The subgradient method

Acknowledgement: slides are based on Prof. Lieven Vandenberghes.

- subgradient method
- convergence analysis
- optimal step size when f^* is known
- alternating projections
- optimality

Subgradient method

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

$$x^{(k)} = x^{(k-1)} - t_k g^{(k-1)}, \quad k = 1, 2, \ldots \qquad \forall g^{(k-1)} \in \partial f(x^{k-1})$$

$g^{(k-1)}$ is any subgradient of f at $x^{(k-1)}$

$$\begin{cases} x^{k+1} = x^k + t_k \underset{\sim}{g^k} \\ \|\nabla f(x^k) - g^k\| \le \underset{\sim}{\varepsilon_k} \end{cases}$$

Step size rules

- fixed step: t_k constant
- fixed length: $t_k\|g^{(k-1)}\|_2 = \|x^{(k)} - x^{(k-1)}\|_2 \overset{=s}{}$ is constant
- diminishing: $t_k \to 0$, $\sum_{k=1}^{\infty} t_k = \infty$ (Learning rate η_t)

Inexact Gradient descent.

Assumptions

- f has finite optimal value $f^\star$, minimizer $x^\star$

- f is convex, $\mathbf{dom}\, f = \mathbf{R}^n$

(∇f)

- f is Lipschitz continuous with constant $G > 0$:

$$|f(x) - f(y)| \le G\|x - y\|_2 \qquad \forall x, y$$

$\Updownarrow$

this is equivalent to $\|g\|_2 \le G$ for all x and $g \in \partial f(x)$ (see next page)

$\sup\{\|y\| \;\; y \in \partial f(x), \forall x\} \le G.$

Proof.

- assume $\|g\|_2 \leq G$ for all subgradients; choose $g_y \in \partial f(y)$, $g_x \in \partial f(x)$:

$$g_x^T(x-y) \geq f(x) - f(y) \geq g_y^T(x-y)$$

x. y

by the Cauchy-Schwarz inequality

$$G\|x-y\|_2 \geq f(x) - f(y) \geq -G\|x-y\|_2$$

- assume $\|g\|_2 > G$ for some $g \in \partial f(x)$; take $y = x + g/\|g\|_2$: $\Rightarrow \|y-x\| = 1$

$|f(y) - f(x)| \leq G \|y-x\|$

$$\begin{aligned} f(y) &\geq f(x) + g^T(y-x) \\ &= f(x) + \|g\|_2 \\ &> f(x) + G \end{aligned}$$

$f(y) - f(x) > G \cdot \|y-x\|$

$2\sum_{i=1}^{k} t_i (f_{best}^{(k)} - f^*) \le 2\sum_{i=1}^{k} t_i (f(x_i) - f^*)$

Analysis

$\le \sum_{i=1}^{k} \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 + \sum_{i=1}^{k} t_i \|g^i\|_2^2$

$\|x_0 - X^*\|_2^2 - \|X_{k+1} - X^*\|_2^2$

- the subgradient method is not a descent method
- the key quantity in the analysis is the distance to the optimal set

$\Rightarrow 2t(f(x) - f^*) \le \|x - x^*\|_2^2 - \|x^+ - X^*\|_2^2 + t^2\|g\|_2^2$

with $x^+ = x^{(i)}$, $x = x^{(i-1)}$, $g = g^{(i-1)}$, $t = t_i$:

$$\begin{aligned}
\|x^+ - x^\star\|_2^2 &= \|x - tg - x^\star\|_2^2 \\
&= \|x - x^\star\|_2^2 - 2tg^T(x - x^\star) + t^2\|g\|_2^2 \\
&\le \|x - x^\star\|_2^2 - 2t\left(f(x) - f^\star\right) + t^2\|g\|_2^2
\end{aligned}$$

$f^{k}_{best} = \min\{f^{k-1}_{best}, f^k\}$

combine inequalities for $i = 1, \ldots, k$, and define $f_{\text{best}}^{(k)} = \min_{0 \le i < k} f(x^{(i)})$:

$$\begin{aligned}
2\left(\sum_{i=1}^{k} t_i\right)\left(f_{\text{best}}^{(k)} - f^\star\right) &\le \|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2 + \sum_{i=1}^{k} t_i^2 \|g^{(i-1)}\|_2^2 \\
&\le \|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^{k} t_i^2 \|g^{(i-1)}\|_2^2
\end{aligned}$$

