COMS 4995-2: Advanced Algorithms (Spring'20) Mar 31, 2020

Lecture 17: Newton's method

Instructor: *Alex Andoni* Scribes: *Haixu Liu*

1 Reminder: Gradient Descent

If a function f is "nice", we can use the following Taylor expansion on it:

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

The gradient descent step can be described as:

$$\delta = -\eta \nabla^2 f(x)$$

We also mentioned two assumptions that can help us find the better bound:

1. β-smooth:

   $$\lambda_{max}(\nabla^2 f(x)) \leq \beta$$

   which means the maximum eigenvalue of the Hessian matrix is upper bounded by β, and we have proved that $\delta = -\frac{1}{\beta}\nabla f(x)$ is the optimal if $\frac{1}{2}\delta^T \nabla^2 f(x)\delta \leq \frac{\beta}{2}||\delta||^2$. This is still not enough, therefore, we introduced a new progress or step: $\delta = -\frac{1}{2\beta}||\nabla f(x)||^2$.

2. convex:

   $$\lambda_{min}(\nabla^2 f(x)) \geq 0$$

   which means the minimum eigenvalue of the Hessian matrix is at least 0. This allows us to use three more properties:

   (a) if $\nabla f(x) = 0$, then we can say x is the global minimum.

   (b) we have proved in the last class that $||\nabla f(x)|| \geq \frac{f(x)\cdot f(x^*)}{||x-x^*||}$, where $x^* = \arg\min f(x)$.

   (c) we can achieve ϵ additive approximation, meaning that we can get $f(x^T) - f(x^*) \leq \epsilon$ using $T = O(\frac{\beta D^2}{\epsilon})$ steps, where $D \approx max||x - x^*||$.

Using the two assumptions we can say that we can find an optimal solution for a given function if it isβ-smooth as well as convex. However, this is not good enough, because we still have ϵ in time complexity which means it is pseudo-polynomial, and we do not have a good control on vector D.

2 α-Strong Convexity

Definition 1. *If function f satisfies $\lambda_{min}(\nabla^2 f(x)) \geq \alpha$, where $\alpha > 0$, we say that f is a α-strong convex function, note that here we assume the minimum eigenvalue of the Hessian matrix is strict positive.*

Now we consider how to remove dependency on $||x^* - x^0||$:

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

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

By applying Taylor Expansion on x^*. Since x^* is optimal, therefore we have: $\nabla f(x^*) = 0$, which implies that:

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

2.1 Progress Per Step

Now we consider how much progress can we make per step during gradient descent, in other word, we need to find an upper bound of the progress using the inequality we have proved before:

$$f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - \frac{1}{2\beta}||\nabla f(x^t)||^2 \tag{4}$$

$$\leq f(x^t) - f(x^*) - \frac{1}{2\beta} \cdot \frac{[f(x^t) - f(x^*)]^2}{||x^t - x^*||^2} \tag{5}$$

$$\leq f(x^t) - f(x^*) - \frac{\alpha}{4\beta} \cdot \frac{[f(x^t) - f(x^*)]^2}{[f(x^t) - f(x^*)]} \tag{6}$$

$$= (1 - \frac{\alpha}{4\beta})(f(x^t) - f(x^*)) \tag{7}$$

Thus, we can say that after T steps

$$f(x^T) - f(x^*) \leq [f(x^0) - f(x^*)] \cdot (1 - \frac{\alpha}{4\beta})^T \tag{8}$$

Remember that our goal is to find an x^T that satisfies $f(x^T) - f(x^*) \leq \epsilon$, it is enough to say that when we have such T:

$$T = \frac{4\beta}{\alpha} \cdot \ln(\frac{f(x^0) - f(x^*)}{\epsilon}) \tag{9}$$

It can be proved that:

$$f(x^T) - f(x^*) \leq [f(x^0) - f(x^*)] \cdot e^{-\frac{\alpha}{4\beta} \cdot T} = \epsilon \tag{10}$$

by plug in T.

Remark 2. *The progress is independent on* $||x^0 - x^*||$.

Remark 3. *The dependence on* ϵ *is logarithmic.*

Remark 4. *T is proportional to* $\frac{\beta}{\alpha} = \frac{\lambda_{max}}{\lambda_{min}}$ *of* $\nabla^2 f(x)$*, here we call* $\frac{\beta}{\alpha}$ *condition number.*

3 Newton Method

Go back to our Taylor expansion:

$$f(x + \delta) = f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\delta^T \cdot \nabla^2 f(y) \cdot \delta \tag{11}$$

Note that what we have done so far is find bounds of the term $\frac{1}{2}\delta^T \cdot \nabla^2 f(y) \cdot \delta$, β-smooth upper bounds this term, and the convexity lower bounds this term. Newton method, however, tries to optimize this quadratic form directly instead of finding bounds.

