1. Gradient method

- gradient method, first-order methods
- quadratic bounds on convex functions
- analysis of gradient method

Approximate course outline

First-order methods

- gradient, conjugate gradient, quasi-Newton methods
- subgradient, proximal gradient methods
- accelerated (proximal) gradient methods

Decomposition and splitting methods

- first-order methods and dual reformulations
- alternating minimization methods
- monotone operators and operator-splitting methods

Interior-point methods

- conic optimization
- primal-dual interior-point methods

Gradient method

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

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

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

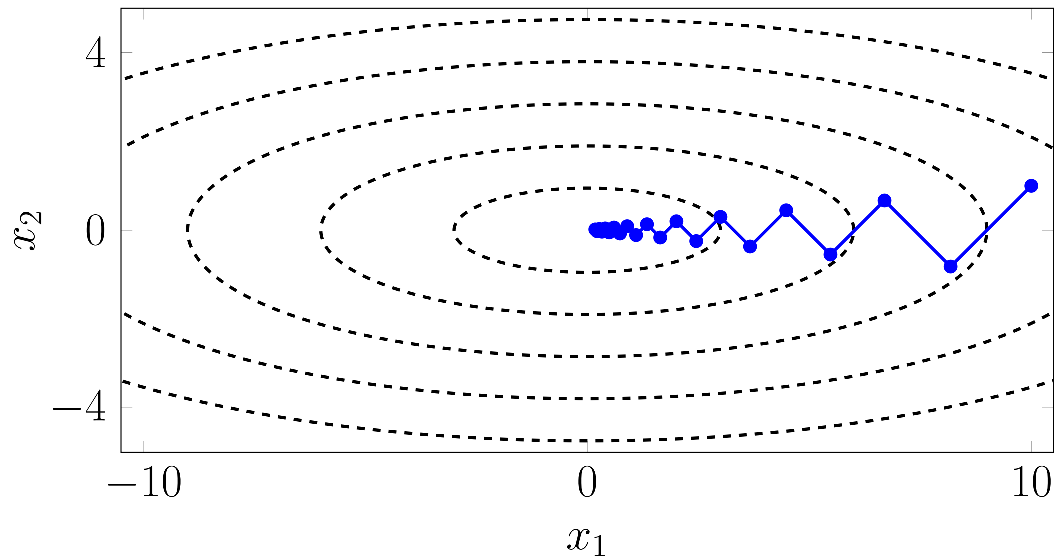
- every iteration is inexpensive
- does not require second derivatives

