E0 306: Deep Learning: Theory and Practice 13 April 2019

Scribe Notes

GD for strongly convex and smooth convex functions

Aditay Tripathi

1 Gradient Descent for L-smooth convex functions

Lemma 1. *If f is $L-smooth$, then for any $x, y \in \mathbb{K}$, we have:*

$$0 \leq f(y) - f(x) - \nabla f(x)^T (y - x) \leq L||x \dot{-} y||^2 \tag{1}$$

This means that the distance of $f(y)$ from its first-order Taylor approximation at x is between 0 and $L||x \dot{-} y||^2$

Proof. From the definition of convexity: $f(x) \geq f(y) + \nabla f(y)^T (x - y)$, and using Cauchy-Schwartz we get:

$$\begin{aligned} f(y) - f(x) &\leq \nabla f(y)^T (y - x) \\ &= \nabla f(x)^T (y - x) + (\nabla f(y) - \nabla f(x))^T (y - x) \\ &\leq \nabla f(x)^T (y - x) + ||\nabla f(y) - \nabla f(x)|| \cdot ||y - x|| \\ &\leq \nabla f(x)^T (y - x) + L \cdot ||x - y||^2 \end{aligned}$$

On the other hand, also from from convexity:

$$f(y) - f(x) \geq \nabla f(x)^T (y - x)$$

Combining this equation with the above equation proves the lemma.

□

Theorem 2. *f is a $L-smooth$ convex function, given an error parameter ϵ, a starting point x_1, it produces a sequence of points $x_1, \ldots, x_T$, such that:*

$$f(x_T) - f(x^*) \leq \epsilon$$

and

$$T = O\left(\frac{LD^2}{\epsilon}\right)$$

where, $D = sup\{||x - x^*|| : f(x) \leq f(x_1)\}$

Proof.

$$\begin{aligned}f(x_{t+1}) &\leq f(x_t) + (x_{t+1} - x_t)^T \nabla f(x_t) + L \cdot ||x_{t+1} - x_t||^2 \\ &\leq f(x_t) - \eta \nabla f(x_t)^T \nabla f(x_t) + L\eta^2 ||\nabla f(x_t)^2||\end{aligned}$$

Because $x_{t+1} = x_t - \eta \nabla f(x_t)$. Choose η that minimizes RHS of the above inequality. $\eta = \frac{1}{2L}$ minimizes the RHS. Therefore,

$$f(x_{t+1}) - f(x_t) \leq \frac{-1}{4L} ||\nabla f(x_t)||^2 \tag{2}$$

Since f is convex:

$$\begin{aligned}f(x^*) &\geq f(x_t) + (x^* - x_t)^T \nabla f(x_t) \\ f(x_t) - f(x^*) &\leq (x_t - x^*)^T \nabla f(x_t) \\ f(x_t) - f(x^*) &\leq ||x_t - x^*|| \cdot ||\nabla f(x_t)||\end{aligned}$$

This means if x_t is not close to x^*, $\nabla f(x_t)$ has to be large.

$$f(x_t) - f(x^*) \leq D ||\nabla f(x_t)||$$

Because $D = sup\{||x - x^*|| : f(x) \leq f(x_1)\}$

If $\frac{\theta}{2} \leq f(x_t) - f(x^*)$, then $||\nabla f(x_t)|| \geq \frac{\theta}{2D}$. Using this fact in eq. (2):

$$f(x_{t+1}) - f(x_t) \leq \frac{-\frac{\theta}{2}^2}{4LD^2} \tag{3}$$

θ indicates how far are we from optimum. If we are θ away from optimum, then the function decreases by amount in eq. (3).

How many steps do we need to go from $f(x_1) - f(x^*) \geq \frac{\theta}{2}$ to $f(x_t) - f(x^*) \leq \frac{\theta}{2}$? If we are θ away from x^*, function decrease is given by eq. (3), then total number of iterations are given by:

$$\begin{aligned}No. of\ iterations &\leq \frac{\frac{\theta}{2}}{\frac{\frac{\theta}{2}^2}{4LD^2}} \qquad (4) \\ &= \frac{8LD^2}{\theta} \qquad (5)\end{aligned}$$

After t to go from $f(x_t) - f(x^*) = \frac{\theta}{2}$ to $f(x_{t+\Delta t}) - f(x^*) < \frac{\theta}{2}$, number of iterations required is given by eq. (5).