$\Rightarrow f_{best}^{(k)} - f^* \le \dfrac{\|X^0 - X^*\|_2^2 + \sum_{i=1}^{k} t_i^2 G^2}{2\sum_{i=1}^{k} t_i} \to 0\ ?$

Fixed step size: $t_i = t$

$$f_{\text{best}}^{(k)} - f^\star \le \frac{\|x^{(0)} - x^\star\|_2^2}{\underbrace{2kt}_{\to 0}} + \boxed{\frac{G^2 t}{2}} \not\to 0$$

$f^* + \frac{G^2 t}{2}$

- does not guarantee convergence of $f_{\text{best}}^{(k)}$
- for large k, $f_{\text{best}}^{(k)}$ is approximately $G^2t/2$-suboptimal $O(G^2)$

f^*

Fixed step length: $t_i = s/\|g^{(i-1)}\|_2$ $\quad t_i \|g^{i-1}\|_2 = s$

$$f_{\text{best}}^{(k)} - f^\star \le \frac{G\|x^{(0)} - x^\star\|_2^2}{2ks} + \frac{Gs}{2}$$

- does not guarantee convergence of $f_{\text{best}}^{(k)}$
- for large k, $f_{\text{best}}^{(k)}$ is approximately $Gs/2$-suboptimal $O(G)$

$$x^{k+1} = x^k - s \frac{g^k}{\|g^k\|_2}$$

Diminishing step size: $\underbrace{t_i \to 0,\ \sum_{i=1}^{\infty} t_i = \infty}$ $O\left(\frac{1}{k}\right)$ $O\left(\frac{1}{\sqrt{k}}\right)$

$t_k \sim O\left(\frac{1}{k}\right)$

$$f_{\text{best}}^{(k)} - f^\star \le \frac{\underbrace{\|x^{(0)} - x^\star\|_2^2}_{\Delta} {}_{\le R^2} + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i} \longrightarrow 0$$

$t_k = \frac{3}{k}$

$\frac{3000}{k}$

Δ

can show that $\underbrace{(\sum_{i=1}^{k} t_i^2)/(\sum_{i=1}^{k} t_i) \to 0}$; hence, $f_{\text{best}}^{(k)}$ converges to $f^\star$

Proof: $\forall \varepsilon > 0$, $\exists N_1$ s.t. $t_i \le \frac{\varepsilon}{G^2}$, $\forall i > N_1$

$\exists N_2$ s.t. $\sum_{i=1}^{N_2} t_i \ge \frac{1}{\varepsilon}\left(R^2 + G^2 \underbrace{\sum_{i=1}^{N_1} t_i^2}_{\Delta}\right)$.

Let $N = \max\{N_1, N_2\}$, For $k > N$.

$$\frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i} = \frac{R^2 + G^2 \sum_{i=1}^{N_1} t_i^2}{2 \sum_{i=1}^{k} t_i} + \frac{G^2 \sum_{i=N_1+1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i}$$

