

1. Gradient method

- first-order methods
- some properties of differentiable convex functions
- convergence theory for gradient method
- limits on convergence rate of first-order methods

Classical gradient method

to minimize a convex differentiable function f : choose $x^{(0)}$ and repeat

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, \dots$$

step size rules

- fixed: t_k constant
- backtracking line search
- exact line search: minimize $f(x - t \nabla f(x))$ over t

advantages of gradient method

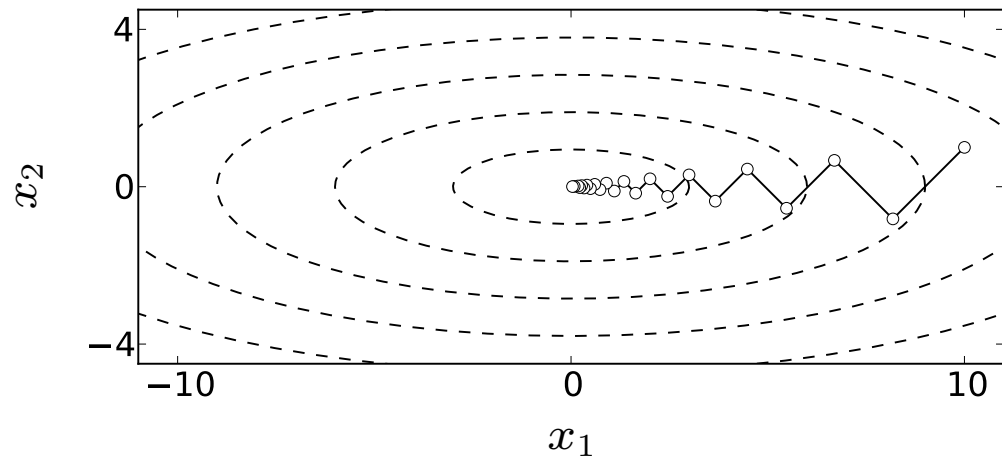
- every iteration is inexpensive
- does not require second derivatives

Quadratic example

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

with exact line search, $x^{(0)} = (\gamma, 1)$

$$\frac{\|x^{(k)} - x^*\|_2}{\|x^{(0)} - x^*\|_2} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k$$

