

CSE203B Convex Optimization:

Chapter 9: Unconstrained Minimization

CK Cheng

Dept. of Computer Science and Engineering
University of California, San Diego

Chapter 9 Unconstrained Minimization

- Introduction
- Taylor's Expansion & Bounds
- Descent Methods
- Newton Method
- Summary

Introduction

Problem: $\min f(x)$ where $f: R^n \rightarrow R$
is convex and twice continuously
differentiable

Theorem: For a convex function $f(x)$, the necessary and sufficient condition for a point x^* to be optimal is $\nabla f(x^*) = 0$.

Remark: keywords Taylor's expansion

Taylor's Expansion & Bounds: Scalar case

$$f(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2} (x - x_0)^T \nabla^2 f(z) (x - x_0)$$

for some z on the segment $[x, x_0]$

(1) Scalar case: $f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(z)(x - x_0)^2$

We simplify the notations $f(x) = \frac{m}{2} (x - x_0)^2 + a(x - x_0) + b$

For fixed m , a , and b , the optimal solution can be derived as:

$$\nabla f(x) = 0 \Rightarrow m(x - x_0) + a = 0 \Rightarrow x - x_0 = -\frac{a}{m}$$

Thus, we have

$$f(x) = \frac{m}{2} \frac{a^2}{m^2} + a \frac{(-a)}{m} + b = \frac{a^2}{2m} - \frac{a^2}{m} + b = \frac{-a^2}{2m} + b$$

Or $f(x) - f(x_0) = -\frac{a^2}{2m}$

a. How far from opt. x^ ? $x^* - x_0 = -\frac{a}{m}$*

b. How much difference from opt. $f(x^)$? $f(x_0) - f(x^*) = \frac{a^2}{2m}$*

Taylor's Expansion & Bounds: Example

$$f(x) = x^2 + 4x + 1$$

For the format

$$f(x) = \frac{m}{2}x^2 + ax + b, \quad m = 2, a = 4, b = 1.$$

Let $x_0 = 0$, we have the answer.

a. *How far?* $x^* - x_0 = -\frac{a}{m} = -2$

b. *How much?* $f(x_0) - f(x^*) = \frac{a^2}{2m} = 4$

Taylor's Expansion & Bounds: Bounds

(2) Vector case:

Assumption A: $\nabla^2 f(x)$ is bounded, i.e. $mI \preceq \nabla^2 f(x) \preceq MI$

Theorem A: We have the following bounds

$$\frac{1}{M} \|\nabla f(x_0)\|_2 \stackrel{\textcircled{4}}{\leq} \|x_0 - x^*\|_2 \stackrel{\textcircled{1}}{\leq} \frac{2}{m} \|\nabla f(x_0)\|_2$$
$$\frac{1}{2M} \|\nabla f(x_0)\|_2^2 \stackrel{\textcircled{3}}{\leq} f(x_0) - p^* \stackrel{\textcircled{2}}{\leq} \frac{1}{2m} \|\nabla f(x_0)\|_2^2$$

Proof: $\textcircled{1}$

$$f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2 \leq f(y)$$
$$\leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$$

(Taylor's Expansion + Assumption A)

Taylor's Expansion & Bounds: Bounds

Proof ①: $\|x - x^*\|_2 \leq \frac{2}{m} \|\nabla f(x)\|_2$

$p^* = f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{m}{2} \|x^* - x\|_2^2$ (Taylor's exp + Assumption A.)

$$\geq f(x) - \|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2} \|x^* - x\|_2^2$$

We shift $f(x)$ to the left hand side.

$$0 \geq p^* - f(x) \geq -\|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2} \|x^* - x\|_2^2$$

Shift $-\|\nabla f(x)\|_2 \|x^* - x\|_2$ to the left,

$$\|\nabla f(x)\|_2 \|x^* - x\|_2 \geq \frac{m}{2} \|x^* - x\|_2^2$$

Therefore we have

1. $\|\nabla f(x)\|_2 \geq \frac{m}{2} \|x^* - x\|_2$
2. $\|x^* - x\|_2 \leq \frac{2}{m} \|\nabla f(x)\|_2$

Taylor's Expansion & Bounds: Bounds

Proof ②: $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2$
(Taylor's exp + assumption A.)
 $\geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$ (Minimization with y)

Thus, we have

$$f(x) - f(y) \leq \frac{1}{2m} \|\nabla f(x)\|_2^2, \quad \forall y$$

Therefore

$$f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|_2^2$$

Taylor's Expansion & Bounds: Bounds

Proof ③: $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2$

(Taylor's exp + assumption A.)

