Inexact Orhant-Wise Quasi-Newton Method

Faguo Wu1,2, Wang Yao1,2, Xiao Zhang1,2,*, Chenxu Wang3 and Zhiming Zheng1,2
1Key Laboratory of Mathematics, Informatics and Behavioral Semantics, Ministry of Education, School of Mathematics and
Systems Science, Beihang University, Beijing, 100191, China.
2Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, 100191, China.
3Big Data Management Department, Information Technology Department China Minsheng Bank
*Corresponding author

Abstract—The Orhant-Wise Limited-memory Quasi-Newton method (OWL-QN), based on the L-BFGS method, is an
effective algorithm for solving the ℓ1-regularized sparse learning problem. In order to deal with the ℓ1-regularization, OWL-QN
restrict the point to an orthant on which the quadratic model is valid and differentiable. In this paper, we propose an Inexact
Orhant-Wise Limited-memory Quasi-Newton method (IOWL-QN). This method, at every iteration, compute an approximate
solution satisfied the inexactness conditions to estimate the exact solution. We give brief proof to the convergence and report the
numerical results.

Keywords—orthant-based; sparse optimization; inexact Newton; proximal gradient

I. INTRODUCTION

The ℓ1-regularized optimization have been applied to
many fields including image deburring, face recognition, linear
and logistic regression. We consider the following problem:

$$\min_{x \in \mathbb{R}^n} \phi(x) = f(x) + \mu \|x\|_1$$  \hspace{1cm} (1)

where $\mu$ is a given constant, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex,
bounded below, continuously differentiable, and the gradient $\nabla f$ is LLipschitz continuous on the set $\{x : f(x) \leq f(x_0)\}$
for some L and some initial point $x_0$. The ℓ1 regularize has
many favorable properties, especially it produces sparse
vectors. But it is not differentiable at zero, that means we
cannot use the gradient based method to solve problem (1).
However, many algorithms have been developed to overcome
this obstacle. The first family of methods, called first-order
algorithms[2,3,7,8,17,26], just take charge of the objective
function of the first order derivative information. Although
problem(1) is non-smooth, but its sub differential is easy to
compute and has been applied in various first order methods
such as subgradient method[16,18], proximal gradient method
ISTA[7,8,17], fast proximal gradient method(FISTA)[2,3]
and let $\mu$ denote the Euclidean norm.

The second family of methods, called second-order
algorithms[12,15,25,28], usually need the Hessian matrix or
approximated Hessian matrix to construct a quadratic model
subproblem, and then solve the subproblem at each iteration.
These methods usually have high computation cost, but can
achieve linear or super-linear convergence rates. Unlike these
second-order methods, the Orhant-Wise Limited-memory
Quasi-Newton method(OWL-QN)[1] restricts the objective
function to a special orthant, generalizes the objective
function to a differentiable one. OWL-QN needs only matrix-vector
multiplications since it use L-BFGS[20] method to
form the inverse of the Hessian matrix. Although OWL-QN
has been proved very effective in practice, but no convergence
analysis was provided[4,22,29]. Pinghua Gong and Jieping
Ye [13] proposed a modified Orhant-Wise Limited Memory
Quasi-Newton Method (mOWL-QN), which establish a
detailed convergence analysis for the OWL-QN type
algorithm and also have similarly convergence as the OWL-QN.
Lee, Sun and Saunders [28] presented an inexact
proximal Newton method to solve problem(1) and establish
several local convergence results. Byrd, Nocedal and
Oztoprak [5] proposed an inexact successive quadratic
approximation method(SQA), this method also use a quadratic
model to approximate the objective function. Instead of
compute the exact solution to the quadratic model, SQA
compute an approximate solution satisfying some inexactness
condition at every iteration. Other relevant inexact algorithms
is describe in [19,21,23].

Motivated by the inexactness condition, we combine it
with the OWL-QN method, that is the Inexact Orhant-Wise
Quasi-Newton Method(IOWL-QN). This paper can be divided
into four sections. In Section 2, we describe the OWL-QN
algorithm and the inexactness condition, and after detail the
IOWL-QN algorithm. In Section 3, we introduce the local and
global convergence analysis of the algorithm. In Section 4, we
report the numeric experiments.

