Lecture 8. Accelerated Gradient Methods

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
University of Utah
Math 5750/6880, Fall 2021

The iteration complexity of (proximal) gradient methods

strongly convex and smooth problems $O(\kappa \log \frac{1}{\epsilon})$.

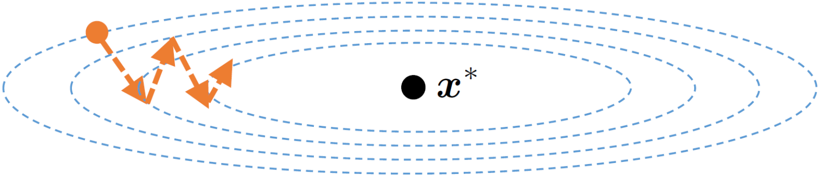
convex and smooth problems $O(\frac{1}{\epsilon})$.

Can one still hope to further accelerate convergence?

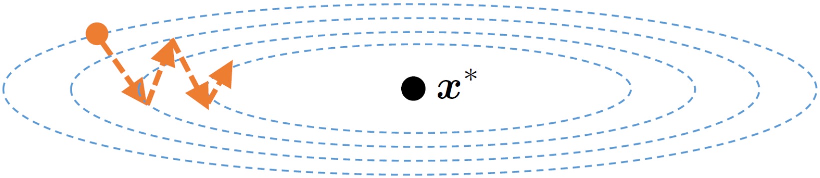
Issues:

1) GD focuses on improving the cost per iteration, which might sometimes be too "short-sighted";

2) GD might sometimes zigzag or experience abrupt changes.



gradient descent



gradient descent

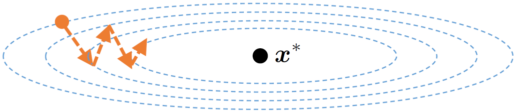
Solutions:

1) exploit information from the history (i.e. past iterates);
2) add buffers (like momentum) to yield smoother trajectory.

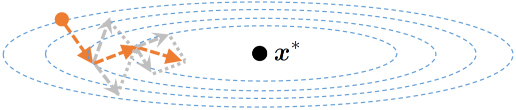
Heavy-ball method

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta_t \nabla f(\boldsymbol{x}^{t}) + \underbrace{\theta_t(\boldsymbol{x}^{t} - \boldsymbol{x}^{t-1})}_{\text{momentum term}},$$

where we add inertia to the "ball" (i.e. include a momentum term) to mitigate zigzagging.



gradient descent



heavy-ball method

Consider

$$\min_{\boldsymbol{x}} \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^*),$$

where $\boldsymbol{Q} \succ 0$ has a condition number κ. One can understand heavy-ball methods through dynamical systems.

Consider the following dynamical system

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

or equivalently

$$\begin{aligned}
\underbrace{\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^{t} - \boldsymbol{x}^* \end{bmatrix}}_{\text{state}} &= \begin{bmatrix} (1+\theta_t)\boldsymbol{I} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{t} - \boldsymbol{x}^* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^* \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\boldsymbol{x}^t) \\ 0 \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta_t \boldsymbol{Q} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}}_{\text{system matrix} := \boldsymbol{H}_t} \begin{bmatrix} \boldsymbol{x}^{t} - \boldsymbol{x}^* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^* \end{bmatrix}
\end{aligned} \tag{1}$$

The convergence of heavy-ball methods depends on the spectrum of the system matrix $\boldsymbol{H}_t$. We need to find appropriate stepsizes η_t and momentum coefficients θ_t to control the spectrum of $\boldsymbol{H}_t$.

Theorem 1. [Convergence of heavy-ball methods for quadratic functions] Suppose f is a L-smooth and μ-strongly convex quadratic function. Set $\eta_t \equiv 4/(\sqrt{L}+\sqrt{\mu})^2$, $\theta_t \equiv \max\{|1-\sqrt{\eta_t L}|, |1-\sqrt{\eta_t \mu}|\}^2$, and $\kappa = L/\mu$. Then

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^t - \boldsymbol{x}^* \end{bmatrix} \right\|_2 \lesssim \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^t \left\| \begin{bmatrix} \boldsymbol{x}^1 - \boldsymbol{x}^* \\ \boldsymbol{x}^0 - \boldsymbol{x}^* \end{bmatrix} \right\|_2$$

Note that the iteration complexity is $O(\sqrt{\kappa}\log\frac{1}{\epsilon})$ (vs. $O(\kappa \log \frac{1}{\epsilon})$ of GD), the convergence rate relies on knowledge of both L and μ.

