Supplementary Material for "Optimal Subsampling for Large Sample Logistic Regression"

HaiYing Wang, Rong Zhu, and Ping Ma

February 3, 2017

S.1 Proofs

In this section we prove the theorems in the paper.

S.1.1 Proof of Theorem 1

We begin by establishing a lemma that will be used in the proof of Theorems 1 and 2.

Lemma 1. *If Assumptions 1 and 2 hold, then conditionally on $\mathcal{F}_n$ in probability,*

$$\tilde{\mathbf{M}}_X - \mathbf{M}_X = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S.1}$$

$$\frac{1}{n}\frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S.2}$$

where

$$\tilde{\mathbf{M}}_X = \frac{1}{n}\frac{\partial^2 \ell^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{1}{nr}\sum_{i=1}^{r}\frac{w_i^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})\mathbf{x}_i^*(\mathbf{x}_i^*)^T}{\pi_i^*}.$$

Proof. Direct calculation yields

$$\text{E}(\tilde{\mathbf{M}}_X|\mathcal{F}_n) = \mathbf{M}_X. \tag{S.3}$$

For any component $\tilde{\mathbf{M}}_X^{j_1j_2}$ of $\tilde{\mathbf{M}}_X$ where $1 \le j_1, j_2 \le d$,

$$\text{Var}\left(\frac{1}{n}\tilde{\mathbf{M}}_X^{j_1j_2}\Big|\mathcal{F}_n\right) = \frac{1}{r}\sum_{i=1}^{n}\pi_i\left\{\frac{w_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})x_{ij_1}x_{ij_2}}{n\pi_i} - \mathbf{M}_X^{j_1j_2}\right\}^2$$

$$
\begin{aligned}
=&\frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})^2(x_{ij_1}x_{ij_2}^T)^2}{\pi_i}-\frac{1}{r}(\mathbf{M}_X^{j_1j_2})^2\\
\leq&\frac{1}{16rn^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^4}{\pi_i}-\frac{1}{r}(\mathbf{M}_X^{j_1j_2})^2\\
=&O_P(r^{-1}),
\end{aligned}
$$

where the second last inequality holds by the fact that $0 < w_i(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \leq 1/4$ and the last equality is from Assumption 2. Using Markov's inequality, this result and (S.3), implies (S.1).

To prove (S.2), direct calculation yields,

$$
\text{E}\left\{\frac{1}{n}\frac{\partial\ell^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial\boldsymbol{\beta}}\middle|\mathcal{F}_n\right\}=\frac{1}{nr}\frac{\partial\ell^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial\boldsymbol{\beta}}=0. \tag{S.4}
$$

From Assumption 2,

$$
\text{Var}\left\{\frac{1}{n}\frac{\partial\ell^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial\boldsymbol{\beta}}\middle|\mathcal{F}_n\right\}=\frac{1}{n^2r}\sum_{i=1}^{n}\frac{\{y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\mathbf{x}_i\mathbf{x}_i^T}{\pi_i}\leq\frac{1}{n^2r}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\pi_i}=O_P(r^{-1}). \tag{S.5}
$$

From (S.4), (S.5) and Markov's inequality, (S.2) follows. □

Now we prove Theorem 1. Note that $t_i(\boldsymbol{\beta}) = y_i \log p_i(\boldsymbol{\beta}) + (1-y_i)\log\{1-p_i(\boldsymbol{\beta})\}$, $t_i^*(\boldsymbol{\beta}) = y_i^* \log p_i^*(\boldsymbol{\beta}) + (1-y_i^*)\log\{1-p_i^*(\boldsymbol{\beta})\}$,

$$
\ell^*(\boldsymbol{\beta})=\frac{1}{r}\sum_{i=1}^{r}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*},\quad\text{and}\quad\ell(\boldsymbol{\beta})=\sum_{i=1}^{n}t_i(\boldsymbol{\beta}).
$$

By direct calculation under the conditional distribution of subsample given $\mathcal{F}_n$,

$$
\text{E}\left\{\frac{\ell^*(\boldsymbol{\beta})}{n}-\frac{\ell(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2=\frac{1}{r}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{t_i^2(\boldsymbol{\beta})}{\pi_i}-\left(\frac{1}{n}\sum_{i=1}^{n}t_i(\boldsymbol{\beta})\right)^2\right]. \tag{S.6}
$$

Note that $|t_i(\boldsymbol{\beta})| \leq \log 4 + 2\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|$. Therefore, from Assumption 1,

$$
\frac{1}{n^2}\sum_{i=1}^{n}\frac{t_i^2(\boldsymbol{\beta})}{\pi_i}-\left(\frac{1}{n}\sum_{i=1}^{n}t_i(\boldsymbol{\beta})\right)^2\leq\frac{1}{n^2}\sum_{i=1}^{n}\frac{t_i^2(\boldsymbol{\beta})}{\pi_i}+\left(\frac{1}{n}\sum_{i=1}^{n}|t_i(\boldsymbol{\beta})|\right)^2=O_P(1). \tag{S.7}
$$

Therefore combing (S.6) and (S.7), $n^{-1}\ell^*(\boldsymbol{\beta})-n^{-1}\ell(\boldsymbol{\beta})\to 0$ in conditional probability given $\mathcal{F}_n$. Note that the parameter space is compact and $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is the unique global maximum of

the continuous convex function $\ell(\boldsymbol{\beta})$. Thus, from Theorem 5.9 and its remark of van der Vaart (1998), conditionally on $\mathcal{F}_n$ in probability,

$$\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\| = o_{P|\mathcal{F}_n}(1) \tag{S.8}$$