Quadratic example

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

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

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

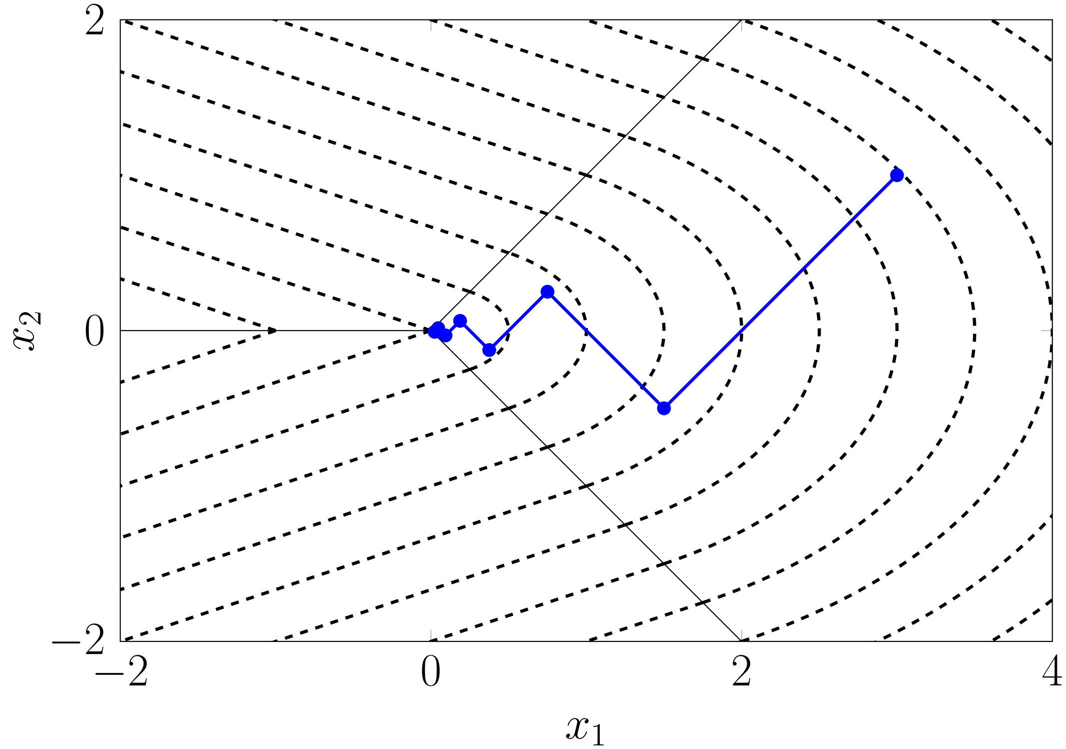


gradient method is often slow; convergence very dependent on scaling

Nondifferentiable example

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

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

- quasi-Newton methods
- conjugate gradient method
- accelerated gradient method

Methods for nondifferentiable or constrained problems

- subgradient method
- proximal gradient method
- smoothing methods
- cutting-plane methods

Outline

- gradient method, first-order methods
- **quadratic bounds on convex functions**
- analysis of gradient method

Convex function

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

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \quad \text{for all } x, y \in \operatorname{dom} f,\ \theta \in [0,1]$$

First-order condition

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

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \operatorname{dom} f$$

Second-order condition

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

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \operatorname{dom} f$$

Strictly convex function

f is *strictly convex* if $\mathrm{dom}\, f$ is a convex set and

$$f(\theta x+(1-\theta)y) < \theta f(x)+(1-\theta)f(y) \quad \text{for all } x, y \in \mathrm{dom}\, f,\ x \neq y, \text{ and } \theta \in (0,1)$$

strict convexity implies that if a minimizer of f exists, it is unique

First-order condition

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

$$f(y) > f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathrm{dom}\, f,\ x \neq y$$

Second-order condition

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

Monotonicity of gradient

a differentiable function f is convex if and only if $\mathrm{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \text{for all } x, y \in \mathrm{dom}\, f$$

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

a differentiable function f is strictly convex if and only if $\mathrm{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) > 0 \quad \text{for all } x, y \in \mathrm{dom}\, f,\, x \neq y$$

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

Proof

- if f is differentiable and convex, then

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

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 $t \geq 0$ and $t \in \mathbf{dom}\, g$, where

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

hence

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)\, dt \quad &\geq \quad g(0) + g'(0) \\ &= \quad f(x) + \nabla f(x)^T(y-x) \end{aligned}$$

this is the first-order condition for convexity

Lipschitz continuous gradient

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y \in \operatorname{dom} f$$

- note that the definition does not assume convexity of f
- we will see that for convex f with $\operatorname{dom} f = \mathbf{R}^n$, this is equivalent to

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

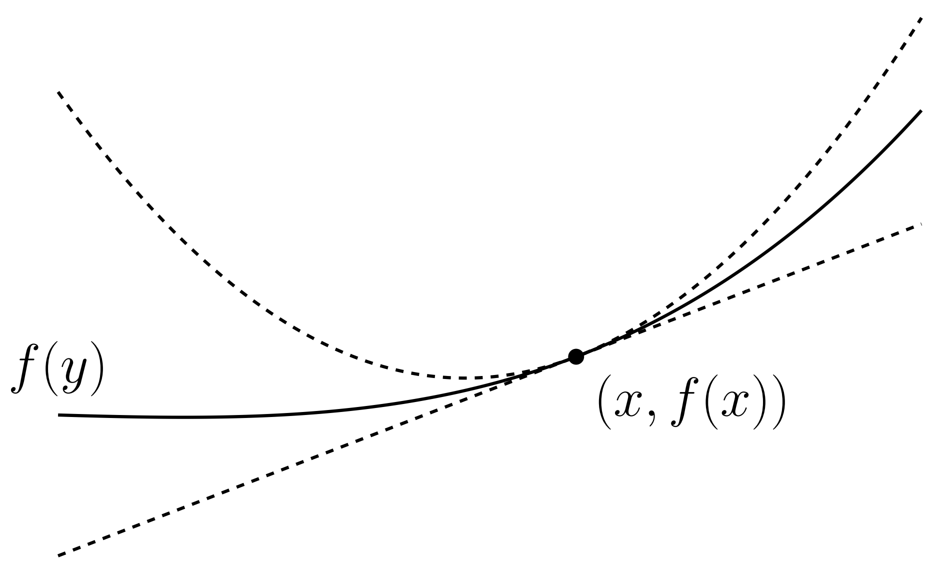
  (*i.e.*, if f is twice differentiable, $\nabla^2 f(x) \preceq LI$ for all x)

