PRINCETON UNIVERSITY
ORFE

Instructor:
Amir Ali Ahmadi
Spring 2016

TA: Georgina Hall

- This lecture:

  Mathematical Background

  - Inner products and norms
  - Positive semidefinite matrices
  - Basics of differential calculus

We also establish our notation.

Inner products and norms

A function $\langle \cdot , \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an <u>inner product</u> if

① $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$ (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 $\langle r x, y \rangle = r \langle x, y \rangle, \ \forall r \in \mathbb{R}$ (homogeneity)

- Additivity in the second argument follows:

$$\langle x, y + z \rangle \overset{②}{=} \langle y + z, x \rangle \overset{③}{=} \langle y, x \rangle + \langle z, x \rangle \overset{②}{=} \langle x, y \rangle + \langle x, z \rangle$$

- Homogeneity in the second argument follows:

$$\langle x, r y \rangle \overset{②}{=} \langle r y, x \rangle \overset{④}{=} r \langle y, x \rangle \overset{②}{=} r \langle x, y \rangle$$

<u>Examples</u>:

- The standard inner product in $\mathbb{R}^n$:

$$\langle x,y\rangle = x^T y = \sum_{i=1}^{n} x_i y_i. \qquad (x,y\in\mathbb{R}^n)$$

- The standard inner product between matrices:

$$\langle X,Y\rangle = Tr(X^T Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij} Y_{ij} \qquad (X,Y\in\mathbb{R}^{m\times n}).$$

Notation. $\mathbb{R}^{m\times n}$: The space of real $m\times n$ matrices.

$Tr(Z)$: The trace of a (square) matrix Z; i.e., $\sum_i Z_{ii}$.

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

- An example of a less standard inner product in $\mathbb{R}^2$:

$$\langle x,y\rangle = 5x_1y_1 + 8x_2y_2 - 6x_1y_2 - 6x_2y_1 \qquad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^T \begin{pmatrix} 5 & -6 \\ -6 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Symmetry ✓

homogeneity ✓

additivity ✓

Positivity: $\langle x,x\rangle = 5x_1^2 + 8x_2^2 - 12x_1x_2 = (x_1-2x_2)^2 + (2x_1-2x_2)^2$

$$\langle x,x\rangle = 0 \Rightarrow \begin{cases} x_1 = 2x_2 \\ x_1 = x_2 \end{cases} \Rightarrow x_1 = x_2 = 0.$$

If $\langle x,y\rangle = 0$, we say x and y are orthogonal.

Given an inner product $\langle \cdot, \cdot \rangle$, define the length of a vector x to be:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

Theorem (Cauchy-Schwarz inequality). $|\langle x, y \rangle| \leq \|x\| \, \|y\|$.

Proof. Suppose first $\|x\| = \|y\| = 1$.

$$\|y-x\|^2 \geq 0 \Rightarrow \langle y-x, y-x \rangle \geq 0 \Rightarrow \langle y, y \rangle + \langle x, x \rangle - 2\langle x, y \rangle \geq 0 \Rightarrow 2 \geq 2\langle x, y \rangle$$

$$\Rightarrow \langle x, y \rangle \leq 1.$$

Now consider general $x, y \in \mathbb{R}^n$; $x, y \neq 0$ (otherwise the inequality is trivial). We know

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

Finally, since (1) holds $\forall x, y$, replace y with $-y$

$$\Rightarrow \langle x, -y \rangle \leq \|x\| \, \|-y\| \Rightarrow \langle x, y \rangle \geq -\|x\| \, \|y\| \quad ②$$

$$① + ② \Rightarrow |\langle x, y \rangle| \leq \|x\| \, \|y\|. \quad \square$$

Norms:

A function $f: \mathbb{R}^n \to \mathbb{R}$ is a norm if

① $f(x) \geq 0 \quad \forall x$, $f(x) = 0 \Leftrightarrow x = 0$ (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:

$\|x\|_2 = \sqrt{\sum x_i^2}$, $\|x\|_1 = \sum |x_i|$, $\|x\|_\infty = \max_i |x_i|$

$\|x\|_p = \left(\sum |x_i|^p\right)^{1/p}$, $p \geq 1$.

Lemma. Let $\langle x, y\rangle$ be any inner product, then $f(x) = \sqrt{\langle x, x\rangle}$ is a norm.

