Lecture 7

W. Hebisch

April 5, 2023

1 Newton method

1.1 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.

1.2 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

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

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
- minus 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 in 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. More precisely, for $a, b > 0$ we have:

$$a^{3/2} + b^{3/2} = \|(a,b)\|_{3/2}^{3/2} \leq (\|(a,0)\|_{3/2} + \|(0,b)\|_{3/2})^{3/2} = (a+b)^{3/2}$$

and we used it for $a = f_1''(x)$, $b = f_2''(x)$.

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

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

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

We have

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

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.

1.3 Estimate for symmetric functions

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

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

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

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

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

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

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.

1.4 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 $W_x = \{y : \|y - x\|_x < 1\}$.

1.4.1 Main estimate

Lemma 1.1 *Let f be as above. f is defined on W_x and for $\|h\|_x < 1$ we have*

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

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

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

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.

1.4.2 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.

1.5 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).

In Newton method we use scalar product $\langle h_1, h_2\rangle_x = \langle \nabla^2 f(x) h_1, h_2\rangle$ dependent on x which is better adapted to f, than usual scalar product.

Our main estimate means that nondegenerate f is well conditioned on compact subsets of W_x. In particular this implies uniform speed of convergence of Newton method. More precisely, recall that gradient of f at x with respect to norm $\|\cdot\|_x$ is given by

$$(\nabla^2 f(x))^{-1}\nabla f(x)$$

Put

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

When $\lambda(f,x)$ is small (say smaller than $\frac{1}{2}$) we have fast (quadratic) convergence. When $\lambda(f,x)$ is bounded from below, then using step staying in W_x we can still get some fixed decay of objective function.

To stay in domain of f it is natural to use damped Newton method, that is put

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

Lemma 1.2 *We have*

$$f(x_i) - f(x_{i+1}) \geq \lambda(f,x_i) - \log(1+\lambda(f,x_i))$$

Note: For $\lambda(f,x_i) = 1$ this predicts decay approximately by 0.3068, for $\lambda(f,x_i) = \frac{1}{2}$ we get 0.0945.

Example: Let $g(x) = -\gamma x - \log(1-x) - x$. Comparing derivative of g to 0 we see that g attains minimal value at $x = \gamma/(1+\gamma)$. Also $g'(0) = -\gamma$, $g''(0) = 1$, $\lambda(g,0) = \gamma$ and damped Newton method started in $x_0 = 0$ will get minimal value of g in single step and decay of objective function is exactly equal to estimate from the lemma. In particular, single step estimate can not be improved and any other choice of step depending only on λ will lead to worse estimate.

Proof of the lemma: Lemma essentially follows from example and main estimate. Namely, let $w = (\nabla^2 f(x_i))^{-1}\nabla f(x_i)$, $u = w/\|w\|_{x_i}$. Put $\gamma = \langle \nabla f(x_i), u\rangle$ and $g(t) = -\gamma t - \log(1-t) - t$.

We have

$$\lambda(f,x_i) = \|w\|_{x_i} = \langle \nabla f(x_i), (\nabla^2 f(x_i))^{-1}\nabla f(x_i)\rangle^{1/2} = \langle \nabla f(x_i), w\rangle^{1/2}$$

so

$$\gamma = \langle \nabla f(x_i), u\rangle = \frac{\langle \nabla f(x_i), w\rangle}{\|w\|_{x_i}} = \frac{\lambda(f,x_i)^2}{\lambda(f,x_i)} = \lambda(f,x_i).$$

Hence, $\lambda(g,0) = \gamma = \|w\|_{x_i} = \lambda(f,x_i)$. By main estimate

$$f(x_i + tu) \leq f(x_i) + g(t).$$

However, in damped Newton method $x_{i+1} = x_i + tu$ with

$$t = \frac{\|w\|_{x_i}}{1 + \lambda(f,x_i)} = \frac{\gamma}{1+\gamma}$$

which is the same t as value produced by damped Newton method applied to g.

