LECTURE 23

LECTURE OUTLINE

- Review of subgradient methods
- Application to differentiable problems - Gradient projection
- Iteration complexity issues
- Complexity of gradient projection
- Projection method with extrapolation
- Optimal algorithms

****

- Reference: The on-line chapter of the textbook

SUBGRADIENT METHOD

- **Problem:** Minimize convex function $f : \Re^n \mapsto \Re$ over a closed convex set X.
- **Subgradient method - constant step α:**

$$x_{k+1} = P_X\Big(x_k - \alpha_k \tilde{\nabla} f(x_k)\Big),$$

where $\tilde{\nabla} f(x_k)$ is a subgradient of f at x_k, and $P_X(\cdot)$ is projection on X.

- Assume $\|\tilde{\nabla} f(x_k)\| \leq c$ for all k.
- **Key inequality:** For all optimal x^*

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha\Big(f(x_k) - f^*\Big) + \alpha^2 c^2$$

- Convergence to a neighborhood result:

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha c^2}{2}$$

- Iteration complexity result: For any $\epsilon > 0$,

$$\min_{0\leq k\leq K} f(x_k) \leq f^* + \frac{\alpha c^2 + \epsilon}{2},$$

where $K = \left\lfloor \frac{\min_{x^*\in X^*} \|x_0 - x^*\|^2}{\alpha\epsilon} \right\rfloor$.

- For $\alpha = \epsilon/c^2$, we need $O(1/\epsilon^2)$ iterations to get within ϵ of the optimal value f^*.

GRADIENT PROJECTION METHOD

- Let f be differentiable and assume

$$\big\|\nabla f(x) - \nabla f(y)\big\| \le L\,\|x-y\|, \qquad \forall\ x, y \in X$$

- **Gradient projection method:**

$$x_{k+1} = P_X\big(x_k - \alpha \nabla f(x_k)\big)$$

- Define the *linear approximation function at* x

$$\ell(y;x) = f(x) + \nabla f(x)'(y-x), \qquad y \in \Re^n$$

- **First key inequality:** For all $x, y \in X$

$$f(y) \le \ell(y;x) + \frac{L}{2}\|y-x\|^2$$

- Using the projection theorem to write

$$\big(x_k - \alpha\nabla f(x_k) - x_{k+1}\big)'(x_k - x_{k+1}) \le 0,$$

and then the 1st key inequality, we have

$$f(x_{k+1}) \le f(x_k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x_{k+1} - x_k\|^2$$

so there is cost reduction for $\alpha \in \big(0, \frac{2}{L}\big)$

ITERATION COMPLEXITY

- Connection with proximal algorithm

$$y = \arg\min_{z \in X} \left\{ \ell(z;x) + \frac{1}{2\alpha}\|z-x\|^2 \right\} = P_X\big(x - \alpha\nabla f(x)\big)$$

- **Second key inequality:** For any $x \in X$, if $y = P_X\big(x - \alpha\nabla f(x)\big)$, then for all $z \in X$, we have

$$\ell(y;x) + \frac{1}{2\alpha}\|y-x\|^2 \le \ell(z;x) + \frac{1}{2\alpha}\|z-x\|^2 - \frac{1}{2\alpha}\|z-y\|^2$$

- **Complexity Estimate:** Let the stepsize of the method be $\alpha = 1/L$. Then for all k

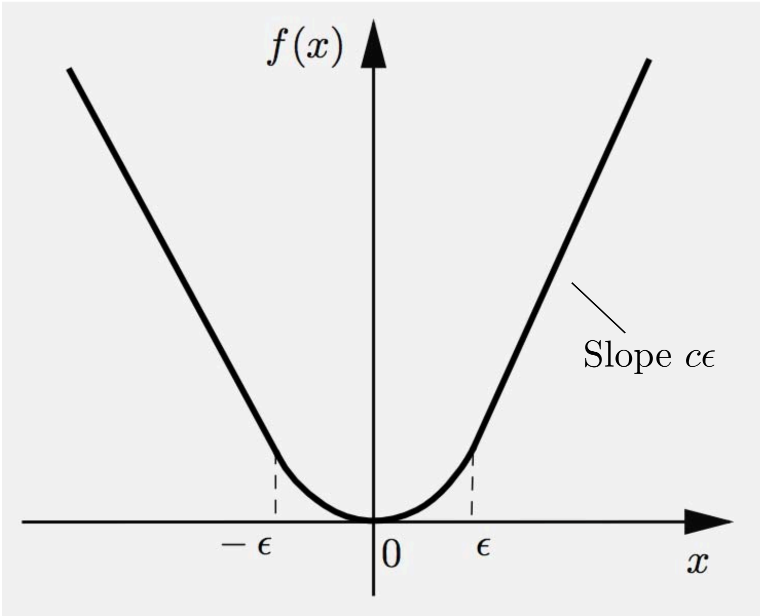
$$f(x_k) - f^* \le \frac{L \min_{x^* \in X^*} \|x_0 - x^*\|^2}{2k}$$

