Accelerated gradient methods

Yuxin Chen
Princeton University, Spring 2018
Outline

- Heavy-ball methods
- Nesterov’s accelerated gradient methods
- Accelerated proximal gradient methods (FISTA)
- Convergence analysis
- Lower bounds
(Proximal) gradient methods

Iteration complexities of (proximal) gradient methods

- strongly convex and smooth problems
  \[ O\left(\kappa \log \frac{1}{\varepsilon}\right) \]

- convex and smooth problems
  \[ O\left(\frac{1}{\varepsilon}\right) \]

Can one still hope to further accelerate convergence?
Issues and possible solutions

Issues:
- GD focuses on improving cost per iteration, which might be too “short-sighted”
- GD might sometimes zigzag / experience abrupt changes

Solutions:
- exploit information from history / past iterates
- add buffer (like momentum) to yield smoother trajectory
Heavy-ball methods

— Polyak ’64
Heavy-ball method

\[ \text{minimize}_{x \in \mathbb{R}^n} \quad f(x) \]

\[ x^{t+1} = x^t - \eta_t \nabla f(x^t) + \theta_t (x^t - x^{t-1}) \]

- add inertia to the “ball” (i.e. include momentum term) to mitigate zigzagging

B. Polyak
Proof of Lemma 2.5

It follows that

\[ x_{t+1} \neq x_t \neq x^* \neq \frac{1}{\alpha} \left( \nabla f(x_t) - \nabla f(x^*) \right), \]

where \( \alpha \) is the step size.

\[ \frac{1}{\alpha} \left( \nabla f(x_t) - \nabla f(x^*) \right) \neq 0, \]

since the smoothness condition holds. This shows that the heavy-ball method accelerates convergence compared to gradient descent.
State-space models

One can understand heavy-ball methods through dynamics of following dynamical system

\[
\begin{bmatrix}
    x^{t+1} \\
    x^t
\end{bmatrix} = \begin{bmatrix}
    (1 + \theta_t)I & -\theta_tI \\
    I & 0
\end{bmatrix} \begin{bmatrix}
    x^t \\
    x^{t-1}
\end{bmatrix} - \begin{bmatrix}
    \eta_t \nabla f(x^t) \\
    0
\end{bmatrix}
\]

or equivalently,

\[
\begin{bmatrix}
    x^{t+1} - x^* \\
    x^t - x^*
\end{bmatrix} = \begin{bmatrix}
    (1 + \theta_t)I & -\theta_tI \\
    I & 0
\end{bmatrix} \begin{bmatrix}
    x^t - x^* \\
    x^{t-1} - x^*
\end{bmatrix} - \begin{bmatrix}
    \eta_t \nabla f(x^t) \\
    0
\end{bmatrix}
\]

with \( x_{\tau} = x^t + \tau(x^* - x^t) \), where last line comes from fundamental theorem of calculus.
System matrix

\[
\begin{bmatrix}
    x^{t+1} - x^* \\
    x^t - x^*
\end{bmatrix}
= 
\begin{bmatrix}
    (1 + \theta_t) I - \eta_t \int_0^1 \nabla^2 f(x_\tau) d\tau & -\theta_t I \\
    I & 0
\end{bmatrix}
\begin{bmatrix}
    x^t - x^* \\
    x^{t-1} - x^*
\end{bmatrix}
:= H_t \text{ (system matrix)}
\]

(7.1)

**implication:** convergence of heavy-ball methods depends on spectrum of system matrix $H_t$

**key idea:** find appropriate stepsize $\eta_t$ and momentum coefficient $\theta_t$ to control spectrum of $H$
Convergence for strongly convex problems

