A MAXIMIN OPTIMAL APPROACH FOR SAMPLING DESIGNS IN TWO-PHASE STUDIES

Ruoyu Wang[1], Qihua Wang[2*] and Wang Miao[3]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health*

[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

and [3] *Department of Probability and Statistics, Peking University*

Supplementary Material

The supplementary material is organized as the follows. In Section S1, we provide some examples about efficient influence functions and the challenge in two-phase sampling design with multi-dimensional parameters. Section S2 contains the regularity conditions and proofs. Estimation under two-phase designs is discussed in Section S3. Some additional simulation results are reported in Section S4.

S1 Examples

S1.1 Examples of full data efficient influence functions

We introduce some examples of full data efficient influence functions for illustration.

Example S1. Let Y be a vector of outcomes which is hard to obtain. Suppose the parameter of interest is the outcome mean $\theta_0 = E[Y]$. Let Z be a vector of inexpensive covariates that is predictive to Y and hence useful in estimating θ_0. In two-phase studies, one can collect $V = Z$ in the first phase and measure $U = Y$ for a subset of subjects in the second

* Corresponding author

phase. In this example, the full data efficient influence function is $\psi = Y - \theta_0$.

Example S2. Let Y be a scalar outcome which is easy to obtain, Z a vector of inexpensive covariates, and X a vector of expensive covariates. Suppose the parameter of interest is the least squares regression coefficient θ_0 of X in the regression of Y on Z, X, which is determined by the estimating equation $E[(X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}(Y - X^{\mathrm{T}}\theta_0 - Z^{\mathrm{T}}\beta_0)] = 0$ where β_0 is the nuisance parameter. In two-phase studies, $V = (Y, Z)$ is collected in the first phase and $U = X$ is measured for a subset of subjects in the second phase. In this case, the full data efficient influence function is $\psi = (E[(X - \alpha_0 Z)(X - \alpha_0 Z)^{\mathrm{T}}])^{-1}(X - \alpha_0 Z)(Y - X^{\mathrm{T}}\theta_0 - Z^{\mathrm{T}}\beta_0)$ where $\alpha_0 = E[XZ^{\mathrm{T}}](E[ZZ^{\mathrm{T}}])^{-1}$ is the population linear regression coefficient of X on Z.

Example S3. Let $T \in \{0, 1\}$ be a binary treatment indicator, and Y the outcome. Suppose the parameter of interest is the average treatment effect, i.e., $\theta_0 = E[Y_1 - Y_0]$, where Y_1 and Y_0 are the potential outcomes under treatments "1" and "0", respectively. In observational studies, one needs to properly adjust for confounders to estimate θ_0 consistently. In practice, some confounders X may be hard to measure, while Y, T, and other confounders Z can be easily accessible. Then, a two-phase study can be conducted, where $V = (Y, T, Z)$ is collected in the first phase, and $U = X$ is measured for a subset of subjects in the second phase. Under the unconfoundness condition $(Y_1, Y_0) \perp\!\!\!\perp T \mid (X, Z)$, the full data efficient influence function is

$$\psi = \frac{TY}{\pi(X,Z)} - \frac{(1-T)Y}{1-\pi(X,Z)} - \left\{\frac{T}{\pi(X,Z)} - 1\right\} m_1(X,Z) + \left\{\frac{1-T}{1-\pi(X,Z)} - 1\right\} m_0(X,Z) - \theta_0,$$

where $\pi(x, z) = P(T = 1 \mid X = x, Z = z)$ is the propensity score, and $m_t(x, z) = E[Y \mid X = x, Z = z, T = t]$ is the outcome regression function for $t = 0, 1$.

The outcome mean estimation in Example S1 is an important problem in survey sampling (Cochran, 2007) and epidemiological studies (McNamee, 2002; Gilbert et al., 2014). Regression problems with expensive covariates in Example S2 are of great interest in modern epidemiological and clinical studies (Zeng and Lin, 2014; Zhou et al., 2014; Tao et al., 2017), because the determination of a disease's risk factor can often boil down to such a regression problem. Example S3 is of practical importance in observational studies (Yang and Ding, 2019). Previous works, e.g., Lin and Chen (2014) and Yang and Ding (2019), focus on the estimation in Example S3 without exploring the sampling rule design. We contribute by establishing the optimal sampling rule for a wide range of problems including Example S3.

S1.2 Example for the issue with a multi-dimensional parameter

Example S4. Suppose $Y \in \{0,1\}$ is an indicator of some disease status and $X \in \{0,1\}$ is the test result of some fallible test for disease status. Suppose $V = X$ and $U = Y$. The prevalence of the disease $\theta_{01} = P(Y = 1)$, sensitivity $\theta_{02} = P(X = 1 \mid Y = 1)$ and specificity $\theta_{03} = P(X = 0 \mid Y = 0)$ of the test are often of primary interest in epidemiological studies. Let $\theta_0 = (\theta_{01}, \theta_{02}, \theta_{03})^{\mathrm{T}}$ be the parameter of interest. It is not hard to show the efficient influence functions of θ_{01}, θ_{02} and θ_{03} are $Y - \theta_{01}$, $\theta_{01}^{-1}(X - \theta_{02})Y$ and $(1 - \theta_{01})^{-1}(1 - X - \theta_{03})(1 - Y)$, respectively. Let $P(X) = P(Y = 1 \mid X)$. The conditional variances $\sigma_1^2(V)$, $\sigma_2^2(V)$ and $\sigma_3^2(V)$ are $P(X)(1 - P(X))$, $\theta_{01}^{-2}P(X)(1 - P(X))(X - \theta_{02})^2$ and $(1 - \theta_{01})^{-2}P(X)(1 - P(X))(1 - X - \theta_{03})^2$, which are different from each other. According to Theorem 1, the optimal sampling rule for θ_{0j} is determined by $\sigma_j^2(\cdot)$. This implies that the optimal sampling rules for different parameters are different from each other. Hence, there is no sampling rule that minimizes the semiparametric efficiency bound for different parameters simultaneously in general.

Suppose $\theta_{01} = 0.2$, $\theta_{02} = 0.8$, $\theta_{03} = 0.6$ and $\varpi = 0.3$. Then some numerical calculations can show that the semiparametric efficiency bound for θ_{03} under $\rho_{\mathrm{sum}}(\cdot; \tau_{\mathrm{sum}})$ and the optimal sampling rule for θ_{01} are approximately 0.35 and 0.43, which are both larger than that under the uniform rule (≈ 0.30).

S2 Technical Details

S2.1 Regularity Conditions

Let F_0 be the distribution of (V, U). We consider the case where the parameter of interest is a general functional of F_0. Throughout this paper, we assume $\rho(\cdot)$ is bounded away from zero and $E[\|\psi\|^2] < \infty$ where $\|\cdot\|$ denotes the Euclidean norm.

As in Newey (1994), we consider inference of a *pathwise differentiable* parameter within a *locally nonparametric* distribution class. Here we briefly review the definitions of "pathwise differentiable" and "locally nonparametric". See Bickel (1982); van der Laan and Robins (2012); Tsiatis (2007) for more background on semiparametric theory. Let $\mathcal{F}$ be a set of joint distributions of (V, U) whose specific definition depends on the problem we consider. Suppose $F_0 \in \mathcal{F}$. A class of distributions $\{F_t : t \in [-1, 1]\}$ is called a submodel of $\mathcal{F}$ if it is contained in $\mathcal{F}$ and the distribution F_t equals to

F_0 when $t = 0$. Suppose F_t has a density $f_t(v, u)$ and let $S(v, u) = d \log f_t(v, u)/dt\big|_{t=0}$ be the score function under the submodel. Suppose the parameter $\theta_0 = \theta(F_0)$ is a functional of F_0 where $\theta(\cdot)$ is a functional defined on $\mathcal{F}$. Then the parameter is pathwise differentiable if there is some function $\phi(V, U)$ with zero mean and finite variance such that $d\theta(F_t)/dt\big|_{t=0} = E[\phi(V, U)S(V, U)]$ for any regular submodel.

Pathwise differentiability is a commonly used regularity condition in semiparametric theory (Bickel, 1982). Here, a regular submodel is a submodel that satisfies certain regularity conditions. See Bickel (1982) for more discussions and the formal definition of a regular submodel. Typical examples of pathwise differentiable parameters including the mean or quantile of a variable, the solution of many commonly used estimating equations among lots of other parameters.

"Locally nonparametric" is a property of the distribution class $\mathcal{F}$. Because all the submodels are required to belong to $\mathcal{F}$, the fewer the restrictions on $\mathcal{F}$, the more submodels, and hence the larger the set of score functions. Here, "locally nonparametric" requires $\mathcal{F}$ to be "general" or "unrestricted" in the sense that the set of score functions can approximate any function of (V, U) with zero mean and finite variance. In a locally nonparametric distribution class, general misspecification is allowed and few restrictions are imposed except for regularity conditions (Newey, 1994). For example, the distribution class which consists of all the distributions with a finite second moment is a locally nonparametric distribution class. For a missing data problem, all the observation distributions with response missing at random also consists of a locally nonparametric class.

S2.2 Proof of Lemma 1

This lemma can be obtained utilizing the techniques in the semiparametric theory for missing data problems (Tsiatis, 2007). To be self-contained, we provide its proof here.