$t_i^2 \le t_i \left(\frac{\varepsilon}{G^2}\right)$

$\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$

Example: 1-norm minimization

$f(x) = h(Ax-b)$, $h(x) = \|x\|_1$

$$\text{minimize } \|Ax - b\|_1$$

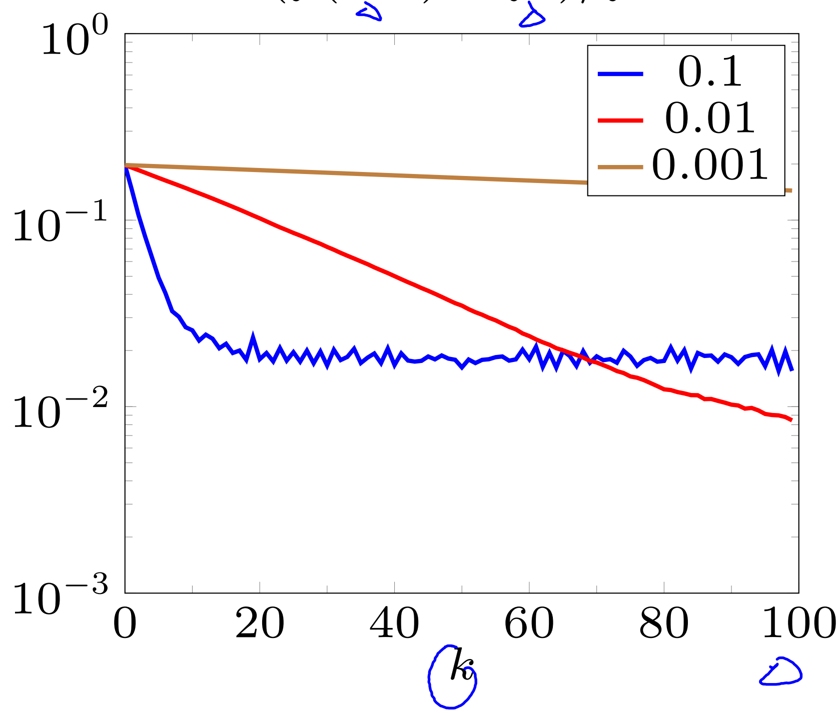
$\Rightarrow \partial f(x) = A^T \partial h(Ax-b)$

- subgradient is given by $A^T \mathbf{sign}(Ax - b)$
- example with $A \in \mathbf{R}^{500 \times 100}$, $b \in \mathbf{R}^{500}$

