The Annals of Statistics
2021, Vol. 49, No. 1, 459–485
https://doi.org/10.1214/20-AOS1964

COMPLEX SAMPLING DESIGNS: UNIFORM LIMIT THEOREMS AND APPLICATIONS

BY QIYANG HAN[1] AND JON A. WELLNER[2]

[1]*Department of Statistics, Rutgers University, qh85@stat.rutgers.edu*
[2]*Department of Statistics, University of Washington, jaw@stat.washington.edu*

In this paper, we develop a general approach to proving global and local uniform limit theorems for the Horvitz–Thompson empirical process arising from complex sampling designs. Global theorems such as Glivenko–Cantelli and Donsker theorems, and local theorems such as local asymptotic modulus and related ratio-type limit theorems are proved for both the Horvitz–Thompson empirical process, and its calibrated version. Limit theorems of other variants and their conditional versions are also established. Our approach reveals an interesting feature: the problem of deriving uniform limit theorems for the Horvitz–Thompson empirical process is essentially no harder than the problem of establishing the corresponding finite-dimensional limit theorems, once the usual complexity conditions on the function class are satisfied. These global and local uniform limit theorems are then applied to important statistical problems including (i) M-estimation, (ii) Z-estimation and (iii) frequentist theory of pseudo-Bayes procedures, all with weighted likelihood, to illustrate their wide applicability.

1. Introduction.

1.1. *Overview.* Over the past thirty years, uniform limit theorems for the empirical process have proved to be a universal tool in various statistical problems based on independent observations; we only refer readers to the textbooks [35, 48, 77, 81] for relevant theoretical developments and various statistical applications.

Our focus here will be uniform limit theorems for the Horvitz–Thompson empirical process arising from complex sampling designs (cf. [70]). Such limit theorems provide fundamental probabilistic tools in statistical applications with survey data, for instance, in combination with the functional delta method (see, e.g., [4, 8, 9, 27] for applications in econometrics), or in semiparametric modeling (see, e.g., [13–15, 50, 56, 57] for applications in biostatistics), just to name a few. Recent years have seen the emergence of interest in further limit theory in this direction (e.g., [7, 11, 16, 17, 26, 68, 69]), but the scope of the existing results in this direction has been somewhat limited, and many of these available results have been derived based on case-by-case analyses. Roughly speaking, there are three approaches so far in the literature:

1. Breslow and Wellner [16, 17] developed theory in the context of two-phase sampling with phase II a simple sampling without replacement sampling design. The key idea therein is to view the Horvitz–Thompson empirical process conditionally as an exchangeably weighted bootstrap empirical process [60]. This idea is further exploited in [69] in the context of calibrated Horvitz–Thompson empirical processes. A similar bootstrap approach is adopted in [68] in the setting of stratified sampling with potential overlaps.

Received April 2019; revised December 2019.
MSC2020 subject classifications. Primary 60E15; secondary 62G05.
Key words and phrases. Complex sampling design, empirical process, uniform limit theorems.

2. Bertail et al. [7] derived a Donsker theorem for the Bernoulli sampling design and other sampling designs that are close enough to the rejective sampling design (= high entropy designs) under a uniform entropy condition on the indexing function class. Their techniques heavily rely on the conditional independence of the inclusion indicators.

3. Conti [26] and Boistard et al. [11] established Donsker theorems over one class $\{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$ under sampling designs with increasing level of generality, by explicit calculations that verify the one-dimensional tightness condition.

The apparent case-by-case complication here is that complex sampling designs typically induce complicated dependence structure between the samples, so in order to use existing techniques from empirical process theory, certain latent independence or exchangeability structure needs to be identified in a case-by-case routine.

On the other hand, some structural commonality is indeed hinted at by the results proved in the above cited papers: uniform laws of large numbers (i.e., Glivenko–Cantelli theorems) and uniform central limit theorems (i.e., Donsker theorems) hold under rather minimal conditions on the indexing function classes. The intriguing question naturally arises:

QUESTION 1.1. Does there exist any general approach to proving uniform limit theorems for the Horvitz–Thompson empirical process under natural conditions, without being confined to a particular form of the sampling design?

A possible solution to this very natural question, however, appears far from obvious from the previously described approaches. The challenges involved here were already noted in Lin [50] as "*...To our knowledge there does not exist a general theory on conditions required for the tightness and weak convergence of Horvitz–Thompson processes...*," dating back to as early as 2000. One of the goals of this paper is to address Question 1.1 in an appropriate general framework that includes a wide variety of sampling designs. Part of the philosophical difficulty in such a general approach is that there is an easily believable impression that any general attempt at establishing global uniform limit theorems for the Horvitz–Thompson empirical process, must necessarily give general recipes for establishing finite-dimensional convergence of the Horvitz–Thompson empirical process. In the specific context of Donsker theorems, this impression pushes one to think about the "right conditions" under which at least central limit theorems hold for a single function under various different sampling designs—a task that usually already requires a case-by-case study.

In this paper, we show that this easily believable impression need not be the rule in the context of uniform limit theorems for Horvitz–Thompson empirical processes, at least in the superpopulation framework adopted in [11, 65] with uniformly positive first-order inclusion probabilities. The major "change of thinking" adopted in the current paper, interestingly, indicates that *the problem of deriving uniform limit theorems for Horvitz–Thompson empirical processes is not really more difficult than that of establishing the corresponding finite-dimensional limit theorems, once the usual complexity conditions on the function class are satisfied*. In the context of Donsker theorems, this amounts to saying that, as long as the Horvitz–Thompson empirical process converges finite-dimensionally, weak convergence at the process level follows almost automatically. Since finite-dimensional convergence is necessary for weak convergence of the process to hold, the real point here is to separate the problem of establishing finite-dimensional convergence of the Horvitz–Thompson empirical process from that of establishing a uniform limit theorem. The approach here is in part inspired by a multiplier inequality developed in a recent work of the authors [40], which holds regardless of the dependence structure among the multipliers, given sufficient independence structure between the multipliers and the samples.

Establishing global uniform limit theorems serves as a first step in understanding the behavior of these Horvitz–Thompson empirical processes. In typical semi/nonparametric applications, it is also of crucial importance to understand the local behavior of these empirical processes. To this end, we further study the local behavior of the Horvitz–Thompson empirical process by characterizing its local asymptotic modulus and proving several ratio-type limit theorems. These local uniform limit theorems show that the Horvitz–Thompson empirical process typically has similar local behavior compared to its empirical process counterpart. Similar global and local uniform limit theorems are established for the calibrated version of the Horvitz–Thompson empirical processes. Some other variants of Horvitz–Thompson empirical processes are discussed. Conditional versions of the uniform limit theorems are also established.

As an illustration and a proof of concept of the utility of our global and local uniform limit theorems (and related techniques), we apply these new tools to a variety of important statistical problems, including (i) M-estimation, or *empirical risk minimization*, in a general nonparametric model, (ii) Z-estimation in a general semiparametric model and (iii) frequentist theory of pseudo-Bayesian procedures (i.e., theory of posterior contraction rates and Bernstein–von Mises type theorems), all based on weighted likelihood. Several concrete examples are illustrated to further demonstrate the applicability of these general results.

The rest of the paper is organized as follows. Section 2 is devoted to a general probabilistic framework for complex sampling designs and detailed illustrations of the theory in the context of a number of examples. Section 3 studies the global and local uniform limit theorems for the Horvitz–Thompson empirical process. Section 4 gives applications of the theory developed in Section 3 to the statistical problems listed above. Proofs are collected in the Appendix (see Supplementary Material [41]).

1.2. *Notation.* For a real-valued measurable function f defined on $(\mathcal{X}, \mathcal{A}, P)$ and $p \geq 1$, $\|f\|_{L_p(P)} \equiv (P|f|^p)^{1/p}$ denotes the usual L_p-norm under P, and $\|f\|_\infty \equiv \|f\|_{L_\infty} \equiv \sup_{x\in\mathcal{X}}|f(x)|$. f is said to be P-centered if $Pf = 0$. $L_p(g, B)$ denotes the $L_p(P)$-ball centered at g with radius B. For simplicity, we write $L_p(B) \equiv L_p(0, B)$.

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of the normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-covering number, and let $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-bracketing number; see page 83 of [81] for more details. To avoid unnecessary measurability digressions, we assume that $\mathcal{F}$ is countable throughout the article. As usual, for any $\phi : \mathcal{F} \to \mathbb{R}$, we write $\|\phi(f)\|_\mathcal{F}$ for $\sup_{f\in\mathcal{F}}|\phi(f)|$.

Throughout the article, $\varepsilon_1, \ldots, \varepsilon_n$ will be i.i.d. Rademacher random variables independent of all other random variables. C_x will denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \leq C_x b$ and $a \geq C_x b$, respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$ [$a \lesssim b$ means $a \leq Cb$ for some absolute constant C]. For two real numbers, a, b, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. For two sequence of nonnegative real numbers $\{a_n\}$, $\{b_n\}$, $a_n \ll (\gg) b_n$ means $\lim_n a_n/b_n = 0(\infty)$. We slightly abuse notation by defining $\log(x) \equiv \log(x \vee e)$ (and similarly for $\log\log(x)$).

2. Sampling designs.

2.1. *Setup.* Let $U_N \equiv \{1, \ldots, N\}$, and $\mathcal{S}_N \equiv \{\{s_1, \ldots, s_n\} : 0 \leq n \leq N, s_i \in U_N, s_i \neq s_j, \forall i \neq j\}$ ($n = 0$ corresponds to the empty set) be the collection of subsets of U_N. We adopt the superpopulation framework as in [65]: Let $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$, $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ be measurable spaces, and $\{(Y_i, Z_i) \in \mathcal{Y} \times \mathcal{Z}\}_{i=1}^N$ be i.i.d. superpopulation samples defined on the probability space $(\mathcal{X}, \mathcal{A}, P_{(Y,Z)}) \equiv (\mathcal{Y} \times \mathcal{Z}, \mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z}, P_{(Y,Z)})$. Here, $Y^{(N)} \equiv (Y_1, \ldots, Y_N)$ is the vector of interest, and $Z^{(N)} \equiv (Z_1, \ldots, Z_N)$ is an auxiliary vector. A sampling design is a function $\mathfrak{p} : \mathcal{S}_N \times \mathcal{Z}^{\otimes N} \to [0, 1]$ such that:

1. for all $s \in \mathcal{S}_N$, $z^{(N)} \mapsto \mathfrak{p}(s, z^{(N)})$ is measurable,
2. for all $z^{(N)} \in \mathcal{Z}^{\otimes N}$, $s \mapsto \mathfrak{p}(s, z^{(N)})$ is a probability measure.

The probability space we work with that includes both the superpopulation and the design-space is the same product space $(\mathcal{S}_N \times \mathcal{X}, \sigma(\mathcal{S}_N) \times \mathcal{A}, \mathbb{P})$ as constructed in [11]. We include the construction here for convenience of the reader: the probability measure $\mathbb{P}$ is uniquely defined through its restriction on all rectangles: for any $(s, E) \in \mathcal{S}_N \times \mathcal{A}$ (note that $\mathcal{S}_N$ is a finite set),

$$\mathbb{P}(s \times E) \equiv \int_E \mathfrak{p}\big(s, z^{(N)}(\omega)\big)\, \mathrm{d}P_{(Y,Z)}(\omega) \equiv \int_E \mathbb{P}_d(s, \omega)\, \mathrm{d}P_{(Y,Z)}(\omega), \tag{2.1}$$

where $\mathbb{P}_d(s, \omega) \equiv \mathfrak{p}(s, z^{(N)}(\omega))$. We also use P to denote the marginal law of Y for notational convenience.

Given $(Y^{(N)}, Z^{(N)})$ and a sampling design $\mathfrak{p}$, let $\{\xi_i\}_{i=1}^N \subset [0, 1]$ be random variables defined on $(\mathcal{S}_N \times \mathcal{X}, \sigma(\mathcal{S}_N) \times \mathcal{A}, \mathbb{P})$ with $\pi_i \equiv \pi_i(Z^{(N)}) \equiv \mathbb{E}[\xi_i | Z^{(N)}]$. We further assume that $\{\xi_i\}_{i=1}^N$ are independent of $Y^{(N)}$ conditionally on $Z^{(N)}$. Typically, we take $\xi_i \equiv \mathbf{1}_{i \in s}$, where $s \sim \mathfrak{p}(\cdot, Z^{(N)})$, to be the indicator of whether or not the ith sample Y_i is observed (and in this case $\pi_i(Z^{(N)}) = \sum_{s \in \mathcal{S}_N : i \in s} \mathfrak{p}(s, Z^{(N)})$), but we do not require this structure a priori. The π_i's are often referred to as the first-order inclusion probabilities, and $\pi_{ij} \equiv \pi_{ij}(Z^{(N)}) \equiv \mathbb{E}[\xi_i \xi_j | Z^{(N)}]$ are the second-order inclusion probabilities.

We define the Horvitz–Thompson empirical measure and empirical process as follows: for $\{\pi_i\}$, $\{\xi_i\}$, $\{Y_i\}$ as above

$$\mathbb{P}_N^\pi(f) \equiv \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_i} f(Y_i), \quad f \in \mathcal{F},$$

and the associated Horvitz–Thompson empirical process

$$\mathbb{G}_N^\pi(f) \equiv \sqrt{N}\big(\mathbb{P}_N^\pi - P\big)(f), \quad f \in \mathcal{F}.$$

The name of such an empirical process goes back to [44], in which $\mathbb{P}_N^\pi(Y) \equiv N^{-1} \sum_{i=1}^N (\xi_i / \pi_i) Y_i$ is used as an estimator for the population mean $P(Y) \equiv \mathbb{E}_{Y \sim P} Y$. The usual empirical measure and empirical process (i.e., with $\xi_i / \pi_i \equiv 1$ for all $i = 1, \ldots, N$) will be denoted by $\mathbb{P}_N$, $\mathbb{G}_N$, respectively.

