# Practical targeted learning from large data sets by survey sampling

Patrice Bertail, Antoine Chambaz, Emilien Joly

▶ **To cite this version:**

Patrice Bertail, Antoine Chambaz, Emilien Joly. Practical targeted learning from large data sets by survey sampling. 2016. hal-01339538

**HAL Id: hal-01339538**
**https://hal.science/hal-01339538**

Preprint submitted on 29 Jun 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Practical targeted learning from large data sets by survey sampling

P. Bertail, A. Chambaz, E. Joly
Modal'X, Université Paris Ouest Nanterre

June 29, 2016

**Abstract**

We address the practical construction of asymptotic confidence intervals for smooth (*i.e.*, pathwise differentiable), real-valued statistical parameters by targeted learning from independent and identically distributed data in contexts where sample size is so large that it poses computational challenges. We observe some summary measure of all data and select a sub-sample from the complete data set by Poisson rejective sampling with unequal inclusion probabilities based on the summary measures. Targeted learning is carried out from the easier to handle sub-sample. We derive a central limit theorem for the targeted minimum loss estimator (TMLE) which enables the construction of the confidence intervals. The inclusion probabilities can be optimized to reduce the asymptotic variance of the TMLE. We illustrate the procedure with two examples where the parameters of interest are variable importance measures of an exposure (binary or continuous) on an outcome. We also conduct a simulation study and comment on its results.

keywords: semiparametric inference; survey sampling; targeted minimum loss estimation (TMLE)

## 1 Introduction

Large data sets are ubiquitous nowadays. They pose computational and theoretical challenges. We consider the particular problem of carrying out inference based on semiparametric models by targeted learning [19, 22] from large data sets. We mainly deal with the fact that the sample size $N$ is, say, huge. Even if we also take advantage of easy to handle summary measures of the observations, we do not consider the specific difficulties yielded by the messiness of real big data. This is why we use the expression "large data sets" instead of "big data".

Confronted with large data sets, many learning algorithms fail to provide an answer in a reasonable time if at all. Following [3], we overcome this computational limitation by *(i)* selecting $n$ among $N$ observations with unequal probabilities and *(ii)* adapting targeted learning from this smaller, tamed data set.

Specifically, our objective is to enable the construction of a confidence interval with given asymptotic level for a statistical parameter $\psi_0 \equiv \Psi(P_0)$ based on a sample $O_1, \ldots, O_N$ of a (huge number)

$N$ of independent and identically distributed (i.i.d.) random variables drawn from $P_0 \in \mathcal{M}$, where $\Psi : \mathcal{M} \to \mathbb{R}$ maps a set $\mathcal{M}$ of measures including possible distributions of $O_1$ to the real line. We focus on the case that the functional $\Psi$ is smooth in the following sense. For every $P \in \mathcal{M}$, there exists a wide class of one-dimensional paths $\{P_t : t \in ]-c, c[\} \subset \mathcal{M}$ with $P_t|_{t=0} = P$ and an influence function $D(P) \in L_0^2(P)$ such that, for all $|t| < c$,

$$\begin{aligned} \Psi(P_t) &= \Psi(P) + \int D(P)(dP_t - dP) + o(t) \\ &= \Psi(P) + \int D(P) dP_t + o(t). \end{aligned} \tag{1}$$

Here, we denote $L_0^2(P)$ the set of centered and square-integrable measurable functions relative to $P$.

Condition (1) trivially holds when $\Psi$ is linear. If, for instance, $\Psi$ is given by $\Psi(P) \equiv \int f dP$ for some measurable function $f$ integrable with respect to (wrt) all elements of $\mathcal{M}$, then (1) holds with $D(P) \equiv f - \Psi(P)$ (without the $o$-term). Even in the very simple example where $f$ is the identity and $\mathcal{M}$ consists of probability measures, hence $\Psi(P) = E_P[O]$, it may be computationally difficult, if not impossible, to build a confidence interval for $\psi_0 = \Psi(P_0)$ *using all observations*, merely because it may be very challenging to *access* to all of them in the context of large data sets.

Typical examples of functionals satisfying (1) include pathwise differentiable functionals as introduced in [24, Section 25.3]. We will give two examples of such functionals. Pathwise differentiability differs slightly from Gateaux, Hadamard and Fréchet differentiability. It is one the of key notions in the theory of semiparametric inference.

We overcome the computational hurdle by resorting to survey sampling, specifically to rejective sampling based on Poisson sampling with unequal inclusion probabilities. It is a particular case of sampling without replacement (we refer to [15] for an overview on sampling without replacement). Survey sampling can also rely on the so called sampling entropy [2, 7, 13], but we do not follow this path. Also known as Sampford sampling, rejective Poisson sampling has been thoroughly studied for the last five decades since the publication of the seminal articles [14, 18]. The key object in the analysis of Sampford sampling is the Horvitz-Thompson (HT) empirical measure. Asymptotic normality of estimators based on the HT empirical measure was first established in [14]. A functional version for the cumulative distribution function was obtained by [26] . Our analysis hinges on the recent study of the HT empirical measure from the viewpoint of empirical processes theory carried out in [3] (we refer the reader to this article for additional references).

For instance [8, 9] show practically how to implement confidence bands for model-assisted estimators of the mean when the variable of interest is functional and storage capacities are limited (with applications to electricity consumption curves). In that case, survey sampling techniques are interesting alternative to signal compression techniques.

The joint use of survey sampling techniques in conjunction with semiparametric models for inference is not new [5, 6]. To the best of our knowledge, however, this is the first attempt to take advantage of survey sampling to enable targeted learning when the data set is so large that computational problems arise. In contrast to naive sub-sampling, sampling designs with unequal probabilities

offer a control over the efficiency of estimators. In this light, we propose an alternative to the so called online version of targeted learning [21].

**Organization.** Section 2 presents our procedure for practical targeted learning from large data sets by survey sampling and the central limit theorem which enables the construction of confidence intervals. Section 3 illustrates Section 2 with two examples, where the parameters of interest are variable importance measures of a (binary or continuous) exposure on an outcome. Section 4 summarizes the results of a simulation study. The proofs are given in appendix.

## 2 Practical targeted learning

Throughout the article, we denote $\mu f \equiv \int f d\mu$ and $\|f\|_{2,\mu} \equiv (\mu f^2)^{1/2}$ for any measure $\mu$ and function $f$ (measurable and integrable wrt $\mu$).

### 2.1 Survey sampling from the large data set and construction of the estimator

**Rejective sampling.** Let $n(N)$ be a deterministic, user-supplied number of observations to select by survey sampling. It is a practical, computationally tractable sample size as opposed to the unpractical, huge $N$. Because our results are asymptotic we impose that, as $N \to \infty$,

$$n(N) \to \infty \quad \text{and} \quad \frac{n(N)}{N} \to 0.$$

In the rest of this article, we will simply denote $n$ for $n(N)$.

We employ a specific survey sampling scheme called *rejective sampling* [14, 3]. The random selection of observations from the complete data set can depend on easily accessible summary measures $V_1, \ldots, V_N \in \mathcal{V}$ attached to $O_1, \ldots, O_N$. Typically, $V_1, \ldots, V_N$ take finitely many different values or are low-dimensional, and the implementation of the database is structured/organized based on the values of $V_1, \ldots, V_N$.

