A Modified Orthant-Wise Limited Memory Quasi-Newton Method with Convergence Analysis

Pinghua Gong GONGP@UMICH.EDU
University of Michigan, Ann Arbor, MI 48109

Jieping Ye JPYE@UMICH.EDU
University of Michigan, Ann Arbor, MI 48109

Abstract

The Orthant-Wise Limited memory Quasi-Newton (OWL-QN) method has been demonstrated to be very effective in solving the ℓ_1-regularized sparse learning problem. OWL-QN extends the L-BFGS from solving unconstrained smooth optimization problems to ℓ_1-regularized (non-smooth) sparse learning problems. At each iteration, OWL-QN does not involve any ℓ_1-regularized quadratic optimization subproblem and only requires matrix-vector multiplications without an explicit use of the (inverse) Hessian matrix, which enables OWL-QN to tackle large-scale problems efficiently. Although many empirical studies have demonstrated that OWL-QN works quite well in practice, several recent papers point out that the existing convergence proof of OWL-QN is flawed and a rigorous convergence analysis for OWL-QN still remains to be established. In this paper, we propose a modified Orthant-Wise Limited memory Quasi-Newton (mOWL-QN) algorithm by slightly modifying the OWL-QN algorithm. As the main technical contribution of this paper, we establish a rigorous convergence proof for the mOWL-QN algorithm. To the best of our knowledge, our work fills the theoretical gap by providing the first rigorous convergence proof for the OWL-QN-type algorithm on solving ℓ_1-regularized sparse learning problems. We also provide empirical studies to show that mOWL-QN works well and is as efficient as OWL-QN.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

Sparse learning with ℓ_1-regularized optimization problems have been extensively studied (Tibshirani, 1996; Efron et al., 2004) and have been successfully applied to many applications including face recognition (Wright et al., 2008), gene selection (Shevade & Keerthi, 2003), signal recovery (Figueiredo et al., 2007; Wright et al., 2009) and image deburring (Bioucas-Dias & Figueiredo, 2007; Beck & Teboulle, 2009). Due to the wide use of the ℓ_1-regularized sparse learning problem, many algorithms have been developed to speedup the optimization. The existing optimization algorithms can be roughly classified into two categories. The first family of algorithms, called first-order algorithms (Beck & Teboulle, 2009; Wright et al., 2009), mainly utilize the gradient or sub-gradient to solve the problem. The first-order algorithms have low computational cost per iteration but usually achieve sub-linear convergence rates. The second family of algorithms, called (quasi-) Newton algorithms (Tseng & Yun, 2009; Friedman et al., 2010), rely on the (approximated) Hessian matrix to achieve linear or super-linear convergence rates but commonly have high computational cost per iteration. In general, (quasi-) Newton algorithms need to estimate the exact or approximated (inverse) Hessian and solve an ℓ_1-regularized quadratic optimization subproblem at each iteration (Schmidt et al., 2009; Friedman et al., 2010; Yuan et al., 2012). Moreover, the subproblem usually does not admit a closed-form solution. Thus, a specific iterative algorithm needs to be designed to solve the subproblem.

Different from the (quasi-) Newton framework presented in Tseng & Yun (2009), the Orthant-Wise Limited memory Quasi-Newton (OWL-QN) method (Andrew & Gao, 2007) does not involve an ℓ_1-regularized quadratic optimization subproblem at each iteration. It generalizes the limited memory quasi-Newton method (Jorge & Stephen, 1999) from solving the unconstrained smooth optimization problem to the ℓ_1-regularized (non-smooth) optimiza-

tion problem; it only involves matrix-vector multiplications without explicitly forming the (inverse) Hessian matrix, making OWL-QN applicable for large-scale problems. Furthermore, OWL-QN is very simple and easy to implement based on the L-BFGS (Jorge & Stephen, 1999) and is often included as a state-of-the-art method for comparison in the ℓ_1-regularized sparse learning problem (Schmidt et al., 2009; Yu et al., 2010; Yuan et al., 2010; Byrd et al., 2012b;a). Yuan et al. (2010) performed an extensive empirical comparison of existing algorithms for solving ℓ_1-regularized problems. Experimental results show that OWL-QN is comparable to the state-of-the-art algorithms. Moreover, Olsen et al. (2012) developed an OWL-QN-type algorithm to solve the sparse inverse covariance estimation problem, achieving a very competitive result, though no convergence analysis was provided.

Although OWL-QN works quite well in practice, several recent papers (Yu et al., 2010; Yuan et al., 2010; Schmidt et al., 2011; Byrd et al., 2012b) pointed out that the convergence proof is flawed. In particular, Schmidt et al. (2011) presented the following comment on the efficiency and the convergence of the OWL-QN algorithm: "*Thus, while the algorithms of Andrew & Gao (2007) and Section 1.5.1 appear to be very effective in practice, it remains to show whether they are globally convergent in general without additional assumptions*." Byrd et al. (2013) also presented a similar comment as follows: "*Although this algorithm*[1] *performed reliably in our tests, its convergence has not been proved*". A key difficulty of the convergence analysis for OWL-QN is that the objective function is not differentiable and the techniques commonly used in convergence analysis of the gradient methods (Bertsekas, 1999) do not work for OWL-QN. In this paper, we first point out the problems of existing convergence analysis for OWL-QN. Then, we propose a modified Orthant-Wise Limited memory Quasi-Newton (mOWL-QN) algorithm by slightly modifying the OWL-QN algorithm. As the main technical contribution of this paper, we establish a rigorous convergence analysis for the mOWL-QN algorithm. Moreover, we present empirical studies to show that mOWL-QN works well and is as efficient as OWL-QN. To our best knowledge, our work fills the theoretical gap by providing the first rigorous convergence proof for the OWL-QN-type algorithm on solving ℓ_1-regularized sparse learning problems.

The rest of the paper is organized as follows: We briefly review the OWL-QN algorithm and point out problems in the existing convergence analysis in Section 2. We propose the mOWL-QN algorithm by slightly modifying the OWL-QN algorithm and provide detailed convergence analysis for the mOWL-QN algorithm in Section 3. We report experimental results to validate the convergence analysis in Section 4 and we conclude in Section 5.

[1]This algorithm here refers to the OWL-QN algorithm.

2. OWL-QN and Problems in the Existing Convergence Analysis

OWL-QN (Andrew & Gao, 2007) is designed to solve the following ℓ_1-regularized optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{x}) = l(\mathbf{x}) + \lambda\|\mathbf{x}\|_1\}, \tag{1}$$

where $\|\mathbf{x}\|_1 = \sum_{i=1}^n$ is the ℓ_1-norm of $\mathbf{x}$ and $l : \mathbb{R}^n \mapsto \mathbb{R}$ is convex, bounded from below and continuously differentiable; the gradient $\nabla l(\mathbf{x})$ is L-Lipschitz continuous for some $L > 0$.

2.1. Basics of OWL-QN

Define a function $\pi : \mathbb{R}^n \mapsto \mathbb{R}^n$ with the i-th entry being:

$$\pi_i(x_i; y_i) = \begin{cases} x_i, & \text{if } \sigma(x_i) = \sigma(y_i), \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{y} \in \mathbb{R}^n$ (y_i is the i-th entry of $\mathbf{y}$) is the parameter of the function π; $\sigma(\cdot)$ is the sign function defined as follows: $\sigma(x_i) = 1$, if $x_i > 0$; $\sigma(x_i) = -1$, if $x_i < 0$ and $\sigma(x_i) = 0$, otherwise. Define the pseudo-gradient $\diamond f(\mathbf{x})$ whose i-th entry is given by:

$$\diamond_i f(\mathbf{x}) = \begin{cases} \nabla_i l(\mathbf{x}) + \lambda, & \text{if } x_i > 0, \\ \nabla_i l(\mathbf{x}) - \lambda, & \text{if } x_i < 0, \\ \nabla_i l(\mathbf{x}) + \lambda, & \text{if } x_i = 0,\ \nabla_i l(\mathbf{x}) + \lambda < 0, \\ \nabla_i l(\mathbf{x}) - \lambda, & \text{if } x_i = 0,\ \nabla_i l(\mathbf{x}) - \lambda > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the objective function in problem (1) is approximated by a quadratic function and a direction is computed by minimizing that quadratic function as follows:

$$\begin{aligned} \mathbf{d}^k &= \arg\min_{\mathbf{d}\in\mathbb{R}^n} \left\{ f(\mathbf{x}^k) + \diamond f(\mathbf{x}^k)^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T B^k \mathbf{d} \right\} \\ &= -H^k \diamond f(\mathbf{x}^k), \end{aligned}$$

where B^k is the (approximated) Hessian matrix at $\mathbf{x} = \mathbf{x}^k$ and $H^k = (B^k)^{-1}$. Here, Andrew & Gao (2007) use the L-BFGS[2] (Jorge & Stephen, 1999) to approximate the inverse Hessian matrix H^k and compute the matrix-vector multiplication $-H^k \diamond f(\mathbf{x}^k)$, which enables OWL-QN to tackle large-scale problems. To guarantee convergence, Andrew & Gao (2007) align the direction as follows:

$$\mathbf{p}^k = \pi(\mathbf{d}^k; \mathbf{v}^k), \text{ where } \mathbf{v}^k = -\diamond f(\mathbf{x}^k).$$

To restrict the next iterate in the same orthant of the previous iterate $\mathbf{x}^k$, Andrew & Gao (2007) propose to project

[2]More details on the L-BFGS are provided in Supplement A.

the point back onto the same orthant of the previous iterate $\mathbf{x}^k$:

$$\mathbf{x}^k(\alpha) = \pi(\mathbf{x}^k + \alpha \mathbf{p}^k; \boldsymbol{\xi}^k), \tag{2}$$

where

$$\xi_i^k = \begin{cases} \sigma(x_i^k), & \text{if } x_i^k \neq 0, \\ \sigma(v_i^k), & \text{if } x_i^k = 0, \end{cases} \tag{3}$$

and α is a step size chosen by the following line search procedure: for constants $\beta, \gamma \in (0,1)$ and $m = 0, 1, \cdots$, find the smallest integer m with $\alpha = \beta^m$ such that

$$f(\mathbf{x}^k(\alpha)) \leq f(\mathbf{x}^k) - \gamma(\mathbf{v}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k). \tag{4}$$

The pseudo code of OWL-QN is given in Algorithm 2.1.