Proof. We show the efficient influence function is

$$h = \frac{R\psi}{\rho(V)} - \left(\frac{R}{\rho(V)} - 1\right)\Pi(V)$$

and the semiparametric efficiency bound follows by straightforward calculation. The observed likelihood of (U, V, R) is

$$f(u \mid v)^r f(v)\rho(v)^r(1 - \rho(v))^{1-r},$$

where $f(v)$ is the density of V and $f(u \mid v)$ is the distribution of U conditional on $V = v$. For any regular submodel $f_t(u \mid v) f_t(v) \rho(v)^r (1 - \rho(v))^{1-r}$ whose distribution is denoted by F_t, the score function is

$$rS(u \mid v) + S(v), \tag{S1}$$

where

$$S(u \mid v) = \frac{d}{dt} \log f_t(u \mid v),$$

and

$$S(v) = \frac{d}{dt} \log f_t(v).$$

We do not consider a submodel for $\rho(v)$ since the sampling rule is determined by the researcher and hence is known in this problem. Because ψ is the full data influence function and $E[S(U \mid V) \mid V] = 0$, we have

$$\begin{aligned} \frac{d\theta(F_t)}{dt} &= E[\psi S(U \mid V)] + E[\psi S(V))] \\ &= E\left[hRS(U \mid V)\right] + E\left[hS(V)\right] \\ &= E\left[h\{RS(U \mid V) + S(V)\}\right]. \end{aligned} \tag{S2}$$

According to (S1), the tangent space under the two-phase design consists of all functions of the form $rS(u \mid v) + S(v)$, where $S(u \mid v)$ and $S(v)$ are the score function of $f(u \mid v)$ and $f(v)$ under some full data submodel. Since the full data model is locally nonparametric, the closure of the tangent space under the two-phase design consists of all score functions of the form (S1), which is

$$\mathcal{T} = \{rs(u, v) + s(v) \ : \ E[s(U, V) \mid V] = 0, \ E[s(V)] = 0\}.$$

It is easy to verify that h belongs to $\mathcal{T}$. This and (S2) implies h is the efficient influence function according to the characterization of the efficient influence function which can be found behind Lemma 25.14 in van der Vaart (1998). □

S2.3 Proof of Theorem 1

Proof. Recall that $\rho_{\mathcal{S}}(\cdot) = \rho(\cdot; \sigma, \tau_{\mathcal{S}})$. By the definition of $\tau_{\mathcal{S}}$, the sampling rule $\rho_{\mathcal{S}}(\cdot)$ satisfies the constraint $E[\rho_{\mathcal{S}}(V)] = E[\rho(V; \sigma, \tau_{\mathcal{S}})] \leq \varpi$. Because the second term in the efficiency bound (3.2) is irrelevant to the sampling rule, to show $\rho(\cdot; \sigma, \tau_{\mathcal{S}})$ is the optimal sampling rule, it suffices to prove

$$E\left[\frac{\sigma^2(V)}{\rho^\star(V)}\right] \geq E\left[\frac{\sigma^2(V)}{\rho(V; \sigma, \tau_{\mathcal{S}})}\right]$$

for any sampling rule $\rho^\star(\cdot)$ satisfying $E[\rho^\star(V)] \le \varpi$. Note that

$$\begin{aligned} E\left[\frac{\sigma^2(V)}{\rho^\star(V)}\right] - E\left[\frac{\sigma^2(V)}{\rho(V;\sigma,\tau_{\mathcal{S}})}\right] &\ge E\left[\frac{\sigma^2(V)}{\rho^2(V;\sigma,\tau_{\mathcal{S}})}\left(\rho(V;\sigma,\tau_{\mathcal{S}}) - \rho^\star(V)\right)\right] \\ &= E\left[\sigma^2(V)\left(\rho(V;\sigma,\tau_{\mathcal{S}}) - \rho^\star(V)\right)1\{\sigma(V) > \tau_{\mathcal{S}}\}\right] \\ &\quad + \tau_{\mathcal{S}}^2 E\left[\left(\rho(V;\sigma,\tau_{\mathcal{S}}) - \rho^\star(V)\right)1\{\sigma(V) \le \tau_{\mathcal{S}}\}\right] \\ &\ge \tau_{\mathcal{S}}^2 E\left[\rho(V;\sigma,\tau_{\mathcal{S}}) - \rho^\star(V)\right] \\ &= \tau_{\mathcal{S}}^2(\varpi - E\left[\rho^\star(V)\right]) \ge 0, \end{aligned}$$

where the first inequality is because $1/z_1 - 1/z_2 \ge (z_2 - z_1)/z_2^2$ for any $z_1, z_2 > 0$. This completes the proof. □

S2.4 Proof of Theorem 2

Proof. Recall that problem (3.6) is

$$\max_{\rho\in\mathcal{P}_{\mathcal{G}}} \min_{j=1,\ldots,d} \left\{ b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)\right\}.$$

By Lemma 1.15 in Rigollet and Hütter (2015), for any $\rho(\cdot)$, we have

$$\begin{aligned} &\min_{j=1,\ldots,d} \left\{ b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)\right\} \\ &= \min_{w\in\mathcal{W}^\dagger} \left\{ \sum_{j=1}^d w_j b_j^{-1}\xi_j - E\left[\frac{\sum_{j=1}^d w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right]\right\}, \end{aligned}$$

where $\mathcal{W}^\dagger = \{w = (w_1,\ldots,w_d) : \sum_{j=1}^d w_j = 1,\ 0 \le w_j \le 1,\ \text{for } j = 1,\ldots,d\}$. Hence (3.6) is equivalent to

$$\max_{\rho\in\mathcal{P}_{\mathcal{G}}} \min_{w\in\mathcal{W}^\dagger} \left\{ \sum_{j=1}^d w_j b_j^{-1}\xi_j - E\left[\frac{\sum_{j=1}^d w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right]\right\}. \tag{S3}$$

Recall that $\mathcal{P}_{\mathcal{G}} := \{\rho(\cdot) : 0 \le \rho(\cdot) \le 1,\ E[\rho(V)] \le \varpi\}$ and $\mathcal{W}^\dagger = \{w = (w_1,\ldots,w_d) : \sum_{j=1}^d w_j = 1,\ 0 \le w_j \le 1,\ \text{for } j = 1,\ldots,d\}$. Let

$$h(\rho, w) = \sum_{j=1}^d w_j b_j^{-1}\xi_j - E\left[\frac{\sum_{j=1}^d w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right].$$

Take the L_2 norm and the Euclidean norm as the norm in $\mathcal{P}$ and $\mathcal{W}^\dagger$, respectively. Then, $\mathcal{P}$, $\mathcal{W}^\dagger$ are compact and $h(\rho, w)$ is continuous with respect to ρ and q. Moreover, $h(\rho, w)$ is convex with respect to ρ and linear (hence concave) w.r.t. w.

Thus, the solution of the optimization problem does not change if we change the order of max and min in (3.6) according to Theorem 3.4 in Sion (1958). Thus, the dual problem

$$\min_{w\in\mathcal{W}^\dagger}\max_{\rho\in\mathcal{P}_\mathcal{G}}\left\{\sum_{j=1}^{d} w_j b_j^{-1}\xi_j - E\left[\frac{\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right]\right\} \tag{S4}$$

shares the same solution as (S3), which also leads to an equivalent problem of (3.6).

According to the above derivations, we can focus on the problem (S4). Notice that the inner optimization problem of (S4)

$$\begin{aligned}&\max_{\rho\in\mathcal{P}_\mathcal{G}}\left\{\sum_{j=1}^{d} w_j b_j^{-1}\xi_j - E\left[\frac{\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right]\right\}\\ &=\sum_{j=1}^{d} w_j b_j^{-1}\xi_j - \min_{\rho\in\mathcal{P}_\mathcal{G}} E\left[\frac{\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right].\end{aligned}$$

Similar arguments to those in the proof of Theorem 1 can show that $\rho(\cdot;\sigma_w,\tau_w)$ minimizes the functional $E\left[\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(V)/\rho(V)\right]$ over $\mathcal{P}_\mathcal{G}$ and the minimum value is

$$\min_{\rho\in\mathcal{P}_\mathcal{G}} E\left[\frac{\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(V)}{\rho(V)}\right] = E\left[\sigma_w(V)\max\{\sigma_w(V),\tau_w\}\right],$$

where $\sigma_w(V)=\sqrt{\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(V)}$ and τ_w is the unique solution of $E[\rho(V;\sigma_w,\tau)]=\varpi$ with respect to τ. This completes the proof of Theorem 2. □

S2.5 Proof of Theorem 3

Proof. We prove the result for $\tilde{\rho}_j(\cdot)$ for $j=1,\ldots,d$. The result for $\tilde{\rho}_\mathcal{C}(\cdot)$ and $\tilde{\rho}_\mathcal{G}(\cdot)$ can be established similarly. For $i=1,\ldots,n$, the expectation of R_{2i} is $(1-R_{1i})\tilde{\rho}_j(V_i)$ conditional on $(R_{11},V_1),\ldots,(R_{1n},V_n)$ and U_j for j with $R_{1j}=1$. Thus conditional on the same variables, the expectation of $\sum_{i=1}^{n}(R_{1i}+R_{2i})$ is $\sum_{i=1}^{n} R_{1i}+\sum_{i=1}^{n}(1-R_{1i})\tilde{\rho}_j(V_i)$. Because $\tilde{\tau}_j$ is the solution of (4.11), we have $\sum_{i=1}^{n}(1-R_{1i})\tilde{\rho}_j(V_i)=(\varpi-\kappa_n)n$. According to the law of iterated conditional expectation, we have $E\left[\sum_{i=1}^{n}(R_{1i}+R_{2i})\right]=\kappa_n n+(\varpi-\kappa_n)n=\varpi n$ which proves Theorem 3. □

S2.6 Proof of Theorem 4

In this and the following proofs, we use M to denote generic positive constants whose values may be different in different places. We first get down to the required regularity conditions. Recall that τ_j is the solution of $E[\rho(V;\sigma_j,\tau)]=\varpi$ for $j=1,\ldots,d$.

