Functional Calibration under Non-Probability Survey Sampling

Zhonglei Wang
Wang Yanan Institute for Studies in Economics and School of Economics,
Xiamen University
and
Xiaojun Mao*
School of Mathematical Sciences, Shanghai Jiao Tong University
and
Jae Kwang Kim
Department of Statistics, Iowa State University

Abstract

Non-probability sampling is prevailing in survey sampling, but ignoring its selection bias leads to erroneous inferences. We offer a unified nonparametric calibration method to estimate the sampling weights for a non-probability sample by calibrating functions of auxiliary variables in a reproducing kernel Hilbert space. The consistency and the limiting distribution of the proposed estimator are established, and the corresponding variance estimator is also investigated. Compared with existing works, the proposed method is more robust since no parametric assumption is made for the selection mechanism of the non-probability sample. Numerical results demonstrate that the proposed method outperforms its competitors, especially when the model is misspecified. The proposed method is applied to analyze the average total cholesterol of Korean citizens based on a non-probability sample from the National Health Insurance Sharing Service and a reference probability sample from the Korea National Health and Nutrition Examination Survey.

Keywords: Data integration; Missing at random; Nonparametric weighting; Reproducing kernel Hilbert space.

*Zhonglei Wang and Xiaojun Mao contribute equally.
1 Introduction

Probability sampling serves as a golden standard to estimate finite population parameters in social science (Elliott and Valliant, 2017; Haziza and Beaumont, 2017), but low response rates and inevitable dropouts have made it “tarnished gold” recently (Keiding and Louis, 2016). Moreover, it is costly and time-consuming to conduct probability sampling, so it is only feasible for well-funded and socially important surveys (Baker et al., 2013; O’Muircheartaigh and Hedges, 2014). On the other hand, due to its feasibility and low cost, non-probability sampling, especially web surveys, has become increasingly popular (Couper, 2000; Couper and Miller, 2008; Dever et al., 2008; Tourangeau et al., 2013; Dever and Valliant, 2014). Nonetheless, a non-probability sample is rarely representative of the target population because of its unknown selection mechanism. If such a selection mechanism is not properly incorporated, it may lead to erroneous inferences. Therefore, adjusting the selection bias for a non-probability sample has become a hot research topic in survey sampling.

If an additional reference probability sample is available, there are primarily two techniques to adjust the selection bias for a non-probability sample. One method involves combining the non-probability sample and the reference probability sample to calculate propensity scores (Rosenbaum and Rubin, 1983). Under a parametric assumption for the response model, Lee (2006) developed a quasi-randomization method to estimate the propensity scores for the pooled sample. The pooled sample is then divided into groups based on the estimated propensity scores, and modified sampling weights for the non-probability sample are calculated for each group. Lee and Valliant (2009) generalized the quasi-randomization method (Lee, 2006) by an additional calibration adjustment (Deville and Särndal, 1992) for the case when marginal population totals of auxiliaries are avail-
able. Valliant and Dever (2011) compared several propensity-score-based estimators and concluded that sampling weights of the reference probability sample should be incorporated when estimating the propensity scores. Also see Rivers (2007), Bethlehem (2010), Brick (2015), Elliott and Valliant (2017) and the references within for more details. The second approach uses calibration to adjust the selection bias of a non-probability sample. Kim and Wang (2019) estimated the “importance weights” for a non-probability sample based on the Kullback-Leibler (KL) divergence, and they only assumed the availability of the marginal population totals for auxiliaries. Chen et al. (2020) proposed to estimate the parameters in the propensity score model based on a calibration constraint (Wu and Sitter, 2001; Beaumont, 2005), and a reference probability sample is used to estimate the population totals of a specific estimating function.

Existing works assume a parametric model either for the propensity scores (Elliott and Haviland, 2007; Chen et al., 2020) or for the sampling weights (Kim and Wang, 2019), so they suffer from model misspecification (Robins et al., 1994; Han and Wang, 2013). In this paper, we present a nonparametric method based on functional calibration in a reproducing kernel Hilbert space (Wahba, 1990, RKHS) for adjusting the selection bias for a non-probability sample. Specifically, uniformly calibrating functions in an RKHS is utilized to get the estimated sampling weights of a non-probability sample, and the reference probability sample is used to estimate the associated population totals as Chen et al. (2020). In addition, we propose to use the KL divergence as a penalty to avoid overfitting. Under some regularity conditions, asymptotic properties of the proposed method are investigated, and numerical results demonstrate the advantages of the proposed method over its alternatives.

The proposed method differs from existing ones in the following aspects. Some existing
methods (Valliant and Dever, 2011; Chen et al., 2020) estimate propensity scores for a non-probability sample, so the corresponding estimator is inefficient if estimated propensity scores are close to zero. To avoid such inefficiency, we propose directly estimating the sampling weights and applying penalties as well. Unlike Chen et al. (2020), we do not make any parametric assumption for calibration. Rather than that, the sampling weights of a non-probability sample are calculated via uniform minimization of a calibration-based objective function over an RKHS. Thus, the proposed method outperforms the method of Chen et al. (2020) and other existing ones in terms of robustness. Since the proposed method uniformly calibrates functions in an RKHS, it is essentially a multitask-oriented learning in the sense that the estimated sampling weights can be used to estimate several different parameters from the non-probability sample. This property is appealing especially when the non-probability sample consists of many survey questions. Up to our knowledge, we are the first to use uniform calibration to adjust the selection bias for a non-probability sample.

Although the proposed method is motivated by Wong and Chan (2018), instead of assuming the auxiliaries to be available for every element in a finite population, we consider the setup when a reference probability sample is used to estimate the population totals for functions in the RKHS. It is worth pointing out that assuming the availability of auxiliary information for each element in a finite population is generally unrealistic under survey sampling. Different from Wong and Chan (2018), we propose a penalty based on a non-parametric density ratio model, and numerical results indicate that the proposed method is more efficient than theirs in terms of computation and estimation; see Section 5 for details. Besides, since the sampling indicators are no longer independent for the reference probability sample under rejective sampling, the theoretical results from Wong and Chan (2018)
are not applicable to our method, and we adopt a different empirical process technique instead. Even though empirical processes have been studied under survey sampling, most of them assumed that the sample size is of the same order with the population size asymptotically \cite{Breslow2007, Conti2014, Bertail2017, Han2021}, but it is rarely the case for a reference probability sample in practice due to the limited budget. \cite{Boistard2017} relaxed that stringent condition on the sample size, but they focused on single-stage sampling designs. In this paper, theoretical properties of the proposed method are investigated without assuming that the sizes of the reference probability sample and the finite population are of the same asymptotic order, and the proposed method applies as long as the reference probability sample is generated by a rejective sampling design, not limited to single-stage sampling.

The remaining of this paper is organized as follows. The motivation of the proposed method is introduced in Section \ref{sec:2}. The proposed method is presented in Section \ref{sec:3}, and its theoretical properties are investigated in Section \ref{sec:4}. Simulation studies are demonstrated in Section \ref{sec:5}. The proposed method is applied to estimate the average total cholesterol of the Korean citizens in Section \ref{sec:6}. Concluding remarks are provided in Section \ref{sec:7}.

\section{Motivation} \label{sec:2}

\subsection{Basic setup} \label{sec:2.1}

To introduce the idea of uniform calibration, assume that the finite population $\mathcal{F}_N = \{(x_i, y_i) : i = 1, 2, \ldots, N\}$ is a random sample of size $N$ from a super-population model,

$$y_i = m(x_i) + \epsilon_i \quad (i = 1, \ldots, N), \quad (2.1)$$
Table 1: Data structure of the two samples. “X” denotes the auxiliary vector, and “Y” denotes the response of interest. “✓” is used if the information is available and “✗” otherwise.

| Sample | Type                  | X | Y | Representativeness |
|--------|-----------------------|---|---|--------------------|
| A      | Non-probability Sample | ✓ | ✓ | No                 |
| B      | Probability Sample    | ✓ | ✗ | Yes                |

where $x_i \in \mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^d$ is a $d$-dimensional compact set, $y_i$ is the response of interest, $m(x_i) = E(y_i \mid x_i)$ is a smooth function (Wahba 1990), $\epsilon_i$ is independent with $x_i$, $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma_i^2$, and $\sigma_1, \ldots, \sigma_N$ are positive constants with respect to the super-population model. We adopt a design-based framework and assume that the finite population $\mathcal{F}_N$ is fixed once it is generated; see Part I of Särndal et al. (2003), Chapter 1 of Fuller (2009b) and Chen et al. (2020) for details. The parameter of interest is the population mean $\bar{Y}_N = N^{-1} \sum_{i=1}^{N} y_i$. For simplicity, assume that the population size $N$ is known.

Let $A$ be a non-probability sample with observations on both the auxiliary vector and the response of interest, and $B$ be a reference probability sample with information on the auxiliary vector only. That is, both $\{(x_i, y_i) : i \in A\}$ and $\{(x_i, \pi_B,i) : i \in B\}$ are available, where $\pi_B,i$ is the first-order inclusion probability of the $i$-th element with respect to the probability sample $B$. Since the selection mechanism of the non-probability sample $A$ is unknown, this sample may not represent the finite population. Table 1 shows the general data structure of the two samples. How to adjust the selection bias of the non-probability sample $A$ using the auxiliary information available from the probability sample $B$ is an important practical problem in survey sampling.

A special case is when the probability sample $B$ is a census, and it was investigated by
However, a census is usually hard to obtain in practice. Thus, we focus on a more general case assuming that the probability sample $B$ is generated by a rejective sampling design. Under rejective sampling, a sample is only acceptable if a certain criterion is satisfied; see Fuller (2009a, Section 1.2.6) and Fuller (2009b) for details. Besides, the corresponding sampling indicators $\{\delta_{B,i} : i = 1, \ldots, N\}$ are negatively associated (Bertail and Clémencçon 2016), where $\delta_{B,i} = 1$ if $i \in B$ and 0 otherwise.

### 2.2 Uniform calibration

To motivate the proposed method, we make a stronger assumption for (2.1) that there exists a positive constant $\sigma_0$ such that $\sigma_i = \sigma_0$ for $i = 1, \ldots, N$. We consider an estimator of the form $\hat{Y} = N^{-1} \sum_{i \in A} \omega_i y_i$, where $\{\omega_i : i \in A\}$ are a set of weights to be determined. Under the super-population model (2.1), we have

$$\hat{Y} - \bar{Y}_N = N^{-1} \left\{ \sum_{i \in A} \omega_i m(x_i) - \sum_{i=1}^N m(x_i) \right\} + N^{-1} \left\{ \sum_{i \in A} \omega_i \epsilon_i - \sum_{i=1}^N \epsilon_i \right\}$$

$$:= C + D. \quad (2.2)$$

Thus, the weights $\{\omega_i : i \in A\}$ are optimal if $Q = C^2 + E\{D^2\}$ is minimized, where the expectation is taken with respect to the super-population model (2.1) conditional on the non-probability sample $A$. If the true mean function satisfied

$$m(x) \in \text{span}\{b_1(x), \ldots, b_L(x)\} \equiv \mathcal{H}_0 \quad (2.3)$$

for some basis functions $b_1(x), \ldots, b_L(x)$, and if $\{x_1, \ldots, x_N\}$ were available, then we could obtain $C^2 = 0$ by imposing

$$\sum_{i \in A} \omega_i [b_1(x_i), \ldots, b_L(x_i)] = \sum_{i=1}^N [b_1(x_i), \ldots, b_L(x_i)] \quad (2.4)$$
as the calibration equation for determining \( \{\omega_i : i \in A\} \). Thus, the optimal weights are those minimizing
\[
\sigma_0^{-2}E\{D^2\} = \sum_{i \in A} (\omega_i - 1)^2 + \text{const}
\]
subject to (2.4). Assumption (2.3) can be restrictive as the mean function is linear in the basis functions. However, we can still use the basis functions in the calibration equation to provide nonparametric calibration estimates by allowing that the basis functions give a uniform approximation of the nonlinear function \( m(x) \) with increasing dimension \( L \). For examples of nonparametric calibration estimation, Montanari and Ranalli (2005) used a single-layer neural network model and Breidt et al. (2005) considered a penalized spline model.

In our setup, instead of observing \( \{x_i : i = 1, \ldots, N\} \), however, only a reference probability sample \( B \) is available. Since \( \sum_{i \in B} \pi_{B,i}^{-1} [b_1(x_i), \ldots, b_L(x_i)] \) is design-unbiased for \( \sum_{i=1}^N [b_1(x_i), \ldots, b_L(x_i)] \) (Horvitz and Thompson, 1952), rather than (2.4), we impose
\[
\sum_{i \in A} \omega_i [b_1(x_i), \ldots, b_L(x_i)] = \sum_{i \in B} \pi_{B,i}^{-1} [b_1(x_i), \ldots, b_L(x_i)]
\]
as the calibration equation. Recall that the calibration (2.5) is justified under (2.3). Now, if assumption (2.3) does not hold, then we may intuitively consider minimizing
\[
Q = \sup_{u \in \mathcal{H}} \left\{ \sum_{i \in A} \omega_i u(x_i) - \sum_{i \in B} \pi_{B,i}^{-1} u(x_i) \right\}^2 + \sigma_0^2 \sum_{i \in A} (\omega_i - 1)^2
\]
directly for some function space \( \mathcal{H} \). The first term of (2.6) achieves the approximate uniform calibration and the second term achieves the weight stabilization. We assume that the function space \( \mathcal{H} \) is an RKHS, so that we can construct certain basis functions in \( \mathcal{H} \) from the sample; see Section S1 of the Supplementary Material for a brief introduction to an RKHS. In the next section, we also propose a different penalty term to stabilize the estimated weights, and the advantage of the new penalty term is shown in Section 5.
2.3 Assumptions

Before closing this section, we make the following assumptions for the non-probability sample $A$

A1. The sampling indicators $\{\delta_{A,i} : i = 1, \ldots, N\}$ are mutually independent, where $\delta_{A,i} = 1$ if $i \in A$ and 0 otherwise.

