Lecture 6

W. Hebisch

November 23, 2021

1 Newton method

1.1 Local convergence

To analyze local convergence, we start with Newton method for equation solving. Let $g : \mathbb{R}^n \to \mathbb{R}^n$. Given x_i we define

$$x_{i+1} = x_i - g'(x_i)^{-1} g(x_i)$$

Assume that x_∞ is a solution, that is $g(x_\infty) = 0$.

Let $B_r = \{x : \|x - x_\infty\| < r\}$ be ball of radius r around x_∞.

Lemma 1.1 *Assume that g' is Lipschitz continuous with constant L on B_r and $\|g'^{-1}(x)\| \le m$. If $x_i \in B_r$, then*

$$\|x_{i+1} - x_\infty\| \le \frac{Lm}{2} \|x_i - x_\infty\|^2$$

If additionally $rLm \le 2$, then $x_{i+1} \in B_r$. Also, when $rLm \le 2$, it is enough to assume that $\|g'^{-1}(x)\| \le m$ for $x \in B_r$.

Proof: Put $\delta_i = x_i - x_\infty$. We compute

$$0 = g(x_\infty) = g(x_i) - \int_0^1 g'(x_i - t\delta_i)\delta_i dt$$

$$= g(x_i) - g'(x_i)\delta_i - \int_0^1 (g'(x_i - t\delta_i) - g'(x_i))\delta_i dt$$

so

$$\|g(x_i) - g'(x_i)\delta_i\| \le \|\delta_i\| \int_0^1 \|g'(x_i - t\delta_i) - g'(x_i)\| dt$$

$$\le \|\delta_i\| \int_0^1 L\|\delta_i\| t dt = \frac{L}{2}\|\delta_i\|^2$$

Now

$$x_{i+1} - x_\infty = x_i - g'(x_i)^{-1}g(x_i) - x_\infty = g'(x_i)^{-1}(g'(x_i)(x_i - x_\infty) - g(x_i))$$

$$= g'(x_i)^{-1}(g'(x_i)\delta_i - g(x_i))$$

so

$$\|x_{i+1} - x_\infty\| \leq \|g'(x_i)^{-1}\|\|g'(x_i)\delta_i - g(x_i)\|$$
$$\leq m\frac{L}{2}\|\delta_i\|^2 = \frac{mL}{2}\|x_i - x_\infty\|^2$$

which is the required estimate.

Since $x_i \in B_r$ we have $\|x_i - x_\infty\| < r$ and if additional assumption is satisfied we have

$$1 \geq \frac{rmL}{2} > \frac{mL\|x_i - x_\infty\|}{2}$$

so

$$\|x_{i+1} - x_\infty\| \leq \frac{mL\|x_i - x_\infty\|}{2}\|x_i - x_\infty\| < \|x_i - x_\infty\|$$

and $\|x_{i+1} - x_\infty\| < \|x_i - x_\infty\| < r$, hence $x_{i+1} \in B_r$.

Finally, we used estimate on $\|g'^{-1}\|$ only for x in line segment joining x_i and x_∞. When $rmL < 2$ this line segment is inside B_r. □

Assuming that g is regular enough and $g'(x_\infty)$ is invertible from the lemma we see that once Newton method is sufficiently close to solution it will converge, moreover, each step approximately doubles accuracy (number of significant bits). So, once we are close to solution 5 or 6 steps usually is enough to get full machine accuracy.

Under reasonable global assumptions it is possible to find out if we are close enough to solution:

Lemma 1.2 *Assume that g' is Lipschitz continuous with constant L and put $m = \|g'(x_0)^{-1}\|$. If*

$$\|g(x_0)\| \leq \frac{3}{16Lm^2},$$

then Newton method starting at x_0 converges to x_∞ and

$$\|x_0 - x_\infty\| \leq \frac{1}{4Lm}.$$

The result remains valid if we only assume that g' is Lipschitz continuous with constant L on ball centered at x_0 and radius $\frac{1}{2Lm}$.

