LECTURE 22

LECTURE OUTLINE

- Gradient projection method
- Iteration complexity issues
- Gradient projection with extrapolation
- Proximal gradient method

****

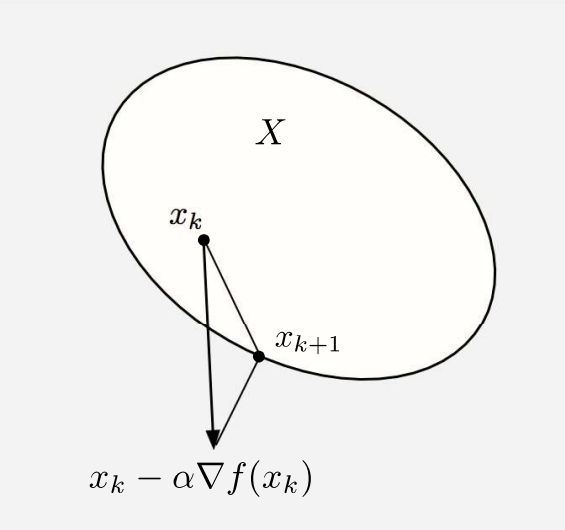
References:

- The on-line chapter of the textbook
- Beck, A., and Teboulle, M., 2010. "Gradient-Based Algorithms with Applications to Signal Recovery Problems, in Convex Optimization in Signal Processing and Communications (Y. Eldar and D. Palomar, eds.), Cambridge University Press, pp. 42-88.
- J. Lee, Y. Sun, M. Saunders, "Proximal Newton-Type Methods for Convex Optimization," NIPS, 2012.

REVIEW OF GRADIENT PROJECTION METHOD

- Let f be continuously differentiable, and X be closed convex.
- Gradient projection method:

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

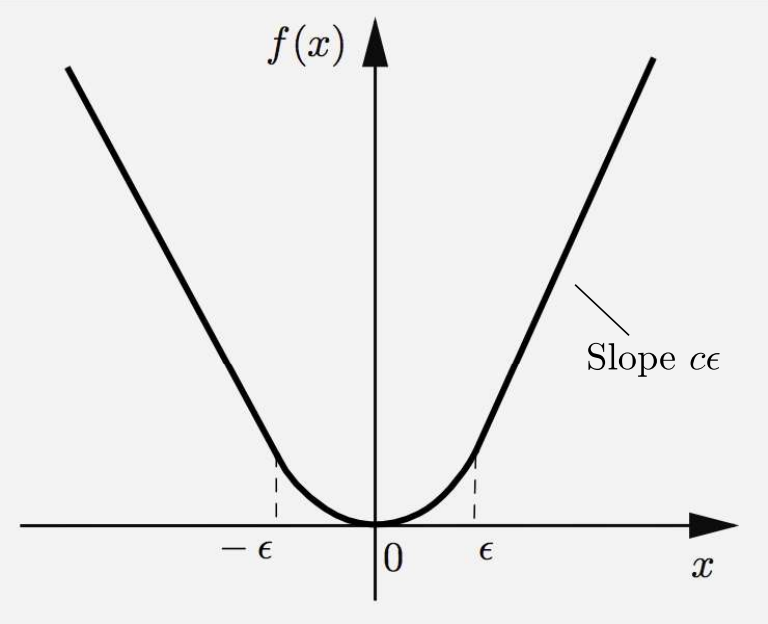


- α_k may be constant or chosen by cost descent-based stepsize rules
- Under gradient Lipschitz assumption

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

iteration complexity $O(1/\epsilon)$, ($O(1/\epsilon)$ iterations for ϵ cost error), i.e., $\min_{k \leq \frac{\text{const}}{\epsilon}} f(x_k) \leq f^* + \epsilon$

SHARPNESS OF COMPLEXITY ESTIMATE



- Unconstrained minimization of

$$f(x) = \begin{cases} \frac{1}{2}|x|^2 & \text{if } |x| \leq \epsilon, \\ \epsilon|x| - \frac{\epsilon^2}{2} & \text{if } |x| > \epsilon \end{cases}$$

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

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

- 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 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 $O(1/k)$ error 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

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \quad \text{(extrapolation step)},$$
$$x_{k+1} = P_X\Big(y_k - \frac{1}{L}\nabla f(y_k)\Big), \quad \text{(grad. projection step)},$$

the method has iteration complexity $O\big(\sqrt{L/\epsilon}\big)$. (Also with "eventually constant" rule for α.)

