Online Supplement to
“Optimal model averaging for divergent-dimensional Poisson regressions with an application on corporate innovation outcomes”

Jiahui Zou Wendun Wang Xinyu Zhang and Guohua Zou

This technical appendix includes the proof of theorems. First, Section A.1 presents some lemmas and their proofs. Then Section A.2 and A.3 provide the proofs of Theorems 1 and 2, respectively.

A.1 Lemmas

To prove the theorems, we first establish Lemmas 1–5. All limiting results below are obtained by letting \( n \) go to infinity unless stated otherwise.

**Proof of Lemma 1.** To show the unbiasedness of our weight-choice criterion, first note that for any \( i = 1, 2, ..., n \),

\[
E_{f(y)} \{ \mu_i \log \hat{\mu}_i(w, y) \} = \sum_{j=1, j\neq i}^{n} \sum_{y_j=0}^{\infty} \sum_{y_i=0}^{\infty} \left[ \log \{ \hat{\mu}_i(w, y) \} \mu_i \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \prod_{j \neq i} \frac{e^{-\mu_j} \mu_j^{y_j}}{y_j!} \right]
\]

\[
= \sum_{j=1, j\neq i}^{n} \sum_{y_j=0}^{\infty} \sum_{y_i=0}^{\infty} \left[ (y_i + 1) \log \{ \hat{\mu}_i(w, y) \} \frac{e^{-\mu_i} \mu_i^{y_i+1}}{(y_i + 1)!} \prod_{j \neq i} \frac{e^{-\mu_j} \mu_j^{y_j}}{y_j!} \right]
\]

\[
= \sum_{j=1, j\neq i}^{n} \sum_{y_j=0}^{\infty} \sum_{y_i=0}^{\infty} \left[ y_i \log \{ \hat{\mu}_i(w, y(y_i - 1)) \} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \prod_{j \neq i} \frac{e^{-\mu_j} \mu_j^{y_j}}{y_j!} \right]
\]

\[
= \sum_{j=1, j\neq i}^{n} \sum_{y_j=0}^{\infty} \sum_{y_i=0}^{\infty} \left[ y_i \log \{ \hat{\mu}_i(w, y(y_i - 1)) \} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \prod_{j \neq i} \frac{e^{-\mu_j} \mu_j^{y_j}}{y_j!} \right]
\]

\[
= E_{f(y)} [y_i \log \{ \hat{\mu}_i(w, y(y_i - 1)) \}] . \quad (A.1)
\]

Combining (A.1) with (5) and (6), we can obtain that

\[
E_{f(y)} \{ C(w) \} = E_{f(y)} \log f(y) + E_{f(y)} \sum_{i=1}^{n} \left[ \hat{\mu}_i(w, y) + \log(y_i!) - y_i \log \{ \hat{\mu}_i(w, y(y_i - 1)) \} \right]
\]
\[ E_f(y) \log f(y) + E_f(y) \sum_{i=1}^{n} \left\{ \hat{\mu}_i(w, y) + \log(y_i!) - \mu_i \log \hat{\mu}_i(w, y) \right\} \]

\[ = E_f(y) \{ KL(w) \}. \]

\[ \Box \]

**Lemma 2** Under Conditions 2 and 3, we have that

\[ \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| = O_P(\sqrt{pn}), \quad \text{and} \quad \left\| \sum_{i=1}^{n} y_i x_i^T \right\| = O_P(\sqrt{pn}), \]

where \( \varepsilon_i = y_i - \mu_i. \)

**Proof of Lemma 2.** Let \( X = (x_1, x_2, \ldots, x_n)^T. \) We first consider \( \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\|. \) By Markov’s Inequality, for some \( \delta > 0, \)

\[ \Pr \left( \frac{1}{\sqrt{pn}} \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| > \delta \right) \leq \frac{1}{p \delta^2} E \| \varepsilon^T X \|^2 \]

\[ = \frac{1}{p \delta^2} E (\varepsilon^T X X^T \varepsilon) \]

\[ = \frac{1}{p \delta^2} \text{tr} \{ X X^T \text{Cov} (\varepsilon) \} \]

\[ = \frac{1}{p \delta^2} \sum_{i=1}^{n} \| x_i \|^2 \text{Var} (y_i) \]

\[ \leq \frac{1}{\delta^2} \max_{1 \leq i \leq n} \frac{\| x_i \|^2}{p} \frac{1}{n} \sum_{i=1}^{n} \text{Var} (y_i) \]

\[ \leq \frac{1}{\delta^2} \max_{1 \leq i \leq n} \frac{\| x_i \|^2}{p} \sqrt{\frac{\| \mu \|^2}{n}} \]

\[ \leq \sqrt{C_2 C_3^2 \delta^{-2}} \to 0 \quad \text{as} \quad \delta \to \infty, \quad (A.2) \]

where the last inequality is from Condition 2 and (12) in Condition 3. By the definition of \( O_P(1), \)

