Unit 5 Gradient Descent and SGD
Friday, October 13, 2023

- Gradient Descent
- Stochastic Gradient Descent
- Convexity, Regret

5.1 Gradient Descent in full-batch setting

- Goal: optimize $f(x)$
- Assume f is differentiable
- Use $\nabla f(x)$ to compute descent direction
- Convexity, Regret

Gradient Descent

- Start from some initial point $x_0 \in \mathbb{R}^d$, $t = 0$, $\eta = 1/L$
- Repeat until convergence: $x_{t+1} = x_t - \eta \nabla f(x_t)$, $t = t+1$

$$\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_d}$$

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_d}\right)^T$$

η is a step size or learning rate

Why is $\nabla f(x)$ a descent direction?

We want to find Δx such that $f(x + \Delta x) < f(x)$

- $\nabla f(x)$ points in the direction of steepest ascent
- $-\nabla f(x)$ points in the direction of steepest descent

We want to find Δx such that $\nabla f(x)^T \Delta x < 0$

$$\nabla f(x)^T \Delta x = \|\nabla f(x)\| \cdot \|\Delta x\| \cdot \cos\theta$$

θ is the angle between $\nabla f(x)$ and Δx

$\cos\theta$ is maximized when $\Delta x = \nabla f(x)$

$\cos\theta$ is minimized when $\Delta x = -\nabla f(x)$

$\cos\theta = 0$ when $\Delta x \perp \nabla f(x)$

$\cos\theta = -1$ when $\Delta x = -\nabla f(x)$

$\cos\theta = 1$ when $\Delta x = \nabla f(x)$

Proof:

$$\Delta x = -\eta \nabla f(x)$$

$$\nabla f(x)^T \Delta x = -\eta \|\nabla f(x)\|^2 < 0$$

Assumption: $\|\nabla f(x)\| \neq 0$

Note: $\|\nabla f(x)\| = 0$ means we are at a stationary point

Assumption: $\eta > 0$

$$\nabla f(x)^T \Delta x = -\eta \|\nabla f(x)\|^2 < 0$$

Gradient Descent

- Start from some initial point $x_0 \in \mathbb{R}^d$, $t = 0$, $\eta = 1/L$
- Repeat until convergence: $x_{t+1} = x_t - \eta \nabla f(x_t)$, $t = t+1$

Convergence: $\|\nabla f(x)\| \leq \epsilon