ASSUMPTION A. Consider the following conditions on the sampling design $\mathfrak{p}$:

(A1) $\min_{1 \le i \le N} \pi_i \ge \pi_0$ holds for some nonrandom $\pi_0 > 0$.
(A2-LLN) $\frac{1}{N} \sum_{i=1}^N (\frac{\xi_i}{\pi_i} - 1) = \mathfrak{o}_{\mathbb{P}}(1)$.
(A2-CLT) $\frac{1}{\sqrt{N}} \sum_{i=1}^N (\frac{\xi_i}{\pi_i} - 1) = \mathcal{O}_{\mathbb{P}}(1)$.

(A1) is a common assumption in the literature. (A2-LLN) says that the weights $\{\xi_i / \pi_i\}$ satisfy a law of large numbers; while (A2-CLT) says that the weights $\{\xi_i / \pi_i\}$ have a $\sqrt{N}$ rate of convergence (so that a uniform central limit theorem for the more complicated Horvitz–Thompson empirical process $\mathbb{G}_N^\pi$ can be possible). As we will see below in the examples, a generic way of verifying these conditions is to obtain a good estimate on the correlations $\{\pi_{ij} - \pi_i \pi_j\}_{i \ne j}$. Conditions on (even higher order) correlations are very common in the literature; cf. [10–12, 18].

2.2. *Examples of sampling designs.*

EXAMPLE 2.1 (Sampling without replacement). A simple random sampling without replacement (SWOR) design $\mathfrak{p}$ is such that for all $z^{(N)} \in \mathcal{Z}^{\otimes N}$, $\mathfrak{p}(\cdot, z^{(N)})$ is the sampling without replacement design with cardinality $n(z^{(N)})$. In this case, the parameter in this sampling design is $n(z^{(N)})$ and $(\xi_1, \ldots, \xi_N)$ is a random permutation of the vector that contains 1 in the first $n(z^{(N)})$ components and 0 otherwise. Then

$$\pi_i\big(z^{(N)}\big) = \mathbb{E}\big[\xi_i | z^{(N)}\big] = \frac{n(z^{(N)})}{N}.$$

Condition (A1) holds if $n(z^{(N)})/N \geq c$ for some constant $c > 0$. Condition (A2) is trivially satisfied since $\sum_{i=1}^N \xi_i = n(z^{(N)})$, and hence

$$\sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right) = \left(\frac{1}{n(z^{(N)})/N} \cdot \sum_{i=1}^N \xi_i\right) - N = 0.$$

EXAMPLE 2.2 (Bernoulli sampling). A Bernoulli sampling design $\mathfrak{p}$ is such that for all $z^{(N)} \in \mathcal{Z}^{\otimes N}$ and $s \in \mathcal{S}_N$,

$$\mathfrak{p}\big(s, z^{(N)}\big) = \prod_{i \in s} \pi_i\big(z^{(N)}\big) \prod_{i \notin s} \big(1 - \pi_i\big(z^{(N)}\big)\big).$$

In other words, conditionally on auxiliary random variables $Z^{(N)}$, the ξ_i's are independent Bernoulli random variables with success probability $\pi_i(Z^{(N)})$, so the parameters in this sampling design are $\{\pi_i(Z^{(N)}) : 1 \leq i \leq N\}$. Note that we allow $\{\pi_i(Z^{(N)})\}$ to be unequal. Condition (A1) holds if $\pi_i(Z^{(N)}) \geq c$ for some constant $c > 0$. Since

$$\mathbb{E}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right)\right)^2 = \mathbb{E}_{(Y^{(N)}, Z^{(N)})}\left[\mathbb{E}_{\xi^{(N)}} \frac{1}{N} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right)^2\right] = \mathcal{O}(1),$$

condition (A2) is satisfied.

EXAMPLE 2.3 (Rejective sampling and high entropy sampling). A rejective sampling design $\mathfrak{r}$ maximizes the entropy functional $\mathfrak{p} \mapsto \sum_{s \in \mathcal{S}_N} \mathfrak{p}(s) \log(\mathfrak{p}(s))$ over all sampling designs of fixed size n with the constraint that the first-order inclusion probabilities equal $(\pi_1, \ldots, \pi_N)$ (cf. [38]). The parameters in this sampling design are n and $\{\pi_i(z^{(N)}) : 1 \leq i \leq N\}$. $\mathfrak{r}$ can also be realized as a conditional Bernoulli sampling design with appropriate success probabilities $(p_1, \ldots, p_N)$: for all $z^{(N)} \in \mathcal{Z}^{\otimes N}$ and $s \in \mathcal{S}_N$,

$$\mathfrak{r}\big(s, z^{(N)}\big) \propto \prod_{i \in s} p_i\big(z^{(N)}\big) \prod_{i \notin s} \big(1 - p_i\big(z^{(N)}\big)\big) \mathbf{1}_{|s|=n},$$

where $\sum_{i=1}^N p_i(z^{(N)}) = n$. The relationship between p_i and π_i is given in, for example, the statement and proof of Theorem 5.1 of [37].

Condition (A1) holds if $\pi_i(Z^{(N)}) \geq c$ for some constant $c > 0$. Let $d_N \equiv \sum_{i=1}^N \pi_i(z^{(N)}) \times (1 - \pi_i(z^{(N)}))$, and suppose that there exists some constant $K > 0$ such that for N large enough

$$\frac{N}{d_N} \leq K. \tag{2.2}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i}-1\right)\right)^2 \\
&\quad=\mathbb{E}_{Y^{(N)},Z^{(N)}}\left[\mathbb{E}_{\xi^{(N)}}\frac{1}{N}\left(\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i}-1\right)^2+\sum_{i\neq j}\left(\frac{\xi_i}{\pi_i}-1\right)\left(\frac{\xi_j}{\pi_j}-1\right)\right)\right] \\
&\quad\lesssim 1+\mathbb{E}_{Y^{(N)},Z^{(N)}}\left[N^{-1}\sum_{i\neq j}|\pi_{ij}-\pi_i\pi_j|\right]=\mathcal{O}(1),
\end{aligned}$$

where in the last inequality we used an old result due to Hajék (cf. Theorem 5.2 of [37]). Hence condition (A2) is satisfied under (2.2).

Assuming (for simplicity) that $0<\inf_i\pi_i\le\sup_i\pi_i<1$. Then Theorems 1 and 2 in [6] showed that high entropy designs satisfy a central limit theorem. More precisely, any sampling design $\mathfrak{p}$ with first-order inclusion probabilities $(\pi_1,\dots,\pi_N)$ and the property that

$$D_{\mathrm{KL}}(\mathfrak{p}\,\|\,\mathfrak{r})=\sum_{s\in\mathcal{S}_N}\mathfrak{p}(s)\log\frac{\mathfrak{p}(s)}{\mathfrak{r}(s)}\to 0$$

satisfies a CLT. An alternative argument can be found in the discussions after Proposition 3.4 below. In particular, all such high entropy designs satisfy conditions (A1)–(A2-CLT) under $0<\inf_i\pi_i\le\sup_i\pi_i<1$. The examples in this regard examined in [6] include Rao–Sampford sampling and successive sampling (under some scaling conditions).

EXAMPLE 2.4 (Stratified sampling). Suppose that U_N is partitioned into $\{U_N(1),\dots,U_N(k)\}$ according to the auxiliary variables $Z^{(N)}$ (we omit such dependence for simplicity). In other words, $\bigcup_{\ell=1}^{k}U_N(\ell)=U_N$, $U_N(\ell)\cap U_N(\ell')=\varnothing$ for $\ell\neq\ell'$ and $|U_N(\ell)|=N_\ell$ with $\sum_{\ell=1}^{k}N_\ell=N$. Let $n_1,\dots,n_k$ be such that $\sum_{\ell=1}^{k}n_\ell=n$. Within each stratum $U_N(\ell)$, we draw $n_\ell\le N_\ell$ samples s_ℓ without replacement. The overall sample is $s=\bigcup_{\ell=1}^{k}s_\ell$. The parameters in this sampling design are the partition $\{U_N(\ell):1\le\ell\le k\}$ and $\{n_\ell(Z^{(N)}):1\le\ell\le k\}$. Similar to the calculations in Example 2.1, since $\sum_{i\in s_\ell}\xi_i=n_\ell$, we have

$$\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i}-1\right)=\sum_{\ell=1}^{k}\left(\frac{1}{n_\ell/N_\ell}\sum_{i\in s_\ell}\xi_i\right)-N=\left(\sum_{\ell=1}^{k}N_\ell\right)-N=0.$$

Hence (A2) is satisfied. (A1) holds if $n_\ell/N_\ell\ge c$ for some constant $c>0$.

EXAMPLE 2.5 (Stratified sampling with overlap). Recently, [68] studied an interesting extension of the stratified sampling design as follows: suppose that $\{U_N(1),\dots,U_N(k)\}\subset U_N$ are k potentially overlapping "data sources" determined by the auxiliary variables $Z^{(N)}$, where k is a fixed integer. Let $N_\ell\equiv|U_N(\ell)|$. For each source $U_N(\ell)$, we draw $n_\ell\le N_\ell$ samples s_ℓ without replacement. The overall sample is $s=\bigcup_{\ell=1}^{k}s_\ell$, which may include duplicate samples due to the overlapping nature of the data sources. The parameters in this sampling design are the same as the above example. This sampling scheme is also known as multiple-frame surveys; cf. [42, 43, 51].

Let $\bar{\pi}_i^{(\ell)}\equiv n_\ell/N_\ell$ if $i\in U_N(\ell)$ be the sampling probability of unit i in the data source $U_N(\ell)$, and let $\bar{\xi}_i^{(\ell)}$ be the indicator of whether or not unit i is sampled in $U_N(\ell)$. Following [68], we consider the following variant of the Horvitz–Thompson empirical measure (or *Hartley empirical measure* as it is named in [68]):

$$\mathbb{P}_N^H(f)\equiv\frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{k}\frac{\bar{\xi}_i^{(\ell)}\rho_i^{(\ell)}}{\bar{\pi}_i^{(\ell)}}\mathbf{1}_{i\in U_N(\ell)}f(Y_i),$$

and the associated (Hartley) empirical process

$$\mathbb{G}_N^H(f) \equiv \sqrt{N}(\mathbb{P}_N^H - P)(f).$$

Here, the weights $\{\rho_i^{(\ell)} \equiv \rho_i^{(\ell)}(z^{(N)}) \in [0,1]\}$ are such that $\sum_{\ell=1}^k \rho_i^{(\ell)}(z^{(N)}) = 1$ and that $\rho_i^{(\ell)} = 0$ if $i \notin U_N(\ell)$. Now letting

$$\pi_i \equiv \prod_{\ell=1}^k \bar{\pi}_i^{(\ell)}, \qquad \xi_i \equiv \sum_{\ell=1}^k \Bigg(\mathbf{1}_{i \in U_N(\ell)} \bar{\xi}_i^{(\ell)} \rho_i^{(\ell)} \prod_{\ell' \neq \ell} \bar{\pi}_i^{(\ell')} \Bigg) \in [0,1], \tag{2.3}$$

we see that the Hartley empirical measure $\mathbb{P}_N^H$ and the associated empirical process $\mathbb{G}_N^H$ reduces to the Horvitz–Thompson empirical measure and empirical process with $\{\pi_i, \xi_i\}$ specified in (2.3).

Condition (A1) holds if $n_\ell / N_\ell \geq c$ for some constant $c > 0$ (by noting that k is a fixed constant that does not depend on $Z^{(N)}$). Now we verify (A2). Note that

$$\begin{aligned} \frac{1}{\sqrt{N}} \sum_{i=1}^N \Bigg(\frac{\xi_i}{\pi_i} - 1 \Bigg) &= \frac{1}{\sqrt{N}} \Bigg[\sum_{i=1}^N \sum_{\ell=1}^k \frac{\bar{\xi}_i^{(\ell)} \rho_i^{(\ell)}}{\bar{\pi}_i^{(\ell)}} \mathbf{1}_{i \in U_N(\ell)} - N \Bigg] \\ &= \sum_{\ell=1}^k \frac{1}{\sqrt{N}} \sum_{i=1}^N \Bigg(\frac{\bar{\xi}_i^{(\ell)}}{\bar{\pi}_i^{(\ell)}} - 1 \Bigg) \rho_i^{(\ell)} \mathbf{1}_{i \in U_N(\ell)} = \mathcal{O}_{\mathbb{P}}(1), \end{aligned}$$

where the last line follows by computing the second moment:

$$\begin{aligned} &\mathbb{E} \Bigg[\frac{1}{\sqrt{N}} \sum_{i=1}^N \Bigg(\frac{\bar{\xi}_i^{(\ell)}}{\bar{\pi}_i^{(\ell)}} - 1 \Bigg) \rho_i^{(\ell)} \mathbf{1}_{i \in U_N(\ell)} \Bigg]^2 \\ &\lesssim 1 + \frac{1}{N} \sum_{i \neq j \in U_N(\ell)} \mathbb{E}_{(Y^{(N)}, Z^{(N)})} \Bigg[\Bigg| \mathbb{E}_{\xi^{(N)}} \Bigg(\frac{\bar{\xi}_i^{(\ell)}}{\bar{\pi}_i^{(\ell)}} - 1 \Bigg) \Bigg(\frac{\bar{\xi}_j^{(\ell)}}{\bar{\pi}_j^{(\ell)}} - 1 \Bigg) \Bigg| \Bigg] = \mathcal{O}(1). \end{aligned}$$

This verifies (A2-CLT).

From the above derivation, it is easy to see that (A1)–(A2-CLT) hold with the sampling without replacement design replaced by Bernoulli sampling and rejective sampling designs.

We also note that different choices of the weights $\{\rho_i^{(\ell)} \equiv \rho_i^{(\ell)}(z^{(N)}) \in [0,1]\}$ lead to different asymptotic variances. Since this issue is not the main concern of this paper, we refer the readers to [68] for the optimal choice of weights in the context of Bernoulli sampling and sampling without replacement designs.