The consistency proved above ensures that $\tilde{\boldsymbol{\beta}}$ is close to $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ as long as r is not small. Using Taylor's theorem (c.f. Chapter 4 of Ferguson 1996),

$$0 = \frac{\dot{\ell}_j^*(\tilde{\boldsymbol{\beta}})}{n} = \frac{\dot{\ell}_j^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + \frac{1}{n}\frac{\partial \dot{\ell}_j^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}^T}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + \frac{1}{n}R_j \tag{S.9}$$

where $\dot{\ell}_j^*(\boldsymbol{\beta})$ is the partial derivative of $\ell^*(\boldsymbol{\beta})$ with respect to β_j, and

$$R_j = (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})^T \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_j^*\{\hat{\boldsymbol{\beta}}_{\text{MLE}} + uv(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v \mathrm{d}u\mathrm{d}v \; (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}).$$

Note that

$$\left\| \frac{\partial^2 \dot{\ell}_j^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\| = \frac{1}{r} \left\| \sum_{i=1}^r \frac{p_i^*(\boldsymbol{\beta})\{1 - p_i^*(\boldsymbol{\beta})\}\{1 - 2p_i^*(\boldsymbol{\beta})\}}{\pi_i^*} x_{ij}^* \mathbf{x}_i^* \mathbf{x}_i^{*T} \right\| \leq \frac{1}{r} \sum_{i=1}^r \frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*}$$

for all $\boldsymbol{\beta}$. Thus

$$\left\| \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_j^*\{\hat{\boldsymbol{\beta}}_{\text{MLE}} + uv(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v \mathrm{d}u\mathrm{d}v \; \right\| \leq \frac{1}{2r} \sum_{i=1}^r \frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*} = O_{P|\mathcal{F}_n}(n), \tag{S.10}$$

where the last equality is from the fact that

$$P\left(\frac{1}{nr} \sum_{i=1}^r \frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*} \geq \tau \middle| \mathcal{F}_n \right) \leq \frac{1}{nr\tau} \sum_{i=1}^r \mathrm{E}\left(\frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*} \middle| \mathcal{F}_n \right) = \frac{1}{n\tau} \sum_{i=1}^n \|\mathbf{x}_i\|^3 \to 0, \tag{S.11}$$

in probability as $\tau \to \infty$ by Assumption 2. From (S.9) and (S.10),

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = -\tilde{\mathbf{M}}_X^{-1} \left\{ \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + O_{P|\mathcal{F}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|^2) \right\}. \tag{S.12}$$

From (S.1) of Lemma 1, $\tilde{\mathbf{M}}_X^{-1} = O_{P|\mathcal{F}_n}(1)$. Combining this with (S.2), (S.8) and (S.12)

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2}) + o_{P|\mathcal{F}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|),$$

which implies that

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S.13}$$

S.1.2 Proof of Theorem 2

Note that

$$\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} = \frac{1}{r}\sum_{i=1}^{r}\frac{\{y_i^* - p_i^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}\mathbf{x}_i^*}{n\pi_i^*} \equiv \frac{1}{r}\sum_{i=1}^{r}\boldsymbol{\eta}_i \tag{S.14}$$

Given $\mathcal{F}_n$, $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_r$ are i.i.d, with mean $\mathbf{0}$ and variance,

$$\text{Var}(\boldsymbol{\eta}_i|\mathcal{F}_n) = r\mathbf{V}_c = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\{y_i - p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\mathbf{x}_i\mathbf{x}_i^T}{\pi_i} = O_P(1). \tag{S.15}$$

Meanwhile, for every $\varepsilon > 0$ and some $\delta > 0$,

$$\begin{aligned}
&\sum_{i=1}^{r}\text{E}\{\|r^{-1/2}\boldsymbol{\eta}_i\|^2 I(\|\boldsymbol{\eta}_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\} \\
&\leq \frac{1}{r^{1+\delta/2}\varepsilon^\delta}\sum_{i=1}^{r}\text{E}\{\|\boldsymbol{\eta}_i\|^{2+\delta} I(\|\boldsymbol{\eta}_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\} \\
&\leq \frac{1}{r^{1+\delta/2}\varepsilon^\delta}\sum_{i=1}^{r}\text{E}(\|\boldsymbol{\eta}_i\|^{2+\delta}|\mathcal{F}_n) \\
&= \frac{1}{r^{\delta/2}}\frac{1}{n^{2+\delta}}\frac{1}{\varepsilon^\delta}\sum_{i=1}^{n}\frac{\{y_i - p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^{2+\delta}\|\mathbf{x}_i\|^{2+\delta}}{\pi_i^{1+\delta}} \\
&\leq \frac{1}{r^{\delta/2}}\frac{1}{n^{2+\delta}}\frac{1}{\varepsilon^\delta}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^{2+\delta}}{\pi_i^{1+\delta}} = o_P(1)
\end{aligned}$$

where the last equality is from Assumption 3. This and (S.15) show that the Lindeberg-Feller conditions are satisfied in probability. From (S.14) and (S.15), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart 1998), conditionally on $\mathcal{F}_n$,

$$\frac{1}{n}\mathbf{V}_c^{-1/2}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = \frac{1}{r^{1/2}}\{\text{Var}(\boldsymbol{\eta}_i|\mathcal{F}_n)\}^{-1/2}\sum_{i=1}^{r}\boldsymbol{\eta}_i \to N(0, \mathbf{I}),$$

in distribution. From Lemma 1, (S.12) and (S.13),

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = -\frac{1}{n}\tilde{\mathbf{M}}_X^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1}) \tag{S.16}$$

From (S.1) of Lemma 1,

$$\tilde{\mathbf{M}}_X^{-1} - \mathbf{M}_X^{-1} = -\mathbf{M}_X^{-1}(\tilde{\mathbf{M}}_X - \mathbf{M}_X)\tilde{\mathbf{M}}_X^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S.17}$$

