EE C227C Convex Optimization and Approximation Lecture 16 - 03/10/2022

Lecture 16: Self-concordant Functions and Newton's Method

Lecturer: Jiantao Jiao *Scribe: Yimeng Wang*

In the last lecture, we pointed out several issues with our initial analysis of Newton's Method. Moreover, we introduced the concept of self-concordant functions which will be the main focus of variants of Newton's Methods. In this lecture, we will refine our definition of self-concordant functions and discuss several nice properties of them. In the end, we will see how these properties come in place while analyzing variants of Newton's methods.

1 Definition of Self-Concordant Functions

Definition 1. *The epigraph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as*

$$Epigraph(f) = \{(x, t) | t \geq f(x),\ x \in dom(f)\}$$

where $dom(f) \triangleq \{x \in \mathbb{R}^n |\ |f(x)| < \infty\}$ is the domain of f.

We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is closed if the epigraph of f is closed. With this notion in mind, we can now give the definition of self-concordant functions.

Definition 2. *(Self-concordant functions) Let $f \in C^3(dom(f))$ (i.e., f''' exists and is continuous) be a closed convex function such that $dom(f)$ is open. Then, we say f is self-concordant if there exists $M_f \geq 0$ such that*

$$|D^3 f(x)[u, u, u]| \leq 2M_f \|u\|^2_{\nabla^2 f(x)}$$

for all $x \in dom(f)$. We say f is standard self-concordant if $M_f = 1$.

Here, $D^3 f(x)[u, u, u]$ is the evaluation of three-dimensional tensor (defined in last lecture) and $\|v\|_x = \sqrt{v^T \nabla^2 f(x) v}$ for $v \in \mathbb{R}^n$. This definition of norm is valid since f is convex. Notice that in real-life optimization problems the domain may not be open. In such cases, we will apply techniques like adding slack variables to address this issue. But we will not be discussing those in this lecture.

The definition above requires function f to be closed. The purpose of this is to rule out some ill-behaved functions that are not of interest. See the Figure belowfor an example of such functions. We can definitely restrict ourselves to continuous functions but closed functions are more general than continuous functions and they are also nice enough to work with.

2 Properties of self-concordant functions

In this section, we will discuss various nice properties of self-concordant functions. The proofs of the results below can be found in Nesterov's book [1].

Lemma 3. *(Theorem 5.1.1 of [1]) Let $f_1 : \mathbb{R}^n \to \mathbb{R}$ be self-concordant with parameter M_1 and $f_2 : \mathbb{R}^n \to \mathbb{R}$ be self-concordant with parameter M_2. Let $\alpha, \beta > 0$. Then the function $f(x) = \alpha f_1(x) + \beta f_2(x)$ is self-concordant with parameter*

$$M_f = \max\left\{\frac{M_1}{\sqrt{\alpha}}, \frac{M_2}{\sqrt{\beta}}\right\}$$

and $dom(f) = dom(f_1) \cap dom(f_2)$.

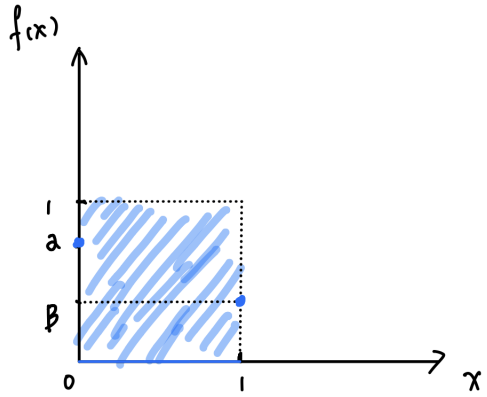

Figure 1: Consider the function $f(x) = \alpha$ when $x = 0$, $f(x) = \beta$ when $x = 1$ and $f(x) = 0$ for all $x \in (0, 1)$. This function is clearly convex but the epigraph (shaded area) is not closed.

As an immediate consequence of the lemma, we have the following corollary.

