CS 224: Advanced Algorithms Spring 2017

Lecture 21 — April 4, 2017

Prof. Jelani Nelson *Scribe: Matthew Deutsch*

1 Overview

Today we'll be going over the Path-following interior point method for solving LPs.

1.1 Recall...

We assume our LP is of the form:

$$\begin{aligned} &\min c^T x \\ &s.t. \quad Ax \geq b \end{aligned}$$

where $A \in \mathbb{R}^{m \times n}$ and $b, c \in \mathbb{R}^n$. We call the feasible polytope $P := \{x : Ax \geq b\}$. We also define the bit-complexity $L \propto 1 + \max \log(||b||_\infty), \log(||c||_\infty), \log(\text{detmax}(A))$, where $\text{detmax}(A)$ is the maximal determinant of any square submatrix of A.

We defined $f_\lambda(x) := \lambda c^T x + p(s(x))$ last week, with $s(x) = Ax - b$, interpreted as being the slacks / distances between x and the constraints / boundaries of P. This week we will explicitly define $p := -\sum_{i=1}^m \ln(s(x_i))$.

1.2 Birds-Eye View of Path Following Algorithm

Intuitively, the way path-following is going to work is like this: We have some polytope P. OPT lives at a vertex. When $\lambda = 0$, the cost function doesn't matter at all and the optimal value is at the center of our shape. If we imagine increasing λ continuously to infinity, we will draw a continuous path to OPT. We call this curve of optimal solutions to f_λ the "central path". The idea of path following is to follow this path to OPT.

We'll make sure we stay close to this path at all times in the algorithm, although we may not be completely lying on it. We'll be close (and we'll define "close" later). The actual curve is continuous, but we're discretizing it by polling it at finitely many λ's, which we'll call $\lambda_0, ..., \lambda_k$.

For some definition of "too fast", this central path doesn't change too fast. We need this so that we have a guarantee that the good solutions to λ_i are not too far away from the good solutions to λ_{i+1}. With this assumption, the way the algorithm will proceed is this:

Start with an Awesome solution to λ_0. Due to the central path not changing too quickly, this solution is already a Good solution for λ_1. We'll use calculus approximation methods in order to turn this Good solution into a new Awesome solution for λ_1. Repeat until rich.

But before we can formalize this we'll need some results from continuous optimization.

2 Continuous Optimization

This lecture will go over

- first order methods (e.g. gradient descent)
- second order methods (e.g. Newton's method. We'll actually use this one.)

2.1 Setup

We have some function f which maps $\mathbb{R}^n \to \mathbb{R}$. We want to find x^*, the minimizer. That is,

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x).$$

In a first-order method, we assume that we can consult an oracle that computes $f(x)$ as well as $\nabla f(x)$. In a second-order method, we also allow our oracle to compute the Hessian matrix, $\nabla^2 f(x)$. To jog your multivariable memory, recall

$$(\nabla^2 f(x))_{i,j} := \frac{\partial^2}{\partial_{x_i} \partial_{x_j}} f(x)$$

For our f_λ that we care about, the gradient is equal to

$$\nabla f_\lambda(x) = \lambda c - A^T S_x^{-1} \mathbf{1}$$

where $(S_x)_{i,j} = 0$ when $i \neq j$ and $(S_x)_{i,j} = s_i(x)$ when $i = j$, and $\mathbf{1}$ is the vector of all 1's.

And our Hessian matrix looks like:

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

3 Gradient Descent

We will start at $x = x_0$, and produce iterates x_k such that $f(x_k) \to f(x^*)$ as $k \to \infty$.

3.1 Idea

The idea is that we will approximate f by its first-order Taylor expansion. Then we will (kind of) minimize that approximated version of the function. We need to make some sort of assumption that the Hessian is controlled, so that our first-order approximation is roughly correct. Namely, we will assume

$$\forall x, \mu I \preceq \nabla^2 f(x) \preceq \beta I \tag{1}$$

Recall: $A \preceq B \iff B - a$ is PSD $\iff \forall z, z^T A z \leq z^T B z$.

