OPTIMAL SUBSAMPLING ALGORITHMS FOR BIG DATA REGRESSIONS

Mingyao Ai[1], Jun Yu[2], Huiming Zhang[1], HaiYing Wang[3]

LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University [1]

School of Mathematics and Statistics, Beijing Institute of Technology[2]

Department of Statistics, University of Connecticut [3]

Supplementary Material

S1 Proofs

To prove Theorem 1, we begin with the following remark and lemma.

Remark 1. By the fact that $\lambda(\theta) := \exp(\psi(\theta))$ is analytic in the interior of Θ (see Theorem 2.7 in Brown (1986)), Cauchy's integral formula tells us that all its higher derivatives exist and are continuous. Therefore, the derivatives $\dot{\psi}(t)$, $\ddot{\psi}(t)$, $\dddot{\psi}(t)$ are continuous in t, and $\dot{\psi}(t)$,$\ddot{\psi}(t)$ are bounded on the compact set, which follows by a well-known property that every real-valued continuous function on a compact set is necessarily bounded.

Lemma 1. *If Assumptions (H.1)–(H.4) and (H.6) hold, then as $r \to \infty$ and $n \to \infty$, conditionally on $\mathcal{F}_n$ in probability,*

$$\check{\mathcal{J}}_X - \mathcal{J}_X = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S1.1}$$

$$\frac{1}{n}L^*(\boldsymbol{\beta}) - \frac{1}{n}L(\boldsymbol{\beta}) = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S1.2}$$

$$\frac{1}{n}\frac{\partial L^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S1.3}$$

where

$$\check{\mathcal{J}}_X = -\frac{1}{n}\frac{\partial^2 L^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{1}{nr}\sum_{i=1}^{r}\frac{\ddot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}^*{}_i))\dot{u}(\boldsymbol{\beta}^T\boldsymbol{x}^*{}_i)\boldsymbol{x}_i^*[\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*]^T}{\pi_i^*} + \frac{1}{nr}\sum_{i=1}^{r}\frac{\ddot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*\boldsymbol{x}^*{}_i^T[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i^*)) - y_i^*]}{\pi_i^*}.$$

Proof. By the definition of conditional expectation and towering property of flirtations, it yields that

$$E(\check{\mathcal{J}}_X|\mathcal{F}_n) = \mathcal{J}_X.$$

For any component $\check{\mathcal{J}}_X^{j_1j_2}$ of $\check{\mathcal{J}}_X$ where $1 \le j_1, j_2 \le p$,

$$\begin{aligned}
&\mathrm{E}\left(\check{\mathcal{J}}_X^{j_1j_2} - \mathcal{J}_X^{j_1j_2}\middle|\mathcal{F}_n\right)^2 \\
=&\frac{1}{r}\sum_{i=1}^{n}\pi_i\left\{\frac{\ddot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i))\dot{u}^2(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}}{n\pi_i}\right. \\
&\left.+\frac{\ddot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)) - y_i]}{n\pi_i} - \mathcal{J}_X^{j_1j_2}\right\}^2 \\
=&\frac{1}{r}\sum_{i=1}^{n}\pi_i\left\{\frac{\ddot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i))\dot{u}^2(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}}{n\pi_i}\right.
\end{aligned} \tag{S1.4}$$

$$
\begin{aligned}
&+\frac{\ddot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i]}{n\pi_i}\Bigg\}^2-\frac{1}{r}\left(\mathcal{J}_X^{j_1j_2}\right)^2\\
\leq&\frac{2}{r}\sum_{i=1}^{n}\pi_i\Bigg[\left\{\frac{\dddot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}}{n\pi_i}\right\}^2\\
&+\left\{\frac{\ddot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i]}{n\pi_i}\right\}^2\Bigg]\\
\leq&\frac{2}{r}\sum_{i=1}^{n}\pi_i\left[O_P(1)\left\{\frac{x_{ij_1}x_{ij_2}}{n\pi_i}\right\}^2+O_P(1)\left\{\frac{x_{ij_1}x_{ij_2}[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i]}{n\pi_i}\right\}^2\right]\\
=&O_{P|\mathcal{F}_n}(\frac{1}{r}),
\end{aligned}
$$

where the last inequality stems form (H.1) and the last equality holds by assumptions (H.3) and (H.6). From Chebyshev's inequality, it is proved that Equation (S1.1) holds.

To prove Equation (S1.3), let $t_i(\boldsymbol{\beta})=y_iu(\boldsymbol{\beta}^T\boldsymbol{x}_i)-\psi(u(\boldsymbol{\beta}^T\boldsymbol{x}_i)$, $t_i^*(\boldsymbol{\beta})=y_i^*u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)-\psi(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))$, then

$$
L^*(\boldsymbol{\beta})=\frac{1}{r}\sum_{i=1}^{r}\frac{t_i^*(\boldsymbol{\beta})}{\tilde{\pi}_i^*},\quad\text{and}\quad L(\boldsymbol{\beta})=\sum_{i=1}^{n}t_i(\boldsymbol{\beta}).
$$

Under the conditional distribution of the subsample given $\mathcal{F}_n$,

$$
E\left\{\frac{L^*(\boldsymbol{\beta})}{n}-\frac{L(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2=\frac{1}{rn^2}\sum_{i=1}^{n}\frac{t_i^2(\boldsymbol{\beta})}{\pi_i}-\frac{1}{r}\left(\frac{1}{n}\sum_{i=1}^{n}t_i(\boldsymbol{\beta})\right)^2.
$$

Combining the facts that the parameter space is compact and $u(t)$ is continuous function, by assumption (H.1) we have that $u(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ are uniformly bounded. Then, it can be shown that by Remark 1 that

$$
\begin{aligned}
|t_i(\boldsymbol{\beta})|&\leq|y_iu(\boldsymbol{\beta}^T\boldsymbol{x}_i)-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))u(\boldsymbol{\beta}^T\boldsymbol{x}_i)|+|\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))u(\boldsymbol{\beta}^T\boldsymbol{x}_i)-\psi(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))|\\
&\leq|[y_i-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))]u(\boldsymbol{\beta}^T\boldsymbol{x}_i)|+|\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))u(\boldsymbol{\beta}^T\boldsymbol{x}_i)|+|\psi(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))|,
\end{aligned}
$$

Therefore, we have

$$\left(\frac{1}{n}\sum_{i=1}^{n} t_i(\boldsymbol{\beta})\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} O_P(1)|y_i - \dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))|\right)^2 + \frac{O_P(1)}{n}\sum_{i=1}^{n}|y_i - \dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))| + O_P(1).$$

From Assumptions (H.1), we have $\sup_n n^{-1}\sum_{i=1}^{n} t_i(\boldsymbol{\beta}) < \infty$. Thus,

$$E\left\{\frac{L^*(\boldsymbol{\beta})}{n} - \frac{L(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2 = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S1.5}$$

Now the desired result (S1.2) follows from Chebyshev's Inequality.

Similarly, we can show that

$$\mathrm{Var}\left(\frac{1}{n}\frac{\partial L^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{\partial\boldsymbol{\beta}}\right) = O_P(r^{-1}).$$

Thus (S1.3) is true. □

S1.1 Proof of Theorem 1

Proof. As $r \to \infty$, by (S1.5), we have that $n^{-1}L^*(\boldsymbol{\beta}) - n^{-1}L(\boldsymbol{\beta}) \to 0$ in conditional probability given $\mathcal{F}_n$. Note that the parameter space is compact and $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ is the unique global maximum of the continuous convex function $L(\boldsymbol{\beta})$. Thus, from Theorem 5.9 and its remark of van der Vaart (1998), by (S1.3) we have

$$\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\| = o_{P|\mathcal{F}_n}(1). \tag{S1.6}$$

as $n \to \infty, r \to \infty$, conditionally on $\mathcal{F}_n$ in probability.