(A.2) further implies that there exists some large \( \delta > 0, \) such that

\[ \Pr \left( \frac{1}{\sqrt{pn}} \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| > \delta \right) \to 0 \quad \text{as} \quad n \to \infty. \quad (A.3) \]

Next, we consider \( \left\| \sum_{i=1}^{n} y_i x_i^T \right\|. \) Under Condition 2 and (12) in Condition 3, we have that

\[ \left\| \sum_{i=1}^{n} \mu_i x_i^T \right\| \leq \sum_{i=1}^{n} |\mu_i| \| x_i \| \]
\[ \sqrt{n} \sum_{i=1}^{n} |\mu_i|^2 \leq \sqrt{C_2} n \sqrt{n \max_{1 \leq i \leq n} \|x_i\|^2} \leq \sqrt{C_2} \sqrt{C_3} np \leq \sqrt{C_2 C_3 \sqrt{pn}}. \quad (A.4) \]

Then combining (A.3) and (A.4), we obtain
\[
\left\| \sum_{i=1}^{n} y_i x_i^T \right\| \leq \left\| \sum_{i=1}^{n} \mu_i x_i^T \right\| + \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| = O(\sqrt{pn}) + O_P(\sqrt{pn}) = O_P(\sqrt{pn}). \]

\[ \square \]

**Lemma 3** Under Conditions 1, 3 and 4, we have that
\[
\max_{1 \leq s \leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} x_{(s),i} \right\| = O_P(\sqrt{Spn}), \quad (A.5)
\]

where \( \varepsilon_{(s),i} = y_i - \exp(x_{(s),i}^T \beta_{(s)}^*) \).

**Proof of Lemma 3.** First, recall that
\[
f_{(s)}(y|\beta_{(s)}) = \prod_{i=1}^{n} \frac{\mu_{(s),i}^{y_i} e^{-\mu_{(s),i}}}{y_i!}, \quad \text{and} \quad \mu_{(s),i} = \exp(x_{(s),i}^T \beta_{(s)}),
\]
\[
i = 1, \ldots, n, \quad y_i = 0, 1, \ldots
\]

Then the KL divergence can be written as
\[
\text{KL}(\beta_{(s)}) = \mathbb{E}_y \log \frac{f(y)}{f_{(s)}(y|\beta_{(s)})} = \mathbb{E}_y \log f(y) + \sum_{i=1}^{n} \mathbb{E}_y \log(y_i!) + \sum_{i=1}^{n} \left\{ \exp(x_{(s),i}^T \beta_{(s)}) - \mu_{(s),i} x_{(s),i}^T \beta_{(s)} \right\}.
\]

Using (15) in Condition 4, we have that
\[
\frac{\partial^2 \text{KL}(\beta_{(s)}^*)}{\partial \beta_{(s)} \partial \beta_{(s)}^*} = \sum_{i=1}^{n} \exp(x_{(s),i}^T \beta_{(s)}^*) x_{(s),i}^T x_{(s),i} > 0.
\]
With the second-order derivative of $KL(\beta^*_s)$ larger than zero, $\beta^*_s$ that leads to the minimum $KL(\beta^*_s)$ satisfies the first-order condition, i.e.,

$$0 = \frac{\partial KL(\beta^*_s)}{\partial \beta^*_s} = \sum_{i=1}^{n} \{ \exp(x^T_{(s),i} \beta^*_s) - \mu_i \} x_{(s),i},$$

which, according to Condition 1, further implies that

$$E \left[ \sum_{i=1}^{n} \{ y_i - \exp(x^T_{(s),i} \beta^*_s) \} x_{(s),i} \right]$$

$$= E \left\{ \sum_{i=1}^{n} (y_i - \mu_i) x_{(s),i} \right\} + \sum_{i=1}^{n} \{ \mu_i - \exp(x^T_{(s),i} \beta^*_s) \} x_{(s),i}$$

$$= 0. \quad (A.6)$$

Next, based on Condition 2 and Markov’s Inequality, for some $\delta > 0$ when $n$ is large enough,

$$\Pr \left( \frac{1}{\sqrt{Spn}} \max_{1 \leq s \leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} x_{(s),i} \right\| > \delta \right)$$

$$\leq \sum_{s=1}^{S} \Pr \left( \frac{1}{\sqrt{Spn}} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} x_{(s),i} \right\| > \delta \right)$$

$$\leq \frac{1}{Spn\delta^2} \sum_{s=1}^{S} E \left\| \varepsilon_{(s)}^T X_{(s)} \right\|^2$$

$$= \frac{1}{Spn\delta^2} \sum_{s=1}^{S} E (\varepsilon_{(s)}^T X_{(s)} X_{(s)}^T \varepsilon_{(s)})$$