Idea of the proof: let $A = g'(x_0)^{-1}$ and $\phi(x) = Ag(x)$. $\phi'(x) = Ag'(x)$, so $\phi'(x_0) = I$ (identity matrix) and ϕ' is Lipschitz continuous with constant Lm. As long as $\|x - x_0\| \leq \frac{1}{2Lm}$ we have $\|\phi'(x) - I\| \leq \frac{1}{2}$, so $\phi'(x)$ is invertible and $\|\phi'(x)\| \leq 2$. If $\|x_0 - x_\infty\| \leq \frac{1}{4Lm}$, then by previous lemma Newton method applied to ϕ is convergent and

$$\|x_0 - x_\infty\| \leq \|x_0 - x_1\| + \|x_1 - x_\infty\| \leq \|\phi(x_0)\| + Lm\|x_0 - x_\infty\|^2$$

$$\leq m\|g(x_0)\| + \frac{1}{4}\|x_0 - x_\infty\|$$

so

$$\frac{3}{4}\|x_0 - x_\infty\| \leq m\|g(x_0)\|$$

and

$$\|x_0 - x_\infty\| \leq \frac{4}{3}m\|g(x_0)\| \leq \frac{1}{4Lm}$$

To drop assumption $\|x_0 - x_\infty\| \leq \frac{1}{4Lm}$ consider equation $\phi(x) - t\phi(x_0) = 0$. For $t = 1$ this has solution in $K = \{x : \|x - x_0\| < \frac{1}{4Lm}\}$, namely x_0. By inverse function theorem set of t such that we have solution in K is open. By compactness set of t such that we have solution in $\bar{K}$ (closure of K) is closed. By previous estimate, for $t > 0$ solution must be in K, so set of t such that we have solution is K is nonempty open and closed subset of $(0, 1]$, so since interval is connected it is whole $(0, 1]$. By compactness for $t = 0$ we get solution such that $\|x - x_0\| \leq \frac{1}{4Lm}$.

We could restate the results for optimization, but we skip details, just plug in ∇f in place of g. Just note that now we assume Lipschitz continuity of second derivative.

1.2 Global convergence

Local convergence means that close to solution $\alpha_i = 1$ is good choice of step size. In fact, assuming regularity there is little motivation to use steps bigger than 1, but to get global convergence we sometimes need step smaller than 1. More precisely, when sufficient decay condition is violated we decrease step size.

In fact, local convergence implies that as long as we get decay we also get global convergence of gradient to zero.

However, there are two difficulties. First, in general (when $\nabla^2 f$ is not positive definite) Newton direction may fail to be decay direction. Second, local convergence assumes that $\nabla^2 f$ is invertible at stationary points. We can avoid both difficulties adding to $\nabla^2 f$ multiple of identity to make it positive definite. Unfortunately, it is hard to give some warranty for such method. Instead, we will assume strong convexity.

Assume that $mI \leq \nabla^2 f(x) \leq MI$ and $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L$. By strong convexity there is optimal point. Since $\nabla^2 f(x)$ is positive definite Newton direction is descent direction and we have global convergence.

For local convergence we have the following results:

Lemma 1.3 *If $\|\nabla f(x_i)\| < \frac{2m^2}{L}$, then pure Newton method starting at x_i is convergent and*

$$\frac{L}{2m^2}\|\nabla f(x_{i+1})\| \leq \left(\frac{L}{2m^2}\|\nabla f(x_i)\|\right)^2$$

We write $h_i = -(\nabla^2 f(x_i))^{-1}\nabla f(x_i)$ so $\nabla^2 f(x_i)h_i + \nabla f(x_i) = 0$ and

$$\nabla f(x_{i+1}) = \nabla f(x_i + h_i) - \nabla f(x_i) - \nabla^2 f(x_i)h_i$$

$$= \int_0^1 (\nabla^2 f(x_i + th_i) - \nabla^2 f(x_i))h_i dt$$

so

$$\|\nabla f(x_{i+1})\| \leq \int_0^1 L\|th_i\|\|h_i\|dt = \frac{L\|h_i\|^2}{2}$$

Since $\|\nabla^2 f(x_i)^{-1}\| \leq \frac{1}{m}$ we have $\|h_i\| \leq \frac{1}{m}\|\nabla f(x_i)\|$ so

$$\|\nabla f(x_{i+1})\| \leq \frac{L\|\nabla f(x_i)\|^2}{2m^2}$$

which gives bound on $\|\nabla f(x_{i+1})\|$. Under assumption this decreases which proves convergence.