3. Theory. In this section, we will be mainly interested in the global and local behavior of the Horvitz–Thompson empirical process. In particular, we prove a Glivenko–Cantelli theorem and a Donsker theorem that provide global information concerning the Horvitz–Thompson empirical process in the limit. As will be seen, our formulation requires almost minimal conditions. We further study local behavior of the Horvitz–Thompson empirical process by characterizing its local asymptotic modulus and several ratio limit theorems. Understanding the local behavior of the Horvitz–Thompson empirical process plays a key role in applications to statistical problems as will be demonstrated in Section 4. Corresponding results for the calibrated version of the Horvitz–Thompson empirical process are also included. We also discuss uniform limit theorems for some variants of the Horvitz–Thompson empirical process and their conditional versions thereof. Finally, we present some positive and negative results on CLTs when the condition (A1) fails.

3.1. *Global and local limit theorems.* First we study the Glivenko–Cantelli theorem. We say that $\mathcal{F}$ is P-Glivenko–Cantelli if and only if $\sup_{f\in\mathcal{F}}|(\mathbb{P}_N - P)(f)| = \mathfrak{o}_{\mathbb{P}}(1)$.

THEOREM 3.1 (Glivenko–Cantelli theorem). *Suppose that* (A1) *and* (A2-*LLN*) *hold. If* $\mathcal{F}$ *is* P*-Glivenko–Cantelli with* $PF < \infty$ *for some measurable envelope* F*, then*

$$\sup_{f\in\mathcal{F}}\bigl|\bigl(\mathbb{P}_N^{\pi} - P\bigr)(f)\bigr| = \mathfrak{o}_{\mathbb{P}}(1).$$

Recall the notion of weak convergence in the Hoffmann–Jørgensen sense: Let $\{X(f)\}_{f\in\mathcal{F}}$ be a bounded process whose finite-dimensional laws correspond to the finite dimensional projections of a tight Borel law on $\ell^\infty(\mathcal{F})$. Let $\{X_N(f)\}_{f\in\mathcal{F}}$ be bounded processes. We say that $X_N \rightsquigarrow X$ in $\ell^\infty(\mathcal{F})$ if and only if $\mathbb{E}^*H(X_N) \to \mathbb{E}H(\tilde{X})$ for all $H \in C_b(\ell^\infty(\mathcal{F}))$, where $C_b(\ell^\infty(\mathcal{F}))$ denotes all bounded continuous functions on $\ell^\infty(\mathcal{F})$, and $\tilde{X}$ is a measurable version of X with separable range (so $H(\tilde{X})$ is measurable). Equivalently, $d_{\mathrm{BL}}(X_N, \tilde{X}) \to 0$, where d_{BL} is the dual bounded Lipschitz metric (cf. p. 246 of [35]). It is also well known that $X_N \rightsquigarrow X$ in $\ell^\infty(\mathcal{F})$ if and only if X_N converges to X finite-dimensionally, and there exists a pseudo-metric d on $\mathcal{F}$ such that for any $\delta_N \to 0$,

$$\sup_{d(f,g)\le\delta_N}\bigl|X_N(f) - X_N(g)\bigr| = \mathfrak{o}_{\mathbb{P}}(1).$$

We refer the readers to [35, 81] for more details. We say that $\mathcal{F}$ is P-Donsker if and only if $\mathbb{G}_N \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ where $\mathbb{G}$ is a P-Brownian bridge process.

THEOREM 3.2 (Donsker theorem). *Suppose that* (A1) *and* (A2-*CLT*) *hold. Further assume that*:

(D1) $\mathbb{G}_N^{\pi}$ *converges finite-dimensionally to a Gaussian process* $\mathbb{G}^{\pi}$.
(D2) $\mathcal{F}$ *is* P*-Donsker.*

Then $\mathbb{G}^{\pi}$ *admits a tight measurable version in* $\ell^\infty(\mathcal{F})$ *for which, using the same notation,*

$$\mathbb{G}_N^{\pi} \rightsquigarrow \mathbb{G}^{\pi} \quad \text{in } \ell^\infty(\mathcal{F}).$$

Clearly, the finite-dimensional convergence condition (D1) above is necessary for a uniform central limit theorem in $\ell^\infty(\mathcal{F})$. (D2) is also minimal. One intriguing feature of Theorem 3.2 is that a uniform central limit theorem follows essentially automatically as long as the *finite-dimensional convergence property of the Horvitz–Thompson empirical process is verified*. A similar phenomenon was also observed in [72] in a univariate non-i.i.d. case.

REMARK 3.3. $\mathcal{F}$ is assumed to be countable for simplicity for Theorems 3.1 and 3.2. For the general uncountable case, we may use outer probability for the statement and proofs of these theorems.

Although being necessary, establishing a finite-dimensional CLT for $\mathbb{G}_N^{\pi}$ and identifying the covariance structure of $\mathbb{G}^{\pi}$ can be a nontrivial problem for general sampling designs; see, for example, [5, 6, 23, 30, 37, 61–64, 82]. Below we exploit one possible strategy, inspired by [11], for identifying the covariance structure of $\mathbb{G}^{\pi}$.

PROPOSITION 3.4. *Suppose* (A1) *and the following conditions hold*:

(F1) *There exists* $q \in [4, \infty]$ *such that for any i.i.d. random variables* $\{V_i\}$ *defined on* $(\mathcal{X}, \mathcal{A}, P_{(Y,Z)})$ *with* $\|V_1\|_{L_q(P_{(Y,Z)})} < \infty$,

$$\frac{1}{S_N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}V_i - \frac{1}{N}\sum_{i=1}^{N}V_i\right) \rightsquigarrow \mathcal{N}(0,1)$$

holds under $\mathbb{P}_d(\cdot, \omega)$ (*notation defined in* (2.1)) *for* $P_{(Y,Z)}$*-a.s.* $\omega \in \mathcal{X}$. *Here,* S_N *is the design-based variance given by*

$$S_N^2 \equiv \frac{1}{N^2}\sum_{1\le i,j\le N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}V_iV_j.$$

(F2) *The* (*essentially*) *first-order inclusion probabilities satisfy*

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\pi_{ii} - \pi_i^2}{\pi_i^2} \to_{P_{(Y,Z)}} \mu_{\pi 1},$$

for some nonrandom $\mu_{\pi 1} \in \mathbb{R}$.

(F3) *The second-order inclusion probabilities satisfy*

$$\sup_{N\in\mathbb{N}}\sup_{1\le i\ne j\le N} N|\pi_{ij} - \pi_i\pi_j| \le K, \qquad \frac{1}{N}\sum_{i\ne j}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} \to_{P_{(Y,Z)}} \mu_{\pi 2},$$

where $K > 0$ *is an absolute constant, and* $\mu_{\pi 2} \in \mathbb{R}$ *is nonrandom.*

If $\mathcal{F}$ *is such that* $\|F\|_{L_q(P)} < \infty$ *for* $q \in [4, \infty]$ *that verifies* (F1), *then* $\mathbb{G}_N^\pi$ *converges finite-dimensionally to a Gaussian process* $\mathbb{G}^\pi$ *whose covariance structure is given by the following*: *for any* $f, g \in \mathcal{F}$,

$$\begin{aligned}\operatorname{Cov}\big(\mathbb{G}^\pi(f), \mathbb{G}^\pi(g)\big) &= (1+\mu_{\pi 1})P(fg) - (1-\mu_{\pi 2})(Pf)(Pg)\\ &= P(fg) - (Pf)(Pg) + \mu_{\pi 1}P(fg) + \mu_{\pi 2}(Pf)(Pg).\end{aligned}$$

The above covariance formula can be inferred from the decomposition

$$\mathbb{G}_N^\pi = \sqrt{N}\big(\mathbb{P}_N^\pi - P\big) = \sqrt{N}(\mathbb{P}_N - P) + \sqrt{N}\big(\mathbb{P}_N^\pi - \mathbb{P}_N\big),$$

where the covariance structure of the second term $\sqrt{N}(\mathbb{P}_N^\pi - \mathbb{P}_N)$ can be deduced from conditions (F1)–(F3). These conditions are also used in [11]: (F1) corresponds to (HT1) in [11], (F2) corresponds to condition (i) in Proposition 3.1 in [11], and (F3) corresponds to (C2) and condition (ii) in Proposition 3.1 in [11]. Combined with Proposition 3.4, we see that Theorem 3.2 extends Proposition 3.2 of [11] in at least the following directions: (i) we work with a general P-Donsker class $\mathcal{F}$ with $\|F\|_{L_q(P)} < \infty$ instead of one particular class $\{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$, and (ii) we weaken conditions for the sampling designs, that is, (C3)–(C4) in [11] are no longer required. We should, however, remind readers that Proposition 3.4 is not exhaustive for identifying the covariance structure of $\mathbb{G}^\pi$ and, therefore, it is possible that the current conditions in Proposition 3.4 can be further weakened via other approaches.

The conditions in Proposition 3.4 are verified in [11] under a slightly different setting, but for the convenience of the reader, we provide some details for various sampling designs. Below we take $q = 4$; see also Table 1 for a summary.

- For sampling without replacement, $\pi_{ii} = \pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$ for $i \ne j$. If $n/N \to \lambda \in (0,1)$, (F1) can be verified using Hajék's rank central limit theorem (cf. [36], or Proposition A.5.3 of [81]), and (F2)–(F3) are satisfied with $\mu_{\pi 1} = \lambda^{-1} - 1$ and $\mu_{\pi 2} = 1 - \lambda^{-1}$. The cases for stratified sampling with/without overlaps can be considered analogously.

TABLE 1
Values of $\mu_{\pi 1}, \mu_{\pi 2}$ for different sampling designs. Here, $\lambda = \lim_N n/N$, $A = \lim_N N^{-1}\sum_{i=1}^N \pi_i^{-1}$, $d = \lim_N N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)$

	SWOR	Bernoulli	Rejective
$\mu_{\pi 1}$	$\lambda^{-1}-1$	$A-1$	$A-1$
$\mu_{\pi 2}$	$1-\lambda^{-1}$	0	$-d^{-1}(1-\lambda)^2$

- For Bernoulli sampling, $\pi_{ii} = \pi_i$ and $\pi_{ij} = \pi_i\pi_j$ for $i \neq j$. If $\{\pi_i\}_{i=1}^N \subset [\varepsilon, 1-\varepsilon](\varepsilon > 0)$, (F1) can be verified using the Lindeberg–Feller central limit theorem, and (F2)–(F3) are satisfied with $\mu_{\pi 1} = \lim_N N^{-1}\sum_{i=1}^N(\pi_i^{-1}-1)$ and $\mu_{\pi 2} = 0$.
- For rejective sampling with first-order inclusion probabilities $\{\pi_i\}_{i=1}^N \subset [\varepsilon, 1-\varepsilon](\varepsilon > 0)$, let $d_N = \sum_{i=1}^N \pi_i(1-\pi_i)$. (F1) can be verified by Theorem 1 of [6]. Using Theorem 1 of [10], (F2)–(F3) are satisfied with $\mu_{\pi 1} = \lim_N N^{-1}\sum_{i=1}^N(\pi_i^{-1}-1)$ and

$$\mu_{\pi 2} = \lim_N\left[-\frac{1}{N}\sum_{i\neq j}\frac{(1-\pi_i)(1-\pi_j)}{d_N} + \mathcal{O}\big(Nd_N^{-2}\big)\right] = -d^{-1}(1-\lambda)^2,$$

  provided $n/N \to \lambda \in (0,1)$ and $d_N/N \to d$. The covariance structure of $\mathbb{G}^\pi$ with high entropy sampling designs is the same as the rejective sampling design, which can be verified using the same arguments in pages 1754–1755 of [11].

Hence, under the assumptions of Proposition 3.4, the covariance formula for $\mathbb{G}^\pi$ can be written more explicitly: for any $f, g \in \mathcal{F}$,

$$\operatorname{Cov}\big(\mathbb{G}^\pi(f), \mathbb{G}^\pi(g)\big) = \begin{cases} \lambda^{-1}\big(P(fg) - (Pf)(Pg)\big) & \text{under SWOR,} \\ A\cdot P(fg) - (Pf)(Pg) & \text{under Bernoulli,} \\ A\cdot P(fg) - \big[1 + d^{-1}(1-\lambda)^2\big](Pf)(Pg) & \text{under Rejective.} \end{cases}$$

Here, $\lambda = \lim_N n/N$, $A = \lim_N N^{-1}\sum_{i=1}^N \pi_i^{-1}$, $d = \lim_N N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)$ (the convergence is all in probability sense).

Our next goal is to study the local behavior of the Horvitz–Thompson empirical process. Although being of crucial importance in applications to semi/nonparametric statistics, to the best knowledge of the authors, this issue has not been addressed in the literature.

We first study *local asymptotic modulus* of the Horvitz–Thompson empirical process, which has been considered historically for VC-type classes of sets and function classes in [3, 33, 34] in the context of usual empirical processes. As will be clear below, one of the strengths of the formulation of our theorems is that finite-dimensional convergence of $\mathbb{G}_N^\pi$ is not required for studying the local behavior of $\mathbb{G}_N^\pi$—we only require that the weights have a $\sqrt{N}$ convergence rate as in (A2-CLT).

Before formally stating the results on the local behavior of the Horvitz–Thompson empirical process, we need some definitions.

DEFINITION 3.5. A *local asymptotic modulus* of the Horvitz–Thompson empirical process indexed by a class of functions $\mathcal{F}$ is an increasing function $\phi(\cdot)$ for which there exist

some $r_N \ll \delta_N \le 1/2$, both nonincreasing with $N \mapsto \sqrt{N}\delta_N$ nondecreasing, such that

$$\sup_{f\in\mathcal{F}: r_N^2 < Pf^2 \le \delta_N^2} \frac{|\mathbb{G}_N^\pi(f)|}{\phi(\sigma_P f)} = \mathcal{O}_{\mathbb{P}}(1). \tag{3.1}$$

Here, $\sigma_P^2(f) = \mathrm{Var}_P(f)$.

DEFINITION 3.6. We say that $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$ if either

$$\sup_Q \log \mathcal{N}\big(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)\big) \lesssim \varepsilon^{-\alpha},$$

where the supremum is over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$; or

$$\log \mathcal{N}_{[\,]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \lesssim \varepsilon^{-\alpha}.$$

The entropy condition is well understood in the literature; we only refer the readers to [35, 77, 81] for various examples in this regard.