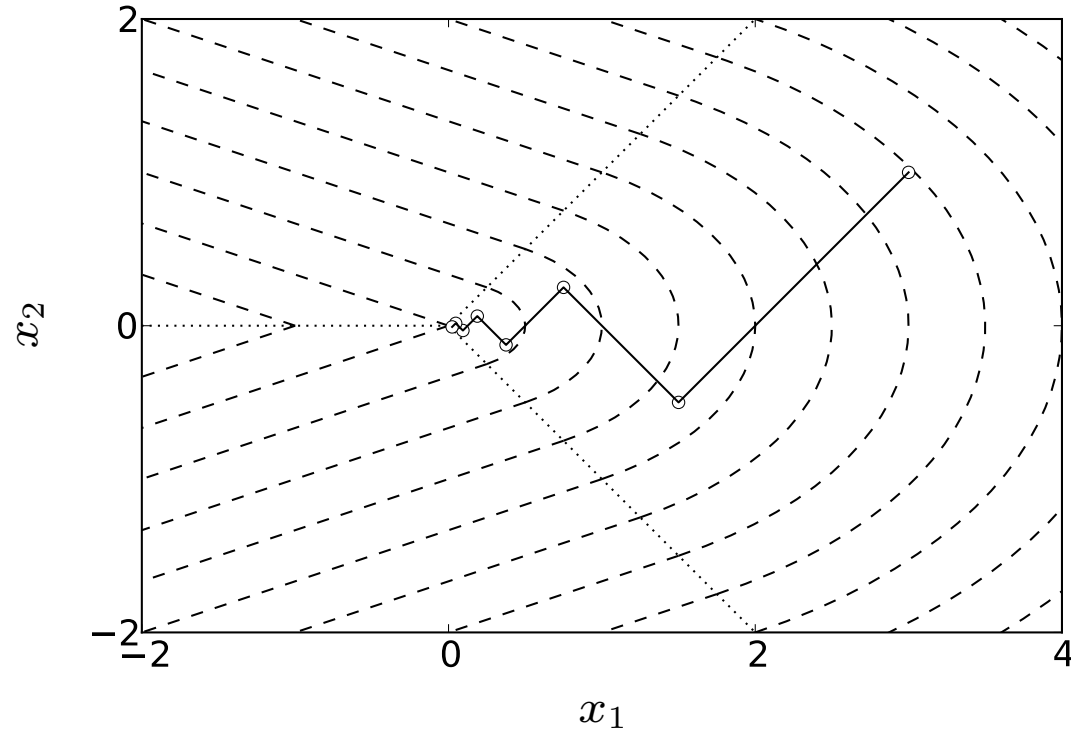


gradient method is often slow; very dependent on scaling

Nondifferentiable example

$$f(x) = \sqrt{x_1^2 + \gamma x_2^2} \quad (|x_2| \leq x_1), \quad f(x) = \frac{x_1 + \gamma|x_2|}{\sqrt{1 + \gamma}} \quad (|x_2| > x_1)$$

with exact line search, $x^{(0)} = (\gamma, 1)$, converges to non-optimal point



gradient method does not handle nondifferentiable problems

First-order methods

address one or both disadvantages of the gradient method

methods for nondifferentiable or constrained problems

- subgradient method (lectures 2 & 3)
- proximal gradient method (lecture 4)
- smoothing methods (lecture 5)

methods with improved convergence

- variable metric methods (lecture 6)
- conjugate gradient method (lecture 7)
- accelerated proximal gradient method (lecture 8)

Convex function

f is convex if $\mathbf{dom} f$ is a convex set and Jensen's inequality holds:

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \forall x, y \in \mathbf{dom} f, \theta \in [0, 1]$$

first-order condition

for (continuously) differentiable f , Jensen's inequality can be replaced with

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \mathbf{dom} f$$

second-order condition

for twice differentiable f , Jensen's inequality can be replaced with

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbf{dom} f$$

Strictly convex function

f is strictly convex if $\mathbf{dom} f$ is a convex set and

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y) \quad \forall x, y \in \mathbf{dom} f, x \neq y, \theta \in (0, 1)$$

first-order condition (for differentiable f): $\mathbf{dom} f$ is convex and

$$f(y) > f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \mathbf{dom} f, x \neq y$$

hence minimizer of f is unique (if a minimizer exists)

second-order condition

note that $\nabla^2 f(x) \succ 0$ is not necessary for strict convexity (cf., $f(x) = x^4$)

Strongly convex function

f is strongly convex with parameter $m > 0$ if

$$f(x) - \frac{m}{2}\|x\|_2^2 \text{ is convex}$$

Jensen's inequality

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) - \frac{m}{2}\theta(1 - \theta)\|x - y\|_2^2$$

first-order condition

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|_2^2 \quad \forall x, y \in \text{dom } f$$