Recall Armijo's rule:

$$f(x_i + \alpha h_i) - f(x_i) \leq \rho\alpha\langle\nabla f(x_i), h_i\rangle.$$

Lemma 1.4 *If $\|\nabla f(x_i)\| \leq \frac{3(1-2\rho)m^2}{L}$, then $\alpha = 1$ is acceptable by Armijo's rule.*

Let $\gamma = \frac{\rho m}{\eta M^2}$

Lemma 1.5 *Assume $\rho \leq \frac{1}{2}$. If step size α in Newton method is selected starting from $\alpha = 1$ and dividing α by η as long as Armijo's condition is violated, then*

$$f(x_{i+1}) - f(x_i) \leq -\gamma\|\nabla f(x_i)\|^2$$

Proof: Put $\lambda = -\langle\nabla f(x_i), h_i\rangle = \langle\nabla^2 f(x)h_i, h_i\rangle$. By strong convexity $\langle h_i, h_i\rangle \leq \frac{1}{m}\lambda$. Using $\nabla^2 f(x) \leq MI$ we have

$$f(x_i + \alpha h_i) \leq f(x_i) + \alpha\langle\nabla f(x_i), h_i\rangle + \frac{M}{2}\alpha^2\|h_i\|^2$$

$$\leq f(x_i) - \alpha\lambda + \frac{M}{2m}\alpha^2\lambda.$$

Now we see that $\alpha = \frac{m}{M}$ satisfies Armijo's rule

$$f(x_i + \alpha h_i) \leq f(x_i) - \alpha\lambda + \frac{1}{2}\alpha\lambda$$

$$= f(x_i) - \frac{1}{2}\alpha\lambda$$

since $\rho \leq \frac{1}{2}$. Therefore we choose at least $\alpha = \frac{m}{\eta M}$ leading to decay

$$f(x_{i+1}) - f(x_i) \leq -\rho\alpha\lambda$$

$$\leq -\frac{\rho m}{\eta M}\lambda$$

Since $\lambda \leq \frac{1}{M}\|\nabla f(x_i)\|^2$ this gives

$$f(x_{i+1}) - f(x_i) \leq -\frac{\rho m}{\eta M^2}\|\nabla f(x_i)\|^2 = -\gamma\|\nabla f(x_i)\|^2.$$

The lemmas together imply global convergence of Newton method with $\rho < 1/2$: as long as $\|\nabla f(x_i)\|$ is big (so that local convergence does not apply) we get steady decay of value of f, in fact, putting $t = \min(3(1-2\rho), 1)\frac{m^2}{L}$ in at most

$$\frac{f(x_0) - f(x_\infty)}{\gamma t^2}$$

steps $\|\nabla f(x_i)\| \geq t$. But once $\|\nabla f(x_i)\| < t$ local convergence holds and we get any fixed accuracy in a fixed number of steps.

1.3 Remarks

Let us state some features, assuming classical algorithms

- need $O(n^2)$ operations and memory to compute and store $\nabla^2 f$
- need $O(n^3)$ operations to compute $(\nabla^2 f(x))^{-1}\nabla f(x)$
- convergence independent of choice of variables
- very fast local convergence

Compare gradient descent

- $O(n)$ operations and storage per step
- very sensitive to bad conditioning

This analysis is not entirely satisfactory. First, far from optimum we only get decay of objective by constant amount. Gradient descent divides objective by a constant which theoretically may by much better. Second, we still have $\frac{m}{M}$ and our estimate predicts very slow convergence when this is small. To put this differently, Newton method is invariant under affine change of coordinates, but our analysis depends on coordinates. When coordinates are badly adapted to the problem, then we will get very pessimistic conclusions.

1.4 Self-concordant functions

We can get better estimates for special classes of functions. We say that convex $f : \mathbb{R} \to \mathbb{R}$ is self-concordant when

$$|f'''(x)| \leq 2f''(x)^{3/2}$$

for all x in domain of f. We say that multivariate f is self-concordant when restriction of f to any line is self-concordant.