- Thus, we need $O(1/\epsilon)$ iterations to get within ϵ of f^*. **Better than nondifferentiable case.**

- Practical implementation/same complexity: Start with some α and reduce it by some factor as many times as necessary to get

$$f(x_{k+1}) \le \ell(x_{k+1}; x_k) + \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2$$

SHARPNESS OF COMPLEXITY ESTIMATE



- Unconstrained minimization of

$$f(x) = \begin{cases} \frac{c}{2}|x|^2 & \text{if } |x| \leq \epsilon, \\ c\epsilon|x| - \frac{c\epsilon^2}{2} & \text{if } |x| > \epsilon \end{cases}$$

- With stepsize $\alpha = 1/L = 1/c$ and any $x_k > \epsilon$,

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) = x_k - \frac{1}{c}c\,\epsilon = x_k - \epsilon$$

- The number of iterations to get within an ϵ-neighborhood of $x^* = 0$ is $|x_0|/\epsilon$.

- The number of iterations to get to within ϵ of $f^* = 0$ is proportional to $1/\epsilon$ for large x_0.

EXTRAPOLATION VARIANTS

- An old method for unconstrained optimization, known as the *heavy-ball* method or gradient method with *momentum*:

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}),$$

where $x_{-1} = x_0$ and β is a scalar with $0 < \beta < 1$.

- A variant of this scheme for constrained problems separates the extrapolation and the gradient steps:

$$y_k = x_k + \beta(x_k - x_{k-1}), \quad \text{(extrapolation step)},$$
$$x_{k+1} = P_X\big(y_k - \alpha\nabla f(y_k)\big), \quad \text{(grad. projection step)}.$$

- When applied to the preceding example, the method converges to the optimum, and reaches a neighborhood of the optimum more quickly

- However, the method still has an $O(1/\epsilon)$ iteration complexity, since for $x_0 >> 1$, we have

$$x_{k+1} - x_k = \beta(x_k - x_{k-1}) - \epsilon$$

so $x_{k+1} - x_k \approx \epsilon/(1-\beta)$, and the number of iterations needed to obtain $x_k < \epsilon$ is $O\big((1-\beta)/\epsilon\big)$.

OPTIMAL COMPLEXITY ALGORITHM

- Surprisingly with a proper more vigorous extrapolation $\beta_k \to 1$ in the extrapolation scheme

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \qquad \text{(extrapolation step)},$$
$$x_{k+1} = P_X\big(y_k - \alpha\nabla f(y_k)\big), \qquad \text{(grad. projection step)},$$

the method has iteration complexity $O\big(1/\sqrt{\epsilon}\big)$.

- Choices that work

$$\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}$$

where the sequence $\{\theta_k\}$ satisfies $\theta_0 = \theta_1 \in (0,1]$, and

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \le \frac{1}{\theta_k^2}, \qquad \theta_k \le \frac{2}{k+2}$$

- One possible choice is

$$\beta_k = \begin{cases} 0 & \text{if } k=0, \\ \frac{k-1}{k+2} & \text{if } k \ge 1, \end{cases} \qquad \theta_k = \begin{cases} 1 & \text{if } k=-1, \\ \frac{2}{k+2} & \text{if } k \ge 0. \end{cases}$$

- Highly unintuitive. Good performance reported.

EXTENSION TO NONDIFFERENTIABLE CASE

- Consider the nondifferentiable problem of minimizing convex function $f : \Re^n \mapsto \Re$ over a closed convex set X.

- Approach: "Smooth" f, i.e., approximate it with a differentiable function by using a proximal minimization scheme.

- Apply optimal complexity gradient projection method with extrapolation. Then an $O(1/\epsilon)$ iteration complexity algorithm is obtained.

- Can be shown that this complexity bound is sharp.

- Improves on the subgradient complexity bound by a an ϵ factor.

- Limited experience with such methods.

- Major disadvantage: Cannot take advantage of special structure, e.g., there are no incremental versions.

6.253 Convex Analysis and Optimization
Spring 2012