Let $h$ be a (measurable) function on $\mathcal{V}$ such that $h(\mathcal{V}) \subset [c(h), \infty)$ for some constant $c(h) > 0$. For each $1 \leq i \leq N$, define

$$p_i \equiv \frac{nh(V_i)}{N}.$$

For $N$ large enough, $p_1, \ldots, p_N \in (0, 1)$. Introduce

- $\varepsilon_1, \ldots, \varepsilon_N$ independently drawn, conditionally on $V_1, \ldots, V_N$, from the Bernoulli distributions with parameters $p_1, \ldots, p_N$, respectively;
- $(\eta_1, \ldots, \eta_N)$ drawn, conditionally on $V_1, \ldots, V_N$, from the conditional distribution of $(\varepsilon_1, \ldots, \varepsilon_N)$ given $\sum_{i=1}^N \varepsilon_i = n$.

The subset of $n$ observations randomly selected by rejective sampling is $\{O_i : \eta_i = 1, 1 \leq i \leq N\}$. It is associated with the so-called HT empirical measure defined by

$$P_{R_N}^{\boldsymbol{p}} \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\eta_i}{p_i} \operatorname{Dirac}(O_i). \tag{2}$$

Note that $P_{R_N}^{\boldsymbol{p}}$ is not necessarily a probability measure. However, if $h \equiv 1$ then $P_{R_N}^{\boldsymbol{p}}$ is a probability measure, and rejective sampling is equivalent to selecting $n$ observations among $O_1, \ldots, O_N$ uniformly.

For computational reasons, it is not desirable that the event "$\sum_{i=1}^{N} \varepsilon_i = n$" be too unlikely. Lemma 3.1 in [14] shows that the conditional probability of the event "$\sum_{i=1}^{N} \varepsilon_i = k$" is maximized when $k$ equals the conditional expectation of $\sum_{i=1}^{N} \varepsilon_i$, in which case the conditional probability is asymptotically equivalent to $(2\pi \sum_{i=1}^{N} p_i(1-p_i))^{-1/2}$. Because the conditional expectation of $n^{-1} \sum_{i=1}^{N} \varepsilon_i$ equals $n^{-1} \sum_{i=1}^{N} p_i = N^{-1} \sum_{i=1}^{N} h(V_i)$, which converges $P_0$-almost surely to $E_{P_0}[h(V)]$, it is thus good practice to choose function $h$ in such a way that $E_{P_0}[h(V)]$ be close to 1. When $V_1, \ldots, V_N$ take finitely many different values, it is easy to estimate accurately $E_{P_0}[h(V)]$ on an independent sample and, therefore, to adapt $h$ so that $E_{P_0}[h(V)] \approx 1$.

**Practical, targeted estimator.** Assume that we have constructed $P_n^* \in \mathcal{M}$ targeted to $\psi_0$ in the sense that

$$P_{R_N}^{\boldsymbol{p}} D(P_n^*) = o_P(1/\sqrt{n}). \tag{3}$$

We define $\psi_n^* \equiv \Psi(P_n^*)$ as our substitution estimator. This construction frames $\psi_n^*$ in the paradigm of the targeted minimum loss estimation methodology [23, 22].

## 2.2 Main theorem

Consider a class $\mathcal{F}$ of functions mapping a measured space $\mathcal{X}$ to $\mathbb{R}$. Set $\delta > 0$ and a semi-metric $d$ or a norm. We denote $N(\varepsilon, \mathcal{F}, d)$ the $\varepsilon$-covering number of $\mathcal{F}$ wrt $d$, *i.e.*, the minimum number of $d$-balls of radius $\varepsilon$ needed to cover $\mathcal{F}$. The corresponding entropy integral for $\mathcal{F}$ evaluated at $\delta$ is $J(\delta, \mathcal{F}, d) \equiv \int_0^\delta \sqrt{\log N(\varepsilon, \mathcal{F}, d)} d\varepsilon$.

Let $\mathcal{R} : \mathcal{M}^2 \to \mathbb{R}$ be given by

$$\mathcal{R}(P, P') \equiv \Psi(P') - \Psi(P) - \int D(P)(dP' - dP) \tag{4}$$

where the influence function $D(P)$ is defined before (1). The real number $\mathcal{R}(P_n^*, P_0)$ can be interpreted as a second-order term in an expansion of $\psi_n^* = \Psi(P_n^*)$ around $P_0$. By (1), we focus on functionals $\Psi$ such that $\mathcal{R}(P, P_t) = o(t)$ for a wide class of one-dimensional paths $\{P_t : t \in ]-c, c[\} \subset \mathcal{M}$ such that $P_t|_{t=0} = P$. This statement is clarified in the examples of Section 3.

We suppose the existence of $\mathcal{F} \subset \{D(P) : P \in \mathcal{M}\}$ satisfying the three following assumptions:

**A1 (complexity)** $\mathcal{F}$ is separable, for every $f \in \mathcal{F}$, $P_0 f^2 h^{-1} < \infty$, and $J(1, \mathcal{F}, \|\cdot\|_{2,P_0}) < \infty$.

**A2 (uniform convergence of empirical metric)** For every $f, f' \in \mathcal{F}$, if

$$\rho_N^2(f, f') \equiv \frac{1}{N} \sum_{i=1}^{N} (f(O_i) - f'(O_i))^2 \tag{5}$$

then, $P_0$-almost surely,

$$\sup_{f,f' \in \mathcal{F}} \left| \frac{\rho_N(f, f')}{\|f - f'\|_{2,P_0}} - 1 \right| \underset{N \to \infty}{\longrightarrow} 0.$$

**A3 (first order convergence)** With $P_0$-probability tending to 1, $D(P_n^*) \in \mathcal{F}$, and there exists $f_1 \in \mathcal{F}$ such that $\|D(P_n^*) - f_1\|_{2,P_0} = o_P(1)$. Moreover, one knows a conservative estimator $\Sigma_n$ of $\sigma_1^2 \equiv P_0 f_1^2 h^{-1}$.

Under **A1**, we can define $\Sigma : \mathcal{F}^2 \to \mathbb{R}$ given by

$$\Sigma(f, f') \equiv P_0 f f' h^{-1}. \tag{6}$$

In particular, $\sigma_1^2$ in **A3** equals $\Sigma(f_1, f_1)$. An additional assumption is needed:

**A4 (second order term)** There exists a real-valued random variable $\gamma_n$ converging in probability to $\gamma_1 \neq 1$ and such that $\gamma_n(\psi_n^* - \psi_0) + \mathcal{R}(P_n^*, P_0) = o_P(1/\sqrt{n})$. Moreover, one knows an estimator $\Gamma_n$ such that $\Gamma_n - \gamma_n = o_P(1)$.

We can now state our main theorem.

**Theorem 1.** *Assume that* **A1**, **A2**, **A3** *and* **A4** *are met. Then it holds that* $(1 - \gamma_n)\sqrt{n}(\psi_n^* - \psi_0)$ *converges in law to the centered Gaussian distribution with variance* $\sigma_1^2$. *Consequently, for any* $\alpha \in (0, 1)$,

$$\left[ \psi_n^* \pm \frac{\xi_{1-\alpha/2} \sqrt{\Sigma_n}}{(1 - \Gamma_n)\sqrt{n}} \right]$$

*is a confidence interval with asymptotic coverage no less than* $(1 - \alpha)$.

