CS 2601 Linear and Convex Optimization

6. Gradient descent (part 1)

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

Fall 2022

Unconstrained optimization problems

Consider an unconstrained, smooth convex optimization problem

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x})$$

where f is convex and differentiable on $\mathbb{R}^n$.

The optimal solution satisfies the first-order optimality condition

$$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$$

In some rare cases, this yields closed-form solutions, e.g.

$$\min_{\boldsymbol{w}} \; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

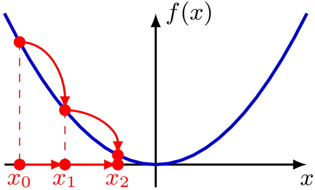
has closed-form solution

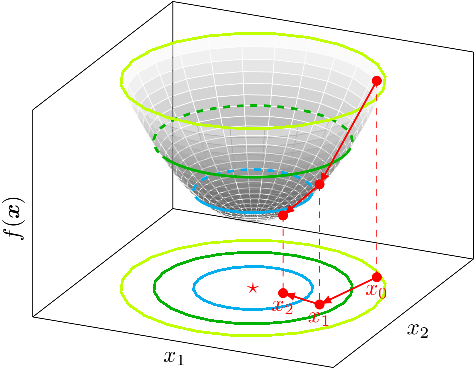
$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

But in most cases we need numerical algorithms.

Descent method

choose initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$
repeat
  choose descent direction $\boldsymbol{d}_k \in \mathbb{R}^n$ and step size $t_k > 0$
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{d}_k \quad$ s.t. $\quad f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$
until stopping criterion is satisfied





Questions

- How to choose $\boldsymbol{d}_k$ and t_k?
- Does $\boldsymbol{x}_k$ converge to $\boldsymbol{x}^*$?

Descent direction

$\boldsymbol{d}_k$ is a descent direction at $\boldsymbol{x}_k$ if for all small enough $t > 0$

$$g(t) \triangleq f(\boldsymbol{x}_k + t\boldsymbol{d}_k) < f(\boldsymbol{x}_k) = g(0)$$

For differentiable f (not necessarily convex),

- if $\boldsymbol{d}_k$ is a descent direction, then $g'(0) = \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) \leq 0$;
- if $g'(0) = \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) < 0$, then $\boldsymbol{d}_k$ is a descent direction.

For convex f, by the first-order condition for convexity,

$$f(\boldsymbol{x}_k) > f(\boldsymbol{x}_k + t\boldsymbol{d}_k) \geq f(\boldsymbol{x}_k) + t\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k).$$

$\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) < 0$ is also necessary for $\boldsymbol{d}_k$ to be a descent direction.

For convex differentiable f,

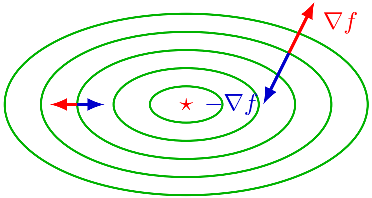
$$\boldsymbol{d}_k \text{ is a descent direction} \iff \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) < 0$$

Gradient descent

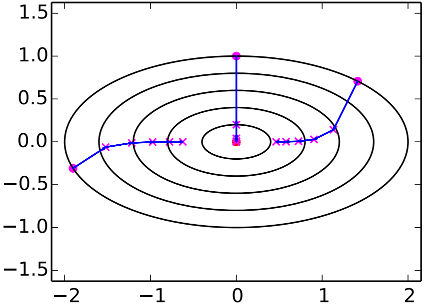
Choose $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$, $\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) = -\|\nabla f(\boldsymbol{x}_k)\|_2^2 < 0$ unless $\nabla f(\boldsymbol{x}_k) = 0$.

Updating rule

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)}$$



level curves of $f(x_1, x_2) = \frac{x_1^2}{4} + x_2^2$



Question. What happens if $\nabla f(\boldsymbol{x}_k) = \mathbf{0}$?

