ORIE 6300: Mathematical Programming I **Fall 2016**

Lecture 21: 11/08

Lecturer: Damek Davis *Scribe: Angela Zhou*

21.1 Overview of IPMs from last time

The idea of interior point methods: given a primal-dual pair:

$$\min\{c^T x | Ax = b, x \geq 0\}$$
$$\max\{b^T y | A^T y + s = c, s \geq 0\}$$

form the primal dual system of C_1, C_2:

$$C_1 : \begin{bmatrix} A & 0 & 0 \\ 0 & A^T & I \end{bmatrix} \begin{bmatrix} x \\ y \\ s \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix}$$
$$C_2 : \begin{bmatrix} x \\ 0 \\ s \end{bmatrix} \geq 0$$
$$x^T(c - A^T y) = c^T x - b^T y = 0$$

This is different from the MAP/DRS setting because of the additional complementary slackness condition.

Then realize that $x^T s = (c - A^T y)$.

Note that IPMs solve a series of relaxed problems, P_v:

$$\left\{ \begin{bmatrix} x \\ y \\ s \end{bmatrix} \subset C_1 \cap C_2^\circ, x \odot s = (x_i \cdot s_i) = v, v > 0 \right\}$$

Suppose that we have a solution to P_v such that $\|v - \mu e\| < C\mu$, where C is some constant.

We want to find a point v_t so that $\|v_t - \mu_t e\| < C\mu_t$, and a solution to P_{v_t}.

Let $v' = \mu_t e$. Given a solution to P_v, called $[x, s]$, the best case is that we solve $v' = x' \odot s'$, $x' = x + \Delta x$, $s' = s + \Delta s$, $y' = y + \Delta y$, $A\Delta x = 0$, $A^T \Delta y + \Delta s = 0$.

It's difficult to solve this exactly, so we linearize instead:

$$x \odot \Delta s + \Delta x \odot s = v' - v$$
$$A\Delta x = 0$$
$$A^T \Delta y + \Delta s = 0$$

Then set $x_t = x + \Delta x$ and $s_t = s + \Delta s$, $y_t = y + \Delta y$. Observe that $v - v^t = \Delta x \odot \Delta s$.

21.2 Proof of main theorem

Theorem 21.1 *If $v' \in B(v, r(v))$, then (x_t, y_t, s_t) is feasible and*

$$\|v_t - v'\| \leq \frac{\|v' - v\|^2}{2r(v)} \tag{21.1}$$

Proof: We will first need some preliminaries:

- (1) $(\forall u, v \in \mathbb{R}^n)\ \|u \odot v\| \leq \|u\| \cdot \|v\|_\infty \leq \|u\|\,\|v\|$
- (2) Young's inequality: $(\forall a, b, \in \mathbb{R})$, $2ab \leq \frac{a^2}{2} + \frac{b^2}{2}$
- (3) $(\forall u \in \mathbb{R}^n_{>0})$, let $\sqrt{u} = (\sqrt{u_0})_{i=1}^n$, $u^{-1} = (u_i^{-1})_{i=1}^n$

Define: $\alpha_x = \sqrt{v}^{-1} s \odot \Delta x$

$\alpha_s = \sqrt{v}^{-1} x \odot \Delta s$

We now show $x + \delta x \in \mathbb{R}^n_{>0}$ and $s + \delta s \in \mathbb{R}^n_{>0}$. It suffices to show that $\left\|x^{-1} \odot \Delta x\right\| < 1$, $\left\|s^{-1} \odot \Delta s\right\| < 1$

Observe that $\alpha_x \circ \alpha_s = v^{-1} \circ (x \circ s) \circ (\Delta x \circ \Delta s) = \Delta x \odot \Delta s$ because $v = \Delta x \odot \Delta s$.

Moreover, $A\Delta x = 0, \Delta s = -A^T \Delta y$. This implies that the vectors are othogonal: $\Delta s^T \Delta x = \Delta y^T (A \Delta x) = 0$.

By (1) we know that $\alpha_x^T \alpha_s = \sum \Delta x_i \Delta s_i = 0$. So $\alpha_x \perp \alpha_s$.

We now prove the inequality:

$$\begin{aligned}
\|v_+ - v'\| &= \|\Delta x \odot \Delta s\| \\
&= \|\alpha_x \odot \alpha_s\| \\
&\leq \|\alpha_x\|\,\|\alpha_s\| \\
&\frac{1}{2}\|\alpha_x\|^2 + \|\alpha_s\|^2 \text{ by Young's inequality} \\
&\leq \frac{1}{2}\|\alpha_x + \alpha_s\|^2 \text{ by orthogonality, (3)} \\
&= \frac{1}{2}\left\|\sqrt{v}^{-1} \odot (s \odot \Delta x + \Delta s \odot x)\right\|^2 \\
&= \frac{1}{2}\left\|\sqrt{v}^{-1} \odot (v' - v)\right\|^2 \\
= \frac{1}{2}\|v' - v\|^2 \left\|\sqrt{v}^{-1}\right\|^2 &\leq \frac{1}{2r(v)}\|v' - v\|^2 \text{ by (2) (****)}
\end{aligned}$$

So we have the inequality and we want to show that the points are feasible in the end.