**Comments.** Assumption **A1** is typical in semiparametric inference, and should be interpreted as a constraint on the complexity of $\mathcal{F}$. Theorem 1 relies on the convergence of an empirical process, see Theorem 2. The proof of Theorem 2 uses a chaining argument, and **A2** allows to upper-bound the resulting *random* term $J(\delta, \mathcal{F}, \rho_N)$ by a *deterministic* term $J(\delta, \mathcal{F}, \|\cdot\|_{2,P_0})$. We say that a class $\mathcal{C}$ has finite uniform entropy integral if it admits an envelope function $F$ and

$$\int_0^\infty \sup_\rho \sqrt{\log N(\epsilon \|F\|_{2,\rho}, \mathcal{C}, \|\cdot\|_{2,\rho})} d\epsilon < \infty,$$

where the supremum is over all probability measures $\rho$ on $\mathcal{O}$ such that $\|F\|_{2,\rho} > 0$. Assumption **A2** can be replaced by the alternative

**A2*** The class $\mathcal{F}$ has a finite uniform entropy integral.

VC-classes of uniformly bounded functions satisfy **A2*** [25, Section 2.6]. Finally, **A3** and **A4** are technical conditions required by the TMLE procedure. The former is not as mild as one may think at first sight, because the conservative estimation of $\sigma_1^2$ is not trivial. For instance, it is not guaranteed in general that the substitution estimator

$$\Sigma_n \equiv P_{R_N}^{\boldsymbol{p}} D(P_n^*)^2 h^{-1} \tag{7}$$

estimates conservatively $\sigma_1^2$. Relying on the non-parametric bootstrap is not a solution either in general.

We argued that $\mathcal{R}(P_n^*, P_0)$ should be interpreted as a second order term. In the simplest examples, this is literally the case and assuming $\mathcal{R}(P_n^*, P_0) = o_P(1/\sqrt{n})$ is natural, see for instance Section 3.1. Sometimes, $\mathcal{R}(P_n^*, P_0)$ must be corrected by adding $\gamma_n(\psi_n^* - \psi_0)$ so that it becomes natural to assume that the corrected expression is $o_P(1/\sqrt{n})$, see for instance Section 3.2.

and **A4** is met with $\gamma_n = 0$, see for instance Section 3.1. Allowing $\gamma_n$ to differ from 0 gives more flexibility. In Section 3, we give additional conditions which imply **A4**.

Knowing the asymptotic variance of $(1-\gamma_n)\sqrt{n}(\psi_n^* - \psi_0)$ allows to discuss further the choice of $h$. Introduce

$$f_2(V) \equiv \sqrt{E_{P_0}\left[f_1(O)^2|V\right]}, \tag{8}$$

which satisfies $\sigma_1^2 = P_0 f_1^2 h^{-1} = P_0 f_2^2 h^{-1}$. The Cauchy-Schwarz inequality yields

$$(P_0 f_2)^2 \leq P_0 f_2^2 h^{-1} \times P_0 h = \sigma_1^2 \times P_0 h, \tag{9}$$

and equality occurs when $f_2$ and $h$ are linearly dependent. Moreover, it should hold that $P_0 h = 1$. In view of (9), the optimal $h$ is $f_2/P_0 f_2$, assuming that $P_0 f_2 > 0$ (otherwise, $\sigma_1^2 = 0$). This argument neglects the second-order dependence of $\gamma_n$ on $h$. In practice, we would first sample $n_0$ data using $h_0 \equiv 1$, use them to estimate $f_2$ and $P_0 f_2$ with $f_{2,n_0}$ and $Z_{2,n_0}$, then finally define $h \equiv f_{2,n_0}/Z_{2,n_0}$ and exclude the sampled data from $\{O_1, \ldots, O_N\}$.

The following expansion taken from the proof of Theorem 1 partly explains why $\sigma_1^2$ is the asymptotic variance of $(1-\gamma_n)\sqrt{n}(\psi_n^* - \psi_0)$: denoting by $P_0^\varepsilon$ the shared distribution of $(O_1, \varepsilon_1), \ldots, (O_N, \varepsilon_N)$, it holds for any $f$ in $\mathcal{F}$ that

$$\begin{aligned}
\mathrm{Var}_{P_0^\varepsilon}\left(\frac{1}{N}\sum_{i=1}^N \frac{f(O_i)\varepsilon_i}{p_i}\right) &= \frac{1}{N}\mathrm{Var}_{P_0^\varepsilon}\left(\frac{f(O_1)\varepsilon_1}{p_1}\right) \\
&= \frac{1}{N}\left(E_{P_0}\left[f^2(O_1)\left(\frac{1}{p_1}-1\right)\right] + \mathrm{Var}_{P_0}\left(f(O)\right)\right).
\end{aligned}$$

If, contrary to facts, we could take $p_1 \equiv 1$ (or, equivalently, $n \equiv N$ and $h \equiv 1$), then the asymptotic variance of the resulting TMLE estimator would be of the form $N^{-1}\mathrm{Var}_{P_0}\left(f(O)\right)$ for some limit $f$, as typically expected. In Section 2.1 $p_1$ is chosen in such a way that $1/p_1$ is typically much larger than 1. Actually, the above RHS expression at $f \equiv f_1$ rewrites

$$\frac{1}{n}\left(P_0 f_1^2 h^{-1} + \frac{n}{N}(P_0 f_1)^2\right) = \frac{1}{n}\left(\sigma_1^2 + o(1)\right). \tag{10}$$

Note the absence of a centering term in $P_0 f_1^2 h^{-1}$.

# 3 Two examples

We illustrate Theorem 1 with the inference of two variable importance measures of an exposure, either binary, in Section 3.1, or continuous, in Section 3.2. In both examples, the $i$th observation $O_i$ writes $(W_i, A_i, Y_i) \in \mathcal{O} \equiv \mathcal{W} \times \mathcal{A} \times [0,1]$. Here, $W_i \in \mathcal{W}$ is the $i$th context, $A_i \in \mathcal{A}$ is the $i$th exposure and $Y_i \in [0,1]$ is the $i$th outcome. In the binary case, $\mathcal{A} \equiv \{0,1\}$. In the continuous case, $\mathcal{A} \ni 0$ is a bounded subset of $\mathbb{R}$ containing 0, which serves as a reference level of exposure. Typically, in biostatistics or epidemiology, $W_i$ could be the baseline covariate describing the $i$th subject, $A_i$ could describe her assignment (*e.g.*, treatment or placebo when $\mathcal{A} = \{0,1\}$ or dose-level when $\mathcal{A} \subset \mathbb{R}$) or exposure (*e.g.*, exposed or not when $\mathcal{A} = \{0,1\}$ or level of exposure when $\mathcal{A} \subset \mathbb{R}$), and $Y_i$ could quantify her biological response.