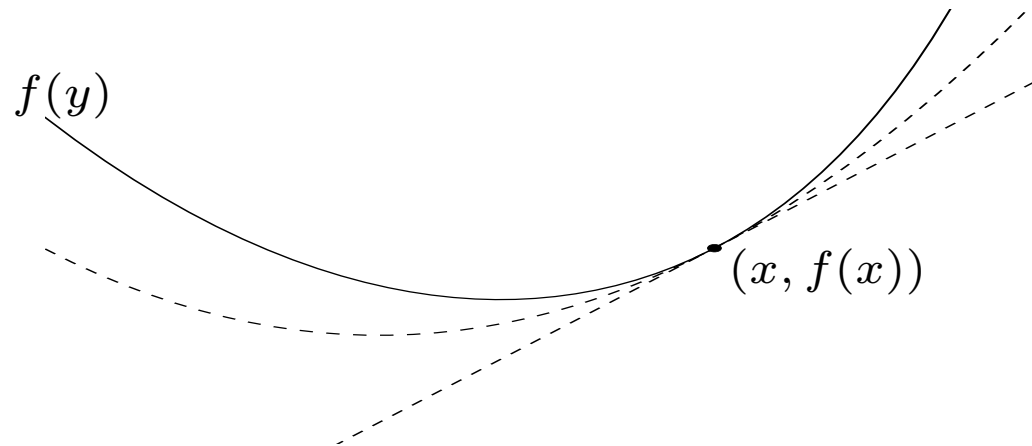
second-order condition

$$\nabla^2 f(x) \succeq mI \quad \forall x \in \text{dom } f$$

Quadratic lower bound

(from 1st order condition) if f is strongly convex with parameter m , then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2 \quad \forall x, y \in \text{dom } f$$



if $\text{dom } f = \mathbf{R}^n$, this implies f has a unique minimizer x^* and

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

Monotonicity of gradient

differentiable f is convex if and only if $\mathbf{dom} f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \forall x, y \in \mathbf{dom} f$$

i.e., $\nabla f : \mathbf{R}^n \rightarrow \mathbf{R}^n$ is a *monotone* mapping

- gradient of strictly convex function is *strictly monotone*:

$$(\nabla f(x) - \nabla f(y))^T (x - y) > 0 \quad \forall x, y \in \mathbf{dom} f, x \neq y$$

- gradient of strongly convex function is *strongly monotone* or *coercive*:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq m \|x - y\|_2^2 \quad \forall x, y \in \mathbf{dom} f$$

proof of monotonicity property

- if f is differentiable and convex, then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad f(x) \geq f(y) + \nabla f(y)^T (x - y)$$

combining the inequalities gives

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0$$

- if ∇f is monotone, then $g'(t) \geq g'(0)$ for $t \geq 0$ where

$$g(t) = f(x + t(y - x)), \quad g'(t) = \nabla f(x + t(y - x))^T (y - x)$$

hence,

$$f(y) = g(1) = g(0) + \int_0^1 g'(t) dt \geq g(0) + g'(0) = f(x) + \nabla f(x)^T (y - x)$$

Functions with Lipschitz continuous gradients

gradient of f is Lipschitz continuous with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \mathbf{dom} f$$

- from the Cauchy-Schwarz inequality, this implies

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

i.e., $Lx - \nabla f(x)$ is monotone; in other words

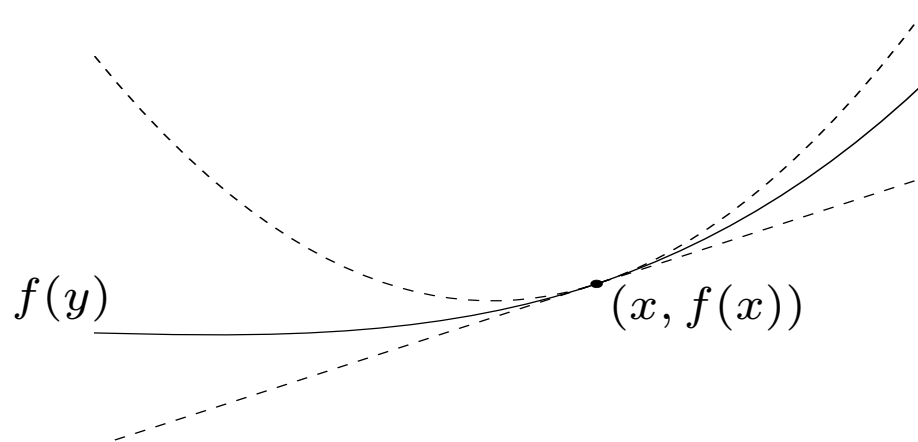
$$\frac{L}{2}\|x\|_2^2 - f(x) \quad \text{is convex}$$

- for twice differentiable f , equivalent to $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbf{dom} f$

