EE C227C Convex Optimization and Approximation	Lecture 15 - 03/08/2022

Lecture 15: Newton's Method and Self-Concordance

Lecturer: Jiantao Jiao *Scribe: Gaurav Ghosal, Lei Shi*

In this lecture, we identify the *correct* perspective of analyzing Newton's method. This allows us to develop the notion of a self-concordant function which turns out to be very important in interior point methods for solving constrained optimization problems

Recall the *classical* analysis of Newton's Method from the previous method. We assumed that $f(x)$ is:

1. Twice differentiable
2. μ-Strongly Convex: $\nabla^2 f(x^*) \succcurlyeq \mu I_n$
3. M-Lipschitz Hessian: $\forall x, y \ \left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq M \left\|x - y\right\|$

In this lecture, we will argue that the this type of analysis does not characterize the true properties of Newton's method well. To begin, we will introduce the Affine Invariance of Newton's Method

1 Affine Invariance of Newton's Method

We will show that Newton's method is invariant to affine changes in the coordinate system of x. Suppose we have a twice differentiable function f and a non-singular matrix $B \in \mathbb{R}^{n \times n}$. Then we define the function $\phi(y) = f(By)$ and introduce the following lemma:

Lemma 1. *Suppose the sequence $\{X_k\}$ and $\{Y_k\}$ are the points generated by Newton's method on f and ϕ, respectively and suppose we have an X_0 and $Y_0 = B^{-1}X_0$. Then $Y_k = B^{-1}X_k$*

Thus, we see that the corresponding objective values attained at each iteration are the same for both functions. This indicates that the trajectory of points selected by Newton's method is somehow invariant to affine transformations of the decision variable.

However, this highlights a weakness of the classical analysis presented in the last lecture. The bound attained in the last lecture is **not** affine invariant. This implies that the classical analysis does not truly capture the meaningful features of newton's method. In order to adapt our analysis, we replace the Lipschitz smoothness assumption on the Hessian.

2 New Assumption

We wish to impose the condition that the Hessian of f does not change very much. It turns out that this can be done in an affine-invariant way using the concept of *self-concordance*. Before we address this concept, however, we need to generalize our notion of derivatives for multivariable functions. This generalization gives rise to the concept of a differentiable form:

2.1 Differential Form

Previously in this course, we have made use of the first order and second order derivatives of multi variable functions. These derivatives take the form of vectors and matrices, respectively. Suppose $f \in C^3$ (that is f is a thrice continuously differentiable function). Then, we have define the third derivative of f in a direction u, at a point x, as

$$f'''(x)[u] = \lim_{\alpha \to 0} \frac{\nabla^2 f(x + \alpha u) - \nabla^2 f(x)}{\alpha}$$

Using this notation, we see that the third derivative of f, alternatively written as $D^3 f$ is a multi-dimensional generalization of matrices, called a *tensor*. This tensor defines a *multi-linear map* in the following way:

$$D^k f(x)[h_1, \ldots, h_k] = \frac{\partial^k}{\partial t_1 \partial t_2 ... \partial t_k} f(x + t_1 h_1 + ... + t_k h_k)\Big|_{t_1=t_2=...=t_k=0}$$

This result coincides well with some of the intuition that we have developed in the first and second derivative cases and gives a general notion of higher-order derivatives in multi variable functions.

2.2 Developing the New Assumption

The previous Lipschitz assumption that we imposed can be expressed using our new differential form notation, as well as the fact that f is thrice differentiable.

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq M \left\|x - y\right\| \forall x, y \Rightarrow \left\|D^3 f(x)[u]\right\| \leq M \left\|u\right\|$$

Noting that $D^3 f(x)[u]$ is a matrix, we observe that we can reduce $D^3 f(x)[u]$ to a scalar by taking the quadratic form with a vector v. This yields a tri-linear form

$$\langle D^3 f(x)[u]v, v\rangle = D^3 f(x)[u, v, v] \leq M \left\|u\right\| \left\|v\right\|^2$$

Observe that the quantity $D^3 f(x)[u, v, v]$ is a generalized directional derivative and so it is affine equivariant. However, the right hand side is not affine equivariant. In particular, recalling that $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$, this can be attributed to the fact that each component contributes equally to the norm. However, in order to be affine equivariant, the contribution of each component needs to be dynamically scaled in conjunction with our affine transformation. This inspires us to change the right-hand side to be affine equivariant and leads us to the following assumption:

$$\left|D^3 f(x)[u, u, u]\right| \leq 2(D^2 f(x)[u, u])^{\frac{3}{2}}$$

and we define this condition to be *self-concordance*. We note that the scaling constant in the first definition is absent in this definition and has been generically replaced by 2. It turns out that it is always possible to scale the a self-concordant function f in a way that it satisfies the definition with the constant 2. The actual value of 2 is selected out of convenience to ensure that the log function is self-concordant.

3 Properties of self-concordant functions

The definition of self-concordance only involves bounds on differential forms of the type $D^3 f(x)[u, u, u]$ and $D^2 f(x)[u, u]$. One might wonder whether we should establish conditions involving different u_1, u_2, u_3 instead of a single u. The following lemma suggests this definition already implies some bound on general tri-linear forms.

