On the Global Convergence of AdaQN method (draft, unfinished)

Baoqian Zhang
zbqmate@outlook.com

January 27, 2017

Abstract

Adaptive quasi Newton method was proposed by Nitish Shirish Keskar and Albert S. Berahas in 2016. Some effective strategies were introduced in that paper, and we obtained favorable performance on USPS dataset with a simple RNN model. In this draft, we try to prove the global convergence of adaQN. However, in this draft, the proof isn't compeleted. We overlook the step acceptance and control part (i.e line 12 to line 16 in the pseudocode).

1 Proposition 1

In the L-BFGS algorithm, the approximations of inverse Hessian matrices B_k are computed on the following form.

$$B_{t,u+1} = B_{t,u} + \frac{y_{t-r+u}^T y_{t-r+u}}{y_{t-r+u}^T s_{t-r+u}} - \frac{B_{t,u} s_{t-r+u} s_{t-r+u}^T B_{t,u}}{s_{t-r+u}^T B_{t,u} s_{t-r+u}} \tag{1}$$

The idea in L-BFGS is that the use of curvature information pairs is restricted to the last r pairs. this restriction is expected to receive little performance penalty for the notion that the previous $t - r$ iterats are likely to carry little curvature information. Here we propose that for a given vector $p = p_0$, we define the sequence of vectors p_k through the following recurtion.

$$p_{u+1} = p_u - \frac{s_{t-u-1}^T p_u y_{t-u-1}}{s_{t-u-1}^T y_{t-u-1}} \tag{2}$$

And now we define the sequence of vectors q_k with initial value $q_0 = B_{t,0}^{-1} p_r$

$$q_{u+1} = q_u + (\frac{s_{t-r+u}^T p_{r-u-1}}{s_{t-r+u}^T y_{t-r+u}} - \frac{y_{t-r+u}^T q_u}{s_{t-r+u}^T y_{t-r+u}}) s_{t-r+u} \tag{3}$$

Then the product $B_t q_r = p$

2 Proof of proposition 1

See the proof of two-loop algorithm in [**?**].

3 Assumption 1

In the convex set S, exist a constant $L > 0$ and Lipschitz condition is satisfied,

$$\| \bigtriangledown f(x) - \bigtriangledown f(y)\| \leq \|x - y\|, \forall x, y \in S \tag{4}$$

In [**?**], it suggests that empirical Fisher Information Matrix is a close approximation of Hessian matrix. Then the update of the approximations of inverse-Hessian matrices B_k can be written as

$$B_{k+1} = \begin{cases} B_k + \frac{y_k^T y_k}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \text{if} y_k^T s_k > \epsilon \cdot s_k^T s_k, \\ B_k \text{otherwise.} \end{cases} \tag{5}$$

For a suffient small positive α used in adaQN algorithm

$$-\sum_{k=0}^{\infty} \bigtriangledown f(x_k)^T s_k < \infty \tag{6}$$

and now we try to proov

$$\lim_{k \to \infty} \| \bigtriangledown f(x_k)\| = 0 \tag{7}$$

4 Theorem 1

If Assumption 1 is satisfied and there are finite number of cases that

$$y_k^T s_k > \epsilon \cdot s_k^T s_k, \tag{8}$$

then

$$\lim_{k \to \infty} \| \bigtriangledown f(x_k)\| = 0 \tag{9}$$

5 Proof of theorem 1

For all k in the finate set that satisfy the condition

$$y_k^T s_k > \epsilon \cdot s_k^T s_k, \tag{10}$$

we choose the largest k_m and B_{k_m} has the property that for all $k > k_m$, $B_k = B_{k_m}$. $B := B_{k_m}$, by

$$-\sum_{k=0}^{\infty} \bigtriangledown f(x_k)^T s_k < \infty, \tag{11}$$

we have

$$-\lim_{k\to\infty} \bigtriangledown f(x_k)^T s_k = 0 \tag{12}$$

and

$$-\lim_{k\to\infty} \alpha \bigtriangledown f{x_k}^T B \bigtriangledown f x_k = -\lim_{k\to\infty} \bigtriangledown f(x_k)^T s_k \tag{13}$$

Since α is a positive constant and B is positive definite, we can safely reach the conclusion of theorem 1.