Corollary 4. *If f is self-concordant with parameter M, then $\phi = \alpha f$ is $\frac{M}{\sqrt{\alpha}}$ self-concordant for $\alpha > 0$.*

Proof Take $f_1 = f$ and $f_2 = 0$. Then apply the lemma. □

Lemma 5. *(Theorem 5.1.2 of [1]) Affine Invariance: Suppose f is M_f self-concordant. Let $L(x) = Ax + b$ be a linear operator. Then $\phi(x) = f(L(x))$ is also self-concordant with $M_\phi = M_f$.*

We are now in the place to write down the main inequalities of self-concordant functions. To this end, recall the definition of Dikin's Ellipsoid:

$$W^o(x; r) = \{y \in \mathbb{R}^n | \|y - x\|_x \leq r\}$$

where $\|h\|_x = \sqrt{h^T \nabla^2 f(x) h}$. The dual norm of $\|\cdot\|_x$ is defined as:

$$\|g\|_x^* = \sqrt{g^T (\nabla^2 f(x))^{-1} g}$$

As a result, by Holder's inequality, we have $|h^T g| \leq \|h\|_x \cdot \|g\|_x^*$.

Theorem 6. *(Theorem 5.1.5, 5.1.8, 5.1.9 of [1]) Let $W(x; r)$ be the closure of $W^o(x; r)$, i.e.*

$$W(x; r) = cl(W^o(x; r)) = \{y \in \mathbb{R}^n \mid \|y - x\|_x \leq r\}$$

Then, we have:

1. *For any $x \in dom(f)$, we have $W^o\left(x; \frac{1}{M_f}\right) \subseteq dom(f)$.*
2. *Six necessary and sufficient equivalent characterizations of self-concordant functions:*
   (a) *For all $x, y \in dom(f)$, the following inequality holds:*

$$\|y - x\|_y \geq \frac{\|y - x\|_x}{1 + M_f \|y - x\|_x}$$

(b) If $\|y-x\|_x < \frac{1}{M_f}$, *then*

$$\|y-x\|_y \leq \frac{\|y-x\|_x}{1+M_f\|y-x\|_x}$$

(c) For all $x, y \in dom(f)$,

$$\langle \nabla f(y) - \nabla f(x), y-x \rangle \ \geq \frac{\|y-x\|_x^2}{1+M_f\|y-x\|_x}$$

(d) For all $x, y \in dom(f)$,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{M_f^2}\omega(M_f\|y-x\|_x)$$

where $\omega(t) = t - \ln(1+t)$

(e) Suppose $x \in dom(f)$. *If* $\|y-x\|_x < \frac{1}{M_f}$, *then*

$$\langle \nabla f(y) - \nabla f(x), y-x \rangle \leq \frac{\|y-x\|_x^2}{1-M_f\|y-x\|_x}$$

(f) Suppose $x \in dom(f)$. *If* $\|y-x\|_x < \frac{1}{M_f}$, *then*

$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{M_f^2}w_*(M_f\|y-x\|_x)$$

where $w_*(t) = -t - \ln(1-t)$

Here one can show that w and w_* are Fenchel duals to each other. We will not prove any of the inequalities above but interested readers can find the proof of each statement in Nesterov's book [1]. We have seen similar inequalities back when we were analyzing first-order methods. Notice that the inequalities above provide both lower and upper bounds of the quantities of interest. These inequalities are important in our analysis of the Interior Point Method and damped Newton Method which will show up later in the class.

The following theorem shows stability of the Hessian of a self-concordant function.

Theorem 7. *(Theorem 5.1.7 of [1]) Let* $x \in dom(f)$ *and* $y \in W^o(x; 1/M_f)$, *then we have,*

$$(1-M_f r)^2\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1-M_f r)^2}\nabla^2 f(x)$$

3 Damped Newton's Method

In this section, we introduce Damped Newton's Method which utilizes some of the inequalities of self-concordant functions introduced above. To this end, we define the local norm of gradient, or the Newton decrement $\lambda_f(x)$ as

$$\lambda_f(x) = \sqrt{(\nabla f(x))^T(\nabla^2 f(x))^{-1})(\nabla f(x))} = \|\nabla f(x)\|_x^*$$

In each iteration of Damped Newton's Method, we make the following update:

$$x_{k+1} = x_k - \frac{1}{1+M_f\lambda_f(x_k)}[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$$

for $k \geq 0$.

Theorem 8. *Given the Damped Newton's Method update, for a self-concordant function f, we have*

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{M_f^2}\omega(M_f\lambda_f(x_k))$$

In other words, Damped Newton's Method is a descent method.

In previous lectures, we mentioned that the original Newton's method is not used in practice since its convergence to the minimizer is not guaranteed. Instead, we combine variants of Newton Method's. Notice that since ω is a monotonically increasing function, by the theorem above, Damped Newton's Method descends fast when $\lambda_f(x_k)$ is large. Therefore, in practice we usually first run Damped Newton's Method up to quadratic convergence regime, which we will define later. Once in the quadratic convergence regime, we will switch to another method with does more aggressive steps than Damped Newton's Method.