Algorithm 1 OWL-QN: Orthant-Wise Limited memory Quasi-Newton

Initialize $\mathbf{x}^0$, $S \leftarrow \{\}, Y \leftarrow \{\}$ and choose $\beta, \gamma \in (0,1)$;
for $k = 0$ **to** maxiter **do**
  Compute $\mathbf{v}^k \leftarrow -\diamond f(\mathbf{x}^k)$;
  Compute $\mathbf{d}^k \leftarrow H^k \mathbf{v}^k$ using L-BFGS with S, Y;
  Alignment: $\mathbf{p}^k \leftarrow \pi(\mathbf{d}^k; \mathbf{v}^k)$;
  Initialize $\alpha \leftarrow 1$;
  while Eq. (4) is not satisfied **do**
    $\alpha \leftarrow \alpha\beta$;
    $\mathbf{x}^k(\alpha) \leftarrow \pi(\mathbf{x}^k + \alpha \mathbf{p}^k; \boldsymbol{\xi}^k)$;
  end while
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k(\alpha)$;
  if some stopping criterion is satisfied **then**
    stop and return $\mathbf{x}^{k+1}$;
  end if
  Update S with $\mathbf{s}^k \leftarrow \mathbf{x}^{k+1} - \mathbf{x}^k$;
  Update Y with $\mathbf{y}^k \leftarrow \nabla l(\mathbf{x}^{k+1}) - \nabla l(\mathbf{x}^k)$;
end for

2.2. Problems in the Existing Convergence Analysis of OWL-QN

Andrew & Gao (2007) utilize the techniques used in the convergence proof of gradient methods (Bertsekas, 1999) to prove the convergence of OWL-QN. However, due to the non-differentiability of the objective function, these techniques do not work for OWL-QN. In the following, we point out the key problems in the existing convergence proof.

First, Andrew & Gao (2007) present the following proposition and establish the convergence proof of OWL-QN based on this proposition.

Proposition 1 *"Define* $\mathbf{q}_\alpha^k = \frac{1}{\alpha}(\pi(\mathbf{x}^k + \alpha \mathbf{p}^k; \boldsymbol{\xi}^k) - \mathbf{x}^k)$. *Then for all* $\alpha \in (0, \infty)$ *and all* i,

$$d_i^k v_i^k \leq p_i^k v_i^k \leq (q_\alpha^k)_i v_i^k \leq 0,$$

Algorithm 2 mOWL-QN: modified Orthant-Wise Limited memory Quasi-Newton

Initialize $\mathbf{x}^0$, $S \leftarrow \{\}, Y \leftarrow \{\}$ and choose $\beta, \gamma \in (0,1), \epsilon > 0, \alpha_0 > 0$;
for $k = 0$ **to** maxiter **do**
  Compute $\mathbf{v}^k \leftarrow -\diamond f(\mathbf{x}^k)$ and
  $\mathcal{I}^k = \{i \in \{1, \cdots, n\} : 0 < |x_i^k| \leq \epsilon^k, x_i^k v_i^k < 0\}$, where $\epsilon^k = \min(\|\mathbf{v}^k\|, \epsilon)$;
  Initialize $\alpha \leftarrow \alpha_0$;
  if $\mathcal{I}^k = \emptyset$ **then**
    (QN-step)
    Compute $\mathbf{d}^k \leftarrow H^k \mathbf{v}^k$ using L-BFGS with S, Y;
    Alignment: $\mathbf{p}^k \leftarrow \pi(\mathbf{d}^k; \mathbf{v}^k)$;
    while Eq. (7) is not satisfied **do**
      $\alpha \leftarrow \alpha\beta$;
      $\mathbf{x}^k(\alpha) \leftarrow \pi(\mathbf{x}^k + \alpha \mathbf{p}^k; \boldsymbol{\xi}^k)$;
    end while
  else
    (GD-step)
    while Eq. (8) is not satisfied **do**
      $\alpha \leftarrow \alpha\beta$;
      $\mathbf{x}^k(\alpha) \leftarrow \arg\min_{\mathbf{x}} \{\nabla l(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^k\|^2 + \lambda\|\mathbf{x}\|_1\}$;
    end while
  end if
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k(\alpha)$;
  if some stopping criterion is satisfied **then**
    stop and return $\mathbf{x}^{k+1}$;
  end if
  Update S with $\mathbf{s}^k \leftarrow \mathbf{x}^{k+1} - \mathbf{x}^k$;
  Update Y with $\mathbf{y}^k \leftarrow \nabla l(\mathbf{x}^{k+1}) - \nabla l(\mathbf{x}^k)$;
end for

and therefore

$$(\mathbf{v}^k)^T \mathbf{q}_\alpha^k \geq (\mathbf{v}^k)^T \mathbf{p}^k \geq (\mathbf{v}^k)^T \mathbf{d}^k."$$

Unfortunately, Proposition 1 is NOT correct (We will correct this proposition in the next section).

Second, Andrew & Gao (2007) claim that $\{\alpha^k\} \to 0$ and $\{\|\mathbf{q}_{\alpha^k\beta^{-1}}^k\|\}$ is bounded [refer to Theorem 2 in Andrew & Gao (2007)]. Obviously, $\{\alpha^k\} \to 0$ indicates that the denominator of $\|\mathbf{q}_{\alpha^k\beta^{-1}}^k\| = \frac{1}{\alpha^k\beta^{-1}}\|\pi(\mathbf{x}^k + \alpha^k\beta^{-1}\mathbf{p}^k; \boldsymbol{\xi}^k) - \mathbf{x}^k\|$ approaches zero. Thus, we can NOT simply conclude that $\{\|\mathbf{q}_{\alpha^k\beta^{-1}}^k\|\}$ is bounded.

Third, Andrew & Gao (2007) apply the mean value theorem to obtain that there exists some $\tilde{\alpha} \in [0, \hat{\alpha}]$ such that

$$\frac{f(\mathbf{x}^k + \hat{\alpha}\hat{\mathbf{q}}^k) - f(\mathbf{x}^k)}{\hat{\alpha}} = f'(\mathbf{x}^k + \tilde{\alpha}^k\hat{\mathbf{q}}^k; \hat{\mathbf{q}}^k).$$

This is NOT necessarily true, due to the non-differentiability of f. In fact, we only obtain that

there exists some $\tilde{\alpha} \in [0, \hat{\alpha}]$ such that

$$\frac{f(\mathbf{x}^k + \hat{\alpha}\hat{\mathbf{q}}^k) - f(\mathbf{x}^k)}{\hat{\alpha}} \in (\hat{\mathbf{q}}^k)^T \partial f(\mathbf{x}^k + \tilde{\alpha}^k \hat{\mathbf{q}}^k).$$

Fourth, based on the following inequality:

$$f'(\mathbf{x}^k + \tilde{\alpha}^k \hat{\mathbf{q}}^k; \hat{\mathbf{q}}^k) > \gamma f'(\mathbf{x}^k; \hat{\mathbf{q}}^k), \tag{5}$$

where $\gamma \in (0,1)$, $\{\tilde{\alpha}^k\}_\kappa \to 0$, $\{\mathbf{x}^k\}_\kappa \to \bar{\mathbf{x}}$ and $\{\hat{\mathbf{q}}^k\}_\kappa \to \bar{\mathbf{q}}$, Andrew & Gao (2007) take limits for Eq. (5) and obtain

$$f'(\bar{\mathbf{x}}; \bar{\mathbf{q}}) > \gamma f'(\bar{\mathbf{x}}, \bar{\mathbf{q}}) \tag{6}$$

and hence $f'(\bar{\mathbf{x}}, \bar{\mathbf{q}}) > 0$, since $\gamma \in (0,1)$. Unfortunately, Eq. (6) is NOT correct, because $f'(\mathbf{x}^k, \hat{\mathbf{q}}^k)$ and $f'(\mathbf{x}^k + \tilde{\alpha}^k \hat{\mathbf{q}}^k, \hat{\mathbf{q}}^k)$ do NOT necessarily converge to $f'(\bar{\mathbf{x}}, \bar{\mathbf{q}})$. Actually, we only have the following upper semi-continuity according to Proposition B.23 in Bertsekas (1999):

$$\limsup_{k\in\kappa, k\to\infty} f'(\mathbf{x}^k; \hat{\mathbf{q}}^k) \le f'(\bar{\mathbf{x}}; \bar{\mathbf{q}}),$$
$$\limsup_{k\in\kappa, k\to\infty} f'(\mathbf{x}^k + \tilde{\alpha}^k \hat{\mathbf{q}}^k; \hat{\mathbf{q}}^k) \le f'(\bar{\mathbf{x}}; \bar{\mathbf{q}}).$$

This problem was also pointed out by Yu et al. (2010).

3. Convergence Analysis for mOWL-QN

We first propose a modified Orthant-Wise Limited memory Quasi-Newton (mOWL-QN) algorithm as in Algorithm 2.1 by slightly modifying the OWL-QN algorithm. At each iteration, mOWL-QN adopts different strategies (QN-step or GD-step) to generate the next iterate depending on if $\mathcal{I}^k = \{i \in \{1, \cdots, n\} : 0 < |x_i^k| \le \min(\|\mathbf{v}^k\|, \epsilon), x_i^k v_i^k < 0\}$ is an empty set. Accordingly, we use different line search criteria for each strategy: for constants $\alpha_0 > 0, \beta, \gamma \in (0,1)$ and $m = 0, 1, \cdots$, find the smallest integer m with $\alpha = \alpha_0 \beta^m$ such that the following inequalities hold:

$$\text{QN-step}: f(\mathbf{x}^k(\alpha)) \le f(\mathbf{x}^k) - \gamma\alpha(\mathbf{v}^k)^T \mathbf{d}^k, \tag{7}$$
$$\text{GD-step}: f(\mathbf{x}^k(\alpha)) \le f(\mathbf{x}^k) - \frac{\gamma}{2\alpha}\|\mathbf{x}^k(\alpha) - \mathbf{x}^k\|^2. \tag{8}$$

Remark 1 *In practice, we set $\epsilon > 0$ as a very small value (e.g., 10^{-12}) such that $\mathcal{I}^k$ is empty and hence QN-step is adopted at almost all iterations. Thus, mOWL-QN not only converges in theory but also works as efficient as OWL-QN in practice. But without the strategy of switching between QN-step and GD-step, it is practically impossible to prove the convergence.*

Next, we provide a rigorous convergence analysis for the mOWL-QN algorithm. We begin the convergence analysis by showing how to correct Proposition 1.