Proof of Theorem 1. In view of (1), it suffices to control the spectrum of $\boldsymbol{H}_t$ (which is time-invariant). Let λ_i be the ith eigenvalue of $\boldsymbol{Q}$ and let $\Lambda := diag(\lambda_1, \cdots, \lambda_n)$, then there exists an orthogonal matrix $\boldsymbol{U}$ such that $\boldsymbol{Q} = \boldsymbol{U}\Lambda\boldsymbol{U}^\top$ and we have

$$\left\| \begin{bmatrix} \boldsymbol{U} & 0 \\ 0 & \boldsymbol{U} \end{bmatrix} \begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta\boldsymbol{Q} & -\theta_t\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{U} & 0 \\ 0 & \boldsymbol{U} \end{bmatrix}^\top \right\|_2 = \left\| \begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta\Lambda & -\theta_t\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \right\|_2.$$

Further, note that the characteristic polynomial of the right-hand side matrix satisfies

$$\begin{aligned} ch\left(\begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta\Lambda & -\theta_t\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}\right) &= det \begin{bmatrix} \lambda\boldsymbol{I} - \big((1+\theta_t)\boldsymbol{I} - \eta\Lambda\big) & \theta_t\boldsymbol{I} \\ -\boldsymbol{I} & \lambda\boldsymbol{I} \end{bmatrix} \\ &= det\Big(\lambda^2\boldsymbol{I} - \lambda\big((1+\theta_t)\boldsymbol{I} - \eta\Lambda\big) + \theta_t\boldsymbol{I}\Big). \end{aligned}$$

The matrix $\lambda^2 \boldsymbol{I} - \lambda\big((1+\theta_t)\boldsymbol{I} - \eta\Lambda\big) + \theta_t \boldsymbol{I}$ is diagonal with each diagonal entry be the characteristic polynomial of the following 2×2 matrix

$$\begin{bmatrix} 1+\theta_t - \eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}.$$

Then the spectral radius (denoted by $\rho(\cdot)$) [1] of $\boldsymbol{H}_t$ obeys

$$\rho(\boldsymbol{H}_t) = \rho\Big(\begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta_t\Lambda & -\theta_t\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \Big) \le \max_{1\le i\le n} \rho\Big(\begin{bmatrix} 1+\theta_t - \eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix} \Big)$$

To finish the proof, it suffices to show

$$\max_i \rho\Big(\begin{bmatrix} 1+\theta_t - \eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix} \Big) \le \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}. \tag{2}$$

[1] The largest absolute value of its eigenvalues

To show (2), note that the two eigenvalues of $\begin{bmatrix} 1+\theta_t-\eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}$ are the roots of

$$z^2-(1+\theta_t-\eta_t\lambda_i)z+\theta_t=0. \tag{3}$$

If $(1+\theta_t-\eta_t\lambda_i)^2 \le 4\theta_t$, then the roots of this equation have the same magnitudes $\sqrt{\theta_t}$ (as they are either both imaginary or there is only one root).

In addition, one can easily check that $(1+\theta_t-\eta_t\lambda_i)^2 \le 4\theta_t$ is satisfied if

$$\theta_t \in [(1-\sqrt{\eta_t\lambda_i})^2, (1+\sqrt{\eta_t\lambda_i})^2], \tag{4}$$

which would hold if one picks $\theta_t = \max\{(1-\sqrt{\eta_t L})^2, (1-\sqrt{\eta_t\mu})^2\}$.

With this choice of θ_t, we have (from (3) and the fact that two eigenvalues have identical magnitudes) [Vieta's formula: $z_1 z_2 = \theta_t$]

$$\rho(\boldsymbol{H}_t) \leq \sqrt{\theta_t}.$$

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

$$\rho(\boldsymbol{H}_t) \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

Nesterov's accelerated gradient methods

$$\boldsymbol{x}^{t+1} = \boldsymbol{y}^t - \eta_t \nabla f(\boldsymbol{y}^t); \quad \boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{t}{t+3}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t).$$

> alternates between gradient updates and proper extrapolation

> each iteration takes nearly the same cost as GD

> not a descent method (i.e. we may not have $f(\boldsymbol{x}^{t+1}) \leq f(\boldsymbol{x}^t)$), we will see it later.