$$= \frac{1}{Spn\delta^2} \sum_{s=1}^{S} \text{tr} \left\{ X_{(s)} X_{(s)}^T \text{Cov}(\varepsilon_{(s)}) \right\}$$

$$= \frac{1}{Spn\delta^2} \sum_{s=1}^{S} \sum_{i=1}^{n} \left\| \Pi_s^T x_i \right\|^2 \text{Var}(\varepsilon_{(s),i})$$

$$\leq \frac{1}{Spn\delta^2} \sum_{s=1}^{S} \sum_{i=1}^{n} \left\| x_i \right\|^2 \text{Var}(y_i)$$

$$\leq \frac{1}{\delta^2} \max_{1 \leq i \leq n} \frac{\left\| x_i \right\|^2}{p} \frac{1}{n} \sum_{i=1}^{n} \text{Var}(y_i)$$

$$\leq \sqrt{C_2 C_3 \delta^{-2}} \to 0 \quad \text{as} \quad \delta \to \infty, \quad (A.7)$$

where $\Pi_s$ is a selection matrix picking the covariates included in the $s^{th}$ model, i.e., $X\Pi_s = X_{(s)}$ and $x^T_{(s)} \Pi_s = x^T_{(s),i}$. The fourth step in (A.7) is based on (A.6), and the last inequality is from Condition 2.
and (12) in Condition 3. From (A.7) we can obtain the desired result (A.5), and this completes the proof.

□

**Lemma 4** Under Conditions 1, 3 and 4,

\[
\max_{1 \leq s \leq S} \left\| \hat{\beta}_{(s)} - \beta^*_{(s)} \right\| = O_p(S^{1/2}p^{1/2}n^{-1/2}),
\]

(A.8)

where \( \hat{\beta}_{(s)} \) is the ML estimator of the \( s^{th} \) candidate model.

**Proof of Lemma 4.** The log-likelihood function of the \( s^{th} \) model can be written as

\[
l_n(\beta_{(s)}) = \sum_{i=1}^{n} \left\{ y_i x_{(s),i}^T \beta_{(s)} - \exp(x_{(s),i}^T \beta_{(s)}) - \log(y_i!) \right\},
\]

(A.9)

and the ML estimator \( \hat{\beta}_{(s)} \) satisfies

\[
\frac{\partial l_n(\hat{\beta}_{(s)})}{\partial \beta_{(s)}} = \sum_{i=1}^{n} \left\{ y_i - \exp(x_{(s),i}^T \hat{\beta}_{(s)}) \right\} x_{(s),i} = 0.
\]

(A.10)

Let \( \mathcal{A}_n(\beta^*_{(s)} | \delta) = \{ \gamma \in \mathbb{R}^p : \sqrt{n} \| \gamma - \beta^*_{(s)} \| / \sqrt{Sp} \leq \delta \} \) and \( \partial \mathcal{A}_n(\beta^*_{(s)} | \delta) \) be the boundary of \( \mathcal{A}_n(\beta^*_{(s)} | \delta) \). By the second-order Taylor expansion of (A.10) at \( \beta^*_{(s)} \), there exists some \( \delta > 0 \), such that when \( n \) is large and \( \gamma_s \in \partial \mathcal{A}_n(\beta^*_{(s)} | \delta) \),

\[
\begin{align*}
& \max_{1 \leq s \leq S} \left\{ l_n(\gamma_s) - l_n(\beta^*_{(s)}) \right\} \\
& = \max_{1 \leq s \leq S} \left\{ \sum_{i=1}^{n} \varepsilon_{(s),i} x_{(s),i}^T (\gamma_s - \beta^*_{(s)}) - \frac{1}{2} (\gamma_s - \beta^*_{(s)})^T \sum_{i=1}^{n} \exp(x_{(s),i}^T \hat{\beta}_{(s)}) x_{(s),i} \gamma_s - \beta^*_{(s)} \right\} \\
& \leq Sp \left[ \frac{1}{\sqrt{Spn}} \max_{1 \leq s \leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} x_{(s),i} \right\| \sqrt{\frac{n}{Sp}} \| \gamma_s - \beta^*_{(s)} \| \right. \\
& \left. - \frac{1}{2} \frac{n}{Sp} \| \gamma_s - \beta^*_{(s)} \| \min_{1 \leq s \leq S} \lambda_{\min} \left\{ I_{(s)}(\hat{\beta}_{(s)}) \right\} \right] \\
& \leq Sp \left\{ \frac{1}{\sqrt{Spn}} \max_{1 \leq s \leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} x_{(s),i} \right\| \delta \| \nu_s \| - \frac{1}{2} \min \min_{1 \leq s \leq S} \delta^2 \| \nu_s \|^2 \right\} \\
& = Sp \left\{ \delta O_P(1) - \frac{1}{2} \min \min_{1 \leq s \leq S} \delta^2 \right\}
\end{align*}
\]

