MA333 Introduction to Big Data Science
Mathematical Preliminary

Zhen Zhang

Southern University of Science and Technology

Outlines

Inner Product and Euclidean Norm

For $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, their inner product is defined as

$$< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T \mathbf{y}.$$

It satisfies

1. (Commutativity) $< \mathbf{x}, \mathbf{y} >=< \mathbf{y}, \mathbf{x} >$;
2. (Scalar Multiplication) $< \lambda \mathbf{x}, \mathbf{y} >= \lambda < \mathbf{x}, \mathbf{y} >=< \mathbf{x}, \lambda \mathbf{y} >$;
3. (Bilinearity) $< \mathbf{x} + \mathbf{y}, \mathbf{z} >=< \mathbf{x}, \mathbf{z} > + < \mathbf{y}, \mathbf{z} >$,
   $< \mathbf{x}, \mathbf{y} + \mathbf{z} >=< \mathbf{x}, \mathbf{y} > + < \mathbf{x}, \mathbf{z} >$;
4. (Positivity) $< \mathbf{x}, \mathbf{x} >\geqslant 0$, and $< \mathbf{x}, \mathbf{x} >= 0$ iff $\mathbf{x} = \mathbf{0}$.

The Euclidean norm (l_2-norm) is $\|\mathbf{x}\| = \sqrt{< \mathbf{x}, \mathbf{x} >}$.

Linear Independency and Orthogonality

- Linear Independency :
  A set of vectors $U = \{bx_1, \ldots, \mathbf{x}_k\}$ is linearly independent if for $\forall i$, $\mathbf{x}_i$ does not lie in the space spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \mathbf{x}_k$. We say U spans a subspace V if V is the span of the vectors in U. U is a basis of V if it is both independent and spans V. The dimension of V is the size of a basis of V (i.e., the number of linearly independent vectors in U).
- Orthogonality :
  We say that U is an orthogonal set if for all $i \neq j$, $< \mathbf{x}_i, \mathbf{x}_j >= 0$. We say that U is an orthonormal set if it is orthogonal and if for every i, $\|\mathbf{x}_i\| = 1$.

Gram-Schmidt Orthogonalization

Given a set of linear independent vectors $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, we can apply Gram-Schmidt orthogonalization to obtain an orthonormal set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ which have the same span as $\operatorname{span} V$. The procedure is as follows :

1. Let $\mathbf{u}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|$;
2. For $j = 2$ to k, project $\mathbf{v}_j$ onto span$\{\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}\}$ and find the perpendicular part $\tilde{\mathbf{u}}_j = \mathbf{v}_j - \sum_{i=1}^{j-1} < \mathbf{u}_i, \mathbf{v}_j > \mathbf{u}_i$, then normalize it to be $\mathbf{u}_j = \tilde{\mathbf{u}}_j / \|\tilde{\mathbf{u}}_j\|$;

This procedure is summarized in the matrix form : $Q = AP$, where $Q = (\mathbf{u}_1 \cdots \mathbf{u}_k) \in \mathbb{R}^{k \times k}$ is an orthogonal matrix whose columns are given by $\mathbf{u}_i$'s, $A = (\mathbf{v}_1 \cdots \mathbf{v}_k) \in \mathbb{R}^{k \times k}$ is a nonsingular matrix whose columns are given by $\mathbf{v}_i$'s, and $P \in \mathbb{R}^{k \times k}$ is an upper tridiagonal matrix whose upper tridiagonal (i, j)-entry is given by $< \mathbf{u}_i, \mathbf{v}_j >$. This is known as the QR factorization : $A = QR$ where $R = P^{-1}$.