3.2 First-order Taylor Expansion

We write

$$f(x_{k+1}) = f(x_k) + < \nabla f(x), x_{k+1} - x_k > + \int_0^1 \int_0^t \langle x_{k+1} - x_k, \nabla^2 f(x_\alpha)(x_{k+1} - x_k)\rangle d\alpha dt. \quad (2)$$

where $x_\alpha = (1-\alpha)x_k + \alpha x_{k+1}$. As it is, by Taylor's theorem this is exact. We note, though that the quantity inside the integral is of the form $u^T \nabla^2 f(x) u$, which allows us to use (1). So this integral is at most $\beta/2 \cdot ||x_{k+1} - x_k||_2^2$. So

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x), x_{k+1} - x_k \rangle + \frac{\beta}{2}||x_{k+1} - x_k||_2^2 \quad (3)$$

And the same logic, with the other side of (1) allows us to conclude that for any $x, y \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{\mu}{2}||y - x||_2^2. \quad (4)$$

3.3 Gradient Descent, the algorithm

Gradient descent works as follows: if the current point is $x = x_k$, we will set x_{k+1} to minimize the RHS of (3). After some calculus/algebra, this sets $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$. Now if we substitute back into (3), we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \quad (5)$$

This is fine and all, but not really what we want to prove. We can show all kinds of relationships between x_k and x_{k+1} until the cows come home, but what we really want to show is that as k gets larger, these x_k approach x^*. In order to get an equation involving x^*, we start by minimizing both sides of the inequality (4):

$$\begin{aligned} f(x^*) &\geq \min_y f(x_k) + \langle \nabla f(x_k), y - x \rangle + \frac{\mu}{2}\|x_k - y\|_2^2 \\ &= f(x_k) - \frac{1}{2\mu}\|\nabla f(x_k)\|_2^2. \end{aligned}$$

and therefore

$$\|\nabla f(x_k)\|_2^2 \geq 2\mu(f(x_k) - f(x^*)). \quad (6)$$

Combining (5) with (6),

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq f(x_k) + \frac{\mu}{\beta}(f(x_k) - f(x^*)) - f(x^*) \\ &= (1 - \frac{\mu}{\beta})(f(x_k) - f(x^*)) \end{aligned}$$

which is exactly what we wanted to show: as k gets larger, the difference between x_k and x^* lowers by a factor of $(1 - \frac{\mu}{\beta})$. In other words, gradient descent reduces the error by a constant factor every $O(\beta/\mu)$ oracle calls. Not bad, but we can do better with Newton.

(Side note: this is not the fastest first-order algorithm we know of. There is an algorithm called "fast gradient descent", that can do it in $O(\sqrt{\beta/\mu})$ oracle calls.)

4 Newton's method

4.1 Birds eye view

For IPM, we will use a *second order* method, i.e. the oracle also allows us access to $\nabla^2 f(x)$ given any point $x \in \mathbb{R}^n$. Recall in our case

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

One can check that

$$\nabla f_\lambda(x) = \lambda c - A^T S_x^{-1} \mathbf{e},$$

where $S_x = diag(s(x))$ and $\mathbf{e}$ is the all ones vector in $\mathbb{R}^m$. Also

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

In our IPM algorithm, we will have some current iterate $x = x_k$ for minimizing $f = f_\lambda$ and will make $\nabla^2 f$ oracle calls. In Newton's method, we estimate by the second-order Taylor expansion

$$f(y) \approx f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2} \left\langle y - x_k, \nabla^2 f(x_k)(y - x_k) \right\rangle$$

and choose $y = x_{k+1}$ to minimize the RHS of the above, so

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

Note that unlike in the case of gradient descent above, we cannot hope that $\|\nabla^2 f_\lambda(x)\| \leq \beta$ everywhere. This is because as the slacks go to zero, the operator norm of the Hessian blows up. We will need a new assumption on our Hessian. We make the following assumption:

There exists an $\epsilon \in (0, 1)$, such that for all k, and for all α between 0 and 1,

$$(1 - \epsilon)\nabla^2 f(x_k) \preceq \nabla^2 f(x_\alpha) \preceq (1 + \epsilon)\nabla^2 f(x_k) \tag{7}$$

for all y of the form $x_k + t(x_{k+1} - x_k)$, where x_{k+1} is the next iterate

What does this condition mean? This lemma gives another way to think about it.