Proof. Positivity follows from definition. Homogeneity:

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

Triangle inequality:

Suppose not $\Rightarrow x, y$ s.t. $\sqrt{\langle x+y, x+y\rangle} > \sqrt{\langle x, x\rangle} + \sqrt{\langle y, y\rangle}$

$\Rightarrow \langle x+y, x+y\rangle > \langle x, x\rangle + \langle y, y\rangle + 2\sqrt{\langle x, x\rangle \langle y, y\rangle}$

$\Rightarrow 2\langle x, y\rangle > 2\sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle}$

Contradicting the Cauchy Schwarz inequality. □

Note: Not every norm comes from an inner product.

Matrix norms: One can also define norms on matrices.

$\|X\|_F = \mathrm{Tr}^{1/2}(X^T X) = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)^{1/2}$. (the Frobenius norm)

$\|X\|_{sav} = \sum\sum |X_{ij}|$ (sum-absolute-value)

$$\|X\|_{max} = \max_{i,j} |X_{ij}|$$

Operator norms.

Let $\|\cdot\|_a$, $\|\cdot\|_b$ be norms on $\mathbb{R}^m$ and $\mathbb{R}^n$. We can define the induced matrix norm on $A \in \mathbb{R}^{m\times n}$ as

$$\|A\|_{a,b} = \max \|Ax\|_a \quad \text{s.t.} \quad \|x\|_b \leq 1$$

This is indeed a norm. Proof of triangle inequality:

$$\|A+B\|_{a,b} = \max_{\|x\|_b \leq 1} \|Ax+Bx\|_a \leq \max_{\|x\|_b \leq 1} \|Ax\|_a + \|Bx\|_a$$

$$\leq \max_{\|x\|_b \leq 1} \|Ax\|_a + \max_{\|x\|_b \leq 1} \|Bx\|_b = \|A\|_{a,b} + \|B\|_{a,b} \quad \square$$

- Notation: $\|A\|_a := \|A\|_{a,a}$; i.e., the same vector norm is used in both spaces.
- Three common induced norms:

$$\|A\|_2 = \sqrt{\lambda_{max}(A^T A)} \qquad \left(\|A\|_2 := \|A\|_{2,2} \right)$$

$$\|A\|_1 = \max_j \sum_i |A_{ij}| \qquad \text{(maximum column sum)}$$

$$\|A\|_\infty = \max_i \sum_j |A_{ij}| \qquad \text{(maximum row sum)}$$

Not every matrix norm is an induced norm: $\|A\|_F$ isn't (why?)

$\|I\|_F = \sqrt{n}$, but identity always has induced norm one (why?)

Induced norms are <u>submultiplicative</u>: $\|AB\| \leq \|A\| \, \|B\|$

Let's first show that $\forall A \in \mathbb{R}^{m \times n}$, $\forall x \in \mathbb{R}^n$ we have $\|Ax\| \leq \|A\| \, \|x\|$.

Suppose not: $\|Ax\| > \|A\| \, \|x\|$. $\Rightarrow$ $\left\| A \frac{x}{\|x\|} \right\| > \|A\|$, contradicting the definition of $\|A\|$ as $\max\limits_{\text{s.t. } \|y\| \leq 1} \|Ay\|$.

Now, $\|AB\| = \max\limits_{\text{s.t. } \|x\| \leq 1} \|ABx\| \leq \max\limits_{\text{s.t. } \|x\| \leq 1} \|A\| \, \|Bx\| = \|A\| \max\limits_{\text{s.t. } \|x\| \leq 1} \|Bx\| = \|A\| \, \|B\|$.

- Not all norms are submultiplicative:

  e.g., $\|A\|_{max} = \max\limits_{i,j} |A_{ij}|$.

  Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \Rightarrow A^2 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$. For any submultiplicative norm $\|\cdot\|$ we must have

  $\|A^2\| \leq \|A\|^2$. But $\|A^2\|_{max} = 2 > \|A\|_{max}^2 = 1$.

- But not every submultiplicative norm is an operator norm: e.g., $\|A\|_F$ is submulticative (why?)

Dual norms.

Let $\|\cdot\|$ be any norm. Its dual norm is defined as

$$\|x\|_* = \max_{y} \; x^T y .$$
$$\text{s.t. } \|y\| \le 1$$

So you can think of this as the operator norm of x^T.

- The dual norm is a norm:

$$\|x+z\|_* = \max_{y \text{ s.t. } \|y\|\le 1} x^T y + z^T y \le \max_{y \text{ s.t. } \|y\|\le 1} x^T y + \max_{y \text{ s.t. } \|y\|\le 1} z^T y$$
$$= \|x\|_* + \|z\|_*.$$

Other properties also easy to check.