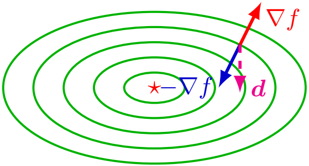
Max-rate descending direction

$-\nabla f(\boldsymbol{x}_k)$ is the direction of fastest rate of decrease of f at $\boldsymbol{x}_k$

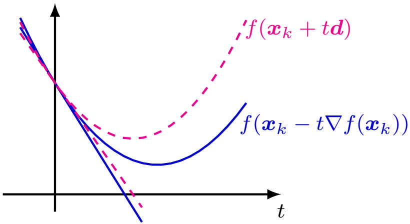
- If $\|\boldsymbol{d}_k\|_2 = 1$,

$$\lim_{t \downarrow 0} \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + t\boldsymbol{d}_k)}{t} = -\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) \leq \|\nabla f(\boldsymbol{x}_k)\|_2$$

with equality iff $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)/\|\nabla f(\boldsymbol{x}_k)\|_2$



level curves of $f(x_1, x_2) = \frac{x_1^2}{4} + x_2^2$



Gradient descent algorithm

initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in \mathbb{R}^n$
while $\|\nabla f(\boldsymbol{x})\| > \delta$ **do**
    $\boldsymbol{x} \leftarrow \boldsymbol{x} - t\nabla f(\boldsymbol{x})$
end while
return $\boldsymbol{x}$

Step size (aka learning rate in machine learning)

- the above algorithm uses constant step size t for all iterations
- there are other methods for choosing t for each iteration, e.g. exact line search, backtracking line search

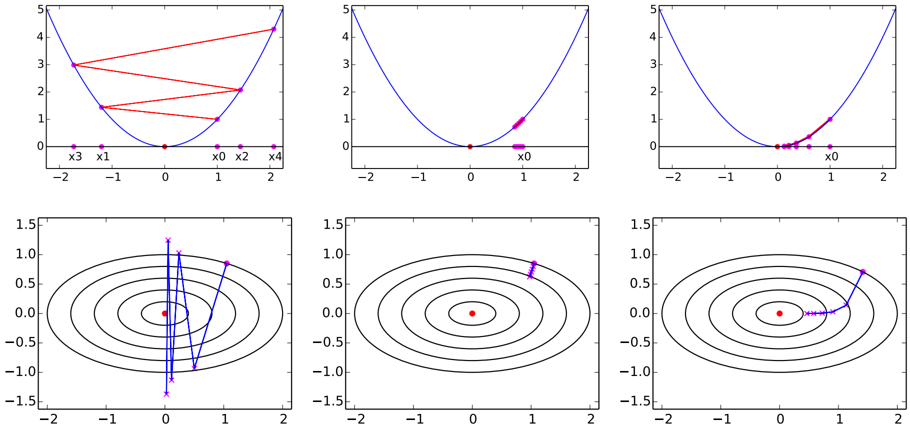
Stopping criterion

- ideally, stop if $\nabla f(\boldsymbol{x}) = 0$ (optimality condition), but impractical
- more practical: stop when $\|\nabla f(\boldsymbol{x})\| \leq \delta$ for some small δ
- other criteria: $|f(\boldsymbol{x}_{\text{new}}) - f(\boldsymbol{x}_{\text{old}})| \leq \delta$, $\frac{|f(\boldsymbol{x}_{\text{new}}) - f(\boldsymbol{x}_{\text{old}})|}{|f(\boldsymbol{x}_{\text{old}})|} \leq \delta$, ...
- in practice, also stop if maximum # of iterations is reached

Large vs. small step size

Consider constant step size. How large should the step size be?

- Too large: may oscillate and diverge
- Too small: may be too slow
- “Just right”: fast convergence



1D example

Consider $f(x) = \frac{1}{2}ax^2$, where $a > 0$.

- gradient step

$$x_{k+1} = x_k - tf'(x_k) = (1 - at)x_k$$

- descent condition

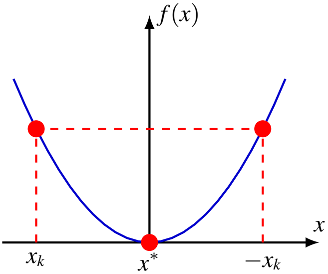
$$f(x_{k+1}) < f(x_k) \iff |1 - at| < 1 \iff 0 < t < \frac{2}{a}$$

- $x_k = (1 - at)^k x_0 \to x^* = 0$ geometrically for such t

Note f satisfies

- $|f'(x) - f'(y)| = a|x - y|$
- $f''(x) = a$

f' is so-called Lipschitz continuous and t is roughly the order of $\frac{1}{a}$.



Lipschitz continuity

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous with Lipschitz constant $L > 0$, or simply L-Lipschitz, if

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Note. Lipschitz continuity can be defined with respect to any norms. But we will assume the norms in the above definition are the 2-norms in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, unless stated otherwise.

Note. Lipschitz continuity implies uniform continuity.

