COMS 4995-2: Advanced Algorithms (Spring'20) Mar 26, 2020

Lecture 16: Gradient Descent

Instructor: *Alex Andoni* Scribes: *Emily Jin, Yihao Li, Jayant Madugula*

1 Gradient Descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ and the objective function be $\min_{x \in \mathbb{R}^n} f(x)$

We have $f(x+\delta) = f(x) + \nabla f(x)^T \cdot \delta + \delta^T \cdot \nabla^2 f(y) \cdot \delta$ for some $y \in [x, x+\delta]$

For small enough δ, we make The "first order approximation" of $f(x)$ by ignoring terms quadratic in δ:

$$f(x+\delta) \approx f(x) + \nabla f(x)^T \cdot \delta$$

The local "greedy" step for delta gives:

$$\begin{aligned} \delta &= \text{argmin}_{||\delta|| \leq \epsilon} f(x) + \nabla f(x)^T \cdot \delta \\ &= -\eta \cdot \nabla f(x)^T \end{aligned}$$

where η is the step size and is defined as $\eta = \text{const}(\epsilon)$.

1.1 Gradient Descent Algorithm

The algorithm itself is quite simple:

1. Fix $x^0 \in \mathbb{R}^n$.
2. For every iteration t, we have
$$x^t = x^{t-1} - \eta \cdot \nabla f^T(x^{t-1})$$
We perform T total iterations.

Given this algorithm, how do we set t and find a sufficient setting for T?
We make two major assumptions.

Definition 1. *Let $\beta > 0$, $y = x + \delta$. f is β-smooth if $\forall x, y \in \mathbb{R}^n$,*

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

Equivalently, $\forall \delta \in \mathbb{R}^n$:

$$\delta^T \cdot \nabla^2 f(x) \cdot \delta = \beta \cdot ||\delta||^2$$

$\Leftrightarrow$ *max eigenvalue of the Hessian is $\leq \beta$, where the Hessian is $\nabla^2 f(x) = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$*

We use the Taylor Expansion of $f(x)$ to get the following upper bound on $f(x+\delta)$:

$$f(x+\delta) = f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\delta^T \cdot \nabla^2 f(y) \cdot \delta \tag{1}$$

$$\leq f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\beta||\delta||^2 \tag{2}$$

We now minimize the latter, without any further constraint on δ. The optimum will satisfy $\delta = -\eta \cdot \nabla f(x)$. We can then solve for the optimal δ in terms of β and the gradient of $f(x)$.

$$\delta \triangleq \text{argmin}_{\delta \in \mathbb{R}^n} f(x) + \nabla f(x)^T \cdot \delta + \frac{\beta}{2}||\delta||^2 \tag{3}$$

$$= \text{argmin}_{\delta \in \mathbb{R}^n} \nabla f(x)^T(-\eta \cdot \nabla f(x)) + \frac{\beta}{2}\eta^2||\nabla f(x)||^2 \tag{4}$$

$$= \text{argmin}_{\delta \in \mathbb{R}^n} ||\nabla f(x)||^2(-\eta + \frac{\beta}{2}\eta^2) \tag{5}$$

$$= \text{argmin}_{\delta \in \mathbb{R}^n} \frac{\beta}{2}\eta^2 - \eta \tag{6}$$

$$= -\frac{1}{\beta} \cdot \nabla f(x) \tag{7}$$

Note that just as $-\frac{1}{\beta} \cdot \nabla f(x)$ is the optimal δ, $\frac{1}{\beta}$ is the optimal η, or step size.
At iteration t:

$$x^t = x^{t-1} - \frac{1}{\beta}\nabla f(x^{t-1})$$

we have $\eta = \frac{1}{\beta}$. We can now plug in $\delta = -\frac{1}{\beta}\nabla f(x)$ into the Taylor expansion to get the upper bound on $f(x+\delta)$:

$$\begin{aligned} f(x+\delta) \leq & f(x) + \nabla f(x)^T \left(-\frac{1}{\beta}\nabla f(x)\right) + \frac{\beta}{2}\| - \frac{1}{\beta}\nabla f(x)\|^2 \\ = & f(x) + \|\nabla f(x)\|^2 \left[-\frac{1}{\beta} + \frac{1}{2\beta}\right] \\ = & f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \end{aligned}$$

If $\nabla f(x) \neq 0$, we know $\|\nabla f(x)\| \neq 0$, therefore $f(x+\delta) < f(x)$, i.e. we are making progress.
If $\nabla f(x) = 0$, we have following cases:

1. Global min, which is our goal
2. Local min
3. Global/local max, can be usually solved (escaped) by some random perturbations
4. Saddle point, in some directions it increase, in some other directions it decreases, and may stay at constant. Can be usually solved by some random perturbations.

To deal with the local minimum, we introduce another assumption: f(x) is convex.