We now show that $x + \Delta x \in \mathbb{R}^n_{>0}$ and $s + \Delta s \in \mathbb{R}^n_{>0}$. It suffices to show that $\left\|x^{-1} \odot \Delta x\right\| < 1$, $\left\|s^{-1} \odot \Delta s\right\| < 1$. (Why? You should work this out for yourself- it's straightforward to show.)

Two other inequalities that we need: If I compute what's in this norm, $x^{-1} \odot \Delta x = (x \odot s)^{-1} \odot \sqrt{v} \odot (\sqrt{v}^{-1} \odot s \odot \Delta x) = \sqrt{v}^{-1} \alpha_x$, and $s^{-1} \odot \Delta s = \sqrt{v^{-1}} \alpha_s$

By Young's inequality (2),

$$\left\|x^{-1} \odot \Delta x\right\| \leq \frac{\alpha_x}{\sqrt{r(v)}}$$

because if you take the inverse of the square root it's clearly less than the minimum of the vector. As we showed in (3) and (****):

$$\begin{aligned}\max\{\|\alpha_x\|^2, \|\alpha_s\|^2\} &\leq \|\alpha_x\|^2 + \|\alpha_s\|^2 \\ &\leq \frac{\|v - v'\|^2}{r(v)}\end{aligned}$$

since the maximum of both elements is less than the sum of both.

So:

$$\left\|x^{-1} \odot \Delta x\right\| \leq \frac{\|v' - v\|}{r(v)} < 1$$

$$\left\|x^{-1} \odot \Delta x\right\| \leq \frac{\|v' - v\|}{r(v)} < 1$$

■

We're really lucky that we don't have to go through and open Nesterov and Nemirovski's book because people have gone through and simplified things.

What's really going on here is that we're applying Newton's method to some system of equations. It should be clear that you should be able to extend these methods to all of convex programming, using Newton's method.

21.3 Nonlinear optimization

Today I'm going to generalize the optimality conditions we have for linear programming to optimizing smooth functions

What does differentiability mean to you? That's my question. Why do we take derivatives, besides doing gradient descent?

You really want a very good linear approximation at that point: just slightly higher than first-order approximation.

Definition 21.2 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\bar{x} \in \mathbb{R}^n$ if $\exists v \in \mathbb{R}^n$ such that*

$$f(x) = f(\bar{x}) + \langle v, x - \bar{x}\rangle + o(x - \bar{x}) \tag{21.2}$$

where $o : \mathbb{R}^n \to \mathbb{R}$ is a function satisfying

$$\lim_{x \to \bar{x}} \frac{o(x - \overline{x})}{\|x - \overline{x}\|} = 0$$

and $o(0) = 0$.

So functions are differentiable if locally they can be locally well approximated by a linear function.

Can there be two such v that approximate f?

Proposition 21.3 (Uniqueness of derivative) *If $\exists v_1, v_2$ such that $\forall x \in \mathbb{R}^n, i = 1, 2$*

$$f(x) = f(\bar{x}) + \langle v_i, x - \bar{x}\rangle + o_i(\|x - \bar{x}\|)$$

Then $v_1 = v_2$.

Proof: Informally, linear approximations that are this good have to be unique.

What we will do is use the inner product and compare.

$$\begin{aligned}\langle v_1 - v_2, \frac{x-\bar{x}}{x-\bar{x}}\rangle &= \frac{(f(x) - f(\bar{x}) - o_1(x-\bar{x})) - (f(x) - f(\bar{x}) - o_2(x-\bar{x}))}{\|x-\bar{x}\|} \\ &= \frac{o_2(x-\bar{x}) - o_1(x-\bar{x})}{\|x-\bar{x}\|}\end{aligned}$$

We want to choose a path between some x and $\bar{x}$ which gives a good direction to compare with, i.e. a function x_ϵ which satisfies $x_\epsilon \to \bar{x}$ as $\epsilon \to 0$. Define

$$x_\epsilon := \bar{x} + \epsilon(v_1 - v_2)$$

.

We plug this x_ϵ into our inner product:

$$\|v_1 - v_2\| = \langle v_1 - v_2, \frac{v_1 - v_2}{\|v_1 - v_2\|}\rangle = \langle v_1 - v_2, \frac{x_\epsilon - \bar{x}}{\|x_\epsilon - \bar{x}\|}\rangle \to 0 \text{ as } \epsilon \to 0$$

where we use the identity $\|x_\epsilon - \bar{x}\| = \epsilon \|v_1 - v_2\|$. Thus, $v_1 = v_2$.