THEOREM 3.7. *Suppose that* (A1) *and* (A2-CLT) *hold and* $\mathcal{F}$ *is a uniformly bounded class satisfying an entropy condition with exponent* $\alpha \in (0,2)$. *Then* $\omega_\alpha(t) = t^{1-\frac{\alpha}{2}}$ *is a local asymptotic modulus for the Horvitz–Thompson empirical process indexed by* $\mathcal{F}$, *that is*, (3.1) *holds with* $\phi = \omega_\alpha$.

The local asymptotic modulus is a key step in understanding the behavior of the Horvitz–Thompson empirical process at a local level. This will be useful in applications in the next section. The local asymptotic modulus ω_α cannot be improved in general; this can be shown for the usual empirical process indexed by *α-full class* (which essentially requires a lower bound for the entropy number in a more local sense; cf. [33]).

One may also invert the above viewpoint by fixing one particular weight function ϕ and asking for the rate of convergence of the corresponding weighted Horvitz–Thompson empirical process. Below are two particular choices: the first one (3.2) uses $\phi(x) = x$, and the second one (3.3) uses (essentially) $\phi(x) = x^2$.

THEOREM 3.8. *Suppose that* (A1) *and* (A2-CLT) *hold and* $\mathcal{F}$ *is a uniformly bounded class satisfying an entropy condition with exponent* $\alpha \in (0,2)$. *Let* $r_N \gtrsim N^{-1/(\alpha+2)}$. *Then*

$$N^{1/(\alpha+2)} \sup_{f\in\mathcal{F}: \sigma_P f \ge r_N} \frac{|(\mathbb{P}_N^\pi - P)(f)|}{\sigma_P f} = \mathcal{O}_{\mathbb{P}}(1). \tag{3.2}$$

If furthermore $\mathcal{F}$ *takes value in* $[0,1]$, *then for any* $L_N \to \infty$,

$$\sup_{f\in\mathcal{F}: Pf \ge L_N \cdot r_N} \left| \frac{\mathbb{P}_N^\pi f}{Pf} - 1 \right| = \mathfrak{o}_{\mathbb{P}}(1). \tag{3.3}$$

Results analogous to (3.2)–(3.3) have been derived in the case of i.i.d. sampling in [55, 73–75, 83] for uniform empirical processes on (subsets of) $\mathbb{R}$ (or $\mathbb{R}^d$), and are further investigated in [3] for VC classes of sets, and extended by [33, 34] who studied more general VC-subgraph classes.

Note that (3.3) can be viewed as a uniform law of large numbers for the weighted Horvitz–Thompson empirical process. We can also establish a central limit theorem for the weighted Horvitz–Thompson empirical process, analogous to the development in [1–3, 33] for the usual empirical process.

THEOREM 3.9. *Suppose that* (A1) *and* (A2-CLT) *hold, and that* $\mathcal{F}$ *is a uniformly bounded class satisfying an entropy condition with exponent* $\alpha \in (0,2)$. *Let* $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *be such that* $\phi(0) = 0$ *and that*

$$\frac{\phi(t)}{t^{1-\frac{\alpha}{2}}(\log\log(1/t))^{1/2}} \to \infty \tag{3.4}$$

as $t \to 0$. *If* $r_N \gtrsim N^{-1/(\alpha+2)}$ *and* $\mathbb{G}_N^\pi$ *converges finite-dimensionally to a Gaussian process* $\mathbb{G}^\pi$, *then* $\mathbb{G}^\pi$ *admits a tight measurable version in* $\ell^\infty(\mathcal{F})$ *for which, using the same notation,*

$$\frac{\mathbb{G}_N^\pi(f)}{\phi(\sigma_P f)}\mathbf{1}_{\sigma_P f > r_N} \rightsquigarrow \frac{\mathbb{G}^\pi(f)}{\phi(\sigma_P f)} \quad \text{in } \ell^\infty(\mathcal{F}).$$

The weight function in the above theorem is required to be only slightly stronger than the local asymptotic modulus by an iterated logarithmic factor. This is very natural: the weight function cannot beat the local asymptotic modulus for a weighted CLT to hold, so the condition (3.4) is optimal up to an iterated logarithmic factor.

REMARK 3.10. The countability assumption on $\mathcal{F}$ in Theorems 3.7–3.9 is used at a technical level via the one-sided Talagrand-type concentration inequality (cf. Proposition C.2). One may assume, for instance, pointwise measurability (cf. Example 2.3.4 of [81]) for $\mathcal{F}$ to handle the uncountable class.

3.2. *Calibration*. In practice, since the Horvitz–Thompson estimator may be severely inefficient, calibration of the weights is often used to improve efficiency [28, 52]. The main purpose of this section, instead of proposing new calibration methods or addressing efficiency issues, rests in demonstrating that our theoretical results are still valid for the Horvitz–Thompson empirical process with calibrated weights.

To illustrate this, we consider one popular calibration method that aims at matching the population mean for the Horvitz–Thompson estimator [28]. Let $\mathcal{Z} \subset \mathbb{R}^d$ be a compact set, and $G : \mathbb{R} \to \mathbb{R}_{\geq 0}$. Let $\hat{\alpha}_N \in \mathcal{A}_c$, where $\mathcal{A}_c$ is a compact set of $\mathbb{R}^d$, be defined via

$$\frac{1}{N}\sum_{i=1}^N \frac{\xi_i G(Z_i^\top \hat{\alpha}_N)}{\pi_i} Z_i = \frac{1}{N}\sum_{i=1}^N Z_i.$$

Then the *calibrated Horvitz–Thompson empirical measure* and *calibrated Horvitz–Thompson empirical process* are defined by

$$\mathbb{P}_N^{\pi,c}(f) \equiv \frac{1}{N}\sum_{i=1}^N \frac{\xi_i G(Z_i^\top \hat{\alpha}_N)}{\pi_i} f(Y_i), \quad f \in \mathcal{F},$$

and

$$\mathbb{G}_N^{\pi,c}(f) \equiv \sqrt{N}(\mathbb{P}_N^{\pi,c} - P)(f), \quad f \in \mathcal{F},$$

respectively.

Our next theorem asserts that as long as $\hat{\alpha}_N$ converges to the "truth" 0 (which can be defined to be another value, but we use 0 for notational convenience) sufficiently fast, the global and local theorems also hold for the calibrated Horvitz–Thompson empirical process.

THEOREM 3.11. *Suppose* $G(0) = 1$, $G'(0) > 0$. *Let* $\mathcal{F}$ *be a class of measurable functions with a measurable envelope* F.

(1) *Let the assumptions in Theorem 3.1 hold with* $PF < \infty$. *If* $\hat{\alpha}_N = \mathfrak{o}_{\mathbb{P}}(1)$, *then the conclusion of Theorem 3.1 holds with* $\mathbb{P}_N^\pi$ *replaced by* $\mathbb{P}_N^{\pi,c}$.

(2) *Let the assumptions in Theorem 3.2 hold with* $PF^2 < \infty$ (*but the finite-dimensional convergence condition is replaced by* $\mathbb{G}_N^{\pi,c}$ *converges finite-dimensionally to some Gaussian process* $\mathbb{G}^{\pi,c}$). *If* $\sqrt{N}\hat{\alpha}_N = \mathcal{O}_{\mathbb{P}}(1)$, *then* $\mathbb{G}^{\pi,c}$ *admits a tight measurable version in* $\ell^\infty(\mathcal{F})$ *for which, using the same notation,*

$$\mathbb{G}_N^{\pi,c} \rightsquigarrow \mathbb{G}^{\pi,c} \quad \text{in } \ell^\infty(\mathcal{F}).$$

(3) *If* $\sqrt{N}\hat{\alpha}_N = \mathcal{O}_{\mathbb{P}}(1)$, *then under the same conditions as in Theorems 3.7, 3.8 and 3.9* (*but the finite-dimensional convergence condition is replaced by* $\mathbb{G}_N^{\pi,c}$ *converges finite-dimensionally to some Gaussian process* $\mathbb{G}^{\pi,c}$), *the respective conclusions hold for the calibrated Horvitz–Thompson empirical process.*

The structural commonality in the above theorem is characterized by the $\sqrt{N}$-rate of the estimate $\hat{\alpha}_N$. Establishing a $\sqrt{N}$-rate for $\hat{\alpha}_N$ is not hard: in fact we can use Theorem 3.3.1 of [81] for such a purpose by verifying the asymptotic equicontinuity of the Horvitz–Thompson empirical process.

Below we exploit one possible strategy for this via the method of Proposition 3.4.

PROPOSITION 3.12. *Assume the conditions of Proposition 3.4 and Theorem 3.11 hold, and that* $\pi_i \equiv \pi_i(Z_i)$ *for* $i = 1, \dots, N$. *Further assume that* G *is continuous with its derivative* G' *continuous in a neighborhood of* 0, *and the map* $\alpha \mapsto P_Z[G(Z^\top\alpha - 1)Z]$ *has a unique zero at* 0, *and* $P_Z(ZZ^\top) \in \mathbb{R}^{d\times d}$ *is invertible. Then*

$$\sqrt{N}\hat{\alpha}_N = -\big(G'(0)\big)^{-1}\big(P_Z\big(ZZ^\top\big)\big)^{-1}\big(\mathbb{G}_N^{\pi} - \mathbb{G}_N\big)Z + \mathfrak{o}_{\mathbb{P}}(1). \tag{3.5}$$

Furthermore, $\mathbb{G}_N^{\pi,c}$ *converges finite-dimensionally to a tight Gaussian process* $\mathbb{G}^{\pi,c}$ *whose covariance structure is given by the following: for any* $f, g \in \mathcal{F}$,

$$\begin{aligned}&\operatorname{Cov}\big(\mathbb{G}^{\pi,c}(f), \mathbb{G}^{\pi,c}(g)\big)\\ &\quad = P(fg) - (Pf)(Pg) + \mu_{\pi 1}P_{(Y,Z)}\big(\mathcal{T}(f)\mathcal{T}(g)\big) + \mu_{\pi 2}\big(P_{(Y,Z)}\mathcal{T}(f)\big)\big(P_{(Y,Z)}\mathcal{T}(g)\big).\end{aligned}$$

Here, the operator $\mathcal{T} : \mathbb{R}^{\mathcal{Y}\times\mathcal{Z}} \to \mathbb{R}^{\mathcal{Y}\times\mathcal{Z}}$ *is defined by*

$$\mathcal{T}(f)(y,z) = f(y) - P_{(Y,Z)}\big((\xi/\pi)f(Y)Z^\top\big)\big(P_Z\big(ZZ^\top\big)\big)^{-1}z.$$

As we will see in the proofs, the asymptotic expansion for $\sqrt{N}\hat{\alpha}_N$ in (3.5) plays a crucial role in identifying the covariance structure of $\mathbb{G}^{\pi,c}$. Although here we only study one particular calibration method that matches the population mean, other calibration methods are also possible. Typically different calibration methods only differ in terms of the exact form of the corresponding operators $\mathcal{T}$; see, for example, [69] for various calibration methods under the (two-phase) stratified sampling design.

3.3. *Other variants.* Our global limit theorems in Theorems 3.1 and 3.2 can be used for several other variants of the Horvitz–Thompson empirical processes studied in [11]. We illustrate this by considering Donsker theorems for the variants as detailed below.

First, consider $\sqrt{n}(\mathbb{P}_N^{\pi} - \mathbb{P}_N)$. We have the following.

COROLLARY 3.13. *Suppose that* (A1) *and* (A2-CLT) *hold, and that* $\mathcal{F}$ *is* P*-Donsker. Further suppose that the conditions in Proposition 3.4 hold, and that* $n/N \to \lambda \in (0,1)$.

Then $\sqrt{n}(\mathbb{P}_N^{\pi} - \mathbb{P}_N)$ *converges weakly in* $\ell^\infty(\mathcal{F})$ *to a Gaussian process* $\bar{\mathbb{G}}^{\pi}$ *whose covariance structure is given by the following*: *for any* $f, g \in \mathcal{F}$,

$$\begin{aligned}\operatorname{Cov}\big(\bar{\mathbb{G}}^{\pi}(f), \bar{\mathbb{G}}^{\pi}(g)\big) &= \lambda\big(\mu_{\pi 1} P(fg) + \mu_{\pi 2}(Pf)(Pg)\big) \\ &= \begin{cases} (1-\lambda)\big(P(fg) - (Pf)(Pg)\big) & \text{under SWOR}, \\ \lambda(A-1)\cdot P(fg) & \text{under Bernoulli}, \\ \lambda\big((A-1)\cdot P(fg) - d^{-1}(1-\lambda)^2(Pf)(Pg)\big) & \text{under Rejective}. \end{cases}\end{aligned}$$

Here, $\lambda = \lim_N n/N$, $A = \lim_N N^{-1}\sum_{i=1}^N \pi_i^{-1}$, $d = \lim_N N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)$.

The covariance formula above is a direct consequence of the assumptions in Proposition 3.4. Furthermore, the above corollary extends Theorem 3.1 of [11] from the one-dimensional case $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$ to a general setting.

Next, consider the Hájek empirical process. Let

$$\mathbb{P}_N^{\pi,H}(f) \equiv \frac{1}{\hat{N}} \sum_{i=1}^N \frac{\xi_i}{\pi_i} f(Y_i), \quad \hat{N} \equiv \sum_{i=1}^N \frac{\xi_i}{\pi_i}$$

be the Hájek empirical measure. We have the following.