(A.11)
where \( \nu_s = \sqrt{n} \left( \gamma - \beta^*_s \right) / \delta \sqrt{sp} \) and \( \tilde{\beta}_s \) lies between \( \gamma_s \) and \( \beta^*_s \). The last second inequality in (A.11) is due to (15) in Condition 4, and the last inequality holds because

\[
\max_{1 \leq s \leq S} \left\| \frac{1}{\sqrt{spn}} \sum_{i=1}^{n} \varepsilon(s,i)x(s,i) \right\| = O_p(1),
\]

according to Lemma 3. Furthermore, (A.11) implies that there exists some \( \delta > 0 \), such that

\[
\text{Pr} \left[ \max_{1 \leq s \leq S} \left\{ l_n(\gamma_s) - l_n(\beta^*_s) \right\} \leq 0 \text{ for all } \gamma_s \in \partial \mathcal{A}_n(\beta^*_s | \delta) \right] \to 1 \text{ as } n \to \infty. \tag{A.12}
\]

Finally, because of the nonnegativity of exponential function \( \exp(\cdot) \) and (15) in Condition 4, we have that

\[
\frac{\partial^2 l_n(\beta^*_s)}{\partial \beta^T(s) \partial \beta^T(s)} = -\sum_{i=1}^{n} \exp(x^T(s,i) \beta(s))x(s,i)x^T(s,i) < 0.
\]

Thus the log-likelihood function is concave, and we can obtained the desired result (A.8).

\[\square\]

**Lemma 5** Under Conditions 1, 3 and 4, we have that

\[
\max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \beta(s)^{(y_{i-1})} - \tilde{\beta}_s \right\| = O_p(p^{1/2}n^{-1}), \tag{A.13}
\]

and

\[
\max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \tilde{\beta}(s)^{(y_{i-1})} - \beta^*_s \right\| = O_p(S^{1/2}p^{1/2}n^{-1/2}). \tag{A.14}
\]

**Proof of Lemma 5.** First, denote \( l_n^{(y_{i-1})}(\beta(s)) \) as the log-likelihood function of the \( s^{th} \) model but using \( y_i - 1 \) instead of \( y_i \), then we have that

\[
l_n^{(y_{i-1})}(\beta(s)) = \sum_{j=1}^{n} \left\{ y_j x^T(s,j) \beta(s) - \exp(x^T(s,j) \beta(s)) - \log(y_j) \right\} - x^T(s,i) \beta(s) + \log(y_i).
\]

Let \( \mathcal{B}_n(\tilde{\beta}(s) | \delta) = \{ \gamma \in \mathbb{R}^p : n\|\gamma - \tilde{\beta}_s\| / \sqrt{p} \leq \delta \} \) and \( \partial \mathcal{B}_n(\tilde{\beta}(s) | \delta) \) be the boundary of \( \mathcal{B}_n(\tilde{\beta}(s) | \delta) \). There exists some \( \delta > 0 \), such that when \( n \) is large enough and \( \gamma_s \in \partial \mathcal{B}_n(\tilde{\beta}(s) | \delta) \), Taylor expansion gives that

\[
\max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\{ l_n^{(y_{i-1})}(\gamma_s) - l_n^{(y_{i-1})}(\tilde{\beta}_s) \right\} = \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\{ \frac{\partial l_n^{(y_{i-1})}(\beta(s))}{\partial \beta_s} (\gamma_s - \tilde{\beta}_s) + \frac{1}{2} \left( \gamma_s - \tilde{\beta}_s \right)^T \frac{\partial^2 l_n^{(y_{i-1})}(\gamma_s)}{\partial \beta_s \partial \beta^T_s} (\gamma_s - \tilde{\beta}_s) \right\}
\]
\[ \begin{align*}
&= \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\{ -x_{(s),i}^T \left( \gamma_s - \hat{\beta}_{(s)} \right) - \frac{1}{2} \left( \gamma_s - \hat{\beta}_{(s)} \right)^T \sum_{j=1}^n \exp \left( x_{(s),j}^T \tilde{\gamma}_s \right) x_{(s),j} x_{(s),j}^T \left( \gamma_s - \hat{\beta}_{(s)} \right) \right\} \\
&= \frac{p}{n} \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\{ -x_{(s),i}^T \frac{n}{\sqrt{p}} \left( \gamma_s - \hat{\beta}_{(s)} \right) - \frac{1}{2} \frac{n}{\sqrt{p}} \left( \gamma_s - \hat{\beta}_{(s)} \right)^T I_s(\gamma_s) \frac{n}{\sqrt{p}} \left( \gamma_s - \hat{\beta}_{(s)} \right) \right\} \\
&\leq \frac{p}{n} \left\{ -\frac{1}{2} C_{\min} \delta^2 \| \nu_s \|^2 + \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \frac{\| x_{(s),i} \|}{\sqrt{p}} \delta \| \nu_s \| \right\} \\
&\leq \frac{p}{n} \left\{ -\frac{1}{2} \delta^2 C_{\min} + \delta C_3 \right\} \\
&< 0,
\end{align*} \]