Definition 2. *f is convex iff $\forall x, y \in \mathbb{R}^n, \lambda \in [0,1]$, there's*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

Claim 3. *if f(x) is a convex function, any minimum of f is global minimum.*

Proof. if f(x) is convex, $\forall \delta \in \mathbb{R}^n, \delta^T \nabla^2 f(x) \delta \geq 0$, i.e. $\nabla^2 f(x)$ is positive semi-definite (and all eigenvalues are positive), By taylor expansion

$$\begin{aligned} f(x+\delta) =& f(x) + \nabla f(x)\delta + \frac{1}{2}\delta^T \nabla^2 f(y)\delta \\ \geq& f(x) + \nabla f(x)\delta \end{aligned}$$

if $\nabla f(x) = 0$, we still have

$$f(x+\delta) \geq f(x)$$

for any δ, which means x is a global minimum. □

Thus for the G.D algorithm, $\delta = -\frac{1}{\beta}\nabla f(x)$

- if $\nabla f(x) \neq 0$, $f(x+\delta) \leq f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2$ from β-smooth
- if $\nabla f(x) = 0$, x is global min by convexity

Time Efficiency: How fast can it be? How should T be defined?

Goal: given $\epsilon > 0$, find x s.t. $f(x) - f(x^*) \leq \epsilon$, where $x^* = \arg\min_x f(x)$, i.e. the optimum. Bounding $\nabla f(x)$?

$$\begin{aligned} f(x^*) =& f(x + x^* - x) \\ =& f(x) + \nabla f(x)^T (x^* - x) + (x^* - x)^T \nabla^2 f(y)(x^* - x) \\ \text{given } \nabla^2 f(y) \text{ positive semi-definite, } \geq& f(x) + \nabla f(x)^T (x^* - x) \end{aligned}$$

Rearrange the equation, we have

$$f(x) - f(x^*) \leq -\nabla f(x)^T (x^* - x) \leq \|\nabla f(x)\| \cdot \|x^* - x\|$$

Thus if $f(x) - f(x^*) > \epsilon$, we have

$$\begin{aligned} \|\nabla f(x)\| \cdot \|x^* - x\| &> \epsilon \\ \|\nabla f(x)\| &> \frac{\epsilon}{\|x^* - x\|} \end{aligned}$$

Theorem 4. *$f(x^T) - f(x^*) \leq \epsilon$ after $T = O(\beta \cdot \frac{D^2}{\epsilon})$ iterations, where $D = \max ||x - x^*||$ for x such that $f(x) \leq f(x^0)$*

Proof. From β-smoothness, we showed that $f(x+\delta) \leq f(x) - \frac{1}{2\beta}||\nabla f(x)||^2$ where $\delta = -\frac{1}{\beta}\nabla f(x)$. Suppose $\Delta = f(x) - f(x^*) > \epsilon$:

$$||\nabla f(x)|| \geq \frac{\Delta}{||x-x^*||} \geq \frac{\Delta}{D}$$

Let $\Delta_t = f(x^t) - f(x^*)$. If we fix Δ_0, how many steps T_1 until $\Delta_{T_1} < \frac{\Delta_0}{2}$? Before this happens, when $t < T_1$:

$$||\nabla f(x)|| \geq \frac{\Delta_0/2}{D}$$

In each iteration, we decrease $f(x)$ by $-\frac{1}{2\beta}||\nabla f(x)||^2$:

$$\begin{aligned}
\Delta_t \leq & \Delta_{t-1} - \frac{1}{2\beta}(\frac{\Delta_0/2}{D})^2 \\
\leq & \Delta_{t-1} - \frac{1}{8\beta D^2} \cdot \Delta_0^2 \\
\text{\# steps } T_1 \text{ is} \leq & \frac{8\beta D^2}{\Delta_0}
\end{aligned}$$

How many steps until...

$$\begin{aligned}
\Delta_{T_2} \leq \frac{\Delta_0}{4} \rightarrow & T_2 \leq 8\beta D^2 \cdot \frac{2}{\Delta_0} \\
\Delta_{T_3} \leq \frac{\Delta_0}{8} \rightarrow & T_3 \leq 8\beta D^2 \cdot \frac{4}{\Delta_0} \\
\Delta_T \leq \epsilon \rightarrow & T \leq 8\beta D^2 \cdot \frac{1}{2\epsilon}
\end{aligned}$$

Total number of steps is $\leq T_1 + T_2 + ... \leq 8\beta D^2 \cdot [\frac{1}{\Delta_0} + \frac{2}{\Delta_0} + ... + \frac{1}{2\epsilon}]$. Since these terms are geometrically increasing, the last epsilon term is the most important. Therefore, the total runtime is $O(\frac{\beta D^2}{\epsilon})$. □

If f is β-smooth and convex, $T = O(\frac{\beta D^2}{\epsilon})$ steps is enough to achieve $f(x) - f(x^*) \leq \epsilon$. Note that the runtime is proportional to $D = \max ||x - x^*||$. This is undesirable: it's like maximum capacity F from the Ford-Fulkerson algorithm (analysis), making this a "pseudo-polynomial time" algorithm.