Supplementary Material for "Sparse learning with non-convex penalty in multi-classification" by Nan Li and Hao Helen Zhang

Before presenting the proofs of the theorems, we first state some regularity conditions, which mostly follow Fan and peng (2004). The reparameterized multinomial log-likelihood $\tilde{\mathcal{L}}$ and its associated true parameter vector $\beta^\star$ are defined in (7) and (8).

1 Regularity Conditions

1. The observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d. with multinomial distribution $(\pi_1, \ldots, \pi_K)$, $1 > \pi_k > 0, \sum_{k=1}^K \pi_k = 1$.

2. The Fisher information matrix $I(\boldsymbol{\beta}) = E\{(\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\beta}})(\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\beta}})^T\}$ is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ for all n observations. For $j, k = 1, 2, ..., d$,

$$E\{(\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_j})(\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_k})\}^2 < C_1 < \infty$$

and

$$E\{(\frac{\partial \tilde{\mathcal{L}}^2}{\partial \beta_j \partial \beta_k})\}^2 < C_2 < \infty.$$

3. There is a sufficient large enough open set w that contains $\boldsymbol{\beta}^\star$ such that for almost all n observations the density admits all third deravatives $\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_j \beta_k \beta_l}$ for all $\boldsymbol{\beta}^\star \in w$, and

$$|\frac{\partial \tilde{\mathcal{L}}^3}{\partial \beta_{kj} \partial \beta_{k_1 j_1} \partial \beta_{k_2 j_2}}| \leq M(\mathbf{x}) < \infty$$

and

$$E_{\boldsymbol{\beta}^\star}[M(\mathbf{x})] < \infty$$

for all $n, k, j, k_1, j_1, k_2, j_2$.

4. Let the first s values of $\boldsymbol{\beta}$ be nonzero, an the rest of $d_n - s_n$ values be zero. Then the $\beta_1, \beta_2, ..., \beta_s$ satisfy

$$\min_{1 \le j \le s_n} |\beta_j|/\lambda \to \infty \text{ as n} \to \infty.$$

2 Proof of Theorem 1

To prove Theorem 1, it is enough to show that for any given $\varepsilon > 0$, there exists a large enough constant C such that

$$P\{\inf_{||\mathbf{u}||=C} \tilde{R}(\boldsymbol{\beta}^\star + \mathbf{u}\sqrt{d/n}) > \tilde{R}(\boldsymbol{\beta}^\star)\} \ge 1 - \varepsilon, \qquad (\star)$$

which implies that with probability at least $1 - \varepsilon$ there exists a local minimum in the ball $\{\boldsymbol{\beta}^\star + \mathbf{u}\sqrt{d/n} : ||\mathbf{u}|| \le C\}$. This in turn implies that there exists a local minimizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\| = O_p(\sqrt{d/n})$, which is exactly what we want to show.

Notice that

$$\tilde{R}(\boldsymbol{\beta}^\star + \mathbf{u}\sqrt{d_n/n}) - \tilde{R}(\boldsymbol{\beta}^\star) = -(\tilde{\mathcal{L}}(\boldsymbol{\beta}^\star + \mathbf{u}\sqrt{d_n/n}) - \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star))$$

$$+ n\sum_{j=1}^{d_n}[J_{\lambda_n}(\max\{|\beta_{1j}^\star + u_{1j}\sqrt{d_n/n}|, |\beta_{2j}^\star + u_{2j}\sqrt{d_n/n}|, \ldots, |\beta_{(K-1),j}^\star + u_{K-1,j}\sqrt{d_n/n}|, |\sum_{l=1}^{K-1}(\beta_{lj}^\star + u_{lj}\sqrt{d_n/n}$$

$$- J_{\lambda_n}(\max\{|\beta_{1j}^\star|, |\beta_{2j}^\star|, \ldots, |\beta_{(K-1),j}^\star|, |\sum_{l=1}^{K-1}\beta_{lj}^\star|\})]$$

$$\geq -(\tilde{\mathcal{L}}(\boldsymbol{\beta}^\star + \mathbf{u}\sqrt{d_n/n}) - \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star))$$

