ORIE 6300 Mathematical Programming I November 10, 2016

Lecture 22

Lecturer: Damek Davis *Scribe: Xueyu Tian*

1 Last Time

1. Finished up IPMs.

2. Defined differentiability at $\bar{x}$

$$(\exists v \in \mathbb{R}^n)(\forall x \in \mathbb{R}^n)\ f(x) = f(\bar{x}) + < v, x - \bar{x} > + o(x - \bar{x}) \tag{1}$$

where:

$$o : \mathbb{R}^n \to \mathbb{R} \text{ s.t. } \lim_{x \to \bar{x}} \frac{1}{||x - \bar{x}||} o(x - \bar{x}) = 0;\ o(0) = 0$$

Proposition 1 *Precisely, one vector can satisfy* (1)*. We write* $\nabla f(\bar{x}) := v$*, whenever* f *is differentiable at* $\bar{x}$*.*

3. Necessary optimality conditions.

Theorem 2 (***necessary optimality***) *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a function, let* $C \subseteq \mathbb{R}^n$ *be a closed convex set. Suppose that* $\bar{x} \in \underset{x \in C}{\operatorname{argmin}}\ f(x)$ *exists and* f *is differentiable at* $\bar{x}$*. Then*

$$-\nabla f(\bar{x}) \in N_c(\bar{x}) \tag{OPT}$$

2 Today

Definition 1 *Points* $\bar{x}$ *satisfying* (OPT) *are called stationary points.*

Observation 1 *Minimizers of* $\inf_{x \in C} f(x)$ *are stationary points, but stationary points are not necessarily minimizers.*

For nonconvex optimization problems, stationary points are all we can hope to find. Later we will see that stationary points of convex problems are global minimizers. So how can we find stationary points?

Theorem 3 $\bar{x}$ *satisfies OPT if, and only if,*

$$(\forall \gamma > 0) \qquad \bar{x} = P_C(\bar{x} - \gamma \nabla f(\bar{x}))$$

.

Proof:

$$\begin{aligned} -\nabla f(\bar{x}) \in N_C(\bar{x}) &\Leftrightarrow -\gamma\nabla f(\bar{x}) \in \gamma N_C(\bar{x}) = N_C(\bar{x}) \\ &\Leftrightarrow (\bar{x} - \gamma\nabla f(\bar{x})) - \bar{x} \in N_C(\bar{x}) \\ &\Leftrightarrow \bar{x} = P_C(\bar{x} - \gamma\nabla f(\bar{x})). \end{aligned}$$

□

Thus, stationary points are exactly elements of $\text{Fix}(P_C \circ (I - \gamma\nabla f))$. This suggests we apply KM iteration to $T = P_C \circ (I - \gamma\nabla f)$. However T is NOT nonexpansive in such a general setting. To get any result, we must restrict to f which have globally Lipschitz continuous gradients.

Algorithm 1 Projected Gradient for f with ∇f L-Lipschitz.

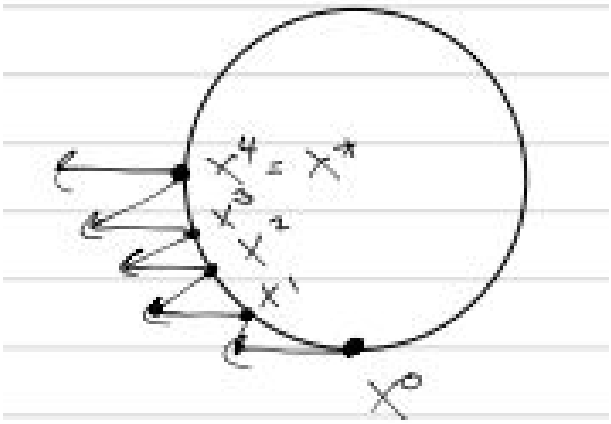
Input: $x \in C, 0 < \gamma < \frac{2}{L}$

1: **loop**

2: $\quad x \leftarrow P_C(x - \lambda\nabla f(x))$

Example 1

$$f(x) = \langle (1,0), x \rangle, \qquad C = B(0,1)$$



Theorem 4 *Suppose* $\inf_{x \in C} f(x) > -\infty$. *Let* $\{x^k\}_{k\in\mathbb{N}}$ *be generated by the projected gradient method. Suppose* $\bar{x}$ *is a limit point of* $\{x^k\}_{k\in\mathbb{N}}$. *Then* $\bar{x}$ *is a stationary point for OPT.*