Note: condition above is invariant under translations and dilations which implies that for functions of single variable self-concordance is affine invariant. But then by definition self-concordance is affine invariant also for multivariate functions

Examples:

- linear function
- positive definite quadratic function
- logarithm

$\exp(x)$ on $\mathbb{R}$ is not self-concordant.

Let us do calculations for $f(x) = -\log(x)$. We have

$$f'(x) = -\frac{1}{x},$$

$$f''(x) = \frac{1}{x^2},$$

$$f'''(x) = -2\frac{1}{x^3}$$

so

$$|f'''(x)| = 2\frac{1}{x^3} = 2(\frac{1}{x^2})^{3/2} = 2f''(x)^{3/2}$$

so indeed $-\log(x)$ is self-concordant.

Note that we have factor 2 on the right hand side is the definition of self-concordant function because we want $-\log(x)$ to be self-concordant. Namely, $-\log(x)$ fails simpler condition

$$|f'''(x)| \leq f''(x)^{3/2}$$

It is easy to check that $-4\log(x)$ satisfies condition above and more generally, if $f(x)$ is self-concordant, then $4f(x)$ satisfies condition above so in principle we could use condition above and multiply all functions by 4. But having 2 in the definition is more natural and leads to simpler theory.

Important property: when f_i are self-concordant and $c_i \geq 1$ then

$$\sum c_i f_i$$

is self-concordant. Namely, it is obvious that $c_i f_i$ are self-concordant. We calculate

$$|(f_1 + f_2)'''(x)| \leq |f_1'''(x)| + |f_2'''(x)| \leq 2(f_1''(x)^{3/2} + f_2''(x)^{3/2})$$
$$\leq 2(f_1''(x) + f_2''(x))^{3/2}$$

where in the last step we used subadditivity of $L^{3/2}$ norm.

Consequently sum of self-concordant functions is self-concordant.

Remark: In general convex combination of self-concordant functions is not self-concordant. Namely, let $f_i(x) = -\log(x_i)$ for $i = 1, 2$. Clearly, by our computation for $-\log(x)$ each f_i is self-concordant. But

$$g(x) = \frac{1}{2}(f_1(x) + f_2(x)) = \frac{-\log(x_1) - \log(x_2)}{2}$$

is not self-concordant. To see this we restrict g to line $x_2 = 1$, that is consider $h(t) = g((t, 1)) = -\log(t)/2$. By our computation for $-\log(x)$ we see that $h(t)$ is not self-concordant so also g is not self-concordant.

From the last property we see that

$$-\log(1 - x^2) = -\log((1 - x)(1 + x)) = -\log(1 - x) - \log(1 + x)$$

is self concordant as sum of self concordant functions. Similarly, minus logarithm of any concave quadratic on real line is self concordant. Consequently also in multidimensional case minus logarithm of any concave quadratic is self concordant. In particular this applies to $-\log(1 - \|x\|^2)$.

More complicated example: $-\log(\det(A))$ is self-concordant on set of strictly positive definite matrices. Namely, a line going trough this set can be written as $A + tB$ where A is strictly positive definite and B is symmetric. Positive definite matrix has square root so we can write

$$A + tB = A^{1/2}(I + tA^{-1/2}BA^{-1/2})A^{1/2} = A^{1/2}(I + tC)A^{1/2}$$

where $C = A^{-1/2}BA^{-1/2}$ is symmetric. C has real eigenvalues $\lambda_1, \ldots, \lambda_m$ and we have

$$\det(I + tC) = \prod_{i=1}^{m}(1 + t\lambda_i)$$

so

$$-\log(\det(A + tB)) = -\log(\det(A)) - \sum_{i=1}^{m}\log(1 + t\lambda_i).$$

Since each of $\log(1 + t\lambda_i)$ is self-concordant as a function of t the whole sum is self-concordant, so $-\log(\det(A + tB))$ is self-concordant as function of t, so $-\log(\det(A))$ is self-concordant as function of A.

Alternative multivariate definition: multivariate f is self-concordant if and only if for each x and h we have

$$|f'''(x)(h, h, h)| \le 2(f''(x)(h, h))^{3/2}$$

This is clear by looking at f on lines of form $x + th$.

