NANJING UNIVERSITY

Mathematical Background

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Nanjing University, China

September 23, 2015

Outline

Outline

Vector and Matrix

Scalar

$$x \in \mathbb{R}$$

Vector

$$\mathbf{x} \in \mathbb{R}^d$$

Matrix

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}, \text{ where } \mathbf{x}_i \in \mathbb{R}^d$$

$$X = [\mathbf{z}_1^\top; \mathbf{x}_2^\top; \ldots; \mathbf{z}_d^\top] \in \mathbb{R}^{d \times n}, \text{ where } \mathbf{z}_i \in \mathbb{R}^n$$

Vector and Matrix

Scalar

$$x \in \mathbb{R}$$

Vector

$$\mathbf{x} \in \mathbb{R}^d$$

Matrix

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}, \text{ where } \mathbf{x}_i \in \mathbb{R}^d$$

$$X = [\mathbf{z}_1^\top; \mathbf{x}_2^\top; \ldots; \mathbf{z}_d^\top] \in \mathbb{R}^{d \times n}, \text{ where } \mathbf{z}_i \in \mathbb{R}^n$$

Vector and Matrix

Scalar

$$x \in \mathbb{R}$$

Vector

$$\mathbf{x} \in \mathbb{R}^d$$

Matrix

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}, \text{ where } \mathbf{x}_i \in \mathbb{R}^d$$

$$X = [\mathbf{z}_1^\top; \mathbf{x}_2^\top; \ldots; \mathbf{z}_d^\top] \in \mathbb{R}^{d \times n}, \text{ where } \mathbf{z}_i \in \mathbb{R}^n$$

Product

Matrix Product

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d \times n}$

$$XY^\top = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i^\top \in \mathbb{R}^{d \times d}$$

Outer Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x}\mathbf{y}^\top \in \mathbb{R}^{d \times d}$$

Inner Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^{d} x_i y_i$$

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d \times n}$

$$\langle X, Y \rangle = \mathrm{trace}(X^\top Y) = \mathrm{trace}(Y^\top X) = \sum_{i,j} X_{ij} Y_{ij} = \sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{y}_i \rangle$$

Product

Matrix Product

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$

$$XY^\top = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i^\top \in \mathbb{R}^{d\times d}$$

Outer Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x}\mathbf{y}^\top \in \mathbb{R}^{d\times d}$$

Inner Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^{d} x_i y_i$$

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$

$$\langle X, Y \rangle = \mathrm{trace}(X^\top Y) = \mathrm{trace}(Y^\top X) = \sum_{i,j} X_{ij} Y_{ij} = \sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{y}_i \rangle$$

Product

Matrix Product

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$

$$XY^\top = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i^\top \in \mathbb{R}^{d\times d}$$

Outer Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x}\mathbf{y}^\top \in \mathbb{R}^{d\times d}$$

Inner Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^{d} x_i y_i$$

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$

$$\langle X, Y \rangle = \mathrm{trace}(X^\top Y) = \mathrm{trace}(Y^\top X) = \sum_{i,j} X_{ij} Y_{ij} = \sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{y}_i \rangle$$

Product

Matrix Product

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$

$$XY^\top = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i^\top \in \mathbb{R}^{d\times d}$$

Outer Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x}\mathbf{y}^\top \in \mathbb{R}^{d\times d}$$

Inner Product

- Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^{d} x_i y_i$$

- Matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$

$$\langle X, Y \rangle = \mathrm{trace}(X^\top Y) = \mathrm{trace}(Y^\top X) = \sum_{i,j} X_{ij} Y_{ij} = \sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{y}_i \rangle$$

Rank

Linear Independence

The vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are said to be linearly independent if the equation

$$\sum_{i=1}^{n} \mathbf{x}_i \alpha_i = 0$$

can only be satisfied by $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Otherwise, they are linearly dependent.

Rank

The rank of a matrix X is the dimension of the vector space spanned by its columns. This is the same as the dimension of the space spanned by its rows. It is also the maximal number of linearly independent columns or rows of X.

Rank

Linear Independence

The vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are said to be linearly independent if the equation

$$\sum_{i=1}^{n} \mathbf{x}_i \alpha_i = 0$$

can only be satisfied by $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Otherwise, they are linearly dependent.

Rank

The rank of a matrix X is the dimension of the vector space spanned by its columns. This is the same as the dimension of the space spanned by its rows. It is also the maximal number of linearly independent columns or rows of X.

