COMS 4995-8: Advanced Algorithms (Spring'21) April 1, 2021

Lecture 22 : $\alpha-$Strong Convexity, Newton's Method

Instructor: *Alex Andoni* *Scribes: Siddharth Bhutoria, Nihar Maheshwari*

1 Gradient Descent : Recap

For $f : \mathbb{R}^n \to \mathbb{R}$ having continuous first and second derivatives and a local improvement δ, we have the following Taylor approximation:

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

$$\delta = -\eta \cdot \nabla f(x)$$

Note: Since this is an approximation ignoring higher order terms, we have $y \in [x, x+\delta]$
Further, we made the following assumptions on f to get a better bound:

1.1 Assumption 1 : f is β-smooth

If f is β-smooth, we have $\lambda_{max}(\nabla^2 f(y)) \leq \beta$. Upon using this to get an upper bound on $f(x+\delta)$ we have

$$\eta = \frac{1}{2\beta}$$

This gives us a guarantee that we are progressing in each iteration of the descent as long as the gradient is not zero ($\nabla f(x) \neq 0$) as we will always then have $f(x+\delta) < f(x)$

1.2 Assumption 2 : f is Convex

If f is convex, we have that the minimum eigen value of the Hessian is at least 0 ($\lambda_{min}(\nabla^2 f(y)) \geq 0$). We further proved the following bound on the number of iterations T for ϵ approximate solution:

$$T = \mathcal{O}\left(\frac{\beta \cdot D^2}{\epsilon}\right)$$

Where:

$$f(x) - f(x^*) \leq \epsilon \text{ where } x^* = \arg\min f(x)$$

$$D = \max_x \|x - x^*\|$$

Note: Drawbacks of this bound:

1. Dependence over $1/\epsilon$: This could be improved
2. $D = \max_x \|x - x^*\|$: This dependency is on the domain. The value of D is proportional to the value of $\|x^*\|$, which may not be polynomial in input size.

2 Assumption 3 : α-Strong Convexity

Definition 1. *A function f is α-strictly convex if it satisfies*

$$\lambda_{min}(\nabla^2 f(y)) \geq \alpha > 0$$

The minimum eigen value of the Hessian is strictly positive in this case. To get a better bound on T, we are looking to remove the $\|x - x^*\|$ dependency with this assumption.

For $\delta = x - x^*$, from the Taylor approximation at x^* and the $f : \alpha-$strictly convex assumption we have:

$$f(x) = f(x^* + (x - x^*)) \geq f(x^*) + \nabla f(x) \cdot \delta + \frac{1}{2}\delta^T \cdot \alpha \cdot \delta \tag{1}$$

$$\Rightarrow f(x) = f(x^* + (x - x^*)) \geq f(x^*) + \nabla f(x) \cdot \delta + \frac{\alpha}{2}\|x - x^*\|^2 \tag{2}$$

Since x^* is the optimal value we have $\nabla f(x^*) = 0$. Therefore

$$f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2 \tag{3}$$

We will use this result ahead to remove the dependency of T on $\|x - x^*\|^2$

2.1 Progress in each Gradient Descent step

Goal: In each iteration as we move from $x^t \to x^{t+1}$, our aim is to quantify the progress made towards the optimal in each iteration.

From previous lecture, since we decrease $f(x)$ by $(1/2\beta) \cdot \|\nabla f(x)\|^2$ in each iteration, using the inequality from before we have that:

$$f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x^t)\|^2$$

$$\Rightarrow f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - \frac{1}{2\beta} \cdot \frac{\big[f(x^t) - f(x^*)\big]^2}{\|x^t - x^*\|^2}$$

Using **(3)** from above, we have

$$\Rightarrow f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - \frac{1}{2\beta} \cdot \frac{\big[f(x^t) - f(x^*)\big]^2}{(2/\alpha) \cdot [f(x^t) - f(x^*)]}$$

$$\Rightarrow f(x^{t+1}) - f(x^*) = \big[f(x^t) - f(x^*)\big] \cdot \left[1 - \frac{\alpha}{4\beta}\right]$$

Note: The distance to the optimal value of the function drops by a factor of $[1 - (\alpha/4\beta)]$ in each iteration

$\therefore$ starting from $x = x^0$, we have in T steps (applying inequality inductively):

$$f(x^T) - f(x^*) \leq \big[(x^0) - f(x^*)\big] \cdot [1 - \frac{\alpha}{4\beta}]^T$$

