CSED490Y: Optimization for Machine Learning

Week 06-1: Proximal gradient descent

Namhoon Lee

POSTECH

Spring 2022

Admin

– 5%

Project proposal

- Due: 11 April (11:59 PM)
- Contents: – Be specific
  - title
  - introduction / motivation / problem / background
  - idea
  - expected result
  - plan / timeline
- Format: LaTeX (max 2 pages)

Submit on PLMS (a link will be created soon).

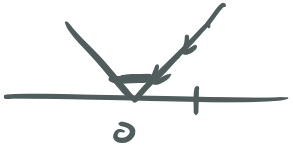
Gradient descent vs subgradient method

$f(x) = 3x^2 + 4x - 2$

$\nabla f(x) = 6x + 4$

$f(x) = |x|$

$g_k = \{-1, 1\}$



Differences between gradient descent and subgradient method:

- Gradient descent improves at every iteration, unlike sub-gradient method.
- Gradient descent can take a big step size: self-tuning property.
- Gradient descent takes bigger steps when far away.

Convergence of subgradient method

Let f be convex and G-Lipschitz continuous:

Subgradient method: $x_{t+1} = x_t - \eta g_t$, $g_t \in \partial f(x_t)$

Convergence:

$$\|x_{t+1} - x^*\|_2^2 = \|x_t - \eta g_t - x^*\|_2^2$$

$$= \|x_t - x^*\|_2^2 - 2\eta \langle g_t, x_t - x^* \rangle + \eta^2 \|g_t\|_2^2$$

$$\leq \|x_t - x^*\|_2^2 - 2\eta (f(x_t) - f(x^*)) + \eta^2 G^2$$

Convergence of subgradient method

Jensen's inequality

$$\|x_{t+1} - x^*\|_2^2 \le \|x_t - x^*\|_2^2 - 2\eta(f(x_t) - f(x^*)) + \eta^2 G^2$$

(a): avg of fn
vs
(b): fn of avg

Convergence (cont'd):

$$f(x_t) - f(x^*) \le \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\eta}{2}G^2$$

(a) $$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \le \frac{1}{2\eta}\cdot\frac{1}{T}\left(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2\right) + \frac{\eta G^2}{2}$$

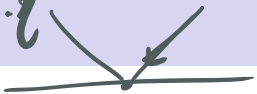
(b) $$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \le \frac{1}{2\eta}\cdot\frac{1}{T}\|x_1 - x^*\|_2^2 + \frac{\eta G^2}{2}$$

$\le R$

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2}$$

Convergence of subgradient method

Take derivative w.r.t. η, set it to be 0



Discussion on η: $f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} \sim \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{T}}$

η fixed → subgradient method doesn't converge to min.