COROLLARY 3.14. *Suppose that* (A1) *and* (A2-CLT) *hold, and that* $\mathcal{F}$ *is* P*-Donsker. Further suppose that the conditions in Proposition* 3.4 *hold, and that* $n/N \to \lambda \in (0,1)$. *Then* $\sqrt{n}(\mathbb{P}_N^{\pi,H} - \mathbb{P}_N)$ *converges weakly to a Gaussian process* $\bar{\mathbb{G}}^{\pi,H}$ *whose covariance structure is given by the following*: *for any* $f, g \in \mathcal{F}$,

$$\begin{aligned}\operatorname{Cov}\big(\bar{\mathbb{G}}^{\pi,H}(f), \bar{\mathbb{G}}^{\pi,H}(g)\big) &= \lambda\mu_{\pi 1}\big(P(fg) - (Pf)(Pg)\big) \\ &= \begin{cases} (1-\lambda)\big(P(fg) - (Pf)(Pg)\big) & \text{under SWOR}, \\ \lambda(A-1)\cdot\big(P(fg) - (Pf)(Pg)\big) & \text{under Bernoulli}, \\ \lambda(A-1)\cdot\big(P(fg) - (Pf)(Pg)\big) & \text{under Rejective}. \end{cases}\end{aligned}$$

Here, $\lambda = \lim_N n/N$, $A = \lim_N N^{-1}\sum_{i=1}^N \pi_i^{-1}$.

As we will see in the proofs, the covariance structure of the limit of $\sqrt{n}(\mathbb{P}_N^{\pi,H} - \mathbb{P}_N)$ is the same as that of

$$f \mapsto \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right)\big(f(Y_i) - Pf\big)$$

up to a factor of λ, which can be determined by the conditions of Proposition 3.4. Furthermore, the above corollary extends Theorem 4.2 of [11], again from the one-dimensional case to a general setting.

REMARK 3.15. Under (F3), since the harmonic mean is less than the arithmetic mean, we have $A^{-1} = \lim_N (N^{-1}\sum_{i=1}^N \pi_i^{-1})^{-1} \le \lim_N (N^{-1}\sum_{i=1}^N \pi_i) = \lim_N \frac{n}{N} = \lambda$, where the next to last equality follows by computing the second moment and using (F3). It then follows that $\lambda(A-1) \ge 1-\lambda$ under (F3).

3.4. *Conditional limit theorems.* In this section, we consider conditional versions of the (global) uniform limit theorems. For clarity of presentation, following [24] and [84], we introduce the following notion.

DEFINITION 3.16. Let $\{\Delta_N\}_{N\in\mathbb{N}}$ be a sequence of random variables defined on $(\mathcal{S}_N \times \mathcal{X}, \sigma(\mathcal{S}_N)\times\mathcal{A}, \mathbb{P})$. We say that Δ_N is of order $\mathfrak{o}_{\mathbb{P}_d}(1)$ in $P_{(Y,Z)}$-probability if for any $\varepsilon, \delta > 0$, we have $P_{(Y,Z)}(\mathbb{P}_{d|(Y,Z)}(|\Delta_N| > \varepsilon) > \delta) \to 0$ as $N \to \infty$.

Below we establish conditional versions of Glivenko–Cantelli and Donsker theorems for $\mathbb{P}_N^\pi - \mathbb{P}_N$.

COROLLARY 3.17 (Conditional Glivenko–Cantelli theorem). *Suppose that* (A1) *and* (A2-*LLN*) *hold. If* $\mathcal{F}$ *is* P*-Glivenko–Cantelli, then*

$$\sup_{f\in\mathcal{F}} \left|\left(\mathbb{P}_N^\pi - \mathbb{P}_N\right)(f)\right| = \mathfrak{o}_{\mathbb{P}_d}(1) \quad \textit{in } P_{(Y,Z)}\textit{-probability}.$$

COROLLARY 3.18 (Conditional Donsker theorem). *Suppose that* (A1) *and* (A2-*CLT*) *hold, and that* $\mathcal{F}$ *is* P*-Donsker. Further suppose that the conditions in Proposition* 3.4 *hold, and that* $n/N \to \lambda \in (0,1)$. *Then*

$$\sqrt{n}\left(\mathbb{P}_N^\pi - \mathbb{P}_N\right) \rightsquigarrow \bar{\mathbb{G}}^\pi \quad \textit{in } \ell^\infty(\mathcal{F}) \textit{ in } P_{(Y,Z)}\textit{-probability}.$$

Here, $\bar{\mathbb{G}}^\pi$ *is a Gaussian process whose covariance structure is given in Corollary* 3.13.

The precise meaning of the above conditional Donsker theorem is that $d_{\mathrm{BL},d}(\sqrt{n}(\mathbb{P}_N^\pi - \mathbb{P}_N), \bar{\mathbb{G}}^\pi) \equiv \sup_{H\in \mathrm{BL}_1(\ell^\infty(\mathcal{F}))} |\mathbb{E}^*_{d|(Y,Z)} H(\sqrt{n}(\mathbb{P}_N^\pi - \mathbb{P}_N)) - \mathbb{E}H(\bar{\mathbb{G}}^\pi)| \to 0$ in $P_{(Y,Z)}$-probability.

3.5. *Positive and negative results for CLTs when* (A1) *fails.* Let

$$Z_N \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right) = \mathbb{G}_N^\pi(1).$$

We present a negative result concerning CLTs for Z_N when (A1) fails.

PROPOSITION 3.19. *Let* e_N *be such that* $e_N \searrow 0$ *and* $Ne_N \nearrow \infty$. *There exists a Bernoulli sampling scheme with equal first-order including probabilities satisfying* $\min_{1\le i\le N} \pi_i = e_N$, *such that a CLT fails for* Z_N.

PROOF. Let the ξ_i's be i.i.d. Bern(e_N) random variables independent of Y_i's. Then $\pi_i = e_N$ for $1 \le i \le N$. First, note that

$$\mathrm{Var}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right)\right] = \frac{1}{N}\sum_{i=1}^N \pi_i^{-2}\mathrm{Var}(\xi_i) = \frac{1-e_N}{e_N}.$$

Hence for $Z \sim \mathcal{N}(0,1)$, we have by Liapunov's CLT and uniform integrability

$$\mathbb{E}\left| e_N^{1/2} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right)\right| \to \mathbb{E}|Z|.$$

This shows that Z_N is not bounded in probability, and hence the CLT fails. □

Therefore, when (A1) fails, the conclusion of Theorem 3.2 with $\sqrt{N}$ normalization does not hold without further conditions.

It is easy to note from the proof above that the problem can be fixed if we change normalization from $\sqrt{N}$ to $\sqrt{n}$. Specifically, if ξ_i's are i.i.d. Bern(e_N) with $e_N \searrow 0$, $Ne_N \nearrow \infty$, then

$$\sqrt{\frac{n}{N}} Z_N = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi_i} - 1 \right) = \sqrt{\frac{n}{N}} \mathbb{G}_N^{\pi}(1) \quad \text{or} \quad e_N^{1/2} \mathbb{G}_N^{\pi}(1)$$

converges to a normal random variable. This phenomenon can be generalized much further to a uniform central limit theorem as follows.

PROPOSITION 3.20. *Let $\{e_N\} \subset (0, 1]$ be such that $Ne_N \nearrow \infty$. Suppose that the first-order probabilities are equal in that $\pi_1 = \cdots = \pi_N = e_N$ with sampling indicators ξ_i's independent from Y_i's, and that $\frac{1}{\sqrt{Ne_N}} \sum_{i=1}^{N} (\xi_i - e_N) = \mathcal{O}_{\mathbb{P}}(1)$. Further assume that*:

1. *$e_N^{1/2} \mathbb{G}_N^{\pi}$ converges finite-dimensionally to a Gaussian process $\mathbb{G}_0^{\pi}$.*
2. *$\mathcal{F}$ is P-Donsker.*

Then $\mathbb{G}_0^{\pi}$ admits a tight measurable version in $\ell^{\infty}(\mathcal{F})$ for which, using the same notation,

$$e_N^{1/2} \mathbb{G}_N^{\pi} \rightsquigarrow \mathbb{G}_0^{\pi} \quad \text{in } \ell^{\infty}(\mathcal{F}).$$

One may wonder to what extent the idea above can be further generalized to the situation where the first-order inclusion probabilities can be unequal. However, as the following further counterexample shows, in such situations a CLT becomes impossible in general.

PROPOSITION 3.21. *Let $\alpha \in (0, 1)$. Then there exists a Bernoulli sampling scheme with unequal first-order inclusion probabilities such that $n/N \searrow 0$, $n/N^{\alpha} \nearrow \infty$ and $\min_{1 \le i \le N} \pi_i \searrow 0$, and the random variable $\sqrt{\frac{n}{N}} \cdot Z_N$ is not bounded in probability.*

Consequently, in sharp contrast to Proposition 3.20, there is no hope for a general Donsker theory with $\sqrt{n}$ normalization if $\min_{1 \le i \le N} \pi_i \searrow 0$ as long as the first-order inclusion probabilities are unequal. The proposition above is actually proving a much more negative phenomenon: although a CLT is possible under equal π_i's for Bernoulli sampling in the whole regime $n/N \searrow 0$, $n \nearrow \infty$, such CLTs are not possible for *any convergence regime of* $n/N \searrow 0$, as soon as one allows unequal π_i's. The failure of CLTs with $\sqrt{n}$ normalization here is particularly striking, as one would heuristically imagine that n is the effective sample size. The main trouble here, however, is that when (A1) fails with unequal first-order inclusion probabilities, the variance pattern of ξ_i's can be arbitrarily complicated.

4. Applications. In this section, we apply the new tools developed in Section 3 in statistical problems including:

1. M-estimation (or *empirical risk minimization*) in a general nonparametric model;
2. Z-estimation in a general semiparametric model;
3. frequentist theory for pseudo-Bayes procedures, namely, theory of pseudo-posterior contraction rates and Bernstein–von Mises type theorems,

where the usual likelihood is replaced by the Horvitz–Thompson weighted likelihood. We will not consider the calibrated version of these problems for simplicity of exposition, given that the corresponding theory has been fully developed in Section 3. These problems are not meant to be exhaustive; they are demonstrated as an illustration and a proof of concept of the new tools.

4.1. *M-Estimation.* Consider the canonical *empirical risk minimization* problem (or "*M*-estimation") based on weighted likelihood:

$$\hat{f}_N^{\pi} \equiv \arg\min_{f\in\mathcal{F}} \mathbb{P}_N^{\pi} f. \tag{4.1}$$

The quality of the estimator defined in (4.1) is evaluated through the *excess risk* of $\hat{f}_N^{\pi}$, denoted $\mathcal{E}_P(\hat{f}_N^{\pi})$, where

$$\mathcal{E}_P(f) \equiv Pf - \inf_{g\in\mathcal{F}} Pg, \quad f \in \mathcal{F}.$$

The problem of studying excess risk of empirical risk minimizers under the usual empirical measure has been extensively studied in the 2000s; we only refer the reader to [33, 46, 47] and references therein. Under the Horvitz–Thompson empirical measure, [25] studied risk bounds for the binary classification problem under sampling designs that are close to the rejective sampling design. Our goal here will be a study of the excess risk for the *M*-estimator based on weighted likelihood as defined in (4.1) for the general empirical risk minimization problem under general sampling designs.

To this end, let $\mathcal{F}_{\mathcal{E}}(\delta) \equiv \{f \in \mathcal{F} : \mathcal{E}_P(f) < \delta^2\}$, let $\rho_P : \mathcal{F}\times\mathcal{F} \to \mathbb{R}_{\geq 0}$ be such that $\rho_P^2(f,g) \geq P(f-g)^2 - (P(f-g))^2$, and $D(\delta) \equiv \sup_{f,g\in\mathcal{F}_{\mathcal{E}}(\delta)} \rho_P(f,g)$. Now we may prove the following theorem.

THEOREM 4.1. *Suppose* (A1) *holds. Suppose that there exist some* $L > 0$, $\kappa \geq 1$ *such that* $D(\delta) \leq L\cdot\delta^{1/\kappa}$, *and that* $\mathcal{F}$ *is uniformly bounded and satisfies an entropy condition with exponent* $\alpha \in (0,2)$. *Then there exist some constants* $\{C_i\}_{i=1}^3$ *only depending on* π_0, L, κ, α *such that for any* $s, t \geq 0$, *with*

$$r_N \geq C_1 N^{-\frac{\kappa}{4\kappa-2+\alpha}} + C_2\left[\left(\frac{s\vee t^2}{N}\right)^{\frac{\kappa}{4\kappa-2}} \vee \frac{s}{N}\right],$$

it holds that

$$\mathbb{P}\big(\mathcal{E}_P\big(\hat{f}_N^{\pi}\big) \geq r_N^2\big) \leq \frac{C_3}{s}e^{-s/C_3} + \mathbb{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i}-1\right)\right| > t\right).$$

As an illustration of Theorem 4.1, we consider below two standard settings, regression and classification, similar to the development in [33]. For simplicity of exposition, we also assume that (A2-CLT) holds.

EXAMPLE 4.2 (Bounded regression). Let $\{(X_i, Y_i) \in \mathcal{X}\times[-1,1]\}_{i=1}^N$ denote the i.i.d. copies of the pairs consisting of covariates X_i and responses Y_i. Our goal is to estimate the regression function $g_0(x) \equiv \mathbb{E}[Y|X=x]$ using the weighted least squares method:

$$\hat{g}_N^{\pi} \equiv \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i}\big(Y_i - g(X_i)\big)^2,$$

where $\mathcal{G}$ is a function class containing functions taking values in $[-1,1]$, and the weights $\{\xi_i, \pi_i\}$ may depend on auxiliary information $Z^{(N)}$. To apply Theorem 4.1, let $\mathcal{F} \equiv \{f_g(x,y) \equiv (y-g(x))^2 : g \in \mathcal{G}\}$. Then following the arguments in page 1208 of [33], we have $\mathcal{E}_{P_{(X,Y)}}(f_g) = \|g - g_0\|_{L_2(P_X)}^2$ and we may take $\kappa = 1$. If $\mathcal{G}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$, it is easy to verify that the same holds for $\mathcal{F}$, and hence Theorem 4.1 yields

$$\big\|\hat{g}_N^{\pi} - g_0\big\|_{L_2(P)}^2 = \mathcal{O}_{\mathbb{P}}\big(N^{-\frac{2}{2+\alpha}}\big),$$

a very typical rate in the regression problem.

