CSE 599: Interplay between Convex Optimization and Geometry **Winter 2018**

Lecture 15: Self-concordant Function

Lecturer: Yin Tat Lee

Disclaimer: *Please tell me any mistake you noticed.*

Today, we will extend the interior point method to optimize any convex function. For more in-depth treatment, please see the structural programming section in [87].

Recall that any convex optimization problem

$$\min_x f(x)$$

can be rewritten as

$$\min_{\{(x,t):\ f(x)\leq t\}} t.$$

Hence, it suffices to study the problem $\min_{x\in K} c^\top x$. Similar to the case of linear programs, we replace the hard constraint $x \in K$ by a soft constraint as follows:

$$\min_x \phi_t(x) \text{ where } \phi_t(x) = tc^\top x + \phi(x).$$

where $\phi(x)$ is a convex function such that $\phi(x) \to +\infty$ as $x \to \partial K$. We call ϕ is a barrier for K. Note that we put the parameter t in front of the cost $c^\top x$ instead of ϕ as the last lecture, it is slightly more convenient here.

To be concrete, you can always think $\phi(x) = -\sum_{i=1}^n \ln x_i$. Similar to before, we define

Definition 15.0.1. The central path $x_t = \arg\min_x \phi_t(x)$.

The interior point method follows the following framework:

1. Find x close to x_1.
2. While t is not tiny,
   (a) Move x closer to x_t
   (b) $t \to (1+h)\cdot t$.

15.1 Self-concordant Function

We first discuss how to move x closer to x_t, the key part of the analysis. It would motivate the definition of ϕ. This is simply based on Newton method:

$$x' = x - (\nabla^2 \phi_t(x))^{-1} \nabla \phi_t(x).$$

The key property of Newton method is that Newton method is invariant under linear transformation. If we define $y = Ax$ for some matrix A and consider the function $g(y) = \phi_t(A^{-1}y)$. Then, the Newton method on y is equivalent to the Newton method on x as follows:

$$\begin{aligned}
y' &= y - (\nabla^2 g(y))^{-1}\nabla g(y) \\
&= Ax - (A^{-\top}\nabla^2\phi_t(x)A^{-1})^{-1}A^{-\top}\nabla\phi_t(x) \\
&= Ax - A\nabla^2\phi_t(x)^{-1}\nabla\phi_t(x) \\
&= Ax'.
\end{aligned}$$

In general, whenever a method uses the k^{th} order information, we need to assume the k^{th} derivative is continuous in certain sense. Otherwise, the k^{th} derivative is meaningless for algorithmic purpose. For Newton method, it is convenient to assume that the Hessian is Lipschitz. Since Newton method is invariant under linear transformation, it only makes sense to impose an assumption that is invariant under linear transformation.

Definition 15.1.1. Given a convex function f. We define $\|v\|_x^2 = v^\top \nabla^2 f(x) v$. And, we call a function f is self-concordant if for any $h \in \mathbb{R}^n$ and for any x in $\mathrm{dom} f$, we have

$$D^3 f(x)[h,h,h] \le 2\|h\|_x^3$$

where $D^k f(x)[h_1, h_2, \cdots, h_k]$ is the directional k^{th} derivative of f on the direction $h_1, h_2, \cdots, h_k$.

Exercise 15.1.2. Equivalently, restricted on any straight line $g(t) = f(x+th)$, we have $g'''(t) \le g''(t)^{3/2}$. Also, this is same as

$$D^3 f(x)[h_1,h_2,h_3] \le 2\|h_1\|_x \|h_2\|_x \|h_3\|_x .$$

Remark. The constant 2 is chosen such that $-\ln(x)$ exactly satisfies the assumption and it is not very important.

Example 15.1.3. $x^\top Ax$, $-\ln x$, $-\ln(1-\sum x_i^2)$, $-\ln\det X$ are self-concordant.

Not too many convex functions are self-concordant. However, for our purpose, it suffices to show that we can construct a self-concordant barrier for any convex set. Next lecture, we will prove the following:

Theorem 15.1.4. *Any convex set has a $O(n)$-self concordant barrier.*

15.1.1 Newton Step

The self-concordant condition says that locally, Hessian does not change too fast. Integrating the self-concordant condition, we have the following:

Lemma 15.1.5. *For any $x \in \mathrm{dom} f$ and any $\|y-x\|_x < 1$, we have that*

$$(1-\|y-x\|_x)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1-\|y-x\|_x)^2}\nabla^2 f(x).$$