Notation In the remainder, we let $g(x_k) = \nabla f(x_k)$, and let $\|\|$ denote the Euclidean norm.

II. PRELIMINARIES

A. L-BFGS Method

Before discussing the orthant-based methods, we give a
short introduction to the quasi-Newton method (QN), this
method is designed to solve the unconstrained optimization of a
smooth function:
quasi-Newton method, like Newton method, iteratively construct a local quadratic approximation to the objective function, but it require only the gradient of the objective function. The most popular quasi-Newton algorithm is the BFGS (Broyden, Fletcher, Goldfarb and Shanno) method [20]. We form the following quadratic model of the objective function at $x_k$:

$$q_k(d) = f(x_k) + g(x_k)^T d + \frac{1}{2} d^T B_k d$$

(3)

The minimizer $d_k$ of this quadratic model is the search direction, and the new iterate is:

$$x_{k+1} = x_k + \alpha_k d_k = x_k - \alpha_k H_k g_k$$

(4)

where $\alpha_k$ is the step length, and $H_k = B_k^{-1}$ is updated at every iteration:

$$s_k = x_{k+1} - x_k, \quad y_k = g(x_{k+1}) - g(x_k)$$

(5)

After the new iterate is completed, the oldest vector pair is replaced by the new pair $\{s_k, y_k\}$. OWL-QN Method

Andrew and Gao [1] proposed an Orthant-Wise Quasi-Newton Method (OWL-QN), which is used to solve the following optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) + r(x)$$

(6)

since the $\ell_1$ regularizer has many (favorable) properties such as sparsity, we consider the following problem:

$$\min_{x \in \mathbb{R}^n} \phi(x) = f(x) + \mu \|x\|_1$$

(7)

where $\mu$ is a given constant, $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, bounded below, continuously differentiable, and the gradient $\nabla f$ is Lipschitz continuous on the set $\{x: f(x) \leq f(x_0)\}$ for some $L$ and some initial point $x_0$.

The steepest descent direction of $\phi(x)$ at $x$, denoted by $-\hat{\nabla} \phi(x)$, is defined as the subgradient with the least norm at point $x$, where

$$\hat{\nabla} \phi(x) = \arg \min_{g \in \hat{\nabla} \phi(x)} \|g\|$$

and hence $x$ is a global minimizer of (5) if and only if $\hat{\nabla} \phi(x) = 0$. For a given sign vector $\xi \in \{-1, 0, 1\}^n$, define

$$\Omega_\xi = \{x \in \mathbb{R}^n : \text{sign}(x_i) = \text{sign}(\xi_i), i = 1, \ldots, n\}$$

we can see that

$$\phi(x) = f(x) + \mu \xi^T x, \forall x \in \Omega_\xi$$

is differentiable on $\Omega_\xi$, and we extend $\phi(x)$ to $\mathbb{R}^n$ denoted as $\hat{\phi}_\xi$, then $\hat{\phi}_\xi$ is a differentiable function on $\mathbb{R}^n$. To explore the reasonable orthant face at iterate $x_k$, let us consider a small step-size along the steepest descent direction $-\hat{\nabla} \phi(x)$, and it defines the orthant:

$$\Omega_\xi = \{d \in \mathbb{R}^n : \text{sign}(d_i) = \text{sign}(\xi_i), i = 1, \ldots, n\}$$

then the objective function in problem (5) is approximated by a quadratic function

$$\min_{d \in \mathbb{R}^n} q_k(d) = f(x_k) + \hat{\nabla} \phi(x_k)^T d + \frac{1}{2} d^T B_k d$$

