2. Gradient methods

- classes of convex functions
- classical gradient method
- complexity analysis of gradient method
- Newton and quasi-Newton methods

Convex function

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

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

first-order condition

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

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall\, x, y \in \mathbf{dom}\, f$$

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 convex and for all $x, y \in \mathbf{dom}\, f$ and $x \neq y$

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y), \quad \forall\, \theta \in (0,1)$$

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

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall\, x, y \in \mathbf{dom}\, f \text{ and } x \neq y$$

hence minimizer of f is unique (if it 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 $\mu > 0$ if

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

Jensen's inequality

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

first-order condition

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

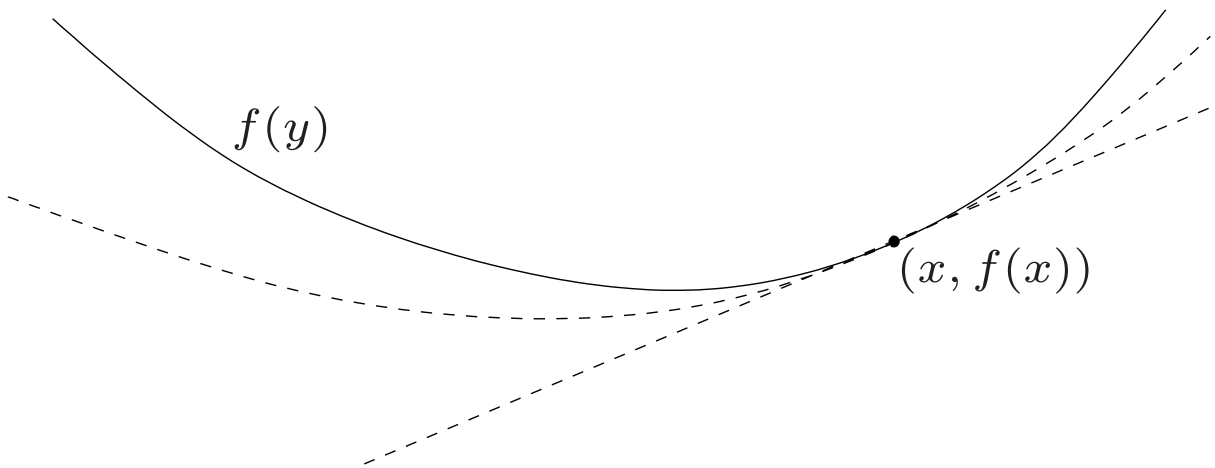
second-order condition

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

Quadratic lower bound

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

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



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

$$\frac{\mu}{2}\|x-x^\star\|_2^2 \ \leq\ f(x) - f(x^\star) \ \leq\ \frac{1}{2\mu}\|\nabla f(x)\|_2^2, \qquad \forall\, x \in \mathbf{R}^n$$

Functions with Lipschitz continuous gradients

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

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

quadratic upper and lower bounds

$$\left| f(y) - f(x) - \nabla f(x)^T (y - x) \right| \;\leq\; \frac{L}{2}\|y - x\|_2^2$$

for convex functions, only the upper bound is useful

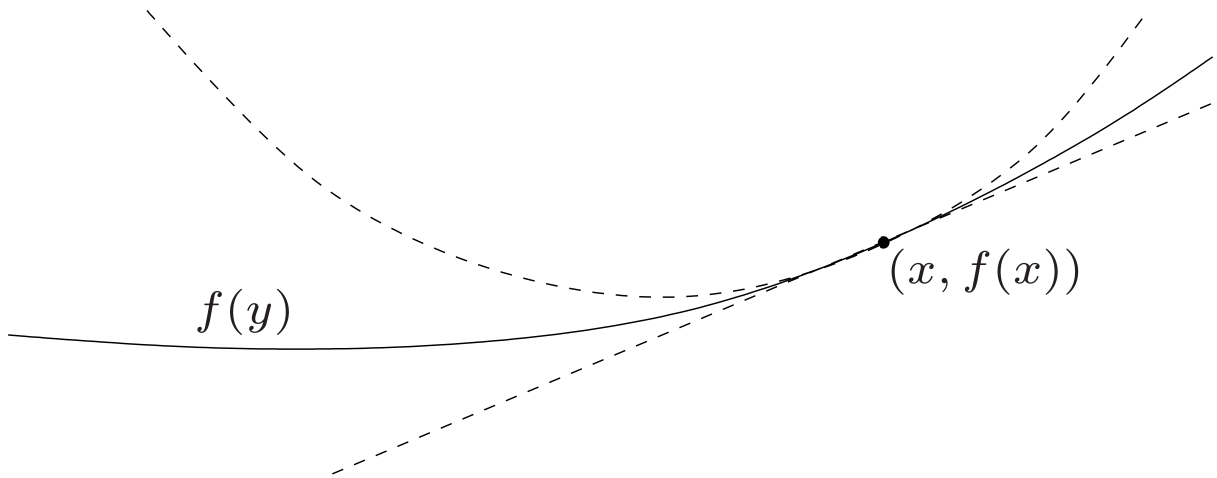
second-order condition (for twice continuously differentiable function)

$$\nabla^2 f(x) \preceq LI, \qquad \forall\, x \in \mathbf{R}^n$$

Quadratic upper bound

if $\nabla f(x)$ is Lipschitz-continuous with parameter $L > 0$, then

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



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

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \;\le\; f(x) - f(x^\star) \;\le\; \frac{L}{2}\|x - x^\star\|_2^2$$

