경사하강법

강요셉

March 25, 2022

이 글은 LaTeX 에서 작성됨

Contents

1 Introduction

머신러닝, 수치해석 등 여러 컴퓨팅 분야에서는 어떤 목적함수(objective function)의 함수값을 최적화(optimization)하는 방법에 대해 연구한다. 이런 최적화 문제에서 대표적인 해결 방법으로 **경사하강법 (Gradient Descent)** 이 제시된다.

이 글에서는 경사하강법의 원리와 타당성을 다변수 해석학을 통해 이해/증명하고, 그에 따른 또다른 방법론을 제시한다.

1.1 List of special symbols

∇f	기울기 벡터	$\mathbb{R}$	실수 집합
$\mathbf{0}$	영벡터	$\int_{\mathbf{X}}$	선적분
$\lvert\cdot\rvert$	절댓값 노름	$\lVert\cdot\rVert$	벡터 노름
$\lVert\cdot\rVert_{op}$	행렬 Operator 노름	$H(f)$	헤세 행렬

2 Method

n-공간에서 정의된 다변수 일급 함수 $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}, \mathbf{X} \mapsto \mathcal{L}(\mathbf{X})$ 의 값을 최소로 만드는 최소점 $\mathbf{X}$를 찾는 문제를 생각해 보자. 임계점 정리에 의하면, $\mathcal{L}$의 최소점은 다음을 만족한다.

$$\nabla\mathcal{L}(\mathbf{X}) = \mathbf{0}$$

경사하강법은 변수의 초기값 $\mathbf{X}_0$ 에서 시작해 기울기 벡터에 비례해 변수를 이동시키며 나타나는 수열 $\{\mathbf{X}_n\}$ 을 통해 최소점을 구한다.

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \alpha(\nabla\mathcal{L}(\mathbf{X}_n))$$

이제 이러한 수열을 다음을 만족시키는 일급 곡선으로 생각해 본다.

$$\mathbf{X}(0) = \mathbf{X}_0 \qquad \mathbf{X}(n\alpha) = \mathbf{X}_n$$

만약 α 가 충분히 작다면, $\frac{\mathbf{X}(t+\alpha)-\mathbf{X}(t)}{\alpha} = \mathbf{X}'(t) + \frac{\mathbf{o}(\alpha)}{\alpha} = -\nabla\mathcal{L}(\mathbf{X}(t))$ 이므로 곡선 $\mathbf{X}$ 는 아래의 미분방정식을 만족시키는 일급 적분곡선 $\tilde{\mathbf{X}}$ 에 수렴하게 된다.

$$\tilde{\mathbf{X}}(0) = \mathbf{X}_0 \qquad \tilde{\mathbf{X}}' = -\nabla\mathcal{L}(\tilde{\mathbf{X}})$$

선택한 α 가 아주 작다는 가정 하에, 이 곡선을 관찰함으로써 수열 $\{\mathbf{X}_n\}$ 의 행동을 관찰할 수 있다. 앞으로는 이 곡선을 관찰함으로써 특수한 조건을 만족하는 함수에서는 이 곡선이 함수의 극소점으로 수렴함을 보인다.

3 Observation

Theorem 3.1. $\mathcal{L}(\tilde{\mathbf{X}}(t))$ 는 단조 감소한다.

Proof. $[t_1, t_2]$ 에서 정의된 곡선 $\mathbf{Y}$ 를 $\mathbf{Y} = \tilde{\mathbf{X}}$ 이라 할 때, 선적분의 기본정리에 의해

$$\mathcal{L}(\tilde{\mathbf{X}}(t_2) - \mathcal{L}(\tilde{\mathbf{X}}(t_1)) = \int_{\mathbf{Y}} \nabla\mathcal{L} \cdot d\mathbf{s} = \int_{t_1}^{t_2} -\|\nabla\mathcal{L}\|^2 dt \leq 0$$

따라서 $\mathcal{L}(\tilde{\mathbf{X}}(t_1)) \geq \mathcal{L}(\tilde{\mathbf{X}}(t_2))$ 이다. □

이 성질은 곡선 $\tilde{\mathbf{X}}$ 를 따라갈 때 함숫값이 줄어든다는 것을 보장해 주지만, 실제로 이 곡선이 함수의 극소점을 찾는것에 도움을 주는가에 대한 질문에 답하진 못한다. 다음 명제는 일단 곡선이 수렴한다면 임계점으로 수렴하는가에 답하기 위해 떠올린 명제이다.