For

$$T = \frac{4\beta}{\alpha} \cdot \ln\left(\frac{f(x^0) - f(x^*)}{\epsilon}\right)$$

We have

$$\left(1 - \frac{\alpha}{4\beta}\right)^T \leq \frac{\epsilon}{f(x^0) - f(x^*)}$$

$$\Rightarrow f(x^T) - f(x^*) \leq \big(f(x^0) - f(x^*)\big) \cdot \left(1 - \frac{\alpha}{4\beta}\right)^T \leq \epsilon$$

Hence we have:

Theorem 2. *For any $\alpha-$strongly convex and $\beta-$smooth function f:*

$$T = \mathcal{O}\left(\frac{\beta}{\alpha} \cdot \ln\left(\frac{f(x^0) - f(x^*)}{\epsilon}\right)\right)$$

Remarks:

1. Here, the number of steps / iterations do not depend on $\|x - x^*\|$. Rather T has a logarithmic dependence on the function values $(f(x^0) - f(x^*))$. This greatly improves over the dependence we had over D^2 initially.
2. Here, T has a logarithmic dependence over $1/\epsilon$ which is again an improvement from the result using Assumption 2.
3. T is further proportional to β/α - The Condition Number

Definition 3. *We define the condition number $\kappa(\nabla^2 f(y))$ as:*

$$\kappa(\nabla^2 f(y)) = \frac{\beta}{\alpha} = \frac{\lambda_{max}(\nabla^2 f(y))}{\lambda_{min}(\nabla^2 f(y))}$$

3 Newton Method

To understand Newton's Method, we will go back to Taylor expansion:

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

Note: So far in Gradient Descent, we found bounds of the term $\frac{1}{2}\delta^T \cdot \nabla^2 f(y) \cdot \delta$, where the α-strong convexity lower bounds this term and β-smooth upper bounds this term. Unlike Gradient Descent, Newton's method try to optimize this quadratic form directly instead of finding the bounds. That's why Newton Method is also known as $2nd$ order method.

3.1 Change of Variables

Goal: Optimize $f(x + \delta)$ as a function of δ.
Intuition: Changing variables shouldn't change the correctness.

We define $\Delta = A \cdot \delta$, where A is a full rank matrix, which means $\delta = A^{-1} \cdot \Delta$

Plugging this δ value into Taylor expansion, we get:

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

Remember, in gradient descent, the number of iteration T depends on the condition number of the hessian term where we divided maximum eigenvalue λ_{max} with minimum eigenvalue λ_{min}.
Further, carefully analysing this strongly convex case we can say that T is proportional to condition number of the matrix $(A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1}$. Ideally, we want to make the condition number as small as possible and the smallest value a condition number can take is 1, which implies that this matrix $(A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1}$ is equal to the identity matrix I i.e.

$$(A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1} = I$$

In this case, we have changed variable δ to Δ, which implies that we have changed the Hessian as well, and therefore, we can get a much better condition number using the matrix A.
Now, let's see how we can generate A:

$$(A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1} = I$$

$$\nabla^2 f(y) = A^T A$$

$$\implies A = (\nabla^2 f(y))^{\frac{1}{2}}$$

where A is defined as long as $\lambda_{min}(\nabla^2 f(y)) \geq 0$.

This signify A is the best "change of variables". Next, we will redo the "best local δ step" analysis to find the best step under this change of variable:

$$\delta = \underset{\delta:\Delta = A\delta}{\arg\min} \nabla f(x)^T \cdot A^{-1}\Delta + \frac{1}{2}\Delta^T \cdot \Delta$$

Fixing the norm of Δ, we realize that the direction of Δ is aligned against $\nabla f(x)^T \cdot A^{-1}$. This implies that:

$$\Delta = -\eta\Big[\nabla f(x)^T A^{-1}\Big]^T = -\eta (A^{-1})^T \nabla f(x)$$

Under the nice function assumption Δ is symmetric which implies A is symmetric

$$\Delta = -\eta (A^{-1})^T \nabla f(x)$$

$$\Delta = -\eta \cdot \Big[\nabla^2 f(y)\Big]^{-\frac{1}{2}} \cdot \nabla f(x)$$

Claim 4. *The above quadratic form gets optimized when $\eta = 1$.*

We know $\delta = A^{-1} \cdot \Delta$ so plugging the value Δ, we will have:

$$\delta = -\eta \cdot \Big[\nabla^2 f(y)\Big]^{-1} \cdot \nabla f(x)$$

Issue: There exists a problem that when we compute the best δ, it depends on the gradient of point x but Hessian of point y, where the point y depends on δ as well and we don't know y. Therefore, we cannot use the solution directly.

