Nonoverlap-Promoting Variable Selection
– Supplementary Material

Pengtao Xie†*, Hongbao Zhang*, Yichen Zhu§ and Eric P. Xing†
†Petuum Inc, USA
§School of Computer Science, Carnegie Mellon University, USA
†School of Mathematical Sciences, Peking University, China

1 Coordinate Descent Algorithm for Learning $W$

In each iteration of the CD algorithm, one basis vector is chosen for update while the others are fixed. Without loss of generality, we assume it is $w_1$. The sub-problem defined over $w_1$ is

$$\min_{w_1} \frac{1}{2} \sum_{i=1}^{n} \| x_i - \sum_{l=2}^{m} a_{il} w_l - a_{i1} w_1 \|^2 + \frac{\lambda_2 + \lambda_3}{2} \| w_1 \|^2 + \frac{\lambda_3}{2} \log \det (W^TW) + u^T w_1 + \frac{\rho}{2} \| w_1 - \bar{w}_1 \|^2$$

To obtain the optimal solution, we take the derivative of the objective function and set it to zero. First, we discuss how to compute the derivative of \( \log \det (W^TW) \) w.r.t. $w_1$. According to the chain rule, we have

$$\frac{\partial \log \det (W^TW)}{\partial w_1} = 2W(W^TW)^{-1}_{:,1}$$

where $(W^TW)^{-1}_{:,1}$ denotes the first column of $(W^TW)^{-1}$. Let $W_{-1} = [w_2, \cdots, w_m]$, then

$$W^TW = \begin{bmatrix} w_1^TW_{-1}w_1 & w_1^TW_{-1} \end{bmatrix}$$

According to the inverse of block matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix}$$

where $\bar{A} = (A - BD^{-1}C)^{-1}$, $\bar{B} = -(A - BD^{-1}C)^{-1}BD^{-1}$, $\bar{C} = -D^{-1}C(A - BD^{-1}C)^{-1}$, $\bar{D} = D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1}$, we have $(W^TW)^{-1}_{:,1}$ equals $[a \ b]^T$ where

$$a = (w_1^TW_{-1}w_1 - w_i^TW_{-1}(W_{-1}^TW_{-1})^{-1}W_{-1}^Tw_1)^{-1}$$

Then

$$b = -(W_{-1}^TW_{-1})^{-1}W_{-1}w_1 a$$

To this end, we obtain the full gradient of the objective function in Eq. (1):

$$\sum_{i=1}^{n} a_{i1} (a_{i1} w_1 + \sum_{l=2}^{m} a_{il} w_l - x_i) + (\lambda_2 + \lambda_3) w_1 - \lambda_3 \frac{Mw_1}{w_1^TMw_1} + \rho (w_1 - \bar{w}_1) + u$$
Setting the gradient to zero, we get
\[
((\sum_{i=1}^{n} a_{i1}^{2} + \lambda_{2} + \lambda_{3} + \rho)I - \lambda_{3}M/(w_{1}^{\top}Mw_{1}))w_{1} = \sum_{i=1}^{n} a_{i1}(x_{i} - \sum_{l=2}^{m} a_{il}w_{l}) - u + \rho w_{1}.
\]
Let \( \gamma = w_{1}^{\top}Mw_{1} \), \( c = \sum_{i=1}^{n} a_{i1}^{2} + \lambda_{2} + \lambda_{3} + \rho \), \( b = \sum_{i=1}^{n} a_{i1}(x_{i} - \sum_{l=2}^{m} a_{il}w_{l}) - u + \rho w_{j} \), then \((cI - \frac{\lambda_{3}}{\gamma}M)w_{1} = b\) and \(w_{1} = (cI - \frac{\lambda_{3}}{\gamma}M)^{-1}b\). Let \( U \Sigma U^{\top} \) be the eigen decomposition of \( M \), we have
\[
w_{1} = \gamma U(\gamma cI - \lambda_{3} \Sigma)^{-1}U^{\top}b.
\]
Then
\[
w_{1}^{\top}Mw_{1} = \gamma^{2}b^{\top}U(\gamma cI - \lambda_{3} \Sigma)^{-1}U^{\top}U(\gamma cI - \lambda_{3} \Sigma)^{-1}U^{\top}b = \gamma^{2}b^{\top}(U^{\top}b)(U^{\top}b)^{-1} = \gamma
\]

The matrix \( A = W_{-1}(W_{-1}^{\top}W_{-1})^{-1}W_{-1}^{\top} \) is idempotent, i.e., \( AA = A \), and its rank is \( m - 1 \). According to the property of idempotent matrix, the first \( m - 1 \) eigenvalues of \( A \) equal to one and the rest equal to zero. Thereafter, the first \( m - 1 \) eigenvalues of \( M = I - A \) equal to zero and the rest equal to one. Based on this property, Eq. (12) can be simplified as
\[
\gamma \sum_{s=m}^{d} \frac{(U^{\top}b)_{s}^{2}}{(rc-\lambda_{3})^{2}} = 1
\]

After simplification, it is a quadratic function where \( \gamma \) has a closed form solution. Then we plug the solution of \( \gamma \) into Eq. (11) to get the solution of \( w_{1} \).