Another key observation is that in the new definition, both sides represent a directional derivative. This is the key observation to understanding that both sides of the inequality are affine invariant.

Lemma 2. *If f is self-concordant, then*

$$|D^3 f(x)[u_1, u_2, u_3]| \leq 2\Pi_{i=1}^3 \sqrt{u_i^\top \nabla^2 f(x) u_i}.$$

For a formal proof, see the "basic inequality" results in Section 2.2 [1]. Of interest is the following definition of local norm:

Definition 3. *Define* $\|y\|_x \triangleq \sqrt{y^\top \nabla^2 f(x) y}$*. This is also denoted as* $\|y\|_{\nabla^2 f(x)}$ *where we follow the tradition that* $\|y\|_A = \sqrt{y^\top A x}$ *for some positive semi-definite* $A \in \mathbb{R}^{n\times n}$*.*

We now present some extremely simple combination rule which can be powerful in proving self-concordance of complex functions and constructing new self-concordant function from known ones.

Lemma 4 (Combination rules)**.** *The following conbmination rules hold for self-concordant functions:*

1. ***Affine substitution.*** *Let* F *be self-concordant on a open convex set* $Q \subset \mathbb{R}^n$*.* $x = Ay + b.y \in \mathbb{R}^k, x \in \mathbb{R}^n$*, then* $Q^+ = \{y \in \mathbb{R}^k \mid Ay + b \in Q\}$ *is an open convex set in* $\mathbb{R}^k$*. The function*

$$F^+(y) = F(Ay + b) : Q^+ \to \mathbb{R}$$

is self-concordant on Q^+*.*

2. ***Linear combination.*** *Let* $F_1, \cdots, F_m$ *be self-concordant functions and* $\alpha_i \geq 1, i = 1, \cdots, m$*. Then* $\sum_{i=1}^m \alpha_i F_i$ *is also self-concordant.*

Proof

Before diving into the formal proof, we present a useful lemma:

Lemma 5. *Suppose* $a_i \geq 1, b_i \geq 0$ *for* $i = 1, \cdots, m$*. We have*

$$\sum_{i=1}^m a_i b_i^{\frac{3}{2}} \leq \left(\sum_{i=1}^m a_i b_i\right)^{\frac{3}{2}}.$$

To see this is true, by rescaling both sides of the inequalities, it suffices to assume $\sum_{i=1}^m a_i b_i = 1$*. Then the lemma translates into an optimization problem:*

$$\max_b \sum_{i=1}^m a_i b_i^{\frac{3}{2}}, \ \text{s.t. } b_i \geq 0, \sum_{i=1}^m a_i b_i = 1.$$

The objective function is convex, hence the maximum is always attained at the boundary. In particular, in this case the constraint set is a simplex, and the maximum is obtained at some $b^\star = (0, \cdot, b_j, \cdots, 0)$ *for some* j*. Now we know* $b_j a_j \leq 1$*, or* $b_j \leq 1/a_j$*. Therefore*

$$\sum_{i=1}^n a_i b_i^{\frac{3}{2}} \leq a_j b_j^{\frac{3}{2}} \leq \frac{1}{\sqrt{a_j}} \leq 1.$$

Proof for Lemma 5 is done.

Now we prove the main statement by showing the following sequence of inequalities hold:

$$\begin{aligned}
&\left|\sum_{i=1}^m \alpha_i D^3 F_i(x)[h,h,h]\right| \\
\leq &\sum_{i=1}^m \alpha_i \left|D^3 F_i(x)[h,h,h]\right| \ \textit{(by triangle inequality)} \\
\leq &\sum_{i=1}^m 2\alpha_i \left\{D^2 F_i(x)[h,h]\right\}^{\frac{3}{2}} \ \textit{(by self-concordance of each } F_i(x)\textit{)} \\
\leq &2\left\{D^2 F(x)[h,h]\right\}^{\frac{3}{2}} \ \textit{(by Lemma 5)} \\
= &2\left\{\sum_{i=1}^m \alpha_i D^2 F_i(x)[h,h]\right\}^{\frac{3}{2}} \ \textit{(since differentiation is a linear operator).}
\end{aligned}$$

$\square$

3. ***Direct sum.*** *Let* $F_i(x_i)$ *be self-concordant functions. Then* $F(x_1, \cdots, x_m) = \sum_{i=1}^m F_i(x_i)$ *is also self-concordant.*

With these combination rules in hand, it is actually easy to show the following example of self-concordant function:

Example 6. The function $F(x) = -\sum_{i=1}^m \log(b_i - a_i^\top x)$ is self-concordant. One can prove this by showing that:

- $f(y) = -\log(y)$ is self-concordant.
- Affine substitution preserves self-concordance, hence $f_i^+(x) = -\log(b_i - a_i^\top x)$ is self-concordant.
- Certain linear combination preserves self-concordance. Here $F(x)$ is a summation of $f_i^+(x)$ with $a_i = 1$, hence F is self-concordant.

References

[1] A. Nemirovski, "Interior point polynomial time methods in convex programming," *Lecture notes*, vol. 42, no. 16, pp. 3215–3224, 2004.