Condition S1. There is some constants $r_j, L_j > 0$ such that $r_j < \tau_j$ and $|E[\rho(V;\sigma_j,\tau_1)] - E[\rho(V;\sigma_j,\tau_2)]| > L_j|\tau_1 - \tau_2|$ for any $\tau_1, \tau_2 \in [\tau_j - r_j, \tau_j + r_j]$ and $j = 1, \ldots, d$, where $\rho(\cdot;\sigma_j,\tau) = 1\{\sigma_j(V) > \tau\} + \sigma_j(V)/\tau 1\{\sigma_j(V) \le \tau\}$.

Condition S2. $\sup_v \Pi_j(v) < \infty$ and $0 < \inf_v \sigma_j(v) \le \sup_v \sigma_j(v) < \infty$ for $j = 1, \ldots, d$.

Condition S1 requires that the budgets under different thresholds are different in a neighborhood of τ_j. Condition S2 is a mild regularity condition. Next, we give the proof of Theorem 4.

Proof. We prove the results for $\tilde{\rho}_j(\cdot)$ for $j = 1, \ldots, d$. The result for $\tilde{\rho}_{\mathcal{S}}(\cdot)$ is a special case of $d = 1$.

We first show $\tilde{\tau}_j$ converges to τ_j for $j = 1, \ldots, d$, where $\tilde{\tau}_j$ is the solution of Equation (4.11) in the main text. Let $\tau_{j,n}$ be the solution of $E[\rho(V;\sigma_j,\tau)] = (\varpi - \kappa_n)/(1-\kappa_n)$. Note that $(\varpi - \kappa_n)/(1-\kappa_n) - \varpi = O(\kappa_n)$. Under Condition S1, we have $|\tau_{j,n} - \tau_j| = O(\kappa_n)$. Next, we show that $|\tilde{\tau}_j - \tau_{j,n}|$ converges to zero. For $\tau \in [\tau_j - r_j, \tau_j + r_j]$, define

$$h_{j,n}(\tau) = \frac{1}{n}\sum_{i=1}^{n}(1 - R_{1i})\left(1\{\sigma(V_i) \ge \tau\} + \frac{\sigma_j(V_i)}{\tau}1\{\sigma_j(V_i) < \tau\}\right)$$

and

$$\tilde{h}_{j,n}(\tau) = \frac{1}{n}\sum_{i=1}^{n}(1 - R_{1i})\left(1\{\tilde{\sigma}_j(V_i) \ge \tau\} + \frac{\tilde{\sigma}_j(V_i)}{\tau}1\{\tilde{\sigma}_j(V_i) < \tau\}\right).$$

By calculating the mean and variance, we have

$$|h_{j,n}(\tau) - E[h_{j,n}(\tau)]| = O_P\left(\frac{1}{\sqrt{n}}\right) \tag{S5}$$

uniformly in $\tau \in [\tau_j - r_j, \tau_j + r_j]$. Moreover, we have $E[h_{j,n}(\tau)] = (1-\kappa_n)E[\rho(V;\sigma_j,\tau)]$ which implies $E[h_{j,n}(\tau)] = \varpi - \kappa_n$. By Condition 1, it is not hard to verify

$$|\tilde{h}_{j,n}(\tau) - h_{j,n}(\tau)| \le \frac{1}{\tau_j - r}\|\tilde{\sigma} - \sigma\|_\infty = O_P\left\{(n\kappa_n)^{-\delta}\right\} \tag{S6}$$

uniformly in $\tau \in [\tau_j - r_j, \tau_j + r_j]$, where δ is a constant determined by the convergence rate of $\|\tilde{\sigma}_j - \sigma_j\|_\infty$ which appears in Condition 1 . Combining (S5) and (S6), we have

$$|\tilde{h}_{j,n}(\tau) - E[h_{j,n}(\tau)]| = O_P\left\{(n\kappa_n)^{-\delta}\right\}.$$

Thus for any $\epsilon > 0$, there is some constant $M > 0$ such that

$$P\left(|\tilde{h}_{j,n}(\tau) - E[h_{j,n}(\tau)]| \ge M(n\kappa_n)^{-\delta}\right) \le \frac{\epsilon}{2}$$

for any $\tau \in [\tau_j - r_j, \tau_j + r_j]$. Define $\tau_{j,n,1} = \tau_{j,n} - M(n\kappa_n)^{-\delta}/(L_j(1-\kappa_n))$ and $\tau_{j,n,2} = \tau_{j,n} + M(n\kappa_n)^{-\delta}/(L_j(1-\kappa_n))$. Because $|\tau_{j,n} - \tau_j| = O(\kappa_n)$, $\kappa_n \to 0$ and $n\kappa_n \to \infty$, we have $\tau_{j,n}$, $\tau_{j,n,1}$, $\tau_{j,n,2} \in [\tau_j - r_j, \tau_j + r_j]$ for sufficiently large n. Hence

$$\begin{aligned}
&P\Big(\tilde{h}_{j,n}(\tau_{j,n,1}) - E[h_{j,n}(\tau_{j,n,1})] < M(n\kappa_n)^{-\delta}, \tilde{h}_{j,n}(\tau_{j,n,2}) - E[h_{j,n}(\tau_{j,n,2})] > -M(n\kappa_n)^{-\delta}\Big) \\
&\geq 1 - P\left(\tilde{h}_{j,n}(\tau_{j,n,1}) - E[h_{j,n}(\tau_{j,n,1})] \geq M(n\kappa_n)^{-\delta}\right) \\
&\quad - P\left(\tilde{h}_{j,n}(\tau_{j,n,2}) - E[h_{j,n}(\tau_{j,n,2})] \leq -M(n\kappa_n)^{-\delta}\right) \\
&\geq 1 - \epsilon
\end{aligned} \tag{S7}$$

for sufficiently large n. According to Condition S1 and the monotonicity of $E[h_{j,n}(\tau)] = (1-\kappa_n)E[\rho(V;\sigma_j,\tau)]$, we have

$$E[h_{j,n}(\tau_{j,n,1})] \leq \varpi - \kappa_n - M(n\kappa_n)^{-\delta}$$

and

$$E[h_{j,n}(\tau_{j,n,2})] \geq \varpi - \kappa_n + M(n\kappa_n)^{-\delta}.$$

Combining this with (S7), we have

$$P\left(\tilde{h}_{j,n}(\tau_{j,n,1}) < \varpi - \kappa_n < \tilde{h}_{j,n}(\tau_{j,n,2})\right) \geq 1 - \epsilon.$$

Recall that $\tilde{\tau}_j$ is the solution of $\tilde{h}_{j,n}(\tau) = \varpi - \kappa_n$. We have $P\left(\tilde{\tau}_j \in [\tau_{j,n,1}, \tau_{j,n,2}]\right) \geq 1 - \epsilon$ by the monotonicity of $\tilde{h}_{j,n}(\tau)$. This implies $|\tilde{\tau}_j - \tau_{j,n}| = O_P\{(n\kappa_n)^{-\delta}\}$ and hence

$$|\tilde{\tau}_j - \tau_j| = O_P\{(n\kappa_n)^{-\delta} + \kappa_n\}. \tag{S8}$$

Under Condition S2, we have

$$\begin{aligned}
&|\tilde{\rho}_j(v) - \rho_j(v)| \\
&= \left|1\{\tilde{\sigma}_j(v) > \tilde{\tau}_j\} + \frac{\tilde{\sigma}_j(v)}{\tilde{\tau}_j}1\{\tilde{\sigma}_j(v) \leq \tilde{\tau}_j\} - 1\{\sigma_j(v) > \tau_j\} - \frac{\sigma_j(v)}{\tau_j}1\{\sigma_j(v) \leq \tau_j\}\right| \\
&\leq \left|1\{\tilde{\sigma}_j(v) > \tilde{\tau}_j\} + \frac{\tilde{\sigma}_j(v)}{\tilde{\tau}_j}1\{\tilde{\sigma}_j(v) \leq \tilde{\tau}_j\} - 1\{\sigma_j(v) > \tilde{\tau}_j\} - \frac{\sigma_j(v)}{\tilde{\tau}_j}1\{\sigma_j(v) \leq \tilde{\tau}_j\}\right| \\
&+ \left|1\{\sigma_j(v) > \tilde{\tau}_j\} + \frac{\sigma_j(v)}{\tilde{\tau}_j}1\{\sigma_j(v) \leq \tilde{\tau}_j\} - 1\{\sigma_j(v) > \tau_j\} - \frac{\sigma_j(v)}{\tau_j}1\{\sigma_j(v) \leq \tau_j\}\right| \\
&\leq M(\|\tilde{\sigma}_j - \sigma_j\|_\infty + |\tilde{\tau}_j - \tau_j|)
\end{aligned}$$

for some constant M with probability approaching one. Combining this with Condition 1 and Equation (S8), we have

$$\|\tilde{\rho}_j - \rho_j\|_\infty = O_P\left\{(n\kappa_n)^{-\delta} + \kappa_n\right\},$$

which completes the proof. □

S2.7 Proof of Theorem 5

In this subsection, we turn to the convergence result of the estimated sampling rule in the multi-dimensional parameter case. For $w \in \mathcal{W}$, define

$$H_j(w) = -b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho_0(V) + \sum_{j=1}^d w_j(\rho_j(V) - \rho_0(V_i))}\right]\right)$$

and

$$H_{\mathcal{C}}(w) = \max_{j=1,\ldots,d} H_j(w),$$

where $\xi_j = E[\sigma_j^2(V)/\rho_0(V)]$ and $b_j = E\left[\sigma_j^2(V)/\rho_0(V)\right] + \mathrm{Var}[\Pi_j(V)]$ for $j = 1, \ldots, d$. Similarly, for $w \in \mathcal{W}^\star$, let

$$H_{\mathcal{G}}(w) = \sum_{j=1}^d w_j b_j^{-1}\xi_j - E\left[\sigma_w(V)\max\{\sigma_w(V), \tau_w\}\right].$$

Then, $w_{\mathcal{C}} = \arg\min_{w\in\mathcal{W}} H_{\mathcal{C}}(w)$ and $w_{\mathcal{G}} = \arg\min_{w\in\mathcal{W}^\dagger} H_{\mathcal{G}}(w)$. The following condition is required to establish the convergence rate of $\tilde{\rho}_{\mathcal{C}}(\cdot)$ and $\tilde{\rho}_{\mathcal{G}}(\cdot)$.