Theorem 2. [Convergence of Nesterov's accelerated gradient method] Suppose f is convex and L-smooth. if $\eta_t \equiv \eta = 1/L$, then

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

Remark. The iteration complexity if $O(\frac{1}{\sqrt{\epsilon}})$, which is much faster than gradient methods.

ODE analogy of Nesterov's accelerated gradient

To develop insight into why Nesterov's method works so well, it's helpful to look at its continuous limits ($\eta_t \to 0$). To begin with, Nesterov's update rule is equivalent to

$$\frac{\boldsymbol{x}^{t+1} - \boldsymbol{x}^t}{\sqrt{\eta}} = \frac{t-1}{t+2} \frac{\boldsymbol{x}^t - \boldsymbol{x}^{t-1}}{\sqrt{\eta}} - \sqrt{\eta} \nabla f(\boldsymbol{y}^t). \tag{5}$$

Let $t = \frac{\tau}{\sqrt{\eta}}$. Set $\boldsymbol{X}(\tau) \approx \boldsymbol{x}^{\tau/\sqrt{\eta}} = \boldsymbol{x}^t$ and $\boldsymbol{X}(\tau + \sqrt{\eta}) \approx \boldsymbol{x}^{t+1}$. Then the Taylor expansion gives

$$\frac{\boldsymbol{x}^{t+1} - \boldsymbol{x}^t}{\sqrt{\eta}} \approx \dot{\boldsymbol{X}}(\tau) + \frac{1}{2} \ddot{\boldsymbol{X}}(\tau) \sqrt{\eta}; \quad \frac{\boldsymbol{x}^t - \boldsymbol{x}^{t-1}}{\sqrt{\eta}} \approx \dot{\boldsymbol{X}}(\tau) - \frac{1}{2} \ddot{\boldsymbol{X}}(\tau) \sqrt{\eta},$$

which combined with (5) yields

$$\dot{\boldsymbol{X}}(\tau) + \frac{1}{2}\ddot{\boldsymbol{X}}(\tau)\sqrt{\eta} \approx \Big(1 - \frac{3\sqrt{\eta}}{\tau}\Big)\Big(\dot{\boldsymbol{X}}(\tau) - \frac{1}{2}\ddot{\boldsymbol{X}}(\tau)\sqrt{\eta}\Big) - \sqrt{\eta}\nabla f(\boldsymbol{X}(\tau))$$

$$\Rightarrow \ddot{\boldsymbol{X}}(\tau) + \frac{3}{\tau}\dot{\boldsymbol{X}}(\tau) + \nabla f(\boldsymbol{X}(\tau)) = 0.$$

What is the ODE limit of the heavy-ball method?

Heavy-ball method

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\nabla f(\boldsymbol{x}^t) + \theta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}),$$

Let $\boldsymbol{m}^t := (\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)/\sqrt{\eta}$ and let $\theta := 1 - \gamma\sqrt{\eta}$, where $\gamma \geq 0$ is another hyperparameter. Then we can rewrite the heavy-ball method as

$$\boldsymbol{m}^{t+1} = (1 - \gamma\sqrt{\eta})\boldsymbol{m}^t - \sqrt{\eta}\nabla f(\boldsymbol{x}^t); \;\; \boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \sqrt{s}\boldsymbol{m}^{t+1}.$$

Let $s \to 0$; we obtain the following system of first-order ODEs

$$\frac{d\boldsymbol{X}(t)}{dt} = \boldsymbol{M}(t); \;\; \frac{d\boldsymbol{M}(t)}{dt} = -\gamma\boldsymbol{M}(t) - \nabla f(\boldsymbol{X}(t)),$$

which can be further written as

$$\ddot{\boldsymbol{X}}(\tau) + \gamma\dot{\boldsymbol{X}}(\tau) + \nabla f(\boldsymbol{X}(\tau)) = 0.$$

Heavy-ball method:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta \nabla f(\boldsymbol{x}^{t}) + \theta(\boldsymbol{x}^{t} - \boldsymbol{x}^{t-1}),$$

and the ODE-limit is

$$\ddot{\boldsymbol{X}}(\tau) + \gamma \dot{\boldsymbol{X}}(\tau) + \nabla f(\boldsymbol{X}(\tau)) = 0.$$

Nesterov's acceleration:

$$\boldsymbol{x}^{t+1} = \boldsymbol{y}^{t} - \eta \nabla f(\boldsymbol{y}^{t}); \quad \boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{t}{t+3}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t}),$$

and the ODE-limit is

$$\ddot{\boldsymbol{X}}(\tau) + \frac{3}{\tau} \dot{\boldsymbol{X}}(\tau) + \nabla f(\boldsymbol{X}(\tau)) = 0.$$

Convergence rate inspired by the ODE analysis

By the standard ODE theory, we can show that

$$f(\boldsymbol{X}(\tau)) - f^{opt} \leq O(\frac{1}{\tau^2}), \tag{6}$$

which somehow explains Nesterov's $O(1/t^2)$ convergence.