■

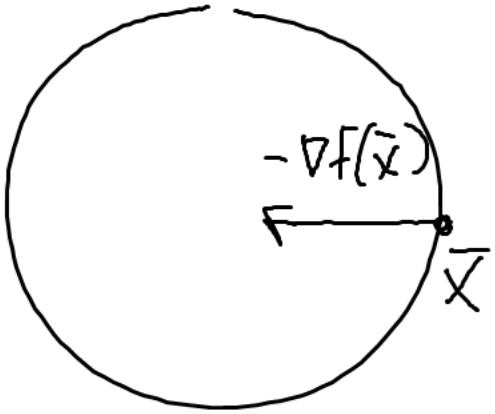
Definition 21.4 *We write $\nabla f(x) = v$ whenever* (21.2) *holds.*

The optimality conditions we saw in linear programming were necessary and sufficient. They are no longer sufficient, but we prove necessity.

Theorem 21.5 (Necessary Optimality Conditions) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, let $C \subseteq \mathbb{R}^n$ be closed and convex. Suppose that $\bar{x} \in \operatorname{argmin}_{x \in C} f(x)$, and that f is differentiable at $\bar{x}$. Then*

$$-\nabla f(\bar{x}) \in N_C(\bar{x})$$

It's easier to understand this via a picture:



Proof: For $x \in C$ we have

$$\langle -\nabla f(\bar{x}), x - \bar{x} \rangle = f(x) - f(\bar{x}) + o(x - \bar{x}) \leq o(x - \bar{x})$$

By definition of the normal cone we then just need to show that this inner product is negative.

Suppose, by way of contradiction, that $\exists \bar{x} \in C$ such that $\langle -\nabla f(\bar{x}), \hat{x} - \overline{x} \rangle > 0$.

Let $\hat{x}_\epsilon = \epsilon \hat{x} + (1 - \epsilon)\overline{x} \in C$.

Note that we have convergence $\hat{x}_\epsilon \to \overline{x}$ as $\epsilon \to 0$, trivially.

Thus, just as in the previous proof: by definition of $\hat{x}$,

$$\begin{aligned} 0 &< \langle -\nabla f(\bar{x}), \frac{\hat{x} - \overline{x}}{\|\hat{x} - \overline{x}\|} \rangle \\ &= \langle -\nabla f(\bar{x}), \frac{\hat{x}_\epsilon - \overline{x}}{\|\hat{x}_\epsilon - \overline{x}\|} \rangle \\ &\leq \frac{o(\hat{x}_\epsilon - \overline{x})}{\|\hat{x}_\epsilon - \overline{x}\|} \to 0 \text{ as } \epsilon \to 0 \end{aligned}$$

This is a contradiction. $\implies -\nabla f(x) \in N_C(\bar{x})$. ■

21.4 Examples

Example 1: Projections

If you're given some $x_0 \in \mathbb{R}^n$ and you want to compute the projection of x_0 onto the set C,

$$P_C(x_0) = \operatorname*{argmin}_{x \in C} \left\{ \frac{1}{2} \|x - x_0\|^2 \right\}$$

If we denote $f(x) = \|x - x_0\|^2$ what is the gradient? $\nabla f(x) = x - x_0$. Recall that our projection inclusion formula said that we had to satisfy $x_0 - P_C(x_0) \in N_C(P_C(x_0))$.

So the projection inclusion formula is a consequence of the optimality conditions.

21.4.1 Crazy nonlinear thing

$f(x) = \sin(x)$, $C = [0, \pi]$, $\nabla f(x) = \cos x$. Recall that

$$N_C(x) = \begin{cases} 0 & x \in (0, \pi) \\ \mathbb{R}_{\geq 0} & x = \pi \\ \mathbb{R}_{\leq 0} & x = 0 \end{cases}$$

Observe that $\cos(x) = 0 \iff x = \frac{\pi}{2}$. Observe that $\pi/2$ is a stationary point since $-\cos(\frac{\pi}{2}) \in N_C(\frac{\pi}{2})$.

We can also check the endpoints: $-\cos(0) = -1 \in N_C(0)$ and $-\cos(\pi) = -1 \in N_C(\pi)$.

Nonnegative matrix factorization

Say that you'd like to factor a matrix into the product of two different matrices. In other words, given $A \in \mathbb{R}^{m \times n}$ we want to find $X \in \mathbb{R}^{k \times m}, Y \in \mathbb{R}^{k \times m}$ such that and $X \geq 0, Y \geq 0$, and $A = X^T Y$.

A common loss function is the squared error.

$$f(X, Y) = \frac{1}{2} \left\| X^T Y - A \right\|^2$$

and if you take the matrix derivative:

$$\nabla f(X, Y) = [Y(X^T Y - A)^T, X(X^T Y - A)]$$

To find the global optimum, we use the fact that the minimizer (X^*, Y^*) satisfies the following:

$$-[Y^*((X^*)^T Y^* - A)^T, X^*((X^*)^T Y^* - A)] \in N_{\mathbb{R}^{m \times k}_{\geq 0}}(X^*) \times N_{\mathbb{R}^{m \times k}_{\geq 0}}(Y^*)$$

So here we have a cubic system of inequalities...you will end up with a lot of stationary points! If you can fit the data exactly that's great, it'll be the stationary point.