Condition S3. The benchmark sampling rule $\rho_0(\cdot)$ is bounded away from zero and satisfies $E[\rho_0(V)] = \varpi$.

Condition S4. (i) $H_{\mathcal{C}}(w)$ has the unique minimum point $w_{\mathcal{C}}$; (ii) for some constants $r_{\mathcal{C}}, L_{\mathcal{C}} > 0$ and any $w \in \mathcal{W}$ such that $\|w - w_{\mathcal{C}}\| \le r_{\mathcal{C}}$, we have $H_{\mathcal{C}}(w) - H_{\mathcal{C}}(w_{\mathcal{C}}) \ge L_{\mathcal{C}}\|w - w_{\mathcal{C}}\|^2$.

Condition S5. (i) $H_{\mathcal{G}}(w)$ has the unique minimum point $w_{\mathcal{G}}$; (ii) for some constants $r_{\mathcal{G}}, L_{\mathcal{G}} > 0$ and any $w \in \mathcal{W}^\dagger$ such that that $\|w - w_{\mathcal{G}}\| \le r_{\mathcal{G}}$, we have $H_{\mathcal{G}}(w) - H_{\mathcal{G}}(w_{\mathcal{G}}) \ge L_{\mathcal{G}}\|w - w_{\mathcal{G}}\|^2$.

Condition S3 is a regularity condition on the benchmark sampling rule $\rho_0(\cdot)$. Condition S4 is a mild regularity condition, which can be satisfied if $H_{\mathcal{C}}(w)$ has a continuous Hessian matrix in a neighborhood of $r_{\mathcal{C}}$ and the Hessian matrix at $r_{\mathcal{C}}$ is positive definite. Condition S5 is similar to Condition S4 with $H_{\mathcal{C}}(w)$ replaced by $H_{\mathcal{G}}(w)$. Now, we are ready to prove Theorem 5.

Proof. We only prove the result for $\tilde{\rho}_{\mathcal{C}}(\cdot)$. The result for $\tilde{\rho}_{\mathcal{G}}(\cdot)$ can be established similarly.

Recall that $\rho_{\mathcal{C}}(\cdot) = \rho_0(\cdot) + \sum_{j=1}^d w_{\mathcal{C},j}\{\rho_j(\cdot) - \rho_0(\cdot)\}$ and $\tilde{\rho}_{\mathcal{C}}(\cdot) = \rho_0(\cdot) + \sum_{j=1}^d \widetilde{w}_{\mathcal{C},j}(\tilde{\rho}_j(\cdot) - \rho_0(\cdot))$, where $w_{\mathcal{C},j}$ and $\widetilde{w}_{\mathcal{C},j}$ are the jth component of $w_{\mathcal{C}}$ and $\widetilde{w}_{\mathcal{C}}$, respectively, for $j = 1, \ldots, d$. Recall that $\rho_j(v) = 1\{\sigma_j(v) > \tau_j\} + \sigma_j(v)1\{\sigma_j(v) \le \tau_j\}/\tau_j$ for $j = 1, \ldots, d$. By Condition S2, we have

$$\frac{1}{M} \le \inf_v \rho_j(v) \le \sup_v \rho_j(v) \le M \tag{S9}$$

for some $M > 1$. Hence

$$\|\tilde{\rho}_{\mathcal{C}} - \rho_{\mathcal{C}}\|_\infty \le \sum_{j=1}^d \|\tilde{\rho}_j - \rho_j\|_\infty + M\|\tilde{w}_{\mathcal{C}} - w_{\mathcal{C}}\|.$$

Theorem 4 has established the convergence rate of $\|\tilde{\rho}_j - \rho_j\|_\infty$. Thus, in order to establish the convergence rate of $\|\tilde{\rho}_{\mathcal{C}} - \rho_{\mathcal{C}}\|_\infty$, it suffices to establish the convergence rate of $\|\widetilde{w}_{\mathcal{C}} - w_{\mathcal{C}}\|$. Define

$$\widetilde{H}_j(w) = -\tilde{b}_j^{-1}\left(\tilde{\xi}_j - \frac{1}{n}\sum_{i=1}^n \frac{\tilde{\sigma}_j^2(V)}{\rho_0(V) + \sum_{j=1}^d w_j(\tilde{\rho}_j(V) - \rho_0(V_i))}\right)$$

and

$$\widetilde{H}_{\mathcal{C}}(w) = \max_{j=1,\ldots,d} \widetilde{H}_j(w),$$

where, for $j = 1, \ldots, d$, $\tilde{b}_j = n^{-1}\sum_{i=1}^n \tilde{\sigma}_j^2(V_i)/\rho_0(V_i) + n^{-1}\sum_{i=1}^n\{\widetilde{\Pi}_j(V_i) - n^{-1}\sum_{i=1}^n \widetilde{\Pi}_j(V_i)\}^2$ and $\tilde{\xi}_j = n^{-1}\sum_{i=1}^n \tilde{\sigma}_j^2(V_i)/\rho_0(V_i)$. Then $\widetilde{w}_{\mathcal{C}} = \arg\min_{w\in\mathcal{W}} \widetilde{H}_{\mathcal{C}}(w)$. Define $\bar{b}_j$ and $\bar{H}_j(w)$ similarly to $\tilde{b}_j$ and $\widetilde{H}_j(w)$ with $\widetilde{\Pi}_j(\cdot)$, $\tilde{\sigma}_j(\cdot)$ and $\tilde{\rho}_j(\cdot)$ in $\tilde{b}_j$ and $\widetilde{H}_j(w)$ replaced by $\Pi_j(\cdot)$, $\sigma_j(\cdot)$ and $\rho_j(\cdot)$. Let $\bar{H}_{\mathcal{C}}(w) = \max_{j=1,\ldots,d} \bar{H}_j(w)$. According to Condition 1, Theorem 4, and inequality (S9), there is some constant $M > 1$ such that $1/M \le \inf_v \tilde{\rho}_j(v) \le \sup_v \tilde{\rho}_j(v) \le M$ with probability approaching one. Conditions S2, S3 imply that $\max_{j=1,\ldots,d} |\tilde{b}_j - \bar{b}_j| \le M(\|\widetilde{\Pi}_j - \Pi_j\|_\infty + \|\tilde{\sigma}_j - \sigma_j\|_\infty + \|\tilde{\rho}_j - \rho_j\|_\infty)$ for some constant M. Then, Conditions S2, S3 and the mean value theorem implies that

$$\sup_{w\in\mathcal{W}} |\widetilde{H}_j(w) - \bar{H}_j(w)| \le M(\|\widetilde{\Pi}_j - \Pi_j\|_\infty + \|\tilde{\sigma}_j - \sigma_j\|_\infty + \|\tilde{\rho}_j - \rho_j\|_\infty)$$

with probability approaching one for some constant M. This implies

$$\sup_{w\in\mathcal{W}} |\widetilde{H}_{\mathcal{C}}(w) - \bar{H}_{\mathcal{C}}(w)| \le$$
$$M\left(\max_{j=1,\ldots,d} \|\widetilde{\Pi}_j - \Pi_j\|_\infty + \max_{j=1,\ldots,d} \|\tilde{\sigma}_j - \sigma_j\|_\infty + \max_{j=1,\ldots,d} \|\tilde{\rho}_j - \rho_j\|_\infty\right)$$

with probability approaching one. Then, according to Condition 1 and Theorem 4, we have

$$\sup_{w\in\mathcal{W}} |\widetilde{H}_{\mathcal{C}}(w) - \bar{H}_{\mathcal{C}}(w)| = O_P\left\{(n\kappa_n)^{-\delta} + \kappa_n\right\}. \tag{S10}$$

Recall that

$$\bar{b}_j = \left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sigma_j^2(V_i)}{\rho_0(V_i)} + \left\{\Pi_j(V_i) - \frac{1}{n}\sum_{i=1}^{n}\Pi_j(V_i)\right\}^2\right]\right)$$

$$\bar{H}_j(w) = \bar{b}_j^{-1}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sigma_j^2(V_i)}{\rho_0(V_i)} - \frac{\sigma_j^2(V_i)}{\rho_0(V_i) + \sum_{j=1}^{d} w_j(\rho_j(V_i) - \rho_0(V_i))}\right\}$$

for any $w \in \mathcal{W}$. According to Condition S2, we have $b_j \geq 1/M$ for some constant M and $j = 1, \ldots, d$. Notice that the function $\sigma_j(v)/\rho_0(v) + \sigma_j(v)/[\rho_0(v) + \sum_{j=1}^{d} w_j(\rho_j(v) - \rho_0(v))]$ is bounded and Lipschitz continuous with respect to w due to (S9), Condition S2, and Condition S3. Then, similar concentration arguments as in the proof of Lemma C.1 in He et al. (2021) can show that

$$\sup_{w\in\mathcal{W}} |\bar{H}_j(w) - H_j(w)| = O_P\left(\frac{1}{\sqrt{n}}\right). \tag{S11}$$

Combining (S10) with (S11), we have

$$\sup_{w\in\mathcal{W}} |\widetilde{H}_j(w) - H_j(w)| = O_P\left\{(n\kappa_n)^{-\delta} + \kappa_n\right\}.$$