Based on Assumption 1 and (S.15), it is verified that,

$$\mathbf{V} = \mathbf{M}_X^{-1}\mathbf{V}_c\mathbf{M}_X^{-1} = \frac{1}{r}\mathbf{M}_X^{-1}\left(r\mathbf{V}_c\right)\mathbf{M}_X^{-1} = O_P(r^{-1}). \tag{S.18}$$

Thus, (S.16), (S.17) and (S.18) yield,

$$\begin{aligned}\mathbf{V}^{-1/2}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\text{MLE}}) &= -\mathbf{V}^{-1/2}n^{-1}\tilde{\mathbf{M}}_X^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})+O_{P|\mathcal{F}_n}(r^{-1/2})\\ &= -\mathbf{V}^{-1/2}\mathbf{M}_X^{-1}n^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})-\mathbf{V}^{-1/2}(\tilde{\mathbf{M}}_X^{-1}-\mathbf{M}_X^{-1})n^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})+O_{P|\mathcal{F}_n}(r^{-1/2})\\ &= -\mathbf{V}^{-1/2}\mathbf{M}_X^{-1}\mathbf{V}_c^{1/2}\mathbf{V}_c^{-1/2}n^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})+O_{P|\mathcal{F}_n}(r^{-1/2}).\end{aligned}$$

The result in (5) of Theorem 1 follows from Slutsky's Theorem(Theorem 6 of Ferguson 1996) and the fact that

$$\mathbf{V}^{-1/2}\mathbf{M}_X^{-1}\mathbf{V}_c^{1/2}(\mathbf{V}^{-1/2}\mathbf{M}_X^{-1}\mathbf{V}_c^{1/2})^T=\mathbf{V}^{-1/2}\mathbf{M}_X^{-1}\mathbf{V}_c^{1/2}\mathbf{V}_c^{1/2}\mathbf{M}_X^{-1}\mathbf{V}^{-1/2}=\mathbf{I}.$$

S.1.3 Proof of Theorems 3 and 4

For Theorem 3,

$$\begin{aligned}\text{tr}(\mathbf{V})=\text{tr}(\mathbf{M}_X^{-1}\mathbf{V}_c\mathbf{M}_X^{-1}) &= \frac{1}{r}\sum_{i=1}^{n}\text{tr}\left[\frac{1}{\pi_i}\{y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\mathbf{M}_X^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{M}_X^{-1}\right]\\ &= \frac{1}{r}\sum_{i=1}^{n}\left[\frac{1}{\pi_i}\{y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\|\mathbf{M}_X^{-1}\mathbf{x}_i\|^2\right]\\ &= \frac{1}{r}\sum_{i=1}^{n}\pi_i\sum_{i=1}^{n}\left[\pi_i^{-1}\{y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\|\mathbf{M}_X^{-1}\mathbf{x}_i\|^2\right]\\ &\geq \frac{1}{r}\left[\sum_{i=1}^{n}|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{M}_X^{-1}\mathbf{x}_i\|\right]^2,\end{aligned}$$

where the last step is from the Cauchy-Schwarz inequality and the equality in it holds if and only if when $\pi_i\propto|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{M}_X^{-1}\mathbf{x}_i\|$.

The proof of Theorem 4 is similar to the proof of Theorem 3 and thus is omit it to save space.

S.1.4 Proof of Theorems 5

Since $r_0r^{-1/2}\to 0$, the contribution of the first step subsample to the likelihood function is a small term with an order $o_{P|\mathcal{F}_n}(r^{-1/2})$ relative the likelihood function. Thus, we can focus on the second step subsample only. Denote

$$\ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})=\frac{1}{r}\sum_{i=1}^{r}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)},$$

where $\pi_i^*(\tilde{\boldsymbol{\beta}}_0)$ has the same expression as π_i^{mVc} except that $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is replaced by $\tilde{\boldsymbol{\beta}}_0$. We first establish two lemmas that will be used in the proof of Theorems 5 and 6.

Lemma 2. *Let the compact parameter space be Θ and $\lambda = \sup_{\boldsymbol{\beta}\in\Theta} \|\boldsymbol{\beta}\|$. Under Assumption 4, for $k_1 \geq k_2 \geq 0$,*

$$\frac{1}{n^2}\sum_{i=1}^{n} \frac{\|\mathbf{x}_i\|^{k_1}}{\pi_i^{k_2}(\tilde{\boldsymbol{\beta}}_0)} \leq \frac{3^{k_2}}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^{k_1-k_2} e^{\lambda k_2 \|\mathbf{x}_i\|} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^{k_2} = O_P(1). \tag{S.19}$$

Proof. From the expression of $\pi_i(\tilde{\boldsymbol{\beta}}_0)$,

$$\frac{1}{n^2}\sum_{i=1}^{n} \frac{\|\mathbf{x}_i\|^{k_1}}{\pi_i^{k_2}(\tilde{\boldsymbol{\beta}}_0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\|\mathbf{x}_i\|^{k_1-k_2}}{|y_i - p_i(\tilde{\boldsymbol{\beta}}_0)|^{k_2}} \frac{1}{n}\sum_{j=1}^{n} |y_j - p_j(\tilde{\boldsymbol{\beta}}_0)|^{k_2} \|\mathbf{x}_j\|^{k_2}. \tag{S.20}$$

For the first term on the right hand side of (S.20),

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \frac{\|\mathbf{x}_i\|^{k_1-k_2}}{|y_i - p_i(\tilde{\boldsymbol{\beta}}_0)|^{k_2}} &\leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^{k_1-k_2} (1 + e^{\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0} + e^{-\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0})^{k_2} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^{k_1-k_2} (1 + 2e^{\|\mathbf{x}_i\|\|\tilde{\boldsymbol{\beta}}_0\|})^{k_2} \\
&\leq \frac{3^{k_2}}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^{k_1-k_2} e^{\lambda k_2 \|\mathbf{x}_i\|}.
\end{aligned} \tag{S.21}$$

Note that

$$\mathrm{E}\{\|\mathbf{x}_i\|^{k_1-k_2} e^{\lambda k_2 \|\mathbf{x}_i\|}\} \leq \{\mathrm{E}(\|\mathbf{x}_i\|^{2(k_1-k_2)})\mathrm{E}(e^{2\lambda k_2 \|\mathbf{x}_i\|})\}^{1/2} \leq \infty. \tag{S.22}$$

Combining (S.20), (S.21) and (S.22), and using the Law of Large Numbers, (S.19) follows. □

The following lemma is similar to Lemma 1.