Eigen Decomposition of a Matrix

Eigenvalue and Eigenvector of a Square Matrix $X \in \mathbb{R}^{d \times d}$

-
$$X\mathbf{v} = \lambda\mathbf{v}$$

Eigen-decomposition of a Matrix

- If eigenvectors of X are linearly independent
$$X = V\Lambda V^{-1}$$
where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is diagonal

Real and Symmetric Matrices

-
$$X = V\Lambda V^\top = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$$
where V is orthogonal matrix ($VV^\top = V^\top V = I$) and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$

Eigen Decomposition of a Matrix

Eigenvalue and Eigenvector of a Square Matrix $X \in \mathbb{R}^{d \times d}$

- $$X\mathbf{v} = \lambda\mathbf{v}$$

Eigen-decomposition of a Matrix

- If eigenvectors of X are linearly independent
$$X = V\Lambda V^{-1}$$
where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is diagonal

Real and Symmetric Matrices

- $$X = V\Lambda V^\top = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$$
where V is orthogonal matrix ($VV^\top = V^\top V = I$) and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$

Eigen Decomposition of a Matrix

Eigenvalue and Eigenvector of a Square Matrix $X \in \mathbb{R}^{d \times d}$

- $$X\mathbf{v} = \lambda \mathbf{v}$$

Eigen-decomposition of a Matrix

- If eigenvectors of X are linearly independent
$$X = V \Lambda V^{-1}$$
where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is diagonal

Real and Symmetric Matrices

- $$X = V \Lambda V^\top = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$$
where V is orthogonal matrix ($VV^\top = V^\top V = I$) and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$

Singular Value Decomposition (SVD)

For a matrix $X \in \mathbb{R}^{m\times n}$ with $m \le n$

- $$X = U\Sigma V^\top = \sum_{i=1}^m \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$
  where $U \in \mathbb{R}^{m\times m}$, $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_m) \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times m}$, $U^\top U = UU^\top = I$, $V^\top V = I$, and $\sigma_i \ge 0$.
- $$\mathrm{rank}(X) = |\{i : \sigma_i > 0\}|$$
- Compact SVD
  $$X = U_r \Sigma_r V_r^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$
  where $U_r \in \mathbb{R}^{m\times r}$, $\Sigma_r = \mathrm{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r\times r}$, $V_r \in \mathbb{R}^{n\times r}$, $U_r^\top U_r = I$, $V_r^\top V_r = I$, and $\sigma_i > 0$.

Singular Value Decomposition (SVD)

For a matrix $X \in \mathbb{R}^{m \times n}$ with $m \leq n$

- $$X = U\Sigma V^\top = \sum_{i=1}^{m} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$
  where $U \in \mathbb{R}^{m\times m}$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times m}$, $U^\top U = UU^\top = I$, $V^\top V = I$, and $\sigma_i \geq 0$.
- $$\mathrm{rank}(X) = |\{i : \sigma_i > 0\}|$$
- Compact SVD
  $$X = U_r \Sigma_r V_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$
  where $U_r \in \mathbb{R}^{m\times r}$, $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$, $V_r \in \mathbb{R}^{n\times r}$, $U_r^\top U_r = I$, $V_r^\top V_r = I$, and $\sigma_i > 0$.

Singular Value Decomposition (SVD)

For a matrix $X \in \mathbb{R}^{m \times n}$ with $m \leq n$

- $$X = U\Sigma V^\top = \sum_{i=1}^{m} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$
  where $U \in \mathbb{R}^{m \times m}$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times m}$, $U^\top U = UU^\top = I$, $V^\top V = I$, and $\sigma_i \geq 0$.
- $$\mathrm{rank}(X) = |\{i : \sigma_i > 0\}|$$
- Compact SVD
  $$X = U_r \Sigma_r V_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$
  where $U_r \in \mathbb{R}^{m \times r}$, $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$, $V_r \in \mathbb{R}^{n \times r}$, $U_r^\top U_r = I$, $V_r^\top V_r = I$, and $\sigma_i > 0$.