where \( \nu_s = n(\gamma_s - \hat{\beta}_{(s)})/\delta \sqrt{p} \) and \( \tilde{\gamma}_s \) lies between \( \gamma_s \) and \( \hat{\beta}_{(s)} \). The last third inequality is due to (15) of Condition 4, and the last second inequality is based on (12) in Condition 3. This implies that there exists some \( \delta > 0 \) satisfying

\[ \Pr \left[ \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\{ I_n^{(y_{n-1})}(\gamma_s) - I_n^{(y_{n-1})}(\hat{\beta}_{(s)}) \right\} \leq 0 \right. \text{ for all } \gamma_s \in \partial B_n(\hat{\beta}_{(s)}|\delta) \left. \right] \]

\[ = \Pr \left[ \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \frac{n}{\sqrt{p}} \left\| \hat{\beta}_{(s)}^{(y_{n-1})} - \beta_{(s)}^{(y_{n-1})} \right\| \leq \delta \right] \rightarrow 1 \text{ as } n \rightarrow \infty, \tag{A.15} \]

where the equality holds because \( I_n^{(y_{n-1})}(\gamma_s) \) is concave and reaches its maximum value at \( \hat{\beta}_{(s)}^{(y_{n-1})} \). (A.15) further implies that

\[ \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \hat{\beta}_{(s)}^{(y_{n-1})} - \beta_{(s)}^{(y_{n-1})} \right\| = O_p(p^{1/2}n^{-1}). \tag{A.16} \]

Combining (A.8) and (A.16), we obtain that

\[ \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \hat{\beta}_{(s)}^{(y_{n-1})} - \beta_{(s)}^{(y_{n-1})} \right\| \leq \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \hat{\beta}_{(s)}^{(y_{n-1})} - \beta_{(s)}^{(y_{n-1})} \right\| + \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \beta_{(s)} - \beta_{(s)}^{(y_{n-1})} \right\| \]

\[ = O_p(p^{1/2}n^{-1}) + O_p(S^{1/2}p^{1/2}n^{-1/2}) \]

\[ = O_p(S^{1/2}p^{1/2}n^{-1/2}). \]

\[ \square \]

### A.2 Proof of Theorem 1

From the proof of Theorem 1' of Wan et al. (2010), to prove (18), it is sufficient to verify that

\[ \sup_{w \in W_n} \frac{|\text{KL}(w) - \text{KL}^*(w)|}{\text{KL}^*(w)} = o_p(1) \tag{A.17} \]
where

\[ e \]

and to (13) in Condition 3. Combining (A.19) with (16) in Condition 4, we have that

\[
\sup_{w \in \mathcal{W}_n} \frac{|\text{KL}(w) - C(w)|}{\text{KL}^*(w)} = o_p(1). \quad (A.18)
\]

First, it is straightforward to show that, for every \( \beta_{(s)} \in \mathcal{R}^p \),

\[
\max_{1 \leq s \leq S} \lambda_{\max}\{I_{(s)}(\beta_{(s)})\} = \max_{1 \leq s \leq S} \frac{1}{n} \sum_{i=1}^{n} \exp \left( x_{(s),i}^T \beta_{(s)} \right) x_{(s),i} x_{(s),i}^T
\]

\[
\geq \max_{1 \leq s \leq S} \frac{1}{p_n} \sum_{i=1}^{n} \exp \left( x_{(s),i}^T \beta_{(s)} \right) e_j^T x_{(s),i} x_{(s),i} e_j
\]

\[
\geq \max_{1 \leq s \leq S} \frac{1}{p_n} \sum_{i=1}^{n} \exp \left( x_{(s),i}^T \beta_{(s)} \right) x_{(s),i} x_{(s),i}^T
\]

\[
\geq \max_{1 \leq s \leq S} \frac{1}{p_n} \sum_{i=1}^{n} \exp \left( x_{(s),i}^T \beta_{(s)} \right) x_{(s),i} C_4 \sqrt{p_n}
\]

\[
= \frac{C_4}{\sqrt{p_n}} \sum_{i=1}^{n} \exp \left( x_{(s),i}^T \beta_{(s)} \right) x_{(s),i}, \quad (A.19)
\]

where \( e_j \) is the column vector whose \( j \)th element is 1 and others are 0, and the last inequality is due to (13) in Condition 3. Combining (A.19) with (16) in Condition 4, we have that

\[
\max_{1 \leq s \leq S} \sup_{\beta_{(s)} \in \partial(\beta_{(s)}^*)} \frac{1}{p_n} \sum_{i=1}^{n} \exp \left( x_{(s),i}^T \beta_{(s)} \right) x_{(s),i} \leq \frac{C_{\max}}{C_4}. \quad (A.20)
\]

