Weighted likelihood estimation under two-phase sampling

Takumi Saegusa
Department of Biostatistics
University of Washington
Seattle, WA 98195-7232
e-mail: tsaegusa@uw.edu

and

Jon A. Wellner
Department of Statistics
University of Washington
Seattle, WA 98195-4322,
e-mail: jaw@stat.washington.edu

Abstract: We develop asymptotic theory for weighted likelihood estimators (WLE) under two-phase stratified sampling without replacement. We also consider several variants of WLE's involving estimated weights and calibration. A set of empirical process tools are developed including a Glivenko-Cantelli theorem, a theorem for rates of convergence of Z-estimators, and a Donsker theorem for the inverse probability weighted empirical processes under two-phase sampling and sampling without replacement at the second phase. Using these general results, we derive asymptotic distributions of the WLE of a finite dimensional parameter in a general semiparametric model where an estimator of a nuisance parameter is estimable either at regular or non-regular rates. We illustrate these results and methods in the Cox model with right censoring and interval censoring. We compare the methods via their asymptotic variances under both sampling without replacement and the more usual (and easier to analyze) assumption of Bernoulli sampling at the second phase.

AMS 2000 subject classifications: Primary 62E20; secondary 62G20, 62D99, 62N01.

Keywords and phrases: calibration, estimated weights, Weighted likelihood, semiparametric model, regular, non-regular: version 39, 17 January 2012.

1. Introduction

Two-phase sampling is a sampling technique that aims at cost reduction and improved efficiency of estimation. At phase I, a large sample is drawn from a population, and information on variables that are easier to measure is collected. These phase I variables may play an important role in statistical analysis such
as exposure in a regression model, or they may simply be auxiliary variables that are correlated with unavailable variables at phase I but are not of interest in themselves. Based on the all variables observed at phase I, the sample space is stratified using only the phase I data. At phase II, a subsample is drawn without replacement from each stratum, and phase II variables that are costly or difficult to obtain are measured. Strata formation is intended either to oversample subjects with important phase I variables, or to effectively sample subjects with phase II variables correlated with phase I variables, or both. This way, two-phase sampling achieves effective access to important variables with less cost and, as a result, enhances efficiency of estimation.

Two-phase sampling was originally introduced in survey sampling by [26] for estimation of the “finite population mean” of some variable. Since then, this sampling method together with the Horvitz-Thompson estimator [14] as a standard estimator has been widely adopted in survey sampling. Much later, two-phase designs were introduced to biostatistical applications where, in contrast to survey sampling, the population is infinite and the parameter in the statistical model is of interest rather than the average of some variables in a finite population. Notable examples of two-phase designs include those given in: [28, 38] (who considered fitting Cox models for the case-cohort design stratifying a cohort by the censoring indicator); [44] (who considered additional stratification on rare exposure in the setting of the case control design); and [2] (who extended the case cohort study of [28] by additionally stratifying on covariates). More recently the broad applicability and importance of two phase designs has been emphasized by [3, 4]. Because of these features, two-phase sampling has recently received more attention from practitioners, including biostatisticians, as an attractive study design.

In this paper, we consider weighted likelihood estimation under two-phase sampling. The main difficulty in this problem is dependence among observations induced by the sampling scheme. The current practice for analysis of this type of data in biostatistics has been to assume independence (i.e. Bernoulli sampling) for theoretical simplicity. Specifically, statistical analysis is often carried out as if observations are obtained from stratified Bernoulli sampling rather than from stratified sampling without replacement at phase II. Because the asymptotic variances under our “without replacement” sampling scheme are shown here to be smaller than under stratified Bernoulli sampling (see also [6]), use of variance estimates based on Bernoulli sampling calculations results in over-estimates of variances and hence conservative conclusions. Despite theoretical difficulties, [6] developed a method to derive asymptotic distributions of the weighted likelihood estimator (WLE) under stratified sampling without replacement. Our goal is to complement and extend their results in several directions.

In addition to obtaining a precise asymptotic variance, there are several advantages to use of the WLE and improvements based on “estimated weights” or calibration. It may be true that one is willing to assume independence because the WLE is generally inefficient under the independence assumption. However, even when this assumption holds, there are not many statistical models where efficient estimators are known (see [30], [29], and [5] for some exceptions). More-
over, efficient estimators, if known, may require sophisticated numerical techniques (e.g., solving integral equations, [25]) or restrictive assumptions that are not imposed when complete data is available (e.g. a parametric covariate distribution [9], discrete covariates [25]). In contrast, the maximum likelihood estimator with complete data is often available in many applications, and the corresponding WLE is obtained simply via a weighted likelihood version of the same likelihood equations. Furthermore, theory for the WLE only requires almost identical conditions to those for the MLE with complete data (see Theorems 3.1 and 3.2 below). Another related advantage of the WLE and their variants involving estimated weights or calibration is robustness to model misspecification: When the underlying model is misspecified, the WLE’s and their relatives continue to estimate the same parameters as would be estimated under model misspecification with complete data. (For example, see [17], [13], [36, 37], [45], and [18]. See also [35] for careful further considerations of this issue.)

Survey sampling has a long history of dealing with dependent observations from stratified sampling without replacement and even more complicated designs. In fact, asymptotic results have been established for more complex survey designs (see [16] and references therein). However, even when a model is postulated, survey statisticians are usually interested in a “finite population parameter”, which is usually defined as the solution of some estimating equation in a finite population [32], not a “super-population parameter” which determines a scientific phenomenon for an infinite population. Asymptotic theory in survey sampling thus usually treats a sequence of finite populations with increasing sample sizes based on conditions regarding designs conditional on observations consisting of a finite population [16]. Because biostatisticians are more interested in super-population parameters, our asymptotic theory is relevant in the usual biostatistical settings.

One notable exception in the survey sampling literature is [32] (see also references therein and [19], [10]). In [32], the authors define the product of the model space and the design space as a probability space, and decompose their normalized estimator into the contributions from a super-population and a sampling design. Two distinct sets of conditions for the model space and design space are used to guarantee the asymptotic normality of each contribution respectively. The former and the latter conditions are familiar to biostatisticians and survey statisticians, respectively, but not vice versa. See [31] for two sets of conditions in an application to the Cox model under cluster sampling.

Our approach relies on the framework and some of the results developed in [6]. In [6], the inverse probability weighted empirical process is decomposed into the usual empirical process (phase I contribution) and the weighted sum of finite sampling empirical processes (phase II contribution). (Compare (10) of [6] with the decomposition (A.8) of [32].) Then conditional on the phase I data, the results for exchangeably weighted bootstrap empirical processes [27], which covers our sampling scheme, is applied to show the weak convergence of the phase II contribution. Despite some similarity to the framework involving model and design spaces, our framework is different from [32] in the following
important points. First of all, we do not need additional conditions for the
phase II contribution unlike design conditions imposed in [32] because the same
conditions for the phase I contribution suffice for the exchangeably weighted
bootstrap empirical process theory to apply in our setting. Second, our method
is more general since our decomposition is at the process level, not the level
of random variables. Third, our formulae for asymptotic variances have more
natural interpretations than the formulae in the framework of [32] that consist
of two incongruent parts, one that depends on the model conditions and the
other that depends on the design conditions. For these reasons, our approach
should be distinguished from those in survey sampling.

The main results of our paper are two $Z$-theorems giving weak sufficient
conditions for asymptotic distributions of the WLE’s in general semiparamet-
ric models. The first theorem covers the case where the nuisance parameter is
estimable at a regular $(\sqrt{n})$ rate, while the second theorem serves for the case
where only estimators with slower than $\sqrt{n}$–convergence rates are available. In
addition to the plain WLE, we include the WLE’s with estimated weights and
(variants of) calibration in the formulations of both theorems. These estimators
are obtained when adjusting weights in the WLE by estimated weights [30], cal-
ibration [12], modified calibration [8] or our new method, centered calibration,
respectively. The first two methods are used in practice to try to gain efficiency
over the WLE (see [20] for a recent review and discussion), and the third method
was recently proposed to improve on the original calibration. It is of some in-
terest to understand which method improves efficiency in comparison to the
plain WLE. The weighted likelihood estimator is already well-studied in cases
with regular rates. [6] derived the asymptotic distribution of the WLE under
stratified Bernoulli sampling and stratified sampling without replacement. [7]
studied the WLE with estimated weights under stratified Bernoulli sampling,
and showed efficiency gains over the plain WLE. [3, 4] obtained in a heuristic
way the asymptotic distributions of the the WLE’s with estimated weights and
the calibrated WLE under stratified sampling without replacement. One of the
difficulties in the derivations in [3, 4] involves the lack of a proof of asymp-
totic equicontinuity of certain stochastic processes under dependence. A similar
difficulty is also recognized by [19] in the context of complex surveys. Direct
application of empirical process theory does not help due to lack of indepen-
dence among observations. Another difficulty, which is also seen in other papers,
concerns (lack of) proofs of consistency of estimators under dependence. When
a nuisance parameter is not estimable at a regular rate, no general consistency,
rates of convergence, or asymptotic normality results are known in the framework
of two-phase designs to the best of our knowledge.

The main contributions of our paper are three-fold. First, we rigorously jus-
tify the results of [3, 4] with weaker conditions. We further extend the result
to the case where the nuisance parameter is not estimable at a regular rate.
The conditions of our theorems are formulated in terms of complete data, not
two-phase sampling data, and, moreover, they are almost identical to those for
the MLE with complete data. Thus, most of them may be already established in
many applications. For the conditions requiring verification, tools from empirical
process theory will be applied. Second, we developed a new method which yields improved efficiency over the plain WLE under our sampling scheme. In fact, our method of centered calibration is the only guaranteed method, among all methods considered in this paper, to gain efficiency under both stratified Bernoulli sampling and stratified sampling without replacement while other methods are warranted only for stratified Bernoulli sampling. Third, we establish general results for the inverse probability weighted (IPW) empirical process, which is defined in the next section. Some results such as Glivenko-Cantelli theorem (Theorem 5.1) and Donsker Theorem (Theorem 5.3) are of interest in their own right. These results, accounting for dependence of observations due to the sampling design, are used to prove our Z-theorems in place of the usual empirical process theory. More importantly, they are useful in verifying the conditions of Z-theorems in applications. For instance, our Theorem 5.2 easily establishes rates of convergence under our “without replacement” sampling scheme. Also, consistency can be verified with the aid of the Glivenko-Cantelli theorem. We illustrate application of the general results with examples in Section 4.

The rest of the paper is organized as follows. In section 2, we introduce our sampling scheme and estimation procedures in a general semiparametric model. The WLE and methods involving adjusted weights intended to improve on the efficiency of the WLE are discussed. Two Z-theorems are presented in section 3 to derive asymptotic distributions of the WLE’s of the finite dimensional parameter. All estimators are compared under Bernoulli sampling and sampling without replacement with different methods of adjusting the weights. We apply our Z-theorems to the Cox model, both with right censoring and interval censoring, in section 4. The WLE of the cumulative baseline hazard function has regular rate of convergence in the first example, while it has cube-root rate in the second example. Section 5 consists of general results for IPW empirical processes. Several open problems are briefly discussed in Section 6. All proofs except those in section 4 and auxiliary results are collected in Section 7.

2. Sampling, Models, and Estimators

We now introduce our sampling scheme. Most of the following notation is based on [6]. Let $W = (X, U) \in W = X \times U$ be the complete data with distribution $P_0$ where $X$ is the vector of the variables of interest with distribution $P_0$ and $U$ is a vector of auxiliary variables. At phase I, only a coarsening $\tilde{X} = \tilde{X}(X)$ of $X$ and the auxiliary variables $U$ are available for all $N$ subjects. The phase I data $V = (\tilde{X}, U) \in V = \tilde{X} \times U$ are used to form the $J$ sampling strata $V_j$ with $\sum_{j=1}^J V_j = V,$ the $j$th of which consists of $N_j$ subjects for $j = 1, \ldots, J$. After stratified sampling, $X$ is fully observed for $n_j$ subjects in the $j$th stratum at phase II. The observed data is $(V, X \xi, \xi)$ where $\xi$ is the indicator of being sampled at phase II. We use a doubly subscripted notation by which $V_{ij}$, for example, denotes $V$ for the $i$th subject in stratum $j$. We denote the stratum probability for the $j$th stratum by $\nu_j \equiv P_0(V \in V_j)$, and the conditional expectation given membership in the $j$th stratum by $P_{0|j}(\cdot) \equiv \tilde{P}_0(\cdot|V \in V_j)$. 
At phase II, samples of size \( n_j \leq N_j \) are drawn at random without replacement from each of the \( J \) strata. The sampling probability is \( P(\xi = 1|V_i) = \pi_0(V_i) = n_j/N_j \) for \( V_i \in V_j \). These sampling probabilities are assumed to be strictly positive; that is, there is a strictly positive constant \( \sigma > 0 \) such that \( 0 < \sigma \leq \pi_0(v) \leq 1 \) for \( v \in V \). We assume that \( n_j/N_j \to p_j > 0 \) for \( j = 1, \ldots, J \) as \( N \to \infty \). Although dependence is induced among the observations \((V_i, \xi_i, X_i, \xi_i)\) by the sampling indicators, the vector of sampling indicators \((\xi_{j1}, \ldots, \xi_{jN_j})\) within strata, \( j = 1, \ldots, J \), are exchangeable for \( j = 1, \ldots, J \), and the \( J \) random vectors \((\xi_{j1}, \ldots, \xi_{jN_j})\) are independent.

One of the most important tools in empirical process theory is the empirical measure. However, the empirical measure is not directly applicable to estimation under two-phase sampling because some observations are not observed at phase II. Instead, we define the inverse probability weighted (IPW) empirical measure by

\[
P_N^\xi = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_0(V_i)} \delta_{X_i} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{\xi_{j,i}}{n_j/N_j} \delta_{X_{j,i}},
\]

where \( \delta_{X_i} \) denotes a Dirac measure placing unit mass on \( X_i \). The identity in the last display is justified by the arguments in Appendix A of [6]. We also define the IPW empirical process by \( G_N^\xi = \sqrt{N}(P_N^\xi - P_0) \) and the phase II empirical process for the \( j \)th stratum by

\[
G_{j,N_j}^\xi \equiv \sqrt{N_j} \left( P_{j,N_j}^\xi - \frac{n_j}{N_j} P_{j,N_j} \right), \quad j = 1, \ldots, J,
\]

where, for \( j \in \{1, \ldots, J\} \), \( P_{j,N_j}^\xi \equiv N_j^{-1} \sum_{i=1}^{N_j} \xi_{j,i} \delta_{X_{j,i}} \) is the phase II empirical measure for the \( j \)th stratum, and \( P_{j,N_j} \equiv N_j^{-1} \sum_{i=1}^{N_j} \delta_{X_{j,i}} \) is the empirical measure for all the data in the \( j \)th stratum; note that the latter empirical measure is not observed. Then, following [6], page 207, we decompose \( G_N^\xi \) as follows:

\[
G_N^\xi = G_N + \sum_{j=1}^{J} \sqrt{N_j} \left( \frac{N_j}{n_j} \right) G_{j,N_j}^\xi, \quad \tag{2.1}
\]

where \( P_N = N^{-1} \sum_{j=1}^{J} N_j P_{j,N_j} \) and \( G_N = \sqrt{N}(P_N - P_0) \). Notice that the phase II empirical processes \( G_{j,N_j}^\xi \) correspond to “exchangeably weighted bootstrap” versions of the stratum-wise complete data empirical processes \( G_{j,N_j} \equiv \sqrt{N_j}(P_{j,N_j} - P_{0|j}) \) where \( P_{0|j} \) is the conditional distribution of \( X \) given membership in the \( j \)th stratum and \( P_{j,N_j} \) is as defined above. This observation allows application of the “exchangeably weighted bootstrap” theory of [27].

### 2.1. Improving efficiency

Efficiency of estimators based on IPW empirical processes can be improved by adjusting weights, either by estimated weights [30] or by calibration [12] via use
of the phase I information; see also [20]. In addition to these two methods, we discuss two variants of calibration, modified calibration [8], and our proposed method, centered calibration.

Let $Z_i \equiv g(V_i)$ be the auxiliary variables for the $i$th subject for a known transformation $g$. For estimated weights through binary regression, the first $J$ elements of $Z_i$ are the membership indicators for the strata, $I_{V_j}(V_i), j = 1, \ldots, J$. Furthermore, observations with $\pi_0(V) = 1$ are dropped from binary regression, and the original weight 1 is used. For notational simplicity, we write $Z_i$ for either method, and assume that sampling probabilities are strictly less than 1 for all strata.

2.1.1. Estimated weights

The method of estimated weights adjusts weights through binary regression on the phase I variables. The sampling probability for the $i$th subject is modelled by

$$p_{\alpha}(\xi_i | Z_i) = G_e(Z_i^T \alpha)^{\xi_i}(1 - G_e(Z_i^T \alpha))^{1 - \xi_i} \equiv \pi_\alpha(V_i)\xi_i\{1 - \pi_\alpha(V_i)\}^{1 - \xi_i},$$

where $\alpha \in \mathcal{A}_e \subset \mathbb{R}^{J+k}$ is a regression parameter and $G_e : \mathbb{R} \mapsto (0, 1]$ is a known function. If $G_e(x) = e^x/(1+e^x)$ for instance, then the adjustment simply involves logistic regression. Let $\hat{\alpha}_N$ be the estimator of $\alpha$ that maximizes the composite likelihood

$$\prod_{i=1}^N p_{\alpha}(\xi_i | Z_i) = \prod_{i=1}^N G_e(Z_i^T \alpha)^{\xi_i}(1 - G_e(Z_i^T \alpha))^{1 - \xi_i}. \quad (2.2)$$

We define the IPW empirical measure with estimated weights by

$$\mathbb{P}_{\pi,e}^N = \frac{1}{N} \sum_{i=1}^N \xi_i \pi_{\hat{\alpha}_N}(V_i) \delta_{X_i},$$

and the IPW empirical process with estimated weights by $\mathbb{G}_{\pi,e}^N = \sqrt{N}(\mathbb{P}_{\pi,e}^N - P_0)$.

2.1.2. Calibration

Calibration adjusts weights so that the inverse probability weighted average from the phase II sample is equated to the phase I average, whereby the phase I information is taken into account for estimation. Consider the problem of choosing the weights $\{w_i\}_{i=1}^N$ subject to the condition

$$\frac{1}{N} \sum_{i=1}^N \xi_i w_i Z_i = \frac{1}{N} \sum_{i=1}^N Z_i; \quad (2.3)$$

since the $Z_i$’s take values in $\mathbb{R}^k$, this is a system of equations in $\mathbb{R}^k$. In general there are many solutions to this system of equations, and the inverse probability weights $1/\pi_0(V_i)$ will typically not satisfy it. Because weights differing greatly from the inverse probability weights are unlikely to improve on the plain weighted likelihood estimates, calibration involves choosing weights closest to
the inverse probability weights in a certain distance measure. Let \( D_i(w, d) \) be a distance measure between the weights \( w \) and \( d \) for the \( i \)th subject, where for every fixed \( d > 0 \), \( D_i(w, d) \) is nonnegative, continuously differentiable with respect to \( w \), and strictly convex in \( w \), and \( (\partial/\partial w)D_i(w, d) \) is strictly increasing and is zero at \( w = d \) (see [12] for various choices of \( D_i \)). The resulting problem is a convex optimization problem: find positive weights \( w_i \) that minimize the average distance \( N^{-1} \sum_{i=1}^{N} D_i(w_i, 1/\pi_0(V_i)) \) with the constraint (2.3).

The method of Lagrange multipliers leads to \((\partial/\partial w_i)D_i(w_i, 1/\pi_0(V_i)) + Z_i^T \alpha = 0\) for the subjects with \( \xi_i = 1 \) where \( \alpha \) is a Lagrange multiplier. The invertibility of \((\partial/\partial w_i)D_i \) leads to the solution \( w_i = G_i(Z_i^T \alpha)/\pi_0(V_i) \) for some function \( G_i \) where \( G_i(0) = 1 \) and \( G_i(0) > 0 \). Substitution in (2.3) gives the calibration equation \( N^{-1} \sum_{i=1}^{N} \xi_i(G_i(Z_i^T \alpha)/\pi_0(V_i))Z_i = N^{-1} \sum_{i=1}^{N} Z_i \). The solution \( \hat{\alpha} \) to the calibration equation determines the calibrated weights \( \hat{w}_i = G_i(Z_i^T \hat{\alpha})/\pi_0(V_i) \).

One easy choice, as in [3, 4, 20], is to take the distance measures \( D_i \) to be the same for all subjects; i.e., \( D_i = D \) and \( G_i = G_c \) for \( i = 1, \ldots, N \). An alternative subject specific choice of the \( Di\)'s, leading to “modified calibration”, will be discussed in the next subsection. In both cases we formulate the calibration in terms of the calibration equation rather than the problem of minimizing a distance with the inverse probability weights. These assumptions simplify the condition on \( G_i \)'s and \( D_i \)'s. In our formulation with equal \( D_i \)'s, we find an estimator \( \hat{\alpha}_N \) that is the solution for \( \alpha \in \mathcal{A}_c \subset \mathbb{R}^k \) of the following calibration equation,

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i G_c(V_i; \alpha)}{\pi_0(V_i)} Z_i = \frac{1}{N} \sum_{i=1}^{N} Z_i,
\]

(2.4)

where \( G_c(V; \alpha) \equiv G(g(V)^T \alpha) = G(Z^T \alpha) \), and \( G \) is a known function with \( G(0) = 1 \) and \( G(0) > 0 \). We call \( \pi_0(V_i) \equiv \pi_0(V_i)/G_c(V; \alpha) \) the calibrated sampling probability for the \( i \)th subject. We define the calibrated IPW empirical measure by

\[
\mathbb{P}_N^{\pi,c} = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_0(V_i)} \delta_{X_i} = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_0(V_i)} G(Z_i^T \hat{\alpha}_N) \delta_{X_i},
\]

and the calibrated IPW process by \( \mathbb{G}_N^{\pi,c} = \sqrt{N}(\mathbb{P}_N^{\pi,c} - P_0) \).

### 2.1.3. Modified calibration

[12] discussed subject-dependent distance measures \( D_i \) when \((\partial/\partial w)D_i = \tilde{D}(w/d)/q_i \) where \( \tilde{D}(x) \) is a continuous, strictly increasing function on \( \mathbb{R} \) with \( \tilde{D}(1) = 0 \) and \((\partial/dx)\tilde{D}(1) = 1 \), independent of the index \( i \), and \( q_i > 0 \). Solving the convex optimization problem in this case with some choice of \( q_i \)'s leads to the calibration equation \( N^{-1} \sum_{i=1}^{N} \xi_i(G(q_i Z_i^T \alpha)/\pi_0(V_i))Z_i = N^{-1} \sum_{i=1}^{N} Z_i \) for the inverse \( G = (D)^{-1} \) of \( D \). Recently the choice \( q_i = (1 - \pi_0(V_i))/\pi_0(V_i) \) was proposed by [8] in a missing response problem. When \( \pi_0(V_i) < 1 \), \( i = 1, \ldots, N \), this choice means that when the sampling probability is larger, the subject contributes...
more to the average distance. Note that \( q_i = 0 \) when \( \pi_0(V_i) = 1 \). Although \( q_i = 0 \) is valid in the calibration equation. One implication of this choice is that we do not modify the weights if subjects are always sampled at phase II. We call the method of choosing weights by solving the calibration equation with \( q_i = (1 - \pi_0(V_i)) / \pi_0(V_i) \) modified calibration.

In modified calibration, we find the estimator \( \hat{\alpha}_N \) that is the solution for \( \alpha \in A_{mc} \subset \mathbb{R}^k \) of the following calibration equation:

\[
\frac{1}{N} \sum_{i=1}^N \frac{\xi_i G_{mc}(V_i; \alpha)}{\pi_0(V_i)} Z_i = \frac{1}{N} \sum_{i=1}^N Z_i, \tag{2.5}
\]

where

\[
G_{mc}(V; \alpha) \equiv G \left( \frac{1 - \pi_0(V)}{\pi_0(V)} g(V)^T \alpha \right) = G \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \alpha \right),
\]

and \( G \) is a known function with \( G(0) = 1 \) and \( \dot{G}(0) > 0 \). We call \( \pi_\alpha(V_i) \equiv \pi_0(V_i) / G_{mc}(V_i; \alpha) \) the calibrated sampling probability with modified calibration for the \( i \)th subject. We define the IPW empirical measure with modified calibration by

\[
P^{\pi,mc}_N = \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_\alpha(N)(V_i)} \delta_{X_i} = \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_0(V_i)} G \left( \frac{1 - \pi_0(V_i)}{\pi_0(V_i)} Z^T \hat{\alpha}_N \right) \delta_{X_i},
\]

and the IPW process with modified calibration by

\[
G_{N}^{\pi,mc} = \sqrt{N}(P^{\pi,mc}_N - P_0).
\]

### 2.1.4. Centered calibration

We propose a new method, centered calibration, that calibrates on centered auxiliary variables with modified calibration. This method in fact improves the plain WLE under our sampling scheme, while retaining the good properties of modified calibration. We discuss advantages of centered calibration and connections to other methods in Section 3.4.