Proof. Let $\alpha(t) = \left\langle \nabla f^2(x+t(y-x))u, u\right\rangle$. Then, we have that $\alpha'(t) = D^3 f(x+t(y-x))[y-x,u,u]$. By the self-concordant, we have

$$|\alpha'(t)| \le 2\|y-x\|_{x+t(y-x)} \|u\|_{x+t(y-x)}^2 . \tag{15.1}$$

For $u = y-x$, we have $|\alpha'(t)| \le 2\alpha(t)^{\frac{3}{2}}$. Hence, we have $\frac{d}{dt}\frac{1}{\sqrt{\alpha(t)}} \ge -1$. Integrating both on t, we have

$$\frac{1}{\sqrt{\alpha(t)}} \ge \frac{1}{\sqrt{\alpha(0)}} - t = \frac{1}{\sqrt{\|x-y\|_x}} - t.$$

Rearranging it gives

$$\alpha(t) \leq \frac{1}{(\frac{1}{\sqrt{\|x-y\|_x}} - t)^2} = \frac{\|x-y\|_x^2}{(1 - t\,\|x-y\|_x)^2}.$$

For general u, (15.1) gives

$$|\alpha'(t)| \leq 2\frac{\|x-y\|_x}{1 - t\,\|x-y\|_x}\alpha(t).$$

Rearranging it gives

$$\left|\frac{d}{dt}\ln\alpha(t)\right| \leq 2\frac{\|x-y\|_x}{1-t\,\|x-y\|_x} = -2\frac{d}{dt}\ln(1 - t\,\|x-y\|_x)$$

Integrating both from $t = 0$ to 1 gives the result. $\square$

Now, we are ready to study convergence of Newton method:

Lemma 15.1.6. *Given a convex function f. Consider the algorithm $x' = x - (\nabla^2 f(x))^{-1}\nabla f(x)$. Suppose that $r = \|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} < 1$, then we have*

$$\|\nabla f(x')\|_{\nabla^2 f(x')^{-1}} \leq \left(\frac{r}{1-r}\right)^2.$$

Remark 15.1.7. Note that $\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} = \left\|\nabla^2 f(x)^{-1}\nabla f(x)\right\|_x$ is the step size of Newton method. This is a measurement of the error.

Proof. Lemma 15.1.5 shows that

$$\nabla^2 f(x') \succeq (1-r)^2\nabla^2 f(x).$$

and hence

$$\|\nabla f(x')\|_{\nabla^2 f(x')^{-1}} \leq \frac{\|\nabla f(x')\|_{\nabla^2 f(x)^{-1}}}{1-r}.$$

To bound $\nabla f(x')$, we calculate that

$$\begin{aligned}
\nabla f(x') &= \nabla f(x) + \int_0^1 \nabla^2 f(x + t(x'-x))(x'-x)dt \\
&= \nabla f(x) - \int_0^1 \nabla^2 f(x + t(x'-x))(\nabla^2 f(x))^{-1}\nabla f(x)dt \\
&= \left(\nabla^2 f(x) - \int_0^1 \nabla^2 f(x + t(x'-x))dt\right)(\nabla^2 f(x))^{-1}\nabla f(x). \qquad (15.2)
\end{aligned}$$

For the first term in the bracket, we use Lemma 15.1.5 to get that

$$(1 - r + \frac{1}{3}r^2)\nabla^2 f(x) \preceq \int_0^1 \nabla^2 f(x + t(x'-x))dt \preceq \frac{1}{1-r}\nabla^2 f(x).$$

Therefore, we have

$$\left\|(\nabla^2 f(x))^{-\frac{1}{2}}\left(\nabla^2 f(x) - \int_0^1 \nabla^2 f(x + t(x'-x))dt\right)(\nabla^2 f(x))^{-\frac{1}{2}}\right\|_{\text{op}} \leq \max(\frac{r}{1-r}, r - \frac{1}{3}r^2) = \frac{r}{1-r}.$$

Putting it into (15.2) gives

$$\|\nabla f(x')\|_{\nabla^2 f(x)^{-1}} \leq \frac{r}{1-r}\left\|(\nabla^2 f(x))^{-\frac{1}{2}}\nabla f(x)\right\| = \frac{r^2}{1-r}.$$

$\square$

Now, we can make our framework more precise:

1. Find x such that $\|\nabla\phi_1(x)\|_{\nabla^2\phi(x)^{-1}} \leq \frac{1}{3}$ (Note that $\nabla^2\phi_1 = \nabla^2\phi$.)
2. While t is not tiny,
   (a) $x \leftarrow x - (\nabla^2\phi(x))^{-1}\nabla\phi_t(x)$. (This decreases $\|\nabla\phi_t(x)\|_{\nabla^2\phi(x)^{-1}}$ to $\frac{1}{4}$.)
   (b) $t \to (1+h)\cdot t$.