For local convergence we have:

Lemma 1.3 *When x_{i+1} is given by damped Newton method*

$$\lambda(f,x_{i+1}) \leq 2\lambda(f,x_i)^2.$$

When x_{i+1} is given by standard Newton method and $\lambda(f,x_i) < 1$, then

$$\lambda(f,x_{i+1}) \leq \left(\frac{\lambda(f,x_i)}{1-\lambda(f,x_i)}\right)^2$$

In particular, damped Newton method has $\lambda(f,x_{i+1}) < \lambda(f,x_i)$ when $\lambda(f,x_i) < \frac{1}{2}$, while standard Newton when $\lambda(f,x_i) < \frac{3-\sqrt{5}}{2} \approx 0.3819$.

The results are quite satisfactory: when $\lambda(f,x_i)$ is large we get steady decay of objective function, once $\lambda(f,x_i) < \frac{1}{2}$ local convergence takes over. But there is a little troubling aspect: decrease of objective function is rather small when say $\frac{1}{2} < \lambda(f,x_i) < 2$. We will see later that this region is particularly interesting for applications.

We can not get better decay of f, but naively we could hope that small number of steps will transition from large $\lambda(f,x_i)$ to quadratic convergence. Below we show by example that this is not the case: we can get many steps when $\lambda(f,x_i)$ is quite close to 1.

Example: Let $f(x) = -\log(x) + \varepsilon x^2$. It is self-concordant and attains minimum at $x_\infty = \frac{1}{\sqrt{2\varepsilon}}$. When $x_0 = 1$ we have $f'(x_0) = -1 + 2\varepsilon$, $f''(x_0) = 1 + 2\varepsilon$. Easy calculation shows that pure Newton method makes step slightly smaller than 1. More complicated calculation shows that damped Newton method makes step slightly smaller than $\frac{1}{2}$. After single step we may rescale our function so we get problem like original, only with changed ε. So we may get several steps like above. In case of pure Newton method improvement of objective function is approximately $\log(2) \approx 0.693$. In case of damped Newton method improvement is approximately $\log(\frac{3}{2}) \approx 0.405$. So theoretically (with very small ε) we may have very large number of steps with only moderate improvement in each step. Practically, we have here numbers of widely varying magnitude and numerical accuracy will limit number of steps.

This was example in dimension one, but we can add arbitrarily many irrelevant variables. If we add quadratic term in extra variables with minimum at 0, we get true multidimensional problem, which however have the same convergence behaviour as our problem in dimension one.

Note that from point of view of classical convergence theory f is very badly conditioned and our result based on self-concordance is much better.

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.

3 Conjugate direction methods

We already know about gradient descent and Newton method.

Gradient descent may require very large number of iterations, but per iteration cost is small.

Newton method typically converges in smaller number of iterations, but in each iteration we need to solve linear system of equations involving second derivative.

Conjugate direction methods have per iteration cost slightly larger than gradient descent, but should converge faster. In this sense are intermediate between gradient descent and Newton method.

Remark: In practice we care about conjugate gradient method, but theory is nicer when we make things slightly more general that is consider conjugate direction methods.

Conjugate direction methods originally were introduced as storage efficient method of exact solving of some linear systems. More precisely we know that for positive A minimization of

$$\frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

is equivalent to solving

$$Ax = b.$$

Original conjugate direction methods minimized quadratic form, in at most n steps reaching exact solution where n is dimension of the problem.

Later it was observed that stopping method earlier one can get approximate solution and frequently one can get good solution in relatively small number of steps.

Conjugate direction methods were generalized to nonlinear problem. In such case method in no longer convergent in finite number of steps.

Let A be a positive definite matrix.

Definition. We say that a sequence of vectors d_i is A-orthogonal (conjugate) if and only if for $i \neq j$

$$\langle Ad_i, d_j\rangle = 0$$

Recall standard linear algebra:

Lemma 3.1 *If vectors d_i are A-conjugate, then they are linearly independent.*

Remark: To better see that this is standard result we can introduce new scalar product by the formula

$$\langle x, y \rangle_A = \langle Ax, y \rangle.$$

Then A-orthogonal simply means orthogonal with respect to this new scalar product.