3.1. Corrected Proposition 1

Proposition 2 *(Corrected) Define $\mathbf{q}_\alpha^k = \frac{1}{\alpha}(\pi(\mathbf{x}^k + \alpha\mathbf{p}^k; \boldsymbol{\xi}^k) - \mathbf{x}^k)$. Then for all $\alpha \in (0, \infty)$ and all i,*

$$p_i^k v_i^k \ge d_i^k v_i^k, \tag{9}$$
$$p_i^k v_i^k \ge (q_\alpha^k)_i v_i^k \ge 0, \tag{10}$$

and therefore

$$(\mathbf{v}^k)^T \mathbf{p}^k \ge (\mathbf{v}^k)^T \mathbf{d}^k \ge 0, \tag{11}$$
$$(\mathbf{v}^k)^T \mathbf{p}^k \ge (\mathbf{v}^k)^T \mathbf{q}_\alpha^k \ge 0. \tag{12}$$

Proof Since $p_i^k = \pi_i(d_i^k; v_i^k)$, Eq. (9) is obvious. By the definition of $\mathbf{q}_\alpha^k$, we know that $(q_\alpha^k)_i = p_i^k$ or $(q_\alpha^k)_i = -x_i^k/\alpha$. We next prove Eq. (10) by considering these two cases separately.

(a) If $(q_\alpha^k)_i = p_i^k$, then we have $(q_\alpha^k)_i v_i^k = p_i^k v_i^k \ge 0$ by recalling $p_i^k = \pi_i(d_i^k; v_i^k)$ and hence Eq. (10) holds.

(b) If $(q_\alpha^k)_i = -x_i^k/\alpha$, then we have $(x_i^k + \alpha p_i^k)\xi_i^k \le 0$ must hold. We next focus on the case (b) in the following three subcases:

(1) If $x_i^k > 0$, then by the definition of ξ_i^k in Eq. (3), we have $\xi_i^k = 1$. Recalling $(x_i^k + \alpha p_i^k)\xi_i^k \le 0$, we have $x_i^k + \alpha p_i^k \le 0$ and hence $p_i^k < 0$. Recalling $p_i^k = \pi_i(d_i^k; v_i^k)$ and $(q_\alpha^k)_i = -x_i^k/\alpha$, we have $v_i^k < 0$ and hence $(q_\alpha^k)_i v_i^k > 0$, $(x_i^k + \alpha p_i^k)v_i^k \ge 0$. Thus, we have $p_i^k v_i^k - (q_\alpha^k)_i v_i^k \ge 0$. Therefore, Eq. (10) holds.

(2) If $x_i^k < 0$, then by the definition of ξ_i^k in Eq. (3), we have $\xi_i^k = -1$. Recalling $(x_i^k + \alpha p_i^k)\xi_i^k \le 0$, we have $x_i^k + \alpha p_i^k \ge 0$ and hence $p_i^k > 0$. Recalling $p_i^k = \pi_i(d_i^k; v_i^k)$ and $(q_\alpha^k)_i = -x_i^k/\alpha$, we have $v_i^k > 0$ and hence $(q_\alpha^k)_i v_i^k > 0$, $(x_i^k + \alpha p_i^k)v_i^k \ge 0$. Thus, we have $p_i^k v_i^k - (q_\alpha^k)_i v_i^k \ge 0$. Therefore, Eq. (10) holds.

(3) If $x_i^k = 0$, then by recalling $p_i^k = \pi_i(d_i^k; v_i^k)$ and $(q_\alpha^k)_i = -x_i^k/\alpha$, we have $(q_\alpha^k)_i v_i^k = 0$ and $p_i^k v_i^k \ge 0$. Therefore, Eq. (10) holds.

Eq. (11) and Eq. (12) readily follow from Eq. (9) and Eq. (10) by noticing that $(\mathbf{v}^k)^T \mathbf{d}^k = (\mathbf{v}^k)^T H^k \mathbf{v}^k \ge 0$ as long as H^k is positive definite (the positive definiteness of H^k is guaranteed by the L-BFGS).

3.2. Convergence Proof of mOWL-QN

We first provide an optimality condition for problem (1), which is directly used to prove the final convergence theorem (Theorem 1).

Proposition 3 *Let $\bar{\mathbf{x}} = \lim_{k\in\mathcal{K}, k\to\infty} \mathbf{x}^k$, $\mathbf{v}^k = -\diamond f(\mathbf{x}^k)$ and $\bar{\mathbf{v}} = -\diamond f(\bar{\mathbf{x}})$, where $\mathcal{K}$ is a subsequence of*

$\{1, 2, \cdots, k, k+1, \cdots\}$. *If* $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = 0$ *for all* $i \in \{1, \cdots, n\}$, *then* $\bar{\mathbf{v}} = \mathbf{0}$ *and hence* $\bar{\mathbf{x}}$ *is a global minimizer of problem (1).*

Proof We firstly use contradiction to prove that if $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = 0$ for all $i \in \{1, \cdots, n\}$, then $\bar{\mathbf{v}} = \mathbf{0}$. Assume that $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = 0$ for all $i \in \{1, \cdots, n\}$ but $\bar{\mathbf{v}} \neq \mathbf{0}$. Then there exists at least one $i \in \{1, \cdots, n\}$ such that $\bar{v}_i = -\diamond_i f(\bar{\mathbf{x}}) \neq 0$. We consider the following two cases:

(1) If $\bar{x}_i \neq 0$, then we have $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = |\bar{v}_i| \neq 0$, leading to a contradiction with that $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = 0$ for all $i \in \{1, \cdots, n\}$.

(2) If $\bar{x}_i = 0$, then $\bar{v}_i = -\diamond_i f(\bar{\mathbf{x}}) \neq 0$ implies that

$$\begin{aligned} &\nabla_i l(\bar{\mathbf{x}}) + \lambda > \nabla_i l(\bar{\mathbf{x}}) - \lambda > 0 \\ \text{or } &\nabla_i l(\bar{\mathbf{x}}) - \lambda < \nabla_i l(\bar{\mathbf{x}}) + \lambda < 0. \end{aligned} \tag{13}$$

By the definition of $v_i^k = -\diamond_i f(\mathbf{x}^k)$, we know that

$$-(\nabla_i l(\mathbf{x}^k) + \lambda) \leq v_i^k \leq -(\nabla_i l(\mathbf{x}^k) - \lambda).$$

Taking limits of the above inequalities, we have

$$\begin{aligned} -(\nabla_i l(\bar{\mathbf{x}}) + \lambda) &\leq \liminf_{k\in\mathcal{K}, k\to\infty} v_i^k \leq -(\nabla_i l(\bar{\mathbf{x}}) - \lambda), \text{ and} \\ -(\nabla_i l(\bar{\mathbf{x}}) + \lambda) &\leq \limsup_{k\in\mathcal{K}, k\to\infty} v_i^k \leq -(\nabla_i l(\bar{\mathbf{x}}) - \lambda), \end{aligned}$$

which together with Eq. (13) imply that

$$\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| \neq 0.$$

This leads to a contradiction with that $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = 0$ for all $i \in \{1, \cdots, n\}$. Therefore, if $\liminf_{k\in\mathcal{K}, k\to\infty} |v_i^k| = 0$ for all $i \in \{1, \cdots, n\}$, then $\bar{\mathbf{v}} = \mathbf{0}$.

To complete the proof, we next prove that $\bar{\mathbf{x}}$ is a global minimizer of problem (1) if and only if $\bar{\mathbf{v}} = \mathbf{0}$. According to the definition of the pseudo gradient, it is easy to verify that $\diamond f(\bar{\mathbf{x}})$ is the minimum norm sub-gradient at point $\mathbf{x} = \bar{\mathbf{x}}$, that is:

$$\diamond f(\bar{\mathbf{x}}) = \arg\min_{\bar{\mathbf{g}} \in \partial f(\bar{\mathbf{x}})} \|\bar{\mathbf{g}}\|. \tag{14}$$

Thus, $\mathbf{0} \in \partial f(\bar{\mathbf{x}}) \Leftrightarrow \diamond f(\bar{\mathbf{x}}) = \mathbf{0} \Leftrightarrow \bar{\mathbf{v}} = \mathbf{0}$ and hence $\bar{\mathbf{x}}$ is a global minimizer of problem (1) if and only if $\mathbf{0} \in \partial f(\bar{\mathbf{x}})$, if and only if $\bar{\mathbf{v}} = \mathbf{0}$.

We subsequently show that we have a similar Lipschitz continuous inequality in the following proposition, which is crucial to prove the final convergence theorem.