Example. $f(x) = ax$ is $|a|$-Lipschitz, $|f(x) - f(y)| = |a| \cdot |x - y|$

Example. $f(x) = |x|$ is 1-Lipschitz, $|f(x) - f(y)| = \big||x| - |y|\big| \leq |x - y|$

Example. $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ is $\|\boldsymbol{a}\|$-Lipschitz, $|\boldsymbol{a}^T\boldsymbol{x} - \boldsymbol{a}^T\boldsymbol{y}| \leq \|\boldsymbol{a}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$ by the Cauchy-Schwarz inequality.

Lipschitz continuity (cont'd)

Example. Let $\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. $f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} = (x_1, 2x_2)^T$ is 2-Lipschitz.

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) = (x_1 - y_1, 2x_2 - 2y_2)^T = (d_1, 2d_2)^T$$

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| = \sqrt{d_1^2 + 4d_2^2} \le 2\sqrt{d_1^2 + d_2^2} = 2\|\boldsymbol{x} - \boldsymbol{y}\|$$

More generally, $f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x}$ with $\boldsymbol{Q} \succeq \boldsymbol{O}$ is $\lambda_{\max}(\boldsymbol{Q})$-Lipschitz, where $\lambda_{\max}(\boldsymbol{Q})$ is the largest eigenvalue of $\boldsymbol{Q}$[1].

Proof. Let $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{y}$. By slide 32 of §2,

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| = \|\boldsymbol{Q}\boldsymbol{d}\| = \sqrt{\boldsymbol{d}^T\boldsymbol{Q}^2\boldsymbol{d}} \le \sqrt{\lambda_{\max}(\boldsymbol{Q}^2)\|\boldsymbol{d}\|^2} = \lambda_{\max}(\boldsymbol{Q})\|\boldsymbol{x} - \boldsymbol{y}\|$$

The last equality uses the fact $\lambda_{\max}(\boldsymbol{Q}^2) = \lambda_{\max}^2(\boldsymbol{Q})$.

---

[1] For general $\boldsymbol{Q}$, $f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x}$ is $\sigma_{\max}(\boldsymbol{Q})$-Lipschitz, where $\sigma_{\max}(\boldsymbol{Q}) = \sqrt{\lambda_{\max}(\boldsymbol{Q}^T\boldsymbol{Q})}$ is the largest singular value of $\boldsymbol{Q}$.

L-smoothness

A function is L-smooth if it is differentiable and its gradient is L-Lipschitz, i.e.

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Note. L upper bounds the rate of change of ∇f

Example. $f(x) = \frac{1}{2}ax^2$ is $|a|$-smooth, since $f'(x) = ax$ is $|a|$-Lipschitz

Example. $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ with $\boldsymbol{Q} \succeq \boldsymbol{O}$ is $\lambda_{\max}(\boldsymbol{Q})$-smooth, since $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x}$ is $\lambda_{\max}(\boldsymbol{Q})$-Lipschitz.

With $\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, we obtain $f(\boldsymbol{x}) = \frac{1}{2}x_1^2 + x_2^2$ is 2-smooth.

Lemma. A twice continuously differentiable convex $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth iff $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, meaning $L\boldsymbol{I} - \nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{O}$, or equivalently $\lambda_{\max}(\nabla^2 f(\boldsymbol{x})) \leq L$.

如果不是凸函数这要加个绝对值

Appendix: Second-order condition for L-smoothness

Lemma. A twice continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth iff for any $\boldsymbol{x}$, $-L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, or equivalently $|\lambda| \leq L$ for all eigenvalues λ of $\nabla^2 f(\boldsymbol{x})$.

Proof. "$\Leftarrow$". Assume $-L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$ for all $\boldsymbol{x}$. By the Mean Value Theorem and slide 30 of §2,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| = \|\nabla^2 f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

"$\Rightarrow$". Assume f is L-smooth. Let $\boldsymbol{d}$ be an eigenvector of $\nabla^2 f(\boldsymbol{x})$ with associated eigenvalue λ. By L-smoothness,

$$\|\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x})\| \leq L\|t\boldsymbol{d}\| = tL\|\boldsymbol{d}\|$$

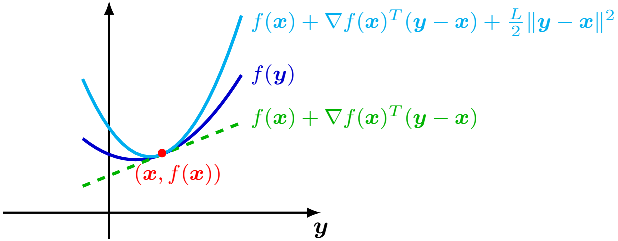
Dividing both sides by t and letting $t \to 0$,

$$|\lambda| \cdot \|\boldsymbol{d}\| = \|\nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| \leq L\|\boldsymbol{d}\| \implies |\lambda| \leq L$$

Quadratic upper bound

Lemma. If f is L-smooth, then

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$



Note. The upper bound does not assume the convexity of f.