Proof: Because f is Lipschitz differentiable, for all $k \in \mathbb{N}$, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}||x^{k+1} - x^k||^2 \\ &= f(x^k) + \frac{1}{\gamma}\langle x^{k+1} - (x^k - \gamma\nabla f(x^k)), x^{k+1} - x^k \rangle - \left(\frac{1}{\gamma} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 \end{aligned} \tag{2}$$

Notice that,

$$x^{k+1} = P_C(x^k - \gamma\nabla f(x^k)) \Leftrightarrow (x^k - \gamma\nabla f(x^k)) - x^{k+1} \in N_C(x^{k+1}).$$

Thus,

$$\frac{1}{\gamma}\langle (x^k - \gamma\nabla f(x^k)) - x^{k+1}, x^k - x^{k+1}\rangle \leq 0.$$

Therefore,

$$f(x^{k+1}) \leq (2) \leq f(x^k) - \left(\frac{1}{\gamma} - \frac{L}{2}\right)||x^{k+1} - x^k||^2.$$

The Fixed-Point Residual (FPR) is then summable,

$$\begin{aligned}\sum_{k=0}^{\infty}||x^k - x^{k+1}||^2 = \lim_{T\to\infty}\sum_{k=0}^{T}||x^k - x^{k+1}||^2 &\leq \frac{1}{\frac{1}{\gamma} - \frac{L}{2}}\lim_{T\to\infty}[f(x^0) - f(x^T)] \\ &\leq \frac{1}{\frac{1}{\gamma} - \frac{L}{2}}[f(x^0) - f^*]\end{aligned}$$

where $f^* = \inf_{x\in C} f(x)$. Thus, $||x^{k+1} - x^k|| \to 0$, as k$\to \infty$.

Suppose some subsequence $\{x^{j_k}\}_{k\in\mathbb{N}} \subseteq \{x^k\}_{k\in\mathbb{N}}$ converges to $\bar{x}$. Then, by continuity, $P_C(x^{jk} - \gamma\nabla f(x^{jk})) \to P_C(\bar{x} - \gamma\nabla f(\bar{x}))$. Moreover, because $x^{k_j} - x^{k_j+1} \to 0$, we have $x^{k_j+1} \to \bar{x}$. So

$$\bar{x} = \lim_{k\to\infty} x^{k_j+1} = \lim_{k\to\infty} P_C(x^{k_j} - \gamma\nabla f(x^{k_j})) = P_C(\bar{x} - \gamma\nabla f(\bar{x})).$$

Thus, $\bar{x}$ is stationary. □

Corollary 5 *For all $T \in \mathbb{N}$, we have*

$$\min_{k=0,1,\ldots,T}||x^{k+1} - x^k|| \leq \sqrt{\frac{f(x^0) - \inf_{x\in C} f(x)}{T\left(\frac{1}{\gamma} - \frac{L}{2}\right)}}$$

Proof:

$$\min_{k=0,1,\ldots,T}||x^{k+1} - x^k||^2 \leq \frac{1}{T}\sum_{k=0}^{T}||x^k - x^{k+1}||^2 \leq \frac{f(x^0) - \inf_{x\in C} f(x)}{T\left(\frac{1}{\gamma} - \frac{L}{2}\right)}$$

which implies the result. □

Significance of the convergence rate. The steplength $x^{k+1} - x^k$ satisfies

$$\begin{aligned}\frac{1}{\gamma}(x^k - x^{k+1}) \in N_C(x^{k+1}) + \nabla f(x^k) &= N_C(x^{k+1}) + \nabla f(x^{k+1}) + \nabla f(x^k) - \nabla f(x^{k+1}) \\ \implies \frac{1}{\gamma}(x^k - x^{k+1}) + (\nabla f(x^{k+1}) - \nabla f(x^k)) &\in N_C(x^{k+1}) + \nabla f(x^{k+1})\end{aligned}$$

So it is a natural measure of stationarity.

