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 $\mathbf{w}_1$. The sub-problem defined over $\mathbf{w}_1$ is

$$\min_{\mathbf{w}_1} \quad \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \sum_{l=2}^{m} a_{il}\mathbf{w}_l - a_{i1}\mathbf{w}_1\|_2^2 + \frac{\lambda_2+\lambda_3}{2}\|\mathbf{w}_1\|_2^2 - \frac{\lambda_3}{2}\mathrm{logdet}(\mathbf{W}^\top\mathbf{W}) + \mathbf{u}^\top\mathbf{w}_1 + \frac{\rho}{2}\|\mathbf{w}_1 - \widetilde{\mathbf{w}}_1\|_2^2 \tag{1}$$

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 logdet$(\mathbf{W}^\top\mathbf{W})$ w.r.t $\mathbf{w}_1$. According to the chain rule, we have

$$\frac{\partial \mathrm{logdet}(\mathbf{W}^\top\mathbf{W})}{\partial \mathbf{w}_1} = 2\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}_{:,1} \tag{2}$$

where $(\mathbf{W}^\top\mathbf{W})^{-1}_{:,1}$ denotes the first column of $(\mathbf{W}^\top\mathbf{W})^{-1}$. Let $\mathbf{W}_{\neg 1} = [\mathbf{w}_2, \cdots, \mathbf{w}_m]$, then

$$\mathbf{W}^\top\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^\top\mathbf{w}_1 & \mathbf{w}_1^\top\mathbf{W}_{\neg 1} \\ \mathbf{W}_{\neg 1}^\top\mathbf{w}_1 & \mathbf{W}_{\neg 1}^\top\mathbf{W}_{\neg 1} \end{bmatrix} \tag{3}$$

According to the inverse of block matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \widetilde{\mathbf{A}} & \widetilde{\mathbf{B}} \\ \widetilde{\mathbf{C}} & \widetilde{\mathbf{D}} \end{bmatrix} \tag{4}$$

where $\widetilde{\mathbf{A}} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$, $\widetilde{\mathbf{B}} = -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}$, $\widetilde{\mathbf{C}} = -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$, $\widetilde{\mathbf{D}} = \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}$, we have $(\mathbf{W}^\top\mathbf{W})^{-1}_{:,1}$ equals $[\mathbf{a} \quad \mathbf{b}^\top]^\top$ where

$$\mathbf{a} = (\mathbf{w}_1^\top\mathbf{w}_1 - \mathbf{w}_1^\top\mathbf{W}_{\neg 1}(\mathbf{W}_{\neg 1}^\top\mathbf{W}_{\neg 1})^{-1}\mathbf{W}_{\neg 1}^\top\mathbf{w}_1)^{-1} \tag{5}$$

$$\mathbf{b} = -(\mathbf{W}_{\neg 1}^\top\mathbf{W}_{\neg 1})^{-1}\mathbf{W}_{\neg 1}^\top\mathbf{w}_1\mathbf{a} \tag{6}$$

Then

$$\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}_{:,1} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_{\neg 1} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \frac{\mathbf{M}\mathbf{w}_1}{\mathbf{w}_1^\top\mathbf{M}\mathbf{w}_1}. \tag{7}$$

where

$$\mathbf{M} = \mathbf{I} - \mathbf{W}_{\neg 1}(\mathbf{W}_{\neg 1}^\top\mathbf{W}_{\neg 1})^{-1}\mathbf{W}_{\neg 1}^\top. \tag{8}$$

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

$$\sum_{i=1}^{n} a_{i1}(a_{i1}\mathbf{w}_1 + \sum_{l=2}^{m} a_{il}\mathbf{w}_l - \mathbf{x}_i) + (\lambda_2+\lambda_3)\mathbf{w}_1 - \lambda_3\frac{\mathbf{M}\mathbf{w}_1}{\mathbf{w}_1^\top\mathbf{M}\mathbf{w}_1} + \rho(\mathbf{w}_1 - \widetilde{\mathbf{w}}_1) + \mathbf{u}. \tag{9}$$