Using Taylor's theorem for random variables (see Ferguson, 1996, Chapter 4),

$$0 = \frac{\dot{L}_j^*(\tilde{\boldsymbol{\beta}})}{n} = \frac{\dot{L}_j^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + \frac{1}{n}\frac{\partial \dot{L}_j^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}^T}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + \frac{1}{n}R_j, \tag{S1.7}$$

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

$$\frac{1}{n}R_j = \frac{1}{n}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})^T \int_0^1 \int_0^1 \frac{\partial^2 \dot{L}_j^*\{\hat{\boldsymbol{\beta}}_{\text{MLE}} + uv(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})\}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} v du dv \ (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}).$$

By calculus, we get

$$\begin{aligned}\frac{\partial^2 \dot{L}_j^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{1}{r}\sum_{i=1}^{r}\{&\frac{\dddot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)[y_i^* - \dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))]}{\pi_i^*} - \frac{\ddot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)\dot{u}(\beta^T\boldsymbol{x}_i^*)\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))}{\pi_i^*} \\ &- \frac{2\ddot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))\ddot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)\dot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)}{\pi_i^*} - \frac{\dddot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))\dot{u}^2(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)}{\pi_i^*}\}x_{ij}^*\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}.\end{aligned}$$

From (H.1) and Remark 1, we have

$$\begin{aligned}&\frac{1}{n}\left\|\frac{\partial^2 \dot{L}_j^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right\|_S \\ =&\frac{1}{nr}\left\|\sum_{i=1}^{r}\left(\frac{\ddot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))}{\pi_i^*} + \frac{2\ddot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))\ddot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)}{\pi_i^*} + \right.\right. \\ &\left.\left.\frac{\dddot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))\dot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)}{\pi_i^*}\right)\dot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)x_{ij}^*\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\right\|_S \\ &+ \frac{1}{rn}\left\|\sum_{i=1}^{r}\frac{\dddot{u}(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)[y_i^* - \dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*))]}{\pi_i^*}\right\|_S \\ \leq&\frac{C_3}{rn}\sum_{i=1}^{r}\frac{\|\boldsymbol{x}_i^*\|^3}{\pi_i^*} + \frac{C_4}{rn}\left\|\sum_{i=1}^{r}\frac{|[y_i^* - \dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i^*)]x_{ij}^*|}{\pi_i^*}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\right\|_S,\end{aligned}$$

for all $\boldsymbol{\beta} \in \Lambda_B$, where C_3 and C_4 are some constants according to Remark 1.

As $\tau \to \infty$, assumption (H.5) gives,

$$P\left(\frac{1}{nr}\sum_{i=1}^{r}\frac{\|\boldsymbol{x}_i^*\|^3}{\pi_i^*} \geq \tau \middle| \mathcal{F}_n\right) \leq \frac{1}{nr\tau}\sum_{i=1}^{r} E\left(\frac{\|\boldsymbol{x}_i^*\|^3}{\pi_i^*}\middle|\mathcal{F}_n\right) = \frac{1}{n\tau}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^3 \xrightarrow{P} 0.$$

Also note that as $\tau \to \infty$,

$$\begin{aligned}
&P\left(\frac{1}{nr}\sum_{i=1}^{r}\left\|\frac{|[y_i-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))]x_{ij}|\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i^*}\right\|_S \geq \tau\middle|\mathcal{F}_n\right)\\
\leq& \frac{1}{nr\tau}E\left(\sum_{i=1}^{r}\left\|\frac{|[y_i-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))]x_{ij}|\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i^*}\right\|_S\middle|\mathcal{F}_n\right)\\
\leq& \frac{1}{n\tau}\sum_{i=1}^{n}\left\||[y_i-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))]x_{ij}|\boldsymbol{x}_i\boldsymbol{x}_i^T\right\|_S\\
\leq& \frac{1}{\tau n}\sum_{i=1}^{n}|[y_i-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))]|\times\left\|x_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T\right\|_S\\
\leq& \frac{1}{\tau}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i-\dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))|^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|x_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T\|_S^2}\\
=&\frac{1}{\tau}O_P(1)\xrightarrow{P}0,
\end{aligned}$$

where the last equality is due to (H.3) and (H.5) by noting that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\|x_{ij}^*\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\right\|_S^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left|x_{ij}^*\right|^2\left\|\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\right\|_S^2\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2\left\|\boldsymbol{x}_i\boldsymbol{x}_i^T\right\|_S^2 \leq \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^6.
\end{aligned}$$

Thus we have

$$\frac{1}{n}\left\|\frac{\partial^2\dot{L}_j^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\|_S = O_{P|\mathcal{F}_n}(1).$$

From (H.1)-(H.3) and Remark 1, it is known that

$$\frac{1}{n}\left\|\int_0^1\int_0^1\frac{\partial^2\dot{L}_j^*\{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}+uv(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}vdudv\right\|$$

$$\leq \frac{1}{n}\int_0^1\int_0^1 \left\| \frac{\partial^2 \dot{L}_j^*\{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} + uv(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \right\| v du dv = O_{P|\mathcal{F}_n}(1).$$

Combining the above equations with the Taylor's expansion (S1.7), we have

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = -\check{\mathcal{J}}_X^{-1}\left\{ \frac{\dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} + O_{P|\mathcal{F}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2) \right\}. \tag{S1.8}$$

From Lemma 1 and Assumption (H.4), it is obvious that $\check{\mathcal{J}}_X^{-1} = O_{P|\mathcal{F}_n}(1)$. Therefore,

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

which implies that $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2})$. □

S1.2 Proof of Theorem 2

Proof. Note that

$$\frac{\dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = \frac{1}{r}\sum_{i=1}^{r} \frac{\{y_i^* - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*))\}\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*}{n\pi_i^*} =: \frac{1}{r}\sum_{i=1}^{r}\eta_i. \tag{S1.9}$$

It can be seen that given $\mathcal{F}_n$, $\eta_1, \ldots, \eta_r$ are i.i.d random variables with mean $\mathbf{0}$ and variance

$$\mathrm{var}(\eta_1|\mathcal{F}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \frac{\{y_i - \dot{\psi}_i(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))\}^2\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i}. \tag{S1.10}$$

Then from (H.7) with $\gamma = 0$, we know that $\mathrm{var}(\eta_i|\mathcal{F}_n) = O_P(1)$ as $n \to \infty$.

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

$$\begin{aligned}
&\sum_{i=1}^{r} E\{\|r^{-1/2}\eta_i\|^2 I(\|\eta_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\} \\
\leq &\frac{1}{r^{1+\gamma/2}\varepsilon^{\gamma}} \sum_{i=1}^{r} E\{\|\eta_i\|^{2+\gamma} I(\|\eta_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\} \\
\leq &\frac{1}{r^{1+\gamma/2}\varepsilon^{\gamma}} \sum_{i=1}^{r} E(\|\eta_i\|^{2+\gamma}|\mathcal{F}_n) \\
= &\frac{1}{r^{\gamma/2}} \frac{1}{\varepsilon^{\gamma}} \frac{1}{n^{2+\gamma}} \sum_{i=1}^{n} \frac{|y_i - \dot{\psi}_i(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))|^{2+\gamma} \|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|^{2+\gamma}}{\pi_i^{1+\gamma}}.
\end{aligned} \tag{S1.11}$$

From (H.7) for some $\gamma > 0$, we obtain

$$\sum_{i=1}^{r} E\{\|r^{-1/2}\eta_i\|^2 I(\|\eta_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\} \leq \frac{1}{r^{\gamma/2}} \frac{1}{\varepsilon^{\gamma}} O_P(1) \cdot O_P(1) = o_P(1),$$