(거짓) 어떤 $\mathbf{P}$ 에 대해서 $\lim_{t\to\infty} \tilde{\mathbf{X}}(t) = \mathbf{P}$ 이면 $\lim_{t\to\infty} \tilde{\mathbf{X}}'(t) = \mathbf{0}$ 이다.
반례) 편의를 위해 $n = 1$ 로 둔다. $\tilde{\mathbf{X}}(t) = \frac{\sin t^2}{t}$ 일때, $\lim_{t\to\infty} \tilde{\mathbf{X}}(t) = 0$ 이지만 $\lim_{t\to\infty} \tilde{\mathbf{X}}'(t) = \lim_{t\to\infty}(2\cos x^2 - \frac{\sin x^2}{x^2})$ 은 수렴하지 않는다.

이는 곡선 $\tilde{\mathbf{X}}$ 가 일반적으로는 임계점으로 수렴하지 않음을 나타내지만, 만약 $\tilde{\mathbf{X}}$ 가 한가지 조건을 만족한다면 참으로 만들 수 있다.

(여기서부터는 $\tilde{\mathbf{X}}''$ 에 대해 다룰것이므로 함수 $\mathcal{L}$ 이 이급 함수임을 추가로 가정한다.)

Theorem 3.2. $\lim_{t\to\infty}\tilde{\mathbf{X}}(t)=\mathbf{P}$ 이고 $\|\tilde{\mathbf{X}}''\|\le M$ 이면 $\lim_{t\to\infty}\tilde{\mathbf{X}}'(t)=\mathbf{0}$ 이다.

이 정리를 증명하기 앞서 한가지 도움정리를 다룬다.

Lemma 3.3. 이급 실함수 f 에 대해 $|f|\le M_0, |f''|\le M_2$ 이면 $|f'|\le 2\sqrt{M_0M_2}$ 이다.

Proof. 테일러 정리에 의해 임의의 $h>0, t$ 에 대해 다음을 만족한다.

$$\xi\in(t,t+2h)\qquad f(t+2h)=f(t)+2hf'(t)+2h^2f''(\xi)$$
$$f'(t)=\frac{f(t+2h)-f(t)}{2h}-hf''(\xi)$$

따라서 다음의 부등식을 얻는다.

$$|f(t)|\le\frac{M_0}{h}+hM_2$$

이제 $h=\sqrt{\frac{M_2}{M_0}}$ 을 대입하면 원하는 결과를 얻는다.

$$|f|\le 2\sqrt{M_0M_2}$$

□

이 정리는 벡터 함수, 즉 곡선에서도 성립한다.

Lemma 3.4. 이급 벡터 함수 $\mathbf{f}$에 대해 $\|\mathbf{f}\|\le M_0, \|\mathbf{f}''\|\le M_2$ 이면 $\|\mathbf{f}'\|\le 2\sqrt{M_0M_2}$ 이다.

Proof. 임의의 실수 t_0 에 대해서 $\mathbf{n}=\frac{1}{\|\mathbf{f}'(t_0)\|}\mathbf{f}'(t_0)$ 이라 하자. $\mathbf{f}'(t_0)\cdot\mathbf{n}=\|\mathbf{f}'(t_0)\|$ 이고, $\|\mathbf{n}\|$=1 이므로 코시-슈바르츠 부등식에 의해 $|\mathbf{f}\cdot\mathbf{n}|\le M_0, |\mathbf{f}''\cdot\mathbf{n}|\le M_2$ 이다. 이제 $\mathbf{f}\cdot\mathbf{n}$ 은 실함수 이므로 앞의 정리를 적용하면 다음이 성립한다.

$$\|\mathbf{f}'(t_0)\|\le 2\sqrt{M_0M_2}$$

t_0 이 임의의 실수였으므로 원하는 결론을 얻는다. □

이제 Theorem 3.2. 를 증명한다.