6 Proposition 2

If Assumption 1 holds, and $\| \bigtriangledown f(x_k)\| \geq \delta$, then we have

$$\sum_{i=1}^{w} \frac{\|B_i s_i\|}{s_i^T B_i s_i} < wM. \tag{14}$$

Where M is a positive constant and w is the number of cases that satisfy the condition $y_k^T s_k > \epsilon \cdot s_k^T s_k$.

7 Proof of proposition 2

By $y_k^T s_k > \epsilon \cdot s_k^T s_k$,

$$\frac{1}{y_i^T s_i} < \frac{1}{\epsilon \cdot \|s_i\|^2} \tag{15}$$

and by Lipschitz condition,

$$L\|s_i\| \geq \|y_i\| \tag{16}$$

we get

$$\frac{\|y_i\|^2}{y_i^T s_i} < \frac{L^2}{\epsilon} := M. \tag{17}$$

Then proposition 2 is prooved.

8 Theorem 2

If Assumption 1 holds and there are infinite number of cases that

$$y_k^T s_k > \epsilon \cdot s_k^T s_k, \tag{18}$$

then

$$\lim_{k\to\infty} \| \bigtriangledown f(x_k)\| = 0 \tag{19}$$

9 Proof of theorem 2

Similar to theorem 3.3 in [**?**], because

$$B_i s_i = -\alpha \bigtriangledown f(x_i) \tag{20}$$

then

$$\sum_{i=0}^{\infty} \alpha \| \bigtriangledown f(x_i)\|^2 \frac{s_i^T B_i s_i}{\|B_i s_i\|^2} = -\sum_{i=0}^{\infty} \bigtriangledown f(x_i)^T s_i < \infty. \tag{21}$$

Suppose $\lim_{k\to\infty} \| \bigtriangledown f(x_k)\| = 0$ does not hold,

$$\sum_{i=0}^{\infty} \alpha \frac{s_i^T B_i s_i}{\|B_i s_i\|^2} < \infty \tag{22}$$

There exist a positive number ζ and a positive interger j,

$$(\prod_{i=j+1}^{j+q} \alpha \frac{s_i^T B_i s_i}{\|B_i s_i\|^2})^{\frac{1}{q}} \leq \frac{1}{q} \sum_{1=j+1}^{j+q} \alpha \frac{s_i^T B_i s_i}{\|B_i s_i\|^2} \leq \frac{\zeta}{q} \tag{23}$$

$$\begin{aligned} (\prod_{i=j+1}^{j+q} \alpha)^{\frac{1}{q}} &\leq \frac{\zeta}{q} \\ &\leq \frac{\zeta}{q} (\prod_{i=j+1}^{j+q} \frac{\|B_i s_i\|^2}{s_i^T B_i s_i})^{\frac{1}{q}} \\ &\leq \frac{\zeta}{q^2} \sum_{i=j+1}^{j+q} \frac{\|B_i s_i\|^2}{s_i^T B_i s_i} \\ &\leq \frac{\zeta}{q^2} \sum_{i=0}^{j+q} \frac{\|B_i s_i\|^2}{s_i^T B_i s_i} \end{aligned} \tag{24}$$

By proposition 2,

$$(\prod_{i=j+1}^{j+q} \alpha)^{\frac{1}{q}} \leq \frac{\zeta}{q^2}(j+q+1)M \tag{25}$$

Because α is a positive constant, when q is suffiently large, that inequality cannot hold. Thus theorem 2 is prooved. The convergence when $\lim_{k\to\infty} \| \bigtriangledown f(x_k)\| = 0$ is stated in [**?**] , and the proof of superlinear convergence of this algorithm seems plausible by adopting the method in it.

References

[1] Aryan Mokhtari, Alejandro Rebeiro, *Global Convergence of Online Limited Memory BFGS*. 2014.

[2] Shun-ichi Amari, *Nature Gradient Works Efficiently in Learning*. 1998.

[3] Masao Fukushima, *On the Global Convergence of BFGS Method for Nonconvex Unconstrained Optimization Problems*. 2013.

[4] Dimitri P. Bertsekas, *Nonlinear Programming*.