A2. The sampling indicator $\delta_{A,i}$ is independent with the response of interest $y_i$ given $x_i$. That is, $\Pr(\delta_{A,i} = 1 \mid x_i, y_i) = \pi_A(x_i)$.

The independence assumption in Assumption A1 is widely adopted for non-probability sampling (Keiding and Louis, 2016; Chen et al., 2020); also see Section 17.2 of Wu and Thompson (2020) for details. The non-informative assumption (Pfeffermann, 1993) in Assumption A2 is also commonly presumed for observational studies with a sample similar as the non-probability one; see Rosenbaum and Rubin (1983) for details.

By Assumption 2 and the Bayes formula, we have

$$\pi_A(x_i) = \frac{\pi_1 f(x_i \mid \delta_{A,i} = 1)}{\pi_1 f(x_i \mid \delta_{A,i} = 1) + \pi_0 f(x_i \mid \delta_{A,i} = 0)} = \frac{\pi_1 f_1(x_i)}{\pi_1 f_1(x_i) + \pi_0 f_0(x_i)}, \quad (2.7)$$

where $\pi_1 = \int \pi_A(x)dx$, and $\pi_0 = 1 - \pi_1$, and $f_1(x_i)$ and $f_0(x_i)$ are the conditional probability densities of $x_i$ given $\delta_{A,i} = 1$ and $\delta_{A,i} = 0$ with respect to a certain dominating measure $\mu$, respectively. For simplicity, we assume the dominating measure $\mu$ to be the Lebesgue measure in this paper. Writing $\omega^*(x) = \{\pi_A(x)\}^{-1}$, by (2.7), we can obtain

$$\omega^*(x) = 1 + \frac{\pi_0}{\pi_1} r^*(x), \quad (2.8)$$

where $r^*(x) = f_0(x)/f_1(x)$. That is, there is a one-to-one correspondence between $\omega^*_i = \omega^*(x_i)$ and $r^*_i = r^*(x_i)$ for $i \in A$. In this paper, we propose to estimate $r^*_i$ instead of $\omega^*_i$, and its advantage is discussed in Remark 2 of the next section.
To regulate the selection probabilities associated with the non-probability sample $A$, we make the following assumption.

A3. There exist two positive constants $0 < C_{A,1} < C_{A,2} < 1$, such that $C_{A,1} \leq \pi_A(x) \leq C_{A,2}$ for $x \in \mathcal{X}$.

The assumption $\pi_A(x) > C_{A,1}$ is slightly stronger than assuming $\pi_A(x) > 0$ for $x \in \mathcal{X}$; see Section 17.2 of [Wu and Thompson (2020)] and Assumption A2 of [Chen et al. (2020)] for comparison. However, such an assumption is required to derive the asymptotic properties of the proposed method; see the proof of Lemma 1 in Section S3 of the Supplementary Material for details. Besides, Assumption A3 implies that $n_B$ asymptotically has the same order as the population size $N$, and it makes sense in practice since a non-probability sample usually corresponds to a big data source. The condition $\pi_A(x) < C_{A,2}$ for $x \in \mathcal{X}$ guarantees that the corresponding density ratio function $r^*(x)$ is positive, so that the loss function (3.1) in the next section is valid for $r^*(x)$. Specifically, by (2.7), we have

$$r^*(x) = \frac{f_0(x)}{f_1(x)} = \frac{\pi_1(1 - \pi_A(x))}{\pi_0 \pi_A(x)},$$

and by Assumption A3, we conclude that there exists two positive constant $0 < C_{r,1} < C_{r,2}$ depending only on $C_{A,1}$ and $C_{A,2}$, such that for $x \in \mathcal{X}$, we have

$$C_{r,1} \leq r^*(x) \leq C_{r,2}.$$  

### 3 Proposed method

In Section 2, we have seen that the propensity score estimation problem reduces to the density ratio estimation problem. Density ratio estimation (DRE), the problem of estimating the ratio of two density functions for two different populations, is a fundamental
problem in machine learning (Sugiyama et al., 2012). By partitioning the sample into two groups based on the response status, we can apply the DRE method and thus obtain the inverse propensity scores. One important method of DRE is so called the maximum entropy method, which minimizes the KL divergence (or negative entropy) subject to a normalization constraint (Nguyen et al., 2010).

Applying the maximum entropy method of Nguyen et al. (2010), the density ratio function $r^*(\mathbf{x})$ can be understood as the maximizer of

$$Q(r) = \int r \log(r) f_1 d\mu - \int r f_1 d\mu + 1,$$

(3.1)

where $r = r(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{X}$. A sample version of (3.1) is

$$Q_A(\gamma) = \frac{1}{n_A} \sum_{i=1}^{N} \delta_{A,i} \{ \log(r_i) - 1 \} + 1,$$

(3.2)

where $\gamma = (r_1, \ldots, r_N)^T$, $r_i = r(\mathbf{x}_i)$ if $i \in A$ and $r_i = 0$ otherwise; see (7.26) of Kim and Shao (2022) for details.

We propose to estimate $\{r_i^* : i \in A\}$ by uniformly calibrating functions in an RKHS $\mathcal{H}$. Consider

$$\hat{\gamma} = \arg \min_{\xi_1 \leq r_i \leq \xi_2} \left[ \sup_{u \in \mathcal{H}} \left\{ \frac{S(\gamma, u)}{\|u\|_2^2} - \lambda_1 \|u\|_\mathcal{H}^2 \right\} - \lambda_2 Q_A(\gamma) \right],$$

(3.3)

where $\xi_1 \leq \xi_2$ are predetermined numbers, $\|u\|_\mathcal{H}$ is the norm associated with the RKHS $\mathcal{H}$, $\|u\|_2^2 = n^{-1} \sum_{i=1}^{N} (\delta_{A,i} + \delta_{B,i}) u(\mathbf{x}_i)^2$, $n = n_A + n_B$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are two tuning parameters, and

$$S(\gamma, u) = \left[ N^{-1} \sum_{i=1}^{N} \delta_{A,i} \left\{ 1 + \left( \frac{N}{n_A} - 1 \right) r_i \right\} u(\mathbf{x}_i) - N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(\mathbf{x}_i) \right]^2.$$

(3.4)

In the optimization problem (3.3), we should choose a sufficiently small $\xi_1$ and a sufficiently large $\xi_2$ to guarantee $\xi_1 \leq C_{r,1} < C_{r,2} \leq \xi_2$ by (2.10). In practice, we can set $\xi_1 = 10^{-8}$
and $\xi_2 = 10^8$, for example. Since $\pi_0 \pi_1^{-1}$ is generally unavailable, we replace it by $Nn_{A}^{-1} - 1$ in (3.4). Different from Wong and Chan (2018), we assume an upper bound for $\{r_i : i = 1, \ldots, N\}$ in the optimization problem (3.3), and such an assumption is used to guarantee the convergence rate of $S(\hat{\gamma}, u)$ for $u \in \mathcal{H}$; see (S4.3) in the Supplementary Material for details. For simplicity, we implicitly assume that the auxiliary vectors $\{x_i : i \in A\}$ and $\{x_i : i \in B\}$ are pairwise distinct. Otherwise, the objective function should be minimized only by distinct auxiliaries in $\{x_i : i \in A\} \cup \{x_i : i \in B\}$, and $n$ is the corresponding size of the pooled set. The values for the two tuning parameters $\lambda_1$ and $\lambda_2$ are determined by five-fold cross validation.

Intuition for the objective function (3.3) is briefly discussed. First, $S(\gamma, u)$ in (3.4) balances the two estimators for the population mean $N^{-1} \sum_{i=1}^{N} u(x_i)$ over $u \in \mathcal{H}$. As discussed in Section 2, since the sampling weights are incorporated for the probability sample $B$, the estimator $N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(x_i)$ is design-unbiased for $N^{-1} \sum_{i=1}^{N} u(x_i)$ and $u \in \mathcal{H}$ (Horvitz and Thompson, 1952). On the other hand, if $1 + (Nn_{A}^{-1} - 1)r_i$ is close to $\omega_i^*$ for $i \in A$, the first term in (3.4) is also approximately design-unbiased, so $S(\gamma, u)$ should be small. However, $S(\gamma, u)$ is not scale invariant, and $S(\gamma, cu) = c^2 S(\gamma, u)$ for $c \in \mathbb{R}$. Thus, to make it scale-invariant, we consider $S(\gamma, u)/\|u\|_2^2$ in the objective function (3.3). Since $\mathcal{H}$ is large, there may not exist $\gamma$ such that $S(\gamma, u) = 0$ holds for every $u \in \mathcal{H}$, so we uniformly balance the two estimators by minimizing $\sup_{u \in \mathcal{H}} \{S(\gamma, u)/\|u\|_2^2\}$. Because we have assumed that $m(x)$ is a smooth function in (2.1), a penalty on the smoothness of a function $u$ is incorporated in the objective function (3.3). To stabilize the estimated weights, we use $-\lambda_2 Q_A(\gamma)$ in (3.2) as another penalty.

**Remark 1.** We highlight the difference between the proposed method and existing ones. Chen et al. (2020) proposed a parametric model for $\text{pr}(\delta_{A,i} = 1 \mid x_i)$, and the model
parameters are estimated by a calibration method based on a pre-specified smooth function \( h(x) \). There exist different choices for \( h(x) \), including basis functions of P-splines (Breidt et al., 2005), neural network estimators (Montanari and Ranalli, 2005), and other modern machine learning methods (Breidt and Opsomer, 2017). Even though the aforementioned works were not proposed to adjust the selection bias for a non-probability sample, their methods can be easily implemented in the framework of Chen et al. (2020). Rather than calibrating the predetermined basis function as existing works, we proposed to uniformly calibrate functions in an RKHS, and the limiting properties of the proposed estimator are also investigated; see Section 4 for details.

Remark 2. Wong and Chan (2018) used \( \lambda_2 N^{-1} \sum_{i=1}^{N} \delta_{A,i} \{1 + (Nn_A^{-1} - 1)r_i\}^2 \) as a penalty to avoid extremely large sampling weights, and it is similar to the second term in (2.6). For such a penalty, as \( \lambda_2 \to \infty \), all estimated sampling weights are close to 1. Then, \( \bar{Y}_N \) is estimated merely by the mean \( n_A^{-1} \sum_{i \in A} y_i \) of the non-probability sample \( A \), and this estimator is biased since the selection mechanism for the non-probability \( A \) is overlooked. To avoid this possible estimation bias when \( \lambda_2 \) is large, we propose \( -\lambda_2 Q_A(\gamma) \) instead. Then, as \( \lambda_2 \to \infty \), we can still get reasonable estimates for the sampling weight due to the fact that the density ratio function \( r^*(x) \) is the maximizer of \( Q(r) \) in (3.1), which can be unbiased estimated by \( Q_A(\gamma) \) in (3.2). Numerical results demonstrate the superior performance of the new objective function compared with Wong and Chan (2018); see Section 5 for details.

By the representer theorem, the solution of the inner optimization of (3.3) lies in the space spanned by \( \{K(x_i, \cdot) : i \in A \cup B\} \), where \( K(x, y) \) is the kernel function associated with the RKHS \( \mathcal{H} \). We can adopt a similar procedure as in Section 2.3 of Wong and Chan (2018) to solve the optimization problem (3.3); see Section S2 of the Supplementary Material for details. Once \( \{\hat{r}_i : i \in A\} \) are obtained, we get \( \hat{\omega}_i = 1 + (Nn_A^{-1} - 1)\hat{r}_i \) for
\( i \in A \), and the parameter of interest \( \hat{Y}_N \) can be estimated by

\[
\hat{Y}_N = N^{-1} \sum_{i \in A} \hat{\omega}_i y_i.
\] (3.5)

Asymptotic properties of the proposed estimator \( \hat{Y}_N \) in (3.5) are discussed in the next section.

**Remark 3.** Hebert-Johnson et al. (2018) and Kim et al. (2022) proposed a multicalibration framework to estimate \( m(x) \) in (2.1), and they showed that their method can be applied to analyze different target (sub-)populations. Even though their methods are not discussed under non-probability sampling, they essentially use uniform calibration. Different from Hebert-Johnson et al. (2018) and Kim et al. (2022), our proposed method can be regarded as multitask-oriented. Although we explicitly propose a regression model in (2.1), the response of interest is not involved in the objective function (3.3). Thus, a single set of the estimated sampling weights \( \{\hat{\omega}_i : i \in A\} \) can be applied to different \( Y \) variables and the internal consistency among survey estimates can be achieved in the non-probability sample.

