CSED490Y: Optimization for Machine Learning

Week 05-1: Subgradient and projected gradient methods

w6

Namhoon Lee

POSTECH

Spring 2022

Admin

"we propose a new idea X for Y" → ok

No need to "invent" something new.

Requirements of group project due tonight:

1. Choose topic
2. Submit the plagiarism statement

Study existing methods/ideas that are relevant to this course, such that you get

1. practical experience.
2. deeper understanding.
   inner-workings
3. something "interesting"

intuition / new phenomenon,

Admin

Requirements of group project due tonight:

1. Choose topic
2. Submit the plagiarism statement

- loss landscape

Some examples

Gradient descent

So far

- GD: algorithm and motivation
- Analysing GD under the smoothness assumption

gradient of f is Lipschitz continuous.

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$$

grad. can't change too quickly.

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

Convergence of gradient descent

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

Convergence rate for smooth function

- Prove from the progress bound.

sublinear

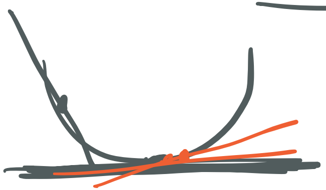
$+$

recap: $\min_{t=\{1,\cdots,T\}} \|\nabla f(x_t)\|^2 \le \boxed{\frac{2LR}{T}} \sim \boxed{O(1/T)} \le \varepsilon$

ε

$T \to \infty$, 0

$\min \|\nabla f(x_t)\|^2 = 0$



Q: How many iterations do we need to run GD to achieve ε-accuracy? $T = O(1/\varepsilon)$

Convergence of gradient descent

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

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \Rightarrow f(x^*) \geq \underline{f(x_t)} + \langle \nabla f(x_t), x^* - x_t \rangle$$

Convergence rate for smooth convex function

- Prove from the convexity and plugging into the progress bound. $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$

$$f(x_t) \leq f(x^*) + \langle \nabla f(x_t), x_t - x^* \rangle$$

$$(a-b)^2 = a^2 - 2a \cdot b + b^2$$

$$f(x_{t+1}) \leq f(x^*) + \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

$$f(x_{t+1}) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{L^2}\|\nabla f(x_t)\|^2 + \|x_t - x^*\|^2 - \|x_t - x^*\|^2 \right)$$

$$= \frac{L}{2}\left(\|x_t - x^*\|^2 - \|\overbrace{x_t - \tfrac{1}{L}\nabla f(x_t)}^{= x_{t+1}} - x^*\|^2 \right)$$

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

$$f(x_{t+1}) \le f(x_t)$$

$$x_{t+1} \approx x^*$$

Convergence rate for smooth convex function

- Prove from the convexity and plugging into the progress bound.

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

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

$$\le \frac{L}{2}\|x_1 - x^*\|^2$$

$$\vee|$$

$$T\left(f(x_{t+1}) - f(x^*)\right)$$

$$\boxed{f(x_{T+1}) - f(x^*) \le \frac{LR^2}{2T}}$$

R

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

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

Summary

- GD algorithm and motivation
- GD convergence rate
- Convergence criterion
- Dimension free

d : # features
parameters

$\boxed{O(1/T)}$ $\Rightarrow$ Nesterov's accelerated method (NAG)

Gradient descent

Summary

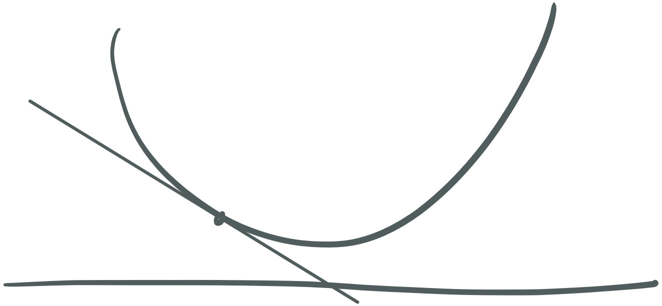
- GD algorithm and motivation
- GD convergence rate
- Convergence criterion
- Dimension free

Next

- strongly convex case
- non-smooth case

Convex function – differentiable case

Recall the C^1 definition:

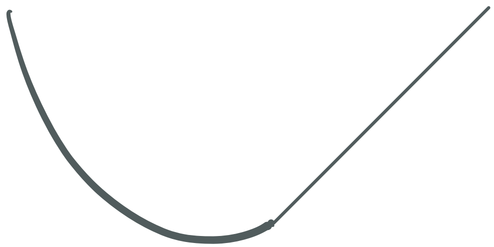


for all x

$$(*) \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall y$$

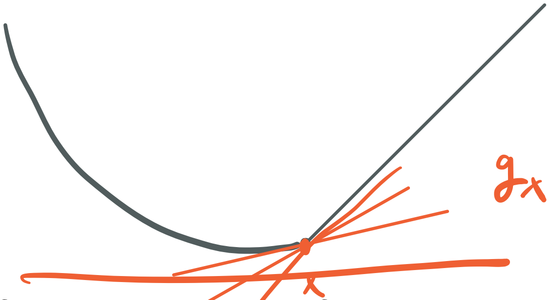
Convex function – non-differentiable case

(A non-differentiable case)



Convex function – non-differentiable case

(A non-differentiable case)



Generalizing to non-nondifferentiable case: A function is convex if $\forall x$, $\exists g$ such that

$$\boxed{f(y) \geq f(x) + \langle g, y - x \rangle \ .}$$

Subgradient and subdifferential

A vector g_x is called a subgradient of a convex function f at x if

$$f(y) \geq f(x) + \langle g_x, y - x \rangle \ , \quad \forall y \ .$$

Subgradient and subdifferential

A vector g is called a subgradient of a convex function f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \ , \quad \forall y \ .$$

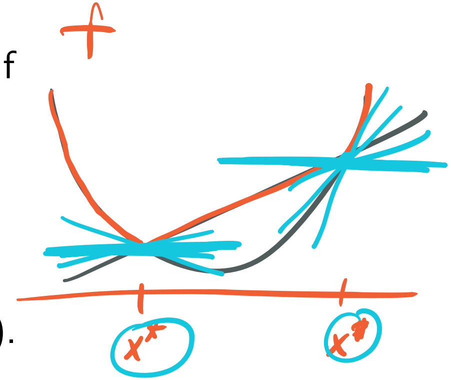
$g_x \in \partial f(x)$

The set of subgradients of f at x is called subdifferential $\boxed{\partial f(x)}$.

Subgradient and subdifferential

A vector g is called a subgradient of a convex function f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \ , \quad \forall y \ .$$



The set of subgradients of f at x is called subdifferential $\partial f(x)$.

- If a function f is differentiable at x, the gradient is the only element in the subdifferential $\partial f(x)$, *i.e.*, $\boxed{g_x = \nabla f(x)}$.
- The optimality condition for non-differentiable function: x^* is a global minimum if $\boxed{0 \in \partial f(x^*)}$.

$\nabla f(x^*) = 0$

Subdifferential example

An absolute value function:

Non-smooth problem

L1 regularized least squares

- You can obtain a sparse solution.
- It can be used for feature selection.

L1 vs L2 regularization

L2 regularized least squares

L1 vs L2 regularization

L2 regularized least squares

L1 regularized least squares

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.