1 Introduction

We're going to build up to interior point methods for convex optimization. Our goal is to solve problems of the form:

$\min c^T x$
s.t. $x \in \mathcal{X}$

for $\mathcal{X} \subseteq \mathbb{R}^n$, convex and compact.

We call a convex function F a barrier if: $\lim_{x \to \partial \mathcal{X}} \to \infty$.

Now for some $t \in \mathbb{R}^+$, let $x^*(t) = \arg\min_{x \in \mathbb{R}^n} tc^T x + F(x)$.

The path $(x^*(t))_{t \in \mathbb{R}^+}$ is the "central path" of the optimization. Intuitively, as t increases, $x^*(t)$ approaches x^*.

Our goal is as follows: we will optimize the function $x^*(t')$ with our initial point of the optimization being a previously computed $x^*(t)$. We will try to balance choosing a t' large (so that we make a lot of progress) while still choosing t' small enought to make sure the optimization will be extremely efficient.

The plan for the next two weeks is:

1. Characterize the region of fast convergence for Newton's method
2. Find the maximum t' we can choose
3. Compute $x^*(0)$ (an initial point of our overall optimization.

We'll do item 1 today.

2 Analysis of Newton's Method

For the rest of this talk, we'll assume f is continuously differentiable enough times that all of the derivatives we write are continuous.

Newton's method is based off of the following observation, the Taylor expansion of f at x is:

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h + o(||h||^2)$$

so we should probably move in the direction that minimizes

$$h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h$$

For $\nabla^2 f(x)$ positive definite, choosing $h = -[\nabla^2 f(x)]^{-1} \nabla f(x)$ suffices (because the derivative of that expression is $h^T \nabla^2 f(x) + \nabla f(x)$).

Thus our update rule is

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

If we start this optimization close enough to the optimum, we will converge quickly:

Theorem 1. *Let f have an M-Lipschitz Hessian i.e. $||\nabla^2 f(x) - \nabla^2 f(y)|| \leq M||x - y||$, and let x^* be a local minimum of f such that $\nabla^2 f(x^*) \succcurlyeq \mu I_n$ for some $\mu > 0$. Let x_0 be a starting point such that $||x_0 - x^*|| \leq \mu/(2M)$.*

Then Newton's method converges to x^* in $\log \log 1/\epsilon$ steps as $||x_{k+1} - x^*|| \leq \frac{M}{\mu} ||x_k - x^*||^2$.

Proof. We begin with the following equation, which is an immediate corollary of FTC.

$$\int_0^1 \nabla^2 f(x + sh) h \, ds = \nabla f(x + h) - \nabla f(x) \tag{1}$$

Plug in x^* for x and $x_k - x^*$ for h to get:

$$\int_0^1 \nabla^2 f(x + s(x_k - x^*)) \cdot (x_k - x^*) \, ds = \nabla f(x_k) - \nabla f(x^*) = \nabla f(x_k) \tag{2}$$

Now consider $x_{k+1} - x^*$. We have

$$\begin{aligned}
x_{k+1} - x^* &= x_k - x^* - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k) && \text{defn of } x_n \\
&= x_k - x^* - [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x^* + s(x_k - x^*))(x_k - x^*) \, ds && \text{by 2} \\
&= [\nabla^2 f(x_k)]^{-1} \int_0^1 \left[\nabla^2 f(x_k) - \nabla^2 f(x^* + s(x_k - x^*))\right] (x_k - x^*) \, ds
\end{aligned}$$

Where the last line follows from

$$x_k - x^* = [\nabla^2 f(x_k)]^{-1} \nabla^2 f(x_k)(x_k - x^*) = [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x_k)(x_k - x^*) \, ds$$

Now applying Cauchy-Schwartz on the last equation we have:

$$||x_{k+1} - x^*|| \leq ||[\nabla^2 f(x_k)]^{-1}|| \cdot \left(\int_0^1 ||\nabla^2 f(x_k) - \nabla^2 f(x^* + s(x_k - x^*))|| \, ds \right) ||x_k - x^*|| \quad (3)$$

We assumed the Hessian was Lipschitz, applying that property, we have that the integral is at most $\frac{M}{2}||x_k - x^*||$.

By Lipschitzness (and the fact that $||A - B|| \leq s$ if and only if $sI_n \succeq A - B \succeq -sI_n$) we have

$$||\nabla^2 f(x_k) - \nabla^2 f(x^*)|| \leq M||x_k - x^*||$$

if and only if

$$M||x_k - x^*||I_n \succeq \nabla^2 f(x_k) - \nabla^2 f(x^*) \succeq -M||x_k - x^*||I_n$$

Take the second of these inequalities, and we can convert to:

$$\nabla^2 f(x_k) \succeq \nabla^2 f(x^*) - M||x_k - x^*||I_n$$

By our assumption on the Hessian at x^* we have:

$$\nabla^2 f(x_k) \succeq \nabla^2 f(x^*) - M||x_k - x^*||I_n \succeq (\mu - M||x_k - x^*||)I_n$$