Proposition 4 *Let* $\nabla l(\mathbf{x})$ *be L-Lipschitz continuous for some* $L > 0$, $\mathbf{v}^k = -\diamond f(\mathbf{x}^k)$, $\mathbf{x}^k(\alpha) = \pi(\mathbf{x}^k + \alpha\mathbf{p}^k; \boldsymbol{\xi}^k)$ *and* $\mathbf{q}_\alpha^k = \frac{1}{\alpha}(\pi(\mathbf{x}^k + \alpha\mathbf{p}^k; \boldsymbol{\xi}^k) - \mathbf{x}^k)$ *with* $\alpha > 0$. *Then we have*

$$\begin{aligned} &(i)\ \nabla l(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \lambda(\|\mathbf{x}^k(\alpha)\|_1 - \|\mathbf{x}^k\|_1) \\ &= -(\mathbf{v}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k), \end{aligned} \tag{15}$$

$$(ii)\ f(\mathbf{x}^k(\alpha)) \leq f(\mathbf{x}^k) - \alpha(\mathbf{v}^k)^T\mathbf{q}_\alpha^k + \frac{\alpha^2 L}{2}\|\mathbf{q}_\alpha^k\|^2. \tag{16}$$

Proof (i) Based on the definition of $\mathbf{x}^k(\alpha)$, we know that $x_i^k(\alpha)x_i^k \geq 0$. We next prove for all $i \in \{1, \cdots, n\}$, the following equality holds by considering two cases:

$$\begin{aligned} &\nabla_i l(\mathbf{x}^k)(x_i^k(\alpha) - x_i^k) + \lambda(|x_i^k(\alpha)| - |x_i^k|) \\ &= -v_i^k(x_i^k(\alpha) - x_i^k). \end{aligned} \tag{17}$$

(a) If $x_i^k \neq 0$, then $x_i^k(\alpha)x_i^k \geq 0$ implies $|x_i^k(\alpha)| - |x_i^k| = \sigma(x_i^k)(x_i^k(\alpha) - x_i^k)$, which together with $\nabla_i l(\mathbf{x}^k) + \lambda\sigma(x_i^k) = -v_i^k$ (by noticing that $x_i^k \neq 0$) implies that Eq. (17) holds.

(b) If $x_i^k = 0$, then we have $x_i^k(\alpha) = \pi_i(\alpha p_i^k; \sigma(v_i^k)) = \alpha p_i^k$. We next focus on the case (b) in the following two subcases:

(1) If $p_i^k \neq 0$, then $|x_i^k(\alpha)| = \alpha\sigma(p_i^k)p_i^k = \sigma(v_i^k)(\alpha p_i^k) = \sigma(v_i^k)x_i^k(\alpha)$. Thus, $|x_i^k(\alpha)| - |x_i^k| = \sigma(v_i^k)(x_i^k(\alpha) - x_i^k)$. Noticing that $p_i^k \neq 0$ implies $v_i^k \neq 0$. According to the definition of v_i^k, we obtain that $\nabla_i l(\mathbf{x}^k) + \lambda\sigma(v_i^k) = -v_i^k$ whenever $x_i^k = 0$ and $v_i^k \neq 0$. Therefore, Eq. (17) holds.

(2) If $p_i^k = 0$, then $|x_i^k(\alpha)| - |x_i^k| = x_i^k(\alpha) - x_i^k = 0$, which implies Eq. (17) holds.

Combining (a) and (b), we obtain that Eq. (17) holds for all $i \in \{1, \cdots, n\}$, which implies that Eq. (15) holds.

(ii) Since $\nabla l(\mathbf{x})$ is L-Lipschitz continuous, we have

$$\begin{aligned} l(\mathbf{x}^k(\alpha)) \leq &l(\mathbf{x}^k) + \nabla l(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) \\ &+ \frac{L}{2}\|\mathbf{x}^k(\alpha) - \mathbf{x}^k\|^2. \end{aligned}$$

It follows that

$$\begin{aligned} f(\mathbf{x}^k(\alpha)) &\leq f(\mathbf{x}^k) + \nabla l(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) \\ &+ \lambda(\|\mathbf{x}^k(\alpha)\|_1 - \|\mathbf{x}^k\|_1) + \frac{L}{2}\|\mathbf{x}^k(\alpha) - \mathbf{x}^k\|^2, \end{aligned}$$

which together with Eq. (15) and $\mathbf{q}_\alpha^k = \frac{1}{\alpha}(\mathbf{x}^k(\alpha) - \mathbf{x}^k)$ implies that Eq. (16) holds.

We next show that both line search criteria in QN-step (Eq. (7)) and GD-step (Eq. (8)) at any iteration k is satisfied in a finite number of trials (The detailed proof is provided in Supplement C).

Proposition 5 *At any iteration k of the mOWL-QN algorithm, if $\mathbf{x}^k$ is not a minimizer of problem (1), then (a) for QN-step, there exists an $\alpha \in [\bar{\alpha}^k, \alpha_0]$ with $0 < \bar{\alpha}^k \leq \alpha_0$ such that the line search criterion in Eq. (7) is satisfied; (b) for GD-step, the line search criterion in Eq. (8) is satisfied whenever $\alpha \geq \beta \min(\alpha_0, (1-\gamma)/L)$. That is, both line search criteria at any iteration k are satisfied in a finite number of trials.*

We finally provide the convergence proof for the mOWL-QN algorithm based on the propositions presented above.

Theorem 1 *The sequence $\{\mathbf{x}^k\}$ generated by the mOWL-QN algorithm has at least a limit point and every limit point of $\{\mathbf{x}^k\}$ is a global minimizer of problem (1).*

Proof It follows from Proposition 5 that both line search criteria in QN-step (Eq. (7)) and GD-step (Eq. (8)) at each iteration can be satisfied in a finite number of trials. Let α^k be the accepted step size at iteration k. Then we have

$$\begin{aligned} & f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \gamma \alpha^k (\mathbf{v}^k)^T \mathbf{d}^k \\ & = \gamma \alpha^k (\mathbf{v}^k)^T H^k \mathbf{v}^k \text{ (QN-step)}, \qquad (18) \\ \text{or } & f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{\gamma}{2\alpha^k} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ & \geq \frac{\gamma}{2\alpha_0} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \text{ (GD-step)}. \qquad (19) \end{aligned}$$

Recalling that H^k is positive definite and $\gamma > 0, \alpha^k > 0$, which together with Eqs.(18),(19) imply that $\{f(\mathbf{x}^k)\}$ is monotonically decreasing. Thus, $\{f(\mathbf{x}^k)\}$ converges to a finite value $\bar{f}$, since f is bounded from below. Due to the boundedness of $\{\mathbf{x}^k\}$ (see Proposition 6 in Supplement B), the sequence $\{\mathbf{x}^k\}$ generated by the mOWL-QN algorithm has at least a limit point $\bar{\mathbf{x}}$. Since f is continuous, there exists a subsequence $\mathcal{K}$ of $\{1, 2 \cdots, k, k+1, \cdots\}$ such that

$$\lim_{k \in \mathcal{K}, k \to \infty} \mathbf{x}^k = \bar{\mathbf{x}}, \qquad (20)$$

$$\lim_{k \to \infty} f(\mathbf{x}^k) = \lim_{k \in \mathcal{K}, k \to \infty} f(\mathbf{x}^k) = \bar{f} = f(\bar{\mathbf{x}}). \qquad (21)$$

In the following, we prove the theorem by contradiction. Assume that $\bar{\mathbf{x}}$ is not a global minimizer of problem (1). Then by Proposition 3, there exists at least one $i \in \{1, \cdots, n\}$ such that

$$\liminf_{k \in \mathcal{K}, k \to \infty} |v_i^k| > 0. \qquad (22)$$

We next consider the following two cases:

(a) There exist a subsequence $\tilde{\mathcal{K}}$ of $\mathcal{K}$ and an integer $\tilde{k} > 0$ such that for all $k \in \tilde{\mathcal{K}}, k \geq \tilde{k}$, GD-step is adopted. Then for all $k \in \tilde{\mathcal{K}}, k \geq \tilde{k}$, we have

$$\begin{aligned} \mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \Big\{ & \nabla l(\mathbf{x}^k)^T (\mathbf{x} - \mathbf{x}^k) \\ & + \frac{1}{2\alpha^k} \|\mathbf{x} - \mathbf{x}^k\|^2 + \lambda \|\mathbf{x}\|_1 \Big\}. \end{aligned}$$

By the optimality condition of the above problem, we have

$$\begin{aligned} & \lambda \|\mathbf{x}\|_1 \geq \lambda \|\mathbf{x}^{k+1}\|_1 \\ & + \left(\frac{1}{\alpha^k} (\mathbf{x}^k - \mathbf{x}^{k+1}) - \nabla l(\mathbf{x}^k) \right)^T (\mathbf{x} - \mathbf{x}^{k+1}), \\ & \forall \mathbf{x} \in \mathbb{R}^n, k \in \tilde{\mathcal{K}}, k \geq \tilde{k}. \qquad (23) \end{aligned}$$

Taking limits with $k \in \tilde{\mathcal{K}}$ for Eq. (19) and considering Eqs. (20), (21), we have

$$\begin{aligned} & \lim_{k \in \tilde{\mathcal{K}}, k \to \infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \leq 0 \\ & \Rightarrow \lim_{k \in \tilde{\mathcal{K}}, k \to \infty} \mathbf{x}^k = \lim_{k \in \tilde{\mathcal{K}}, k \to \infty} \mathbf{x}^{k+1} = \bar{\mathbf{x}}. \qquad (24) \end{aligned}$$

Taking limits with $k \in \tilde{\mathcal{K}}$ for Eq. (23) and considering Eq. (24) and $\alpha^k \geq \beta \min(\alpha_0, (1-\gamma)/L)$ (Proposition 5), we have

$$\begin{aligned} & \lambda \|\mathbf{x}\|_1 \geq \lambda \|\bar{\mathbf{x}}\|_1 - \nabla l(\bar{\mathbf{x}})^T (\mathbf{x} - \bar{\mathbf{x}}), \forall \mathbf{x} \in \mathbb{R}^n \\ & \Rightarrow -\nabla l(\bar{\mathbf{x}}) \in \lambda \partial \|\bar{\mathbf{x}}\|_1, \end{aligned}$$

which leads to a contradiction with the assumption that $\bar{\mathbf{x}}$ is not a global minimizer of problem (1).