## 3.1 Variable importance measure of a binary exposure

In this section, $\mathcal{A} \equiv \{0,1\}$ and $\psi_0$ equals

$$\psi_0^b \equiv E_{P_0}\left[E_{P_0}\left[Y|A=1,W\right] - E_{P_0}\left[Y|A=0,W\right]\right] \tag{11}$$

(the superscript "$b$" stands for "binary"). Now, let $\mathcal{M}$ be the subset of the set of finite measures on $\mathcal{O} \equiv \mathcal{W} \times \{0,1\} \times [0,1]$ equipped with the Borel $\sigma$-field such that every $P \in \mathcal{M}$ puts mass on all events of the form $B_1 \times \{a\} \times B_2$ ($a = 0, 1$, $B_1$ and $B_2$ Borel sets of $\mathcal{W}$ and $[0,1]$). It contains the set of all possible data-generating distributions for $O_1$ such that the conditional distribution of $A$ given $W$ is not deterministic, including $P_0$. For each $P \in \mathcal{M}$, we denote $P_W$, $P_{A|W}$ and $P_{Y|A,W}$ the marginal measure of $W$ and conditional measures of $A$ and $Y$ given $W$ and $(A,W)$, respectively. (The conditional measure $P_{A|W}$ is $P(\mathcal{O})$ times the conditional law of $A$ given $W$ under the probability distribution $P/P(\mathcal{O})$. The conditional measure $P_{Y|A,W}$ is defined analogously.) We see $\psi_0^b$ as the value at $P_0$ of the functional $\Psi^b$ characterized over $\mathcal{M}$ by

$$\Psi^b(P) \equiv \int_{\mathcal{W}} \left( \int_{[0,1]} y \left( dP_{Y|A=1,W=w}(y) - dP_{Y|A=0,W=w}(y) \right) \right) dP_W(w). \tag{12}$$

In particular, if $P$ is a possible data-generating *distribution* for $O_1$ (*i.e.*, if $P(\mathcal{O}) = 1$), then

$$\Psi^b(P) = E_P\left[E_P\left[Y|A=1,W\right] - E_P\left[Y|A=0,W\right]\right].$$

Moreover, under additional causal assumptions, $\Psi^b(P)$ can be interpreted as the additive causal effect of the exposure on the response, see [17, 22].

Two infinite-dimensional features of every $P \in \mathcal{M}$ will play an important role in the analysis. Namely, for each $P \in \mathcal{M}$ and $(w,a) \in \mathcal{W} \times \mathcal{A}$, we introduce and denote $g_P(0|w) \equiv P_{A|W=w}(\{0\})$, $g_P(1|w) \equiv P_{A|W=w}(\{1\})$, and $Q_P(a,w) \equiv \int_{[0,1]} y dP_{Y|A=a,W=w}(y)$. In particular if $P(\mathcal{O}) = 1$, then $g_P(1|W) = P(A=1|W)$ is the conditional probability that the binary exposure equal one and $Q_P(A,W) = E_P\left[Y|A,W\right]$ is the conditional expectation of the response given exposure and context.

**Pathwise differentiability.** The functional $\Psi^b$ is pathwise differentiable at each $P \in \mathcal{M}$ wrt the maximal tangent space $L_0^2(P)$ (the space of functions $s : \mathcal{O} \to \mathbb{R}$ such that $Ps = 0$ and $Ps^2 < \infty$) in the following sense [22, Chapter 5 and Section A.3]:

**Lemma 1.** *Fix $P \in \mathcal{M}$ and introduce the influence curve $D^b(P) \in L_0^2(P)$ given by $D^b(P) \equiv D_1^b(P) + D_2^b(P)$ with*

$$\begin{aligned} D_1^b(P)(O) &\equiv Q_P(1, W) - Q_P(0, W) - \Psi^b(P), \\ D_2^b(P)(O) &\equiv (Y - Q_P(A, W))\frac{2A-1}{g_P(A|W)}. \end{aligned}$$

*For every uniformly bounded $s \in L_0^2(P)$ and every $t \in ]-\|s\|_\infty^{-1}, \|s\|_\infty^{-1}[$, define $P_{s,t} \in \mathcal{M}$ by setting*

$$\frac{dP_{s,t}}{dP} = 1 + ts.$$

*It holds that $t \mapsto \Psi^b(P_{s,t})$ is differentiable at 0 (as a function from $\mathbb{R}$ to $\mathbb{R}$) with a derivative at 0 equal to $PD^b(P)s$.*

*The asymptotic variance of any regular estimator of $\Psi^b(P_0)$ is larger than the Cramér-Rao lower-bound $P_0 D^b(P_0)^2$. Moreover, for any $P, P' \in \mathcal{M}$,*

$$PD^b(P') = \Psi^b(P) - \Psi^b(P') + P(2A-1)(Q_{P'} - Q_P)\left(\frac{1}{g_P} - \frac{1}{g_{P'}}\right). \tag{13}$$

*Consequently if $PD^b(P') = 0$, then $\Psi^b(P') = \Psi^b(P)$ whenever $g_{P'} = g_P$ **or** $Q_{P'} = Q_P$.*

The last statement is called a "double-robustness property". Let $\mathcal{R}^b : \mathcal{M}^2 \to \mathbb{R}$ be given by

$$\mathcal{R}^b(P, P') \equiv \Psi^b(P') - \Psi^b(P) - (P' - P)D^b(P), \tag{14}$$

as in (4). In particular,

$$\begin{aligned} \mathcal{R}^b(P, P_{s,t}) &= \Psi^b(P_{s,t}) - \Psi^b(P) - (P_{s,t} - P)D^b(P) \\ &= \Psi^b(P_{s,t}) - \Psi^b(P) - tPD^b(P)s = o(t), \end{aligned}$$

showing that (1) is met.

Furthermore, (13) and $PD^b(P) = 0$ imply

$$\mathcal{R}^b(P, P') = P'(2A-1)(Q_{P'} - Q_P)\left(\frac{1}{g_{P'}} - \frac{1}{g_P}\right).$$

In the context of this example, **A4** is fulfilled with $\gamma_n \equiv 0$ (hence $\Gamma_n \equiv 0$ and $\gamma_1 = 0$) when

$$\mathcal{R}^b(P_n^*, P_0) = P_0(2A-1)(Q_{P_n^*} - Q_{P_0})\left(\frac{1}{g_{P_n^*}} - \frac{1}{g_{P_0}}\right) = o_P(1/\sqrt{n}). \tag{15}$$

Through the product, we will draw advantage of the synergistic convergences of $Q_{P_n^*}$ to $Q_{P_0}$ and $g_{P_n^*}$ to $g_{P_0}$ (by the Cauchy-Schwarz inequality for example). Note that if $g_{P_0}$ is known, then we can impose that $g_{P_n^*} = g_{P_0}$ and $\mathcal{R}^b(P_n^*, P_0) = 0$ exactly.

**Construction of the targeted estimator.** Let $\mathcal{Q}^w$ and $\mathcal{G}^w$ be two user-supplied classes of functions mapping $\mathcal{A} \times \mathcal{W}$ to $[0,1]$. We impose that the elements of $\mathcal{Q}^w$ are uniformly bounded away from 0 and 1. Similarly, we impose that the elements of $\mathcal{G}^w$ are uniformly bounded away from 0. Let $\ell$ be the logistic loss function given by

$$-\ell(u,v) \equiv u \log(v) + (1-u)\log(1-v)$$

(all $u, v \in [0,1]$ with conventions $\log(0) = -\infty$ and $0 \log(0) = 0$).

