

1. Gradient method

- classical gradient method
- convergence analysis
- Nesterov's accelerated gradient method
- optimality of Nesterov's method

1-1

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

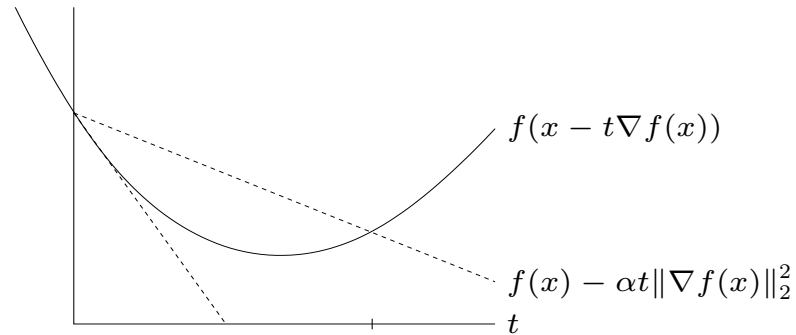
- fixed: t_k constant
- backtracking line search
- exact line search: minimize $f(x - t \nabla f(x))$ over t
- diminishing: $t_k \rightarrow 0$, $\sum_{k=1}^{\infty} t_k = \infty$

we will study fixed and backtracking line search

Backtracking line search

initialize t at some positive value \hat{t} (for example, $\hat{t} = 1$); take $t := \beta t$ until

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



- $0 < \beta < 1$; we will take $\alpha = 1/2$ (mostly to simplify proofs)
- variation: use $\hat{t} = t_{k-1}$ to initialize backtracking at iteration k

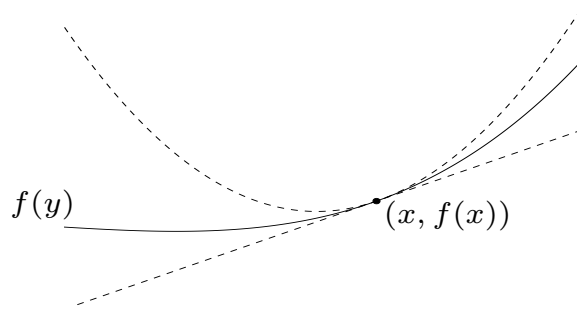
Assumptions

1. f has finite optimal value f^* , minimizer x^*
2. f is convex, $\text{dom } f = \mathbf{R}^n$
3. $\nabla f(x)$ is Lipschitz continuous with constant $L > 0$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

for twice differentiable functions, this means $\nabla^2 f(x) \preceq LI$ for all x

Upper and lower bound



- affine lower bound from convexity

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

- quadratic upper bound from Lipschitz property

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

proof of upper bound (define $v = y - x$)

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T v + \int_0^1 (\nabla f(x + tv) - \nabla f(x))^T v dt \\ &\leq f(x) + \nabla f(x)^T v + \int_0^1 \|\nabla f(x + tv) - \nabla f(x)\|_2 \|v\|_2 dt \\ &\leq f(x) + \nabla f(x)^T v + \int_0^1 Lt \|v\|_2^2 dt \\ &= f(x) + \nabla f(x)^T v + \frac{L}{2} \|v\|_2^2 \end{aligned}$$

Analysis for constant step size

from quadratic upper bound 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 $0 < t \leq 1/L$,

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

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

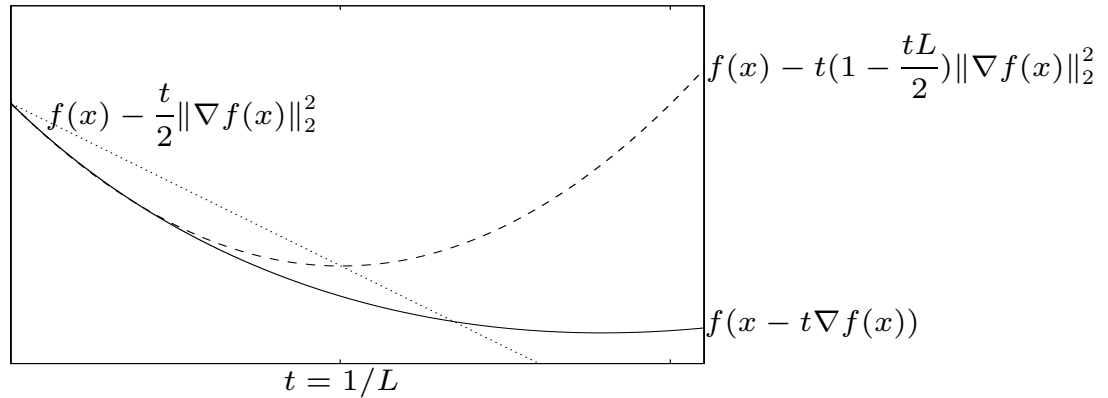
$$\begin{aligned} \sum_{i=1}^k (f(x^{(i)}) - f^*) &\leq \frac{1}{2t} \sum_{i=1}^k \left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right) \\ &= \frac{1}{2t} \left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2 \right) \\ &\leq \frac{1}{2t} \|x^{(0)} - x^*\|_2^2 \end{aligned}$$

