Advanced Graph Algorithms and Optimization: Convexity and Second Derivatives, Gradient Descent and Acceleration

2021년 발표된 Minimum Cost Flow가 Almost-Linear Time에 풀린다는 결과는 현대적인 Convex Optimization 기법들이 전통적인 알고리즘 문제를 해결하는 아주 강력한 도구를 제공함을 보여주는 상징적인 순간이라고 할 수 있다. 요즘은 기계학습의 인기가 워낙 엄청나기 때문에 Convex Optimization은 대개 기계 학습의 관점에서 다뤄지지만, Convex Optimization은 현대 전산학 전반에 있어서 중요도가 아주 높으며, 그래프 알고리즘적인 관점에서 Convex Optimization을 다뤘을 때 얻어갈 수 있는 것이 아주 많다.

최근 관심 있는 학생들과 함께 ETH Zurich의 Advanced Graph Algorithms and Optimization 강의를 스터디하고 있는데, 이 강의의 Chapter 3인 "Convexity and Second Derivatives, Gradient Descent and Acceleration" 를 정리해서 소개하려고 한다. 기초적인 내용에 속해서 어렵지 않게 읽을 수 있고, 굳이 알고리즘 분야에 관심이 없더라도 기계 학습 등 다양한 분야에 있어서 관련있는 주제일 것이라고 생각한다.

Semi-definiteness of a matrix and Courant-Fischer Theorem

대칭 행렬 A 가 있을 때 다음과 같은 것을 정의한다.

Definition 1. 대칭 행렬 $A \in \mathbb{R}^{n \times n}$ 에 대해

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

위 Semi-definiteness라는 개념은 대칭 행렬의 고윳값 (eigenvalue) 과 연관이 있다.

Theorem 1. 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

Theorem 1은 나도 기계 학습 시간에 증명 없이 배웠던 기억이 있다. 이 사실의 증명은 전혀 자명하지 않은데, 이 과정을 따라가는 것이 유익할 것 같아서 이 글에서 다뤄보려고 한다.

벡터 $x \neq 0$ 의 행렬 A 에 대한 *Rayleigh quotient* 는 다음과 같다:

$$\frac{x^T A x}{x^T x}$$

만약 이 벡터가 A 의 eigenvector일 경우, *Rayleigh quotient* 는 대응되는 Eigenvalue가 된다. $Ax = \lambda x$ 일 때:

$$\frac{x^T A x}{x^T x} = \frac{x^T \lambda x}{x^T x} = \lambda$$

이를 증명하기 위해 두 개의 Theorem을 소개한다. Theorem 2는 잘 알려져 있기 때문에 딱히 증명하지는 않는다 (증명이 쉬운 거랑은 별개인 거 같다. 증명은 여기 잘 나와 있다.)

Theorem 2 (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 A의 i 번째 eigenvalue이다.

Theorem 3 (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 (2 -> 3) 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

$x^T A x$

$= x^T A (\sum_i c_i v_i)$

$= x^T (\sum_i \lambda_i c_i v_i)$

$= \sum_{i,j} c_i c_j \lambda_j v_i^T v_j$

$= \sum_i c_i^2 \lambda_i$

첫 번째로, $S = span\{v_1, \ldots, v_k\}$ 에서 $\min_{x \in S} \frac{x^T A x}{x^T x} \geq \lambda_k$ 임을 증명한다. 위 Lemma에 의해 모든 $x \in S$ 에 대해서

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

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

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

■

Proof (3 -> 1). *Rayleigh quotient* 를 최대화하는 x 는 λ_1 에 대응되는 eigenvector v_1 이다. 이는 $k = 1$ 을 넣고 마지막 등식을 관찰하면 된다. 비슷하게, $k = n$ 을 넣으면, 최소화하는 x 역시 eigenvector v_n 이다. $x^T x > 0$ 이니, Theorem 3은 Theorem 1을 함의한다.

Second Derivatives

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

Definition 2. 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:

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

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:

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

달리 말해, 모든 δ 에 대해 다음이 성립한다:

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

여기서 o 의 정의는 $\lim_{\delta \to 0} \frac{o(||\delta||_2^2)}{||\delta||_2^2} = 0$ 이다.

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

Theorem 4 (Taylor's Theorem) $f : S \to \mathbb{R}$ 이 $[x, y]$ 에서 두 번 연속적으로 미분 가능하다면 어떠한 $z \in [x, y]$ 에 대해 $f(y) = f(x) + \Delta f(x)^T (y - x) + \frac{1}{2}(y - x)^T H_f(z)(y - x)$

다음 사실을 쉽게 관찰할 수 있다:

Proposition 5. x 가 $f : S \to \mathbb{R}$ 의 local extremum 이라면 x 는 고정점이다: $\Delta f(x) = 0$ 이다.