Setting the gradient to zero, we get

$$((\textstyle\sum_{i=1}^n a_{i1}^2 + \lambda_2 + \lambda_3 + \rho)\mathbf{I} - \lambda_3 \mathbf{M}/(\mathbf{w}_1^\top \mathbf{M}\mathbf{w}_1))\mathbf{w}_1 = \sum_{i=1}^n a_{i1}(\mathbf{x}_i - \sum_{l=2}^m a_{il}\mathbf{w}_l) - \mathbf{u} + \rho\widetilde{\mathbf{w}}_1. \tag{10}$$

Let $\gamma = \mathbf{w}_1^\top \mathbf{M}\mathbf{w}_1$, $c = \sum_{i=1}^n a_{i1}^2 + \lambda_2 + \lambda_3 + \rho$, $\mathbf{b} = \sum_{i=1}^n a_{i1}(\mathbf{x}_i - \sum_{l=2}^m a_{il}\mathbf{w}_l) - \mathbf{u} + \rho\widetilde{\mathbf{w}}_j$, then $(c\mathbf{I} - \frac{\lambda_3}{\gamma}\mathbf{M})\mathbf{w}_1 = \mathbf{b}$ and $\mathbf{w}_1 = (c\mathbf{I} - \frac{\lambda_3}{\gamma}\mathbf{M})^{-1}\mathbf{b}$. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$ be the eigen decomposition of $\mathbf{M}$, we have

$$\mathbf{w}_1 = \gamma \mathbf{U}(\gamma c\mathbf{I} - \lambda_3\mathbf{\Sigma})^{-1}\mathbf{U}^\top\mathbf{b}. \tag{11}$$

Then

$$\begin{aligned} &\mathbf{w}_1^\top \mathbf{M}\mathbf{w}_1 \\ &= \gamma^2 \mathbf{b}^\top \mathbf{U}(\gamma c\mathbf{I} - \lambda_3\mathbf{\Sigma})^{-1}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}(\gamma c\mathbf{I} - \lambda_3\mathbf{\Sigma})^{-1}\mathbf{U}^\top\mathbf{b} \\ &= \gamma^2 \mathbf{b}^\top \mathbf{U}(\gamma c\mathbf{I} - \lambda_3\mathbf{\Sigma})^{-1}\mathbf{\Sigma}(\gamma c\mathbf{I} - \lambda_3\mathbf{\Sigma})^{-1}\mathbf{U}^\top\mathbf{b} \\ &= \gamma^2 \sum_{s=1}^d \frac{(\mathbf{U}^\top\mathbf{b})_s^2 \Sigma_{ss}}{(rc - \lambda_3\Sigma_{ss})^2} = \gamma \end{aligned} \tag{12}$$

The matrix $\mathbf{A} = \mathbf{W}_{\neg 1}(\mathbf{W}_{\neg 1}^\top \mathbf{W}_{\neg 1})^{-1}\mathbf{W}_{\neg 1}^\top$ is idempotent, i.e., $\mathbf{A}\mathbf{A} = \mathbf{A}$, and its rank is $m-1$. According to the property of idempotent matrix, the first $m-1$ eigenvalues of $\mathbf{A}$ equal to one and the rest equal to zero. Thereafter, the first $m-1$ eigenvalues of $\mathbf{M} = \mathbf{I} - \mathbf{A}$ equal to zero and the rest equal to one. Based on this property, Eq.(12) can be simplified as

$$\gamma \sum_{s=m}^d \frac{(\mathbf{U}^\top\mathbf{b})_s^2}{(rc - \lambda_3)^2} = 1 \tag{13}$$

After simplification, it is a quadratic function where γ has a closed form solution. Then we plug the solution of γ into Eq.(11) to get the solution of $\mathbf{w}_1$.

2 Proofs

2.1 Proof of Equation (7) in the Main Paper

Proof. Let $\mathbf{V}\mathbf{\Pi}\mathbf{V}^\top$ be the eigen-decomposition of the Gram matrix $\mathbf{G} = \mathbf{W}^\top\mathbf{W}$, where $[\mathbf{v}_1, \cdots, \mathbf{v}_m]$ are the eigenvectors and $\pi_1, \cdots, \pi_m$ are the eigenvalues. Then $\mathbf{G} - \mathbf{I} = \mathbf{V}(\mathbf{\Pi} - \mathbf{I})\mathbf{V}^\top = \sum_{j=1}^m (\pi_j - 1)\mathbf{v}_j\mathbf{v}_j^\top$. By Cauchy-Schwarz inequality, we have $\|\mathbf{v}_j\mathbf{v}_j^\top\|_1 \leq (\mathbf{v}_j^\top\mathbf{v}_j)\cdot m = m$. Thus,