(8)

where $B_k$ is the Hessian matrix at $x_k$, then we can get a direction which is the solution of problem (7)

$$d_k = \arg \min_{d \in \mathbb{R}^n} q_k(d) = -H_k^{-1} \hat{\nabla} \phi(x_k)$$

here $H_k$ is the inverse of Hessian matrix $B_k$, and we use the L-BFGS method to get the approximate inverse matrix cite [book03]. In order to ensure the search point do not leave $\Omega_\xi$, we project it back onto $\Omega_\xi^k$ at each iteration, that is

$$x_{k+1} = \pi(x_k + \alpha d_k; \xi_k)$$

and use the backtracking line search to choose step size $\alpha$, choose $\beta, \gamma \in (0, 1)$ and for $n = 0, 1, 2, \ldots$, accept the first step size $\alpha = \beta^n$ such that

$$f\left(\pi(x + \alpha d_k; \xi_k)\right) \leq f(x_k) - \gamma \nu_k^T (\pi(x + \alpha d_k; \xi_k) - x_k)$$
III. INEXACT OWL-QN

A. Inexactness Condition

Byrd, Nocedal and Oztoprak [5] proposed an Inexact Newton-like Method, we also call it proximal Newton Method. This method compute an inexact solution of the piecewise quadratic model satisfying some inexactness conditions at every iteration, consider the following optimization problem:

\[ \min_{x \in \mathbb{R}^n} \phi(x) = f(x) + \mu \|x\|_1 \]

First, we consider the smooth unconstrained case (\( \mu = 0 \)):

\[ \min_{x \in \mathbb{R}^n} f(x) \]

for this problem, Dembo, Eisenstat and Steihaug [9] propose an inexact newton method, compute an approximate solution \( \hat{x} \) at each iteration, satisfy the condition:

\[ \| g(x_k) + B_k (\hat{x} - x_k) \| \leq \eta_k \| g(x_k) \|, \quad 0 < \eta_k < 1 \]

Motivate by this, we consider the following inexactness condition for the unsmooth case (\( \mu \neq 0 \)):

\[ \| F_q(x_k; \hat{x}) \| \leq \eta_k \| F_q(x_k; \hat{x}) \|, \quad 0 < \eta_k < 1 \]

B. Termination Condition

For problem (9), the soft thresholding step \( x_{\text{ista}} \) [5] is a proper estimate, that is

\[ x_{\text{ista}} = \arg \min_{x} g(x) + \|x - x_k\| + \frac{1}{2\tau} \|x - x_k\|^2 + \mu \|x\|_1 \]

in order to obtain an proper measure of the optimality, we introduce the following lemma:

Lemma 3.1 \( x_k \) is a solution of problem (9) if and only if

\[ x_k = x_{\text{ista}} \]

Proof. By [24], we know that \( x_{\text{ista}} \) a solution of problem (9) if and only if

\[ x_k = \text{prox}_{\mu ||\cdot||_1}(x_k - \tau g(x_k)) \]

\[ \triangleq x_{\text{ista}} \]

from Lemma 2.1 we know that \( x_k = x_{\text{ista}} \) is a measure of the optimality, but it is hard to test, so we introduce another Lemma:

Lemma 3.2 Define \( F(x) = g(x) - P_{[-\mu, \mu]}(g(x) - x / \tau) \), where \( P_{[-\mu, \mu]}(\cdot) \) denotes the component t-wise projection of \( x \) onto the interval \([-\mu, \mu]\)

IV. CONVERGENCE

Based on above analysis, we consider the convergence of this algorithm. We omit the proof since these convergence analysis can be directly obtained from [5].

A. Global Convergence

Theorem 4.1 Suppose that \( f \) is a smooth function that is bounded below and with Lipschitz continuous gradient i.e. there is a constant \( M > 0 \) such that

\[ \| g(x) - g(y) \| \leq M \| x - y \| \]

for all \( x, y \). Let \( \{x_k\} \) be the sequence of iterates generated by Algorithm 2, and suppose that there exist constants \( 0 < \lambda \leq \Lambda \) such that the sequence \( \{B_k\} \) satisfies

\[ \lambda_{\text{min}}(B_k) \geq \lambda > 0 \text{and} \lambda_{\text{max}}(B_k) \leq \Lambda \]

for all \( k \). Then \( \lim_{k \to \infty} F(x_k) = 0 \)

where \( \lambda_{\text{min}}(B_k) \) and \( \lambda_{\text{max}}(B_k) \) are the smallest and the largest eigenvalues of \( B_k \).