We first estimate $Q_{P_0}$ and $g_{P_0}$ with $Q_n$ and $g_n$ built upon $P^{\boldsymbol{p}}_{R_N}$, $\mathcal{Q}^w$ and $\mathcal{G}^w$. For instance, one could simply minimize (weighted) empirical risks and define

$$\begin{aligned} Q_n &\equiv \operatorname*{argmin}_{Q \in \mathcal{Q}^w} P^{\boldsymbol{p}}_{R_N} \ell(Y, Q(A,W)) = \operatorname*{argmin}_{Q \in \mathcal{Q}^w} \sum_{i=1}^N \frac{\eta_i}{p_i} \ell(Y_i, Q(A_i, W_i)), \\ g_n &\equiv \operatorname*{argmin}_{g \in \mathcal{G}^w} P^{\boldsymbol{p}}_{R_N} \ell(A, g(A|W)) = \operatorname*{argmin}_{g \in \mathcal{G}^w} \sum_{i=1}^N \frac{\eta_i}{p_i} \ell(A_i, g(A_i|W_i)) \end{aligned}$$

(assuming that the argmins exist). Alternatively, one could prefer minimizing cross-validated (weighted) empirical risks. This is beyond the scope of this article but will be studied in future work. We also estimate the marginal distribution $P_{0,W}$ of $W$ under $P_0$ with

$$P^{\boldsymbol{p}}_{R_N,W} \equiv \frac{1}{N} \sum_{i=1}^N \frac{\eta_i}{p_i} \operatorname{Dirac}(W_i). \tag{16}$$

Let $P_n^0$ be a measure such that $Q_{P_n^0} = Q_n$ and $P^0_{n,W} = P^{\boldsymbol{p}}_{R_N,W}$. Then

$$\Psi^b(P_n^0) = \frac{1}{N} \sum_{i=1}^N \frac{\eta_i}{p_i} \left(Q_n(1,W_i) - Q_n(0,W_i)\right) \tag{17}$$

is an estimator of $\psi_0^b$, whose construction is not tailored/targeted to $\psi_0^b$. It is now time to target the inference procedure.

Targeting the inference procedure consists in modifying $P_n^0$ in such a way that the resulting $P_n^*$ satisfies (3) with $D^b$ substituted for $D$. We first note that, by construction of $P_n^0$,

$$P^{\boldsymbol{p}}_{R_N} D_1^b(P_n^0) = P^{\boldsymbol{p}}_{R_N,W} D_1^b(P_n^0) = 0.$$

This equality is equivalent to (17).

The construction of $P_n^*$ based on $P_n^0$ reduces to ensuring $P^{\boldsymbol{p}}_{R_N} D_2^b(P_n^*) = o_P(1/\sqrt{n})$. We achieve this objective by fluctuating the conditional measure of $Y$ given $(A,W)$ only. For this, we introduce the one-dimensional parametric model $\{Q_n(t) : t \in \mathbb{R}\}$ given by

$$\operatorname{logit} Q_n(t)(A,W) = \operatorname{logit} Q_n(A,W) + t \frac{2A-1}{g_n(A|W)}.$$

This parametric model fluctuates $Q_n$ in the direction of $\frac{2A-1}{g_n(A|W)}$ in the sense that $Q_n(0) = Q_n$ and

$$\frac{d}{dt} \ell(Y, Q_n(t)(A,W)) = (Y - Q_n(t)(A,W)) \frac{2A-1}{g_n(A|W)} \tag{18}$$

for all $t \in \mathbb{R}$. The optimal move along the fluctuation is indexed by

$$t_n \equiv \arg\min_{t \in \mathbb{R}} P^{\boldsymbol{p}}_{R_N} \ell(Y, Q_n(t)(A, W)) \tag{19}$$

(note that the random function $t \mapsto P^{\boldsymbol{p}}_{R_N} \ell(Y, Q_n(t)(A, W))$ is strictly convex).

Define $Q^*_n \equiv Q_n(t_n)$ and let $P^*_n$ be any element $P$ of $\mathcal{M}$ such that $Q_P = Q^*_n$, $g_P = g_n$ and $P_W = P^0_{n,W} = P^{\boldsymbol{p}}_{R_N,W}$. Our final estimator is

$$\psi^*_n \equiv \Psi^b(P^*_n) = \frac{1}{N} \sum_{i=1}^{N} \frac{\eta_i}{p_i} \left( Q^*_n(1, W_i) - Q^*_n(0, W_i) \right).$$

By definition of $t_n$ and (18), we have $P^{\boldsymbol{p}}_{R_N} D^b_1(P^*_n) = 0$ (just like $P^{\boldsymbol{p}}_{R_N} D^b_1(P^0_n) = 0$) and

$$P^{\boldsymbol{p}}_{R_N} \left. \frac{d}{dt} \ell(Y, Q_n(t)(A, W)) \right|_{t=t_n} P^{\boldsymbol{p}}_{R_N} D^b_2(P^*_n) = 0$$

(whereas it is very unlikely that $P^{\boldsymbol{p}}_{R_N} D^b_2(P^0_n)$ be equal to zero). Consequently, (3) is met because

$$P^{\boldsymbol{p}}_{R_N} D^b(P^*_n) = 0.$$

Theorem 1 is tailored to the present setting in Section 3.3.

### 3.2 Variable importance measure of a continuous exposure

In this section, $\mathcal{A} \subset \mathbb{R}$ is a bounded subset of $\mathbb{R}$ containing 0, which serves as a reference value. Moreover, we assume that $P_{0,A|W}(A \neq 0|W) > 0$ $P_{0,W}$-almost surely and the existence of a constant $c(P_0) > 0$ such that $P_{0,A|W}(A = 0|W) \geq c(P_0)$ $P_{0,W}$-almost surely. Introduced in [12, 10], the true parameter of interest is

$$\begin{aligned} \psi^c_0 &\equiv \arg\min_{\beta \in \mathbb{R}} E_{P_0} \left[ \left( Y - E_{P_0}[Y|A = 0, W] - \beta A \right)^2 \right] \\ &= \arg\min_{\beta \in \mathbb{R}} E_{P_0} \left[ \left( E_{P_0}[Y|A, W] - E_{P_0}[Y|A = 0, W] - \beta A \right)^2 \right] \end{aligned} \tag{20}$$

(the superscript "$c$" stands for "continuous").

Let $\mathcal{M}$ be the set of finite measures $P$ on $\mathcal{O} \equiv \mathcal{W} \times \mathcal{A} \times [0, 1]$ equipped with the Borel $\sigma$-field such that there exists a constant $c(P) > 0$ guaranteeing that the marginal measure of $\{w \in \mathcal{W} : P_{A|W=w}(\mathcal{A} \setminus \{0\}) > 0 \text{ and } P_{A|W=w}(\{0\}) \geq c(P)\}$ under $P_W$ equals $P(\mathcal{O})$. In particular, $P_0 \in \mathcal{M}$ by the above assumption.