In centered calibration, we find the estimator \( \hat{\alpha}_N \) that is the solution for \( \alpha \in A_{cc} \subset \mathbb{R}^k \) of the following calibration equation:

\[
\frac{1}{N} \sum_{i=1}^N \frac{\xi_i G_{cc}(V_i; \alpha)}{\pi_0(V_i)} (Z_i - \overline{Z}_N) = 0, \tag{2.6}
\]

where

\[
G_{cc}(V; \alpha) \equiv G \left( \frac{1 - \pi_0(V)}{\pi_0(V)} (Z - \overline{Z}_N)^T \alpha \right),
\]

with \( \overline{Z}_N = N^{-1} \sum_{i=1}^N Z_i \) suppressed in the definition, and \( G \) is a known function with \( G(0) = 1 \) and \( \dot{G}(0) > 0 \). We call \( \pi_\alpha(V_i) \equiv \pi_0(V_i) / G_{cc}(V_i; \alpha) \) the calibrated
sampling probability with centered calibration for the $i$th subject. We define the IPW empirical measure with centered calibration by

$$P_{π,cc}^N = \frac{1}{N} \sum_{i=1}^{N} \xi_i \delta_{\hat{X}_i}, \quad \xi_i = \frac{1}{\hat{\pi}_0(V_i)} G_{cc}(V_i; \hat{\alpha}_N) \delta_{\hat{X}_i},$$

and the IPW process with centered calibration by $G_{π,cc}^N = \sqrt{N} (P_{π,cc}^N - P_0)$.

### 2.2. Estimators

We study the asymptotic distribution of the weighted likelihood estimator of a finite dimensional parameter $\theta$ in a general semiparametric model $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ where $\Theta \subset \mathbb{R}^p$ and the nuisance parameter space $H$ is a subset of some Banach space $\mathcal{B}$. Let $P_0 = P_{\theta_0,\eta_0}$ denote the true distribution.

The maximum likelihood estimator with complete data is often obtained as a solution of the infinite dimensional likelihood equations. In such models, the WLE under two-phase sampling is obtained by solving the corresponding infinite dimensional inverse probability weighted likelihood equations. Specifically, the WLE $(\hat{\theta}_N, \hat{\eta}_N)$ is a solution of the following weighted likelihood equations

$$\Psi_{1}(\theta, \eta) = \mathbb{P}_{N}^\pi \hat{\ell}_{\theta, \eta} = o_P \left( N^{-1/2} \right),$$

$$\|\Psi_{2}(\theta, \eta)h\|_H = \|\mathbb{P}_{N}^\pi (B_{\theta, \eta}h - P_{\theta, \eta}B_{\theta, \eta}h)\|_H = o_P \left( N^{-1/2} \right), \quad (2.7)$$

where $\hat{\ell}_{\theta, \eta} \in L_2^0(P_{\theta, \eta})^p$ is the score function for $\theta$, and the score operator $B_{\theta, \eta} : H \mapsto L_2^0(P_{\theta, \eta})$ is the bounded linear operator mapping a direction $h$ in some Hilbert space $H$ of one-dimensional submodels for $\eta$ along which $\eta \rightarrow \eta_0$. The corresponding WLE with estimated weights $(\hat{\theta}_{N,e}, \hat{\eta}_{N,e})$, the calibrated WLE $(\hat{\theta}_{N,cc}, \hat{\eta}_{N,cc})$, the WLE with modified calibration $(\hat{\theta}_{N,mc}, \hat{\eta}_{N,mc})$, and the WLE with centered calibration $(\hat{\theta}_{N,cc}, \hat{\eta}_{N,cc})$ are obtained by replacing $\mathbb{P}_N^\pi$ by $\mathbb{P}_{N,e}^\pi$, $\mathbb{P}_{N,cc}^\pi$, $\mathbb{P}_{N,mc}^\pi$ or $\mathbb{P}_{N,cc}^\pi$ in $(2.7)$, respectively. Let $\hat{\ell}_0 = \hat{\ell}_{\theta_0,\eta_0}$ and $B_0 = B_{\theta_0,\eta_0}$.

### 3. Asymptotics for the WLE in general semiparametric models

We consider two cases: in the first case the nuisance parameter $\eta$ is estimable at a regular (i.e., $\sqrt{n}$) rate and, for ease of exposition, $\eta$ is assumed to be a measure. In the second case $\eta$ is only estimable at a non-regular (slower than $\sqrt{n}$) rate. Our theorem (Theorem 3.2) concerning the second case nearly covers the former case, but requires slightly more smoothness and a separate proof of the rate of convergence for an estimator of $\eta$. On the other hand, our theorem (Theorem 3.1) concerning the former case includes a proof of the (regular) ($\sqrt{n}$) rate of convergence, and hence is of interest by itself.
3.1. Conditions for estimating weights and (modified and centered) calibration

To derive asymptotic distributions of WLE's with estimated weights and (modified and centered) calibration, we need to establish asymptotic results on estimators of $\alpha$ with estimated weights and (modified and centered) calibration. To this end, we assume the following. Throughout this paper, we may assume both Conditions 3.1 and 3.2 at the same time, but it should be understood that the former condition is used exclusively for the estimators regarding estimated weights and the latter condition is imposed only for estimators regarding (modified and centered) calibration. Moreover, it should be understood that Conditions 3.2(a)(i) and 3.2(d)(i) are assumed for the estimators regarding calibration, Conditions 3.2(a)(ii) and 3.2(d)(ii) are imposed for the estimators regarding estimated weights and the latter condition is imposed only for estimators regarding modified calibration and that Conditions 3.2(a)(iii) and 3.2(d)(iii) are used for the estimators regarding centered calibration, respectively.

**Condition 3.1 (Estimated weights).** (a) The estimator $\hat{\alpha}_N$ is a maximizer of the composite likelihood (2.2).
(b) $Z \in \mathbb{R}^{J+k}$ is not concentrated on a $(J+k)$-dimensional affine space of $\mathbb{R}^{J+k}$ and has bounded support.
(c) $G_e : \mathbb{R} \mapsto [0,1]$ is a twice continuously differentiable, monotone function.
(d) $S_0 \equiv P_0 \left( \{ G_e(Z^T \alpha_0) \}^2 \pi_0(V) (1 - \pi_0(V))^{-1} Z^{\otimes 2} \right)$ is finite and nonsingular, where $G_e$ is a derivative of $G_e$.
(e) The “true” parameter $\alpha_0 = (\alpha_{0,1}, \ldots, \alpha_{0,J+k})$ is given by $\alpha_{0,j} = G_e^{-1}(p_j)$, for $j = 1, \ldots, J$, and $\alpha_{0,j} = 0$, for $j = J + 1, \ldots, J + k$. The parameter $\alpha$ is identifiable, that is, $p_0 = p_{\alpha}$ almost surely implies $\alpha = \alpha_0$.
(f) For a fixed $p_j \in (0,1)$, $n_j$ satisfies $n_j = [N_j p_j]$ for $j = 1, \ldots, J$.

**Condition 3.2 ((modified and centered) Calibration).** (a) (i) The estimator $\hat{\alpha}_N = \hat{\alpha}^*_N$ is a solution of the calibration equation (2.4). (ii) The estimator $\hat{\alpha}_N = \hat{\alpha}^c_N$ is a solution of the calibration equation (2.5). (iii) The estimator $\hat{\alpha}_N = \hat{\alpha}^c_N$ is a solution of the calibration equation (2.6).
(b) The distribution of $Z \in \mathbb{R}^k$ is not concentrated at 0 and has bounded support.
(c) $G$ is a strictly increasing continuously differentiable function on $\mathbb{R}$ such that $G(0) = 1$ and for all $x$, $-\infty < m_1 \leq G(x) \leq M_1 < \infty$ and $0 < G'(x) \leq M_2 < \infty$, where $G'$ is the derivative of $G$.
(d) (i) $P_0 Z^{\otimes 2}$ is finite and positive definite. (ii) $P_0 [\pi_0(V)^{-1} (1 - \pi_0(V)) Z^{\otimes 2}]$ is finite and positive definite. (iii) $P_0 [\pi_0(V)^{-1} (1 - \pi_0(V)) (Z - \mu Z)^{\otimes 2}]$ is finite and positive definite where $\mu_Z = P Z$.
(e) The “true” parameter $\alpha_0 = 0$.

Condition 3.1 (f) may seem unnatural at first, but in practice the phase II sample size $n_j$ can be chosen by the investigator so that the sampling probability $p_j$ can be understood to be automatically chosen to satisfy $n_j = [N_j p_j]$. The other parts of Condition 3.1 are standard in binary regression, and Condition 3.2 is similar to Condition 3.1.
Asymptotic properties of $\hat{\alpha}_N$ for all cases (estimated weights and (modified and centered) calibration) are proved in [34].

3.2. Regular rate for a nuisance parameter

We assume the following conditions.

**Condition 3.3** (Consistency). The estimator $(\hat{\theta}_N, \hat{\eta}_N)$ is consistent for $(\theta_0, \eta_0)$ and solves the weighted likelihood equations (2.7), where $\mathbb{P}_N$ may be replaced by $\mathbb{P}_N^\pi, c \mathbb{P}_N^\pi, mc$ or $\mathbb{P}_N^\pi, oc$ for the estimators with estimated weights, calibration, modified calibration or centered calibration.

**Condition 3.4** (Asymptotic equicontinuity). Let $F_1(\delta) = \{ \hat{\ell}_{\theta, \eta} : |\theta - \theta_0| + ||\eta - \eta_0|| < \delta \}$ and $F_2(\delta) = \{ B_{\theta, \eta} h - P_{\theta, \eta} B_{\theta, \eta} h : h \in \mathcal{H}, |\theta - \theta_0| + ||\eta - \eta_0|| < \delta \}$. There exists a $\delta_0 > 0$ such that (1) $F_k(\delta_0), k = 1, 2$ are $F_0$-Donsker and sup$_{h \in \mathcal{H}} P_0|f_j - f_0,j|^2 \rightarrow 0$, as $|\theta - \theta_0| + ||\eta - \eta_0|| \rightarrow 0$, for every $f_j \in F_j(\delta_0), j = 1, 2$, where $f_{0,1} = \hat{\ell}_{\theta, \eta_0}$ and $f_{0,2} = B_0 h - P_0 B_0 h$, (2) $F_k(\delta_0), k = 1, 2$, have integrable envelopes.

**Condition 3.5.** The map $\Psi = (\Psi_1, \Psi_2) : \Theta \times \mathcal{H} \rightarrow \mathbb{R}^p \times \ell^\infty(\mathcal{H})$ with components

$$
\Psi_1(\theta, \eta) \equiv P_0 \Psi N, 1(\theta, \eta) = P_0 \hat{\ell}_{\theta, \eta}, \\
\Psi_2(\theta, \eta) h \equiv P_0 \Psi N, 2(\theta, \eta) = P_0 B_{\theta, \eta} h - P_{\theta, \eta} B_{\theta, \eta} h, \ h \in \mathcal{H},
$$

has a continuously invertible Fréchet derivative map $\hat{\Psi}_0 = (\hat{\Psi}_{1,1}, \hat{\Psi}_{1,2}, \hat{\Psi}_{2,1}, \hat{\Psi}_{2,2})$ at $(\theta_0, \eta_0)$ given by $\hat{\Psi}_{ij}(\theta_0, \eta_0) h = P_0 \hat{\ell}_{i,j, \theta_0, \eta_0, h}$, $i, j \in \{1, 2\}$ in terms of $L_2(P_0)$ derivatives of $\psi_{1, \theta, \eta, h} = \hat{\ell}_{\theta, \eta}$ and $\psi_{2, \theta, \eta, h} = B_{\theta, \eta} h - P_{\theta, \eta} B_{\theta, \eta} h$; that is,

$$
\sup_{h \in \mathcal{H}} \left\{ P_0 \left( \psi_{1, \theta, \eta_0, h} - \psi_{1, \theta_0, \eta_0, h} - \hat{\psi}_{1, \theta_0, \eta_0, h}(\theta - \theta_0) \right)^2 \right\}^{1/2} = o(||\theta - \theta_0||), \\
\sup_{h \in \mathcal{H}} \left\{ P_0 \left( \psi_{2, \theta, \eta_0, h} - \psi_{2, \theta_0, \eta_0, h} - \hat{\psi}_{2, \theta_0, \eta_0, h}(\eta - \eta_0) \right)^2 \right\}^{1/2} = o(||\eta - \eta_0||).
$$

Furthermore, $\hat{\Psi}_0$ admits a partition

$$(\theta - \theta_0, \eta - \eta) \mapsto \left( \begin{array}{c} \hat{\Psi}_{11} \\ \hat{\Psi}_{12} \\ \hat{\Psi}_{21} \\ \hat{\Psi}_{22} \end{array} \right) \left( \begin{array}{c} \theta - \theta_0 \\ \eta - \eta_0 \end{array} \right),$$

where

$$
\hat{\Psi}_{11}(\theta - \theta_0) = - P_{\theta_0, \eta_0} \hat{\ell}_{\theta_0, \eta_0} \ell_{\theta_0, \eta_0}^T (\theta - \theta_0), \\
\hat{\Psi}_{12}(\eta - \eta_0) = - \int B_{\eta_0, \eta_0}^* \ell_{\theta_0, \eta_0} d(\eta - \eta_0), \\
\hat{\Psi}_{21}(\theta - \theta_0) h = - P_{\theta_0, \eta_0} B_{\theta_0, \eta_0} \ell_{\theta_0, \eta_0}^T (\theta - \theta_0), \\
\hat{\Psi}_{22}(\eta - \eta_0) h = - \int B_{\theta_0, \eta_0}^* B_{\theta_0, \eta_0} h d(\eta - \eta_0),
$$

and $B_{\theta_0, \eta_0}^* B_{\theta_0, \eta_0}$ is continuously invertible.
Let $\tilde{I}_0 = P_0[(I - B_0(B_0^T B_0)^{-1} B_0^T)\ell_0 \ell_0^T]$ be the efficient information for $\theta$ and $\tilde{e}_0 = \tilde{I}_0^{-1}(I - B_0(B_0^T B_0)^{-1} B_0^T)\ell_0$ be the efficient influence function for $\theta$ for the semiparametric model with complete data.

**Theorem 3.1.** Under the Conditions 3.1-3.5,
\[
\sqrt{N}(\hat{\theta}_N - \theta_0) = \sqrt{N}P_0^{\pi_n}(\tilde{e}_0 + o_P(1)) \rightsquigarrow Z \sim N_p(0, \Sigma),
\]
\[
\sqrt{N}(\hat{\theta}_{N,c} - \theta_0) = \sqrt{N}P_0^{\pi_n}(\tilde{e}_0 + o_P(1)) \rightsquigarrow Z_c \sim N_p(0, \Sigma_c),
\]
\[
\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) = \sqrt{N}P_0^{\pi_n,mc}(\tilde{e}_0 + o_P(1)) \rightsquigarrow Z_{mc} \sim N_p(0, \Sigma_{mc}),
\]
\[
\sqrt{N}(\hat{\theta}_{N,cc} - \theta_0) = \sqrt{N}P_0^{\pi_n,cc}(\tilde{e}_0 + o_P(1)) \rightsquigarrow Z_{cc} \sim N_p(0, \Sigma_{cc}),
\]

where
\[
\Sigma = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0j}(\tilde{e}_0),
\]
(3.8)
\[
\Sigma_c = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0j}((I - Q_c)\tilde{e}_0),
\]
(3.9)
\[
\Sigma_c = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0j}((I - Q_c)\tilde{e}_0),
\]
(3.10)
\[
\Sigma_{mc} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0j}((I - Q_{mc})\tilde{e}_0),
\]
(3.11)
\[
\Sigma_{cc} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0j}((I - Q_{cc})\tilde{e}_0),
\]
(3.12)

and (recall Conditions 3.1 and 3.2)

\[
Q_{cf} \equiv P_0[\pi_0^{-1}(V)f\hat{G}_c(Z^T \alpha_0)Z]S_0^{-1}(1 - \pi_0(V))^{-1}\hat{G}_c(Z^T \alpha_0)Z,
\]
\[
Q_{cf} \equiv P_0[fZ^T\{P_0Z^\otimes 2\}]^{-1}Z,
\]
\[
Q_{mcf} \equiv P_0[\pi_0^{-1}(V) - 1)fZ^T\{P_0[\pi_0^{-1}(V) - 1]Z^\otimes 2\}]^{-1}Z,
\]
\[
Q_{ccf} \equiv P_0[\pi_0^{-1}(V) - 1)f(Z - \mu_Z)^2\{P_0[(\pi_0^{-1}(V) - 1)(Z - \mu_Z)^\otimes 2]\}^{-1}(Z - \mu_Z).
\]

**Remark 3.1.** Our conditions in Theorem 3.1 are the same as those in [7] except the integrability condition. Our Condition 3.4 (2) requires existence of integrable envelopes for class of scores while the condition (A1*) in [7] requires square integrable envelopes. Note that this integrability condition is required only for the WLE with estimated weights and (modified and centered) calibration, as in [6].

### 3.3. Non-regular rate for a nuisance parameter

Set

\[
B_{\theta,q}[h] = (B_{\theta,q}h_1, \ldots, B_{\theta,q}h_p)^T
\]
for $h = (h_1, \ldots, h_p)^T$ where $h_k \in H$ for each $k = 1, \ldots, p$. We assume the following conditions.

**Condition 3.6** (Consistency and rate of convergence). An estimator $(\hat{\theta}_N, \hat{\eta}_N)$ of $(\theta_0, \eta_0)$ satisfies $|\hat{\theta}_N - \theta_0| = o_P(1)$, and $||\hat{\eta}_N - \eta_0|| = O_P(N^{-\beta})$ for some $\beta > 0$.

**Condition 3.7** (Positive information). There is an $h^* = (h_1^*, \ldots, h_p^*)$, where $h_k^* \in H$ for $k = 1, \ldots, p$, such that

$$P_0 \left\{ (\ell_0 - B_0[h^*]) B_0 h \right\} = 0$$

for all $h \in H$. Furthermore, the efficient information $I_0 = P_0 \left( (\ell_0 - B_0[h^*]) \right)^2$ for $\theta$ for the semiparametric model with complete data is finite and nonsingular. Denote the efficient influence function for the semiparametric model with complete data by $\ell_0 \equiv I_0^{-1}(\ell_0 - B_0[h^*]).$

**Condition 3.8** (Asymptotic equicontinuity). (1) For any $\delta_N \downarrow 0$ and $C > 0$,

$$\sup_{|\theta - \theta_0| \leq \delta_N, ||\eta - \eta_0|| \leq CN^{-\beta}} |G_N(\hat{\theta}_N, \eta) - \hat{\theta}_0| = o_P(1),$$

$$\sup_{|\theta - \theta_0| \leq \delta_N, ||\eta - \eta_0|| \leq CN^{-\beta}} |G_N(B_{\theta, \eta}[h^*])| = o_P(1).$$

(2) There exists a $\delta > 0$ such that the classes $\{\ell_{\theta, \eta} : |\theta - \theta_0| + ||\eta - \eta_0|| \leq \delta\}$ and $\{B_{\theta, \eta}[h^*] : |\theta - \theta_0| + ||\eta - \eta_0|| \leq \delta\}$ are $P_0$-Glivenko-Cantelli and have integrable envelopes. Moreover, $\ell_{\theta, \eta}$ and $B_{\theta, \eta}[h^*]$ are continuous with respect to $(\theta, \eta)$ either pointwise or in $L_1(P_0)$.

**Condition 3.9** (Smoothness of the model). For some $\alpha > 1$ satisfying $\alpha \beta > 1/2$ and for $(\theta, \eta)$ in the neighborhood $\{ (\theta, \eta) : |\theta - \theta_0| \leq \delta_N, ||\eta - \eta_0|| \leq CN^{-\beta} \}$,

$$P_0 \left\{ (\ell_{\theta, \eta} - \ell_0 + \ell_0 B_0(\theta - \theta_0) + B_0[\eta - \eta_0]) \right\} = o(||\theta - \theta_0|| + O(||\eta - \eta_0||^\alpha)), $$

$$P_0 \left\{ (B_{\theta, \eta} - B_0)[h^*] + B_0[h^*] (\ell_0 (\theta - \theta_0) + B_0[\eta - \eta_0]) \right\} = o(||\theta - \theta_0|| + O(||\eta - \eta_0||^\alpha)).$$

In the previous section, we required that the WLE solves the weighted likelihood equations (2.7) for all $h \in H$. Here, we only assume that the WLE $(\hat{\theta}_N, \hat{\eta}_N)$ satisfies the weighted likelihood equations

$$\Psi_{N,1}(\theta, \eta, \alpha) = P_N(\hat{\theta}_N, \eta) = o_P(\left( N^{-1/2} \right)), $$

$$\Psi_{N,2}(\theta, \eta, \alpha) [h^*] = P_N(B_{\theta, \eta}[h^*]) = o_P(\left( N^{-1/2} \right)).$$

The corresponding WLE with estimated weights, $(\hat{\theta}_{N,c}, \hat{\eta}_{N,c})$, the calibrated WLE $(\tilde{\theta}_{N,c}, \tilde{\eta}_{N,c})$, the WLE $(\tilde{\theta}_{N,mc}, \tilde{\eta}_{N,mc})$ with modified calibration and the
WLE (\(\hat{\theta}_{N,cc}, \hat{\eta}_{N,cc}\)) with centered calibration satisfy (3.13) with \(\mathbb{P}_N^\pi\) replaced by \(\mathbb{P}_{N,c}^\pi, \mathbb{P}_{N,c}^\tau, \mathbb{P}_{N,mc}^\tau\) or \(\mathbb{P}_{N,cc}^\tau\), respectively.

**Theorem 3.2.** Suppose that the WLE is a solution of (3.13) where \(\mathbb{P}_N^\pi\) may be replaced by \(\mathbb{P}_{N,c}^\pi, \mathbb{P}_{N,c}^\tau, \mathbb{P}_{N,mc}^\tau\) or \(\mathbb{P}_{N,cc}^\tau\) for the estimators with estimated weights, calibration, modified calibration and centered calibration. Under the Conditions 3.1, 3.2 and 3.6-3.9.

\[
\begin{align*}
\sqrt{N}(\hat{\theta}_{N} - \theta_0) &= \frac{\sqrt{N}\mathbb{P}_N^\pi \hat{\theta}_0 + o_p(1)} {\sim} Z \sim N_p(0, \Sigma), \\
\sqrt{N}(\hat{\theta}_{N,c} - \theta_0) &= \frac{\sqrt{N}\mathbb{P}_{N,c}^\tau \hat{\theta}_0 + o_p(1)} {\sim} Z_e \sim N_p(0, \Sigma_e), \\
\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) &= \frac{\sqrt{N}\mathbb{P}_{N,mc}^\pi \hat{\theta}_0 + o_p(1)} {\sim} Z_{mc} \sim N_p(0, \Sigma_{mc}), \\
\sqrt{N}(\hat{\theta}_{N,cc} - \theta_0) &= \frac{\sqrt{N}\mathbb{P}_{N,cc}^\tau \hat{\theta}_0 + o_p(1)} {\sim} Z_{cc} \sim N_p(0, \Sigma_{cc}),
\end{align*}
\]

where \(\Sigma, \Sigma_c, \Sigma_e, \Sigma_{mc}\) and \(\Sigma_{cc}\) are as defined in (3.8) - (3.12) of Theorem 3.1, but now \(\hat{\theta}_0\) and \(\hat{\ell}_0\) are defined in Condition 3.7, and \(Q_e, Q_c, Q_{mc}\) and \(Q_{cc}\) are defined in Theorem 3.1.

**Remark 3.2.** Our conditions are identical to those of the Z-theorem of [15] except the Condition 3.8 (2). This additional condition is not stringent. First, the Glivenko-Cantelli condition is usually assumed to prove consistency of estimators before deriving asymptotic distributions. Second, a stronger \(L_2(P_0)\)-continuity condition is standard as is seen in Condition 3.4 (See also 25.8 of [41] for a nice discussion of regularity conditions for efficient score equations with complete data). Note that the \(L_1(P_0)\)-continuity condition is only required for the WLE’s with estimated weights and (modified and centered) calibration. Another way to understand the relative weakness of the Condition 3.8 (2) is to compare it with standard conditions for bootstrapping \(Z\)-estimators because the IPW empirical process is closely related to the exchangeably weighted bootstrap empirical process. See, for example, conditions A.4 and A.5 in [43]. The differentiability condition A.4, which implies continuity, corresponds to our \(L_1(P_0)\)-continuity condition. However, we do not impose a condition similar to the weak \(L_2(P_0)\) condition A.5. In fact, our Lemma 5.4 in Section 5 with the Glivenko-Cantelli condition can be used to relax condition A.5 of [43].

### 3.4. Comparisons of methods

We compare asymptotic variances of five WLE’s in view of improvement by adjusting weights and change of design. To make these comparisons clearly, we first need to give a clear statement of the result corresponding to Theorem 3.1 for stratified Bernoulli sampling.

#### 3.4.1. Stratified Bernoulli sampling

We present asymptotic normality of the WLE’s, \(\hat{\theta}_{N,Bern}, \hat{\theta}_{N,c,Bern, \hat{\theta}_{N,mc,Bern, \hat{\theta}_{N,cc,Bern}}\) under stratified Bernoulli sampling where all subjects are independent with the sampling probability \(p_j\) if \(V \in V_j\).
Theorem 3.3. Suppose the Conditions 3.1 (except 3.1(f)) and 3.2 hold. Let \( \xi_i \in \{0, 1\} \) and \( \xi \) be i.i.d. with \( E[\xi|V] = \pi_0(V) = \sum_{j=1}^J p_j I(V \in V_j) \).