Hessian에 대해서도 위와 같은 논증을 할 수 있다. 이 때 위에서 배운 정의들이 도움이 된다.

Theorem 6. $f : S \to \mathbb{R}$ 이 $x \in S$ 에서 두 번 미분가능하다고 하자. x 가 local minimum 이라면 $H_f(x)$ 는 positive semi-definite하다.

Proof. Proposition 5에 의해 $\Delta(x) = 0$ 이 성립하니, 다음과 같은 전개를 할 수 있다:

$f(x + \lambda d) = f(x) + \lambda^2 \frac{1}{2} d^T H_f(x) d + o(\lambda^2 ||d||_2^2)$

x 가 local minimum이니 다음이 성립한다:

$0 \leq \lim_{\lambda \to 0^+} \frac{f(x+\lambda d) - f(x)}{\lambda^2} = \frac{1}{2} d^T H_f(x) d$

고로 모든 d 에 대해 $d^T H_f(x) d$ 는 0 이상이고 $H_f(x)$ 는 positive semi-definite하다.

Theorem 6은 Local minimum이 고정점이며 고로 positive semi-definite Hessian을 가진다는 내용이다. 고정점에 대해서 이 사실의 역도 *거의* 참이지만, 약간의 차이는 있다.

Theorem 7. $f : S \to \mathbb{R}$ 이 $x \in S$ 에서 두 번 미분가능하다고 하고 x 가 고정점이라고 하자. $H_f(x)$ 가 positive definite하다면 x 가 local minimum 이다.

Proof. $\Delta(x) = 0$ 이 성립하니, 다음과 같은 전개를 할 수 있다:

$f(x+\delta) = f(x) + \frac{1}{2}\delta^T H_f(x)\delta + o(||\delta||_2^2)$

$H_f(x)$ 가 positive definite이니, 최소 Eigenvalue λ_{min} 이 0 초과이다. 고로 $\delta^T H_f(x)\delta \geq \lambda_{min}||\delta||_2^2$ 이다. 만약 $||\delta||_2^2$ 가 충분히 작다면, o 항의 크기를 원하는 만큼 줄일 수 있으니 (1/4 이하라고 하자), $f(x+\delta) - f(x) \geq \frac{1}{4}\lambda_{min}||\delta_2||^2 > 0$ 이 된다. 고로 x 는 local minimum 이다.

단락을 들어오면서 고등학교 때 배운 이계도함수와 볼록성의 관계를 잠시 언급했다. 이제 이 정의의 n 차원 버전을 다룬다.

Definition 3. 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:

$f(\theta x_1 + (1-\theta)x_2) < \theta f(x_1) + (1-\theta)f(x_2)$

Definition 4. 등호가 허용될 경우 S 는 **convex** 하다.

Lemma 8. 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.

Proof. $\to$: 실수 θ 에 대해 $z = \theta x + (1-\theta)y$ 라 하자. 다음이 성립한다

- $f(y) \geq f(z) + \Delta f(z)^T(y-z)$
- $f(x) \geq f(z) + \Delta f(z)^T(x-z)$

연립하면 $\theta f(y) + (1-\theta)f(x) \geq f(z) + \Delta f(z)^T 0 = f(\theta y + (1-\theta)x)$ 이다.

$\leftarrow$: Convexity 가정에 의해 임의의 θ 에 대해 다음 사실이 성립한다:

$f(\theta y + (1-\theta)x) \leq (1-\theta)f(x) + \theta f(y)$

양변에서 $f(x)$ 를 빼고 θ 로 나누면

$\frac{f(x+\theta(y-x)) - f(x)}{\theta} \leq (f(y) - f(x))$

$\theta \to 0^+$ 로 보내면, $\Delta f(x)^T(y-x) \leq f(y) - f(x)$ 가 된다.

Theorem 9. 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.

Proof.

- Theorem 4에 의해서, 어떠한 $z \in [x, y]$ 에 대해 $f(y) = f(x) + \Delta f(x)^T(y-x) + \frac{1}{2}(y-x)^T H_f(z)(y-x)$ 가 성립한다. $H_f(z)$ 가 positive semi-definite하기 때문에 $f(y) \geq f(x) + \Delta f(x)^T(y-x)$ 이다. Lemma 8에 의해
- strict 경우 역시 동일하다.
- 어떤 작은 $\epsilon > 0$ 과 임의의 $d \in \mathbb{R}^n$ 에 대해, $x + \epsilon d \in S$ 이다. f 의 Taylor expansion에 의해
  $f(x+\epsilon d) = f(x) + \epsilon \Delta f(x)^T d + \frac{\epsilon^2}{2} d^T H_f(x) d + o(\epsilon^2 ||d||_2^2)$ 이다. Lemma 8에 의해
  $f(x+\epsilon d) \geq f(x) + \epsilon \Delta f(x)^T d$ 이니, 임의의 $d \in \mathbb{R}^n$ 에 대해 $\frac{\epsilon^2}{2} d^T H_f(x) d + o(\epsilon^2 ||d||_2^2) \geq 0$ 이다. 양변을 ϵ^2 로 나누고 $\epsilon \to 0^+$ 로 보내면 위 식이 증명된다.