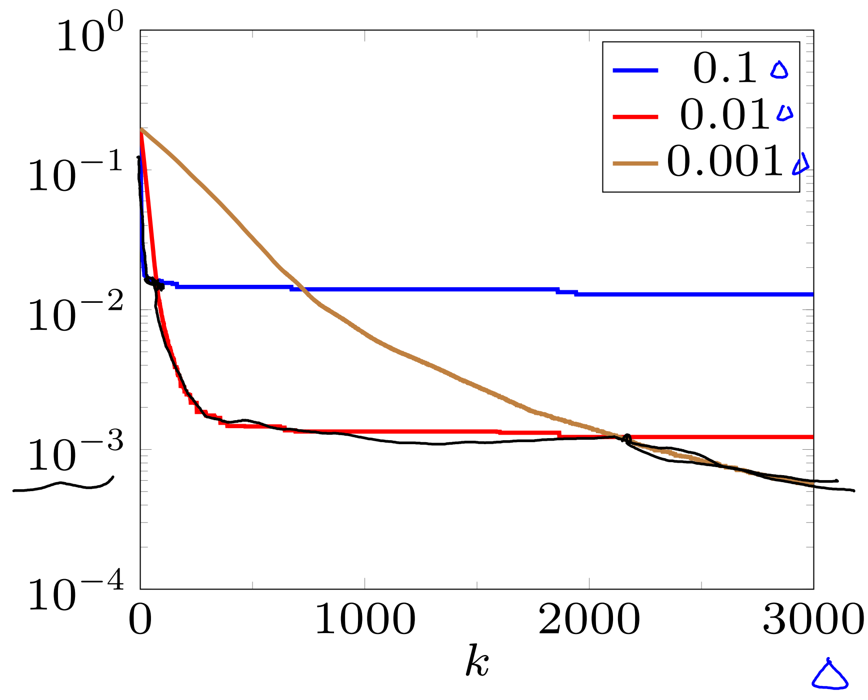
$\frac{\text{G.S.}}{2}$

Fixed steplength $t_k = s/\|g^{(k-1)}\|_2$ for $s = 0.1, 0.01, 0.001$

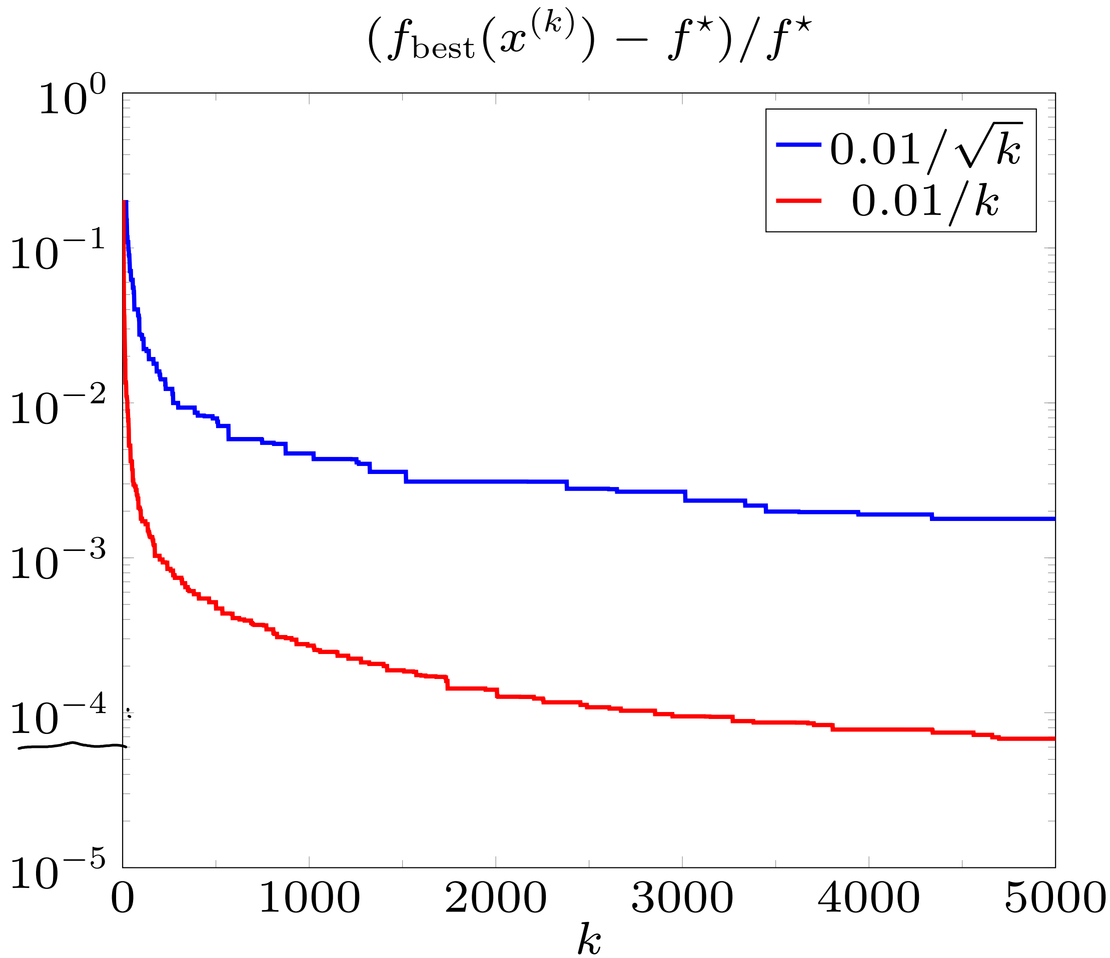
$(f(x^{(k)}) - f^\star)/f^\star$



$(f_{\text{best}}(x^{(k)}) - f^\star)/f^\star$



Diminishing step size: $t_k = 0.01/\sqrt{k}$ and $t_k = 0.01/k$



Optimal step size for fixed number of iterations

from page 5-5: if $s_i = t_i \|g^{(i-1)}\|_2$ and $\|x^{(0)} - x^\star\|_2 \leq R$:

$$\min_{s>0} \frac{R^2}{2ks/G} + \frac{s}{2/G}$$

$$f_{\mathrm{best}}^{(k)} - f^\star \leq \frac{R^2 + \sum_{i=1}^{k} s_i^2}{2\sum_{i=1}^{k} s_i/G}$$

$$\Longleftrightarrow \quad \frac{R^2}{2ks} = \frac{s}{2}$$

$$\Longrightarrow \quad s = \frac{R}{\sqrt{k}}$$

- for given k, bound is minimized by fixed step length $s_i = s = R/\sqrt{k}$

GD: $O\left(\frac{1}{k}\right)$

- resulting bound after k steps is

$$\boxed{f_{\mathrm{best}}^{(k)} - f^\star \leq \frac{GR}{\sqrt{k}}}$$

- guarantees accuracy $f_{\mathrm{best}}^{(k)} - f^\star \leq \epsilon$ in $k = O(1/\epsilon^2)$ iterations