Classical gradient method

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

$$x^{(k+1)} = x^{(k)} - t_k \nabla f(x^{(k)}), \qquad k = 0, 1, 2, \ldots$$

step size rules

- exact line search: $t_k = \underset{t}{\operatorname{argmin}}\, f(x^{(k)} - t\nabla f(x^{(k)}))$
- fixed: t_k constant
- backtracking line search (most practical)

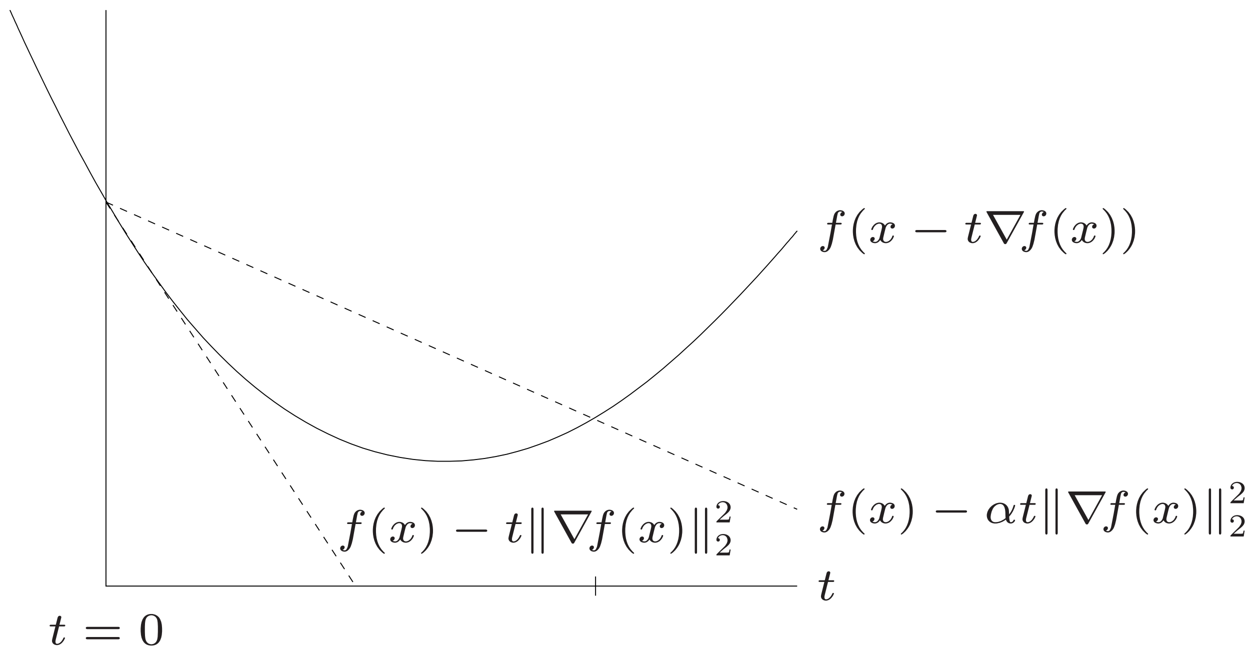
advantages of gradient method

- every iteration is inexpensive
- does not require second derivatives

Backtracking line search

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

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



two parameters: $0 < \beta < 1$ and $0 < \alpha \leq 0.5$

Analysis of gradient method

$$x^{(k+1)} = x^{(k)} - t_k \nabla f(x^{(k)}), \qquad k = 0, 1, 2, \ldots$$

with fixed step size or backtracking line search

assumptions

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

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

recall quadratic upper bound: $f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$,
plug in $y = x - t\nabla f(x)$ to obtain