**Theorem 7.1 (Convergence of heavy-ball methods for strongly convex functions)**

Suppose $f$ is $\mu$-strongly convex and $L$-smooth. Set $\kappa = L/\mu$, 
$\eta_t \equiv 4/((\sqrt{L} + \sqrt{\mu})^2$, $\theta_t \equiv \max \{ |1 - \sqrt{\eta_t L}|, |1 - \sqrt{\eta_t \mu}| \}^2$. Then

\[
\left\| \begin{bmatrix} x^{t+1} - x^* \\ x^t - x^* \end{bmatrix} \right\|_2 \leq \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^t \left\| \begin{bmatrix} x^1 - x^* \\ x^0 - x^* \end{bmatrix} \right\|_2
\]

- iteration complexity: $O(\sqrt{\kappa} \log \frac{1}{\varepsilon})$
- significant improvement over GD: $O(\sqrt{\kappa} \log \frac{1}{\varepsilon})$ vs. $O(\kappa \log \frac{1}{\varepsilon})$
- relies on knowledge of both $L$ and $\mu$
Proof of Theorem 7.1

In view of (7.1), it suffices to control spectrum of $H_t$. Let $\lambda_i$ denote $i$th eigenvalue of \( \int_0^1 \nabla^2 f(x_\tau) d\tau \) and set $\Lambda := \begin{bmatrix} \lambda_1 & \cdots & \cdots & \lambda_n \end{bmatrix}$, then

$$
\|H_t\| = \left\| \begin{bmatrix} (1 + \theta_t)I - \eta_t \Lambda & -\theta_t I_t \\ I & 0 \end{bmatrix} \right\|
\leq \max_{1 \leq i \leq n} \left\| \begin{bmatrix} 1 + \theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix} \right\|
$$

To finish proof, it suffices to show

$$
\max_i \left\| \begin{bmatrix} 1 + \theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix} \right\| \leq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \quad (7.2)
$$
Proof of Theorem 7.1

To show (7.2), note that two eigenvalues of
\[
\begin{bmatrix}
1 + \theta_t - \eta_t \lambda_i & -\theta_t \\
1 & 0
\end{bmatrix}
\]
are roots of
\[
z^2 - (1 + \theta_t - \eta_t \lambda_i)z + \theta_t = 0
\]
If \((1 + \theta_t - \eta_t \lambda_i)^2 \leq 4\theta_t\), then roots of this equation have same magnitudes \(\sqrt{\theta_t}\) (as they are either both imaginary or there is only one root).

In addition, one can easily check that \((1 + \theta_t - \eta_t \lambda_i)^2 \leq 4\theta_t\) is satisfied if
\[
\theta_t \in \left[ (1 - \sqrt{\eta_t \lambda_i})^2, (1 + \sqrt{\eta_t \lambda_i})^2 \right], \quad (7.3)
\]
which would hold if one picks \(\theta_t = \max \left\{ (1 - \sqrt{\eta_t L})^2, (1 - \sqrt{\eta_t \mu})^2 \right\}\)
Proof of Theorem 7.1

With this choice of $\theta_t$, we have

$$\|H_t\| \leq \sqrt{\theta_t}$$

Finally, setting $\eta_t = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}$ ensures $1 - \sqrt{\eta_t L} = -(1 - \sqrt{\eta_t \mu})$, which yields

$$\theta_t = \max \left\{ \left(1 - \frac{2\sqrt{L}}{\sqrt{L} + \sqrt{\mu}}\right)^2, \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2 \right\} = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^2$$

This in turn establishes

$$\|H_t\| \leq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$
Nesterov’s accelerated gradient methods
For strongly convex $f$, including momentum terms allows to improve iteration complexity from $O(\kappa \log \frac{1}{\varepsilon})$ to $O(\sqrt{\kappa} \log \frac{1}{\varepsilon})$

Can we obtain improvement for convex case (without strong convexity) as well?
Nesterov’s idea

— Nesterov ’83

\[ x^{t+1} = y^t - \eta_t \nabla f(y^t) \]
\[ y^{t+1} = x^{t+1} + \frac{t}{t+3}(x^{t+1} - x^t) \]