2 Proofs

2.1 Proof of Equation (7) in the Main Paper

**Proof.** Let \( VII^{\top} \) be the eigen-decomposition of the Gram matrix \( G = W^{\top}W \), where \( [v_{1}, \cdots, v_{m}] \) are the eigenvectors and \( \pi_{1}, \cdots, \pi_{m} \) are the eigenvalues. Then \( G - I = V(\Pi - I)V^{\top} = \sum_{j=1}^{m}(\pi_{j} - 1)v_{j}v_{j}^{\top} \). By Cauchy-Schwarz inequality, we have \( \|v_{j}v_{j}^{\top}\|_{1} \leq (v_{j}^{\top}v_{j}) \cdot m = m \). Thus,
\[
\|G - I\|_{1} = \sum_{j=1}^{m}(\pi_{j} - 1)\|v_{j}v_{j}^{\top}\|_{1} \leq \sum_{j=1}^{m}|\pi_{j} - 1|\|v_{j}v_{j}^{\top}\|_{1} \leq \sum_{j=1}^{m}|\pi_{j} - 1|m = m\mathcal{C}(W)
\]

2.2 Proof of Lemma 1 in the Main Paper

**Proof.** Let \( \mathcal{U} = \{ u : (x, y) \to \|W(x - y)\|^{2}_{2} \} \) be the set of hypothesis \( u(x, y) = \|W(x - y)\|^{2}_{2} \), and \( \mathcal{R}(\mathcal{U}) \) be the Rademacher complexity \( \mathcal{I} \) of \( \mathcal{U} \) which is defined as:
\[
\mathcal{R}(\mathcal{U}) = \mathbb{E}_{S_{N}, \sigma} \sup_{u \in \mathcal{U}} \frac{1}{N} \sum_{n=1}^{N} \sigma_{n}\|W(x_{n} - y_{n})\|^{2}_{2},
\]
where \( S_{N} = ((x_{1}, y_{1}, t_{1}), (x_{2}, y_{2}, t_{2}), \cdots, (x_{N}, y_{N}, t_{N})) \) are the training examples, \( \sigma_{n} \in \{-1, 1\} \) are the Rademacher variables, and \( \sigma = (\sigma_{1}, \sigma_{2}, \cdots, \sigma_{N}) \).

Lemma 3 shows that the generalization error can be bounded using the Rademacher complexity. Its proof is adapted from \( \mathcal{I} \). Readers only need to notice \( x + 1 \) is an upper bound of \( \log(1 + \exp(x)) \) for \( x > 0 \).

**Lemma 3.** With probability at least \( 1 - \delta \), we have
\[
L(u) - \hat{L}(u) \leq 2\mathcal{R}(\mathcal{U}) + \sup_{x, y, W' \in \mathcal{W}} (\|W'(x - y)\|^{2}_{2} + 1)\sqrt{\frac{2\log(1/\delta)}{N}}.
\]
We then bound $\mathcal{R}(U)$ and $\sup_{x,y, w' \in W} \|W' (x - y)\|_2^2$. The result is in the following lemma.

**Lemma 4.** Suppose $\sup_{(x,y)} \|x - y\|_2 \leq B_0$, then we have

$$\mathcal{R}(U) \leq \frac{2B_0^2 \sqrt{m}}{\sqrt{N}} (\bar{\mathcal{C}}(W) + 1),$$

and

$$\sup_{x,y, w' \in W} \|W' (x - y)\|_2^2 \leq (\bar{\mathcal{C}}(W) + m)B_0^2$$

**Proof:** We first give bound on $\mathcal{R}(U)$. Let $\mathcal{R}(\mathcal{S}) = \{ s : (x,y) \rightarrow \sum_{j=1}^m |\langle w_j, x - y \rangle|, W \in \mathcal{W} \}$ be the set of hypothesis $s(x,y) = \sum_{j=1}^m |\langle w_j, x - y \rangle|$. Denote $|\langle w_j, x_n - y_n \rangle| = |\langle w_j, a_{n,j}(x_n - y_n) \rangle|$, where $a_{n,j} \in \{-1, 1\}$. Then

$$\mathcal{R}(\mathcal{S}) = \mathbb{E}_{\mathcal{S}_N, \sigma} \sup_{W} \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^m \sigma_n a_{n,j} (x_n - y_n).$$

We first bound $\mathcal{R}(\mathcal{S})$.