Convergence rate inspired by the ODE analysis

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

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

$$\underbrace{=}_{(i)} 2\tau(f(\boldsymbol{X}) - f^{opt}) - 2\tau\langle \boldsymbol{X} - \boldsymbol{X}^*, \nabla f(\boldsymbol{X})\rangle \underbrace{\leq}_{convexity} 0$$

where (i) follows by replacing $\tau\ddot{\boldsymbol{X}} + 3\dot{\boldsymbol{X}}$ with $-\tau\nabla f(\boldsymbol{X})$. This means E is non-decreasing in τ, and hence

$$f(\boldsymbol{X}(\tau)) - f^{opt} \underbrace{\leq}_{def\ \ of\ \ E} \frac{E(\tau)}{\tau^2} \leq \frac{E(0)}{\tau^2} = O(\frac{1}{\tau^2}).$$

Extend Nesterov's acceleration to composite models

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) := f(\boldsymbol{x}) + h(\boldsymbol{x}) \quad s.t. \quad \boldsymbol{x} \in \mathbb{R}^n,$$

where f is convex and smooth and h is convex (may not be differentiable). Let $F^{opt} := \min_{\boldsymbol{x}} F(\boldsymbol{x})$ be the optimal cost.

FISTA (Fast iterative shrinkage-thresholding algorithm)

$$\boldsymbol{x}^{t+1} = prox_{\eta_t h}(\boldsymbol{y}^t - \eta_t \nabla f(\boldsymbol{y}^t))$$
$$\boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$$

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

We can show that $\frac{\theta_t - 1}{\theta_{t+1}} = 1 - \frac{3}{t} + o(\frac{1}{t})$ Homework..
We can also show that $\theta_t \geq \frac{t+2}{2}$. (Math induction.)

Theorem 3. [Convergence of accelerated proximal gradient methods for convex problems] Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

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

Remark. The algorithm is fast if prox can be efficiently implemented.

Remark. The algorithm is particularly useful for ℓ_1-regularization problem in e.g. image processing (total variation in the wavelet space) and compressed sensing.

Lemma 1. [Fundamental inequality for proximal method] Let

$$\boldsymbol{y}^{+} = prox_{\frac{1}{L}h}(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y})),$$

then

$$F(\boldsymbol{y}^{+}) - F(\boldsymbol{x}) \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^{+}\|_2^2.$$

To proof Theorem 3, we follow: 1) build a discrete-time version of "Lyapunov function"; 2) "Lyapunov function" is non-increasing when Nesterov's momentum coefficients are adopted.

Proof of Lemma 1. More precisely, we have

$$F(\boldsymbol{y}^+) - F(\boldsymbol{x}) \leq \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}^+\|_2^2 - \underbrace{g(\boldsymbol{x},\boldsymbol{y})}_{\geq 0 \ by \ convexity}$$

where $g(\boldsymbol{x},\boldsymbol{y}) := f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle$.

Define $\phi(\boldsymbol{z}) = f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{z}-\boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2 + h(\boldsymbol{z})$. It is easily seen that $\boldsymbol{y}^+ = \arg\min_{\boldsymbol{z}} \phi(\boldsymbol{z})$. Two important properties:

1. Since $\phi(\boldsymbol{z})$ is L-strongly convex, one has

$$\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{y}^+) + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}^+\|_2^2.$$

2. From smoothness,

$$\phi(\boldsymbol{y}^+) = \underbrace{f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{y}^+-\boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{y}^+-\boldsymbol{y}\|_2^2}_{\text{upper bound on } f(\boldsymbol{y}^+) \text{ (L-smoothness)}} + h(\boldsymbol{y}^+) \geq f(\boldsymbol{y}^+) + h(\boldsymbol{y}^+) = F(\boldsymbol{y}^+).$$

Taken collectively, these yield

$$\phi(\boldsymbol{x}) \geq F(\boldsymbol{y}^+) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^+\|_2^2,$$

which together with the definition of $\phi(\boldsymbol{x})$ gives

$$\underbrace{f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + h(\boldsymbol{x})}_{=f(\boldsymbol{x})+h(\boldsymbol{x})-g(\boldsymbol{x},\boldsymbol{y})=F(\boldsymbol{x})-g(\boldsymbol{x},\boldsymbol{y})} + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \geq F(\boldsymbol{y}^+) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^+\|_2^2$$

which finishes the proof.

