Convexity and Second Derivatives, Gradient Descent and Acceleration

구재현

March 9, 2023

Semi-definiteness of a matrix and Courant-Fischer Theorem

Semi-definiteness of a matrix

Definition (Positive semi-definiteness)

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해

- $x^T Ax > 0$ 이 모든 $x \in \mathbb{R}^n \setminus \{0\}$ 에 대해 성립하면 **positive definite.**
- $x^T Ax \geq 0$ 이 모든 $x \in \mathbb{R}^n$ 에 대해 성립하면 **positive semi-definite**.
- $A, -A$ 가 모두 positive semi-definite하지 않으면 A 는 **indefinite**.

Semi-definiteness of a matrix

Definition (Positive semi-definiteness)

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해

- $x^T Ax > 0$ 이 모든 $x \in \mathbb{R}^n \setminus \{0\}$ 에 대해 성립하면 **positive definite.**
- $x^T Ax \geq 0$ 이 모든 $x \in \mathbb{R}^n$ 에 대해 성립하면 **positive semi-definite**.
- $A, -A$ 가 모두 positive semi-definite하지 않으면 A 는 **indefinite**.

Semi-definiteness라는 개념은 대칭 행렬의 고윳값과 연관이 있다.

Theorem

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해

- *모든 eigenvalue 가 0 초과 $\iff$ A 는 positive definite*
- *모든 eigenvalue 가 0 이상 $\iff$ A 는 positive semi-definite*

Semi-definiteness of a matrix

Theorem

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해

- *모든 eigenvalue 가 0 초과 $\iff$ A 는 positive definite*
- *모든 eigenvalue 가 0 이상 $\iff$ A 는 positive semi-definite*

이 사실의 증명은 자명하지 않으나, 잠시 따라가 본다.

Semi-definiteness of a matrix

Theorem

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해

- *모든 eigenvalue 가 0 초과 $\iff$ A 는 positive definite*
- *모든 eigenvalue 가 0 이상 $\iff$ A 는 positive semi-definite*

이 사실의 증명은 자명하지 않으나, 잠시 따라가 본다.

Definition (Rayleigh quotient)

벡터 $x \neq 0$ 의 행렬 A 에 대한 *Rayleigh quotient* 는 $\frac{x^T A x}{x^T x}$ 이다.

벡터가 A 의 eigenvector일 경우, *Rayleigh quotient* 는 대응되는 Eigenvalue가 된다.

다시 말해 $Ax = \lambda x$ 일 때: $\frac{x^T A x}{x^T x} = \frac{x^T \lambda x}{x^T x} = \lambda$.

Courant-Fischer Theorem

Theorem (Spectral Theorem)

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해 행렬 $V \in \mathbb{R}^{n \times n}$ 그리고 대각 행렬 $\Lambda \in \mathbb{R}^{n \times n}$ 가 존재하여

- $A = V \Lambda V^T$
- $V^T V = I$ *이며,* v_i *는* A *의* i *번째 eigenvector 이다.*
- $\Lambda_{i,i}$ *는* A*의* i *번째 eigenvalue이다.*

Spectral Theorem은 잘 알려져 있으니 증명을 생략한다.

(증명도 잘 알려져 있나?)

Courant-Fischer Theorem

Theorem (Courant-Fischer Theorem)

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ *의 eigenvalue가* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ *일 경우,*

$\lambda_k = \max_{S \subseteq \mathbb{R}^n, \dim(S)=k} \min_{x \in S - \{0\}} \frac{x^T A x}{x^T x} =$

$\min_{T \subseteq \mathbb{R}^n, \dim(T)=n-k+1} \max_{x \in S - \{0\}} \frac{x^T A x}{x^T x}$

Courant-Fischer Theorem

Theorem (Courant-Fischer Theorem)

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ *의 eigenvalue가* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ *일 경우,*

$\lambda_k = \max_{S \subseteq \mathbb{R}^n, \dim(S)=k} \min_{x \in S - \{0\}} \frac{x^T A x}{x^T x} =$

$\min_{T \subseteq \mathbb{R}^n, \dim(T)=n-k+1} \max_{x \in S - \{0\}} \frac{x^T A x}{x^T x}$

Spectral Theorem에서 Courant-Fischer Theorem을 유도할 수 있다.