Gradient Descent

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 이동* 하는 것을 반복하는 게 잘 알려진 Gradient descent의 방법이다. 이 단락에서는 *얼마나 이동해야 하는지*, 그리고 이를 반복하면 *얼마나 빠른* 알고리즘을 얻을 수 있는지, 이 두 가지의 기술적 분석에 집중한다.

Definition 5. 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-smooth function에 대해서 다음 두 가지 사실이 알려져 있다. 두 사실은 증명 없이 사용한다.

Proposition 10.

- $f : S \to \mathbb{R}$ 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.

Proposition 10의 명제를 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. 고로 $x_1 = x_0 - \frac{1}{\beta} \Delta f(x)$ 로 두면 $f(x_1) \leq f(x_0) - \frac{||\Delta f(x)^2||_2^2}{2\beta}$ 이다.

임의의 x_0 에서 시작해서 위와 같이 움직이는 것을 반복하는 알고리즘을 수행했을 때의 결과를 분석해 보자. 이를 위해서 $gap_i = f(x_i) - f(x^*)$ 라는 값을 정의하자. 이 때 x^* 는 f 의 global minimizer이다. global minimizer가 꼭 unique할 필요는 없다. 단지 연결된 영역에 있을 뿐이다.

다음 사실을 알 수 있다.

- $gap_{i+1} - gap_i \leq -\frac{||\Delta f(x)^2||_2^2}{2\beta}$
- $f(x^*) \geq f(x_i) + \Delta f(x_i)^T (x^* - x_i)$ (Lemma 8)
- $gap_i \leq \Delta f(x_i)^T (x_i - x^*)$
- $gap_i \leq ||\Delta f(x_i)||_2 ||x_i - x^*||_2$ (Cauchy-Schwarz)

다음 명제가 성립하는데, 이에 대한 증명은 생략한다.

Proposition 11. Gradient Descent를 위에서 설명한 것과 같이 돌릴 때 모든 i 에 대해 $||x_i - x^*||_2 \leq ||x_0 - x^*||_2$ 이다.

이 사실을 사용하면 위 첫번째 식과 네번째 식을 결합할 수 있다:

$gap_{i+1} - gap_i \leq -\frac{1}{2\beta} (\frac{gap_i}{||x_0 - x^*||_2})^2$

Theorem 12. $f : \mathbb{R}^n \to \mathbb{R}$ 을 β-smooth convex function 이라고 할 때, x_0을 초기 시작 점, 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:

- $f(x_i) - f(x^*) \leq \frac{2\beta ||x_0 - x^*||_2^2}{i+1}$

Proof. 수학적 귀납법을 사용하자.

- $i = 0$ 일 때는 Proposition 10에 의해서 자명하다.
- $i > 0$ 일 때, $gap_{i+1} \leq \frac{2\beta||x_0 - x^*||_2^2}{i+1} - \frac{1}{2\beta}(\frac{2\beta|x_0 - x^*||_2}{(i+1)})^2 = 2\beta||x_0 - x^*||_2^2(\frac{i}{(i+1)^2}) \leq \frac{2\beta||x_0 - x^*||_2^2}{i+2}$

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. 이후 책에서 이 알고리즘에 대해 다시 다루지 않기 때문에, 이 알고리즘에 대해서는 high-level idea와 결론 정도만 짧게 설명한다.

Gradient Descent는 초기 시작점 x_0 을 잡아 $x_{i+1} = x_i - \frac{1}{\beta}\Delta f(x_i)$ 라는 점화식을 계산한다고 생각할 수 있다. Accelerated Gradient Descent에서 사용하는 점화식은 추가적으로 두 수열 $y_0, y_1, \ldots, y_k \in \mathbb{R}^n, v_0, v_1, \ldots, v_k \in \mathbb{R}^n$ 을 사용한다. 이 때

- y_i 는 $f(y_i)$ 를 최소화하는 것을 목표로 구성되며, 최종적으로 얻게 되는 y_k 는 우리의 알고리즘의 output이 된다. (정확하게 그렇지는 않으나 대략 그렇다). 달리 말해, y_i는 $f(x^*)$ 의 upper bound이다.
- v_i 는 $f(x^*)$ 의 lower bound 를 잡기 위해서 구성된다.

