CS 798 - Convexity and Optimization, Winter 2017, Waterloo

Lecture 14: Self-concordant barriers

We discuss the theory of self-concordant barriers in solving convex programs using interior point methods, and also discuss ideas how to go beyond the $\tilde{O}(\sqrt{m})$ iteration bound for solving linear programs.

Recap

Last time, we see the analysis of interior point method for linear programming.

We use the log-barrier functions, which make the computations involving the gradient and the Hessian explicit, and this allows us to derive some properties to establish the convergence results.

But other than that, the analysis is quite general, which only require two abstract properties of the barrier functions:

① The Hessian of the barrier function is smooth: $e^{-0.1} H_y \lesssim H_x \lesssim e^{0.1} H_y$ when $\|y-x\|_{H_x} \le \delta$

② The gradient of the barrier function is bounded: $\|\nabla \phi(x)\|_{H^{-1}} \le \sqrt{\beta}$ for any x ($\beta = m$ for log-barrier)

Ignoring the initial term and some constant, these would imply that $\bar{t} \gtrsim t(1+\frac{\delta}{\sqrt{\beta}})$ and the number of iteration is bounded by $\tilde{O}(\frac{\sqrt{\beta}}{\delta})$ iterations.

Likewise, we haven't used much about linear programming, except that helped us to establish the properties of the gradient and the Hessian (as we have explicit formulas).

Plan

Today, we will study some more abstract definitions that would imply the two key properties we want from a barrier function, and this allows us to use the interior point method to solve convex programs.

Also, we will discuss how to improve the parameters of the barrier functions, to obtain faster convergence results for linear programming.

Structure

The first property relates to the (quadratic) convergence of Newton's method in inner iterations.

The second property relates to how large we can increase t in the outer iterations.

We will address these two properties separately, one in each section.

Self-concordance functions

Roughly speaking, self-concordant functions is a class of functions in which we can prove a satisfactory convergence result for Newton's method.

Recall the traditional analysis of convergence of Newton's method in L12: To prove the quadratic convergence results, the assumption is that $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le M\|x-y\|_2$.

This is not a good assumption for the following reasons: recall that Newton's method is affine invariant, but the smoothness parameter M is not. So, it is very unsatisfying that we don't have an affine invariant assumption for the convergence result.

Also, for our purpose, this assumption is not satisfied for barrier functions which blow up in the boundary.

What we are looking for is an affine invariant assumption that also incorporates barrier functions and also implies the first property that we want.

We start by considering functions of a single variable.

Definition A convex function $f:\mathbb{R}\to\mathbb{R}$ is self-concordant if $|f'''(x)| \le 2f''(x)^{\frac{3}{2}}$.

Some observations / remarks before we go on:

- It is about how the second derivative would change given the current second derivative. This is close to what we want in the first property and we will elaborate more about this later. It allows a function to blow up (to incorporate barrier functions) but not so drastically.
- The constant 2 is not so important. If a function f satisfies $|f'''(x)| \le k f''(x)^{\frac{3}{2}}$, then the scaled function $\tilde{f}(x) = \frac{k^2}{4} f(x)$ is self-concordant (check).
- A self-concordant function is affine invariant. Suppose $g(y) = f(ay+b) =: f(x)$. Then g is self-concordant if and only if f is. To see this, $g'(y) = af'(x)$, $g''(y) = a^2 f''(x)$, $g'''(y) = a^3 f'''(x)$, and so $|g'''(y)| = |a^3 f'''(x)| \le 2a^3 f''(x)^{\frac{3}{2}} = 2g''(y)^{\frac{3}{2}}$.

Definition A convex function $f:\mathbb{R}^n\to\mathbb{R}$ is self-concordant if it is self-concordant along every line, i.e. the function $g(t) = f(x+tv)$ is self-concordant for every x and v.

Equivalently, a function $f:\mathbb{R}^n\to\mathbb{R}$ is self-concordant if $|\nabla^3 f(x)[v,v,v]| \le 2\left(v^T\nabla^2 f(x)v\right)^{\frac{3}{2}} = 2\|v\|^3_{\nabla^2 f(x)}$ $\forall x, v$

where $\nabla^3 f(x)$ is the 3-tensor with $(\nabla^3 f(x))_{i,j,k} = \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k}$ and $\nabla^3 f(x)[v,v,v] = \sum_{ijk} (\nabla^3 f(x))_{ijk} v_i v_j v_k$.

