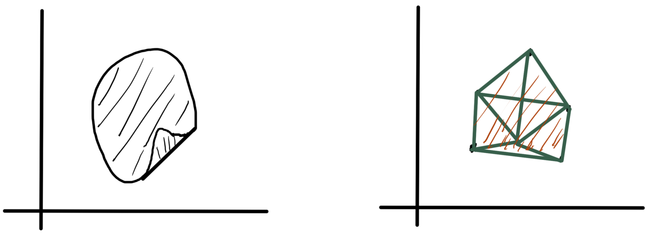
$$\sum_{i=1}^{n} \theta_i = 1$$

Definition 1.8: Convex hull.

Convex hull of a set C is a set of all convex combinations of points in C, denoted

$$\text{conv}(C) = \left\{ \sum_{i=1}^{n} \theta_i \mathbf{x}_i | \mathbf{x}_i \in C, \theta_i \geq 0, \sum_{i=1}^{n} \theta_i = 1 \right\}. \tag{1.34}$$

A non-convex set can be converted into a convex hull by filling in all the combinations of points connecting points in the set, as sketched in fig. 1.3.



(a) (b)

Figure 1.3: Convex hulls.

Definition 1.9: Cones.

A set C is a cone if $\forall \mathbf{x} \in C$ and $\forall \theta \geq 0$ we have $\theta \mathbf{x} \in C$.

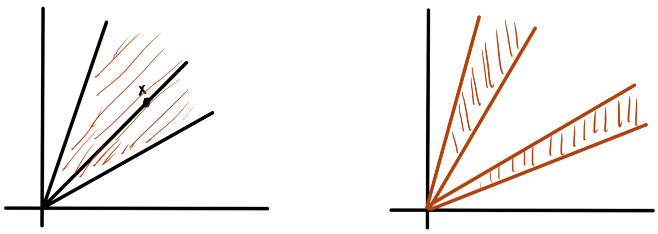
This scales out if $\theta > 1$ and scales in if $\theta < 1$.
A convex cone is a cone that is also a convex set. A conic combination is

Table 1.1: Affine, Convex, and Conic properties.

	$\theta_i \ge 0$	$\sum \theta_i = 1$
Affine	No	Yes
Convex	Yes	Yes
Conic	Yes	No

$$\sum_{i=1}^{n} \theta_i \mathbf{x}_i, \theta_i \ge 0.$$

A convex and non-convex 2D cone is sketched in fig. 1.4



(a) (b)

Figure 1.4: Convex and non-convex cone.

A comparison of properties for different set types is tabulated in table 1.1.

1.9 Hyperplanes and half spaces

Definition 1.10: Hyperplane.

A hyperplane is defined by

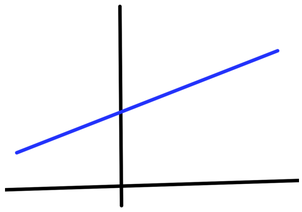
$$\left\{ \mathbf{x} | \mathbf{a}^{\mathrm{T}} \mathbf{x} = \mathbf{b}, \mathbf{a} \ne 0 \right\}.$$

A line and plane are examples of this general construct as sketched in fig. 1.5.
An alternate view is possible should one find any specific $\mathbf{x}_0$ such that $\mathbf{a}^{\mathrm{T}} \mathbf{x}_0 = \mathbf{b}$

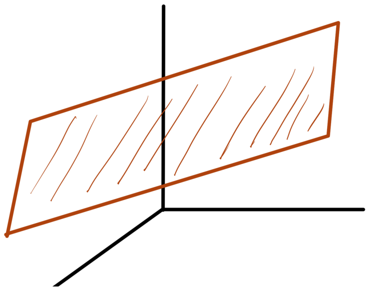
$$\left\{ \mathbf{x} | \mathbf{a}^{\mathrm{T}} \mathbf{x} = b \right\} = \left\{ \mathbf{x} | \mathbf{a}^{\mathrm{T}} (\mathbf{x} - \mathbf{x}_0) = 0 \right\} \tag{1.35}$$

This shows that $\mathbf{x} - \mathbf{x}_0 = \mathbf{a}^{\perp}$ is perpendicular to $\mathbf{a}$, or

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{a}^{\perp}. \tag{1.36}$$



(a)



(b)

Figure 1.5: Hyperplanes.

This is the subspace perpendicular to $\mathbf{a}$ shifted by $\mathbf{x}_0$, subject to $\mathbf{a}^T\mathbf{x}_0 = \mathbf{b}$. As a set

$$\mathbf{a}^{\perp} = \left\{ \mathbf{v} | \mathbf{a}^T\mathbf{v} = 0 \right\}. \tag{1.37}$$

1.10 Half space

Definition 1.11: Half space.

The half space is defined as

$$\left\{ \mathbf{x} | \mathbf{a}^T\mathbf{x} = \mathbf{b} \right\} = \left\{ \mathbf{x} | \mathbf{a}^T(\mathbf{x} - \mathbf{x}_0) \leq 0 \right\}.$$

This can also be expressed as $\{\mathbf{x} | \langle \mathbf{a}, \mathbf{x} - \mathbf{x}_0 \rangle \leq 0\}$.