Lemma 3. *If Assumption 4 holds, then conditionally on $\mathcal{F}_n$ in probability,*

$$\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0} - \mathbf{M}_X = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S.23}$$

$$\frac{1}{n}\frac{\partial \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S.24}$$

where

$$\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0} = \frac{1}{n}\frac{\partial^2 \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{1}{nr}\sum_{i=1}^{r} \frac{w_i^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})\mathbf{x}_i^*(\mathbf{x}_i^*)^T}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}.$$

Proof. Direct calculation yields,

$$\mathrm{E}(\tilde{\mathbf{M}}_X|\mathcal{F}_n) = \mathrm{E}_{\tilde{\boldsymbol{\beta}}_0}\{\mathrm{E}(\tilde{\mathbf{M}}_X|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0)\} = \mathrm{E}_{\tilde{\boldsymbol{\beta}}_0}(\mathbf{M}_X|\mathcal{F}_n) = \mathbf{M}_X, \tag{S.25}$$

where $\mathrm{E}_{\tilde{\boldsymbol{\beta}}_0}$ means the expectation is taken with respect to the distribution of $\tilde{\boldsymbol{\beta}}_0$ given $\mathcal{F}_n$. For any component $\tilde{\mathbf{M}}_X^{j_1 j_2}(\tilde{\boldsymbol{\beta}}_0)$ of $\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0}$ where $1 \le j_1, j_2 \le d$,

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{n}\tilde{\mathbf{M}}_X^{j_1 j_2}\middle|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right) \\
=&\frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})^2(x_{ij_1}x_{ij_2}^T)^2}{\pi_i(\tilde{\boldsymbol{\beta}}_0)} - \frac{1}{r}(\mathbf{M}_X^{j_1 j_2})^2 \\
\le& \frac{1}{16rn^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^4}{\pi_i(\tilde{\boldsymbol{\beta}}_0)} - \frac{1}{r}(\mathbf{M}_X^{j_1 j_2})^2
\end{aligned} \tag{S.26}$$

From Lemma 2, and (S.26),

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{n}\tilde{\mathbf{M}}_X^{j_1 j_2}\middle|\mathcal{F}_n\right) &= \mathrm{E}_{\tilde{\boldsymbol{\beta}}_0}\left\{\mathrm{Var}\left(\frac{1}{n}\tilde{\mathbf{M}}_X^{j_1 j_2}\middle|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right)\right\} \\
&\le \frac{3}{16r}\frac{1}{n}\sum_{j=1}^{n}\|\mathbf{x}_j\|\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^3 e^{\lambda\|\mathbf{x}_i\|} = O_P(r^{-1}),
\end{aligned} \tag{S.27}$$

Using Markov's inequality, (S.23) follows from (S.25) and (S.27).

Analogously, we obtain that

$$\mathrm{E}\left\{\frac{1}{n}\frac{\partial \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{\partial \boldsymbol{\beta}}\middle|\mathcal{F}_n\right\} = 0, \tag{S.28}$$

and

$$\mathrm{Var}\left\{\frac{1}{n}\frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{\partial \boldsymbol{\beta}}\middle|\mathcal{F}_n\right\} = O_P(r^{-1}). \tag{S.29}$$

From (S.28), (S.29) and Markov's inequality, (S.24) follows. □

Now we prove Theorem 5. By direct calculation,

$$\begin{aligned}
&\mathrm{E}\left\{\frac{\ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{n} - \frac{\ell(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\}^2 \\
&= \frac{1}{r}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{t_i^2(\boldsymbol{\beta})}{\pi_i(\tilde{\boldsymbol{\beta}}_0)} - \left(\frac{1}{n}\sum_{i=1}^{n}t_i(\boldsymbol{\beta})\right)^2\right]
\end{aligned}$$

$$\leq \frac{1}{r}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{(\log 4+2\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|)^2}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}-\left(\frac{1}{n}\sum_{i=1}^{n}t_i(\boldsymbol{\beta})\right)^2\right]. \tag{S.30}$$

Therefore, from Lemma 2 and (S.30),

$$\mathrm{E}\left\{\frac{\ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{n}-\frac{\ell(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2=O_P(r^{-1}). \tag{S.31}$$

Therefore combing (S.31) and the fact that $\mathrm{E}\{\ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})|\mathcal{F}_n\}=\ell(\boldsymbol{\beta})$, we have $n^{-1}\ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})-n^{-1}\ell(\boldsymbol{\beta})\to 0$ in conditional probability given $\mathcal{F}_n$. Thus, conditionally on $\mathcal{F}_n$,

$$\|\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|=o_{P|\mathcal{F}_n}(1) \tag{S.32}$$

The consistency proved above ensures that $\breve{\boldsymbol{\beta}}$ is close to $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ as long as r is large enough. Using Taylor's theorem (c.f. Chapter 4 of Ferguson 1996),

$$0=\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\breve{\boldsymbol{\beta}})}{n}=\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}+\frac{1}{n}\frac{\partial\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{\partial\boldsymbol{\beta}^T}(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})+\frac{1}{n}R_{\tilde{\boldsymbol{\beta}}_0,j} \tag{S.33}$$

where

$$R_{\tilde{\boldsymbol{\beta}}_0,j}=(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})^T\int_0^1\int_0^1\frac{\partial^2\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}\{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}+uv(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}v\mathrm{d}u\mathrm{d}v\ (\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}).$$

