Lecture 20

20.1 Local Convergence of Newton's Method 1
20.2 Bound for the Distance to the Solution 4

20.1 Local Convergence of Newton's Method

In this lecture, we establish the local quadratic convergence of the classic Newton's method. Using self-concordant analysis, it is possible to provide an *affine-invariant characterization* of the *region of quadratic convergence*, which is essential for the design of interior-point methods.

We consider differentiable strictly convex function $f : Q \to \mathbb{R}$ defined on an open convex set $Q \subseteq \mathbb{R}^n$. We assume that Q is a natural *domain* of f, so the function blows up when approaching the boundary, and that f is self-concordant with constant $M > 0$ on this set (see previous lecture).

20.1.1 Dual Local Norm

We use the local norm $\|\cdot\|_x$, induced by the Hessian, for the primal vectors, for any $x \in Q$:

$$\|h\|_x \;\; := \;\; \langle \nabla^2 f(x) h, h \rangle^{1/2}, \qquad h \in \mathbb{R}^n.$$

For the dual objects (gradients), we have to use the *dual norm* to it, which with abuse of notation we denote by the same symbol. Thus, for a linear form $\langle g, \cdot \rangle$, $g \in \mathbb{R}^n$, its local norm at $x \in Q$ is

$$\|\langle g, \cdot \rangle\|_x \;\; \equiv \;\; \|g\|_{x,*} \;\; \equiv \;\; \max_{h \in \mathbb{R}^n \,:\, \|h\|_x \leq 1} \langle g, h \rangle \;\; = \;\; \langle g, \nabla^2 f(x)^{-1} g \rangle \;\; = \;\; \|\nabla^2 f(x)^{-1/2} g\|_2,$$

where denotes the standard Euclidean norm for vectors, and the spectral norm for matrices.

The most important case for us is when $g := \nabla f(x)$. Denote,

$$\boxed{\lambda(x) \;\; := \;\; \|\langle \nabla f(x), \cdot \rangle\|_x \;\; = \;\; \langle \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \rangle^{1/2},}$$

which is sometimes called *the Newton decrement*.

20.1.2 Newton's Step

Consider one iteration of Newton's method:

$$x^+ \;\; = \;\; x - \nabla^2 f(x)^{-1} \nabla f(x) \qquad \Leftrightarrow \qquad \nabla f(x) + \nabla^2 f(x)(x^+ - x) \;\; = \;\; 0. \tag{20.1}$$

Then, the Newton decrement is equal to the length of Newton's step in local norm:

$$\|x^+ - x\|_x \;\; = \;\; \langle \nabla^2 f(x)(x^+ - x), x^+ - x \rangle^{1/2} \;\; = \;\; \langle \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \rangle^{1/2} \;\; = \;\; \lambda(x).$$

Therefore, when

$$\lambda(x) \;\; < \;\; \tfrac{2}{M}, \tag{20.2}$$

we conclude that $x^+ \in \mathcal{E}_x$, where

$$\mathcal{E}_x \;=\; \left\{ y \,:\, \|y-x\|_x < \tfrac{2}{M} \right\} \;\subseteq\; Q$$

is Dikin's ellipsoid (see Proposition 19.2.1 in the previous lecture). Hence, by preconditioning the gradient direction with the inverted Hessian (20.1), we remain in the domain without any auxiliary projections, under condition (20.2).

In the last lecture we have proved the following lemma, that is a key to analyze local behavior of the Newton's method.

Lemma 20.1.1. *Let $x \in Q$ and assume that $y \in \mathcal{E}_x$. Then, the Hessians are comparable:*

$$\left(1 - \tfrac{M}{2}\|y-x\|_x\right)^2 \nabla^2 f(x) \;\preceq\; \nabla^2 f(y) \;\preceq\; \left(1 - \tfrac{M}{2}\|y-x\|_x\right)^{-2} \nabla f(x). \tag{20.3}$$

We are ready to establish the main result about Newton's method.

Theorem 20.1.2. *Let*

$$\lambda(x) \;\leq\; \tfrac{1}{M}. \tag{20.4}$$

Then,

$$\lambda(x^+) \;\leq\; M\lambda(x)^2 \tag{20.5}$$

Proof. Denote $r := \lambda(x) \overset{(20.4)}{\leq} \frac{1}{M} < \frac{2}{M}$. Therefore, the condition of Lemma 20.1.1 is satisfied.