Quadratic upper bound

if ∇f is Lipschitz-continuous with parameter L , then

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbf{dom} f$$



if $\mathbf{dom} f = \mathbf{R}^n$ and f has a minimizer x^* , implies

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

Co-coercivity of gradient

if f is convex and ∇f is Lipschitz continuous with parameter L , then

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y \in \mathbf{dom} f$$

property is known as *co-coercivity* of ∇f (with parameter $1/L$)

proof: the convex functions

$$g(z) = f(z) - \nabla f(x)^T z, \quad h(z) = f(z) - \nabla f(y)^T z$$

have Lipschitz-continuous gradients, and minimizers $z = x$ and $z = y$, resp.

therefore, from page 1-13,

$$g(x) \leq g(y) - \frac{1}{2L} \|\nabla g(y)\|_2^2, \quad h(y) \leq h(x) - \frac{1}{2L} \|\nabla h(x)\|_2^2$$

combining these two inequalities shows co-coercivity

extension: if in addition f is strongly convex with parameter m , then

$$g(x) = f(x) - \frac{m}{2}\|x\|_2^2$$

is convex and ∇g is Lipschitz continuous with parameter $L - m$

the co-coercivity property for g gives

$$\begin{aligned} & (\nabla f(x) - \nabla f(y))^T (x - y) \\ & \geq \frac{mL}{m + L}\|x - y\|_2^2 + \frac{1}{m + L}\|\nabla f(x) - \nabla f(y)\|_2^2 \end{aligned}$$

for all $x, y \in \text{dom } f$

Analysis of gradient method

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, \dots$$

with fixed step size or backtracking line search

assumptions

1. f is convex and differentiable with $\text{dom } f = \mathbf{R}^n$
2. $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$
3. optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*

Analysis for constant step size

from quadratic upper bound (page 1-13) with $y = x - t\nabla f(x)$:

$$f(x - t\nabla f(x)) \leq f(x) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x)\|_2^2$$

therefore, if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 1/L$,

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

take $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t_i = t$, and add the bounds for $i = 1, \dots, k$:

$$\begin{aligned} \sum_{i=1}^k (f(x^{(i)}) - f^*) &\leq \frac{1}{2t} \sum_{i=1}^k \left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right) \\ &= \frac{1}{2t} \left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2 \right) \\ &\leq \frac{1}{2t} \|x^{(0)} - x^*\|_2^2 \end{aligned}$$

since $f(x^{(i)})$ is non-increasing,

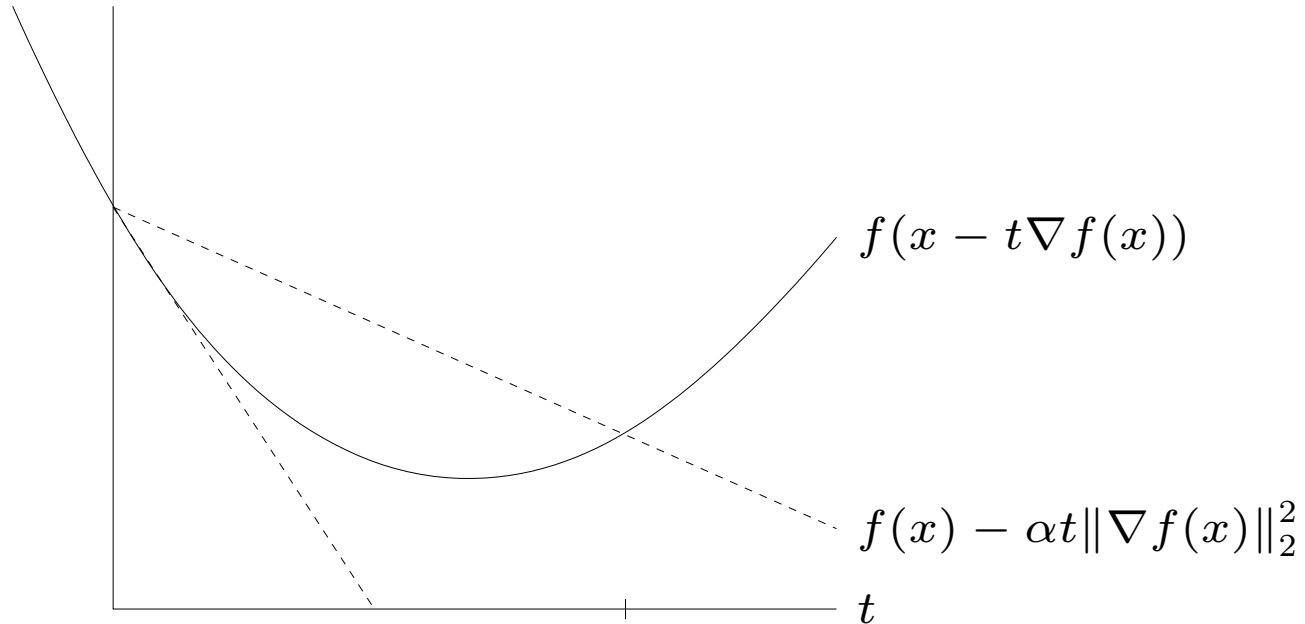
$$f(x^{(k)}) - f^* \leq \frac{1}{k} \sum_{i=1}^k (f(x^{(i)}) - f^*) \leq \frac{1}{2kt} \|x^{(0)} - x^*\|_2^2$$