We assumed the initial point x_0 was close to the optimum (i.e. $||x_0 - x^*|| \leq \mu/(2M)$). We claim that $||x_k - x^*|| \leq \mu/(2M)$ as well (this is easy to verify inductively, as (we will show) we are getting closer to x^*. Using this assumption we have:

$$\nabla^2 f(x_k) \succeq (\mu - M||x_k - x^*||)I_n \succeq \left(\mu - M\frac{\mu}{2M}\right) I_n = \frac{\mu}{2} I_n$$

Returning to Equation 3, and plugging in the bounds on the terms we have:

$$||x_{k+1} - x^*|| \leq \left(\frac{2}{\mu}\right) \cdot \left(\frac{M}{2}||x_k - x^*||\right) \cdot ||x_k - x^*|| = \frac{M}{\mu}||x_k - x^*||^2$$

□

3 Self-Concordant Functions

Now that we've done this analysis, we're going to argue that the previous analysis was dumb. Consider an invertable matrix A. Let f be the map of Newton's method, and let ϕ map $y = Ax$ to $f(A^{-1}y)$. That is

$$x^+ = x - [\nabla^2 f(x)]^{-1} \nabla f(x) \text{ and } y^+ = y - [\nabla^2 \phi(y)]^{-1} \nabla \phi(y)$$

One can show that $y^+ = Ax^+$. That is even after an affine transformation, Newton's method follows the same trajectory. (Newton's method is the only algorithm we've seen in reading group with this property). But with this observation, the assumption we made in the last section that the Hessian is Lipschitz seems a little silly. It assumes we're using some fixed inner-product, but we just showed we can change the inner product without changing the execution of the algorithm. Thus we'd like a different theorem that works for any inner-product.

What's the right way to measure a norm now? Well first we need it to not change when we do a linear transformation, it should only care about the local geometry. That sounds like it should involve the Hessian somehow. How should it change with the Hessian? Well let's consider taking the norm of the gradient. We want a smaller gradient to indicate being close to x^*. As a thought experiment, let's say that the gradient has some value k in two directions. In one direction the Hessian is large, in the other it's small. Which is more concerning to us? It's the one where the Hessian is small – in that direction, we'd need more steps to decrease the gradient to 0, thus our norm should be larger in directions where the gradient is smaller.

Thus we'll use the following norm.

$$||h||_x := \sqrt{h^T \nabla^2 f(x) h}$$

Definition 1. *For $\mathcal{X}$ a convex set with non-empty interior, and let f be a closed, thrice-continuously differentiable on int($\mathcal{X}$). We say f is self-concordant (with constant M) if for all $x \in$ int($\mathcal{X}$) and all $h \in \mathbb{R}^n$:*

$$\nabla^3 f(x)[h,h,h] \leq M||h||_x^3$$

For intuition $f(x) = -\log x$ for $x > 0$ is self-concordant with constant 2.

A tedious limit argument can relate our new definition to barriers:

Lemma 1. *If f is self-concordant then f is also a barrier.*

Proof. (from Nesterov)

Consider a sequence $\{x_k\} \subseteq \text{dom} f$ such that $\lim_k x_k \to \overline{x}$ To show f is a barrier we need to show $\lim f(x_k) \to \infty$. Note that $\{f(x_k)\}$ is bounded below. By convexity of f, $f(x_k) \geq f(x_0) + \langle f'(x_0), x_k - x_0 \rangle$ which is at most some constant. Now suppose for contradiction, the sequence $\{f(x_k)\}$ is bounded above. By convexity of f, this sequence actually has a limit, call it $\overline{f}$. Now we have that the sequence $z_k = (x_k, f(x_k)) \to \overline{z} = (\overline{x}, \overline{f})$.

$z_k \in \text{epi} f$ but $\overline{x}$ is not in the domain of f, thus $\overline{z}$ is not in the epigraph of f. But we assumed f was a closed function, a contradiction. □

Our next definition will be used to define the region of very fast convergence for Newton's method.

Definition 2. *Let f be a self-concordant function with constant 2 on $\mathcal{X}$. for $x \in int(\mathcal{X})$ we say $\lambda_f(x) = ||\nabla f(x)||_x^* = \sqrt{\nabla f(x)^T (\nabla^2 f(x))^{-1} \nabla f(x)}$ is the **Newton decrement** of f at x.*

The usefulness of this definition is the following lemma.

Lemma 2. *If x is such that $\lambda_f(x) < 1$ and $x^* = \arg\min f(x)$ then*

$$||x - x^*||_x \leq \frac{\lambda_f(x)}{1 - \lambda_f(x)}$$

We won't prove this, because Seb just skips it and the proof is long. Note that one way to interpret this lemma is that when λ_f is $O(\epsilon)$, then $||x - x^*||_x$ is also $O(\epsilon)$.

The lemma above is the key to proving the following result about Newton's Method:

Theorem 2. *If f is self-concordant with constant 2 on $\mathcal{X}$ and $x \in int(\mathcal{X})$ such that $\lambda_f(x) \leq 1/4$ then*

$$\lambda_f\left(x - [\nabla^2 f(x)]^{-1} \nabla f(x)\right) \leq 2\lambda_f(x)^2$$

That is, if we inititalize Newton's Method with x_0 fitting the hypothesis then the iterates satisfy $\lambda_f(x_{k+1}) \leq 2\lambda_f(x_k)^2$. Now we see item 1 of our TODO list is satisfied by a self-concordant function for the barrier.