This shows that the Lindeberg-Feller conditions are satisfied in probability. From (S1.9) and (S1.10), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart, 1998), conditionally on $\mathcal{F}_n$,

$$\frac{1}{n} V_c^{-1/2} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) = \frac{1}{r^{1/2}} \{\mathrm{var}(\eta_i|\mathcal{F}_n)\}^{-1/2} \sum_{i=1}^{r} \eta_i \to N(0, I),$$

in distribution. From Lemma 1, (S1.8) and Theorem 1, we have

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = -\frac{1}{n} \check{\mathcal{J}}_X^{-1} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1}). \tag{S1.12}$$

From (S1.1) in Lemma 1, it follows that

$$\check{\mathcal{J}}_X^{-1} - \mathcal{J}_X^{-1} = -\mathcal{J}_X^{-1}(\check{\mathcal{J}}_X - \mathcal{J}_X)\check{\mathcal{J}}_X^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S1.13}$$

Based on Assumption (H.4) and (S1.10), it can be proved that

$$V = \mathcal{J}_X^{-1} V_c \mathcal{J}_X^{-1} = \frac{1}{r} \mathcal{J}_X^{-1} (r V_c) \mathcal{J}_X^{-1} = O_P(r^{-1}).$$

Thus,

$$
\begin{aligned}
&V^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) = -V^{-1/2} n^{-1} \check{\mathcal{J}}_X^{-1} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}) \\
=& - V^{-1/2} \mathcal{J}_X^{-1} n^{-1} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) - V^{-1/2}(\check{\mathcal{J}}_X^{-1} - \mathcal{J}_X^{-1}) n^{-1} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}) \\
=& - V^{-1/2} \mathcal{J}_X^{-1} V_c^{1/2} V_c^{-1/2} n^{-1} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}).
\end{aligned}
$$

So the result in (2.4) of Theorem 2 follows by applying Slutsky's Theorem (Theorem 6, Section 6 of Ferguson, 1996) and the fact that

$$
V^{-1/2} \mathcal{J}_X^{-1} V_c^{1/2} (V^{-1/2} \mathcal{J}_X^{-1} V_c^{1/2})^T = V^{-1/2} \mathcal{J}_X^{-1} V_c^{1/2} V_c^{1/2} \mathcal{J}_X^{-1} V^{-1/2} = I.
$$

□

S1.3 Proof of Theorem 3

Proof. Note that

$$
\begin{aligned}
&\mathrm{tr}(V) = \mathrm{tr}(\mathcal{J}_X^{-1} V_c \mathcal{J}_X^{-1}) \\
=& \frac{1}{n^2 r} \sum_{i=1}^{n} \mathrm{tr} \left[\frac{1}{\pi_i} \{ y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)) \}^2 \mathcal{J}_X^{-1} \dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i [\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i]^T \mathcal{J}_X^{-1} \right] \\
=& \frac{1}{n^2 r} \sum_{i=1}^{n} \left[\frac{1}{\pi_i} \{ y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)) \}^2 \| \mathcal{J}_X^{-1} \dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i \|^2 \right] \\
=& \frac{1}{n^2 r} (\sum_{i=1}^{n} \pi_i) \sum_{i=1}^{n} \left[\pi_i^{-1} \{ y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)) \}^2 \| \mathcal{J}_X^{-1} \dot{u}(\boldsymbol{\beta}^T \boldsymbol{x}_i) \boldsymbol{x}_i \|^2 \right] \\
\geq& \frac{1}{n^2 r} \left[\sum_{i=1}^{n} |y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))| \| \mathcal{J}_X^{-1} \dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i \| \right]^2,
\end{aligned}
$$

where the last inequality follows from the Cauchy-Schwarz inequality, and the equality in it holds if and only if

$$\pi_i \propto |y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))| \|\mathcal{J}_X^{-1} \boldsymbol{x}_i\| I\{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))| \|\mathcal{J}_X^{-1} \boldsymbol{x}_i\| > 0\}.$$

Here we define $0/0 = 0$, and this is equivalent to removing data points with $|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))| = 0$ in the expression of V_c. □

S1.4 Proof of Theorems 5 and 6

Let $\|A\|_F := (\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2)^{1/2}$ denote the Frobenius norm. For a given $m \times n$ matrix A and an $n \times p$ matrix B, we want to get an approximation to the product AB. In the following fast Monte Carlo algorithm in Drineas et al. (2006), we do r independent trials. In each trial we randomly sample an element of $\{1, 2, \cdots, n\}$ with given discrete distribution $P =: \{p_i\}_{i=1}^n$. Then we extract an $m \times r$ matrix C from the columns of A, and extract an $r \times n$ matrix R from the corresponding rows of B. If the P is chosen appropriately in the sense that CR is a nice approximation to AB, then the F-norm matrix concentration inequality in Lemma 2 holds with high probability.

Lemma 2. *(Theorem 2.1 in Drineas et al. (2006)) Let $A^{(i)}$ be the i-th row of $A \in R^{m \times n}$ as row vector and $B_{(j)}$ be the j-th column of $B \in R^{n \times p}$ as column vector. Suppose sampling probabilities $\{p_i\}_{i=1}^n, (\sum_{i=1}^n p_i = 1)$ are*

such that

$$p_i \geq \beta \frac{\left\|A^{(i)}\right\| \left\|B_{(j)}\right\|}{\sum_{j=1}^{n} \left\|A^{(i)}\right\| \left\|B_{(j)}\right\|}$$

for some $\beta \in (0, 1]$. *Construct* C *and* R *with Algorithm 1 in Drineas et al. (2006), and assume that* $\varepsilon \in (0, 1/3)$. *Then, with probability at least* $1 - \varepsilon$, *we have*

$$\|AB - CR\|_F \leq \frac{4\sqrt{\log(1/\varepsilon)}}{\beta\sqrt{c}} \|A\|_F \|B\|_F.$$

Now we prove Theorems 5 and 6 by applying the above Lemma 2.

Proof. Note the fact that the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ of the parameter vector $\boldsymbol{\beta}$ satisfy the following estimation equation

$$\boldsymbol{X}^T[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\boldsymbol{\beta}))]\dot{u}(\boldsymbol{X}^T\boldsymbol{\beta}) = \mathbf{0}, \tag{S1.14}$$

where $\dot{\psi}(u(\boldsymbol{X}^T\boldsymbol{\beta}))$ denotes the $n \times n$ diagonal matrix whose i-th element in its diagonal is $\dot{\psi}(u(\boldsymbol{x}_i^T\boldsymbol{\beta}))$.

Without of loss of generality, we only show the case with probability $\boldsymbol{\pi}^{\mathrm{mV}}$, since the proof for $\boldsymbol{\pi}^{\mathrm{mVc}}$ is quite similar. Let S be an $n \times r$ matrix whose i-th column is $1/\sqrt{r\pi_{j_i}^{\mathrm{mV}}}\boldsymbol{e}_{j_i}$, where $\boldsymbol{e}_{j_i} \in \mathbb{R}^n$ denotes the all-zeros vector except that its j_i-th entry is set to one. Here j_i denotes the j_i-th data point chosen from the i-th independent random subsampling with probabilities $\boldsymbol{\pi}^{\mathrm{mV}}$. Then $\tilde{\boldsymbol{\beta}}$ satisfies the following equation

$$\boldsymbol{X}^T SS^T[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\boldsymbol{\beta}))]\dot{u}(\boldsymbol{X}^T\boldsymbol{\beta}) = \mathbf{0}. \tag{S1.15}$$