- Dual of common norms: $\|x\|_{1*} \overset{①}{=} \|x\|_\infty$, $\|x\|_{2*} \overset{②}{=} \|x\|_2$, $\|x\|_{\infty*} \overset{③}{=} \|x\|_1$.

Proof of ③

$$\|x\|_{\infty *} = \max_{y,\ \|y\|_\infty \le 1} x^T y$$

$$y_{opt} = \text{sign}(x) \Rightarrow \text{optimal value} = \|x\|_1.$$

Proof of ②:

$$\|x\|_{2*} = \max_{y,\ \|y\|_2 \le 1} x^T y$$

Cauchy-Schwarz $\Rightarrow$ $x^T y \le \|x\|\,\|y\| \le \|x\|$.

But $y = x$ acheives this bound.

Proof of ①: Exercise.

Positive semidefinite matrices

Given a matrix $A \in \mathbb{R}^{n\times n}$, we'll be looking often at the quadratic form $x^T A x$. Whenever you see $x^T A x$, w.l.o.g. you may assume A is symmetric.

Here's why:

$$A = \underbrace{\left(\frac{A + A^T}{2}\right)}_{\text{Symmetric part of } A}^{B} + \underbrace{\left(\frac{A - A^T}{2}\right)}_{\text{Anti-symmetric part of } A}^{C}, \qquad \text{Note: } x^T C x = 0.$$

<u>Notation.</u> $S^{n\times n}$: the space of symmetric (real) $n\times n$ matrices.

<u>Definition.</u> A matrix $A \in S^{n\times n}$ is said to be

- Positive semidefinite (psd) if $x^T A x \geq 0 \ \forall x \in \mathbb{R}^n$. Notation: $A \succeq 0$.
- Positive definite (pd) if $x^T A x > 0 \ \forall x \in \mathbb{R}^n, x \neq 0$. " : $A \succ 0$.
- Negative semidefinite (nsd) if $-A$ is psd. " : $A \preceq 0$.
- Negative definite (nd) if $-A$ is pd. " : $A \prec 0$.
- Indefinite, if it's neither psd nor nsd.

<u>Example</u>: $\begin{bmatrix} 5 & 1 \\ 1 & -2 \end{bmatrix}$ is indefinite: Consider $x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

<u>Notation comment</u>: $A \succeq 0$ means A is psd; $A \geq 0$ means $A_{ij} \geq 0 \ \forall ij$.

Eigenvalue characterization

Thm. Eigenvalues of a real symmetric matrix are real.

Proof. Let $Ax = \lambda x$, where $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n$.

$\Rightarrow$ $x^* A x = \lambda x^* x$ ①, where x^* is the conjugate transpose.

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

$$x^* A^T x = \bar{\lambda} x^* x \Rightarrow x^* A x = \bar{\lambda} x^* x \quad ② \qquad (\bar{\lambda} \text{ is the conjugate of } \lambda)$$

$$① + ② \Rightarrow (\lambda - \bar{\lambda}) x^* x = 0 \underset{\substack{\text{evec} \\ \text{non-zero}}}{\Longrightarrow} \lambda = \bar{\lambda} \Rightarrow \lambda \text{ is real.} \quad \square$$

Thm. A is psd $\Leftrightarrow$ all eigenvalues of A are nonnegative.

A is pd $\Leftrightarrow$ " " " " " positive.

Proof. We only prove the "psd case". The pd claim is similar.

($\Rightarrow$) Suppose some eigenvalue λ is negative.

$$Ax = \lambda x \Rightarrow x^T A x = \underbrace{\lambda}_{<0} \underbrace{x^T x}_{>0} < 0 \Rightarrow A \text{ not psd.}$$

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

Pick any $x \in \mathbb{R}^n$.

$$x^T A x = (\alpha_1 v_1 + \cdots + \alpha_n v_n)^T A (\alpha_1 v_1 + \cdots + \alpha_n v_n)$$

$$\underset{\substack{v_i^T v_j = 0 \\ i \neq j}}{=} \sum_{i=1}^{n} \alpha_i^2 v_i^T A v_i = \sum \underbrace{\alpha_i^2}_{\geq 0} \underbrace{\lambda_i}_{\geq 0} \underbrace{v_i^T v_i}_{> 0} \geq 0. \quad \square$$

Sylvester's characterization.

Thm. $A \succeq 0 \Leftrightarrow$ All $2^n - 1$ principal minors are nonnegative.

