Matrix Calculus ¶

- See matrix calculus entry of Wikipedia.
- See the notes about matrix calculus in the cs229 linear algebra notes. Very nice complement to the notes below.

Basic relations

It is in fact unnecessary to memorize these relations as they can be obtained by two simple ways.

- Dimensional analysis. Most relation can be obtained this way without element-wise examination. See the analysis for a relatively complicated example below (No. 7).
- Examining typical element.

(1) Two vectors $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$. Let $y = \psi(x)$. Then conventionally we have the definition

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix},$$

which is called the Jacobian matrix of the transformation $\psi()$. Depending the dimension of y and x, $\frac{\partial y}{\partial x}$ might be a matrix, a column or a row vector.

(2)

$$y = Ax \Rightarrow \frac{\partial y}{\partial x} = A,$$

where matrix A does not depend on x.

(3)

$$y = Ax \Rightarrow \frac{\partial y}{\partial z} = \frac{\partial y}{\partial x}\frac{\partial x}{\partial z} = A\frac{\partial x}{\partial z},$$

where x is function of z and A does not depend on both x and z.

(4)

$$\alpha = y^T Ax \Rightarrow \frac{\partial \alpha}{\partial x} = y^T A \quad \text{and} \quad \frac{\partial \alpha}{\partial y} = x^T A^T,$$

(5)

$$\alpha = x^T Ax \Rightarrow \frac{\partial \alpha}{\partial x} = x^T (A^T + A) = 2x^T A \quad (A = A^T)$$

This relation can be obtained by simple dimension analysis. It is just a special case of (7). However, here provide another way of examining typical elements.

$$\alpha = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{i,j} x_j$$

Take derivative w.r.t x_k gives

$$\frac{\partial \alpha}{\partial x_k} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\partial x_i}{\partial x_k} a_{i,j} x_j + x_i a_{i,j} \frac{\partial x_j}{\partial x_k} \right) = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i = \sum_{j=1}^{n} x_j a_{jk}^* + \sum_{i=1}^{n} x_i a_{ik} = x^T (A^T + A)$$

(6)

$$\alpha = y^T x \Rightarrow \frac{\partial \alpha}{\partial z} = \frac{\partial \alpha}{\partial y} \frac{\partial y}{\partial z} + \frac{\partial \alpha}{\partial x} \frac{\partial x}{\partial z} = x^T \frac{\partial y}{\partial z} + y^T \frac{\partial x}{\partial z} = 2x^T \frac{\partial x}{\partial z} \quad (x = y),$$

where x, y are functions of vector z. Here is the proof.

$$\alpha = \sum_{j=1}^{n} x_j y_j \Rightarrow \frac{\partial \alpha}{\partial z_k} = \sum_{j=1}^{n} \left(x_j \frac{\partial y_j}{z_k} + y_j \frac{\partial x_j}{z_k} \right)$$

The relation can also be obtained by dimensional analysis.

(7)

$$\alpha = y^T Ax \Rightarrow \frac{\partial \alpha}{\partial z} = x^T A^T \frac{\partial y}{\partial z} + y^T A \frac{\partial x}{\partial z} = 2x^T A \frac{\partial x}{\partial z} \quad (x = y, A = A^T),$$

where x, y depend on z. The equation can be proved using the result in (6) using $\alpha = y^T Ax \equiv w^T x$.

A simple way to apply this relation. Assuming the dimensions of x, y, z, A are respectively $n \times 1, m \times 1, k \times 1, m \times n$. The dimension of α is 1×1. We first have

$$\frac{\partial \alpha}{\partial z} = \frac{\partial (y^T Ax)}{\partial x} \frac{\partial x}{\partial z} + \frac{\partial (y^T Ax)}{\partial y} \frac{\partial y}{\partial z}$$

The dimension of the LHS is $1 \times k$. Thus we must arrange order or take transpose of the RHS expressions as proved earlier in order to have the same dimensions. In the analysis, be familiar with the conventional arrangement of Jacobian matrix. For example, from the dimensions of x, y, z, we know the dimension of $\frac{\partial y}{\partial z}$ is $m \times k$.

(8)

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1},$$

where A is a $m \times n$ matrix and α is a scalar parameter.

Considerations of using dimension analysis

- Using dimension analysis can directly write many equations involving matrix calculus without element-wise calculations. However, although this trick is applicable for many cases, it is not for all. For example, it is suitable for $\frac{\partial y}{\partial x}$ when x, y are either scalar or vector,but not for the cases when x, y are matrix. For example in the definition of gradient of $\nabla_A f(A)$, its dimension is given by the dimension of A, rather than by the rule we used before.
- Different notations are used in literatures. For 1D vector, people usually do not distinguish x and x^T . So the format of equations in different places such as here and cs229 notes is different.
- Note gradient and Hessian are defined only for scalar function. We cannot calculate $\nabla_x(Ax)$ as Ax is a vector. When calculating gradients, it is possible to use the dimension analysis to obtain equations. For example, for $f = x^T Ax$, $\nabla_x x^T Ax = 2Ax$. For Hessian, however, we must use the original definition to calculate its element and then obtain the matrix-vector version. We **cannot take Hessian as taking gradient twice**, as the gradient of vector is not defined. See details in cs229.
- When taking gradient like $\nabla f(Ax)$, we need make explicitly we differentiate relative to what. The above expression might have two interpretations (cs229 notes).

Applications

(1) Gradients of the Determinant

$$\nabla_A |A| = |A| A^{-T}$$
$$\nabla_A \log |A| = \frac{1}{|A|}\nabla_A |A| = A^{-T}$$

See proof in the determinant section in this write up. The formula above has been used in the derivation of ICA.

