ORF 523	Lecture 2	Princeton University
Instructor: A.A. Ahmadi		Scribe: G. Hall

Any typos should be emailed to a_a_a@princeton.edu.

Today, we review basic math concepts that you will need throughout the course.

- Inner products and norms
- Positive semidefinite matrices
- Basic differential calculus

1 Inner products and norms

1.1 Inner products

1.1.1 Definition

Definition 1 (Inner product)**.** *A function* $\langle ., . \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is an inner product if*

1. $\langle x, x \rangle \geq 0, \langle x, x \rangle = 0 \Leftrightarrow x = 0$ *(positivity)*
2. $\langle x, y \rangle = \langle y, x \rangle$ *(symmetry)*
3. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ *(additivity)*
4. $\langle rx, y \rangle = r\langle x, y \rangle$ *for all* $r \in \mathbb{R}$ *(homogeneity)*

Homogeneity in the second argument follows:

$$\langle x, ry \rangle = \langle ry, x \rangle = r\langle y, x \rangle = r\langle x, y \rangle$$

using properties (2) and (4) and again (2) respectively, and

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

using properties (2), (3) and again (2).

1.1.2 Examples

- The standard inner product is

$$\langle x, y\rangle = x^T y = \sum x_i y_i,\ x, y \in \mathbb{R}^n.$$

- The standard inner product between matrices is

$$\langle X, Y\rangle = \mathrm{Tr}(X^T Y) = \sum_i \sum_j X_{ij} Y_{ij}$$

  where $X, Y \in \mathbb{R}^{m\times n}$.

Notation: Here, $\mathbb{R}^{m\times n}$ is the space of real $m \times n$ matrices. $\mathrm{Tr}(Z)$ is the trace of a real square matrix Z, i.e., $\mathrm{Tr}(Z) = \sum_i Z_{ii}$.

Note: The matrix inner product is the same as our original inner product between two vectors of length mn obtained by stacking the columns of the two matrices.

- A less classical example in $\mathbb{R}^2$ is the following:

$$\langle x, y\rangle = 5x_1y_1 + 8x_2y_2 - 6x_1y_2 - 6x_2y_1$$

  Properties (2), (3) and (4) are obvious, positivity is less obvious. It can be seen by writing

$$\langle x, x\rangle = 5x_1^2 + 8x_2^2 - 12x_1x_2 = (x_1 - 2x_2)^2 + (2x_1 - 2x_2)^2 \geq 0$$
$$\langle x, x\rangle = 0 \Leftrightarrow x_1 - 2x_2 = 0 \text{ and } 2x_1 - 2x_2 = 0 \Leftrightarrow x_1 = 0 \text{ and } x_2 = 0.$$

1.1.3 Properties of inner products

Definition 2 (Orthogonality)**.** *We say that x and y are orthogonal if*

$$\langle x, y\rangle = 0.$$

Theorem 1 (Cauchy Schwarz)**.** *For $x, y \in \mathbb{R}^n$*

$$|\langle x, y\rangle| \leq ||x||\ ||y||,$$

where $||x|| := \sqrt{\langle x, x\rangle}$ is the length of x (it is also a norm as we will show later on).

Proof: First, assume that $||x|| = ||y|| = 1$.

$$||x-y||^2 \geq 0 \Rightarrow \langle x-y, x-y\rangle = \langle x,x\rangle + \langle y,y\rangle - 2\langle x,y\rangle \geq 0 \Rightarrow \langle x,y\rangle \leq 1.$$

Now, consider any $x, y \in \mathbb{R}^n$. If one of the vectors is zero, the inequality is trivially verified. If they are both nonzero, then:

$$\left\langle \frac{x}{||x||}, \frac{y}{||y||} \right\rangle \leq 1 \Rightarrow \langle x,y\rangle \leq ||x|| \cdot ||y||. \tag{1}$$

Since (1) holds $\forall x, y$, replace y with $-y$:

$$\langle x, -y\rangle \leq ||x|| \cdot ||-y||$$
$$\langle x, -y\rangle \geq -||x|| \cdot ||y||$$

using properties (1) and (2) respectively. □

1.2 Norms

1.2.1 Definition

Definition 3 (Norm)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is a norm if*

1. $f(x) \geq 0$, $f(x) = 0 \Leftrightarrow x = 0$ *(positivity)*
2. $f(\alpha x) = |\alpha| f(x),\ \forall \alpha \in \mathbb{R}$ *(homogeneity)*
3. $f(x+y) \leq f(x) + f(y)$ *(triangle inequality)*

Examples:

