| | |
|---|---|
| **Mathematical Programming II** | **Lecture 21** |
| **ORIE 6310 Spring 2014** | **April 15, 2014** |
| **Scribe: Jiayi Guo** | |

Recall the algorithm for minimizing a convex function $f$ over $B(0, R)$, where we assume $f$ has range at most one on this ball. Let $x_*$ minimize $f$ over the ball.

Start with $x_0 = 0$.

At iteration $k$, if $\|x_k\| > R$, choose $v_k = x_k$; otherwise compute a subgradient $g_k$ of $f$ at $x_k$, and stop if $g_k = 0$. Otherwise, $v_k = g_k$.

Set $x_{k+1} := x_k - \epsilon R \frac{v_k}{\|v_k\|}$.

**Theorem 1** *This algorithm generates a solution within $\epsilon$ of the minimum within $\epsilon^{-2}$ iterations.*

**Proof:** We have as before

$$\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 = \|x_{k+1} - x_k\|^2 + 2(x_{k+1} - x_k)^T(x_k - x_*)$$

In fact, we'll replace $x_*$ by $\bar{x} = (1 - \epsilon)x_* = (1 - \epsilon)x_* + \epsilon \cdot 0$, which is the center of the ball $\{x_*\} + \epsilon B(0, R) \subseteq B(0, R)$, and $f(x) \leq f(x_*) + \epsilon$ for all $x$ in this small ball. From the identity above, with $\bar{x}$ instead of $x_*$,

$$\|x_{k+1} - \bar{x}\|^2 - \|x_k - \bar{x}\|^2 = \epsilon^2 R^2 - \frac{2\epsilon R}{\|v_k\|} v_k^T (x_k - \bar{x}).$$

Note: $v_k^T(x_k - \bar{x}) = v_k^T(x_k - (\bar{x} + \frac{\epsilon R v_k}{\|v_k\|})) + \epsilon R \|v_k\|$:

1. if $x_k \notin B(0, R)$, $v_k^T(x_k - (\bar{x} + \frac{\epsilon R v_k}{\|v_k\|})) \geq 0$, because $v_k = x_k$ (separating hyperplane);
2. if $x_k \in B(0, R)$,

$$\begin{aligned} v_k^T(x_k - (\bar{x} + \frac{\epsilon R v_k}{\|v_k\|})) &\geq f(x_k) - f(\bar{x} + \frac{\epsilon R v_k}{\|v_k\|}) \\ &\geq f(x_k) - (f(x_*) + \epsilon) \\ &\geq 0 \quad \textbf{if } f(x_k) \geq f(x_*) + \epsilon. \end{aligned}$$

Hence, as long as $f(x_k) \geq f(x_*) + \epsilon$, we have

$$\|x_{k+1} - \bar{x}\|^2 - \|x_k - \bar{x}\|^2 \leq \epsilon^2 R^2 - 2\frac{\epsilon R}{\|v_k\|} \cdot \epsilon R \|v_k\| = -\epsilon^2 R^2.$$

So $\|x_l - \bar{x}\|^2 \leq \|x_0 - \bar{x}\|^2 - l\epsilon^2 R^2 \leq R^2 - l\epsilon^2 R^2$ as long as $\min_{0 \leq k \leq l} f(x_k) > f(x_*) + \epsilon$ whence $l \leq \epsilon^{-2}$. If $l = \epsilon^{-2}$ with all iterates up to $x_l$ having too large $f$, then $x_l = \bar{x}$ and $f(x_l) \leq f(x_*) + \epsilon$, a contradiction. So this inequality holds within $\epsilon^{-2}$ steps.
$\square$

This algorithm seems stupid, taking very small steps independent of $f$, but it is within a constant factor of optimal!

WLOG we'll assume $R = \frac{1}{2}$ and consider functions in the class

$$\mathcal{F}_0 := \{f(x) \equiv \max_{1 \le j \le M} \{\sigma_j(e_j^T x) + \tau_j : \sigma_j \in \{-1, +1\}, \tau_j \in (0, \delta)\}\}.$$

**Theorem 2** *Every first-order oracle algorithm to minimize convex functions $f$ with range at most 1 on $B(0, \frac{1}{2})$ with $n \ge \frac{1}{4\epsilon^2}$ will take at least $\lfloor \frac{1}{4\epsilon^2} \rfloor$ steps on some such function to get within $\epsilon$ of its minimum.*

**Proof:** Choose $M = \lfloor \frac{1}{4\epsilon^2} \rfloor - 1$, so $M \le n$ and $\delta := -\epsilon + \frac{1}{2\sqrt{M}} > 0$.

The algorithm generates $x_1 \in \Re^n$ independent of $f$. Let $j_1$ be the index of its largest component in absolute value.

