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

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

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

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

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

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

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

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

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 $\alpha\ge\beta\min(\alpha_0,(1-\gamma)/L)$.

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 $i\in\{1,\cdots,n\}$, $d_i^kv_i^k\le 0$ by the definition of $\mathbf{p}^k$. Thus, we have

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

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

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

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 $\mathbf{p}^k = \mathbf{0}$.