3.1 Change of Variables

Assume $\Delta = A \cdot \delta$, where A is a linear matrix, and it is full rank, which means $\delta = A^{-1} \cdot \Delta$.

Plug this equation into Taylor expansion:

$$f(x + \delta) = f(x) + \nabla f(x)^T A^{-1} \Delta + \frac{1}{2}\Delta^T (A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1} \cdot \Delta \tag{12}$$

The strongly convex case analysis says that T is proportional to condition number of matrix $(A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1}$. Therefore, ideally we want this matrix to have condition number 1, this implies that this matrix $(A^{-1})^T \cdot \nabla^2 f(y) \cdot A^{-1}$ is equal to identity matrix I. In this case, we change variable δ to Δ, which means we change the Hessian as well, and therefore, we can get a much better condition number using matrix A.
Now we need to consider how to generate A:

$$\nabla^2 f(y) = A^T A \tag{13}$$

$$\implies A = (\nabla^2 f(y))^{\frac{1}{2}} \tag{14}$$

This means A is the best "change of variables". Redo the "best local δ step" analysis and find the best step:

$$\delta = \arg\min_{\delta:\Delta=A\delta} \nabla f(x)^T \cdot A^{-1}\Delta + \frac{1}{2}\Delta^T \cdot \Delta \tag{15}$$

if we fixed the norm of Δ, the direction of Δ is aligned against $\nabla f(x)^T \cdot A^{-1}$. Therefore, we consider:

$$\Delta = -\eta[\nabla f(x)^T A^{-1}]^T = -\eta(A^{-1})^T \nabla f(x) \tag{16}$$

Since $\delta = A^{-1}\Delta$, we have:

$$\Delta = -\eta[\nabla f(x)^T A^{-1}]^T = -\eta(A^{-1})^T \nabla f(x) \tag{17}$$

$$= -\eta[\nabla^2 f(y)]^{-1} \cdot \nabla f(x) \tag{18}$$

Claim 5. *When $\eta = 1$ is the optimizer of the whole quadratic form, but there is still 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. Therefore, the solution above cannot be used directly.*

Remark 6. *Assume that $\nabla^2 f(y) \approx \nabla^2 f(x)$ when y is close to x, then*

$$\delta = \arg\min_{\delta:\Delta=A\delta, ||A|| \leq \epsilon} \nabla f(x)^T \cdot A^{-1}\Delta + \frac{1}{2}\Delta^T \cdot (A^{-1})^T \nabla^2 f(x) A^{-1}\Delta \tag{19}$$

$$= -\eta[\nabla^2 f(x)]^{-1} \nabla f(x) \tag{20}$$

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

Remark 7. *Gradient descent step depends on $\nabla^2 f(x)$, which is called second-order method, this means that we are not only using the assumption on the Hessian here, we are actually using the value of the Hessian.*

Remark 8. *Gradient descent step is computationally harder because when we compute*

$$\delta = -\eta[\nabla^2 f(x)]^{-1} \cdot \nabla f(x) \tag{21}$$

we need to compute an inverse of a matrix, this can be $O(n^3)$.

4 Theorems for Newton's Method

Theorem 9. *Suppose $\exists r > 0$ s.t, $\forall x$ the distance from x^* is at most r:*

1. *$\lambda_{min}(\nabla^2 f(x)) \geq \alpha$, means 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.*

Set $x^1 \triangleq x^0 + n(x)$, where $n(x) = -\eta[\nabla^2 f(x)]^{-1}\nabla f(x)$ is Newton step, then we can have:

$$||x^1 - x^*|| \leq \frac{L}{2\alpha} \cdot ||x^0 - x^*||^2 \tag{22}$$

where x^0 is at distance at most r from x^*.

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

To show why this is useful, we define a new variable $\gamma \triangleq \frac{2\alpha}{L}$, the above inequality is equal to:

$$||\frac{x^1 - x^*}{\gamma}|| \leq ||\frac{x^0 - x^*}{\gamma}||^2 \tag{23}$$

$$\implies ||\frac{x^T - x^*}{\gamma}|| \leq ||\frac{x^0 - x^*}{\gamma}||^{2^T} \tag{24}$$

Since 2^T is a very large number, if we set $||\frac{x^0-x^*}{\gamma}||$ to be a large number, we actually do not have any bound here, therefore, we need $||\frac{x^0-x^*}{\gamma}|| \leq \frac{1}{2}$, which implies:

$$T = O(\log\log\frac{1}{\epsilon}) \tag{25}$$

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

Observation 10. *A big caveat of Newton's method is that $||x^0 - x^*|| \leq \frac{\gamma}{2} = \frac{\alpha}{2}$*

Observation 11. *Newton's method works for "warm start" x^0.*

5 Next Time

Next time we will talk about Interior Point Method for linear programming.