Positive Semi-definite (PSD) Matrices

A symmetric matrix $X \in \mathbb{R}^{d \times d}$ is PSD if

- $\mathbf{x}^\top X \mathbf{x} \geq 0, \forall \mathbf{x} \in \mathbb{R}^d$
- $X = UU^\top$
- all the eigenvalues are nonnegative

Norm

Vector Norm

- $\|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|$, $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$, $\|\mathbf{x}\|_\infty = \max |x_i|$,
- $\|\mathbf{x}\|_0 = |\{i : x_i \neq 0\}|$, which is non-convex

Matrix Norm $X = U\Sigma V^\top$

- Nuclear Norm: $\|X\|_* = \sum_{i=1}^{n} \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_1$
- Frobenius norm: $\|X\|_F = \sqrt{\sum_{ij} X_{ij}^2} = \sqrt{\mathrm{trace}(X^\top X)} = \|(\sigma_1, \ldots, \sigma_n)\|_2$
- Spectral Norm: $\|X\|_2 = \max_i \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_\infty$
- Rank: $\mathrm{rank}(X) = |\{i : \sigma_i > 0\}| = \|(\sigma_1, \ldots, \sigma_n)\|_0$, which is non-convex

Norm

Vector Norm

- $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$, $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$, $\|\mathbf{x}\|_\infty = \max |x_i|$,
- $\|\mathbf{x}\|_0 = |\{i : x_i \neq 0\}|$, which is non-convex

Matrix Norm $X = U\Sigma V^\top$

- Nuclear Norm: $\|X\|_* = \sum_{i=1}^n \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_1$
- Frobenius norm: $\|X\|_F = \sqrt{\sum_{ij} X_{ij}^2} = \sqrt{\mathrm{trace}(X^\top X)} = \|(\sigma_1, \ldots, \sigma_n)\|_2$
- Spectral Norm: $\|X\|_2 = \max_i \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_\infty$
- Rank: $\mathrm{rank}(X) = |\{i : \sigma_i > 0\}| = \|(\sigma_1, \ldots, \sigma_n)\|_0$, which is non-convex

Questions

Eigen-decomposition versus SVD

For a real and symmetric matrix A, what is the difference between its Eigen-decomposition and SVD?

Outline

Analysis I

- Function
- Continuous
- Differentiable
- Derivative
- Gradient/Subgradient
- Chain rule

The Challenge

How to evaluate the gradient of functions where the variable is a vector or a matrix?

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \quad \nabla f(\mathbf{x}) = \mathbf{w}$$
$$f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}, \quad \nabla f(\mathbf{x}) = 2A\mathbf{x}$$

A good reference is “The Matrix Cookbook”.

Outline

Probability and Statistics

- Random Variable
- Independent
- Expectation
- Variance/Covariance
- Probability Density Function
- Common Distributions
- Concentration Inequalities

Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent random variables. Assume that the X_i are almost surely bounded, that is,

$$\Pr(X_i \in [a_i, b_i]) = 1, \; 1 \le i \le n$$

Denote

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, we have

$$\Pr\left(\overline{X} - \mathrm{E}[\overline{X}] \ge t\right) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$\Pr\left(\overline{X} - \mathrm{E}[\overline{X}] \le -t\right) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Corollary 1

Assume $b_i - a_i \le c$. With a probability at least $1 - 2\delta$, we have

$$-\sqrt{\frac{c}{2n}\log\frac{1}{\delta}} \le \overline{X} - \mathrm{E}[\overline{X}] \le \sqrt{\frac{c}{2n}\log\frac{1}{\delta}}$$

Hoeffding's inequality

Let $X_1, \dots, X_n$ be independent random variables. Assume that the X_i are almost surely bounded, that is,

$$\Pr(X_i \in [a_i, b_i]) = 1,\ 1 \le i \le n$$

Denote

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, we have

$$\Pr\left(\overline{X} - \mathrm{E}[\overline{X}] \ge t\right) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$\Pr\left(\overline{X} - \mathrm{E}[\overline{X}] \le -t\right) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Corollary 1

Assume $b_i - a_i \le c$. With a probability at least $1 - 2\delta$, we have

$$-\sqrt{\frac{c}{2n} \log \frac{1}{\delta}} \le \overline{X} - \mathrm{E}[\overline{X}] \le \sqrt{\frac{c}{2n} \log \frac{1}{\delta}}$$

Outline

Convex Optimization

- Convexity
- Optimal Solution
- The Lagrange Dual Problem
- KKT Optimality Conditions
- Gradient Descent
- Stochastic Optimization

A good reference is “Stephen Boyd and Lieven Vandenberghe. Convex Optimization, Cambridge University Press, 2004.”