1 Overview

In this lecture we will describe a path-following implementation of the Interior Point Method to a general LP problem using Newton’s method [Gon92]. In the previous lecture we described gave an introduction to interior point and also discussed a first-order optimization method (gradient descent).

As in last lecture, our LP takes the form: Minimize $c^T x$ such that $Ax \geq b$.

2 Recap of Gradient Descent

Gradient descent works by repeatedly approximately minimizing

$$f_\lambda(x) = \lambda c^T x - \sum_{i=1}^{m} \ln(s(x)_i),$$

where $s(x) = Ax - b$ is the slack vector. Gradient descent is a first order method, meaning that we are supplied with an oracle that, given $x$ provides us access to $f(x)$ and the gradient $\nabla f(x)$, where $f = f_\lambda$.

3 Newton’s Method

For IPM we will use a second order method, which also provides oracle access to compute the Hessian $\nabla^2 f(x)$ at any query point $x \in \mathbb{R}^n$ (recall that this is the matrix of second partial derivatives of $f$). As before, want to minimize some function $f : \mathbb{R}^n \to \mathbb{R}$ (which for us will eventually be $f_\lambda$), but now we have an oracle giving us $f, \nabla f, \nabla^2 f$. As usual, we start with some initial $x_{\text{init}} \in \mathbb{R}^n$ and repeatedly apply an update rule, gradually decreasing our error. Given some iterate $x_0 \in \mathbb{R}^n$, the update rule to form the next iterate $x_1$ is

$$x_1 \leftarrow x_0 - (\nabla^2 f(x_0))^{-1}\nabla f(x_0).$$

Note that this rule comes from solving for a minimum in the second-order Taylor approximation

$$f(y) \approx f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle + \frac{1}{2} \langle y - x_0, \nabla^2 f(x_0)(y - x_0) \rangle.$$ 

Now we wish to understand how well this method behaves, i.e. how fast it decreases our error over iterations. The theorem below and its proof are from some notes Aaron Sidford wrote for himself and shared with me.
Theorem 1. For \( t \in [0,1] \) define
\[
x_t = x_0 + t(x_1 - x_0),
\]
Suppose \( \exists \varepsilon > 0 \) s.t. \( \forall t \in [0,1] \),
\[
(1 - \varepsilon)\nabla^2 f(x_0) \preceq \nabla^2 f(x_t) \preceq (1 + \varepsilon)\nabla^2 f(x_0),
\]
where \( A \preceq B \) is the Loewner order given by \( B - A \) being positive semidefinite (PSD). Then
\[
\|\nabla f(x_1)\|_{(\nabla^2 f(x_1))^{-1}} \leq \frac{\varepsilon}{1 - \varepsilon}\|\nabla f(x_0)\|_{(\nabla^2 f(x_0))^{-1}}.
\]
(Definition: For a PSD \( A \), we define \( \|x\|_A := \sqrt{x^T Ax} \)).

Observations: Theorem 1 gives us a bound on the rate of convergence of Newton’s method. Given that the Hessian \( \nabla^2 f \) doesn’t change too much along the optimization path \( x_t \), Theorem 1 tells us that the \( \nabla f \) decreases by a small multiplicative constant \( \varepsilon/(1 - \varepsilon) \). Of course having small gradient is equivalent to being close to the optimum, since \( \nabla f = 0 \iff f \) is minimized. In fact, throughout Newton’s method we measure proximity to the optimum by the size of the gradient vector.

We can rewrite the condition of Theorem 1 as follows:
\[
(1 - \varepsilon)A \preceq B \preceq (1 + \varepsilon)A \iff -\varepsilon A \preceq B - A \preceq \varepsilon A
\]
\[
\iff \forall x \in \mathbb{R}^n, -\varepsilon x^T Ax \leq x^T (B - A)x \leq \varepsilon x^T Ax
\]
\[
\iff \forall y \in \mathbb{R}^n, -\varepsilon y^T y \leq y^T A^{-1/2}(B - A)A^{-1/2}y \leq \varepsilon y^T y
\]
\[
\iff -\varepsilon I \preceq A^{-1/2}(B - A)A^{-1/2} \preceq \varepsilon I
\]
For the third equivalence, we substituted \( y = A^{1/2}x \). This form will prove more amenable to our calculations. In particular, if we apply these calculations to (1), we find that given the conditions of the theorem,
\[
-\varepsilon I \preceq (\nabla^2 f(x_0))^{-1/2}(\nabla^2 f(x_0) - \nabla^2 f(x_t))(\nabla^2 f(x_0))^{-1/2} \preceq \varepsilon I.
\]
(2)
Let us now prove the convergence theorem.