Courant-Fischer Theorem

앞 등식만 증명한다 (뒤는 똑같다).

v_i와 $\lambda_i = \Lambda_{i,i}$ 는 Orthonormal basis이기 때문에 $x = \sum_i v_i^T x v_i$ 라고 쓸 수 있다. $c_i = v_i^T x$ 라 두면

$$x^T A x = x^T (\sum_i \lambda_i c_i v_i) = \sum_{i,j} c_i c_j \lambda_j v_i^T v_j = \sum_i c_i^2 \lambda_i$$

Courant-Fischer Theorem

앞 등식만 증명한다 (뒤는 똑같다).

v_i와 $\lambda_i = \Lambda_{i,i}$ 는 Orthonormal basis이기 때문에 $x = \sum_i v_i^T x v_i$ 라고 쓸 수 있다. $c_i = v_i^T x$ 라 두면

$x^T Ax = x^T(\sum_i \lambda_i c_i v_i) = \sum_{i,j} c_i c_j \lambda_j v_i^T v_j = \sum_i c_i^2 \lambda_i$

$S = span\{v_1, \ldots, v_k\}$ 에서 $\min_{x \in S} \frac{x^T Ax}{x^T x} \geq \lambda_k$ 임을 증명한다.

위 Lemma에 의해 모든 $x \in S$ 에 대해서 $\frac{x^T Ax}{x^T x} = \frac{\sum_i \lambda_i c_i^2}{\sum_i c_i^2} \geq \frac{\sum_i \lambda_k c_i^2}{\sum_i c_i^2} = \lambda_k$

고로 $\lambda_k \leq \max_{S \subseteq \mathbb{R}^n, \dim(S)=k} \min_{x \in S - \{0\}} \frac{x^T Ax}{x^T x}$ 이다.

Courant-Fischer Theorem

이제 반대 방향을 증명한다.

다시 말해, 모든 $dim(S) = k$ 인 subspace S 에 대해 $\min_{x \in S} \frac{x^T A x}{x^T x} \leq \lambda_k$ 임을 보여야 한다.

Courant-Fischer Theorem

이제 반대 방향을 증명한다.

다시 말해, 모든 $dim(S) = k$ 인 subspace S 에 대해 $\min_{x \in S} \frac{x^T A x}{x^T x} \leq \lambda_k$ 임을 보여야 한다.

$T = span\{v_k, \ldots, v_n\}$ 이라 하자. $dim(T) = n - k + 1$ 이니 $dim(S \cap T) \geq 1$ 이고 고로 $S \cap T$ 에는 0 이 아닌 원소가 존재한다.

따라서 $\min_{x \in S} \frac{x^T A x}{x^T x} \leq \min_{x \in S \cap T} \frac{x^T A x}{x^T x} \leq \max_{x \in T} \frac{x^T A x}{x^T x}$

임의의 $x \in T$ 는 $x = \sum_{i=k}^{n} c_i v_i$ 로 표현된다.

$$\frac{x^T A x}{x^T x} = \frac{\sum_i \lambda_i c_i^2}{\sum_i c_i^2} \leq \frac{\sum_i \lambda_k c_i^2}{\sum_i c_i^2} = \lambda_k$$

Courant-Fischer Theorem implies first Theorem

Theorem (Courant-Fischer Theorem)

대칭 행렬 $A \in \mathbb{R}^{n \times n}$ *의 eigenvalue가* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ *일 경우,*
$\lambda_k = \max_{S \subseteq \mathbb{R}^n, \dim(S)=k} \min_{x \in S-\{0\}} \frac{x^T A x}{x^T x} =$
$\min_{T \subseteq \mathbb{R}^n, \dim(T)=n-k+1} \max_{x \in S-\{0\}} \frac{x^T A x}{x^T x}$

Proof of the first theorem

*Rayleigh quotient*를 최대화하는 x 는 λ_1 에 대응되는 eigenvector v_1 이다. 이는 Courant-Fischer Theorem에 $k = 1$ 을 넣고 마지막 등식을 관찰하면 된다. 비슷하게, $k = n$ 을 넣으면, 최소화하는 x 역시 eigenvector v_n 이다. $x^T x > 0$ 이니, 모든 eigenvalue 가 0 초과 $\iff$ A 는 positive definite.