We have

$$|f'''(x)(h_1, h_2, h_3)| \le 2\left(f''(x)(h_1, h_1)f''(x)(h_2, h_2)f''(x)(h_3, h_3)\right)^{1/3}$$

Remark: Existence of derivatives leads to somewhat tricky theoretical problems. In practice we work with rather regular functions, so we assume existence of third derivative and then prove bounds.

1.5 Estimate for symmetric functions

The last inequality follows from general property: when A is a real k-linear symmetric form then

$$\sup_{\|x_i\| \le 1} |A(x_1, x_2, \ldots, x_k)| \le \sup_{\|x\| \le 1} |A(x, x, \ldots, x)|$$

This in turn follows from properties of bilinear forms: if $\|h_1\| = \|h_2\| = 1$ and

$$|A(h_1, h_2)| = \sup_{\|x_1\| \le 1, \|x_2\| \le 1} |A(x_1, x_2)|$$

then

$$|A(h_1, h_2)| = |A(h_1, h_1)|$$

Note that it is enough to prove the last claim for two dimensional space (subspace spanned by h_1, h_2).

Symmetric form can then be written as

$$A(h_1, h_2) = \langle Bh_1, h_2 \rangle$$

where B is real symmetric matrix. B has two real eigenvalues λ_1, λ_2. When $\lambda_1 = \lambda_2 = \lambda$, then $|A(h_1, h_2)| = |\lambda||\langle h_1, h_2 \rangle|$ and the claim follows from properties of scalar product. When $\lambda_1 \neq \lambda_2$, then h_1 and h_2 maximizing $A(h_1, h_2)$ must be multiple of a single eigenvector and again claim follows. Having claim for bilinear forms by induction we prove that

$$\sup_{\|x_i\| \le 1} |A(x_1, x_2, \dots, x_k)|$$

is attained when all x_i are equal, which gives claim for multilinear forms.

Note: the estimate above is specific to real scalars and euclidean norm $\| \cdot \|$. Similar results hold for arbitrary norm and complex scalars, but at the cost of adding on right hand side a constant bigger than 1.

1.6 Back to self-concordant functions

Recall that we argued that scalar product $\langle h_1, h_2 \rangle_x = \langle \nabla^2 f(x) h_1, h_2 \rangle$ dependent on x is probably better adapted to f, than usual scalar product. For self-concordant functions we can show this in precise way. To avoid trivial difficulties we will assume that values of self-concordant function f go to infinity when arguments go to boundary of the domain. This ensures that self-concordant function is defined on maximal possible domain. Let $W_x = \{y : \|y - x\|_x < 1\}$.

1.6.1 Main estimate

Lemma 1.6 *Let f be as above. f is defined on W_x and for $\|h\|_x < 1$ we have*

$$f(x) + \langle \nabla f(x), h \rangle + \phi(-\|h\|_x) \le f(x + h) \le f(x) + \langle \nabla f(x), h \rangle + \phi(\|h\|_x)$$

where $\phi(s) = -log(1 - s) - s = \sum_{i=2}^{\infty} \frac{s^i}{i}$. Moreover,

$$(1 + \|h\|_x)^{-2} \nabla^2 f(x) \le \nabla^2 f(x + h) \le (1 - \|h\|_x)^{-2} \nabla^2 f(x).$$

Lower bounds remain valid as long as $x + h$ is in domain of f.

Proof: Let $u = \frac{h}{\|h\|_x}$. Put

$$\psi(s) = \inf\{t : f''(x+su) \leq tf''(x)\}$$

Note: In single variable we could use $f''(x+su)/f''(x)$, but above f'' is a quadratic form, so we need more complicated condition above.

For one variable real function g put

$$(\delta_+ g)(s) = \limsup_{r\to 0_+} \frac{g(s+r)-g(s)}{r}$$

Similarly define δ_- with lim sup replaced by lim inf. By self-concordance of f we have