Concepts in Matrix

- Kernel and Range :
  Given a matrix $A \in \mathbb{R}^{n \times d}$, the range of A (Range(A)) is the span of its columns and the kernel of A (Ker(A)) is the subspace of all vectors that satisfy $A\mathbf{x} = \mathbf{0}$. The rank of A is the dimension of its range and is denoted by $\mathrm{rank}(A)$ or $\mathrm{r}(A)$ for short.
- Symmetric and Definite Matrix :
  A is symmetric if $A = A^T$. A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is positive definite if for all $\mathbf{x} \in \mathbb{R}^d$, $< \mathbf{x}, A\mathbf{x} > \geqslant 0$, and equality holds if and only if ("iff") $\mathbf{x} = \mathbf{0}$. This definition can be relaxed to give semidefiniteness : A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is positive semidefinite if for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^T A\mathbf{x} \geqslant 0$. In particular, all the eigenvalues of a positive definite (resp. semidefinite) matrix are positive (resp. nonnegative). And $A = BB^T$ for some matrix B. (See next slides for eigen-decomposition)

Eigenvalues and Eigenvectors

Let $A \in \mathbb{R}^{d \times d}$ be a squared matrix. A nonzero vector $\mathbf{x}\mathbb{R}^d$ is an eigenvector of A with a corresponding eigenvalue λ if $A\mathbf{x} = \lambda\mathbf{x}$.

Theorem

(Eigen-decomposition or Spectral Decomposition) If $A \in \mathbb{R}^{d \times d}$ is a symmetric matrix of rank k, then there exists an orthogonal basis of $\mathbb{R}^d$, $\mathbf{x}_1, \ldots, \mathbf{x}_d$, such that each $\mathbf{x}_i$ is an eigenvector of A. Furthermore, A can be written as $A = \sum_{i=1}^{d} \lambda_i \mathbf{x}_i \mathbf{x}_i^T$, where each λ_i is the eigenvalue corresponding to the eigenvector $\mathbf{x}_i$. In matrix form, this is $A = UDU^T$, where the columns of U are the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_d$, and $D = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$ is a diagonal matrix. Finally, $\mathrm{r}(A)$ is the number of nonzero λ_i's, and the corresponding eigenvectors span the range of A. The eigenvectors corresponding to the zero eigenvalues span the null space of A.

Singular Values Decomposition (SVD)

Let $A \in \mathbb{R}^{m \times n}$ be a matrix of rank r. Unit (nonzero) vector $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{u} \in \mathbb{R}^m$ are called right and left singular vectors of A with corresponding singular values σ if $A\mathbf{v} = \sigma \mathbf{v}$ and $\mathbf{u}^T A = \sigma \mathbf{u}^T$.

Theorem

(SVD) Let $A \in \mathbb{R}^{m \times n}$ be a matrix of rank r. Then there exist orthonormal sets of right and left singular vectors of A, say $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ respectively, and the corresponding singular values $\sigma_1, \ldots, \sigma_r$, such that $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. In matrix form, this is $A = UDV^T$, where the columns of U are the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$, the columns of V are the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$, and $D = \mathrm{diag}\{\sigma_1, \ldots, \sigma_d\}$ is a diagonal matrix.

Corollary

The squared matrices $A^T A \in \mathbb{R}^{n \times n}$ and $AA^T \in \mathbb{R}^{m \times m}$ have (a subset of) the eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ respectively, corresponding the the same eigenvalues $\sigma_1^2, \ldots, \sigma_r^2$.

Reyleigh Quotient

Theorem

Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank r. Define $\mathbf{v}_1 = \underset{\mathbf{v}\in\mathbb{R}^n:\|\mathbf{v}\|=1}{\arg\max} \|A\mathbf{v}\|$, $\mathbf{v}_2 = \underset{\substack{\mathbf{v}\in\mathbb{R}^n:\|\mathbf{v}\|=1\\ <\mathbf{v},\mathbf{v}_1>=0}}{\arg\max} \|A\mathbf{v}\|, \ldots, \mathbf{v}_r = \underset{\substack{\mathbf{v}\in\mathbb{R}^n:\|\mathbf{v}\|=1\\ \forall i<r, <\mathbf{v},\mathbf{v}_i>=0}}{\arg\max} \|A\mathbf{v}\|$. *Then* $\mathbf{v}_1, \ldots, \mathbf{v}_r$ *is an orthonormal set of right singular vectors of A.*

