Optimization for Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Outline

Composite problems

$$min_{\mathbf{x}} F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$$

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

Proximal gradient descent

Define the proximal operator

$$\mathrm{prox}_h(\mathbf{x}) \triangleq \arg\min_{\mathbf{z}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + h(\mathbf{z}) \right\}$$

for any convex function h.

In each iteration, the proximal gradient descent method computes

$$\mathbf{x}_{t+1} = \mathrm{prox}_{\eta_t h}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)).$$

Convergence analysis

Lemma

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

$$F(\mathbf{y}^+) - F(\mathbf{x}) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 - \frac{L}{2}\left\|\mathbf{x} - \mathbf{y}^+\right\|_2^2 - g(\mathbf{x}, \mathbf{y})$$

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

- Take $\mathbf{x} = \mathbf{y} = \mathbf{x}_t$, we get $F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t)$.
- Take $\mathbf{x} = \mathbf{x}^*$, $\mathbf{y} = \mathbf{x}_t$, we get $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|_2$.

Convergence for convex problems

Suppose f is convex and L-smooth. The proximal graident descent with stepsize $\eta_t = 1/L$ obeys

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \frac{L \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2^2}{2t}.$$

- Achieves better iteration complexity ($O(1/\varepsilon)$) than subgradient method ($O(1/\varepsilon^2)$).

Convergence for strongly convex problems

Suppose f is μ-strongly convex and L-smooth. The proximal graident descent with stepsize $\eta_t = 1/L$ obeys

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \le \left(1 - \frac{\mu}{L}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 .$$

- Achieves linear convergence $O(\kappa \log \frac{1}{\varepsilon})$.

Outline

(Proximal) gradient methods

Iteration complexities of (proximal) gradient methods

- strongly convex and smooth problems

$$O\left(\kappa \log \frac{1}{\epsilon}\right)$$

- convex and smooth problems

$$O\left(\frac{1}{\epsilon}\right)$$

Can one still hope to further accelerate convergence?

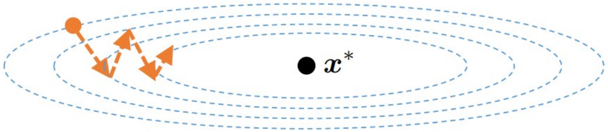
Polyak's heavy-ball method

Heavy ball Method (HB):

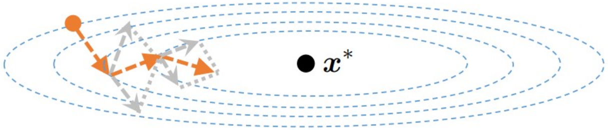
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) + \underbrace{\theta_t(\mathbf{x}_t - \mathbf{x}_{t-1})}_{\text{momentum term}}$$

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

Polyak's heavy-ball method



gradient descent



heavy-ball method

Polyak's heavy-ball method

Theorem (Convergence of heavy ball methods)

Suppose f is a L-smooth and μ-strongly convex quadratic function. If we choose $\eta_t = 4/(\sqrt{L} + \sqrt{\mu})^2$, $\theta_t = \max\{|1 - \sqrt{\eta_t L}|, |1 - \sqrt{\eta_t \mu}|\}^2$ and $\kappa = L/\mu$, then

$$\left\| \begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}^* \\ \mathbf{x}_t - \mathbf{x}^* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^t \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}^* \\ \mathbf{x}_0 - \mathbf{x}^* \end{bmatrix} \right\|_2$$

- only have convergence guarantee for quadratic function
- significant improvement over GD: $O\left(\sqrt{\kappa} \log \frac{1}{\epsilon}\right)$ v.s. $O\left(\kappa \log \frac{1}{\epsilon}\right)$

Can we obtain improvement for more general convex cases as well?

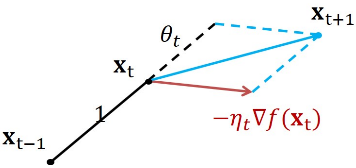
Nesterov's idea

Nesterov's accelerated gradient (NAG) method:

$$\mathbf{y}_t = \mathbf{x}_t + \theta_t(\mathbf{x}_t - \mathbf{x}_{t-1})$$
$$\mathbf{x}_{t+1} = \mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t)$$

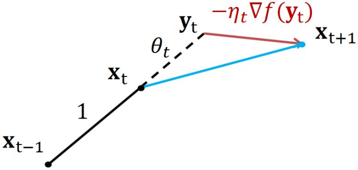
- alternates between gradient updates and proper extrapolation
- not a descent method (i.e. we may not have $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$)
- one of the most beautiful and mysterious results in optimization