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

let $x^+ = x - t\nabla f(x)$ and assume $0 < t \le 1/L$,

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

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

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

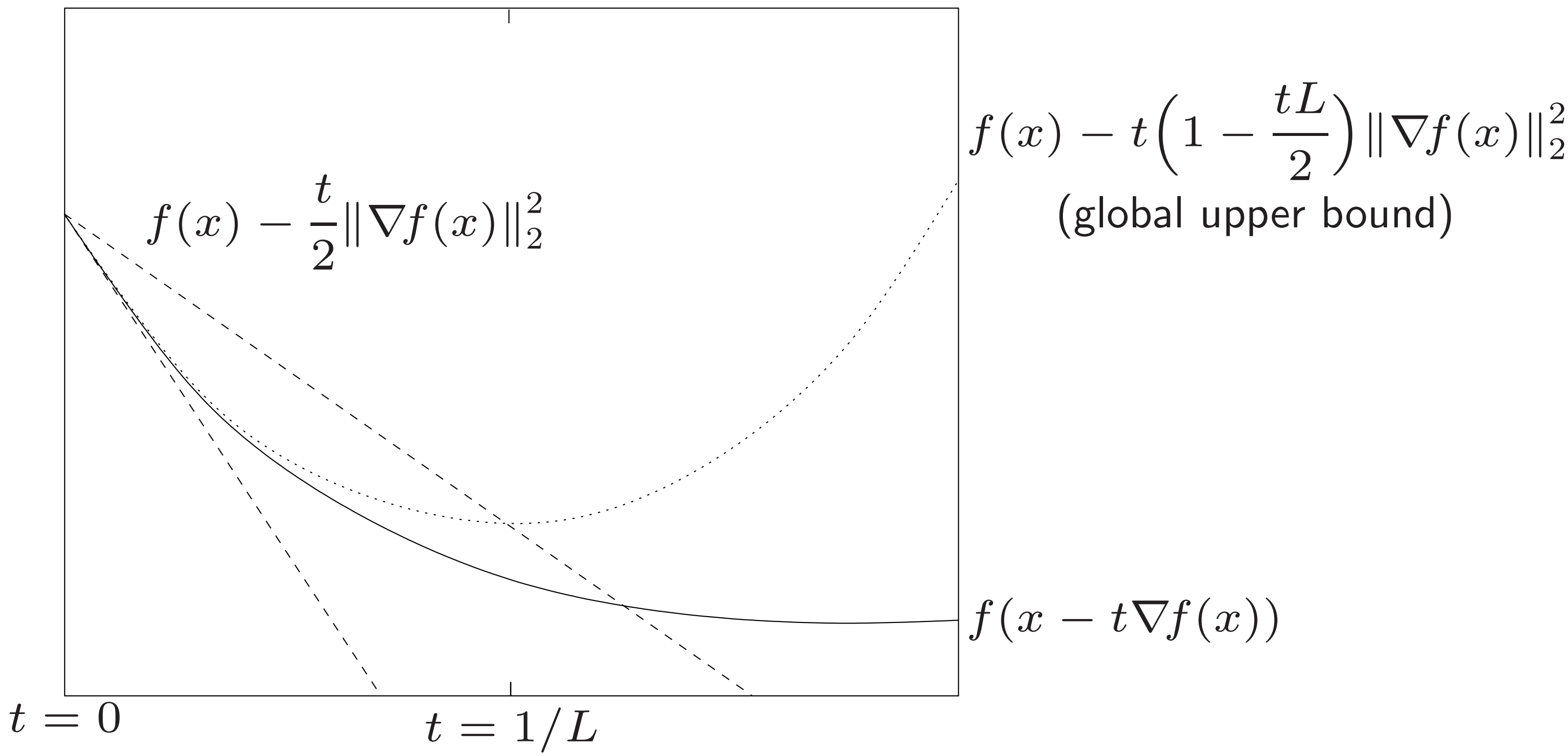
since $f(x^{(i)})$ is non-increasing,

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

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

Analysis for backtracking line search

line search with $\alpha = 1/2$ and $0 < \beta < 1$



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

convergence analysis

- from page 2–11:

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

- add the upper bounds to obtain

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

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

Analysis for strongly convex functions

faster convergence rate with additional assumption of strong convexity

analysis for exact line search: recall from quadratic upper bound

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

use $x^+ = \operatorname{argmin}_t f(x - t\nabla f(x))$ to obtain

$$f(x^+) \;\leq\; f\left(x - \frac{1}{L}\nabla f(x)\right) \;\leq\; f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

subtract $f^\star$ from both sides

$$f(x^+) - f^\star \;\leq\; f(x) - f^\star - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

now use strong convexity: $f(x) - f^\star \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$

$$f(x^+) - f^\star \;\leq\; \left(1 - \frac{\mu}{L}\right)(f(x) - f^\star)$$

therefore

$$f(x^{(k)}) - f^\star \;\leq\; \left(1 - \frac{\mu}{L}\right)^k \left(f(x^{(0)}) - f^\star\right)$$

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

$$\frac{\log\left((f(x^{(0)}) - f^\star)/\epsilon\right)}{\log(1 - \mu/L)^{-1}} \;\approx\; \frac{L}{\mu}\log\left(\frac{f(x^{(0)}) - f^\star}{\epsilon}\right)$$