(b) There exists an integer $\hat{k} > 0$ such that for all $k \in \mathcal{K}, k \geq \hat{k}$, QN-step is adopted. According to Remark 5 (in Supplement B), we know that the smallest eigenvalue of H^k is uniformly bounded from below by a positive constant, which together with Eq. (22) implies

$$\liminf_{k \in \mathcal{K}, k \to \infty} (\mathbf{v}^k)^T H^k \mathbf{v}^k > 0. \qquad (25)$$

Taking limits with $k \in \mathcal{K}$ for Eq. (18), we have

$$\lim_{k \in \mathcal{K}, k \to \infty} \gamma \alpha^k (\mathbf{v}^k)^T H^k \mathbf{v}^k \leq 0,$$

which together with $\gamma \in (0, 1), \alpha^k \in (0, \alpha_0]$ and Eq. (25) implies that

$$\lim_{k \in \mathcal{K}, k \to \infty} \alpha^k = 0. \qquad (26)$$

Eq. (22) implies that there exist an integer $\check{k} > 0$ and a constant $\bar{\epsilon} > 0$ such that $\epsilon^k = \min(\|\mathbf{v}^k\|, \epsilon) \geq \bar{\epsilon}$ for all $k \in \mathcal{K}, k \geq \check{k}$. Notice that for all $k \in \mathcal{K}, k \geq \hat{k}$, QN-step is adopted. Thus, we have $\mathcal{I}^k = \{i \in$

$\{1,\cdots,n\}: 0<|x_i^k|\leq\epsilon^k, x_i^k v_i^k<0\}=\emptyset$ for all $k\in\mathcal{K}, k\geq\hat{k}$. We also notice that, if $|x_i^k|\geq\bar{\epsilon}$, there exists a constant $\bar{\alpha}_i>0$ such that $x_i^k(\alpha)=\pi_i(x_i^k+\alpha p_i^k;\xi_i^k)=x_i^k+\alpha p_i^k$ for all $\alpha\in(0,\bar{\alpha}_i]$, as $\{p_i^k\}$ is bounded (Proposition 7 in Supplement B). Therefore, we conclude that, for all $k\in\mathcal{K}, k\geq\bar{k}=\max(\check{k},\hat{k})$ and for all $i\in\{1,\cdots,n\}$, at least one of the following three cases must happen:

$$\begin{aligned}
&x_i^k=0\Rightarrow x_i^k(\alpha)=\pi_i(x_i^k+\alpha p_i^k;\xi_i^k)\\
&\qquad\qquad=x_i^k+\alpha p_i^k,\forall\alpha>0,\\
&\text{or } |x_i^k|>\epsilon^k\geq\bar{\epsilon}\Rightarrow x_i^k(\alpha)=\pi_i(x_i^k+\alpha p_i^k;\xi_i^k)\\
&\qquad\qquad=x_i^k+\alpha p_i^k,\forall\alpha\in(0,\bar{\alpha}_i],\\
&\text{or } x_i^k v_i^k\geq 0\Rightarrow x_i^k p_i^k\geq 0\Rightarrow\\
&\quad x_i^k(\alpha)=\pi_i(x_i^k+\alpha p_i^k;\xi_i^k)=x_i^k+\alpha p_i^k,\forall\alpha>0.
\end{aligned}$$

It follows that there exists a constant $\bar{\alpha}>0$ such that

$$\begin{aligned}
&\mathbf{q}_\alpha^k=\frac{1}{\alpha}(\mathbf{x}^k(\alpha)-\mathbf{x}^k)=\mathbf{p}^k,\\
&\forall k\in\mathcal{K}, k\geq\bar{k},\alpha\in(0,\bar{\alpha}].
\end{aligned}\tag{27}$$

Thus, considering Eq. (11) and $|p_i^k|=|\pi_i(d_i^k;v_i^k)|\leq|d_i^k|$ for all $i\in\{1,\cdots,n\}$, we have

$$\begin{aligned}
&\|\mathbf{q}_\alpha^k\|^2=\|\mathbf{p}^k\|^2\leq\|\mathbf{d}^k\|^2=(\mathbf{v}^k)^T(H^k)^2\mathbf{v}^k,\\
&\forall k\in\mathcal{K}, k\geq\bar{k},\alpha\in(0,\bar{\alpha}],
\end{aligned}\tag{28}$$

$$\begin{aligned}
&(\mathbf{v}^k)^T\mathbf{q}_\alpha^k=(\mathbf{v}^k)^T\mathbf{p}^k\geq(\mathbf{v}^k)^T\mathbf{d}^k=(\mathbf{v}^k)^TH^k\mathbf{v}^k,\\
&\forall k\in\mathcal{K}, k\geq\bar{k},\alpha\in(0,\bar{\alpha}].
\end{aligned}\tag{29}$$

According to Proposition 7 (in Supplement B), we know that the largest eigenvalue of H^k is uniformly bounded from above by some positive constant M. Thus, we have

$$\begin{aligned}
(\mathbf{v}^k)^T(H^k)^2\mathbf{v}^k\leq\frac{2}{\alpha L}(\mathbf{v}^k)^TH^k\mathbf{v}^k\\
-\left(\frac{2}{\alpha L}-M\right)(\mathbf{v}^k)^TH^k\mathbf{v}^k,\forall k,
\end{aligned}$$

which together with Eqs. (28), (29) and $\mathbf{d}^k=H^k\mathbf{v}^k$ implies

$$\begin{aligned}
&\|\mathbf{q}_\alpha^k\|^2\leq\frac{2}{\alpha L}(\mathbf{v}^k)^T\mathbf{q}_\alpha^k-\left(\frac{2}{\alpha L}-M\right)(\mathbf{v}^k)^T\mathbf{d}^k,\\
&\forall k\in\mathcal{K}, k\geq\bar{k},\alpha\in(0,\bar{\alpha}].
\end{aligned}\tag{30}$$

Considering Eqs. (16), (30), we have

$$\begin{aligned}
&f(\mathbf{x}^k(\alpha))\leq f(\mathbf{x}^k)-\alpha\left(1-\frac{\alpha LM}{2}\right)(\mathbf{v}^k)^T\mathbf{d}^k,\\
&\forall k\in\mathcal{K}, k\geq\bar{k},\alpha\in(0,\bar{\alpha}],
\end{aligned}$$

which together with Eq. (11) implies that the line search criterion in QN-step (Eq. (7)) is satisfied if

$$\begin{aligned}
&1-\frac{\alpha LM}{2}\geq\gamma\ ,0<\alpha\leq\alpha_0 \text{ and } 0<\alpha\leq\bar{\alpha},\\
&\forall k\in\mathcal{K}, k\geq\bar{k}.
\end{aligned}$$

Considering the backtracking form of the line search in QN-step (Eq. (7)), we conclude that the line search criterion in QN-step (Eq. (7)) is satisfied whenever

$$\begin{aligned}
&\alpha^k\geq\beta\min(\min(\bar{\alpha},\alpha_0),2(1-\gamma)/(LM))>0,\\
&\forall k\in\mathcal{K}, k\geq\bar{k}.
\end{aligned}$$

This leads to a contradiction with Eq. (26).

Conbiming (a) and (b), we conclude that $\bar{\mathbf{x}}=\lim_{k\in\mathcal{K},k\to\infty}\mathbf{x}^k$ is a global minimizer of problem (1), which together with Eq. (21) implies that every limit point of $\{\mathbf{x}^k\}$ is a global minimizer of problem (1).

Remark 2 *The challenge of proving the convergence of OWL-QN without any modification is: if there exists a subsequence $\mathcal{K}$ such that $\{x_i^k\}_\mathcal{K}$ converges to zero, it is possible that for a large enough $k\in\mathcal{K}$, $|x_i^k|$ is arbitrarily small but x_i^k is never equal to zero. In this case, Eq. (27) cannot be obtained (note that for $k\in\mathcal{K}$, $k+1$ may not be in $\mathcal{K}$), while Eq. (27) is critical to the subsequent proof. To exclude the above case in the QN-step, we propose to switch the iteration to the GD-step when $|x_i^k|$ is very small. We also change $(\mathbf{v}^k)^T\mathbf{q}_\alpha^k$ [note that $\mathbf{q}_\alpha^k=(\mathbf{x}^k(\alpha)-\mathbf{x}^k)/\alpha$] in Eq. (4) to $(\mathbf{v}^k)^T\mathbf{d}^k$ in Eq. (7) for the line search in the QN-step. Such a modification is needed to obtain Eq. (26) which is the contradiction we will show in the subsequent proof. If Proposition 1 is correct, we do not need this modification, because $(\mathbf{v}^k)^T\mathbf{q}_\alpha^k\geq(\mathbf{v}^k)^T\mathbf{p}^k\geq(\mathbf{v}^k)^T\mathbf{d}^k$ can also lead to Eq. (26).*

Remark 3 *Note that even for the QN-step, we use a totally different proof framework from OWL-QN (Andrew & Gao, 2007). In particular, the Lipschitz-continuous-like inequality in Proposition 4 is new for non-smooth problems and it is critical to the convergence proof.*

4. Experiments

In this section, we validate the convergence analysis by applying the mOWL-QN algorithm to solve the following ℓ_1-regularized logistic regression problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{f(\mathbf{x})=\frac{1}{N}\sum_{i=1}^N\log(1+\exp(-y_i\mathbf{a}_i^T\mathbf{x}))+\lambda\|\mathbf{x}\|_1\right\},$$

where N is the number of samples; $\lambda>0$ is the regularized parameter; $\mathbf{a}_i\in\mathbb{R}^n$ is the i-th sample; $y_i\in\{1,-1\}$ is the label of the sample $\mathbf{a}_i$.

Table 1. Data set statistics.

No.	1	2	3	4
datasets	kdd2010a	kdd2010b	real-sim	rcv1
♯ samples N	510,302	748,401	72,309	677,399
dimensionality n	20,216,830	29,890,095	20,958	47,236

We include mOWL-QN and OWL-QN algorithms in comparison. Experiments are conducted on four large scale data sets which are summarized in Table 1. These data sets are high-dimensional and sparse and can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html.

Both algorithms are implemented in Matlab and executed on an Intel(R) Core(TM)2 i7-3770 CPU (@3.4GHz) with 32GB memory. We choose the starting points $\mathbf{x}^0$ for both algorithms using the same random vector whose entries are independently sampled from the standard Gaussian distribution. We terminate both algorithms if the relative change of two consecutive objective function values is less than 10^{-5} or the number of iterations exceeds 500. We set $\gamma = 10^{-2}, \beta = 0.2, \alpha_0 = 1, \epsilon = 10^{-12}$ and the number of unrolling steps (in L-BFGS) as $m = 10$.