Thus, for any $\epsilon > 0$, there is some $M > 0$ such that $P(\sup_{w\in\mathcal{W}} |\widetilde{H}_j(w) - H_j(w)| \geq M\{(n\kappa_n)^{-\delta} + \kappa_n\}) \leq \epsilon$. Because $\widetilde{w}_{\mathcal{C}}$ is the minimum point of $\widetilde{H}_{\mathcal{C}}(w)$, we have $\widetilde{H}_{\mathcal{C}}(\widetilde{w}_{\mathcal{C}}) \leq \widetilde{H}_{\mathcal{C}}(w_{\mathcal{C}})$ and hence

$$H_{\mathcal{C}}(\widetilde{w}_{\mathcal{C}}) - H_{\mathcal{C}}(w_{\mathcal{C}}) \leq \widetilde{H}_{\mathcal{C}}(\widetilde{w}_{\mathcal{C}}) - \widetilde{H}_{\mathcal{C}}(w_{\mathcal{C}}) + 2\sup_{w\in\mathcal{W}} |\widetilde{H}_j(w) - H_j(w)|$$

$$\leq 2\sup_{w\in\mathcal{W}} |\widetilde{H}_j(w) - H_j(w)|.$$

Thus

$$H_{\mathcal{C}}(\widetilde{w}_{\mathcal{C}}) - H_{\mathcal{C}}(w_{\mathcal{C}}) \leq 2M\{(n\kappa_n)^{-\delta} + \kappa_n\} \tag{S12}$$

with probability at least $1-\epsilon$. Because $H_{\mathcal{C}}(w)$ is continuous with respect to w, $\mathcal{W}$ is a compact set and $w_{\mathcal{C}}$ is the unique minimum point of $H_{\mathcal{C}}(w)$, we have $\inf_{w\in\mathcal{W},\|w-w_{\mathcal{C}}\|\geq r_{\mathcal{C}}}\{H_{\mathcal{C}}(w) - H_{\mathcal{C}}(w_{\mathcal{C}})\} > 0$ for $r_{\mathcal{C}}$ in Condition S4. For sufficiently large n such that $2M\{(n\kappa_n)^{-\delta} + \kappa_n\} < \inf_{w\in\mathcal{W},\|w-w_{\mathcal{C}}\|\geq r_{\mathcal{C}}}\{H_{\mathcal{C}}(w) - H_{\mathcal{C}}(w_{\mathcal{C}})\}$, we have $\|\widetilde{w}_{\mathcal{C}} - w_{\mathcal{C}}\| \leq r_{\mathcal{C}}$ with probability at least $1-\epsilon$ according to (S12). Then inequality (S12) and Condition S4 imply $\|\widetilde{w}_{\mathcal{C}} - w_{\mathcal{C}}\| \leq \sqrt{2M\{(n\kappa_n)^{-\delta} + \kappa_n\}/L_{\mathcal{C}}}$ with probability at least $1-\epsilon$. Because ϵ is arbitrary, we have $\|\widetilde{w}_{\mathcal{C}} - w_{\mathcal{C}}\| = O_P(\sqrt{(n\kappa_n)^{-\delta} + \kappa_n})$. Combining this with Condition S2 and the fact that $\|\tilde{\rho}_j - \rho_j\|_\infty = O_P\{(n\kappa_n)^{-\delta} + \kappa_n\}$, we conclude that $\|\tilde{\rho}_{\mathcal{C}} - \rho_{\mathcal{C}}\|_\infty = O_P(\sqrt{(n\kappa_n)^{-\delta} + \kappa_n})$. □

S3 Estimation

S3.1 Estimate the Conditional Mean and Variance of the Efficient Influence Function

The full data efficient influence function depends on θ_0 and may also depend on some unknown nuisance parameters, e.g., α_0 and β_0 in Example S2 and $m_1(\cdot)$, $m_0(\cdot)$ and $\pi(\cdot)$ in Example S3. Thus, we need to estimate these unknown quantities. We write the efficient influence function as $\psi(V, U; \theta_0, \eta_0)$ where η_0 is the nuisance parameter. The nuisance parameter η_0 can be estimated using the pilot sample. Denote the resulting estimator by $\tilde{\eta}$. Then θ_0 can be estimated by $\tilde{\theta}$ which is the solution of the estimating equation

$$\sum_{i=1}^{n} R_{1i}\psi(V_i, U_i; \theta, \tilde{\eta}) = 0,$$

where R_{1i} is the inclusion indicator for the pilot sample. Then we obtain the estimates $\tilde{\psi}_i = \psi(V_i, U_i; \tilde{\theta}, \tilde{\eta})$ $(i : R_{1i} = 1)$ for the full data efficient influence function of observations in the pilot sample.

In order to estimate the optimal sampling rule and construct efficient estimator according to the efficient influence function in Lemma 1, one needs to estimate the conditional mean $\Pi_j(\cdot)$ and the standard deviation $\sigma_j(\cdot)$ for $j = 1, \ldots, d$. These quantities can be estimated by fitting a heteroscedastic parametric regression model using the pilot sample and the estimated $\tilde{\psi}_i$'s. In practice, it may be hard to model $\Pi_j(\cdot)$ and $\sigma_j(\cdot)$ for $j = 1, \ldots, d$. If plausible parametric models are not available, we recommend to estimate them nonparametrically. Many methods are available for this task, including kernel smoothing (Fan and Yao, 1998; Fan and Gijbels, 2018), sieve methods (Huang, 1998, 2001; Chen, 2007) and kernel ridge regression (Mendelson, 2002). A sieve method that can estimate the conditional mean and standard deviation simultaneously is proposed in the next section. The proposed method is computationally efficient and performs well even when the dimension of the first-phase variable is moderately high.

S3.2 New Nonparametric Estimators for the Conditional Mean and Variance

If a plausible model for the conditional mean function $\Pi_j(v)$ is unavailable for $j = 1, \ldots, d$, we can approximate it with a linear combination of some basis functions such as polynomials, wavelets, or splines. Let $p(v) = (p_1(v), \ldots, p_K(v))^{\mathrm{T}}$ be a vector of basis functions which can change with n. One can increase K with n to make the approximation more and more accurate as the sample size increases. Then, we approximate $\Pi_j(v)$ by $\gamma_1^{\mathrm{T}} p(v)$ for some γ_1. The conditional standard deviation $\sigma_j(v)$ can be approximate similarly by $\gamma_2^{\mathrm{T}} p(v)$ for some γ_2. However, this approximation is not guaranteed to be non-negative. An infeasible negative sampling rule is obtained if we plug the negative approximation into the expression of the optimal sampling rule in Theorem 1. Using the truncated version $\max\{\gamma_2^{\mathrm{T}} p(v), 0\}$ can avoid this problem but leads the function not differentiable with respect to γ_2 which makes optimization difficult. Hence we propose to use a transformation function $\Lambda(v) = \log(1+\exp(v))$ and use $\Lambda(\gamma_2^{\mathrm{T}} p(v))$ to approximate $\sigma_j(v)$. The function $\Lambda(v)$ is a smooth approximation of the truncation function $\max\{v, 0\}$. Moreover, it is convex, Lipchitz continuous, differentiable, and non-negative. These nice properties benefit the optimization.

Then, the remaining problem is to determine γ_1 and γ_2. The discussion behind Theorem 1 can show that $\sigma_j(\cdot)$ minimizes $E[(\psi_j - \Pi_j(V))^2 / f_2(V)]$ over all positive $f_2(\cdot)$ such that $E[f_2(V)] \leq E[\sigma_j(V)]$. This motivates us to consider the penalized objective formulation of the constrained optimization problem $E[(\psi_j - \Pi_j(V))^2 / f_2(V)] + E[f_2(V)]$. One can verify that this objective function is minimized if $f_2(\cdot) = \sigma_j(\cdot)$. On the other hand, it is straightforward to show, for any given positive function $f_2(\cdot)$, the conditional mean function $\Pi_j(\cdot)$ minimizes the weighted least squares objective function $E[(\psi_j - f_1(V))^2 / f_2(V)] + E[f_2(V)]$ over all $f_1(\cdot)$. This inspires us to recover the conditional mean and variance simultaneously by minimizing $E[(\psi_j - f_1(V))^2 / f_2(V)] + E[f_2(V)]$ with respect to $f_1(\cdot)$ and $f_2(\cdot)$. By replacing the expectation with sample mean and plugging in the estimates and approximations, we obtain the objective function

$$\mathcal{L}_{nj}(\gamma_1, \gamma_2) = \frac{1}{n\kappa_n} \sum_{i=1}^{n} R_{1i} \left\{ \frac{(\tilde{\psi}_{ij} - \gamma_1^{\mathrm{T}} p(V_i))^2}{\Lambda(\gamma_2^{\mathrm{T}} p(V_i))} + \Lambda(\gamma_2^{\mathrm{T}} p(V_i)) \right\}, \tag{S13}$$

where $\tilde{\psi}_{ij}$ is the jth component of $\tilde{\psi}_i$ for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. Let $\tilde{\gamma}_{1j}$, $\tilde{\gamma}_{2j}$ be the minimum point of (S13). Then $\Pi_j(V)$ and $\sigma_j(V)$ can be estimated by $\widetilde{\Pi}_j(V) = \tilde{\gamma}_{1j}^{\mathrm{T}} p(V)$ and $\tilde{\sigma}_j(V) = \Lambda(\tilde{\gamma}_{2j}^{\mathrm{T}} p(V))$, respectively. The proposed objective function has the following block-wise convex property.