Comparison between HB and NAG



Heavy ball

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) + \theta_t(\mathbf{x}_t - \mathbf{x}_{t-1})$$



NAG

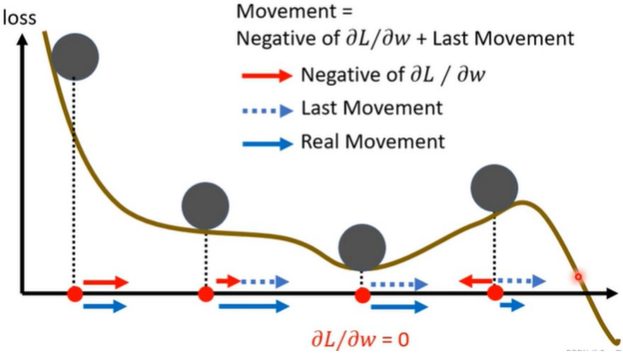
$$\begin{cases} \mathbf{y}_t = \mathbf{x}_t + \theta_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \\ \mathbf{x}_{t+1} = \mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t) \end{cases}$$

History

- Polyak invented HB momentum in 1964 (and discussed the physics analogy)
- Nesterov invented NAG in 1983
  - Even though Nesterov was Polyak's student, he seems not to have mentioned the physics analogy
- Sutskever et al. (2013)[1] popularized momentum methods in machine learning and revived the momentum interpretation.

[1] On the importance of initialization and momentum in deep learning. ICML 2013.

Momentum methods for nonconvex problems



Convergence rate of NAG

Suppose f is μ-strongly convex and L-smooth. If we choose $\eta_t = \eta = 1/L$ and $\theta_t = \theta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, then

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{t-1} \left[f(\mathbf{x}_1) - f(\mathbf{x}^*) + \frac{\mu}{2} \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2\right].$$

Convergence rate of NAG

Suppose f is convex and L-smooth. If we choose $\eta_t = \eta = 1/L$ and $\theta_t = \frac{\lambda_t - 1}{\lambda_{t+1}}$ where $\lambda_0 = 1$ and $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$. Then

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{2L \left\| \mathbf{x}_1 - \mathbf{x}^* \right\|_2^2}{(t+1)^2}.$$

- A simpler choice the θ_t is $\theta_t = \frac{t}{t+3}$.

Extension to composite models

Fast iterative shrinkage-thresholding algorithm (FISTA, Beck & Teboulle '09):

$$\begin{aligned}\mathbf{y}_t &= \mathrm{prox}_{\eta_t h}(\mathbf{x}_t + \theta_t(\mathbf{x}_t - \mathbf{x}_{t-1})) \\ \mathbf{x}_{t+1} &= \mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t)\end{aligned}$$

- has same convergence property as the convex problems
- fast if prox can be efficiently implemented

Outline

Lower bounds

Interestingly, no first-order methods can improve upon Nesterov's results in general.

More precisely, there exists convex and L-smooth function f s.t.

$$f(\mathbf{x}_t) - f^* \geq \frac{3L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{32(t+1)^2}$$

as long as $\underbrace{\mathbf{x}_k \in \mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{k-1})\}}_{\text{definition of first-order methods}}$ for all $1 \leq k \leq t$.

Example

$$\min_{\mathbf{x}\in\mathbb{R}^{2n+1}} f(\mathbf{x}) = \frac{L}{4}\left(\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{e}_1^\top \mathbf{x}\right)$$

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

- f is convex and L-smooth
- the optima $\mathbf{x}^*$ is given by $x_i^* = 1 - \frac{i}{2n+2}(1 \le i \le n)$.

Lower bounds for strongly convex functions

There exists μ-strongly convex and L-smooth function f s.t.

$$f(\mathbf{x}_t) - f^* \geq \frac{\mu}{4} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2t} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

as long as $\underbrace{\mathbf{x}_k \in \mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{k-1})\}}_{\text{definition of first-order methods}}$ for all $1 \leq k \leq t$.

Example

$$\min_{\mathbf{x}\in\mathbb{R}^{2n+1}} f(\mathbf{x}) = \frac{L-\mu}{4}\left(\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{e}_1^\top\mathbf{x}\right) + \frac{\mu}{2}\|\mathbf{x}\|_2^2$$

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

- f is μ-strongly convex and L-smooth
- the optima $\mathbf{x}^*$ is given by $x_i^* = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i (1 \le i \le n)$.