ORIE 6300 Mathematical Programming I	November 2, 2016

Recitation 11

Lecturer: Calvin Wylie | *Topic: Woo-Hyung Cho*

Smooth Convex Optimization [1]

Recall that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if $\forall x, y$ and $t \in [0, 1]$, $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$

Lemma 1 *If f is convex and continuously differentiable, then $f(x) + \langle \nabla f(x), y - x\rangle \leq f(y)$ $(\forall x, y)$.*

Proof: Let $x, y \in \mathbb{R}^n$ and $t \in [0, 1)$. Let $x_t = tx + (1-t)y$.

If f is convex, $f(x_t) \leq tf(x) + (1-t)f(y)$. Since $t \neq 1$, we can divide by $1-t$:

$$
\begin{aligned}
f(y) &\geq \frac{1}{1-t}(f(x_t) - tf(x)) \\
&= f(x) + \frac{1}{1-t}(f(x_t) - f(x)) \\
&= f(x) + \frac{1}{1-t}(f(x + (1-t)(y-x)) - f(x))
\end{aligned}
$$

Let $t \to 1$, then $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$.

Proof in the reverse direction is left to the readers as an exercise.

$\square$

Lemma 2 *If f is convex and continuously differentiable, then $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$ $(\forall x, y)$, i.e., ∇f is monotone.*

Proof: If Lemma 1 holds, Lemma 2 holds. We only prove one direction.

Let $x, y \in \mathbb{R}^n$. By Lemma 1,

$$
f(x) + \langle \nabla f(x), y - x\rangle \leq f(y) \tag{1}
$$
$$
f(y) + \langle \nabla f(y), x - y\rangle \leq f(x) \tag{2}
$$

Add (1) and (2) to get $\langle \nabla f(x), y - x\rangle + \langle \nabla f(y), x - y\rangle \leq 0$. Then $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$.

$\square$

Theorem 3 *If f is continuously differentiable (but not necessarily convex), and ∇f is L-Lipschitz, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, then*

$$
|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|^2
$$

$\phi_1(x) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$ is an upper bound on f. Likewise, $\phi_2(x) = f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y - x\|^2$ offers a lower bound.

[1] Based on Nesterov, Yurii. *Introductory lectures on convex optimization: A basic course.*

Proof: By the fundamental theorem of calculus,

$$
\begin{aligned}
f(y) - f(x) &= \int_0^1 \frac{d}{dt} f(x + t(y-x))dt \\
&= \int_0^1 \langle \nabla f(x + t(y-x)), y - x\rangle dt
\end{aligned}
$$

Then,

$$
\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &= |\int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle dt| \\
&\leq \int_0^1 |\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle| dt \\
&\leq \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\| \|y - x\| dt \quad (Cauchy-Schwarz) \\
&\leq \int_0^1 L\|\nabla x + t(y-x) - x\| \|y - x\| dt \quad (Lipschitz) \\
&= L\|y-x\|^2 \int_0^1 t dt \\
&= \frac{L}{2}\|y-x\|^2
\end{aligned}
$$

□

Theorem 4 *Let f be convex and continuously differentiable. Then x^* is a global minimizer of f iff $\nabla f(x^*) = 0$.*

Proof: ($\leftarrow$) The proof follows immediately from Lemma 1: $f(x^*) + \langle \nabla f(x^*), y - x^*\rangle \leq f(y)$ $(\forall y)$. If $\nabla f(x^*) = 0$, then $f(x^*) \leq f(y)$ $(\forall y)$. Hence, $f(x^*)$ is the global minimizer.