Proof. $\lim_{t\to\infty} \tilde{\mathbf{X}}(t) = 0$ 이므로 임의의 양수 ϵ 에 대해서 (N, ∞) 에서 $\|\tilde{\mathbf{X}}\| \le \frac{\epsilon^2}{4M}$ 인 N이 존재한다. 앞의 정리에 의해서 (N, ∞) 에서는 $\|\tilde{\mathbf{X}}'\| \le 2\sqrt{(\frac{\epsilon^2}{4M})M} = \epsilon$ 이므로 $\lim_{t\to\infty} \tilde{\mathbf{X}}'(t) = 0$ 이다. □

Remark. 미분방정식 $\tilde{\mathbf{X}}' = -\nabla\mathcal{L}(\tilde{\mathbf{X}})$ 을 t에 대해 미분하면

$$\tilde{\mathbf{X}}'' = -\begin{bmatrix} \frac{\partial^2\mathcal{L}}{\partial x_1^2} & \cdots & \frac{\partial^2\mathcal{L}}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2\mathcal{L}}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2\mathcal{L}}{\partial x_n^2} \end{bmatrix} (\tilde{\mathbf{X}}') = H(\mathcal{L})(\nabla\mathcal{L})$$

($H(\mathcal{L})$ 은 **헤세 행렬 (Hessian matrix)**)
따라서 만약 함수 $\mathcal{L}$ 이 정의역의 모든 점에서 $\|H(\mathcal{L})(\nabla\mathcal{L})\| \le M$ 이라면 조건 $\|\tilde{\mathbf{X}}''\| \le M$ 을 만족함을 알 수 있고, 특히 $\|H(\mathcal{L})(\nabla\mathcal{L})\| \le \|H(\mathcal{L})\|_{op}\|\nabla\mathcal{L}\|$ 이므로 $\|H(\mathcal{L})\|_{op}, \|\nabla\mathcal{L}\|$ 이 유계(bounded)라면 조건 $\|\tilde{\mathbf{X}}''\| \le M$ 을 만족한다.

이는 "…를 만족하는 함수는 …를 만족한다."라는 서술을 가능하게 해준다.
(기존의 조건인 $\|\tilde{\mathbf{X}}''\| \le M$ 은 곡선의 초기값에 따라 변할수도 있는 조건이므로 이러한 서술에 적절하지 않다.)

앞의 정리는 곡선이 수렴하는 점은 임계점임을 말해주지만, 그것이 극소점임을 말해주진 않는다. 하지만 '좋은' 조건이 주어진다면 그러한 임계점이 극소점임을 알 수 있다.

Definition 3.1. 함수 f 가 정의역의 임의의 점 A, B 와 $0 \le \lambda \le 1$ 에 대해 $f(\lambda A + (1-\lambda)B) \le \lambda f(A) + (1-\lambda)f(B)$ 를 만족시키면 f 를 **볼록(convex)함수** 라고 한다.

Lemma 3.5. 다변수 볼록함수 f 와 벡터 $\mathbf{P}, \mathbf{v}$ 에 대해 $g(t) = f(\mathbf{P} + t\mathbf{v})$ 는 볼록함수이다.

Proof. 임의의 $a < b, 0 \le \lambda \le 1$ 에 대해서

$$\begin{aligned} g(\lambda a + (1-\lambda)b) &= f(\mathbf{P} + \lambda a\mathbf{v} + (1-\lambda)b\mathbf{v}) \\ &= f(\lambda(\mathbf{P} + a\mathbf{v}) + (1-\lambda)(\mathbf{P} + b\mathbf{v})) \\ &\le \lambda f(\mathbf{P} + a\mathbf{v}) + (1-\lambda)f(\mathbf{P} + b\mathbf{v}) \\ &= \lambda g(a) + (1-\lambda)g(b) \end{aligned}$$

따라서 g 는 볼록함수이다. □

Lemma 3.6. 일변수 일급 함수 f 가 볼록함수이면 f' 은 단조 증가한다.