### 4 Asymptotic theory

Since \( \mathcal{X} \) is compact, we set \( \mathcal{X} = [0,1]^d \) for simplicity. A Sobolev space is commonly used when the underlying regression function is smooth, and by Proposition 12.31 of Wainwright (2019), we consider a tensor product RKHS \( \mathcal{H} = \bigotimes_{j=1}^d \mathcal{H}_j \), where \( \mathcal{H}_j \) is an \( l \)-th order Sobolev space,

\[
W^{l,2}[0,1] = \{ f : f, f^{(1)}, \ldots, f^{(l-1)} \text{ are absolutely continuous, } f^{(l)} \in L^2([0,1]) \},
\]

and \( f^{(k)} \) is the \( k \)-th derivative of a function \( f \) for \( k = 1, \ldots, l \). The corresponding reproducing kernel of \( \mathcal{H} \) is \( K(z_1, z_2) = \prod_{j=1}^d K_s(z_{1j}, z_{2j}) \), where \( z_i = (z_{i1}, \ldots, z_{id})^T \in [0,1]^d \) for
$i = 1, 2$, and $K_s(\cdot, \cdot)$ is the reproducing kernel of $W^{l,2}[0, 1]$ \cite{Wahba1990} Section 1.2. See Section 12.2 of \cite{Wainwright2019} for discussion about other reproducing kernels. For $u \in \mathcal{H}$, we also assume $\|u\|_{\mathcal{H}}^2 < \infty$ for the sequential analysis.

To investigate the theoretical properties of the proposed method, we adopt the asymptotic framework of \cite{IsakiFuller1982} and consider a sequence of finite populations and samples. Besides, we make the following additional assumptions.

**A4.** The true regression function $m \in \mathcal{H}$ and $d/l < 2$.

**A5.** There exist positive constants $C_{\sigma,1} < C_{\sigma,2}$, $\delta$ and $C_{\delta}$ with respect to $N$, such that $C_{\sigma,1} \leq \sigma_i^2 \leq C_{\sigma,2}$ and $E\{|\epsilon_i|^{2+\delta}\} < C_{\delta}$ for $i = 1, \ldots, N$. Besides, the errors terms $\{\epsilon_i : i = 1, \ldots, N\}$ are independent with the sampling indicators $\{\delta_{B,i} : i = 1, \ldots, N\}$ for the probability sample $B$.

**A6.** The rejective sampling design satisfies $N^{-1} \sum_{i=1}^{N} (\delta_{B,i} \pi_{B,i}^{-1} - 1) y_i = O_p(n_B^{-1/2})$.

**A7.** There exist positive constants $C_{B,1} \leq C_{B,2}$ with respect to $n_B$ and $N$, such that $C_{B,1} \leq \pi_{B,i} N n_B^{-1} \leq C_{B,2}$ for $i = 1, \ldots, N$. Besides, $n_B N^{-1} = o(1)$ and $N^{1/2} n_B^{-1} = o(1)$.

**A8.** There exists a positive constant $M^*$ such that $\|r^*\|_{\mathcal{H}} \leq M^*$.

In Assumption A4 we assume that the true regression function $m$ lies in $\mathcal{H}$, so it can be well approximated by a certain function in $\mathcal{H}$. Besides, the assumption $d/l < 2$ regulates the complexity of the RKHS to guarantee theoretical properties of the proposed method; see Lemma S6 of \cite{WongChan2018} for details. The first part of Assumption A5 is a common condition to show the limiting distribution of the proposed estimator. Since the responses of interest $\{y_1, \ldots, y_N\}$ are not available in the probability sample $B$, it is reasonable to postulate independence between the response of interest and the sampling
indicators for the probability sample \( B \) in the second part of Assumption A5. Assumption A6 guarantees the convergence rate of the estimator \( N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i} y_i \), and it is also a common assumption in survey sampling; see Section 1.3.2 of Fuller (2009b) for details. Assumption A7 is widely used to regulate \( \pi_{B,i} \); see Theorem 1.3.5 of Fuller (2009b) and condition C5 of Chen et al. (2020) for details. Rather than assuming that \( n_B \) has asymptotically the same order as \( N \) as in Breslow and Wellner (2007), Han and Wellner (2021) and other references on empirical process for survey sampling, we make a more practically reasonable assumption \( n_B = o(N) \) for a probability sample in Assumption A7. The technical condition \( N^{1/2} n_B^{-1} = o(1) \) guarantees the convergence rate in Lemma 1 below; see Lemma S5 in Section S3 of the Supplementary Material for details. In Assumption A7, we implicitly assume that \( n_B \) is non-stochastic, and we should use \( o_p(1) \) instead if such an assumption fails. Even though we can show that \( r^*(x) \) is bounded by Assumption A3, Assumption A8 is a condition on its smoothness, and a similar condition is also implicitly assumed in (10) of Nguyen et al. (2010).

**Lemma 1.** Suppose that Assumptions A1–A4 and Assumption A7 hold. Then, there exists a positive constant \( c \), such that for all \( T \geq c \),

\[
P\left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_N(\gamma^*, u)}{\|u\|_{d_{\mathcal{H}}}} \geq T^2 \right\} \leq c \exp\left(-\frac{16T^2}{c^2}\right),
\]

where \( \gamma^* = (r_{1}^*, \ldots, r_{N}^*)^{T} \) and \( \tilde{H}_N = \{ u \in \mathcal{H} : \|u\|_2 = 1 \} \).

The proof of Lemma 1 is relegated to Section S3 of the Supplementary Material. Lemma 1 is a counterpart of Lemma S1 of Wong and Chan (2018), and it establishes the convergence rate of \( S_N(\gamma^*, u) \) when the true density ratios \( \{ r_i^* : i = 1, \ldots, N \} \) are available. In addition, it serves as a building block to investigate the consistency and the limiting distribution of the proposed estimator. Rather than assuming the availability of
\{ x_i : i = 1, \ldots, N \} \) in \cite{Wong:Chan:2018}, we consider the case when only a probability sample \( \{ x_i : i \in B \} \) is available. Besides, under rejective sampling, the sampling indicators \( \{ \delta_{B,i} : i = 1, \ldots, N \} \) are negatively associated, so we develop a different proof to incorporate the design features from the probability sample \( B \).

**Theorem 1.** Suppose that Assumptions A1–A8, \( \lambda_1 \approx n_B^{-1} \) and \( \lambda_2 \approx n_B^{-1} \) hold. Then, we have
\[
N^{-1} \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - 1) y_i = O_p(n_B^{-1/2}).
\]

Theorem 1 establishes the consistency of the estimator in (3.5), and its proof is in Section S4 of the Supplementary Material. By Theorem 1, \( \hat{Y}_N \) in (3.5) achieves a parametric convergence rate \( O_p(n_B^{-1/2}) \), even though we do not assume any parametric models for \( m(x) \) in (2.1) and the selection mechanism for the non-probability sample \( A \). The proposed estimator in (3.5) is consistent by Theorem 1, but it is hard to derive an unbiased variance estimator for it. Instead, we propose to use the bootstrap variance estimator discussed in \cite{Kim:et:al:2019}; see Section S5 of the Supplementary Material for details.

**Theorem 2.** Suppose that Assumptions A1–A8 hold. Let \( h = \hat{m} - m \in \mathcal{H} \) such that \( \|h\|_{\mathcal{H}} = O_p(1), \|h\|_2 = o_p(1), \lambda_2\|h\|_2 = o_p(n_B^{-1}), \lambda_1 = o(n_B^{-1}) \) and \( \lambda_1^{-1}\|h\|_2^{2(2-d)/d} = o_p(n_B) \), where \( \hat{m} \) is a kernel estimator of \( m \). Then, we have
\[
B_N^{-1} \left\{ \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) y_i - \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) \hat{m}(x_i) \right\} \to N(0,1)
\]
in distribution, where \( B_N^2 = \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1})^2 \sigma_i^2 \). In addition, \( B_N \approx N^2 n_B^{-1/2} \) in probability.

Theorem 2 establishes the limiting distribution for the proposed method, and its proof is relegated to Section S6 of the Supplementary Material. We have validated the convergence
rate of \(\hat{Y}_N\) in Theorem 1, but it is hard to get its limiting distribution. Instead, we propose the following “calibrated” estimator,

\[
\hat{Y}_{\text{prop}} = N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} \hat{m}(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \{y_i - \hat{m}(x_i)\}. \tag{4.1}
\]

By Theorem 2 and the fact that \(N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} y_i\) is design-unbiased for \(\bar{Y}_N\), we conclude that \(\hat{Y}_{\text{prop}}\) is asymptotically unbiased. The proposed estimator \(\hat{Y}_{\text{prop}}\) is similar to a doubly robust estimator (Chen et al., 2020), but it is more attractive since that we do not make any model assumption for the regression model \(m(x)\) in (2.1) and the response model \(\pi_A(x)\).

The corresponding variance estimator of \(\hat{Y}_{\text{prop}}\) is established in the following corollary.

**Corollary 1.** Suppose that the assumptions in Theorem 2 hold. Then, a plug-in variance estimator of \(\hat{Y}_{\text{prop}}\) in (4.1) is

\[
\hat{V}_{\text{prop}} = \hat{V} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} \hat{m}(x_i) \right\} + N^{-2} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i^2 \{y_i - \hat{m}(x_i)\}^2,
\]

where \(\hat{V} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} z_i \right\}\) is a design-based variance estimator of \(N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} z_i\) for a fixed sequence \(\{z_1, \ldots, z_N\}\).

The proof of Corollary 1 is relegated to Section S7 of the Supplementary Material. The first term of \(\hat{V}_{\text{prop}}\) estimates the variability due to probability sampling, and the second term estimates \(N^{-2} B_N^2\) in Theorem 2.

## 5 Simulation study

In this section, the performance of the proposed estimator (3.5) is compared with its alternatives in terms of estimating the population mean \(\bar{Y}_N\). The finite population \(\{(y_i, x_i) : i = 1, \ldots, N\}\) and the two samples \(A\) and \(B\) are generated by the following setups.
Linear. $m(x_i) = 10 + 2x_{1i} + 2x_{2i}$ and $\epsilon_i \sim N(0, 1)$, where $x_i = (x_{1i}, x_{2i})^T$, $x_{1i} = z_{1i}$, $x_{2i} = 0.3x_{1i} + z_{2i}$, $z_{ki} = 2(\xi_{ki} - 0.5)$ for $k = 1, 2$, $\xi_{ki} \sim \text{Beta}(3, 3)$, and $\text{Beta}(\alpha, \beta)$ is a beta distribution with two shape parameters $\alpha$ and $\beta$. A non-probability sample $A$ is generated by Assumptions 1–2, where $\pi_A(x_i) \propto \{m(x_i) - m_{\text{min}} + 0.25\}$, $\sum_{i=1}^N \pi_A(x_i) = n_A$, $m_{\text{min}} = \min\{m(x_i) : i = 1, \ldots, N\}$, and $n_A$ is the expected size of the non-probability sample $A$.

Nonlinear. $m(x_i) = 3 + 2z_{1i} + z_{2i}$ and $\epsilon_i \sim N(0, 0.5^2)$, where $x_i = (x_{1i}, x_{2i})^T$, $x_{1i} = |z_{1i}| \exp(-z_{1i})$, $x_{2i} = |z_{2i}| \exp(z_{2i})$, and $z_{1i}$ and $z_{2i}$ are independently generated by a truncated normal distribution restricted on the interval $[-3, 3]$ with mean zero and standard deviation one. The response probability of the non-probability sample $A$ is logit $\{c_A \pi_A(x_i; \theta_0)\} = 1 - 0.8z_{1i} - 0.8z_{2i}$, where $c_A$ is chosen such that $\sum_{i=1}^N \pi_A(x_i; \theta_0) = n_A$.

For each setup, we conduct Poisson sampling to generate a probability sample $B$, and the corresponding including probability satisfies $\pi_{B,i} \propto \log\{m(x_i) - m_{\text{min}} + 2\}$ and $\sum_{i=1}^N \pi_{B,i} = n_{B0}$, where $n_{B0}$ is the expected size of the probability sample $B$. The linear model setup is similar to [Chen et al. (2020)], but the selection mechanism for the non-probability sample $A$ is not based on a logistic regression model. A nonlinear regression model is considered in the second setup, and it is similar to [Wong and Chan (2018)]. Even though we adopt a logistic regression model for the selection mechanism of the non-probability sample $A$, it is not linear in $x_i$.

We consider $(N, n_{A0}, n_{B0}) \in \{(5000, 1000, 100), (10000, 2000, 200)\}$, and the following estimators are compared:

1. Naive sample mean (NSM): $\hat{Y}_{NSM} = \frac{n_A^{-1}}{N} \sum_{i \in A} y_i$.

2. Quasi-randomization estimator [Elliott and Valliant, 2017] with $N$ known (EV1):

   $\hat{Y}_{EV1} = N^{-1} \sum_{i \in A} \tilde{w}_i y_i$, where $\tilde{w}_i = \tilde{d}_i \tilde{p}_i$, $\tilde{d}_i$ is obtained by a linear regression model
for $d_{B,i}$ against $x_i$ based on the probability sample $B$, \( \hat{p}_i = \hat{P}(\delta_{A,i} \mid x_i, \delta_{A,i} + \delta_{B,i} \geq 1) / \hat{P}(\delta_{B,i} \mid x_i, \delta_{A,i} + \delta_{B,i} \geq 1) \), and \( \hat{P}(\delta_{A,i} \mid x_i, \delta_{A,i} + \delta_{B,i} \geq 1) \) and \( \hat{P}(\delta_{B,i} \mid x_i, \delta_{A,i} + \delta_{B,i} \geq 1) \) are estimated by a logistic regression model; see \cite{ElliottValliant2017} and \cite{KimShao2022} for details.