Note that

$$\left\|\frac{\partial^2\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\|\leq\frac{1}{r}\sum_{i=1}^{r}\frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}$$

for all $\boldsymbol{\beta}$. Thus

$$\left\|\int_0^1\int_0^1\frac{\partial^2\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}\{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}+uv(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}v\mathrm{d}u\mathrm{d}v\right\|\leq\frac{1}{2r}\sum_{i=1}^{r}\frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}=O_{P|\mathcal{F}_n}(n), \tag{S.34}$$

where the last equality is from the fact that

$$P\left(\frac{1}{nr}\sum_{i=1}^{r}\frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}\geq\tau\middle|\mathcal{F}_n\right)\leq\frac{1}{nr\tau}\sum_{i=1}^{r}\mathrm{E}\left(\frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}\middle|\mathcal{F}_n\right)=\frac{1}{n\tau}\sum_{i=1}^{n}\|\mathbf{x}_i\|^3\to 0, \tag{S.35}$$

in probability as $\tau\to\infty$. From (S.33) and (S.34),

$$\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}=-\big(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0}\big)^{-1})\left\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}+O_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2)\right\}. \tag{S.36}$$

From (S.23) of Lemma 2, $(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} = O_{P|\mathcal{F}_n}(1)$. Combining this with (S.25), (S.32) and (S.36)

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2}) + o_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|),$$

which implies that

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S.37}$$

S.1.5 Proof of Theorem 6

Denote

$$\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} = \frac{1}{r}\sum_{i=1}^{r} \frac{\{y_i^* - p_i^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}\mathbf{x}_i^*}{n\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} \equiv \frac{1}{r}\sum_{i=1}^{r} \boldsymbol{\eta}_i^{\tilde{\boldsymbol{\beta}}_0} \tag{S.38}$$

Given $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, $\boldsymbol{\eta}_1^{\tilde{\boldsymbol{\beta}}_0}, ..., \boldsymbol{\eta}_r^{\tilde{\boldsymbol{\beta}}_0}$ are i.i.d, with mean $\mathbf{0}$ and variance

$$\text{Var}(\boldsymbol{\eta}_i|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0) = r\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0} = \frac{1}{n^2}\sum_{i=1}^{n} \frac{\{y_i - p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2 \mathbf{x}_i\mathbf{x}_i^T}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}. \tag{S.39}$$

Meanwhile, for every $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{i=1}^{r} \text{E}\{\|r^{-1/2}\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\beta}}_0}\|^2 I(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\varepsilon)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} \\
&\leq \frac{1}{r^{3/2}\varepsilon}\sum_{i=1}^{r} \text{E}\{\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\beta}}_0}\|^3 I(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\varepsilon)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} \leq \frac{1}{r^{3/2}\varepsilon}\sum_{i=1}^{r} \text{E}(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\beta}}_0}\|^3|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0) \\
&= \frac{1}{r^{1/2}}\frac{1}{n^3}\sum_{i=1}^{n} \frac{\{y_i - p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^3\|\mathbf{x}_i\|^3}{\pi_i^2(\tilde{\boldsymbol{\beta}}_0)} \leq \frac{1}{r^{1/2}}\frac{1}{n^3}\sum_{i=1}^{n} \frac{\|\mathbf{x}_i\|^3}{\pi_i^2(\tilde{\boldsymbol{\beta}}_0)} = o_P(1)
\end{aligned}$$

where the last equality is from Lemma 2. This and (S.39) show that the Lindeberg-Feller conditions are satisfied in probability. From (S.38) and (S.39), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart 1998), conditionally on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$,

$$\frac{1}{n}(\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = \frac{1}{r^{1/2}}\{\text{Var}(\boldsymbol{\eta}_i|\mathcal{F}_n)\}^{-1/2}\sum_{i=1}^{r}\boldsymbol{\eta}_i \to N(0, I),$$

in distribution.

Now we exam the distance between $\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0}$ and $\mathbf{V}_c$. First,

$$\|\mathbf{V}_c - \mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0}\| \leq \frac{1}{rn^2}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2\left|\frac{1}{\pi_i} - \frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\right| \tag{S.40}$$

For the last term in the above equation,

$$\begin{aligned}
&\left|\frac{1}{\pi_i}-\frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\right| \\
&\le \left|\frac{\sum_{j=1}^n |y_j-p_j(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_j\|}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_i\|}-\frac{\sum_{j=1}^n |y_j-p_j(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{x}_j\|}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_i\|}\right| \\
&\quad + \left|\frac{\sum_{j=1}^n |y_j-p_j(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{x}_j\|}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_i\|}-\frac{\sum_{j=1}^n |y_j-p_j(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{x}_j\|}{|y_i-p_i(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{x}_i\|}\right| \\
&\le \frac{\sum_{j=1}^n |p_j(\tilde{\boldsymbol{\beta}}_0)-p_j(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_j\|}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_i\|}+\left|\frac{1}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|}-\frac{1}{|y_i-p_i(\tilde{\boldsymbol{\beta}}_0)|}\right|\frac{\sum_{j=1}^n \|\mathbf{x}_j\|}{\|\mathbf{x}_i\|} \qquad \text{(S.41)}
\end{aligned}$$

Note that

$$|p_j(\tilde{\boldsymbol{\beta}}_0)-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})| \le \|\mathbf{x}_i\|\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\text{MLE}}\|, \qquad \text{(S.42)}$$

and

$$\begin{aligned}
\left|\frac{1}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|}-\frac{1}{|y_i-p_i(\tilde{\boldsymbol{\beta}}_0)|}\right| &= \left|e^{(2y_i-1)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\text{MLE}}}-e^{(2y_i-1)\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0}\right| \\
&\le e^{\lambda\|\mathbf{x}_i\|}\|\mathbf{x}_i\|\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\text{MLE}}\|. \qquad \text{(S.43)}
\end{aligned}$$

