LECTURE 23

LECTURE OUTLINE

- Incremental methods
- Review of large sum problems
- Review of incremental gradient methods
- Incremental subgradient-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 - Lagrangian relaxation, integer programming.

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

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

  - Classification (logistic regression, support vector machines)
  - Max-likelihood

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

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

- More (many constraint problems, etc ...)

INCREMENTAL GRADIENT METHOD

- Problem: Minimization of $f(x) = \sum_{i=1}^{m} f_i(x)$ over a closed convex set X (f_i differentiable).
- Operates in cycles: If x_k is the vector obtained after k cycles, the vector x_{k+1} obtained after one more cycle is $x_{k+1} = \psi_{m,k}$, where $\psi_{0,k} = x_k$, and

$$\psi_{i,k} = P_X\big(\psi_{i-1,k} - \alpha_k \nabla f_{i,k}(\psi_{i-1,k})\big), \qquad i = 1, \ldots, m$$

- Does NOT compute the (expensive) gradient of f, which is $\sum_i \nabla f_i$.
- Interesting issues of ordering the processing of components.
- Randomization of selection of component f_i is possible. Connection with stochastic gradient method.
- Diminishing stepsize needed for convergence.
- Example: Consider

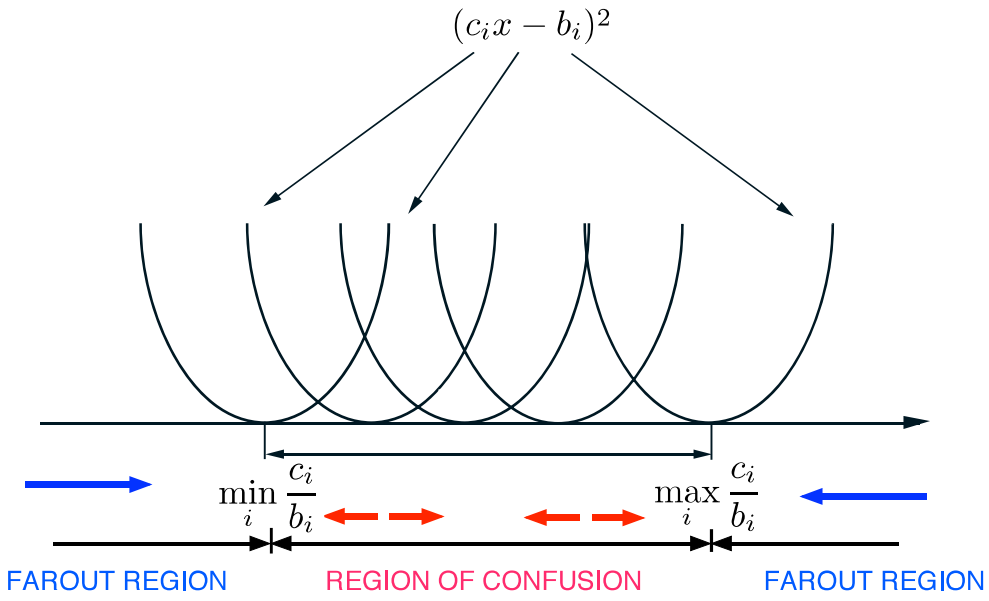
$$\min_{x \in \Re} \frac{1}{2}\big\{(1-x)^2 + (1+x)^2\big\}$$

For a constant stepsize the incremental gradient method oscillates.

COMPARE W/ NONINCREMENTAL GRADIENT

- Two complementary performance issues:
  - Progress when far from convergence. Here the incremental method can be much faster.
  - Progress when close to convergence. Here the incremental method can be inferior.
- Example: Scalar case

$$f_i(x) = \tfrac{1}{2}(c_i x - b_i)^2, \qquad x \in \Re$$



- Interesting issues of batching/shaping the region of confusion.
- Hybrids between incremental and nonincremental gradient methods. Aggregated gradient method.

INCREMENTAL SUBGRADIENT METHODS

- Problem: Minimize

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

over a closed convex set X, where $f_i : \Re^n \mapsto \Re$ are convex, and possibly nondifferentiable.

- We first consider incremental subgradient methods which move x along a subgradient $\tilde{\nabla} f_i$ of a component function 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: 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 $\alpha_k = \alpha_{k+1} = \cdots = \alpha_{k+m-1}$)
- Key inequality: For all $y \in X$ and all k that mark the beginning of a cycle

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

Progress if $-2\alpha_k\Big(f(x_k) - f(y)\Big) + \alpha_k^2 m^2 c^2 < 0$.

- 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

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

- 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 $\alpha_k \equiv \alpha$:

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

(with probability 1) - improvement by a factor m over the cyclic order case.

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$. (with probability 1). Use of the supermartingale convergence theorem.
- 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.

SUBGRADIENT-PROXIMAL 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

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

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 (split iteration, similar to proximal gradient).

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.

- Consider a combination of subgradient and proximal incremental iterations

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

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

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

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

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 the constant $\beta = 1/m + 4$.

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

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

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

CONVERGENCE: RANDOMIZED ORDER

- Convergence and convergence rate results are qualitatively similar to incremental subgradient case.
- 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).

- Faster convergence for randomized stepsize rule - improvement by a factor m over the cyclic order case.
- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$. (with probability 1). Use of the supermartingale convergence theorem.

EXAMPLE I

- ℓ_1-Regularization for least squares

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

- 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 with the shrinkage operation

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

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

EXAMPLE II

- Incremental constraint projection methods for

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in \cap_{i=1}^{m} X_i, \end{aligned} \tag{1}$$

- Convert to the problem

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) + c \sum_{i=1}^{m} \text{dist}(x; X_i) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned} \tag{2}$$

where c is a positive penalty parameter.

- Then for f Lipschitz continuous and c sufficiently large, problems (1) and (2) are equivalent (their minima coincide).

- Apply incremental subgradient-proximal:

$$y_k = x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k),$$

$$x_{k+1} \in \arg \min_{x \in \Re^n} \left\{ c\, \text{dist}(x; X_{j_k}) + \frac{1}{2\alpha_k} \|x - y_k\|^2 \right\}.$$

The second iteration can be implemented in "closed form," using projection on X_{j_k}.