3. Quasi-randomization estimator (Elliott and Valliant, 2017) with $N$ estimated (EV2):
   \[ \hat{Y}_{EV2} = \left( \sum_{i \in A} \tilde{w}_i \right)^{-1} \sum_{i \in A} \tilde{w}_i y_i, \]
   where \( \tilde{w}_i \) is estimated in the same way as EV1.

4. Doubly robust estimator (Chen et al., 2020) with $N$ known (DR1); see Section S8 of the Supplementary Material for details.

5. Doubly robust estimator (Chen et al., 2020) with $N$ estimated (DR2).

6. HT estimator (3.5) with the proposed penalty (HTKL).

7. Balancing estimator of Wong and Chan (2018) adapted to survey sampling (BSS).
   That is, instead of using the KL-divergence as the penalty term, we consider
   \[ \hat{\gamma} = \arg \min_{\xi_1 \leq r_i \leq \min\{\xi_2, CN\}} \left[ \sup_{u \in \mathcal{H}} \left\{ S(\gamma, u) / \|u\|_2^2 - \lambda_1 \|u\|_\infty^2 \right\} + \lambda_2 Q_2(\gamma) \right], \quad (5.1) \]
   where \( Q_2(\gamma) = n_A^{-1} \sum_{i \in A} \{1 + (Nn_A^{-1} - 1)r_i\}^2 \).

8. Proposed estimator in (4.1) (Prop).

For the two doubly robust estimators, we make an assumption as in \cite{Chenetal2020} that the underlying regression model $m(x)$ is linear in the auxiliary vector $x$ and the response model $\pi_A(x)$ is logistic. It is worth pointing out that the BSS estimator has not been proposed by other researchers yet, and we use an $l_2$ penalty for the sampling weights in the proposed method for comparison.

We conduct $M = 1000$ Monte Carlo simulations for each model setup, and Figure 1 shows the Monte Carlo bias of the corresponding estimators, where the Monte Carlo bias
is $\hat{Y}_N^{(m)} - \bar{Y}_N^{(m)}$ for $m = 1, \ldots, M$. $\hat{Y}_N^{(m)}$ is a specific estimator with respect to the $m$-th Monte Carlo simulation, and $\bar{Y}_N^{(m)}$ is the corresponding finite population mean. Regardless of model setups, NSM is biased since it fails to incorporate the selection mechanism for the non-probability sample. When the regression model is correctly specified, EV2 and DR2 with population size estimated are more efficient than the others. HT_KL is slightly more efficient than EV1, DR1, BSS and Prop, but it has a positive bias, especially when the size of the non-probability sample is large. Prop is nearly as efficient as EV1, DR1 and BSS. However, when the regression model $m(x)$ and the response model $\pi_A(x)$ are wrongly specified, all estimators other than BSS and Prop are biased, regardless of the sample sizes. A similar phenomenon for the doubly robust estimators was also discussed by [Kang and Schafer (2007)]. Besides, the efficiency gain by EV2 and DR2 is less compared with their counterparts. Although we have established the consistency of HT_KL in Theorem 1, its finite sample performance is questionable when the true model is complex. On the contrary, both BSS and Prop are unbiased. Compared with BSS, Prop is slightly more efficient, especially for the nonlinear model setup.

We also compare the computation efficiency of BSS and Prop, and the average computation time required by each estimator is shown in Table 2. Regardless of the model setups, the proposed estimator (4.1) is more computationally efficient than BSS.

The coverage rates of the interval estimator with 95% confidence level is also investigated for Prop, and we use Corollary 1 to estimate its variance. Specifically, under Poisson sampling, a plug-in variance estimator is

$$\hat{V}_{prop} = N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-2} (1 - \pi_{B,i}) \hat{m}(x_i)^2 + \hat{\sigma}^2 N^{-2} \sum_{i=1}^{N} \delta_{A,i} \hat{u}_i^2,$$

where $\hat{m}(x)$ is obtained by the generalized additive model [Wood (2003)], and $\hat{\sigma}^2$ is the sample variance of $\{y_i - \hat{m}(x_i) : i \in A\}$. The coverage rates are close to 0.95 in different
Figure 1: Boxplots for the Monte Carlo bias of different estimators under different setups.

The horizontal dashed line corresponds to no bias.

model setups, especially when the sample sizes are large.

Remark 4. Although the performance of HT_KL is questionable under the nonlinear model setup, we still consider the performance of its bootstrap variance estimator, and the number of bootstrap replication is $B = 200$. We relegate the simulation results to Section S9 of the Supplementary Material, and its performance is satisfactory in terms of relative bias.

Table 2: Computation efficiency of BSS and Prop in terms of average computation time based on 1000 Monte Carlo simulations (unit: second).

| Sizes           | Linear | Nonlinear |
|-----------------|--------|-----------|
|                 | BSS    | Prop      | BSS    | Prop    |
| $(5000, 1000, 100)$ | 35.84  | 7.26      | 31.11  | 5.22    |
| $(10000, 2000, 200)$ | 66.59  | 31.03     | 137.69 | 42.47   |
Table 3: Coverage rate of the interval estimator with 95% confidence level based on 1 000 Monte Carlo simulations.

| Model     | (5 000, 1 000,100) | (10 000, 2 000,200) |
|-----------|--------------------|---------------------|
| Linear    | 0.960              | 0.959               |
| Nonlinear | 0.967              | 0.966               |

6 Application

We compare the performance of the proposed estimator and its alternatives based on a non-probability sample \( A \) from the National Health Insurance Sharing Service (NHISS) and a probability sample \( B \) from the Korea National Health and Nutrition Examination Survey (KNHANES). National Health Insurance was implemented in 1963 by the Medical Insurance Act, and whole Korean citizens are virtually enrolled in building a healthcare system; see Choi et al. (2015), Jee and Kim (2019) and the references within for details. KNHANES, on the other hand, is a national survey conducted by the Korea Centers for Disease Control and Prevention since 1998, and it is mainly adopted to assess the health and nutrition status of Korean citizens and provide health-related statistics in Korea. Therefore, the sample of KNHANES is nationally representative, and health-related information, including socioeconomic status, quality of life, health-related behaviors, and healthcare utilization, has been collected; see Kweon et al. (2014) and the references within for details.

In this section, a non-probability sample \( A \) contains \( n_A = 20\,000 \) elements randomly selected from an NHISS dataset. Demographic information, including age and gender, and health-related information, such as total cholesterol (mg/dL), hemoglobin (HGB), triglyceride (TG), and high-density lipoprotein cholesterol (HDL, mg/dL), is available. The probability sample \( B \) is a subset of the blood test results in the 2014 KNHANES, and it was
obtained by a multi-stage clustered probability design with sample size \( n_B = 4,929 \). The probability sample \( B \) contains the health-related information as that in the non-probability sample. We are interested in estimating the average total cholesterol for different age and gender groups by incorporating information from the two samples.

Even though the average total cholesterol can be estimated by \( \hat{N}^{-1} \sum_{i \in B} \pi_{B,i}^{-1} y_i \) with \( \hat{N} = \sum_{i \in B} \pi_{B,i}^{-1} \), we treat it as unavailable and use it as a benchmark to evaluate the performance of different methods, where \( y_i \) is the total cholesterol for the \( i \)th person. That is, we only assume \( \{(x_i, y_i) : i \in A\} \) and \( \{(x_i, \pi_{B,i}) : i \in B\} \) are available, where \( x_i \) contains the covariates, including HGB, TG and HDL. The population size \( N \) is not available, and we use \( \hat{N} \) instead. Both samples can be categorized into three age groups, including 20–40, 40–60, and more than 60 years old. Table 4 summarizes the marginal means of the covariates within each age and gender group, and we conclude that there exists a difference for the covariates in the two samples.

For each age and gender group, consider a regression model (2.1) for the proposed method, and the corresponding benchmark is \( \hat{Y}_{BM} = (\sum_{i \in D} \pi_{B,i}^{-1})^{-1} \sum_{i \in D} \pi_{B,i}^{-1} y_i \), where \( D \subset B \) consists of elements in the group. We also consider NSM, EV2 and DR2 in Section 5 for comparison, and different methods are evaluated by the estimation error \( \hat{Y} - \hat{Y}_{BM} \), where \( \hat{Y} \) is a specific estimator.

Figure 2 summarizes the estimation errors of different methods for each age and gender group, and we can reach the following conclusions. NSM overestimates the average total cholesterol for each age and gender group, and its performance is questionable. Even though HT_KL performs better than NSM, it is still much worse than EV2, DR2, BSS and Prop. Prop and BBS perform at least as well as EV2 and DR2 in all groups, and they outperform EV2 and DR2 for some groups. For example, in the female group with age 20–40,
Table 4: Marginal means of covariates for the non-probability sample $A$ and the probability sample $B$ for different domains. “20–40” stands for the group with age between 20 and 40, “40–60” for the group with age between 40 and 60, and “60+” for the group with age more than 60.

| Gender | Covariate | 20–40 | 40–60 | 60+ |
|--------|-----------|-------|-------|-----|
|        |           | $A$   | $B$   | $A$ | $B$ |
| Female | HDL       | 63.82 | 57.28 | 59.74 | 54.72 | 55.20 | 49.90 |
|        | TG        | 83.72 | 89.16 | 110.68 | 118.31 | 128.28 | 137.86 |
|        | HGB       | 12.95 | 13.06 | 12.96 | 13.24 | 12.89 | 13.18 |
| Male   | HDL       | 53.14 | 48.19 | 51.78 | 47.12 | 51.09 | 46.95 |
|        | TG        | 147.21 | 160.74 | 162.87 | 184.98 | 133.50 | 141.65 |
|        | HGB       | 15.47 | 15.68 | 15.18 | 15.41 | 14.39 | 14.58 |

estimation errors of Prop and BSS are less than EV2 and DR2, and a similar observation holds for the male group with age 20–40.

7 Concluding remarks

We propose a uniform function calibration method to estimate the sampling weights of a non-probability sample based on a probability sample, which is generated by a rejective sampling design. Compared with existing methods, the proposed method does not make any parametric assumption either for the regression model or the response model, so it can be widely adopted in practice. Besides, different from existing works, a KL-divergence-
Figure 2: Estimation error of different methods for each age and gender group. “NSM” is the naive sample mean estimator using the non-probability sample, “EV2” is the method considered by Elliott and Valliant (2017), “DR2” is the one proposed by Chen et al. (2020), “HT_KL” is the estimator in (3.5), “BSS” is the balancing estimator of Wong and Chan (2018) adopted to survey sampling, and “Prop” is the proposed method. Based penalty is proposed to improve the performance of the proposed method. Consistency and the asymptotic normality of the proposed estimator are established under regularity conditions. Numerical results show that the proposed method outperforms its alternatives, especially when both regression and response models are wrongly specified. The proposed method can be viewed as a “soft” calibration method, since we do not require that $S(\hat{\gamma}, u) = 0$ holds for every $u \in \mathcal{H}$. In survey sampling, however, we may would like to achieve “hard” calibration for certain functions of the covariates, it would be an interesting project to incorporate a “hard” calibration component in the proposed objective function.

References

Athreya, K. B. and Lahiri, S. N. (2006). Measure Theory and Probability Theory. Springer, New York.
Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2):90–143.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):445–458.

Bertail, P., Chautru, E., and Clémençon, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. *Scandinavian Journal of Statistics*, 44(1):97–111.

Bertail, P. and Clémençon, S. (2016). Sharp exponential inequalities in survey sampling: conditional Poisson sampling schemes. *arXiv:1610.03776*, pages 1–25.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2):161–188.

Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2017). Functional central limit theorems for single-stage sampling designs. *Annals of Statistics*, 45(4):1728 – 1758.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102.
Brick, J. M. (2015). Compositional model inference. In Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, pages 299–307, American Statistical Association, Alexandria, VA.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532):2011–2021.

Choi, Y., Kim, J.-H., Yoo, K.-B., Cho, K. H., Choi, J.-W., Lee, T. H., Kim, W., and Park, E.-C. (2015). The effect of cost-sharing in private health insurance on the utilization of health care services between private insurance purchasers and non-purchasers: A study of the Korean health panel survey (2008–2012). BMC Health Services Research, 15(1):1–11.

Conti, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. Sankhya B, 76(2):234–259.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. Public Opinion Quarterly, 64(4):464–494.

Couper, M. P. and Miller, P. V. (2008). Web survey methods: Introduction. Public Opinion Quarterly, 72(5):831–835.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? In Survey Research Methods, volume 2, pages 47–60.