Since $\nabla^3 f(x)$ is symmetric, it can be proved that this is equivalent to

$$\left|\nabla^3 f(x)[v_1,v_2,v_3]\right| \le 2\|v_1\|_{\nabla^2 f(x)}\|v_2\|_{\nabla^2 f(x)}\|v_3\|_{\nabla^2 f(x)} \quad \forall x, v_1, v_2, v_3.$$

Examples: We will first see some simple examples, and then some non-trivial examples later.

- $f(x) = -\log x$ is self-concordant. This is why the constant is 2.
- $f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$ is self-concordant, as the sum of concordant functions is concordant and the composition of affine transformation preserves self-concordance.

  So, the log-barrier function is self-concordant.
- $f(x) = -\log\det X$ is self-concordant on $\operatorname{dom} f = S^n_{++}$, by restricting on a line (exercise).

  This is the barrier function used for semidefinite programming.

Properties of self-concordance functions

There are many nice properties of self-concordance functions, but we focus on deriving the first property.

To study how the Hessian changes, the key is to consider the following function:

$$\psi(t) = \frac{1}{\sqrt{u^T \nabla^2 f(x+tu) u}} = \frac{1}{\|u\|_{\nabla^2 f(x+tu)}}.$$

By the self-concordance assumption, we have $|\psi'(t)| = \left|\frac{-\nabla^3 f(x+tu)[u,u,u]}{2\left(u^T\nabla^2 f(x+tu)u\right)^{\frac{3}{2}}}\right| \le 1 \quad \forall t$

For two points x, y, we would like to see how the Hessian norm changes.

We can set $u = y - x$, so that $\psi(1) = \frac{1}{\|y-x\|_{\nabla^2 f(y)}}$ and $\psi(0) = \frac{1}{\|y-x\|_{\nabla^2 f(x)}}$.

Suppose $\|y-x\|_{\nabla^2 f(x)} < 1$, then $\psi(0) > 1$, and as $|\psi'(t)| \le 1$, we have $\psi(1) \ge \psi(0) - 1$.

This implies that $\frac{1}{\|y-x\|_{\nabla^2 f(y)}} \ge \frac{1}{\|y-x\|_{\nabla^2 f(x)}} - 1$, which is equivalent to $\|y-x\|_{\nabla^2 f(y)} \le \frac{\|y-x\|_{\nabla^2 f(x)}}{1 - \|y-x\|_{\nabla^2 f(x)}}$.

We are now ready to derive the Hessian smoothness property.

Theorem If $\|y-x\|_{\nabla^2 f(x)} \le 1$, then $(1 - \|y-x\|_{\nabla^2 f(x)})^2 \nabla^2 f(x) \le \nabla^2 f(y) \le (1 - \|y-x\|_{\nabla^2 f(x)})^{-2} \nabla^2 f(x)$.

proof We would like to see how the quadratic form changes when we move from x to y.

We consider the function $g(t) = u^T \nabla^2 f(x + t(y-x)) u$ so that $g(0) = u^T \nabla^2 f(x) u$ and $g(1) = u^T \nabla^2 f(y) u$.

Let $y_t = x + t(y-x)$ be an intermediate point.

The rate of change $|g'(t)| = \left|\nabla^3 f(y_t)[y-x, u, u]\right| \le 2\|y-x\|_{\nabla^2 f(y_t)}\|u\|^2_{\nabla^2 f(y_t)}$ by self-concordance

$= \frac{2}{t}\|y_t - x\|_{\nabla^2 f(y_t)}\, g(t)$ by definition of y_t

$\le \frac{2}{t}\left(\frac{\|y_t - x\|_{\nabla^2 f(x)}}{1 - \|y_t - x\|_{\nabla^2 f(x)}}\right) g(t)$ by discussion above the thm.