Let $\|\cdot\|_F$ denote the Frobenius norm, we have

$$
\begin{aligned}
&\sigma_{\min}(\boldsymbol{X}^T SS^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))\|[\dot{\psi}(u(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})) - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\| \\
&\leq \|\boldsymbol{X}^T SS^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\dot{\psi}(u(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})) - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|_F \\
&\leq \|\boldsymbol{X}^T SS^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\dot{\psi}(u(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})) - \boldsymbol{y}]\|_F \\
&+ \|\boldsymbol{X}^T SS^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|_F \\
&= \|\boldsymbol{X}^T SS^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|_F \quad [by\ (S1.15)] \\
&\leq \|\boldsymbol{X}^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|_F \\
&+ \left\|\boldsymbol{X}^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))] - \boldsymbol{X}^T SS^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\right\|_F \\
&\leq \|\boldsymbol{X}^T \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|_F \\
&+ \frac{4\kappa(\mathcal{J}_X^{-1})\sqrt{\log(1/\epsilon)}}{\sqrt{r}}\|\boldsymbol{X}\|_F\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\| \\
&\leq \sigma_{\max}(\boldsymbol{X})\sqrt{p}\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\| \\
&+ \frac{4\kappa(\mathcal{J}_X^{-1})\sqrt{\log(1/\epsilon)}}{\sqrt{r}}\sigma_{\max}(\boldsymbol{X})\sqrt{p}\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\| \\
&\leq [1 + \frac{4\kappa(\mathcal{J}_X^{-1})\sqrt{\log(1/\epsilon)}}{\sqrt{r}}]\sigma_{\max}(\boldsymbol{X})\sqrt{p}\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\| \\
&\leq C_{\dot{u}}[1 + \frac{4\kappa(\mathcal{J}_X^{-1})\sqrt{\log(1/\epsilon)}}{\sqrt{r}}]\sigma_{\max}(\boldsymbol{X})\sqrt{p}\|[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|
\end{aligned}
$$

where the fourth last inequality follows from Lemma 2 by putting $A = \boldsymbol{X}^T\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}), B = \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})[\boldsymbol{y}-\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))], C = \boldsymbol{X}^T\boldsymbol{S}, R = \boldsymbol{S}^T\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})(\boldsymbol{y}-\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})))$ and $\beta = 1/\kappa(\mathcal{J}_X^{-1})$, and last equality stems from (H.1) and Remark 1 with $C_{\dot{u}} = \sup_{r\in K\subset\Theta}|\dot{u}(r)|$.

Hence,

$$\begin{aligned}&\|\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))-\dot{\psi}(u(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))\|\\ \leq& C_{\dot{u}}\frac{[1+\frac{4\kappa(\mathcal{J}_X^{-1})\sqrt{\log(1/\epsilon)}}{\sqrt{r}}]\sqrt{p}\sigma_{\max}(\boldsymbol{X})}{\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T)}\|[\boldsymbol{y}-\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|. \qquad \text{(S1.16)}\end{aligned}$$

Then by following the facts that

$$\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))\leq\sigma_{\max}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T)$$

and $\sigma^2_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}^T)=\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))\geq 0.5\sigma^2_{\min}(\boldsymbol{X})$,

it holds that

$$\sigma_{\min}(\dot{u}^2(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T)\geq 0.5\sigma^2_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})/\sigma_{\max}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}). \quad \text{(S1.17)}$$

Combing the result (S1.17) with (S1.16), the desired result holds

$$\begin{aligned}&\|\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))-\dot{\psi}(u(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))\|\\ &\leq 2C_{\dot{u}}[1+\frac{4\alpha\sqrt{\log(1/\epsilon)}}{\sqrt{r}}]\sqrt{p}\kappa^2(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})\|[\boldsymbol{y}-\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}))]\|. \qquad \text{(S1.18)}\end{aligned}$$

Now, we turn to prove Theorem 6.

Note that

$$p_i\geq\alpha\frac{\min_j|y_j-\dot{\psi}(u(\hat{\boldsymbol{\beta}}^T_{\mathrm{MLE}}\boldsymbol{x}_j))||\dot{u}(\hat{\boldsymbol{\beta}}^T_{\mathrm{MLE}}\boldsymbol{x}_j)|}{\sqrt{\sum_j|y_j-\dot{\psi}(u(\hat{\boldsymbol{\beta}}^T_{\mathrm{MLE}}\boldsymbol{x}_j))|^2|\dot{u}^2(\hat{\boldsymbol{\beta}}^T_{\mathrm{MLE}}\boldsymbol{x}_j)|}}\frac{\|\boldsymbol{x}_i\|}{\sum_j\|\boldsymbol{x}_j\|}=\delta\frac{\|\boldsymbol{x}_i\|}{\sum_j\|\boldsymbol{x}_j\|},$$

with some $0<\delta:=\frac{\alpha\gamma}{\sqrt{\sum_j|y_j-\dot{\psi}(u(\hat{\boldsymbol{\beta}}^T_{\mathrm{MLE}}\boldsymbol{x}_j))|^2|\dot{u}^2(\hat{\boldsymbol{\beta}}^T_{\mathrm{MLE}}\boldsymbol{x}_j)|}}\leq 1$.

According to the Weyl inequality, we have

$$
\begin{aligned}
&|\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^T\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}) - \sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))| \\
\leq &\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^T\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})X - \dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\|_S \\
\leq &\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\|_S\|(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^TSS^T\boldsymbol{X})\|_S\|\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\|_S \\
\leq &c_dC_{\dot{u}}^2\|(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^TSS^T\boldsymbol{X})\|_F \\
\leq &c_d\frac{4\sqrt{\log(1/\epsilon)}C_{\dot{u}}^2}{\delta\sqrt{r}}\|\boldsymbol{X}\|_F^2 \\
\leq &c_d\frac{4\sqrt{\log(1/\epsilon)}C_{\dot{u}}^2}{\delta\sqrt{r}}p\sigma_{\max}^2(\boldsymbol{X}).
\end{aligned}
$$

Using the above inequality, if we set

$$
r > 64c_d^2C_{\dot{u}}^2\log(1/\epsilon)\sigma_{\max}^4(\boldsymbol{X})p^2/(\delta^2\sigma_{\min}^4(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})),
$$

it holds that

$$
\begin{aligned}
&|\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})) - \sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))| \\
\leq &0.5\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})).
\end{aligned}
$$

Thus the following equation holds with probability at least $1-\varepsilon$:

$$
\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}^TSS^T\boldsymbol{X}\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})) \geq 0.5\sigma_{\min}^2(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}).
$$

□

S1.5 Proof of Theorem 7

For the average weighted log-likelihood in Step 2 of two-step algorithm, we have

$$L^{\text{two-step}}_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) := \frac{1}{r+r_0}\sum_{i=1}^{r+r_0}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} = \frac{1}{r+r_0}[\sum_{i=1}^{r_0}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} + \sum_{i=r_0+1}^{r+r_0}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}]$$
$$= \frac{r_0}{r+r_0}\cdot\frac{1}{r_0}\sum_{i=1}^{r_0}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} + \frac{r}{r+r_0}\cdot\frac{1}{r}\sum_{i=r_0+1}^{r+r_0}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)},$$

where $\pi_i^*(\tilde{\boldsymbol{\beta}}_0)$ in the first item stands for the initial subsampling strategy which satisfies (H.5).

For the sake of brevity, we begin with the case with probability $\boldsymbol{\pi}^{\text{mV}}$. Denote the log-likelihood in the first and second steps by

$$L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) = \frac{1}{r_0}\sum_{i=1}^{r_0}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}, \quad \text{and} \quad L^{*}_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) = \frac{1}{r}\sum_{i=1}^{r}\frac{t_i^*(\boldsymbol{\beta})}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)},$$

respectively, where $\pi_i(\tilde{\boldsymbol{\beta}}_0) = \check{\pi}_i^{\text{opt}}$ in $L^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$, and it has been calculated in the two-step algorithm in Section 4.

To proof of Theorem 7, we begin with the following Lemma 3.