From (S.40), (S.41), (S.42) and (S.43),

$$\|\mathbf{V}_c-\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0}\| \le \frac{\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\text{MLE}}\|}{r}C_1 = O_{P|\mathcal{F}_n}(r^{-1}r_0^{-1/2}), \qquad \text{(S.44)}$$

where

$$C_1=\frac{1}{n}\sum_{i=1}^n\frac{\|\mathbf{x}_i\|}{|y_i-p_i(\hat{\boldsymbol{\beta}}_{\text{MLE}})|}\frac{1}{n}\sum_{i=1}^n\|\mathbf{x}_i\|^2+\frac{1}{n}\sum_{i=1}^n\|\mathbf{x}_i\|e^{\lambda\|\mathbf{x}_i\|}\frac{1}{n}\sum_{i=1}^n\|\mathbf{x}_i\|=O_P(1).$$

From Lemma 3, (S.36) and (S.37),

$$\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\text{MLE}} = -\frac{1}{n}(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})+O_{P|\mathcal{F}_n}(r^{-1}) \qquad \text{(S.45)}$$

From (S.23) of Lemma 3,

$$(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1}-\mathbf{M}_X^{-1} = -\mathbf{M}_X^{-1}(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0}-\mathbf{M}_X)(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2}). \qquad \text{(S.46)}$$

From (S.18), (S.45), (S.44) and (S.46),

$$\mathbf{V}^{-1/2}(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\text{MLE}})$$

$$
\begin{aligned}
&= -\mathbf{V}^{-1/2} n^{-1} (\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} \dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}) \\
&= -\mathbf{V}^{-1/2} \mathbf{M}_X^{-1} n^{-1} \dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) - \mathbf{V}^{-1/2} \{(\tilde{\mathbf{M}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} - \mathbf{M}_X^{-1}\} n^{-1} \dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}) \\
&= -\mathbf{V}^{-1/2} \mathbf{M}_X^{-1} (\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0})^{1/2} (\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0})^{-1/2} n^{-1} \dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}).
\end{aligned}
$$

The result in Theorem 1 follows from Slutsky's Theorem(Theorem 6 of Ferguson 1996) and the fact that

$$
\begin{aligned}
\mathbf{V}^{-1/2} \mathbf{M}_X^{-1} (\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0})^{1/2} (\mathbf{V}^{-1/2} \mathbf{M}_X^{-1} (\mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0})^{1/2})^T =& \mathbf{V}^{-1/2} \mathbf{M}_X^{-1} \mathbf{V}_c^{\tilde{\boldsymbol{\beta}}_0} \mathbf{M}_X^{-1} \mathbf{V}^{-1/2} \\
=& \mathbf{V}^{-1/2} \mathbf{M}_X^{-1} \mathbf{V}_c \mathbf{M}_X^{-1} \mathbf{V}^{-1/2} + O_{P|\mathcal{F}_n}(r_0^{-1/2} r^{-1/2}) \\
=& \mathbf{I} + O_{P|\mathcal{F}_n}(r_0^{-1/2} r^{-1/2}),
\end{aligned}
$$

which is obtained using (S.44).

S.1.6 Proofs for nonrandom covariates

To prove the theorems for the case of nonrandom covariates, we need to use the following two assumptions to replace Assumptions 1 and 4, respectively.

Assumption S.1. *As* $n \to \infty$, $\mathbf{M}_X = n^{-1} \sum_{i=1}^n w_i(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \mathbf{x}_i \mathbf{x}_i^T$ *goes to a positive-definite matrix in probability and* $\limsup_n n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^3 < \infty$.

Assumption S.2. *The covariate distribution satisfies that* $n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ *converges to a positive definite matrix, and* $\limsup_n n^{-1} \sum_{i=1}^n e^{a\|\mathbf{x}_i\|} < \infty$ *for any* $a \in \mathbb{R}$.

Note that $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is random, so the condition on $\mathbf{M}_X$ holds in probability in Assumption S.1. π_i's could be functions of the responses, and the optimal π_i's are indeed functions of the responses. Thus Assumptions 2 and 3 involve random terms and remain unchanged.

The proof of Lemma 1 does not require the condition that $n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^3 = O_P(1)$, so it is automatically valid for nonrandom covariates. The proof of Theorem 1 requires $n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^3 = O_P(1)$ in (S.11). If it is replaced with $\limsup_n n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^3 < \infty$, (S.11) still holds. Thus Theorem 1 is valid if Assumptions 2 and S.1 are true.

Theorem 2 is built upon Theorem 1 and does not require additional conditions besides Assumption 3. Thus it is valid under Assumptions 2, 3 and S.1.

Theorems 3 and 4 are proved by the application of Cauchy-Schwarz inequality, and they are valid regardless whether the covariates are random or nonrandom.

To prove Theorems 5 and 6 for nonrandom covariates, we first prove Lemma 2. From Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^{k_1-k_2}e^{\lambda k_2\|\mathbf{x}_i\|} &\le \left\{\left(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^{2(k_1-k_2)}\right)\left(\frac{1}{n}\sum_{i=1}^{n}e^{2\lambda k_2\|\mathbf{x}_i\|}\right)\right\}^{1/2} \\ &\le \left\{\frac{\{2(k_1-k_2)\}!}{n}\sum_{i=1}^{n}e^{\|\mathbf{x}_i\|}\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}e^{2\lambda k_2\|\mathbf{x}_i\|}\right\}^{1/2}\end{aligned}$$

Thus, under Assumption S.2,

$$\limsup_{n}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^{k_1-k_2}e^{\lambda k_2\|\mathbf{x}_i\|} \le \infty. \tag{S.47}$$

Combining (S.20), (S.21) and (S.47), Lemma 2 follows. With the results in Lemma 2, the proofs of Lemma 3 and Theorem 5, and Theorem 6 are the same as those in Section S.1.4, and Section S.1.5, respectively, except that $(n\tau)^{-1}\sum_{i=1}^{n}\|\mathbf{x}_i\|^3 \to 0$ deterministically instead of in probability in (S.35).