Dever, J. A. and Valliant, R. (2014). Estimation with non-probability surveys and the question of external validity. In Proceedings of Statistics Canada Symposium, pages 1–8.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87(418):376–382.
Elliott, M. N. and Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology, 33*(2):211–5.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science, 32*(2):249–264.

Fuller, W. A. (2009a). *Sampling Statistics.* Wiley, Hoboken.

Fuller, W. A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika,* 96(4):933–944.

Han, P. and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika,* 100(2):417–430.

Han, Q. and Wellner, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. *Annals of Statistics,* 49(1):459 – 485.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2nd edition.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science,* 32(2):206–226.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning,* pages 1939–1948.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47(260):663–685.
Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Jee, H. and Kim, J.-H. (2019). Gender difference in colorectal cancer indicators for exercise interventions: The National Health Insurance Sharing Service-derived big data analysis. *Journal of Exercise Rehabilitation*, 15(6):811–818.

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *Annals of Statistics*, 1(11):286–295.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). *Statistical Science*, 22(4):523–539.

Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):319–376.

Kim, J.-K., Rao, J. N. K., and Wang, Z. (2019). Hypotheses testing from complex survey data using bootstrap weights: A unified approach. *arXiv: 1902.08944*, pages 1–81.

Kim, J. K. and Shao, J. (2022). *Statistical Methods for Handling Incomplete Data*. CRC Press, Boca Raton, 2nd edition.

Kim, J. K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87(S1):S177–S191.

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., and Reingold, O. (2022). Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):1–6.
Kweon, S., Kim, Y., Jang, M.-j., Kim, Y., Kim, K., Choi, S., Chun, C., Khang, Y.-H., and Oh, K. (2014). Data resource profile: The Korea national health and nutrition examination survey (KNHANES). *International Journal of Epidemiology*, 43(1):69–77.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2):329–349.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–343.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100:1429–1442.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

O’Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(2):195–210.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings*, pages 1–26, American Statistical Association, Alexandria, VA.
Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer, New York.

Sugiyama, M., Suzuko, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York.

Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013). *The Science of Web Surveys*. Oxford University Press, New York.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137.

van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge University Press.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, Cambridge.

Wong, R. K. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.
Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Wu, C. and Thompson, M. E. (2020). *Sampling Theory and Practice*. Springer, Gewerbestrasse.

Yuan, D.-M., Wei, L.-R., and Lei, L. (2014). Conditional central limit theorems for a sequence of conditional independent random variables. *Journal of the Korean Mathematical Society*, 51(1):1–15.
Supplemental Material for “Functional Calibration under Non-Probability Survey Sampling”

S1 Brief introduction to RKHS

A symmetric bivariate function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a positive semidefinite kernel function, if for all integer $n \geq 1$ and elements $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the $n \times n$ matrix $M$ is positive semidefinite, where $K(x_i, x_j)$ serves as its $(i, j)$-th entry for $i, j = 1, \ldots, n$; see Definition 12.6 of [Wainwright (2019)] for details.

Let $K(x, y)$ be a positive semidefinite kernel function, and consider a functional space $\mathcal{H}^\dagger = \left\{ f : f(\cdot) = \sum_{i=1}^{n} \alpha_j K(\cdot, x_j) \text{ for some } n \geq 1, \{\alpha_1, \ldots, \alpha_n\} \subset \mathbb{R}, \{x_1, \ldots, x_n\} \subset \mathcal{X} \right\}$. Then, by Theorem 12.11 of [Wainwright (2019)], the complement of $\mathcal{H}^\dagger$, say $\mathcal{H}$, is an RKHS with the reproducing kernel $K(x, y)$.

Furthermore, suppose that the kernel function $K(x, y)$ has the following eigen-decomposition:

$$K(x, y) = \sum_{j=1}^{\infty} \mu_j \psi_j(x)\psi_j(y),$$

where $\{\mu_j : j = 1, 2, \ldots\}$ are non-negative eigenvalues satisfying $\sum_{j=1}^{\infty} \mu_j^2 < \infty$, and $\{\psi_j(x) : j = 1, 2, \ldots\}$ are the corresponding eigenfunctions. Then, for any $f \in \mathcal{H}$, there exist $\{c_j : j = 1, 2, \ldots\}$ such that

$$f(x) = \sum_{j=1}^{\infty} c_j \psi_j(x),$$

and the corresponding norm associated with $\mathcal{H}$ is defined as

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} c_j^2 / \mu_j.$$
S2  Numerical solution of the optimization problem

Consider $u(x) = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$, and denote $u = M\alpha$, where $u = (u(x_1), \ldots, u(x_n))^T$, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, and $M$ is an $n \times n$ Gram matrix with $(i,j)$-th element being $K(x_i, x_j)$. Assume that the eigen-decomposition of $M$ is

$$M = \begin{pmatrix} P_1 & P_2 \end{pmatrix} \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} P_1^T \\ P_2^T \end{pmatrix}, \quad (S2.1)$$

where $Q_1$ is a diagonal matrix consisting of the positive eigenvalues, and $Q_2$ is a zero matrix. Notice that $Q_2$ may be a null matrix. Then, $S(\gamma, u) = N^{-2} \alpha^T M A(\gamma) M \alpha$ and $\|u\|_2^2 = n^{-1} \alpha^T M^2 \alpha$, where $A(\gamma) = w(\gamma) w(\gamma)^T$, and $w(\gamma) = (w_1(\gamma), \ldots, w_n(\gamma))^T$. Denote $\beta = n^{-1/2} Q_1 P_1^T \alpha$, the inner optimization problem becomes

$$\sup_{\beta: \|\beta\|_2 \leq 1} \beta^T \left\{ \frac{n}{N^2} P_1^T A(\gamma) P_1 - n \lambda_1 Q_1^{-1} \right\} \beta. \quad (S2.2)$$

Then, we can use a similar procedure in Section 2.3 of [Wong and Chan (2018)] to solve the optimization problem (3.3).

S3  Proof of Lemma 1

First, we present a definition of negative association as in Definition 2.1 of Joag-Dev and Proshan [1983].

**Definition 1.** Random variables $X_1, \ldots, X_N$ are said to be negatively associated if for every pair of disjoint subsets $A_1, A_2$ of $\{1, \ldots, N\},$

$$\text{Cov}\{f_1(X_i, i \in A_1), f_2(X_i, i \in A_2)\} \leq 0, \quad (S3.1)$$
where \( f_1 \) and \( f_2 \) are increasing functions in every variable.

It can be easily shown that if both \( f_1 \) and \( f_2 \) are decreasing functions in every variable, we still get (S3.1) for negatively associated random variables since \(-f_1\) and \(-f_2\) are increasing and \( \text{Cov}\{-f_1(X_i, i \in A_1), -f_2(X_i, i \in A_2)\} = \text{Cov}\{f_1(X_i, i \in A_1), f_2(X_i, i \in A_2)\} \).

**Lemma S1.** Suppose that Assumption A7 holds. Then, we have

\[
E\{\exp(W_{B,i}^2)\} - 1 \leq \sigma_0^2,
\]

uniformly for \( i = 1, \ldots, N \), where \( W_{B,i} = (\delta_{B,i} - 1) n_B N^{-1} \) and \( \sigma_0^2 = \exp\{\max\{1, (C_{B,2} - 1)^2, C_{B,1}\}\} - 1 \).

**Proof of Lemma S1.** Consider

\[
E\{\exp(W_{B,i}^2)\} = \pi_{B,i} \exp\{(\pi_{B,i}^{-1} - 1)^2 n_B^2 N^{-2}\} + (1 - \pi_{B,i}) \exp(n_B^2 N^{-2}) \leq \pi_{B,i} \exp(\max\{(C_{B,2} - 1)^2, C_{B,1}\}) + (1 - \pi_1) e \leq \exp(\max\{1, (C_{B,2} - 1)^2, C_{B,1}\}),
\]

where first inequality holds by Assumption A7. By (S3.2), we have proved Lemma S1.

The next lemma is a straightforward result from Definition so we omit its proof.

**Lemma S2.** For any \( m \geq 2 \) and mutually disjoint subsets \( A_1, \ldots, A_m \) of \( \{1, \ldots, N\} \) and negatively associated random variables \( X_1, \ldots, X_N \), we have

\[
E\left\{\prod_{k=1}^m f_k(X_i : i \in A_k)\right\} \leq \prod_{k=1}^m E\{f_k(X_i : i \in A_k)\},
\]

where \( f_1, \ldots, f_m \) are decreasing and non-negative functions in every variable.

The next lemma shows a Hoeffding’s inequality (van de Geer, 2000, Lemma 3.5) under rejective sampling.
Lemma S3. Suppose Assumption A7 holds. Then, there exist positive constants $C_{B,3}$ and $C_{B,4}$, such that for any $a \geq 0$ and $\{\gamma_i : i = 1, \ldots, N\} \subset \mathbb{R}$, we have

$$P \left( \left| \sum_{i=1}^{N} W_{B,i} \gamma_i \right| \geq a \right) \leq C_{B,3} \exp \left\{ -\frac{a^2}{C_{B,4} \sum_{i=1}^{N} \gamma_i^2} \right\}, \quad (S3.4)$$

where $W_{B,i}$ is defined in Lemma S1.

Proof of Lemma S3. Since the probability sample $B$ is generated by a rejective sampling design, the corresponding sampling indicators are negatively associated by Theorem 3 of Bertail and Clémençon (2016). Given the sequence $\{\gamma_i : i = 1, \ldots, N\}$, denote $I = \{i : \gamma_i \geq 0\}$. Then, for any positive constant $a$, we have

$$P \left( \left| \sum_{i=1}^{N} W_{B,i} \gamma_i \right| \geq a \right) \leq P \left( \left| \sum_{i \in I} W_{B,i} \gamma_i \right| \geq \frac{a}{2} \right) + P \left( \left| \sum_{i \notin I} W_{B,i} \gamma_i \right| \geq \frac{a}{2} \right)$$

$$\leq P \left( \sum_{i \in I} W_{B,i} \gamma_i \geq \frac{a}{2} \right) + P \left( \sum_{i \notin I} W_{B,i} \gamma_i \leq -\frac{a}{2} \right)$$

$$+ \exp \left( -\beta \sum_{i \notin I} W_{B,i} \gamma_i \right)$$

(S3.5)

Assume $P \left( \left| \sum_{i \in I} W_{B,i} \gamma_i \right| \geq a/2 \right) = 0$ if $I = \emptyset$ and $P \left( \left| \sum_{i \notin I} W_{B,i} \gamma_i \right| \geq a/2 \right) = 0$ if $I = \{1, \ldots, N\}$.

Without loss of generality, assume that $I \neq \emptyset$ and $I \neq \{1, \ldots, N\}$. For any $\beta > 0$, we have

$$P \left( \left| \sum_{i \in I} W_{B,i} \gamma_i \right| \geq \frac{a}{2} \right) \leq \exp \left( -\frac{\beta a}{2} \right) \prod_{i \in I} E \left\{ \exp (\beta W_{B,i} \gamma_i) \right\}$$

\begin{align*}
&\leq \exp \left( -\frac{\beta a}{2} \right) \prod_{i \in I} E \left\{ \exp (\beta W_{B,i} \gamma_i) \right\} \\
&\leq \exp \left\{ 2(1 + \sigma_0^2) \beta^2 \sum_{i \in I} \gamma_i^2 - \frac{\beta a}{2} \right\} \\
&\leq \exp \left\{ 2(1 + \sigma_0^2) \beta^2 \sum_{i=1}^{N} \gamma_i^2 - \frac{\beta a}{2} \right\}, \quad (S3.6)
\end{align*}
where the first inequality holds by Cramer’s inequality \cite{Athreya2006} Corollary 3.1.5, the second inequality holds by Property P2 of \cite{Joag-Dev1983} and the fact that $\beta \gamma_i > 0$ for $i \in \mathcal{I}$, the third inequality by Lemma 8.1 of \cite{vandeGeer2000} and Lemma S1, the last inequality holds since $2(1 + \sigma_0^2) > 0$, and $\sigma_0^2$ is defined in Lemma S1. If we set

$$\beta = \frac{a}{8(1 + \sigma_0^2) \sum_{i=1}^{N} \gamma_i^2},$$

by (S3.6), we have

$$P\left( \sum_{i \in \mathcal{I}} W_{B,i} \gamma_i \geq \frac{a}{2} \right) \leq \exp \left\{ - \frac{a^2}{32(1 + \sigma_0^2) \sum_{i=1}^{N} \gamma_i^2} \right\}. \quad (S3.7)$$

Next, for any $\beta > 0$, consider

$$P\left( \sum_{i \in \mathcal{I}} W_{B,i} \gamma_i \leq -\frac{a}{2} \right) = P\left( \sum_{i \in \mathcal{I}} W_{B,i} \tilde{\gamma}_i \geq \frac{a}{2} \right)$$

$$\leq \exp \left( -\beta \frac{a}{2} \right) E \left\{ \exp \left( \beta \sum_{i \in \mathcal{I}} W_{B,i} \tilde{\gamma}_i \right) \right\}$$

$$\leq \exp \left( -\beta \frac{a}{2} \right) \prod_{i \in \mathcal{I}} E \left\{ \exp \left( \beta W_{B,i} \tilde{\gamma}_i \right) \right\}$$

where $\tilde{\gamma}_i = -\gamma_i$, and the last inequality holds by Lemma S2 since $\beta \tilde{\gamma}_i \leq 0$ for $i \in \mathcal{I}$. Then, we can use a similar argument leading to (S3.7) to get

$$P\left( \sum_{i \in \mathcal{I}} W_{B,i} \gamma_i \leq -\frac{a}{2} \right) \leq \exp \left\{ - \frac{a^2}{32(1 + \sigma_0^2) \sum_{i=1}^{N} \gamma_i^2} \right\}. \quad (S3.8)$$

Besides, we can also get

$$P\left( \sum_{i \not\in \mathcal{I}} W_{B,i} \gamma_i \geq \frac{a}{2} \right) \leq \exp \left\{ - \frac{a^2}{32(1 + \sigma_0^2) \sum_{i=1}^{N} \gamma_i^2} \right\}, \quad (S3.9)$$

since $\gamma_i \leq 0$ for $i \not\in \mathcal{I}$.

$$P\left( \sum_{i \not\in \mathcal{I}} W_{B,i} \gamma_i \leq -\frac{a}{2} \right) = P\left( \sum_{i \not\in \mathcal{I}} W_{B,i} \tilde{\gamma}_i \geq \frac{a}{2} \right).$$

38
Then, we can use a similar procedure as (S3.6)–(S3.7) to verify
\[
P\left( \sum_{i \notin \mathcal{I}} W_{B,i} \gamma_i \leq -\frac{a}{2} \right) \leq \exp\left\{ -\frac{a^2}{32(1 + \sigma_0^2) \sum_{i=1}^{N} \gamma_i^2} \right\},
\] (S3.10)
since \( \tilde{\gamma}_i \geq 0 \) for \( i \notin \mathcal{I} \).