$$\geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 \text{ (Minimization with } y)$$

Let $y = x - \frac{1}{M} \nabla f(x)$, we have

$$\begin{aligned} f\left(x - \frac{1}{M} \nabla f(x)\right) &\leq f(x) + \nabla f(x)^T \frac{-1}{M} \nabla f(x) + \frac{M}{2} \left\| \frac{1}{M} \nabla f(x) \right\|_2^2 \\ &= f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2 \end{aligned}$$

Shift the terms on the left and right, we have

$$\begin{aligned} \frac{1}{2M} \|\nabla f(x)\|_2^2 &\leq f(x) - f\left(x - \frac{1}{M} \nabla f(x)\right) \\ &\leq f(x) - f(x^*) \end{aligned}$$

Taylor's Expansion & Bounds: Bounds

(4) Proof: $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$

(Taylor's exp. + assumption A)

(i) Let $x = x^*$, we have $\nabla f(x^*) = 0$,
thus, we can write the above eq.

$$f(y) \leq f(x^*) + \frac{M}{2} \|y - x^*\|_2^2$$

$$\text{or } f(y) - p^* \leq \frac{M}{2} \|y - x^*\|_2^2$$

(ii) From (3), we have

$$\frac{1}{2M} \|\nabla f(x_o)\|_2^2 \leq f(x_o) - p^*$$

(iii) From (i)&(ii), we have

$$\frac{1}{2M} \|\nabla f(x_o)\|_2^2 \leq \frac{M}{2} \|x_o - x^*\|_2^2$$

Therefore, we have

$$\frac{1}{M} \|\nabla f(x_o)\|_2 \leq \|x_o - x^*\|_2$$

Taylor's Expansion & Bounds

Remark:

(1) If $\|\nabla f(x)\|_2 \leq (2m\epsilon)^{\frac{1}{2}}$

We have $\|x - x^*\|_2 \leq \frac{2}{m} (2m\epsilon)^{\frac{1}{2}}$

$$f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|_2^2}{2m} = \epsilon$$

(2) The bounds can be used to design algorithms.

prove the convergence.

(3) If $M \gg m$ (e.g. 10^{10})

Impact on the bounds become very loose

→ Efficiency of gradient descent approaches.

(4) Quadratic obj. with sparse matrix (A)

$$\frac{1}{2} x^T A x + b^T x + c$$

is a preferred formulation in terms of algorithm efficiency.

Taylor's Expansion & Bounds: Scalar case

$$f(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2} (x - x_0)^T \nabla^2 f(z) (x - x_0)$$

for some z on the segment $[x, x_0]$

(2) Vector case:

We simplify the notations $f(x) = \frac{1}{2} (x - x_0)^T H (x - x_0) + a^T (x - x_0) + b$

For fixed m , a , and b , the optimal solution can be derived as:

$$\nabla f(x) = 0 \Rightarrow H(x - x_0) + a = 0 \Rightarrow x - x_0 = -H^{-1}a$$

Thus, we have

$$f(x) = \frac{1}{2} a^T H^{-T} H H^{-1} a - a^T H^{-1} a + b = \frac{-1}{2} a^T H^{-1} a + b$$

a. How far from opt. x^ ? $x^* - x_0 =$*

b. How much difference from opt. $f(x^)$? $f(x_0) - f(x^*) =$*

II. Descent Methods

Given convex function, twice continuously differentiable $f(x)$
and an initial point $x_0 \in \text{dom } f$.

Repeat

1. Determine a descent direction Δx ($\nabla f(x)^T \Delta x < 0$)
2. Line Search, choose a step size $t > 0$.
3. Update $x = x + t\Delta x$

Until stopping criterion is met.

