Optimization Methods in Machine Learning
Lecture 13: Proximal and Optimal Proximal Gradient Methods

Katya Scheinberg

Lehigh University

Spring 2015

Logistic Loss

Recall the logisitic loss

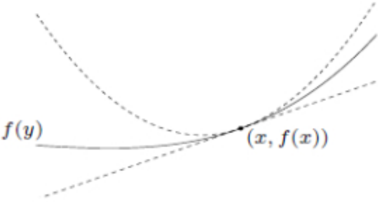
with L_2 norm:

$$\min_w f(w) = \frac{1}{n} \sum \log(1 + e^{-y_i(w^T x_i)}) + \lambda \|w\|_2^2$$

with L_1 norm:

$$\min_w f(w) = \frac{1}{n} \sum \log(1 + e^{-y_i(w^T x_i)}) + \lambda \|w\|_1$$

Proximal Gradient Method



Use quadratic approximation in each iteration

$$w^k = \arg\min_w : f(u^k) + \nabla f(u)^T (w - u^k) + \frac{1}{2\mu_k} \|w - u^k\|^2$$

Also let

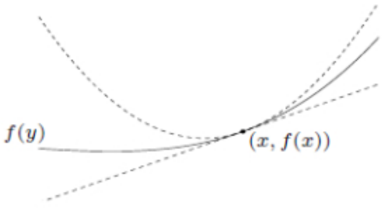
$$q(w) = \frac{1}{2\mu_k} \|w - u^k\|^2$$

Proximal Gradient Method

- Set $u^0 = 0, t_1 = 1$.
- $w^{k+1} = w^k - \mu_k \nabla f(w^k)$
- Use line search to find $\mu_k \geq \mu_{k+1}$ such that $f(w^{k+1}) \leq q(w^{k+1})$

Convergence rate: $\frac{1}{k}$.

Optimal Proximal Gradient Method



Use quadratic approximation in each iteration

$$w^k = \arg\min_w : f(u^k) + \nabla f(u)^T (w - u^k) + \frac{1}{2\mu_k}\|w - u^k\|^2$$

Also let

$$q(w) = \frac{1}{2\mu_k}\|w - u^k\|^2$$

Optimal Proximal Gradient Method

- Set $u^0 = 0, t_1 = 1$.
- $w^k = u^k - \mu_k \nabla f(u^k)$
- Use line search to find $\mu_k \geq \mu_{k+1}$ such that $f(w^{k+1}) \leq q(w^{k+1})$
- $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$
- $u^{k+1} = w^k + \frac{t_k - 1}{t_{k+1}}(w_k - w_{k-1})$

Convergence rate: $\frac{1}{k^2}$.

Comparison

- **Convergence Rate**
  Proximal Gradient Method: $\frac{1}{k}$
  Optimal Proximal Gradient Method: $\frac{1}{k^2}$
  Optimal proximal gradient method is faster than proximal gradient method.
- **Selection of μ_k**
  Selection in optimal proximal gradient method is more restrictive than that in proximal gradient method.

Second Order Method

Let

$$\ell(w) = \frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i)}),$$

so

$$f(w) = \ell(w) + \lambda\|w\|_1.$$

We use the second order term to approximate f,

$$q(w) = \ell(u^k) + \nabla\ell(u^k)^T(w - u^k) + \frac{1}{2}(w - u^k)^T\nabla^2\ell(u^k)(w - u^k) + \lambda\|w\|_1,$$

from where we can use Newton method.

Second Order Method

If the Hessian is expensive to compute, then we can use,

$$q(w) = \ell(u^k) + \nabla \ell(u^k)^T (w - u^k) + \frac{1}{2}(w - u^k)^T \nabla^2 H_k (w - u^k) + \lambda \|w\|_1,$$

where H_k is the Hessian approximation in k^{th} iteration.
Normally, second order method works better than first order method.

General Form

We want to solve the following problem,

$$\min_{x \in \mathbb{R}^n} \quad f(x) + g(x),$$

where $f(x)$ is convex and smooth, $g(x)$ is convex and simple. If the problem

$$\min \quad \frac{1}{2}\|z - y\|^2 + \lambda g(y)$$

is easy, then we can state that $g(y)$ is simple.
If $g(y) = \lambda\|x\|_1$, then

$$y^* = \begin{cases} z - \lambda, \; \textit{if} \; z > \lambda \\ 0, \; \textit{if} \; -\lambda \leq z \leq \lambda \\ z + \lambda, \; \textit{if} \; z < -\lambda \end{cases}$$

General Form

Let

$$q(x) = f(x^k) + \nabla f(x^k)^T(y - x^k) + \frac{1}{2\mu_k}\|y - x^k\|^2 + g(y)$$

The following two problems are equivalent:

$$\min_x \quad q(x).$$

$$\min_x \quad \|x^k - u_k \nabla f(x^k) - y\|^2 + \mu_k g(y)$$

Extensions

Consider $g(y) = \lambda \sum_i \|y^i\|, \forall i$, then the problem

$$\min_y \quad \frac{1}{2}\|z - y\|^2 + \lambda \sum_i \|y^i\|$$

is equivalent to solve for each i separately because all variables y_i are independent, so

$$\min_{y_i} \quad \frac{1}{2}\|z^i - y^i\|^2 + \lambda\|y^i\|,$$

so

$$y^{i*} = \frac{r^i}{\|r^i\|} \max(0, \|r^i\| - \lambda)$$

Extensions

Consider one example that we have a group of identical features and want to study the effect,

$$Ax = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0.01 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = b,$$

where x_1, x_2 are identical.
Obviously, both $(1, 0, 1)^T$ and $(0.5, 0.5, 1.01)$ are solutions, however, if we solve the following problem,

$$\min_x \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|(x_1, x_2)^T\| + \lambda\|x_3\|,$$

we will find the unique optimal solution $(0.5, 0.5, 1.01)$.