Acceleration

Ernest K. Ryu and Wotao Yin

Large-Scale Convex Optimization via Monotone Operators

Acceleration

Theorem 1 establishes a $\mathcal{O}(1/k)$ rate on the squared norm of the fixed-point residual and a similar $\mathcal{O}(1/k)$ rate can be established for the other setups. This rate can be improved (at least in the worst-case rate).

In optimization, an acceleration is a modification of a base method that improves the convergence rate. An improvement from $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$ is most common for first-order algorithms.

Accelerations is an active topic of research. We keep the discussion minimal and discuss Nesterov's AGM and APPM/OHM.

Accelerated gradient method

Consider

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x),$$

where f is convex and L-smooth. The method

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \frac{k-1}{k+2}(x^{k+1} - x^k),$$

where $x^0 = y^0$, is Nesterov's accelerated gradient method (AGM).

Theorem 17.

Assume the convex, L-smooth function f has a minimizer $x^\star$. Then AGM converges with the rate

$$f(x^k) - f(x^\star) \leq \frac{2L\|x^0 - x^\star\|^2}{k^2}$$

for $k = 1, 2, \ldots$.

Proof of Theorem 17

Equivalent form of AGM:

$$
\begin{aligned}
x^{k+1} &= y^k - \frac{1}{L}\nabla f(y^k) \\
z^{k+1} &= z^k - \frac{k+1}{2L}\nabla f(y^k) \\
y^{k+1} &= \left(1 - \frac{2}{k+2}\right) x^{k+1} + \frac{2}{k+2} z^{k+1},
\end{aligned}
$$

where $x^0 = y^0 = z^0$. (cf. Exercise 12.1).

Proof of Theorem 17

Preliminary observations. Define

$$\theta_k = \frac{k+1}{2}$$

for $k = -1, 0, 1, \ldots$. It is straightforward to verify

$$\theta_k^2 - \theta_k \leq \theta_{k-1}^2 \tag{1}$$

for $k = 0, 1, \ldots$. We will use the inequalities

$$f(x^{k+1}) - f(y^k) + \frac{1}{2L}\|\nabla f(y^k)\|^2 \leq 0 \tag{2}$$

$$f(y^k) - f(x^k) \leq \langle \nabla f(y^k), y^k - x^k \rangle \tag{3}$$

$$f(y^k) - f(x^\star) \leq \langle \nabla f(y^k), y^k - x^\star \rangle. \tag{4}$$

The first, (2), follows from L-smoothness, which implies $f(x) - \frac{L}{2}\|x - y^k\|^2$ is concave as a function of x, which in turn implies

$$f(x) - \frac{L}{2}\|x - y^k\|^2 \leq f(y^k) + \langle \nabla f(y^k), x - y^k \rangle.$$

We plug in $x = x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$ to get (2). The second and third inequalities, (3) and (4), follow from convexity of f.

Proof of Theorem 17

Define

$$V^k = \theta_{k-1}^2 \left(f(x^k) - f(x^\star)\right) + \frac{L}{2}\|z^k - x^\star\|^2.$$

If we establish $V^k \leq V^{k-1} \leq \cdots \leq V^0$, then $V^k \leq V^0$ implies

$$\theta_{k-1}^2(f(x^k) - f(x^\star)) \leq V^k \leq V^0 = \frac{2L}{k^2}\|z^0 - x^\star\|^2.$$

Proof of Theorem 17

$$
\begin{aligned}
&V^{k+1} - V^k \\
&= \theta_k^2 \left(f(x^{k+1}) - f(x^\star) + \frac{1}{2L}\|\nabla f(y^k)\|^2 \right) - \theta_{k-1}^2 (f(x^k) - f(x^\star)) \\
&\qquad - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&\overset{(2)}{\leq} \theta_k^2 \left(f(y^k) - f(x^\star) \right) - \theta_{k-1}^2 (f(x^k) - f(x^\star)) - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&= (\theta_k^2 - \theta_k)(f(y^k) - f(x^k)) + \theta_k (f(y^k) - f(x^k)) + (\theta_k^2 - \theta_{k-1}^2)(f(x^k) - f(x^\star)) \\
&\qquad - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&\overset{(1)}{\leq} (\theta_k^2 - \theta_k)(f(y^k) - f(x^k)) + \theta_k (f(y^k) - f(x^\star)) - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&\overset{(3),(4)}{\leq} (\theta_k^2 - \theta_k) \langle \nabla f(y^k), y^k - x^k \rangle + \theta_k \langle \nabla f(y^k), y^k - x^\star \rangle - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&= \theta_k \langle \nabla f(y^k), (1-\theta_k) x^k + \theta_k y^k - z^k \rangle \overset{\text{def. of } z^k}{=} 0,
\end{aligned}
$$