By (S3.5) and (S3.7)–(S3.10), we have proved Lemma S3 with \( C_{B,3} = 4 \) and \( C_{B,4} = 32(1 + \sigma_0^2) \).

Let \( \mathcal{H}_1 = \{ u \in \mathcal{H} : \|u\|_{\mathcal{H}} = 1 \} \). By Lemma S7 of Wong and Chan (2018), there exists a constant \( R \) such that
\[
\sup_{u \in \mathcal{H}_1} \|u\|_{\infty} \leq R.
\] (S3.11)
Denote \( H_\infty(\epsilon, \mathcal{H}_1) \) to be the uniform entropy for \( \mathcal{H}_1 \); see Definition 2.3 of van de Geer (2000) for details about the uniform entropy.

**Lemma S4.** Suppose Assumption A4 holds. There exists a constant \( C_{B,5} \), such that for any \( n_B \leq N \) and \( S \geq S_0 \), we have
\[
\sum_{s=S_0}^{S} 2^{-s} R H_{\infty}^{1/2}(2^{-s}(n_B N^{-1})^{1/2} R, \mathcal{H}_1) \leq C_{B,5} N^{1/2} n_B^{-1/2},
\] (S3.12)
where \( S_0 = \max\{s : R \leq 2^{-s} n_B N^{-1} \leq 2R\} \).

**Proof of Lemma S4.** By Assumption A4 and Lemma S6 of Wong and Chan (2018), there exists a constant \( C_\mathcal{H} \) such that for \( \epsilon > 0 \),
\[
H_\infty(\epsilon, \mathcal{H}_1) \leq C_\mathcal{H} \epsilon^{-d/4}.
\] (S3.13)
Consider

\[
\sum_{s=S_0}^{S} 2^{-s} R H_{\infty}^{1/2}(2^{-s}(n_B N^{-1})^{1/2} R, \mathcal{H}_1) = \left( N n_B^{-1} \right)^{1/2} \sum_{s=S_0}^{S} 2^{-s}(n_B N^{-1})^{1/2} R H_{\infty}^{1/2}(2^{-s}(n_B N^{-1})^{1/2} R, \mathcal{H}_1) \leq 2 \left( N n_B^{-1} \right)^{1/2} \int_0^{2R} H_{\infty}^{1/2}(\epsilon, \mathcal{H}_1) d\epsilon \leq 2 \left( N n_B^{-1} \right)^{1/2} C_{4,1}^{1/2}(2R)^{1-d/(2l)} \left( 1 - d/(2l) \right) = C_{B,5}(N n_B^{-1})^{1/2}, \quad (S3.14)
\]

where the second inequality holds by \((S3.13)\) and \(C_{B,5} = 2^{2-d/(2l)} C_{4,1}^{1/2} R^{1-d/(2l)} \{1 - d/(2l)\}^{-1}\).

Thus, we have proved Lemma \[\text{S4}\] by \((S3.14)\). \[\Box\]

**Lemma S5.** Suppose Assumption A4 and Assumption A7 hold. Then, for all \(0 \leq \epsilon < \delta\) and \(K > 1\), there exists \(N_0 = N_0(\delta, \epsilon)\), such that for \(N \geq N_0\), we have

\[
P \left[ \left\{ \sup_{u \in \mathcal{H}_1} \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i}u(x_i) \right| \geq \delta \right\} \bigcap \left\{ \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} \right| \leq K \right\} \right] \leq C_{B,6} \exp \left\{ - \frac{n_B(\delta - \epsilon)^2}{C_{B,6} R^2} \right\},
\]

where \(C_{B,6}\) only depends on \(\sigma^2_0\) defined in Lemma \[\text{S1}\].

**Proof of Lemma S5.** By \((S3.13)\), there exists a finite \(N_s\) such that \(\{ u_s^j : j = 1, \ldots, N_s \}\) is a minimal \(\{2^{-s}(n_B N^{-1})^{1/2} R\}\)-covering set of \(\mathcal{H}_1\) for \(s = S_0, S_0 + 1, \ldots, S\) in terms of the \(\|\cdot\|_{\infty}\) norm, where \(S_0 = \max\{s : R \leq 2^{-s} n_B N^{-1} \leq 2R\}\) and \(S = \min\{s \geq 1 : 2^{-s}(n_B N^{-1})^{1/2} R \leq \epsilon/(2K)\}\).

By Assumption A7 and Lemma \[\text{S4}\], there exists \(N_0 = N_0(\delta, \epsilon)\), such that when \(N \geq N_0\),

\[
n_B^{1/2}(\delta - \epsilon) \geq 12 C_{B,4}^{1/2} R \sum_{s=S_0+1}^{S} 2^{-s} H_{\infty}^{1/2}(2^{-s}(n_B N^{-1})^{1/2} R, \mathcal{H}_1) \vee \left\{ (1152 \log 2)^{1/2} C_{B,4}^{1/2} R \right\},
\]

where \(a \vee b = \max\{a, b\}\).
We adopt the notation convenience from Section 3.2 of van de Geer (2000) and index functions in \( \mathcal{H} \) by \( \Theta \): \( \mathcal{H}_1 = \{ u_\theta : \theta \in \Theta \} \). Then, for any \( u_\theta \in \mathcal{H}_1 \), there exists \( u_\theta^S \) such that \( \|u_\theta - u_\theta^S\|_\infty \leq \epsilon/(3K) \). Thus, we have

\[
\begin{align*}
&\left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} \left\{ u_\theta(x_i) - u_\theta^S(x_i) \right\} \right| \\
&\leq \frac{1}{n_B} \left| \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} \sum_{i=1}^{n_B} \left\{ u_\theta(x_i) - u_\theta^S(x_i) \right\} \right| + \frac{1}{N} \left| \sum_{i=1}^{N} \left\{ u_\theta(x_i) - u_\theta^S(x_i) \right\} \right| \\
&\leq \left\{ \frac{1}{n_B} \left( \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} \sum_{i=1}^{n_B} \right) + 1 \right\} \max_{i=1,\ldots,N} \left| \left\{ u_\theta(x_i) - u_\theta^S(x_i) \right\} \right| \\
&\leq \frac{K + 2}{3K} \epsilon \leq \epsilon
\end{align*}
\]

(S3.16)
on the event \( \{ n_B^{-1} \sum_{i=1}^{N} W_{B,i} \leq K \} \), where the first inequality holds by the definition of \( W_{B,i} \) in Lemma S1, the third inequality is due to the fact that the event \( \{ n_B^{-1} \sum_{i=1}^{N} W_{B,i} \leq K \} \) implies \( \{ n_B^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} n_B N^{-1} \leq K + 1 \} \), and the last inequality holds since \( K > 1 \).

Thus, it is enough to show the exponential inequality for

\[
\begin{align*}
P \left\{ \sup_{\theta \in \Theta} \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} u_\theta^S(x_i) \right| \geq \delta - \epsilon \right\}.
\end{align*}
\]

(S3.17)
Define \( u_\theta^{S_0} = 0 \) for \( u \in \mathcal{H}_1 \), and we have \( u_\theta^S = \sum_{s=S_0+1}^{S} (u_\theta^s - u_\theta^{s-1}) \). For any sequence \( \{ \eta_s : s = S_0 + 1, \ldots, S \} \) satisfying \( \sum_{s=S_0+1}^{S} \eta_s \leq 1 \), we have

\[
\begin{align*}
P \left[ \sup_{\theta \in \Theta} \left| \frac{1}{n_B} \sum_{i=1}^{N} \sum_{s=S_0+1}^{S} W_{B,i} \left\{ u_\theta^s(x_i) - u_\theta^{s-1}(x_i) \right\} \right| \geq \delta - \epsilon \right] \\
\leq \sum_{s=S_0+1}^{S} P \left[ \sup_{\theta \in \Theta} \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} \left\{ u_\theta^s(x_i) - u_\theta^{s-1}(x_i) \right\} \right| \geq \eta_s (\delta - \epsilon) \right] \\
\leq \sum_{s=S_0+1}^{S} C_{B,3} \exp \left\{ 2H_\infty (2^{-s} (n_B N^{-1})^{1/2} R, \mathcal{H}_1) - \frac{n_B \eta_s^2 (\delta - \epsilon)^2}{9C_{B,4} 2^{-2s} R^2} \right\},
\end{align*}
\]

(S3.18)
where the last inequality holds by Lemma S3 and

\[
\|u_\theta^s - u_\theta^{s-1}\|_\infty \leq \|u_\theta^s - u_\theta\|_\infty + \|u_\theta^{s-1} - u_\theta\|_\infty \\
\leq 2^{-s} (n_B N^{-1})^{1/2} R + 2^{-s+1} (n_B N^{-1})^{1/2} R \leq 3 \{2^{-s} (n_B N^{-1})^{1/2} R\}. 
\]

41
Now, we consider
\[ \eta_s = \frac{6RC_B^{1/2}2^{-s}H_\infty^{1/2}(2^{-s}(n_B N^{-1})^{1/2}R, \mathcal{H}_1)}{n_B^{1/2}(\delta - \epsilon)} \sqrt{2^{-s}S_0(s - S_0)^{1/2}} \]

Then, by \( \text{(S3.15)} \) and a similar argument in the proof of Lemma 3.2 of van de Geer (2000), we can show that
\[ \sum_{s=S_0+1}^{S} \eta_s \leq 1. \]
Thus, we can show that
\[ \sum_{s=S_0+1}^{S} C_{B,3} \exp \left\{ 2H_\infty(2^{-s}(n_B N^{-1})^{1/2}R, \mathcal{H}_1) - \frac{n_B \eta_s^2(\delta - \epsilon)^2}{9C_{B,4}2^{-2s}R^2} \right\} \]
\[ \leq \sum_{s=S_0+1}^{S} C_{B,3} \exp \left\{ -\frac{n_B \eta_s^2(\delta - \epsilon)^2}{18C_{B,4}2^{-2s}R^2} \right\}. \]

Thus, we can use a similar argument in the proof of Lemma 3.2 of van de Geer (2000) to conclude the proof of Lemma S5 by \( \text{(S3.15)} \) with \( C_{B,6} = \max\{2C_{B,3}, C_{B,4}\} \).