EXAMPLE 4.3 (Classification). Let $\{(X_i, Y_i) \in \mathcal{X} \times \{0,1\}\}_{i=1}^N$ denote the i.i.d. copies of the pairs consisting of covariates X_i and responses Y_i. A classifier $g : \mathcal{X} \to \{0,1\}$ over a class $\mathcal{G}$ has a generalization error $P_{(X,Y)}(Y \neq g(X))$. The excess risk for a classifier g over $\mathcal{G}$ under law $P_{(X,Y)}$ is given by

$$\mathcal{E}_{P_{(X,Y)}}(g) \equiv P_{(X,Y)}\big(Y \neq g(X)\big) - \inf_{g' \in \mathcal{G}} P_{(X,Y)}\big(Y \neq g'(X)\big).$$

It is known that for a given law $P_{(X,Y)}$ on (X, Y), the minimal generalized error is attained by a Bayes classifier $g_0(x) \equiv \mathbf{1}_{\eta(x) \geq 1/2}$ where $\eta(x) \equiv \mathbb{E}[Y|X = x]$; cf. [29]. In the setting of complex sampling design, it is natural to estimate g_0 by minimizing the weighted training error:

$$\hat{g}_N^\pi \equiv \arg\min_{g \in \mathcal{G}} \sum_{i=1}^N \frac{\xi_i}{\pi_i} \mathbf{1}_{Y_i \neq g(X_i)},$$

where $g_0 \in \mathcal{G}$ is a collection of classifiers. To apply Theorem 4.1, let $\mathcal{F} \equiv \{f_g \equiv \mathbf{1}_{y \neq g(x)} : g \in \mathcal{G}\}$. Suppose the following margin condition (cf. [54, 76]) holds for some $c > 0$, $\kappa \geq 1$: for all $g \in \mathcal{G}$,

$$\mathcal{E}_{P_{(X,Y)}}(g) \geq c\Pi^\kappa\big(g(X) \neq g_0(X)\big), \tag{4.2}$$

where Π is the marginal law of X under P. Following page 1212 of [33], we may choose $D(\delta) \lesssim \delta^{1/\kappa}$, and hence if the collection of classifiers $\mathcal{G}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$, using $(f_{g_1} - f_{g_2})^2 \leq (g_1 - g_2)^2$, we see that $\mathcal{F}$ also satisfies the same entropy condition, and hence

$$P_{(X,Y)}\big(Y \neq \hat{g}_N^\pi(X)\big) - \inf_{g' \in \mathcal{G}} P_{(X,Y)}\big(Y \neq g'(X)\big) = \mathcal{O}_{\mathbb{P}}\big(N^{-\frac{\kappa}{2\kappa - 1 + \alpha/2}}\big),$$

a very typical rate in the classification problem.

4.2. *Z-Estimation.* The method of Z-estimation that produces estimators by finding those values of the parameters which zero out a set of estimating equations is well understood by now under the usual empirical measure; see [78, 81] for a comprehensive treatment. With the Horvitz–Thompson empirical measure, [16, 17, 68, 69] considered weighted likelihood estimation under stratified sampling designs, both with and without overlaps. The goal of this section is to give a unified theoretical treatment for the Z-estimation problem under general sampling designs.

Let $\hat{\theta}_N^\pi \in \Theta$ solve the (possibly infinite-dimensional) estimating equations based on weighted likelihood:

$$\mathbb{P}_N^\pi \psi_{\hat{\theta}_N^\pi, h} = 0 \quad \text{for all } h \in \mathcal{H},$$

while the "truth" $\theta_0 \in \Theta$ solves the population equations

$$P\psi_{\theta_0, h} = 0 \quad \text{for all } h \in \mathcal{H}.$$

Let $\Psi_N, \Psi : \Theta \to \ell^\infty(\mathcal{H})$ be given by $\Psi_N(\theta)(h) \equiv \mathbb{P}_N^\pi \psi_{\theta,h}$ and $\Psi(\theta)(h) \equiv P\psi_{\theta,h}$. We assume that $\mathcal{H}$ is countable without loss of generality.

THEOREM 4.4. *Suppose that* (A1) *and* (A2-CLT) *hold, and that the following conditions hold*:

(Z1) *The map* Ψ *is Fréchet differentiable at* θ_0 *with a continuously invertible derivative* $\dot{\Psi}_{\theta_0}$.

(Z2) *The stochastic equicontinuity condition holds*:

$$\|\mathbb{G}_N(\psi_{\hat{\theta}_N^\pi,h} - \psi_{\theta_0,h})\|_{\mathcal{H}} = \mathfrak{o}_{\mathbb{P}}(1 + \sqrt{N}\|\hat{\theta}_N^\pi - \theta_0\|)$$

and $\{\psi_{\theta_0,h} : h \in \mathcal{H}\}$ *is a* P*-Glivenko–Cantelli class.*

If $\hat{\theta}_N^\pi \to_{\mathbf{P}} \theta_0$, *then*

$$\sqrt{N}(\hat{\theta}_N^\pi - \theta_0) = -\dot{\psi}_{\theta_0}^{-1}\mathbb{G}_N^\pi \psi_{\theta_0,\cdot} + \mathfrak{o}_{\mathbb{P}}(1).$$

This theorem is comparable to the standard Z-Theorem 3.3.1 in [81], but here we work in the context of Z-estimation under weighted likelihood. Note that our conditions are are almost identical to the standard Z-Theorem, many examples for which Theorem 4.4 applies can be found in Section 3.3 of [81] (see also [78, 79]). In particular, (Z2) is imposed for the usual empirical process $\mathbb{G}_N$, and can be easily checked if a Donsker property for the class $\{\psi_{\theta,h} - \psi_{\theta_0,h} : \|\theta - \theta_0\| \le \delta, h \in \mathcal{H}\}$ holds. We omit these details here.

Now consider estimation of a finite-dimensional parameter in the presence of an infinite-dimensional nuisance parameter, that is, estimation in a semiparametric model. Following [24, 53], we use the following general semiparametric framework: Consider a model $\{P_{\theta,\eta} : (\theta, \eta) \in \mathbb{R}^d \times \mathcal{H}\}$, where $\mathcal{H}$ is an infinite dimensional Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$. Suppose that the true parameter is (θ_0, η_0). An estimator $(\hat{\theta}_N^\pi, \hat{\eta}_N^\pi)$ of (θ_0, η_0) usually takes the form

$$(\hat{\theta}_N^\pi, \hat{\eta}_N^\pi) := \arg\sup \mathbb{P}_N^\pi m_{\theta,\eta}, \tag{4.3}$$

where $m_{\theta,\eta}$ is often the log likelihood function (for $n = 1$). However, here we will work with a more general Z-estimation framework.

For any fixed $\eta \in \mathcal{H}$, let $\eta(t)$ be a smooth curve at $t = 0$ with $\eta(0) = \eta$ and $a = (\partial/\partial t)\eta(t)|_{t=0}$ for some $a \in \mathcal{H}$. Denote $\mathcal{A} \subset \mathcal{H}$ the collection for all such admissible a's. Now let $m_\theta(\theta, \eta) = \partial_\theta m(\theta, \eta) \in \mathbb{R}^d$, $m_\eta(\theta, \eta)[a] = (\partial/\partial t)m(\theta, \eta(t))|_{t=0}$ with $\partial_t \eta(t)|_{t=0} = a \in \mathcal{A}$. The second derivatives can be defined in a similar fashion. Suppose further the following orthogonality condition hold: there exists $A^* = (a_1^*, \ldots, a_d^*) \in \mathcal{A}^d$ so that for any $A \in \mathcal{A}^d$, it holds that

$$P_{\theta_0,\eta_0}\big(m_{\theta\eta}(\theta_0, \eta_0)[A] - m_{\eta\eta}[A^*][A]\big) = 0. \tag{4.4}$$

This condition is commonly adopted in semiparametric literature to handle the case when nuisance parameter is not $\sqrt{n}$-estimable; see, for example, Condition 2, page 555 in [45].[1]

Define the *efficient score function* $\tilde{m}(\theta, \eta) = m_\theta(\theta, \eta) - m_\eta(\theta, \eta)[A^*]$ (since if m is the log likelihood function, $\tilde{m}$ typically becomes the efficient score function). Then (4.4) can be rewritten as following: for any $A \in \mathcal{A}^d$,

$$P_{\theta_0,\eta_0}\tilde{m}_\eta(\theta_0, \eta_0)[A] = 0. \tag{4.5}$$

We assume that the true parameter (θ_0, η_0) zeros out the population estimating equation:

$$P_{\theta_0,\eta_0}\tilde{m}(\theta_0, \eta_0) = 0. \tag{4.6}$$

To allow some flexibility in the framework, the estimators $(\hat{\theta}_N^\pi, \hat{\eta}_N^\pi)$ are assumed to approximately zero out the Horvitz–Thompson empirical estimating equation:

$$\mathbb{P}_N^\pi \tilde{m}(\hat{\theta}_N^\pi, \hat{\eta}_N^\pi) = \mathfrak{o}_{\mathbb{P}}(N^{-1/2}). \tag{4.7}$$

It is easy to see that the above condition is satisfied if (4.3) holds. Note here our general condition also includes the case where $\hat{\eta}_N^\pi$ may depend on $\hat{\theta}_N^\pi$, for example, profile likelihood estimation.

[1] See also condition A3 in [85], page 2138; condition (4) in [53], page 196; condition (4) in [24], page 2887.

THEOREM 4.5. *Suppose that* (A1) *holds, and that* (4.5)–(4.7) *hold. Further assume the following conditions*:

(S1) *The matrix* $I_{\theta_0,\eta_0} \equiv -P_{\theta_0,\eta_0}\tilde{m}_\theta(\theta_0,\eta_0) \in \mathbb{R}^{d\times d}$ *is nonsingular.*

(S2) $\|\hat{\theta}_N^\pi - \theta_0\| \vee \|\hat{\eta}_N^\pi - \eta_0\|_{\mathcal{H}} = \mathcal{O}_{\mathbb{P}}(N^{-\beta})$ *holds for some* $\beta > 1/4$.

(S3) *The model is smooth in the sense that*

$$\begin{aligned}&\big\|P_{\theta_0,\eta_0}\big(\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0) - \tilde{m}_\theta(\theta_0,\eta_0)(\theta-\theta_0)\big)\big\|\\ &\quad= \mathcal{O}\big(\|\theta-\theta_0\|^2 \vee \|\eta-\eta_0\|_{\mathcal{H}}^2\big)\end{aligned}$$

holds for (θ,η) *close enough to* (θ_0,η_0).

(S4) *For any* $C > 0$,

$$\sup_{\|\theta-\theta_0\|\vee\|\eta-\eta_0\|_{\mathcal{H}}\le CN^{-\beta}} \big|\mathbb{G}_N\big(\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0)\big)\big| = \mathfrak{o}_{\mathbb{P}}(1).$$

Then

$$\sqrt{N}\big(\hat{\theta}_N^\pi - \theta_0\big) = I_{\theta_0,\eta_0}^{-1}\mathbb{G}_N^\pi\tilde{m}(\theta_0,\eta_0) + \mathfrak{o}_{\mathbb{P}}(1).$$

Conditions (S1)–(S4) are all standard assumptions in semiparametric literature, and can be verified in numerous models, including the Cox model with right censored/current status data, partially linear model, panel count data (with covariates), etc. Here, we only consider the partially linear model; detailed verifications for other models can be found in, for example, [24, 53, 68, 69, 85].

EXAMPLE 4.6 (Partially linear model). Consider the following model:

$$Y_i = X_i^\top\theta_0 + f_0(W_i) + e_i, \quad i = 1,\dots,N,$$

where Y_i's are the responses, $\{(X_i, W_i) \in [-1,1]^d \times [0,1]\}$'s are i.i.d. covariates, and e_i's are i.i.d. normal errors independent of the covariates. The "true signal" $\theta_0 \in \mathbb{R}^d$ and $f_0 : [0,1] \to \mathbb{R}$ is a "smooth" function. For ease of exposition, we will consider the parameter space $\Xi \equiv \{(\theta, f) : \|\theta\|_1 \le 1, \|f\|_\infty \le 1, J(f) \le M\}$ for some $M > 0$, and here $J^2(f) := \int_0^1 (f''(t))^2\,\mathrm{d}t$. Now with $\bar{\lambda}_N \asymp N^{-2/5}$, let

$$(\hat{\theta}_N^\pi, \hat{f}_N^\pi) := \arg\min_{(\theta,f)\in\Xi}\big[\mathbb{P}_N^\pi\big(Y - X^\top\theta - f(W)\big)^2 + \bar{\lambda}_N^2 J^2(f)\big]. \tag{4.8}$$

To put the model into our framework, let $m(\theta, f) := -(y - x^\top\theta - f(w))^2$. Then for any admissible a, b, we have

$$\begin{aligned}&m_\theta(\theta,f) = 2x\big(y - x^\top\theta - f(w)\big), \qquad m_f(\theta,f)[a] = 2a(w)\big(y - x^\top\theta - f(w)\big),\\ &m_{\theta f}(\theta,f)[b] = -2xb(w), \qquad m_{ff}(\theta,f)[a][b] = -2a(w)b(w).\end{aligned}$$

Now let $A^*(W) = \mathbb{E}[X|W] \in \mathbb{R}^d$. Then a direct calculation verifies (4.4). Thus we can take

$$\tilde{m}(\theta,f) = 2\big(y - x^\top\theta - f(w)\big)\big(x - \mathbb{E}[X|W = w]\big). \tag{4.9}$$

(4.6) is immediately verified; (4.7) can also be verified by taking partial derivatives in the definition (4.8) and noting that $\bar{\lambda}_N^2 = \mathfrak{o}(N^{-1/2})$. Now we verify (S1)–(S4). (S1) will be satisfied if the matrix $I_{\theta_0,\eta_0} \equiv 2\mathbb{E}[(X - \mathbb{E}[X|W])X^\top] = 2\mathbb{E}[(X - \mathbb{E}[X|W])^{\otimes 2}]$ is nonsingular. (S2) can be verified with $\beta = 2/5$ along the lines of Lemma 25.88 in [80] with the tools developed in Section 3. (S3) is trivially satisfied since $\tilde{m}$ is linear in θ and f. (S4) is also easy to verify. Hence we have shown that under the same conditions as in Lemma 25.88 of [80],

$$\sqrt{N}\big(\hat{\theta}_N^\pi - \theta_0\big) = I_{\theta_0,\eta_0}^{-1}\mathbb{G}_N^\pi\tilde{m}(\theta_0,\eta_0) + \mathfrak{o}_{\mathbb{P}}(1).$$

4.3. *Frequentist theory for pseudo-Bayesian procedures.* Suppose the i.i.d. superpopulation variables of interest $\{Y_i\}_{i=1}^N$ have law P_{f_0} where f_0 belongs to a statistical model $\mathcal{F}$ and $\{P_f\}_{f\in\mathcal{F}}$ is dominated by a σ-finite measure μ. A Bayesian approach assigns a prior Π_N on the model $\mathcal{F}$ and makes estimation/inference based on the posterior distribution. In the case where all the superpopulation $\{Y_i\}_{i=1}^N$ are available, by Bayes' formula, the posterior distribution, that is, a random measure on $\mathcal{F}$, is defined as follows: for a measurable subset $B \subset \mathcal{F}$,

$$\Pi_N\big(B|Y^{(N)}\big) \equiv \frac{\int_B \prod_{i=1}^N p_f(Y_i)\,\mathrm{d}\Pi_N(f)}{\int \prod_{i=1}^N p_f(Y_i)\,\mathrm{d}\Pi_N(f)} = \frac{\int_B \exp(N\mathbb{P}_N \log p_f)\,\mathrm{d}\Pi_N(f)}{\int \exp(N\mathbb{P}_N \log p_f)\,\mathrm{d}\Pi_N(f)}, \tag{4.10}$$

where $p_f(\cdot)$ denotes the probability density function of P_f with respect to the dominating measure μ.