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 $O(1/\epsilon)$

Analysis for backtracking line search



- line search (page 1–3) with $\alpha = 1/2$ gives step size

$$t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$$

- variation: take $t_0 > 0$, initialize line search at iteration k with $\hat{t} = t_{k-1}$

$$t_k \geq t_{\min} = \min\{t_0, \beta/L\}$$

convergence analysis

- from page 1–7:

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

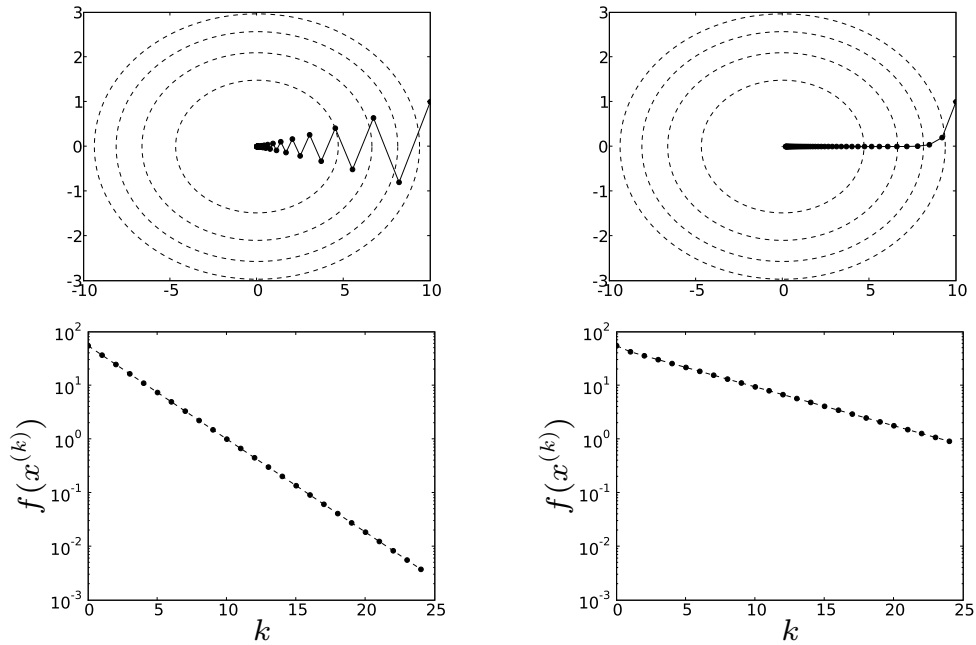
- add the upper bounds to get

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

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

Quadratic example

$f(x_1, x_2) = (x_1^2 + Lx_2^2)/2$; left: $t_k = 1.8/L$; right: $t_k = 0.8/L$



Gradient method

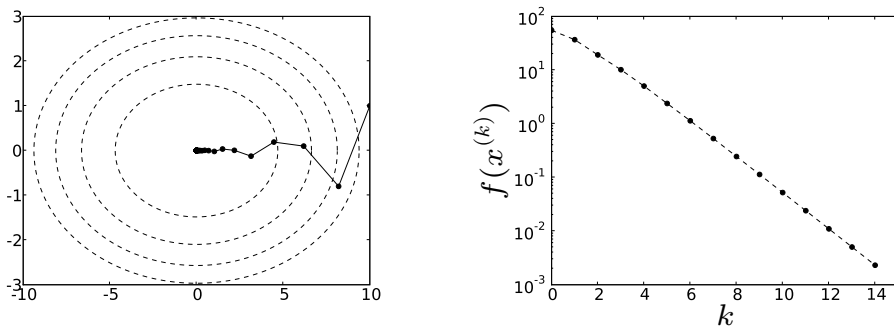
1-11

Two-step methods

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

- adds 'momentum' term to improve convergence
- examples: 'heavy ball' method (t_k, s_k fixed), conjugate gradients

quadratic example: $t_k = 1.8/L, s_k = 0.3$



Gradient method

1-12

Nesterov's gradient method

choose $x^{(0)}$ and set $y^{(0)} = x^{(0)}$

repeat for $k = 1, 2, \dots$

$$\begin{aligned}x^{(k)} &= y^{(k-1)} - t_k \nabla f(y^{(k-1)}) \\y^{(k)} &= x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)})\end{aligned}$$