conclusion: #iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(1/\epsilon)$

Backtracking line search

initialize t_k at $\hat{t} > 0$ (for example, $\hat{t} = 1$); take $t_k := \beta t_k$ until

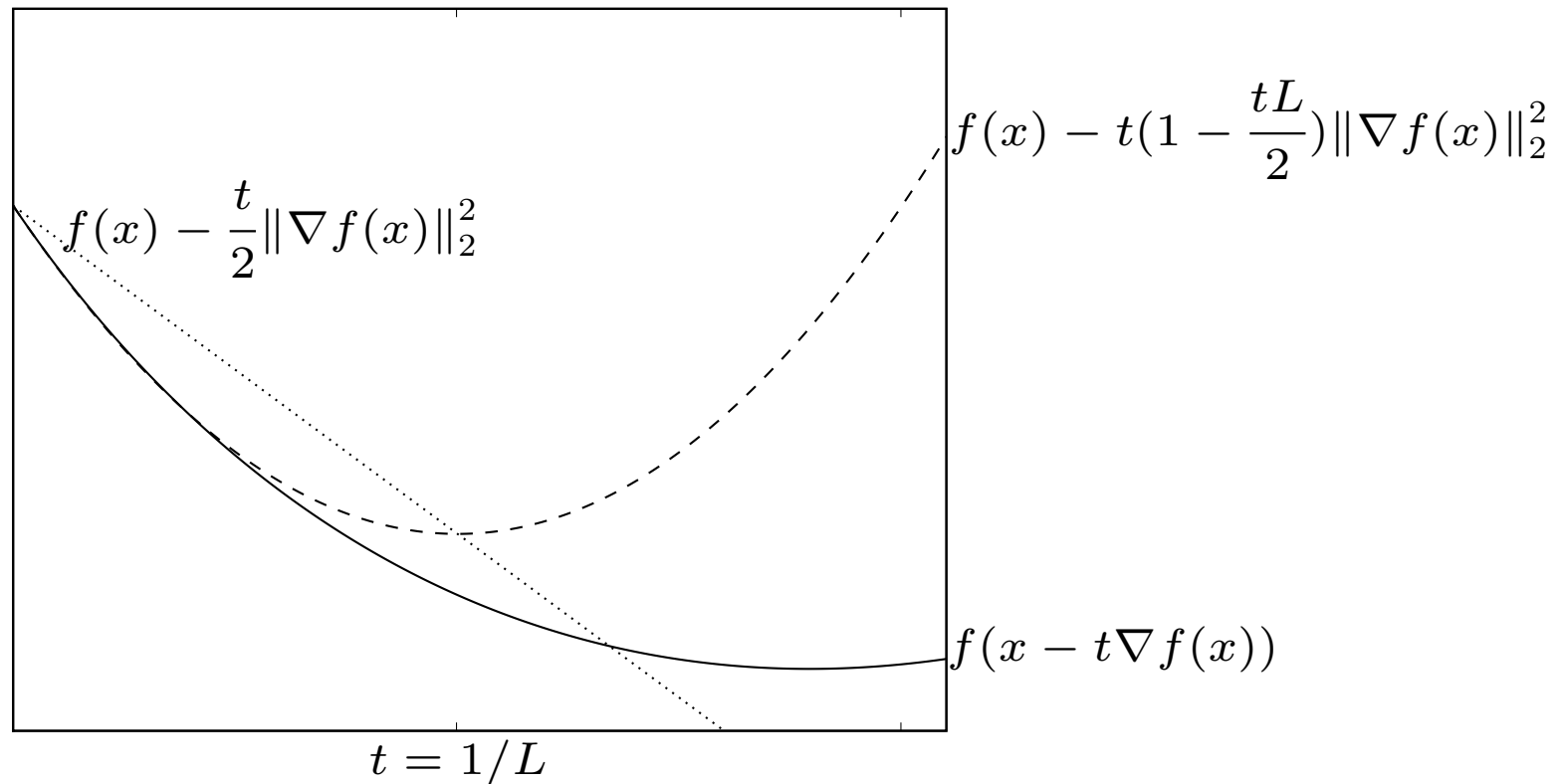
$$f(x - t_k \nabla f(x)) < f(x) - \alpha t_k \|\nabla f(x)\|_2^2$$