$$\leq \frac{2}{t}\left(\frac{\|y_t - x\|_{\nabla^2 f(x)}}{1 - \|y_t - x\|_{\nabla^2 f(x)}}\right) g(t) \quad \text{by discussion above the thm.}$$

$$= \frac{2\|y - x\|_{\nabla^2 f(x)}}{1 - t\|y - x\|_{\nabla^2 f(x)}} \cdot g(t).$$

This implies that $\frac{-2\|y-x\|_{\nabla^2 f(x)}}{1 - t\|y-x\|_{\nabla^2 f(x)}} \leq \frac{g'(t)}{g(t)} \leq \frac{2\|y-x\|_{\nabla^2 f(x)}}{1 - t\|y-x\|_{\nabla^2 f(x)}}$

$$\Rightarrow \quad 2\ln\left(1 - t\|y-x\|_{\nabla^2 f(x)}\right)' \leq \ln(g(t))' \leq -2\ln\left(1 - t\|y-x\|_{\nabla^2 f(x)}\right)'$$

Integrating in t from 0 to 1, $2\ln\left(1 - \|y-x\|_{\nabla^2 f(x)}\right) \leq \ln(g(1)) - \ln(g(0)) \leq -2\ln\left(1 - \|y-x\|_{\nabla^2 f(x)}\right)$

Exponentiating, we get $\left(1 - \|y-x\|_{\nabla^2 f(x)}\right)^2 \leq \frac{g(1)}{g(0)} \leq \left(1 - \|y-x\|_{\nabla^2 f(x)}\right)^{-2}$, which implies the theorem. □

Another interesting property is that the set $\{y \in \mathbb{R}^n \mid \|y-x\|_{\nabla^2 f(x)} \leq 1\}$ is contained in $\operatorname{dom} f$.

This is a generalization of a statement about Dikin ellipsoid is contained in the polytope, which is a special case when the function f is the log-barrier function.

Using the theorem, one can prove the quadratic convergence of Newton's method for self-concordant functions (see Theorem 4.1.14 of Nesterov).

For our purpose, we can use the theorem to prove the lemma in the inner iteration in L13, and that is what we need for the interior point method to work.

Self concordant barriers

Consider a convex program $\min f_0(x)$

s.t. $x \in Q$ where Q is a convex set.

Without loss of generality, we can assume that $f_0(x) = \langle c, x \rangle$, as otherwise we can just consider

$\min \alpha$ s.t. $x \in Q'$ where $Q' = Q \cap \{f_0(x) \leq \alpha\}$ which is still convex.

Using the interior point method, we need a barrier function $\phi(x)$ s.t. $\phi(x) = \infty$ for $x \in \partial Q$, the boundary of Q and also infinity outside Q.

If it is self-concordant, then Newton's method would work.

Now we focus on the second property, to bound $\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} \quad \forall x$.

Definition Let f be a self-concordant function. We say it is a β-self-concordant barrier if

$$\sup_{u \in \mathbb{R}^n} \left[2\langle \nabla f(x), u \rangle - u^T \nabla^2 f(x) u \right] \leq \beta.$$

Maximizing u will give $\|\nabla f(x)\|^2_{\nabla^2 f(x)^{-1}} \leq \beta$, which is exactly what we wanted, by definition.

By setting u with λu and maximizing the LHS, the definition is equivalent to

$\langle \nabla f(x), u \rangle^2 \le \beta u^T \nabla^2 f(x) u$, which can be written as $\nabla^2 f(x) \succeq \frac{1}{\beta} \nabla f(x) \nabla f(x)^T$.

We have defined what we wanted, so are we done?

There is one more place that we used the special property of log-barrier functions, in arguing that

$\langle c, x_t^* \rangle - \langle c, x^* \rangle \le \varepsilon$ when $t \ge \frac{m}{\varepsilon}$.

Theorem Let $f_t(x) := tc^T x + \phi(x)$ where ϕ is a β-self-concordant barrier.

Then $\langle c, x_t^* \rangle - \langle c, x^* \rangle \le \frac{\beta}{t}$.

proof By the optimality condition, x_t^* satisfies $\nabla f_t(x) = tc + \nabla \phi(x) = 0$.

So, $\langle c, x_t^* \rangle - \langle c, x^* \rangle = \langle c, x_t^* - x^* \rangle = \frac{1}{t} \langle \nabla \phi(x_t^*), x^* - x_t^* \rangle$.