First, we can move from the new norm, to the old norm, using the lemma:

$$\lambda(x^+) \;=\; \|\nabla f(x^+)\|_{x^+} \;\overset{(20.3)}{\leq}\; \tfrac{1}{1-\frac{M}{2}r}\|\nabla f(x^+)\|_x.$$

Then, using the definition of Newton's step and the main theorem of calculus, we get:

$$\begin{aligned} \nabla f(x^+) &\overset{(20.1)}{=} \nabla f(x^+) - \nabla f(x) - \nabla^2 f(x)(x^+ - x) \\ &= (G - H)(x^+ - x), \end{aligned}$$

where $G = \int\limits_0^1 \nabla^2 f(x + \tau(x^+ - x))d\tau$ and $H = \nabla^2 f(x)$. Hence, we obtain:

$$\begin{aligned} \|\nabla f(x^+)\|_x &= \|(G-H)(x^+ - x)\|_x \\ &= \|H^{-1/2}(G-H)H^{-1/2}H^{1/2}(x^+ - x)\|_2 \\ &\leq \|H^{-1/2}(G-H)H^{-1/2}\|_2 \cdot \|H^{1/2}(x^+ - x)\|_2 \\ &= \|H^{-1/2}(G-H)H^{-1/2}\|_2 \cdot r \end{aligned}$$

We have the following lower bounds:

$$G \;=\; \int\limits_0^1 \nabla^2 f(x + \tau(x^+ - x))d\tau \;\overset{(20.3)}{\succeq}\; H \cdot \int\limits_0^1 (1 - t\tfrac{M}{2}r)^2 dt \;=\; \left(1 - \tfrac{Mr}{2} + \tfrac{1}{12}M^2r^2\right)H,$$

and

$$H^{-1/2}(G-H)H^{-1/2} \;\succeq\; \tfrac{Mr}{2}\left(\tfrac{Mr}{6} - 1\right)I.$$

The corresponding upper bound is:

$$G \overset{(20.3)}{\preceq} H \int_0^1 \frac{dt}{(1-tMr/2)^2} = \frac{1}{1-Mr/2} H,$$

and

$$H^{-1/2}(G-H)H^{-1/2} \preceq \Bigl(\frac{1}{1-Mr/2} - 1\Bigr) I = \frac{Mr/2}{1-Mr/2}.$$

Thus,

$$\|H^{-1/2}(G-H)H^{-1/2}\|_2 \leq \frac{Mr}{2} \max\Bigl\{ \frac{1}{1-Mr/2}, 1 - \frac{Mr}{6} \Bigr\} \leq \frac{Mr/2}{1-Mr/2}.$$

Combining all ingredients together, we get:

$$\lambda(x^+) \leq \frac{M}{2-Mr} r^2 = \frac{M}{2-M\lambda(x)} \lambda(x)^2 \overset{(20.4)}{\leq} M\lambda(x)^2,$$

which completes the proof. □

20.1.3 Discussion

We observe that inequality (20.5) leads to a very quick progress of the method.

Corollary 20.1.3. *Denote*

$$\delta_k = M\lambda(x_k).$$

We have

$$\delta_{k+1} \leq \delta_k^2.$$

Hence, after $k \geq 0$ iterations, starting from $\delta_0 = M\lambda(x_k) \leq \frac{1}{2}$ we obtain

$$\delta_k \leq \delta_0^{2^k} \leq \bigl(\tfrac{1}{2}\bigr)^{2^k}. \tag{20.6}$$

The convergence rate (20.6) is called *quadratic* convergence. It is very fast: with each iteration, the number of the correct digits in the solution doubles! To obtain a point x_k satisfying:

$$\lambda(x_k) \leq \varepsilon,$$

for a given $\varepsilon > 0$, it follows from (20.6) that it is sufficient to perform just

$$k = 1 + \Bigl\lceil \log_2 \log_2 \tfrac{1}{M\varepsilon} \Bigr\rceil$$

Newton step, provided that $x_0 \in \mathcal{Q}$, where

$$\mathcal{Q} := \Bigl\{ x \,:\, \lambda(x) = \langle \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \rangle^{1/2} \leq \tfrac{1}{2M} \Bigr\} \subseteq Q$$

is the *region of quadratic convergence.*

Note that set $\mathcal{Q}$ is affine-invariant as it does not depend on the choice of the coordinate system in our space. At the same time, if we fix any particular norm $\|\cdot\|$, and assume that function f is strongly convex with parameter $\mu > 0$ with respect to this norm:

$$\langle \nabla^2 f(x)h, h \rangle \geq \mu \|h\|^2, \qquad \forall h \in \mathbb{R}^n,\ x \in Q. \tag{20.7}$$

Then,

$$\lambda(x)^2 \;=\; \langle \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \rangle \;\leq\; \|\nabla f(x)\|_* \|\nabla^2 f(x)^{-1} \nabla f(x)\| \overset{(20.7)}{\leq} \tfrac{\lambda(x)}{\mu^{1/2}} \|\nabla f(x)\|_*,$$

and we obtain an upper bound on the Newton decrement:

$$\lambda(x) \;\leq\; \tfrac{1}{\mu^{1/2}} \|\nabla f(x)\|_*. \tag{20.8}$$

Assuming additionally that the Hessian is Lipschitz, for some constant $L > 0$:

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \;\leq\; L\|y - x\|, \qquad \forall x, y \in Q,$$

we can set the constant of self-concordance for f as $M = \frac{L}{\mu^{3/2}}$ (see Lecture 18).

