Notes for Lecture 2

Scribe: Tingkai Jia

1 Some Examples of Matrix Functions

Here are three examples of gradient calculations of matrix functions:

Example 1. $\mathbf{X} \in \mathbb{R}^{m \times n}, f(\mathbf{X}) = \|\mathbf{X}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2$

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} 2x_{11} & \cdots & 2x_{1n} \\ \vdots & \ddots & \vdots \\ 2x_{m1} & \cdots & 2x_{mn} \end{bmatrix} = 2\mathbf{X}$$

Example 2. $\mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}$$

Example 3. $\mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{A} \in \mathbb{R}^{m \times n}, f(\mathbf{X}) = tr(\mathbf{A}^T\mathbf{X})$

$$f(\mathbf{X}) = \mathrm{tr}\left(\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \right) = \sum_{i=1}^n \sum_{j=1}^m a_{ij} \cdot x_{ij}$$

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \mathbf{A}$$

2 Logistic Regression

Consider the loss function of logistic regression:

$$\mathbf{x} \in \mathbb{R}^m, f(\mathbf{x}) = \ln\left(\sum_{i=1}^n \exp(\mathbf{a}_i^T\mathbf{x} + b_i) \right).$$

Let $h(\mathbf{y}) = \sum_{i=1}^{m} \exp(y_i)$ and $g(\mathbf{y}) = \log(h(\mathbf{y}))$. Then we have $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$, where $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_m]^\top$ and $\mathbf{b} = [b_1, \dots, b_m]^\top$. By the chain rule, we can obtain:

$$g(\mathbf{y}) = g'(h(\mathbf{y}))\nabla h(\mathbf{y}) = \frac{1}{\sum_{i=1}^{m} \exp(y_i)} \begin{bmatrix} \exp(y_1) \\ \exp(y_2) \\ \vdots \\ \exp(y_m) \end{bmatrix}$$

$$\nabla f(\mathbf{x}) = \mathbf{A}^T \nabla g(\mathbf{A}\mathbf{x}+\mathbf{b}) = \mathbf{A}^T \frac{1}{\sum_{i=1}^{n} \exp(\mathbf{a}_i^T \mathbf{x} + b_i)} \begin{bmatrix} \exp(\mathbf{a}_1^T \mathbf{x} + b_1) \\ \exp(\mathbf{a}_2^T \mathbf{x} + b_2) \\ \vdots \\ \exp(\mathbf{a}_m^T \mathbf{x} + b_n) \end{bmatrix} = \frac{1}{\mathbf{1}^\top \mathbf{z}} \mathbf{A}^\top \mathbf{z}$$

where $z_i = \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)$.

3 A Property of Convex Sets

Property 1. *If $\mathcal{S}$ and $\mathcal{T}$ are convex sets, then $\mathcal{S} + \mathcal{T} = \{\mathbf{s} + \mathbf{t} | \mathbf{s} \in \mathcal{S}, \mathbf{t} \in \mathcal{T}.\}$*

Proof. Let $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}$ and $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{T}$, we have $\theta\mathbf{s}_1 + (1-\theta)\mathbf{s}_2 \in \mathcal{S}, \theta\mathbf{t}_1 + (1-\theta)\mathbf{t}_2 \in \mathcal{T}$ and $\mathbf{s}_1 + \mathbf{t}_1 \in \mathcal{S} + \mathcal{T}, \mathbf{s}_2 + \mathbf{t}_2 \in \mathcal{S} + \mathcal{T}$. Consider $\theta(\mathbf{s}_1 + \mathbf{t}_1) + (1-\theta)(\mathbf{s}_2 + \mathbf{t}_2) = \theta\mathbf{s}_1 + (1-\theta)\mathbf{s}_2 + \theta\mathbf{t}_1 + (1-\theta)\mathbf{t}_2 \in \mathcal{S} + \mathcal{T}$. Since we have shown that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S} + \mathcal{T}$ and $\lambda \in [0,1]$, $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \mathcal{S} + \mathcal{T}$, it follows that $\mathcal{S} + \mathcal{T}$ is convex. □

4 Strict Separation Theorem

Theorem 1. *Suppose $\mathcal{C}$ and $\mathcal{D}$ are nonempty disjoint convex sets. If $\mathcal{C}$ is closed and $\mathcal{D}$ is compact, there exists $\mathbf{a} \neq 0$ and b s.t.*

$$\mathbf{a}^T\mathbf{x} < b \text{ for } \mathbf{x} \in \mathcal{C}, \mathbf{a}^T\mathbf{x} > b \text{ for } \mathbf{x} \in \mathcal{D}.$$

Remark 1. *In the theorem, we must restrict both sets $\mathcal{C}$ and $\mathcal{D}$ to be closed and one of them to be bounded. Below are some relevant counterexamples:*

- *Both $\mathcal{C}$ and $\mathcal{D}$ are closed and unbounded:*

$$\mathcal{C} = \left\{(x,y) \;\middle|\; y \geq \frac{1}{x}, x > 0\right\}, \quad \mathcal{D} = \{(x,y) \mid y \leq 0\}.$$

- *$\mathcal{C}$ is open and $\mathcal{D}$ is compact:*

$$\mathcal{C} = \{(x,y) \mid x \in (0,1)\}, \quad \mathcal{D} = \{x,y) \mid y \in [1,2]\}.$$