S.2 Additional numerical results

In this section, we provide additional numerical results for rare events data and unconditional MSEs.

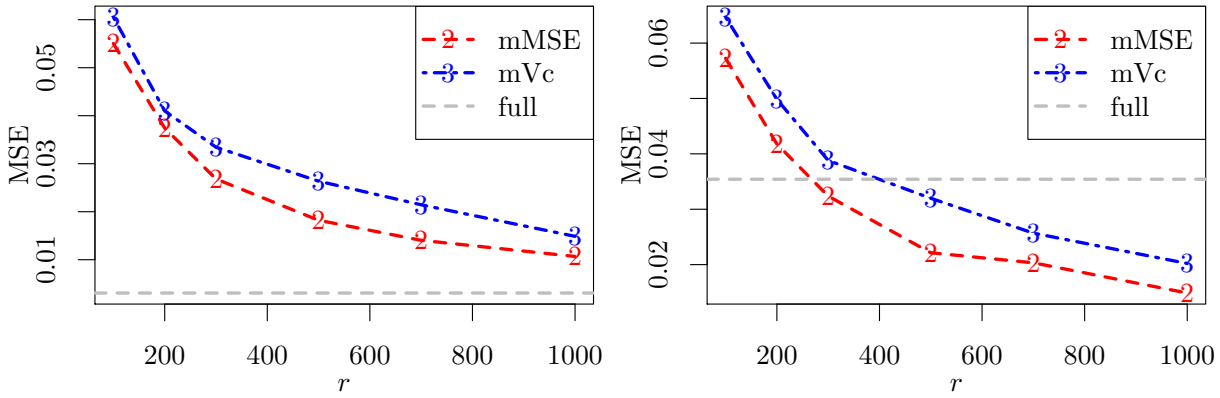
S.2.1 Further numerical evaluations for rare events data

To further investigate the performance of the proposed method for more extreme rare events data, we adopt the model setup with a univariate covariate in King & Zeng (2001), namely,

$$P(y=1|x) = \frac{1}{1+\exp(-\beta_0-\beta_1 x)}.$$

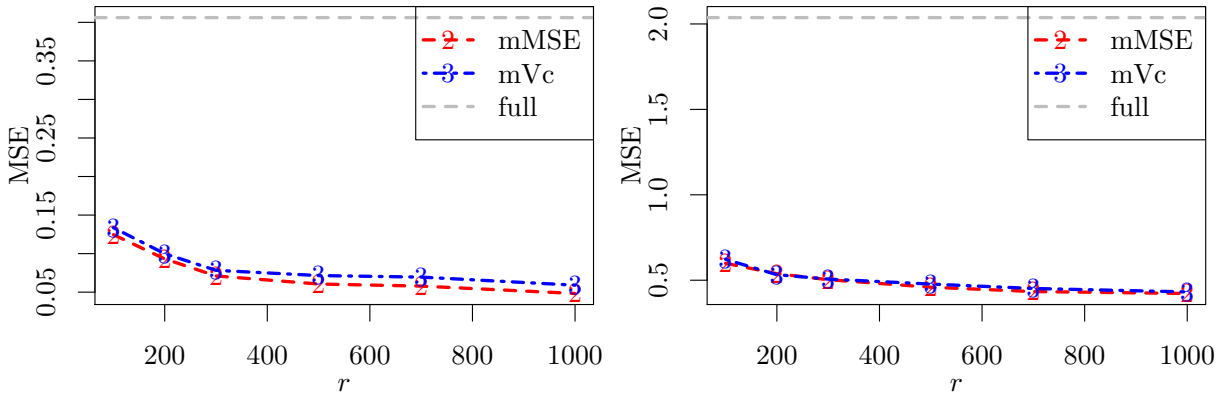
Following King & Zeng (2001), we assume that the covariate x follows a standard normal distribution and consider different values of β_0 and a fixed value of $\beta_1 = 1$. The full data sample size is set to $n = 10^6$ and β_0 is set to $-7, -9.5, -12.5$, and -13.5, generating responses with the percentages of 1's equaling 0.1493%, 0.0111%, 0.0008%, and 0.0002%

respectively. For the last case there are only two 1's (0.0002%) in the full data of $n = 10^6$, and this is a very extreme case of rare events data. For comparison, we also calculate the MSE of the full data approach using 1000 Bootstrap sample (the gray dashed line). Results are reported in Figure S.1. It is seen that as the rare event rate gets closer to 0, the performance of the OSMAC methods relative to the full data Bootstrap gets better. When the rare event rate is 0.0002%, for the full data Bootstrap approach, there are 110 cases out of 1000 Bootstrap samples that the MLE are not found, while this occurs for 18, 2, 4, and 1 cases when $r_0 = 200$, and $r = 200, 500, 700$, and 1000, respectively.



(a) 0.1493% of y_i's are 1

(b) 0.0111% of y_i's are 1



(c) 0.0008% of y_i's are 1

(d) 0.00002% of y_i's are 1

Figure S.1: MSEs for rare event data with different second step subsample size r and a fixed first step subsample size $r_0 = 200$, where the covariate follows the standard normal distribution.