- Choices that work

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

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

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

- One possible choice is

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

- Highly unintuitive. Good practical 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.
- "Smooth" f, i.e., approximate it with a differentiable function f_ϵ with Lipschitz constant $O(1/\epsilon)$ by using a proximal minimization scheme.
- The smoothed function satisfies

$$f_\epsilon(x) \leq f(x) \leq f_\epsilon(x) + O(\epsilon)$$

- Apply optimal complexity gradient projection method with extrapolation. Then an $O(1/\epsilon)$ complexity algorithm is obtained.
- Can be shown that this complexity bound is sharp.
- Improves on the subgradient complexity bound by an ϵ factor.
- Limited practical experience with such methods.

CRITIQUE OF THE OPTIMAL ALGORITHM

- Requires gradient Lipschitz assumption
- Chooses the stepsize α_k in the basis of the worst possible curvature information (same Lipschitz constant assumed in all directions).
- Compares well relative to competitors for some difficult problems (singular Hessian, but under Lipschitz gradient assumption).
- Not so well for other difficult problems (Lipschitz gradient assumption not holding) or easier problems (nonsingular Hessian) for which it has to compete with conjugate gradient and quasi-Newton methods
- A weak point: Cannot take advantage of special structure, e.g., there are no incremental versions.
- A strong point: Its favorable complexity estimate carries over to combinations with proximal algorithms.

PROXIMAL GRADIENT METHOD

- Minimize $f(x) + h(x)$ over $x \in X$, where X: closed convex, f, h: convex, f is differentiable.
- Proximal gradient method:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ \ell(x; x_k) + h(x) + \frac{1}{2\alpha} \|x - x_k\|^2 \right\}$$

where $\ell(x; x_k) = f(x_k) + \nabla f(x_k)'(x - x_k)$

- Equivalent definition of proximal gradient:

$$z_k = x_k - \alpha \nabla f(x_k)$$

$$x_{k+1} \in \arg\min_{x \in X} \left\{ h(x) + \frac{1}{2\alpha} \|x - z_k\|^2 \right\}$$

- Simplifies the implementation of proximal, by using gradient iteration to deal with the case of an inconvenient component f
- Important example: h is the ℓ_1 norm - use the shrinkage operation to simplify the proximal
- The gradient projection and extrapolated variant analysis carries through, with the same iteration complexity

PROXIMAL GRADIENT METHOD ANALYSIS

- Recall descent lemma: For all $x, y \in X$

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

where

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

- Recall three-term inequality: For all $y \in \Re^n$,

$$\begin{aligned}\|x_{k+1} - y\|^2 \leq\ & \|x_k - y\|^2 \\ & - 2\alpha_k\big(\ell(x_{k+1}; x_k) + h(x_{k+1}) - \ell(y; x_k) - h(y)\big) \\ & - \|x_k - x_{k+1}\|^2\end{aligned}$$

- Eventually constant stepsize rule: Keep using same α, as long as

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

- As soon as this condition is violated, reduce α by a certain factor, and repeat the iteration as many times as is necessary for Eq. (1) to hold.

RATE OF CONVERGENCE RESULT

• Assume ∇f satisfies the Lipschitz condition and the set of minima X^* of f over X is nonempty. If $\{x_k\}$ is a sequence generated by the proximal gradient method using any stepsize rule such that

$$\alpha_k \downarrow \overline{\alpha},$$

for some $\overline{\alpha} > 0$, and for all k,

$$f(x_{k+1}) \leq \ell(x_{k+1}; x_k) + \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|^2,$$

then $\lim_{k\to\infty} d(x_k) = 0$, and

$$f(x_k) + h(x_k) - \min_{x\in X} \Big\{ f(x) + h(x) \Big\} \leq \frac{\overline{\alpha}\, d(x_0)^2}{2k}, \quad \forall\ k,$$

where

$$d(x) = \min_{x^*\in X^*} \|x - x^*\|, \qquad x \in \Re^n$$

SCALED PROXIMAL GRADIENT METHODS

- Idea: Instead of gradient, use scaled gradient, quasi-Newton, or Newton:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ \ell(x; x_k) + h(x) + \tfrac{1}{2}(x - x_k)' H_k (x - x_k) \right\},$$

where H_k is a positive definite symmetric matrix.

- Can use $H_k = \nabla^2 f(x_k)$ (fast convergence) but the proximal minimization may become complicated.

- Lots of room for new methods ...