$$\delta_+\psi(s) \leq 2\psi(s)^{3/2}.$$

Namely, let $A(s) = f''(x+su)$. By self-concordance of f we have

$$|A'(s)(v,v)| = |f'''(x+su)(v,v,u)|$$

$$\leq 2f''(x+su)(v,v)(f''(x+su)(u,u))^{1/2}$$

By definition of ψ we have

$$f''(x+su)(v,v) \leq \psi(s)f''(x)(v,v)$$

so

$$|A'(s)(v,v)| \leq 2\psi(s)^{3/2}f''(x)(v,v)(f''(x)(u,u))^{1/2}$$

But

$$f''(x)(u,u)^{1/2} = \|u\|_x = 1$$

so

$$|A'(s)(v,v)| \leq 2\psi(s)^{3/2}f''(x)(v,v) = 2\psi(s)^{3/2}\|v\|_x^2.$$

Now $A(s+t) = A(s) + tA'(s) + o(t)$ so for $t > 0$ we have

$$A(s+t)(v,v) \leq A(s)(v,v) + tA'(s)(v,v) + o(t)\|v\|_x^2$$

$$\leq \psi(s)\|v\|_x^2 + 2t\psi(s)^{3/2}\|v\|_x^2 + o(t)\|v\|_x^2$$

Since $\|v\|_x^2 = f''(x)(v,v)$ this means

$$A(s+t)(v,v) \leq (\psi(s) + 2t\psi(s)^{3/2} + o(t))f''(x)(v,v)$$

that is

$$A(s+t) \leq (\psi(s) + 2t\psi(s)^{3/2} + o(t))f''(x).$$

Consequently

$$\psi(s+t) \leq (\psi(s) + 2t\psi(s)^{3/2} + o(t))$$

which gives inequality

$$\delta_+\psi(s) = \limsup_{t\to 0_+} \frac{\psi(s+t)-\psi(s)}{t} \leq 2\psi(s)^{3/2}.$$

Similarly we get inequality for $\delta_-\psi(s)$ so

$$-2\psi(s)^{3/2} \leq \delta_-\psi(s) \leq \delta_+\psi(s) \leq 2\psi(s)^{3/2}$$

and

$$\delta_-\psi(s)^{-1/2} \geq -\frac{\delta_+\psi(s)}{2\psi(s)^{3/2}} \geq -1.$$

Since $\psi(0) = 1$ this implies

$$\psi(s)^{-1/2} \geq 1 - s.$$

Hence

$$\psi(s) \leq (1-s)^{-2}$$

so

$$f''(x+h) = f''(x + \|h\|_x u) \leq (1 - \|h\|_x)^{-2} f''(x)$$

which gives upper estimate on $f''(x+h)$, when $x+h$ is in domain of f.

In similar way we prove lower bound on $f''(x+h)$. Integrating twice upper and lower bounds for $f''(x+su)$ we get bounds for f. Since we assume that f goes to infinity at boundary of the domain upper bound implies that for $\|h\|_x < 1$ we have $x+h$ in the domain. This ends the proof when $\|h\|_x > 0$. When $\|h\|_x = 0$ we choose h_n so that $\|h_n\|_x > 0$ and $h = \lim h_n$ and get estimate as a limit.

Remark: ϕ above is self-concordant so bounds are sharp.

1.6.2 Self-concordant functions, nondegeneracy

In general it may happen that for some nonzero h we have $\|h\|_x = 0$. Under assumption of our main estimate it follows that for all y in the domain of f we have $\|h\|_y = 0$. In other words, space $F = \{h : \|h\|_x = 0\}$ is independent of x. Moreover, f is sum of linear function and function that is invariant under translations by vectors from F.

We say that f is nondegenerate if the space $F = \{0\}$. Under assumption of main estimate this is always the case when domain of f does not contain any line. In the sequel we assume that f is nondegenerate.

1.6.3 Newton method, self-concordant functions

Our main estimate means that nondegenerate f is well conditioned on compact subsets of W_x. This implies strong results about convergence of Newton method for self-concordant functions. In particular this implies uniform speed of convergence of Newton method (bad conditioning is not a problem).

2 Further reading

Stephen Boyd, Lieven Vandenberghe, Convex Optimization, chapter 9.

David G. Luenberger, Yinyu Ye, Linear and Nonlinear Programming, chapter 8.

A. Nemirovski, INTERIOR POINT POLYNOMIAL TIME METHODS IN CONVEX PROGRAMMING, lecture notes, chapter 2.

Yurii Nesterov, Introductory lectures on convex optimization, Springer 2004, chapter 1, chapter 4.1.

Jorge Nocedal, Stephen J. Wright, Numerical Optimization, chapter 3.