We report the objective function value vs. CPU time plots in Figure 1. We observe that our results agree with empirical studies of OWL-QN algorithms in existing literature (Schmidt et al., 2009; Yu et al., 2010; Yuan et al., 2010; Byrd et al., 2012a;b; 2013), that is, the OWL-QN algorithm works very well in practice, though a rigorous convergence proof has not been established so far. We also observe that mOWL-QN and OWL-QN have very similar convergence behaviors since the convergence curves of mOWL-QN and OWL-QN almost overlap with each other. The underlying reason is that, for a very small ϵ, mOWL-QN adopts QN-step at almost all iterations[3]. Moreover, both line search criteria in Eq. (4) and Eq. (7) accept the unit step size (i.e., the line search terminates in only one trial).

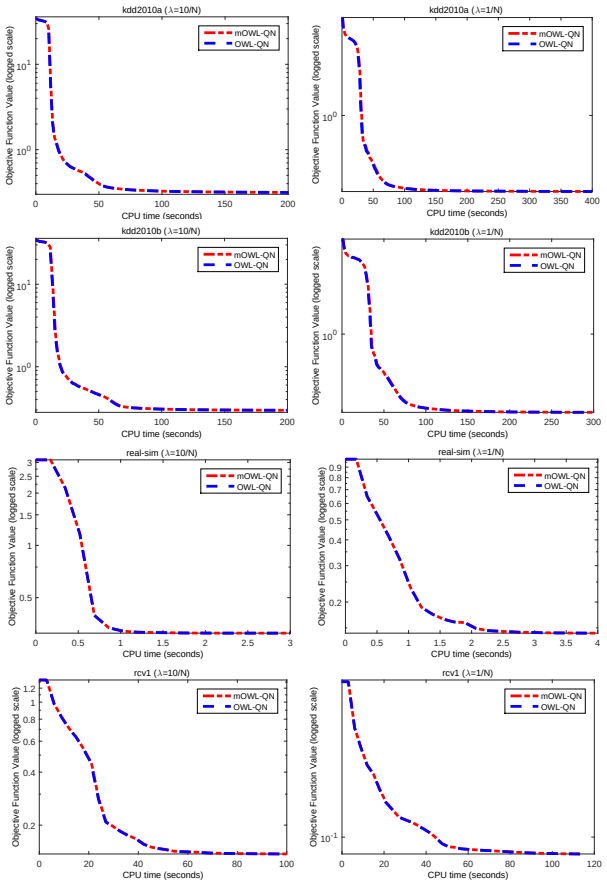

Figure 1. Objective function value (logged scale) vs. CPU time (seconds) plots using OWL-QN and mOWL-QN algorithms with different values of λ.

5. Conclusions

In this paper, we establish a detailed convergence analysis for the mOWL-QN algorithm which is based on a slight modification of a well-known algorithm called OWL-QN. To the best of our knowledge, this is the first work to provide a rigorous convergence proof and fills the theoretical gap for the OWL-QN-type algorithm.

There are several interesting directions that we will explore in the future. First, we plan to extend similar ideas to other sparsity-inducing problems including non-convex regularized problems. Second, we plan to analyze the convergence rate of the mOWL-QN algorithm in the future. Third, we plan to develop parallel and distributed variants of the algorithm to deal with much larger data sets.

Acknowledgements

We would like to thank reviewers for their constructive comments. This work is supported in part by research grants from NIH (R01 LM010730, U54 EB020403) and NSF (IIS- 0953662, IIS-1421057, IIS-1421100).

[3]The main purpose of the current experiments is to show that the modified algorithm is almost identical to the original algorithm when ϵ is set to be very small. Meanwhile, the modified algorithm has a rigorous convergence guarantee.

References

Andrew, G. and Gao, J. Scalable training of ℓ_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 33–40, 2007.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bertsekas, D.P. *Nonlinear Programming*. Athena Scientific, 1999.

Bioucas-Dias, J.M. and Figueiredo, M.A.T. A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, 2007.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5): 1190–1208, 1995.

Byrd, R. H., Chin, G. M., Nocedal, J., and Oztoprak, F. A family of second-order methods for convex ℓ_1-regularized optimization. Technical report, Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 2012a.

Byrd, R. H., Chin, G. M., Nocedal, J., and Wu, Y. Sample size selection in optimization methods for machine learning. *Mathematical Programming*, 134(1):127–155, 2012b.

Byrd, R.H., Nocedal, J., and Oztoprak, F. An inexact successive quadratic approximation method for convex ℓ_1 regularized optimization. *arXiv preprint arXiv:1309.3529*, 2013.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32(2):407–451, 2004.

Figueiredo, M.A.T., Nowak, R.D., and Wright, S.J. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal on Selected Topics in Signal Processing*, 1(4): 586–597, 2007.

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

Jorge, N. and Stephen, J. *Numerical Optimization*. Springer, 1999.

Olsen, P., Oztoprak, F., Nocedal, J., and Rennie, S. Newton-like methods for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 764–772, 2012.

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.

Schmidt, M., Kim, D., and Sra, S. *Optimization on Machine Learning*, chapter Projected Newton-type Methods in Machine Learning. MIT Press, 2011.

Shevade, S.K. and Keerthi, S.S. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246, 2003.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Tseng, P. and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., and Ma, Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2008.

Wright, S.J., Nowak, R., and Figueiredo, M. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Yu, J., Vishwanathan, S., Günter, S., and Schraudolph, N. A quasi-newton approach to nonsmooth convex optimization problems in machine learning. *The Journal of Machine Learning Research*, 11:1145–1200, 2010.

Yuan, G., Chang, K., Hsieh, C., and Lin, C. A comparison of optimization methods and software for large-scale ℓ_1-regularized linear classification. *The Journal of Machine Learning Research*, 11:3183–3234, 2010.

Yuan, G., Ho, C., and Lin, C. An improved glmnet for ℓ_1-regularized logistic regression. *The Journal of Machine Learning Research*, 13:1999–2030, 2012.

Supplementary Material for "A Modified Orthant-Wise Limited Memory Quasi-Newton Method with Convergence Analysis"

A. BFGS and L-BFGS

For self-containedness, we briefly review the update of the inverse Hessian matrix in the BFGS and L-BFGS (Jorge & Stephen, 1999). Assume that we are given an approximate inverse Hessian matrix H^k at $\mathbf{x} = \mathbf{x}^k$. BFGS updates the inverse Hessian matrix H^{k+1} at $\mathbf{x} = \mathbf{x}^{k+1}$ as:

$$H^{k+1} = (V^k)^T H^k V^k + \rho^k \mathbf{s}^k (\mathbf{s}^k)^T, \tag{31}$$

where $V^k = I - \rho^k \mathbf{y}^k (\mathbf{s}^k)^T$, $\mathbf{s}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$, $\mathbf{y}^k = \nabla l(\mathbf{x}^{k+1}) - \nabla l(\mathbf{x}^k)$, $\rho^k = ((\mathbf{y}^k)^T \mathbf{s}^k)^{-1}$. It is easy to verify that $H^{k+1} \succ 0$, if $H^k \succ 0$ and $\rho^k > 0$ (Jorge & Stephen, 1999).

L-BFGS updates the inverse Hessian matrix by unrolling the update from BFGS back to m steps:

$$\begin{aligned}
H^k &= (V^{k-1})^T H^{k-1} V^{k-1} + \rho^{k-1} \mathbf{s}^{k-1} (\mathbf{s}^{k-1})^T \\
&= (V^{k-1})^T (V^{k-2})^T H^{k-2} V^{k-2} V^{k-1} \\
&\quad + (V^{k-1})^T \mathbf{s}^{k-2} \rho^{k-2} (\mathbf{s}^{k-2})^T V^{k-1} \\
&\quad + \rho^{k-1} \mathbf{s}^{k-1} (\mathbf{s}^{k-1})^T \\
&= \left(U^{k,m}\right)^T H^{k-m} U^{k,m} \\
&\quad + \rho^{k-m} \left(U^{k,m-1}\right)^T \mathbf{s}^{k-m} (\mathbf{s}^{k-m})^T U^{k,m-1} \\
&\quad + \rho^{k-m+1} \left(U^{k,m-2}\right)^T \mathbf{s}^{k-m+1} (\mathbf{s}^{k-m+1})^T U^{k,m-2} \\
&\quad + \cdots \\
&\quad + \rho^{k-2} (V^{k-1})^T \mathbf{s}^{k-2} (\mathbf{s}^{k-2})^T V^{k-1} \\
&\quad + \rho^{k-1} \mathbf{s}^{k-1} (\mathbf{s}^{k-1})^T,
\end{aligned} \tag{32}$$

where $U^{k,m} = V^{k-m} V^{k-m+1} \cdots V^{k-1}$. For the L-BFGS, we need *not* explicitly store the approximated inverse Hessian matrix. Instead, we only require matrix-vector multiplications at each iteration, which can be implemented by a two-loop recursion with a time complexity of $O(mn)$ (Jorge & Stephen, 1999). Thus, we only store $2m$ vectors of length n: $\mathbf{s}^{k-1}, \mathbf{s}^{k-2}, \cdots, \mathbf{s}^{k-m}$ and $\mathbf{y}^{k-1}, \mathbf{y}^{k-2}, \cdots, \mathbf{y}^{k-m}$ with a storage complexity of $O(mn)$, which is very useful when n is large. In practice, L-BFGS updates H^{k-m} as $\mu^k I$, where $\mu^k = (\mathbf{s}^k)^T \mathbf{y}^k / \|\mathbf{y}^k\|^2$.

B. Properties of L-BFGS

We first show that some key sequences are bounded, which are critical for establishing some important properties of L-BFGS.

