Advances in Intelligent Systems Research, volume 143
2nd International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2018)

Inexact Orthant-Wise Quasi-Newton Method

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

***Abstract*—The Orthant-Wise Limited-memory Quasi-Newton method (OWL-QN), based on the L-BFGS method, is an effective algorithm for solving the ‘1-regularized sparse learning problem. In order to deal with the ‘1-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 Orthant-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 gadient

I. INTRODUCTION

The ℓ_1 -regularized optimization have been applied to many fields including image deburring, face recognition, linear and logistic regression. We consider the following problem:

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

where μ is a given constant, $f:\mathbb{R}^n \mapsto \mathbb{R}$ is convex, bounded below, continuously differentiable, and the gradient ∇f is LLipschitz continuous on the set $\{x: f(x) \le f(x_0)\}$ for some L and some initial point x_0. The ℓ_1 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] etc. These algorithms have low computation cost at each iteration, and can usually achieve sub-linear convergence rates. Other relevant algorithms is describe in [10, 11, 27].

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 Orthant-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 Orthant-Wise Limited Memory Quasi-Newton Method (mOWL-QN), which establish a detailed convergence analysis for the OWK-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 Orthant-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 $\|\cdot\|$ 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:

$$\min_{x\in\mathbb{R}^n} f(x) \tag{2}$$

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 :

$$q_k(d) = f(x_k) + g(x_k)^{\mathrm{T}} d + \frac{1}{2} d^{\mathrm{T}} B_k d \tag{3}$$

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

$$x_{k+1} = x_k + \alpha_k d_k = x_k - \alpha_k H_k g_k \tag{4}$$

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

$$s_k = x_{k+1} - x_k, \quad y_k = g(x_{k+1}) - g(x_k) \tag{5}$$

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:

$$\min_{x\in\mathbb{R}^n} \phi(x) = f(x) + r(x) \tag{6}$$

since the ℓ_1 regularizer has many (favorable) properties such as sparsity, we consider the following problem:

$$\min_{x\in\mathbb{R}^n} \phi(x) = f(x) + \mu \| x\|_1 \tag{7}$$

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

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

$$\Diamond\phi(x) = \arg\min_{g\in\partial\phi(x)} \| g\|$$

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

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

we can see that

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

is differentiable on Ω_ξ, and we extend $\phi(x)$ to $\mathbb{R}^n$ denoted as ϕ_ξ, then ϕ_ξ 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 $-\Diamond\phi(x)$, and it defines the orthant:

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

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

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

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

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

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 Ω_ξ, we project it back onto Ω_ξ^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 α, choose $\beta, \gamma \in (0,1)$ and for $n = 0,1,2,\cdots,$ accept the first step size $\alpha = \beta^n$ such that

$$f\left(\pi\left(x + \alpha d_k; \xi_k\right)\right) \le f(x_k) - \gamma v_k^{\mathrm{T}} \left(\pi(x + \alpha d_k; \xi_k) - x_k\right)$$

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:

$$\| g(x_k) + B_k(\hat{x} - x_k)\| \le \eta_k \| g(x_k)\|, \quad 0 < \eta_k < 1$$

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

$$\| F_q(x_k; \hat{x})\| \le \eta_k \| F_q(x_k; \hat{x})\|, \quad 0 < \eta_k < 1$$

B. Termination Condition

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

$$x_{\text{ista}} = \arg\min_x g(x_k)^{\mathrm{T}}(x - x_k) + \frac{1}{2\tau}\| x - x_k\|^2 + \mu\| x\|_1$$

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_{ista} a solution of problem (9) if and only if

$$x_k = \text{prox}_{\mu\|\cdot\|_1}(x_k - \tau g(x_k))$$
$$\doteq x_{\text{ista}}$$

from Lemma2.1 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

$$\| g(x) - g(y)\| \le M\| x - y\|$$

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

$$\lambda_{\min}(B_k) \ge \lambda > 0 \text{?and?} \lambda_{\max} \le \Lambda$$

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

where $\lambda_{\min}(B_k)$ and $\lambda_{\max}(B_k)$ 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 $F_q(x_k; y)$ is defined by \eqref{eq07}, and that x_{k+1} is computed by solving

$$F_q(x_k; x_{k+1}) = r_k$$

(1) if $\eta_k \leq \bar{\eta}$ for all 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 $\tilde{\eta}$, $\eta_k \leq \tilde{\eta} \| F(x_k) \|$ then the convergence rate is Q-quadratic.

V. NUMERICAL RESULTS

A. Algorithm for Solving the Subproblem

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

(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

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

is given by a logistic function

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

C. Result

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

$$\| F(x_k) \|_\infty \leq \mathrm{TOL}$$

where F is defined in **Lemma3.2** :

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

Maximum number of outer iterations was 3000, in the OWL and IOWL 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^{-4}, 10^{-6} and 10^{-8}.