$$+ n\sum_{j=1}^{s}[J_{\lambda_n}(\max\{|\beta_{1j}^\star + \frac{u_{1j}\sqrt{d_n}}{\sqrt{n}}|, |\beta_{2j}^\star + \frac{u_{2j}\sqrt{d_n}}{\sqrt{n}}|, \ldots, |\beta_{(K-1),j}^\star + \frac{u_{K-1,j}\sqrt{d_n}}{\sqrt{n}}|, |\sum_{l=1}^{K-1}(\beta_{lj}^\star + \frac{u_{lj}\sqrt{d_n}}{\sqrt{n}})|\})$$

$$- J_{\lambda_n}(\max\{|\beta_{1j}^\star|, |\beta_{2j}^\star|, \ldots, |\beta_{(K-1),j}^\star|, |\sum_{l=1}^{K-1}\beta_{lj}^\star|\})]$$

follow the proof in Fan and Peng (2004)

$$= -\tilde{\mathcal{L}}'(\boldsymbol{\beta}^\star)^T\frac{\mathbf{u}\sqrt{d_n}}{\sqrt{n}} + \frac{d_n}{2n}\mathbf{u}^T\tilde{\mathcal{L}}''(\boldsymbol{\beta}^\star)\mathbf{u}\{1 + o_p(1)\} + D_3$$

$$= -\tilde{\mathcal{L}}'(\boldsymbol{\beta}^\star)^T\frac{\mathbf{u}\sqrt{d_n}}{\sqrt{n}} + \frac{d_n}{2}\mathbf{u}^T I(\boldsymbol{\beta}^\star)\mathbf{u}\{1 + o_p(1)\} + D_3$$

$$= D_1 + D_2 + D_3$$

Note that $\tilde{\mathcal{L}}'(\boldsymbol{\beta}^\star)^T = O_p(\sqrt{d_n n})$, thus D_2 is asymptotic positive and dominates D_1 by choosing a sufficiently large C. Since the supSCAD penalty is flat for coefficients of magnitude larger than $a\lambda_n$ as $n \to \infty$,

$$D_3 = n\sum_{j=1}^{s_n}[J_{\lambda_n}(\max\{|\beta_{1j}^\star + \frac{u_{1j}}{\sqrt{n}}|, |\beta_{2j}^\star + \frac{u_{2j}}{\sqrt{n}}|, \ldots, |\beta_{(K-1),j}^\star + \frac{u_{K-1,j}}{\sqrt{n}}|, |\sum_{l=1}^{K-1}(\beta_{lj}^\star + \frac{u_{lj}}{\sqrt{n}})|\})$$
$$- J_{\lambda_n}(\max\{|\beta_{1j}^\star|, |\beta_{2j}^\star|, \ldots, |\beta_{(K-1),j}^\star|, |\sum_{l=1}^{K-1}\beta_{lj}^\star|\})] = 0.$$

Based on the above, $\tilde{R}(\boldsymbol{\beta}^\star + \mathbf{u}\sqrt{d_n/n}) - \tilde{R}(\boldsymbol{\beta}^\star)$ is dominated by D_2. Hence, by choosing a sufficient large C $(\star)$ holds.

3 Proof of Lemma 1

As long as the $\max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}$ is zero, then each component in $\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}$ is zero.

It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_+$ satisfying $||\boldsymbol{\beta}_+ - \boldsymbol{\beta}_+^\star|| = O_p(\sqrt{d_n/n})$ and any constant C, for $j = s_n + 1, \ldots, d$,

$$\frac{\partial \tilde{R}(\boldsymbol{\beta})}{\partial \beta_{kj}^r} > 0, \text{ for } 0 \leq \beta_{kj} \leq C\sqrt{d_n/n} \text{ and } \beta_{kj} = \max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}$$
$$\frac{\partial \tilde{R}(\boldsymbol{\beta})}{\partial \beta_{kj}^l} < 0, \text{ for } -C\sqrt{d_n/n} \leq \beta_{kj} \leq 0 \text{ and } \beta_{kj} = -\max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}$$

where $\frac{\partial \tilde{R}}{\partial \beta_{kj}^r}$ and $\frac{\partial \tilde{R}}{\partial \beta_{kj}^l}$ denote the right and left hand partial derivative respectively.

$$\frac{\partial \tilde{R}}{\partial \beta_{kj}} = -\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_{kj}} + nJ'_{\lambda_n}(|\beta_{kj}|)sgn(\beta_{kj}) \text{ when } |\beta_{kj}| = \max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}$$