1) Suppose that the WLE is a solution of (3.13) where \( \mathbb{P}_N^\pi \) may be replaced by \( \mathbb{P}_N^\pi, \mathbb{P}_N^{\pi,c}, \mathbb{P}_N^{\pi,mc} \) or \( \mathbb{P}_N^{\pi,cc} \) for the estimators with estimated weights, calibration, modified calibration and centered calibration. Under the same conditions as in Theorem 3.1,

\[
\sqrt{N} (\hat{\theta}^\text{Bern}_N - \theta_0) = \sqrt{N} \mathbb{P}_N^\pi \hat{\theta}_0 + o_p \to \mathcal{N} (0, \Sigma^\text{Bern}),
\]

\[
\sqrt{N} (\hat{\theta}^\text{Bern}_{N,c} - \theta_0) = \sqrt{N} \mathbb{P}_N^{\pi,c} \hat{\theta}_0 + o_p \to \mathcal{N} (0, \Sigma^\text{Bern}),
\]

\[
\sqrt{N} (\hat{\theta}^\text{Bern}_{N,mc} - \theta_0) = \sqrt{N} \mathbb{P}_N^{\pi,mc} \hat{\theta}_0 + o_p \to \mathcal{N} (0, \Sigma^\text{Bern}),
\]

\[
\sqrt{N} (\hat{\theta}^\text{Bern}_{N,cc} - \theta_0) = \sqrt{N} \mathbb{P}_N^{\pi,cc} \hat{\theta}_0 + o_p \to \mathcal{N} (0, \Sigma^\text{Bern}),
\]

where

\[
\Sigma^\text{Bern} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} (\hat{\ell}_0)^{\otimes 2},
\]

(3.14)

\[
\Sigma^\text{Bern} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_c) \hat{\ell}_0)^{\otimes 2},
\]

(3.15)

\[
\Sigma^\text{Bern} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_{mc}) \hat{\ell}_0)^{\otimes 2},
\]

(3.16)

\[
\Sigma^\text{Bern} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_{cc}) \hat{\ell}_0)^{\otimes 2},
\]

(3.17)

\[
\Sigma^\text{Bern} = I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_{cc}) \hat{\ell}_0)^{\otimes 2},
\]

(3.18)

where \( Q_c, Q_{mc}, Q_{mc} \) and \( Q_{cc} \) are defined in Theorem 3.1.

2) Under the same conditions as in Theorem 3.2, the same conclusion in (1) holds with \( I_0 \) and \( \hat{\ell}_0 \) replaced by those defined in Condition 3.7.

Comparing the variance-covariance matrices in Theorem 3.3 to those in Theorems 3.1 and 3.2, we obtain the following corollary comparing designs. All estimators have smaller variances under sampling without replacement.

Corollary 3.1.

\[
\Sigma = \Sigma^\text{Bern} - \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} (\hat{\ell}_0)^{\otimes 2},
\]

\[
\Sigma = \Sigma^\text{Bern} - \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_c) \hat{\ell}_0)^{\otimes 2},
\]

\[
\Sigma = \Sigma^\text{Bern} - \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_{mc}) \hat{\ell}_0)^{\otimes 2},
\]

\[
\Sigma = \Sigma^\text{Bern} - \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0|j} ((I - Q_{cc}) \hat{\ell}_0)^{\otimes 2}.
\]
\[ \Sigma_c = \Sigma_{c,\text{Bern}} - \sum_{j=1}^{J} \nu_j \left( \frac{1 - p_j}{p_j} \right) \{ P_{0j}(I - Q_c) \} \tilde{\ell}_0, \]

\[ \Sigma_{mc} = \Sigma_{mc,\text{Bern}} - \sum_{j=1}^{J} \nu_j \left( \frac{1 - p_j}{p_j} \right) \{ P_{0j}(I - Q_{mc}) \} \tilde{\ell}_0, \]

\[ \Sigma_{cc} = \Sigma_{cc,\text{Bern}} - \sum_{j=1}^{J} \nu_j \left( \frac{1 - p_j}{p_j} \right) \{ P_{0j}(I - Q_{cc}) \} \tilde{\ell}_0. \]

Variance formulae (3.15), (3.17) and (3.18) have the following alternative representations which show the efficiency gains over the plain WLE under Bernoulli sampling.

**Corollary 3.2.** Under the same conditions as in Theorem 3.3,

\[ \Sigma_{e,\text{Bern}} = \Sigma_{\text{Bern}} - \text{Var} \left( \frac{\xi - \pi_0(V)}{\pi_0(V)} Q_e \tilde{\ell}_0 \right), \]

\[ \Sigma_{mc,\text{Bern}} = \Sigma_{\text{Bern}} - \text{Var} \left( \frac{\xi - \pi_0(V)}{\pi_0(V)} Q_{mc} \tilde{\ell}_0 \right), \]

\[ \Sigma_{cc,\text{Bern}} = \Sigma_{\text{Bern}} - \text{Var} \left( \frac{\xi - \pi_0(V)}{\pi_0(V)} Q_{cc} \tilde{\ell}_0 \right). \]

Thus modified calibration and centered calibration yield improved efficiency over the plain WLE, while (ordinary) calibration does not yield a guaranteed improvement in general. We do not have similar formulas for sampling without replacement except for the special case involving within-stratum calibration described in part (2) of Corollary 3.3 below.

### 3.4.2. Within-stratum adjustment of weights

[3] proposed calibration within each stratum to improve the calibrated WLE. Let \( Z^{(j)} = I(V \in V_j) Z^T \) and \( \tilde{Z} = (Z^{(1)}, \ldots, Z^{(J)})^T \), and consider calibration on \( \tilde{Z} \). The calibration equation (2.4) becomes

\[ \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i G_c(\tilde{Z}_i; \alpha)}{\pi_0(V_i)} Z_{ij} I(V_i \in V_j) = \frac{1}{N} \sum_{i=1}^{N} Z_{ij} I(V_i \in V_j), \quad j = 1, \ldots, J, \]

where \( \alpha \in \mathbb{R}^{nK} \). We call this special case *within-stratum calibration*. We define *within-stratum modified and centered calibration* analogously.

We also call the method of adjusting weights *within-stratum estimated weights* when binary regression is done within each stratum. Recall that the first \( J \) elements of \( Z \) for estimated weights are stratum membership indicators and the rest are other auxiliary variables, say \( Z^{(2)} \). Within-stratum estimated weights uses \( \tilde{Z} = (Z^{(1)}, \ldots, Z^{(J)})^T \) where \( Z^{(j)} = I(V \in V_j)(Z^{(2)})^T \) with 1 included in \( Z^{(2)} \). The “true” parameter \( \tilde{\alpha}_0 \) has zero for all elements except having \( G_e^{-1}(p_j) \) for the element corresponding to \( I(V \in V_j), \ j = 1, \ldots, J. \)
The following corollary summarizes within-stratum adjustment of weights under stratified Bernoulli sampling and sampling without replacement.

**Corollary 3.3.** (1) (Bernoulli) Under the same conditions as in Theorem 3.3 with Z replaced by \( \tilde{Z} \) and \( \tilde{\alpha}_0 \) replaced by \( \tilde{\alpha}_0 \) for within-stratum estimated weights,

\[
\Sigma_{c}^{\text{Bern}} = \Sigma_{o}^{\text{Bern}} - \sum_{j=1}^{J} \nu_j \frac{1}{p_j} P_{0lj} \left( Q_{c}^{(j)} \tilde{\ell}_0 \right)^{\otimes 2}, \tag{3.19}
\]

\[
\Sigma_{cc}^{\text{Bern}} = \Sigma_{mc}^{\text{Bern}} = \Sigma_{o}^{\text{Bern}} - \sum_{j=1}^{J} \nu_j \frac{1}{p_j} P_{0lj} \left( Q_{cc}^{(j)} \tilde{\ell}_0 \right)^{\otimes 2}, \tag{3.20}
\]

where

\[
Q_{c}^{(j)} f = \frac{P_{0lj} \left[ f \hat{G}_{c}(\tilde{Z}^T \tilde{\alpha}_0)(Z^{[2]} \tilde{\alpha}_0)^{\otimes 2} \right]^{-1} \hat{G}_{c}(\tilde{Z}^T \tilde{\alpha}_0)I(V \in V_j)Z^{[2]}}{P_{0lj} \left[ P_{0lj} (Z^{[2]} \tilde{\alpha}_0)^{\otimes 2} \right]^{-1} I(V \in V_j)Z},
\]

\[
Q_{cc}^{(j)} f = \frac{P_{0lj} \left[ f(Z - \mu_{Z,j})^{\otimes 2} \right]^{-1} I(V \in V_j)Z}{P_{0lj} \left[ P_{0lj} (Z - \mu_{Z,j})^{\otimes 2} \right]^{-1} I(V \in V_j)(Z - \mu_{Z,j})},
\]

with \( \mu_{Z,j} = E[I(V \in V_j)Z] \) for \( j = 1, \ldots, J \).

(2) (without replacement) Under the same conditions as in Theorem 3.1 or Theorem 3.2 with Z is replaced by \( \tilde{Z} \),

\[
\Sigma_{cc} = \Sigma - \sum_{j=1}^{J} \nu_j \frac{1}{p_j} \text{Var}_{0lj} \left( Q_{cc}^{(j)} \tilde{\ell}_0 \right). \tag{3.22}
\]

**3.4.3. Comparisons**

We summarize Corollaries 3.1-3.3. All estimators have reduced variance under the sampling without replacement design in comparison to Bernoulli sampling. Every method of adjusting weights improves efficiency over the plain WLE in a certain design and with a certain range of adjustment of weights (within-stratum or “across-strata” adjustment). However, particularly notable among all methods is centered calibration. While other methods gain efficiency only under stratified Bernoulli sampling, centered calibration improves efficiency over the plain WLE under both sampling schemes. There is no known method of “across-strata” adjustment that is guaranteed to gain efficiency over the plain WLE under stratified sampling without replacement.

There are close connections among all methods. When the auxiliary variables have mean zero, then centered and modified calibration are essentially the same. Within-stratum calibration and within-stratum modified calibration give the
same asymptotic variance. For \( Z \) and \( \alpha_0 \) defined for estimated weights and \( \tilde{Z} \) and \( \tilde{\alpha}_0 \) defined for within-stratum estimated weights, modified calibration based on \((1-\pi_0(V))^{-1}G_e(Z^T\alpha_0)Z \) and within-stratum calibration based on \( G_e(Z^T\alpha_0) \) perform in the same way as the estimated weights and within-stratum estimated weights, respectively. Because of these connections among methods, there is no single method superior to others in each scenario. Performance depends on choice and transformation of auxiliary variables, the true distribution \( P_0 \), and the design. For our sampling scheme, within-stratum centered calibration is the only guaranteed method to gain efficiency while other methods may perform even worse than the plain WLE.

4. Examples

To prove asymptotic normality of WLE’s, consistency and rate of convergence need to be established in order to apply our \( Z \)-theorems in the previous section. To this end, general results on IPW empirical processes discussed in the next section will be useful. We illustrate this in the Cox proportional hazards models with right censoring and interval censoring under two-phase sampling.

Let \( T \sim F \) be a failure time, and \( X \) be a vector of covariates with bounded supports in the regression model. The Cox model ([11]) specifies the relationship

\[
\Lambda(t|x) = \exp(\theta^T x)\Lambda(t),
\]

where \( \theta \in \Theta \subset \mathbb{R}^p \) is the regression parameter, \( \Lambda \in H \) is the (baseline) cumulative hazard function. Here the space \( H \) for the nuisance parameter \( \Lambda \) is the set of nonnegative, nondecreasing cadlag functions defined on the positive line. The true parameters are \( \theta_0 \) and \( \Lambda_0 \).

In addition to \( X \), let \( U \) be a vector of auxiliary variables collected at phase I which are correlated with the covariate \( X \). For simplicity of notation, we assume that the covariates \( X \) are only observed for the subject sampled at phase II. Thus, if some covariates \( X \) are available at phase I, then we include an identical copy \( X' \) of \( X \) in the vector of \( U \).

4.1. Cox model with right censored data

Under right censoring, we only observe the minimum of the failure time \( T \) and the censoring time \( C \sim G \). Define the observed time \( Y = T\wedge C \) and the censoring indicator \( \Delta = I(T \leq C) \). The phase I data is \( V = (Y, \Delta, U) \), and the observed data is \( (Y, \Delta, X, U, \xi) \) where \( \xi \) is the sampling indicator.

We assume the following conditions.

**Condition 4.1.** The finite-dimensional parameter space \( \Theta \) is compact and contains the true parameter \( \theta_0 \) as an interior point.

**Condition 4.2.** The failure time \( T \) and the censoring time \( C \) are conditionally independent given \( X \), and that there is \( \tau > 0 \) such that \( P(T > \tau) > 0 \) and
\( P(C \geq \tau) = P(C = \tau) > 0 \). Both \( T \) and \( C \) have continuous conditional densities given the covariates \( X = x \).

**Condition 4.3.** The covariate \( X \) has bounded support. For any measurable function \( h \), \( P(X \neq h(Y)) > 0 \).

Let \( \lambda(t) = (d/dt)\Lambda(t) \) be the baseline hazard function. With complete data, the density of \((Y, \Delta, X)\) is

\[
p_{\theta, \Lambda}(y, \delta, x) = \left( \lambda(y)e^{\theta^T x - \Lambda(y)e^{\theta^T x}}(1 - G)(y|x) \right) \delta \left( e^{-\Lambda(y)e^{\theta^T x}}g(y|x) \right)^{1-\delta} p_X(x),
\]

where \( p_X \) is the marginal density of \( X \) and \( g(\cdot|x) \) is the conditional density of \( C \) given \( X = x \). The score for \( \theta \) is given by

\[
\dot{\ell}_{\theta, \Lambda}(y, \delta, x) = x \left( \delta - e^{\theta^T x} \Lambda(y) \right),
\]

and the score operator \( B_{\theta, \Lambda} : \mathcal{H} \mapsto L_2(P_{\theta, \Lambda}) \) is defined on the unit ball \( \mathcal{H} \) in the space \( BV[0, \tau] \) such that

\[
B_{\theta, \Lambda} h(y, \delta, x) = \delta h(y) - e^{\theta^T x} \int_{[0, y]} h d\Lambda.
\]

Because the likelihood based on the density above does not yield the MLE with complete data, we define the log likelihood for one observation with complete data by

\[
\ell_{\theta, \Lambda}(y, \delta, x) = \log \left\{ \left( e^{\theta^T x} \Lambda(y) \right)^{\delta} e^{-\Lambda(y)e^{\theta^T x}} \right\} = \delta \Lambda(y) + \delta \theta^T x - e^{\theta^T x} \Lambda(y),
\]

where \( \Lambda\{t\} \) is the (point) mass of \( \Lambda \) at \( t \). Then maximizing the weighted log likelihood \( \mathbb{P}_N^\tau \ell_{\theta, \Lambda} \) reduces to solving the system of equations \( \mathbb{P}_N^\tau \dot{\ell}_{\theta, \Lambda} = 0 \) and \( \mathbb{P}_N^\tau B_{\theta, \Lambda} h = 0 \) for every \( h \in \mathcal{H} \). The efficient score for \( \theta \) with complete data is given by

\[
\dot{\ell}^*_{\theta, \Lambda_0}(y, \delta, x) = \delta \left( x - \frac{M_1}{M_0}(y) \right) - e^{\theta^T x} \int_{[0, y]} \delta \left( x - \frac{M_1}{M_0}(t) \right) d\Lambda_0(t),
\]

and the efficient information for \( \theta \) with complete data is

\[
\tilde{I}_{\theta_0, \Lambda_0} = E \left[ (\dot{\ell}^*_{\theta_0, \Lambda_0}) \otimes^2 \right] = E e^{\theta^T X} \int_0^\tau \left( X - \frac{M_1}{M_0}(y) \right) \otimes^2 (1 - G)(y|X) d\Lambda_0(y),
\]

where

\[
M_k(s) = P_{\theta_0, \Lambda_0} X^k e^{\theta^T X} I(Y \geq s), \quad k = 0, 1.
\]

**Theorem 4.1** (Consistency). Under Conditions 3.1, 3.2, 4.1-4.3, the WLE’s are consistent for \((\theta_0, \Lambda_0)\).
Proof. We only consider the WLE with modified calibration. Proofs for the other four estimators are similar. Our proof closely follows the consistency proof for the MLE with complete data in [39].

Because of the assumption on \( \tau \), we restrict our attention to the interval \([0, \tau]\). For a bounded function \( h \in L_2(\Lambda) \), define a perturbation \( d\hat{\Lambda}_{N,mc,t} = (1 + th)d\Lambda_{N,mc} \) of \( \hat{\Lambda}_{N,mc} \). The weighted log likelihood with modified calibration, \( \mathbb{P}_N^{\pi,mc} e^{\hat{\theta}_N^{\pi,mc}X} \), evaluated at \( (\hat{\theta}_{N,mc}, \hat{\Lambda}_{N,mc,t}) \) viewed as a function of \( t \) is maximal at \( t = 0 \) by the definition of the WLE with modified calibration. Thus, differentiating at \( t = 0 \) we obtain \( \mathbb{P}_N^{\pi,mc} B_{\hat{\theta}_{N,mc},\hat{\Lambda}_{N,mc}} h = 0 \), or

\[
\mathbb{P}_N^{\pi,mc} \Delta h(Y) = \mathbb{P}_N^{\pi,mc} e^{\hat{\theta}_N^{\pi,mc}X} \int_{[0,Y]} h d\hat{\Lambda}_{N,mc} = \int \mathbb{P}_N^{\pi,mc} e^{\hat{\theta}_N^{\pi,mc}X} I_{[Y \geq s]} h(s) d\hat{\Lambda}_{N,mc}(s).
\]

Let \( \hat{M}_{N,0}(s) = \mathbb{P}_N^{\pi,mc} e^{\hat{\theta}_N^{\pi,mc}X} I(Y \geq s) \). Replacing \( h \) in the above display by \( h/\hat{M}_{N,0} \) yields

\[
\hat{\Lambda}_{N,mc} h = \int \frac{h(s)}{\hat{M}_{N,0}(s)} \mathbb{P}_N^{\pi,mc} e^{\hat{\theta}_N^{\pi,mc}X} I(Y \geq s) d\hat{\Lambda}_{N,mc}(s) = \mathbb{P}_N^{\pi,mc} \frac{\Delta h(Y)}{\hat{M}_{N,0}(Y)}.
\]

Similar reasoning via \( P_0 B_0 h = 0 \) leads to \( \Lambda_0 h = P_0 \Delta h(Y)/M_0(Y) \). Let

\[
\tilde{\Lambda}_N h = \mathbb{P}_N^{\pi,mc} \Delta h(Y)/M_0(Y).
\]

Since \( P(T > \tau) > 0 \) and \( P(C = \tau) > 0 \), we have for \( s \leq \tau \) that \( M_0(s) \geq M_0(\tau) > 0 \). The function \( (y, \delta) \mapsto \delta h(y)/M_0(y) \) is bounded, and therefore \( \{\delta h(y)/M_0(y) : h \in \mathcal{H}\} \) is Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem [40] and the fact that \( \mathcal{H} \) is Glivenko-Cantelli. Thus, \( \|\hat{\Lambda}_N\|_{\mathcal{H}} \rightarrow P^* \|P_{0,\Lambda_0} \Delta h(Y)/M_0(Y)\|_{\mathcal{H}} = \|\Lambda_0\|_{\mathcal{H}} \). Moreover, since \( \hat{\Lambda}_{N,mc}\{Y_i\} = \hat{\Lambda}_{N,mc}\delta_{Y_i} = N^{-1}((\sum_{i=1}^n \delta_{\hat{\Lambda}_{N,mc}(V_i)}(\Delta_i)) / (\hat{M}_{N,0}(Y_i))) \), and similarly

\[
\hat{\Lambda}_N\{Y_i\} = N^{-1}(\sum_{i=1}^n \delta_{\hat{\Lambda}_N(V_i)}(\Delta_i) / M_0(Y_i)),
\]

it follows that

\[
\hat{\Lambda}_{N,mc}\{Y_i\}/\hat{\Lambda}_N\{Y_i\} = M_0(Y_i)/\hat{M}_{N,0}(Y_i).
\]

Since the weighted log likelihood with modified calibration evaluated at \( (\hat{\theta}_{N,mc}, \hat{\Lambda}_{N,mc}) \) is larger than at \( (\theta_0, \Lambda_N) \), we have

\[
0 \leq \mathbb{P}_N^{\pi,mc} \left( \ell_{\hat{\theta}_{N,mc},\hat{\Lambda}_{N,mc}} - \ell_{\theta_0,\Lambda_N} \right) = (\hat{\theta}_{N,mc} - \theta_0)^T \mathbb{P}_N^{\pi,mc} \Delta X - \mathbb{P}_N^{\pi,mc} \left( e^{\hat{\theta}_{N,mc}^{\pi,mc}X} \hat{\Lambda}_N(Y) - e^{\theta_0^{\pi,mc}X} \hat{\Lambda}_N(Y) \right) + \mathbb{P}_N^{\pi,mc} \Delta \log \frac{M_0(Y)}{\hat{M}_{N,0}(Y)}.
\]
We take the limit of this on $N$. Because $\Theta$ is compact, there is a subsequence of $\{\theta_N\}$ that converges to $\theta_\infty \in \Theta$. It follows by Theorem 5.1 that along the convergent subsequence of $\{\theta_N\}$

$$
(\hat{\theta}_N - \theta_0)^T \mathbb{P}_N^{\pi,mc} \Delta X \rightarrow_p (\theta_\infty - \theta_0)^T P_{\hat{\theta}_0,\Lambda_0} \Delta X.
$$

For the second term, note that $\hat{\Lambda}_N(\tau) = \mathbb{P}_N^{\pi,mc} e^{\theta_0 X} I(Y = \tau) \leq \mathbb{P}_N^{\pi,mc} e^{\theta_0 \bar{X}} \hat{\Lambda}_N(Y) = \mathbb{P}_N^{\pi,mc} \Delta \leq 1$. Here we use the Glivenko-Cantelli preservation theorem again so that

$$
\mathbb{P}_N^{\pi,mc} e^{\theta_0 \bar{X}} \hat{\Lambda}_N(Y) - e^{\theta_0 \bar{X}} \hat{\Lambda}_N(Y)
$$

along a subsequence of $\hat{\theta}_N, mc$.

For the third term, note that $\{\hat{M}, M_0\}$ is a subset of the class of monotone, bounded, cadlag functions, which is Glivenko-Cantelli, and hence so is it. Note also that $\hat{M}_{0}(\tau) = \mathbb{P}_N^{\pi,mc} e^{\theta_0 \bar{X}} I(Y = \tau)$ is bounded away from zero with probability tending to 1 since $P(T > \tau) > 0$ and $P(C = \tau) > 0$. Since $\hat{M}_{0}(t) \geq \hat{M}_{0}(\tau)$ for $t \leq \tau$, the set $\{\delta \log(M_0(y)/\hat{M}_{0}(y))\}$ is Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem again so that

$$
\mathbb{P}_N^{\pi,mc} \Delta \log(M_0(Y)/\hat{M}_{0}(Y)) = P_{\hat{\theta}_0,\Lambda_0} \Delta \log(M_0(Y)/\hat{M}_{0}(Y)) + o_p(1) \quad (4.24)
$$

by Theorem 5.1.

The set $\{\delta h(Y)/\hat{M}_{0}(y) : h \in \mathcal{H}\}$ is Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem [40] so that $\|\hat{\Lambda}_N\|_\mathcal{H} = \|P_{\hat{\theta}_0,\Lambda_0} \Delta h(Y)/\hat{M}_{0}(Y)\|_\mathcal{H} + o_p(1)$ by Theorem 5.1. Since we have by Theorem 5.1 that

$$
\hat{M}_{0}(s) = \mathbb{P}_N^{\pi,mc} e^{\theta_0 \bar{X}} I(Y \geq s) \rightarrow_p P_{\hat{\theta}_0,\Lambda_0} e^{\theta_0 \bar{X}} I(Y \geq s) \equiv M_{\infty,0}(s)
$$

uniformly in $s$, it follows by the dominated convergence theorem that

$$
\|\hat{\Lambda}_N\|_\mathcal{H} = \|P_{\hat{\theta}_0,\Lambda_0} \Delta h(Y)/\hat{M}_{0}(Y)\|_\mathcal{H} + o_p(1)
$$

along a subsequence of $\hat{\theta}_N$.

Apply the dominated convergence theorem to replace $\hat{\Lambda}_N, \bar{\Lambda}_N$, and $\hat{M}_{0,0}$ by $\Lambda_\infty, \Lambda_0$ and $M_{\infty,0}$ in (4.23) and (4.24) and conclude

$$
0 \geq (\theta_\infty - \theta_0)^T P_{\hat{\theta}_0,\Lambda_0} \Delta X - P_{\hat{\theta}_0,\Lambda_0} \left( e^{\theta_0 \bar{X}} \Lambda_\infty(Y) - e^{\theta_0 \bar{X}} \Lambda_0(Y) \right)
$$

+ $P_{\hat{\theta}_0,\Lambda_0} \Delta \log \frac{M_0(Y)}{M_{\infty}(Y)}$. 