Y. Nesterov

- alternates between gradient updates and proper extrapolation
- each iteration takes nearly same cost as GD
- not a descent method (i.e. we may not have \( f(x^{t+1}) \leq f(x^t) \))
- one of most beautiful and mysterious results in optimization ...
Convergence of Nesterov’s accelerated gradient method

Suppose $f$ is convex and $L$-smooth. If $\eta_t \equiv \eta = 1/L$, then

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

- iteration complexity: $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$
- much faster than gradient methods
- we’ll provide proof for (more general) proximal version later
Nesterov’s momentum coefficient \( \frac{t}{t+3} = 1 - \frac{3}{t} \) is particularly mysterious.
Interpretation using differential equations

spring force
\[ F = -\nabla f(X) \]

time-varying damping force
\[ -\frac{3}{\tau} \dot{X} \]

To develop insights into why Nesterov’s method works so well, it’s helpful to look at its continuous limits \((\eta_t \to 0)\), which is given by second-order ordinary differential equations (ODE)

\[
\ddot{X}(\tau) + \frac{3}{\tau} \dot{X}(\tau) + \nabla f(X(\tau)) = 0
\]

— Su, Boyd, Candes ’14
Heuristic derivation of ODE

To begin with, Nesterov’s update rule is equivalent to
\[
\frac{x^{t+1} - x^t}{\sqrt{\eta}} = \frac{t - 1}{t + 2} \frac{x^t - x^{t-1}}{\sqrt{\eta}} - \sqrt{\eta} \nabla f(y^t) \tag{7.4}
\]

Let \( t = \frac{\tau}{\sqrt{\eta}} \). Set \( X(\tau) \approx x^\tau/\sqrt{\eta} = x^t \) and \( X(\tau + \sqrt{\eta}) \approx x^{t+1} \). Then Taylor expansion gives
\[
\frac{x^{t+1} - x^t}{\sqrt{\eta}} \approx \dot{X}(\tau) + \frac{1}{2} \ddot{X}(\tau) \sqrt{\eta}
\]
\[
\frac{x^t - x^{t-1}}{\sqrt{\eta}} \approx \dot{X}(\tau) - \frac{1}{2} \ddot{X}(\tau) \sqrt{\eta}
\]

which combined with (7.4) yields
\[
\dot{X}(\tau) + \frac{1}{2} \ddot{X}(\tau) \sqrt{\eta} \approx \left(1 - \frac{3\sqrt{\eta}}{\tau}\right) \left(\dot{X}(\tau) - \frac{1}{2} \ddot{X}(\tau) \sqrt{\eta}\right) - \sqrt{\eta} \nabla f(X(\tau))
\]
\[
\implies \dot{X}(\tau) + \frac{3}{\tau} \dddot{X}(\tau) + \nabla f(X(\tau)) \approx 0
\]
\[ \ddot{X} + \frac{3}{\tau} \dot{X} + \nabla f(X) = 0 \] (7.5)

It is not hard to show that this ODE obeys

\[ f(X(\tau)) - f^{\text{opt}} \leq O \left( \frac{1}{\tau^2} \right) \] (7.6)

which somehow explains Nesterov’s $O(1/t^2)$ convergence
Proof of (7.6)

Define $E(\tau) = \tau^2 (f(X) - f^{\text{opt}}) + 2\|X + \frac{\tau}{2} \dot{X} - X^*\|_2^2$. This obeys

Lyapunov function / energy function

\[
\dot{E} = 2\tau (f(X) - f^{\text{opt}}) + \tau^2 \langle \nabla f(X), \dot{X} \rangle + 4 \langle X + \frac{\tau}{2} \dot{X} - X^*, \frac{3}{2} \dot{X} + \frac{\tau}{2} \ddot{X} \rangle
\]

\[(i) \quad 2\tau (f(X) - f^{\text{opt}}) - 2\tau \langle X - X^*, \nabla f(X) \rangle \leq 0 \quad \text{(by convexity)}
\]

where (i) follows by replacing $\tau \ddot{X} + 3 \dot{X}$ with $-\tau \nabla f(X)$