Quadratic upper bound

suppose ∇f is Lipschitz continuous with parameter L and $\mathrm{dom}\, f$ is convex

- then $g(x) = (L/2)x^T x - f(x)$, with $\mathrm{dom}\, g = \mathrm{dom}\, f$, is convex
- convexity of g is equivalent to a quadratic upper bound on f:

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \text{for all } x, y \in \mathrm{dom}\, f$$



Proof

- Lipschitz continuity of ∇f and the Cauchy-Schwarz inequality imply

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq L\|x - y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

  this is monotonicity of the gradient

$$\nabla g(x) = Lx - \nabla f(x)$$

- hence, g is a convex function if its domain $\operatorname{dom} g = \operatorname{dom} f$ is convex
- the quadratic upper bound is the first-order condition for convexity of g

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) \quad \text{for all } x, y \in \operatorname{dom} g$$

Consequence of quadratic upper bound

if $\mathrm{dom}\, f = \mathbf{R}^n$ and f has a minimizer $x^\star$, then

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^\star) \leq \frac{L}{2}\|x - x^\star\|_2^2 \quad \text{for all } x$$

- right-hand inequality follows from quadratic upper bound at $x = x^\star$
- left-hand inequality follows by minimizing quadratic upper bound

$$\begin{aligned} f(x^\star) &\leq \inf_{y \in \mathrm{dom}\, f} \left(f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \right) \\ &= f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \end{aligned}$$

minimizer of upper bound is $y = x - (1/L)\nabla f(x)$ because $\mathrm{dom}\, f = \mathbf{R}^n$

Co-coercivity of gradient

if f is convex with $\operatorname{dom} f = \mathbf{R}^n$ and $(L/2)x^Tx - f(x)$ is convex then

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \text{for all } x, y$$

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

- co-coercivity implies Lipschitz continuity of ∇f (by Cauchy-Schwarz)
- hence, for differentiable convex f with $\operatorname{dom} f = \mathbf{R}^n$

$$\begin{aligned} \text{Lipschitz continuity of } \nabla f \quad & \Rightarrow \quad \text{convexity of } (L/2)x^Tx - f(x) \\ & \Rightarrow \quad \text{co-coercivity of } \nabla f \\ & \Rightarrow \quad \text{Lipschitz continuity of } \nabla f \end{aligned}$$

therefore the three properties are equivalent

Proof of co-coercivity: define two convex functions f_x, f_y with domain $\mathbf{R}^n$

$$f_x(z) = f(z) - \nabla f(x)^T z, \qquad f_y(z) = f(z) - \nabla f(y)^T z$$

the functions $(L/2)z^T z - f_x(z)$ and $(L/2)z^T z - f_y(z)$ are convex

- $z = x$ minimizes $f_x(z)$; from the left-hand inequality on page 1-14,

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

- similarly, $z = y$ minimizes $f_y(z)$; therefore

$$f(x) - f(y) - \nabla f(y)^T (x - y) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

combining the two inequalities shows co-coercivity

Strongly convex function

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

$$g(x) = f(x) - \frac{m}{2}x^T x \quad \text{is convex}$$

Jensen's inequality: Jensen's inequality for g is

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

Monotonicity: monotonicity of ∇g gives

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

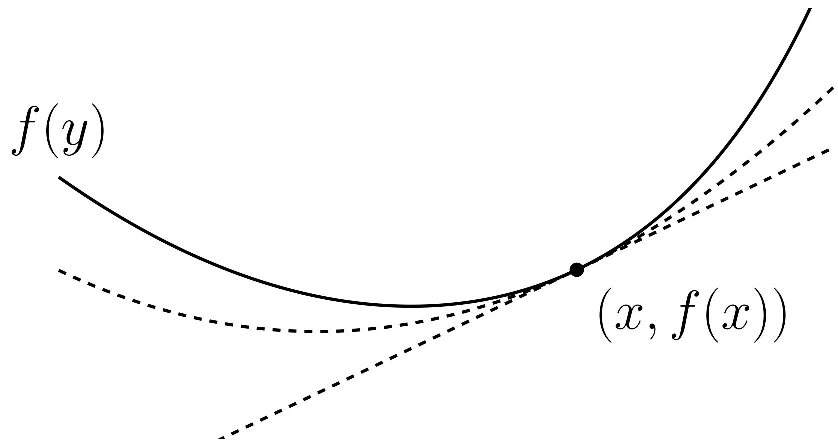
this is called *strong monotonicity (coercivity)* of ∇f