Now, it suffices to prove that after step b, we have $\|\nabla\phi_t(x)\|_{\nabla^2\phi(x)^{-1}} \leq \frac{1}{3}$.

Lemma 15.1.8. *We have that*

$$\left\|\nabla\phi_{(1+h)t}(x)\right\|_{\nabla^2\phi(x)^{-1}} \leq (1+h)\left\|\nabla\phi_t(x)\right\|_{\nabla^2\phi(x)^{-1}} + h\left\|\nabla\phi(x)\right\|_{\nabla^2\phi(x)^{-1}}.$$

Proof. Note that $\nabla\phi_{(1+h)t}(x) = (1+h)tc + \nabla\phi(x)$. Hence, $\nabla\phi_{(1+h)t}(x) = (1+h)\nabla\phi_t(x) - h\nabla\phi(x)$. ☐

Using that $\|\nabla\phi_t(x)\|_{\nabla^2\phi(x)^{-1}} \leq \frac{1}{4}$, we can take step

$$h = \frac{1}{3 + 12\max_x \|\nabla\phi(x)\|_{\nabla^2\phi(x)^{-1}}}$$

and still maintain $\left\|\nabla\phi_{(1+h)t}(x)\right\|_{\nabla^2\phi(x)^{-1}} \leq \frac{1}{3}$.

Due to this, we define the following:

Definition 15.1.9. We call ϕ is a ν-self concordant barrier on K if $\|\nabla\phi(x)\|^2_{\nabla^2\phi(x)^{-1}} \leq \nu$ for all $x \in K$ and that $D^3\phi(x)[h,h,h] \leq D^2\phi(x)[h,h]^{3/2}$ for all $x \in K$ and h.

Now, we only need to handle the issue of computing initial point and how large t need to be.

15.1.2 Initial Point

If we have an interior point $\overline{x} \in K$, then we can set temporarily $c = -\nabla\phi(\overline{x})$. For this c, $\overline{x}$ is on the central path with $t = 1$. Now, we gradually decrease t until t is very close to 0. Then, we can switch back to the original cost. Since t is very small when we switch the cost, it does not change the gradient too much.

Unfortunately, if we do not have an interior point $x \in K$ but only with access to $\phi(x)$, there is no way to solve $\min_{x\in K} c^\top x$ because we would always get $\phi(x) = +\infty$ until we can find an interior point.

15.1.3 Termination Condition

Intuitively, we should think $\phi_t(x_t)$ tends to optimum as $t \to \infty$. We first need a lemma showing that the gradient of ϕ is small.

Lemma 15.1.10 (Duality Gap)**.** *Suppose that ϕ is a ν-self concordant barrier. For any $x, y \in K$, we have that*

$$\langle\nabla\phi(x), y - x\rangle \leq \nu.$$

Proof. Let $\alpha(t) = \langle\nabla\phi(x_t), y - x\rangle$ where $x_t = x + t(y-x)$. Then, we have

$$\alpha'(t) = \left\langle\nabla^2\phi(x_t)(y-x), y-x\right\rangle.$$

Note that
$$\alpha(t) \leq \|\nabla\phi(x_t)\|_{\nabla^2\phi(x_t)} \|y-x\|_{\nabla^2\phi(x_t)^{-1}} \leq \sqrt{v}\,\|y-x\|_{\nabla^2\phi(x_t)^{-1}}.$$

Hence, we have $\alpha'(t) \geq \frac{1}{v}\alpha(t)^2$. If $\alpha(0) \leq 0$, then we are done. Otherwise, α is increasing and hence $\alpha(1) > 0$. Since $\frac{1}{\alpha(1)} \leq \frac{1}{\alpha(0)} - \frac{1}{v}$. So, $\alpha(0) \leq v$. □

Lemma 15.1.11 (Duality Gap)**.** *Suppose that ϕ is a ν-self concordant barrier, we have that*
$$\langle c, x_t\rangle \leq \langle c, x^*\rangle + \frac{\nu}{t}.$$

Proof. Let x^* be a minimizer of $c^\top x$ on K. By optimality, we have $tc + \nabla\phi(x) = 0$. Therefore, we have
$$\langle c, x_t\rangle - \langle c, x^*\rangle = \frac{1}{t}\langle\nabla\phi(x_t), x^* - x_t\rangle \leq \frac{\nu}{t}.$$
□

Hence, it suffices to end with $t = \nu/\varepsilon$, which is exactly same as the previous lecture.