This means $E$ is non-decreasing in $\tau$, and hence

$f(X(\tau)) - f^{\text{opt}} \overset{\text{defn of } E}{\leq} \frac{E(\tau)}{\tau^2} \leq \frac{E(0)}{\tau^2} = O\left(\frac{1}{\tau^2}\right)$
\[ \ddot{X} + \frac{3}{\tau} \dot{X} + \nabla f(X) = 0 \]

- 3 is smallest constant that guarantees $O(1/\tau^2)$ convergence, and can be replaced by any other $\alpha \geq 3$
- in some sense, 3 minimizes pre-constant in convergence bound $O(1/\tau^2)$ (see Su, Boyd, Candes ’14)
Numerical example

taken from UCLA EE236C

minimize \( \mathbf{x} \) \( \log \left( \sum_{i=1}^{m} \exp(\mathbf{a}_i^\top \mathbf{x} + b_i) \right) \)

with randomly generated problems and \( m = 2000, n = 1000 \)

\[ f(\mathbf{x}) - f_{\text{opt}} \]

\[ 0 \quad 50 \quad 100 \quad 150 \quad 200 \]

\[ 10^{-6} \quad 10^{-5} \quad 10^{-4} \quad 10^{-3} \quad 10^{-2} \quad 10^{-1} \quad 10^0 \]

\text{gradient} \quad \text{FISTA}

Example

加速正則化 Gradient Methods
Extension to composite models

\[
\begin{align*}
\text{minimize}_x & \quad F(x) := f(x) + h(x) \\
\text{subject to} & \quad x \in \mathbb{R}^n
\end{align*}
\]

- \( f \): convex and smooth
- \( h \): convex (may not be differentiable)

let \( F^{\text{opt}} := \min_x F(x) \) be optimal cost
FISTA (Beck & Teboulle ’09)

Fast iterative shrinkage-thresholding algorithm

\[ x^{t+1} = \text{prox}_{\eta_t h}(y^t - \eta_t \nabla f(y^t)) \]

\[ y^{t+1} = x^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}} (x^{t+1} - x^t) \]

where \( y^0 = x^0 \), \( \theta_0 = 1 \) and \( \theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} \)

- momentum coefficient originally proposed by Nesterov ’83
Momentum coefficient

\[ \theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} \]

with \( \theta_0 = 1 \)

coefficient \( \frac{\theta_{t-1}}{\theta_{t+1}} = 1 - \frac{3}{t} + o\left(\frac{1}{t}\right) \) (homework)

- asymptotically equivalent to \( \frac{t}{t+3} \)
Momentum coefficient

\[ \theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} \]

with \( \theta_0 = 1 \)

Fact 7.2

For all \( t \geq 1 \), one has \( \theta_t \geq \frac{t+2}{2} \) (homework)
Convergence analysis
Theorem 7.3 (Convergence of accelerated proximal gradient methods for convex problems)

Suppose $f$ is convex and $L$-smooth. If $\eta_t \equiv 1/L$, then

$$F(x^t) - F^{\text{opt}} \leq \frac{2L\|x^0 - x^*\|_2^2}{(t + 1)^2}$$

- improved iteration complexity (i.e. $O(1/\sqrt{\varepsilon})$) than proximal gradient method (i.e. $O(1/\varepsilon)$)

- fast if prox can be efficiently implemented
Recall following fundamental inequality shown in last lecture:

**Lemma 7.4**

Let \( y^+ = \text{prox}_{\frac{1}{L} h}(y - \frac{1}{L} \nabla f(y)) \), then

\[
F(y^+) - F(x) \leq \frac{L}{2} \| x - y \|_2^2 - \frac{L}{2} \| x - y^+ \|_2^2
\]
Proof of Theorem 7.6