Optimal step size when $f^\star$ is known

- right-hand side in first inequality of page 5-5 is minimized by

$$t_i = \frac{f(x^{(i-1)}) - f^\star}{\|g^{(i-1)}\|_2^2}$$

Quadratic w.r.t. t_i

$$\|x^i - x^*\|_2^2 \le \underbrace{\|x^{i-1} - x^*\|_2^2 - 2t_i(f^{i-1} - f^*) + t_i^2 \|g^{i}\|^2}_{\min}$$

- optimized bound is

$$\frac{\left(f(x^{(i-1)}) - f^\star\right)^2}{\|g^{(i-1)}\|_2^2} \le \|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2$$

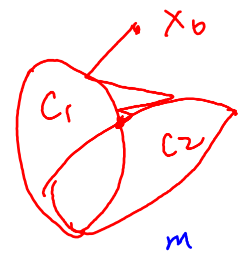
- applying recursively (with $\|x^{(0)} - x^\star\|_2 \le R$ and $\|g^{(i)}\|_2 \le G$) gives

$$f_{\text{best}}^{(k)} - f^\star \le \frac{GR}{\sqrt{k}}$$

$O\left(\frac{1}{\sqrt{k}}\right) \to$ best ?

Exercise: find point in intersection of convex sets

Alternating projection.



find a point in the intersection of m closed convex sets $C_1, \ldots, C_m$:

$$\text{minimize} \quad f(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

Find $x \in \bigcap_{j=1}^{m} C_j$

where $f_j(x) = \inf_{y \in C_j} \|x - y\|_2$ is Euclidean distance of x to C_j

≥ 0

- $f^\star = 0$ if the intersection is nonempty

$f_j(\hat{x}) > 0$

- (from p. 4-14): $g \in \partial f(\hat{x})$ if $g \in \partial f_j(\hat{x})$ and C_j is farthest set from $\hat{x}$
- (from p. 4-20) subgradient $g \in \partial f_j(\hat{x})$ follows from projection $P_j(\hat{x})$ on C_j:

$$g = 0 \quad (\text{if } \hat{x} \in C_j), \qquad g = \frac{1}{\|\hat{x} - P_j(\hat{x})\|_2}(\hat{x} - P_j(\hat{x})) \quad (\text{if } \hat{x} \notin C_j)$$

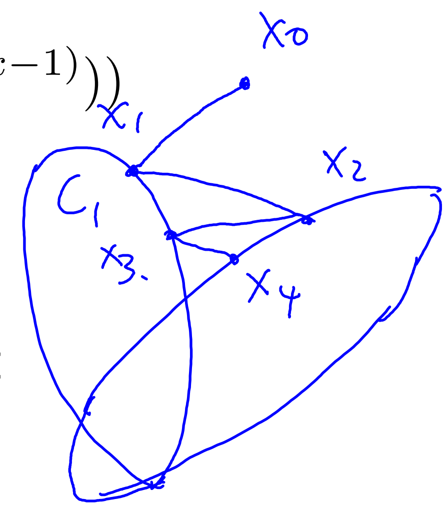
note that $\|g\|_2 = 1$ if $\hat{x} \notin C_j$

Subgradient method

$t_i = \frac{f^{i-1} - f^* = 0}{\|g^{i-1}\|_2^2}$

- optimal step size (page 5-11) for $f^\star = 0$ and $\|g^{(i-1)}\|_2 = 1$ is $t_i = f(x^{(i-1)})$
- at iteration k, find farthest set C_j (with $f(x^{(k-1)}) = f_j(x^{(k-1)})$), and take

$$
\begin{aligned}
x^{(k)} &= x^{(k-1)} - \frac{f(x^{(k-1)})}{f_j(x^{(k-1)})}(x^{(k-1)} - P_j(x^{(k-1)})) \\
&= P_j(x^{(k-1)})
\end{aligned}
$$



at each step, we project the current point onto the farthest set

- a version of the *alternating projections* algorithm $O\left(\frac{1}{\sqrt{k}}\right)$
- for $m = 2$, projections alternate onto one set, then the other
- later, we will see faster versions of this that are almost as simple

$\bigcap_{i=1}^{m} \{x \mid f_i(x) \le 0\}$

(1) $A_i x + b_i$

$\{f(x) \le 0\}$ feasible.