Remark :(Reyleigh Quotient) If $A \in \mathbb{R}^{n\times n}$ is a squared matrix, then its eigenvalues can be found as the solution to the following optimization problems :

$$\lambda_1 = \max_{\mathbf{v}\in\mathbb{R}^n:\|\mathbf{v}\|=1} \mathbf{v}^T A\mathbf{v}, \quad \lambda_2 = \max_{\substack{\mathbf{v}\in\mathbb{R}^n:\|\mathbf{v}\|=1\\ <\mathbf{v},\mathbf{v}_1>=0}} \mathbf{v}^T A\mathbf{v},$$

$$\ldots, \quad \lambda_n = \max_{\substack{\mathbf{v}\in\mathbb{R}^n:\|\mathbf{v}\|=1\\ \forall i<n, <\mathbf{v},\mathbf{v}_i>=0}} \mathbf{v}^T A\mathbf{v}.$$

Power Method - Dominant Eigenvalue

Assume the eigenvalues of A can be sorted according to their magnitudes : $|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant |\lambda_n| \geqslant 0$. If The corresponding eigenvectors $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ form a basis of $\mathbb{R}^n$, then any vector can be expressed as $\mathbf{x} = \sum_{i=1}^n \beta_i \mathbf{v}_i$. Multiplying $\mathbf{x}$ by A on the left for n times, we have an idea

$$A^k\mathbf{x} = \sum_{i=1}^{n} \beta_i \lambda_i^k \mathbf{v}_i = \lambda_1^k \sum_{i=1}^{n} \beta_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \mathbf{v}_i \sim \lambda_1^k \beta_1 \mathbf{v}_1, \quad k \to \infty$$

1. For any nonzero vector $\mathbf{x}$, let $\mathbf{y}^{(0)} = \mathbf{x}$;
2. For $k = 0, 1, \dots$: compute the smallest integer p_k such that satisfying $y_{p_k}^{(k)} = \|\mathbf{y}^{(k)}\|_\infty$, then compute $\mathbf{x}^{(k)} = \mathbf{y}^{(k)}/y_{p_k}^{(k)}$, $\mathbf{y}^{(k+1)} = A\mathbf{x}^{(k)}$, $\mu^{(k+1)} = y_{p_k}^{(k+1)}$.

It can be shown that $\lim_{k\to\infty} \mu^{(k)} = \lambda_1$ and $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{v}_1/\|\mathbf{v}_1\|_\infty$.

Other methods : QR factorization, Householder transformations

Linear Systems

A system of m linear algebraic equations in n unknown variables can be written in the matrix form : $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. This system has solutions iff $\mathrm{r}([A, \mathbf{b}]) = \mathrm{r}(A)$.

Theorem (Solvability Condition)

1. *If $\mathrm{Ker}(A) = \{0\}$, then $A\mathbf{x} = \mathbf{b}$ either has a unique solution or has no solution. It has a solution iff $\mathbf{b} \perp \mathrm{Ker}(A^T)$.*
2. *If $\mathrm{Ker}(A) \neq \{0\}$, then $A\mathbf{x} = \mathbf{b}$ either has infinitely many solutions or has no solution. It has a solution iff $\mathbf{b} \perp \mathrm{Ker}(A^T)$.*

If $A \in \mathbb{R}^{n\times n}$ is a square matrix, we have a simple rule : the system has a unique solution iff $\det A \neq 0$. If the solution exists, we can solve it by $\mathbf{x} = A^{-1}\mathbf{b}$, where A^{-1} is the inverse of A satisfying $A^{-1}A = AA^{-1} = I$.

Moreover, we can find it by Cramer's rule : $x_i = \frac{\det A_i}{\det A}$, where A_i is the matrix obtained from A by replacing its i-th column with $\mathbf{b}$. The direct application of this formula requires $O(n!)$ arithmetic operations to find $\det A$, which is unacceptable for large n.

Gaussian Elimination

Gaussian Elimination is an algorithm that can reduce the computational complexity of solving linear systems to $O(n^3)$. It is equivalent to perform an elementary row transformation for A to obtain an upper or lower triangular matrix.

Another way to view Gaussian elimination is the LU decomposition : The k-th row transformation can be represented by a left multiplication by $M^{(k)}$, where $M^{(k)}$ is a lower triangular matrix with its diagonal entries being all 1's ; after n operations, A is transformed to an upper triangular matrix U, i.e., $M^{(n)} \cdots M^{(2)} M^{(1)} A = U$; since the inverse of a lower triangular matrix is also a lower triangular matrix, we have $A = LU$, where $L = (M^{(1)})^{-1} (M^{(2)})^{-1} \cdots (M^{(n)})^{-1}$ with its diagonal entries being all 1's.