- The 2-norm: $||x|| = \sqrt{\sum_i x_i^2}$
- The 1-norm: $||x||_1 = \sum_i |x_i|$
- The inf-norm: $||x||_\infty = \max_i |x_i|$
- The p-norm: $||x||_p = (\sum_i |x_i|^p)^{1/p}$, $p \geq 1$

Lemma 1. *Take any inner product $\langle ., . \rangle$ and define $f(x) = \sqrt{\langle x, x\rangle}$. Then f is a norm.*

Proof: Positivity follows from the definition.
For homogeneity,

$$f(\alpha x) = \sqrt{\langle \alpha x, \alpha x \rangle} = |\alpha| \sqrt{\langle x, x \rangle}$$

We prove triangular inequality by contradiction. If it is not satisfied, then $\exists x, y$ s.t.

$$\begin{aligned} \sqrt{\langle x+y, x+y \rangle} &> \sqrt{\langle x, x \rangle} + \sqrt{\langle y, y \rangle} \\ \Rightarrow \langle x+y, x+y \rangle &> \langle x, x \rangle + 2\sqrt{\langle x, x \rangle \langle y, y \rangle} + \langle y, y \rangle \\ \Rightarrow 2\langle x, y \rangle &> 2\sqrt{\langle x, x \rangle \langle y, y \rangle} \end{aligned}$$

which contradicts Cauchy-Schwarz.

Note: Not every norm comes from an inner product.

1.2.2 Matrix norms

Matrix norms are functions $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ that satisfy the same properties as vector norms. Let $A \in \mathbb{R}^{m \times n}$. Here are a few examples of matrix norms:

- The Frobenius norm: $||A||_F = \sqrt{\text{Tr}(A^T A)} = \sqrt{\sum_{i,j} A_{i,j}^2}$
- The sum-absolute-value norm: $||A||_{sav} = \sum_{i,j} |X_{i,j}|$
- The max-absolute-value norm: $||A||_{mav} = \max_{i,j} |A_{i,j}|$

Definition 4 (Operator norm)**.** *An operator (or induced) matrix norm is a norm*

$$||.||_{a,b} : \mathbb{R}^{m \times n} \to \mathbb{R}$$

defined as

$$\begin{aligned} ||A||_{a,b} = &\max_x ||Ax||_a \\ &s.t.\ ||x||_b \leq 1, \end{aligned}$$

where $||.||_a$ is a vector norm on $\mathbb{R}^m$ and $||.||_b$ is a vector norm on $\mathbb{R}^n$.

Notation: When the same vector norm is used in both spaces, we write

$$\begin{aligned} ||A||_c = &\max ||Ax||_c \\ &\text{s.t. } ||x||_c \leq 1. \end{aligned}$$

Examples:

- $||A||_2 = \sqrt{\lambda_{\max}(A^T A)}$, where $\lambda_{\max}$ denotes the largest eigenvalue.
- $||A||_1 = \max_j \sum_i |A_{ij}|$, i.e., the maximum column sum.
- $||A||_\infty = \max_i \sum_j |A_{ij}|$, i.e., the maximum row sum.

Notice that not all matrix norms are induced norms. An example is the Frobenius norm given above as $||I||_* = 1$ for any induced norm, but $||I||_F = \sqrt{n}$.

Lemma 2. *Every induced norm is submultiplicative, i.e.,*

$$||AB|| \leq ||A||\ ||B||.$$

Proof: We first show that $||Ax|| \leq ||A||\ ||x||$. Suppose that this is not the case, then

$$\begin{aligned} &||Ax|| > ||A||\ |x|| \\ \Rightarrow &\frac{1}{||x||}||Ax|| > ||A|| \\ \Rightarrow &\left|\left|A\frac{x}{||x||}\right|\right| > ||A|| \end{aligned}$$

but $\frac{x}{||x||}$ is a vector of unit norm. This contradicts the definition of $||A||$.

Now we proceed to prove the claim.

$$||AB|| = \max_{||x||\leq 1} ||ABx|| \leq \max_{||x||\leq 1} ||A||\ ||Bx|| = ||A|| \max_{||x||\leq 1} ||Bx|| = ||A||\ ||B||.$$

□

Remark: This is only true for induced norms that use the same vector norm in both spaces. In the case where the vector norms are different, submultiplicativity can fail to hold. Consider e.g., the induced norm $||\cdot||_{\infty,2}$, and the matrices

$$A = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

In this case,

$$||AB||_{\infty,2} > ||A||_{\infty,2} \cdot ||B||_{\infty,2}.$$

Indeed, the image of the unit circle by A (notice that A is a rotation matrix of angle $\pi/4$) stays within the unit square, and so $||A||_{\infty,2} \leq 1$. Using similar reasoning, $||B||_{\infty,2} \leq 1$.