1. build discrete-time version of “Lyapunov function”

2. *magic happens!*
   - “Lyapunov function” is non-increasing when Nesterov’s momentum coefficient is adopted
Proof of Theorem 7.6

Key lemma: monotonicity of certain “Lyapunov function”

**Lemma 7.5**

Let $u^t = \theta_{t-1} x^t - (x^* + (\theta_{t-1} - 1)x^{t-1})$. Then

$$
\|u^{t+1}\|_2^2 + 2\frac{2}{L}\theta_t^2 (F(x^{t+1}) - F^{opt}) \leq \|u^t\|_2^2 + 2\frac{2}{L}\theta_{t-1}^2 (F(x^t) - F^{opt})
$$

- quite similar to $2\|X + \frac{\tau}{2} \dot{X} - X^*\|_2^2 + \tau^2 (f(X) - f^{opt})$
  (Lyapunov function) as discussed before (think about $\theta_t \approx t/2$)
Proof of Theorem 7.6

With Lemma 7.5 in place, one has

\[
\frac{2}{L} \theta_{t-1}^2 (F(x^t) - F^{\text{opt}}) \leq \|u^1\|_2^2 + \frac{2}{L} \theta_0^2 (F(x^1) - F^{\text{opt}}) \\
= \|x^1 - x^*\|_2^2 + \frac{2}{L} (F(x^1) - F^{\text{opt}})
\]

To bound RHS of this inequality, we use Lemma 7.4 and \(y^0 = x^0\) to get

\[
\frac{2}{L} (F(x^1) - F^{\text{opt}}) \leq \|y^0 - x^*\|_2^2 - \|x^1 - x^*\|_2^2 = \|x^0 - x^*\|_2^2 - \|x^1 - x^*\|_2^2
\]

\[
\iff \|x^1 - x^*\|_2^2 + \frac{2}{L} (F(x^1) - F^{\text{opt}}) \leq \|x^0 - x^*\|_2^2
\]

As a result,

\[
\frac{2}{L} \theta_{t-1}^2 (F(x^t) - F^{\text{opt}}) \leq \|x^1 - x^*\|_2^2 + \frac{2}{L} (F(x^1) - F^{\text{opt}}) \leq \|x^0 - x^*\|_2^2,
\]

\[
\implies F(x^t) - F^{\text{opt}} \leq \frac{L\|x^0 - x^*\|_2^2}{2\theta_{t-1}^2} \leq \frac{2L\|x^0 - x^*\|_2^2}{(t+1)^2}
\]
Proof of Lemma 7.5

Take \( x = \frac{1}{\theta_t} x^* + (1 - \frac{1}{\theta_t}) x^t \) and \( y = y^t \) in Lemma 7.4 to get

\[
F(x^{t+1}) - F\left(\theta_t^{-1} x^* + (1 - \theta_t^{-1}) x^t\right)
\leq \frac{L}{2} \left\| \theta_t^{-1} x^* + (1 - \theta_t^{-1}) x^t - y^t \right\|_2^2 - \frac{L}{2} \left\| \theta_t^{-1} x^* + (1 - \theta_t^{-1}) x^t - x^{t+1} \right\|_2^2
\]

\[
= \frac{L}{2\theta_t^2} \left\| x^* + (\theta_t - 1) x^t - \theta_t y^t \right\|_2^2 - \frac{L}{2\theta_t^2} \left\| x^* + (\theta_t - 1) x^t - \theta_t x^{t+1} \right\|_2^2
= -\mathbf{u}^{t+1}
\]

\[
\overset{(i)}{=} \frac{L}{2\theta_t^2} \left( \left\| \mathbf{u}^t \right\|_2^2 - \left\| \mathbf{u}^{t+1} \right\|_2^2 \right),
\]

(7.8)

where (i) follows from definition of \( \mathbf{u}^t \) and \( y^t = x^t + \frac{\theta_t^{-1} - 1}{\theta_t} (x^t - x^{t-1}) \)
Proof of Lemma 7.5 (cont.)