Lemma 2. [Monotonicity of certain "Lyapunov function"]
Let

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

Then

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

Remark. Note that this is quite similar to $2\|\boldsymbol{X} + \frac{\tau}{2}\dot{\boldsymbol{X}} - \boldsymbol{X}^*\|_2^2 + \tau^2(f(\boldsymbol{X}) - f^{opt})$, think about $\theta_t \approx t/2$.

Proof of Lemma 2. Take $\boldsymbol{x} = \frac{1}{\theta_t}\boldsymbol{x}^* + (1 - \frac{1}{\theta_t})\boldsymbol{x}^t$ and $\boldsymbol{y} = \boldsymbol{y}^t$ (based on FISTA $\boldsymbol{x}^{t+1} = prox_{\frac{1}{L}h}(\boldsymbol{y}^t - \frac{1}{L}\nabla f(\boldsymbol{y}^t))$, we have $\boldsymbol{x}^{t+1} = prox_{\frac{1}{L}h}(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y}))$) in Lemma 1 to get

$$\begin{aligned}
& F(\boldsymbol{x}^{t+1}) - F(\theta_t^{-1}\boldsymbol{x}^* + (1-\theta_t^{-1})\boldsymbol{x}^t) \\
& \le \frac{L}{2}\|\theta_t^{-1}\boldsymbol{x}^* + (1-\theta_t^{-1})\boldsymbol{x}^t - \boldsymbol{y}^t\|_2^2 - \frac{L}{2}\|\theta_t^{-1}\boldsymbol{x}^* + (1-\theta_t^{-1})\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|_2^2 \\
& = \frac{L}{2\theta_t^2}\|\boldsymbol{x}^* + (\theta_t - 1)\boldsymbol{x}^t - \theta_t\boldsymbol{y}^t\|_2^2 - \frac{L}{2\theta_t^2}\|\underbrace{\boldsymbol{x}^* + (\theta_t - 1)\boldsymbol{x}^t - \theta_t\boldsymbol{x}^{t+1}}_{=-\boldsymbol{u}^{t+1}}\|_2^2 \qquad (7) \\
& \underbrace{=}_{(i)} \frac{L}{2\theta_t^2}(\|\boldsymbol{u}^t\|_2^2 - \|\boldsymbol{u}^{t+1}\|_2^2),
\end{aligned}$$

where (i) follows from the definition of $\boldsymbol{u}^t$ and $\boldsymbol{y}^t = \boldsymbol{x}^t + \frac{\theta_{t-1}-1}{\theta_t}(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$.

We will also lower bound (7). By convexity of F,

$$F\Big(\theta_t^{-1}\boldsymbol{x}^* + (1-\theta_t^{-1})\boldsymbol{x}^t\Big) \le \theta_t^{-1}F(\boldsymbol{x}^*) + (1-\theta_t^{-1})F(\boldsymbol{x}^t) = \theta_t^{-1}F^{opt} + (1-\theta_t^{-1})F(\boldsymbol{x}^t)$$

$$\Rightarrow F\Big(\theta_t^{-1}\boldsymbol{x}^* + (1-\theta_t^{-1})\boldsymbol{x}^t\Big) - F(\boldsymbol{x}^{t+1}) \le (1-\theta_t^{-1})(F(\boldsymbol{x}^t) - F^{opt}) - (F(\boldsymbol{x}^{t+1}) - F^{opt})$$

Combining this with (7) (last equation) and $\theta_t^2 - \theta_t = \theta_{t-1}^2$ yields

$$\begin{aligned}\frac{L}{2}(\|\boldsymbol{u}^t\|_2^2 - \|\boldsymbol{u}^{t+1}\|_2^2) &\ge \theta_t^2(F(\boldsymbol{x}^{t+1}) - F^{opt}) - (\theta_t^2 - \theta_t)(F(\boldsymbol{x}^t) - F^{opt}) \\ &= \theta_t^2(F(\boldsymbol{x}^{t+1}) - F^{opt}) - \theta_{t-1}^2(F(\boldsymbol{x}^t) - F^{opt}),\end{aligned}$$

thus finishing the proof.