This implies that $||A||_{\infty,2}||B||_{\infty,2} \leq 1$. However, $||AB||_{\infty,2} \geq \sqrt{2}$, as $||ABx||_\infty = \sqrt{2}$ for $x = (1,0)^T$.

Example of a norm that is not submultiplicative:

$$||A||_{mav} = \max_{i,j} |A_{i,j}|$$

This can be seen as any submultiplicative norm satisfies

$$||A^2|| \leq ||A||^2.$$

In this case,

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } A^2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

So $||A^2||_{mav} = 2 > 1 = ||A||^2_{mav}$.

Remark: Not all submultiplicative norms are induced norms. An example is the Frobenius norm.

1.2.3 Dual norms

Definition 5 (Dual norm)**.** *Let $||.||$ be any norm. Its dual norm is defined as*

$$\begin{aligned} ||x||_* = &\max x^T y \\ &s.t.\ ||y|| \leq 1. \end{aligned}$$

You can think of this as the operator norm of x^T.
The dual norm is indeed a norm. The first two properties are straightforward to prove. The triangle inequality can be shown in the following way:

$$||x+z||_* = \max_{||y||\leq 1}(x^T y + z^T y) \leq \max_{||y||\leq 1} x^T y + \max_{||y||\leq 1} z^T y = ||x||_* + ||z||_*$$

□

Examples:

1. $||x||_{1*} = ||x||_\infty$

2. $||x||_{2*} = ||x||_2$

3. $||x||_{\infty *} = ||x||_1$.

Proofs:

- The proof of (1) is left as an exercise.
- Proof of (2): We have

$$||x||_{2*} = \max_y x^T y$$
$$\text{s.t. } ||y||_2 \leq 1.$$

  Cauchy-Schwarz implies that

$$x^T y \leq ||x||\ ||y|| \leq ||x||$$

  and $y = \frac{x}{||x||}$ achieves this bound.

- Proof of (3): We have

$$||x||_{\infty *} = \max_y x^T y$$
$$\text{s.t. } ||y||_\infty \leq 1$$

  So $y_{opt} = \text{sign}(x)$ and the optimal value is $||x||_1$.

2 Positive semidefinite matrices

We denote by $S^{n\times n}$ the set of all symmetric (real) $n \times n$ matrices.

2.1 Definition

Definition 6. *A matrix $A \in S^{n\times n}$ is*

- *positive semidefinite (psd) (notation: $A \succeq 0$) if*

$$x^T Ax \geq 0,\ \forall x \in \mathbb{R}^n.$$

- *positive definite (pd) (notation: $A \succ 0$) if*

$$x^T Ax > 0,\ \forall x \in \mathbb{R}^n,\ x \neq 0.$$

- *negative semidefinite if $-A$ is psd. (Notation: $A \preceq 0$)*
- *negative definite if $-A$ is pd. (Notation: $A \prec 0$.)*

Notation: $A \succeq 0$ means A is psd; $A \geq 0$ means that $A_{ij} \geq 0$, for all i, j.

Remark: Whenever we consider a quadratic form $x^T Ax$, we can assume without loss of generality that the matrix A is symmetric. The reason behind this is that any matrix A can be written as

$$A = \left(\frac{A + A^T}{2}\right) + \left(\frac{A - A^T}{2}\right)$$

where $B := \left(\frac{A+A^T}{2}\right)$ is the symmetric part of A and $C := \left(\frac{A-A^T}{2}\right)$ is the anti-symmetric part of A. Notice that $x^T Cx = 0$ for any $x \in \mathbb{R}^n$.

Example: The matrix

$$M = \begin{pmatrix} 5 & 1 \\ 1 & -2 \end{pmatrix}$$

is indefinite. To see this, consider $x = (1, 0)^T$ and $x = (0, 1)^T$.

2.2 Eigenvalues of positive semidefinite matrices

Theorem 2. *The eigenvalues of a symmetric real-valued matrix A are real.*

Proof: Let $x \in \mathbb{C}^n$ be a nonzero eigenvector of A and let $\lambda \in \mathbb{C}$ be the corresponding eigenvalue; i.e., $Ax = \lambda x$. By multiplying either side of the equality by the conjugate transpose x^* of eigenvector x, we obtain

$$x^* Ax = \lambda x^* x, \tag{2}$$

We now take the conjugate of both sides, remembering that $A \in S^{n \times n}$:

$$x^* A^T x = \bar{\lambda} x^* x \Rightarrow x^* Ax = \bar{\lambda} x^* x \tag{3}$$

Combining (2) and (3), we get

$$\lambda x^* x = \bar{\lambda} x^* x \Rightarrow x^* x(\lambda - \bar{\lambda}) = 0 \Rightarrow \lambda = \bar{\lambda},$$

since $x \neq 0$.