- roughly proportional to *condition number* L/μ when it is large
- slightly tighter bound exists (smaller constant in iteration bound)
- distance to optimum $\|x^{(k)} - x^\star\|_2$ also decreases geometrically

Numerical examples

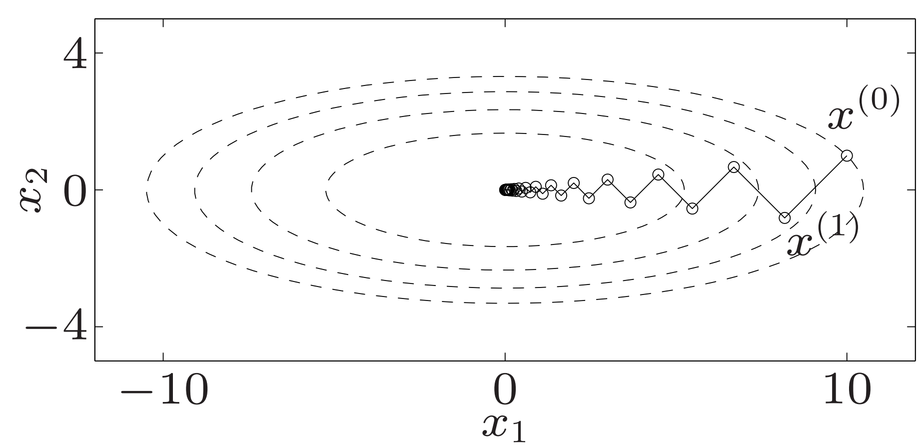
quadratic example

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

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

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

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

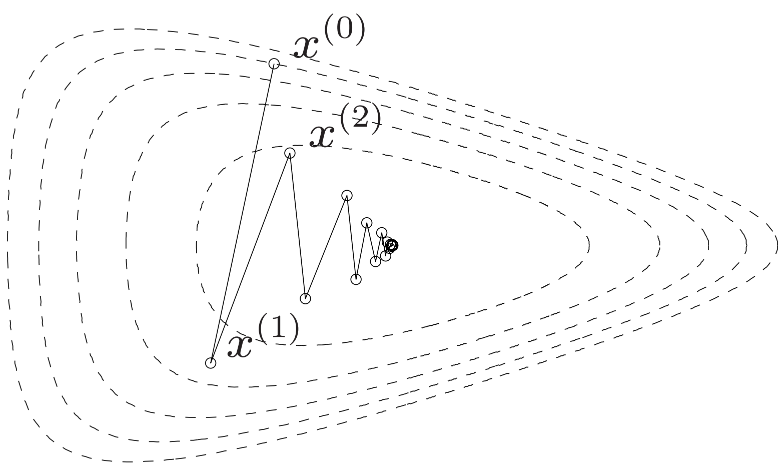


gradient method is often very slow; very much dependent on scaling

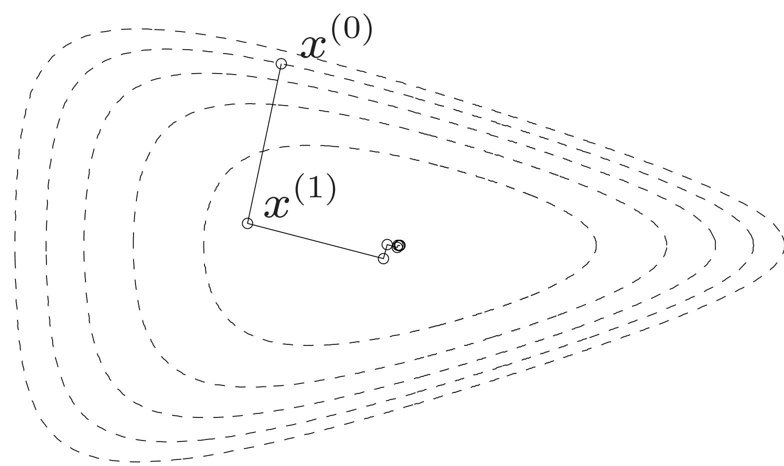
nonquadratic example

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$

$(\alpha = 0.1,\ \beta = 0.7)$



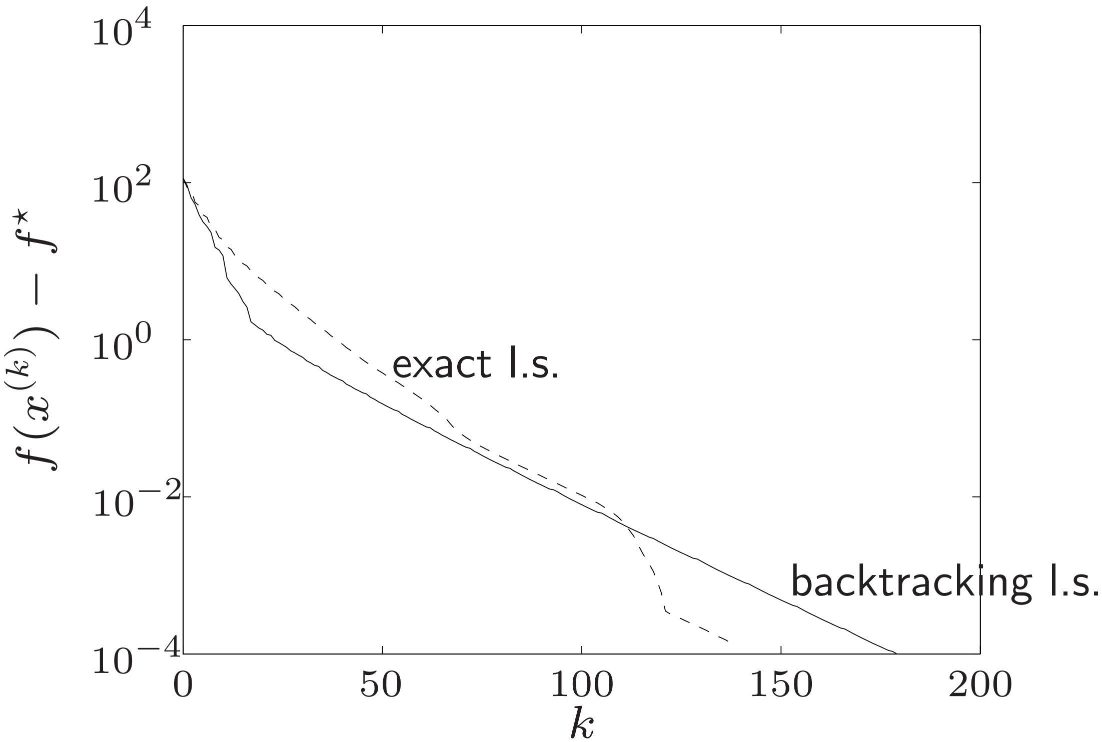
backtracking line search



exact line search

a problem in $\mathbf{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$



linear convergence, i.e., a straight line on a semilog plot

Newton's method

assume $f(x)$ is twice continuously differentiable and convex

(pure) Newton method

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

damped Newton method

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

- *advantages:* fast convergence, affine invariance
- *disadvantages:* requires second derivatives, solution of linear equation

can be too expensive for large-scale applications

Classical convergence analysis

assumptions

- f strongly convex with parameter μ
- $\nabla^2 f$ is Lipschitz continuous with parameter $M > 0$:

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

(M measures how well f can be approximated by a quadratic function)

outline: there exist constants $\eta \in (0, \mu^2/M)$, $\gamma > 0$ such that

- if $\|\nabla f(x)\|_2 \geq \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$
- if $\|\nabla f(x)\|_2 < \eta$, then

$$\frac{M}{2\mu^2}\|\nabla f(x^{(k+1)})\|_2 \ \leq \ \left(\frac{M}{2\mu^2}\|\nabla f(x^{(k)})\|_2\right)^2$$

damped Newton phase ($\|\nabla f(x)\|_2 \geq \eta$)

- most iterations require backtracking steps
- at each iteration, function value decreases by at least γ

quadratically convergent phase ($\|\nabla f(x)\|_2 < \eta$)

- all iterations use step size $t = 1$
- $\|\nabla f(x)\|_2$ converges to zero quadratically:

$$\frac{M}{2\mu^2}\|\nabla f(x^l)\|_2 \;\leq\; \left(\frac{M}{2\mu^2}\|\nabla f(x^k)\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}, \qquad l \geq k$$

- quadratic convergence for $f(x^{(k)}) - f^\star$ and $\|x^{(k)} - x^\star\|_2$

conclusion: number of iterations until $f(x) - f^\star \leq \epsilon$ is bounded above by

$$\frac{f(x^{(0)}) - f^\star}{\gamma} \;+\; \log_2 \log_2(\epsilon_0/\epsilon)$$

Convergence rate and complexity bound

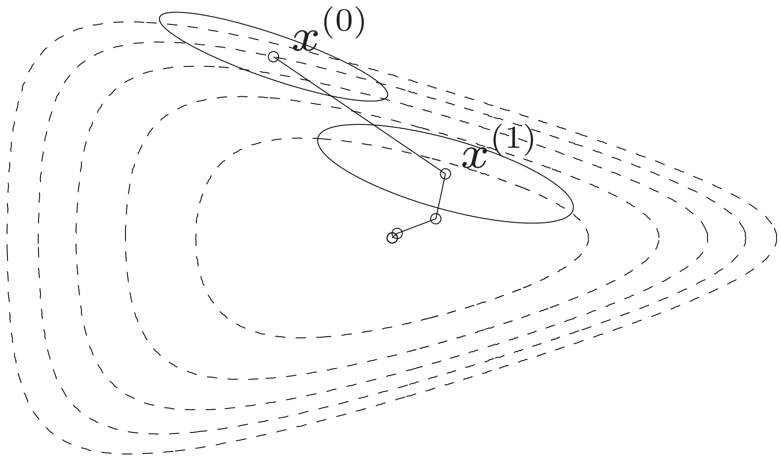
	convergence rate	complexity bound	dependence on c
sublinear rate	$r_k \leq \dfrac{c}{k^p}$	$\left(\dfrac{c}{\epsilon}\right)^{1/p}$	strong
linear rate	$r_k \leq c(1-q)^k$	$\dfrac{1}{q}\left(\log c + \log \dfrac{1}{\epsilon}\right)$	weak
quadratic rate	$r_{k+1} \leq c r_k^2$	$\log\log \dfrac{1}{\epsilon}$	very weak

