ML & Opt Lecture 10

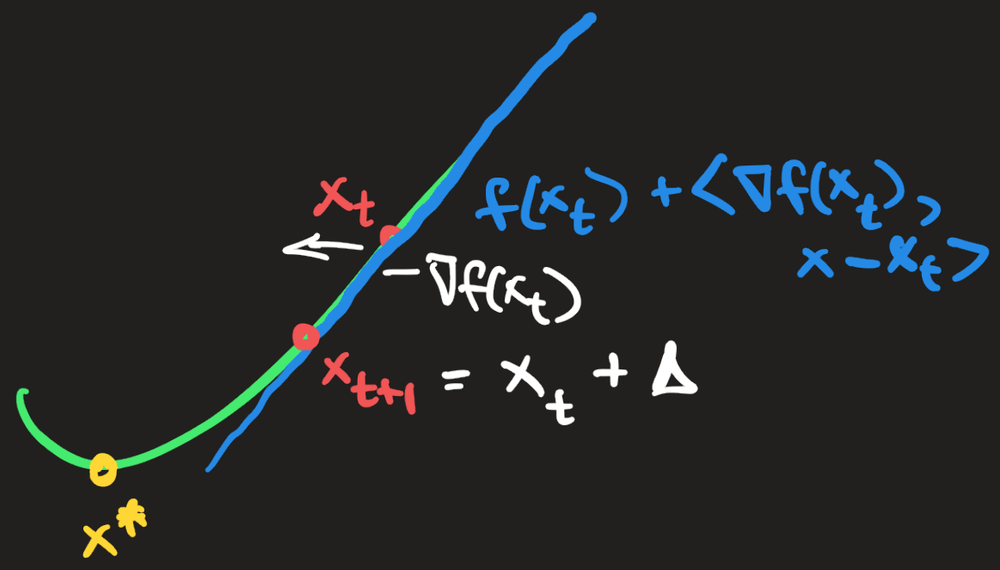
- Gradient Descent & Projected Gradient Descent
- Subgradient Descent & " Subgradient Descent
- Example

Gradient Descent

f – cvx objective (smooth)

want solve

$$x_* \in \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x)$$



The basic idea is to note that by T.S. expansion argument:

$$f(x_{t+1}) \approx f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle$$

$$= f(x_t) + \langle \nabla f(x_t), \Delta \rangle$$

since $\Delta = -\alpha \nabla f(x_t)$ pts in the neg grad dir,

$$f(x_{t+1}) \approx f(x_t) - \alpha \langle \nabla f(x_t), \nabla f(x_t) \rangle$$

$$= f(x_t) - \alpha \|\nabla f(x_t)\|_2^2$$

$$< f(x_t), \quad \text{when } \alpha \text{ is small}$$

Gradient Descent Alg

For solving

$$(U) \quad \min_{x \in \mathbb{R}^d} f(x)$$ where f is smooth & convex

Input

∇f – gradient oracle

T – iterations

x_0 – starting guess

$\alpha_1, \dots, \alpha_T$ – step sizes for each iteration

Output

$x_1, \dots, x_T \in \mathbb{R}^d$ estimates of x^*

Alg

For $t = 0$ to $T-1$

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

For now we want to know GD converges under some reasonable assumptions of f, that tell us:

- how to choose our stepsizes
- how many steps T to take to reach approximate convergence

ε-suboptimality → our focus for cvx optim

Three notions of approximate convergence

1) $f(x_t) - f(x_*) < \varepsilon$ — approx conv in function value

2) $\|x_t - x_*\|_2^2 < \varepsilon$ — " " " parameter

3) $\|\nabla f(x_t)\|_2 < \varepsilon$ — approx conv of gradient to zero

usual convergence condition for nonCvx optim

ε-stationarity

A reasonable assumption for GD to work is that f is β-smooth:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2,$$

i.e. ∇f is Lipschitz-continuous

Ex of a function that is β-smooth

$$f(x) = \|Ax - b\|_2^2 \qquad \nabla f(x) = A^T(Ax - b)$$

$$\|\nabla f(x) - \nabla f(y)\|_2 = \|A^TA(x-y)\|_2 \leq \|A^TA\|_2 \|x-y\|_2$$

so f is β-smooth where $\beta = \|A^TA\|_2$

Ex of a function that is <u>not</u> β-smooth for <u>any</u> β

$$f(x) = e^x \qquad \nabla f(x) = e^x$$

is there a β s.t. $|e^x - e^y| \leq \beta |x-y|$?