We see $\psi^c_0$ as the value at $P_0$ of the functional $\Psi^c$ characterized over $\mathcal{M}$ by

$$\Psi^c(P) \equiv \arg\min_{\beta \in \mathbb{R}} \int_{\mathcal{A} \times \mathcal{W}} \left( Q_P(a, w) - Q_P(0, w) - \beta a \right)^2 dP_{A|W=w}(a) dP_W(w), \tag{21}$$

using the notation of Section 3.1. By Proposition 1 in [12], for each $P \in \mathcal{M}$,

$$\Psi^c(P) = \frac{\int_{\mathcal{A} \times \mathcal{W}} a (Q_P(a, w) - Q_P(0, w)) dP_{A|W=w}(a) dP_W(w)}{\int_{\mathcal{A} \times \mathcal{W}} a^2 dP_{A|W=w}(a) dP_W(w)}.$$

If $P$ is a *distribution*, then

$$\Psi^c(P) = \frac{E_P\left[A(Q_P(A,W) - Q_P(0,W))\right]}{E_P\left[A^2\right]}.$$

For clarity, we introduce some notation. For each $P \in \mathcal{M}$ and $(w,a) \in \mathcal{W} \times \mathcal{A}$, $\mu_P(w) \equiv \int_{\mathcal{A}} a dP_{A|W=w}(a)$, and $g_P(0|w) \equiv P_{A|W=w}(\{0\})$, $\zeta^2(P) \equiv \int_{\mathcal{A}} a^2 dP_{A|W=w}(a)$. If $P(\mathcal{O}) = 1$, then $\mu_P(W) = E_P[A|W]$, $g_P(0|W) = P(A=0|W)$, and $\zeta^2(P) = E_P\left[A^2\right]$.

**Pathwise differentiability.** A result similar to Lemma 1 [see 12, Proposition 1] guarantees that $\Psi^c$ is pathwise differentiable like $\Psi^b$ with influence curves $D^c(P) \equiv D_1^c(P) + D_2^c(P) \in L_0^2(P)$,

$$\begin{aligned}
\zeta^2(P) D_1^c(P)(O) &\equiv A\left(Q_P(A,W) - Q_P(0,W) - A\Psi^c(P)\right), \\
\zeta^2(P) D_2^c(P)(O) &\equiv (Y - Q_P(A,W))\left(A - \frac{\mu_P(W)\mathbf{1}\{A=0\}}{g_P(0|W)}\right)
\end{aligned}$$

(all $P \in \mathcal{M}$). Let $\mathcal{R}^c : \mathcal{M}^2 \to \mathbb{R}$ be characterized by

$$\mathcal{R}^c(P,P') \equiv \Psi^c(P') - \Psi^c(P) - (P' - P)D^c(P).$$

as in (4) and (14). As in the previous example, $\mathcal{R}^c$ satisfies (1) and, for every $P, P' \in \mathcal{M}$,

$$\begin{aligned}
\mathcal{R}^c(P,P') = \left(1 - \frac{\zeta^2(P')}{\zeta^2(P)}\right)\left(\Psi^c(P') - \Psi^c(P)\right) \\
+ \frac{1}{\zeta^2(P)} P'\left(\left(Q_{P'}(0,\cdot) - Q_P(0,\cdot)\right)\left(\mu_{P'} - \mu_P \frac{g_{P'}(0|\cdot)}{g_P(0|\cdot)}\right)\right). \quad (22)
\end{aligned}$$

Introduce

$$\gamma_n \equiv 1 - \frac{\zeta^2(P_0)}{\zeta^2(P_n^*)} \quad \text{and} \quad \Gamma_n \equiv 1 - \frac{\zeta_n^2(P_0)}{\zeta_n^2(P_n^*)}$$

where $\zeta_n^2(P_0)$ and $\zeta_n^2(P_n^*)$ estimate $\zeta^2(P_0)$ and $\zeta^2(P_n^*)$. With these choices, (22) guarantees that **A4** is fulfilled in the context of this example when $\zeta^2(P_n^*)$ converges in probability to a finite real number such that $\gamma_1 \neq 1$ and

$$\frac{1}{\zeta^2(P_n^*)} P_0\left(\left(Q_{P_0}(0,\cdot) - Q_{P_n^*}(0,\cdot)\right)\left(\mu_{P_0} - \mu_{P_n^*}\frac{g_{P_0}(0|\cdot)}{g_{P_n^*}(0|\cdot)}\right)\right) = o_P(1/\sqrt{n}).$$

Through the product, we will draw advantage of the synergistic convergences of $Q_{P_n^*}(0,\cdot)$ to $Q_{P_0}(0,\cdot)$ and $(\mu_{P_n^*}, g_{P_n^*})$ to $(\mu_{P_0}, g_{P_0})$ (by the Cauchy-Schwarz inequality for example).

**Construction of the targeted estimator.** Let $\mathcal{Q}^w$, $\mathcal{M}^w$ and $\mathcal{G}^w$ be three user-supplied classes of functions mapping $\mathcal{A} \times \mathcal{W}$, $\mathcal{W}$ and $\mathcal{W}$ to $[0,1]$, respectively. We first estimate $Q_{P_0}$, $\mu_{P_0}$ and $g_{P_0}$ with $Q_n$ and $\mu_n$ and $g_n$ built upon $P_{R_N}^{\boldsymbol{p}}$, $\mathcal{Q}^w$, $\mathcal{M}^w$ and $\mathcal{G}^w$. For instance, one could simply minimize (weighted) empirical risks and define

$$Q_n \equiv \underset{Q \in \mathcal{Q}^w}{\operatorname{argmin}} P_{R_N}^{\boldsymbol{p}} \ell(Y, Q(A,W)) = \underset{Q \in \mathcal{Q}^w}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{\eta_i}{p_i} \ell(Y_i, Q(A_i, W_i)),$$

$$\mu_n \equiv \underset{\mu\in\mathcal{M}^w}{\operatorname{argmin}} P^{\boldsymbol{p}}_{R_N}\ell(A,\mu(W)) = \underset{\mu\in\mathcal{M}^w}{\operatorname{argmin}} \sum_{i=1}^N \frac{\eta_i}{p_i}\ell(A_i,\mu(W_i)),$$

$$g_n \equiv \underset{g\in\mathcal{G}^w}{\operatorname{argmin}} P^{\boldsymbol{p}}_{R_N}\ell(\mathbf{1}\{A=0\},g(0|W)) = \underset{g\in\mathcal{G}^w}{\operatorname{argmin}} \sum_{i=1}^N \frac{\eta_i}{p_i}\ell(\mathbf{1}\{A_i=0\},g(0|W_i))$$

(assuming that the argmins exist). Alternatively, one could prefer minimizing cross-validated (weighted) empirical risks. We also estimate the marginal distribution $P_{0,W}$ of $W$ under $P_0$ with

$$P^{\boldsymbol{p}}_{R_N,W} \equiv \frac{1}{N}\sum_{i=1}^N \frac{\eta_i}{p_i}\,\mathrm{Dirac}(W_i), \tag{23}$$

and the real-valued parameter $\zeta^2(P_0)$ with $\zeta^2(P^{\boldsymbol{p}}_{R_N,X})$ where $P^{\boldsymbol{p}}_{R_N,X}$ is defined as in (23) with $X$ and $X_i$ substituted for $W$ and $W_i$.