LU Decomposition

Theorem

An $n \times n$ nonsingular matrix A can be decomposed uniquely in the form $A = LU$, where

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \cdots & l_{n,n-1} & 1 \end{pmatrix}, U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1,n} \\ 0 & u_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & u_{n-1,n} \\ 0 & \cdots & 0 & u_{n,n} \end{pmatrix}.$$

The computational complexity for LU decomposition is $O(n^3)$. If A is symmetric, we have Cholesky decomposition $A = LL^T$, where L is a lower diagonal matrix.

Iterative Solver

We introduce two commonly used iterative methods : Jacobi iteration and Gauss-Seidel iteration. First we write $A = D - L - U$, where $D = \mathrm{diag}\{a_{11}, \ldots, a_{nn}\}$, $L = \{l_{ij}\}$ and $U = \{u_{ij}\}$ are the lower and upper diagonal parts of $-A$ respectively. That means $l_{ij} = -a_{ij}$ for $i > j$ and 0 for $i \leqslant j$, $u_{ij} = -a_{ij}$ for $i < j$ and 0 for $i \geqslant j$.

- Jacobi iteration : Rewrite the linear system as $D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b}$, if D^{-1} exists ($a_{ii} \neq 0$), then we can build the iteration $\mathbf{x}^{(k)} = D^{-1}(L + U)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}$, $k = 1, 2, \ldots$
- Gauss-Seidel iteration : Rewrite the linear system as $(D - L)\mathbf{x} = U\mathbf{x} + \mathbf{b}$, if $(D - L)^{-1}$ exists ($a_{ii} \neq 0$), then we can build the iteration $\mathbf{x}^{(k)} = (D - L)^{-1}U\mathbf{x}^{(k-1)} + (D - L)^{-1}\mathbf{b}$, $k = 1, 2, \ldots$

Both are easy to implement in component form and can be written as $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$, with $T = D^{-1}(L + U)$ for Jacobi iteration and $(D - L)^{-1}U$ for Gauss-Seidel iteration. (Fixed point iteration !)

Vector Norms

Vector Norm is a non-negative real-valued function on $\mathbb{R}^n$, usually denoted by $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$, with the following properties :

1. (Positivity) $\|\mathbf{x}\| \geqslant 0$ for all $\mathbf{x} \in \mathbb{R}^n$; $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$;
2. (Homogeneity) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for $\forall\alpha \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$;
3. (Triangle Inequality) $\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$ for $\forall\mathbf{x},\mathbf{y} \in \mathbb{R}^n$.

Examples :

- l_2-norm : $\|\mathbf{x}\|_2 = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$;
- l_1-norm : $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$;
- l_∞-norm : $\|\mathbf{x}\|_\infty = \max_{1\leqslant i\leqslant n} |x_i|$;

Theorem

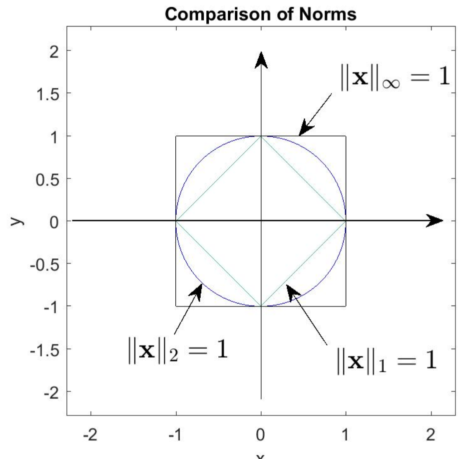
Define l_p-norm as $\|\mathbf{x}\|_p = (\sum_{i=1}^{n} x_i^p)^{\frac{1}{p}}$, it is really a norm for $p \leqslant 1$.

Vector Norms (Cont')

Remark : i) l_p-norm is not a norm for $0 < p \leqslant 1$, since the triangular inequality is not satisfied. It is called semi-norm.
ii) Useful to define l_0-norm : $\|\mathbf{x}\|_0 = \#\{1 \leqslant i \leqslant n : x_i \neq 0\}$.
Induced Distances : $\mathrm{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$, e.g., l_2-distance is $\|\mathbf{x} - \mathbf{y}\|_2 = (\sum_{i=1}^{n}(x_i - y_i)^2)^{\frac{1}{2}}$.