Proposition S1. *For $j = 1, \ldots, d$ and any give γ_1, $\mathcal{L}_{nj}(\gamma_1, \gamma_2)$ is convex with respect to γ_2; for $j = 1, \ldots, d$ and any give γ_2, $\mathcal{L}_{nj}(\gamma_1, \gamma_2)$ is convex with respect to γ_1.*

Proof. For any given γ_1, the Hessian of $\mathcal{L}_{nj}(\gamma_1, \gamma_2)$ with respect to γ_2 is

$$\frac{1}{n\kappa_n}\sum_{i=1}^{n} R_{1i}p(V_i)p(V_i)^{\mathrm{T}}\Big\{(\tilde{\psi}_i - \gamma_1^{\mathrm{T}}p(V_i))^2\times$$
$$\Big[2\Lambda(\gamma_2^{\mathrm{T}}p(V_i))^{-3}\Lambda^{(1)}(V_i)^2 - \Lambda(\gamma_2^{\mathrm{T}}p(V_i))^{-2}\Lambda^{(2)}(V_i)^2\Big] + \Lambda^{(2)}(\gamma_2^{\mathrm{T}}p(V_i))\Big\}$$

where $\Lambda^{(1)}(v) = \exp(v)/\{1+\exp(v)\}$ and $\Lambda^{(2)}(v) = \exp(v)/\{1+\exp(v)\}^2$. This matrix is positive semi-definite because $\Lambda^{(2)}(v) > 0$ and $2\Lambda(v)^{-3}\Lambda^{(1)}(v)^2 - \Lambda(v)^{-2}\Lambda^{(2)}(v) > 0$. Thus $\mathcal{L}_{nj}(\gamma_1, \gamma_2)$ is convex with respect to γ_2. For any given γ_2, the Hessian of $\mathcal{L}_{nj}(\gamma_1, \gamma_2)$ with respect to γ_1 is

$$\frac{1}{n\kappa_n}\sum_{i=1}^{n} R_{1i}p(V_i)p(V_i)^{\mathrm{T}}\frac{2}{\Lambda(\gamma_2^{\mathrm{T}}p(V_i))},$$

which is also positive semi-definite. This completes the proof. □

Proposition S1 shows $\mathcal{L}_{nj}(\gamma_1, \gamma_2)$ is block-wise convex with respect to γ_1 and γ_2 for $j = 1, \ldots, d$ and hence the optimization problem (S13) can be solved efficiently by routine optimization algorithms. So far we have defined a nonparametric estimator for $\Pi_j(V)$, $\sigma_j(V)$ based on the sieve method (Chen, 2007). For our numerical experiments, the estimators $\widetilde{\Pi}_j$ and $\tilde{\sigma}_j$ are employed. In the simulation, we introduce a small regularization $0.1(d_V+1)(\|\gamma_1\|^2+\|\gamma_2\|^2)$ into the loss function (S13) to further enhance the stability of the solution, where d_V is the dimension of the first-phase variable V. In the simulation studies and real data analysis, we normalize all first-phase variables to the range $[0, 1]$ using min-max normalization and then use a second-order polynomial with full interactions between variables as the basis functions. This simple choice performs well in our numerical experiments, and we recommend it for practical use. However, we do not claim that this choice of basis functions is optimal in any sense. Identifying the optimal basis functions for the problem considered here remains an interesting direction for future research.

The idea to recover the mean and variance simultaneously also appears in parametric heterogeneous regression literature and is shown to perform well in finite sample (Daye et al., 2012; Spady and Stouli, 2018). Our estimator is an extension of the idea to the nonparametric literature. The proposed method has several nice properties compared to other nonparametric conditional mean and variance estimators, such as kernel smoothing, sieve least squares, and kernel ridge regression. First, it is computationally efficient as the conditional mean and variance can be estimated simultaneously by

solving the optimization problem (S13). Second, heteroscedasticity is considered in fitting the conditional mean model, which can deliver efficiency gains when fitting a model with many parameters and limited observations (Daye et al., 2012). Third, the conditional variance estimator is always positive. This is also a desirable property (Yu and Jones, 2004) which is not possessed by some classic existing methods, e.g., local linear kernel smoothing, sieve least squares, and kernel ridge regression.

S3.3 Estimate the Parameters of Interest

Define $R_{2i} = 0$ for subjects with $R_{1i} = 1$. With some abuse of notation, let $R_i = R_{1i} + R_{2i}$ be the overall sampling indicator for the second phase sampling. Under the sampling procedure proposed in Section S3, V_i is measured for all the n subjects, and U_i is measured for subjects with $R_i = 1$. If a nonrandom sampling rule $\rho(\cdot)$ is used to select the subsequent sample in the second phase, we have the inclusion probability $P(R_i = 1 \mid V_i) = \kappa_n + (1 - \kappa_n)\rho(V_i)$ for $i = 1, \ldots, n$. Denote the sampling rule adopted in the second phase by $\tilde{\rho}(\cdot)$, where $\tilde{\rho}(\cdot)$ depends on the pilot sample and hence is random. However, it converges to some nonrandom sampling rule according to Theorem 5. Thus the inclusion probability can be approximated by $\rho_n(V_i) = \kappa_n + (1 - \kappa_n)\tilde{\rho}(V_i)$. Let the inverse probability weighted estimator $\hat{\theta}_{\text{ipw}}$ be the solution of the estimating equation

$$\sum_{i=1}^{n} \frac{R_i \psi(V_i, U_i; \theta, \tilde{\eta})}{\rho_n(V_i)} = 0.$$

The inverse probability weighted estimator $\hat{\theta}_{\text{ipw}}$ is $\sqrt{n}$-consistent under certain regularity conditions but may not be efficient (Tsiatis, 2007). Based on $\hat{\theta}_{\text{ipw}}$, an efficient estimator can be obtained through one-step estimation (Bickel, 1982). Let $\widetilde{\Pi}(\cdot)$ be an estimate of $\Pi(\cdot)$ based on the pilot sample. According to the efficient influence function given in Lemma 1, the one-step estimator is defined by

$$\begin{aligned}\hat{\theta} &= \hat{\theta}_{\text{ipw}} + \sum_{i=1}^{n} \frac{R_i \psi(V_i, U_i; \hat{\theta}_{\text{ipw}}, \tilde{\eta})}{\rho_n(V_i)} - \sum_{i=1}^{n} \left(\frac{R_i}{\rho_n(V_i)} - 1 \right) \widetilde{\Pi}(V_i) \\ &= \hat{\theta}_{\text{ipw}} - \sum_{i=1}^{n} \left(\frac{R_i}{\rho_n(V_i)} - 1 \right) \widetilde{\Pi}(V_i),\end{aligned}$$

The one-step estimator $\hat{\theta}$ is asymptotically normal and efficient under appropriate empirical process conditions (van der Laan and Robins, 2012; van der Vaart and Wellner, 1996).

As noted by an anonymous reviewer, the use of the pilot sample breaks the i.i.d. structure across observations because the distribution of the sampling indicator in the second phase depend on the pilot sample. This complicates the theoretical

analysis of the proposed one-step estimator $\hat{\theta}$. However, we note that, conditional on the pilot sample, the remaining observations are i.i.d., and the pilot sample typically constitutes only a small fraction of the total dataset. Intuitively, this suggests that the overall data distribution does not deviate substantially from the i.i.d. setting. Although not able to rigorously prove, we conjecture that this deviation does not cause serious issues. Our numerical results indicate that, despite of the non-i.i.d. structure, the proposed estimator $\hat{\theta}$ performs well in practice, suggesting that the non-i.i.d. structure and potential overfitting are not major concerns—at least in our numerical experiments. The simulations in Section S4.4 show that $\hat{\theta}$ has negligible bias and achieves higher finite-sample efficiency than several alternative estimators. Based on these observations, we recommend the one-step estimator $\hat{\theta}$ for practical use, while leaving a rigorous investigation of its theoretical properties to future research.

S4 Additional Simulations

S4.1 Response Mean

In this section, we consider the response mean estimation problem where covariates are inexpensive and the response is hard to obtain. As in the main text, we set $n = 5000$. Let Z be a q-dimensional covariate vector with independent $U[-2.5, 2.5]$ components, where $U[-2.5, 2.5]$ is the uniform distribution on $[-2.5, 2.5]$. Suppose the response

$$Y = \theta_0 + \zeta_q^{\mathrm{T}} Z + \nu_1(Z)\epsilon,$$

where $\theta_0 = 1$, $\zeta_q = (0.5/\sqrt{q}, \ldots, 0.5/\sqrt{q})^{\mathrm{T}}$, $\nu_1(z) = \sqrt{0.1 + (2\zeta_q^{\mathrm{T}} z)^4}$, ϵ is the standard normal error. In this example, we let $V = Z$ be the vector of first-phase variables and $U = Y$ be the second-phase variable. The parameter of interest is the response mean $\theta_0 = E[Y]$. We take $\kappa_n = \varpi/\{1 + \log(\varpi n/q)\}$ in this simulation. Figure S1 is the boxplot based on the results of 500 simulations.

[Insert Fig. S1 about here]

As can be seen in Fig. S1, the estimation efficiency is improved under $\tilde{\rho}_{\mathcal{S}}$ compared to the uniform rule. The improvement is observed for different combinations of n and q, and is particularly pronounced when $q = 5$. The REs are

1.5038, 1.7409, 1.5103, and 2.1777 when $(n, q) = (2000, 1), (2000, 5), (5000, 1)$, and $(5000, 5)$, which indicates that the efficiency is significantly improved under the proposed optimal sampling rule compared to that under the uniform rule.