Let $P_n^0$ be a measure such that $Q_{P_n^0} = Q_n$, $\mu_{P_n^0} = \mu_n$, $g_{P_n^0} = g_n$, $\zeta^2(P_n^0) = \zeta^2(P^{\boldsymbol{p}}_{R_N,X})$, $P^0_{n,W} = P^{\boldsymbol{p}}_{R_N,W}$, *and from which we can sample $A$ conditionally on $W$*. Picking up such a $P_n^0$ is an easy technical task, see [12, Lemma 5] for a computationally efficient choice. Then the initial estimator $\Psi^b(P_n^0)$ of $\psi_0^b$ can be computed with high accuracy by Monte-Carlo. It suffices to sample a large number $B$ (say $B = 10^7$) of independent $(A^{(b)}, W^{(b)})$ by *(i)* sampling $W^{(b)}$ from $P^0_{n,W} = P^{\boldsymbol{p}}_{R_N,W}$ then *(ii)* sampling $A^{(b)}$ from the conditional distribution of $A$ given $W = W^{(b)}$ under $P_n^0$ repeatedly for $b = 1,\ldots,B$ and to make the approximation

$$\Psi^c(P_n^0) \approx \frac{B^{-1}\sum_{b=1}^B A^{(b)}(Q_n(A^{(b)},W^{(b)}) - Q_n(0,W^{(b)}))}{\zeta^2(P_n^0)}. \tag{24}$$

However, the construction $\Psi^c(P_n^0)$ is not tailored/targeted to $\psi_0^c$ yet. It is now time to target the inference procedure.

Targeting the inference procedure consists in modifying $P_n^0$ in such a way that the resulting $P_n^*$ satisfies (3) with $D^c$ substituted for $D$. We proceed iteratively. Suppose that $P_n^k$ has been constructed for some $k \geq 0$. We fluctuate $P_n^k$ with the one-dimensional parametric model $\{P_n^k(t) : t \in \mathbb{R}, t^2 \leq c(P_n^k)/\|D^c(P_n^k)\|_\infty\}$ characterized by

$$\frac{dP_n^k(t)}{dP_n^k} = 1 + tD^c(P_n^k).$$

Lemma 1 in [12] shows how $Q_{P_n^k(t)}$, $\mu_{P_n^k(t)}$, $g_{P_n^k(t)}$, $\zeta^2(P_n^k(t))$ and $P^k_{n,W}(t)$ depart from their counterparts at $t = 0$. The optimal move along the fluctuation is indexed by

$$t_n^k \equiv \arg\max_t P^{\boldsymbol{p}}_{R_N}\log\left(1 + tD^c(P_n^k)\right),$$

*i.e.*, the maximum likelihood estimator of $t$ (note that the random function $t \mapsto P^{\boldsymbol{p}}_{R_N}\log(1+tD^c(P_n^k))$ is strictly concave). It results in the $(k+1)$-th update of $P_n^0$, $P_n^{k+1} \equiv P_n^k(t_n^k)$.

Contrary to what happened in the first example, see Section 3.1, there is no guarantee that a $P_n^{k+1}$ will coincide with its predecessor $P_n^k$. In this light, the updating procedure in Section 3.1 converged in one single step. Here, we assume that the iterative updating procedure converges (in $k$) in the sense

that, for $k_n$ large enough, $P^{\boldsymbol{p}}_{R_N} D^c(P_n^{k_n}) = o_P(1/\sqrt{n})$. We set $P_n^* \equiv P_n^{k_n}$. It is actually possible to come up with a one-step updating procedure (*i.e.*, an updating procedure such that $P_n^k = P_n^{k+1}$ for all $k \geq 1$) in this example too by relying on so-called universally least favorable models [20]. We adopt this multi-step updating procedure for simplicity.

We can assume without loss of generality that we can sample $A$ conditionally on $W$ from $P_n^*$. The final estimator is computed with high accuracy like $\Psi^c(P_n^0)$ previously: with $Q_n^* \equiv Q_{P_n^*}$, we sample $B$ independent $(A^{(b)}, W^{(b)})$ by *(i)* sampling $W^{(b)}$ from $P^*_{n,W}$ then *(ii)* sampling $A^{(b)}$ from the conditional distribution of $A$ given $W = W^{(b)}$ under $P_n^*$ repeatedly for $b = 1, \ldots, B$ and make the approximation

$$\psi_n^* \equiv \Psi^c(P_n^*) \approx \frac{B^{-1} \sum_{b=1}^B A^{(b)} (Q_n^*(A^{(b)}, W^{(b)}) - Q_n^*(0, W^{(b)}))}{\zeta^2(P_n^*)}. \tag{25}$$

Theorem 1 is tailored to the present setting in Section 3.3.

### 3.3 Tailoring the main theorem in the settings of Sections 3.1 and 3.2

Consider the following assumptions for the study of $\psi_n^*$ in the setting of Section 3.1:

**A1**$^b$ The classes $\mathcal{Q}^w$ and $\mathcal{G}^w$ are separable, $P_0 Q^2 h^{-1} < \infty$ and $P_0 g^2 h^{-1} < \infty$ for all $(Q, g) \in \mathcal{Q}^w \times \mathcal{G}^w$, and $J(1, \mathcal{Q}^w, \|\cdot\|_{2,P_0}) < \infty$, $J(1, \mathcal{G}^w, \|\cdot\|_{2,P_0}) < \infty$. Moreover, **A2**$^*$ is met by $\mathcal{Q}^w$ and $\mathcal{G}^w$.

**A2**$^b$ There exists $P_1 \in \mathcal{M}$ such that $\|D^b(P_n^*) - D^b(P_1)\|_{2,P_0} = o_P(1)$. Moreover, $\|Q_n^* - Q_{P_0}\|_{2,P_0} \times \|g_n - g_{P_0}\|_{2,P_0} = o_P(1/\sqrt{n})$ and one knows a conservative estimator $\Sigma_n$ of $P_0 D^b(P_1)^2 h^{-1}$.

The assumptions required for the study of $\psi_n^*$ in the setting of Section 3.2 are very similar:

**A1**$^c$ There exists a set $\mathcal{F} \subset \{D(P) : P \in \mathcal{M}\}$ such that **A1** and **A2** are verified.

**A2**$^c$ There exist $\zeta_-^2 > 0$ and $P_1 \in \mathcal{M}$ with $\zeta^2(P_1) \geq \zeta_-^2 > 0$ such that

$$\begin{gathered}
\zeta^2(P_n^*) = \zeta^2(P_1) + O_P(1/\sqrt{n}), \\
\|D^c(P_n^*) - D^c(P_1)\|_{2,P_0} = o_P(1), \\
\|Q_n^* - Q_{P_0}\|_{2,P_0} \times (\|\mu_n^* - \mu_{P_0}\|_{2,P_0} + \|g_n - g_{P_0}\|_{2,P_0}) = o_P(1/\sqrt{n}).
\end{gathered}$$

Moreover, $\Gamma_n - \gamma_n = o_P(1)$ and one knows a conservative estimator $\Sigma_n$ of $P_0 D^b(P_1)^2 h^{-1}$.

In **A2**$^b$, $Q_{P_1}$ and $g_{P_1}$ should be interpreted as the limits of $Q_{P_n^*}$ and $g_{P_n^*}$. Likewise, $Q_{P_1}$, $\mu_{P_1}$ and $g_{P_1}$ in **A2**$^c$ should be interpreted as the limits of $Q_{P_n^*}$, $\mu_{P_n^*}$ and $g_{P_n^*}$.

