NORTHWESTERN UNIVERSITY, EECS DEPARTMENT

EECS 495 Homework 3

Zhiyuan Wang

February 25, 2014

1 Accelerated proximal gradient for the lasso problem

the proximal gradient step:

Since $\|A\mathbf{x}-\mathbf{b}\|_2^2$ is L-Lipschitz by (5.2.3) we get:

$$\mathbf{x}^* = \min_{\mathbf{x}} \left\|\mathbf{x}-\mathbf{y}\right\|_2^2 + \frac{2\lambda}{L}\|\mathbf{x}\|_1 \tag{1.1}$$

where $\mathbf{y} = \mathbf{x}^{k-1} - \frac{1}{L}\bigtriangledown f(\mathbf{x}^{k-1})$, $L = d_{max}(A^T A)$

This problem is separable by each entry of x_i:

$$x_i^* = \operatorname*{argmin}_x \,(x_i - y_i)^2 + \frac{2\lambda}{L}|x_i| \tag{1.2}$$

$$\bigtriangledown f(x_i) = 2(x_i - y_i + \frac{\lambda}{L}sign(x_i)) \tag{1.3}$$

If $y_i = 0$, then clearly $x_i^* = 0$. If $y_i > 0$, there must be $x_i > 0$, otherwise let $\bigtriangledown f(x_i) = 0$, then for $\forall\lambda > 0$, $y_i = x_i - \frac{\lambda}{L} < 0$, that contradicts the assumption. Likewise, when $y_i < 0$, $x_i < 0$.

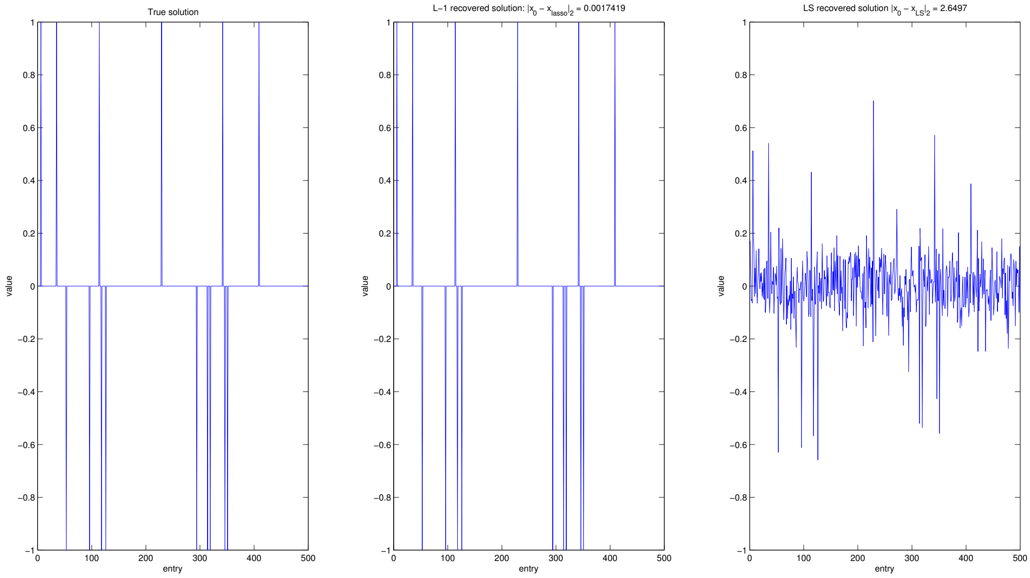
Thus,

$$x_i^k = \begin{cases} [y_i - \frac{\lambda}{L}]^+ & \text{if } y > 0 \\ 0 & \text{if } y = 0 \\ -[-y_i - \frac{\lambda}{L}]^+ & \text{if } y < 0 \end{cases} \tag{1.4}$$

$$
\begin{aligned}
x_i^k &= \left[|y_i| - \frac{\lambda}{L}\right]^+ sign(y_i) \\
&= \left[\left|\left(\mathbf{x}^{k-1} - \frac{1}{d_{max}(A^T A)} A^T(A\mathbf{x}^{k-1} - \mathbf{b})\right)_i\right| - \frac{\lambda}{L}\right]^+ sign\left(\left(\mathbf{x}^{k-1} - \frac{1}{d_{max}(A^T A)} A^T(A\mathbf{x}^{k-1} - \mathbf{b})\right)_i\right)
\end{aligned}
\tag{1.5}
$$

the accelerated proximal gradient step:

$$
\begin{aligned}
x_i^k &= \left[\left|\left(\mathbf{y}^{k-1} - \frac{1}{d_{max}(A^T A)} A^T(A\mathbf{y}^{k-1} - \mathbf{b})\right)_i\right| - \frac{\lambda}{L}\right]^+ sign\left(\left(\mathbf{y}^{k-1} - \frac{1}{d_{max}(A^T A)} A^T(A\mathbf{y}^{k-1} - \mathbf{b})\right)_i\right) \\
\mathbf{y}^k &= \mathbf{x}^k + \frac{k}{k+3}(\mathbf{x}^k - \mathbf{x}^{k-1})
\end{aligned}
\tag{1.6}
$$



2 The L-Lipschitz constant for logistic loss

$$
\bigtriangledown_{\mathbf{x}}^3 \mathscr{H} = \sum_{n=1}^{N} Q \bigtriangledown f_{\mathbf{x}}(\mathbf{a}^n)(1 - 2f_{\mathbf{x}}(\mathbf{a}^n))R \tag{2.1}
$$

