Optimization Methods in Machine Learning
Lecture 22

Katya Scheinberg

Lehigh University

katyas@lehigh.edu

Splitting, alternating linearization and alternating direction methods

Augmented Lagrangian

$$\begin{aligned} \min \quad & f_0(x), \\ \text{s.t.} \quad & f_i(x) = 0, \ \ i = 1, \dots, m \end{aligned}$$

Augmented Lagrangian function

$$L(x, y) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{m} \frac{1}{2\mu_i} \|f_i(x)\|^2$$

Augmented Lagrangian method

$$\begin{aligned} & \text{For } k = 1, 2, \dots \\ & x^k = \text{argmin}_x L(x, \lambda^k) \\ & \lambda_i^{k+1} = \lambda_i^k - \frac{1}{\mu_i} f_i(x^k), \ \ i = 1, \dots, m \end{aligned}$$

Alternating directions (splitting) method

- Consider:

$$\min_{x} F(x) = f(x) + g(x)$$

$$\Updownarrow$$

$$\begin{aligned} \min_{x,y} \quad & f(x) + g(y) \\ \text{s.t.} \quad & y = x \end{aligned}$$

- Relax constraints via Augmented Lagrangian technique

$$\min_{x,y} f(x) + g(y) + \lambda^\top (y - x) + \frac{1}{2\mu}||y - x||^2 = Q_\lambda(x, y)$$

Assume that *f(x)* and *g(y)* are both such that the above functions are easy to optimize in x or y

Alternating direction method (ADM)

- $x^{k+1} = \min_x Q_\lambda(x, y^k)$
- $y^{k+1} = \min_y Q_\lambda(x^{k+1}, y)$
- $\lambda^{k+1} = \lambda^k + \frac{1}{\mu}(y^{k+1} - x^{k+1})$

Widely used method without complexity bounds

Combettes and Wajs, ‘05

Eckstein and Bertsekas, ‘92,

Eckstein and Svaiter, ’08

Glowinski and Le Tallec, ‘89

Kiwiel, Rosa, and Ruszczynski, ’99

Lions and Mercier ‘79

A slight modification of ADM

- $x^{k+1} = \min_x Q_\lambda(x, y^k)$
- $\lambda^{k+\frac{1}{2}} = \lambda^k + \frac{1}{\mu}(y^k - x^{k+1})$
- $y^{k+1} = \min_y Q_\lambda(x^{k+1}, y)$
- $\lambda^{k+1} = \lambda^{k+\frac{1}{2}} + \frac{1}{\mu}(y^{k+1} - x^{k+1})$

This turns out to be equivalent to……

Goldfarb, Ma and S, '10

Alternating linearization method (ALM)

- $x^{k+1} = \min_x Q_g(x, y^k)$
- $y^{k+1} = \min_y Q_f(x^{k+1}, y)$

$$Q_g(x, y) = f(x) + \nabla g(y)^\top (x - y) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

$$Q_f(x, y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

Goldfarb, Ma, S, ‘10

Convergence rate for ALM

- $x^{k+1} = \min_x Q_g(x, y^k)$
- $y^{k+1} = \min_y Q_f(x^{k+1}, y)$

Th: If $\mu \leq$ *1/L* then in *O(L/ϵ)* iterations finds ϵ -optimal solution

Goldfarb, Ma, S, '10

Convergence rate for fast ALM

- $x^k := \min_x Q_g(x, z^k)$
- $y^k := \min_y Q_f(x^k, y)$
- $t_{k+1} := (1 + \sqrt{1 + 4t_k^2})/2$
- $z^{k+1} := y^k + \frac{t_k - 1}{t_{k+1}}[y^k - y^{k-1}]$

Th: If $\mu \leq$ *1/L* then in $O(\sqrt{L/\epsilon})$ iterations finds ϵ -optimal solution

Goldfarb, Ma, S, '10

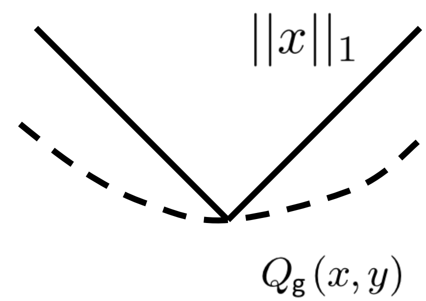
Alternating linearization method for nonsmooth g

$$\min_x F(x) = \min_x f(x) + g(x)$$

$$|\nabla f(x) - \nabla f(y)| \leq L||x-y||$$

This is not true for $||x||_1$!!!

$$|\nabla g(x) - \nabla g(y)| \leq L||x-y||$$



$Q_g(x,y)$ may not be an upper approximation of F(x)!