(4.25)
Since $M_0/M_\infty = d\Lambda_\infty/d\Lambda_0$, (4.25) is in fact minus one times the Kullback-Leibler divergence
\[
K(P_{\theta_0, \Lambda_0}, P_{\theta_\infty, \Lambda_\infty}) = P_{\theta_0, \Lambda_0} \log \left\{ p_{\theta_0, \Lambda_0}/p_{\theta_\infty, \Lambda_\infty} \right\} \geq 0,
\]
for the complete data model. Thus, (4.25) is exactly zero. But since $K(P_{\theta_0, \Lambda_0}, P_{\theta, \Lambda})$ is strictly positive unless $(\theta, \Lambda) = (\theta_0, \Lambda_0)$ by the identifiability of parameters, we must have $(\theta_\infty, \Lambda_\infty) = (\theta_0, \Lambda_0)$. This is true for any subsequence of $\theta_{N,mc}$, and the result follows.

We apply our Z-theorem (Theorem 3.1) to show the asymptotic normality of the WLE’s.

**Theorem 4.2** (Asymptotic normality). Under Conditions 3.1, 3.2, 4.1-4.3,
\[
\sqrt{N}(\hat{\theta}_N - \theta_0) = \sqrt{N}P^c_N \hat{I}_{\theta_0, \Lambda_0} + o_p(1) \rightarrow_d N(0, \Sigma),
\]
\[
\sqrt{N}(\hat{\theta}_{N,e} - \theta_0) = \sqrt{N}P^e_N \hat{I}_{\theta_0, \Lambda_0} + o_p(1) \rightarrow_d N(0, \Sigma_e),
\]
\[
\sqrt{N}(\hat{\theta}_{N,c} - \theta_0) = \sqrt{N}P^c_N \hat{I}_{\theta_0, \Lambda_0} + o_p(1) \rightarrow_d N(0, \Sigma_c),
\]
\[
\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) = \sqrt{N}P^{mc}_N \hat{I}_{\theta_0, \Lambda_0} + o_p(1) \rightarrow_d N(0, \Sigma_{mc}),
\]
where $\hat{I}_{\theta_0, \Lambda_0} = I_{\theta_0, \Lambda_0}^{-1} \hat{\ell}_0^* \hat{\ell}_0^*$ is the efficient influence function for $\theta$ with complete data, and $\Sigma$, $\Sigma_e$, $\Sigma_c$, $\Sigma_{mc}$ and $\Sigma_{cc}$ are given in Theorem 3.1.

**Proof.** We verify the conditions of Theorem 3.1. Condition 3.3 holds by Theorem 4.1. Conditions 3.4 and 3.5 hold under the present hypotheses as was shown in [41], section 25.12.

For variance estimation regarding $\hat{\theta}_N$, $\hat{I}_N \equiv P^e_N \left\{ \hat{\ell}^*_{\theta_0, \Lambda_N} \right\}^{\otimes 2}$ can be used to estimate $I_0$. Letting $\hat{I}_0 = I_N^{-1} \hat{\ell}_0^* \hat{\ell}_0^*$, we can estimate $\text{Var}(\hat{\theta}_N)$ by $\hat{P}_j \hat{I}_0 \hat{P}_j^{\otimes 2} - \left\{ \hat{P}_j \hat{I}_0 \right\}^{\otimes 2}$ where $\hat{P}_j \hat{I}_0 \equiv P^e_N \hat{I}_0 I(V \in \mathcal{V}_j)$ and $\hat{P}_j \hat{I}_0^{\otimes 2} \equiv P^e_N \hat{I}_0^{\otimes 2} I(V \in \mathcal{V}_j)$. The other four cases are similar.

### 4.2. Cox model with interval censored data

Let $Y$ be a censoring time that is assumed to be conditionally independent of a failure time $T$ given a covariate vector $X$. Under the case 1 interval censoring, we do not observe $T$ but $(Y, \Delta)$ where $\Delta \equiv I(T \leq Y)$. The phase I data is $V = (Y, \Delta, U)$ and the observed data is $(Y, \Delta, \xi X, U, \xi)$ where $\xi$ is the sampling indicator.

With complete data, the log likelihood for one observation is given by
\[
\ell(\theta, F) = \delta \log \left\{ 1 - F(y)^{\exp(\theta^T x)} \right\} + (1 - \delta) \log F(y)^{\exp(\theta^T x)}
\]
\[
= \delta \log \left\{ 1 - \exp(-\Lambda(y) \exp(\theta^T x)) \right\} - (1 - \delta) \exp(\theta^T x) \Lambda(y)
\]
\[
= \ell(\theta, \Lambda),
\]
where $F$, $F$, and $A$ are related by $F = 1 - F = \exp(-\Lambda)$. The WLE $(\hat{\theta}_N, \hat{A}_N)$ of $(\theta, A)$ maximizes $P_N \ell(\theta, A)$.

The score for $\theta$ and the score operator $B_{\theta, A}$ for $A$ with complete data are

$$\ell_{\theta, A} = x \exp(\theta^T x) A(y) (\delta r(y, x; \theta, A) - (1 - \delta)),
B_{\theta, A}^T[h] = \exp(\theta^T x) h(y) \{\delta r(y, x; \theta, A) - (1 - \delta)\}.$$  

where

$$r(y, x; \theta, A) = \frac{\exp(-\exp(\theta^T x) A(y))}{1 - \exp(-\exp(\theta^T x) A(y))}$$

The efficient score for $\theta$ with complete data is given by

$$\ell_{\theta, A}^{(x)} = e^{\theta^T x} Q(y, \delta, x; \theta, A_0) \Lambda_0(y) \left\{ x \left( \frac{e^{\theta^T x} O(Y|X)|Y = y}}{E[s^{\theta^T x} O(Y|X)|Y = y]} \right) - \frac{E[s^{\theta^T x} O(Y|X)|Y = y]}{E[s^{\theta^T x} O(Y|X)|Y = y]} \right\},$$

where $Q(y, \delta, x; \theta, A) = \delta r(y, x; \theta, A) - (1 - \delta)$ and $O(y|x) = r(y, x; \theta, A_0)$. The efficient information for $\theta$ with complete data is

$$\hat{I}_{\theta, A_0} = E[(\ell_{\theta, A_0}^{(x)})^2] = E[R(Y, X) \left\{ X - \frac{E[R(Y, X)|Y]}{E[R(Y, X)|Y]} \right\}]$$

where $R(Y, X) = \Lambda_0^2(Y|X) O(Y|X)$. See [15] for further details.

We impose the same assumptions made for complete data in [15].

**Condition 4.4.** The finite-dimensional parameter space $\Theta$ is compact and contains the true parameter $\theta_0$ as its interior point.

**Condition 4.5.** (a) The covariate $X$ has bounded support; that is, there exists $x_0$ such that $|X| \leq x_0$ with probability 1. (b) For any $\theta \neq \theta_0$, the probability $P(\theta^T X \neq \theta_0^T X) > 0$.

**Condition 4.6.** $F_0(0) = 0$. Let $\tau_{F_0} = \inf\{t : F_0(t) = 1\}$. The support of $Y$ is an interval $S[Y] = [l_Y, u_Y]$, and $0 < l_Y \leq u_Y < \tau_{F_0}$.

**Condition 4.7.** The cumulative hazard function $\Lambda_0$ has strictly positive derivative on $S[Y]$, and the joint function $G(y, x)$ of $(Y, X)$ has bounded second order (partial) derivative with respect to $y$.

### 4.2.1. Characterization of the WLE

We characterize the WLE’s before studying their asymptotic properties. Let $n = \sum_{i=1}^N \xi_i$ be the number of observations sampled at phase II. Let $Y_1, \ldots, Y_n$ be the order statistics of $Y_1, \ldots, Y_N$ with $\xi_i = 1, i = 1, \ldots, N$. Let $\Delta(i), X(i), U(i)$, and $\xi(i)$ correspond to $Y(i)$; for example, if $Y(i) = Y_j$, then $\Delta(i) = \Delta_j$. Let $\pi(i) = \pi_0(V(i))$. Because only fully observed subjects contribute to the weighted likelihood, $A_N(Y_i)$ for subjects with $\xi_i = 0$ does not matter in the maximization.
In fact, $\hat{\Lambda}_N(Y(i)) = \hat{\Lambda}_N(Y(i-1))$ for subjects with $\xi_{(i)} = 0$ for $i \geq 2$. The WLE $\hat{\Lambda}_N$ of $\Lambda$ corresponds to $z = (\Lambda(1), \ldots, \Lambda(N))$ that maximizes
\[
\phi(\theta, z) = \sum_{i=1}^{n} \frac{1}{\pi(i)} \left[ \log \left( 1 - \exp \left( -e^{\theta^T X(i)} \right) \right) x_i - (1 - \Delta_{(i)}) e^{\theta^T X(i)} x_i \right]
\]
at $\hat{\theta}_N$ subject to $0 \leq x_1 \leq \cdots \leq x_n$. The monotonicity constraint on $x$ is imposed to guarantee that an estimate of $\Lambda$ is nondecreasing. Note that $\phi(\hat{\theta}, z)$ is concave in $x$.

Without loss of generality, we can assume that $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$. If $\Delta_{(1)} = 0$ or $\Delta_{(n)} = 1$, then $\hat{\Lambda}_N(Y(1)) = 0$ or $\hat{\Lambda}_N(Y(n)) = \infty$, so that the first or the last summmand in $\phi$ is zero. Hence ignoring these terms does not change the maximization of the weighted likelihood.

**Proposition 4.1.** Assume that $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$. Then the WLE $(\hat{\theta}_N, \hat{\Lambda}_N)$ satisfies
\[
P_N \hat{\Lambda}_N(Y) \exp(\hat{\theta}_N^T X) X Q(Y, \Delta, X, \hat{\theta}_N, \hat{\Lambda}_N(Y)) = 0,
\]
\[
\sum_{j \geq 1} \frac{\xi_{(j)}}{\pi_{(j)}} Q(Y_{(j)}, \Delta_{(j)}, X_{(j)}; \hat{\theta}_N, \hat{\Lambda}_N) \exp(\hat{\theta}_N^T X_{(j)}) \leq 0, \text{ for } i = 1, \ldots, n,
\]
\[
P_N Y, \Delta, X; \hat{\theta}_N, \hat{\Lambda}_N) \exp(\hat{\theta}_N^T X) \hat{\Lambda}_N(Y) = 0.
\]
Moreover, the corresponding (in)equalities holds for the WLE’s with estimated weights and (modified and centered) calibration.

**Proof.** The first equation is simply the weighted score equation for $\theta$.

For the second inequality, let $1_j$ be the vector which has 1's as its last $j$ components and zeros as its first $n - j$ components. Let $\hat{\Lambda}_N = (\hat{\Lambda}_N(Y(i)))_{i=1}^{n}$. For $\epsilon > 0$, the vector $\hat{\Lambda}_N + \epsilon 1_j$ satisfies the monotonicity constraint. It follows by the definition of the WLE that
\[
0 \geq \lim_{\epsilon \downarrow 0} \frac{\phi(\hat{\theta}_N, \hat{\Lambda}_N + \epsilon 1_j) - \phi(\hat{\theta}_N, \hat{\Lambda}_N)}{\epsilon}
\]
\[
= \sum_{i=1}^{n} \frac{1}{\pi(i)} \left[ \frac{\Delta_{(i)} e^{-e^{\hat{\theta}_N^T X(i)} \hat{\Lambda}_N(Y(i))} e^{\theta^T X(i)} - (1 - \Delta_{(i)}) e^{\theta^T X(i)}}{1 - e^{-e^{\theta^T X(i)} \hat{\Lambda}_N(Y(i))}} \right] I(i \geq j).
\]
Relabeling $i$ and $j$ gives the desired result. Note that the assumption that $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$ guarantees that the above derivative is finite.

The last equality follows for the same reason that
\[
\lim_{h \rightarrow 0} \frac{\phi(\hat{\theta}_N, \hat{\Lambda}_N + h \hat{\Lambda}_N) - \phi(\hat{\theta}_N, \hat{\Lambda}_N)}{h} = 0.
\]
Note that adding terms associated with $\xi_i = 0$ does not contribute to the sum in the above derivative.

For the other four estimators, change weights appropriately. \qed
4.2.2. Consistency

We prove consistency of the WLE’s in the metric given by
\[ d((\theta_1, \Lambda_1), (\theta_2, \Lambda_2)) \equiv \|\theta_1 - \theta_2\| + \|\Lambda_1 - \Lambda_2\|_{P_Y, 2}, \]
where \(\|\cdot\|\) for \(\theta\) is the Euclidean distance,
\[ \|\Lambda_1 - \Lambda_2\|_{P_Y, 2} = \int (\Lambda_1(y) - \Lambda_2(y))^2 dP_Y, \]
and \(P_Y\) is the marginal probability measure of the censoring variable \(Y\). The idea of our proof is first to show the consistency in the Kullback-Leibler divergence. To this end, we use the Glivenko-Cantelli theorem for the IPW empirical processes (Theorem 5.1) in Section 5. Then noting that the Kullback-Leibler divergence bounds the Hellinger distance, we apply the inequality of Lemma A5 of [23] which bounds the metric \(d\) by the Hellinger distance.

**Theorem 4.3** (Consistency). **Under Conditions 3.1, 3.2, 4.4-4.7, the WLE’s are consistent in the metric \(d\).**

**Proof.** We only prove consistency for the WLE. Proofs for the other four estimators are similar. Instead of directly working on \(H\), let \(\hat{H}\) be the set of all subdistribution functions defined on \([0, \infty)\). We denote the WLE of \(F\) as \(\hat{F}_N = 1 - \exp\left(-\hat{\Lambda}_N \right)\).

Define the set \(\mathcal{F}\) of functions by
\[ \mathcal{F} \equiv \{ f(\theta, F) = \delta(1 - \overline{F}(y)^{\exp(\theta^T x)}) + (1 - \delta)\overline{F}(y)^{\exp(\theta^T x)} : \theta \in \Theta, F \in \hat{H} \}. \]

Boundedness of \(X\) and compactness of \(\Theta \subset \mathbb{R}^p\) imply that the set \{\(\exp(\theta^T x) : \theta \in \Theta\}\} is Glivenko-Cantelli. The set \(\hat{H}\) is also Glivenko-Cantelli since it is a subset of the set of bounded monotone functions. Thus, it follows from boundedness of functions in \(\mathcal{F}\) and the Glivenko-Cantelli preservation theorem ([40]) that \(\mathcal{F}\) is Glivenko-Cantelli.

Let \(0 < \alpha < 1\) be a fixed constant. It follows by concavity of the function \(u \mapsto \log u\) and Jensen’s inequality that
\[ P_0 \left[ \log \left( 1 + \alpha \left( \frac{f(\theta, F)}{f(\theta_0, F_0)} - 1 \right) \right) \right] \leq \log \left( P_0 \left[ 1 + \alpha \left( \frac{f(\theta, F)}{f(\theta_0, F_0)} - 1 \right) \right] \right) \]
\[ = \log \left( 1 - \alpha + \alpha P_0 \frac{f(\theta, F)}{f(\theta_0, F_0)} \right) \leq 0, \]
where the first equality holds if and only if \(1 + \alpha(f(\theta, F)/f(\theta_0, F_0) - 1)\) is constant on \(S[Y]\), in other words, \((\theta, F) = (\theta_0, F_0)\) on \(S[Y]\) by the identifiability condition 4.5. Note also that by monotonicity of the logarithm
\[ P_0 \left[ \log \left( 1 + \alpha \left( \frac{f(\theta, F)}{f(\theta_0, F_0)} - 1 \right) \right) \right] \geq P_0 \left[ \log \left( 1 + \alpha \left( 0 - 1 \right) \right) \right] = \log(1 - \alpha). \]
Thus, the set
\[ G = \left\{ \log \left( 1 + \alpha \left( \frac{f(\theta, F)}{f(\theta_0, F_0)} - 1 \right) \right) : f(\theta, F) \in \mathcal{F} \right\} \]
has an integrable envelope. To see this, form a sequence \((\theta_n, F_n)\) such that
\[ g_n = \log \left( 1 + \alpha \left( \frac{f(\theta_n, F_n)}{f(\theta_0, F_0)} - 1 \right) \right) \]
Then \(\{g_n - \log(1 - \alpha)\}_{n \in \mathbb{N}}\) is a monotone increasing sequence of nonnegative functions. By the monotone convergence theorem,
\[ P_0 g_n - \log(1 - \alpha) \to P_0 G - \log(1 - \alpha) \leq -\log(1 - \alpha). \]
Moreover, the set \(G\) is Glivenko-Cantelli by a Glivenko-Cantelli preservation theorem [40].

Now, by the concavity of the map \(u \mapsto \log u\), and the definition of the WLE, we have
\[ P_0^\pi N \log \left( 1 + \alpha \left( \frac{f(\theta_n, F_n)}{f(\theta_0, F_0)} - 1 \right) \right) \]
\[ \geq P_0^\pi N \left( (1 - \alpha) \log(1) + \alpha \log \frac{f(\hat{\theta}_N, \hat{F}_N)}{f(\hat{\theta}_0, \hat{F}_0)} \right) \]
\[ = \alpha \left\{ P_0^\pi N \log f(\hat{\theta}_N, \hat{F}_N) - P_0^\pi N \log f(\theta_0, F_0) \right\} \geq 0. \]
Since \(\Theta\) and \(\hat{H}\) are compact, there exists a subsequence of \((\hat{\theta}_N, \hat{F}_N)\) that converges to \((\theta_\infty, F_\infty) \in \Theta \times \hat{H}\). Along this subsequence it follows by Theorem 5.1 that
\[ 0 \leq P_0^\pi N \log \left( 1 + \alpha \left( \frac{f(\hat{\theta}_N, \hat{F}_N)}{f(\hat{\theta}_0, \hat{F}_0)} - 1 \right) \right) \]
\[ \to P_\star P_{0_0, F_0} \left[ \log \left( 1 + \alpha \left( \frac{f(\theta_\infty, F_\infty)}{f(\theta_0, F_0)} - 1 \right) \right) \right] \leq 0. \]
Thus, we have
\[ P_{0_0, F_0} \log \left( 1 + \alpha \left( \frac{f(\theta_\infty, F_\infty)}{f(\theta_0, F_0)} - 1 \right) \right) = 0. \]
This is possible only at \((\theta_\infty, F_\infty) = (\theta_0, F_0)\) because \((\theta, F) \mapsto P[\log \left( 1 + \alpha(f(\theta, F)/f(\theta_0, F_0) - 1) \right)]\) attains its maximum only at \((\theta_0, F_0)\). Hence conclude that \((\hat{\theta}_N, \hat{F}_N)\) converges to \((\theta_0, F_0)\) in the sense of Kullback-Leibler divergence. Since the Kullback-Leibler divergence bounds the Hellinger distance, it follows by Lemma A5 of [23] that \(d \left( (\hat{\theta}_N, \hat{A}_N), (\theta_0, A_0) \right) = o_P(1). \)

\[ \square \]
4.2.3. Rate of convergence

We prove the rate of convergence of the WLE is $N^{1/3}$. We apply the rate theorem (Theorem 5.2) in Section 5. Since we proved the consistency of $(\hat{\theta}_N, \hat{\Lambda}_N)$ to $(\theta_0, \Lambda_0)$ on $S[Y]$, under the Condition 4.6 we can restrict a parameter space of $\Lambda$ to

$$H_M \equiv \{ \Lambda \in H : M^{-1} \leq \Lambda \leq M, \text{ on } S[Y] \},$$

where $M$ is a positive constant such that $M^{-1} \leq \Lambda_0 \leq M$ on $S[Y]$. Define $\mathcal{M} \equiv \{ \ell(\theta, \Lambda) : \theta \in \Theta, \Lambda \in H_M \}$.

**Theorem 4.4** (Rate of convergence). *Under Conditions 4.4-4.7,*

$$d \left( (\hat{\theta}_N, \hat{\Lambda}_N), (\theta_0, \Lambda_0) \right) = O_{P^*} \left( N^{-1/3} \right).$$

*This holds if we replace the WLE by the WLE’s with estimated weights and (modified and centered) calibration assuming Conditions 3.1 and 3.2.*

**Proof.** Since the rate of convergence for the WLE is easier to verify than the other four estimators, we only prove the theorem for the WLE with modified calibration. The cases for the WLE’s with estimated weights and (centered) calibration are similar.

We proceed by verifying the conditions in Theorem 5.2. The bound (5.29) follows by Lemma 5.2 in Section 5 and Lemma A5 of [23].

For the bound (5.30), we follow the proof of (5.28) in [15]. Since $\hat{\alpha}_N$ is consistent, we can specify the small neighborhood $A_{mc,0}$ of a zero vector such that $G_{mc}(z; \alpha)$ is contained in a small interval that contains 1 and consists of strictly positive numbers. Thus, multiplying the log likelihood by a uniformly bounded quantity, $G_{mc}(z; \alpha)$ only require a slight modification of Huang’s proof of his Lemma 3.1 to obtain

$$\sup_Q \log N \left\{ \epsilon, \mathcal{G}_M, L_2(Q) \right\} \lesssim \epsilon^{-1},$$

for $\epsilon$ small enough where the supremum is taken over the all discrete probability measures, and $\mathcal{G}_M = \{ G_{mc}(z; \alpha) \ell(\theta, \Lambda) : \alpha \in A_{mc,0}, \ell(\theta, \Lambda) \in \mathcal{M} \}$. Thus, it follows by Lemma 3.2.2 of [42] that

$$E^* \| G_N \|_{GM_\delta} \lesssim \delta^{1/2} \left( 1 + \frac{\delta^{1/2}}{\delta^2 \sqrt{N} M} \right) \equiv \phi_N(\delta),$$

where the set $GM_\delta$ is

$$\{ m(\theta, \Lambda, \alpha) - m(\theta_0, \Lambda_0, \alpha) : m(\theta, \Lambda, \alpha) \in GM, d((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \delta \}.$$ 

Apply Theorem 5.2 to conclude $r_N = N^{1/3}$. 

□
4.2.4. Asymptotic normality of the estimators

We apply Theorem 3.2 to derive the asymptotic distributions of the WLE’s.

**Theorem 4.5 (Asymptotic normality).** Under Conditions 3.1, 3.2, 4.4-4.7,

\[
\sqrt{N}(\hat{\theta}_N - \theta_0) = \sqrt{N}P_N^\pi \ell_{\theta_0,\Lambda_0} + o_P(1) \rightsquigarrow N(0, \Sigma),
\]

\[
\sqrt{N}(\hat{\theta}_{N,e} - \theta_0) = \sqrt{N}P_N^{\pi,e} \ell_{\theta_0,\Lambda_0} + o_P(1) \rightsquigarrow N(0, \Sigma_e),
\]

\[
\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) = \sqrt{N}P_N^{\pi,mc} \ell_{\theta_0,\Lambda_0} + o_P(1) \rightsquigarrow N(0, \Sigma_{mc}),
\]

\[
\sqrt{N}(\hat{\theta}_{N,cc} - \theta_0) = \sqrt{N}P_N^{\pi,cc} \ell_{\theta_0,\Lambda_0} + o_P(1) \rightsquigarrow N(0, \Sigma_{cc}),
\]

where \( \ell_{\theta_0,\Lambda_0} \) is the efficient influence function with complete data and \( \Sigma, \Sigma_e, \Sigma_{mc}, \Sigma_{cc} \) are given in Theorem 3.2.

**Proof.** We give a proof for the WLE with modified calibration by verifying the conditions of Theorem 3.2. The cases for the other four estimators are similar.

Condition 3.6 is satisfied with \( \beta = 1/3 \) by Theorems 4.3 and 4.4. Conditions 3.7-3.9 are verified by [15] with

\[
\mathcal{h}^*(y) = \Lambda_0(y)E[\theta_0 \theta_0 T X O(Y|X)|Y = y]/E[\exp(2\theta_0^T X O(Y|X)|Y = y)].
\]

Since \( P_N^{\pi,mc} \ell_{\theta_{N,mc},\hat{\Lambda}_{N,mc}} = 0 \) by Proposition 4.1, it remains to show that \( P_N^{\pi,mc} B_{\theta_{N,mc},\hat{\Lambda}_{N,mc},\Lambda_0} \mathcal{h}^* = 0 \). Let \( g_0 \equiv \mathcal{h}^* \circ \Lambda_0^{-1} \) be the composition of \( \mathcal{h}^* \) and the inverse of \( \Lambda_0 \). Note that \( \Lambda_0 \) is a strictly increasing continuous function by our assumption. Since \( g_0(\hat{\Lambda}_{N,mc}(y)) \) is a right continuous function and has exactly the same jump points as \( \hat{\Lambda}_{N,mc}(y) \), by characterization of \( \Lambda_{N,mc} \) in Proposition 4.1,

\[
P_N^{\pi,mc} g_0 \left( \hat{\Lambda}_{N,mc}(Y) \right) e^{\theta_{N,mc}^T X Q(Y, \Delta, X; \hat{\theta}_{N,mc}, \hat{\Lambda}_{N,mc})} = 0.
\]

By Conditions 4.5-4.7, \( \mathcal{h}^* \) has bounded derivative. This and the assumption that \( \Lambda_0 \) has strictly positive derivative by the Condition 4.7 imply that \( g_0 \) has bounded derivative, too. So, noting that \( \mathcal{h}^* = g_0 \circ \Lambda_0 \), we have

\[
P_N^{\pi,mc} B_{\theta_{N,mc},\hat{\Lambda}_{N,mc},\Lambda_0} \mathcal{h}^* = P_N^{\pi,mc} \ell_{\theta_{N,mc},\hat{\Lambda}_{N,mc}} \mathcal{h}^* + P_{\theta_0,\Lambda_0} \left( g_0 \circ \Lambda_0(Y) - g_0(\hat{\Lambda}_{N,mc}(Y)) \right) e^{\theta_{N,mc}^T X Q(Y, \Delta, X; \hat{\theta}_{N,mc}, \hat{\Lambda}_{N,mc})} + P_{\theta_0,\Lambda_0} \left( g_0 \circ \Lambda_0(Y) - g_0(\hat{\Lambda}_{N,mc}(Y)) \right) e^{\theta_{N,mc}^T X Q(Y, \Delta, X; \hat{\theta}_{N,mc}, \hat{\Lambda}_{N,mc})}.
\]
[15] showed that the second term in the display is $o_{P^*}(N^{-1/2})$. We show that the first term in the display is also $o_{P^*}(N^{-1/2})$. Let $C > 0$ be an arbitrary constant. Define for a fixed constant $\eta > 0$

$$D(\eta) \equiv \{ \psi(y, x; \theta, \Lambda) : d((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \eta, \Lambda \in H_M \},$$

where $\psi(y, \delta, x; \theta, \Lambda) \equiv \{ g_0 \circ \Lambda_0(y) - g_0(\Lambda(y)) \} e^{\theta^T x} Q(y, \delta, x; \theta, \Lambda)$. Because Huang (1996) showed that $D(\eta)$ is Donsker for every $\eta > 0$ and that $\|G_N\|_{D(CN^{-1/3})} = o_{P^*}(1)$, it follows by Lemma 5.4 with $F_N$ replaced by $D(CN^{-1/3})$ that $\|G_{\pi,mc}^N\|_{D(CN^{-1/3})} = o_{P^*}(1)$. This completes the proof. □

Unlike the previous example, $\ell^*_\theta, \Lambda$ depends on additional unknown functions, and the method used in the previous example does not work to estimate asymptotic variances in the present case. See the discussion in Section 6.