where the first equality follows from

$$
\frac{L}{2} \left\| z^k - x^\star - \frac{\theta_k}{L} \nabla f(y^k) \right\|^2 - \frac{L}{2} \|z^k - x^\star\|^2 = -\theta_k \langle \nabla f(y^k), z^k - x^\star \rangle + \frac{\theta_k^2}{2L} \|\nabla f(y^k)\|^2.
$$

□

Comparison with gradient descent

Gradient descent with stepsize $\alpha = 1/L$

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$$

converges at rate $\mathcal{O}(1/k)$. To see this, define

$$V^k = k(f(x^k) - f(x^\star)) + \frac{L}{2}\|x^k - x^\star\|^2$$

and note

$$\begin{aligned}
&V^{k+1} - V^k \\
&= (k+1)(f(x^{k+1}) - f(x^k)) + f(x^k) - f(x^\star) + \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x^k)\|^2 - \frac{2}{L}\langle \nabla f(x^k), x^k - x^\star\rangle\right) \\
&\leq -\frac{k+1}{2L}\|\nabla f(x^k)\|^2 + \langle \nabla f(x^k), x^k - x^\star\rangle + \frac{1}{2L}\|\nabla f(x^k)\|^2 - \langle \nabla f(x^k), x^k - x^\star\rangle \\
&= -\frac{k}{2L}\|\nabla f(x^k)\|^2 \leq 0,
\end{aligned}$$

where inequality follows from analogues of (2) and (4). $V^k \leq V^0$ implies

$$f(x^k) - f(x^\star) \leq \frac{L}{2k}\|x^0 - x^\star\|^2 - \frac{L}{2k}\|x^k - x^\star\|^2 \leq \frac{L}{2k}\|x^0 - x^\star\|^2.$$

Constructing the Lyapunov function

The non-increasing quantity V^k is called a Lyapunov function, energy function, or potential function, and the style of proof relying on such quantities is called a Lyapunov analysis. Convergence proofs based on Lyapunov analyses tend to be more concise. Constructing a V^k is a highly non-trivial art, and we briefly outline the process for GD and AGM.

Imagine analyzing GD, and we suspect the convergence rate is $f(x^k) - f(x^\star) = \mathcal{O}(1/k)$. We define $W^k = k(f(x^k) - f(x^\star))$ and, through some analysis, find

$$W^{k+1} - W^k \leq \frac{L}{2}\|x^k - x^\star\|^2 - \frac{L}{2}\|x^{k+1} - x^\star\|^2.$$

So we define $V^k = k(f(x^k) - f(x^\star)) + \frac{L}{2}\|x^k - x^\star\|^2$ and present a Lyapunov analysis.

Constructing the Lyapunov function

We may try to prove a faster rate for GD by defining $W^k = t_k^2(f(x^k) - f(x^\star))$ with a yet unspecified t_k-sequence and analyzing $W^{k+1} - W^k$. If $t_k = \mathcal{O}(k)$, then perhaps we can establish an $O(1/k^2)$ rate. However, such an effort does not leads nowhere.

For AGM, we again define $W^k = t_{k-1}^2(f(x^k) - f(x^\star))$ and analyze $W^{k+1} - W^k$. We can show

$$W^{k+1} - W^k \leq \frac{L}{2}\|z^k - x^\star\|^2 - \frac{L}{2}\|z^{k+1} - x^\star\|^2.$$

for $t_k^2 - t_k \leq t_{k-1}^2$ and $t_k \geq 0$. An admissible sequence is $t_k = (k+1)/2$. So we define $V^k = t_k^2(f(x^k) - f(x^\star)) + \frac{L}{2}\|z^k - x^\star\|^2$ and present a Lyapunov analysis.

Accelerated proximal point and optimized Halpern

Consider

$$\underset{x\in\mathbb{R}^n}{\text{find}} \quad 0 \in \mathbb{A}x,$$

where $\mathbb{A}$ is maximal monotone. The method

$$y^{k+1} = \mathbb{J}_{\mathbb{A}} x^k$$
$$x^{k+1} = y^{k+1} + \frac{k}{k+2}(y^{k+1} - y^k) - \frac{k}{k+2}(y^k - x^{k-1}),$$

where $y^0 = x^0$, is the accelerated proximal point method (APPM).