B. Local Convergence

Theorem 4.2 If \( \nabla^2 f(x) \) is Lipschitz continuous and positive define at \( x^* \), \( \tau < 1/\| \nabla^2 f(x^*) \| \)

Then there is a neighborhood of \( x^* \) such that, if \( x_0 \) lies in that neighborhood, the iteration that defines \( x_{k+1} \) as the unique solution to

\[ F_q(x_k; x_{k+1}) = 0 \]

converges quadratic ally to \( x^* \).

Theorem 4.3 Suppose that \( \nabla^2 f(x) \) is Lipschitz continuous and positive define at \( x^* \), \( \tau < 1/\| \nabla^2 f(x^*) \| \), and the
function $F'_k(x_k; y)$ is defined by \eqref{eq07}, and that $x_{k+1}$ is computed by solving

$$F'_k(x_k; x_{k+1}) = r_k$$

(1) if $\eta_k \leq \eta \forall k$ and if $x_0 \in \mathcal{N}$ then the sequence \{x_k\} converges Q-linearly to $x^*$;

(2) In addition if $\eta_k \to 0$, then the convergence rate of \{x_k\} is Q-superlinear;

(3) if for some $\eta$, $\eta_k \leq \eta \|F(x_k)\|$ then the convergence rate is Q-quadratic.

V. NUMERICAL RESULTS

A. Algorithm for Solving the Subproblem

$$\min_{x \in \mathbb{R}^*} \phi(x) = f(x) + \mu \|x\|_1$$

(1) TFOCS: we get the exact solution by TFOCS package, N83 solver.

(2) FISTA: this is the FISTA\[9\] algorithm applied to original problem(5).

(3) OWL-QN: we use the OWL-QN method to solve the original problem(5) directly, in which we use the L-BFGS method to get the approximate inverse of Hessian matrix, with memory parameter mem.

(4) IOWL-QN: This is the IOWL-QN method, we use the OWL-QN method to solve the subproblem (11), In which we use the L-BFGS method to get the approximate inverse of Hessian matrix, with memory parameter mem.

B. Logistic Regression Problems

In our numerical experiments, the function $f(x)$ in

$$\min_{x \in \mathbb{R}^*} \phi(x) = f(x) + \mu \|x\|_1$$

is given by a logistic function

$$\min_{x \in \mathbb{R}^*} \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i x^T z_j)) + \mu \|x\|_1$$

C. Result

All the algorithms are implemented in Matlab, the initial point was produced randomly in all experiments, and the iteration was terminated if

$$\|F(x_k)\|_2 \leq TOL$$

where F is defined in Lemma3.2 :

$$F(x) = g(x) - P_{\Omega_{\mu, \rho}}(g(x) - x / \tau)$$

Maximum number of outer iterations was 3000, in the OWL and IOWL-QN method, the parameter $\eta_k = \max{1 / k, 0.1}$, and we set $\theta = 0.1$. We choose the TFOCS solution as the exact solution, and employ three levels of accuracy TOL: $10^{-6}$, $10^{-8}$ and $10^{-10}$.

The numerical results are presented in Table 2 and Table 3. The objective function value are reported in Figure 1 and 3. The termination condition $\|F(x)\|_2$ in are reported in Figure 2 and 4. We observe from the numerical results that, both the IOWL-QN method and the OWL-QN method perform the best. The OWL-QN method requires the least number of total iterations, though a convergence proof has not been established so far, it works well in the practice. The IOWL-QN method cost less CPU time and outer iterations, and almost always accepts the unit step length ($\alpha = 1$). The FISTA method does not perform very well in the test, mainly because FISTA is not efficient in this log regression problem.

VI. CONCLUSION

In this paper, we presented an algorithm IOWL-QN based on the well-known OWL-QN method. Since the convergence proof of OWL-QN has not been established yet, we provide a brief convergence analysis to the IOWL-QN. We provide numeric results for solving the logistic regression problem, and found that the IOWL-QN was obviously faster than the FISTA for this problem. There are several directions that we can explore in the future. It would be interesting to explore other inexactness conditions. Another direction to explore would be extend this method to other $\ell_1$-regularized problems.

ACKNOWLEDGMENT

This work was supported by the Major Program of National Natural Science Foundation of China (11290141).

REFERENCES