5. General results for IPW empirical processes

The IPW empirical measure and IPW empirical process inherit important properties from the empirical measure and empirical process, respectively. We emphasize the similarity between empirical processes and IPW empirical processes.

5.1. Glivenko-Cantelli theorem

The next theorem states that the Glivenko-Cantelli property for complete data is preserved under two-phase sampling.

**Theorem 5.1.** Suppose that $F$ is $P_0$-Glivenko-Cantelli. Then

$$\|P_N^\pi - P_0\|_F \rightarrow P^* 0 \quad (5.26)$$

where $\|\|_F$ is the supremum norm. This also holds if we replace $P_N^\pi$ by $P_N^{\pi,e}$, $P_N^{\pi,c}$, $P_N^{\pi,mc}$ or $P_N^{\pi,cc}$, assuming Conditions 3.1 and 3.2.

5.2. Rate of convergence

The rate of convergence of an $M$-estimator with complete data is often established via maximal inequalities for the empirical processes. If we follow the same line of reasoning, it is natural to derive the maximal inequalities for IPW empirical processes, though this may require some efforts. Fortunately, these maximal inequalities for empirical processes (or slight modifications of them) suffice to establish the same rate of convergence under two-phase sampling.

**Theorem 5.2.** Let $M = \{ m_\theta : \theta \in \Theta \}$ be the set of criterion functions and define $M_\delta = \{ m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta \}$ for some fixed $\delta > 0$ where $d$ is a semimetric on the parameter space $\Theta$.

(1) Suppose that for every $\theta$ in a neighborhood of $\theta_0$,

$$P_0(m_\theta - m_{\theta_0}) \lesssim -d^2(\theta, \theta_0); \quad (5.27)$$
here \( a \lesssim b \) means \( a \leq Kb \) for some constant \( K \in (0, \infty) \). Assume that there exists a function \( \phi_N \) such that \( \delta \mapsto \phi_N(\delta)/\delta^\alpha \) is decreasing for some \( \alpha < 2 \) (not depending on \( N \)) and for every \( N \),

\[
E^*\|G_N\|_{\mathcal{M}_5} \lesssim \phi_N(\delta),
\]

where \( G_N \) is the empirical process. If the WLE \( \tilde{\theta}_N \) satisfying \( P_N^* m_{\tilde{\theta}_N} \geq P_N^* m_{\theta_0} - O_P(r_N^2) \) converges in outer probability to \( \theta_0 \), then \( r_N d(\tilde{\theta}_N, \theta_0) = O_P(1) \) for every sequence \( r_N \) such that \( r_N^2 \phi_N(1/r_N) \leq \sqrt{N} \) for every \( N \).

(2) Suppose the Condition 3.2 holds. Suppose also that for every \( \theta \in \Theta \) in a neighborhood of \( \theta_0 \),

\[
P_0\{G_{mc}(V; \alpha)(m_\theta - m_{\theta_0})\} \lesssim -d^2(\theta, \theta_0) + |\alpha - \alpha_0|^2.
\]

Assume that

\[
E^*\|G_N\|_{G, \mathcal{M}_5} \lesssim \phi_N(\delta),
\]

where \( G, \mathcal{M}_5 \equiv \{G_{mc}(::\alpha); f : |\alpha| \leq \delta, \alpha \in \mathcal{A}_N, f \in \mathcal{M}_5 \} \) for some \( \mathcal{A}_N \subset \mathcal{A}_{mc} \). Then the WLE with modified calibration, \( \tilde{\theta}_{N,mc} \), satisfying \( P_N^{*,mc} m_{\tilde{\theta}_{N,mc}} \geq P_N^{*,mc} m_{\theta_0} - O_P(r_N^{-2}) \) has the same rate of convergence as the WLE in part (1) if it is consistent.

(3) Suppose the Conditions 3.1 and 3.2 hold. Under the same conditions of (2) with \( G_{mc} \) replaced by \( \pi_0/G_{\#} \), for \( \# \in \{e, c, cc\} \) the same conclusions hold for the WLE with estimated weights, \( \hat{\theta}_{N,\#} \), satisfying \( P_N^{\#,\#} m_{\hat{\theta}_{N,\#}} \geq P_N^{\#,\#} m_{\theta_0} - O_P(r_N^{-2}) \) with \( \# \in \{e, c, cc\} \) respectively.

**Remark 5.1.** The key to establishing a general theorem for the rate of convergence is to make use of the boundedness of the weights in the IPW empirical process and also deal with the dependence of the weights. In treating independent bootstrap weights in the weighted bootstrap [21], Lemmas 1-3, require the boundedness of bootstrap weights, because the product of an unbounded weight and a bounded function is no longer bounded. Our theorem exploits the boundedness of sampling indicators in the IPW empirical processes by applying a multiplier inequality for the case of bounded weights (Lemma 5.1) to cover more general cases.

The following is a multiplier inequality for bounded exchangeable weights. Note that the sum of stochastic processes in the second term is divided by \( n^{1/2} \) rather than \( k^{1/2} \).

**Lemma 5.1.** For i.i.d. stochastic processes \( Z_1, \ldots, Z_n \), every bounded, exchangeable random vector \( (\xi_1, \ldots, \xi_n) \) with each \( \xi_i \in [l, u] \) that is independent of \( Z_1, \ldots, Z_n \), and any \( 1 \leq n_0 \leq n \),

\[
E \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i Z_i \right\|^* \leq \frac{2(n_0 - 1)}{n} \sum_{i=1}^{n} E^* \|Z_i\|_{F} E \max_{l \leq i \leq n} \frac{\xi_i}{\sqrt{n}} + 2(u - l) \max_{n_0 \leq k \leq n} E \left\| \frac{1}{\sqrt{n}} \sum_{i=n_0}^{k} Z_i \right\|^*.
\]
The bound (5.30) is not difficult to verify in the presence of the bound (5.28) since $G_{mc}(:, \alpha)$ is a bounded monotone function indexed by a finite dimensional parameter. The bound (5.29) may be verified through the lemma below for some applications including the Cox model with current status data.

**Lemma 5.2.** Suppose Conditions 3.1 and 3.2 hold. Let $m_\theta$ be the log likelihood log $p_\theta$ where $p_\theta$ is the density with dominating measure $\mu$, and $d$ is the Hellinger distance. Then the bound (5.29) and the corresponding bounds for the WLE's with estimated weights and (centered) calibration hold.

### 5.3. Donsker theorem

The next theorem yields weak convergence of the IPW empirical processes under sampling without replacement.

**Theorem 5.3.** Suppose that $F$ with $\|P_0\|_F < \infty$ is $P_0$-Donsker and Conditions 3.1 and 3.2 hold. Then,

\[
G^\pi_N \rightsquigarrow G^\pi \equiv G + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{\frac{1 - p_j}{p_j}} G_j, \quad (5.31)
\]

\[
G^{\pi,c}_N \rightsquigarrow G^{\pi,c} \equiv G + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{\frac{1 - p_j}{p_j}} G_j(\cdot - Q_c), \quad (5.32)
\]

\[
G^{\pi,mc}_N \rightsquigarrow G^{\pi,mc} \equiv G + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{\frac{1 - p_j}{p_j}} G_j(\cdot - Q_{mc}), \quad (5.33)
\]

\[
G^{\pi,cc}_N \rightsquigarrow G^{\pi,cc} \equiv G + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{\frac{1 - p_j}{p_j}} G_j(\cdot - Q_{cc}), \quad (5.34)
\]

in $\ell^\infty(F)$ where the $P_0$-Brownian bridge process, $G$, indexed by $F$ and the $P_0|j$-Brownian bridge processes, $G_j$, indexed by $F$ are all independent.

**Remark 5.2.** The integrability hypothesis $\|P_0\|_F < \infty$ is only required for the IPW empirical processes with estimated weights and (modified and centered) calibration.

If the index set $F$ is Donsker, then it follows by the previous theorem and Lemma 2.3.11 of [42] that asymptotic equicontinuity in probability and in mean follows for the metric that depends on the limit process. In applications, it is of interest to have these results for the original metric $\rho_{P_0}(f, g) = \sigma_{P_0}(f - g)$.

**Theorem 5.4.** Let $F$ be Donsker and define $F_\delta = \{ f - g : f, g \in F, \rho_{P_0}(f, g) < \delta \}$ for some fixed $\delta > 0$. Then, for every sequence $\delta_N \downarrow 0$,

\[
E^* \|G^\pi_N\|_{F_{\delta_N}} \to 0,
\]
and consequently, $\|G_N^\pi\|_{{\mathcal F}_{k_N}} = o_{P^*}(1)$. Moreover, $\|G_N^\pi_\#\|_{{\mathcal F}_{k_N}} = o_{P^*}(1)$ for $\# \in \{e, c, mc, cc\}$ assuming Conditions 3.1 and 3.2.

We end this section with two important lemmas. The first lemma is an extension of Lemma 3.3.5 of [42] and will be used in our proof of Theorem 3.1 to verify asymptotic equicontinuity.

**Lemma 5.3.** Suppose $F = \{\psi_{\theta,h} - \psi_{\theta_0,h} : \|\theta - \theta_0\| < \delta, h \in H\}$ is $P_0$-Donsker for some $\delta > 0$ and that $\sup_{h \in H} P_0(\psi_{\theta,h} - \psi_{\theta_0,h})^2 \to 0$, as $\theta \to \theta_0$. If $\hat{\theta}_N$ converges in outer probability to $\theta_0$, then $\|G_N^\pi(\psi_{\hat{\theta}_N,h} - \psi_{\theta_0,h})\|_H = o_{P^*}(1)$.

This also holds if we replace $G_N^\pi$ by $G_N^{\pi,e}$, $G_N^{\pi,c}$, $G_N^{\pi,mc}$ or $G_N^{\pi,cc}$ assuming Conditions 3.1 and 3.2, hold and $\|P_0\|_F < \infty$.

The second lemma is used to verify asymptotic equicontinuity in the proof of Theorem 3.2, the first part for the IPW empirical process and the second part for the other four IPW empirical processes with estimated weights and (modified and centered) calibration.

**Lemma 5.4.** Let $F_N$ be a sequence of decreasing classes of functions such that $\|G_N\|_{F_N} = o_{P^*}(1)$. Assume that there exists an integrable envelope for $F_{N_0}$ for some $N_0$. Then $E\|G_N\|_{F_N} \to 0$ as $N \to \infty$. As a consequence, $\|G_N^\pi\|_{F_N} = o_{P^*}(1)$.

Suppose, moreover, that $F_N$ is $P_0$-Glivenko-Cantelli with $\|P_0\|_{F_{N_1}} < \infty$ for some $N_1$, and that every $f \in F_N$ converges to zero either pointwise or in $L_1(P_0)$ as $N \to \infty$. Then $\|G_N^\pi,e\|_{F_N} = o_{P^*}(1)$, $\|G_N^\pi,c\|_{F_N} = o_{P^*}(1)$, $\|G_N^{\pi,mc}\|_{F_N} = o_{P^*}(1)$ and $\|G_N^{\pi,cc}\|_{F_N} = o_{P^*}(1)$, assuming Conditions 3.1 and 3.2.

### 6. Discussion

We developed asymptotic theory for weighted likelihood estimation under stratified sampling without replacement. To deal with difficulties due to the dependence of observations, we established general results for the IPW empirical processes. These results and some methods of proof in this paper may be applicable to other estimation procedures. For instance, the weighted Kaplan-Meier estimator can be shown to be asymptotically Gaussian by the functional delta method and the Donsker theorem for the IPW empirical processes [33]. Another application is the weighted estimating equations approach. Beyond these rather obvious applications, it is interesting to study the asymptotic behavior of estimators proposed under independence assumptions that are not involved with inverse probability weighting. In particular, whether or not some known efficient estimators under Bernoulli sampling such as those proposed by [25] is “efficient” under our sampling scheme is an open problem. (See [22] for the definition of efficiency with non i.i.d. data.)
There are several other open problems. [24] developed a method for computing the observed information corresponding to the finite dimensional parameter in the context of the profile likelihood with complete data. This method is particularly useful when the asymptotic variance is not given as a closed formula nor an expectation of a known function. One of examples they considered is in fact the Cox model with interval censoring. Their method is not directly applicable to two-phase sampling, and a general method of variance estimation will be useful. Another open problem is to study other complex survey designs. Stratified sampling without replacement is sufficiently simple for the existing bootstrap empirical process theory to apply. Other complex designs may provide interesting theoretical challenges, perhaps in connection with bootstrap empirical process theory. In Section 4, we followed [41] in imposing somewhat stronger conditions than necessary. Those assumptions allow us to stay within (IPW) empirical process theory, but removing unnecessary conditions raises the question of sorting out connections between empirical process theory and martingale theory as used for example in [1] to study Cox’s partial likelihood estimators with complete data. Empirical process methods will undoubtedly need to be used for further study of the behavior of the WLE methods under model miss-specification.

7. Appendix

We repeatedly use the notation for empirical measures and processes introduced in Section 2 following [6]. The fundamental idea of [6] is to view $G^\xi_{j,N_j}$ as the exchangeably weighted bootstrap empirical process corresponding to $G_{j,N_j} \equiv \sqrt{N_j} (P_{j,N_j} - P_{0,j})$ for $j = 1, \ldots, J$. The processes $G^\xi_{j,N_j}$ converge weakly to $\sqrt{p_j (1 - p_j)} G_j$ for independent $P_{0|j}$-Brownian bridge processes $G_j$, $j = 1, \ldots, J$, in $\ell^\infty(\mathcal{F})$ for Donsker classes $\mathcal{F}$.

Asymptotic linearity and the limiting distributions of $\hat{\alpha}_N$ in binary regression and (modified and centered) calibration are given by the following proposition. The proof requires a Glivenko-Cantelli theorem for $\mathbb{P}^\pi_N$ whose proof is independent of Proposition 7.1.

**Proposition 7.1.** Under the Condition 3.1, $\hat{\alpha}_N$ is consistent for $\alpha_0$, and

$$\sqrt{N}(\hat{\alpha}_N - \alpha_0) = S_0^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^{N} \frac{\dot{G}_i(Z_i^T \alpha_0) Z_i}{\pi_0(V_i)(1 - \pi_0(V_i))} (\xi_i - \pi_0(V_i)) + o^*_p(1)$$

$$\sim S_0^{-1} \sum_{j=1}^{J} \sqrt{\frac{\nu_j}{p_j (1 - p_j)}} G_j \dot{G}_e(Z_i^T \alpha_0) Z_i,$$

where $G_j$ are independent $P_{0|j}$-Brownian bridge processes.
Under the Condition 3.2, both \( \hat{\alpha}_N^c \), \( \hat{\alpha}_N^{mc} \) and \( \hat{\alpha}_N^{cc} \) are consistent, and

\[
\sqrt{N}(\hat{\alpha}_N^c - \alpha_0) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{G}(0)^{-1} \left\{ P_0 Z \otimes \mathbf{1} \right\}^{-1} Z_i \left( \frac{\xi_i - \pi_0(V_i)}{\pi_0(V_i)} \right) + o_p(1)
\]

\[
\sim -\hat{G}(0)^{-1} \left\{ P_0 Z \otimes \mathbf{1} \right\}^{-1} \sum_{j=1}^{J} \sqrt{p_j} \frac{1 - p_j}{p_j} G_j Z,
\]

\[
\sqrt{N}(\hat{\alpha}_N^{mc} - \alpha_0) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{G}(0)^{-1} \left\{ P_0 \frac{1 - \pi_0(V)}{\pi_0(V)} Z \otimes \mathbf{1} \right\}^{-1} Z_i \left( \frac{\xi_i - \pi_0(V_i)}{\pi_0(V_i)} \right) + o_p(1)
\]

\[
\sim -\hat{G}(0)^{-1} \left\{ P_0 \frac{1 - \pi_0(V)}{\pi_0(V)} Z \otimes \mathbf{1} \right\}^{-1} \sum_{j=1}^{J} \sqrt{p_j} \frac{1 - p_j}{p_j} G_j Z,
\]

and

\[
\sqrt{N}(\hat{\alpha}_N^{cc} - \alpha_0) = -\hat{G}(0)^{-1} \left\{ P_0 \frac{1 - \pi_0(V)}{\pi_0(V)} (Z - \mu Z) \otimes \mathbf{1} \right\}^{-1}
\]

\[
\times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (Z_i - \mu Z) \left( \frac{\xi_i - \pi_0(V_i)}{\pi_0(V_i)} \right) + o_p(1)
\]

\[
\sim -\hat{G}(0)^{-1} \left\{ P_0 \frac{1 - \pi_0(V)}{\pi_0(V)} (Z - \mu Z) \otimes \mathbf{1} \right\}^{-1} \sum_{j=1}^{J} \sqrt{p_j} \frac{1 - p_j}{p_j} G_j (Z - \mu Z),
\]

where the \( P_{ij} \)-Brownian bridge processes, \( G_j \), are independent.

**Proof.** We first consider estimated weights. Define \( M_N(\alpha) \equiv \mathbb{P}_N m_\alpha \) and \( M(\alpha) = \mathbb{P}_0 m_\alpha \), where \( m_\alpha(Z, \xi) = \log \left( \{ p_\alpha(\xi | Z) + p_\alpha(\xi | Z) \} / 2 \right) \). We again apply Theorem 5.7 of [41] for a consistency proof. Because \( p_\alpha(\xi | Z) \) is a valid marginal density of a single observation \( \xi \) given \( Z \), the argument of [41], page 66, can be used to verify the second condition of the theorem. We verify the first condition of Theorem 5.7 of [41]. Let \( \hat{G}_e(z; \alpha) \equiv \{ G_e(z^T \alpha) + G_e(z^T \alpha) \} / 2 \). Then \( m_\alpha(z, \xi) = \xi \log \hat{G}_e(z; \alpha) + (1 - \xi) \log (1 - \hat{G}_e(z; \alpha)) \). We rewrite \( \mathbb{P}_N m_\alpha \) as

\[
\mathbb{P}_N m_\alpha = \frac{1}{N} \sum_{i=1}^{N} \xi_i \log \hat{G}_e(Z_i; \alpha) + (1 - \xi_i) \log \left( 1 - \hat{G}_e(Z_i; \alpha) \right)
\]

\[
= \sum_{j=1}^{J} \left\{ \frac{N_j}{N} \sum_{i=1}^{N_j} \frac{1}{n_j} \frac{1}{n_j} \log \hat{G}_e(Z_i i; \alpha) \right\}
\]

\[
+ \sum_{j=1}^{J} \left\{ \frac{N_j}{N} \left( 1 - \frac{n_j}{N_j} \right) \sum_{i=1}^{N_j} \frac{1}{1 - n_j} \log \left( 1 - \hat{G}_e(Z_i i; \alpha) \right) \right\}.
\]

imart-generic ver. 2011/05/20 file: wlecozv-arxiv3.tex date: January 19, 2012
Thus, if we establish that both $S_{0,j} \equiv \left\{ \log \left( 1 - \hat{G}_e(z^T \alpha) \right) : \alpha \in \mathbb{R}^{J+k}, V \in \mathcal{V}_j \right\}$ and $S_{1,j} \equiv \left\{ \log \hat{G}_e(z^T \alpha) : \alpha \in \mathbb{R}^{J+k}, V \in \mathcal{V}_j \right\}$ are $P_0$-Glivenko-Cantelli for $j = 1, \ldots, J$, it follows from Theorem 5.1 applied to sampled subjects and nonsampled subjects in each stratum separately that $\mathbb{P}_N m_\alpha$ converges in probability to

$$P_0m_\alpha = \sum_{j=1}^{J} \nu_j p_j P_0 \left( \log \hat{G}_e(Z^T \alpha) \bigg| V \in \mathcal{V}_j \right) + \sum_{j=1}^{J} \nu_j (1 - p_j) P_0 \left( \log \left( 1 - \hat{G}_e(Z^T \alpha) \right) \bigg| V \in \mathcal{V}_j \right),$$

uniformly in $\alpha$. Note that the method of estimated weights does not estimate the sampling probability for the subjects in a stratum if the sampling probability is 1. Thus, we can assume that $G_e(z^T \alpha_0) \leq \sigma' < 1$. Hence we have $\log(\sigma/2) \leq \log \hat{G}_e(z^T \alpha) \leq 0$ and $\log(1 - \sigma')/2 \leq \log \left( 1 - \hat{G}_e(Z^T \alpha) \right) \leq 0$ for all $j = 1, \ldots, J$ and $\alpha \in \mathbb{R}^{J+k}$. This implies that all sets $S_{k,j}, k = 1, 0, 1$, have integrable envelopes. Now it suffices to show that all sets are VC subgraph classes. Note first that $\{ z^T \alpha : \alpha \in \mathbb{R}^{J+k} \}$ is a VC subgraph class by Lemma 2.6.15 of [42]. Note also that $G_e$ and the logarithm are monotone functions. Because a map by a monotone function, addition and multiplication all preserve the property of $\hat{\alpha}_N$, it follows from Theorem 2.6.17 of [42], our claim follows and hence the first condition is verified. Since we have by concavity of the logarithm and the property of $\hat{\alpha}_N$ that

$$M_N(\hat{\alpha}_N) \geq \frac{1}{2} \mathbb{P}_N \log p_{\hat{\alpha}_N}(\xi|V) + \frac{1}{2} \mathbb{P}_N \log p_{\alpha_0}(\xi|V)$$

$$\geq \frac{1}{2} \mathbb{P}_N \log p_{\alpha_0}(\xi|V) + \frac{1}{2} \mathbb{P}_N \log p_{\alpha_0}(\xi|V) = M_N(\alpha_0),$$

consistency follows from Theorem 5.7 of [41].

We apply Theorem 3.3.1 of [42] to show asymptotic normality of $\hat{\alpha}_N$. Define

$$\Phi_{N,e}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{G}_e(z^T \alpha)Z_i}{G_e(Z^T \alpha) (1 - G_e(Z^T \alpha))} \left( \xi_i - G_e(Z^T \alpha) \right) \equiv \mathbb{P}_N \phi_{\alpha}(\xi, V),$$

and

$$\Phi_e(\alpha) = P_0 \left\{ \frac{\hat{G}_e(z^T \alpha)Z}{G_e(z^T \alpha) (1 - G_e(z^T \alpha))} \left( \sum_{j=1}^{J} p_j I(V \in \mathcal{V}_j) - G_e(z^T \alpha) \right) \right\}.$$

Note that $\Phi_{N,e}(\hat{\alpha}_N) = 0$ because $(\partial/\partial \alpha) \mathbb{P}_N \log p_{\alpha} = \Phi_{N,e}(\alpha)$. Note also that $\Phi_e(\alpha_0) = 0$ since $G_e(z^T \alpha_0) = p_j$ when $V \in \mathcal{V}_j$. It follows by the decomposition
(10) of the inverse probability weighted empirical processes in [6] that

\[
\sqrt{N} \left( \Phi_{N,e}(\alpha_0) - \Phi_e(\alpha_0) \right) = \sqrt{N} \left( \Phi_{N,e}(\alpha_0) - \Phi_e(\alpha_0) \right)
\]

\[
= \sqrt{N} \sum_{j=1}^{J} \frac{1}{J} \frac{N_j}{N} \frac{\pi_0(V)}{G_0(Z \alpha_0)} \frac{G_j(Z \alpha_0)}{1 - G_j(Z \alpha_0)} \left( \hat{G}_j(Z \alpha_0) - \frac{\hat{G}_j(Z \alpha_0)}{G_j(Z \alpha_0)} \right)
\]

\[
+ \sqrt{N} \sum_{j=1}^{J} \frac{1}{J} \frac{N_j}{N} \frac{\pi_0(V)}{G_0(Z \alpha_0)} \frac{G_j(Z \alpha_0)}{1 - G_j(Z \alpha_0)} \left( \hat{G}_j(Z \alpha_0) - \frac{\hat{G}_j(Z \alpha_0)}{G_j(Z \alpha_0)} \right).
\]

Since \( \pi_0(V) = n_j / N_j \) and \( G_j(Z \alpha_0) = p_j \) when \( V \in \mathcal{V}_j \), the first term converges to

\[
\sum_{j=1}^{J} \sqrt{N_j} \left( \frac{n_j}{N_j} - p_j \right) \sqrt{\frac{N_j}{N} \frac{1}{p_j(1-p_j)} \frac{1}{N} \sum_{i=1}^{N_j} \hat{G}_j(Z_{j,i} \alpha_0) Z_{j,i}.
\]

The second term can be written as

\[
\sum_{j=1}^{J} \sqrt{N_j} \left( \frac{n_j}{N_j} - p_j \right) \sqrt{\frac{N_j}{N} \frac{1}{p_j(1-p_j)} \frac{1}{N} \sum_{i=1}^{N_j} \hat{G}_j(Z_{j,i} \alpha_0) Z_{j,i}.
\]

Since \( n_j = [N_j p_j] \) by assumption, it is easy to see that \(-N_j^{-1/2} \leq \sqrt{N_j} (n_j / N_j - p_j) \leq 0\), and hence \( \sqrt{N_j} (n_j / N_j - p_j) \to 0 \). Since \( N_j^{-1} \sum_{i=1}^{N_j} \hat{G}_j(z_{j,i} \alpha_0) = O_{P}(1) \) by the weak law of large numbers and \( \sqrt{N_j / N} \to \sqrt{p_j} \), the second term converges to zero in probability. The weak convergence of \( \sqrt{N} \left( \Phi_{N,e} - \Phi_e(\alpha_0) \right) \) follows from Slutsky’s theorem.