By Lemma S3 and setting \( K \) sufficient large, we have
\[ P \left( \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} \right| > K \right) \leq C_{B,6} \exp \left\{ -\frac{n_B(\delta - \epsilon)^2}{C_{B,6}R^2} \right\}, \quad \text{(S3.19)} \]
where the related quantities are defined in Lemma S5. Thus, by Lemma S5 and \( \text{(S3.19)} \), we have
\[ P \left\{ \sup_{u \in \mathcal{H}_1} \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i}u(x_i) \right| \geq \delta \right\} \]
\[ \leq P \left\{ \sup_{u \in \mathcal{H}_1} \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i}u(x_i) \right| \geq \delta \right\} \cap \left\{ \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} \right| \leq K \right\} + P \left( \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} \right| > K \right) \]
\[ \leq 2C_{B,6} \exp \left\{ -\frac{n_B(\delta - \epsilon)^2}{C_{B,6}R^2} \right\}. \quad \text{(S3.20)} \]
Thus, take \( C_{B,7} = 2C_{B,6} \) and \( \epsilon \) sufficiently small, we conclude
\[ P \left\{ \sup_{u \in \mathcal{H}_1} \left| \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i}u(x_i) \right| \geq \delta \right\} \leq C_{B,7} \exp \left\{ -\frac{n_B\delta^2}{C_{B,7}R^2} \right\}. \quad \text{(S3.21)} \]
Denote
\[ S_{N,A}(\gamma, u) = \left( N^{-1} \sum_{i=1}^{N} \left[ \delta_{A,i} \left\{ 1 + \left( \frac{N}{n_A} - 1 \right) r_i \right\} - 1 \right] u(x_i) \right)^2, \quad (S3.22) \]
and
\[ S_{N,B}(u) = \left\{ N^{-1} \sum_{i=1}^{N} (\delta_{B,i} \pi_{B,i}^{-1} - 1) u(x_i) \right\}^2. \quad (S3.23) \]

**Proof of Lemma 1.** By the fact that \( S_{N}(\gamma^*, u) \leq 2S_{N,A}(\gamma^*, u) + 2S_{N,B}(u) \), we have
\[
P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,A}(\gamma^*, u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\} \leq P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,B}(u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\}, \quad (S3.24)
\]
where \( S_{N,A}(\gamma, u) \) and \( S_{N,B}(u) \) are in (S3.22) and (S3.23). By Assumptions A1–A4, following a similar argument of Lemma S1 of Wong and Chan (2018), we can show
\[
P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,A}(\gamma^*, u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\} \leq C_{A,3} \exp \left( -\frac{16T^2}{C_{A,3}^2} \right), \quad (S3.25)
\]
where \( C_{A,3} \) is a constant. Since \( n_B \leq N \), we have
\[
P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,A}(r^*, u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\} \leq P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,A}(r^*, u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\}. \quad (S3.26)
\]
By (S3.25)–(S3.26), we have
\[
P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,A}(r^*, u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\} \leq C_{A,3} \exp \left( -\frac{16T^2}{C_{A,3}^2} \right). \quad (S3.27)
\]
Next, we investigate the second part on the right hand side of (S3.24). By
\[ S_{N,B}(u) = \left[ \frac{1}{n_B} \sum_{i=1}^{N} W_{B,i} u(x_i) \right]^2, \quad (S3.28) \]
where \( W_{B,i} \) is defined in Lemma S1.

By (S3.21) and a similar proof for Lemma 8.4 of van de Geer (2000), there exists a constant \( C_{B,8} \) such that
\[
P \left\{ \sup_{u \in \tilde{H}_N} \frac{n_B S_{N,B}(u)}{\|u\|_d^d} \geq \frac{T^2}{4} \right\} \leq C_{B,8} \exp \left( -\frac{16T^2}{C_{B,8}^2} \right). \quad (S3.29)
\]
By (S3.27) and (S3.29), we have validated Lemma 1 by setting \( c = C_{A,3} + C_{B,8} \). □
S4 Proof of Theorem 1

Recall $n = n_A + n_B$, and $\|u\|_2^2 = n^{-1} \sum_{i=1}^{n} u(x_i)^2$. Notice that we have assumed $\|u\|_H < \infty$ for $u \in \mathcal{H}$.

**Lemma S6.** Suppose Assumption A4 holds. Then, for $u \in \mathcal{H}$, we have

$$E(\|u\|_2^2) < \infty, \quad E(\|u\|_4^2) < \infty.$$  

**Proof of Lemma S6.** By Lemma S7 of Wong and Chan (2018), there exists a constant $R$ such that $\sup_{u \in \mathcal{H}_1} \|u\|_\infty \leq R$. Thus, for $u \in \mathcal{H}$, $\|u\|_\infty \leq R \|u\|_H$. Since we have assumed $\|u\|_H < \infty$ for $u \in \mathcal{H}$ in Section 4, we conclude that $\|u\|_\infty < \infty$, so we have proved Lemma S6.

Denote $u^* = \arg \max_{u \in \tilde{\mathcal{H}}_N} \{S(\gamma^*, u) - \lambda_1 \|u\|_H\}$, and its existence is shown in Appendix S2 where $\tilde{\mathcal{H}}_N = \{u \in \mathcal{H} : \|u\|_2 = 1\}$. Then, for any $u \in \mathcal{H}$, we have

$$S(\hat{\gamma}, u) - \lambda_1 \|u\|_H^2 - \lambda_2 Q_A(\hat{\gamma}) \|u\|_2^2 \leq \{S(\gamma^*, u^*) - \lambda_1 \|u^*\|_H^2 - \lambda_2 Q_A(\gamma^*)\} \|u\|_2^2. \quad (S4.1)$$

**Lemma S7.** Suppose Assumptions A3–A7 hold. If $\lambda_1 \asymp n_B^{-1}$ and $\lambda_2 \asymp n_B^{-1}$, we have

$$S(\hat{\gamma}, u) = O_p(n_B^{-1}) \|u\|_2^2$$

for $u \in \mathcal{H}$.

**Proof of Lemma S7.** By (S4.1), we have

$$S(\hat{\gamma}, u) + \lambda_1 \|u^*\|_H^2 \|u\|_2^2 + \lambda_2 Q_A(\gamma^*) \|u\|_2^2 \leq S(\gamma^*, u^*) \|u\|_2^2 + \lambda_1 \|u\|_H^2 + \lambda_2 Q_A(\hat{\gamma}) \|u\|_2^2. \quad (S4.2)$$

By Lemma 1, we can use a similar argument in the proof of Lemma S3 of Wong and Chan (2018) to reach the following result:

Case (i): Suppose that $S(\gamma^*, u^*) \|u\|_2^2$ is the largest on the right-hand side of (S4.2). If $\|u\|_2^2 > 0$, we have $S(\hat{\gamma}, u) \leq \lambda_1^{-d/(2l-d)} O_p(n_B^{-2l/(2l-d)}) \|u\|_2^2$. If $\|u\|_2^2 = 0$, we can still get the same result.
Case (ii): Suppose that \( \lambda_1 \| u \|_H^2 \) is the largest on the right-hand side of (S4.2). Then, we have \( S(\hat{\gamma}, u) \leq 3\lambda_1 \| u \|_H^2 \).

Case (iii): Suppose that \( \lambda_2 Q_A(\hat{\gamma}) \| u \|_2^2 \) is the largest on the right-hand side of (S4.2). Then, we have \( S(\hat{\gamma}, u) \leq 3\lambda_2 Q_A(\hat{\gamma}) \| u \|_2^2 \).

By (S3.11) and the proof of Lemma S6, \( \| u \|_2 \leq \| u \|_\infty \leq R \| u \|_H < \infty \). Then, we have

\[
S(\hat{\gamma}, u) = O_p \left[ \max \{ \lambda_1^{-d/(2l-d)}n_B^{-2l/(2l-d)} \| u \|_2^2, \lambda_1 \| u \|_H^2, \lambda_2 Q_A(\hat{\gamma}) \| u \|_2^2 \} \right].
\] (S4.3)

By the proposed optimization problem (3.3), \( \hat{r}_i < \xi_2 \) for \( i = 1, \ldots, N \), so \( Q_A(\hat{\gamma}) = O_p(1) \). Thus, by the conditions on \( \lambda_1 \) and \( \lambda_2 \), we have \( S(\hat{\gamma}, u) = O_p(n_B^{-1}) \| u \|_2^2 \). Thus, we have complete the proof of Lemma S7.

Based on Lemma S6, we can also use a similar argument as the proof for Lemma S3 of Wong and Chan (2018) to show that \( E \{ n_B S(\hat{\gamma}, u) \} < \infty \).

**Proof of Theorem 1.** Consider

\[
N^{-1} \sum_{i=1}^N (\delta_{\hat{A},i} \hat{w}_i - 1) y_i = N^{-1} \sum_{i=1}^N (\delta_{\hat{A},i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) y_i + N^{-1} \sum_{i=1}^N (\delta_{B,i} \pi_{B,i}^{-1} - 1) y_i
\]

\[
= N^{-1} \sum_{i=1}^N (\delta_{\hat{A},i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) m_0(x_i) + N^{-1} \sum_{i=1}^N (\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) \epsilon_i
\]

\[+ N^{-1} \sum_{i=1}^N (\delta_{B,i} \pi_{B,i}^{-1} - 1) y_i. \] (S4.4)

By Assumption 4 and Lemma S7, we have

\[
N^{-1} \sum_{i=1}^N (\delta_{\hat{A},i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) m_0(x_i) = O_p(n_B^{-1/2}). \] (S4.5)

Since \( \{ \epsilon_i : i = 1, \ldots, N \} \) are independent with the sampling indicators \( \{(\delta_{A,i}, \delta_{B,i}) : i = 1, \ldots, N\} \),
as well as the weights \( \{(\hat{w}_i, \pi_{B,i}) : i = 1, \ldots, N\} \), by Assumption A5, we have

\[
\text{var} \left\{ N^{-1} \sum_{i=1}^{N} (\delta_{A,i}\hat{w}_i - \delta_{B,i}\pi_{B,i}) \epsilon_i \right\} \\
\leq C_{\sigma,2} E \left\{ N^{-2} \sum_{i=1}^{N} (\delta_{A,i}\hat{w}_i - \delta_{B,i}\pi_{B,i})^2 \right\} \\
\leq 2C_{\sigma,2} E \left( N^{-2} \sum_{i=1}^{N} \delta_{B,i}\pi_{B,i}^{-2} \right) + 2C_{\sigma,2} E \left( N^{-2} \sum_{i=1}^{N} \delta_{A,i}\hat{w}_i^2 \right). \quad (S4.6)
\]

To show the order of (S4.6), we first consider

\[
E \left( N^{-2} \sum_{i=1}^{N} \delta_{B,i}\pi_{B,i}^{-2} \right) = N^{-2} \sum_{i=1}^{N} \pi_{B,i}^{-1} \\
\leq C_{B,1}^{-1} N^{-2} \sum_{i=1}^{N} n_B^{-1} \\
= O(n_B^{-1}), \quad (S4.7)
\]

where the first inequality holds by Assumption A7.

In addition, \( n_A^{-1}(N - n_A) \rightarrow P(\delta = 1)^{-1}P(\delta = 0) \) almost surely by strong law of large number. Then, by Assumption A3, there exists a constant \( C_{A,4} \), such that \( n_A^{-1}(N - n_A) < C_{A,4} \) for \( N \geq N_0 \), where \( N_0 \) is determined by \( C_{A,4} \). Then, for \( N \geq N_0 \), since \( \hat{w}_i = 1 + (N - n_A)/n_A\hat{r}_i \), we have

\[
N^{-2} \sum_{i=1}^{N} \delta_{A,i}\hat{w}_i^2 \leq N^{-2} \sum_{i=1}^{N} \delta_{A,i}(1 + C_{A,4}\hat{r}_i)^2 \\
\leq 2N^{-2} \sum_{i=1}^{N} \delta_{A,i} + 2C_{A,4}^2 N^{-2} \sum_{i=1}^{N} \delta_{A,i}\hat{r}_i^2 \\
\leq 2N^{-1} + 2C_{A,4}^2 N^{-2} \sum_{i=1}^{N} \xi_2^2 \\
= 2N^{-1}(1 + C_{A,4}^2\xi_2^2). \quad (S4.8)
\]

Thus, by Assumption A7 and (S4.8), we have

\[
E \left( N^{-2} \sum_{i=1}^{N} \delta_{A,i}\hat{w}_i^2 \right) = O(n_B^{-1}). \quad (S4.9)
\]
Thus, by (S4.6)–(S4.7) and (S4.9), we have shown that
\[
\text{var} \left\{ N^{-1} \sum_{i=1}^{N} \left( \delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1} \right) \epsilon_i \right\} = O(n_B^{-1}). \tag{S4.10}
\]

By (S4.4), (S4.5), (S4.10) and Assumption A6, we complete the proof of Theorem 1.

\[\square\]

S5 Bootstrap variance estimator

Since the inclusion probabilities are unavailable for the non-probability sample \(A\), we consider a bootstrap variance estimator (Kim et al., 2019) only taking into consideration the design features associated with the probability sample \(B\).

For \(b = 1, \ldots, B\), let the bootstrap version of (3.5) be
\[
\hat{Y}_N^{(b)*} = N^{-1} \sum_{i \in A} \hat{\omega}_i^{(b)*} y_i,
\]
where \(B\) is the number of bootstrap replications, \(\hat{\omega}_i^{(b)*} = 1 + (Nn_A^{-1} - 1) \hat{r}_i^{(b)*}\) for \(i \in A\), \(\hat{\gamma}^{(b)*} = (\hat{\gamma}_1^{(b)*}, \ldots, \hat{\gamma}_N^{(b)*})\) with \(\hat{\gamma}_i^{(b)*} = 0\) for \(i \notin A\) is obtained by
\[
\hat{\gamma}^{(b)*} = \arg \min_{\xi_1 \leq r_i \leq \xi_2} \mathbb{E} \left[ \sup_{u \in H} \left\{ \frac{S^{(b)*}(\gamma, u)}{\|u\|_2^2} - \lambda_1 \left\| u \right\|_2^2 \right\} - \lambda_2 Q_A(\gamma) \right],
\]
\[
S^{(b)*}(\gamma, u) = \left[ N^{-1} \sum_{i=1}^{N} \delta_{A,i} \left\{ 1 + \left( \frac{N}{n_A} - 1 \right) \hat{r}_i \right\} u(x_i) - N^{-1} \sum_{i=1}^{N} \delta_{B,i} d_{B,i}^{(b)*} u(x_i) \right]^2,
\]
the set of bootstrap weights \(\{d_{B,i}^{(b)*} : i \in B\}\) satisfies
\[
E^{*} \{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} d_{B,i}^{(b)*} u(x_i) \} = N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(x_i),
\]
\[
\text{var}^{*} \{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} d_{B,i}^{(b)*} u(x_i) \} = \hat{V} \{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(x_i) \},
\]
\(E^{*}(\cdot)\) and \(\text{var}^{*}(\cdot)\) are the conditional expectation and variance with respect to the bootstrap procedure given the probability sample \(B\), and \(\hat{V} \{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(x_i) \}\) is a design-unbiased variance estimator of
\[ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(x_i). \] Then, a bootstrap variance estimator for (3.5) can be the sample variance of \( \{ \hat{Y}^{(b)}_N : b = 1, \ldots, B \} \). The bootstrap variance estimator is reasonable if we can safely ignore the variability with respect to the non-probability sample \( A \), for example, when \( n_B = o_p(n_A) \) by Assumption 3 and Assumption 7.