Second-order condition: $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbf{dom}\, f$

Quadratic lower bound

from 1st order condition of convexity of g:

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



- implies sublevel sets of f are bounded
- if f is closed (has closed sublevel sets), it has a unique minimizer $x^\star$ and

$$\frac{m}{2}\|x - x^\star\|_2^2 \leq f(x) - f(x^\star) \leq \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad \text{for all } x \in \operatorname{dom} f$$

Extension of co-coercivity

- if f is strongly convex and ∇f is Lipschitz continuous, then the function

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

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

- co-coercivity of g gives

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

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

Outline

- gradient method, first-order methods
- quadratic bounds on convex functions
- **analysis of gradient method**

Analysis of gradient method

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

with fixed step size or backtracking line search

Assumptions

1. f is convex and differentiable with $\mathrm{dom}\, f = \mathbf{R}^n$

2. $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$

3. optimal value $f^\star = \inf_x f(x)$ is finite and attained at $x^\star$

Analysis for constant step size

- from quadratic upper bound (page 1-12) with $y = x - t\nabla f(x)$:

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

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

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

second line follows from convexity of f

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

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

- since $f(x^{(i)})$ is non-increasing (see (1))

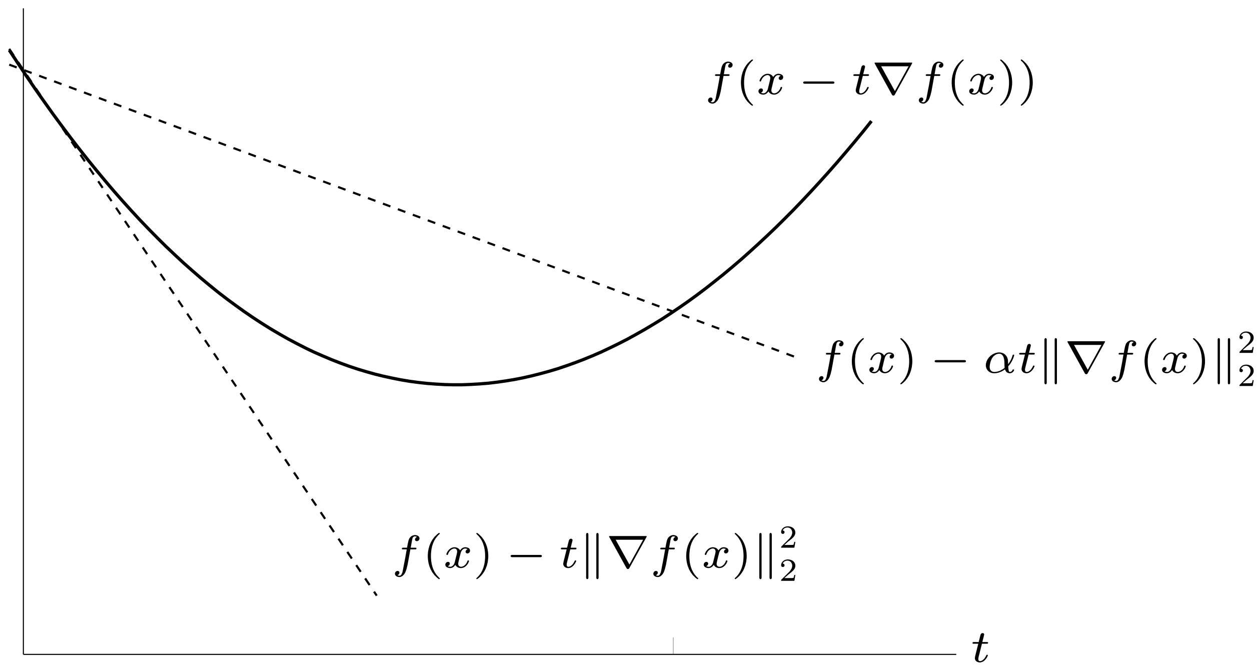
$$f(x^{(k)}) - f^\star \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2kt}\|x^{(0)} - x^\star\|_2^2$$

