MIT 6.7220/15.084 — Nonlinear Optimization Apr 16-18th 2024

Lecture 14A-B
Self-concordant functions

Instructor: Prof. Gabriele Farina (✉ gfarina@mit.edu)*

In this lecture, we return to Hessian preconditioning and Newton's method, and extend the strong guarantees—including the quadratic convergence rate to optimality—to a class of functions, called *self-concordant functions*. Crucially, the set of self-concordant functions includes certain functions that "shoot to infinity" at the boundary of their domain, such as the function $-\log(x)$ on the domain $(0, \infty)$. The material we will cover today serves as a foundation for the analysis of interior-point methods, which we will discuss in the next lecture.

1 Two shortcomings of the standard analysis

In Lecture 12, we have taken a look at the standard analysis of Newton's method. The key result we established was that the method converges double exponentially fast (in distance) to a local minimum of a function f if the Hessian of f is M-Lipschitz continuous and the curvature of the function around the minimum is sufficiently large (say, $\nabla^2 f(x_\star) \succcurlyeq \mu I$). However, this analysis has two shortcomings.

1.1 Inability to handle log functions

The first shortcoming is practical. The analysis we gave in Lecture 12 breaks down for functions that "shoot to infinity" at the boundary of their domain. This is the case—for example—of functions such as $f(x) = x - \log(x)$, $f(x) = -\log(x) - \log(1-x)$, *et cetera*. These functions have enormous applications in optimization, but their Hessian is *not* Lipschitz continuous. Yet, as we show next, Newton's method converges to the minimum at a double exponential rate.

Example 1.1.

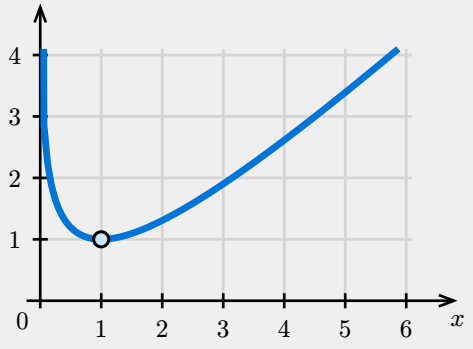
Consider the function $f : (0, \infty) \to \mathbb{R}$

$$f(x) = x - \log(x),$$

whose minimum is at $x = 1$. The update of Newton's method is

$$\begin{aligned} x_{t+1} &= x_t - [f''(x_t)]^{-1} f'(x_t) \\ &= x_t - x_t^2 \cdot \left(1 - \frac{1}{x_t}\right) = 2x_t - x_t^2. \end{aligned}$$

So, we have $1 - x_{t+1} = 1 + x_t^2 - 2x_t = (1 - x_t)^2$.



*These notes are class material that has not undergone formal peer review. The TAs and I are grateful for any reports of typos.

This shows that not only does Newton's method converge to the minimum at a quadratic rate, but it does so with *equality*. One can check very easily that changing the parameterization from x to kx also rescales the radius of convergence by k.

不用Lipschitz条件，怎么描述smoothness

The above hints to the fact that requiring Lipschitz continuity of the Hessian is not the most natural condition for studying the quadratic convergence of Newton's method. Of course, if Lipschitz continuity has to go, one has to wonder: "*what other condition ensuring smoothness can we impose?*" We will propose one later today, called *self-concordance*.

对自变量施加一个affine transformation，收敛半径会同时改变

1.2 Lack of affine invariance in the radius of fast convergence

The second shortcoming is *conceptual*. For $\eta = 1$, Newton's method repeatedly moves to the minimum of the quadratic approximation of f around the current iterate. This point is independent of the choice of coordinates used to parameterize the space, and only depends on the function itself. Hence, the radius of fast convergence is also independent of the choice of coordinates.

However, the radius of fast convergence we gave in Lecture 12 was *not* affine invariant. For example, consider a function $f(x)$. If we now make a transformation of coordinates and consider the function $f(Ax)$, the constants M and μ change, and in general we arrive at a different radius of fast convergence.

Part of the issue is that even the requirement $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_s \leq M\|x - y\|_2$ is *not* affine invariant, in that reparameterization of x and y lead to a different constant. Ideally, we would like to impose a condition that is fully affine invariant. Self-concordance satisfies this *desideratum*.

2 Self-concordant functions

With these two shortcomings in mind, we introduce the concept of *self-concordant functions*, which are a class of strictly convex functions (that is, with positive definite Hessians) that are smooth and have a well-behaved Hessian. The definition provides an affine-invariant condition focused on *bounding the approximation error of the second-order expansion of the function around each point*.