For asymptotic equiconvergence of the process, it suffices to consider a compact subset \( \mathcal{A}_{e,0} \subset \mathbb{R}^{J+k} \) where \( \alpha_0 \) is its interior point since \( \hat{\alpha}_N \) is consistent. Let

\[
\phi_{\alpha,1}(v) = \frac{\pi_0(v) z \psi}{G_e(z \alpha_0) \{1 - G_e(z \alpha_0)\}} \left( \hat{G}_e(z \alpha_0) - \frac{\hat{G}_e(z \alpha_0)}{G_e(z \alpha_0)} \right)^2,
\]

\[
\phi_{\alpha,2}(v) = \frac{z \psi}{1 - G_e(z \alpha_0)} \left( \hat{G}_e(z \alpha_0) - \frac{\hat{G}_e(z \alpha_0)}{1 - G_e(z \alpha_0)} \right)^2.
\]

Taylor’s theorem gives

\[
\phi_{\alpha}(\xi, v) - \phi_{\alpha,0}(\xi, v) = \phi_{\alpha,1}(v) (\xi - \alpha_0) + \phi_{\alpha,2}(v)(\xi - \alpha_0).
\]
where \( \alpha_j^*, j = 1, 2 \), are some convex combinations of \( \alpha \) and \( \alpha_0 \). Thus,

\[
\sqrt{N}(\Phi_{N,e} - \Phi_e)(\alpha) - \sqrt{N}(\Phi_{N,e} - \Phi_e)(\alpha_0) \\
= \sqrt{N}(P_N - P_0)(\phi_\alpha - \phi_{\alpha_0}) + \sqrt{N}P_0(\phi_\alpha - \phi_{\alpha_0}) \\
- \sqrt{N}\Phi_e(\alpha) + \sqrt{N}\Phi_e(\alpha_0) \\
= (\mathbb{P}_N - P_0)\phi_{\alpha_1^*} \sqrt{N}(\alpha - \alpha_0) + (\mathbb{P}_N - P_0)\phi_{\alpha_2^*} \sqrt{N}(\alpha - \alpha_0) \\
+ P_0\phi_{\alpha_1^*}(\xi - \sum_{j=1}^J p_j I(V \in V_j)) \sqrt{N}(\alpha - \alpha_0). \\
\tag{7.36}
\]

To show this is \( o_p(1) + \sqrt{N}(\alpha - \alpha_0) \), we first show that the set \( T_k = \{ \phi_{\alpha,k} : \alpha \in A_{e,0}, k = 1, 2 \} \) are Glivenko-Cantelli. It is easy to see that \( \{ z^T \alpha : \alpha \in A_{e,0} \} \) is Glivenko-Cantelli. Since \( G_e \in C^2 \) by assumption, \( \phi_{\alpha,k}, k = 1, 2 \), are uniformly bounded in \( \alpha \in A_{e,0} \). Thus, the sets \( T_k, k = 1, 2 \) are both Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem [40]. For the third term in (7.36), apply the dominated convergence theorem with \( P_0(\xi | V) = \sum_{j=1}^J (n_j/N_j)I(V \in V_j) \to \sum_{j=1}^J p_j I(V \in V_j) \).

Since \( \Phi(\alpha_0) = -S_0 \), apply Theorem 3.3.1 of [42] to obtain

\[
\sqrt{N}(\alpha_N - \alpha_0) \\
= S_0^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N \frac{\hat{G}_e(Z_i^T \alpha_0)Z_i}{\hat{G}_e(Z_i^T \alpha_0)(1 - \hat{G}_e(Z_i^T \alpha_0))} (\xi_i - \hat{G}_e(Z_i^T \alpha_0)) + o_p(1) \\
\sim S_0^{-1} \sum_{j=1}^J \sqrt{\frac{p_j}{p_j(1 - p_j)}} g_j \hat{G}_e(Z^T \alpha_0)Z.
\]

This completes the proof.

Next we consider modified calibration with \( \hat{\alpha}_N = \hat{\alpha}_N^{mc} \). The cases for (centered) calibration (i.e., \( \hat{\alpha}_N = \hat{\alpha}_N^c \) and \( \hat{\alpha}_N = \hat{\alpha}_N^{mc} \)) are similar. Define \( \Phi_{N,mc}(\alpha) \equiv \mathbb{P}_N^* G_{mc}(V; \alpha)Z - \mathbb{P}_N Z \) and \( \Phi_{mc}(\alpha) \equiv P_0([G_{mc}(V; \alpha) - 1]Z) \). Note that \( \Phi_{N,mc}(\hat{\alpha}_N) = 0 \) and \( \Phi_{mc}(0) = 0 \). We apply Theorem 5.7 of [41] for a consistency proof. For the first condition of the theorem, we have

\[
\sup_{\alpha \in \mathbb{R}^k} \| \Phi_{N,mc}(\alpha) - \Phi_{mc}(\alpha) \| \\
= \sup_{\alpha \in \mathbb{R}^k} \left\| \frac{1}{N} \sum_{i=1}^N \left( \frac{\xi_i}{\pi_0(V_i)} G_{mc}(V_i; \alpha) - 1 \right) Z_i - P_0 \{ G_{mc}(V; \alpha) - 1 \} Z \right\| \\
\leq \sup_{\alpha \in \mathbb{R}^k} \left\| \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_0(V_i)} G_{mc}(V_i; \alpha)Z_i - P_0 G_{mc}(\alpha)Z \right\| \\
+ \sup_{\alpha \in \mathbb{R}^k} \left\| \frac{1}{N} \sum_{i=1}^N Z_i - P_0 Z \right\|,
\]

imsart-generic ver. 2011/05/20 file: wlecoyv-arxiv3.tex date: January 19, 2012
where \( \| \cdot \| \) is the Euclidean norm. Since \( \alpha \) is a vector in \( \mathbb{R}^k \) and \( G \) is monotone, \( \{ G_{mc}(\cdot; \alpha) : \alpha \in \mathbb{R}^k \} \) is a VC subgraph by Lemmas 2.6.15 and 2.6.18 of [42]. Boundedness of \( G \) implies that the set \( \{ G_{mc}(v; \alpha)z : \alpha \in \mathbb{R}^k \} \) is \( P_0 \)-Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem [40]. Then the first term is \( o_{P^*}(1) \) by Theorem 5.1. The second term is \( o_{P^*}(1) \) by the weak law of large numbers.

The second condition of the theorem is that for any \( \epsilon > 0, \inf_{|\alpha| > \epsilon} \| \Phi_{mc}(\alpha) \| > 0 \). Suppose, to the contrary, that \( \inf_{|\alpha| > \epsilon} \| \Phi_{mc}(\alpha) \| = 0 \) for some \( \epsilon > 0 \). Then there exists a sequence \( \{ \alpha^{(m)} \} \subset \mathbb{R}^k \) with \( |\alpha^{(m)}| > \epsilon \) for each \( m = 1, 2, \ldots \), such that

\[
\| \Phi_{mc}(\alpha^{(m)}) \| \to 0.
\]

Let \( \Phi_{j,\epsilon}(\alpha), j = 1, \ldots, k, \) be the \( j \)th element of \( \Phi_{mc}(\alpha) \). Since the norm \( \| \cdot \| \) is the Euclidean norm, each element \( \Phi_{j,\epsilon}(\alpha^{(m)}) \) converges to zero. If \( \alpha^{(m)} \) converges to \( \alpha^{(\infty)} \) with \( |\alpha^{(\infty)}| < \infty \), then by the dominated convergence theorem and Taylor’s theorem,

\[
0 = P_0 \left[ \left( G_{mc}(V; \alpha^{(\infty)}) - 1 \right) Z \right] = P_0 \left[ (\pi_0(V)^{-1} - 1) G_{mc}(V; \alpha^*) Z^\otimes 2 \alpha^{(\infty)} \right]
\]

for some \( \alpha^* \) with \( |\alpha^*| \leq |\alpha^{(\infty)}| \). Because \( P_0(\pi_0(V)^{-1} - 1) G_{mc}(V; \alpha^*) Z^\otimes 2 \) is positive definite by assumption, \( \alpha^{(\infty)} \) must be zero, which contradicts the fact that \( |\alpha^{(\infty)}| \geq \epsilon \).

We assume that some elements of \( \alpha^{(m)} \) diverge. Then, a further subsequence \( \alpha^{(m')} \) converges to some \( \alpha^{(\infty)} \) whose elements are extended real numbers. Define a unit vector \( \beta^{(\infty)} \equiv \lim_{m' \to \infty} \alpha^{(m')}/|\alpha^{(m')}| \). Then we have for each \( Z \) on the set \( \{ \pi_0(V) < 1 \} \) that

\[
G_{mc}^{(\infty)}(Z) = \lim_{m' \to \infty} G \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \alpha^{(m')} / \alpha^{(m')} \| \alpha^{(m')} \| \right) Z^T \beta^{(\infty)} = \begin{cases} M_1 Z^T \beta^{(\infty)} & \text{if } Z^T \beta^{(\infty)} > 0 \\ m_1 Z^T \beta^{(\infty)} & \text{if } Z^T \beta^{(\infty)} < 0 \\ 0 & \text{if } Z^T \beta^{(\infty)} = 0 \end{cases}.
\]

It follows by the dominated convergence theorem applied to each element of the vector of \( \Phi_{mc}(\alpha) \) that

\[
0 = \lim_{m' \to \infty} \Phi_{mc} \left( \alpha^{(m')} \right)^T \beta^{(\infty)} = P_0 \lim_{m' \to \infty} \left( G_{mc} \left( V; \alpha^{(m')} \right) - 1 \right) Z^T \beta^{(\infty)} = (M_1 - 1) P_0 I \{ Z^T \beta^{(\infty)} > 0, \pi_0(V) < 1 \} Z^T \beta^{(\infty)} + (m_1 - 1) P_0 I \{ Z^T \beta^{(\infty)} < 0, \pi_0(V) < 1 \} Z^T \beta^{(\infty)}.
\]

However, this is strictly positive since \( m_1 < 1 \) and \( M_1 > 1 \), which is a contradiction. This completes the proof that \( \hat{\alpha}_N \to P^* \) 0.
We apply Theorem 3.3.1 of [42] to show the asymptotic normality of $\hat{\alpha}$. For asymptotic equicontinuity condition, it follows by Taylor’s theorem that
\[
\sqrt{N}(\Phi_{N,mc} - \Phi_{mc})(\hat{\alpha}) - \sqrt{N}(\Phi_{N,mc} - \Phi_{mc})(\alpha_0) = G_N^\alpha[G_{mc}(V; \alpha{}) - G_{mc}(V; \alpha)Z] \\
= (P_N^\alpha - P_0)(\pi_0(V)^{-1} - 1)\hat{G}_{mc}(V; \alpha^*)Z^{\otimes 2}\sqrt{N}(\hat{\alpha} - \alpha_0)
\]
for some $\alpha^*$ with $|\alpha^* - \alpha_0| \leq |\hat{\alpha} - \alpha_0|$. This term is $o_P(1 + \sqrt{N}|\hat{\alpha} - \alpha_0|)$ if $(P_N^\alpha - P_0)(\pi_0(V)^{-1} - 1)Z^{\otimes 2}\hat{G}_{mc}(V; \alpha) \to P$, 0, uniformly in $\alpha$. Let $A_{mc,1} \subset \mathbb{R}^k$ be a compact neighborhood of zero. Since $\hat{\alpha}$ is consistent, it suffices to show that the set $\{(\pi_0^{-1}(V) - 1)Z^{\otimes 2}\hat{G}_{mc}(Z; \alpha) : \alpha \in A_{mc,1}\}$ is Glivenko-Cantelli. Since $|\pi_0^{-1}(V) - 1|$ and $Z$ are bounded, the VC subgraph class $\{(\pi_0^{-1}(V) - 1)\alpha : \alpha \in A_{mc,1}\}$ (Lemma 2.6.15 of [42]) is $P_0$-Glivenko-Cantelli. Because $\hat{G}$ is continuous and bounded, the set $\{\hat{G}_{mc}(Z; \alpha) : \alpha \in A_{mc,1}\}$ is Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem of [40]. Apply the Glivenko-Cantelli preservation theorem of [40] again to conclude $\{(\pi_0^{-1}(V) - 1)Z^{\otimes 2}\hat{G}_{mc}(Z; \alpha) : \alpha \in A_{mc,1}\}$ is Glivenko-Cantelli. Hence, asymptotic equicontinuity follows from Theorem 5.1. We show the weak convergence of the process $\sqrt{N}(\Phi_{N,mc} - \Phi_{mc})(\alpha)$ at $\alpha_0 = 0$. Since $G_{mc}(v; \alpha_0) = 1$, it follows from the decomposition (10) of the inverse probability weighted empirical processes in [6] that
\[
\sqrt{N}(\Phi_{N,mc} - \Phi_{mc})(\alpha_0) = \sqrt{N}\Phi_{N,mc}(0) = \sqrt{N}(P_N^\alpha - P_N)Z \\
= \sum_{j=1}^J \sqrt{\frac{N_j}{N}} \frac{N_j}{n_j} G_j^{\xi} Z \\
\sim \sum_{j=1}^J \sqrt{\frac{1 - p_j}{p_j}} G_j Z \quad (\text{by Theorem 5.3}).
\]
The Fréchet derivative of $\Phi_{mc}(\alpha_0)$ is
\[
\hat{\Phi}_{mc}(\alpha)|_{\alpha = \alpha_0} = \frac{\partial}{\partial \alpha} P_0(G_{mc}(V; \alpha) - 1)Z \bigg|_{\alpha = \alpha_0} = \hat{G}(0)P_0(\pi_0(V)^{-1} - 1)Z^{\otimes 2}.
\]
Thus, by Theorem 3.3.1 of [42] we obtain
\[
\sqrt{N}\hat{\alpha} = -\hat{\Phi}_{mc}(0)\sqrt{N}(\Phi_{N,mc} - \Phi_{mc})(0) + o_{P*}(1) \\
\sim -\hat{G}(0)^{-1} \left\{ P_0(\pi_0(V)^{-1} - 1)Z^{\otimes 2}\right\}^{-1} \sum_{j=1}^J \sqrt{\frac{1 - p_j}{p_j}} G_j Z.
\]

Here we give proofs of the theorems in Section 5.

**Proof of Theorem 5.1.** First consider $P_N^\alpha$. By the decomposition (10) of the inverse probability weighted empirical processes in [6], we have
\[
\|P_N^\alpha - P_0\|_F \leq \|P_N - P_0\|_F + \sum_{j=1}^J \sqrt{\frac{N_j}{N}} \frac{N_j}{n_j} \left\| P_j^{\xi} \frac{n_j}{N_j} P_j^{\eta_j} \right\|_F.
\]
The first term is $o_{P^*}(1)$ since $\mathcal{F}$ is Glivenko-Cantelli. Since $(N_j/N)(N_j/n_j) \to P^* \nu_j/p_j$, each summand in the second term is $o_{P^*}(1)$ by the bootstrap Glivenko-Cantelli theorem, which is an easy corollary to Lemma 3.6.16 of [42].

Consider $\mathbb{P}_N^{\pi,e}$. Because $\hat{\alpha}_N \to P^* \alpha_0$ by Proposition 7.1, it suffices to consider a compact neighborhood $K \subset \mathbb{R}^{J+k}$ of $\alpha_0$. Since $Z$ is bounded and $G_e$ is continuous, $\{\pi_0(V)\}^{-1} = \{G_e(\alpha^T Z)\}^{-1}$ is bounded in this neighborhood. Because $\alpha$ is a vector in $\mathbb{R}^{J+k}$ and $G_e$ is monotone, $\{G_e(\alpha)\}^{-1} : \alpha \in K$ is a VC subgraph class by Lemmas 2.6.15 and 2.6.18 of [42]. Boundedness of $G_e$ implies that the set

$$\{\pi_0\{G_e(\cdot)\}^{-1} f : f \in \mathcal{F}, \alpha \in K\}$$

is $P_0$-Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem [40]. Since $\hat{\alpha}_N \to P^* \alpha_0$, we have by (5.26) that $\|\mathbb{P}_N^{\pi,e} - P_0\|_F \to P^* 0$, by recognizing that

$$\mathbb{P}_N^{\pi,e} = \frac{1}{N} \sum_{i=1}^N \xi_i \pi_0(V_i) \left\{ \frac{\pi_0(V_i)}{G_e(\hat{\alpha}_N^T Z_i) \delta_{X_i}} \right\}.$$

Consider $\mathbb{P}_N^{\pi,mc}$. The cases for $\mathbb{P}_N^{\pi,c}$ and $\mathbb{P}_N^{\pi,cc}$ are similar. We verified in the proof of Proposition 7.1 that $\{G_{mc}(\cdot; \alpha) : \alpha \in \mathbb{R}^k\}$ is a VC subgraph class. Boundedness of $G$ implies that the set

$$\{G_{mc}(\cdot; \alpha) f : f \in \mathcal{F}, \alpha \in \mathbb{R}^k\}$$

is $P_0$-Glivenko-Cantelli by the Glivenko-Cantelli preservation theorem [40]. Since $\hat{\alpha}_N$ converges to zero in probability by Proposition 7.1, the result follows by (5.26).

Several lemmas are required for the proof of Theorem 5.2.

**Lemma 7.1.** Let $\mathcal{F}$ be a class of functions with $P_0|f| < \infty$ for every $f \in \mathcal{F}$. Then,

$$E^* \left\| \sqrt{\frac{N_j}{N}} I(N_j > 0) G_{j,N} \right\|_F \lesssim E^* \| G_N \|_F, \quad \text{for each} \quad j = 1, \ldots, J.$$

**Proof.** Let $\epsilon_i$, $i = 1, \ldots, N$, be independent Rademacher variables, independent of $X_i$, $i = 1, \ldots, N$, and $N_j$. It follows from the symmetrization inequality (Lemma 2.3.6) of [42]

$$E^* \| G_N \|_F \gtrsim E^* \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i f(X_i) \right\|_F.$$

---

*imsart-generic ver. 2011/05/20 file: wlecoxz-arxiv3.tex date: January 19, 2012*
Rewrite this and use Jensen’s inequality again with $E[e f(X)] = 0$ to obtain
\[
E^* \left\| \sum_{j=1}^J I(N_j > 0) \sqrt{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^{N_j} \xi_j i f(X_{j,i}) \right\|_F \\
\geq E^* \left\| I(N_j > 0) \sqrt{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^{N_j} \xi_j i f(X_{j,i}) \right\|_F.
\]

Here we implicitly change the law. This can be justified by Proposition A.1 of [6].

Now applying the Lemma 2.3.6 of [42] to the $j$th stratum, this is further bounded below, up to some constant, by
\[
E^* \left\| I(N_j > 0) \sqrt{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^{N_j} (f(X_{j,i}) - P_{0j} f) \right\|_F \\
= E^* \left\| I(N_j > 0) \sqrt{N_j/N_{j,N_j}} \right\|_F.
\]

The following is a multiplier inequality for bounded exchangeable weights. Note that the sum of stochastic processes in the second term is divided by $n^{1/2}$ rather than $k^{1/2}$.

\textbf{Proof of Lemma 5.1.} This follows the proof of Lemma 3.6.7 of [42] up to the last line. Since the $\xi_i$'s can be split into their positive and negative parts, we only consider the case where they are nonnegative. Thus for any $1 \leq n_0 \leq n$,
\[
E \left\| \sum_{i=1}^n \xi_i Z_i \right\|_F^* \leq E \left\| \sum_{i=1}^{n_0-1} \xi_{(i)} Z_i \right\|_F^* + E \left\| \sum_{i=n_0}^n \xi_{(i)} Z_i \right\|_F^* \\
\leq E \left( \max_{1 \leq i \leq n} \xi_i \right) \frac{n_0 - 1}{n} \sum_{i=1}^n E^* ||Z_i||_F + E \left\| \sum_{i=n_0}^n \xi_{(i)} Z_i \right\|_F^*,
\]

where $\xi_{(i)}$, $i = 1, \ldots, n$, are the reverse order statistics of $\xi_i$, $i = 1, \ldots, n$. To bound the second term, we substitute $\xi_{(i)} = \sum_{k=i}^n (\xi(k) - \xi_{(k+1)})$ with $\xi_{(n+1)} = 0$, and change the order of summation to obtain
\[
E \left\| \sum_{i=n_0}^n \xi_{(i)} Z_i \right\|_F^* = E \left\| \sum_{i=n_0}^n \sum_{k=i}^n (\xi(k) - \xi_{(k+1)}) Z_i \right\|_F^* \\
= E \left\| \sum_{k=n_0}^n (\xi(k) - \xi_{(k+1)}) \sum_{i=n_0}^k Z_i \right\|_F^*.
\]
It follows from the triangle inequality and the independence of the $\xi$'s and $Z_i$'s that this is bounded by

$$
\sum_{k=n_0}^n E^* \left( \left\| \left( \xi_k - \xi_{k+1} \right) \sum_{i=n_0}^k Z_i \right\|_F^* \right) \\
= \sum_{k=n_0}^n E^* \left\{ \left( \xi_k - \xi_{k+1} \right) \left\| \sum_{i=n_0}^k Z_i \right\|_F^* \right\} \\
= \sum_{k=n_0}^n E^* (\xi_k - \xi_{k+1}) E^* \left\| \sum_{i=n_0}^k Z_i \right\|_F^* \\
\leq \sum_{k=n_0}^n E^* (\xi_k - \xi_{k+1}) \max_{n_0 \leq k \leq n} E^* \left\| \sum_{i=n_0}^k Z_i \right\|_F^* \\
= E^* \sum_{k=n_0}^n (\xi_k - \xi_{k+1}) \max_{n_0 \leq k \leq n} E^* \left\| \sum_{i=n_0}^k Z_i \right\|_F^* \\
\leq (u - l) \max_{n_0 \leq k \leq n} E^* \left\| \sum_{i=n_0}^k Z_{R_i} \right\|_F^*
$$

using the boundedness of the $\xi_i$'s in the last line. The proof for the negative parts of the $\xi_i$'s is similar and the inequality follows. \qed

**Lemma 7.2.** For an arbitrary set $\mathcal{F}$ of integrable functions,

$$E^* \| G_N^\tau \|_F \lesssim E^* \| G_N \|_F.$$

**Proof.** We decompose $G_N^\tau$ as in (2.1): thus

$$E^* \| G_N^\tau \|_F = E^* \left\| G_N + \sum_{j=1}^J \sqrt{\frac{N_j}{N}} \left( \frac{N_j}{n_j} \right) G_{j,N_j}^\tau \right\|_F \\
\leq E^* \| G_N \|_F + \sum_{j=1}^J E^* \left\| \sqrt{\frac{N_j}{N}} \left( \frac{N_j}{n_j} \right) G_{j,N_j}^\tau \right\|_F.$$

It therefore suffices to show that each $E^* \| m_j G_{j,N_j} \|_F$ is bounded up to some constant by $E^* \| G_N \|_F$ where $m_j \equiv (N_j/N)^{1/2}(N_j/n_j)$.