Lemma 3. *If Assumptions (H.1)–(H.4) holds, then as $n \to \infty$, conditionally on $\mathcal{F}_n$ in probability,*

$$\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0} - \mathcal{J}_X = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S1.19}$$

$$\frac{1}{n}\frac{\partial L^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{S1.20}$$

where

$$\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0} = -\frac{1}{n}\frac{\partial^2 L^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{1}{nr}\sum_{i=1}^{r}\frac{\ddot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}^*{}_i))\dot{u}(\boldsymbol{\beta}^T\boldsymbol{x}^*{}_i)\boldsymbol{x}_i^*[\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*]^T}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} + \frac{1}{nr}\sum_{i=1}^{r}\frac{\ddot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*\boldsymbol{x}^*{}_i^T[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*)) - y_i^*]}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}.$$

Proof. Using the same arguments in Lemma 1, we have

$$\mathrm{E}\left(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0,j_1j_2} - \mathcal{J}_X^{j_1j_2}\Big|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0\right)^2 \le \frac{O_P(1)}{r}\Bigg[\sum_{i=1}^{n}\frac{\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)(x_{ij_1}x_{ij_2})^2}{n^2\pi_i(\tilde{\boldsymbol{\beta}}_0)} + \sum_{i=1}^{n}\frac{\{\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)) - y_i]\}^2}{n^2\pi_i(\tilde{\boldsymbol{\beta}}_0)}\Bigg]. \tag{S1.21}$$

Now we substitute expression of $\pi_i(\tilde{\boldsymbol{\beta}}_0)$ in the two-step algorithm: $\tilde{\boldsymbol{\pi}}^{\mathrm{mV}}$ and $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$. Here we only give the proof of the case $\tilde{\boldsymbol{\pi}}^{\mathrm{mV}}$, and the proof of the case $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$ is analogous thus we omit it. For the first terms in (S1.21), note that $\sigma_{\max}(\tilde{\mathcal{J}}_X^{-1}), \sigma_{\min}(\tilde{\mathcal{J}}_X^{-1})$ are bounded from Lemma 1 and (H.4), it implies

$$\begin{aligned}
&\sum_{i=1}^{n}\frac{\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)(x_{ij_1}x_{ij_2})^2}{n^2\pi_i(\tilde{\boldsymbol{\beta}}_0)}\\
&\le \sum_{i=1}^{n}\frac{\left\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|^2\sum_{j=1}^{n}\max(|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|,\delta)\left\|\mathcal{J}_X^{-1}\dot{u}(\tilde{\boldsymbol{\beta}}_0^Tx_i)x_i\right\|}{n^2\max(|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|,\delta)\left\|\mathcal{J}_X^{-1}\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|}\\
&\le \sum_{i=1}^{n}\frac{\left\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|^2\sum_{j=1}^{n}\max(|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|,\delta)\sigma_{\max}(\mathcal{J}_X^{-1})\left\|\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|}{n^2\delta\sigma_{\min}(\mathcal{J}_X^{-1})\left\|\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|}\\
&\le \kappa(\mathcal{J}_X^{-1})\sum_{i=1}^{n}\frac{\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|}{n^2\delta}\Bigg[\sum_{j=1}^{n}\frac{|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|\|\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_j\|}{n}
\end{aligned}$$

$$+\sum_{j=1}^{n}\frac{\delta\|\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_j\|}{n}\Big]$$

$$\leq \kappa(\mathcal{J}_X^{-1})\sum_{i=1}^{n}\frac{\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T x_i)x_i\|}{n\delta}\Bigg[\sqrt{\sum_{j=1}^{n}\frac{|y_j-\dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|^2}{n}}\sqrt{\sum_{j=1}^{n}\frac{\|\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_j\|^2}{n}}$$

$$+\sum_{j=1}^{n}\frac{\delta\|\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_j\|}{n}\Bigg]$$

$$\leq O_P(1)\kappa(\mathcal{J}_X^{-1})\sum_{i=1}^{n}\frac{\|\boldsymbol{x}_i\|}{n\delta}\left[\sqrt{\sum_{j=1}^{n}\frac{|y_j-\dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|^2}{n}}\sqrt{\sum_{j=1}^{n}\frac{\|\boldsymbol{x}_j\|^2}{n}}+\sum_{j=1}^{n}\frac{\delta\|\boldsymbol{x}_j\|}{n}\right]$$

$$=O_P(1).$$

where the last equality is from (H.3) and (H.5).

For the second terms in (S1.21), we have

$$\sum_{i=1}^{n}\frac{(\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)x_{ij_1}x_{ij_2}[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i])^2}{n^2\pi_i(\tilde{\boldsymbol{\beta}}_0)}$$

$$\leq\sum_{i=1}^{n}\frac{\left\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|^2\left|\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i\right|^2}{n^2\delta\left\|\mathcal{J}_X^{-1}\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|}$$

$$\times\sum_{j=1}^{n}(|y_j-\dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|+\delta)\left\|\mathcal{J}_X^{-1}\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j)\boldsymbol{x}_j\right\|$$

$$\leq\kappa(\mathcal{J}_X^{-1})\sum_{i=1}^{n}\Bigg[\frac{\left\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|\left|\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i\right|^2}{n\delta}$$

$$\times\frac{\sum_{j=1}^{n}(|y_j-\dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|+\delta)\left\|\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j)\boldsymbol{x}_j\right\|}{n}\Bigg]$$

$$=\kappa(\mathcal{J}_X^{-1})\sum_{i=1}^{n}\frac{\left\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|\left|\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i\right|^2}{n\delta}O_P(1)$$

$$\leq \frac{\kappa(\mathcal{J}_X^{-1})}{\delta}\sqrt{\sum_{i=1}^{n}\frac{\left\|\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\right\|^2}{n}}\sqrt{\sum_{i=1}^{n}\frac{\left|\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))-y_i\right|^4}{n}}O_P(1)$$

$$= O_P(1)$$

where the second last and last equality is from (H.3) and (H.5).

Direct calculation yields

$$E\Big(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0,j_1j_2}-\mathcal{J}_X^{j_1j_2}|\mathcal{F}_n\Big)^2 = E_{\tilde{\boldsymbol{\beta}}_0}E\Big(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0,j_1j_2}-\mathcal{J}_X^{j_1j_2}|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0\Big)^2 = O_P(r^{-1})$$

where $E_{\tilde{\boldsymbol{\beta}}_0}$ means that the expectation is taken with respect to the distribution of $\tilde{\boldsymbol{\beta}}_0$ given $\mathcal{F}_n$.

On the other hand, following the same arguments in Lemma 1, we can have

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

Then $E\left\{n^{-1}L^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})-n^{-1}L(\boldsymbol{\beta})|\mathcal{F}_n\right\}^2 = O_P(r^{-1})$.

Similarly, we can see that $\mathrm{Var}(n^{-1}\partial L^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})/\partial\boldsymbol{\beta}) = O_P(r^{-1})$. Thus, the desired result holds. □

Now we prove Theorem 7.

Proof. Using the same arguments in Lemma 3 , we have

$$E\left\{\frac{L_{\tilde{\boldsymbol{\beta}}_0}^{\text{two-step}}(\boldsymbol{\beta})}{n}-\frac{L(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2 \le 2(\frac{r_0}{r+r_0})^2 E\left\{\frac{L_{\tilde{\boldsymbol{\beta}}_0}^{*0}(\boldsymbol{\beta})}{n}-\frac{L(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2$$
$$+2(\frac{r}{r+r_0})^2 E\left\{\frac{L_{\tilde{\boldsymbol{\beta}}_0}^{*}(\boldsymbol{\beta})}{n}-\frac{L(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2 = O_P(r^{-1}).$$

Therefore $E\{n^{-1}L_{\tilde{\boldsymbol{\beta}}_0}^{\text{two-step}}(\boldsymbol{\beta})-n^{-1}L(\boldsymbol{\beta})|\mathcal{F}_n\}^2 \to 0$ as $r_0/r \to 0, r \to \infty$ and $n^{-1}L_{\tilde{\boldsymbol{\beta}}_0}^{\text{two-step}}(\boldsymbol{\beta}) - n^{-1}L(\boldsymbol{\beta}) \to 0$ in conditional probability given $\mathcal{F}_n$. Also note that the parameter space is compact and $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is the unique global maximum of the continuous convex function $L(\boldsymbol{\beta})$. Thus, from Theorem 5.9 and its remark of van der Vaart (1998), we have

