CSED490Y: Optimization for Machine Learning

Week 06-2: Proximal gradient descent

Namhoon Lee

POSTECH

Spring 2022

Admin

Reminder:

- Lectures on campus (Enginnering Bldg 2, Room 109) starting next week.
- Midterm exam on Monday 25 April.

Smoothness

Recall β-smoothness:

Smoothness

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

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

proof — directly from smoothness

Smoothness

Co-coercivity: for f β-smooth convex,

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

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

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

① + ② $\Rightarrow$

Smoothness

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

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

First, $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is $(\beta - \alpha)$-smooth. $\Rightarrow \langle \nabla g(x) - \nabla g(y), x-y \rangle \geq \frac{1}{\beta-\alpha}\|\nabla g(x) - \nabla g(y)\|_2^2$

convex

$$\left(\begin{aligned} & g(y) \leq g(x) + \langle \nabla g(x), y-x \rangle + \frac{\beta-\alpha}{2}\|y-x\|_2^2 \\ & f(y) - \frac{\alpha}{2}\|y\|_2^2 \leq f(x) - \frac{\alpha}{2}\|x\|^2 + \langle \nabla f(x) - \alpha x, y-x \rangle + \frac{\beta-\alpha}{2}\|y-x\|_2^2 \\ & f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|y-x\|_2^2 \quad - \quad \boxed{\beta\text{-smooth}} \end{aligned} \right)$$

Smoothness

$$g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$$

Extension of co-coercivity (cont'd):

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

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

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

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

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

Convergence of GD for smooth and strongly convex functions

(proof) $x_{t+1} = x_t - \eta \nabla f(x_t)$, $\boxed{\eta = \frac{2}{\alpha+\beta}}$

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

$$= \|x_t - x^*\|_2^2 - 2\eta \underline{\langle \nabla f(x_t) - \nabla f(x^*), x_t - x^* \rangle} + \eta^2 \|\nabla f(x_t)\|_2^2$$

$$\leq \underline{\|x_t - x^*\|_2^2} - 2\eta \left(\frac{\alpha\beta}{\alpha+\beta} \underline{\|x_t - x^*\|^2} + \frac{1}{\alpha+\beta} \|\underline{\nabla f(x_t)} - \underline{\nabla f(x^*)}\|_2^2 \right) + \eta^2 \|\underline{\nabla f(x_t)}\|_2^2$$

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

Convergence of GD for smooth and strongly convex functions

$\eta = \frac{2}{\alpha+\beta}$

(proof – cont'd)

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

$= \frac{(\beta-\alpha)^{\cancel{2}}}{(\beta+\alpha)^{\cancel{2}}} \boxed{\|x_t - x^*\|_2^{\cancel{2}}} \le \boxed{\frac{\beta-\alpha}{\beta+\alpha}\|x_{t-1} - x^*\|}$

$\|x_{t+1} - x^*\|_2 \le \left(\frac{\beta-\alpha}{\beta+\alpha}\right)^{t} \|x_1 - x^*\|_2$

< 1

$\boxed{f(x) = 3x^2 + 4x - 2}$

$|1 - 6\eta| < 1$

$\varepsilon \sim \rho^t \Longleftrightarrow t \sim \log(1/\varepsilon)$

Summary

f α-strongly convex & β-smooth

- $\varepsilon \sim \rho^t$ $\quad t \sim \log(1/\varepsilon)$ $\quad$ 1
- subgradient $\sim \frac{1}{\sqrt{T}}$ $\quad \sim \frac{1}{\varepsilon^2}$ $\quad$ 3
- GD smooth $\frac{1}{T}$ $\quad \frac{1}{\varepsilon}$ $\quad$ 2

Projected gradient method

So far we have seen unconstrained optimization problems:

$$\min_{x \in \mathbb{R}^d} f(x)$$

- any $x \in \mathbb{R}^n$ can be a solution.

Projected gradient method

So far we have seen unconstrained optimization problems:

$$\min_{x \in \mathbb{R}^d} f(x)$$

- any $x \in \mathbb{R}^n$ can be a solution.

For constrained optimization problem:

$$\min_{x \in \mathbb{C}} f(x)$$

x^*

- now x must be in the set $\mathbb{C}$.

Projected gradient method

So far we have seen unconstrained optimization problems:

$$\min_{x \in \mathbb{R}^d} f(x)$$

- any $x \in \mathbb{R}^n$ can be a solution.

For constrained optimization problem:

$$\min_{x \in \mathbb{C}} f(x)$$

- now x must be in the set $\mathbb{C}$.

GD is the standard way to solve the unconstrained optimization problems.

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

Q: Can we apply GD to solve the constrained optimization problem?

Projected gradient method

So far we have seen unconstrained optimization problems:

$$\min_{x \in \mathbb{R}^d} f(x)$$

- any $x \in \mathbb{R}^n$ can be a solution.

For constrained optimization problem:

$$\min_{x \in \mathbb{C}} f(x)$$

- now x must be in the set $\mathbb{C}$.

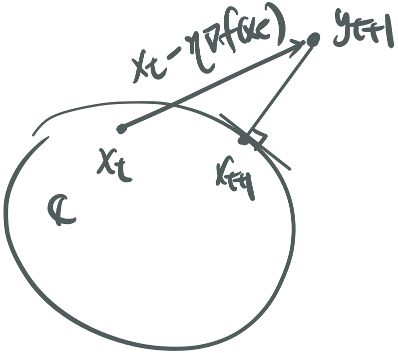
GD is the standard way to solve the unconstrained optimization problems.

