CSED490Y: Optimization for Machine Learning

Week 04-2: Gradient descent

Namhoon Lee

POSTECH

Spring 2022

Admin

Group project

- Submit the plagiarism pledge form (available on PLMS)

Admin

Group project

- Submit the plagiarism pledge form (available on PLMS)

Switching to in-person class

- (when) Commencing on 11 April
- (where) Engineering building 2, Room 109

Admin

Group project

- Submit the plagiarism pledge form (available on PLMS)

Switching to in-person class

- (when) Commencing on 11 April
- (where) Engineering building 2, Room 109

Office hours

- Thursdays 5-6pm (by appointment)

Admin

Group project

- Submit the plagiarism pledge form (available on PLMS)

Switching to in-person class

- (when) Commencing on 11 April
- (where) Engineering building 2, Room 109

Office hours

- Thursdays 5-6pm (by appointment)

Quick poll

Gradient descent

Cost of solving (regularized) least squares

- $\mathcal{O}(nd^2 + d^3)$ vs $\mathcal{O}(ndt)$

How many iterations does GD require?

Gradient descent :

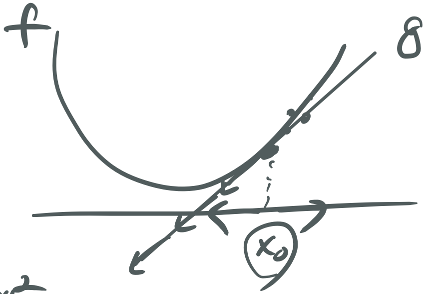
$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

$\nabla f(x_t) = 0$: global solution $x_t = x^*$ for f convex

$\nabla f(x_t) = 0$ for non-convex: stationary, critical

GD algorithm

- An iterative algorithm to find a minimum.
- Update the current iterate by taking a step into the negative direction of gradient.
- Stop when it isn't making any progress in practice.



- Another way to motivate GD: function approximation.

$$\min f(x) \text{ s.t. } x \in \mathbb{R}^d \simeq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2\eta}\|x - x_0\|^2$$

$$\arg\min_x f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2\eta}\|x - x_0\|^2$$

- Need to analyze convergence behaviour.

$$\nabla f(x_0) + \frac{1}{\eta}(x - x_0) = 0 \Rightarrow x = x_0 - \eta \nabla f(x_0)$$

Gradient descent

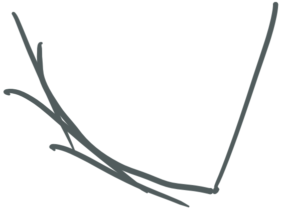
- Assumption: Lipschitz continuity of objective gradient (or "smoothness")
  - (definition)
  - (meaning)
  - (illustration)

A cnt. diff. func. f is called L-smooth if

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \|x - y\|_2 \quad \forall \{x, y\}$$

$L > 0$

↳ gradients don't change arbitrarily quickly.



Smoothness

$\int_0^1 F'(t)\,dt = F(1) - F(0)$

$\int_0^1 \nabla f((1-t)x + ty)^T (y-x)\,dt = f(y) - f(x)$

An important consequence of Lipschitz continuous objective gradient:

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

$$⊛ \quad f(y) \le f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2 \,.$$

Proof (recall ftc):

$$f(y) = f(x) + \int_0^1 \nabla f((1-t)x + ty)^T (y-x)\,dt$$

$$= f(x) + \nabla f(x)^T (y-x) + \int_0^1 \left(\nabla f((1-t)x+ty) - \nabla f(x)\right)^T (y-x)\,dt$$

$x \cdot y \le \|x\|\,\|y\|$

$$\overset{C.S}{\le} f(x) + \nabla f(x)^T (y-x) + \int_0^1 \|\nabla f((1-t)x+ty) - \nabla f(x)\|\,\|y-x\|\,dt$$

$$\le f(x) + \nabla f(x)^T (y-x) + \int_0^1 L t \|(x-y)\|^2\,dt$$

$\int_0^1 t\,dt = \frac{1}{2}$

$$= f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x-y\|^2$$

Smoothness

An important consequence of Lipschitz continuous objective gradient:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 \ .$$

Illustration:

Gradient descent progress bound

Under the quadratic upper bound, we are interested in how much progress gradient descent can make at each step.

Gradient descent progress bound

Under the quadratic upper bound, we are interested in how much progress gradient descent can make at each step.

Consider gradient descent with $\eta = 1/L$.

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t) \ .$$

Gradient descent progress bound

Under the quadratic upper bound, we are interested in how much progress gradient descent can make at each step.

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

Consider gradient descent with $\eta = 1/L$.

$y \to x_{t+1}$

$x \to x_t$

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)\ .$$

smoothness $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2 \quad \forall \{y, x\}$

Plugging this into the bound gives

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

$$= f(x_t) - \frac{1}{L}\|\nabla f(x_t)\|^2 + \frac{1}{2L}\|\nabla f(x_t)\|^2$$

$$= f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

Convergence of gradient descent

$\epsilon \to O(1/T)$

$f(x^*)$

$O(1/\epsilon^2)$ = grid search

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

$f^* \le f(x_t)$

Convergence rate for smooth function

- Prove from the progress bound.

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

$\|\nabla f(x_t)\| \sim$ error $\Rightarrow 0$

$T \sim O(\frac{1}{\epsilon})$

0.1, 0.01, ...

sum both sides for $t = 1, \dots, T$

$\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \le 2L \sum_{t=1}^{T} (f(x_t) - f(x_{t+1})) = 2L((f(x_1) - f(x_2)) + (f(x_2) - f(x_3)) + \cdots$

$= 2L(f(x_1) - f(x_T)) \le 2L(f(x_1) - f^*)$

$T \min_{t=\{1,\dots,T\}} \|\nabla f(x_t)\|^2 \le 2L(f(x_1) - f^*) \Rightarrow \min_t \|\nabla f(x_t)\|^2 \le \frac{2LR}{T} \to 0$

R, ϵ, $T \to \infty$

Convergence of gradient descent

Convergence rate for smooth convex function

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

Gradient descent

Summary

- GD algorithm and motivations
- GD Convergence rates

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.