$$\|\mathbf{G} - \mathbf{I}\|_1 = \|\sum_{j=1}^m (\pi_j - 1)\mathbf{v}_j\mathbf{v}_j^\top\|_1 \leq \|\sum_{j=1}^m |\pi_j - 1| \|\mathbf{v}_j\mathbf{v}_j^\top\|_1 \leq \|\sum_{j=1}^m |\pi_j - 1| m = m\mathcal{C}(\mathbf{W})$$

□

2.2 Proof of Lemma 1 in the Main Paper

Proof. Let $\mathcal{U} = \{u : (\mathbf{x}, \mathbf{y}) \to \|\mathbf{W}(\mathbf{x} - \mathbf{y})\|_2^2\}$ be the set of hypothesis $u(\mathbf{x}, \mathbf{y}) = \|\mathbf{W}(\mathbf{x} - \mathbf{y})\|_2^2$, and $\mathcal{R}(\mathcal{U})$ be the Rademacher complexity (1) of $\mathcal{U}$ which is defined as:

$$\mathcal{R}(\mathcal{U}) = \mathbb{E}_{S_N,\sigma} \sup_{u \in \mathcal{U}} \frac{1}{n} \sum_{n=1}^N \sigma_n \|\mathbf{W}(\mathbf{x}_n - \mathbf{y}_n)\|_2^2,$$

where $S_N = ((\mathbf{x}_1, \mathbf{y}_1, t_1), (\mathbf{x}_2, \mathbf{y}_2, t_2) \cdots (\mathbf{x}_n, \mathbf{y}_n, t_N))$ 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 (1). Readers only need to notice $x + 1$ is an upper bound of $\log(1 + \exp(\mathbf{x}))$ for $x > 0$.

Lemma 3. *With probability at least $1 - \delta$, we have*

$$L(u) - \hat{L}(u) \leq 2\mathcal{R}(\mathcal{U}) + \sup_{\mathbf{x},\mathbf{y},\mathbf{W}' \in \mathcal{W}} (\|\mathbf{W}'(\mathbf{x} - \mathbf{y})\|_2^2 + 1)\sqrt{\frac{2\log(1/\delta)}{N}}. \tag{14}$$

We then bound $\mathcal{R}(\mathcal{U})$ and $\sup_{\mathbf{x},\mathbf{y},\mathbf{W}'\in\mathcal{W}} \|\mathbf{W}'(\mathbf{x}-\mathbf{y})\|_2^2$. The result is in the following lemma.

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

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

and

$$\sup_{\mathbf{x},\mathbf{y},\mathbf{W}'\in\mathcal{W}} \|\mathbf{W}'(\mathbf{x}-\mathbf{y})\|_2^2 \le (\widetilde{\mathcal{C}}(\mathcal{W})+m)B_0^2$$

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

$$\mathcal{R}(\mathcal{S}) = \mathbb{E}_{S_N,\sigma} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\sum_{n=1}^N \sigma_n \sum_{j=1}^m \langle \mathbf{w}_j, a_{n,j}(\mathbf{x}_n-\mathbf{y}_n)\rangle.$$

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

$$\begin{aligned}
\mathcal{R}(\mathcal{S}) &= \mathbb{E}_{S_N,\sigma} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\sum_{j=1}^m \langle \mathbf{w}_j, \sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n)\rangle \\
&= \mathbb{E}_{S_N,\sigma} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\langle \sum_{j=1}^m \mathbf{w}_j, \sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n)\rangle \\
&\le \mathbb{E}_{S_N,\sigma} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\|\sum_{j=1}^m \mathbf{w}_j\|_2 \|\sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n)\|_2 \\
&= \mathbb{E}_{S_N,\sigma} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\sqrt{\langle \sum_{j=1}^m \mathbf{w}_j, \sum_{j=1}^m \mathbf{w}_j\rangle}\sqrt{\langle \sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n), \sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n)\rangle} \qquad (15)
\end{aligned}$$