Next, by the definition of KL divergence in (5) and the differential mean value theorem, we can obtain that, when \( n \) is large enough,

\[
\sup_{w \in \mathcal{W}_n} |\text{KL}(w) - \text{KL}^*(w)|
\]

\[
= \sup_{w \in \mathcal{W}_n} \left| \sum_{i=1}^{n} \left[ \tilde{\mu}_i(w, y) - \mu_i \log \{\tilde{\mu}_i(w, y)\} \right] - \sum_{i=1}^{n} \left[ \mu_i^*(w) - \mu_i \log \{\mu_i^*(w)\} \right] \right|
\]

\[
\leq \sup_{w \in \mathcal{W}_n} \sum_{i=1}^{n} \left| \tilde{\mu}_i(w, y) - \mu_i^*(w) \right| + \sup_{w \in \mathcal{W}_n} \sum_{i=1}^{n} \left| \mu_i \log \{\tilde{\mu}_i(w, y)\} - \log \{\mu_i^*(w)\} \right|
\]

\[
\leq \sup_{w \in \mathcal{W}_n} \sum_{i=1}^{n} \left| \sum_{s=1}^{S} w_s x_{(s),i}^T \beta_{(s)} \right| \sum_{s=1}^{S} w_s x_{(s),i}^T \Pi_s \left( \beta_{(s)} - \beta_{(s)}^* \right)
\]
according to Condition 5. The last second equality is due to Lemmas 2, 4 and (A.20).

Finally, using Conditions 2–4 and the triangle inequality, we have that

\[
\begin{align*}
\sup_{w \in W_n} |\text{KL}(w) - C(w)| &
\leq \sup_{w \in W_n} \left| \sum_{i=1}^{n} y_i \log \{ \tilde{\mu}_i(w, y^{(y_i-1)}) \} - \mu_i \log \{ \tilde{\mu}_i(w, y) \} \right| \\
&+ \sup_{w \in W_n} \left| \sum_{i=1}^{n} y_i \left[ \log \{ \tilde{\mu}_i(w, y^{(y_i-1)}) \} - \log \{ \mu_i^*(w) \} \right] \right| \\
&+ \sup_{w \in W_n} \left| \sum_{i=1}^{n} \mu_i \left[ \log \{ \tilde{\mu}_i(w, y) \} - \log \{ \mu_i^*(w) \} \right] \right| \\
&\leq \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| \sup_{w \in W_n} \left\| \sum_{s=1}^{S} w_s \Pi_s \beta^*_s \right\| \\
&+ \sup_{w \in W_n} \left\| \sum_{i=1}^{n} y_i x_i^T \sum_{s=1}^{S} w_s \left( \Pi_s \tilde{\beta}^{(y_i-1)}_s - \Pi_s \beta^*_s \right) \right\| \\
&+ \sup_{w \in W_n} \left\| \sum_{s=1}^{S} w_s \left( \Pi_s \hat{\beta}_s - \Pi_s \beta^*_s \right) \right\| \\
&\leq \left\| \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| \max_{1 \leq s \leq S} \| \beta^*_s \|
\end{align*}
\]

where $\tilde{\beta}_s$ lies between $\hat{\beta}_s$ and $\beta^*_s$, and thus falls into the neighborhood $O(\beta^*_s, \delta)$ if $n$ is large, according to Condition 5. The last second equality is due to Lemmas 2, 4 and (A.20).
\[\begin{aligned}
&+ \sum_{i=1}^{n} y_i x_i^T \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \hat{\beta}^{(y_i-1)} - \beta_{(s)}^* \right\| \\
&+ \sum_{i=1}^{n} \mu_i \| x_i \| \max_{1 \leq s \leq S} \left\| \hat{\beta}_{(s)} - \beta_{(s)}^* \right\| \\
&= O_p(pn^{1/2}) + \left\{ O(p^{1/2}n) + O_p(p^{1/2}n^{1/2}) \right\} O_p(S^{1/2}p^{1/2}n^{-1/2}) \\
&\quad + O(p^{1/2}n)O_p(S^{1/2}p^{1/2}n^{-1/2}) \\
&= O_p(S^{1/2}pn^{1/2}),
\end{aligned}\]  

(A.22)

where \( \mu_i(x) = \exp(\sum_{s=1}^{S} w_s x_{(s),i}^T \beta_{(s)}^*) \), and the last second equality is due to Lemmas 2, 4 and 5 and (14) in Condition 3.