$A \succ 0 \Leftrightarrow$ All n leading principal minors are positive.

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

2x2:

$Q = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$, $\left[Q \succ 0 \Leftrightarrow \begin{matrix} a > 0 \\ \underbrace{ac - b^2}_{\det Q} > 0 \end{matrix} \right]$, $\left[Q \succeq 0 \Leftrightarrow \begin{matrix} a \geq 0, \ c \geq 0 \\ ac - b^2 \geq 0 \end{matrix} \right]$

3x3:

$Q = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$, $\left[Q \succ 0 \Leftrightarrow \begin{matrix} a > 0 \\ ad - b^2 > 0 \\ \det Q > 0 \end{matrix} \right]$, $\left[Q \succeq 0 \Leftrightarrow \begin{matrix} a \geq 0, \ d \geq 0, \ f \geq 0 \\ ad - b^2 \geq 0, \ af - c^2 \geq 0, \ df - e^2 \geq 0 \\ \det Q \geq 0 \end{matrix} \right]$

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

Differential Calculus

- <u>Continuity</u>. A function $f: \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $a \in \mathbb{R}^n$ if

$$\forall \epsilon > 0, \exists \delta > 0 \quad \text{s.t.} \quad \|x - a\| \leq \delta \Rightarrow \|f(x) - f(a)\| \leq \epsilon.$$

where the choice of the norm is yours.

- <u>Jacobians, gradients, and Hessians</u>

∘ Let $f: \mathbb{R}^n \to \mathbb{R}$. $\quad \frac{\partial f}{\partial x_i} = \lim_{\epsilon \to 0} \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}$, where e_i is the i^{th} standard basis vector.

∘ For $f: \mathbb{R}^n \to \mathbb{R}^m$, the <u>Jacobian</u> matrix J_f is the $m \times n$ matrix of first partial derivatives:

$$J_f(x) = \begin{pmatrix} \partial f_1/\partial x_1 & \cdots & \partial f_1/\partial x_n \\ \vdots & & \\ \partial f_m/\partial x_1 & \cdots & \partial f_m/\partial x_n \end{pmatrix}.$$

∘ For $f: \mathbb{R}^n \to \mathbb{R}$, the <u>gradient</u> vector ∇f is defined as

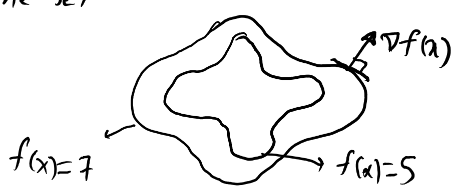
$$\nabla f(x) = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}$$

∘ For $f: \mathbb{R}^n \to \mathbb{R}$, the <u>Hessian</u> $\nabla^2 f$ is the symmetric matrix of partial derivatives:

$$\left[\nabla^2 f(x)\right]_{ij} = \frac{\partial f}{\partial x_i \partial x_j}$$

∘ For a function $f: \mathbb{R}^n \to \mathbb{R}$, the <u>$\alpha$-level set</u> is the set

$$S_\alpha = \{x \mid f(x) = \alpha\}.$$



Basic functions we encounter frequently:

- Linear: $f(x) = c^T x$, $c \in \mathbb{R}^n$, $c \neq 0$
- Affine: $f(x) = c^T x + b$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}$

  $\nabla f(x) = c$, $\nabla^2 f(x) = 0$.

- Quadratic: $f(x) = x^T Q x + c^T x + b$

  $\nabla f(x) = 2Qx + c$

  $\nabla^2 f(x) = 2Q$

Differentiation Rules:

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

  then, $J_h(x) = f^T(x) J_g(x) + g^T(x) J_f(x)$, $\nabla_h(x) = J_h^T(x)$.

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

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

- 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,

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

Taylor expansion:

Let $f \in C^m$ (m times continously differentiable)

Taylor expansion around a (in one variable):

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

where $h := b - a$, and the "little o" is defined as follows:

$f(x) = o(g(x))$ if $\lim_{x \to 0} \frac{|f(x)|}{|g(x)|} = 0$ ("f goes to zero faster than g".)

In many dimensions, the following pop up often:

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

1st order: $f(x) = f(x_0) + \nabla f^T(x_0)(x - x_0) + o(\|x - x_0\|)$

2nd 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 Appendeix A of [BV04].

References

- [BV04] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge Press, 2004.

- [CZ13] E.K.P. Chong and S.H. Zak. An Introduction to Optimization. Fourth Edition. Wiley, 2013.