Proof. 임의의 $a < b$ 와 $c \in (a, b)$ 에 대해서 $\lambda = \frac{c-b}{a-b}$ 로 두면 다음의 부등식을 얻는다.

$$f(c) \le \frac{c-b}{a-b}f(a) + \frac{a-c}{a-b}f(b)$$
$$\frac{f(c)-f(a)}{c-a} \le \frac{f(b)-f(a)}{b-a} \le \frac{f(b)-f(c)}{b-c}$$

이제 c 를 a, b 로 보내면 f 가 미분 가능이므로 각각 $f'(a), f'(b)$ 로 수렴하고, 위의 부등식에 의해 증명이 완료된다.

$$f'(a) \le \frac{f(b)-f(a)}{b-a} \le f'(b)$$

□

Theorem 3.7. 다변수 일급 함수 f가 볼록함수이면 임의의 벡터 $\mathbf{P}, \mathbf{v}$에 대해 $\nabla f(\mathbf{P}) \cdot \mathbf{v} = D_{\mathbf{v}} f(\mathbf{P}) \leq f(\mathbf{P}+\mathbf{v}) - f(\mathbf{P})$ 이다.

Proof. Lemma 3.5. 에 의해 $g(t) = f(\mathbf{P} + t\mathbf{v})$ 가 볼록함수 임을 알고, 위의 부등식에 의해 다음이 성립하므로 증명이 완료된다.

$$D_{\mathbf{v}} f(P) = g'(0) \leq \frac{g(1) - g(0)}{1 - 0} = f(\mathbf{P} + \mathbf{v}) - f(\mathbf{P})$$

□

Corollary 3.7.1. 다변수 일급 함수 f가 볼록함수이고 $\nabla f(\mathbf{P}) = \mathbf{0}$ 이라면 f는 $\mathbf{P}$에서 극소이다.

Proof. 임의의 벡터 $\mathbf{v}$에 대해 $g(t) = f(\mathbf{P} + t\mathbf{v})$ 라고 하자.

$$\nabla f(\mathbf{P}) \cdot \mathbf{v} = 0 \leq f(\mathbf{P} + \mathbf{v}) - f(\mathbf{P})$$
$$f(\mathbf{P}) \leq f(\mathbf{P} + \mathbf{v})$$

이므로 f는 $\mathbf{P}$에서 극소이다. □

Theorem 3.2. Corollary 3.7.1. 그리고 Remark.에서 다룬 내용들을 종합하면 다음과 같은 정리를 얻는다.

Theorem 3.8. $\|H(\mathcal{L})\|_{op}, \|\nabla \mathcal{L}\|$ 이 유계인 이급 볼록함수 $\mathcal{L}$에서 $\tilde{\mathbf{X}}$가 수렴한다면, 그 점은 극소점이다.

Theorem 3.9.

4 Conclusion

Theorem 4.1. $\|H(\mathcal{L})\|_{op}, \|\nabla \mathcal{L}\|$ 이 유계인 이급 볼록함수 $\mathcal{L}$에서 극소점이 존재한다면 $\tilde{\mathbf{X}}$는 $\mathcal{L}$의극소점으로 수렴한다.

Proof. □

Corollary 4.1.1. 컴팩트(compact) 집합 U 위에서 정의된 이급 볼록함수 $\mathcal{L}$에서 극소점이 존재한다면 $\tilde{\mathbf{X}}$는 $\mathcal{L}$의 극소점으로 수렴한다.

Proof. 함수 $\mathcal{L}$이 이급 함수 이였으므로 $\mathcal{L}, \frac{\partial \mathcal{L}}{\partial x_i}, \frac{\partial^2 \mathcal{L}}{\partial x_i \partial x_j}$ 은 연속이고, 연속함수에서 컴팩트 집합의 상은 컴팩트 집합이므로 유계이다. 따라서 $\|H(\mathcal{L})\|_{op}, \|\nabla \mathcal{L}\|$ 또한 유계이므로 Theorem 4.1.에 의해 $\tilde{\mathbf{X}}$는 $\mathcal{L}$의 극소점으로 수렴한다. $\square$