Second Derivatives and Convexity

Second Derivative

Definition (Hessian Matrix)

함수 $f : S \to \mathbb{R}$ 에 대해 $x \in S$ 의 **Hessian matrix** $H_f(x)$ ($\Delta^2 f(x)$ 라고도 씀) 은 다음과 같이 정의된다:

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x(1)^2} & \frac{\partial^2 f(x)}{\partial x(1)\partial x(2)} & \cdots & \frac{\partial^2 f(x)}{\partial x(1)\partial x(n)} \\ \frac{\partial^2 f(x)}{\partial x(2)\partial x(1)} & \frac{\partial^2 f(x)}{\partial x(2)^2} & \cdots & \frac{\partial^2 f(x)}{\partial x(2)\partial x(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x(n)\partial x(1)} & \frac{\partial^2 f(x)}{\partial x(n)\partial x(2)} & \cdots & \frac{\partial^2 f(x)}{\partial x(n)^2} \end{bmatrix}$$

고등학교 시간에 배운 볼록 함수의 정의는 이계도함수 $f''(x)$ 가 음이 아님을 뜻한다. 이 정의를 다변수에 대해서 확장하면서 함수의 *볼록함* 그리고 *최적성* 에 대해서 논하는 것이 목표이다.

Second Derivative

Definition (Twice differentiable)

f 가 $x \in S$ 에서 두 번 미분 가능하다는 것을 이 글에서는 다음과 같이 정의한다. 어떠한 $\Delta f(X) \in \mathbb{R}^n$ 과 $H_f(x) \in \mathbb{R}^{n \times n}$ 이 존재하여 다음 등식이 성립한다면, f 는 S 에서 두 번 미분 가능하다:

$\lim_{\delta \to 0} \frac{||f(x+\delta) - f(x) - (\Delta f(x)^T \delta + \frac{1}{2}\delta^T H_f(x)\delta)||_2}{||\delta||_2^2} = 0$

f 가 $S \subseteq \mathbb{R}^n$ 에서 *두 번 연속적으로 미분가능* 함은 두 번 미분 가능하며 도함수와 Hessian이 S 에서 연속임을 뜻한다.

Theorem (Taylor's Theorem)

$f : S \to \mathbb{R}$ 이 $[x, y]$ 에서 두 번 연속적으로 미분 가능하다면 어떠한 $z \in [x, y]$ 에 대해

$f(y) = f(x) + \Delta f(x)^T (y - x) + \frac{1}{2}(y - x)^T H_f(z)(y - x)$

Second Derivative

다항함수에서 고정점은 도함수가 0이고 이계도함수가 0 이상이다.
다차원에서도 같은 이야기를 할 수 있다.

Theorem. x 가 $f : S \to \mathbb{R}$ 의 local extremum 이라면 x 는 고정점이다: $\Delta f(x) = 0$ 이다.

Theorem. $f : S \to \mathbb{R}$ 이 $x \in S$ 에서 두 번 미분가능하다고 하자. x 가 local minimum 이라면 $H_f(x)$ 는 positive semi-definite하다.

Second Derivative

다항함수에서 고정점은 도함수가 0이고 이계도함수가 0 이상이다.
다차원에서도 같은 이야기를 할 수 있다.

Theorem. x 가 $f : S \to \mathbb{R}$ 의 local extremum 이라면 x 는 고정점이다: $\Delta f(x) = 0$ 이다.

Theorem. $f : S \to \mathbb{R}$ 이 $x \in S$ 에서 두 번 미분가능하다고 하자. x 가 local minimum 이라면 $H_f(x)$ 는 positive semi-definite하다.

위의 역도 *대충* 성립한다.

Theorem. $f : S \to \mathbb{R}$ 이 $x \in S$ 에서 두 번 미분가능하다고 하고 x 가 고정점이라고 하자. $H_f(x)$ 가 positive definite하다면 x 가 local minimum 이다.