Using Lipschitz continuity of ∇f, we can easy get rates on

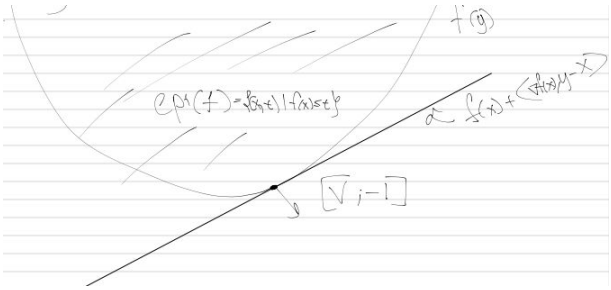
$$\min_{k=0,1,\ldots,T} ||(x^k - x^{k+1}) + \nabla f(x^{k+1}) - \nabla f(x^k)|| \leq \min_{k=0,1,\ldots,T} (1+L)||x^k - x^{k+1}||$$
$$\leq (1+L)\sqrt{\frac{f(x^0) - \inf_{x \in C} f(x)}{T(\frac{1}{\gamma} - \frac{L}{2})}}$$

This is a pretty terrible rate, but with convexity, we can get MUCH faster rates, and much stronger guarantees. Let's play with convexity for a bit first.

Definition 2 *A continuously differentiable function f on $\mathbb{R}^n$ (notation $f \in \mathfrak{F}(\mathbb{R}^n)$) is called convex if*

$$(\forall x, y \in \mathbb{R}^n)\ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{3}$$

The picture to have in mind is the following



Proposition 6 *Convexity is equivalent to the following*

$$(\forall x, y \in \mathbb{R}^n)(\alpha \in [0,1]) \qquad f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y)$$

Proof: See thm 2.1.2 from Nesterov's book. □

For convex functions, stationary points are global minimizers.

Notation. We let $\mathcal{F}(\mathbb{R}^n)$ denote the set of continuously differentiable, convex functions on $\mathbb{R}^n$.

Theorem 7 (Sufficient Optimality Conditions) *Let $f \in \mathcal{F}(\mathbb{R}^n)$ and let $C \in \mathbb{R}^n$ be a closed convex set. Then*

$$-\nabla f(x) \in N_C(x) \Leftrightarrow x \in \operatorname*{argmin}_{y \in C} f(y)$$

Proof: We proved necessity last time.

Let's prove sufficiency. Suppose $-\nabla f(x) \in N_C(x)$. Then

$$(\forall y \in C)\ \langle -\nabla f(x), y - x \rangle \leq 0 \Rightarrow\ \langle \nabla f(x), y - x \rangle \geq 0.$$

Thus,

$$(\forall y \in C)\ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \geq f(x)$$

□

When $C = \mathbb{R}^n$, so all stationary points satisfy $\nabla f(x) = 0$.

Example 2 (Linearly constrained optimization) *Let* $C = \{x \mid Ax = b\}$. *Then*

$$N_C(x) = \begin{cases} A^T\mathbb{R}^n & \text{if } Ax = b; \\ \emptyset & \text{if otherwise.} \end{cases}$$

The optimality conditions of the problem $\inf_{x \in C} f(x)$ *become:*

$$\begin{aligned} \exists y \text{ s.t. } \nabla f(x) + A^T y &= 0 \\ Ax &= b \end{aligned}$$

If f is convex, these conditions are necessary and sufficient.

The conditions should remind you of Lagrange multipliers. What about nonsmooth convex functions? Consider the minimization problem.

$$\min \ f(x) + g(x)$$

where g is nonsmooth, but convex. Can we reformulate it into a form that the projected gradient method applies to?
Yes!

$$\min\{f(x) + t\}$$

$$(x, t) \in \text{epi}(g) = \{x| \ g(x) \leq t\}$$

epi(g) is a closed, convex set and $f(x) + t$ is a smooth function.

Our optimality conditions become

$$\begin{bmatrix} -\nabla f(\bar{x}) \\ -1 \end{bmatrix} \in N_{\text{epi}(g)}(\bar{x}, t).$$

Notice that if g is a continuous function, then $g(x) < t \Rightarrow \text{int}(\text{epi}(g)) \implies N_{\text{epi}(g)}(x, t) = 0$.
Thus, optimality must occur at $(\bar{x}, t) = (\bar{x}, g(\bar{x}))$. Then from homework 3, problem 3, we find that

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

i.e.,

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x})$$

where $\partial g(x) = \{v|(\forall x) \ g(x) \geq g(\bar{x}) + \langle v, x - \bar{x}\rangle\}$. Thus

Theorem 8 *Suppose, f and g are convex. Then*

$$\bar{x} \in \text{argmin}\{f(x) + g(x)\} \Leftrightarrow \ 0 \in \nabla f(\bar{x}) + \partial g(\bar{x}).$$