Proof. We can write (by the Fundamental Theorem of Calculus)
\[
\nabla f(x_1) = \nabla f(x_0) + \int_0^1 \nabla^2 f(x_t)(x_t - x_0)dt
\]
\[
= \int_0^1 (\nabla^2 f(x_0) - \nabla^2 f(x_t))(\nabla^2 f(x_0))^{-1}\nabla f(x_0)dt
\]
Now we want to bound \( \|\nabla f(x_1)\|_{(\nabla^2 f(x_1))^{-1}} \), so plugging in the above integral, we get
Recall that we defined the objective function to be $f_\lambda(x) = \lambda e^T x - \sum_{i=1}^m \ln(s(x)_i)$.

But actually the Hessians $\nabla^2 f(x_t)$ are uniformly bounded within a factor of $\varepsilon$ by $\nabla^2 f(x_0)$ spectrally. Explicitly, the condition of the Theorem tells us that $(\nabla^2 f(x_1))^{-1}$ has the same eigenvalues up to a factor of $\varepsilon$ as the same matrix at $x_0$. Thus we can replace the norm with respect to $(\nabla^2 f(x_1))^{-1}$ to that with respect to $(\nabla^2 f(x_0))^{-1}$, and increase the norm by at most a $1/(1 - \varepsilon)$ factor. Bringing the norm inside the integral also only increases the right hand side. Thus

$$\|\nabla f(x_1)\|((\nabla^2 f(x_1))^{-1} \leq \frac{1}{1 - \varepsilon} \int_0^1 \|((\nabla^2 f(x_0))^{-1/2}\nabla f(x_0))((\nabla^2 f(x_0))^{-1/2}(\nabla^2 f(x_0) - \nabla^2 f(x_t))(\nabla^2 f(x_0))^{-1/2})^2 dt$$

The last being a purely algebraic identity using the definition of the norm $\| \cdot \|_A$ as defined above. Call the matrix inside the square of the norm

$$M = ((\nabla^2 f(x_0))^{-1/2}(\nabla^2 f(x_0) - \nabla^2 f(x_t))(\nabla^2 f(x_0))^{-1/2}$$

and notice that it is exactly of the form described at the end of section 2.1. Thus inequalities (2) say exactly

$$-\varepsilon I \leq M \leq \varepsilon I \implies 0 \leq M^2 \leq \varepsilon^2 I$$

It is a general fact that if $0 \leq A \leq B$ then $\|x\|_A \leq \|x\|_B$ for all $x \in \mathbb{R}^n$. Thus, we can transform the last inequality to give

$$\|\nabla f(x_1)\|((\nabla^2 f(x_1))^{-1} \leq \frac{1}{1 - \varepsilon} \int_0^1 \|((\nabla^2 f(x_0))^{-1/2}\nabla f(x_0))\|_{\varepsilon^2 I} dt$$

$$\leq \frac{\varepsilon}{1 - \varepsilon} \|((\nabla^2 f(x_0))^{-1/2}\nabla f(x_0))\|_2$$

$$= \frac{\varepsilon}{1 - \varepsilon} \sqrt{(\nabla f(x_0))^T(\nabla^2 f(x_0))^{-1}(\nabla f(x_0))}.$$
and so we can calculate the gradient and Hessian explicitly, using the definition \( s(x) = Ax - b \) of the slack vector:

\[
\nabla f_\lambda(x) = \lambda c - A^T S_x^{-1} 1 \\
\nabla^2 f_\lambda(x) = A^T S_x^{-2} A
\]

Here \( 1 \in \mathbb{R}^m \) is the all-1's vector. Now, Newton’s method runs as before. We start with some fixed \( \lambda_k \), run Newton’s method a few steps until we get an approximate solution for \( \lambda_k \), then increment the \( \lambda \) to \( \lambda_k + 1 \) and iterate. To measure the speed of approximation, we introduce the concept of centrality.