Convexity

Theorem (Convexity)

Convex set $S \subseteq \mathbb{R}^n$ 위의 함수 $f : S \to \mathbb{R}$이 임의의 두 점 $x_1, x_2 \in S$ 과 실수 $\theta \in (0, 1)$ 에 대해서 다음을 만족한다면 S 가 **strictly convex** *하다고 한다: $f(\theta x_1 + (1-\theta)x_2) < \theta f(x_1) + (1-\theta) f(x_2)$. 등호가 허용될 경우 S 는* **convex** *하다.*

Lemma

Open, convex set $S \subseteq \mathbb{R}^n$ 위의 미분 가능한 함수 $f : S \to \mathbb{R}$ 가 주어질 때, $x, y \in S$ 에 대해 $f(y) \geq f(x) + \Delta f(x)^T (y - x)$ 가 항상 성립함과 f 가 convex임이 동치이다.

Convexity

Theorem (Convexity and Semi-definiteness)

Open, convex set $S \subseteq \mathbb{R}^n$ 위의 두번 미분 가능한 함수 $f : S \to \mathbb{R}$ 에 대해 다음이 성립한다:

- *$H_f(x)$ 가 모든 $x \in S$ 에 대해 positive semi-definite 할 경우 f 는 S 위에서 convex 하다.*
- *$H_f(x)$ 가 모든 $x \in S$ 에 대해 semi-definite 할 경우 f 는 S 위에서 strictly convex 하다.*
- *f 가 convex할 경우 $H_f(x)$ 는 모든 $x \in S$ 에 대해 positive semi-definite 하다.*

이 챕터에서 나온 내용의 증명이 궁금하다면, 2023년 5월 소멤 과제를 참고하면 좋다.

Gradient Descent

Naive Gradient Descent

Gradient Descent의 개념은 대부분 알고 있을 것이다. 여기서는 Gradient Descent의 이론적 분석에 대해서 간단히 짚고 넘어간다.

Naive Gradient Descent

Gradient Descent의 개념은 대부분 알고 있을 것이다. 여기서는 Gradient Descent의 이론적 분석에 대해서 간단히 짚고 넘어간다.

모르시는 분들에 대해 설명드리자면~

Convex function f 에 대해서 $min_{x \in \mathbb{R}^n} f(x)$ 를 찾는 문제를 생각해 보자.

Convex function에서는 Local minimum이 Global minimum이니, 현재 있는 x 가 최소가 아니라면 $\Delta f(x) \neq 0$ 이고, 고로 $-\Delta f(x)$ 방향으로 움직이면서 무조건 해를 개선시킬 수 있다.

이 방향으로 *적당히 짧은 거리를 반복적으로 이동* 하는 것이 Gradient Descent 방법이다.

중요한 것은 **어떻게 이동하면 언제 수렴하는지** 의 문제이다.

β-gradient Lipschitz

β-gradient Lipschitz (β-smooth) 개념은 도함수가 얼마나 *평탄한지* 를 나타내는 indicator이다.

Definition (β-gradient Lipschitz)

Convex, open set $S \subseteq \mathbb{R}^n$ 에서 미분 가능한 함수 $f : S \to \mathbb{R}$ 가 주어질 때, 모든 $x, y \in S$에 대해 다음이 성립하면 f 를 β-gradient Lipschitz (혹은 β-smooth) 라고 한다: $||\Delta f(x) - \Delta f(y)||_2 \leq \beta ||x - y||_2$

β-gradient Lipschitz

β-gradient Lipschitz (β-smooth) 개념은 도함수가 얼마나 *평탄한지* 를 나타내는 indicator이다.

Definition (β-gradient Lipschitz)

Convex, open set $S \subseteq \mathbb{R}^n$ 에서 미분 가능한 함수 $f : S \to \mathbb{R}$ 가 주어질 때, 모든 $x, y \in S$에 대해 다음이 성립하면 f 를 β-gradient Lipschitz (혹은 β-smooth) 라고 한다: $||\Delta f(x) - \Delta f(y)||_2 \leq \beta ||x - y||_2$

Lemma

- $f : S \to \mathbb{R}$ 이 두 번 연속적으로 미분가능하다고 하자. f 가 β-smooth 함은 모든 $x \in S$ 에 대해 $||H_f(x)|| \leq \beta$ 임과 동치이다.
- $f : S \to \mathbb{R}$ 이 β-smooth function 이라고 하자. 모든 x, y 에 대해 $f(y) \leq f(x) + \Delta f(x)^T (y - x) + \frac{\beta}{2} ||x - y||_2^2$ 이다.

