CSC 576: Mathematical Foundations I

Ji Liu
Department of Computer Sciences, University of Rochester

September 20, 2016

1 Notations and Assumptions

In most cases (if without local definitions), we use

- Greek alphabets such as α, β, and γ to denote real numbers;
- Small letters such as x, y, and z to denote vectors;
- Capital letters to denote matrices, e.g., A, B, and C.

Other notations:

- $\mathbb{R}$ is the one dimensional Euclidean space;
- $\mathbb{R}^n$ is the n dimensional *vector* Euclidean space;
- $\mathbb{R}^{m\times n}$ is the $m \times n$ dimensional *matrix* Euclidean space;
- $\mathbb{R}_+$ denotes the range $[0, +\infty)$;
- $1_n \in \mathbb{R}^n$ denotes a vector with 1 in all entries;
- For any vector $x \in \mathbb{R}^n$, we use $|x|$ to denote the absolute vector, that is, $|x|_i = |x_i|\ \forall i = 1, \cdots, n$;
- $\odot$ denotes the component-wise product, that is, for any vectors x and y, $(x \odot y)_i = x_i y_i$.

Some assumptions:

- Unless explicit (local) definition, we always assume that all vectors are column vectors.

2 Vector norms, Inner product

A function $f:\ x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}_+$ is called a “norm”, if the following three conditions are satisfied

- (Zero element) $f(x) \geq 0$ and $f(x) = 0$ if and only if $x = 0$;
- (Homogeneous) For any $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$, $f(\alpha x) = |\alpha| f(x)$;
- (Triangle inequality) Any $x, y \in \mathbb{R}^n$ satisfy $f(x) + f(y) \geq f(x + y)$.

The ℓ_2 norm "$\|\cdot\|_2$" (a special "$f(\cdot)$") in $\mathbb{R}^n$ is defined as

$$\|x\|_2 = (|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)^{\frac{1}{2}}.$$

Because of ℓ_2 is the most commonly used norm (also known as Euclidean norm), we denote it as $\|\cdot\|$ sometimes for short. (Think about it how about $f([x_1, x_2]) = 2x_1^2 + x_2^2$?)

A general ℓ_p norm ($p \geq 1$) is defined as

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}.$$

Note that for $p < 1$, it is not a "norm" since the triangle inequality is violated. ℓ_∞ norm is defined as

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \cdots, |x_n|\}.$$

One may notice that the ℓ_∞ norm is the limit of the ℓ_p norm, that is, for any $x \in \mathbb{R}^n$, $\|x\|_\infty = \lim_{p\to+\infty} \|x\|_p$. In addition, people use $\|x\|_0$ to denote the ℓ_0 "norm".

The inner product $\langle\cdot,\cdot\rangle$ in $\mathbb{R}^n$ is defined as

$$\langle x, y\rangle = \sum_i x_i y_i.$$

One can show that $\langle x, x\rangle = \|x\|^2$. Two vectors x and y are orthogonal if $\langle x, y\rangle = 0$. That is one reason why ℓ_2 norm is so special.

If $p \geq q$, then for any $x \in \mathbb{R}^n$ we have $\|x\|_p \leq \|x\|_q$. In particular, we have

$$\|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty.$$

To bound from the order sides, we have

$$\|x\|_1 \leq \sqrt{n}\|x\|_2 \quad \|x\|_2 \leq \sqrt{n}\|x\|_\infty.$$

Proof. To see the first one, we have

$$\|x\|_1 = \langle 1_n,\ |x|\rangle \leq \|1_n\|_2 \||x|\|_2 = \sqrt{n}\|x\|_2$$

where the last inequality uses the Cauchy inequality. I leave the proof of the second inequality in your homework. □

Given a norm "$\|\cdot\|_A$", its dual norm is defined as

$$\|x\|_{A^*} = \max_{\|y\|_A \leq 1} \langle x, y\rangle = \max_{\|y\|_A = 1} \langle x, y\rangle = \max_z \frac{\langle x, z\rangle}{\|z\|_A}.$$