In the current superpopulation setup with complex sampling designs, we may naturally replace the usual empirical measure $\mathbb{P}_N$ in (4.10) by the Horvitz–Thompson empirical measure $\mathbb{P}_N^\pi$ to define the *pseudo-posterior distribution with weighted likelihood* as follows: for a measurable subset $B \subset \mathcal{F}$,

$$\Pi_N^\pi\big(B|D^{(N)}\big) \equiv \frac{\int_B \prod_{i=1}^N p_f(Y_i)^{\xi_i/\pi_i}\,\mathrm{d}\Pi_N(f)}{\int \prod_{i=1}^N p_f(Y_i)^{\xi_i/\pi_i}\,\mathrm{d}\Pi_N(f)} = \frac{\int_B \exp(N\mathbb{P}_N^\pi \log p_f)\,\mathrm{d}\Pi_N(f)}{\int \exp(N\mathbb{P}_N^\pi \log p_f)\,\mathrm{d}\Pi_N(f)}. \tag{4.11}$$

Recall here $D^{(N)} \equiv (Y^{(N)}, Z^{(N)}, \xi^{(N)}, \pi^{(N)})$. Note that since $\prod_{i=1}^N p_f(Y_i)^{\xi_i/\pi_i}$ is not a true likelihood because of the power ξ_i/π_i, the resulting expression is not a posterior distribution in the usual sense, and hence we call it a *pseudo-posterior based on weighted likelihood*. Bayesian inference based on (4.11) in the complex sampling setting is initiated in [71], and is further developed in [49]. As we will see below, one particular advantage of the pseudo-posterior distribution with weighted likelihood defined above is that we may obtain a complete frequentist theory for pseudo-Bayes procedures analogous to that based on observing the whole superpopulation $\{Y_i\}_{i=1}^N$.

4.3.1. *Pseudo-posterior contraction rate theory.* We say that the pseudo-posterior distribution with weighted likelihood, namely $\Pi_N^\pi(\cdot|D^{(N)})$, contracts at a rate δ_N with respect to a metric d if

$$P_{f_0}\Pi_N^\pi\big(f \in \mathcal{F} : d^2(f, f_0) > L_N\delta_N^2 | D^{(N)}\big) \to 0$$

for any $L_N \to \infty$.

Our first goal in this section is to develop some useful results in deriving such pseudo-posterior contraction rates for the pseudo-posterior distribution using weighted likelihood. We will use (essentially the same) machinery developed in [39] (which we find easier to adapt in the current context than the standard machinery [31, 32]). For some $v > 0$, $c \in [0, \infty)$ let

$$\psi_{v,c}(\lambda) = v\lambda^2 \cdot \mathbf{1}_{|\lambda|\le 1/c} + \infty \cdot \mathbf{1}_{|\lambda|>1/c}$$

denote the local quadratic function.

THEOREM 4.7. *Suppose* (A1) *holds and the following conditions hold*:

(B1) (*Local Gaussianity condition*) *There exist some constants* $c_1 > 0$ *and* $\kappa = (\kappa_g, \kappa_\Gamma) \in (0, \infty) \times [0, \infty)$ *such that for all* $n \in \mathbb{N}$, *and* $f_0, f_1 \in \mathcal{F}$,

$$P_{f_0} \exp\left[\lambda\left(\log\frac{p_{f_0}}{p_{f_1}} - P_{f_0}\log\frac{p_{f_0}}{p_{f_1}}\right)\right] \le c_1 \exp\big[\psi_{\kappa_g d^2(f_0,f_1),\kappa_\Gamma}(\lambda)\big].$$

Here, $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ is a symmetric function satisfying

$$c_2 \cdot d^2(f_0, f_1) \leq P_{f_0} \log \frac{p_{f_0}}{p_{f_1}} \leq c_3 \cdot d^2(f_0, f_1)$$

for some constants $c_2, c_3 > 0$.

(B2) (*Local entropy condition*) *There exist some $\{\delta_N\}_{N\in\mathbb{N}}$ such that*

$$1 + \sup_{\varepsilon > \delta_N} \log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F} : d(f, f_0) \leq 2\varepsilon\}, d\big) \leq c_4 N\delta_N^2,$$

where $c_4 \in (0, 1)$, $c_5 \in (0, 1/4)$ depend on $\{c_i\}_{i=1}^3$.

(B3) (*Prior mass condition*) *For all $j \in \mathbb{N}$,*

$$\frac{\Pi_N(\{f \in \mathcal{F} : j\delta_N < d(f, f_0) \leq (j+1)\delta_N\})}{\Pi_N(d(f, f_0) \leq \delta_N)} \leq \exp\big(c_6 j^2 N\delta_N^2\big),$$

where $c_6 > 0$ is a small enough constant depending on $\{c_i\}_{i=1}^3$.

Then

$$P_{f_0} \Pi_N^{\pi}\big(f \in \mathcal{F} : d^2(f, f_0) > C_1\delta_N^2 | D^{(N)}\big) \leq C_2 \exp\big(-N\delta_N^2/C_2\big).$$

Here, $C_1, C_2 > 0$ only depend on $\{c_i\}_{i=1}^3$ and κ.

The local Gaussianity condition (B1) can be easily verified in a wide range of experiments including regression/density estimation/Gaussian autoregression/Gaussian time series/covariance matrix estimation, etc. (B2)–(B3) are standard conditions in the literature. In particular, (B3) allows the exact $\sqrt{N}$ parametric pseudo-posterior contraction rate, which will be useful below. It is also possible to consider hierarchical priors to formulate a similar theorem as in [39]—in essence all examples therein can be considered here (except for regression where random design instead of fixed design is needed to maintain the i.i.d. property of the superpopulation $\{Y_i\}_{i=1}^N$).

4.3.2. *Bernstein–von Mises theorem.* Next, we will be interested in a more precise limiting distribution of the pseudo-posterior distribution with weighted likelihood, that is, a Bernstein–von Mises type theorem. To this end, we work with a finite-dimensional model, where Θ is a compact subset of $\mathbb{R}^d$. Let $\theta_0 \in \Theta$, an interior point of Θ, be the true parameter. Let $\mathcal{N}_{\mu,\Sigma}$ denote the d-dimensional normal distribution with mean μ and covariance matrix Σ.

THEOREM 4.8. *Suppose that* (A1) *and* (A2-*CLT*) *hold. Further assume the following conditions*:

(Bv1) (*Experiment*) *The map $\theta \mapsto \log p_\theta(x) = \ell_\theta(x)$ is differentiable at θ_0 for all x with derivative $\dot{\ell}_{\theta_0}(x)$, and for θ_1, θ_2 close enough to θ,*

$$\big|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)\big| \leq m(x)\|\theta_1 - \theta_2\|$$

holds for some P_{θ_0}-square integrable function m. Furthermore, the log-likelihood ratio $\{\log \frac{p_\theta}{p_{\theta_0}}\}_{\theta\in\Theta}$ satisfies the local Gaussianity condition, and is twice differentiable under P_{θ_0} with a nonsingular Hessian I_{θ_0}: for θ close enough to θ_0,

$$P_{\theta_0} \log \frac{p_{\theta_0}}{p_\theta} = \frac{1}{2}(\theta - \theta_0) I_{\theta_0}(\theta - \theta_0) + \mathfrak{o}\big(\|\theta - \theta_0\|^2\big).$$

(Bv2) (*Prior*) *The prior Π has a Lebesgue density bounded away from 0 and ∞ on Θ.*

Then the pseudo-posterior distribution with weighted likelihood Π_N^π converges to a sequence of normal distributions in the total variational distance:

$$\sup_B \left|\Pi_N^\pi\left(\sqrt{N}(\theta-\theta_0)\in B|D^{(N)}\right)-\mathcal{N}_{I_{\theta_0}^{-1}\mathbb{G}_N^\pi\dot{\ell}_{\theta_0},I_{\theta_0}^{-1}}(B)\right|=\mathfrak{o}_{\mathbb{P}}(1).$$

Note that in finite-dimensional problems, the efficient score $\tilde{m}$ in Theorem 4.5 can usually be taken as $\dot{\ell}_{\theta_0}$. Then under the regularity conditions as in Theorem 4.5, we have the usual interpretation of the Bernstein–von Mises theorem in our context of weighted likelihood estimation: the sequence of pseudo-posterior distributions with weighted likelihood resembles that of progressively sharpened normal distributions centered at the maximum weighted likelihood estimator $\hat{\theta}_N^\pi$:

$$\sup_B \left|\Pi_N^\pi\left(\theta\in B|D^{(N)}\right)-\mathcal{N}_{\hat{\theta}_N^\pi,N^{-1}I_{\theta_0}^{-1}}(B)\right|=\mathfrak{o}_{\mathbb{P}}(1).$$

4.3.3. *Inference using the Bernstein–von Mises theorem.* The Bernstein–von Mises theorem in the i.i.d. sampling models justifies the frequentist validity of the credible sets of the posterior distribution for the purpose of inference. The situation in the complex sampling setting is however more subtle. As will be clear from the discussion below, the structure of the Bernstein–von Mises theorem in Theorem 4.8 shows that: (1) vanilla credible sets may not lead to valid frequentist inference procedure; (2) but at the same time suggests the construction of a corrected credible set with asymptotically valid coverage.

To see (1), suppose $C_N=C_N(D^{(N)})\subset\Theta$ is a (vanilla) $(1-\alpha)$ credible set, that is, $\Pi_N^\pi(\theta\in C_N|D^{(N)})=1-\alpha$. Then by the Bernstein–von Mises theorem in Theorem 4.8, $\mathcal{N}_{0,I}((NI_{\theta_0})^{1/2}(C_N-\bar{\theta}_N^\pi))\to 1-\alpha$ in $\mathbb{P}$-probability. Here, $\bar{\theta}_N^\pi\equiv\theta_0+N^{-1/2}I_{\theta_0}^{-1}\mathbb{G}_N^\pi\dot{\ell}_{\theta_0}$. In other words, $C_N=\bar{\theta}_N^\pi+I_{\theta_0}^{-1/2}B_N/\sqrt{N}$ for some random B_N such that $\mathcal{N}_{0,I}(B_N)\to 1-\alpha$ in $\mathbb{P}$-probability. By Proposition 3.4, we have $\mathbb{G}_N^\pi\dot{\ell}_{\theta_0}\to_d\mathcal{N}(0,(1+\mu_{\pi 1})I_{\theta_0})$, and hence the frequentist coverage for the credible set C_N is

$$\begin{aligned}\mathbb{P}_{\theta_0}(\theta_0\in C_N)&=\mathbb{P}\big(\theta_0\in\bar{\theta}_N^\pi+I_{\theta_0}^{-1/2}B_N/\sqrt{N}\big)=\mathbb{P}\big(I_{\theta_0}^{1/2}\sqrt{N}\big(\theta_0-\bar{\theta}_N^\pi\big)\in B_N\big)\\&=\mathbb{P}\big(-I_{\theta_0}^{-1/2}\mathbb{G}_N^\pi\dot{\ell}_{\theta_0}\in B_N\big)=\mathbb{E}\mathcal{N}_{0,I}\big(B_N/(1+\mu_{\pi 1})^{1/2}\big)+\mathfrak{o}(1),\end{aligned}\tag{4.12}$$

which does not converge to $1-\alpha$ in general as $N\to\infty$ as long as $\mu_{\pi 1}\neq 0$.