Rewrite $G_{j,N_j}^\tau$ as

$$G_{j,N_j}^\tau = \sqrt{N_j} \left( \frac{1}{N_j} \sum_{i=1}^{N_j} \xi_{j,i} \right) \left( \delta_{X_{j,i}} - p_{j,N_j} \right).$$

Now we condition on $N \equiv (N_1, \ldots, N_J)$, and write $E_N$ for $E(\cdot | N)$. Since $\xi_{j,i} \in \{0, 1\}$, it follows by the multiplier inequality of Lemma 5.1 applied conditionally.
with \( n_0 = 1 \) and \( Z_i = m_j(\delta_X, j - \mathbb{P}, j, N_j) \) that \( E_N^* \|m_jG^j_{N_j} \|_F \) is bounded by

\[
(1 - 0) \max_1 \leq k \leq N_j \frac{E_N^*}{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^k m_j(\delta_X, j - \mathbb{P}, j, N_j) \|_F ^* \\
= \max_1 \leq k \leq N_j \frac{E_N^*}{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^k (\delta_X, j - \mathbb{P}, j, N_j) \|_F ^* 
\]

Note that \( N_j/n_j \leq \sigma^{-1} \) for some \( \sigma > 0 \) by assumption so that we can replace \( N_j/n_j \) by \( \sigma^{-1} \) in the last display to obtain an upper bound. Then, apply the triangle inequality to further bound this by

\[
\max_1 \leq k \leq N_j \frac{E_N^*}{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^k (\delta_X, j - P_{0j}) \|_F ^* + \max_1 \leq k \leq N_j \frac{E_N^*}{N_j} \frac{1}{\sqrt{N_j}} \sum_{i=1}^k (P_{0j} - \mathbb{P}, j, P_{0j}) \|_F ^* 
\]

Since \( \delta_X, j - P_{0j} \) has mean zero, it follows by Jensen’s inequality that the first term is bounded by

\[
E_N^* \frac{1}{\sqrt{N_j}} \sum_{i=1}^N (\delta_X, j - P_{0j}) \|_F ^* = E_N^* \frac{1}{\sqrt{N_j}} \sum_{i=1}^N \mathbb{G}_j, N_j \|_F ^* 
\]

The second term is bounded by \( E_N^* \frac{1}{\sqrt{N_j}} \sum_{i=1}^N \mathbb{G}_j, N_j \|_F ^* \). Now compute unconditionally and apply Lemma 7.1 to find that both terms are bounded by \( E_N^* \|G_N \|_F \).

**Proof of Theorem 5.2.** It follows by Lemma 7.2 and the assumption on \( E_N^* \|G_N \|_M \) that

\[
E_N^* \|G_N \|_M \leq E_N^* \|G_N \|_M \leq \phi_N(\delta). 
\]

By application of Theorem 3.2.5 of [42], we conclude that the conclusion of (1) of the theorem holds.

For the second statement, note that Theorem 3.2 of [24] holds in a general setting where \( P_0m_{\theta, \eta} \) and \( P_nm_{\theta, \eta} \) are replaced by the deterministic function \( M(\theta, \eta) \) and the stochastic process \( M_n(\theta, \eta) \), respectively. Our parameters \( \alpha \) and \( \theta \) play roles of their \( \theta \) and \( \eta \), respectively. Our choice of \( M \) and \( M_N \) is \( P_0G_{mc}(V; \alpha)\theta \) and \( P_{N,mc}m_{\theta} \). The condition 5.29 corresponds to (3.5) of [24]. The condition 5.30 together with Lemma 7.2 verifies their (3.6). Apply their Theorem 3.2 to obtain \( d(\hat{\theta}_{N,mc}, \theta_0) \leq O_P^*(\delta_N^{-1} + |\hat{\alpha}_N - \alpha_0|) = O_P^*(\delta_N^{-1}) \). The cases for \( \hat{\theta}_{N,c}, \hat{\theta}_{N,c} \) and \( \hat{\theta}_{N,cc} \) are similar.

**Proof of Lemma 5.2.** We consider modified calibration. Other three cases are similar. Because \( G(0) = 1 \) and \( Z \) is bounded, consistency of \( \hat{\alpha}_N \) implies that there exists \( \mathcal{A}_{mc,2} \subseteq \mathcal{A}_{mc} \) such that for some fixed constant \( C > 0 \), \( G_{mc}(v; \alpha) \geq \)

\[
\]
C and \( \hat{G}_{mc}(v; \alpha) \geq C \) for every \( \alpha \in A_{mc, 2} \) and \( P(\hat{\pi}_N \in A_{mc, 2}) \to 1 \). Then, for arbitrary \( \alpha \in A_{mc, 2} \),

\[
P_0 G_{mc}(V; \alpha)(m_\theta - m_{\theta_0}) = P_0 G_{mc}(V; \alpha) \log \frac{p_\theta}{p_{\theta_0}} \\
\leq 2 P_0 G_{mc}(V; \alpha) \left( \sqrt{\frac{p_\theta}{p_{\theta_0}}} - 1 \right) \\
= \int G_{mc}(v; \alpha) \left\{ -(p_\theta^{1/2} - p_{\theta_0}^{1/2})^2 + p_\theta - p_{\theta_0} \right\} d\mu \\
\leq -C \int (p_\theta^{1/2} - p_{\theta_0}^{1/2})^2 d\mu + \int \left\{ G_{mc}(v; \alpha) - 1 \right\} (p_\theta - p_{\theta_0}) d\mu \\
= -Ch^2(p_\theta, p_{\theta_0}) + \int \hat{G}_{mc}(v; \alpha^*)(\pi_0^{-1}(v) - 1)v^T(p_\theta - p_{\theta_0}) d\mu(\alpha - \alpha_0),
\]

where \( \alpha^* \) is some convex combination of \( \alpha \) and \( \alpha_0 \). Because the integral in the last display is a bounded row vector, the second term in the last display is bounded by \( |\alpha - \alpha_0|^2 \) up to some constant. Thus, the condition (5.29) holds.

The following lemma is useful when showing asymptotic equicontinuity of processes involving \( \mathbb{P}_N^c, \mathbb{P}_N^c, \mathbb{P}_{mc}^c \) and \( \mathbb{P}_{mc}^c \).

**Lemma 7.3.** Suppose Conditions 3.2 and 3.1 hold. Let \( \mathcal{F} \) be a Glivenko-Cantelli class. Then

\[
\sup_{f \in \mathcal{F}} \left| \sqrt{N}(\mathbb{P}_N - P_0) \left\{ \frac{\xi}{\pi_{\hat{\pi}_N}(V)} f - \frac{\xi}{\pi_{\pi_0}(V)} f \right\} \right| = o_P(1), \tag{7.37}
\]

where \( \pi_{\pi_0} \) is either an estimated or calibrated probability (with modified centered calibration).

**Proof.** We only consider modified calibration. The cases for estimated weights and (centered) calibration are similar. It follows by Taylor’s theorem that

\[
\sup_{f \in \mathcal{F}} \left| \sqrt{N}(\mathbb{P}_N - P_0) \left\{ \frac{\xi}{\pi_{\hat{\pi}_N}(V)} f - \frac{\xi}{\pi_{\pi_0}(V)} f \right\} \right| \\
= \sup_{f \in \mathcal{F}} \left| (\mathbb{P}_N - P_0) \left\{ (\pi_0^{-1}(V) - 1)Z^T \hat{G}_{mc}(Z; \alpha^*) f \right\} \right| \sqrt{N}|\hat{\pi}_N - \pi_0|,
\]

for some \( \alpha^* \) with \( |\alpha^* - \alpha_0| \leq |\hat{\pi}_N - \pi_0| \). Because \( \sqrt{N}(\hat{\pi}_N - \pi_0) = O_P(1) \) by Proposition 7.1, it follows that (7.37) is \( o_P(1) \) by Theorem 5.1 and Proposition 7.1 if the set \( \{(\pi_0^{-1}(V) - 1)Z^T \hat{G}(\pi_0^{-1}(V) - 1)Z^T \alpha : \alpha \in A_{mc, 3}, f \in \mathcal{F}\} \) is \( P_0 \)-Glivenko-Cantelli where \( A_{mc, 3} \subset A_{mc} \) is some compact set containing \( \alpha_0 = 0 \). This is easily verified in the same way as in the proof of Proposition 7.1.

**Proof of Theorem 5.3.** The result (5.31) follows from [6]. Consider the IPW em-
pirical process with modified calibration. It follows by Taylor’s theorem that
\[
G_{\pi,mc}N^f - G_{\pi}N^f = G_{\pi}N^f \left( \xi - \frac{\xi}{\pi_0}(V) \right) f + \sqrt{N} P_0 \left( \frac{\xi}{\pi_0}(V) - \frac{\xi}{\pi_0}(V) \right) f
\]
\[
+ P_0 \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \hat{G}_{mc}(V; \alpha^*) f \right) \sqrt{N}(\hat{\alpha}_N - \alpha_0),
\]
where $\alpha^*$ is some convex combination of $\hat{\alpha}_N$ and $\alpha_0$. The first term is $o_p(1)$ by Lemma 7.3. Since $(\pi_0(V)^{-1} - 1) Z^T \hat{G}_{mc}$ is bounded and $f$ is integrable, it follows from the dominated convergence theorem that
\[
P_0 \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \hat{G}_{mc}(V; \alpha^*) f \right) \to P_0 \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \hat{G}(0) f \right).
\]
Apply the result (5.31) and Proposition 7.1 to conclude the finite-dimensional convergence
\[
G_{\pi,mc}N^f = G_{\pi}N^f + (G_{\pi,mc}N^f - G_{\pi}N^f)
\]
\[
\to_d Gf + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{1 - p_j} G_j f
\]
\[
- P_0 \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \hat{G}(0) f \right) \hat{G}(0)^{-1} \left( P_0 \frac{1 - \pi_0(V)}{\pi_0(V)} Z \otimes 2 \right)^{-1}
\]
\[
\times \sum_{j=1}^J \sqrt{\nu_j} \sqrt{1 - p_j} G_j Z
\]
\[
= Gf + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{1 - p_j} G_j f - \sum_{j=1}^J \sqrt{\nu_j} \sqrt{1 - p_j} G_j Q_{mc} f
\]
\[
= Gf + \sum_{j=1}^J \sqrt{\nu_j} \sqrt{1 - p_j} G_j (f - Q_{mc} f).
\]
Next, we prove asymptotic equicontinuity of $G_{\pi,mc}N^f$ with respect to the metric $\rho_{mc}$ defined by
\[
\rho_{mc}^2(f, g) = P_0(f - g)^2 + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0|j}(f - g).
\]
First recall that $G_{\pi}N^f$ is asymptotically equicontinuous with respect to the metric $\rho$ defined by
\[
\rho^2(f, g) = \sigma_0^2(f - g) + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \text{Var}_{0|j}(f - g).
\]
The part \( \sigma_0^2(f - g) \) corresponds to the empirical process \( \mathbb{G}_N \equiv \sqrt{N}(P_N - P_0) \) in the decomposition (2.1) of the inverse probability weighted empirical processes. However, this empirical process \( \mathbb{G}_N \) is asymptotically equicontinuous with respect to the \( L_2(P) \)-metric with an assumption \( \|P_0\|_F < \infty \) in view of Problem 2.1.2 of [42]. Thus, \( \mathbb{G}_N^\pi \) is asymptotically equicontinuous with respect to \( \rho_{mc} \). Now, it remains to verify the asymptotic equicontinuity of \( \mathbb{G}_N^{\pi,mc} - \mathbb{G}_N^\pi \).

Let \( h_N \in F_{\delta_N} \equiv \{ f - g : f, g \in F, \rho_{mc}(f, g) \leq \delta_N \} \) for an arbitrary sequence \( \delta_N \downarrow 0 \). In view of (7.38)

\[
(\mathbb{G}_N^{\pi,mc} - \mathbb{G}_N^\pi)h_N = o_{P^*}(1) + P_0 \left( \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \hat{G}_{mc}(V; \alpha^*)h_N \right) O_{P^*}(1),
\]

where \( \alpha^* \) is some convex combination of \( \hat{\alpha}_N \) and \( \alpha_0 \). Because each element of a vector \( (\pi_0(V)^{-1} - 1)Z^T \hat{G}_{mc}(V; \alpha^*) \) is bounded, it follows from the Cauchy-Schwarz inequality that each element of \( P_0 \{ (\pi_0(V)^{-1} - 1)Z^T \hat{G}_{mc}(V; \alpha^*)h_N \} \) is bounded up to some constant by \( P_0(h_N^2) \). Since \( \rho_{mc}(f, g) \rightarrow 0 \) implies \( P_0(f - g)^2 \rightarrow 0 \), we have \( P_0h_N^2 \rightarrow 0 \) as \( N \rightarrow \infty \). This verifies the asymptotic equicontinuity of \( \mathbb{G}_N^{\pi,mc} \) and hence completes showing its weak convergence.

The cases for \( \mathbb{G}_N^{\pi,e} \), \( \mathbb{G}_N^{\pi,cc} \) and \( \mathbb{G}_N^{\pi,cc} \) follow analogously. \( \square \)

Proof of Theorem 5.4. Since \( F \) is Donsker, it follows by Lemma 2.3.11 of [42] that \( E^* \| \mathbb{G}_N \|_{F_{\delta_N}} \rightarrow 0 \) for every sequence \( \delta_N \downarrow 0 \). Thus, the result follows from Lemma 7.2. Apply Markov’s inequality to obtain \( \| \mathbb{G}_N^\pi \|_{F_{\delta_N}} = O_{P^*}(1) \). For the second statement, consider the expansion (7.38) of \( \mathbb{G}_N^{\pi,mc} f - \mathbb{G}_N^\pi f \) with \( f \in F_{\delta_N} \). The first term is \( o_{P^*}(1) \) by Lemma 7.3. Since \( f \) converges to zero in \( L_2(P_0) \), the second term is \( o_{P^*}(1) \) by the dominated convergence theorem and Proposition 7.1. Apply the triangle inequality to conclude \( \| \mathbb{G}_N^{\pi,mc} \|_{F_{\delta_N}} = O_{P^*}(1) \).

The proofs for \( \mathbb{G}_N^{\pi,e} \), \( \mathbb{G}_N^{\pi,cc} \) and \( \mathbb{G}_N^{\pi,cc} \) are similar. \( \square \)

Proof of Lemma 5.3. Without loss of generality, assume that \( \hat{\theta}_N \) takes its values in \( \Theta_\delta \equiv \{ \theta \in \Theta : \| \theta - \theta_0 \| < \delta \} \) because of consistency of \( \hat{\theta}_N \) to \( \theta_0 \). Define a function \( f : \ell^\infty(\Theta_\delta \times \mathcal{H}) \times \Theta_\delta \rightarrow \ell^\infty(\mathcal{H}) \) by \( f(z, \theta)h = z(\theta, h) \). Note that \( f \) is continuous at every point \( (z, \theta_0) \) such that \( \| z(\theta, h) - z(\theta_0, h) \|_{\mathcal{H}} \rightarrow 0 \), as \( \theta \rightarrow \theta_0 \). To see this, suppose \( z_N \rightarrow z \) and \( \theta_N \rightarrow \theta_0 \). Then, for a fixed \( \epsilon > 0 \), there exists \( n_0 \) such that \( \| z_N - z \| < \epsilon \) and \( \| \theta_N - \theta_0 \| < \epsilon \) for \( N \geq N_0 \). For \( N \geq N_0 \), we have

\[
\| f(z_N, \theta_N) - f(z, \theta_0) \|_{\mathcal{H}} \\
\leq \| f(z_N, \theta_N) - f(z_0, \theta_N) \|_{\mathcal{H}} + \| f(z_0, \theta_N) - f(z_0, \theta_0) \| \\
\leq \sup_{\theta \in \Theta_\delta, h \in \mathcal{H}} | z_N(\theta, h) - z(\theta, h) | + \| z(\theta_N, h) - z(\theta_0, h) \|_{\mathcal{H}} \\
< 2\epsilon.
\]

Define a stochastic process \( Z_N \) indexed by \( \Theta_\delta \times \mathcal{H} \) by

\[
Z_N(\theta, h) = \mathbb{G}_N^\pi(\psi_{\theta, h} - \psi_{\theta_0, h}).
\]

Because \( \{ \psi_{\theta, h} - \psi_{\theta_0, h} : \| \theta - \theta_0 \| < \delta, \theta \in \Theta, h \in \mathcal{H} \} \) is Donsker, Theorem 5.3 implies that the sequence \( Z_N \) converges in \( \ell^\infty(\Theta_\delta \times \mathcal{H}) \) to a tight Gaussian process.
rem, Z given by
\[ Z = \mathcal{G} + \sum_{j=1}^{J} \sqrt{\nu_j} \sqrt{1 - \frac{p_j}{p_j^2}}, \]
This process has continuous sample paths with respect to the semimetric ρ given by
\[ \rho^2((\theta_1, h_1), (\theta_2, h_2)) = P (\psi_{\theta_1, h_1} - \psi_{\theta_\theta, h_1} - \psi_{\theta_2, h_2} + \psi_{\theta_0, h_2})^2 \]
because (Θ_δ × H, ρ) is totally bounded and Z is uniformly ρ-continuous. To see the latter, note that
\[ \rho^2((\theta_1, h_1), (\theta_2, h_2)) \geq P \left\{ (\psi_{\theta_1, h_1} - \psi_{\theta_\theta, h_1} - \psi_{\theta_2, h_2} + \psi_{\theta_0, h_2})^2 \mid V \in \mathcal{V}_j \right\} \nu_j \]
for each j = 1, ..., J. By assumption
\[ \sup_{h \in \mathcal{H}} \rho^2((\theta, h), (\theta_0, h)) = \sup_{h \in \mathcal{H}} P (\psi_{\theta, h} - \psi_{\theta_\theta, h} + 0)^2 \to 0, \]
as θ → θ_0. Thus, f is continuous at almost all sample paths of Z.

By Slutsky’s theorem, (Z_N, ˆθ_N) → (Z, ˆθ_0). By the continuous mapping theorem, Z_N( ˆθ_N) = f(Z_N, ˆθ_N) → f(Z, ˆθ_0) = 0 in ℓ∞(H).

The other cases for G_N, G_N^c, G_N^mc and G_N^cc follow analogously; see the proof of Theorem 5.3.

With the results of Section 5 in hand, we are ready to prove the main theorems.

**Proof of Theorem 3.1.** The asymptotic distributions of ˆθ_N is derived in [6]. Here we derive the asymptotic distribution of ˆθ_N,mc that is a solution of the calibrated weighted likelihood equations with modified calibration
\[ \Psi_{1,mc}(\theta, \eta, \alpha) = (\Psi_{1,mc}(\theta, \eta, \alpha), \Psi_{2,mc}(\theta, \eta, \alpha)) \]
\[ \Psi_{1,mc}(\theta, \eta, \alpha) = P_0 G_{mc}(V; \alpha) \hat{\ell}_{\theta, \eta}, \]
\[ \Psi_{2,mc}(\theta, \eta, \alpha) = P_0 G_{mc}(V; \alpha) (B_{\theta, \eta} h - P_{\theta, \eta} B_{\theta, \eta} h), \]
for all h ∈ H with α = ˆ\alpha_N. Let \( \Psi_{mc}(\theta, \eta, \alpha) = (\Psi_{1,mc}(\theta, \eta, \alpha), \Psi_{2,mc}(\theta, \eta, \alpha)) \)
The derivative map of \( \Psi_{mc} \) with respect to (θ, η) at (θ_0, η_0, α) has components
\[ P_0 \{ G_{mc}(V; \alpha) \psi_{ij, 0, \eta_0, \alpha_0}, i, j = 1, 2. \]
Our proof proceed by verifying the conditions of Theorem 1 of [7]. The weak convergence of \( \sqrt{N}(\Psi_{N,j,mc} - \Psi_{j,mc})(\theta_0, \eta_0, \alpha_0) \) follows from Theorem 5.3. The asymptotic equicontinuity conditions
\[ \sup_{\theta \in \Theta, \eta \in \mathcal{H}} \left\| \sqrt{N}(\Psi_{N,j,mc} - \Psi_{j,mc})(\theta, \eta, \hat{\alpha}_N) - \sqrt{N}(\Psi_{N,j,mc} - \Psi_{j,mc})(\theta, \eta, \alpha_0) \right\| = o_p(1), \]
for \( j = 1, 2 \), follows from Lemma 7.3. The other asymptotic equicontinuity condition
\[
\left\| \sqrt{N} (\Psi_{N,j,mc} - \Psi_{j,mc})(\hat{\theta}_{N,mc}, \hat{\eta}_{N,mc}, \alpha_0) - \sqrt{N} (\Psi_{N,j,mc} - \Psi_{j,mc})(\theta_0, \eta_0, \alpha_0) \right\|_\mathcal{H} = o_P(1),
\]
for \( j = 1, 2 \), follows from the Condition 3.4 and Lemma 5.3. Thus conditions (2) and (3) of [7] are satisfied.

The Fréchet differentiability of the map \((\theta, \eta) \mapsto \Phi_{j,mc}(\theta, \eta, \alpha)\) uniformly over the neighborhood of \(\alpha_0\) follows by the Condition 3.5 and boundedness of \(G\):
\[
\left\| \Psi_{mc}(\theta, \eta, \alpha)h - \Psi_{mc}(\theta_0, \eta_0, \alpha)h - \dot{\Psi}_{mc}( \theta, \eta, (\theta_0, \eta_0) ) \right\|_\mathcal{H} = \sup_{h \in \mathcal{H}} \left\{ E \left[ G_{mc}(V; \alpha) \left( \psi_{\theta,\eta,h} - \psi_{\theta_0,\eta_0,h} - \dot{\psi}_{\theta_0,\eta_0,h}(\theta, \eta, (\theta_0, \eta_0)) \right) \right] \right\} \\
\leq \left\{ E G^2_{mc}(V; \alpha) \right\}^{1/2} \sup_{h \in \mathcal{H}} \left[ E \left\{ \psi_{\theta,\eta,h} - \psi_{\theta_0,\eta_0,h} - \dot{\psi}_{\theta_0,\eta_0,h}(\theta, \eta, (\theta_0, \eta_0)) \right\}^2 \right]^{1/2} = o_P(1).
\]
The Fréchet derivative \(\dot{\Psi}_{mc}\) of the map \(\alpha \mapsto \{ \Psi_{mc}(\theta, \eta, \alpha)h : h \in \mathcal{H} \}\) is
\[
\frac{\partial}{\partial \alpha} \Psi_{mc}(\theta, \eta, \alpha)h = \frac{\partial}{\partial \alpha} E [G_{mc}(V; \alpha) \psi_{\theta,\eta,h}] = E \left[ \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \dot{\Gamma}_{mc}(V; \alpha) \psi_{\theta,\eta,h} \right].
\]

Now proceed in the same way as [7] to obtain
\[
\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) \\
= \sqrt{N}(\hat{\theta}_N - \theta_0) + E \left[ \bar{\ell}_{\theta_0,\eta_0} \frac{1 - \pi_0(V)}{\pi_0(V)} Z^T \dot{\Gamma}(0) \right] \sqrt{N}(\hat{\alpha}_N - \alpha_0) + o_P(1).
\]
Because \(\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^\top \bar{\ell}_{\theta_0,\eta_0} + o_P(1)\) (16 of [6]), it follows from (7.38) and consistency and asymptotic normality of \(\hat{\alpha}_N\) that \(\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) = \mathbb{G}_N^\top \bar{\ell}_{\theta_0,\eta_0} + o_P(1)\). Apply Theorem 5.3 to complete the proof.

The other three cases are similar. \(\Box\)

**Lemma 7.4.** Let \(Z_1, Z_2, \ldots\) be i.i.d. stochastic processes indexed by \(\mathcal{F}_N\) with \(E^*\|Z_1\|_{\mathcal{F}_N}\) uniformly bounded in \(N\). Suppose that \(\|S_N\|_{\mathcal{F}_N} \equiv \|\sum_{i=1}^N Z_i\|_{\mathcal{F}_N} = o_P(1)\). Then
\[
E^*\|S_N\|_{\mathcal{F}_N} \to 0, \quad N \to \infty.
\]

**Proof.** Fix \(\epsilon > 0\). Let \(Y_i\) be independent copies of \(Z_i\) and define \(\bar{T}_N = \sum_{i=1}^N Y_i\), and \(U_N = \bar{T}_N - S_N\). Since \(\|U_N\|_{\mathcal{F}_N} = o_P(1)\), \(\limsup_N P(\|U_N\|_{\mathcal{F}_N} \geq x \sqrt{N}) \leq \limsup_N P(\|U_N\|_{\mathcal{F}_N} \geq x) = 0\) by the portmanteau theorem. This implies that there exists \(N_0\) such that for \(N \geq N_0\)
\[
P^*(\|U_N\|_{\mathcal{F}_N} > x \sqrt{N}) \leq \epsilon/x^2.
\]
Since $U_N$ is a sum of independent symmetric processes, we can apply Lévy’s inequality to obtain
\[ P^* \left( \max_{1 \leq i \leq n} \| Z_i - Y_i \|_{F_N} > x\sqrt{N} \right) \leq 2P^* \left( \| U_N \|_{F_N} > x\sqrt{N} \right) \leq 2\epsilon/x^2. \]

In view of Problem 2.3.2 of [42], for every $N \geq N_0$,
\[ x^2 N P^* \left( \| Z_1 - Y_1 \|_{F_N} > x\sqrt{N} \right) \leq 4\epsilon. \]

Note that on the event that $\| Z_1 \|_{F_N} > x$, we have
\[ \beta_N(x) \equiv P^* \left( \| Y_1 \|_{F_N} < x/2 \right) \leq P^* \left( \| Z_1 - Y_1 \|_{F_N} > x/2 \right). \]

Integrating both sides with respect to $Z$ gives
\[ \beta_N(x) P^* \left( \| Z_1 \|_{F_N} > x \right) \leq P^* \left( \| Z_1 - Y_1 \|_{F_N} > x/2 \right). \]

By Markov’s inequality,
\[ \beta_N(x) = 1 - P^* \left( \| Y_1 \|_{F_N} \geq x/2 \right) \geq 1 - 2x^{-1}E\| Y_1 \|_{F_N} \]
Since $E\| Y_1 \|_{F_N}$ is uniformly bounded in $N$, it follows that, for $x$ sufficiently large, $\beta_N(x)^{-1}$ is uniformly bounded in $N$ and, therefore, $P^* \left( \| Z_1 \|_{F_N} > x\sqrt{N} \right)$ is bounded by $P^* \left( \| Z_1 - Y_1 \|_{F_N} > x\sqrt{N} \right)$ up to some constant for every $N$. Hence this proves that $P^* \left( \| Z_1 \|_{F_N} > x \right) = o(x^{-2})$.