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

Using Taylor's theorem,

$$\begin{aligned}0 &= \frac{\dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{\text{two-step}}(\breve{\boldsymbol{\beta}})}{n} = \frac{r_0}{r+r_0}\frac{\dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{*0}(\breve{\boldsymbol{\beta}})}{n} + \frac{r}{r+r_0}\frac{\dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{*}(\breve{\boldsymbol{\beta}})}{n}\\ &= \frac{r}{r+r_0}\left\{\frac{\dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{*}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + \frac{1}{n}\frac{\partial \dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{*}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}^T}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + \frac{1}{n}R_{\tilde{\boldsymbol{\beta}}_0,j}\right\}\\ &\quad + \frac{r_0}{r+r_0}\frac{\dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{*0}(\breve{\boldsymbol{\beta}})}{n},\end{aligned}$$

where $\dot{L}_{\tilde{\boldsymbol{\beta}}_0,j}^{*}(\boldsymbol{\beta})$ is the partial derivative of $L_{\tilde{\boldsymbol{\beta}}_0,j}^{*}(\boldsymbol{\beta})$ with respect to β_j.

By similar argument in the Proof of Theorem 1, the Lagrange remainder have the rate

$$\frac{1}{n}R_{\tilde{\boldsymbol{\beta}}_0,j} := \frac{1}{n}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})^T \int_0^1\int_0^1 \frac{\partial^2 \dot{L}_j^*\{\hat{\boldsymbol{\beta}}_{\text{MLE}} + uv(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} v du dv\ (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})$$

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

Note that the subsampling probabilities in the first stage satisfies the condition (H.1)-(H.7), thus from Theorem 2, it holds that

$$\frac{\dot{L}^{*0}_{\tilde{\boldsymbol{\beta}}_0,j}(\breve{\boldsymbol{\beta}})}{n} = \frac{\dot{L}^{*0}_{\tilde{\boldsymbol{\beta}}_0,j}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + \frac{1}{n}\frac{\partial \dot{L}^{*0}_{\tilde{\boldsymbol{\beta}}_0,j}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}^T}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + O_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|^2).$$

Therefore

$$\frac{1}{n}\frac{\partial L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\breve{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\frac{\partial L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} + \frac{1}{n}\frac{\partial^2 L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + O_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|^2).$$

From Lemma 1, it is clear to see that

$$\frac{1}{n}\frac{\partial L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r_0^{-1/2})$$

for the first step, since π_i^* is prespecified and satisfied (H.6), and

$$\frac{r_0}{r + r_0}\frac{1}{n}\frac{\partial L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}} = \frac{r_0}{r} O_{P|\mathcal{F}_n}(r_0^{-1/2}) = o_{P|\mathcal{F}_n}(r^{-1/2}),$$

since $r_0/r \to 0$. This step holds due to the fact that $\frac{\sqrt{r_0}}{r} O_{P|\mathcal{F}_n}(1) = \frac{\sqrt{r_0}}{\sqrt{r}} O_{P|\mathcal{F}_n}(1) O_{P|\mathcal{F}_n}(r^{-1/2}) = o(1) O_{P|\mathcal{F}_n}(r^{-1/2})$. Let

$$\breve{\mathcal{J}}_X := \frac{r}{r + r_0}\frac{1}{n}\frac{\partial^2 L^{*}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} + \frac{r_0}{r + r_0}\frac{1}{n}\frac{\partial^2 L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}.$$

Combine Lemmas 1 and 3, we have

$$\begin{aligned}
\breve{\mathcal{J}}_X - \mathcal{J}_X =& \frac{r}{r + r_0}\left(\breve{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0} - \mathcal{J}_X\right) + \frac{r_0}{r + r_0}\left(\frac{1}{n}\frac{\partial^2 L^{*0}_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} - \mathcal{J}_X\right) \\
=& \frac{r}{r + r_0} O_{P|\mathcal{F}_n}(r^{-1/2}) + \frac{r_0}{r + r_0} O_{P|\mathcal{F}_n}(r_0^{-1/2}) = O_{P|\mathcal{F}_n}(r^{-1/2}),
\end{aligned}$$

since $r_0/r \to 0$.

Hence,

$$\begin{aligned}\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} =& -(\breve{\mathcal{J}}_X)^{-1}\left\{\frac{1}{n}\dot{L}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2) + o_{P|\mathcal{F}_n}(r^{-1/2})\right\},\\ =& O_{P|\mathcal{F}_n}(r^{-1/2}) + o_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|)\end{aligned}$$

as $r_0/r \to 0$, by noting $(\breve{\mathcal{J}}_X)^{-1} = O_{P|\mathcal{F}_n}(1)$ from (H.5). Therefore, the desired result follows by noting

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

□

S1.6 Proof of Theorem 8

Proof. For the sake of brevity, we begin with the case with probability $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$. Denote

$$\frac{\dot{L}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = \frac{1}{r}\sum_{i=1}^{r}\frac{\{y_i^* - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*))\}\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*}{n\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} =: \frac{1}{r}\sum_{i=1}^{r}\eta_i^{\tilde{\boldsymbol{\beta}}_0}. \tag{S1.22}$$

It can be shown that given $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, $\eta_1^{\tilde{\boldsymbol{\beta}}_0}, \ldots, \eta_r^{\tilde{\boldsymbol{\beta}}_0}$ are i.i.d random variables with zero mean and variance

$$\mathrm{var}(\eta_i^{\tilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0) = rV_c^{\tilde{\boldsymbol{\beta}}_0} = \frac{1}{n^2}\sum_{i=1}^{n}\pi_i(\tilde{\boldsymbol{\beta}}_0)\frac{\{y_i^* - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*))\}^2\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i^2(\tilde{\boldsymbol{\beta}}_0)}.$$

Meanwhile, for every $\varepsilon > 0$,

$$
\begin{aligned}
&\sum_{i=1}^{r} E\{\|r^{-1/2}\eta_i^{\tilde{\boldsymbol{\beta}}_0}\|^2 I(\|\eta_i^{\tilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\varepsilon)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} \\
\leq &\frac{1}{r^{3/2}\varepsilon}\sum_{i=1}^{r} E\{\|\eta_i^{\tilde{\boldsymbol{\beta}}_0}\|^3 I(\|\eta_i^{\tilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\varepsilon)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} \\
\leq &\frac{1}{r^{3/2}\varepsilon}\sum_{i=1}^{r} E(\|\eta_i^{\tilde{\boldsymbol{\beta}}_0}\|^3|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0) \\
\leq &\frac{1}{r^{1/2}}\frac{1}{n^3}\sum_{i=1}^{n} \frac{\{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\}^3\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|^3}{\pi_i^2(\tilde{\boldsymbol{\beta}}_0)} \\
\leq &\frac{1}{r^{1/2}}\frac{1}{n}\sum_{i=1}^{n} \frac{\{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\}^2\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|}{\delta} \\
&\times \left(\frac{1}{n}\sum_{j=1}^{n} \max(|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))|, \delta)\|\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j)\boldsymbol{x}_j\|\right)^2 \\
\leq &\frac{1}{r^{1/2}}\frac{1}{n}\sum_{i=1}^{n} \frac{\{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\}^2\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|}{\delta} \\
&\times \left(\frac{1}{n}\sum_{j=1}^{n} (|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))| + \delta)\|\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j)\boldsymbol{x}_j\|\right)^2.
\end{aligned}
$$

From (H.1), (H.3) and (H.5),

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \frac{\{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\}^2\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|}{\delta} \\
\leq &\delta^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\}^4\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|^2\right)^{1/2} \\
= &O_P(1),
\end{aligned}
$$

by Holder's inequality.