### 4.1 Centrality

Define the centrality of \( x \) for \( f_\lambda \) as \( \delta_k(x) = \|\nabla f_\lambda\| (\nabla^2 f_\lambda(x))^{-1} \). If \( x \) is the actual optimum for \( \lambda_k \) then the centrality is 0 since the gradient will vanish there.

Define \( x \) to be finely central if \( \delta_k(x) \leq \frac{1}{100} \), and coarsely central if \( \delta_k(x) \leq \frac{1}{3} \).

We will show that if \( x \) is finely central for a given \( k \), then when we increment \( \lambda_k \) to \( \lambda_k + 1 \) (but not by much), the same \( x \) will be coarsely central for \( \lambda_k + 1 \). We then run Newton’s method to go from a coarsely central point to a finely central point. We will pick the increment on \( \lambda \) so that any finely central point for \( \lambda_k \) remains coarsely central for \( \lambda_k + 1 \).

Note Theorem 1 is telling us exactly that if the Hessians aren’t changing much along the optimization path, then \( \delta_k(x_1) \leq \varepsilon / (1 - \varepsilon) \delta_k(x_0) \), so we will need a constant number of such steps to get from a coarsely central point to a finely central point.

### 4.2 Centrality Implies Conditions of Newton’s Method

How does the Hessian \( \nabla^2 f_\lambda \) change as we change \( x \)? We need it to not change too much on the line from one iterate of Newton’s method to the next in order to apply Theorem 1. Note that

\[
\nabla^2 f_\lambda(x) = A^T S_x^{-2} A
\]

doesn’t depend on \( \lambda_k \), just the slacks \( S_x \). Thus we only need to understand how the slack matrix and its eigenvalues change. In particular, suppose we are at some iterate \( x_0 \) of Newton’s method and update to \( x_1 \), and we again define \( x_t = x_0 + t(x_1 - x_0) \) for \( t \in [0, 1] \) Then

\[
(1 - \varepsilon)S_{x_0} \preceq S_{x_t} \preceq (1 + \varepsilon)S_{x_0} \iff -\varepsilon S_{x_0} \preceq S_{x_t} - S_{x_0} \preceq \varepsilon S_{x_0} \\
\iff -\varepsilon I \preceq S_{x_0}^{-1}(S_{x_t} - S_{x_0}) \preceq \varepsilon I \\
\iff \|S_{x_0}^{-1}(s(x_t) - s(x_0))\|_\infty \leq \varepsilon
\]

But the \( \ell_\infty \) norm can be bounded by the \( \ell_2 \) norm,
\[ \|S_{x_0}^{-1}(s(x_t) - s(x_0))\|_\infty \leq \|S_{x_0}^{-1}(s(x_t) - s(x_0))\|_2 = \|S_{x_0}^{-1}A(x_t - x_0)\|_2 \]

by the definition of the slack vectors \( s(x) = Ax - b \). Then

\[
\|S_{x_0}^{-1}A(x_t - x_0)\|_2 = t \cdot \|S_{x_0}^{-1}A(\nabla^2 f(x_0))^{-1}\nabla f(x_0)\|_2 = t \cdot \|\nabla f(x_0)\|((\nabla^2 f(x_0))^{-1} = t \cdot \delta_k(x_0)
\]

which is exactly the centrality of \( x_0 \), which we assumed to be \( \leq 1/3 \). Thus we can take \( \varepsilon = 1/3 \) in Theorem 1.

### 4.3 Incrementing \( \lambda \)

We have just shown that given a coarsely central point for \( \lambda_k \), we can make it finely central in \( O(1) \) Newton steps. Now, assuming that \( \delta_k(\tilde{x}_k) < 1/100 \), we want to find the largest step \( \lambda_{k+1} \) such that \( \tilde{x}_k \) is still coarsely central for the new \( \lambda_{k+1} \), i.e. \( \delta_{k+1}(\tilde{x}_k) < 1/3 \).