In the following, we consider the problem of multi-dimensional parameter estimation and evaluate the effectiveness of the sampling rule $\tilde{\rho}_{\mathcal{C}}$ and $\tilde{\rho}_{\mathcal{G}}$. We consider a two-dimensional response variable $Y = (Y_1, Y_2)^{\mathrm{T}}$. Suppose

$$Y_1 = \theta_{01} - \zeta_q^{\mathrm{T}} Z + \nu_1(Z)\epsilon_1,$$

$$Y_2 = \theta_{02} + \sin(\zeta_q^{\mathrm{T}} Z) + \nu_2(Z)\epsilon_2,$$

where $\theta_{01} = 1$, $\theta_{02} = 0$, $\nu_1(z)$ is introduced in the scalar case, $\nu_2(z) = \exp(2\zeta_q^{\mathrm{T}} z)$, ϵ_1 and ϵ_2 are independent standard normal errors, and Z is defined in the same way as in the scalar case. We consider the estimation of the two-dimensional response mean $\theta_0 = (\theta_{01}, \theta_{02})^{\mathrm{T}} = (E[Y_1], E[Y_2])^{\mathrm{T}}$ under two-phase designs. Table S1 reports the bias and standard error (SE) of the one-step estimator with different sampling rules based on 500 simulations, and the RE compared to the uniform rule.

[Insert Table S1 about here]

In all cases, the SEs under the sampling rules $\tilde{\rho}_{\mathcal{C}}$ and $\tilde{\rho}_{\mathcal{G}}$ are smaller than that under the uniform rule. In some cases, the improvement is very significant with a RE close to 3.

S4.2 Linear Regression Coefficient

In this subsection, we consider the problem of estimating linear regression coefficients when response variables and a part of covariates are measured in the first phase and other covariates are measured in the second phase. Define Z in the same way as in Section S4.1. Suppose

$$X = \sin(\zeta_q^{\mathrm{T}} Z) + \nu_2(Z)\epsilon_x,$$

$$Y = \zeta_q^{\mathrm{T}} Z + \theta_0 X + \epsilon_y,$$

where $\theta_0 = 1$, $\nu_2(z)$ is defined in the last Section S4.1, and ϵ_x and ϵ_y are independent and follow a standard normal distribution. Let $V = (Y, Z^{\mathrm{T}})^{\mathrm{T}}$ be the first-phase variable vector and $U = X$ be the second-phase variable. The estimation

of the regression coefficient θ_0 is considered in this simulation. We take $\kappa_n = \varpi/[1+\log\{\varpi n/(q+1)\}]$ in this simulation. Figure S2 contains boxplots of the estimator in 500 simulations across different combinations of n and q.

[Insert Fig. S2 about here]

As can be seen in Fig. S2, the accuracy of the estimator is improved under $\tilde{\rho}_{\mathcal{S}}$ compared to that under the uniform rule. The improvements are observed across different combinations of n and q, and is particularly pronounced when $(n, q) = (5000, 5)$. The REs are 2.1175, 2.4225, 2.3096, and 2.3206 when $(n, q) = (2000, 1), (2000, 5), (5000, 1)$, and $(5000, 5)$.

Next, we consider the case with a two-dimensional regression coefficient of interest. The covariate vector Z is defined in the same way as in Section S4.1. Let $X = (X_1, X_2)^{\mathrm{T}}$ be a two-dimensional covariate vector which satisfies

$$X_1 = \zeta_q^{\mathrm{T}} Z + \nu_1(Z)\epsilon_{x1},$$

$$X_2 = -\zeta_q^{\mathrm{T}} Z + \nu_2(Z)\epsilon_{x2},$$

where $\nu_1(z)$ and $\nu_2(z)$ are defined in Section S4.1, ϵ_{x1}, ϵ_{x2} are independent standard normal variables. The response variable Y satisfies

$$Y = \zeta_q^{\mathrm{T}} Z + \theta_0^{\mathrm{T}} X + \epsilon_y,$$

where $\theta_0^T = (\theta_{01}, \theta_{02}) = (0, 1)$ and ϵ_y follows a standard normal distribution.

Table S2 reports the bias, SE of the one-step estimator, and the RE compared to the uniform rule based on 500 simulations.

[Insert Table S2 about here]

As seen in Table S2, the SEs under under the sampling rules $\tilde{\rho}_{\mathcal{C}}$ and $\tilde{\rho}_{\mathcal{G}}$ are smaller than those under the uniform rule in all cases. The REs are larger than two in some cases.

S4.3 Sampling Design with Different Priorities under the Multi-dimensional Setting in Section 5

In this section, under the multi-dimensional setting in Section 5, we illustrate the numerical effect of the priority parameter $a = (a_1, 1 - a_1)$ discussed in Remark 1. We set $(n, q) = (5000, 1)$. Table S3 presents the bias, SE of the estimator, and the RE compared to the uniform rule, based on 500 simulations. We consider the one-step estimator under the estimated optimal rule for θ_{01}, θ_{02}, and the G-opt rule with $a_1 = 0.05, 0.5$ and 0.95. The results under the uniform sampling rule are also reported for reference. Table S3 shows that biases of the estimator are small under all sampling rules. Notably, the SE for estimating θ_{01} decreases as a_1 increases, accompanied by a modest increase in the SE for estimating θ_{02}. Notably, the SEs for both parameter components are smaller than those obtained under the uniform sampling rule, no matter what value a_1 takes. This desirable property is not achieved for the estimated optimal rule for θ_{01} or θ_{02}.

[Insert Table S3 about here]

S4.4 Comparison with Alternative Estimators

In this section, we further investigate the numerical performance of the proposed one-step estimator $\hat{\theta}$ and compare it with several alternative estimators. There are multiple alternative ways to construct an efficient estimator in two-phase studies. For example, one may construct an estimator, based on the efficient influence function and the one-step estimation technique, using all observations except those in the pilot sample. Denote the resulting estimator by $\hat{\theta}_{\mathrm{ex}}$. Conditional on the pilot sample, the remaining observations are i.i.d. Thus, the asymptotic properties of $\hat{\theta}_{\mathrm{ex}}$ can be established using standard arguments conditional on the pilot sample. In addition, the efficiency loss lead by excluding the pilot sample is asymptotically negligible because $\kappa_n \to 0$. Another reasonable approach is to perform an inverse-variance-weighted meta-analysis combining the pilot sample estimator $\tilde{\theta}$ and the estimator $\hat{\theta}_{\mathrm{ex}}$. Denote the resulting estimator by $\hat{\theta}_{\mathrm{ivw}}$. Figure S3 presents boxplots of different average treatment effect estimators under the optimal sampling rule for a scalar parameter $\tilde{\rho}_{\mathcal{S}}$ over 500 simulations with $q = 1$ and $n = 2000, 5000, 20000, 50000$. For reference, the results for the one-step estimator under the uniform sampling rule are also included.

[Insert Fig. S3 about here]

Figure S3 shows that, under $\tilde{\rho}_{\mathcal{S}}$, all three estimators have higher efficiency than the one-step estimator under the uniform sampling rule. The performance of the three estimators under $\tilde{\rho}_{\mathcal{S}}$ is similar when $n = 50000$. However, for smaller sample sizes ($n = 2000, 5000, 50000$), the proposed one-step estimator $\hat{\theta}$ demonstrates better finite-sample efficiency than both $\hat{\theta}_{\text{ex}}$ and $\hat{\theta}_{\text{ivw}}$ under $\tilde{\rho}_{\mathcal{S}}$. In addition, the confidence interval based on $\hat{\theta}$ and normal approximation performs well in the simulation. Coverage rates of the confidence intervals constructed using $\hat{\theta}$ and the influence function-based standard error estimator are 94.4%, 93.8%, 96.8%, and 96.8% when $n = 2000, 5000, 20000, 50000$, respectively.

References

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, 647–671.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Cochran, W. G. (2007). *Sampling Techniques* (3rd ed.). John Wiley & Sons.

Daye, Z. J., J. Chen, and H. Li (2012). High-dimensional heteroscedastic regression with an application to eqtl data analysis. *Biometrics 68*(1), 316–326.

Fan, J. and I. Gijbels (2018). *Local polynomial modelling and its applications*. Routledge.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*(3), 645–660.

Gilbert, P. B., X. Yu, and A. Rotnitzky (2014). Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in Medicine 33*(6), 901–917.

He, X., X. Pan, K. M. Tan, and W.-X. Zhou (2021). Smoothed quantile regression with large-scale inference. *Journal of Econometrics*.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics 26*(1), 242–272.

Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica*, 173–197.

Lin, H.-W. and Y.-H. Chen (2014). Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *American Journal of Epidemiology 180*(3), 308–317.

McNamee, R. (2002). Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in Medicine 21*(23), 3609–3625.

Mendelson, S. (2002). Geometric parameters of kernel machines. In *Proceedings of the Fifteenth Annual Conference on Computational Learning Theory,*, pp. 29–43.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 1349–1382.

Rigollet, P. and J.-C. Hütter (2015). High dimensional statistics. *Lecture Notes for Course 18S997 813*(814), 46.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics 8*(1), 171–176.

Spady, R. and S. Stouli (2018). Simultaneous mean-variance regression. *arXiv:1804.01631*.

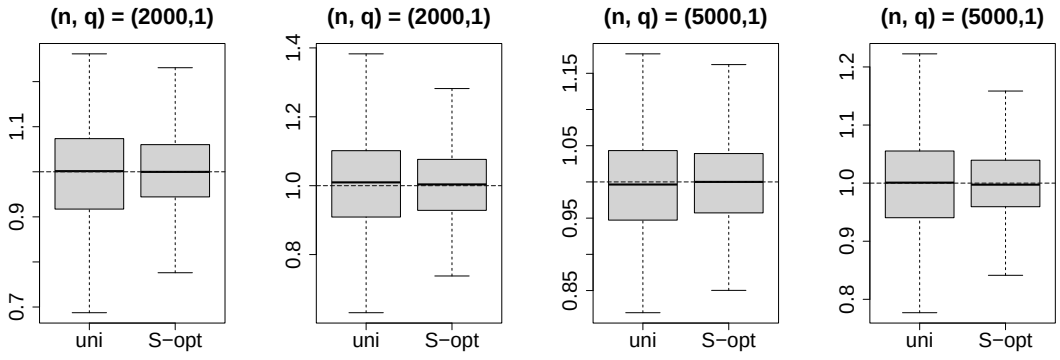
Tao, R., D. Zeng, and D.-Y. Lin (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association 112*(520), 1468–1476.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

van der Laan, M. J. and J. M. Robins (2012). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York, NY: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, New York.