Similarly, it can be shown

$$\frac{1}{n}\sum_{j=1}^{n}(|y_j - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j))| + \delta)\|\dot{u}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_j)\boldsymbol{x}_j\| = O_P(1),$$

from (H.1), (H.3) and (H.5).

Hence

$$\sum_{i=1}^{r} E\{\|r^{-1/2}\eta_i^{\tilde{\boldsymbol{\beta}}_0}\|^2 I(\|\eta_i^{\tilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\varepsilon)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} = o_{P|\mathcal{F}_n}(1).$$

This shows that the Lindeberg-Feller conditions are satisfied in probability. By the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart, 1998), conditionally on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$,

$$\frac{1}{n}(V_c^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) = \frac{1}{r^{1/2}}\{\mathrm{var}(\eta_i|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0)\}^{-1/2}\sum_{i=1}^{r}\eta_i \to N(0, I),$$

in distribution.

The distance between $V_c^{\tilde{\boldsymbol{\beta}}_0}$ and V_c is

$$\begin{aligned}
&\|V_c - V_c^{\tilde{\boldsymbol{\beta}}_0}\| \\
&\le \frac{1}{r}\sum_{i=1}^{n}\left\|\frac{1}{\pi_i^{\mathrm{mVc}}} - \frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\right\|\frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))\}^2\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\|\boldsymbol{x}_i\|^2}{n} \\
&= \frac{1}{r}\sum_{i=1}^{n}\left\|1 - \frac{\pi_i^{\mathrm{mVc}}}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\right\|\frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))\}^2\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\|\boldsymbol{x}_i\|^2}{n\pi_i^{\mathrm{mVc}}} \\
&\le \frac{1}{r}\sum_{i=1}^{n}\left\|1 - \frac{\pi_i^{\mathrm{mVc}}}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\right\|\frac{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|}{n} \\
&\le \frac{1}{r}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|1 - \frac{\pi_i^{\mathrm{mVc}}}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\right\|^2\right)^{1/2}\left(\sum_{i=1}^{n}\frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))\}^2\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\|\boldsymbol{x}_i\|^2}{n}\right)^{1/2} \\
&= o_{P|\mathcal{F}_n}(r^{-1}),
\end{aligned}$$

where the last equation follows from the facts that

$$\left\| 1 - \frac{\pi_i^{\mathrm{mVc}}}{\pi_i(\tilde{\boldsymbol{\beta}}_0)} \right\|^2 \leq \frac{(\pi_i^{\mathrm{mVc}} - \pi_i(\tilde{\boldsymbol{\beta}}_0))^2}{\delta^2} = o_P(1),$$

and

$$\sum_{i=1}^{n} \frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))\}^2 \dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i) \|\boldsymbol{x}_i\|^2}{n} = O_P(1).$$

Here the first equality in the fact above holds for the continues mapping theorem and the second equality holds from (H.1), (H.3), (H.5) and Cauchy's inequality.

Utilizing the facts

$$(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} - \breve{\mathcal{J}}_X^{-1} = -\breve{\mathcal{J}}_X^{-1}(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0} - \mathcal{J}_X + \mathcal{J}_X - \breve{\mathcal{J}}_X)(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2}),$$

we have $(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} - (\breve{\mathcal{J}}_X)^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2})$ from Lemma 3 and Theorem 7. Thus

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = -\frac{1}{n}(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} \dot{L}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1}) \tag{S1.23}$$

Based on Equation (S1.19), we further have

$$(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} - \mathcal{J}_X^{-1} = -\mathcal{J}_X^{-1}(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0} - \mathcal{J}_X)(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2}).$$

Therefore

$$\begin{aligned} & V^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) \\ =& -V^{-1/2}\frac{1}{n}(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} \dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}) \end{aligned}$$

$$= - V^{-1/2}\mathcal{J}_X^{-1}\frac{1}{n}\dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) - V^{-1/2}\{(\check{\mathcal{J}}_X^{\tilde{\boldsymbol{\beta}}_0})^{-1} - \mathcal{J}_X^{-1}\}\frac{1}{n}\dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2})$$
$$= - V^{-1/2}\mathcal{J}_X^{-1}(V_c^{\tilde{\boldsymbol{\beta}}_0})^{1/2}(V_c^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\frac{1}{n}\dot{L}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + O_{P|\mathcal{F}_n}(r^{-1/2}).$$

It can be shown that

$$\begin{aligned}
&V^{-1/2}\mathcal{J}_X^{-1}(V_c^{\tilde{\boldsymbol{\beta}}_0})^{1/2}(V^{-1/2}\mathcal{J}_X^{-1}(V_c^{\tilde{\boldsymbol{\beta}}_0})^{1/2})^T \\
=&V^{-1/2}\mathcal{J}_X^{-1}(V_c^{\tilde{\boldsymbol{\beta}}_0})\mathcal{J}_X^{-1}V^{-1/2} \\
=&V^{-1/2}\mathcal{J}_X^{-1}(V_c)\mathcal{J}_X^{-1}V^{-1/2} + o_{P|\mathcal{F}_n}(r^{-1/2}) \\
=&I + o_{P|\mathcal{F}_n}(r^{-1/2}).
\end{aligned}$$

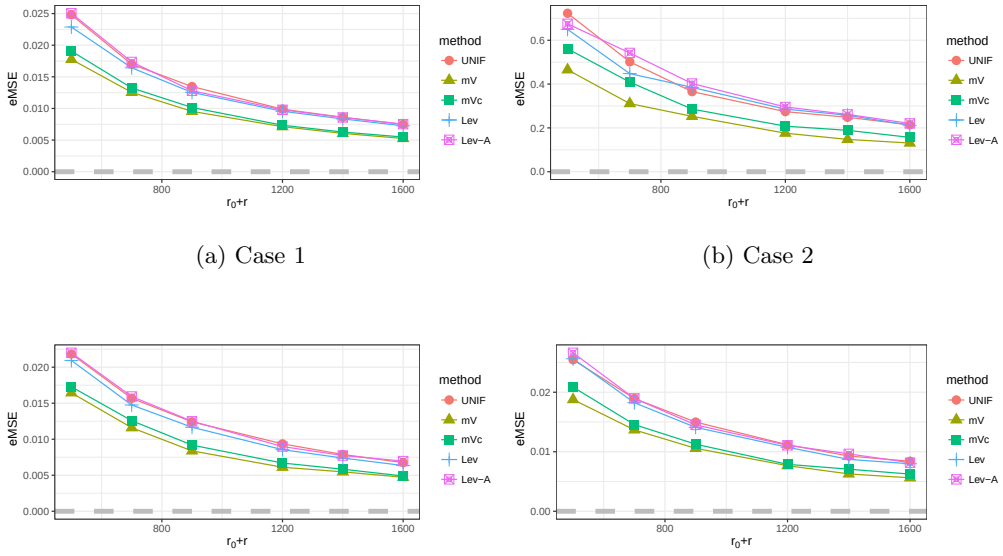
The desired result follows by Slutsky's theorem.

As for the case $\pi_i(\tilde{\boldsymbol{\beta}}_0) = \tilde{\pi}_i^{\mathrm{mV}}$ in $L^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$, $\tilde{\pi}_i^{\mathrm{mV}}$ has the same expression as π_i^{mV} except that $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$, is replaced by $\tilde{\boldsymbol{\beta}}_0$. Also note that $\pi_i(\tilde{\boldsymbol{\beta}}_0) \geq \kappa(\tilde{\mathcal{J}}_X)^{-1}\tilde{\pi}_i^{\mathrm{mVc}}$. The rest of the proof is the same as that of $\tilde{\pi}_i^{\mathrm{mVc}}$ with minor modifications. □

S2 Additional Simulation Results

In terms of the allocation between r_0 and r, it is clear to see that the two-step approach works the best when r_0/r is around 0.2 from the simulation result in Figure 3 of the main text. To well demonstrate our methods, we compare different $r_0 + r$ with fixed $r_0/r = 0.2$.