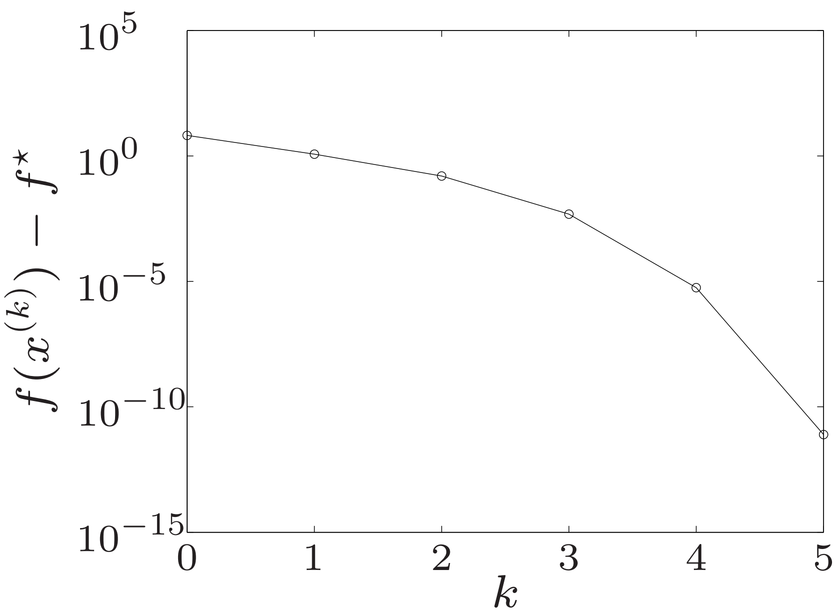
r_k can be $f(x^{(k)}) - f^\star$, $\|x^{(k)} - x^\star\|_2$, or $\|\nabla f(x^{(k)})\|_2$; c is some constant

- complexity bound is inverse function of rate of convergence
- interpretation through amount of work for each correct digit

Examples for Newton's method

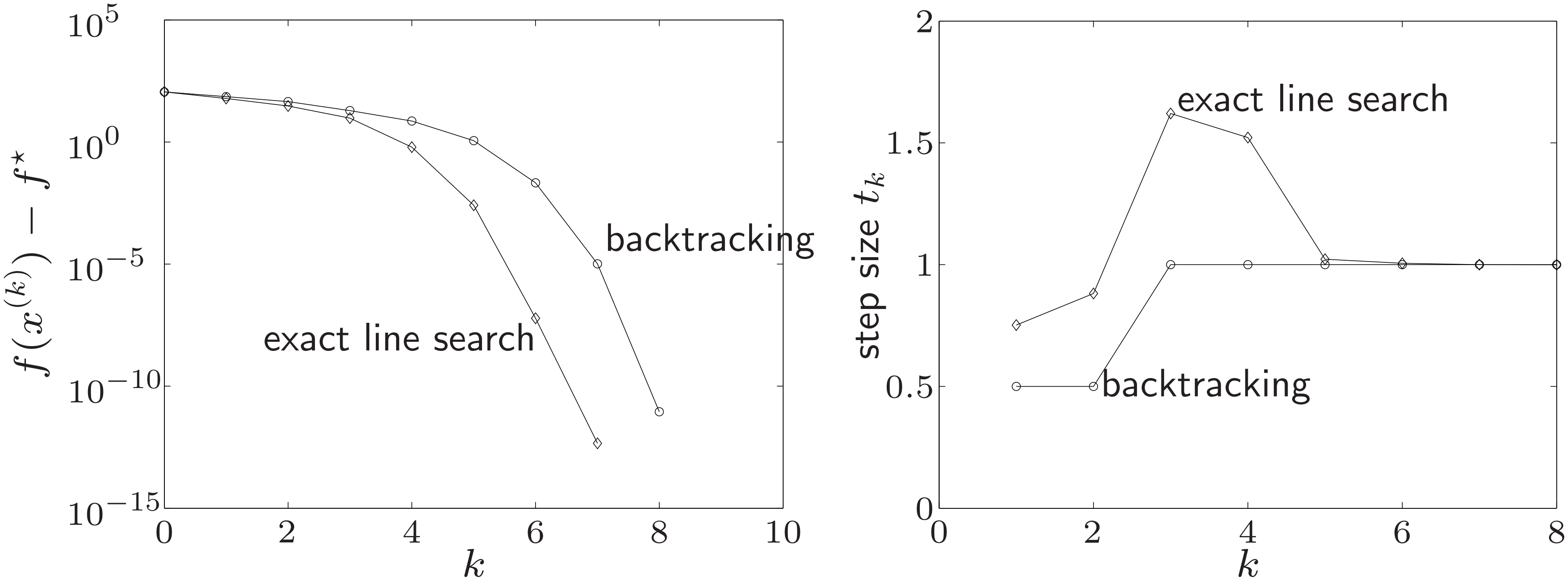
example in $\mathbf{R}^2$ (page 2–18)