But how can we possibly come up with this step size $\eta = \frac{1}{1+M_f\lambda_f(x_k)}$? It turns out to be a consequence of the inequalities of self-concordant functions. Set $\eta > 0$ to be the learning rate, i.e.

$$x_{k+1} = x_k - \eta[\nabla^2 f(x_k)]^{-1}\nabla f(x_k).$$

Recall that $\omega_*(t) = -t - \ln(1-t)$. Then $\omega_*'(t) = \frac{t}{1-t}$. For simplicity of notation, denote $g = \nabla f(x_k)$ and $H = \nabla^2 f(x_k)$. Notice we have that $g^T H^{-1} g = \lambda_f^2(x_k)$. Then by (f) from the self-concordant inequalities, we have:

$$\begin{aligned} f(x_{k+1}) - f(x_k) \leq\ & \langle \nabla f(x), x_{k+1} - x_k\rangle + \frac{1}{M_f^2} w_*(M_f\|x_{k+1} - x_k\|_{x_k}) \\ =& < g, -\eta H^{-1} g > + \frac{1}{M_f^2} w_*(M_f\sqrt{\eta(H^{-1}g)^T H H^{-1} g\eta}) \\ =& -\eta g^T H^{-1} g + \frac{1}{M_f^2} w_*(\eta M_f \lambda) \end{aligned}$$

Let $h(\eta) = -\eta g^T H^{-1} g + \frac{1}{M_f^2} w_*(\eta M_f\lambda)$. Since h is a convex function of η, we can set its derivative to 0 and solve for η to obtain the tightest bound:

$$\frac{d}{d\eta}h(\eta) = 0 \Rightarrow \eta = \frac{1}{1+M_f\lambda}$$

which gives us the update rule of Damped Newton's Method. This is an example of how properties of self-concordant functions can motivate algorithm design.

4 General Newton's Method

For general Newton's Method, we have the following update rule:

$$x_{k+1} = x_k - \frac{1}{1+\xi_k}[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$$

Variants of Newton's method use different values of ξ_k:

- Standard Newton's Method: $\xi_k = 0$
- Damped Newton's Method: $\xi_k = M_f\lambda_{f,k}$
- Intermediate Newton's Method: $\xi_k = \frac{M_f^2\lambda_{f,k}^2}{1+M_f\lambda_{f,k}} = M_f\lambda_{f,k}\frac{M_f\lambda_{f,k}}{1+M_f\lambda_{f,k}}$

Notice that the ξ_k in Intermediate Newton's Method is smaller than the ξ_k in damped Newton's Method. Hence in practice, we tend to use Intermediate Newton's Method in the quadratic convergence regime which is defined as: $\{k|\lambda_k < c\}$ for some threshold value c that depends on other parameters including M_f.

Let's now describe the local convergence of different variants of the Newton's Method. Notice that we can measure the convergence of these schemes in different ways. We can for instance, estimate the rate of convergence for $f(x_k) - f(x_f^*)$ or for the local norm of the gradient $\lambda_f(x_f) = \|\nabla f(x_k)\|_{x_k}^*$ or the *local distance to the minimum* $\|x_k - x_f^*\|_{x_k}$. Finally, we can look at the distance to the minimum in a fixed metric $r_*(x_k) = \|x_k - x_f^*\|_{x_f^*}$. One result is that all these metrics are equivalent as shown in the following theorem:

Theorem 9. *(Theorem 5.2.1 of [1]) Suppose $\lambda_f(x) < \frac{1}{M_f}$. Then*

(1) $\omega(M_f\lambda_f(x)) \leq M_f^2(f(x) - f(x_f^*)) \leq \omega_*(M_f\lambda_f(x))$

(2) $\omega'(M_f\lambda_f(x)) \leq M_f\|x - x_f^*\|_x \leq \omega_*'(M_f\lambda_f(x))$

(3) $\omega(M_f r_*(x)) \leq M_f^2(f(x) - f(x_f^*)) \leq \omega_*(M_f r_*(x))$

where x_f^* is the global minimizer of f.

Next time, we will examine how λ_f behaves in variants of Newton's Method and self-concordant barriers.

References

[1] Y. Nesterov, *Lectures on convex optimization.* Springer, 2018.