Lemma 1. *Suppose* $(1 - \epsilon)A \preceq B \preceq (1 + \epsilon)A$. *Then* $-\epsilon I \preceq A^{1/2} B A^{-1/2}$.

Proof of Lemma 1. Well, before we begin the proof, this lemma statement has some notation that we need to define properly. If we can write $A = Q\Lambda Q^T$ for some orthogonal Q and diagonal Λ (as is always possible in our case), then when we write $f(A)$ or $A^{1/2}$ we mean $Qf(\Lambda)Q^T$ or $Q\Lambda^{1/2}Q^T$, respectively.

The proof of this lemma is fairly straightforward:

$$(1 - \epsilon)A \preceq B \preceq (1 + \epsilon)A$$

if and only if

$$-\epsilon A \preceq B - A \preceq \epsilon A$$

which definitionally is equivalent to

$$\forall z, -\epsilon z^T A z \leq z^T (B - A) z \leq \epsilon z^T A z.$$

Let $q = A^{1/2} z$. Substituting into the above equation yields

$$\forall q, -\epsilon q^T q \leq q^T A^{1/2} (B - A) A^{-1/2} q \leq \epsilon q^T q$$

which is definitionally equivalent to

$$-\epsilon I \preceq A^{1/2} (B - A) A^{-1/2} \preceq \epsilon I$$

□

This somewhat arcane result will become bizarrely applicable in a few short moments. We define more notation in order to make provable statements about "closeness":

$$||x||_A = \sqrt{x^T A x} = ||A^{1/2} x||_2.$$

and specifically we define

$$\delta(x) = ||\nabla f(x)||_{(\nabla^2 f(x))^{-1}}.$$

Later when we finish with Newton and return to using it for interior point, we will use δ_λ to denote the case where $f = f_\lambda$.

The goal for the rest of this lecture is to show that under Newton's method, we always have

$$\delta(x) \leq \frac{\epsilon}{1 - \epsilon} \delta(x_k)$$

In other words, under our new, weird definition of "closer", the gradient is always getting closer to 0.

4.2 The Final Stretch

Step 1: Use the fundamental theorem of calculus:

$$\nabla f(x_{k+1}) = \nabla f(x_k) + \int_0^1 \nabla^2 f(x_\alpha) \cdot (x_{k+1} - x_k) d\alpha$$

The fundamental theorem of calculus is thankfully beyond the scope of this lecture to prove.

We'll rewrite the RHS in a funny way using that $\int_0^1 d\alpha = 1$:

$$\begin{aligned} \nabla f(x_{k+1}) &= \int_0^1 \nabla^2 f(x_k) (\nabla^2 f(x_k))^{-1} \nabla f(x_k) d\alpha - \int_0^1 \nabla^2 f(x_\alpha) (\nabla^2 f(x_k))^{-1} \nabla f(x_k) d\alpha \\ &= \int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x_\alpha)) (\nabla^2 f(x_k))^{-1} \nabla f(x_k) d\alpha \end{aligned}$$

Therefore

$$\underbrace{\|\nabla f(x_{k+1})\|_{(\nabla^2 f(x_{k+1}))^{-1}}}_{\delta(x_{k+1})} = \| \int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x_\alpha))(\nabla^2 f(x_k))^{-1} \nabla f(x_k) d\alpha \|_{(\nabla^2 f(x_{k+1}))^{-1}}$$

$$\leq \int_0^1 \|(\nabla^2 f(x_k) - \nabla^2 f(x_\alpha))(\nabla^2 f(x_k))^{-1} \nabla f(x_k)\|_{(\nabla^2 f(x_{k+1}))^{-1}} d\alpha$$

At this point we would like to replace the $(\nabla^2 f(x_{k+1}))^{-1}$-norm on the RHS with the $(\nabla^2 f(x_k))^{-1}$-norm. Equation (7) relates the $(\nabla^2 f(x_{k+1}))$-norm to the $(\nabla^2 f(x_k))$-norm, but how can we use that to relate the norms given by the inverse matrices? We use the following lemma (proof suggested by Michael Cohen).