- backtracking parameters $\alpha = 0.1$, $\beta = 0.7$
- converges in only 5 steps
- quadratic local convergence

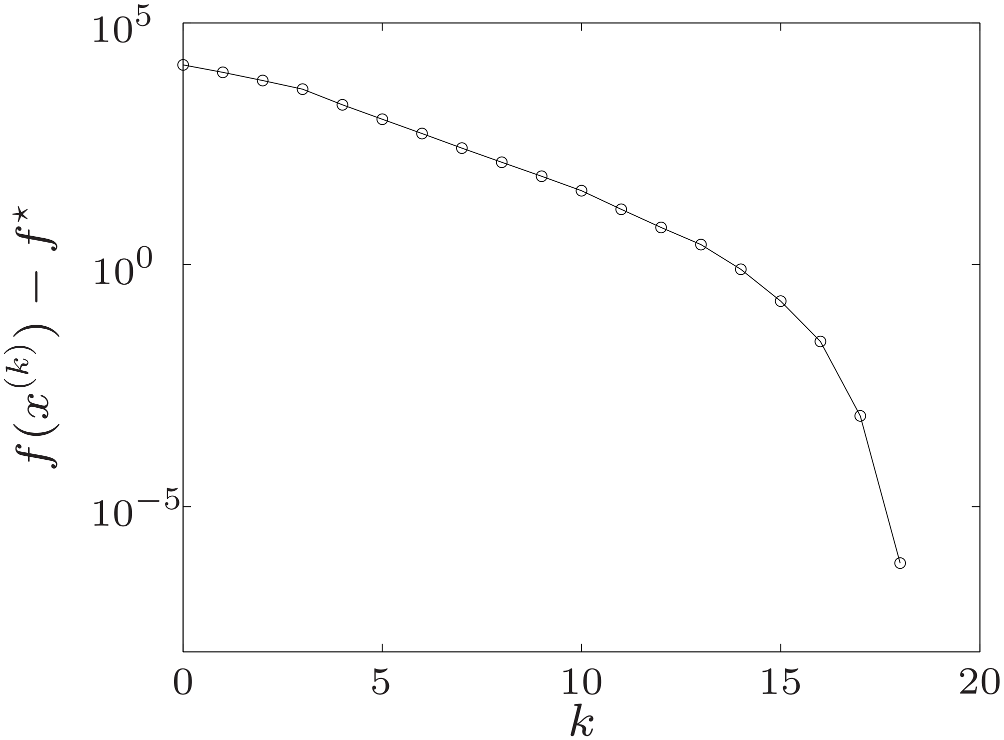
example in $\mathbf{R}^{100}$ (page 2–19)



- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$
- backtracking line search almost as fast as exact l.s. (and much simpler)
- clearly shows two phases in algorithm

example in $\mathbf{R}^{10000}$ (with sparse a_i)

$$f(x) = -\sum_{i=1}^{10000} \log(1 - x_i^2) - \sum_{i=1}^{100000} \log(b_i - a_i^T x)$$



- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$.
- performance similar as for small examples

Approximation

majority of general nonlinear optimization methods are based on

nonincreasing seq.: generate a sequence $\{x^{(k)}\}_{k=0}^{\infty}$ such that

$$f(x^{(k+1)}) \leq f(x^{(k)}), \qquad k = 0, 1, 2, \ldots$$

- if $f(x)$ is bounded below, then the sequence $\{f(x^{(k)})\}_{k=0}^{\infty}$ converges
- we always improve the objective function

another view:

approximation: replace original complex objective by a simplified one

- local approximation: first-order and second-order approximations
- global perspectives are necessary for optimal methods (next lecture)

An approximation perspective

$$x^{(k+1)} \;=\; \underset{y}{\operatorname{argmin}}\; \phi_{t_k}(x^{(k)}; y)$$

where $\phi_{t_k}(x^{(k)}; y)$ is an approximation of f near $x^{(k)}$, with parameter t_k

gradient method

$$\phi_t^{\mathrm{grad}}(x; y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}\|y - x\|_2^2$$