Naive Gradient Descent

앞 Lemma를 사용하여 적당한 이동 거리를 찾아보자. 현재 점이 x 라면, 적당한 d 를 찾아서 $f(x+d)$ 의 상한을 충분히 작게 하고 싶은 것이 목표이다.

$\Delta f(x)^T d + \frac{\beta}{2}||d||_2^2$ 를 미분하면, $d = -\frac{1}{\beta}\Delta f(x)$ 일 때 위 값이 $-\frac{||\Delta f(x)^2||_2^2}{2\beta}$ 로 최소화됨을 알 수 있다. Gradient Descent의 점화식을 $x_{i+1} = x_i - \frac{1}{\beta}\Delta f(x_i)$ 로 두자.

Naive Gradient Descent

앞 Lemma를 사용하여 적당한 이동 거리를 찾아보자. 현재 점이 x 라면, 적당한 d 를 찾아서 $f(x+d)$ 의 상한을 충분히 작게 하고 싶은 것이 목표이다.

$\Delta f(x)^T d + \frac{\beta}{2}||d||_2^2$ 를 미분하면, $d = -\frac{1}{\beta}\Delta f(x)$ 일 때 위 값이 $-\frac{||\Delta f(x)^2||_2^2}{2\beta}$ 로 최소화됨을 알 수 있다. Gradient Descent의 점화식을 $x_{i+1} = x_i - \frac{1}{\beta}\Delta f(x_i)$ 로 두자.

x^* 를 최적해라고 할 때, $(f(x_i) - f(x^*))(i+1)$ 가 비감소함수 라는 사실을 증명할 수 있다. 따라서:

Theorem (Gradient Descent is linearly fast)

$f : \mathbb{R}^n \to \mathbb{R}$ 을 β-smooth convex function 이라고 할 때, x_0을 초기 시작점, x^* 을 f의 global minimizer라고 할 때, $x_{i+1} = x_i - \frac{1}{\beta}\Delta f(x_i)$ 를 반복하는 Gradient Descent 알고리즘은 i 번째 반복에서 다음을 만족한다: $f(x_i) - f(x^*) \leq \frac{2\beta||x_0 - x^*||_2^2}{i+1}$

Accelerated Gradient Descent

Accelerated Gradient Descent는 Gradient Descent보다 더 빠르게 수렴하는 변형으로, Nesterov가 1983년 제안하였다.

Gradient Descent의 Gap은 $O(\frac{1}{T})$ 의 속도로 수렴하는 반면, Accelerated Gradient Descent의 Gap은 $O(\frac{1}{T^2})$ 의 속도로 수렴한다는 차이가 있다.

간단히 말하자면, Gradient Descent의 방향을 조정할 때 *최악의 경우* 볼 수 있는 손해를 줄이는 조정을 하면서 진행한다. 이를 위해 점화식이 조금 더 복잡하다 (두 보조 수열을 잡는다).

Accelerated Gradient Descent

Accelerated Gradient Descent는 Gradient Descent보다 더 빠르게 수렴하는 변형으로, Nesterov가 1983년 제안하였다.

Gradient Descent의 Gap은 $O(\frac{1}{T})$ 의 속도로 수렴하는 반면, Accelerated Gradient Descent의 Gap은 $O(\frac{1}{T^2})$ 의 속도로 수렴한다는 차이가 있다.

간단히 말하자면, Gradient Descent의 방향을 조정할 때 *최악의 경우* 볼 수 있는 손해를 줄이는 조정을 하면서 진행한다. 이를 위해 점화식이 조금 더 복잡하다 (두 보조 수열을 잡는다).

이후 책에서 이 알고리즘에 대해 다시 다루지 않는다. 별로 중요하지 않은 거 아닐까? 자세한 내용과, 앞 Gradient Descent에 대해 빼놓은 증명들이 궁금하다면 2023년 5월 소멤 과제를 참고하면 된다.

(거기도 딱히 자세히 나와 있지는 않다 ㅋ)