Lemma 2. *Suppose P, Q are real, symmetric positive definite matrices. Then $P \preceq Q$ iff $PQ^{-1} \preceq I$. Thus in particular, $P \preceq Q$ iff $P^{-1} \succeq Q^{-1}$.*

Proof. We first prove the first claim. First, $P \preceq Q$ iff $Q^{-1/2} P Q^{-1/2} \preceq I$. We thus need to show $Q^{-1/2} P Q^{-1/2} \preceq I$ iff $PQ^{-1} \preceq I$. To say $M \preceq I$ means all eigenvalues of M are at most 1. Note eigenvalues do not change under a similarity transformation (that is, $X^{-1} M X$ has the same eigenvalues as M for X of full rank). But then, assuming either condition we can obtain the other condition by perform such a similarity transformation with $X = Q^{1/2}$.

For the second claim, $PQ^{-1} \preceq I$ iff $Q^{-1} P \preceq I$, implying $P \preceq Q$ iff $Q^{-1} \preceq P^{-1}$. □

Then by (7) and Lemma 2 we can change the norm on the RHS, yielding

$$\delta(x_{k+1}) \leq \frac{1}{1-\epsilon} \int_0^1 ||(\nabla^2 f(x_k) - \nabla^2 f(x_\alpha))(\nabla^2 f(x_k))^{-1} \nabla f(x_k)||_{(\nabla^2 f(x_k))^{-1}} d\alpha$$

The integrand on the RHS can be rewritten as

$$= \frac{1}{1-\epsilon} ||(\nabla^2 f(x_k))^{-1/2} \nabla f(x_k)||_{D^2}$$

where $D = (\nabla^2 f(x_\alpha))^{-1/2} (\nabla^2 f(x_k) - \nabla^2 f(x_\alpha)) (\nabla^2 f(x_k))^{-1/2}$. Well would you look at that, the matrix D is exactly of the form $A^{1/2}(B - A)A^{1/2}$ that we had in the lemma we had earlier! This is a very emotional moment. From the lemma, we have $-\epsilon I \preceq D \preceq \epsilon I \implies 0 \preceq D^2 \preceq \epsilon^2 I$. This allows us to bound D^2 by ϵI. But the I-norm is just the ℓ_2 norm, so our expression becomes:

$$\leq \frac{\epsilon}{1-\epsilon} \int_0^1 ||(\nabla^2 f(x_k))^{-1/2} \nabla f(x_k)||_2 d\alpha$$

which is equal to

$$\frac{\epsilon}{1-\epsilon} ||\nabla f(x_k)||_{(\nabla^2 f(x_k))^{-1}} = \frac{\epsilon}{1-\epsilon} \delta(x_k).$$

as desired. Whew.

We have thus shown the following theorem.

Theorem 4.1. *Let $U \subset \mathbb{R}^n$ be open and convex, and such that $f : U \to \mathbb{R}$ is twice differentiable with $\nabla^2 f(x)$ positive definite for all $x \in U$. Fix $x_0 \in \mathbb{R}^n$ and define inductively for $k \geq 0$, $x_{k+1} = x_k - (\nabla^2 f(x))^{-1} \nabla f(x_k)$. Suppose (7) holds for all $k \geq 0$. Then for all $k \geq 0$,*

$$\delta(x_{k+1}) \leq \frac{\epsilon}{1-\epsilon} \delta(x_k).$$