Line Search : $t = \arg \min_{t>0} f(x + t\Delta x)$

Backtracking line search ($\alpha \in (0, 1/2), \beta \in (0, 1)$)

Start at $t = 1$, repeat $t := \beta t$

until $f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$

Stopping criterion $\|\nabla f(x)\|_2 \leq \eta$ $\eta = (2m\epsilon)^{\frac{1}{2}}$ (Theorem A (2))

II. Descent Methods: Example

Problem: $\min f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \quad \gamma > 0$

$$x^0 = (\gamma, 1), f(x^0) = \frac{\gamma(\gamma+1)}{2}, \nabla f(x^0) = (\gamma, \gamma)$$

Thus, $x^1 = (\gamma, 1) - t(\gamma, \gamma) = (\gamma(1-t), 1-t\gamma)$

and $\nabla f(x^1) = (\gamma(1-t), \gamma(1-t\gamma))$

1. To opt $f(x^1)$ with respect to variable t ,

we have $f(x^1) = \frac{1}{2}(\gamma^2(1-t)^2 + \gamma(1-t\gamma)^2)$

$$\frac{\partial f(x^1)}{\partial t} = \gamma^2(1-t) + \gamma(1-t\gamma)\gamma = 0$$

Thus, $t = \frac{2\gamma^2}{\gamma^2 + \gamma^3} = \frac{2}{1+\gamma}$, and $x^1 = \left(\frac{\gamma(\gamma-1)}{1+\gamma}, \frac{1-\gamma}{1+\gamma}\right) = \left(\frac{10 \times 9}{11}, -\frac{9}{11}\right)$

2. We repeat the process to step k , $x^k = \left(\gamma \left(\frac{\gamma-1}{\gamma+1}\right)^k, \left(\frac{1-\gamma}{1+\gamma}\right)^k\right)$

3. Equal potential plot

$$f(x^k) = \frac{\gamma(\gamma+1)}{2} \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^0) = \left(\frac{1-m/M}{1+m/M}\right)^{2k} f(x^0)$$

II. Descent Methods: Descent for various norms

1. Problem: Min $f(x)$

2. For each iteration, we try the steepest descent in terms of a given norm.

$$\begin{aligned} \text{Min } \nabla f(x)^T \Delta x \\ \text{s.t. } \|\Delta x\| \leq 1 \end{aligned}$$

3. We show the step of

i. Quadratic norm

ii. L1 norm

II. Descent Methods: Descent for quadratic norm

1. Problem: Min $f(x)$
2. For each iteration, we try the steepest descent in terms of a given norm.

$$\begin{aligned} \text{Min}_{\Delta x} \quad & \nabla f(x)^T \Delta x \\ \text{s.t.} \quad & \|\Delta x\|_P \leq 1 \end{aligned}$$

$$\|\Delta x\|_P = (\Delta x^T P \Delta x)^{1/2}, P \in S_{++}^n$$

$$\text{Lagrangian } L(\Delta x, \lambda) = \nabla f(x)^T \Delta x + \lambda (\|\Delta x\|_P - 1), \lambda \geq 0$$

$$\text{We can derive: } \Delta x_{nsd} = -(\nabla f(x)^T P^{-1} \nabla f(x))^{-1/2} P^{-1} \nabla f(x)$$

$$\text{Or } \Delta x_{sd} = -P^{-1} \nabla f(x)$$

II. Descent Methods: Descent for quadratic norm

The coordinate change has effects on the descent direction.

Example: $\min f(x) = \frac{1}{2}x^T P x + q^T x, P \in S_{++}^n$

Affine transform: $\bar{x} = P^{1/2} x$

II. Descent Methods: Descent for L1 norm

1. Problem: Min $f(x)$

2. For each iteration, we try the steepest descent in terms of a given norm.

$$\begin{aligned} \text{Min } \nabla f(x)^T \Delta x < 0 \\ \text{s.t. } \|\Delta x\|_1 \leq 1 \end{aligned}$$

$$\text{Lagrangian } L(\Delta x, \lambda) = \nabla f(x)^T \Delta x + \lambda (\|\Delta x\|_1 - 1), \lambda \geq 0$$

$$\text{We can derive: } \Delta x_{nsd} = -\text{sign}\left(\frac{\partial f(x)}{\partial x_i}\right) e_i,$$

where i is the index for which $\|\nabla f(x)\|_\infty = |\nabla f(x)_i|$

$$\text{Or } \Delta x_{sd} = -\frac{\partial f(x)}{\partial x_i} e_i$$

Gradient descent method: Convergence analysis

$$\tilde{f}(t) \equiv f(x - t\nabla f(x)) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$
$$\tilde{f}(t_{exact}) \leq \tilde{f}\left(t = \frac{1}{M}\right) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \quad (\min_t f(x) - \nabla f(x))$$

