CSED490Y: Optimization for Machine Learning

Week 07-2: Proximal gradient descent

Namhoon Lee

POSTECH

Spring 2022

Proximal gradient method

Minimizing composite function:

differentiable

non-differentiable

$$\min_x f(x) = g(x) + h(x)$$

Proximal gradient:

$$x_{t+1} = \text{prox}_{\eta h}(x_t - \eta \nabla g(x_t))$$

How fast is it? → rate of convergence

Gradient mapping

Define gradient mapping:

- $G_\eta(x) \overset{\text{def.}}{=} \frac{1}{\eta}(x - \text{prox}_{\eta h}(x - \eta \nabla g(x)))$.

Then we can rewrite the proximal gradient method into something that looks more like a gradient descent update step:

- $x_{t+1} = x_t - \eta G_\eta(x_t)$.

- G_η is called the gradient map of proximal gradient method, and we treat this as if it's a gradient, but G_η is not a (sub)gradient of f in general.
- We do this to make analyzing convergence behavior easier.

Gradient mapping and optimality condition

$x = \text{prox}_{\eta h}(y) = \arg\min_x \frac{1}{2\eta}\|x-y\|_2^2 + h(x)$

$\Leftrightarrow \quad 0 \in \frac{1}{\eta}(x-y) + \partial h(x)$

$\Leftrightarrow \boxed{y - x \in \eta\, \partial h(x)}$

Optimal solutions are the only fixed points of the prox grad update

$G_\eta(x_t) \notin \partial f(x_t)$

$$x_{t+1} = x_t - \eta G_\eta(x_t)$$

i.e., $G_\eta(x) = 0$ when x is a minimizer of f.

$G_\eta(\hat{x}) = 0 \quad \Leftrightarrow \quad \frac{1}{\eta}(\hat{x} - \text{prox}_{\eta h}(\hat{x} - \eta \nabla g(\hat{x}))) = 0$

$\Leftrightarrow \quad \hat{x} = \text{prox}_{\eta h}(\hat{x} - \eta \nabla g(\hat{x}))$

$\Leftrightarrow \quad \hat{x} - \eta \nabla g(\hat{x}) - \hat{x} \in \eta\, \partial h(\hat{x})$

$\Leftrightarrow \quad 0 \in \nabla g(\hat{x}) + \partial h(\hat{x})$

$\min_x f(x) = g(x) + h(x)$

Key lemma

$\min_x f(x) = g(x) + h(x)$ — convex

g: β-smooth, α-strongly convex.

Lemma

$$f(x - \eta G_\eta(x)) \le f(z) + \langle G_\eta(x), x - z\rangle - \frac{\eta}{2}\|G_\eta(x)\|_2^2 - \frac{\alpha}{2}\|x - z\|_2^2$$

Setting $z = x_t$, $x = x_t$ gives

$$f(x_{t+1}) = f(x_t - \eta G_\eta(x_t)) \le f(x_t) - \frac{\eta}{2}\|G_\eta(x_t)\|_2^2$$

which looks the progress bound for smooth functions.

- $f(x_{t+1}) \le f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$

Key lemma – proof (1/3)

$$f(x-\eta G_\eta(x)) = \underbrace{g(x-\eta G_\eta(x))}_{\text{(A)}} + \underbrace{h(x-\eta G_\eta(x))}_{\text{(B)}} \leq \ ?$$

$\eta = \frac{1}{\beta}$

(A) (1) β-smooth : $f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|y-x\|_2^2$

$$g(x-\eta G_\eta(x)) \leq g(x) + \langle \nabla g(x), -\eta G_\eta(x) \rangle + \frac{\eta}{2}\|G_\eta(x)\|_2^2$$

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

$$g(x-\eta G_\eta(x)) \leq g(z) - \langle \nabla g(x), z-x \rangle - \frac{\alpha}{2}\|z-x\|_2^2 + \langle \nabla g(x), -\eta G_\eta(x) \rangle + \frac{\eta}{2}\|G_\eta(x)\|_2^2$$

Key lemma – proof (2/3)

(3): $h(x - \eta G_\eta(x))$

(1) $x - \eta G_\eta(x) = x - \eta \frac{1}{\eta}(x - \mathrm{prox}_{\eta h}(x - \eta \nabla g(x))) = \mathrm{prox}_{\eta h}(x - \eta \nabla g(x))$

$x - \eta G_\eta(x) = \mathrm{prox}_{\eta h}(x - \eta \nabla g(x))$

$\Leftrightarrow x - \eta \nabla g(x) - (x - \eta G_\eta(x)) \in \eta \partial h(x - \eta G_\eta(x))$