If $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, this is intuitive from the second-order Taylor expansion

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^T\nabla^2 f(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{x})$$

for some $\boldsymbol{z}$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$. (Check $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$)

Proof

First prove the 1D case. Let $g(t)$ be L_g-smooth, $|g'(t) - g'(s)| \leq L_g|t - s|$.

$$\begin{aligned} g(1) &= g(0) + \int_0^1 g'(t)dt \\ &= g(0) + g'(0) + \int_0^1 [g'(t) - g'(0)]dt \\ &\leq g(0) + g'(0) + \int_0^1 L_g t dt \quad \text{since } |g'(t) - g'(0)| \leq L_g t \\ &= g(0) + g'(0) + \frac{1}{2}L_g \end{aligned}$$

For the general case, apply the above to $g(t) = f(\boldsymbol{x} + t\boldsymbol{d})$ with $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{x}$ and $L_g = L\|\boldsymbol{d}\|^2$. By the Cauchy-Schwarz inequality

$$\begin{aligned} |g'(t) - g'(s)| &= \left|[\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x} + s\boldsymbol{d})]^T \boldsymbol{d}\right| \\ &\leq \|\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x} + s\boldsymbol{d})\| \cdot \|\boldsymbol{d}\| \quad \text{Cauchy-Schwarz} \\ &\leq (t - s)L\|\boldsymbol{d}\|^2 \quad f \text{ is } L\text{-smooth} \end{aligned}$$

Consequence of quadratic upper bound

For L-smooth f, the sequence $\{\boldsymbol{x}_k\}$ produced by gradient descent satisfies

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2$$

Proof. Plugging in $\boldsymbol{x} = \boldsymbol{x}_k$ and $\boldsymbol{y} = \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$ in the quadratic upper bound,

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - t\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{L}{2}t^2\|\nabla f(\boldsymbol{x}_k)\|^2 \\ &= f(\boldsymbol{x}_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 \end{aligned}$$

Note. If $\nabla f(\boldsymbol{x}_k) \neq 0$ and $0 < t < \frac{2}{L}$, then $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$, so gradient descent with step size $t \in (0, 2/L)$ is indeed a descent method.

Note. We can lower bound the decrease in function value in each step. In particular, for $0 < t \leq \frac{1}{L}$,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{t}{2}\|\nabla f(\boldsymbol{x}_k)\|^2$$

Convergence analysis

Theorem. If f is convex and L-smooth, and $\boldsymbol{x}^*$ is a minimum of f, then for step size $t \in (0, \frac{1}{L}]$, the sequence $\{\boldsymbol{x}_k\}$ produced by the gradient descent algorithm satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2tk}$$

Notes.

- $f(\boldsymbol{x}_k) \downarrow f^*$ as $k \to \infty$.
- Any limiting point of $\boldsymbol{x}_k$ is an optimal solution.
- The rate of convergence is $O(1/k)$, i.e. # of iterations to guarantee $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \epsilon$ is $O(1/\epsilon)$. For $\epsilon = 10^{-p}$, $k = O(10^p)$, exponential in the number of significant digits!
- Faster convergence with larger t; best $t = \frac{1}{L}$, but L is unknown.
- Good initial guess helps.

Proof

不要求

1. By the basic gradient step $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$,

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k) - \boldsymbol{x}^*\|^2 \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + t^2\|\nabla f(\boldsymbol{x}_k)\|^2 + 2t\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}^* - \boldsymbol{x}_k)\end{aligned}$$

2. By the last inequality on slide 15, the second term is upper bounded by

$$t^2\|\nabla f(\boldsymbol{x}_k)\|^2 \le 2t[f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})]$$

3. By the first-order condition for convexity, the third term is upper bounded by

$$2t\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}^* - \boldsymbol{x}_k) \le 2t[f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)]$$

4. Plugging 2 and 3 into 1,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + 2t[f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k+1})]$$

Proof (cont'd)

5. Rearranging and using the descent property $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$, the suboptimality gap is upper bounded by

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2}{2t}$$

   for $k \leq N - 1$.

6. Summing over k from 0 to $N - 1$,

$$N[f(\boldsymbol{x}_N) - f(\boldsymbol{x}^*)] \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_N - \boldsymbol{x}^*\|^2}{2t} \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2t}$$

   so

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2Nt}$$