Choose $\sigma_{j_1} = \bar{\sigma}_{j_1}$ so that $\sigma_{j_1} e_{j_1}^T x_1 \ge 0$ and choose $\tau_{j_1} = \bar{\tau}_{j_1} = \frac{\delta}{2}$.

Now consider

$$\mathcal{F}_k = \{f \in \mathcal{F}_0 : \sigma_{j_i} = \bar{\sigma}_{j_i}, \tau_{j_i} = \bar{\tau}_{j_i} \ \textbf{for} \ 1 \le i \le k, \ \textbf{and} \ \tau_j < \frac{\delta}{2^k} \ \textbf{for} \ j \notin \{j_1, .., j_k\}\}.$$

Note that $f(x_1)$ and $g(x_1) \in \partial f(x_1)$ are the same for all $f \in \mathcal{F}_1$. After $k-1$ iterations, we have $x_k$, and set $j_k$ to be the index of the largest in absolute value component of $x_k$, not in $\{j_1, ..., j_{k-1}\}$.

Choose $\sigma_{j_k} = \bar{\sigma}_{j_k}$ so that $\sigma_{j_k} e_{j_k}^T x_k \ge 0$ and $\tau_{j_k} = \bar{\tau}_{j_k} = \frac{\delta}{2^k}$.

After $M$ steps, $\mathcal{F}_M$ is just a single function.

For $1 \le k \le M$, $f(x_k) \ge \bar{\sigma}_{j_k}(e_{j_k}^T x_k) + \bar{\tau}_{j_k} \ge \frac{\delta}{2^M} > 0$.

But consider $\bar{x}$ with $e_j^T \bar{x} = \begin{cases} -\bar{\sigma}_j \frac{1}{2\sqrt{M}} & j = 1, 2.., M \\ 0 & ow. \end{cases}$

Then $\|\bar{x}\| = \frac{1}{2}$, so $\bar{x}$ lies inside $B(0, \frac{1}{2})$. Also $f(\bar{x}) = \max\{(-\frac{1}{2\sqrt{M}} + \tau_j)\} \le -\frac{1}{2\sqrt{M}} + \delta = -\epsilon$.

So we cannot have generated a solution within $\epsilon$ of the minimum, so we need at least $\lfloor \frac{1}{4\epsilon^2} \rfloor$ steps. □

The "stupid" algorithm always take steps of size $\epsilon R$, so needs to chose $\epsilon$ in advance. But if we choose $\lambda_k = \frac{R}{\sqrt{k+1}}$ (to satisfy the conditions of Polyak's convergence result) we also need only a little more than $O(\frac{1}{\epsilon^2})$ steps.

Also if $f_* = \min f(B(0, R))$ is known, in practice, $\lambda_k = \frac{f(x_k) - f_*}{\|g_k\|}$ is much better.

Now we turn to smooth convex functions.

We'll look at $C^{1,1}$ functions, i.e., continuously differentiable with Lipschitz continuous gradients.

For all $x, y \in \Re^n$, assume

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

**Proposition 1** *If $f$ as above is convex, then for all $x, y \in \Re^n$,*
$f(x) + \nabla f(x)^T(y - x) \leq f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}L\|y - x\|^2.$

**Proof:** The LH inequality follows by convexity. For the RH inequality,

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \nabla f(x + \lambda(y - x))^T(y - x)d\lambda \\
&= f(x) + \nabla f(x)^T(y - x) + \int_0^1 (\nabla f(x + \lambda(y - x)) - \nabla f(x))^T(y - x)d\lambda \\
&\leq f(x) + \nabla f(x)^T(y - x) + \int_0^1 \|\nabla f(x + \lambda(y - x)) - \nabla f(x)\| \cdot \|y - x\|d\lambda \\
&\quad \text{by Cauchy-Schwarz} \\
&\leq f(x) + \nabla f(x)^T(y - x) + L\|y - x\|^2 \int_0^1 \lambda d\lambda \\
&= f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}L\|y - x\|^2.
\end{aligned}$$

□

Note: given $x$, the RHS in the proposition is minimized by $y = x_+ = x - \frac{1}{L}\nabla f(x)$.

Remark: we will assume that $L$ is known so we can make this update. Extensions of this algorithm keep an estimate of $L$ and adjust it during the iterations.

If we update $x$ this way, then

$$\begin{aligned}
f(x_+) &\leq f(x) + \nabla f(x)^T(x_+ - x) + \frac{1}{2}L\|x_+ - x\|^2 \\
&= f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 \\
&= f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \\
&= f(x) - \frac{1}{2}L\|x_+ - x\|^2.
\end{aligned} \tag{1}$$

**Theorem 3** *If we turn the update into an algorithm for functions in this class with $\|x_0 - x_*\| \leq R$ for some $x_*$ minimizing $f$, then the algorithm produces an iterate within $\epsilon$ of the minimum in $\frac{LR^2}{\epsilon}$ steps.*