使每个点Taylor Expansion近似质量足够好

In other words, instead of focusing on the Lipschitz continuity of the Hessian, we focus on the *quality* of the approximation of the function by the second-order Taylor expansion around each point.

2.1 Intrinsic norms

Before we can give a meaningful quantification of the approximation error of the second-order expansion in our setting, special care must be taken to describe the radius around which the approximation is good in an affine-invariant way. Furthermore, one needs to use care close to the boundary of the domain. In the theory of self-concordant functions, both of these issues are elegantly resolved by moving away from the Euclidean norm and using instead the notion of *intrinsic norm* at a point x in the domain of f, defined as.

$$\boxed{\|v\|_x := \sqrt{\langle v, \nabla^2 f(x) v\rangle}}\,.$$

The name *intrinsic* norm comes from the following two important facts:

- the intrinsic norm is affine-invariant, in that if all points x are now replaced with $x = Ax'$, the vectors v with Av', and the function $g(x') := f(Ax')$ is introduced, then

$$\|v'\|_{x'} = \sqrt{\langle v', \nabla^2 g(x') v'\rangle} = \sqrt{\langle v', A^\top \nabla^2 f(x) A v'\rangle} = \sqrt{\langle Av', \nabla^2 f(x) Av'\rangle} = \|v\|_x.$$

- the intrinsic norm is insensitive to the choice of reference inner product $\langle \cdot, \cdot \rangle$ for the space. This is a consequence of the previous point, since all inner products are equivalent up to change of coordinates. 所有内积空间在坐标变化时都是等价的

2.2 Definition of self-concordance

We are now ready to define self-concordance. For this lecture, we will in particular focus on the notion of *strong nondegenerate* self-concordance, which is a stronger version of the definition, and is the definition that plays the central role in the theory of interior point methods.

Definition 2.1. Let $\Omega \subseteq \mathbb{R}^n$ be open and convex. A twice-differentiable function $f : \Omega \to \mathbb{R}$ is said to be *strongly nondegenerate self-concordant* if:

1. The Hessian of f is positive definite everywhere on Ω;
2. The ellipsoid $W(x) := \left\{ y \in \mathbb{R}^n : \|y - x\|_x^2 < 1 \right\}$ is contained in Ω for all $x \in \Omega$; and
3. Inside of the ellipsoid $W(x)$, the function f is *almost quadratic*:

$$\left(1 - \|y - x\|_x\right)^2 \nabla^2 f(x) \preccurlyeq \nabla^2 f(y) \preccurlyeq \frac{1}{\left(1 - \|y - x\|_x\right)^2} \nabla^2 f(x) \qquad \forall x \in \Omega, y \in W(x),$$

which is equivalent to the statement

$$\left(1 - \|y - x\|_x\right) \|v\|_x \leq \|v\|_y \leq \frac{\|v\|_x}{1 - \|y - x\|_x} \qquad \forall x \in \Omega, y \in W(x), v \in \mathbb{R}^n.$$

The following equivalent characterization is often used too.

Theorem 2.1. If $f : \Omega \to \mathbb{R}$ is three-time differentiable with positive definite Hessian everywhere on Ω, then strong nondegenerate self-concordance is equivalent to f satisfying the following two properties:

1. for any $x_0 \in \Omega$ and $d \in \mathbb{R}^n$, the restriction $\varphi(\gamma) := f(x_0 + \gamma d)$ of f to the segment $\{\gamma : x_0 + \gamma d \in \Omega\}$ satisfies

$$\varphi'''(\gamma) \leq 2\varphi''(\gamma)^{3/2}; \qquad \text{and}$$

2. any sequence $\{x_k\}$ converging to a point on the boundary of Ω is such that $f(x_k) \to +\infty$.

Remark 2.1. In this lecture, we will use the term *self-concordant* to mean *strongly nondegenerate self-concordant*. In different contexts, self-concordance without qualifications refers to only condition 1 in Theorem 2.1.

2.3 Notable example: the log function

As a sanity check, let's show that the function $f(x) = -\log(x)$ on the domain $\Omega := (0, \infty)$ is self-concordant. We will show the multidimensional version of this result.

Example 2.1. The function $f(x) = -\sum_{i=1}^n \log(x_i)$ on the domain $\Omega := \mathbb{R}^n_{>0}$ is self-concordant.