Let $\bigtriangledown_{\mathbf{x}}^3 \mathscr{H} = 0$, $f_{\mathbf{x}}(\mathbf{a}^n) = 0, \frac{1}{2}, 1$. $\bigtriangledown_{\mathbf{x}}^2 \mathscr{H}$ get maxima at $f_{\mathbf{x}}(\mathbf{a}^n) = \frac{1}{2}$.

$$
\begin{aligned}
\bigtriangledown_x^2 \mathscr{H} &\preceq \sum_{n=1}^{N} \frac{1}{4} (\mathbf{a}^n)^T \mathbf{a}^n \\
&\preceq \frac{1}{4} d_{max}(A^T A) I
\end{aligned}
\tag{2.2}
$$

$$L = \frac{1}{4} d_{max}(A^T A) \tag{2.3}$$

where $d_{max}(A^T A)$ is the largest eigenvalue of $A^T A$.

3 Sparse logistic regression applications

I tried to search some interesting stuff at IEEE Xplore, Google Scholar and Microsoft Academic Search, but nothing really caught my eyes, so I answered the other questions.

4 Nonnegative Matrix Factorization

The subproblem of minimizing over X is given as

$$\begin{aligned} &\min_X \ \frac{1}{2}\|AX - B\|_F^2 = \frac{1}{2}\sum_{i=1}^{N} \left\|\mathbf{a}^i X - \mathbf{b}^i\right\|_2^2 \\ &subject\ to\ X \geq 0 \end{aligned} \tag{4.1}$$

$$\begin{aligned} \nabla g(X) &= \sum_{i=1}^{N} (\mathbf{a}^i)^T (\mathbf{a}^i X - \mathbf{b}^i) \\ &= A^T(AX - B) \end{aligned} \tag{4.2}$$

$$\begin{aligned} \nabla^2 g(X) &= \sum_{i=1}^{N} (\mathbf{a}^i)^T \mathbf{a}^i \\ &= A^T A \end{aligned} \tag{4.3}$$

$$X^k = \left[X^{k-1} - \tfrac{1}{L} A^T (AX^{k-1} - B)\right]^+ \tag{4.4}$$

where $L = d_{max}(A^T A)$.

The subproblem of minimizing over A is given as

$$\begin{aligned} &\min_X \ \frac{1}{2}\|AX - B\|_F^2 = \frac{1}{2}\sum_{j=1}^{P} \|A\mathbf{x}_j - \mathbf{b}_j\|_2^2 \\ &subject\ to\ A \geq 0 \end{aligned} \tag{4.5}$$

$$\begin{aligned} \nabla g(A) &= \sum_{j=1}^{P} \mathbf{x}_j (A\mathbf{x}_j - \mathbf{b}_j) \\ &= X^T(AX^T - B^T) \end{aligned} \tag{4.6}$$

$$\begin{aligned} \nabla^2 g(A) &= \sum_{j=1}^{P} \mathbf{x}_j \mathbf{x}_j^T \\ &= XX^T \end{aligned} \tag{4.7}$$

$$A^k = \left[A^{k-1} - \tfrac{1}{L} X^T (A^{k-1} X^T - B^T)\right]^+ \tag{4.8}$$

where here $L = d_{max}(XX^T)$

5 Robust Face Recognition

$$\min_{\mathbf{x},\mathbf{e}} \mu \| A\mathbf{x} + \mathbf{e} - \mathbf{b}\|_2^2 + \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \tag{5.1}$$

where μ is dependent on ϵ. The subproblem of minimizing over $\mathbf{x}$ is given as

$$\min_{\mathbf{x}} \frac{1}{2} \| A\mathbf{x} - (\mathbf{b} - \mathbf{e})\|_2^2 + \frac{1}{2\mu L_x} \|\mathbf{x}\|_1 \tag{5.2}$$

$$\bigtriangledown^2 f(\mathbf{x}) = A^T A \tag{5.3}$$

$$\mathbf{x}^k = \mathcal{T}_{\frac{1}{2\mu L_x}} (\mathbf{x}^{k-1} - \frac{1}{L_x} A^T [A\mathbf{x} - (\mathbf{b} - \mathbf{e})]) \tag{5.4}$$

$$= \left[\left| \mathbf{x}^{k-1} - \frac{1}{d_{max}(A^T A)} A^T [A\mathbf{x}^{k-1} - (\mathbf{b} - \mathbf{e})] \right| - \frac{1}{2\mu L_x} \right]^+ sign(\mathbf{x}^{k-1} - \frac{1}{d_{max}(A^T A)} A^T [A\mathbf{x}^{k-1} - (\mathbf{b} - \mathbf{e})])$$

where $L_x = d_{max}(A^T A)$.

The subproblem of minimizing over $\mathbf{e}$ is given as

$$\min_{\mathbf{e}} \frac{1}{2} \|\mathbf{e} - (\mathbf{b} - A\mathbf{x})\|_2^2 + \frac{1}{2\mu L_e} \|\mathbf{e}\|_1 \tag{5.5}$$

$$\bigtriangledown^2 f(\mathbf{e}) = I \tag{5.6}$$

$$\mathbf{e}^k = \mathcal{T}_{\frac{1}{2\mu L_e}} (\mathbf{e}^{k-1} - \frac{1}{L_e} [\mathbf{e}^{k-1} - (\mathbf{b} - A\mathbf{x})]) \tag{5.7}$$

$$= \left[\left| \mathbf{b} - A\mathbf{x} \right| - \frac{1}{2\mu} \right]^+ sign(\mathbf{b} - A\mathbf{x})$$

where $L_e = 1$.