(2) Gradient and Hessian

$$\nabla_x b^T x = b$$
$$\nabla_x x^T Ax = 2Ax \quad (A = A^T)$$
$$\nabla_x^2 x^T Ax = 2A(A = A^T)$$

The hessian above is a symmetric matrix and positive semi-definite. **We can use dimension analysis to calculate the first two equations. But we need element-wise derivations for the Hessian.**. Also, when using dimension analysis, the format of above equations might be slightly different. For example, the first two equations will become $\nabla_x b^T x = b^T$ and, $\nabla_x x^T Ax = 2x^T A$.

(3) Least square.

$$||Ax - b||^2 = (Ax - b)^T(Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$

Taking the gradient with respect to x we have,

$$\nabla_x(x^T A^T Ax - 2b^T Ax + b^T b) = 2A^T Ax - 2A^T b$$

Setting this last expression equal to zero and solving for x gives the normal equations

$$x = (A^T A)^{-1} A^T b$$

With the dimension analysis approach, we can obtain the normal equation easily without another way using trace property, as in cs229 notes. To solve linear regression problems, we now have three ways:

- Using projection operator.
- Using the matrix calculus as above.
- Using gradient descent.

(4) Eigenvalues as Optimization

$$\max_{x \in \mathbb{R}^n} x^T Ax \quad \text{subject to } ||x||_2^2 = 1$$

for a symmetric matrix $A \in \mathbb{S}^n$. A standard way of solving optimization problems with equality constraints is by forming the Lagrangian, $L(x, \lambda) = x^T Ax - \lambda x^T x$, where λ is called the Lagrange multiplier associated with the equality constraint. It can be established that for x^* to be a optimal point to the problem, the gradient of the Lagrangian has to be zero at x^* (this is not the only condition, but it is required). That is,

$$\nabla_x L(x, \lambda) = \nabla_x(x^T Ax - \lambda x^T x) = 2A^T x - 2\lambda x = 0$$

Notice that this is just the linear equation $Ax = \lambda x$. This shows that the only points which can possibly maximize (or minimize) $x^T Ax$ assuming $||x||_2^2 = 1$ are the eigenvectors of A.

Norms

Vector norms

A norm is a FUNCTION that assigns a strictly positive length or size to each vector in a vector space except for the zero vector, which is assigned a length of zero. See the definition in the link https://en.wikipedia.org/wiki/Norm_(mathematics (https://en.wikipedia.org/wiki/Norm_(mathematics)).
l_p norms has the definition of

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

When $p = 1, 2, \infty$, we have the special l_1, l_2, l_∞ norms

$$\|x\|_1 = \left(\sum_{i=1}^{n} |x_i| \right)$$

$$\|x\|_2 = \sqrt{\left(\sum_{i=1}^{n} x_i^2 \right)} = x^T x$$

$$\|x\|_\infty = \max_i |x_i|$$

Matrix norms

http://mathworld.wolfram.com/MatrixNorm.html (http://mathworld.wolfram.com/MatrixNorm.html)
https://en.wikipedia.org/wiki/Matrix_norm (https://en.wikipedia.org/wiki/Matrix_norm)
Given a square complex or real matrix A, a matrix norm $\|A\|$ is a nonnegative number associated with A having the properties.

- $\|A\| > 0$ when $\|A\| \neq 0$ and $\|A\| = 0$ iff $A = 0$,
- $\|kA\| = |k| \, \|A\|$ for any scalar k,
- $\|A + B\| \leq \|A\| + \|B\|$,
- $\|AB\| \leq \|A\| \, \|B\|$

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A, then

$$\frac{1}{\|A^{-1}\|} \leq |\lambda| \leq \|A\|$$

Matrix norms induced by vector norms

Suppose a vector norm is defined on K^m. Any $m \times n$ matrix induces a linear operator from K^n to K^m with respect to the standard basis, and one defines the corresponding induced norm or operator norm on the space

$$\|A\| = \sup \left(\frac{\|Ax\|}{\|x\|} : x \in K^n \text{ with } x \neq 0 \right)$$

In particular, if the p norm for vectors is used for both spaces K^n and K^m, then the induced operator norm (matrix norm) is:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

When p takes special values of $1, 2, \infty$, we have the following special matrix norms.

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|,$$

which is simply the maximum absolute column sum of the matrix;

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|,$$

which is simply the maximum absolute row of the matrix;

$$\|A\|_2 = \sigma_{max}(A),$$

where $\sigma_{max}(A)$ represents the largest singular value of matrix A.

An important inequality for $p = 2$

$$\|A\|_2 = \sigma_{max}(A) \le \|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}},$$

where $\|A\|_F$ is the Frobenius norm. Equality holds if and only if the matrix A is a rank-one matrix or a zero matrix. This inequality can be derived from the fact that the trace of a matrix is equal to the sum of its eigenvalues.

"Entrywise" and Schatten matrix norms

"Entrywise" matrix norms treat an $m \times n$ matrix as a vector of size mn, and use one of the familiar vector norms. For example, using the p-norm for vectors, $p \ge 1$, we have

$$\|A\|_p = \|\mathrm{vec}(A)\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p \right)^{\frac{1}{p}}$$

This is a different norm from the induced p norm and the Schatten p norm (see Wiki), but the notation is the same.

Geometric significance of matrix norms

If we choose $p = 2$ for the induced matrix norm, then $\|A\|_2 = \sigma_{max}(A)$, where $\sigma_{max}(A)$ represents the largest singular value of matrix A. We know the geometric significance of singular value decomposition is 'rotate' -> 'stretch' -> 'rotate'. The stretching part is from the diagonal matrix Σ, where each diagonal element represents the stretching or scaling factor to the axes. So $\|A\|_2$ is just the largest scaling factor of A to a vector, or to a basis. This can also be seen in its original definition

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$