Solution. The function f is twice differentiable with positive definite Hessian

$$\nabla^2 f(x) = \operatorname{diag}\left(\frac{1}{x_1^2}, ..., \frac{1}{x_n^2}\right) \succ 0 \qquad \forall x \in \Omega = \mathbb{R}^n_{>0}.$$

The ellipsoid $W(x)$ is therefore equivalent to

$$W(x) = \left\{ y \in \mathbb{R}^n : \sum_{i=1}^n \left(\frac{y_i - x_i}{x_i}\right)^2 < 1 \right\} = \left\{ y \in \mathbb{R}^n : \sum_{i=1}^n \left(\frac{y_i}{x_i} - 1\right)^2 < 1 \right\}.$$

It is immediate to check that the condition implies that $y_i > 0$ for all $i = 1, ..., n$. So, $W(x) \subseteq \Omega$.

We now check the third condition. If $y \in W(x)$, then for any $v \in \mathbb{R}^n$ we have

$$\|v\|_y^2 = \sum_{i=1}^n \left(\frac{v_i}{y_i}\right)^2 = \sum_{i=1}^n \left(\frac{v_i}{x_i}\right)^2 \left(\frac{x_i}{y_i}\right)^2 \leq \|v\|_x^2 \max_i \left(\frac{x_i}{y_i}\right)^2.$$

Using the fact that

$$\frac{y_i}{x_i} \geq 1 - \left|\frac{y_i}{x_i} - 1\right| \geq 1 - \sqrt{\sum_{i=1}^n \left(\frac{y_i}{x_i} - 1\right)^2} = 1 - \|y - x\|_x \qquad \text{for all } i = 1, ..., n,$$

we find that

$$\max_i \left(\frac{x_i}{y_i}\right)^2 = \left(\max_i \left(\frac{x_i}{y_i}\right)\right)^2 \leq \frac{1}{\left(1 - \|y - x\|_x\right)^2} \qquad \Longrightarrow \qquad \|v\|_y^2 \leq \frac{\|v\|_x^2}{\left(1 - \|y - x\|_x\right)^2}.$$

On the other hand, we have

$$\|v\|_y^2 = \sum_{i=1}^n \left(\frac{v_i}{y_i}\right)^2 = \sum_{i=1}^n \left(\frac{v_i}{x_i}\right)^2 \left(\frac{x_i}{y_i}\right)^2 \geq \|v\|_x^2 \min_i \left(\frac{x_i}{y_i}\right)^2.$$

Using the fact that

$$\frac{y_i}{x_i} \leq 1 + \left|\frac{y_i}{x_i} - 1\right| \leq 1 + \sqrt{\sum_{i=1}^n \left(\frac{y_i}{x_i} - 1\right)^2} = 1 + \|y - x\|_x \qquad \text{for all } i = 1, ..., n,$$

we find that

$$\min_i \left(\frac{x_i}{y_i}\right)^2 = \left(\min_i \left(\frac{x_i}{y_i}\right)\right)^2 \geq \frac{1}{\left(1 + \|y - x\|_x\right)^2} \qquad \Longrightarrow \qquad \|v\|_y^2 \geq \frac{\|v\|_x^2}{\left(1 + \|y - x\|_x\right)^2}.$$

Finally, using the fact that $\frac{1}{1+z} \geq 1 - z$ (valid for all $z > -1$), the statement follows. □

2.4 Composition properties of self-concordant functions

In general, checking whether a function is self-concordant is a nontrivial task. Usually, one does not use the definition; rather, the following composition rules are used.

■ **Sum of self-concordant functions**. The set of self-concordant functions is closed under addition.

Theorem 2.2. Let $f_1 : \Omega_1 \to \mathbb{R}$ and $f_2 : \Omega_2 \to \mathbb{R}$ be self-concordant functions whose domains satisfy $\Omega_1 \cap \Omega_2 \neq \emptyset$. Then, the function $f + g : \Omega_1 \cap \Omega_2 \to \mathbb{R}$ is self-concordant.

[▷ You should try to prove this!]

■ **Addition of an affine function**. Addition of an affine function to a self-concordant functions does not affect the self-concordance property, since self-concordance depends only on the Hessian of the function, and the addition of affine functions does not affect the Hessian.

Theorem 2.3. Let $f : \Omega \to \mathbb{R}$ be self-concordant function. Then, the function $g(x) \coloneqq f(x) + \langle a, x\rangle + b$ is self-concordant on Ω.

■ **Affine transformation**. The composition of a self-concordant function with an injective affine transformation preserves the self-concordance.

Theorem 2.4. Let $f : \Omega \to \mathbb{R}$ be self-concordant and $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ represent an injective transformation.[1] Then, assuming $\Omega' \coloneqq \{x \in \mathbb{R}^m : Ax + b \in \Omega\} \neq \emptyset$, the affinely-transformed function $g(x) \coloneqq f(Ax + b)$ is self-concordant on the domain Ω'.