Remark 5. *Assuming that $\nabla^2 f(y) \approx \nabla^2 f(x)$ i.e. y is close to x, then*

$$\delta = \underset{\delta:\Delta=A\delta}{\arg\min} \nabla f(x)^T \cdot A^{-1}\Delta + \frac{1}{2}\Delta^T \cdot (A^{-1})^T \nabla^2 f(x) A^{-1}\Delta$$

$$= -\eta \Big[\nabla^2 f(x)\Big]^{-1} \nabla f(x)$$

Here, instead of using Hessian of y, we are using Hessian of x. In this case the optimal A to be used is $\Big[\nabla^2 f(x)\Big]^{\frac{1}{2}}$

Remark 6. *Computing δ is harder because when we compute*

$$\delta = -\eta [\nabla^2 f(x)]^{-1} \cdot \nabla f(x)$$

we have to compute matrix inverse, and that can take $O(n^3)$.

4 Theorems for Newton's Method

$$n(x) \triangleq [\nabla^2 f(x)]^{-1} \cdot \nabla f(x)$$

Theorem 7. *Suppose $\exists r > 0$ s.t, $\forall x$ the distance from x^* is $\leq r$:*

1. *$\lambda_{min}(\nabla^2 f(x)) \geq \alpha$, implies the function is α-strong convex.*
2. *$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L \cdot ||x - y||$, where $||A||$ is spectral norm as well as the max singular value.*

Then $\forall x^0$ at distance $\leq r$ from x^*, we have

$$x^1 \triangleq x^0 + n(x)$$

$$||x^1 - x^*|| \leq \frac{L}{2\alpha} \cdot ||x^0 - x^*||^2$$

This theorem suggests that if we update x with Newton step, the distance between the updated point and the optimal point will drop quadratically.

To show the significance, let's define a new variable $\gamma = \frac{2\alpha}{L}$, then we can re-write the above inequality as follows:

$$||\frac{x^1 - x^*}{\gamma}|| \leq \frac{L^2}{(2\alpha)^2} ||x^0 - x^*||^2 = ||\frac{x^0 - x^*}{\gamma}||^2$$

$$\implies ||\frac{x^T - x^*}{\gamma}|| \leq ||\frac{x^0 - x^*}{\gamma}||^{2^T}$$

Since 2^T is a very large number, if we set $||\frac{x^0-x^*}{\gamma}||$ to some large number, we actually do not have any bound, therefore, we define $||\frac{x^0-x^*}{\gamma}|| \leq 0.9$, which implies:

$$T = O(\log\log\frac{1}{\epsilon})$$

is enough for $||\frac{x^T-x^*}{\gamma}|| \leq \epsilon$.

Remark 8. *A big caveat of Newton's method is that* $||x^0 - x^*|| \leq \frac{9\gamma}{10} = \frac{18\alpha}{10L}$*, which means that* x^0 *our starting position is close to the optimal point* x^*

Remark 9. *Newton's method works for "warm start"* x^0*.*

5 Interior Point Method

Consider a linear program of the form,

$$\min_{x\in K} f(x) \tag{4}$$

Example:-

$$\begin{aligned} K &= \{x : Ax \leq b\} \\ f(x) &= c^T x \end{aligned} \tag{5}$$

where $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

We want to relax the minimization by defining a new function such that we have an unconstraint optimization framework similar to what we had in Gradient Descent.

Idea: We have already seen one way to turn this into an unconstrained problem, by replacing the objective function with one that evaluates to $f(x)$ for $x \in K$ and to $+\infty$ otherwise.

$$DefineF(x) = \begin{cases} f(x) & \text{if } x \in K \\ +\infty & \text{if } x \notin K \end{cases}$$

But there is an issue with such a function, as function isn't smooth/continuous and so it does not work well with the gradient descent method or Newton's method. The boundaries for the function acts as some repellant force and so we should penalize if the x comes too close to boundary or it goes in opposite direction.

We require a smoother function and next time we will see how to make the function smoother and solve this linear program using Interior Point Method.