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

$$\mathcal{R}(\mathcal{S}) \le \mathbb{E}_{S_N} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\sqrt{\sum_{j,k=1}^m |\langle \mathbf{w}_j, \mathbf{w}_k\rangle|}\sqrt{\mathbb{E}_\sigma\langle \sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n), \sum_{n=1}^N \sigma_n a_{n,j}(\mathbf{x}_n-\mathbf{y}_n)\rangle} \qquad (16)$$

Combining Eq.(16) with the inequality $\sum_{j,k=1}^m |\langle \mathbf{w}_j, \mathbf{w}_k\rangle - \delta_{j,k}| \le m\mathcal{C}(\mathbf{W})$, we have

$$\begin{aligned}
\mathcal{R}(\mathcal{S}) &\le \mathbb{E}_{S_N} \sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{N}\sqrt{m\mathcal{C}(\mathbf{W})+m}\sqrt{\sum_{n=1}^N \|\mathbf{x}_n-\mathbf{y}_n\|_2} \\
&\le \frac{\sqrt{m}}{\sqrt{N}} \sup_{\mathbf{W}\in\mathcal{W}} \sqrt{(\mathcal{C}(\mathbf{W})+1)}B_0
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{R}(\mathcal{U}) &\le 2\sup_{\mathbf{w},\mathbf{x}}\langle \mathbf{w},\mathbf{x}\rangle \mathcal{R}(\mathcal{S}) \\
&\le 2\sup_{\mathbf{w}}\|\mathbf{w}\|_2 B_0 \mathcal{R}(\mathcal{S}) \\
&\le 2\sup_{\mathbf{w}}\|\mathbf{w}\|_1 B_0 \mathcal{R}(\mathcal{S}) \\
&\le 2\sup_{\mathbf{W}'\in\mathcal{W}}\|\mathbf{W}'\|_1 B_0 \mathcal{R}(\mathcal{S}) \\
&\le \frac{2B_0^2\sqrt{m}}{\sqrt{N}} \sup_{\mathbf{W}'\in\mathcal{W}}\|\mathbf{W}'\|_1 \sqrt{\widetilde{\mathcal{C}}(\mathcal{W})+1}
\end{aligned}$$

Next we give bound on $\sup_{\mathbf{x},\mathbf{y},\mathbf{W}'\in\mathcal{W}} \|\mathbf{W}'(\mathbf{x}-\mathbf{y})\|_2^2$.

$$\begin{aligned}
\sup_{\mathbf{x},\mathbf{y},\mathbf{W}'\in\mathcal{W}} \|\mathbf{W}'(\mathbf{x}-\mathbf{y})\|_2^2 &\leq \sup_{\mathbf{W}'\in\mathcal{W}} \sum_{j=1}^{m} \langle \mathbf{w}'_j, \mathbf{w}'_j \rangle \sup_{(\mathbf{x},\mathbf{y})} \|\mathbf{x}-\mathbf{y}\|_2^2 \\
&= \sup_{\mathbf{W}'\in\mathcal{W}} \mathrm{tr}(\mathbf{W}'^{\top}\mathbf{W}') \sup_{(\mathbf{x},\mathbf{y})} \|\mathbf{x}-\mathbf{y}\|_2^2 \\
&\leq (\widetilde{\mathcal{C}}(\mathcal{W}) + m) B_0^2
\end{aligned}$$

□

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

$$\begin{aligned}
\Omega_{ldd}(\mathbf{W}) &= \mathrm{tr}(\mathbf{W}^{\top}\mathbf{W}) - \log\det(\mathbf{W}^{\top}\mathbf{W}) - m \\
&= \sum_{j=1}^{m} g(\pi_j - 1) \\
&\geq \sum_{j=1}^{m} g(|\pi_j - 1|) \\
&\geq g(\frac{1}{m}\sum_{j=1}^{m} |\pi_j - 1|)m \\
&= g(\mathcal{C}(\mathbf{W})/m)m
\end{aligned}$$

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

$$g(\mathcal{C}(\mathbf{W})/m)m \leq \Omega_{ldd}(\mathbf{W}).$$

Thus, we have

$$\mathcal{C}(\mathbf{W}) \leq g^{-1}(\Omega_{ldd}(\mathbf{W})/m)m.$$

□

References

[1] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.