Now we apply the Hoffmann-Jørgensen inequality to obtain
\[ E^* \| S_N \|_{F_N} \leq E^* \max_{i \leq N} \| Z_i \|_{F_N} + G_N^{-1}(u) \]
for an absolute constant $u$ where
\[ G_N(t) = P^* \left( \| S_N \|_{F_N} \leq t \right). \]

Since $P^* \left( \| Z_1 \|_{F_N} > x \right) = o(x^{-2})$, $E^* \max_{i \leq N} \| Z_i \|_{F_N} \to 0$ in view of Problem 2.3.3 of [42]. The second term goes to zero since $\| S_N \|_{F_N} = o_{P^*}(1)$. This completes the proof. \hfill $\square$

**Proof of Lemma 5.4.** Define $G_N = \{ N^{-1/2}f : f \in F_N \}$. We apply Lemma 7.4 with $Z_i$ and $F_N$ in Lemma 7.4 replaced by $\delta_X$, $P_0$ and $G_N$, respectively. The uniform boundedness condition of Lemma 7.4 is satisfied, because $E^* \| \delta_X - P_0 \|_{F_N} < \infty$ for $N \geq N_0$, and this expectation is decreasing in $N \geq N_0$. Thus,
\[ E^* \| G_N \|_{F_N} = E^* \sum_{i=1}^N (\delta_X, - P) \| G_N \to 0. \]

Apply Lemma 7.2, and Markov’s inequality to obtain
\[ \| G_N \|_{F_N} = o_{P^*}(1). \]

For the IPW process with modified calibration, consider the expansion (7.38) of $(\theta^*_N - \theta^*_N mc)$. Then the first term is $o_{P^*}(1)$ by Lemma 7.3. Suppose that $f = f_N \in F_N$ converges to zero pointwise. Since $(\pi_0(V)^{-1} - 1)Z\tilde{G}_{mc}$ is bounded, the second term in the expansion (7.38) is $o_{P^*}(1)$ by the dominated convergence theorem and Proposition 7.1. Suppose instead that $f = f_N \in F_N$ converges to
zero in $L_1(P_0)$. Then the same conclusion that the second term in the expansion (7.38) is $o_{P^*}(1)$ follows directly. Apply the triangle inequality to conclude $||G_{N,mc}^\pi||_{F_N} = o_{P^*}(1)$.

The proofs for $G_{N}^{\pi,c}, G_{N}^{\pi,c}$ and $G_{N}^{\pi,cc}$ are similar. □

Proof of Theorem 3.2. We only consider the WLE with modified calibration, $\theta_{N,mc}$. The other four cases are similar.

We evaluate the stochastic order of

\[ \sqrt{N}P_{N} \ell_{\theta_{N,mc},\hat{\eta}_{N,mc}} = o_{P^*}(N^{-1/2}) \]

by assumption and $P_0 \ell_{\hat{\theta}_0,\eta_0} = 0$, we have

\[ \sqrt{N}P_{N} \ell_{\theta_{N,mc},\hat{\eta}_{N,mc}} = -G_{N,mc}^\pi(\hat{\theta}_{N,mc},\hat{\eta}_{N,mc} - \hat{\theta}_{\theta_0,\eta_0}) + o_{P^*}(1). \]

Let $\delta_N \downarrow 0$ be arbitrary and define $F_N = \{ \ell_{\theta,\eta} - \ell_{\theta_0,\eta_0} : |\theta - \theta_0| \leq \delta_N, ||\eta - \eta_0|| \leq N^{-\beta} \}$. Then $f \in F_N$ converges to zero either pointwise or in $L_1(P_0)$ by Condition 3.8 as $N \to \infty$. Moreover, it follows from Condition 3.8 that $||G_N||_{F_N} = o_{P^*}(1)$ and that there exists some $N_0$ that $F_N$ is Glivenko-Cantelli for $N \geq N_0$. Apply Lemma 5.4 to obtain $||G_{N,mc}^\pi||_{F_N} = o_{P^*}(1)$ and conclude

\[ \sqrt{N}P_{N} \ell_{\theta_{N,mc},\hat{\eta}_{N,mc}} = o_{P^*}(1). \]

Similarly, $\sqrt{N}P_{N} B_{\theta_0,\eta_0}[h^*]+\sqrt{N}P_{0}B_{\theta_{N,mc},\hat{\eta}_{N,mc},\hat{\eta}_{0}}[h^*] = o_{P^*}(1)$. These stochastic orders and Condition 3.9 imply that

\[ P_0 \left\{ -\ell_{\theta_0,\eta_0}(\hat{\theta}_{N,mc} - \theta_0) + B_{\theta_0,\eta_0}([\hat{\eta}_{N,mc} - \eta_0]) \right\} + o \left( \|\hat{\eta}_{N,mc} - \eta_0\|^\alpha \right) + P_{N,mc}^\pi \ell_{\theta_0,\eta_0} = o_{P^*}(N^{-1/2}), \]

(7.39)

and, furthermore, that

\[ P_0 \left\{ -B_{\theta_0,\eta_0}[h^*] \right\} \left( \ell_{\theta_0,\eta_0} \right) \left( \hat{\theta}_{N,mc} - \theta_0 \right) = o \left( \|\hat{\theta}_{N,mc} - \theta_0\|^\alpha \right) + \|P_{N,mc}^\pi \left( \ell_{\theta_0,\eta_0} - B_{\theta_0,\eta_0}[h^*] \right) \]

(7.40)

By Condition 3.6 and $\alpha \beta > 1/2$, $\sqrt{N}O_{P^*}(\|\hat{\eta}_{N} - \eta_0\|^\alpha) = o_{P^*}(1)$. So by Condition 3.7 and taking the difference of (7.39) and (7.40), we have

\[ -P_0 \left\{ \left( \ell_{\theta_0,\eta_0} - B_{\theta_0,\eta_0}[h^*] \right) \left( \hat{\theta}_{N,mc} - \theta_0 \right) ^\top \right\} \left( \hat{\theta}_{N,mc} - \theta_0 \right) + o \left( \|\hat{\theta}_{N,mc} - \theta_0\|^\alpha \right) \]

\[ +o_{P}(N^{-1/2}) - o_{P}(N^{-1/2}) + P_{N,mc}^\pi \left( \ell_{\theta_0,\eta_0} - B_{\theta_0,\eta_0}[h^*] \right) = o_{P}(N^{-1/2}) - o_{P}(N^{-1/2}), \]

or

\[ -I_0(\hat{\theta}_{N,mc} - \theta_0) = P_{N,mc}^\pi \left( \ell_{\theta_0,\eta_0} - B_{\theta_0,\eta_0}[h^*] \right) + o_{P}(N^{-1/2}). \]
It follows by the invertibility of $I_0$ that
\[
\sqrt{N} \left( \hat{\theta}_{N,mc} - \theta_0 \right) = -\sqrt{N} \mathbb{P}^{\pi,mc}_{N} I_0^{-1} \left( \hat{\theta}_{0,m0} - B_{\theta_0,m0} \mathbb{H}^* \right) + o_P(1).
\]

Now, we recognize that the summand inside $\mathbb{P}^{\pi,mc}_{N}$ is the efficient influence function for $\theta$ and apply Theorem 5.3.

\textit{Proof of Theorem 3.3.} Theorem 3.1 for cases for $\hat{\theta}_{N,mc}^{Bern}$ and $\hat{\theta}_{N,e}^{Bern}$ are proved in [6, 7]. We only consider the WLE with modified calibration, $\hat{\theta}_{N,mc}$. The other four estimators for both theorems are similar.

Under stratified Bernoulli sampling, independence of sampling indicators allows us to proceed in the same as in the proofs of Theorems 3.1 and 3.2 to conclude $\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) = \sqrt{N} \mathbb{P}^{\pi,mc}_{N} \hat{\theta}_0 + o_P(1)$ and asymptotic linearity of $\hat{\alpha}_N$ in Proposition 7.1. In view of (7.38), $\sqrt{N}(\hat{\theta}_{N,mc} - \theta_0) = \sqrt{N} \mathbb{P}_{N} f + o_P(1)$ where
\[
f(X, V, \xi) = \frac{\xi}{\pi_0(V)} \hat{\theta}_0 - \frac{\xi - \pi_0(V)}{\pi_0(V)} Q_{mc} \hat{\theta}_0.
\]  

Apply the central limit theorem and compute
\[
\Sigma_{mc}^{Bern} = \text{Var}(f)
\]
\[
= \text{Var} \left( E \left[ \frac{\xi}{\pi_0(V)} \hat{\theta}_0 - \frac{\xi - \pi_0(V)}{\pi_0(V)} Q_{mc} \hat{\theta}_0 \mid X, V \right] \right) + E \left[ \text{Var} \left( \frac{\xi}{\pi_0(V)} \hat{\theta}_0 - \frac{\xi - \pi_0(V)}{\pi_0(V)} Q_{mc} \hat{\theta}_0 \mid X, V \right) \right]
\]
\[
= \text{Var}(\hat{\theta}_0) + E \left[ \text{Var} \left( \frac{\xi}{\pi_0(V)} (I-Q_{mc}) \hat{\theta}_0 \mid X, V \right) \right]
\]
\[
= I_0^{-1} + E \left[ \frac{1 - \pi_0(V)}{\pi_0(V)} \left( I - Q_{mc} \right) \hat{\theta}_0 \right] \otimes^2
\]
\[
= I_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} P_{0j} \left( I - Q_{mc} \right) \hat{\theta}_0 \right] \otimes^2.
\]

\textit{Proof of Corollary 3.2.} We only consider the WLE with modified calibration, $\hat{\theta}_{N,mc}$. The other two cases are similar.

Let $Q_{mc} \hat{\theta}_0 = AZ$ where $A = A_1 A_2$ with $A_1 = P_0[(\pi_0^{-1}(V) - 1) \hat{\theta}_0 Z^T]$ and $A_2 = \{P_0[(\pi_0^{-1}(V) - 1) Z^2]\}^{-1}$. Recall that $\Sigma^{Bern} = \text{Var}\{\xi/\pi_0(V)\hat{\theta}_0\}$. In view of (7.41), it suffices to show that $\text{Cov}\{(\xi/\pi_0(V))\hat{\theta}_0, (\xi/\pi_0(V) - 1) AZ\}$ is equal.
to \( \text{Var}((\xi/\pi_0(V)-1)AZ) \). This is true since
\[
\text{Cov}\left\{ \frac{\xi}{\pi_0(V)}\tilde{\ell}_0, \frac{\xi-\pi_0(V)}{\pi_0(V)}AZ \right\} = E\left\{ \frac{\xi}{\pi_0(V)}\tilde{\ell}_0, \frac{\xi-\pi_0(V)}{\pi_0(V)}Z \right\} A^T
\]
\[
= E\left[ \tilde{\ell}_0 Z E\left\{ \frac{\xi}{\pi_0(V)}\frac{\xi-\pi_0(V)}{\pi_0(V)}|X,V \right\} \right] A^T
\]
\[
= E\left[ \frac{1-\pi_0(V)}{\pi_0(V)}\tilde{\ell}_0 Z \right] A^T = A_1A_2A_1^T,
\]
and
\[
\text{Var}\left( \frac{\xi-\pi_0(V)}{\pi_0(V)}AZ \right) = A\text{Var}\left( \frac{\xi-\pi_0(V)}{\pi_0(V)}Z \right) A^T
\]
\[
= AE\left[ \text{Var}\left( \frac{\xi-\pi_0(V)}{\pi_0(V)}Z|X,V \right) \right] A^T
\]
\[
+ A\text{Var}\left( Z E\left[ \frac{\xi-\pi_0(V)}{\pi_0(V)}|X,V \right] \right) A^T
\]
\[
= AE\left[ Z \otimes \frac{1-\pi_0(V)}{\pi_0(V)} \right] A^T + 0 = A_1A_2A_1^T.
\]

\[\square\]

**Proof of Corollary 3.3.** (1) We first consider stratified Bernoulli sampling. The case for \( \hat{\theta}_{N, e} \) was proved in [3]. We only consider the WLE with modified calibration, \( \hat{\theta}_{N, mc} \). The other two cases, (3.19) and (3.21) corresponding to \( \hat{\theta}_{N, e} \) and \( \hat{\theta}_{N, ec} \), are similar.

For \( \tilde{Z} \equiv (Z^{(1)}, \ldots, Z^{(J)})^T \) with \( Z^{(j)} \equiv I(V \in V_j)Z^T \), we compute \( \tilde{A}_1 \equiv P_0[(\pi_0^{-1}(V) - \tilde{\ell}_0 \tilde{Z})^T] \) and \( \tilde{A}_2 \equiv \left\{ P_0[(\pi_0^{-1}(V) - \tilde{\ell}_0 \tilde{Z})^T] \right\}^{-1} \). Note that \( Q_{mc} \tilde{\ell}_0 = \tilde{A}_1 \tilde{A}_2 \tilde{Z} \). The matrix \( \tilde{A}_1 = [\tilde{A}_{1,1}, \ldots, \tilde{A}_{1,J}] \) is a partitioned matrix where
\[
\tilde{A}_{1,j} \equiv P_0\left( \frac{1-\pi_0(V)}{\pi_0(V)}\tilde{\ell}_0 Z^{(j)} \right) = \nu_j P_{0|j} \left( \frac{1-p_j}{p_j} \tilde{\ell}_0 Z^T \right) \in \mathbb{R}^{p \times k}.
\]
and the matrix \( \tilde{A}_2 \) is the block diagonal matrix the jth block of which is
\[
\tilde{A}_{2,j} \equiv \left\{ P_0\left( \frac{1-\pi_0(V)}{\pi_0(V)}[(Z^{(j)})^T]^\otimes 2 \right) \right\}^{-1} = \left\{ \nu_j P_{0|j} \left( \frac{1-p_j}{p_j} Z^\otimes 2 \right) \right\}^{-1} \in \mathbb{R}^{k \times k}.
\]
Thus, the matrix \( \tilde{A} \equiv \tilde{A}_1 \tilde{A}_2 \) is a partitioned matrix \( \tilde{A} = [\tilde{A}_1, \ldots, \tilde{A}_J] \) where
\[
\tilde{A}_j = \tilde{A}_{1,j} \tilde{A}_{2,j} = P_{0|j} \left( \tilde{\ell}_0 Z^T \right) \left\{ P_{0|j} Z^\otimes 2 \right\}^{-1}.
\]
It follows by the definition of the \( Z^{(j)} \)'s that
\[
P_{0|j} \left( (I - Q_{mc}) \tilde{\ell}_0 \right)^\otimes 2 = P_{0|j} \left( \tilde{\ell}_0 - \tilde{A} \tilde{Z} \right)^\otimes 2
\]
\[
= P_{0|j} \left( \tilde{\ell}_0 - \tilde{A}_j Z \right)^\otimes 2 = P_{0|j} \left( (I - Q_{c}^{(j)}) \tilde{\ell}_0 \right)^\otimes 2.
\]
Since
\[ P_{0ij} \left( \tilde{A}_j Z \right)^\otimes 2 \left( \tilde{A}_j Z \right)^T = \tilde{A}_j P_{0ij} Z \otimes 2 \tilde{A}_j^T = P_{0ij} \left( \tilde{\ell}_0 Z^T \right) \left( P_{0ij} Z \otimes 2 \right)^{-1} P_{0ij} \left( \tilde{\ell}_0 Z^T \right)^T, \]

and
\[ P_{0ij} \left( \tilde{\ell}_0 Z^T \right) \tilde{A}_j^T = P_{0ij} \left( \tilde{\ell}_0 Z^T \right) \left( P_{0ij} Z \otimes 2 \right)^{-1} P_{0ij} \left( \tilde{\ell}_0 Z^T \right)^T, \]
it follows that
\[ P_{0ij} \left( \left( I - Q^{(j)}_c \tilde{\ell}_0 \right) \right)^\otimes 2 = P_{0ij} \tilde{\ell}_0^\otimes 2 - P_{0ij} \left( Q^{(j)}_c \tilde{\ell}_0 \right)^\otimes 2. \]

Substitution of this into (3.17) gives (3.20).

(2) Next, we consider the second part of Corollary 3.3 concerning stratified sampling without replacement. For \( \tilde{Z} = (Z^{(1)}, \ldots, Z^{(j)})^T \) with \( Z^{(j)} \equiv I(V \in V_j)^Z^T \), we compute \( \tilde{B}_1 \equiv P_0 \left[ \tilde{\sigma}_0^{-1} (V) - 1 \right] \tilde{\ell}_0 (\tilde{Z} - \mu_Z)^T \) and \( \tilde{B}_2 \equiv P_0 \left[ \tilde{\sigma}_0^{-1} (V) - 1 \right] (\tilde{Z} - \mu_Z)^{\otimes 2} \). Note that \( \tilde{Q}_c \tilde{\ell}_0 = \tilde{B}_1 \tilde{B}_2 \tilde{Z} \) and \( \mu_Z = (\mu_{Z,1}^T, \ldots, \mu_{Z,j}^T)^T \). The matrix \( \tilde{B}_1 = [\tilde{B}_{1,1}, \ldots, \tilde{B}_{1,j}] \) is a partitioned matrix where
\[ \tilde{B}_{1,j} \equiv P_0 \left( \frac{1 - \tilde{\sigma}_0 (V)}{\tilde{\sigma}_0 (V)} \tilde{\ell}_0 (Z^{(j)} - \mu_Z^{(j)}) \right) = \nu_j P_{0ij} \left( \frac{1 - p_j}{p_j} \tilde{\ell}_0 (Z - \mu_Z)^T \right). \]

and the matrix \( \tilde{B}_2 \) is the block diagonal matrix the \( j \)th block of which is
\[ \tilde{B}_{2,j} \equiv \left\{ P_0 \frac{1 - \tilde{\sigma}_0 (V)}{\tilde{\sigma}_0 (V)} (Z^{(j)} - \mu_Z^{(j)})^{\otimes 2} \right\}^{-1} = \left\{ \nu_j P_{0ij} \frac{1 - p_j}{p_j} (Z - \mu_Z)^{\otimes 2} \right\}^{-1}. \]

Thus, the matrix \( \tilde{B} \equiv [\tilde{B}_1, \ldots, \tilde{B}_j] \) is a partitioned matrix \( \tilde{B} = [\tilde{B}_{1,j} \tilde{B}_{2,j}] \) where
\[ \tilde{B}_j = \tilde{B}_{1,j} \tilde{B}_{2,j} = P_{0ij} \left( \tilde{\ell}_0 (Z - \mu_Z)^T \right) \left( P_{0ij} (Z - \mu_Z)^{\otimes 2} \right)^{-1}. \]

It follows by the definition of \( Z^{(j)} \)’s that
\[ \text{Var}_{0ij} \left\{ \left( I - Q_{cc} \right) \tilde{\ell}_0 \right\} = \text{Var}_{0ij} \left\{ \tilde{\ell}_0 - \tilde{B} (\tilde{Z} - \mu_Z) \right\} \]
\[ = \text{Var}_{0ij} \left\{ \tilde{\ell}_0 - \tilde{B}_j (Z - \mu_Z) \right\} = \text{Var}_{0ij} \left\{ (I - Q^{(j)}_c) \tilde{\ell}_0 \right\}. \]

Then, since
\[ \text{Var}_{0ij} \left( \tilde{B}_j (Z - \mu_Z) \right) = \tilde{B}_j \text{Var}_{0ij} (Z) \tilde{B}_j^T \]
\[ = P_{0ij} \left( \tilde{\ell}_0 (Z - \mu_Z)^T \right) \left( \text{Var}_{0ij} (Z) \right)^{-1} P_{0ij} \left( \tilde{\ell}_0 (Z - \mu_Z)^T \right)^T, \]

and
\[ \text{Cov}_{0ij} \left( \tilde{\ell}_0, \tilde{B}_j (Z - \mu_Z) \right) = P_{0ij} \left( \tilde{\ell}_0 (Z - \mu_Z)^T \right) \tilde{B}_j^T \]
\[ = P_{0ij} \left( \tilde{\ell}_0 (Z - \mu_Z)^T \right) \left( \text{Var}_{0ij} (Z) \right)^{-1} P_{0ij} \left( \tilde{\ell}_0 (Z - \mu_Z)^T \right)^T, \]
it follows that
\[
\text{Var}_{0|j} \left\{ (I - Q_c^{(j)} \tilde{\ell}_0) \right\} = \text{Var}_{0|j} \left( \tilde{\ell}_0 \right) - \text{Var}_{0|j} \left\{ Q_c^{(j)} \tilde{\ell}_0 \right\}.
\]
Substitution of this last identity into (3.12) gives (3.22).

Acknowledgements: We owe thanks to Kwun Chuen (Gary) Chan for suggesting the modified calibration method introduced in subsection 2.1.3. We would also like to thank Norman Breslow for many helpful conversations concerning problems connected with two-phase sampling and for a helpful critique of the first version of this manuscript.

References

[1] Andersen, P. K. and Gill, R. D. (1982). Cox’s regression model for counting processes: a large sample study. *Ann. Statist.* 10 1100–1120. With Discussion.

[2] Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* 6 39–58.

[3] Breslow, N. E., Lumley, T., Ballantyne, C., Chambless, L. and Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Stat. Biosc.* 1 32-49.

[4] Breslow, N. E., Lumley, T., Ballantyne, C., Chambless, L. and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort data. *American J. Epidemiol.* 169 1398-1405.

[5] Breslow, N. E., McNeney, B. and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.* 31 1110–1139.

[6] Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.* 34 86–102.

[7] Breslow, N. E. and Wellner, J. A. (2008). A Z-theorem with estimated nuisance parameters and correction note for: “Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression” [Scand. J. Statist. 34 (2007), no. 1, 86–102]. *Scand. J. Statist.* 35 186–192.

[8] Chan, K. C. G. (2010). Personal communication.

[9] Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.* 94 896–908.

[10] Chen, J. and Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statist. Sinica* 17 1047–1064.

[11] Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* 34 187–220. With discussion.
[12] DeVille, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382.

[13] Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.* **54** 127–138.

[14] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685.

[15] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24** 540–568.

[16] Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96.

[17] Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *J. Roy. Statist. Soc. Ser. B* **36** 1–37. With discussion.

[18] Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*. Wiley.

[19] Lin, D. Y. (2000). On fitting Cox’s proportional hazards models to survey data. *Biometrika* **87** 37–47.

[20] Lumley, T., Shaw, P. A. and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* **79** 200–232.

[21] Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.* **96** 190–217.

[22] McNeney, B. and Wellner, J. A. (2000). Application of convolution theorems in semiparametric models with non-i.i.d. data. *J. Statist. Plann. Inference* **91** 441–480. Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998).

[23] Murphy, S. A. and van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25** 1471–1509.

[24] Murphy, S. A. and van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli* **5** 381–412.

[25] Nan, B. (2004). Efficient estimation for case-cohort studies. *Canad. J. Statist.* **32** 403–419.

[26] Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.* **33** 101–116.

[27] Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086.

[28] Prentice, R. L. (1986). A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika* **73** pp. 1-11.

[29] Robins, J. M., Hsieh, F. S. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. Roy. Statist. Soc. Ser. B* **57** 409–424.

[30] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866.

[31] Rubin-Bleuer, S. (2011). The proportional hazards model for survey data
from independent and clustered super-populations. *Journal of Multivariate Analysis* **101** 884–895.

[32] RUBIN-BLEUER, S. and SCHIOPU KRATINA, I. (2005). On the two-phase framework for joint model and design-based inference. *Ann. Statist.* **33** 2789–2810.

[33] SAEGUSA, T. (2012). Improving efficiency and robustness under two-phase sampling PhD thesis, University of Washington.

[34] SAEGUSA, T. and WELLNER, J. A. (2012). Supplementary Materials for “Weighted likelihood estimation under two-phase sampling”.

[35] SCOTT, A. and WILD, C. (2002). On the robustness of weighted methods for fitting models to case-control data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 207–219.

[36] SCOTT, A. J. and WILD, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. Roy. Statist. Soc. Ser. B* **48** 170–182.

[37] SCOTT, A. J. and WILD, C. J. (1989). Hypothesis testing in case-control studies. *Biometrika* **76** 806–808.

[38] SELF, S. G. and PRENTICE, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16** 64–81.

[39] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999).* Lecture Notes in Math. **1781** 331–457. Springer, Berlin.

[40] VAN DER VAART, A. and WELLNER, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High Dimensional Probability, II (Seattle, WA, 1999).* *Progr. Probab.* **47** 115–133. Birkhäuser Boston, Boston, MA.

[41] VAN DER VAART, A. W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics **3**. Cambridge University Press, Cambridge.

[42] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer Series in Statistics. Springer-Verlag, New York.

[43] WELLNER, J. A. and ZHAN, Y. (1996). Bootstrapping Z-estimators Technical Report No. # 308, Department of Statistics, University of Washington.

[44] WHITE, J. E. (1986). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidemiol.* **115** 119–128.

[45] XIE, Y. and MANSKI, C. F. (1989). The logit model and response-based samples. *Sociol. Meth. Res.* **17** 283–302.