LECTURE 22

LECTURE OUTLINE

- Incremental methods
- Review of large sum problems
- Review of incremental gradient and subgradient methods
- Combined incremental subgradient and proximal methods
- Convergence analysis
- Cyclic and randomized component selection
- References:

(1) D. P. Bertsekas, "Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey", Lab. for Information and Decision Systems Report LIDS-P-2848, MIT, August 2010

(2) Published versions in Math. Programming J., and the edited volume "Optimization for Machine Learning," by S. Sra, S. Nowozin, and S. J. Wright, MIT Press, Cambridge, MA, 2012.

LARGE SUM PROBLEMS

- Minimize over $X \subset \Re^n$

$$f(x) = \sum_{i=1}^{m} f_i(x), \qquad m \text{ is very large,}$$

where X, f_i are convex. Some examples:

- **Dual cost of a separable problem.**

- **Data analysis/machine learning:** x is parameter vector of a model; each f_i corresponds to error between data and output of the model.
  - Least squares problems (f_i quadratic).
  - ℓ_1-regularization (least squares plus ℓ_1 penalty):

$$\min_x \gamma \sum_{j=1}^{n} |x^j| + \sum_{i=1}^{m} (c_i' x - d_i)^2$$

The nondifferentiable penalty tends to set a large number of components of x to 0.

- **Min of an expected value** $\min_x E\{F(x, w)\}$ **- Stochastic programming:**

$$\min_x \left[F_1(x) + E_w\{\min_y F_2(x, y, w)\} \right]$$

- **More** (many constraint problems, distributed incremental optimization ...)

INCREMENTAL SUBGRADIENT METHODS

- The special structure of the sum

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

can be exploited by incremental methods.

- We first consider incremental subgradient methods which **move x along a subgradient $\tilde{\nabla} f_i$ of a component function** f_i NOT the (expensive) subgradient of f, which is $\sum_i \tilde{\nabla} f_i$.

- At iteration k select a component i_k and set

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

with $\tilde{\nabla} f_{i_k}(x_k)$ being a subgradient of f_{i_k} at x_k.

- **Motivation is faster convergence.** A cycle can make much more progress than a subgradient iteration with essentially the same computation.

CONVERGENCE PROCESS: AN EXAMPLE

- **Example 1:** Consider

$$\min_{x \in \Re} \frac{1}{2} \left\{ (1-x)^2 + (1+x)^2 \right\}$$

- Constant stepsize: Convergence to a limit cycle
- Diminishing stepsize: Convergence to the optimal solution
- **Example 2:** Consider

$$\min_{x \in \Re} \left\{ |1-x| + |1+x| + |x| \right\}$$

- Constant stepsize: Convergence to a limit cycle that depends on the starting point
- Diminishing stepsize: Convergence to the optimal solution
- What is the effect of the order of component selection?

CONVERGENCE: CYCLIC ORDER

- Algorithm

$$x_{k+1} = P_X\big(x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k)\big)$$

- Assume all subgradients generated by the algorithm are bounded: $\|\tilde{\nabla} f_{i_k}(x_k)\| \leq c$ for all k
- Assume components are chosen for iteration in cyclic order, and stepsize is constant within a cycle of iterations (for all k with $i_k = 1$ we have $\alpha_k = \alpha_{k+1} = \ldots = \alpha_{k+m-1}\big)$
- **Key inequality:** For all $y \in X$ and all k that mark the beginning of a cycle

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 c^2$$

- Result for a constant stepsize $\alpha_k \equiv \alpha$:

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

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$.

CONVERGENCE: RANDOMIZED ORDER

- Algorithm

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

- Assume component i_k chosen for iteration in randomized order (independently with equal probability)
- Assume all subgradients generated by the algorithm are bounded: $\|\tilde{\nabla} f_{i_k}(x_k)\| \leq c$ for all k
- Result for a constant stepsize $\alpha_k \equiv \alpha$:

$$\lim \inf_{k\to\infty} f(x_k) \leq f^* + \alpha \frac{mc^2}{2}$$

(with probability 1)

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$. (with probability 1)
- In practice, randomized stepsize and variations (such as randomization of the order within a cycle at the start of a cycle) often work much faster

PROXIMAL-SUBGRADIENT CONNECTION

- **Key Connection:** The proximal iteration

$$x_{k+1} = \arg\min_{x \in X} \left\{ f(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

can be written as

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

where $\tilde{\nabla} f(x_{k+1})$ is *some* subgradient of f at x_{k+1}.

- Consider an incremental proximal iteration for $\min_{x \in X} \sum_{i=1}^m f_i(x)$

$$x_{k+1} = \arg\min_{x \in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

- **Motivation:** Proximal methods are more "stable" than subgradient methods
- **Drawback:** Proximal methods require special structure to avoid large overhead
- This motivates a combination of incremental subgradient and proximal

INCR. SUBGRADIENT-PROXIMAL METHODS

- Consider the problem

$$\min_{x \in X} F(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} F_i(x)$$

where for all i,

$$F_i(x) = f_i(x) + h_i(x)$$

X, f_i and h_i are convex.

- We consider combinations of subgradient and proximal incremental iterations

$$z_k = \arg\min_{x \in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

$$x_{k+1} = P_X\big(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k)\big)$$

- Variations:
  - Min. over $\Re^n$ (rather than X) in proximal
  - Do the subgradient without projection first and then the proximal

- **Idea:** Handle "favorable" components f_i with the more stable proximal iteration; handle other components h_i with subgradient iteration.

CONVERGENCE: CYCLIC ORDER

- Assume all subgradients generated by the algorithm are bounded: $\|\tilde{\nabla} f_{i_k}(x_k)\| \leq c$, $\|\tilde{\nabla} h_{i_k}(x_k)\| \leq c$ for all k, plus mild additional conditions
- Assume components are chosen for iteration in cyclic order, and stepsize is constant within a cycle of iterations
- **Key inequality:** For all $y \in X$ and all k that mark the beginning of a cycle:

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k \big(F(x_k) - F(y)\big) + \beta \alpha_k^2 m^2 c^2$$

where β is a (small) constant

- Result for a constant stepsize $\alpha_k \equiv \alpha$:

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

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$.

CONVERGENCE: RANDOMIZED ORDER

- Result for a constant stepsize $\alpha_k \equiv \alpha$:

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

(with probability 1)

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$. (with probability 1)

EXAMPLE

- ℓ_1-Regularization for least squares with large number of terms

$$\min_{x\in\Re^n}\left\{\gamma\,\|x\|_1+\frac{1}{2}\sum_{i=1}^{m}(c_i'x-d_i)^2\right\}$$

- Use incremental gradient or proximal on the quadratic terms
- Use proximal on the $\|x\|_1$ term:

$$z_k=\arg\min_{x\in\Re^n}\left\{\gamma\,\|x\|_1+\frac{1}{2\alpha_k}\|x-x_k\|^2\right\}$$

- Decomposes into the n one-dimensional minimizations

$$z_k^j=\arg\min_{x^j\in\Re}\left\{\gamma\,|x^j|+\frac{1}{2\alpha_k}|x^j-x_k^j|^2\right\},$$

and can be done in closed form

$$z_k^j=\begin{cases} x_k^j-\gamma\alpha_k & \text{if } \gamma\alpha_k\le x_k^j,\\ 0 & \text{if } -\gamma\alpha_k<x_k^j<\gamma\alpha_k,\\ x_k^j+\gamma\alpha_k & \text{if } x_k^j\le-\gamma\alpha_k.\end{cases}$$

- Note that "small" coordinates x_k^j are set to 0.

6.253 Convex Analysis and Optimization
Spring 2012