Sample Size Selection in Optimization Methods for Machine Learning

Xialiang Dou, Mingrui Zhang, Dongming Huang

22. December 2014

Assumption:

- J is L-Lipchistz and λ-strong convex.
- In every step, $\|g_k - \nabla J(w_k)\| \leq \theta \|g_k\|$

Then,

$$J(w_k) - J^* \leq \frac{1}{\lambda} \|\nabla J(w_k)\|_2^2$$

$$J(w_k) - J^* \geq \frac{\lambda}{2L^2} \|\nabla J(w_k)\|_2^2$$

Without loss of generality, we assume $J^* = 0$. We take step length $\alpha = (1-\theta)/L$.
Then

$$J(w_{k+1}) \leq (1 - \beta\lambda/L)J(w_k)$$

, where $\beta = \frac{1-\theta}{2}$

Proof.
by triangle inequality:

$$\|\nabla J_k\| \leq (1+\theta)g_k, \|\nabla J_k\| \geq (1-\theta)g_k$$

$$\begin{aligned} 2\nabla J_k^T g_k &\geq (1-\theta^2)\|g_k\|^2 + \|\nabla J_k\|^2 \\ &\geq [1-\theta^2-(1-\theta)^2]\|g_k\| \\ J(w_k) &\leq \frac{1}{\lambda}\|\nabla J(w_k)\|_2^2 \text{ (strong convex)} \end{aligned}$$

$$\implies \frac{L}{2}\frac{(1-\theta^2)(1-\theta)}{L^2}\|g_k\|^2 - \frac{1-\theta}{L}\nabla J_k^T g_k \leq -\frac{\lambda(1-\theta)}{2L}J(w_k)$$

$w_{k+1} = w_k - \alpha g_k$, note that $\alpha = (1-\theta)/L$

$$
\begin{aligned}
J(w_{k+1}) &\leq J(w_k) - \alpha \nabla J_k^T g_k + \frac{L}{2}\|\alpha g_k\|^2 \text{ (Lipschitz)} \\
&\leq J(w_k) - \frac{\lambda(1-\theta)}{2L} J(w_k)
\end{aligned}
$$

Assumption:

- $\|Var(\nabla l(w; x_i, y_i))\|_1 \leq \omega$
- $n_k = [a^k]$

Explanation:

$$\begin{aligned} J(w_k) &\geq \frac{\lambda}{2L^2}\|\nabla J(w_k)\|_2^2 \text{ (Lipschitz and strong convex)} \\ (1 - \frac{\beta\lambda}{L})^k J(w_0) &\geq \frac{\lambda}{2L^2}\|\nabla J(w_k)\|_2^2 \\ &\geq C\frac{Var(S_k)}{n_k} \end{aligned}$$

$w_{k+1} = w_k - \frac{1}{L} g_k$, also $E(g_k) = \nabla J_k$

$$
\begin{aligned}
J(w_{k+1}) &\leq J(w_k) - \frac{1}{L}\nabla J(w_k)^T g_k + \frac{1}{2L}\|g_k\|^2 \text{ (Lipschitz)} \\
EJ(w_{k+1}) &\leq J(w_k) - \frac{1}{L}\|\nabla J(w_k)\|^2 + \frac{1}{2L}\|\nabla J(w_k)\|^2 + \frac{1}{2L} Var(g_k) \\
EJ(w_{k+1}) &\leq (1 - \frac{\lambda}{2L}) J(w_k) + \frac{\omega}{2Ln_k}, \text{ note } n_k = [a^k]
\end{aligned}
$$

$$\|Var(g_k)\|_1 \leq \frac{\|Var(\nabla l))\|_1}{n_k}$$

Notice that the expectation are taken condition on $J(w_k)$. Then

$$E(J_{w_{k+1}} - J_{w_k}) < Cp^k, \text{ where } p = max(1 - \lambda/(4L), 1/a)$$

Algorithm Name	Bound	Algorithm Description
Dynamic Sample Gradient Method	$O\left(m\kappa\omega/\lambda\epsilon\right)$	(4.23), (4.24)
Fixed Sample Gradient Method	$O(m^2\kappa\epsilon^{-1/\bar{\alpha}}\log^2\frac{1}{\epsilon})$	(4.32)
Stochastic Gradient Method	$O(m\bar{\nu}\kappa^2/\epsilon)$	(4.33)

Where κ is the condition number $\frac{\lambda}{L}$, m is the problem size.