$$\mathcal{R}(\mathcal{S}) = \mathbb{E}_{\mathcal{S}_N, \sigma} \sup_{W} \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^m \langle w_j, \sum_{n=1}^N \sigma_n a_{n,j} (x_n - y_n) \rangle$$

$$= \mathbb{E}_{\mathcal{S}_N, \sigma} \sup_{W} \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^m \sigma_n a_{n,j} (x_n - y_n)$$

$$\leq \mathbb{E}_{\mathcal{S}_N, \sigma} \sup_{W} \frac{1}{N} \left[ \sum_{j=1}^m \|w_j\|_2 \sum_{j=1}^m \|w_j\|_2 \right] \left[ \sum_{n=1}^N \sigma_n a_{n,j} (x_n - y_n), \sum_{n=1}^N \sigma_n a_{n,j} (x_n - y_n) \right]$$

Applying Jensen’s inequality to Eq. (15), we have

$$\mathcal{R}(\mathcal{S}) \leq \mathbb{E}_{\mathcal{S}_N} \sup_{W} \frac{1}{N} \sqrt{\sum_{j,k=1}^m |\langle w_j, w_k \rangle|} \sqrt{\mathbb{E}_{\sigma} \sum_{n=1}^N \sigma_n a_{n,j} (x_n - y_n), \sum_{n=1}^N \sigma_n a_{n,j} (x_n - y_n)} \tag{16}$$

Combining Eq. (16) with the inequality $\sum_{j,k=1}^m |\langle w_j, w_k \rangle| - \delta_{j,k} \leq mC(W)$, we have

$$\mathcal{R}(\mathcal{S}) \leq \mathbb{E}_{\mathcal{S}_N} \sup_{W} \frac{1}{N} \sqrt{mC(W) + m} \sum_{n=1}^N \|x_n - y_n\|_2$$

$$\leq \frac{\sqrt{m}}{\sqrt{N}} \sup_{W} \sqrt{|\mathcal{C}(W)| + 1} B_0$$

Let $w$ denote any column vector of $W \in \mathcal{W}$ and $x$ denote any data example. According to the composition property of Rademacher complexity (Theorem 12 in [1]), we have

$$\mathcal{R}(U) \leq 2 \sup_{w,x} (w, x) \mathcal{R}(\mathcal{S})$$

$$\leq 2 \sup_w \|w\|_2 B_0 \mathcal{R}(\mathcal{S})$$

$$\leq 2 \sup_w \|w\|_1 B_0 \mathcal{R}(\mathcal{S})$$

$$\leq 2 \sup_{W \in \mathcal{W}} \|W\|_1 B_0 \mathcal{R}(\mathcal{S})$$

$$\leq \frac{2B_0^2 \sqrt{m}}{\sqrt{N}} \sup_{W \in \mathcal{W}} \|W\|_1 \sqrt{|\mathcal{C}(W)| + 1}$$
Next we give bound on \( \sup_{\mathbf{x}, \mathbf{y}, \mathbf{W}^\prime \in \mathcal{W}} \| \mathbf{W}'(\mathbf{x} - \mathbf{y}) \|_2^2 \).

\[
\sup_{\mathbf{x}, \mathbf{y}, \mathbf{W}^\prime \in \mathcal{W}} \| \mathbf{W}'(\mathbf{x} - \mathbf{y}) \|_2^2 \leq \sup_{\mathbf{W}^\prime \in \mathcal{W}} m \sum_{j=1}^m \langle \mathbf{w}'_j, \mathbf{w}'_j \rangle \sup_{(\mathbf{x}, \mathbf{y})} \| \mathbf{x} - \mathbf{y} \|_2^2 \\
= \sup_{\mathbf{W}^\prime \in \mathcal{W}} \text{tr}(\mathbf{W}'^\top \mathbf{W}') \sup_{(\mathbf{x}, \mathbf{y})} \| \mathbf{x} - \mathbf{y} \|_2^2 \\
\leq (\bar{C}(\mathbf{W}) + m) B_0^2
\]

Combining Lemma 4 with Lemma 3, we complete the proof of Lemma 1.

### 2.3 Proof of Lemma 2 in the Main Paper

**Proof.** The function \( g(x) = x - \log(x + 1) \) is decreasing on \((-1, 0]\), increasing on \([0, \infty)\), \( g(0) = 0 \), and \( g(-t) > g(t) \) for \( \forall 0 \leq t < 1 \). We have

\[
\Omega_{\text{ldd}}(\mathbf{W}) = \text{tr}(\mathbf{W}^\top \mathbf{W}) - \log \det(\mathbf{W}^\top \mathbf{W}) - m \\
= \sum_{j=1}^m g(\pi_j - 1) \\
\geq \sum_{j=1}^m g(|\pi_j - 1|) \\
\geq g(\frac{1}{m} \sum_{j=1}^m |\pi_j - 1|) m \\
= g(\bar{C}(\mathbf{W}) / m) m
\]

The first inequality is due to \( g(-t) > g(t) \), and the second inequality can be attained by Jensen’s inequality. Finally we have

\[ g(\bar{C}(\mathbf{W}) / m) m \leq \Omega_{\text{ldd}}(\mathbf{W}). \]

Thus, we have

\[ \bar{C}(\mathbf{W}) \leq g^{-1}(\Omega_{\text{ldd}}(\mathbf{W}) / m) m. \]

**References**