$\Leftrightarrow G_\eta(x) - \nabla g(x) \in \partial h(x - \eta G_\eta(x))$

(2) $h(x - \eta G_\eta(x)) \le h(z) - \langle G_\eta(x) - \nabla g(x), z - (x - \eta G_\eta(x)) \rangle$

Key lemma – proof (3/3)

$f(x) = \underbrace{g(x)}_{\text{A}} + \underbrace{h(x)}_{\text{B}}$

$$\text{A} = g(x - \eta G_\eta(x)) \le \underbrace{g(z)}_{①} - \underbrace{\langle \nabla g(x), z - x \rangle}_{④} - \underbrace{\frac{\alpha}{2}\|z - x\|_2^2}_{⑥} + \underbrace{\langle \nabla g(x), -\eta G_\eta(x) \rangle}_{⑤} + \underbrace{\frac{\eta}{2}\|G_\eta(x)\|_2^2}_{③}$$

$$\text{B} = h(x - \eta G_\eta(x)) \le \underbrace{h(z)}_{①} - \underbrace{\langle G_\eta(x) - \nabla g(x)}_{②⑤}, \underbrace{z - (x - \eta G_\eta(x))}_{②④} \underbrace{\rangle}_{③⑤} \quad (④①)$$

$$\text{A} + \text{B} \le f(z) + \langle G_\eta(x), x - z \rangle - \frac{\eta}{2}\|G_\eta(x)\|_2^2 - \frac{\alpha}{2}\|x - z\|_2^2 \qquad \blacksquare$$

Convergence for smooth functions

g : β-smooth. $\quad d=0$

$$f(x-\eta G_\eta(x)) \le f(z) + \langle G_\eta(x), x-z\rangle - \frac{\eta}{2}\|G_\eta(x)\|_2^2 - \frac{\alpha}{2}\|x-z\|_2^2$$

substitute $x \to x_t$, $z \to x^*$

$$f(x_{t+1}) \le f(x^*) + \langle G_\eta(x_t), x_t - x^*\rangle - \frac{\eta}{2}\|G_\eta(x_t)\|_2^2$$

$$\Leftrightarrow \quad f(x_{t+1}) - f(x^*) \le \langle G_\eta(x_t), x_t - x^*\rangle - \frac{\eta}{2}\|G_\eta(x_t)\|_2^2$$

$$= \frac{1}{2\eta}\left(2\eta\langle G_\eta(x_t), x_t - x^*\rangle - \eta^2\|G_\eta(x_t)\|_2^2 - \|x_t - x^*\|_2^2 + \|x_t - x^*\|_2^2\right)$$

sum over T iterations, both sides & divide by T

$$= \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_t - x^* - \eta G_\eta(x_t)\|_2^2\right)$$

$$\sim O\left(\frac{1}{T}\right) \qquad = \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)$$

Convergence for smooth and strongly convex functions

g – β-smooth & α-strongly convex $\qquad \min_x f(x) = g(x) + h(x)$

$$0 \le f(x_{t+1}) - f(x^*) \le \langle G_\eta(x_t), x_t - x^* \rangle - \frac{\eta}{2}\|G_\eta(x_t)\|_2^2 - \frac{\alpha}{2}\|x_t - x^*\|_2^2$$

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

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

$$\Leftarrow \|x_{t+1} - x^*\|_2^2 \le (1 - \eta\alpha)^t \|x_1 - x^*\|_2^2 \quad : \sim O(\rho^t)$$

$$R^2$$

Example: ISTA

L1-regularized least squares or Lasso:

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1$$

- Recall that this can achieve sparse solutions.
- Apply subgradient method? It will converge with $\mathcal{O}(1/\sqrt{t})$.
- f is a composite function of g smooth and h non-smooth functions, so consider using proximal gradient method.

Example: ISTA

(cont'd) Apply proximal gradient:

$$\begin{aligned} x_{t+1} &= \text{prox}_{\eta h}(y_{t+1}) \\ &= \arg\min_x \frac{1}{2\eta}\|x - y_{t+1}\|_2^2 + \lambda\|x\|_1 \end{aligned}$$

where $y_{t+1} = x_t - \eta\nabla g(x_t) = x_t - \eta(A^\top(Ax_t - y))$.

The solution to the prox will be

$$x^+ : x_i = \begin{cases} x_i + \eta\lambda & \text{for } y_i < -\eta\lambda \\ 0 & \text{for } |y_i| \leq \eta\lambda \\ x_i - \eta\lambda & \text{for } y_i > \eta\lambda \ . \end{cases}$$

"A fast iterative shrinkage-thresholding algorithm for linear inverse problems" (Beck and Teboulle 2009)

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.

References I

 Beck, Amir and Marc Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences*.