■ **Consequences**. Putting together the previous result together with Example 2.1, obtain the following important corollary.

Corollary 2.1. Let $\Omega \coloneqq \{x \in \mathbb{R}^n : a_i^\top x > b_i \text{ for all } i\}$ be a nonempty open polyhedral set containing no lines, where $a_i, b_i \in \mathbb{R}^n$. A function of the form

$$f(x) = c^\top x - \sum_{i=1}^{m} \log(a_i^\top x - b_i),$$

where $c \in \mathbb{R}^n$, is self-concordant on Ω.

2.5 Existence and uniqueness of the minimum

In addition, we mention the following property.

Theorem 2.5. Let $f : \Omega \to \mathbb{R}$ be self-concordant and lower bounded. Then, f attains a unique minimum.

The proof of Theorem 2.5 is typically derived from the convergence of Newton's method, which we will discuss in Section 3 (in particular, Theorem 3.1 and its corollary play an important role).

3 Newton's method applied to self-concordant functions

In this section, we discuss how the guarantees of Newton's method can be extended to self-concordant functions, while gaining affine invariance at the same time.

For notational convenience, in the rest of the section we use $n(x)$ to denote the Newton direction (*i.e.*, second-order descent direction) at a generic point $x \in \Omega$, which is defined as usual as

$$n(x) \coloneqq -\big[\nabla^2 f(x)\big]^{-1} \nabla f(x).$$

[1] Injectivity is necessary to preserve the positive definiteness of the Hessian.

Remark 3.1. The intrinsic norm of the descent direction $\|n(x)\|_x$ at any point x is a crucial quantity to study Newton's method for self-concordant functions. It is sometimes called the *Newton decrement*, and indicated as $\lambda(x) := \|n(x)\|_x$. We will avoid the notation $\lambda(x)$ to minimize the set of notation.

3.1 Proximity to the minimum

A neat property of self-concordant functions is that it is possible to bound the distance from the minimum of the function simply by looking at the (intrinsic) norm of the Newton direction at any point x. In particular, if the norm is sufficiently small, then the minimum must be near. Formally, we have the following result.

Theorem 3.1. Let $f : \Omega \to \mathbb{R}$ be self-concordant. If a point $x \in \Omega$ is such that $\|n(x)\|_x \leq 1/9$, then there exists a minimum z of f within distance

$$\|z - x\|_x \leq 3 \cdot \|n(x)\|_x.$$

In fact, the result above can be further strengthened to a larger intrinsic radius of $1/4$ (instead of $1/9$), as we remark next.

Remark 3.2. With a bit of additional work, the result in Theorem 3.1 can be strengthened as follows: If a point $x \in \Omega$ is such that $\|n(x)\|_x \leq 1/4$, then there exists a minimum z of f within distance

$$\|z - x\|_x \leq \|n(x)\|_x + \frac{3\|n(x)\|_x^2}{\left(1 - \|n(x)\|_x\right)^3}.$$

3.2 Recovering the quadratic convergence rate

For self-concordant functions, the following affine-invariant guarantee can be established.

Theorem 3.2. Let $f : \Omega \to \mathbb{R}$ be self-concordant. If a point $x_t \in \Omega$ is such that $\|n(x_t)\|_{x_t} < 1$, then

$$\|n(x_{t+1})\|_{x_{t+1}} \leq \left(\frac{\|n(x_t)\|_{x_t}}{1 - \|n(x_t)\|_{x_t}} \right)^2.$$

The proof is somewhat elaborate (it involves a few steps) but it is not particularly difficult. The key idea is to use the self-concordance property to bound the Hessian at the next iterate in terms of the Hessian at the current iterate. You can find a detailed proof in the references cited below.

Combined with the result in the previous subsection, this gives a quadratic convergence rate.

4 Further readings

The short book by Renegar, J. [Ren01] and the monograph by Nesterov, Y. [Nes18] (Chapter 5) provide a comprehensive introduction to self-concordant functions and their applications in optimization.

I especially recommend the book by Renegar, J. [Ren01] for a concise yet rigorous account.

[Ren01] J. Renegar, *A Mathematical View of Interior-point Methods in Convex Optimization*. Philadelphia, PA, USA: SIAM, 2001. doi: 10.1137/1.9780898718812.

[Nes18] Y. Nesterov, *Lectures on Convex Optimization*. Springer International Publishing, 2018. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-91578-4