Mathematical Programming II	Lecture 20
ORIE 6310 Spring 2014	April 10, 2014
Scribe: James Dong	

1 Subgradient Method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Recall that the subdifferential of f is the set of subgradients:

$$\partial f(x) := \left\{ g \in \mathbb{R}^n \; : \; f(y) \geq f(x) + g^T (y - x), \; \forall \, y \in \mathbb{R}^n \right\}.$$

Assume that at each x, we can compute a *single* subgradient $g = g(x)$, e.g., as in the Lagrangian relaxation approach to hard problems.

While the subdifferential tells us a lot about the behavior of f around x, a single subgradient doesn't reveal very much.

Consider the directional derivative of f at x in direction d:

$$f'(x; d) := \lim_{\lambda \downarrow 0} \left\{ \frac{f(x + \lambda d) - f(x)}{\lambda} =: q(x, d, \lambda) \right\}.$$

Note that because q is nondecreasing in λ, the limit exists and $f'(x; d) = \inf_{\lambda > 0} \{ q(x, d, \lambda) \}$

Note also that $q(x, d, \lambda) \geq g^T d \; \forall \, g \in \partial f(x)$.

Hence, $f'(x; d) \geq g^T d \; \forall \, g \in \partial f(x)$, so that $f'(x; d) \geq \max \{ g^T d : g \in \partial f(x) \}$

In fact, it can be shown (see Borwein and Lewis, "Convex Analysis and Nonlinear Optimization: Theory and Examples") that

$$f'(x; d) = \max \{ g^T d : g \in \partial f(x) \}.$$

So in particular, if $g \in \partial f(x)$ and $g \neq 0$, then g is an ascent direction for f at x since $f'(x; g) \geq g^T g > 0$. However, $-g$ may not be a descent direction:
$f'(x; -g) = \max \{ (-g)^T h \; : \; h \in \partial f(x) \} \geq -g^T g < 0$ might be positive. See Figure 2 of Lecture 19.

Theorem 1 *x is a global minimizer of f if and only if $0 \in \partial f(x)$.*

Proof:

The "if" component is trivial through the definition of a subgradient.

We prove the "only if" portion through contradiction, "constructively." Suppose 0 is not in the closed convex set $\partial f(x)$. Then there is some $0 \neq d \in \mathbb{R}^n$ such that $0^T d = 0 > \max \{ g^T d \; : \; g \in \partial f(x) \} = f'(x; d)$.

So x is not even a local minimizer since d is a descent direction. $\square$

Corollary 1 *If $0 \notin \partial f(x)$, then $d = -\arg\min\{\|g\| \ : \ g \in \partial f(x)\}$ is a descent direction for f at x.*

2 Two Alternatives

1. Try to find more points in $\partial f(x)$ until we can get a descent direction:
   (a) Bundle methods (Lemaréchal, 1970s),
   (b) Gradient Sampling Methods (Burke, Lewis, Overton).
2. Move in the direction $-g$ even if it is not a descent direction.

We will show how to choose a very general step size rule that is independent of f.

General Subgradient Algorithm (N.Z. Shor, 1960s)

Choose $x_0 \in \mathbb{R}^n$ and a sequence $\{\lambda_k\}$ of positive scalars. At iteration k, compute a subgradient g_k of f at x_k. Stop if $g_k = 0$. Otherwise, set $x_{k+1} := x_k - \lambda_k \frac{g_k}{\|g_k\|}$.

Theorem 2 *(B.T. Polyak, 1967) Suppose $X_* := \{x_* \in \mathbb{R}^n \ : \ f(x) \geq f(x_*) \ \forall \ x \in \mathbb{R}^n\} \neq \emptyset$. Then, as long as $\sum_{k=0}^{\infty} \lambda_k = \infty$ and $\sum_{k=0}^{\infty} \lambda_k^2 < \infty$, for any $x_0 \in \mathbb{R}^n$, $\liminf_k f(x_k) = \min f(\mathbb{R}^n)$ for $\{x_k\}$ generated by the subgradient method.*

Proof: Choose any $x_* \in X_*$ and look at $\|x_k - x_*\|$ before and after a step.

$$\begin{aligned} \|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 &= \|x_{k+1}\|^2 - \|x_k\|^2 - 2x_{k+1}^T x_* + 2x_k^T x_* \\ &= \|x_{k+1} - x_k\|^2 + 2(x_{k+1} - x_k)^T (x_k - x_*) \\ &= \lambda_k^2 - 2\lambda_k \frac{g_k}{\|g_k\|}^T (x_k - x_*). \end{aligned}$$

By the subgradient inequality, $g_k^T (x_k - x_*) \geq f_k - f_* \geq 0$, where $f_\bullet := f(x_\bullet)$.

So $\|x_\ell - x_*\|^2 - \|x_0 - x_*\|^2 \leq \sum_{k=0}^{\ell-1} \lambda_k^2$.

Hence, $\{x_\ell\}$ lies in some bounded set since $\sum_{k=0}^{\infty} \lambda_k^2 < \infty$. So all $\|g_\ell\|$'s are uniformly bounded (true but not obvious: proof omitted).

So $\|g_k\| \leq \Gamma \ \forall \ k$ for some Γ.

Now assume $f_k \geq f_* + \varepsilon$, for some $\varepsilon > 0$, $\forall \ k \geq K_1$. Also, $\lambda_k \downarrow 0$ so $\lambda_k \leq \frac{\varepsilon}{\Gamma} \ \forall \ k \geq K_2 \geq K_1$. Then for $k \geq K_2$,

$$\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 \leq \lambda_k^2 - \frac{2\lambda_k}{\|g_k\|}\varepsilon \leq \frac{\varepsilon\lambda_k}{\Gamma} - \frac{2\varepsilon\lambda_k}{\Gamma} = -\frac{\varepsilon\lambda_k}{\Gamma}.$$

So $\|x_\ell - x_*\|^2 - \|x_{K_2} - x_*\|^2 \le -\frac{\varepsilon}{\Gamma}\sum_{K_2}^{\ell-1}\lambda_k$.

So $\|x_\ell - x_*\|^2 \le \|x_{K_2} - x_*\|^2 - \frac{\varepsilon}{\Gamma}\sum_{K_2}^{\ell-1}\lambda_k \to -\infty$ as $\ell \to \infty$..

We have obtained our contradiction and $\liminf f_k = f_*$ as claimed. $\square$

To examine the complexity of the subgradient method, assume $f : \mathbb{R}^n \to \mathbb{R}$ is convex and we want to minimize f on $B(0,R)$. Assume $\max\{f(x) : x \in B(0,R)\} - \min\{f(x) : x \in B(0,R)\} \le 1$.

We will derive an algorithm with complexity $O(\varepsilon^{-2})$. Note that this is independent of n! This does not contradict our earlier lower bound of $\Omega\left(n \ln \frac{1}{\varepsilon}\right)$ because ε is sufficiently large. This bound is valid for all $\varepsilon < 1/2$ for $G = [-1,+1]^n$ but it is only valid for $\varepsilon < \frac{1}{n^3}$ for $G = B(0,R)$.

The algorithm uses step size $\lambda_k = \varepsilon R$.

Algorithm

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

Theorem 3 *Using this algorithm,* $\min_{0 \le k \le \ell} f(x_k) \le \min\{f(x) : x \in B(0,R)\} + \varepsilon$ *within* $\frac{1}{\varepsilon^2}$ *iterations.*

Proof of the theorem to follow in Lecture 21.