알고리즘은 i 를 늘려가면서 $f(x^*)$ 가 속할 수 있는 구간을 줄여나갈 것이다. 매 iteration에서 y_{i+1}, v_{i+1} 을 고를 때의 목표는, lower bound를 늘려가면서 upper bound 역시 줄여가는 것이 된다.

먼저, y_i 수열을 잡는 것은 기본 버전과 상당히 유사하다:

$y_i = x_i - \frac{1}{\beta}\Delta f(x_i)$

위에서 본 것과 같이 이 경우 $f(y_i) \leq f(x_i) - \frac{||\Delta f(x)^2||_2^2}{2\beta}$ 이 된다. 이제 Upper bound를 $U_i = f(y_i)$ 라고 하자.

Lower bound를 잡는 과정에서 잠시 직관에 대한 설명을 곁들인다. Gradient Descent의 수렴을 보이기 위해서 우리가 결합할 두 명제는

- Gradient가 클 경우 최적해에 빠르게 접근한다
- Gradient가 작을 경우 이미 최적해에 가깝다 (gap 이라는 값은 현재 Gradient에 비례한다)

후자의 명제를 표현하는 식은 전 단락에서 유도했던 이 식이다:

$f(x^*) \geq f(x_i) - ||\Delta f(x_i)||_2||x_i - x^*||_2$

여기서 한 가지 관찰할 점은, 이 식이 꼭 i 가 증가할수록 좋은 lower bound를 주지는 않는다는 것이다. 오히려, 예전에 계산했던 x_j 에서 더 좋은 하한을 줄 가능성도 있다. 이러한 점을 감안하여 예전에 계산했던 x_j 의 값을 적절히 안배하는 분석을 사용해 보자.

각 스텝에 대해서 Weight $a_i > 0$ 을 주어서, 계산했던 하한들을 weighted average로 취해 보자. 이렇게 할 경우, 임의의 $v \in \mathbb{R}^n$ 에 대한 답의 하한은:

$f(v) \geq \frac{1}{A_i}\sum_{j \leq i} a_j(f(x_j) + \langle \Delta f(x_j), v - x_j \rangle)$

여기서 $\langle a, b \rangle$ 은 두 벡터 a, b 의 내적을 뜻하며, $A_i = \sum_{j \leq i} a_j$ 이다.

$\phi(v) = \frac{\beta}{2}||v - x_0||_2^2$ 라는 regularization term을 도입하자. 책에는 이 부분이 *somewhat magical trick* 이라고 나와있는데, 이 글에서는 증명을 모두 생략하기 때문에 이 term이 왜 magical한지 알 수 있는 여지가 없을 것이다. 이제 항을 이런 식으로 쓸 수 있다:

$f(x^*) \geq \frac{1}{A_i}(\phi(x^*) + \sum_{j \leq i} a_j f(x_j) + \langle a_j \Delta f(x_j), x^* - x_j \rangle) - \frac{\phi(x^*)}{A_i}$

$f(x^*) \geq \min_{v \in \mathbb{R}^n} \frac{1}{A_i}(\phi(v) + \sum_{j \leq i} a_j f(x_j) + \langle a_j \Delta f(x_j), v - x_j \rangle) - \frac{\phi(x^*)}{A_i}$

우변의 항을 L_i 라고 하고, 우변의 항을 최소화하는 벡터를 v_i 라고 하자. 최종 알고리즘에서는 $(U_i - L_i)(i+1)(i+2)$ 가 i 에 대한 비증가 수열임을 증명할 수 있다. 이는 $O(\frac{1}{T^2})$ 의 속도로 알고리즘이 수렴하는 증명이 된다.

Theorem 13. $f : \mathbb{R}^n \to \mathbb{R}$ 을 β-smooth convex function 이라고 할 때, x_0을 초기 시작 점, x^* 을 f의 global minimizer라고 할 때, 다음을 반복적으로 계산하는 Accelerated Gradient Descent 알고리즘은 i 번째 반복에서 $f(x_i) - f(x^*) \leq \frac{2\beta||x_0 - x^*||_2^2}{(i+1)(i+2)}$ 을 보장한다.

- $a_i = \frac{i+1}{2}, A_i = \frac{(i+1)(i+2)}{4}$
- $v_0 = x_0 - \frac{1}{2\beta}\Delta f(x_0)$
- $y_i = x_i - \frac{1}{\beta}\Delta f(x_i)$
- $x_{i+1} = \frac{A_i y_i + a_{i+1} v_i}{A_{i+1}}$
- $v_{i+1} = v_i - \frac{a_{i+1}}{\beta}\Delta f(x_{i+1})$