CSED490Y: Optimization for Machine Learning

Week 05-2: Subgradient and projected gradient methods

Namhoon Lee

POSTECH

Spring 2022

Gradient descent example

$f'(x^*) = 6x^* + 4 = 0$ $\quad \boxed{x^* = -\frac{2}{3}}$

$1 + r + r^2 + \cdots + r^{t-1} = \frac{1-r^t}{1-r}$

min

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

$$\boxed{x_{t+1}} = x_t - \eta \nabla f(x_t)$$
$$= x_t - \eta(6x_t + 4)$$
$$= (1-6\eta)x_t - 4\eta$$
$$= (1-6\eta)(x_{t-1} - \eta\nabla f(x_{t-1})) - 4\eta$$
$$= (1-6\eta)(x_{t-1} - \eta(6x_{t-1}+4)) - 4\eta$$
$$= (1-6\eta)((1-6\eta)x_{t-1} - 4\eta) - 4\eta$$
$$= (1-6\eta)^2 x_{t-1} - 4\eta(1-6\eta) - 4\eta$$

$\cdots$

$$= (1-6\eta)^t x_1 - 4\eta(1-6\eta)^{t-1} - 4\eta(1-6\eta)^{t-2} \cdots - 4\eta$$

$$= (1-6\eta)^t x_1 - 4\eta \cdot \frac{1-(1-6\eta)^t}{1-(1-6\eta)}$$

$t \to \infty \quad |1-6\eta| < 1$

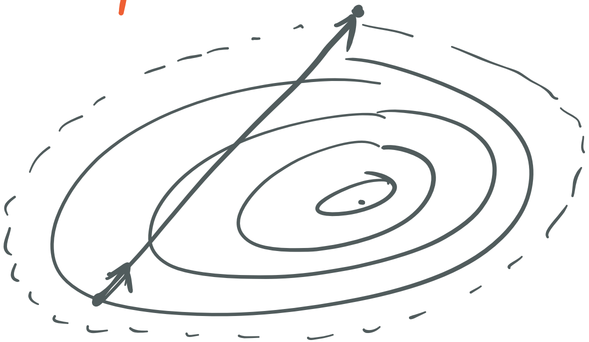
$$= \frac{-4\eta}{1-(1-6\eta)} = \boxed{-\frac{2}{3}}$$

$$= (1-6\eta)^t\left(x_1 + \frac{2}{3}\right) - \frac{2}{3}$$

Gradient descent example

From the previous example

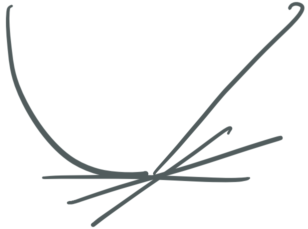
- solution $-\frac{2}{3}$
- convergence : linear : $\varepsilon \sim \rho^t$ or e^{-ct}
- stepsize $|1-6\eta| < 1$; stepsize small enough
- initial guess x_1



Subgradient and subdifferential

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

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



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 $\partial f(x)$.

Subgradient and subdifferential

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

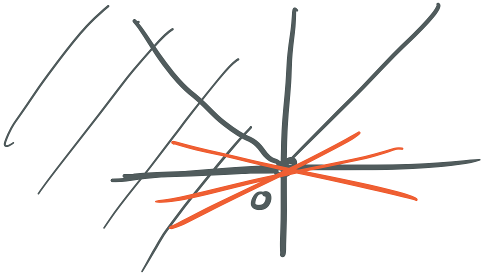
$$f(y) \geq f(x) + \langle g, y - x \rangle \ , \quad \forall y \ .$$

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.*, $g_x = \nabla f(x)$.
- The optimality condition for non-differentiable function: x^* is a global minimum if $0 \in \partial f(x^*)$.

Subdifferential example

An absolute value function: $f(x) = |x|$



$$\boxed{\partial |x| = \begin{cases} -1, & x < 0 \\ [-1, 1], & x = 0 \\ +1, & x > 0 \end{cases}}$$