Assuming that Condition 5 is satisfied, the above two results (A.21) and (A.22) imply (A.17) and (A.18), respectively. This completes the proof.

### A.3 Proof of Theorem 2

Assume that the \( s_0^{th} \) model is a correct candidate model, we know that \( \| \hat{\beta}_s - \beta_{\text{true}} \| = O_p(p^{1/2}n^{-1/2}) \).

Let \( w_{s_0}^o \) denotes the weight vector whose \( s_0^{th} \) element is 1 and others are 0. We note that

\[\begin{aligned}
\mathcal{C}(w_{s_0}^o) &= \log f(y) + \sum_{i=1}^{n} \left\{ \exp(x_i^T \hat{\beta}_{s_0}) + \log(y_i!) - y_i x_i^T \hat{\beta}_{(y_i-1)}^* \right\} \\
&= \log f(y) + \sum_{i=1}^{n} \left\{ \exp(x_i^T \beta_{\text{true}}) + \log(y_i!) - y_i x_i^T \beta_{\text{true}} \right\} \\
&\quad + \sum_{i=1}^{n} \left\{ \exp(x_i^T \beta_{\text{true}}) x_i^T (\hat{\beta}_{s_0} - \beta_{\text{true}}) + \frac{1}{2} (\hat{\beta}_{s_0} - \beta_{\text{true}})^T \exp(x_i^T \beta) x_i x_i^T (\hat{\beta}_{s_0} - \beta_{\text{true}}) \\
&\quad - y_i x_i^T (\hat{\beta}_{s_0}^{(y_i-1)} - \beta_{\text{true}}) \right\} \\
&= \log f(y) + \sum_{i=1}^{n} \left\{ \exp(x_i^T \beta_{\text{true}}) + \log(y_i!) - y_i x_i^T \beta_{\text{true}} \right\} \\
&\quad + \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} \varepsilon_i x_i^T \sqrt{\frac{n}{p}} (\hat{\beta}_{s_0} - \beta_{\text{true}}) \\
&\quad + \frac{1}{2} \frac{n}{p} (\hat{\beta}_{s_0} - \beta_{\text{true}})^T I_n(\tilde{\beta}) \frac{n}{p} (\hat{\beta}_{s_0} - \beta_{\text{true}}) \\
&\quad - \sum_{i=1}^{n} \left( \frac{1}{\sqrt{pn}} y_i x_i^T \right) \frac{n}{p} (\hat{\beta}_{s_0}^{(y_i-1)} - \hat{\beta}_{s_0}) \right\},
\end{aligned}\]  

(A.23)

where \( I_n(\tilde{\beta}) = n^{-1} \sum_{i=1}^{n} \exp(x_i^T \tilde{\beta}) x_i x_i^T \), \( \varepsilon_i = y_i - \exp(x_i^T \beta_{\text{true}}) \), and \( \tilde{\beta} \) lies between \( \hat{\beta}_{s_0} \) and \( \beta_{\text{true}} \).
For notation convenience, we define

$$\eta_n = p \left[ - \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} \varepsilon_i x_i^T \sqrt{\frac{n}{p}} (\hat{\beta}_{s_0} - \beta_{true}) \\
+ \frac{1}{2} \sqrt{\frac{n}{p}} (\hat{\beta}_{s_0} - \beta_{true})^T I_n(\tilde{\beta}) \sqrt{\frac{n}{p}} (\hat{\beta}_{s_0} - \beta_{true}) \\
- \sum_{i=1}^{n} \left( \frac{1}{\sqrt{pn}} y_i x_i^T \right) \frac{n}{\sqrt{p}} (\hat{\beta}_{s_0}^{(y_i-1)} - \hat{\beta}_{s_0}) \right]$$

and under Conditions 2–4, we have that

$$|\eta_n| \leq \sqrt{\frac{n}{p}} \left\| \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| \| \hat{\beta}_{s_0} - \beta_{true} \| + \frac{n}{p} \left\| I_n(\tilde{\beta}) \right\| \| \hat{\beta}_{s_0} - \beta_{true} \|^2 \\
+ \frac{n}{\sqrt{pn}} \sqrt{\frac{n}{p}} \sum_{i=1}^{n} y_i x_i^T \right\| \max_{1 \leq i \leq n} \| \hat{\beta}_{s_0}^{(y_i-1)} - \hat{\beta}_{s_0} \| \\
= O_p(1) + C_{\max} O_p(1) + O_P(1) \\
= O_P(1), \quad (A.24)$$

where the last second equality is implied by (16) in Condition 4, Lemmas 2 and 5, and the fact that $\| \hat{\beta}_{s_0} - \beta_{true} \| = O_P(p^{1/2} n^{-1/2})$.