Several important properties about the dual norm are

- The dual norm's dual norm is itself, that is, $\|x\|_{(A^*)^*} = \|x\|_A$;
- The ℓ_2 norm is self-dual, that is, the dual norm of the ℓ_2 norm is still the ℓ_2 norm;
- The dual norm of the ℓ_p norm ($p \geq 1$) is ℓ_q norm where p and q satisfy $1/p + 1/q = 1$. Particularly, ℓ_1 norm and ℓ_∞ norm are dual to each other.
- (Holder inequality): $\langle x, y\rangle \leq \|x\|_A \|y\|_{A^*}$

3 Linear space, subspace, linear transformation

A set S is a linear space if

- $0 \in S$;
- given any two points $x \in S$, $y \in S$ in S and any two scalars $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, we have

$$\alpha x + \beta y \in S.$$

Note that $\emptyset$ is not a linear space. Examples: vector space $\mathbb{R}^n$, matrix space $\mathbb{R}^{m \times n}$. How about the following things:

- 0; (no)
- $\{0\}$; (yes)
- $\{x \mid Ax = b\}$ where A is a matrix and b is a vector. ($b = 0$ yes; otherwise, no)

Let S be a linear space. A set S' is a subspace if S' is a linear space and also a subset of S. Actually, "subspace" is equivalent to "linear space", because any subspace is a linear space and any linear space is a subspace. They are indeed talking about the same thing.

Let S be a linear space. A function $L(\cdot)$ is a linear transformation if given any two points $x, y \in S$ and two scalars $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, one has

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y).$$

For vector space, there exists a 1-1 correspondence between a linear transformation and a matrix. Therefore, we can simply say "a matrix is a linear transformation".

- Prove that $\{L(x) \mid x \in S\}$ is a linear space if S is a linear space and L is a linear transformation.
- Prove that $\{x \mid L(x) \in S\}$ a linear space assuming S is a linear space, and L is a linear transformation.

How to express a subspace? The most intuitive way is to use a bunch of vectors. A subspace can be expressed by

$$\text{span}\{x_1, x_2, \cdots, x_n\} = \left\{ \sum_{i=1}^{n} \alpha_i x_i \mid \alpha_i \in \mathbb{R} \right\} = \{X\alpha \mid \alpha\},$$

which is called the range space of matrix X. A subspace can be also represented by the null space of X by

$$\{\alpha \mid X\alpha = 0\}.$$

4 Eigenvalues / eigenvectors, rank, SVD, inverse

The transpose of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as $A^T \in \mathbb{R}^{n \times m}$:

$$(A^T)_{ij} = A_{ji}.$$

One can verify that

$$(AB)^T = B^T A^T.$$

A matrix $B \in \mathbb{R}^{n \times n}$ is the inverse of an invertible matrix $A \in \mathbb{R}^{n \times n}$ if

$$AB = I \quad \text{and} \quad BA = I.$$

B can be denoted as A^{-1}. A has the inverse is equivalent to that A has a full rank (the definition for "rank" will be clear very soon.) Note that the inverse of a matrix is unique. One can also verify that if both A and B are invertible, then

$$(AB)^{-1} = B^{-1} A^{-1}.$$

The "transpose" and the "inverse" are exchangeable:

$$(A^T)^{-1} = (A^{-1})^T.$$

When we write A^{-1}, we have to make sure that A is invertible.

Given a square matrix $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$ ($x \neq \mathbf{0}$) is called its eigenvector and $\lambda \in \mathbb{R}^n$ is called its eigenvalue, if the following relationship is satisfied

$$Ax = \lambda x. \quad \text{(The effect of applying the linear transformation } A \text{ on } x \text{ is nothing but scaling it.)}$$

Note that

- If $\{\lambda, x\}$ is a pair of eigenvalue-eigenvector, then so is $\{\lambda, \alpha x\}$ for any $\alpha \neq 0$.
- One eigenvalue may correspond to multiple different eigenvectors. "Different" means eigenvectors are different after normalization.