(damped) Newton's method

$$\phi_t^{\mathrm{Newton}}(x; y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}(y - x)^T \nabla^2 f(x)(y - x)$$

role of line search: choose appropriate parameter t for approximation

Variable metric method

$$x^{(k+1)} = \underset{y}{\operatorname{argmin}} \ \phi_{t_k}(x^{(k)}; y)$$

where

$$\phi_{t_k}(x^{(k)}; y) = f(x^{(k)}) + \nabla f(x^{(k)})^T (y - x^{(k)}) + \frac{1}{2t_k}(y - x^{(k)})^T H_k (y - x^{(k)})$$

- better approximation than gradient method

$$\{H_k\} : H_k \to \nabla^2 f(x^\star)$$

- less expensive than Newton's method

(low-rank) updates of $\{H_k\}$ or $\{H_k^{-1}\}$ only involve gradients

- *variable metric*: steepest descent direction with quadratic norm

$$\|z\|_{H_k} = \sqrt{z^T H_k z}$$

Variable metric methods

given initial point $x^{(0)}$ and $H_0 \succ 0$

repeat for $k = 0, 1, 2, \ldots$ until a stopping criterion is satisfied

1. compute quasi-Newton direction

$$\Delta x = -H_k^{-1} \nabla f(x^{(k)})$$

2. determine step size t_k (e.g., via backtracking line search)
3. update $x^{(k+1)} = x^{(k)} + t_k \Delta x$ and call oracle for $\nabla f(x^{(k+1)})$
4. compute H_{k+1} based on current information set

- different methods use different rules for updating H_k in step 4
- can directly propagate H_k^{-1} to simplify calculation of Δx

Secant condition (quasi-Newton rule)

$$\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = H_{k+1}\left(x^{(k+1)} - x^{(k)}\right)$$

interpretation: for any quadratic function

$$f(x) = \alpha + \langle h, x\rangle + \frac{1}{2}\langle Hx, x\rangle$$

we have $\nabla f(x) = Hx + h$, and therefore for any $x, y \in \mathbf{R}^n$,

$$\nabla f(x) - \nabla f(y) = H(x - y)$$

Broyden-Fletcher-Goldfard-Shanno (BFGS)

BFGS update

$$H_{k+1} = H_k - \frac{H_k s s^T H_k}{s^T H_k s} + \frac{y y^T}{y^T s}$$

where

$$s = x^{(k+1)} - x^{(k)}, \qquad y = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$$

inverse update

$$H_{k+1}^{-1} = \left(I - \frac{s y^T}{y^T s}\right) H_k^{-1} \left(I - \frac{y s^T}{y^T s}\right) + \frac{s s^T}{y^T s}$$

- satisfies secant condition with unit step size
- $y^T s > 0$ preserves positive definiteness, thus ensures descent direction
- cost of update or inverse update is $O(n^2)$ arithmetic operations

Convergence result

global convergence

if f is strongly convex, then BFGS with backtracking line search converges to the optimum for any $x^{(0)}$ and $H_0 \succ 0$

local convergence

if f is strongly convex and $\nabla^2 f(x)$ is Lipschitz continuous, then local convergence is *superlinear*: for sufficiently large k,

$$\|x^{(k+1)} - x^\star\|_2 \leq c_k \|x^{(k)} - x^\star\|_2$$

where $c_k \to 0$ (cf., quadratic local convergence of Newton's method)

Low-memory quasi-Newton methods

main disadvantage of quasi-Newton method is need to store H_k or H_k^{-1}

limited-memory BFGS (L-BFGS): do not store H_k^{-1} explicitly

- instead store m (say, $m = 30$) most recent values of

$$s_j = x^{(j)} - x^{(j-1)}, \qquad y_j = \nabla f(x^{(j)}) - \nabla f(x^{(j-1)})$$

- evaluate $\Delta x = -H_k^{-1}\nabla f(x^{(k)})$ recursively, using

$$H_j^{-1} = \left(I - \frac{s_j y_j^T}{y_j^T s_j}\right) H_{j-1}^{-1} \left(I - \frac{y_j s_j^T}{y_j^T s_j}\right) + \frac{s_j s_j^T}{y_j^T s_j}$$

  for $j = k, k-1, \ldots, j-m+1$, assuming, for example, $H_{k-m}^{-1} = I$

- cost per iteration is $O(mn)$; storage is $O(mn)$

References

- S. Boyd and L. Vandenberghe, *Convex Optimization* (2004), Chapter 9.

- J. Nocedal and S. J. Wright, *Numerical Optimization (2nd Edition)* (2006), Chapters 3 and 6.

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), Sections 1.2, 1.3 and 2.1.

- L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems*, UCLA.