Conclusion: number of iterations to reach $f(x^{(k)}) - f^\star \leq \epsilon$ 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

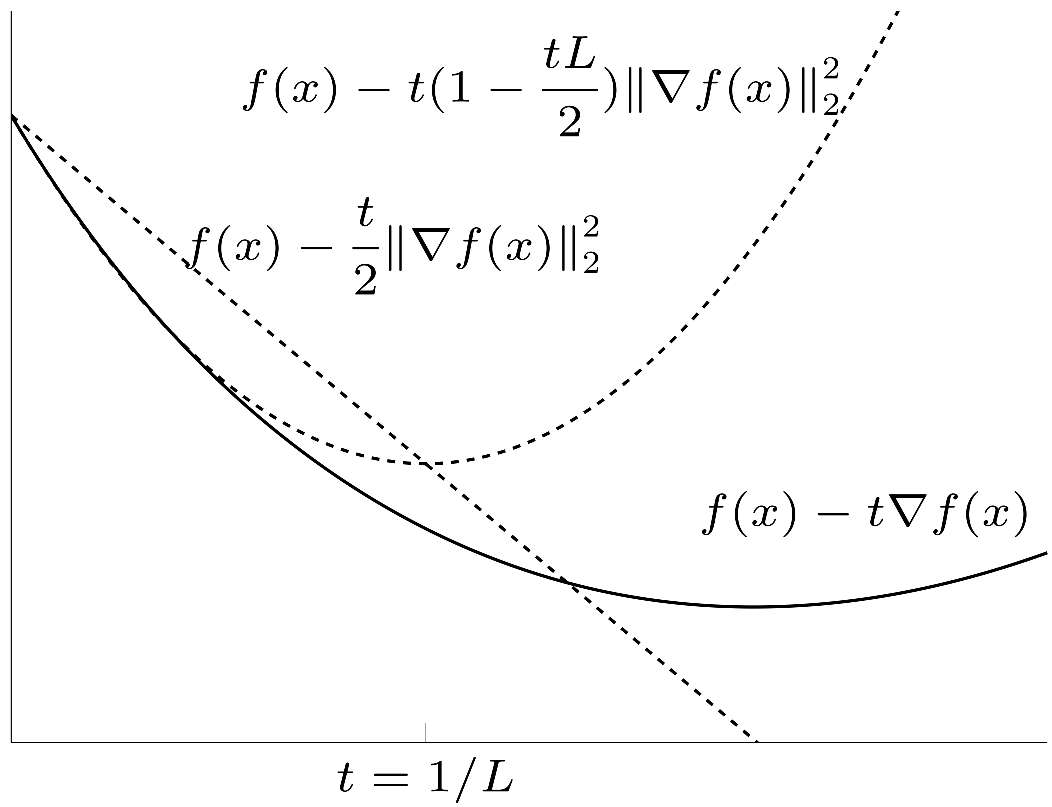
$$f(x - t_k \nabla f(x)) < f(x) - \alpha t_k \|\nabla f(x)\|_2^2$$



$0 < \beta < 1$; 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

- as on page 1-21:

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

the first line follows from the line search condition

- add the upper bounds to get

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

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

Gradient method for strongly convex functions

better results exist if we add strong convexity to the assumptions on p. 1-20

Analysis for constant step size

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

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

(step 3 follows from result on p. 1-19)

Distance to optimum

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

- implies (linear) convergence
- for $t = 2/(m+L)$, get $c = \left(\frac{\gamma - 1}{\gamma + 1}\right)^2$ with $\gamma = L/m$

Bound on function value (from page 1-14)

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

Conclusion: number of iterations to reach $f(x^{(k)}) - f^\star \le \epsilon$ is $O(\log(1/\epsilon))$

Limits on convergence rate of first-order methods

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

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

Problem class: any function that satisfies the assumptions on page 1-20

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

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

- suggests $1/k$ rate for gradient method is not optimal
- recent fast gradient methods have $1/k^2$ convergence (see later)

References

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), section 2.1 (the result on page 1-28 is Theorem 2.1.7 in the book)
- B. T. Polyak, *Introduction to Optimization* (1987), section 1.4
- the example on page 1-5 is from N. Z. Shor, *Nondifferentiable Optimization and Polynomial Problems* (1998), page 37