If the matrix A is symmetric, then any two eigenvectors (corresponding to different eigenvalues) are orthogonal, that is, if $A^T = A$, $Ax_1 = \lambda_1 x_1$, $Ax_2 = \lambda_2 x_2$, and $\lambda_1 \neq \lambda_2$, then

$$x_1^T x_2 = 0.$$

Proof. Consider $x_1^T A x_2$. We have

$$x_1^T A x_2 = x_1^T (Ax_2) = x_1^T (Ax_2) = x_1^T (\lambda_2 x_2) = \lambda_2 x_1^T x_2,$$

and

$$x_1^T A x_2 = (x_1^T A) x_2 = (A^T x_1)^T x_2 \underbrace{=}_{A = A^T} (Ax_1)^T x_2 = \lambda_1 x_1^T x_2.$$

Therefore, we have

$$\lambda_2 x_1^T x_2 = \lambda_1 x_1^T x_2.$$

Since $\lambda_1 \neq \lambda_2$, we obtain $x_1^T x_2 = 0$. $\square$

A matrix $A \in \mathbb{R}^{m\times n}$ is a "rank-1" matrix, if A can be expressed as

$$A = xy^T$$

where $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, and $x \neq 0$, $y \neq 0$. The rank of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$\begin{aligned}\operatorname{rank}(A) &= \min\left\{ r \mid A = \sum_{i=1}^{r} x_i y_i^T,\ x_i \in \mathbb{R}^m, y_i \in \mathbb{R}^n \right\} \\ &= \min\left\{ r \mid A = \sum_{i=1}^{r} B_i,\ B_i \text{ is a "rank-1" matrix} \right\}.\end{aligned}$$

Examples: $[1,1;1,1]$, $[1,1;2,2]$, and many natural images have the low rank property. "Low rank" implies that the contained information is few.

We say "$U \in \mathbb{R}^{m\times n}$ has orthogonal columns" if $U^T U = I$, that is, any two columns $U_{i\cdot}$ and $U_{j\cdot}$ of U satisfies

$$U_{i\cdot}^T U_{j\cdot} = 0 \quad \text{if} \quad i \neq j; \quad \text{otherwise} \quad U_{i\cdot}^T U_{j\cdot} = 1.$$

Swapping any two columns in U to get U', U' still satisfies $U'^T U' = I$.

- $\|Ux\| = \|x\| \quad \forall x$.
- $\|U^T y\| \leq \|y\| \quad \forall y$.

If U is a square matrix and has orthogonal columns, then we call it "orthogonal matrix". It has some nice properties

- $U^{-1} = U^T$ (which means that $UU^T = U^T U = I$.)
- U^T is also an orthogonal matrix.
- The effect of applying the transformation U on a vector x is to rotate x, that is, $\|Ux\| = \|x\| = \|U^T x\|$.

"SVD" is short for "singular value decomposition", which is the most important concept in linear algebra and matrix analysis. SVD almost explores all structures of a matrix. Given *any* matrix $A \in \mathbb{R}^{m\times n}$, it can be decomposed into

$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i U_{i\cdot} V_{i\cdot}^T$$

where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ have orthogonal columns, and $\Sigma = \operatorname{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$ is a diagonal matrix with positive diagonal elements. σ_i's are called singular values, which are positive and are arranged in the decreasing order.

- $\operatorname{rank}(A) = r$;
- $\|Ax\| \leq \sigma_1 \|x\|$. Why?

A matrix $B \in \mathbb{R}^{n\times n}$ is positive semi-definite (PSD), if the following things are satisfied

- B is symmetric;
- $\forall x \in \mathbb{R}^n$, we have $x^T Bx \geq 0$.

The positive definite matrix is defined by adding one more condition

- $x^T Bx = 0 \;\Leftrightarrow x = 0$.

We can also use an equivalent definition for PSD matrices in the following: A matrix $B \in \mathbb{R}^{n\times n}$ is positive semi-definite (PSD), if the SVD of B can be written as

$$B = U\Sigma U^T$$

where $U^T U = I$ and Σ is a diagonal matrix with nonnegative diagonal elements. Examples of PSD matrices: I, $A^T A$.

Assume matrices A and B are invertible. We have the following identity:

$$B^{-1} = A^{-1} - B^{-1}(B - A)A^{-1}.$$

The Sherman-Morrison-Woodbury Formula is very useful to calculate the matrix inverse:

$$(A + UV^\top)^{-1} = A^{-1} - A^{-1}U(I + V^\top A^{-1}U)^{-1}V^\top A^{-1}.$$

This result is especially important from the perspective of computation. A special case would be that U and V are two vectors u and v. Then it is in form of

$$(A + uv^\top)^{-1} = A^{-1} - (1 + v^\top A^{-1}u)^{-1}A^{-1}uv^\top A^{-1},$$

which can be calculated with complexity $O(n^2)$ if A^{-1} is known.