Corollary 20.1.4. *For any norm $\|\cdot\|$, consider the region:*

$$\mathcal{G} \;=\; \mathcal{G}_{\|\cdot\|} \;=\; \left\{ x \;:\; \|\nabla f(x)\|_* \;\leq\; \tfrac{\mu^2}{2L} \right\}. \tag{20.9}$$

Then, all points from $\mathcal{G}$ are in the region of quadratic convergence of Newton's method:

$$\mathcal{G} \;\subseteq\; \mathcal{Q}.$$

Proof. Indeed, for any $x \in \mathcal{G}$:

$$\lambda(x) \overset{(20.8)}{\leq} \tfrac{1}{\mu^{1/2}} \|\nabla f(x)\|_* \overset{(20.9)}{\leq} \tfrac{\mu^{3/2}}{2L} \;=\; \tfrac{1}{2M},$$

thus $x \in \mathcal{Q}$. □

When using a fixed norm, it is much easier to prove directly that (20.9) is the region of quadratic convergence for Newton's method. However, we obtain it as a simple direct consequence of Theorem 20.1.2. Affine-invariant characterization of $\mathcal{Q}$ is crucial for analyzing the path-following scheme.

20.2 Bound for the Distance to the Solution

We have proved local quadratic convergence in terms of the quantity $\lambda(x)$. But what about other possible accuracy measures? We can think of the functional residual $f(x) - f^\star$, or distance to the solution, e.g. $\|x - x^\star\|_x$. It appears that, locally, *all these measures are equivalent* (see Theorem 5.2.1 in [Nes18] for the exact bounds):

$$f(x) - f^\star \;\approx\; \|x - x^\star\|_x \;\approx\; \lambda(x). \tag{20.10}$$

So since we can make $\lambda(x)$ extremely small, we can also make any of these measures as small as we want.

Let us prove one inequality quantifying (20.10), which will be important for our further analysis of the path-following scheme.

Proposition 20.2.1. *Let for some point $x \in Q$ we have $\lambda(x) \leq \frac{1}{2M}$. Then, the minimizer $x^\star$ of f exists, and it holds:*

$$\|x - x^\star\|_x \;\leq\; 3\lambda(x). \tag{20.11}$$

Proof. Consider the closed ellipsoid $B := \{y \ : \ \|y - x\|_x \ \leq \ 3\lambda(x)\} \subset \mathcal{E}_x$. Our goal is to show $x^\star \in B$. We can assume $\lambda(x) \neq 0$, since otherwise $x = x^\star$ and the statement holds.

First, for any $y \in \mathcal{E}_x$, using the main theorem of calculus, we have

$$\begin{aligned} \langle \nabla f(y) - \nabla f(x), y - x \rangle &= \int\limits_0^1 \langle \nabla^2 f(x + \tau(y - x))(y - x), y - x \rangle d\tau \\ &\overset{(20.3)}{\geq} \|y - x\|_x^2 \cdot \int\limits_0^1 (1 - \tau \tfrac{M}{2} \|y - x\|_x)^2 d\tau \\ &\geq \|y - x\|_x^2 \cdot \int\limits_0^1 (1 - \tau)^2 d\tau \ = \ \tfrac{1}{3}\|y - x\|_x^2. \end{aligned} \tag{20.12}$$

Consider points from the boundary of the ellipsoid, $y \in S := \partial B$, thus $\|y - x\|_x = 3\lambda(x)$. Then, for such points:

$$\begin{aligned} \langle \nabla f(y), y - x \rangle &\overset{(20.12)}{\geq} \langle \nabla f(x), y - x \rangle + \tfrac{1}{3}\|y - x\|_x^2 \\ &\geq -\lambda(x)\|y - x\|_x + \tfrac{1}{3}\|y - x\|_x^2 \ = \ 0. \end{aligned} \tag{20.13}$$

Then, for any $z \in Q \setminus B$, there exists $\alpha \in (0, 1)$ such that $y = \alpha x + (1 - \alpha)z \in S$. By convexity, we obtain

$$f(z) \ \geq \ f(y) + \langle \nabla f(y), z - y \rangle \ = \ f(y) + \tfrac{\alpha}{1-\alpha} \langle \nabla f(y), y - x \rangle \overset{(20.13)}{\geq} f(y).$$

Therefore, the minimum of f over the compact set B is its global minimum, which proves the required statement. □

Literature

[Nes18] Yurii Nesterov. *Lectures on convex optimization.* Springer, 2018.