Suppose we perform a multiplicative increase \( \lambda_{k+1} = (1 + \alpha)\lambda_k \). We want to make \( \alpha \) as large as possible while maintaining coarse centrality. If \( \tilde{x} \) is finely central for \( \lambda_k \), what can we say about the centrality for \( \lambda_{k+1} \)?

\[
\delta_{k+1}(\tilde{x}) = \|\nabla f_{\lambda_{k+1}}(\tilde{x})\|((A^T S_{x_2}^{-1} A)^{-1} = \|\lambda_k (1 + \alpha)c - A^T S_{x_2}^{-1}1\|((A^T S_{x_2}^{-1} A)^{-1} = \|(1 + \alpha)(\lambda_k c - A^T S_{x_2}^{-1}1) + \alpha A^T S_{x_2}^{-1}1\|((A^T S_{x_2}^{-1} A)^{-1} \leq (1 + \alpha)\delta_k(\tilde{x}) + \alpha\|A^T S_{x_2}^{-1}1\|((A^T S_{x_2}^{-1} A)^{-1}
\]

by the triangle inequality, the left side term just being the original centrality \( \delta_k(\tilde{x}) \leq 1/100 \). To bound the other term, we have

\[
\|A^T S_{x_2}^{-1}1\|((A^T S_{x_2}^{-1} A)^{-1} = 1^T S_{x_2}^{-1} A((A^T S_{x_2}^{-1} A)^{-1} A^T S_{x_2}^{-1}1
\]

In fact, the matrix between the two \( 1 \)'s is an orthogonal projection onto the column space of \( S_{x_2}^{-1} A \).

It follows that the second error term is at most \( \sqrt{1^T 1} = \sqrt{3} \) in absolute value, which implies

\[
\delta_{k+1}(\tilde{x}) \leq (1 + \alpha)\frac{1}{100} + \alpha \sqrt{3}
\]

Choosing \( \alpha = O(1/\sqrt{3}) \) we are done. With \( L \) defined as in last lecture, recall we said we will start with some \( \lambda_0 \) which is initially \( \exp(-CL) \) and can end when \( \lambda \) is at least \( m \exp(CL) \). Thus we need \( \lambda_k = (1 + \alpha)^k \lambda_0 \) to be at least \( m \exp(CL) \) which is satisfied for \( k = O(\sqrt{m}(L + \log m)) \), the number of iterations in our final IPM.
5 Initialization

It remains to find a coarsely or finely central initial point for some \( \lambda_0 \) to begin with. First, alter the original LP by adding a free variable \( z \):

Minimize \( c^T x + N z \) s.t. \( Ax + 1z \geq b, \ 0 \leq z \leq 2^{L+1}, \ -2^{L+1} \leq x \leq 2^{L+1} \).

It is known that if \( N \geq \exp( CL ) \) then original LP bounded and feasible \( \implies \) new LP’ has optimum with \( z = 0 \) (and this will give an optimum for the original LP). See [LS13, Lemma 41] for details of the proof. Thus, to solve the original LP it suffices to solve LP’.

LP’ has an obvious interior point: \( x' = (x, z) \) where \( x = 0, z = \|b\|_{\infty}, \) but we also need it to be coarsely central for some starting \( \lambda_0 \) to apply the IPM mentioned above. Set \( \lambda = 1, \) and note

\[
\nabla f_1(x') = c - A^T S^{-1} x'.
\]

Therefore \( x' \) is perfectly central for the modified cost vector \( c' = A^T S^{-1} x' \). Now run the IPM mentioned in this lecture in reverse using this new cost \( c' \), gradually decrementing \( \lambda \) instead of incrementing it, making \( \lambda \) small enough that our solution is essentially independent of the cost function (i.e. so that we are very near the analytic center of the feasible region) Once we have an approximate analytic center \( \tilde{x} \), this can be used as our starting solution for a very small \( \lambda = \exp(-CL) \) with the orginal cost function \( c \).

References