To complete the proof, we show that $\langle \nabla \phi(x), y - x \rangle < \beta$ for any x, y.

Consider the function $g(s) = \langle \nabla \phi(x + s(y-x)), y - x \rangle$ for $s \in [0,1]$.

We would like to prove $g(0) < \beta$.

Assume $g(0) > 0$ as otherwise we have nothing to prove.

Then $g'(s) = (y-x)^T \nabla^2 \phi(x + s(y-x))(y-x)$

$\ge \frac{1}{\beta} \langle \nabla \phi(x + s(y-x)), y - x \rangle^2$ by the equivalent definition of β-self-concordance

$= \frac{1}{\beta} \cdot g(s)^2$.

This mean $g(s)$ is increasing and in particular $g(s) > g(0) > 0$.

Also, integrating $\frac{g'(s)}{g(s)^2} \ge \frac{1}{\beta}$ from $s \in [0,t]$, we have $-\frac{1}{g(t)} + \frac{1}{g(0)} \ge \frac{t}{\beta}$,

and therefore $g(0) < \frac{\beta}{t}$. for any $t \in [0,1]$, and hence $g(0) < \beta$. □

Dikin ellipsoid Another interesting property is that $\| x - x_{ac}^* \|_{H_{x_{ac}}} \le 2\beta$ contains dom f, where x_{ac} is the minimizer of $\phi(x)$. This is a generalization of the result we have seen in Lob, in the special case for the log-barrier function. This is another intuition for the interior point method to converge in $\tilde{O}(\sqrt{\beta})$ steps.

Conclusion

I think this completes the proof of the interior point method for any convex program, using the same proof as in last lecture, and assuming we have a β-self-concordant barrier for the convex set Q, so that the properties that we needed from the barrier function are satisfied.

Then, ignoring the term corresponding to the initial point, the number of iterations is $\tilde{O}(\frac{\sqrt{\beta}}{\delta})$.

Barriers

How to construct a good barrier function for a convex set Q?

For SDP, one can use $\phi(X) = -\log\det(X)$ which is an n-self-concordant barrier for $X \in S^n_+$.

In general, it is a difficult question.

Universal barrier

Nesterov and Nemirovski proved that every convex body $Q \subseteq \mathbb{R}^n$ has a $O(n)$-self-concordant barrier.

Given an interior point $x \in Q$, let $P(x) = \{ s \in \mathbb{R}^n \mid \langle s, y-x \rangle \leq 1 \ \forall y \in Q \}$ be the polar set of the convex body centered at x.

Theorem There are absolute constants c_1 and c_2 such that $\phi(x) = c_1 \ln \text{vol}_n(P(x))$ is a $(c_2 n)$-self-concordant barrier.

For linear programs, this implies a barrier with only $\tilde{O}(\sqrt{n})$ iterations.

The problem, however, is that $\phi(x)$ is expensive to compute.

It is at least as hard as solving the convex program itself, and so it is not useful for algorithmic purposes.

In general, it is optimal and it is a very important result in convex optimization.

Volumetric barrier

The first result that go beyond the log-barrier for linear program is by Vaidya.

A major problem of the log-barrier function is that it puts equal weight on each constraint, so if a constraint is repeated many times it affects the central path.

One idea is to give weight to the constraints so that only a total weight of n is given.

The volumetric barrier is defined as $V(x) = -\frac{1}{2}\log\det(\nabla^2 \phi(x))$ where $\phi(x)$ is the log-barrier.

Note that $V(x)$ can be computed in one determinant computation.

Recall that $\nabla^2 \phi(x) = \sum_{i=1}^m \frac{a_i a_i^T}{(a_i^T x - b_i)^2}$.

Let S_x be the diagonal matrix with $(S_x)_{i,i} = \frac{1}{a_i x - b_i}$, and $A_x = S_x A$.

Then $\nabla^2 \phi(x) = A_x^T A_x$.