Also consider

$$\underset{x\in\mathbb{R}^n}{\text{find}} \quad x = \mathbb{T}x,$$

where $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive. We call

$$x^{k+1} = \frac{1}{k+2}x^0 + \frac{k+1}{k+2}\mathbb{T}x^k$$

the optimized Halpern method (OHM).

Accelerated proximal point and optimized Halpern

With $\mathbb{T} = \mathbb{R}_{\mathbb{A}}$, finding elements of $\operatorname{Zer} \mathbb{A}$ and $\operatorname{Fix} \mathbb{T}$ are equivalent (cf. Exercise 10.1). The two methods APPM and OHM are equivalent (cf. Exercise 12.2).

Theorem 18.

Assume the maximal monotone operator $\mathbb{A}$ has a zero $x^\star$. Then APPM/OHM converges with the rate

$$\|x^{k-1} - \mathbb{J}_{\mathbb{A}} x^{k-1}\|^2 \leq \frac{\|x^0 - x^\star\|^2}{k^2}$$

for $k = 1, 2, \ldots$.

We can equivalently state this result as

$$\|\mathbb{T} x^{k-1} - x^{k-1}\|^2 \leq \frac{4\|x^0 - x^\star\|^2}{k^2}.$$

Proof of Theorem 18

Define $\tilde{\mathbb{A}}y^k = x^{k-1} - y^k$, which implies $\tilde{\mathbb{A}}y^k \in \mathbb{A}y^k$. Define

$$V^k = k^2\|\tilde{\mathbb{A}}y^k\|^2 + k\langle\tilde{\mathbb{A}}y^k, y^k - x^0\rangle$$

for $k = 0, 1, \ldots$.

Then

$$V^{k+1} - V^k = -k(k+1)\langle\tilde{\mathbb{A}}y^{k+1} - \tilde{\mathbb{A}}y^k, y^{k+1} - y^k\rangle,$$

which can be verified by plugging in

$$y^{k+1} = \frac{1}{k+1}x^0 + \frac{k}{k+1}(y^k - \tilde{\mathbb{A}}y^k) - \tilde{\mathbb{A}}y^{k+1}$$

and performing basic (albeit somewhat tedious) calculations to check that all terms vanish. By monotonicity of $\mathbb{A}$, we have $V^{k+1} \leq V^k$.

Proof of Theorem 18

Since $V^k \leq V^0 = 0$, we have

$$\begin{aligned}
0 &\geq V^k \\
&= k^2\|\tilde{\mathbb{A}}y^k\|^2 + k\langle\tilde{\mathbb{A}}y^k, x^\star - x^0\rangle + k\langle\tilde{\mathbb{A}}y^k, y^k - x^\star\rangle \\
&= \frac{k^2}{2}\|\tilde{\mathbb{A}}y^k\|^2 - \frac{1}{2}\|x^\star - x^0\|^2 + \frac{1}{2}\|k\tilde{\mathbb{A}}y^k + x^\star - x^0\|^2 + k\langle\tilde{\mathbb{A}}y^k, y^k - x^\star\rangle \\
&\geq \frac{k^2}{2}\|\tilde{\mathbb{A}}y^k\|^2 - \frac{1}{2}\|x^\star - x^0\|^2,
\end{aligned}$$

where the second equality follows from

$$k\langle\tilde{\mathbb{A}}y^k, x^\star - x^0\rangle = \frac{1}{2}\|k\tilde{\mathbb{A}}y^k + x^\star - x^0\|^2 - \frac{k^2}{2}\|\tilde{\mathbb{A}}y^k\|^2 - \frac{1}{2}\|x^\star - x^0\|^2$$

and the final inequality follows from monotonicity. □

When does an acceleration accelerate?

In optimization (and more generally in applied mathematics and computer science), convergence rates are usually established in the worst case. If an unaccelerated method actually converges at an $\mathcal{O}(1/k)$ rate, then an $\mathcal{O}(1/k^2)$ acceleration is a speedup. However, if the observed convergence is already faster than $\mathcal{O}(1/k^2)$, then there is no guarantee of improvement. The acceleration may even slow down the convergence.

In practice, an acceleration sometimes provides a speedup. An "acceleration" should be tried it out with the expectation that it may improve or worsen the convergence.