Theorem

$\forall \mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1$.



Matrix Norms

Matrix Norm is a non-negative real-valued function on $\mathbb{R}^{n\times m}$, usually denoted by $\|\cdot\| : \mathbb{R}^{n\times m} \to \mathbb{R}$, with the following properties :

1. (Positivity) $\|A\| \geqslant 0$ for all $A \in \mathbb{R}^{n\times m}$; $\|A\| = 0$ iff $A = 0$;
2. (Homogeneity) $\|\alpha A\| = |\alpha|\|A\|$ for $\forall \alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times m}$;
3. (Triangle Inequality) $\|A + B\| \leqslant \|A\| + \|B\|$ for $\forall A, B \in \mathbb{R}^{np\times m}$;
4. $\|AB\| \leqslant \|A\|\|B\|$ for $\forall A, B \in \mathbb{R}^{n\times m}$.

Theorem

If $\|\cdot\|$ is a vector norm on $\mathbb{R}^n$, then $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$ is a matrix norm (called natural norm).

Corollary

$\|A\mathbf{x}\| \leqslant \|A\|\|\mathbf{x}\|$ *for* $\forall A \in \mathbb{R}^{n\times m}$ *and* $\mathbf{x} \in \mathbb{R}^n$.

Matrix Norm (Cont')

Examples :

- l_1-norm : $\|A\|_1 = \max\limits_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$;
- l_∞-norm : $\|A\|_\infty = \max\limits_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|$;

l_2-norm is not trivial. For a symmetric matrix A, define its spectral radius as $\rho(A) = \max_{1\leqslant i\leqslant n} \lambda_i$, where $\lambda_i (i = 1, \ldots, n)$ are the eigenvalues of A. Then

Theorem

1. $\|A\|_2 = \sqrt{\rho(A^T A)}$;
2. $\rho(A) \leqslant \|A\|$ *for any natural norm* $\|\cdot\|$.

Theorem (Convergence of Jacobi and Gauss-Seidel Iterations)

The Jacobi and Gauss-Seidel iterations converge to the unique solution of $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ *iff* $\rho(T) < 1$. *Moreover, we have the error estimate* $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leqslant \frac{\|T\|^k}{1-\|T\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$.

Matrix Calculus

By convention, the lowercase letter a denotes a scalar, the bold letter $\mathbf{x} = (x_1, \ldots, x_n)^T$ denotes a column vector, and the uppercase letter $A = (a_{ij})$ denotes an $m \times n$ matrix. Assume $\mathbf{x}$ (or x) is independent variables, $\mathbf{a}$, $\mathbf{b}$, etc. are constant vectors, A, B, etc. are constant matrices, $f(x)$, $g(x)$, $\mathbf{u}(\mathbf{x})$, and $\mathbf{v}(\mathbf{x})$ are (scalar or vector valued) functions of $\mathbf{x}$ (or x)

- Vector-by-vector formula : (resulting in matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \left(\frac{\partial y_i}{\partial x_j}\right)$)
  1. Linear vector-valued functions : $\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = 0$, $\frac{\partial (A\mathbf{x})}{\partial \mathbf{x}} = A$, $\frac{\partial (\mathbf{x}^T A)}{\partial \mathbf{x}} = A^T$,
  2. Nonlinear vector-valued functions : $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \left(\frac{\partial u_i}{\partial x_j}\right)$ is Jacobian, $\frac{\partial (a\mathbf{u}(\mathbf{x}) + b\mathbf{v}(\mathbf{x}))}{\partial \mathbf{x}} = a\frac{\partial (\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}} + b\frac{\partial (\mathbf{v}(\mathbf{x}))}{\partial \mathbf{x}}$, $\frac{\partial (f(\mathbf{x})\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}} = f(\mathbf{x})\frac{\partial (\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial (f(\mathbf{x}))}{\partial \mathbf{x}}$, $\frac{\partial (A\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}} = A\frac{\partial (\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}}$
  3. Chain rule : $\frac{\partial \mathbf{g}(\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial (\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}}$,