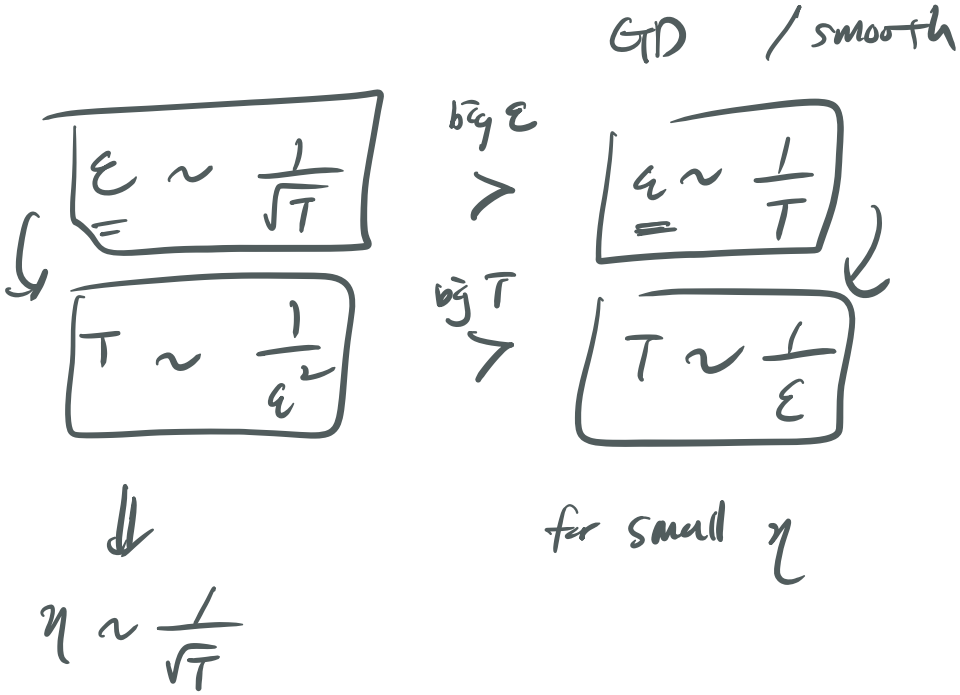
η NOT fixed → η big 2nd term big

η small 1st term big $\eta \sim \frac{1}{\sqrt{T}}$

$\boxed{\eta \sim \frac{1}{\sqrt{T}}}$ $-\frac{R^2}{2T\eta^2} + \frac{G^2}{2} = 0 \Leftrightarrow \frac{R^2}{2T} = \frac{G^2}{2}\eta^2 \Leftrightarrow \eta = \sqrt{\frac{R^2}{G^2 T}}$

Subgradient method

Summary:



Strong convexity → quadratic lower bound

f is α-strongly convex, if $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex

$$g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle$$

$$f(y) - \frac{\alpha}{2}\|y\|_2^2 \geq f(x) - \frac{\alpha}{2}\|x\|_2^2 + \langle \nabla f(x) - \alpha x, y - x \rangle$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2$$

What does strong convexity imply?

$f(x) = x^T Q x$

$\nabla^2 f(x) = 2Q$

f twice differentiable: lower bound on Hessian

H eigenvalues $\geq \alpha$

$\alpha = 2 \cdot \lambda_{min}(Q)$

$$\|\nabla^2 f(x)\| \geq \alpha I$$

Strong convexity

A bound on suboptimality of any point: if f is α-strongly convex,

$$\frac{\alpha}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

(a) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2$ ←

$f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|_2^2$ ■

(b) min of $f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2$

$0 + \nabla f(x) + \alpha(y^* - x) = 0$

$y^* = x - \frac{1}{\alpha}\nabla f(x)$

Strong convexity

Coercivity: for f α-strongly convex,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|_2^2$$

(proof – from monotonicity)

f α-strongly convex $\Rightarrow$ $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex $\Rightarrow$ monotone.

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq 0$$

$$\langle \nabla f(x) - \alpha x - (\nabla f(y) - \alpha y), x - y \rangle \geq 0$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|_2^2 \quad \blacksquare$$

Smoothness

Recall β-smoothness: $\|\nabla f(x) - \nabla f(y)\|_2 \le \beta \|x - y\|_2$

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|_2^2$$

Smoothness

A bound on suboptimality of any point: if f is β-smooth,

$$\frac{1}{2\beta}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{\beta}{2}\|x - x^*\|_2^2$$

(a)

(b)

Smoothness

Co-coercivity: for f β-smooth convex,

$$\frac{1}{2\beta}\|\nabla f(x)\|_2^2 \le f(x) - f(x^*)$$

① + ② $\Rightarrow$ $\langle \nabla f(x) - \nabla f(y), x - y\rangle \ge \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|_2^2$

(proof) $f_x(z) = f(z) - \langle \nabla f(x), z\rangle$ $\quad z^* = x$

$f_y(z) = f(z) - \langle \nabla f(y), z\rangle$ $\quad z^* = y$

① $f(y) - (f(x) + \langle \nabla f(x), y - x\rangle) = f(y) - \langle \nabla f(x), y\rangle - (f(x) - \langle \nabla f(x), x\rangle)$

② $f(x) - (f(y) + \langle \nabla f(y), x - y\rangle) = f_x(y) - f_x(x) \ge \frac{1}{2\beta}\|\nabla f_x(y)\|_2^2$

$\ge \frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|_2^2$ $\quad = \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_2^2$

Smoothness

Extension of co-coercivity: for f α-strongly convex and β-smooth,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x-y\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(x) - \nabla f(y)\|_2^2$$

First, $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is $(\beta - \alpha)$-smooth.

Smoothness

Extension of co-coercivity (cont'd):

Convergence of GD for smooth and strongly convex functions

Summary

Thank you

Any questions?

Credits

A lot of material in this course is borrowed or derived from the following:

- Numerical Optimization, Jorge Nocedal and Stephen J. Wright.
- Convex Optimization, Stephen Boyd and Lieven Vandenberghe.
- Convex Optimization, Ryan Tibshirani.
- Optimization for Machine Learning, Martin Jaggi and Nicolas Flammarion.
- Optimization Algorithms, Constantine Caramanis.
- Advanced Machine Learning, Mark Schmidt.