Proof of Theorem 3. With Lemma 2, one has

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

To bound the RHS of this inequality, we use Lemma 1 and $\boldsymbol{y}^0 = \boldsymbol{x}^0$ $(\boldsymbol{y}^+ = \boldsymbol{x}^1)$ and take $\boldsymbol{x} = \boldsymbol{x}^*$ to get

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

$$\Leftrightarrow \|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_1^2 + \frac{2}{L}(F(\boldsymbol{x}^*) - F^{opt}) \leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2$$

As a result,

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

Hence,

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

Lower bounds

Interestingly, no first-order methods can improve upon Nesterov's results in general. More precisely, $\exists$ convex and L-smooth function f s.t.

$$f(\boldsymbol{x}^t) - f^{opt} \geq \frac{3L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{32(t+1)^2},$$

as long as $\underbrace{\boldsymbol{x}^k \in \boldsymbol{x}^0 + span\{\nabla f(\boldsymbol{x}^0), \cdots, \nabla f(\boldsymbol{x}^{k-1})\}}_{\text{def. of first-order methods}}$ for all $1 \leq k \leq t$.

Example

Consider $\min_{\boldsymbol{x}\in\mathbb{R}^{(2n+1)}} \frac{L}{4}\left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{e}_1^\top \boldsymbol{x}\right)$ where

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

Note that f is convex and L-smooth and the optimizer $\boldsymbol{x}^*$ is given by $x_i^* = 1 - \frac{i}{2n+2}(1 \le i \le n)$ obeying

$$f^{opt} = \frac{L}{8}\left(\frac{1}{2n+2} - 1\right) \quad and \quad \|\boldsymbol{x}^*\|_2^2 \le \frac{2n+2}{3}.$$

Example

Also, $\nabla f(\boldsymbol{x}) = \frac{L}{4}\boldsymbol{A}\boldsymbol{x} - \frac{L}{4}\boldsymbol{e}_1$ and $\underbrace{span\{\nabla f(\boldsymbol{x}^0), \cdots, \nabla f(\boldsymbol{x}^{k-1})\}}_{:=\mathcal{K}_k} = span\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k\}$ if $\boldsymbol{x}^0 = 0$. That is, every iteration of first-order methods expands the search space by at most one dimension.

If we start with $\boldsymbol{x}^0 = 0$, then

$$f(\boldsymbol{x}^n) \geq \inf_{\boldsymbol{x}\in\mathcal{K}_n} f(\boldsymbol{x}) = \frac{L}{8}\left(\frac{1}{n+1} - 1\right) \Rightarrow \frac{f(\boldsymbol{x}^n) - f^{opt}}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2} \geq \frac{\frac{L}{8}\left(\frac{1}{n+1} - \frac{1}{2n+2}\right)}{\frac{1}{3}(2n+2)} = \frac{3L}{32(n+1)^2}.$$

Numerical example

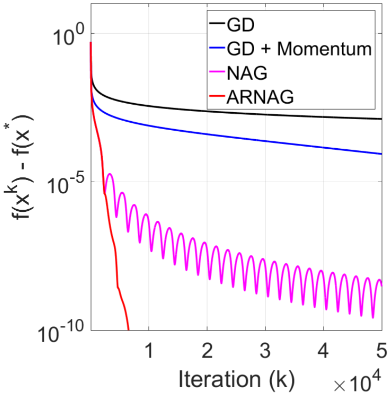
Consider

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{L} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{e}_1,$$

where

$$\boldsymbol{L} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}_{1000\times 1000},$$

and $\boldsymbol{e}_1$ is a 1000-dim vector whose first entry is 1 and all the other entries are 0.



Nesterov's acceleration is not a monotonic method! We can further accelerated it via restart, resulting in linear convergence with some further assumption. See V. Roulet, A. d'Aspremont, "Sharpness, Restart and Acceleration", NeurIPS 2017.

Nesterov's method for strongly convex problems

$$
\begin{aligned}
\boldsymbol{x}^{t+1} &= prox_{\eta_t h}(\boldsymbol{y}^t - \eta_t \nabla f(\boldsymbol{x}^t)) \\
\boldsymbol{y}^{t+1} &= \boldsymbol{x}^{t+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)
\end{aligned}
$$

Theorem 4. [Convergence of accelerated proximal gradient methods for strongly convex case] Suppose f is μ-strongly convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$
F(\boldsymbol{x}^t) - F^{opt} \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \left(F(\boldsymbol{x}^0) - F^{opt} + \frac{\mu \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{2}\right).
$$