Theorem 15.1.12. *Given a ν-self concordant barrier, we can solve $\min_{x\in K} c^\top x$ up to ε error in $O^*(\sqrt{\nu})$ iterations.*

15.2 Examples

First of all, we need some self-concordant barriers for some simplex convex sets:

Lemma 15.2.1. *We have the following self-concordant barriers. We use ν-sc as a short form for ν-self-concordant barrier.*

- $-\ln x$ *is* 1*-sc for* $\{x \geq 0\}$.
- $-\ln\cos(x)$ *is* 1*-sc for* $\{|x| \leq \frac{\pi}{2}\}$.
- $-\ln(t^2 - \|x\|^2)$ *is* 2*-sc for* $\{t \geq \|x\|_2\}$.
- $-\ln\det X$ *is n-sc for* $\{X \in \mathbb{R}^{n\times n}, X \succeq 0\}$.
- $-\ln x - \ln(\ln x + t)$ *is* 2*-sc for* $\{x \geq 0, t \geq -\ln x\}$.
- $-\ln t - \ln(\ln t - x)$ *is* 2*-sc for* $\{t \geq e^x\}$.
- $-\ln x - \ln(t - x\ln x)$ *is* 2*-sc for* $\{x \geq 0, t \geq x\ln x\}$.
- $-2\ln t - \ln(t^{2/p} - x^2)$ *is* 4*-sc for* $\{t \geq |x|^p\}$ *for* $p \geq 1$.
- $-\ln x - \ln(t^p - x)$ *is* 2*-sc for* $\{t^p \geq x \geq 0\}$ *for* $0 < p \leq 1$.
- $-\ln t - \ln(x - t^{-1/p})$ *is* 2*-sc for* $\{x > 0, t \geq x^{-p}\}$ *for* $p \geq 1$.
- $-\ln x - \ln(t - x^{-p})$ *is* 2*-sc for* $\{x > 0, t \geq x^{-p}\}$ *for* $p \geq 1$.

In general, one can build a self-concordant barriers by combining the barriers above.

Lemma 15.2.2. *If ϕ_1 and ϕ_2 are ν_1 and ν_2-self concordant barriers on K_1 and K_2 with respectively, then $\phi_1 + \phi_2$ is a $\nu_1 + \nu_2$ self concordant barrier on $K_1 \cap K_2$.*

Lemma 15.2.3. *If ϕ is a ν-self concordant barrier on K, then $\phi(Ax)$ is a ν-self concordant on $\{y : Ay \in K\}$.*

Exercise 15.2.4. Using the lemmas above, prove that $-\sum_{i=1}^m \ln(a_i^\top x - b_i)$ is a m-self concordant barrier on $\{Ax \geq b\}$.

Here let me consider two examples:

15.2.1 ℓ_p regression

Consider the problem

$$\min_x \sum_{i=1}^m |a_i^\top x - b_i|^p$$

where $x \in \mathbb{R}^n$. Rewriting it, we have

$$\min_{x\in\mathbb{R}^n, s,t\in\mathbb{R}^m} \sum_i t_i \text{ subjects to } Ax - b = s, t \geq |s|^p.$$

Hence, the regularized function is simply

$$\min_{Ax-b=s} \mu \sum_i t_i - 2\sum \ln t_i - \sum_i \ln(t_i^{2/p} - s_i^2).$$

This gives a $O^*(\sqrt{m})$ iterations algorithm to solve ℓ_p regression. Recently, it is improved to [86].

15.2.2 John ellipsoid

Recall that we proved the following theorem in Lecture 5.

Theorem 15.2.5. *Given any convex set K. Define $a_i^\top x \leq 1$ be the separating hyperplanes of K, namely $K = \bigcap_{i\in I}\{x : a_i^\top x \leq 1\}$. Then, $J(K)$ uniquely exists and is given by $J(K) = \{x : \left\|G^{-1}(x-v)\right\|_2 \leq 1\}$ where G is the maximizer of the problem*

$$\max_{G\succeq 0, v\in\mathbb{R}^n} \ln\det G \quad \textit{subjects to} \quad \|Ga_i\|_2 \leq 1 - a_i^\top v \textit{ for all } i \in I.$$

Now, we show how to solve it using interior point method. We can rewritten the problem as

$$\min \tau \quad \text{subjects to} \quad -\ln\det G \leq \tau, \|Ga_i\|_2 \leq 1 - a_i^\top v, G \succeq 0.$$

The barrier for that set is

$$-\ln(\tau + \ln\det G) - \sum_i \ln\left((1 - a_i^\top v)^2 - \|Ga_i\|_2^2\right) - \ln\det G.$$

References

[86] Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, and Yuanzhi Li. An homotopy method for lp regression provably beyond self-concordance and in input-sparsity time. *arXiv preprint arXiv:1711.01328*, 2017.

[87] Yu Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 1998.