By Taylor expansion,

$$\begin{aligned}\frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\beta})}{\partial \beta_{kj}} = \frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star)}{\partial \beta_{kj}} &+ \sum_{j_1=1}^{d}\sum_{k_1=1}^{K-1} \frac{\partial^2 \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star)}{\partial \beta_{kj}\partial \beta_{k_1 j_1}}(\beta_{k_1 j_1} - \beta_{k_1 j_1}^\star) \\ &+ \sum_{j_1=1}^{d_n}\sum_{k_1=1}^{K-1}\sum_{j_2=1}^{d_n}\sum_{k_2=1}^{K-1} \frac{\partial^3 \tilde{\mathcal{L}}(\boldsymbol{\beta}^*)}{\partial \beta_{kj}\partial \beta_{k_1 j_1}\partial \beta_{k_2 j_2}}(\beta_{k_1 j_1} - \beta_{k_1 j_1}^\star)(\beta_{k_2 j_2} - \beta_{k_2 j_2}^\star)\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\star$. Note that $\frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star)}{\partial \beta_{kj}} = O_p(\sqrt{d_n n})$ and $\frac{1}{n}\frac{\partial^2 \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star)}{\partial \beta_{kj}\partial \beta_{k_1 j_1}} = E\{\frac{\partial^2 \tilde{\mathcal{L}}(\boldsymbol{\beta}^\star)}{\partial \beta_{kj}\partial \beta_{k_1 j_1}}\} + o_p(1)$. We have

$$\frac{\partial \tilde{R}(\boldsymbol{\beta})}{\partial \beta_{kj}^r} = n\lambda_n\{O_p(\sqrt{d_n/n}/\lambda_n) + \frac{J'_{\lambda_n}(|\beta_{kj}|)}{\lambda_n}\} \text{ if } \beta_{kj} > 0$$
$$\frac{\partial \tilde{R}(\boldsymbol{\beta})}{\partial \beta_{kj}^l} = n\lambda_n\{O_p(\sqrt{d_n/n}/\lambda_n) - \frac{J'_{\lambda_n}(|\beta_{kj}|)}{\lambda_n}\} \text{ if } \beta_{kj} < 0$$

In both cases, the second term dominates the first term.

If $|\kappa_j| = |\sum_{m=1}^{K-1} \beta_{mj}| = \max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}$, then

$$\frac{\partial \tilde{R}}{\partial \kappa_j} = -\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_{kj}} \frac{\partial \beta_{kj}}{\partial \kappa_j} + n J'_{\lambda_n}(|\kappa_j|) sgn(\kappa_j) = -\frac{\partial \tilde{\mathcal{L}}}{\partial \kappa_j} + n J'_{\lambda_n}(|\kappa_j|) sgn(\kappa_j).$$

Therefore,

$$\frac{\partial \tilde{R}(\boldsymbol{\beta})}{\partial \kappa_j^r} = n\lambda_n \{O_p(\sqrt{d_n/n}/\lambda_n) + \frac{J'_{\lambda_n}(|\kappa_j|)}{\lambda_n}\} \text{ if } \kappa_j > 0$$
$$\frac{\partial \tilde{R}(\boldsymbol{\beta})}{\partial \kappa_j^l} = n\lambda_n \{O_p(\sqrt{d_n/n}/\lambda_n) - \frac{J'_{\lambda_n}(|\kappa_j|)}{\lambda_n}\} \text{ if } \kappa_j < 0$$

The second term still dominates the first term. Thus the result of Lemma 1 follows.

4 Proof of Theorem 2

Proof: Part 1 holds by lemma 1. For Part 2

$$\begin{aligned}
\lambda_n &\leq a^{-1} \max_{1 \leq j \leq s} \max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\} \\
&\Rightarrow \max_{1 \leq j \leq s} \max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\} \geq a\lambda_n \\
&\Rightarrow J_{\lambda_n}(\max\{|\beta_{1j}|, |\beta_{2j}|, \ldots, |\beta_{K-1,j}|, |\sum_{m=1}^{K-1} \beta_{mj}|\}) = J_{\lambda_n}(a\lambda_n) = \frac{(a+1)^2 \lambda_n^2}{2}.
\end{aligned}$$

Therefore $\arg\min \tilde{R}(\boldsymbol{\beta}) = \arg\min -\tilde{\mathcal{L}}(\boldsymbol{\beta})$. The desired result follows.

5 Algorithm 1 for supSCAD multinomial logistic regression

Algorithm 1

Initialize $\boldsymbol{\beta}^{(0)}$.

for (t = 0 to MaxIteration)

Compute $Q(\boldsymbol{\beta}^{(t)})$ in (9).