Now to go from $\frac{\theta}{2^i}$ to $\frac{\theta}{2^{i+1}}$, number of iterations is less than $O\left(\frac{2^i LD^2}{\theta}\right)$. Initially $\theta =$

$f(x_1) - f(x^*)$. Total number of iterations needed to converge are:

$$
\begin{aligned}
&= \sum_{i=0}^{\log_2(\theta/\epsilon)} O\left(\frac{2^i LD^2}{\theta}\right) \\
&= \frac{LD^2}{\theta} \sum_{i=0}^{\log_2(\theta/\epsilon)} 2^i \\
&= O\left(\frac{LD^2}{\theta}\right) [2^{\log_2 \frac{\theta}{\epsilon}}] \\
&= O\left(\frac{LD^2}{\theta}\right) \cdot \frac{\theta}{\epsilon} \\
&= O\left(\frac{LD^2}{\epsilon}\right)
\end{aligned}
$$

□

Definition 1.1. *A twice differentable function is said to be l-strongly convex for $l > 0$ if:*

$$\nabla^2 f(x) \geq lI \ \ \forall x$$

In other words $\nabla^2 f(x) - lI$ is positive semi-definite.

Lemma 3. *If f is $L-strongly\ convex$, then $\forall x, y$:*

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

Theorem 4. *There is an algorithm exists with $\eta = \frac{2}{l(t+1)}$ for a function f which is $l-$strongly convex and $G-$Lipschitz such that given an error parameter ϵ, a starting point x_1, produces a sequence of points $x_1, \ldots, x_t$ that satisfies the following:*

$$f(x_T) - f(x^*) \leq \epsilon$$

where, $T = O\left(\frac{G^2}{l\epsilon}\right)$

Proof. Using Lemma 3.

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq (x_t - x^*)^T \nabla f(x_t) - \frac{l}{2}||x_t - x^*||^2 \\
&= \frac{1}{\eta_t}(x_t - x^*)^T (x_t - x_{t+1}) - \frac{l}{2}||x_t - x^*||^2 \\
&= \frac{1}{2\eta_t}\left[||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right] - \frac{l}{2}||x_t - x^*||^2
\end{aligned}
$$

We know that:

$$||x_t - x_{t+1}||^2 \leq \eta_t^2 G^2$$

Using this equation:

$$f\left(x_t\right)-f\left(x^*\right) \leq \frac{1}{2\eta_t}\left[\eta_t^2 G^2+\|x_t-x^*\|^2-\|x_{t+1}-x^*\|^2\right]-\frac{l}{2}\|x_t-x^*\|^2,$$

for every $t=1,...,T$. Summing all the $t-multiples$ of above equation:

$$\sum_{t=1}^{T} t\left(\left(x_t\right)-f\left(x^*\right)\right) \leq \frac{G^2}{2}\sum_{t=1}^{T} t\eta_t+\sum_{t=2}^{T}\|x_t-x^*\|^2\cdot\left(\frac{t}{2\eta_t}-\frac{lt}{2}-\frac{t-1}{2\eta_{t-1}}\right)$$
$$+\|x_1-x^*\|^2\cdot\left(\frac{1}{2\eta_1}-\frac{l}{2}\right)-\|x_{T+1}-x^*\|^2\cdot\frac{T}{2\eta_T}$$

We bound the last term by just zero. Now, to make the sum telescoping, we would like to get $\frac{t}{2\eta_2}-\frac{lt}{2}-\frac{t-1}{2\eta_{t-1}}=0$ for every $t=2,\ldots,T$. As for the term $\frac{1}{2\eta_1}-\frac{l}{2}$, we would also prefer to remove it, so as not to have any dependence on $\|x_1-x^*\|^2$. Solving these equations, yield: $\eta_t=\frac{2}{l(t+1)}$.