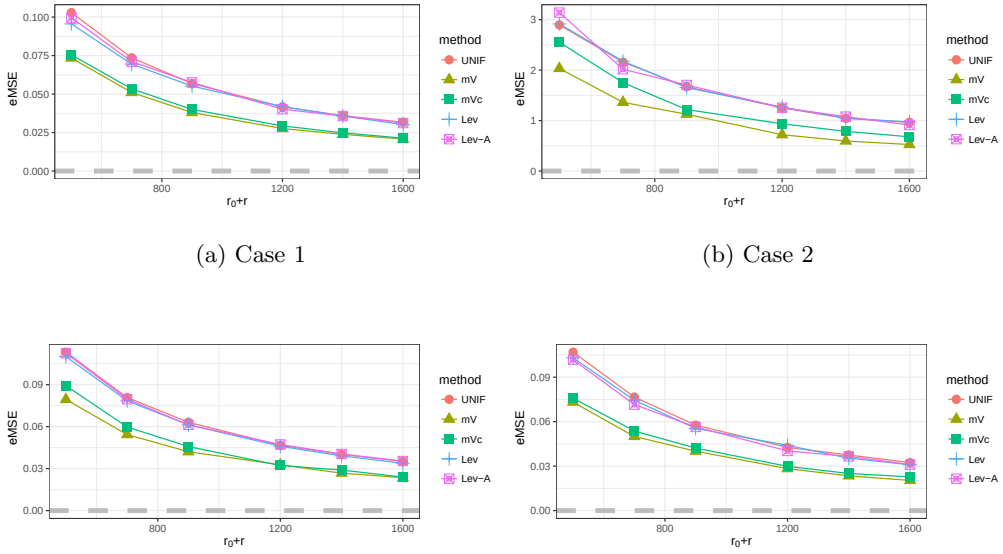
In each of the settings described in Section 5.1 of the article, we reevaluated the performance of $\tilde{\pi}_i^{\mathrm{mV}}$ and $\tilde{\pi}_i^{\mathrm{mVc}}$ when r_0/r is fixed at 0.2. For comparison, the uniform subsampling, leverage subsampling and adjusted leverage subsampling methods are also considered. Inline with the setting in the main text, the sample size $r_0 + r$ is selected as 500, 700, 900, 1200, 1400, and 1600. We report the results for the Poisson regression and the negative binomial regression in Figures S1 and S2, respectively.



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Figure S1: The eMSEs for the Poisson regression with different subsample size $r_0 + r$ and fixed $r_0/r = 0.2$. The distributions of the covariates are listed at the beginning of Section 5.



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure S2: The eMSEs for the NBR with different subsample size $r_0 + r$ and fixed $r_0/r = 0.2$. The distributions of the covariates are listed at the beginning of Section 5.

From Figures S1 and S2, we can see that our methods are slightly better than the cases that r_0 is fixed at 200. However, this improvement is not significant.

To explore influential factors on subsample sizes that have been discussed in Section 3.2 in terms of estimation accuracy, we consider additional four cases for the Poisson regression models listed as below.

Case S1: The true value of $\boldsymbol{\beta}$ is a 7×1 vector of 0.5 and the covariates matrix $X = \Sigma_n^{-1/2}\tilde{X}$. Here $\tilde{X}$ is the centralized version of a $n \times 7$ matrix whose elements are i.i.d., generated from $U([-1, 1])$, and Σ_n is the sample covariance matrix of $\tilde{X}$ so that X has a sample covariance matrix as I_p and a condition number as 1.

Case S2: The true value of $\boldsymbol{\beta}$ is a 14×1 vector whose first seven elements are set to be 0.5 and rest are set to be 0.1. The covariates matrix $X = \Sigma_n^{-1/2}\tilde{X}$, where $\tilde{X}$ is the centralized version of a $n \times 14$ matrix whose elements are i.i.d. generated from $U([-1, 1])$ and Σ_n is the sample covariance matrix of $\tilde{X}$ so that the condition number of X is 1 and the signal to noise ratio is nearly the same as that in Case S1.

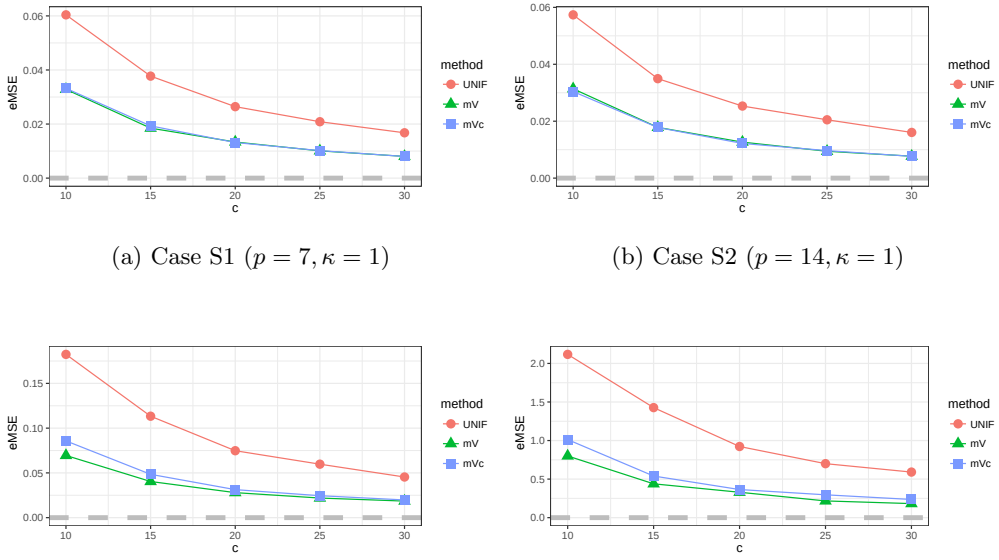
Case S3: This case is the same as the Case S2 except that x_{i2} in Case S2 is replaced with $x_{i2} = x_{i1} + \varepsilon_i$ where $\varepsilon_i \overset{\text{i.i.d}}{\sim} U([-0.4, 0.4])$ for

$i = 1, \ldots, n$. For this setup, the condition number of X is around 5.

Case S4: This case is the same as the Case S2 except that x_{i2} in Case S2 is replaced with $x_{i2} = x_{i1} + \varepsilon_i$ where $\varepsilon_i \overset{\text{i.i.d}}{\sim} U([-0.1, 0.1])$ for $i = 1, \ldots, n$. For this setup, the condition number of X is around 26.

To exclude the pilot subsampling effect, the ideal case that $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is given before conducting the subsampling strategy is considered. Although this setting is hard to satisfy, the simulation provides some key insights for Theorem 5 and it is also valuable for the two step Algorithm. The sample size r is selected as 10, 15, 20, 25 and 30 times of the dimension respectively. For comparison, the uniform subsampling method is also demonstrated. The eMSEs are reported in Figures S3.

Through the simulation results reported in Figures S3(a) and S3(b), we can see that the cases with $r = 10p,\ 20p$ exhibit similar performance when the conditional numbers of the covariate matrix are fixed at one. And for the same dimensional case, the eMSEs become larger as the conditional number of the covariate matrix increasing. These echo the results discussed in Section 3.2.



(a) Case S1 ($p = 7, \kappa = 1$)

(b) Case S2 ($p = 14, \kappa = 1$)

(c) Case S3 ($p = 14, \kappa \approx 5$)

(d) Case S4 ($p = 14, \kappa \approx 26$)

Figure S3: The eMSEs for Poisson regression with different subsample size $r = cp$. The different distributions of covariates are listed in the beginning of Section S2.

References

Brown, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics, Hayward, California: Lecture Notes-Monograph Series, vol. 9.

Drineas, P., R. Kannan, and M. W. Mahoney (2006). Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing 36*, 132–157.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2006). Sampling algorithms for l_2 regression and applications. *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, 1127–1136.

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

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