Fortunately, the vanilla credible set C_N can be corrected as follows. Let $\hat{\theta}_N^\pi=\arg\sup_{\theta\in\Theta}\mathbb{P}_N^\pi\log p_\theta$ be the maximum weighted likelihood estimator as in Section 4.2. Then under regularity conditions, $\hat{\theta}_N^\pi=\theta_0+N^{-1/2}I_{\theta_0}^{-1}\mathbb{G}_N^\pi\dot{\ell}_{\theta_0}+\mathfrak{o}_{\mathbb{P}}(1)$. Now for any $C_N=C_N(D^{(N)})$ such that $\Pi_N^\pi(\theta\in C_N|D^{(N)})=1-\alpha$, let

$$C_N^*\equiv\hat{\theta}_N^\pi+(1+\mu_{\pi 1})^{1/2}\big(C_N-\hat{\theta}_N^\pi\big).\tag{4.13}$$

Note again that $C_N=\hat{\theta}_N^\pi+I_{\theta_0}^{-1/2}B_N/\sqrt{N}$ for some random B_N such that $\mathcal{N}_{0,I}(B_N)\to 1-\alpha$ in $\mathbb{P}$-probability, so we have

$$\begin{aligned}\mathbb{P}_{\theta_0}(\theta_0\in C_N^*)&=\mathbb{P}_{\theta_0}\big(\theta_0\in\hat{\theta}_N^\pi+(1+\mu_{\pi 1})^{1/2}\big(C_N-\hat{\theta}_N^\pi\big)\big)\\&=\mathbb{P}_{\theta_0}\big(I_{\theta_0}^{1/2}\sqrt{N}\big(\theta_0-\hat{\theta}_N^\pi\big)\in(1+\mu_{\pi 1})^{1/2}\cdot I_{\theta_0}^{1/2}\sqrt{N}\big(C_N-\hat{\theta}_N^\pi\big)\big)\\&=\mathbb{P}_{\theta_0}\big(-(1+\mu_{\pi 1})^{-1/2}I_{\theta_0}^{-1/2}\mathbb{G}_N^\pi\dot{\ell}_{\theta_0}\in B_N\big)+\mathfrak{o}(1)\\&\to 1-\alpha.\end{aligned}$$

Hence the corrected credible set C_N^* (4.13) has the correct coverage.

The construction of the corrected credible set in (4.13) is generic, regardless of different sampling schemes as long as $\mu_{\pi 1}$ is known (cf. Table 1). Such a unified construction of corrected credible sets based on the pseudo-posterior distributions could bring significant advantage for the purpose of inference in the complex sampling setting, in that it alleviates complicated bootstrap methods whose design architectures must adapt to the dependence structure in each and every different sampling schemes (e.g., [66, 67] in the context of two-phase sampling). In Appendix D, we present an illustrative example and simulation results in the context of one-dimensional Gaussian location model with a Gaussian prior for the phenomenon discussed above.

Finally, we remark that there are interesting recent developments in semiparametric and nonparametric Bernstein–von Mises theorems; cf. [19–22, 58, 59]. It is an interesting open question to extend the Bernstein–von Mises theorem and the correction scheme (4.13) to these more complicated settings with complex sampling designs.

Acknowledgments. The authors would like to thank Thomas Lumley for several useful suggestions. Helpful and constructive comments from an Associate Editor and two referees are greatly appreciated.

The research of Q. Han was supported in part by NSF Grant DMS-1916221. The research of J. A. Wellner was supported in part by NSF Grant DMS-1566514, NI-AID grant 2R01 AI291968-04, a Simons Fellowship via the Newton Institute (INI-program STS 2018), Cambridge University and the Saw Swee Hock Visiting Professorship of Statistics at the National University of Singapore (in 2019).

SUPPLEMENTARY MATERIAL

Supplement: Proofs (DOI: 10.1214/20-AOS1964SUPP; .pdf). In the supplement, we provide proofs for the results in the main paper.

REFERENCES

[1] ALEXANDER, K. S. (1985). Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, *Vol. II* (*Berkeley*, *Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 475–493. Wadsworth, Belmont, CA. MR0822047

[2] ALEXANDER, K. S. (1987). The central limit theorem for weighted empirical processes indexed by sets. *J. Multivariate Anal.* **22** 313–339. MR0899666 https://doi.org/10.1016/0047-259X(87)90093-5

[3] ALEXANDER, K. S. (1987). Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Related Fields* **75** 379–423. MR0890285 https://doi.org/10.1007/BF00318708

[4] BARRETT, G. F. and DONALD, S. G. (2009). Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *J. Bus. Econom. Statist.* **27** 1–17. MR2484980 https://doi.org/10.1198/jbes.2009.0001

[5] BERGER, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *J. Statist. Plann. Inference* **74** 149–168. MR1665125 https://doi.org/10.1016/S0378-3758(98)00107-4

[6] BERGER, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz–Thompson estimator. *J. Statist. Plann. Inference* **67** 209–226. MR1624693 https://doi.org/10.1016/S0378-3758(97)00107-9

[7] BERTAIL, P., CHAUTRU, E. and CLÉMENÇON, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. *Scand. J. Stat.* **44** 97–111. MR3619696 https://doi.org/10.1111/sjos.12243

[8] BHATTACHARYA, D. (2007). Inference on inequality from household survey data. *J. Econometrics* **137** 674–707. MR2354960 https://doi.org/10.1016/j.jeconom.2005.09.003

[9] BHATTACHARYA, D. and MAZUMDER, B. (2011). A nonparametric analysis of black–white differences in intergenerational income mobility in the United States. *Quant. Econ.* **2** 335–379.

[10] BOISTARD, H., LOPUHAÄ, H. P. and RUIZ-GAZEN, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electron. J. Stat.* **6** 1967–1983. MR3020253 https://doi.org/10.1214/12-EJS736

[11] BOISTARD, H., LOPUHAÄ, H. P. and RUIZ-GAZEN, A. (2017). Functional central limit theorems for single-stage sampling designs. *Ann. Statist.* **45** 1728–1758. MR3670194 https://doi.org/10.1214/16-AOS1507

[12] BREIDT, F. J. and OPSOMER, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *Ann. Statist.* **28** 1026–1053. MR1810918 https://doi.org/10.1214/aos/1015956706

[13] BRESLOW, N., 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. MR2001644 https://doi.org/10.1214/aos/1059655907

[14] BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Stat. Biosci.* **1** 32–49.

[15] BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* **169** 1398–1405.

[16] 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. Stat.* **34** 86–102. MR2325244 https://doi.org/10.1111/j.1467-9469.2006.00523.x

[17] 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; MR2325244]. *Scand. J. Stat.* **35** 186–192. MR2391566 https://doi.org/10.1111/j.1467-9469.2007.00574.x

[18] CARDOT, H., CHAOUCH, M., GOGA, C. and LABRUÈRE, C. (2010). Properties of design-based functional principal components analysis. *J. Statist. Plann. Inference* **140** 75–91. MR2568123 https://doi.org/10.1016/j.jspi.2009.06.012

[19] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753 https://doi.org/10.1007/s00440-010-0316-5

[20] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856 https://doi.org/10.1214/13-AOS1133

[21] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246

[22] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

[23] CHAUVET, G. (2015). Coupling methods for multistage sampling. *Ann. Statist.* **43** 2484–2506. MR3405601 https://doi.org/10.1214/15-AOS1348

[24] CHENG, G. and HUANG, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.* **38** 2884–2915. MR2722459 https://doi.org/10.1214/10-AOS809

[25] CLÉMENÇON, S., BERTAIL, P. and PAPA, G. (2016). Learning from survey training samples: Rate bounds for Horvitz–Thompson risk minimizers. In *Asian Conference on Machine Learning* 142–157.

[26] CONTI, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B* **76** 234–259. MR3302272 https://doi.org/10.1007/s13571-014-0083-x

[27] DAVIDSON, R. (2009). Reliable inference for the Gini index. *J. Econometrics* **150** 30–40. MR2525992 https://doi.org/10.1016/j.jeconom.2008.11.004

[28] DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

[29] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[30] FULLER, W. A. (2011). *Sampling Statistics* **560**. Wiley.

[31] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[32] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274 https://doi.org/10.1214/009053606000001172

[33] GINÉ, E. and KOLTCHINSKII, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143–1216. MR2243881 https://doi.org/10.1214/009117906000000070

[34] GINÉ, E., KOLTCHINSKII, V. and WELLNER, J. A. (2003). Ratio limit theorems for empirical processes. In *Stochastic Inequalities and Applications. Progress in Probability* **56** 249–278. Birkhäuser, Basel. MR2073436

[35] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[36] HÁJEK, J. (1961). Some extensions of the Wald–Wolfowitz–Noether theorem. *Ann. Math. Stat.* **32** 506–523. MR0130707 https://doi.org/10.1214/aoms/1177705057

[37] HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Stat.* **35** 1491–1523. MR0178555 https://doi.org/10.1214/aoms/1177700375

[38] HÁJEK, J. (1981). *Sampling from a Finite Population*. *Statistics*: *Textbooks and Monographs* **37**. Dekker, New York. Edited by Václav Dupač, With a foreword by P. K. Sen. MR0627744

[39] HAN, Q. (2017). Bayes model selection. arXiv preprint, arXiv:1704.07513.

[40] HAN, Q. and WELLNER, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.* **47** 2286–2319. MR3953452 https://doi.org/10.1214/18-AOS1748

[41] HAN, Q. and WELLNER, J. A. (2021). Supplement to "Complex sampling designs: Uniform limit theorems and applications." https://doi.org/10.1214/20-AOS1964SUPP

[42] HARTLEY, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section* **19** 203–206. American Statistical Association, Washington, DC.

[43] HARTLEY, H. O. (1974). Multiple frame methodology and selected applications. *Sankhyā* **36** 118.

[44] 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. MR0053460

[45] HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24** 540–568. MR1394975 https://doi.org/10.1214/aos/1032894452

[46] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442 https://doi.org/10.1214/009053606000001019

[47] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math.* **2033**. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871 https://doi.org/10.1007/978-3-642-22147-7

[48] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5

[49] LEÓN-NOVELO, L. G. and SAVITSKY, T. D. (2019). Fully Bayesian estimation under informative sampling. *Electron. J. Stat.* **13** 1608–1645. MR3939589 https://doi.org/10.1214/19-ejs1538

[50] LIN, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87** 37–47. MR1766826 https://doi.org/10.1093/biomet/87.1.37

[51] LOHR, S. and RAO, J. N. K. (2006). Estimation in multiple-frame surveys. *J. Amer. Statist. Assoc.* **101** 1019–1030. MR2324141 https://doi.org/10.1198/016214506000000195

[52] LUMLEY, T., SHAW, P. A. and DAI, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *Int. Stat. Rev.* **79** 200–220. https://doi.org/10.1111/j.1751-5823.2011.00138.x

[53] MA, S. and KOSOROK, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.* **96** 190–217. MR2202406 https://doi.org/10.1016/j.jmva.2004.09.008

[54] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

[55] MASON, D. M., SHORACK, G. R. and WELLNER, J. A. (1983). Strong limit theorems for oscillation moduli of the uniform empirical process. *Z. Wahrsch. Verw. Gebiete* **65** 83–97. MR0717935 https://doi.org/10.1007/BF00534996

[56] NAN, B., KALBFLEISCH, J. D. and YU, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Ann. Statist.* **37** 2351–2376. MR2543695 https://doi.org/10.1214/08-AOS657

[57] NAN, B. and WELLNER, J. A. (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statist. Sinica* **23** 1155–1180. MR3114709

[58] NICKL, R. (2017). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation, arXiv preprint arXiv:1707.01764.

[59] NICKL, R. and SÖHL, J. (2019). Bernstein–von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat.* **13** 3513–3571. MR4013745 https://doi.org/10.1214/19-ejs1609

[60] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301

[61] ROSÉN, B. (1965). Limit theorems for sampling from finite populations. *Ark. Mat.* **5** 383–424. MR0177437 https://doi.org/10.1007/BF02591138

[62] ROSÉN, B. (1967). On the central limit theorem for a class of sampling procedures. *Z. Wahrsch. Verw. Gebiete* **7** 103–115. MR0210181 https://doi.org/10.1007/BF00536324
[63] ROSÉN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Ann. Math. Stat.* **43** 373–397; ibid. 43 (1972), 748–776. MR0321223 https://doi.org/10.1214/aoms/1177692620
[64] ROSÉN, B. (1974). Asymptotic theory for Des Raj's estimator. I, II. *Scand. J. Stat.* **1** 71–83; ibid. 1 (1974), no. 3, 135–144. MR0375560
[65] 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. MR2253102 https://doi.org/10.1214/009053605000000651
[66] SAEGUSA, T. (2014). Bootstrapping two-phase sampling. arXiv preprint, arXiv:1406.5580.
[67] SAEGUSA, T. (2015). Variance estimation under two-phase sampling. *Scand. J. Stat.* **42** 1078–1091. MR3426311 https://doi.org/10.1111/sjos.12152
[68] SAEGUSA, T. (2019). Large sample theory for merged data from multiple sources. *Ann. Statist.* **47** 1585–1615. MR3911123 https://doi.org/10.1214/18-AOS1727
[69] SAEGUSA, T. and WELLNER, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.* **41** 269–295. MR3059418 https://doi.org/10.1214/12-AOS1073
[70] SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. *Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6
[71] SAVITSKY, T. D. and TOTH, D. (2016). Bayesian estimation under informative sampling. *Electron. J. Stat.* **10** 1677–1708. MR3522657 https://doi.org/10.1214/16-EJS1153
[72] SHORACK, G. R. (1973). Convergence of reduced empirical and quantile processes with application to functions of order statistics in the non-I.I.D. case. *Ann. Statist.* **1** 146–152. MR0336776
[73] SHORACK, G. R. and WELLNER, J. A. (1982). Limit theorems and inequalities for the uniform empirical process indexed by intervals. *Ann. Probab.* **10** 639–652. MR0659534
[74] STUTE, W. (1982). The oscillation behavior of empirical processes. *Ann. Probab.* **10** 86–107. MR0637378
[75] STUTE, W. (1984). The oscillation behavior of empirical processes: The multivariate case. *Ann. Probab.* **12** 361–379. MR0735843
[76] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131
[77] VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. MR1739079
[78] 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. MR1915446
[79] VAN DER VAART, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Stat. Neerl.* **49** 9–30. MR1333176 https://doi.org/10.1111/j.1467-9574.1995.tb01452.x
[80] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256
[81] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
[82] VÍŠEK, J. Á. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling. In *Contributions to Statistics* 263–275. Reidel, Dordrecht. MR0561274
[83] WELLNER, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. Verw. Gebiete* **45** 73–88. MR0651392 https://doi.org/10.1007/BF00635964
[84] WELLNER, J. A. and ZHAN, Y. (1996). Bootstrapping Z-estimators. Technical Report 308, Univ. Washington Dept. Statistics.
[85] WELLNER, J. A. and ZHANG, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35** 2106–2142. MR2363965 https://doi.org/10.1214/009053607000000181