Specifically, if the probability sample \( B \) is generated by a Poisson sampling as shown in Section 5, then a design-unbiased variance estimator of \( N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} u(x_i) \) is

\[ \hat{V}_B = \sum_{i \in B} \frac{1 - \pi_{B,i}}{\pi_{B,i}^2} u^2(x_i). \]

Then, the distribution to generate the bootstrap weights \( d_{B,i}^* \) can be normal with mean \( \pi_{B,i}^{-1} \) and variance \( (1 - \pi_{B,i}) \pi_{B,i}^{-2} \).

**S6 Proof of Theorem 2**

**Lemma S8.** Suppose that Assumption A7 holds. Then, we have

\[ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} = O_p(1). \]

**Proof of Lemma S8** Since \( P(\delta_{B,i}) = \pi_{B,i} \), we have

\[ E \left( N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} \right) = 1. \quad (S6.1) \]

Consider

\[
\begin{align*}
\text{var} \left( N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} \right) &= N^{-2} \sum_{i=1}^{N} \text{var}(\delta_{B,i} \pi_{B,i}^{-1}) + N^{-2} \sum_{i \neq j} \pi_{B,i}^{-1} \pi_{B,j}^{-1} \text{cov}(\delta_{B,i}, \delta_{B,j}) \\
&\leq N^{-2} \sum_{i=1}^{N} \text{var}(\delta_{B,i} \pi_{B,i}^{-1}) \\
&= N^{-2} \sum_{i=1}^{N} (1 - \pi_{B,i}) \pi_{B,i}^{-1} \\
&\leq N^{-2} C_{B,1} N^2 n_B^{-1} \\
&= O(n_B^{-1}), \quad (S6.2)
\end{align*}
\]
where the first inequality holds since \( \{ \delta_{B,i} : i = 1, \ldots, N \} \) are negatively associated, and the second inequality holds by Assumption A7. By Assumption A7, \( n_B^{-1} \to 0 \), so we have proved Lemma S8 by (S6.1)–(S6.2).

**Lemma S9.** Suppose that Assumptions A1–A8 hold. Then, we have

\[
B_N^{-1} \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1}) \epsilon_i \to N(0, 1),
\]

where \( B_N^2 = \sum_{i=1}^{N}(\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1})^2 \sigma_i^2 \). Besides, \( B_N^2 \approx N^2 n_B^{-1} \) in probability.

**Proof of Lemma S9.** Denote \( \mathcal{A}_N = \{ (\delta_{A,i}, \mathbf{x}_i) : i \in A \} \cup \{ (\delta_{B,i}, \mathbf{x}_i) : i \in B \} \). Then, given \( \mathcal{A}_N \), \( B_N^2 \) is the conditional variance of \( \sum_{i=1}^{N}(\delta_{A,i} \hat{w}_i - 1) \epsilon_i \).

We first consider the stochastic order of \( B_N \). On the one hand, we have

\[
B_N^2 \leq C_{\sigma,2} \sum_{i=1}^{N}(\delta_{A,i} \hat{w}_i - \delta_{B,i} \pi_{B,i}^{-1})^2 \\
\leq 2C_{\sigma,2} \sum_{i=1}^{B} \delta_{A,i} \hat{w}_i^2 + 2C_{\sigma,2} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-2} \\
= O_p(N^2 n_B^{-1}),
\]

(S6.3)

where the last equality holds by Assumption A7, (S4.7) and (S4.8). On the other hand, consider

\[
B_N^2 = \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i^2 \sigma_i^2 - 2 \sum_{i=1}^{N} \delta_{A,i} \delta_{B,i} \hat{w}_i \pi_{B,i}^{-1} \sigma_i^2 + \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-2} \sigma_i^2 \\
\geq C_{\sigma,1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i^2 - 2C_{\sigma,1} \max\{ \hat{w}_i : i = 1, \ldots, N \} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} + C_{\sigma,1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-2} \\
\geq C_{\sigma,1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i^2 - 2C_{\sigma,1} \max\{ \hat{w}_i : i = 1, \ldots, N \} O_p(N) + C_{\sigma,1} C_{B,1}^{-2} N^2 n_B^{-1} \\
\geq C_{\sigma,1} C_{B,1}^{-2} N^2 n_B^{-1} + o_p(N^2 n_B^{-1})
\]

(S6.4)

where the second inequality holds by Lemma S8 and the last inequality holds by the condition that \( \max\{ \hat{w}_i : i = 1, \ldots, N \} \) is bounded since \( \hat{w}_i = 1 + (N n_A^{-1} - 1) \hat{r}_i, \hat{r}_i \leq \xi_2 \).
and \((Nn_A^{-1} - 1) < C_{A4}\), where \(C_{A4}\) is discussed in the proof of Theorem 1. Thus, by (S6.3)–(S6.4), we have shown that \(B_N^2 \asymp N^2 n_B^{-1}\) in probability.

For any \(\eta > 0\), consider

\[
N^{-1} \sum_{i=1}^{N} \left( \delta_{A,i} \hat{w}_i - \delta_{B,i} \hat{\pi}_{B,i}^{-1} \right) y_i - N^{-1} \sum_{i=1}^{N} \left( \delta_{A,i} \hat{w}_i - \delta_{B,i} \hat{\pi}_{B,i}^{-1} \right) \hat{m}(x_i)
\]

\[
= N^{-1} \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \hat{\pi}_{B,i}^{-1}) \epsilon_i + N^{-1} \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \hat{\pi}_{B,i}^{-1}) \left( m_0(x_i) - \hat{m}(x_i) \right).
\]

(S6.6)

Lemma S9 validates the central limit theorem for the first part of (S6.6). By Lemma 1 and a similar argument in the proof of Lemma S3 of [Wong and Chan (2018)], we can show that

\[
N^{-1} \sum_{i=1}^{N} (\delta_{A,i} \hat{w}_i - \delta_{B,i} \hat{\pi}_{B,i}^{-1}) \left( m(x_i) - \hat{m}(x_i) \right) = o_p(n_B^{-1}).
\]

(S6.7)

By the stochastic order of \(B_N\) in Lemma S9, we have proved Theorem 2.
Proof of Corollary 1

Denote
\[ \tilde{\theta} = N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \{ y_i - m(x_i) \} \]
\[ = N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \epsilon_i. \] (S7.1)

Then, by (S6.7), we have
\[ |\hat{\theta}_{\text{prop}} - \tilde{\theta}| = \left| N^{-1} \sum_{i=1}^{N} (\delta_{B,i} \pi_{B,i}^{-1} - \delta_{A,i} \hat{w}_i) \{ \hat{m}(x_i) - m(x_i) \} \right| = o_p(n_B^{-1/2}). \] (S7.2)

Since the asymptotic order of \( \hat{\theta}_{\text{prop}} \) is \( O_p(n_B^{-1/2}) \) by Assumption A6 and Theorem 2, it is enough to investigate the variance of \( \tilde{\theta} \) in (S7.1) by (S7.2).

Consider
\[ \text{var}(\tilde{\theta}) = \text{var} \left[ E \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \epsilon_i \mid \mathcal{A}_N \right\} \right] \]
\[ + E \left[ \text{var} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \epsilon_i \mid \mathcal{A}_N \right\} \right]. \] (S7.3)

Since \( \epsilon_i \) is independent with \( \{ \delta_{A,i} : i = 1, \ldots, N \} \) and \( \{ \delta_{B,i} : i = 1, \ldots, N \} \) by Assumption A6, we have
\[ E \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \epsilon_i \mid \mathcal{A}_N \right\} = N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i). \]

Thus, we conclude that
\[ \text{var} \left[ E \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \epsilon_i \mid \mathcal{A}_N \right\} \right] \]
\[ = \text{var} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi_{B,i}^{-1} m(x_i) \right\}. \] (S7.4)
Next, consider
\[
\text{var} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi^{-1}_{B,i} m(x_i) + N^{-1} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i \epsilon_i \mid A_N \right\} = N^{-1} \sum_{i=1}^{N} \hat{w}_i^2 \sigma_i^2. \tag{S7.5}
\]
Thus, by (S7.2)–(S7.5), a plug-in variance estimator of \( \hat{\theta}_{\text{prop}} \) is
\[
\hat{V} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi^{-1}_{B,i} \hat{m}(x_i) \right\} + N^{-2} \sum_{i=1}^{N} \delta_{A,i} \hat{w}_i^2 \{ y_i - \hat{m}(x_i) \}^2, \tag{S7.6}
\]
where \( \hat{V} \left\{ N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi^{-1}_{B,i} \hat{m}(x_i) \right\} \) is a design-based variance of \( N^{-1} \sum_{i=1}^{N} \delta_{B,i} \pi^{-1}_{B,i} \hat{m}(x_i) \) treating \( \{ \hat{m}(x_i) : \delta_{B,i} = 1 \} \) as non-stochastic. Thus, we have finished the proof of Corollary 1.

**S8 Doubly robust estimator by Chen et al. (2020)**

Consider a logistic model for the non-probability sample \( A, \pi_{A,i} = \pi_A(x_i; \theta_0) \), where \( \text{logit} \{ \pi_A(x_i; \theta_0) \} = x_i^T \theta_0 \), and \( \theta_0 \) is the true parameter. An estimator of \( \theta_0 \), say \( \hat{\theta} \), is obtained by solving
\[
\sum_{i \in A} x_i - \sum_{i \in B} \pi^{-1}_{B,i} \pi_A(x_i; \theta) x_i = 0. \tag{S8.1}
\]
The corresponding estimator is termed as “maximum pseudo-likelihood estimator” by [Chen et al. (2020)](2020).

To overcome the model mis-specification for the sampling mechanism associated with the non-probability sample, [Chen et al. (2020)](2020) also proposed two double robust estimators by assuming a parametric model for \( m_0(x) = m(x; \beta_0) \), where \( \beta_0 \) is an unknown parameter to be estimated. Since missing at random is assumed, we can obtain a consistent estimator of \( \beta_0 \), say \( \hat{\beta} \), using a standard approach. Then, the doubly robust estimators are
\[
\hat{Y}_1 = N^{-1} \sum_{i \in A} \{ \pi_A(x_i; \hat{\theta}) \}^{-1} \{ y_i - m(x_i; \hat{\beta}) \} + N^{-1} \sum_{i \in B} \pi^{-1}_{B,i} m(x_i; \hat{\beta}) \tag{S8.2}
\]
Table 5: Relative bias of the bootstrap variance estimator for HT\_KL based on 1000 Monte Carlo simulations under different simulation setups. The number of replication is $B = 200$.

| Model       | (5000, 1000, 100) | (10000, 2000, 200) |
|-------------|-------------------|-------------------|
| Linear      | -0.026            | 0.031             |
| Nonlinear   | 0.032             | 0.032             |

and

\[
\hat{Y}_2 = \hat{N}^{-1} \sum_{i \in A} \{\pi_A(x_i; \hat{\theta})\}^{-1} \{y_i - m(x_i; \hat{\beta})\} + \hat{N}^{-1} \sum_{i \in B} \pi_{B,i}^{-1} m(x_i; \hat{\beta}) ,
\]

(S8.3)

where $\hat{N} = \sum_{i \in A} \{\pi_A(x_i; \hat{\theta})\}^{-1}$. The only difference between (S8.2) and (S8.3) is that a true population size $N$ is used for (S8.2), but its estimator is used for (S8.3).

S9 Additional simulation result

Under a certain simulation setup, denote $\hat{Y}^{(m)}_N$ and $\hat{V}^{(m)}$ to be the HT\_KL estimator and its bootstrap variance estimator for the $m$-th Monte Carlo simulation for $m = 1, \ldots, M$, where $M = 1000$ in the simulation study; see Section S5 for details about the bootstrap variance estimator. Let $\hat{V}$ be the sample variance of $\{\hat{Y}^{(m)}_N : m = 1, \ldots, M\}$. Then, the relative bias of the bootstrap variance estimator is

\[
\frac{M^{-1} \sum_{m=1}^M \hat{V}^{(m)} - \hat{V}}{\hat{V}}.
\]

Table 5 shows the relative bias of the bootstrap variance estimator for HT\_KL under different simulation setups. The relative bias of the bootstrap variance estimator is small regardless of the simulation setups. Thus, the variance of HT\_KL can be reasonably estimated by bootstrap.