We will also lower bound (7.7). By convexity of $F$,

$$F\left(\theta_t^{-1} x^* + (1 - \theta_t^{-1}) x^t\right) \leq \theta_t^{-1} F(x^*) + (1 - \theta_t^{-1}) F(x^t)$$

$$= \theta_t^{-1} F^{\text{opt}} + (1 - \theta_t^{-1}) F(x^t)$$

$$\iff F\left(\theta_t^{-1} x^* + (1 - \theta_t^{-1}) x^t\right) - F(x^{t+1}) \leq (1 - \theta_t^{-1}) (F(x^t) - F^{\text{opt}}) - (F(x^{t+1}) - F^{\text{opt}})$$

Combining this with (7.8) and $\theta_2^2 - \theta_t = \theta_{t-1}^2$ yields

$$\frac{L}{2} \left(\|u^t\|_2^2 - \|u^{t+1}\|_2^2\right) \geq \theta_t^2 (F(x^{t+1}) - F^{\text{opt}}) - (\theta_t^2 - \theta_t) (F(x^t) - F^{\text{opt}})$$

$$= \theta_t^2 (F(x^{t+1}) - F^{\text{opt}}) - \theta_{t-1}^2 (F(x^t) - F^{\text{opt}}),$$

thus finishing proof.
Convergence for strongly convex case

\[ x^{t+1} = \text{prox}_{\eta_t h}(y^t - \eta_t \nabla f(y^t)) \]

\[ y^{t+1} = x^{t+1} + \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} (x^{t+1} - x^t) \]

**Theorem 7.6 (Convergence of accelerated proximal gradient methods for strongly convex case)**

Suppose \( f \) is \( \mu \)-strongly convex and \( L \)-smooth. If \( \eta_t \equiv 1/L \), then

\[ F(x^t) - F^{\text{opt}} \leq \left( 1 - \frac{1}{\sqrt{\kappa}} \right)^t \left( F(x^0) - F^{\text{opt}} + \frac{\mu \| x^0 - x^* \|^2}{2} \right) \]
A practical issue

Fast convergence requires knowledge of $\kappa = L/\mu$

- in practice, estimating $\mu$ is typically very challenging

A common observation: ripples / bumps in the traces of cost values
Rippling behavior

Numerical example: take $y^{t+1} = x^{t+1} + \frac{1 - \sqrt{q}}{1 + \sqrt{q}}(x^{t+1} - x^t)$; $q^* = 1/\kappa$

- when $q > q^*$: we underestimate momentum $\quad\rightarrow$ slower convergence
- when $q < q^*$: we overestimate momentum ($\frac{1 - \sqrt{q}}{1 + \sqrt{q}}$ is large)
  $\quad\rightarrow$ overshotting / rippling behavior

period of ripples is often proportional to $\sqrt{L/\mu}$

O'Donoghue, Candes '12
Adaptive restart (O’Donoghue, Candes ’12)

When certain criterion is met, restart running FISTA with

\[ x^0 \leftarrow x^t \]
\[ y^0 \leftarrow x^t \]
\[ \theta_0 = 1 \]

- take current iteration as new starting point
- erase all memory of previous iterates and restart momentum back to zero
Numerical comparisons of adaptive restart schemes

- function scheme: restart when $f(x^t) > f(x^{t-1})$
- gradient scheme: restart when $\langle \nabla f(y^{t-1}), x^t - x^{t-1} \rangle > 0$

- restart when momentum lead us towards a bad direction
• with overestimated momentum (e.g. $q = 0$), one sees spiralling trajectory

• adaptive restart helps mitigate this issue
Lower bounds
Optimality of Nesterov’s method

Interestingly, no first-order methods can improve upon Nesterov’s result in general.