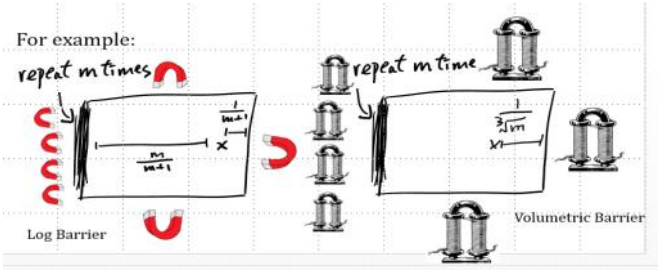
To check that it is a good barrier function, we need to compute its gradient and Hessian.

Gradient $\nabla V(x) = A_x^T \sigma_x$ where $\sigma_{x,i} = \frac{a_i^T \nabla^2 \phi(x)^{-1} a_i}{(a_i x - b_i)^2}$.

Hessian $\nabla^2 V(x) \approx A_x^T \Sigma_x A_x$ where Σ_x is a diagonal matrix with $(\Sigma_x)_{i,i} = \sigma_{x,i}$.

It is not difficult to check that $\sigma_{x,i} \le 1 \ \forall i$ and $\sum_{i=1}^m \sigma_{x,i} = n$.

So, we can interpret the Hessian as a reweighting of the constraints so that the total weight is n.



(from Yin Tat)

With some calculations, one can prove that:

- $\| \nabla V(x) \|_{\nabla^2 V(x)^{-1}} \le \sqrt{n}$,
- if $\| x - y \|_{\nabla^2 V(x)} \le m^{-\frac{1}{2}}$, then $\nabla^2 V(x) \approx \nabla^2 V(y)$.

The first part is good but the second part is not good, and we don't gain anything yet.

The next idea is to use a combination of the two barriers: $V(x) + \frac{n}{m}\phi(x)$.

Then $\| \nabla V(x) \|_{\nabla^2 V(x)^{-1}} = O(\sqrt{n})$ and if $\| x-y \|_{\nabla^2 V(x)} \le \left(\frac{n}{m}\right)^{\frac{1}{4}}$ then $\nabla^2 V(x) \approx \nabla^2 V(y)$.

So, $\beta = O(n)$ and $\delta = \left(\frac{n}{m}\right)^{\frac{1}{4}}$, and this implies $\tilde{O}((mn)^{\frac{1}{4}})$ iterations.

John ellipsoid barrier

Lee and Sidford recently gave a $\tilde{O}(n)$-self-concordant barrier which is also efficiently computable.

Their initial idea is to recurse on Vaidya's idea.

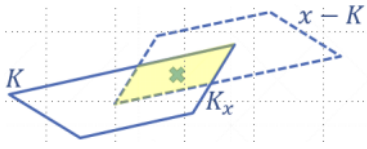
Call Vaidya's barrier function $V_1(x) := -\frac{1}{2}\log\det \nabla^2 \phi(x) + \frac{n}{m}\phi(x)$

They consider $V_k(x) = -\frac{1}{2}\log\det \nabla^2 V_{k-1}(x) + \frac{n}{m} V_{k-1}(x)$, and consider the fixed point $V_\infty(x)$.

So, they keep refining the weight until to a fixed point.

By that time, $\nabla^2 V_\infty(x)$ becomes the John ellipsoid of $K_x := K \cap (x - k)$,

which is a symmetric polytope and $\nabla^2 V_\infty(x)$ is a $\sqrt{n}$-approximation.



(from Yin Tat)

What they do is to compute $\nabla^2 V_{\log(n)}(x)$, which is already a $\tilde{O}(\sqrt{n})$-approximation.

The main problem is to deal with the smoothness issue where Vaidya's volumetric barrier also has, and what they do is to add some magical smoothener terms to establish the Hessian smoothness. Amazingly, these weights can be computed in near linear time using random sampling and linear equation solvers, where for min-cost flow these are Laplacian equations as we saw in L12. As a corollary, this implies a $\tilde{O}(m\sqrt{n})$ exact algorithm for min-cost flow, improving the state of the art for this classical combinatorial problem.

References

- Yin Tat Lee, personal communications.
- Nesterov, Introductory lectures on convex optimization, chapter 4.
- Vishnoi, A mini-course on convex optimization, chapter 4.
- [BV 9.6], [Bubeck, chapter 5]