Note that from Lemma 2 and Conditions 3 and 4, when $n$ is large enough,

$$\max_{1 \leq i \leq n} \left\| \hat{\beta}(\hat{w})^{(y_i-1)} - \hat{\beta}(\tilde{w}) \right\| = \max_{1 \leq i \leq n} \left\| \sum_{s=1}^{S} \hat{w}_s \Pi_s \left( \hat{\beta}_{(s)}^{(y_i-1)} - \hat{\beta}_{(s)} \right) \right\| \\
\leq \max_{1 \leq i \leq n} \max_{1 \leq s \leq S} \left\| \hat{\beta}_{(s)}^{(y_i-1)} - \hat{\beta}_{(s)} \right\| \\
= O_p \left( p^{1/2} n^{-1} \right). \quad (A.25)$$

Let $\hat{\beta}(\tilde{w}) = \sum_{s=1}^{S} \hat{w}_s \Pi_s \hat{\beta}_{(s)}$, where $\tilde{w}$ minimizes $C(w)$. Combining with (A.23) and using Taylor expansion, when $n$ is large enough we have that

$$C(\hat{w}) - C(w_{s_0}^*) = p \left[ - \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} \varepsilon_i x_i^T \sqrt{\frac{n}{p}} \left( \hat{\beta}(\tilde{w}) - \beta_{true} \right) \\
+ \sqrt{\frac{n}{p}} \left( \hat{\beta}(\tilde{w}) - \beta_{true} \right)^T I_n(\tilde{\beta}) \sqrt{\frac{n}{p}} \left( \hat{\beta}(\tilde{w}) - \beta_{true} \right) \\
- \sum_{i=1}^{n} \left( \frac{1}{\sqrt{pn}} y_i x_i^T \right) \frac{n}{\sqrt{p}} \left( \hat{\beta}(\tilde{w})^{(y_i-1)} - \hat{\beta}(\tilde{w}) \right) \right] - p\eta_n$$
\[
\begin{align*}
&= p \left[ - \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} \varepsilon_i x_i^T \nu + \nu^T I_n(\tilde{\beta}) \nu \\
&\quad - \sum_{i=1}^{n} \left( \frac{1}{\sqrt{pn}} y_i x_i^T \right) \frac{n}{\sqrt{p}} \left\{ \tilde{\beta}(\hat{w})(y_{i-1}) - \tilde{\beta}(\hat{w}) \right\} \right] - pn_n \\
&\geq p \left\{ C_{\min} \| \nu \|^2 - \left\| \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} \varepsilon_i x_i^T \right\| \| \nu \| \\
&\quad - \left\| \frac{1}{\sqrt{pn}} \sum_{i=1}^{n} y_i x_i^T \right\| \frac{n}{\sqrt{p}} \max_{1 \leq i \leq n} \left\| \tilde{\beta}(\hat{w})(y_{i-1}) - \tilde{\beta}(\hat{w}) \right\| - \left| \eta_n \right| \right\} \\
&= p \left\{ C_{\min} \| \nu \|^2 - \| \nu \| O_P(1) - O_P(1) - O_P(1) \right\}, \\
\end{align*}
\]
where \( \nu = \sqrt{n} \{ \tilde{\beta}(\hat{w}) - \beta_{\text{true}} \} / \sqrt{p} \), and \( \tilde{\beta} \) lies between \( \hat{\beta}(\hat{w}) \) and \( \beta_{\text{true}} \). The last inequality is due to (15) in Condition 4, and the last equality is based on (A.24), (A.25) and Lemma 2.

From (A.26), we know that there exists some large \( \delta > 0 \), such that
\[
\Pr \{ C(\hat{w}) - C(w^o) > 0 \| \nu \| > \delta \} \to 1 \quad \text{as} \quad n \to \infty.
\]
However, in fact \( \Pr \{ C(\hat{w}) - C(w^o) > 0 \} = 0 \), because \( \hat{w} \) minimizes \( C(w) \). This implies that
\[
0 = \Pr \{ C(\hat{w}) - C(w^o) > 0 \} \\
= \Pr \{ C(\hat{w}) - C(w^o) > 0 \| \nu \| > \delta \} \Pr (\| \nu \| > \delta) \\
+ \Pr \{ C(\hat{w}) - C(w^o) > 0 \| \nu \| \leq \delta \} \Pr (\| \nu \| \leq \delta) \\
\to \Pr (\| \nu \| > \delta) + \Pr \{ C(\hat{w}) - C(w^o) > 0 \| \nu \| \leq \delta \} \Pr (\| \nu \| \leq \delta) \\
\quad \text{as} \quad n \to \infty,
\]
which further implies that \( \Pr (\| \nu \| > \delta) \to 0 \) since the probabilities cannot be negative. Therefore, we can conclude that
\[
\| \hat{\beta}(\hat{w}) - \beta_{\text{true}} \| = O_P(p^{1/2}n^{-1/2}).
\]
This completes the proof.