$$\begin{aligned}\sum_{t=1}^{T} t\left(\left(x_t\right)-f\left(x^*\right)\right) &\leq \frac{G^2}{2}\sum_{t+1}^{T} \\ &= \sum_{t=1}^{T}\frac{G^2 t}{l\left(l+1\right)} \\ &\leq \sum_{t=1}^{T}\frac{G^2}{l} \\ &= \frac{G^2 T}{l}\end{aligned}$$

Normalizing by $\left(1+\ldots+T\right)=\frac{T(T+1)}{2}$ and using the convexity property of f, we get:

$$f\left(\frac{2}{T\left(T+1\right)}\sum_{t=1}^{T} t * x_t\right)-f\left(x^*\right) \leq \frac{2G^2}{l\left(T+1\right)}$$

If we choose $T=O\left(\frac{G^2}{l\epsilon}\right)$ and $\eta_t=\frac{2}{l(t+1)}$, we can get ϵ error in T steps. □

2 Second order local minima

In convex functions, proving $\nabla f\left(x\right)=0$ is sufficient for finding maximum and minimum of a function, but it is not true for general functions. And set of points $x:\nabla f\left(x\right)=0$ are called stationary points.

Definition 2.1. *A point x is a second order local minima of a function f if $\nabla f\left(x\right)=0$ and $\nabla^2 f\left(x\right)$ is positive semi-definite.*

In an Gradient descent step:

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

If we take the first order Taylor series approximation of $f(x_{t+1})$, we get:

$$\begin{aligned} f(x_{t+1}) &\approx f(x_t) + (x_{t+1} - x_t)^T \nabla f(x_t) \\ &= f(x_t) - \eta ||\nabla f(x_t)||^2 \end{aligned}$$

This approximation works only for small η. The second order Taylor series expansion of $f(x)$ is:

$$\begin{aligned} f(x_{t+1}) \approx\ & f(x_t) + (\nabla f(x_t))^T (x_{t+1} - x_t) \\ &+ \frac{1}{2}(x_{t+1} - x_t)^T \left(\nabla^2 f(x_t)\right)(x_{t+1} - x_t) \end{aligned}$$

However, if function f is $L - smooth$,i.e:

$$||\nabla f(x) - \nabla f(x')|| \leq L||x - x'||$$

and

$$-LI \preceq \nabla^2 f(x) \preceq LI$$

From Lemma 1:

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

Lemma 5 (Descent Lemma)**.** *If $\eta \leq 1/2L$, then:*

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq (\nabla f(x_t))^T (x_{t+1} - x_t) + L||x_{t+1} - x_t||^2 \\ &= -\eta||\nabla f(x_t)||^2 + L\eta^2||\nabla f(x_t)||^2 \\ &\leq -\frac{1}{2}\eta||\nabla f(x_t)||^2 \end{aligned}$$

The final inequality we obtained is at $\eta = 1/2L$.

We can also write the Gradient descent using pre-conditioners update equation as :

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

where, H is the pre-conditioner matrix. Sometimes taking good values of H, η makes the optimization problem easier. It can be seen as shaping the gradient; varying the emphasis in different directions.

Example: If $f(x)$ is a quadratic function then $f(x) = x^T Ax + b^T x + c$, where $A \in \mathbb{R}^{nXn}, b \in \mathbb{R}^n$ *and* $c \in \mathbb{R}$. Then Gradient update equation with pre-conditioner is given as follows:

$$x_2 = x_1 - \eta H^{-1}(2Ax_1 + b)$$

To converge in one step $\nabla f(x_2) = 0$, therefore:

$$2A\left(x_1 - \eta H^{-1}(2Ax_1 + b)\right) + b = 0$$

If $H = A$, $\eta = \frac{1}{2}$, then the above equation becomes 0. Hence it converges in 1 gradient descent step.

In general we can define $H_t = \nabla^2 f(x_t)$, then the Gradient descent with pre-conditioner update become Newton's iterations i.e.:

$$x_{t+1} = x_t - \eta\left(\nabla^2 f(x_t)\right)^{-1} \nabla f(x_t)$$