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 $\| F(x) \|_\infty$ 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 ℓ_1-regularized problems.

ACKNOWLEDGMENT

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

REFERENCES

[1] Andrew G, Gao J, Scalable training of L1 -regularized log-linear models, Proceedingsof the 24th International Conference on Machine Learning. ACM, 33-40 (2007)

[2] Beck A, Teboulle M, A fast iterative shrinkage-thresholding algorithm for linear inverseproblems, SIAM Journal on Imaging Sciences, 2(1):183-202 (2009)

[3] Beck A, Teboulle M, Gradient-based algorithms with applications to signal recovery,in: Y. Eldar and D. Palomar(Eds.), Convex Optimization in Signal Processing and Communications(2009)

[4] Byrd R H, Chin G M, Nocedal J, and Oztoprak F, A family of second-order methodsfor convex ‘1 regularized optimization. Technical report, Industrial Engineering andManagement Sciences, Northwestern University, Evanston, IL (2012)

[5] Byrd R H, Nocedal J, Oztoprak F, An inexact successive quadratic approximationmethod for L-1 regularized optimization, Mathematical Programming(2015)

[6] Byrd R H, Nocedal J, Schnabel R B, Representation of quasi-Newton matrices andtheir use in limited memory methods, Computer Science(1992)

[7] Combettes P L, Va A, Wajs E R, Signal Recovery by Proximal Forward-BackwardSplitting, Siam Journal on Multiscale Modeling & Simulation, 4(4):1168-1200, (2005)

[8] Combettes P L, Pesquet j Ch, Proximal splitting methods in signal processsing, Fixed-Point Algorithms for Inverse Problems in Science and Engineering (2011)

[9] Dembo R S, Eisenstat S C, Steihaug T, Inexact-Newton methods, SIAM J. Numer.Anal. 19(2), 400-408 (1982)

[10] Donoho D L, De-noising by soft-thresholding, Information Theory, IEEE Transactionson, 41(3):613-627 (1995)

[11] Duchi J, Singer Y, Effcient online and batch learning using forward backward splitting,The Journal of Machine Learning Research, 10:2899-2934 (2009)

[12] Facchinei F, Pang J S, Finite-dimensional variational inequalities and complementarity problems, volume 1, Springer Verlag (2003)

[13] Pinghua Gong, Jieping Ye, A Modified Orthant-Wise Limited Memory Quasi-Newton Method with Convergence Analysis, Proceedings of The 32nd International Conference on Machine Learning, pp. 276-284 (2015)

[14] Lee J D, Sun Y, Saunders M A. Proximal Newton-type methods for minimizing composite functions[J]. Siam Journal on Optimization, 2012, 24.

[15] Miffin R, Semismooth and semiconvex functions in constrained optimization, SIAMNumber footnotes separately in superscripts. Journal on Control and Optimization, 15(6):959-972 (1977)

[16] Nesterov, IU. E, Introductory lectures on convex optimization: basic course, KluwerAcademic (2004)

[17] Parikh N, Boyd S, Proximal Algorithms, Foundations & Trends in Optimization,1(3):127-239 (2013)

[18] Polyak B T, Introduction to Optimization, Optimization Software, section 5.1 (1987)

[19] Le Roux, N., Schmidt, M.W., Bach, F., Convergence rates of inexact proximal-gradient methods for convex optimization. In: NIPS, pp. 1458-1466 (2011)

[20] Nocedal J, Wright S J, Numerical optimization, Springer (1999)

[21] S Salzo, S Villa, Inexact and accelerated proximal point algorithms. J. Convex Anal.19(4), 1167-1192 (2012)

[22] Schmidt M, Berg E, Friedlander M, and Murphy K, Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In International Conference on Artificial Intelligence and Statistics (2009)

[23] Tappenden R, Richtarik P, Gondzio J, Inexact coordinate descent: complexity and preconditioning, arXiv preprint arXiv:1304.5530 (2013)

[24] Parikh N, Boyd S, Proximal Algorithms, Foundations & Trends in Optimization,1(3):127-239 (2014)

[25] Tseng P, Yun S, A coordinate gradient descent method for nonsmooth separable minimization, Mathematical Programming, 117(1-2):387-423 (2009)

[26] Wright S J, Nowak R D, and Figueiredo M, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing, 57(7):2479-2493 (2009)

[27] Wright S J, Nowak R D, and Figueiredo M A T, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing, 57(7):2479-2493 (2009)

[28] Friedman J, Hastie T, Tibshirani R, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software, 33(1):1 (2010)

[29] Yuan G, Chang K, Hsieh C, and Lin C, A comparison of optimization methods and software for large-scale ‘1-regularized linear classification. The Journal of Machine Le arning Research, 11:3183-3234 (201