More precisely, \( \exists \) convex and \( L \)-smooth function \( f \) s.t.

\[
f(x^t) - f^{\text{opt}} \geq \frac{3L\|x^0 - x^*\|_2^2}{32(t + 1)^2}
\]

as long as \( x^k \in x^0 + \text{span}\{\nabla f(x^0), \cdots, \nabla f(x^{k-1})\} \) for all \( 1 \leq k \leq t \)

— Nemirovski, Yudin ’83
Example

\[
\text{minimize}_{x \in \mathbb{R}^{(2n+1)}} \quad f(x) = \frac{L}{4} \left( \frac{1}{2} x^\top A x - e_1^\top x \right)
\]

where \( A = \begin{bmatrix} 2 & -1 \\ -1 & 2 & -1 \\ & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \ddots & \ddots \\ -1 & 2 & -1 \\ -1 & 2 \end{bmatrix} \in \mathbb{R}^{(2n+1) \times (2n+1)} \)

- \( f \) is convex and \( L \)-smooth
- optimizer \( x^* \) is given by \( x^*_i = 1 - \frac{i}{2n+2} \) \((1 \leq i \leq n)\) obeying

\[
f^{\text{opt}} = \frac{L}{8} \left( \frac{1}{2n+2} - 1 \right) \quad \text{and} \quad \|x^*\|_2^2 \leq \frac{2n+2}{3}
\]
Example

\[
\begin{align*}
    &\text{minimize} \quad x \in \mathbb{R}^{(2n+1)} \\
    &f(x) = \frac{L}{4} \left( \frac{1}{2} x^\top A x - e_1^\top x \right)
\end{align*}
\]

where \( A = \begin{bmatrix}
2 & -1 & \cdot & \cdot & \cdot \\
-1 & 2 & -1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
-1 & 2 & -1 & \cdot & \cdot \\
-1 & 2 & \cdot & \cdot & \cdot \\
\end{bmatrix} \in \mathbb{R}^{(2n+1) \times (2n+1)} \)

- \( \nabla f(x) = \frac{L}{4} A x - \frac{L}{4} e_1 \)
- \( \text{span}\{\nabla f(x^0), \cdots, \nabla f(x^{k-1})\} = \text{span}\{e_1, \cdots, e_k\} \) if \( x^0 = 0 \)
  - \( \text{span}\{\nabla f(x^0), \cdots, \nabla f(x^{k-1})\} := \mathcal{K}_k \)
  - every iteration of first-order methods expands search space by \textit{at most} one dimension
Example (cont.)

If we start with \( x^0 = 0 \), then

\[
    f(x^n) \geq \inf_{x \in K_n} f(x) = \frac{L}{8} \left( \frac{1}{n + 1} - 1 \right)
\]

\[
    \implies \quad \frac{f(x^n) - f^{\text{opt}}}{\|x^0 - x^*\|^2_2} \geq \frac{L}{8} \left( \frac{1}{n + 1} - \frac{1}{2n + 2} \right) = \frac{3L}{32(n + 1)^2}
\]
### Summary: accelerated proximal gradient

<table>
<thead>
<tr>
<th></th>
<th>stepsize rule</th>
<th>convergence rate</th>
<th>iteration complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex &amp; smooth problems</td>
<td>$\eta_t = \frac{1}{L}$</td>
<td>$O\left(\frac{1}{t^2}\right)$</td>
<td>$O\left(\frac{1}{\sqrt{\varepsilon}}\right)$</td>
</tr>
<tr>
<td>strongly convex &amp; smooth problems</td>
<td>$\eta_t = \frac{1}{L}$</td>
<td>$O\left((1 - \frac{1}{\sqrt{\kappa}})^t\right)$</td>
<td>$O\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$</td>
</tr>
</tbody>
</table>
Reference


Reference

[10] "Optimization methods for large-scale systems, EE236C lecture notes,” L. Vandenberghe, UCLA.
[12] "Introductory lectures on convex optimization: a basic course,” Y. Nesterov, 2004