Optimality of the subgradient method

can the $f_{\text{best}}^{(k)} - f^\star \leq GR/\sqrt{k}$ bound on page 5-10 be improved?

Problem class

- f is convex, with a minimizer $x^\star$ ✓
- we know a starting point $x^{(0)}$ with $\|x^{(0)} - x^\star\|_2 \leq R$ ✓
- we know the Lipschitz constant G of f on $\{x \mid \|x - x^{(0)}\|_2 \leq R\}$
- f is defined by an oracle: given x, oracle returns $f(x)$ and a subgradient

Algorithm class: k iterations of any method that chooses $x^{(i)}$ in

$$x^{(0)} + \operatorname{span}\{g^{(0)}, g^{(1)}, \ldots, g^{(i-1)}\}$$

$$x^{k+1} = x^k + t_k g^k$$

$$= x^{k-1} + t_{k-1} g^{k-1} + t_k g^k.$$

Test problem and oracle

$= \max_{1\le i\le k} \{e_i^T x\} + \frac{1}{2}\|x\|_2^2$

$$f(x) = \max_{i=1,\ldots,k} x_i + \frac{1}{2}\|x\|_2^2, \qquad x^{(0)} = 0$$

$0 \in \partial f(x) = \mathrm{conv}\{e_j \mid j \in I(x)\} + x$

- solution: $x^\star = -\frac{1}{k}(\underbrace{1,\ldots,1}_{k}, \underbrace{0,\ldots,0}_{n-k})$ and $f^\star = -\frac{1}{2k}$

$0 \in \partial f(x^*)$

- $R = \|x^{(0)} - x^\star\|_2 = 1/\sqrt{k}$ and $G = 1 + 1/\sqrt{k}$
- oracle returns subgradient $e_{\hat{\jmath}} + x$ where $\hat{\jmath} = \min\{j \mid x_j = \max_{i=1,\ldots,k} x_i\}$

Iteration: for $i = 0, \ldots, k-1$, entries $x_{i+1}^{(i)}, \ldots, x_k^{(i)}$ are zero; therefore

$$f_{\text{best}}^{(k)} - f^\star = \min_{i<k} f(x^{(i)}) - f^\star \geq -f^\star = \frac{GR}{2(1+\sqrt{k})}$$

$f(x^{(i)}) \geq 0$

Conclusion: $O(1/\sqrt{k})$ bound cannot be improved

$g^{(0)} = e_1 + x^{(0)} = e_1 \Rightarrow t_1 = \frac{f^0 - f^*}{\|g^0\|_2^2}$

$\Rightarrow x^1 = x^{(0)} - t_1 e_1 \Rightarrow x^2 = x^{(1)} - t_2(e_1 + \alpha e_2)$

Summary: subgradient method

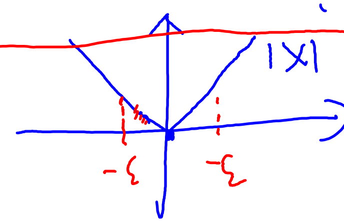
$\boxed{\|x^k - x^{k-1}\| < \varepsilon \quad (X)}$

- handles general nondifferentiable convex problem

if $x \neq 0$ $\partial |x_i| = \mathrm{sgn}(x_i)$

$\Rightarrow \|\partial x_i\| = 1$

- often leads to very simple algorithms
- convergence can be very slow $\longrightarrow$ Acceleration!!!
- no good stopping criterion $(\|\nabla f(x)\| \le \varepsilon)$



pratical stopping creterion!!!

- theoretical complexity: $O(1/\epsilon^2)$ iterations to find ϵ-suboptimal point
- an 'optimal' 1st-order method: $O(1/\epsilon^2)$ bound cannot be improved

References

- S. Boyd, lecture notes and slides for EE364b, Convex Optimization II
- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004)

  §3.2.1 with the example on page 5-15 of this lecture