The Sylvester's determinant theorem is

$$\det(I_m + AB) = \det(I_n + BA).$$

5 Matrix norms (spectral norm, nuclear norm, Frobenius norm)

The Frobenius norm (F-norm) of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$\|A\|_F = \left(\sum_{1\leq i\leq m, 1\leq j\leq n} |A_{i,j}|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1} \sigma_i^2\right)^{\frac{1}{2}}$$

If A is a vector, one can verify that $\|A\|_F = \|A\|_2$.

The inner product $\langle\cdot,\cdot\rangle$ in $\mathbb{R}^{m\times n}$ is defined as

$$\langle X, Y\rangle = \sum_{i,j} X_{ij}Y_{ij} = \text{trace}(X^T Y) = \text{trace}(YX^T) = \text{trace}(XY^T) = \text{trace}(Y^T X).$$

An important property for trace(AB):

$$\text{trace}(AB) = \text{trace}(BA) = \text{trace}(A^T B^T) = \text{trace}(B^T A^T).$$

One may notice that $\langle X, X\rangle = \|X\|_F^2$.

The spectral (trace) norm of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$\|A\|_{\text{spec}} = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=1, \|y\|=1} y^T Ax = \sigma_1(A)$$

The nuclear norm of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$\|A\|_{\text{tr}} = \sum_i \sigma_i(A) = \text{trace}(\Sigma)$$

where Σ is the diagonal matrix of SVD of $A = U\Sigma V^T$.

An important relationship

$$\|A\|_{\text{spec}} \le \|A\|_F \le \|A\|_{\text{tr}} \quad \text{and} \quad \text{rank}(A)\|A\|_{\text{spec}} \ge \sqrt{\text{rank}(A)}\|A\|_F \ge \|A\|_{\text{tr}}.$$

The dual norm for a matrix norm $\|\cdot\|_A$ is defined as

$$\|Y\|_{A^*} := \max_{\|X\|\le 1} \frac{\langle X, Y\rangle}{\|X\|_A} = \max_X \langle X, Y\rangle. \tag{1}$$

We have the following properties (think about why it is true):

$$\|X\|_{\text{spec}^*} = \|X\|_{\text{tr}}, \quad \|X\|_{F^*} = \|X\|_F.$$

6 Matrix and Vector Differential

Let $f(X) : \mathbb{R}^{m\times n} \to \mathbb{R}$ be a function with respect to matrix $X \in \mathbb{R}^{m\times n}$. It is differential (or gradient) is defined as

$$\frac{\partial f(X)}{\partial X} = \begin{bmatrix} \frac{\partial f(X)}{\partial X_{11}} & \cdots & \frac{\partial f(X)}{\partial X_{1j}} & \cdots & \frac{\partial f(X)}{\partial X_{1n}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(X)}{\partial X_{i1}} & \cdots & \frac{\partial f(X)}{\partial X_{ij}} & \cdots & \frac{\partial f(X)}{\partial X_{in}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(X)}{\partial X_{m1}} & \cdots & \frac{\partial f(X)}{\partial X_{mj}} & \cdots & \frac{\partial f(X)}{\partial X_{mn}} \end{bmatrix}.$$

We provide a few examples in the following

$$\begin{aligned} f(X) = \text{trace}(A^T X) = \langle A, X\rangle \quad & \frac{\partial f(X)}{\partial X} = A \\ f(X) = \text{trace}(X^T AX) \quad & \frac{\partial f(X)}{\partial X} = (A + A^T)X \\ f(X) = \frac{1}{2}\|AX - B\|_F^2 \quad & \frac{\partial f(X)}{\partial X} = A^T(AX - B) \\ f(X) = \frac{1}{2}\text{trace}(B^T X^T XB) \quad & \frac{\partial f(X)}{\partial X} = XBB^T \\ f(X) = \frac{1}{2}\text{trace}(B^T X^T AXB) \quad & \frac{\partial f(X)}{\partial X} = \frac{1}{2}(A + A^T)XBB^T \end{aligned}$$