A. $\tilde{f}(t_{exact}) - p^* \leq f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|_2^2$

B. $\frac{1}{2M}\|\nabla f(x)\|_2^2 \geq \frac{m}{M}(f(x) - p^*)$ since $\frac{\|\nabla f(x)\|_2^2}{2m} \geq f(x) - p^*$

C. From B, we have

$$f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|_2^2 \leq f(x) - p^* - \frac{m}{M}(f(x) - p^*)$$
$$= (f(x) - p^*)\left(1 - \frac{m}{M}\right)$$

D. We can conclude from A & C

$$f(x^{k+1}) - p^* \leq \left(1 - \frac{m}{M}\right)(f(x^k) - p^*) \leq \left(1 - \frac{m}{M}\right)^k (f(x^0) - p^*)$$

To achieve $f(x^*) - p^* \leq \epsilon$,

we need $\frac{\log((f(x^0) - p^*)/\epsilon)}{\log(1/c)}$ steps, where $c = 1 - \frac{m}{M} < 1$,

Gradient descent method : Convergence analysis

$$\log(1/c) = -\log(1 - m/M) \approx m/M \text{ for large } M/m$$

Remark: when $M/m > 100$

the method can be very slow.

Newton Step

Use the approximation of 2nd order Taylor's Exp.

$$f(x + v) \approx f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

We would like to derive

$$\nabla_v f(x + v) = 0 \rightarrow \nabla f(x) + \nabla^2 f(x) v = 0$$

Thus, we have $v = -\nabla^2 f(x)^{-1} \nabla f(x)$

$$\begin{aligned} f(x + v) &= f(x) + (-1) \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) + \\ &\quad \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \\ &= f(x) - \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \end{aligned}$$

Input $x \in \text{dom } f$, $\epsilon > 0$

Repeat: 1. $\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$, $\lambda^2(x) = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$

2. Quit if $\lambda^2 / 2 \leq \epsilon$

3. Line Search t

4. $x := x + t \Delta x_{nt}$

Newton Method : Convergence analysis

Assumptions: $S = \{x \in \text{dom } f \mid f(x) \leq f(x_0)\}$

f strongly convex on S with constant m , s.t. $\nabla^2 f(x) \geq mI, \forall x \in S$

$\nabla^2 f$ is Lipschitz continuous on S with constant L , i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$$

Outlines: $\exists \eta \in (0, m^2/L)$, two cases.

1. Damped Newton Phase: ($t < 1$)

$$\|\nabla f(x)\|_2 \geq \eta \text{ then } f(x^{k+1}) - f(x^k) \leq -\alpha\beta\eta^2 m/M^2$$

2. Pure Newton Phase (Quadratically Convergent Stage): ($t = 1$)

$\|\nabla f(x)\|_2 < \eta$ then

$$\begin{aligned} \frac{L}{m^2} \|\nabla f(x^{k+1})\|_2 &\leq \left(\frac{L}{2m^2} \|\nabla f(x^k)\|_2 \right)^2 \\ &\leq \left(\frac{L}{2m^2} \|\nabla f(x^l)\|_2 \right)^{2^{k+1-l}} \leq \left(\frac{1}{2} \right)^{2^{k+1-l}} \quad k+1 \geq l \end{aligned}$$

Newton Method: Affine Invariant

Problem: $\min f(x)$

Theorem: Newton's step is invariant to affine transform.

Proof: Let $x = Ty, T \in R^{nn}, f(x) = f(Ty) = \bar{f}(y)$

For the x coordinate system, we have.

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

Therefore, we have the invariant results

$$x + \Delta x_{nt} = T(y + \Delta y_{nt}).$$

For the y coordinate system, we have.

1. $\nabla_y \bar{f}(y) = T^T \nabla_x f(Ty),$

$$\nabla_y^2 \bar{f}(y) = T^T \nabla^2 f(Ty) T$$

2. The Newton step at $y,$

$$\Delta y_{nt}$$

$$= -\nabla_y^2 \bar{f}(y)^{-1} \nabla_y \bar{f}(y)$$

$$= -(T^T \nabla^2 f(x) T)^{-1} (T^T \nabla f(x))$$

$$= -T^{-1} \nabla^2 f(x)^{-1} \nabla f(x)$$

$$= T^{-1} \Delta x_{nt}$$

Summary

1. Gradient Descent Method: (**minimization solution**)
 1. Vector operations per iteration
 2. Linear convergence rate
2. Newton's Method: (**equality solution**)
 1. Matrix operations per iteration
 2. Quadratic convergence rate (near the solution)
3. Gradient Descent Method Variations:
 1. Conjugate gradient method
 2. Nesterov gradient descent method
 3. Quasi-Newton method