Theorem 3.

$$A \succeq 0 \Leftrightarrow \text{ all eigenvalues of } A \text{ are } \geq 0$$
$$A \succ 0 \Leftrightarrow \text{ all eigenvalues of } A \text{ are } > 0$$

Proof: We will just prove the first point here. The second one can be proved analogously. ($\Rightarrow$) Suppose some eigenvalue λ is negative and let x denote its corresponding eigenvector. Then

$$Ax = \lambda x \Rightarrow x^T Ax = \lambda x^T x < 0 \Rightarrow A \not\succeq 0.$$

($\Leftarrow$) For any symmetric matrix, we can pick a set of eigenvectors $v_1, \ldots, v_n$ that form an orthogonal basis of $\mathbb{R}^n$. Pick any $x \in \mathbb{R}^n$.

$$\begin{aligned} x^T Ax &= (\alpha_1 v_1 + \ldots + \alpha_n v_n)^T A(\alpha_1 v_1 + \ldots + \alpha_n v_n) \\ &= \sum_i \alpha_i^2 v_i^T A v_i = \sum_i \alpha_i^2 \lambda_i v_i^T v_i \geq 0 \end{aligned}$$

where we have used the fact that $v_i^T v_j = 0$, for $i \neq j$.

2.3 Sylvester's characterization

Theorem 4.

$$A \succeq 0 \Leftrightarrow \textit{ All } 2^n - 1 \textit{ principal minors are nonnegative.}$$
$$A \succ 0 \Leftrightarrow \textit{ All } n \textit{ leading principal minors are positive.}$$

Minors are determinants of subblocks of A. Principal minors are minors where the block comes from the same row and column index set. Leading principal minors are minors with index set $1, \ldots, k$ for $k = 1, \ldots, n$. Examples are given below.

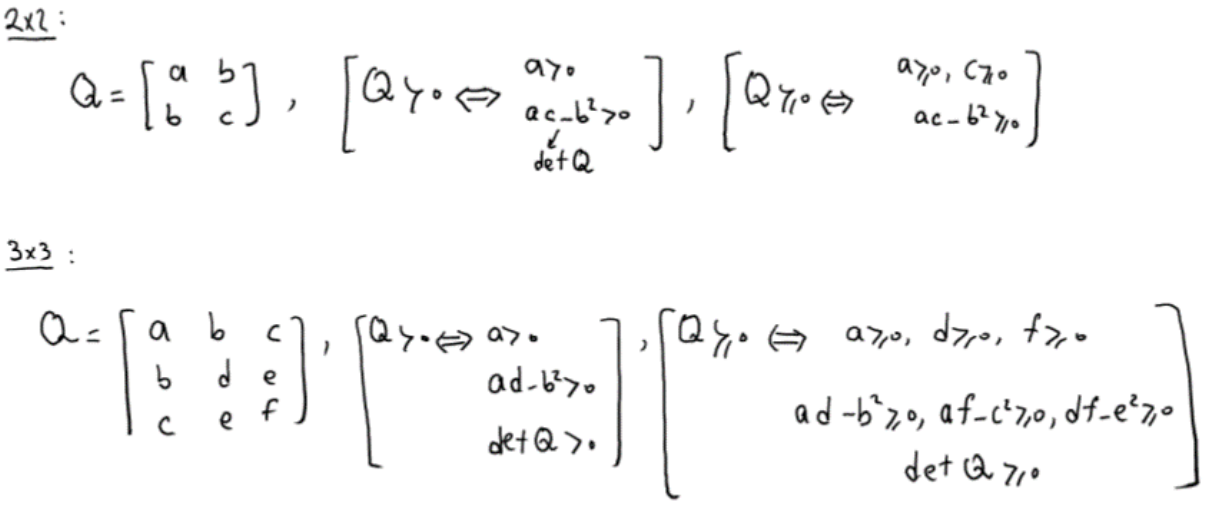

Figure 1: A demonstration of the Sylverster criteria in the 2×2 and 3×3 case.

Proof: We only prove ($\Rightarrow$). Principal submatrices of psd matrices should be psd (why?). The determinant of psd matrices is nonnegative (why?).

3 Basic differential calculus

You should be comfortable with the notions of continuous functions, closed sets, boundary and interior of sets. If you need a refresher, please refer to [1, Appendix A].