- can use constant step t_k or backtracking (see later)
- not a descent method: $f(x^{(k+1)})$ can be greater than $f(x^{(k)})$
- momentum term added to x before calculating gradient step
- published in 1983; other variants have appeared since then

Analysis for constant step size

notation: define $\theta_0 = 1$, $v^{(0)} = x^{(0)}$,

$$v^{(k)} = \frac{k+1}{2}x^{(k)} - \frac{k-1}{2}x^{(k-1)}, \quad \theta_k = \frac{2}{k+2} \quad (k \geq 1)$$

- gives expression of $y^{(k)}$ as a convex combination of $v^{(k)}$ and $x^{(k)}$:

$$y^{(k)} = \theta_k v^{(k)} + (1 - \theta_k)x^{(k)}$$

- satisfies $v^{(k+1)} = v^{(k)} + (1/\theta_k)(x^{(k+1)} - y^{(k)})$

improvement in one iteration (with fixed step size $t \in (0, 1/L]$)

with $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $y = y^{(i-1)}$, $v = v^{(i-1)}$, $v^+ = v^{(i)}$, $\theta = \theta_{i-1}$

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

therefore for $i \geq 1$

$$\frac{1}{\theta_{i-1}^2} (f(x^{(i)}) - f^*) + \frac{1}{2t} \|v^{(i)} - x^*\|_2^2 \leq \frac{1 - \theta_{i-1}}{\theta_{i-1}^2} (f(x^{(i-1)}) - f^*) + \frac{1}{2t} \|v^{(i-1)} - x^*\|_2^2$$

Gradient method

1-15

note that

$$\frac{1 - \theta_{i-1}}{\theta_{i-1}^2} = \frac{(i+1)(i-1)}{4} \leq \frac{i^2}{4} = \frac{1}{\theta_{i-2}^2} \quad (i \geq 2)$$

combine the bounds for $i = 1, \dots, k$, to get

$$\frac{1}{\theta_{k-1}^2} (f(x^{(k)}) - f^*) + \frac{1}{2t} \|v^{(k)} - x^*\|_2^2 \leq \frac{1 - \theta_0}{\theta_0^2} (f(x^{(0)}) - f^*) + \frac{1}{2t} \|v^{(0)} - x^*\|_2^2$$

since $\theta_0 = 1$ and $v^{(0)} = x^{(0)}$ this reduces to

$$f(x^{(k)}) - f^* \leq \frac{\theta_{k-1}^2}{2t} \|x^{(0)} - x^*\|_2^2 = \frac{2}{(k+1)^2 t} \|x^{(0)} - x^*\|_2^2$$

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

Gradient method

1-16

Analysis for backtracking line search

use t_{k-1} to initialize line search at iteration k

- implies $t_k \leq t_{k-1}$ and $t_k \geq t_{\min} = \min\{t_0, \beta/L\}$ (see page 1-9)
- improvement in one iteration (page 1-15 and 1-16):

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

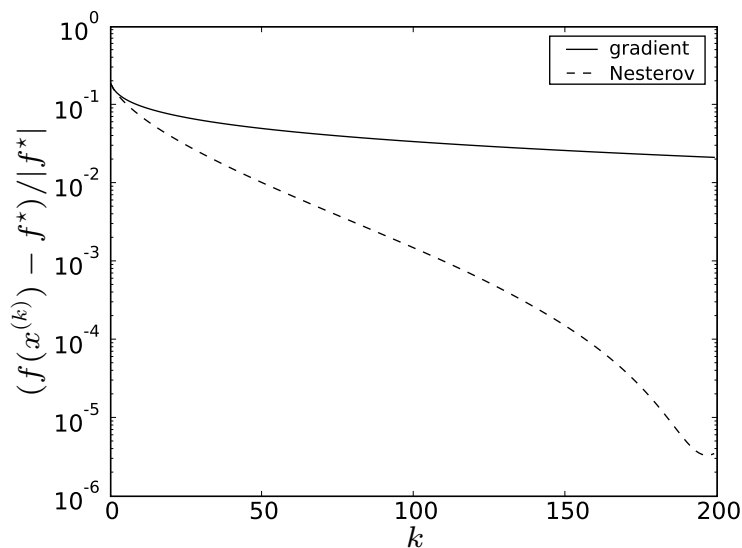
conclusion. same $1/k^2$ bound as with constant step size:

$$f(x^{(k)}) - f^* \leq \frac{\theta_{k-1}^2}{2t_k} \|x^{(0)} - x^*\|_2^2 = \frac{2}{(k+1)^2 t_{\min}} \|x^{(0)} - x^*\|_2^2$$

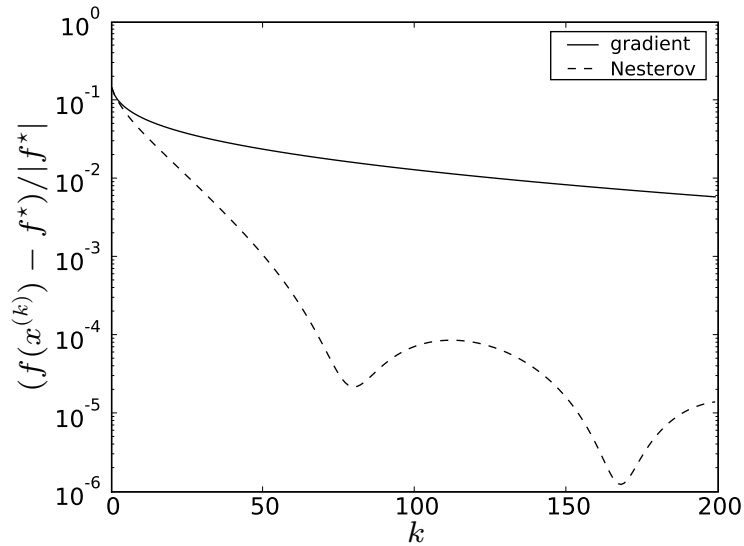
Example

$$\text{minimize } \log \sum_{i=1}^m \exp(a_i^T x + b_i)$$

randomly generated data with $m = 2000$, $n = 1000$, fixed step size



another instance



Optimality of Nesterov's method

define a **first order method** as any iterative algorithm that selects $x^{(k)}$ in

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

optimality

no first-order method improves the $1/k^2$ convergence rate (uniformly, over all convex functions with Lipschitz continuous gradients)

Example

$$f(x) = \frac{1}{2}(x_1^2 + \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 + x_n^2) - x_1 = \frac{1}{2}x^T Ax - x_1$$

with

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{bmatrix}$$

satisfies $0 \prec A \preceq 4I$ for all n

optimal solution

$$x_i^* = 1 - \frac{i}{n+1}, \quad i = 1, \dots, n, \quad f^* = -\frac{n}{2(n+1)}$$

$$\|x^*\|_2^2 = \frac{1}{(n+1)^2}(n^2 + (n-1)^2 + \cdots + 2^2 + 1) \leq \frac{n+1}{3}$$

first-order algorithm started at $x^{(0)} = 0$:

$$x^{(k)} \in x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \dots, \nabla f(x^{(k-1)})\} \subseteq \text{span}\{e_1, \dots, e_k\}$$

therefore

$$f(x^{(k)}) \geq \inf_{x_{k+1}=\dots=x_n=0} f(x) = \frac{-k}{2(k+1)}$$

conclusion: if $n = 2k + 1$,

$$\frac{f(x^{(k)}) - f^*}{\|x^{(0)} - x^*\|_2^2} \geq \frac{3}{n+1} \left(\frac{n}{2(n+1)} - \frac{k}{2(k+1)} \right) = \frac{3}{8(k+1)^2}$$

for every k , $x^{(0)}$, L there exists a convex quadratic function f with

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

if $x^{(k)}$ is constructed by any first order method

References

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004)
 - section 2.2. on optimal methods
 - page 80: basically the algorithm of page 1–13 of this lecture
 - section 2.1.2: the example of page 1–21
- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, to appear in *SIAM J. Imaging Sciences*
 - the convergence analysis on page 1–7 and 1–15
- B. T. Polyak, *Introduction to Optimization* (1987)
 - section 3.2.1 on multistep gradient methods