($\rightarrow$) For a proof by contradiction, suppose $\nabla f(x^*) \neq 0$ and let $d = -\nabla f(x^*)$. Then $\langle d, \nabla f(x^*)\rangle < 0$.
Now recall the mean value theorem: $(\forall x, y \in \mathbb{R}^n)$ $f(y) = f(x) + \langle \nabla f(x + t(y-x)), y - x\rangle$ for some $t \in (0, 1)$. Since ∇f is continuous, $\langle \nabla f(x^* + td), d\rangle < 0$ $(\forall\ 0 \leq t \leq T)$ for some T. For $\bar{t} \in [0, T]$, $f(x^* + \bar{t}d) = f(x^*) + \langle \nabla f(x^* + t\bar{t}d), \bar{t}d\rangle$ holds for some $t \in (0, 1)$, where $\langle \nabla f(x^* + t\bar{t}d), \bar{t}d\rangle < 0$. This shows that x^* is not a global minimizer, and we have a contradiction. Therefore, $\nabla f(x^*) = 0$. □

Theorem 5 *Let f be convex and continuously differentiable. Let ∇f be L-Lipschitz continuous. Then*

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle$$

This is called the "co-coercivity condition."

Proof: Let $y \in \mathbb{R}^n$ and define $g(x) = f(x) - \langle \nabla f(y), x\rangle$. Note that $\nabla g(y) = \nabla f(y) - \nabla f(y) = 0$, i.e., y minimizes g. Because $g(y) \leq g(\cdot)$, $g(y) \leq g(x - \frac{1}{L}\nabla g(x))$ also holds $\forall x$. We apply Theorem 3.

$$
\begin{aligned}
g(x - \frac{1}{L}\nabla g(x)) &\leq g(x) + \langle \nabla g(x), -\frac{1}{L}\nabla g(x)\rangle + \frac{L}{2}\| - \frac{1}{L}\nabla g(x)\|^2 \\
&= g(x) - \frac{1}{L}\|\nabla g(x)\|^2 + \frac{1}{2L}\|\nabla g(x)\|^2 \\
&= g(x) - \frac{1}{2L}\|\nabla g(x)\|^2
\end{aligned}
$$

We use the definition $g(x) = f(x) - \langle \nabla f(y), x\rangle$ and the inequality $g(y) \leq g(x) - \frac{1}{2L}\|\nabla g(x)\|^2$ to derive

$$
f(y) - \langle \nabla f(y), y\rangle - f(x) + \langle \nabla f(y), x\rangle \leq -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2
$$

Interchanging x and y,

$$
f(x) - \langle \nabla f(x), x\rangle - f(y) + \langle \nabla f(x), y\rangle \leq -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2
$$

We add the two inequalities to get

$$
\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle
$$

□

Corollary 6 $I - \frac{2}{L}\nabla f$ *is non-expansive.*

Proof:

$$
\begin{aligned}
\|(x - \frac{2}{L}\nabla f(x)) - (y - \frac{2}{L}\nabla f(y))\|^2 &= \|(x-y) - \frac{2}{L}(\nabla f(x) - \nabla f(y))\|^2 \\
&= \|x-y\|^2 + \frac{4}{L^2}\|\nabla f(x) - \nabla f(y)\|^2 - \frac{4}{L}\langle x-y, \nabla f(x) - \nabla f(y)\rangle \\
&= \|x-y\|^2 + \frac{4}{L}(\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 - \langle x-y, \nabla f(x) - \nabla f(y)\rangle) \\
&\leq \|x-y\|^2 \quad \text{by Theorem 5}
\end{aligned}
$$

□

A KM iteration

$$
\begin{aligned}
x^{k+1} &= \frac{1}{2}(I - \frac{2}{L}\nabla f)(x^k) + \frac{1}{2}x^k \\
&= x^k - \frac{1}{L}\nabla f(x^k)
\end{aligned}
$$

performs a gradient descent with step size $\frac{1}{L}$. If we apply the KM algorithm iteratively, the sequence x^k converges to a fixed-point x^* such that $x^* = (I - \frac{2}{L}\nabla f)(x^*)$, which implies $\nabla f(x^*) = 0$. Steepest gradient descent converges to a minimizer when the step size is chosen between 0 and $\frac{2}{L}$.