Proposition 6 *The sequence $\{\mathbf{x}^k\}$ generated by the mOWL-QN algorithm is bounded. Let $\mathbf{s}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$, $\mathbf{y}^k = \nabla l(\mathbf{x}^{k+1}) - \nabla l(\mathbf{x}^k)$. Then $\{\mathbf{s}^k\}$, $\{\mathbf{y}^k\}$ and $\{\mathbf{v}^k\}$ are also bounded.*

Proof Proposition 5 guarantees that both line search criteria in QN-step (Eq. (7)) and GD-step (Eq. (8)) can be satisfied in a finite number of trials with some $\alpha^k > 0$. By Eqs. (11), (7), (8), we have

$$\begin{aligned}
f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) &\geq \gamma \alpha^k (\mathbf{v}^k)^T \mathbf{d}^k \geq 0 \text{ (QN-step)}, \\
\text{or } f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) &\geq \frac{\gamma}{2\alpha^k} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \geq 0 \text{ (GD-step)},
\end{aligned} \tag{33}$$

which imply that $\{f(\mathbf{x}^k)\}$ is decreasing. Hence for all $k \geq 1$, $f(\mathbf{x}^k) \leq f(\mathbf{x}^0)$. Assume that $\{\mathbf{x}^k\}$ is unbounded. Then there exists a subsequence $\{\mathbf{x}^k\}_{\tilde{\mathcal{K}}}$ such that $\{\|\mathbf{x}^k\|_1\}_{\tilde{\mathcal{K}}} \to \infty$. Recall that $l(\mathbf{x})$ is bounded from below (see Section 2). Thus, we have $\{f(\mathbf{x}^k)\}_{\tilde{\mathcal{K}}} \to \infty$, which leads to a contradiction with that $f(\mathbf{x}^k) \leq f(\mathbf{x}^0), \forall k \geq 1$. Therefore, $\{\mathbf{x}^k\}$ is bounded, which immediately imply that $\{\mathbf{s}^k\}$ is also bounded. Recalling that $\nabla l(\mathbf{x})$ is L-Lipschitz continuous, we obtain that $\|\mathbf{y}^k\| \leq L\|\mathbf{x}^k - \mathbf{x}^{k+1}\|$ and hence $\{\mathbf{y}^k\}$ is bounded. Since $-\mathbf{v}^k \in \partial f(\mathbf{x}^k)$, then based on the Proposition B.24(b) in Bertsekas (1999), we obtain that $\{\mathbf{v}^k\}$ is bounded.

Based on Proposition 6, we present the following important properties of L-BFGS.

Proposition 7 *In the course of the inversion Hessian matrix update using L-BFGS, let $\{H^0\}$ and $\{H^{k-m}\}$ be bounded and positive definite, and $\{\mathbf{x}^k\}$, $\{\mathbf{s}^k\}$, $\{\mathbf{v}^k\}$, $\{\mathbf{y}^k\}$ and $\{\rho^k\}$ be bounded, where $\mathbf{s}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$, $\mathbf{y}^k = \nabla l(\mathbf{x}^{k+1}) - \nabla l(\mathbf{x}^k)$ and $\rho^k = ((\mathbf{y}^k)^T\mathbf{s}^k)^{-1}$. Then there exists a positive constant M such that for all $\mathbf{x} \in \mathbb{R}^n$ and all $k \geq 1$: $\mathbf{x}^T H^k \mathbf{x} \leq M\|\mathbf{x}\|^2$. That is, the eigenvalues of H^k are uniformly bounded from above by M. Moreover, $\{\mathbf{d}^k\}$ and $\{\mathbf{p}^k\}$ are bounded.*

Proof When $k \leq m$ (m is the unrolling steps of L-BFGS), L-BFGS is equivalent to BFGS and H^k is updated by the recursive relationship in Eq. (31). When $k > m$, H^k is updated by the recursive relationship in Eq. (32). Thus, Eqs. (31), (32) and the boundedness of $\{H^0\}$, $\{H^{k-m}\}$, $\{\mathbf{s}^k\}$, $\{\mathbf{y}^k\}$, $\{\mathbf{v}^k\}$ and $\{\rho^k\}$ immediately imply that $\{\|H^k\|_F\}$ is bounded. That is, there exist an $M > 0$ such that $\|H^k\|_F \leq M$ for all $k \geq 1$. Thus, for all $k \geq 1$, $\lambda_{\max}(H^k) \leq \|H^k\|_F \leq M$, where $\lambda_{\max}(H^k)$ is the largest eigenvalue of H^k. That is, there exists a positive constant M such that for all $\mathbf{x} \in \mathbb{R}^n$ and all $k \geq 1$: $\mathbf{x}^T H^k \mathbf{x} \leq M\|\mathbf{x}\|^2$. Thus, the eigenvalues of H^k are uniformly bounded from above by M. It easily follows that $\{\mathbf{d}^k\}$ and $\{\mathbf{p}^k\}$ are bounded by noticing that $\{\mathbf{v}^k\}$ is bounded.

Remark 4 *We discuss how to guarantee that the conditions in Proposition 7 are satisfied in practical L-BFGS updates. We usually choose H^0 and H^{k-m} as multiple identity matrices such that $\{H^0\}$ and $\{H^{k-m}\}$ are bounded and positive definite. Proposition 6 guarantees that $\{\mathbf{x}^k\}$, $\{\mathbf{s}^k\}$, $\{\mathbf{v}^k\}$ and $\{\mathbf{y}^k\}$ are bounded. To guarantee that $\{\rho^k\}$ is also bounded, we adopt a similar strategy presented in Byrd et al. (1995); Andrew & Gao (2007): choose a small positive constant δ and perform L-BFGS updates only when $(\mathbf{s}^k)^T\mathbf{y}^k \geq \delta$.*

Remark 5 *To guarantee the eigenvalues of H^k are uniformly bounded from below by a positive constant, we can add a small positive diagonal matrix νI to H^k (e.g., $\nu = 10^{-12}$). Thus, the eigenvalues of H^k are both uniformly bounded from below by ν and uniformly bounded from above by M, respectively.*

C. Proof of Proposition 5 and Auxiliary Propositions

We present the following proposition which is useful to prove Proposition 5.

Proposition 8 *At the point $\mathbf{x} = \mathbf{x}^k$ with the vector $\mathbf{v}^k = -\diamond f(\mathbf{x}^k)$, if $\mathbf{p}^k = \pi(\mathbf{d}^k; \mathbf{v}^k)$ is a non-zero vector, then $f'(\mathbf{x}^k; \mathbf{p}^k) = -(\mathbf{v}^k)^T\mathbf{p}^k < 0$, where $f'(\mathbf{x}^k; \mathbf{p}^k)$ denotes the directional derivative of $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^k$ along the direction $\mathbf{p}^k$ defined as follows:*

$$f'(\mathbf{x}^k; \mathbf{p}^k) = \lim_{\alpha \downarrow 0} \frac{f(\mathbf{x}^k + \alpha\mathbf{p}^k) - f(\mathbf{x}^k)}{\alpha}. \tag{34}$$

Proof According to the property of the directional derivative of a convex function (Bertsekas, 1999), we have

$$f'(\mathbf{x}^k; \mathbf{p}^k) = \max_{\mathbf{g}^k \in \partial f(\mathbf{x}^k)} (\mathbf{g}^k)^T\mathbf{p}^k = \sum_{i=1}^{n} \max_{g_i^k \in \partial_i f(\mathbf{x}^k)} g_i^k p_i^k.$$

Noticing that $\diamond_i f(\mathbf{x}^k) = \nabla_i l(\mathbf{x}^k) + \lambda\sigma(x_i^k)$ is the unique element of $\partial_i f(\mathbf{x}^k)$ whenever $x_i^k \neq 0$, we have

$$\begin{aligned} f'(\mathbf{x}^k; \mathbf{p}^k) &= \sum_{i \in \mathcal{A}_k} \diamond_i f(\mathbf{x}^k) p_i^k + \sum_{i \in \mathcal{A}_k^c} \max_{g_i^k \in \partial_i f(\mathbf{x}^k)} g_i^k p_i^k. \\ &= \sum_{i \in \mathcal{A}_k} \diamond_i f(\mathbf{x}^k) p_i^k + \sum_{i \in \mathcal{A}_k^c} \max_{g_i^k \in \partial_i f(\mathbf{x}^k)} g_i^k \sigma(v_i^k)|p_i^k|, \end{aligned}$$

where $\mathcal{A}_k = \{i : x_i^k \neq 0\}$, $\mathcal{A}_k^c = \{i : x_i^k = 0\}$ and the last equality is due to $p_i^k = \pi_i(d_i^k; v_i^k)$. We now focus on $x_i^k = 0$ in the following three cases:

(1) If $v_i^k > 0$, then $\diamond_i f(\mathbf{x}^k) = \nabla_i l(\mathbf{x}^k) + \lambda < 0$ and hence $\nabla_i l(\mathbf{x}^k) - \lambda \leq g_i^k \leq \nabla_i l(\mathbf{x}^k) + \lambda < 0$. Thus, we should choose $g_i^k = \diamond_i f(\mathbf{x}^k)$ to make $g_i^k \sigma(v_i^k)|p_i^k|$ achieve the maximum value.

(2) If $v_i^k < 0$, then $\diamond_i f(\mathbf{x}^k) = \nabla_i l(\mathbf{x}^k) - \lambda > 0$ and hence $0 < \nabla_i l(\mathbf{x}^k) - \lambda \leq g_i^k \leq \nabla_i l(\mathbf{x}^k) + \lambda$. Thus, we should choose $g_i^k = \diamond_i f(\mathbf{x}^k)$ to make $g_i^k \sigma(v_i^k)|p_i^k|$ achieve the maximum value.

(3) If $v_i^k = 0$, then $g_i^k \sigma(v_i^k)|p_i^k| = 0$ for any $g_i^k \in \partial_i f(\mathbf{x}^k)$.

Combining the above three cases, we have:

$$\begin{aligned} f'(\mathbf{x}^k;\mathbf{p}^k) &= \sum_{i\in\mathcal{A}_k} \diamond_i f(\mathbf{x}^k) p_i^k + \sum_{i\in\mathcal{A}_k^c} \diamond_i f(\mathbf{x}^k)\sigma(v_i^k)|p_i^k| \\ &= \sum_{i\in\mathcal{A}_k} \diamond_i f(\mathbf{x}^k) p_i^k + \sum_{i\in\mathcal{A}_k^c} \diamond_i f(\mathbf{x}^k) p_i^k \\ &= \diamond f(\mathbf{x}^k)^T \mathbf{p}^k = -(\mathbf{v}^k)^T \mathbf{p}^k < 0, \end{aligned}$$

where the last inequality follows from that $\mathbf{p}^k = \pi(\mathbf{d}^k;\mathbf{v}^k)$ and the condition $\mathbf{p}^k \neq \mathbf{0}$.