Idea: with line search can accept different μ values, including zero, for g

Goldfarb, Ma, S, '10

Examples of applications of alternating linearization method

Sparse Inverse Covariance Selection

$$\max_{X \succ 0} (\ln\det(X) - Tr(AX)) - \rho \|X\|_1$$

$$\underbrace{\qquad\qquad}_{f(x)} \qquad \underbrace{\qquad}_{g(x)}$$

$$X^{k+1} := \operatorname{argmin}_X \{ f(X) + \frac{1}{2\mu_{k+1}} \|X - (Y^k + \mu_{k+1}\Lambda^k)\|_F^2 \}$$

Eigenvalue decomposition O(n^3) ops. Same as one gradient of *f(X)*

$$Y^{k+1} := \operatorname{argmin}_Y \{ g(Y) + \frac{1}{2\mu_{k+1}} \|Y - (X^{k+1} - \mu_{k+1}(A - (X^{k+1})^{-1}))\|_F^2 \}$$

Shrinkage O(n^2) ops

Sparse Inverse Covariance Selection

$$\max_{X \succ 0} (\ln\det(X) - Tr(AX)) - \lambda \|X\|_1$$

$$\underbrace{\ln\det(X) - Tr(AX)}_{f(x)} \qquad \underbrace{\lambda \|X\|_1}_{g(x)}$$

$$X^{k+1} := \operatorname{argmin}_X \{ f(X) + \frac{1}{2\mu_{k+1}} \|X - (Y^k + \mu_{k+1}\Lambda^k)\|_F^2 \}$$

$V\mathrm{Diag}(d)V^\top$ - the spectral decomposition of $Y^k + \mu_{k+1}(\Lambda^k - A)$

$$\gamma_i = \left(d_i + \sqrt{d_i^2 + 4\mu_{k+1}} \right) / 2, \quad i = 1, \ldots, p$$

$$X^{k+1} := V\mathrm{Diag}(\gamma)V^\top$$

Eigenvalue decomposition O(n^3) ops. Same as one gradient of *f(X)*

Lasso or group Lasso

$$\min_x \|Ax - b\|^2 + \rho\|x\|_1$$

$$\underbrace{}_{f(x)} \quad \underbrace{}_{g(x)}$$

$$x^{k+1} := \text{argmin}_x\{f(x) + \frac{1}{2\mu_{k+1}}\|x - (y^k + \mu_{k+1}\lambda^k)\|^2\}$$

Eigenvalue decomposition O(n^3) ops. Same as one gradient of *f(X)*

$$y^{k+1} := \text{argmin}_y\{g(y) + \frac{1}{2\mu_{k+1}}\|y - (x^{k+1} - \mu_{k+1}A^\top(Ax - b))\|^2\}$$

Shrinkage O(n^2) ops

Robust PCA

$$\min_X \|X\|_* + \rho\|M - X\|_1$$

f(x) *g(x)*

$$X^{k+1} := \text{argmin}_X \{f(X) + \frac{1}{2\mu_{k+1}}\|X - (Y^k + \mu_{k+1}\Lambda^k)\|_F^2\}$$

Eigenvalue decomposition O(n^3) ops. Same as one gradient of *f(X)*

$$Y^{k+1} := \text{argmin}_Y \{g(Y) + \frac{1}{2\mu_{k+1}}\|Y - (X^{k+1} - \mu_{k+1}\Lambda^{k+\frac{1}{2}})\|_F^2\}$$

Shrinkage O(n^2) ops

Recall Collaborative Prediction?

$$\min_{X \in \mathrm{R}^{n \times m}} f(X) + ||X||_*$$

$$\min_Y Q_f(X, Y)$$

$$\Updownarrow$$

$$\min_Y \left[\frac{1}{2\mu} \|Y - Z\|_F^2 + \|Y\|_* \right]$$

$$\Updownarrow$$

$$Z = P\mathrm{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_n\} Q^\top$$

$$Y^* = P\mathrm{diag}\{\sigma_1^*, \sigma_2^*, \ldots, \sigma_n^*\} Q^\top, \; \sigma_i^* = \begin{cases} \sigma_i - \mu & \text{if } \sigma_i > \mu \\ 0 & \text{if } -\mu \le \sigma_i \le \mu \\ \sigma_i + \mu & \text{if } \sigma_i < -\mu \end{cases}$$

Closed form solution!
O(n^3) effort