No, b/c there exists $z \in [x,y]$ s.t. $|e^x - e^y| = |e^z(x-y)| \geq e^x|x-y|$

so for any β I can find x & $y = x+1$ so that $\beta < e^x$, so $|e^x - e^y| > \beta|x-y|$

Consequences of β-smoothness

1) $\|\nabla f(x)\|_2 \leq \beta \|x - x_*\|_2$

(take $y = x^*$ in the def of β-smoothness & note $\nabla f(x^*) = 0$)

2) This implies a concrete bound on how good the approximation

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) \text{ is.}$$

Precisely,

$$\left| f(y) - \left[f(x) + \nabla f(x)^T (y - x) \right] \right| \leq \frac{\beta}{2} \|x - y\|_2^2$$

Chap 3 of Bubeck's Convex Optimization

Now assume f is β-smooth. Take $\alpha_t = \frac{1}{\beta}$

Descent Lemma With this choice, GD ensures that

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta} \|\nabla f(x_t)\|_2^2$$

Pf

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

$$\Rightarrow \quad f(x_{t+1}) - \left[f(x_t) + \langle \nabla f(x_t), -\frac{1}{\beta}\nabla f(x_t)\rangle\right]$$

$$= f(x_{t+1}) - \left[f(x_t) - \frac{1}{\beta}\|\nabla f(x_t)\|_2^2\right]$$

$$\leq \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2 \quad = \frac{\beta}{2}\left\|\frac{1}{\beta}\nabla f(x_t)\right\|_2^2$$

$$= \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

Thus,

$$f(x_{t+1}) - f(x_t) + \frac{1}{\beta}\|\nabla f(x_t)\|_2^2 \leq \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

$$\Rightarrow$$

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

as claimed.

Fact If f is convex and β-smooth, then gradient descent with $\alpha_t = \frac{1}{\beta}$ guarantees that

$$f(x_T) - f(x_*) \leq \frac{2\beta\|x_0 - x_*\|_2^2}{T}$$

We say GD takes $T = O\left(\frac{\beta}{\varepsilon}\right)$ iterations to get to an ε-suboptimal minimizer (iteration complexity)

Alternatively, GD has convergence rate $O\left(\frac{\beta}{T}\right)$

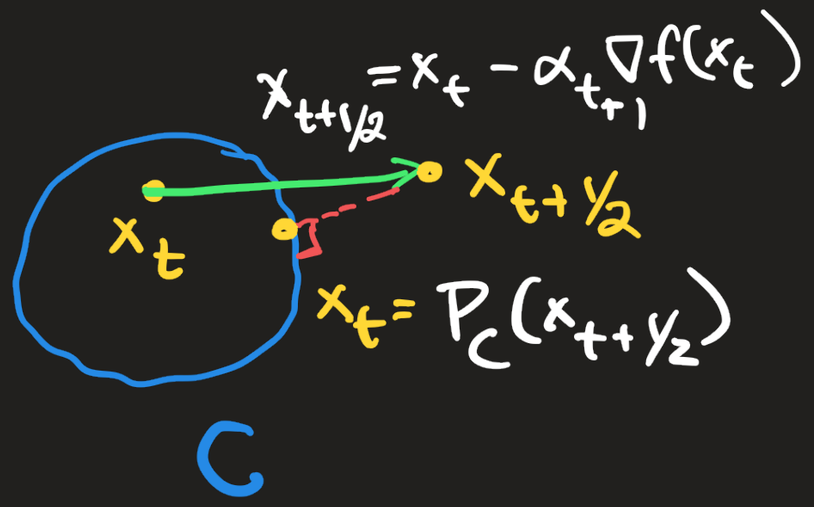
What about constrained optimization?

Projected Gradient Descent

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

Inputs

∇f – gradient oracle
T – iterations
x_0 – initial pt
$\alpha_1, \dots, \alpha_T$ – step sizes
P_C – projection operator onto C



Alg

for t in 0 to T−1:

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

What about nonsmooth, constrained convex optimization?

Subgradient Descent Alg

f – cvx function

$x_* = \underset{x \in C}{\operatorname{argmin}} f(x)$

Inputs

∂f – oracle for a subgradient

T – # of iterations

x_0 – starting pt

$\alpha_1, \dots, \alpha_T$ – stepsizes

P_C – projector onto C

Alg

for $t = 0, \dots, T-1$:

$g_t \leftarrow$ a vector in $\partial f(x_t)$

$x_{t+1} \leftarrow P_C(x_t - \alpha_t g_t)$

Return

$$\hat{x} = \frac{1}{T} \sum_{t=1}^{T} x_t \quad \text{(Polyak average)}$$

Output

$\hat{x} \approx x^*$

Guarantees

$f(\hat{x}) - f(x^*) \leq$

$O\left(\frac{\log T}{T}\right)$