$0 < \beta < 1$; we will take $\alpha = 1/2$ (mostly to simplify proofs)

Analysis for backtracking line search

line search with $\alpha = 1/2$ if f has a Lipschitz continuous gradient



selected step size satisfies $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

convergence analysis

- from page 1-17:

$$\begin{aligned} f(x^{(i)}) &\leq f^* + \frac{1}{2t_i} \left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right) \\ &\leq f^* + \frac{1}{2t_{\min}} \left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right) \end{aligned}$$

- add the upper bounds to get

$$f(x^{(k)}) - f^* \leq \frac{1}{k} \sum_{i=1}^k (f(x^{(i)}) - f^*) \leq \frac{1}{2kt_{\min}} \|x^{(0)} - x^*\|_2^2$$

conclusion: same $1/k$ bound as with constant step size

Analysis for strongly convex functions

better results exist if we add strong convexity to the assumptions on p.1-16

analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 2/(m + L)$:

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

(step 3 follows from result on p.1-15)

distance to optimum

$$\|x^{(k)} - x^*\|_2^2 \leq c^k \|x^{(0)} - x^*\|_2^2, \quad c = 1 - t \frac{2mL}{m + L}$$

- implies (linear) convergence
- for $t = 2/(m + L)$, get $c = ((\kappa - 1)/(\kappa + 1))^2$ with $\kappa = L/m$

bound on function value (from page 1-13)

$$f(x^{(k)}) - f^* \leq \frac{L}{2} \|x^{(k)} - x^*\|_2^2 \leq \frac{c^{2k} L}{2} \|x^{(0)} - x^*\|_2^2$$

conclusion: #iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(\log(1/\epsilon))$

Limits on convergence rate of first-order methods

first-order method: any iterative algorithm that selects $x^{(k)}$ in

$$x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \dots, \nabla f(x^{(k-1)})\}$$

problem class: any function that satisfies the assumptions on p. 1-16

theorem (Nesterov): for every integer $k \leq (n - 1)/2$ and every $x^{(0)}$, there exist functions in the problem class such that for any first-order method

$$f(x^{(k)}) - f^* \geq \frac{3}{32} \frac{L \|x^{(0)} - x^*\|_2^2}{(k + 1)^2}$$

conclusion: room to improve $1/k$ rate of gradient method to $1/k^2$ (lec.8)

References

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), section 2.1.
- B. T. Polyak, *Introduction to Optimization* (1987), section 1.4.
- the example on page 1-4 is from N. Z. Shor, *Nondifferentiable Optimization and Polynomial Problems* (1998), page 37.