Use DCA/LLA to solve the problem (10); denote the solution by $\boldsymbol{\beta}^{(t+1)}$.

Evaluate the objective function given in (7) at $\boldsymbol{\beta}^{(t+1)}$.

if (the stopping criterion is satisfied), Break;

end

end

6 Algorithm 2 for supSCAD Multicategory Support Vector Machine

First we introduce a set of slack variables

$$\delta_{ik} = I(y_i \neq k) \text{ and } \xi_{ik} = [\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}_i + 1]_+, \text{ for } i = 1, \ldots, n, \ k = 1, \ldots, K,$$

$$\eta_j = \|\boldsymbol{\beta}_{(j)}\|_\infty = \max_{k=1,\ldots,K} |\beta_{kj}|, \text{ for } j = 1, \ldots, d,$$

and new constraints $|\beta_{kj}| \leq \eta_j$, for $k = 1, \ldots, K, j = 1, \ldots, d$. Then applying DCA and LLA, the non-convex minimization supSCAD MSVM can be solved via a sequence of LP

problems,

$$
\begin{aligned}
\textbf{DCA} \min_{\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\eta}} & \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\xi_{ik} + \lambda\sum_{j=1}^{d}\eta_j - \sum_{j=1}^{d}J'_{\lambda,2}(\eta_j^{(t)})(\eta_j - \eta_j^{(t)}) \\
\textbf{LLA} \min_{\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\eta}} & \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\xi_{ik} + \sum_{j=1}^{d}J'_{\lambda}(\eta_j^{(t)})\cdot\eta_j \\
& \text{subject to } \sum_{k=1}^{K}\beta_{k0} = 0, \quad \sum_{k=1}^{K}\beta_{kj} = 0, \quad j = 1,\ldots,d, \\
& \xi_{ik} \geq \beta_{k0} + \boldsymbol{\beta}_k^T\mathbf{x}_i + 1, \xi_{ik} \geq 0, i = 1,\ldots,n \text{ and } k = 1,\ldots,K \\
& \boldsymbol{\beta}_{(j)} \leq \eta_j\mathbf{1}_K, -\boldsymbol{\beta}_{(j)} \leq \eta_j\mathbf{1}_K, j = 1,\ldots,d.
\end{aligned} \tag{1}
$$

Algorithm 2:

1. **Initialization:** $\boldsymbol{\beta}_0^{(0)} = 0, \boldsymbol{\beta}^{(0)} = 0, \boldsymbol{\xi}^{(0)} = 0, \boldsymbol{\eta}^{(0)} = 0, t = 0.$
2. **Repeat:** solve $\boldsymbol{\beta}_0^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\xi}^{(t+1)}$, and $\boldsymbol{\eta}^{(t+1)}$ from (1), $t = t + 1$.
3. **Stop:** $\boldsymbol{\beta}_0^{(t+1)}$ and $\boldsymbol{\beta}^{(t+1)}$ meet the rule of convergence.

7 Algorithm for supSCAD MPSVM

Similarly, we can use DCA/LLA to supSCAD MPSVM by solving a series of QP problems:

$$
\begin{aligned}
\textbf{DCA} \min_{\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\eta}} & \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}[\beta_{k0} + \boldsymbol{\beta}_k^T\mathbf{x}_i + 1]^2 + \lambda\sum_{j=1}^{d}\eta_j - \sum_{j=1}^{d}J'_{\lambda,2}(\eta_j^{(t)})(\eta_j - \eta_j^{(t)}) \\
\textbf{LLA} \min_{\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\eta}} & \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}[\beta_{k0} + \boldsymbol{\beta}_k^T\mathbf{x}_i + 1]^2 + \sum_{j=1}^{d}J'_{\lambda}(\eta_j^{(t)})\cdot\eta_j \\
& \text{subject to } \sum_{k=1}^{K}\beta_{k0} = 0, \quad \sum_{k=1}^{K}\beta_{kj} = 0, j = 1,\ldots,d, \\
& \boldsymbol{\beta}_{(j)} \leq \eta_j\mathbf{1}_K, \ -\boldsymbol{\beta}_{(j)} \leq \eta_j\mathbf{1}_K, j = 1,\ldots,d.
\end{aligned}
$$

References

[1] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, **32(3)**, 928-961.