3.1 Partial derivatives, Jacobians, and Hessians

Definition 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$.*

- *The partial derivative of f with respect to x_i is defined as*

$$\frac{\partial f}{\partial x_i} = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}.$$

- *The gradient of f is the vector of its first partial derivatives:*

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

- *Let* $f : \mathbb{R}^n \to \mathbb{R}^m$*, in the form* $f = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$*. Then the Jacobian of* f *is the* $m \times n$ *matrix of first derivatives:*

$$J_f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

- *Let* $f : \mathbb{R}^n \to \mathbb{R}$*. Then the Hessian of* f*, denoted by* $\nabla^2 f(x)$*, is the* $n \times n$ *symmetric matrix of second derivatives:*

$$(\nabla^2 f)_{ij} = \frac{\partial f}{\partial x_i \partial x_j}.$$

3.2 Level Sets

Definition 8 (Level sets)**.** *The* α*-level set of a function* $f : \mathbb{R}^n \to \mathbb{R}$ *is the set*

$$S_\alpha = \{x \in \mathbb{R}^n \mid f(x) = \alpha\}.$$

Definition 9 (Sublevel sets)**.** *The* α*-sublevel set of a function* $f : \mathbb{R}^n \to \mathbb{R}$ *is the set*

$$\bar{S}_\alpha = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}.$$

Lemma 3. *At any point* x*, the gradient is orthogonal to the level set.*

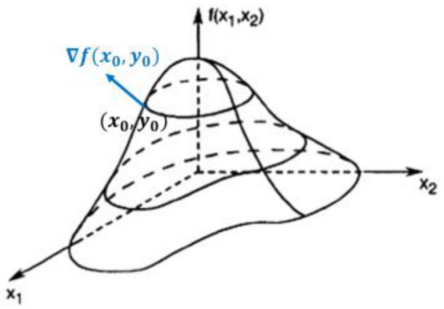

Figure 2: Illustration of Lemma 3

3.3 Common functions

We will encounter the following functions from $\mathbb{R}^n$ to $\mathbb{R}$ frequently. It is also useful to remember their gradients and Hessians.

- Linear functions:

$$f(x) = c^T x, c \in \mathbb{R}^n, c \neq 0.$$

- Affine functions:

$$f(x) = c^T x + b, c \in \mathbb{R}^n, b \in \mathbb{R}$$
$$\nabla f(x) = c, \nabla^2 f(x) = 0.$$

- Quadratic functions

$$f(x) = x^T Q x + c^T x + b$$
$$\nabla f(x) = 2Qx + c$$
$$\nabla^2 f(x) = 2Q.$$

3.4 Differentiation rules

- Product rule. Let $f, g : \mathbb{R}^n \to \mathbb{R}^m$, $h(x) = f^T(x)g(x)$ then

$$J_h(x) = f^T(x)J_g(x) + g^T(x)J_f(x) \text{ and } \nabla h(x) = J_h^T(x)$$

- Chain rule. Let $f : \mathbb{R} \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}$, $h(t) = g(f(t))$ then

$$h'(t) = \nabla f^T(f(t)) \begin{pmatrix} f_1'(t) \\ \vdots \\ f_n'(t) \end{pmatrix}.$$

Important special case: Fix $x, y \in \mathbb{R}^n$. Consider $g : \mathbb{R}^n \to \mathbb{R}$ and let

$$h(t) = g(x + ty).$$

Then,

$$h'(t) = y^T \nabla g(x + ty).$$

3.5 Taylor expansion

- Let $f \in C^m$ (m times continuously differentiable). The Taylor expansion of a univariate function around a point a is given by

$$f(b) = f(a) + \frac{h}{1!}f'(a) + \frac{h^2}{2!}f''(a) + \ldots + \frac{h^m}{m!}f^{(m)}(a) + o(h^m)$$

  where $h := b - a$. We recall the "little o" notation: we say that $f = o(g(x))$ if

$$\lim_{x \to 0} \frac{|f(x)|}{|g(x)|} = 0.$$

  In other words, f goes to zero faster than g.

- In multiple dimensions, the first and second order Taylor expansions of a function $f : \mathbb{R}^n \to \mathbb{R}$ will often be useful to us:

  First order: $f(x) = f(x_0) + \nabla f^T(x_0)(x - x_0) + o(||x - x_0||)$.

  Second order: $f(x) = f(x_0) + \nabla f^T(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0) + o(||x - x_0||^2)$.

Notes

For more background material see [1, Appendix A].

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, http://stanford.edu/ boyd/cvxbook/, 2004.

[2] E.K.P Chong and S.H. Zak. *An Introduction to Optimization, Fourth Edition*. Wiley, 2013.