Based on Proposition 8, we prove Proposition 5 as follows:

Proposition 5 (a) For QN-step, let's define

$$\mathcal{B}_k = \{i : x_i^k p_i^k < 0\} \text{ and } \bar{\alpha}_1^k = \begin{cases} \min_{i\in\mathcal{B}_k} \frac{|x_i^k|}{|p_i^k|}, & \text{if } \mathcal{B}_k \neq \emptyset, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then for all $\alpha \in (0, \bar{\alpha}_1^k)$, we have

$$\mathbf{x}^k(\alpha) = \pi(\mathbf{x}^k + \alpha\mathbf{p}^k; \boldsymbol{\xi}^k) = \mathbf{x}^k + \alpha\mathbf{p}^k. \tag{35}$$

Define

$$s(\alpha) = f(\mathbf{x}^k + \alpha\mathbf{p}^k),\ h(\alpha) = \frac{s(\alpha) - s(0)}{\alpha}.$$

Since f is convex, $s(\alpha)$ is convex. Let $0 < \alpha \leq \alpha'$. Then the convexity of $s(\alpha)$ leads to

$$s(\alpha) \leq \frac{\alpha}{\alpha'} s(\alpha') + \frac{\alpha' - \alpha}{\alpha'} s(0).$$

Thus,

$$\frac{s(\alpha) - s(0)}{\alpha} \leq \frac{s(\alpha') - s(0)}{\alpha'},$$

which indicates that $h(\alpha)$ is an increasing function in the interval $(0, \infty)$. Recalling the definition of the directional derivative in Eq. (34), $\gamma \in (0, 1)$ and Proposition 8, we have

$$\lim_{\alpha\downarrow 0} \frac{s(\alpha) - s(0)}{\alpha} = -(\mathbf{v}^k)^T \mathbf{p}^k \leq -(\mathbf{v}^k)^T \mathbf{d}^k < -\gamma(\mathbf{v}^k)^T \mathbf{d}^k,$$

where the first inequality follows from Eq. (11) and the last inequality follows from $\gamma \in (0, 1)$ and $(\mathbf{v}^k)^T \mathbf{d}^k > 0$ whenever $\mathbf{x}^k$ is not a global minimizer of problem (1) [see Eq. (11) and Proposition 9]. Thus, there exists an $\bar{\alpha}_2^k \in (0, \min(\alpha_0, \bar{\alpha}_1^k))$ such that

$$\frac{s(\alpha) - s(0)}{\alpha} \leq -\gamma(\mathbf{v}^k)^T \mathbf{d}^k,\ \forall 0 < \alpha \leq \bar{\alpha}_2^k. \tag{36}$$

Recall that $h(\alpha)$ is continuous and increasing in the interval $(0, \infty)$. Thus, considering Eq. (36) and the backtracking form of the line search in QN-step (Eq. (7)), there exists an α with $\alpha \geq \bar{\alpha}^k = \beta\bar{\alpha}_2^k > 0$ such that

$$\frac{s(\alpha) - s(0)}{\alpha} \leq -\gamma(\mathbf{v}^k)^T \mathbf{d}^k. \tag{37}$$

Substituting the definition of $s(\alpha)$ into Eq. (37) and considering that Eq. (35) holds for all $\alpha \in (0, \bar{\alpha}_1^k)$, we obtain that there exists an $\alpha \in [\bar{\alpha}^k, \alpha_0]$ such that the line search criterion in Eq. (7) is satisfied.

(b) For GD-step, we have

$$\nabla l(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha)-\mathbf{x}^k)+\frac{1}{2\alpha}\|\mathbf{x}^k(\alpha)-\mathbf{x}^k\|^2+\lambda\|\mathbf{x}^k(\alpha)\|_1\leq\lambda\|\mathbf{x}^k\|_1. \tag{38}$$

Noticing that $\nabla l(\mathbf{x})$ is Lipschitz continuous with constant L, we have

$$l(\mathbf{x}^k(\alpha))\leq l(\mathbf{x}^k)+\nabla l(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha)-\mathbf{x}^k)+\frac{L}{2}\|\mathbf{x}^k(\alpha)-\mathbf{x}^k\|^2,$$

which together with Eq. (38) and $f(\mathbf{x})=l(\mathbf{x})+\lambda\|\mathbf{x}\|_1$ implies that

$$f(\mathbf{x}^k(\alpha))\leq f(\mathbf{x}^k)-\frac{1-\alpha L}{2\alpha}\|\mathbf{x}^k(\alpha)-\mathbf{x}^k\|^2.$$

Thus, the line search in Eq. (8) is satisfied if

$$\gamma\leq 1-\alpha L \text{ and } 0<\alpha\leq\alpha_0.$$

Considering the backtracking form of the line search in GD-step (Eq. (8)), we obtain that the line search criterion in Eq. (8) is satisfied whenever $\alpha\geq\beta\min(\alpha_0,(1-\gamma)/L)$.

D. More Optimality Conditions for Problem (1)

Proposition 9 *Let* $\mathbf{d}^k=H^k\mathbf{v}^k$, $\mathbf{p}^k=\pi(\mathbf{d}^k;\mathbf{v}^k)$, $\mathbf{q}_\alpha^k=\frac{1}{\alpha}(\pi(\mathbf{x}^k+\alpha\mathbf{p}^k;\boldsymbol{\xi}^k)-\mathbf{x}^k)$. *Then for all* $\alpha\in(0,\infty)$, $\mathbf{x}^k$ *is a global minimizer of problem (1)* $\Leftrightarrow\mathbf{d}^k=\mathbf{0}\Leftrightarrow\mathbf{v}^k=\mathbf{0}\Leftrightarrow\mathbf{p}^k=\mathbf{0}\Leftrightarrow\mathbf{q}_\alpha^k=\mathbf{0}$.

Proof Based on Proposition 3 and its proof, we know that $\mathbf{x}^k$ is a global minimizer of problem (1) if and only if $\mathbf{v}^k=\mathbf{0}$. Thus, we only need to prove the following equivalence to complete the proof of Proposition 9:

$$\mathbf{d}^k=\mathbf{0}\Leftrightarrow\mathbf{v}^k=\mathbf{0}\Leftrightarrow\mathbf{p}^k=\mathbf{0}\Leftrightarrow\mathbf{q}_\alpha^k=\mathbf{0}.$$

(i) We first prove $\mathbf{d}^k=\mathbf{0}\Leftrightarrow\mathbf{v}^k=\mathbf{0}$.

This equivalence immediately follows from that $\mathbf{d}^k=H^k\mathbf{v}^k$ and H^k is positive definite.

(ii) We next prove $\mathbf{v}^k=\mathbf{0}\Leftrightarrow\mathbf{p}^k=\mathbf{0}$.

- If $\mathbf{v}^k=\mathbf{0}$, then $\mathbf{p}^k=\mathbf{0}$ by the definition of $\mathbf{p}^k$.
- If $\mathbf{p}^k=\mathbf{0}$, then for all $i\in\{1,\cdots,n\}$, $d_i^kv_i^k\leq 0$ by the definition of $\mathbf{p}^k$. Thus, we have

$$(\mathbf{v}^k)^TH^k\mathbf{v}^k=\sum_{i=1}^n d_i^kv_i^k\leq 0.$$

On the other hand, due to the positive definiteness of H^k, we have

$$(\mathbf{v}^k)^TH^k\mathbf{v}^k\geq 0.$$

Thus, $(\mathbf{v}^k)^TH^k\mathbf{v}^k=0$ and hence $\mathbf{v}^k=\mathbf{0}$.

(iii) We finally prove $\mathbf{p}^k=\mathbf{0}\Leftrightarrow\mathbf{q}_\alpha^k=\mathbf{0}$.

- If $\mathbf{p}^k=\mathbf{0}$, then $\mathbf{q}_\alpha^k=\frac{1}{\alpha}(\pi(\mathbf{x}^k;\boldsymbol{\xi}^k)-\mathbf{x}^k)$. We consider the following two cases:
  (1) If $x_i^k=0$, then $(q_\alpha^k)_i=(0-0)/\alpha=0$.
  (2) If $x_i^k\neq 0$, then $(q_\alpha^k)_i=(x_i^k-x_i^k)/\alpha=0$.

Combing the above two cases, we obtain that $\mathbf{q}_\alpha^k = \mathbf{0}$.

- If $\mathbf{q}_\alpha^k = \mathbf{0}$, then $\pi(\mathbf{x}^k + \alpha\mathbf{p}^k; \boldsymbol{\xi}^k) = \mathbf{x}^k$. We consider the following two cases:
  (1) If $x_i^k = 0$, then $\pi_i(x_i^k + \alpha p_i^k; \xi_i^k) = 0$. Thus, $(0 + \alpha p_i^k)\xi_i^k = \alpha p_i^k \sigma(v_i^k) \leq 0$, which together with $p_i^k = \pi_i(d_i^k; v_i^k)$ implies $p_i^k \sigma(v_i^k) = |p_i^k| \leq 0$. Therefore, $p_i^k = 0$.
  (2) If $x_i^k \neq 0$, then $\pi_i(x_i^k + \alpha p_i^k; \xi_i^k) = x_i^k$. By the definition of $\pi_i(\cdot)$, we have $\pi_i(x_i^k + \alpha p_i^k; \xi_i^k) = x_i^k + \alpha p_i^k$ or 0. Thus, by recalling that $x_i^k \neq 0$ and $\pi_i(x_i^k + \alpha p_i^k; \xi_i^k) = x_i^k$, we must have $x_i^k + \alpha p_i^k = x_i$. Therefore, $p_i^k = 0$.

  Combing the above two cases, we obtain that $\mathbf{p}^k = \mathbf{0}$.