ORIE 4741: Linear Algebra and Gradient Descent

Chengrun Yang

September 30, 2019

Outline

Full rank matrices

Pseudoinverse

Gradient descent for least squares problem

Outline

Full rank matrices

Pseudoinverse

Gradient descent for least squares problem

Full rank matrices

Claim: The followings are equivalent for any matrix $A \in \mathbb{R}^{m \times n}$:

1. $(Ax = 0 \Leftrightarrow x = 0)$
2. A has full column rank
3. $A^\top A$ is invertible

Full rank matrices

Claim: The followings are equivalent for any matrix $A \in \mathbb{R}^{m \times n}$:

1. $(Ax = 0 \Leftrightarrow x = 0)$
2. A has full column rank
3. $A^\top A$ is invertible

Equivalence of 1 and 2:

Proof.

Write A as the concatenation of column vectors $(a_1, a_2, \cdots, a_n)$. $Ax = 0$ can then be written as $\sum_{i=1}^{n} a_i x_i = 0$. Thus $(Ax = 0 \Leftrightarrow x = 0)$ is equivalent to the columns of A being linearly independent, i.e. A has full column rank. $\square$

Full rank matrices

Claim: The followings are equivalent for any matrix $A \in \mathbb{R}^{m \times n}$:

1. $(Ax = 0 \Leftrightarrow x = 0)$
2. A has full column rank
3. $A^\top A$ is invertible

Equivalence of 1 and 2:

Proof.

Write A as the concatenation of column vectors $(a_1, a_2, \cdots, a_n)$. $Ax = 0$ can then be written as $\sum_{i=1}^{n} a_i x_i = 0$. Thus $(Ax = 0 \Leftrightarrow x = 0)$ is equivalent to the columns of A being linearly independent, i.e. A has full column rank. $\square$

Question: equivalence of 1, 2 and 3?

Outline

Full rank matrices

Pseudoinverse

Gradient descent for least squares problem

Pseudoinverse

Definition: for any matrix $A \in \mathbb{R}^{m \times n}$, a pseudoinverse of A is defined as a matrix $A^\dagger \in \mathbb{R}^{n \times m}$ if it satisfies all the following:

- $AA^\dagger A = A$
- $A^\dagger A A^\dagger = A^\dagger$
- $(AA^\dagger)^\top = AA^\dagger$
- $(A^\dagger A)^\top = A^\dagger A$

Pseudoinverse

Definition: for any matrix $A \in \mathbb{R}^{m \times n}$, a pseudoinverse of A is defined as a matrix $A^\dagger \in \mathbb{R}^{n \times m}$ if it satisfies all the following:

- $AA^\dagger A = A$
- $A^\dagger AA^\dagger = A^\dagger$
- $(AA^\dagger)^\top = AA^\dagger$
- $(A^\dagger A)^\top = A^\dagger A$

If A has full column rank, $A^\dagger = (A^\top A)^{-1} A^\top$. Thus:

- $A^\dagger A = I_n$
- $AA^\dagger \neq I_m$ (not necessarily equal)

But . . .

Pseudoinverse

Definition: for any matrix $A \in \mathbb{R}^{m \times n}$, a pseudoinverse of A is defined as a matrix $A^\dagger \in \mathbb{R}^{n \times m}$ if it satisfies all the following:

- $AA^\dagger A = A$
- $A^\dagger A A^\dagger = A^\dagger$
- $(AA^\dagger)^\top = AA^\dagger$
- $(A^\dagger A)^\top = A^\dagger A$

If A has full column rank, $A^\dagger = (A^\top A)^{-1} A^\top$. Thus:

- $A^\dagger A = I_n$
- $AA^\dagger \neq I_m$ (not necessarily equal)

But ...

Claim: If $y \in \text{range}(A)$, then $AA^\dagger y = y$.

Outline

Full rank matrices

Pseudoinverse

Gradient descent for least squares problem

Convexity

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}$ being convex if the domain of f (denoted as $\mathbf{dom}(f)$) is a convex set and $\forall x, y \in \mathbf{dom}(f)$ and $\theta \in [0, 1]$, $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$.

Convexity

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}$ being convex if the domain of f (denoted as $\mathbf{dom}(f)$) is a convex set and $\forall x, y \in \mathbf{dom}(f)$ and $\theta \in [0, 1]$, $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$.

Equivalent definitions:

- (First-order Convexity Condition) Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable in $\mathbf{dom}(f)$. Then f is convex if and only if $\mathbf{dom}(f)$ is convex and $\forall x, y \in \mathbf{dom}(f)$, $f(y) \geq f(x) + \nabla f(x)^T (y - x)$.
- (Second-order Convexity Condition) Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable in $\mathbf{dom}(f)$. Then f is convex if and only if $\mathbf{dom}(f)$ is convex, $\forall x \in \mathbf{dom}(f)$, $\nabla^2 f \succeq 0$ (positive semi-definite).

Convergence rate of smooth functions

A function f is smooth if and only if $\forall x, y \in \mathbf{dom}(f)$, $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} ||x - y||^2$.

Convergence rate of smooth functions

A function f is smooth if and only if $\forall x, y \in \textbf{dom}(f)$, $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}||x-y||^2$.

Theorem: Under the following conditions:

1. $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable with $\textbf{dom}(f) = \mathbb{R}^n$
2. f is smooth with parameter $\beta > 0$
3. Optimal value $p^* = \inf_x f(x)$ is finite and is attained at x^*

If we perform gradient descent updates $x^{(k+1)} = x - t\nabla f(x^{(k)})$ on f with a constant step size t that satisfies $0 < t \leq \frac{1}{\beta}$, the number of steps taken to achieve $f(x^{(k)}) - p^* \leq \epsilon$ is $O(\frac{1}{\epsilon})$.

Convergence on least squares problem

Our problem: minimize$||y - Xw||^2$

First and second-order derivatives:
$\nabla_w ||y - Xw||^2 = 2X^\top (Xw - y)$
$\nabla_w^2 ||y - Xw||^2 = 2X^\top X$

Properties of the least squares problem:

- Convexity: $\nabla_w^2 ||y - Xw||^2 \succeq 0$
- Smoothness: $||\nabla_w ||y - Xw_1||^2 - \nabla_w ||y - Xw_2||^2|| \leq 2||X^\top X||_{\text{op}} ||w_1 - w_2||_2$

Thus if step size t satisfies $0 \leq t \leq \frac{1}{2||X^\top X||_{\text{op}}}$, we can get a convergence rate of $O(\frac{1}{k})$ with respect to the number of steps k.