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

Q: Can we apply GD to solve the constrained optimization problem?

Idea: use projection!

Projected gradient method



✓ Step 1: Update x_t by GD

$$\boxed{y_{t+1}} = x_t - \eta \nabla f(x_t)$$

Step 2: Project onto the set $\mathbb{C}$

$$\boxed{x_{t+1}} = \mathrm{proj}_{\mathbb{C}}(y_{t+1})$$

If the updated point gets outside $\mathbb{C}$, project it back to the set.

Projected gradient method

The projection operator $\boxed{\mathrm{proj}_{\mathbb{C}}(\cdot)}$ is an optimization problem by itself:

$$\underline{\mathrm{proj}_{\mathbb{C}}(x_0)} = \arg\min_{x \in \mathbb{C}} \frac{1}{2} \|x - \underline{x_0}\|_2^2$$

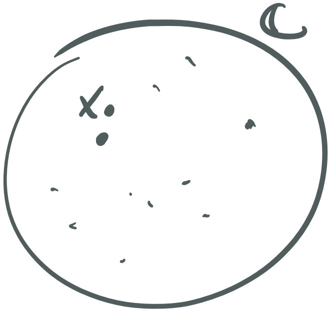
i.e., given a point x_0, find a point $x \in \mathbb{C}$ that is closest to x_0.

Projected gradient method

The projection operator $\mathrm{proj}_{\mathbb{C}}(\cdot)$ is an optimization problem by itself:

$$\mathrm{proj}_{\mathbb{C}}(x_0) = \arg\min_{x \in \mathbb{C}} \frac{1}{2}\|x - x_0\|_2^2$$

i.e., given a point x_0, find a point $x \in \mathbb{C}$ that is closest to x_0.



When $x_0 \in \mathbb{C}$:

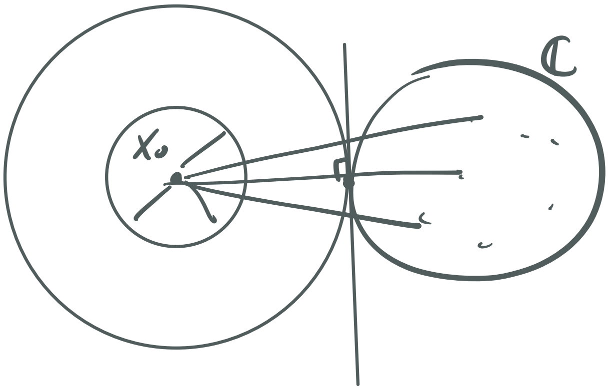
- The closest point to x_0 in $\mathbb{C}$ is x_0 itself.

Projected gradient method

The projection operator $\text{proj}_{\mathbb{C}}(\cdot)$ is an optimization problem by itself:

$$\boxed{\text{proj}_{\mathbb{C}}(x_0) = \arg\min_{x \in \mathbb{C}} \frac{1}{2}\|x - x_0\|_2^2}$$

i.e., given a point x_0, find a point $x \in \mathbb{C}$ that is closest to x_0.



When $x_0 \in \mathbb{C}$:

- The closest point to x_0 in $\mathbb{C}$ is x_0 itself.

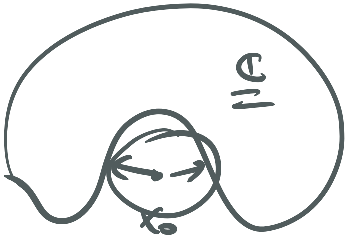
When $x_0 \notin \mathbb{C}$:

- The closest point to x_0 in $\mathbb{C}$ is is the point where the norm ball touches $\mathbb{C}$.

Projected gradient method

Projected gradient method:

$$x_{t+1} = \text{proj}_{\mathbb{C}}(x_t - \eta \nabla f(x_t))$$



Note:

- PGD has one more step than GD: the projection.
- PGD is an "economic" algorithm if the problem is easy to solve.
- If $\mathbb{C}$ is a convex set, the projection has a unique solution; otherwise the solution may not be unique.
- Projected gradient method is a special case of proximal gradient method.

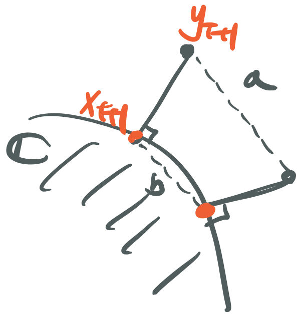
Convergence of projected subgradient method

$\|g_t\| \le G$

$\boxed{\min f(x) \quad \text{s.t. } x \in C}$

Recall subgradient method: $y_{t+1} = x_t - \eta g_t$, $\quad g_t \in \partial f(x_t)$

$x_{t+1} = \mathrm{proj}_C(y_{t+1})$



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

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

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

$a \ge b$

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

$\|y_{t+1} - x^*\|_2^2 \ge \|x_{t+1} - x^*\|_2^2$

$\boxed{\varepsilon \sim \frac{1}{\sqrt{T}}}$

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

Discussion

Comparing the convergence rates between GD and PGD:

- For f convex and Lipschitz continous, both GD and PGD converge $\mathcal{O}(1/\sqrt{t})$.
- For f convex and smooth, both GD and PGD converge $\mathcal{O}(1/t)$.
- For f strongly convex and smooth, both GD and PGD converge $\mathcal{O}(\rho^t)$.

i.e., the theoretical convergence rate of PGD will be the same as that of GD.

Projected gradient method is only efficient if the projection step is cheap or simple.

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.