(a) (b) (c) (d)

Figure S1: Boxplots for the mean estimation with different combinations of n and q; dashed lines are the true values.

Yang, S. and P. Ding (2019). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*.

Yu, K. and M. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association 99*(465), 139–144.

Zeng, D. and D. Lin (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association 109*(505), 371–383.

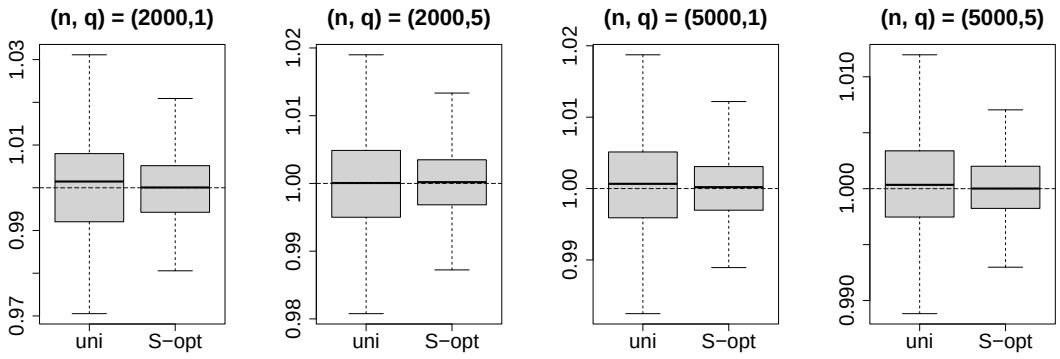
Zhou, H., W. Xu, D. Zeng, and J. Cai (2014). Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 1*(76), 197–215.

Table S1: Bias, SE, and RE in two-dimensional mean estimation

(n, q)	Rule	Estimate of θ_{01}			Estimate of θ_{02}		
		Bias	SE	RE	Bias	SE	RE
$(2000, 1)$	uni	0.0017	0.1224	1.0000	-0.0012	0.1601	1.0000
	C-opt	-0.0028	0.0897	1.8620	-0.0001	0.1201	1.777
	G-opt	0.0007	0.0965	1.6088	0.0050	0.1240	1.667
$(2000, 5)$	uni	0.0068	0.1441	1.0000	0.0254	0.2305	1.0000
	C-opt	0.0024	0.1031	1.9535	0.0053	0.1526	2.2816
	G-opt	0.0060	0.1068	1.8205	0.0036	0.1522	2.2936
$(5000, 1)$	uni	0.0008	0.0719	1.0000	0.0047	0.0981	1.0000
	C-opt	-0.0038	0.0601	1.4312	0.0019	0.0754	1.6928
	G-opt	0.0010	0.0599	1.4408	-0.0003	0.0722	1.8461
$(5000, 5)$	uni	-0.0056	0.0889	1.0000	-0.0047	0.1639	1.0000
	C-opt	-0.0006	0.0665	1.7871	-0.0018	0.0947	2.9954
	G-opt	0.0010	0.0620	2.0560	-0.0012	0.0962	2.9027

Table S2: Bias, SE, and RE in two-dimension regression coefficient estimation

(n, q)	Rule	Estimate of θ_{01}			Estimate of θ_{02}		
		Bias	SE	RE	Bias	SE	RE
$(2000, 1)$	uni	-0.0005	0.0145	1.0000	0.0010	0.0110	1.0000
	C-opt	-0.0012	0.0126	1.3243	4e-04	0.0092	1.4296
	G-opt	-0.0003	0.0133	1.1886	8e-04	0.0088	1.5625
$(2000, 5)$	uni	0.0000	0.0126	1.0000	0.0003	0.0076	1.0000
	C-opt	0.0005	0.0105	1.4400	0.0004	0.0057	1.7778
	G-opt	0.0000	0.0103	1.4965	0.0003	0.0057	1.7778
$(5000, 1)$	uni	0.0001	0.0094	1.0000	0.0009	0.0071	1.0000
	C-opt	-0.0003	0.0079	1.4158	0.0005	0.0055	1.6664
	G-opt	-0.0003	0.0078	1.4523	0.0003	0.0054	1.7287
$(5000, 5)$	uni	0.0008	0.0075	1.0000	0.0003	0.0047	1.0000
	C-opt	0.0001	0.0062	1.4633	0.0001	0.0032	2.1572
	G-opt	0.0003	0.0060	1.5625	0.0001	0.0032	2.1572



(a) (b) (c) (d)

Figure S2: Boxplots for linear regression coefficient estimation with different combinations of n and q; dashed lines are the true values.

Table S3: Bias, SE, and RE in two-dimensional average treatment effect estimation with $(n, q) = (5000, 1)$

Rule	Estimate of θ_{01}			Estimate of θ_{02}		
	Bias	SE	RE	Bias	SE	RE
uni	0.0273	0.3113	1.0000	-0.0101	0.2787	1.0000
S-opt (θ_{01})	0.0206	0.2463	1.5975	-0.0024	0.2977	0.8764
S-opt (θ_{02})	0.0516	0.3639	0.7318	0.0009	0.1851	2.2671
G-opt ($a_1 = 0.05$)	0.0427	0.2814	1.2238	0.0026	0.1982	1.9773
G-opt ($a_1 = 0.5$)	0.0254	0.2585	1.4502	0.0058	0.2222	1.5732
G-opt ($a_1 = 0.95$)	0.0262	0.2517	1.5296	-0.0053	0.2262	1.5181

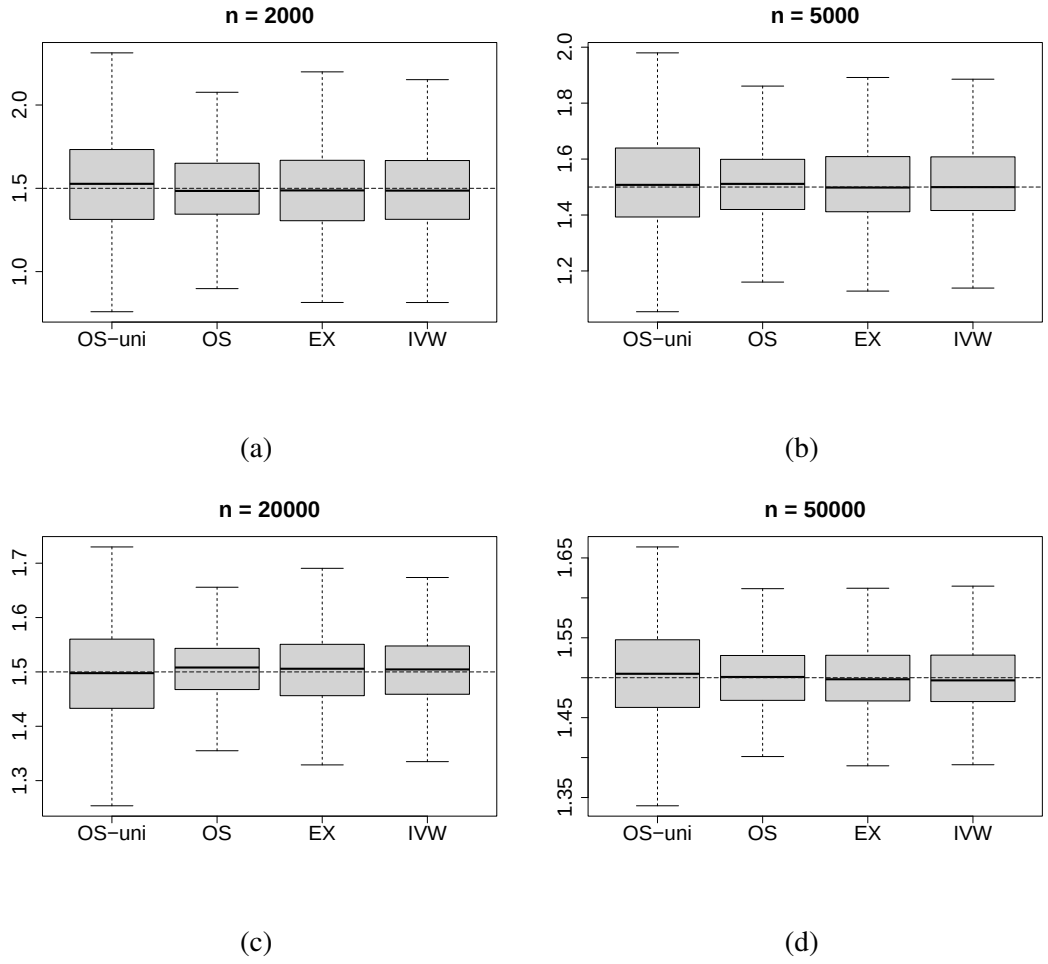


Figure S3: Boxplots for the average treatment effect estimation under $q = 1$ and $n = 2000, 5000, 20000, 50000$; OS-uni: $\hat{\theta}$ under the uniform sampling rule; OS: $\hat{\theta}$ under $\tilde{\rho}_s$; EX: $\hat{\theta}_{\text{ex}}$ under $\tilde{\rho}_s$; IVW: $\hat{\theta}_{\text{ivw}}$ under $\tilde{\rho}_s$; dashed lines are the true values.