Matrix Calculus

- Scalar-by-vector : (resulting in row vector $\frac{\partial y}{\partial \mathbf{x}} = (\nabla_{\mathbf{x}} y)^T$)
  Some of the formula can be obtained from the previous page by letting the numerator be of dimension one, the others are :
  1. Inner product : $\frac{\partial(\mathbf{a}^T\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial \mathbf{x}} = a^T$, $\frac{\partial \mathbf{a}^T\mathbf{u}(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}^T\frac{\partial(\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}}$,
     $\frac{\partial(\mathbf{u}(\mathbf{x})^T A\mathbf{v}(\mathbf{x}))}{\partial \mathbf{x}} = \mathbf{u}^T A\frac{\partial(\mathbf{v}(\mathbf{x}))}{\partial \mathbf{x}} + \mathbf{v}^T A^T\frac{\partial(\mathbf{u}(\mathbf{x}))}{\partial \mathbf{x}}$
  2. Quadratic forms : $\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T(A + A^T)$, $\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{x}^T A$ if A is symmetric, $\frac{\partial(\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b})}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)$
     $\frac{\partial(A\mathbf{x}+\mathbf{b})^T C(D\mathbf{x}+\mathbf{e})}{\partial \mathbf{x}} = (D\mathbf{x} + \mathbf{e})^T C^T A + (A\mathbf{x} + \mathbf{b})^T CD$
  3. l_2 norm : $\frac{\partial\|\mathbf{x}-\mathbf{a}\|}{\partial \mathbf{x}} = \frac{(\mathbf{x}-\mathbf{a})^T}{\|\mathbf{x}-\mathbf{a}\|}$
  4. 2nd order derivative (resulting in a matrix) :
     $\frac{\partial^2(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T} = (A + A^T)$, $\frac{\partial^2(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T} = 2A$ if A is symmetric,
     $\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T} = H = (\frac{\partial f}{\partial x_i \partial x_j})$ is the Hessian matrix.

Trace and Frobenius inner product

Trace is defined as the sum of the diagonal entries in a matrix : $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$

- $\mathrm{tr}(A) = \mathrm{tr}(A^T)$
- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$
- $\frac{\partial \mathrm{tr}(AB)}{\partial A} = B^T$, $\frac{\partial \mathrm{tr}(ABA^T C)}{\partial A} = CAB + C^T AB^T$
- $a = \mathrm{tr}(a)$ for scalar a, as a result, $< \mathbf{x}, \mathbf{y} >= \mathrm{tr}(\mathbf{x}^T \mathbf{y}) = \mathrm{tr}(\mathbf{y}\mathbf{x}^T)$ (useful formula)

The Frobenius inner product is defined for matrices : $< A, B >_F = \mathrm{tr}(AB^T) = \sum_{i,j=1}^{n} a_{ij} b_{ij}$. The induced norm is called Frobenius norm : $\|A\|_F = \sqrt{\mathrm{tr}(AA^T)} = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$.

A last useful formula : $\frac{\mathrm{d}}{\mathrm{d}t} \log\det(A(t)) = \mathrm{tr}(A(t)^{-1} A'(t))$

Jaccard distance

Let f be a nonnegative, monotone, submodular set function on X. The generalized Jaccard distance $J_{\delta,f}(A,B) = 1 - \frac{f(A \cap B)}{f(A \cup B)}$, when $f(A) = |A|$, we obtain the standard Jaccard distance $J_\delta(A,B) := 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \triangle B|}{|A \cup B|}$.
The Jaccard distance J_δ is known to fulfill all properties of a metric, notably, the triangle inequality, can be proved by following steps :

Lemma

For all sets $A, B, C \subseteq X$, it holds that

$$f(A \cap C) \cdot f(B \cup C) + f(A \cup C) \cdot f(B \cap C) \leq f(C) \cdot (f(A) + f(B))$$

Jaccard distance

Corollary

For all sets $S, T \subseteq X$, it holds that

$$f(S \cap T) \cdot f(S \cup T) \leq f(S) \cdot f(T).$$

Theorem

For all sets $A, B, C \subseteq X$, it holds that

$$J_{\delta,f}(A, B) \leq J_{\delta,f}(A, C) + J_{\delta,f}(C, B).$$

Outlines

References

- Numerical Analysis, 9th Edition, by Richard L. Burden, J. Douglas Faires, Brooks/Cole, 2011.