S.2.2 Numerical results on unconditional MSEs

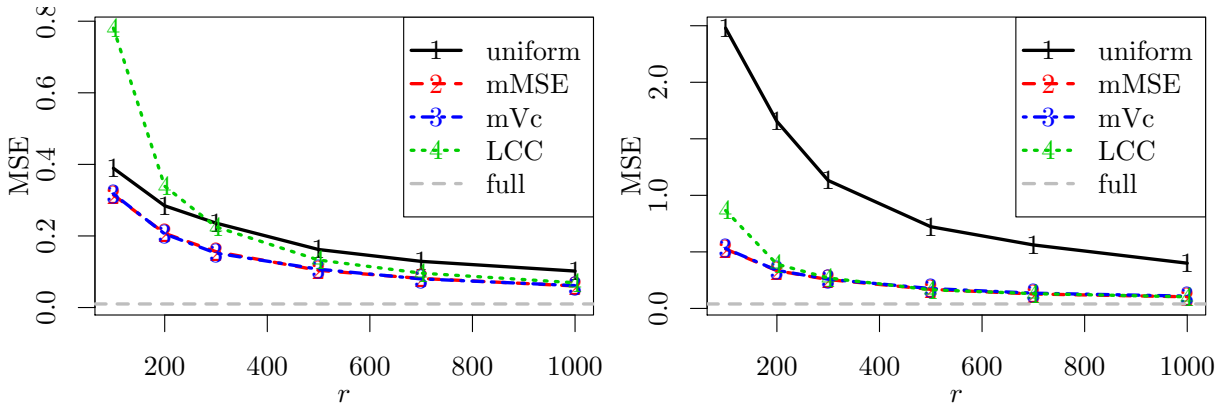
To calculate unconditional MSEs, we generate the full data in each repetition and then apply the subsampling methods. This way, the resultant MSEs are the unconditional MSEs. The exactly same configurations in Section 5 are used. Results are presented in Figure S.2. It is seen that the unconditional results are very similar to the conditional results, even for the imbalanced case of nzNormal data sets. For extreme imbalanced data or rare events data, the conditional MSE and the unconditional MSE can be different, as seen in the results in Section S.2.1.

References

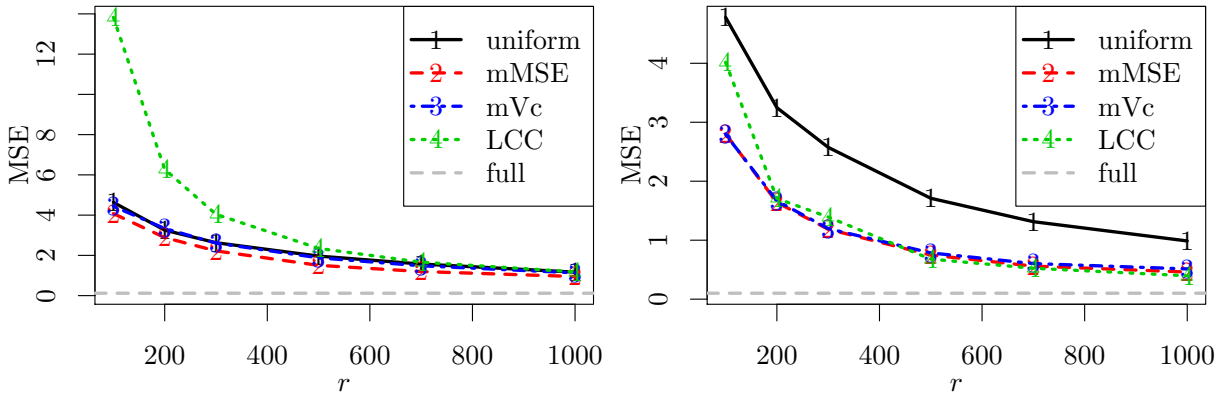
Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman and Hall.

King, G. & Zeng, L. (2001), 'Logistic regression in rare events data', *Political analysis* **9**(2), 137–163.

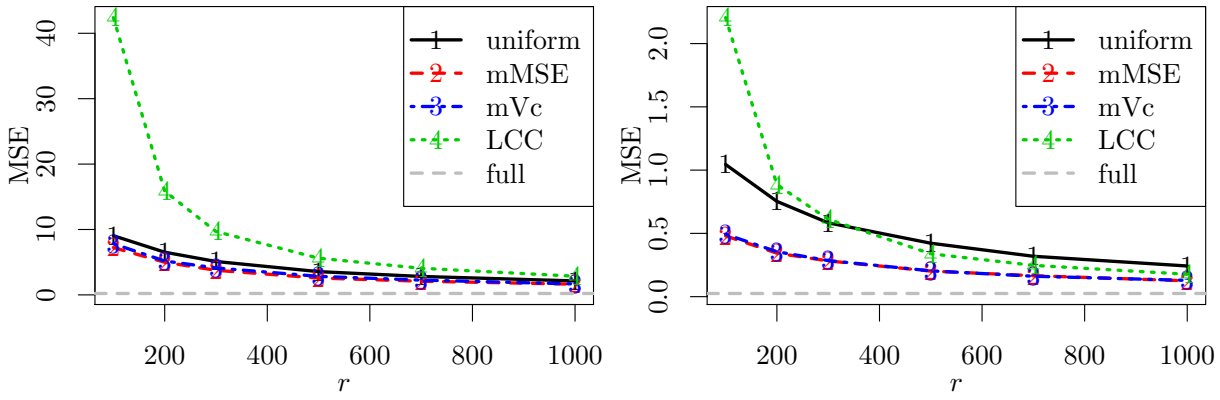
van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press, London.



(a) mzNormal (b) nzNormal



(c) ueNormal (d) mixNormal



(e) T_3 (f) EXP

Figure S.2: Unconditional MSEs for different second step subsample size r with the first step subsample size being fixed at $r_0 = 200$.