Non-smooth problem

$\hat{y} = w^T x = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$

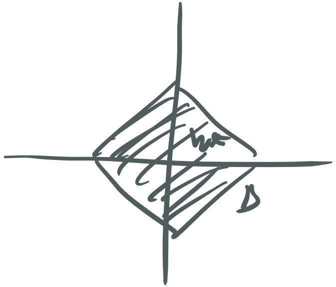
X, y : dataset

$w^* = (x \; x \; 0 \; 0 \; 0 \; x \; \cdots \; 0 \; 0)$

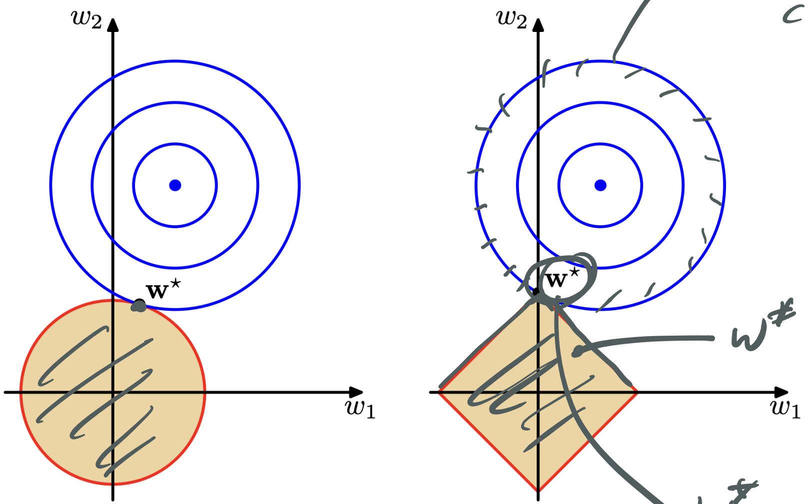
L1 regularized least squares

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|\underbrace{Xw}_{\hat{y}} - y\|^2 + \boxed{\lambda \|w\|_1} \iff \begin{array}{ll} \min & \frac{1}{2} \|Xw - y\|^2 \\ \text{s.t.} & w \in \mathbb{R}^d \\ & \boxed{\|w\|_1 \leq B} \end{array} \quad w^*$$

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



L1 vs L2 regularization



level sets

contours of model obj.

w^*

L2   L1

$w^* = (w_1^*, w_2^*)$ $w^* = (0, w_2^*)$

L1 vs L2 regularization

L2 regularized least squares

$f(w)$

$$\min_w \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|_2^2$$

$$\boxed{w_j^* = 0}$$

$$\underline{\underline{w^T x}}$$

$$\nabla f(w) = X^T(\underline{Xw^* - y}) + \lambda w^* = 0$$

$$\nabla_j f(w) = \underbrace{X_j^T}_{\text{data features}}(\underbrace{Xw^* - y}_{\text{? error, residual}}) + \lambda \underbrace{w_j^*}_{=0} = \underline{0}$$

L1 vs L2 regularization

L2 regularized least squares

L1 regularized least squares

$$f(w)$$

$$\min_w \frac{1}{2}\|Xw-y\|^2 + \lambda\|w\|_1$$

$$\partial f(w) = X^T(Xw-y) + \lambda\partial\|w\|_1$$

$$0 \in \partial_j f(w) = \underline{X_j^T(Xw^*-y)} + \lambda[-1,1]$$

$$\|w\|_1 = \sum_i |w_i|$$

$$w_j^* = 0$$

$$\partial|w_j| = \underline{[-1, 1]} \text{ at } w_j = 0$$

$$-X_j^T(Xw^*-y) \in \lambda[-1,1]$$

Subgradient method

Subgradient method:

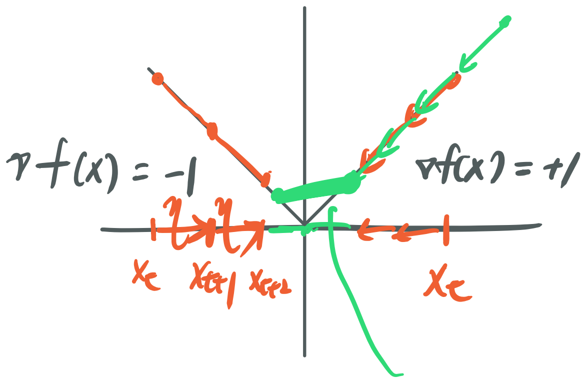
$$x_{t+1} = x_t - \eta g_{x_t}$$

where $g_{x_t} \in \partial f(x_t)$.

- Gradient descent for non-differentiable cases.
- Applicable to the previous example of absolute value function.

Subgradient method example

$f(x) = |x|$



$x_t < 0: \quad x_{t+1} = x_t + \eta$

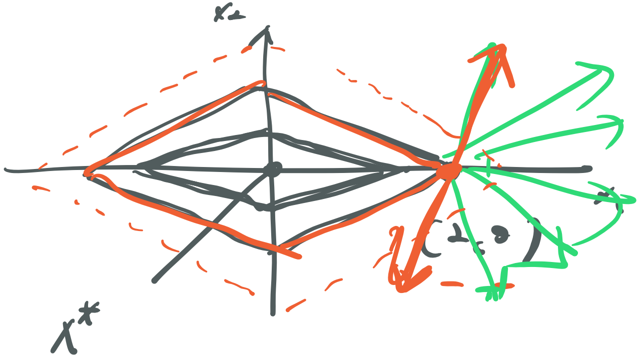
$x_t > 0: \quad x_{t+1} = x_t - \eta$

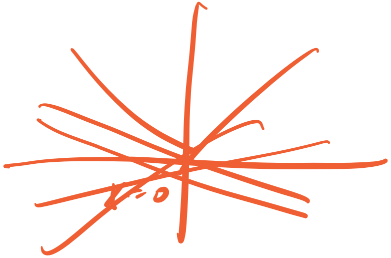
Q: what happens when you get close to the min?

Subgradient "method" example

$$f(x_1, x_2) = |x_1| + \overset{4}{\cancel{2}}|x_2|$$

$$\partial f(x_1, x_2) = \boxed{(1, 4[-1,1])^T} \quad \underline{x_1 > 0,\ x_2 = 0}$$





$\partial|x| = [-1, 1]$

Gradient descent vs subgradient method

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.

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.