**Corollary 1.** *Set $\alpha \in (0,1)$. In the setting of Section 3.1 and under* **A1**$^b$, **A2**$^b$,

$$\left[ \psi_n^* \pm \frac{\xi_{1-\alpha/2} \sqrt{\Sigma_n}}{\sqrt{n}} \right]$$

*is a confidence interval for $\psi_0^b$ with asymptotic coverage no less than $(1-\alpha)$. In the setting of Section 3.2 and under* $\mathbf{A1}^c$, $\mathbf{A2}^c$,

$$\left[\psi_n^* \pm \frac{\xi_{1-\alpha/2}\sqrt{\Sigma_n}}{(1-\Gamma_n)\sqrt{n}}\right]$$

*is a confidence interval for $\psi_0^c$ with asymptotic coverage no less than $(1-\alpha)$.*

# 4 Simulation study

We illustrate the methodology with the inference of the variable importance measure of a continuous exposure presented in Section 3.2. We consider three data-generating distributions $P_{0,1}$, $P_{0,2}$ and $P_{0,3}$ of a data-structure $O = (W, A, Y)$. The three distributions differ only in terms of the conditional variance of $Y$ given $(A, W)$, but do so drastically. Specifically, $O = (W, A, Y)$ drawn from $P_{0,j}$ $(j = 1, 2, 3)$ is such that

- $W \equiv (V, W_1, W_2)$ with $P_0(V = 1) = 1/6$, $P(V = 2) = 1/3$, $P(V = 3) = 1/2$ and, conditionally on $V$, $(W_1, W_2)$ is a Gaussian random vector with mean $(0, 0)$ and variance $\left(\begin{smallmatrix} 1 & -0.2 \\ -0.2 & 1 \end{smallmatrix}\right)$ (if $V = 1$), $(1, 1/2)$ and $\left(\begin{smallmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{smallmatrix}\right)$ (if $V = 2$), $(1/2, 1)$ and $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ (if $V = 3$);
- conditionally on $W$, $A = 0$ with probability 80% if $W_1 \geq 1.1$ and $W_2 \geq 0.8$ and 10% otherwise; moreover, conditionally on $W$ and $A \neq 0$, $A - 1$ is drawn from the $\chi^2$-distribution with 1 degree of freedom and non-centrality parameter $\sqrt{(W_1 - 1.1)^2 + (W_2 - 0.8)^2}$;
- conditionally on $(W, A)$, $Y$ is a Gaussian random variable with mean $E_{P_0}[Y|A, W] \equiv A(W_1 + W_2)/6 + W_1 + W_2/4 + \exp((W_1 + W_2)/10)$ and standard deviation
  - 1.5 (if $V = 1$), 1 (if $V = 2$) and 0.5 (if $V = 3$) for $j = 1$;
  - 1 (if $V = 1$), 5 (if $V = 2$) and 10 (if $V = 3$) for $j = 2$;
  - 50 (if $V = 1$), 10 (if $V = 2$) and 1 (if $V = 3$) for $j = 3$.

The unique true parameter is $\psi_0^c = \Psi^c(P_{0,1}) = \Psi^c(P_{0,2}) = \Psi^c(P_{0,3})$. It equals approximately 0.1204.

For $B = 10^3$ and each $j = 1, 2, 3$, we repeat independently the following steps:

1. simulate a data set of $N = 10^7$ independent observations drawn from $P_{0,j}$;
2. extract $n_0 \equiv 10^3$ observations from the data set by survey sampling with $h_0 \equiv 1$, and based on these observations:
   (a) apply the procedure described in Section 3.2 and retrieve $D^c(P_{n_0}^{k_{n_0}})$;
   (b) set $f_{n_0,1} \equiv D^c(P_{n_0}^{k_{n_0}})$ and regress $f_{n_0,1}(O)^2$ on $V$, call $f_{n_0,2}$ the square root of the resulting conditional expectation, see (8);
   (c) estimate the marginal distribution of $V$, estimate $P_0 f_{n_0,2}$ with $\pi_{n_0,2}$ and set $h \equiv f_{n_0,2}/\pi_{n_0,2}$;

3. for each $n$ in $\{10^3, 5\times10^3, 10^4, 5\times10^4, 10^5\}$, successively, extract by survey sampling with $h$ a sub-sample of $n$ observations from the data set (deprived of the observations extracted in step 2) and, based on these observations, apply the procedure described in Section 3.2. We use $\Sigma_n$ given in (7) to estimate $\sigma_1^2$, although we are not sure in advance that it is a conservative estimator.

We thus obtain $15 \times B$ estimates of $\psi_0^c$ and their respective confidence intervals.

To give an idea of what is the optimal $h$ in each case, we save the result of step 2 in the above list in the first of the $B$ simulations under $P_{0,1}$, $P_{0,2}$ and $P_{0,3}$. So, the optimal $h$ equals approximately

- $h_1$ given by $(h_1(1), h_1(2), h_1(3)) \approx (1.03, 0.67, 1.21)$ under $P_{0,1}$;
- $h_2$ given by $(h_2(1), h_2(2), h_2(3)) \approx (0.30, 0.60, 1.50)$ under $P_{0,2}$;
- $h_3$ given by $(h_3(1), h_3(2), h_3(3)) \approx (4.66, 0.53, 0.09)$ under $P_{0,3}$

Note how different are $h_1$, $h_2$ and $h_3$ (to facilitate the comparisons, $h_1$, $h_2$ and $h_3$ are renormalized to satisfy $P_{0,j}h_j = 1$ for $j = 1, 2, 3$).

Applying the TMLE procedure is straightforward thanks to the `R` package called `tmle.npvi` [11, 10]. Note, however, that it is necessary to compute $\Gamma_n$ and $\Sigma_n$. Specifically, we fine-tune the TMLE procedure by setting `iter` (the maximum number of iterations of the targeting step) to `7` and `stoppingCriteria` to `list(mic=0.01, div=0.01, psi=0.05)`. Moreover, we use the default `flavor` called `"learning"`, thus notably rely on parametric linear models for the estimation of the infinite-dimensional parameters $Q_{P_0}$, $\mu_{P_0}$ and $g_{P_0}$ and their fluctuation. We refer the interested reader to the package's manual and vignette for details.

Sampford's sampling method [18] implements the survey sampling described in Section 2.1. However, when the ratio $n/N$ is close to 0 or 1, this acceptance-rejection algorithm typically takes too much time to succeed. In our setting, this is the case when $n/N$ differs from $10^{-3}$. To circumvent that issue, we approximate the survey sampling described in Section 2.1 with a Pareto sampling [see Algorithm 2 in 4, Section 5].

The results are summarized in Table 1. We first focus on the empirical bias of the TMLE and $p$-values of the Shapiro-Wilk test of normality of its distribution. In all settings, the empirical bias decreases as $n$ grows (under $P_{0,1}$, the empirical biases for $n = 5\times10^4$ and $n = 10^5$ equal 0.0044 and 0.0036 when relying on $h_1$ or $h_0$). Under each $P_{0,j}$ and for every sub-sample size, the empirical bias is smaller when relying on $h_j$ than on $h_0$, approximately twice smaller under $P_{0,3}$. As expected due to our choices of conditional standard deviations of $Y$ given $(A, W)$, the empirical bias is larger under $P_{0,3}$ than under $P_{0,2}$ and larger under $P_{0,2}$ than under $P_{0,1}$. Except under $P_{0,3}$ when relying on $h_0$, for every $n \geq 5\times10^3$, the $p$-values of the Shapiro-Wilk test of normality are coherent with the convergence in law of the TMLE to a Gaussian distribution. Under $P_{0,3}$ and when relying on $h_0$, there is more evidence of a departure from a Gaussian distribution. Inspecting the results of the simulations studies reveals that this is mostly due to slightly too heavy tails.

We now focus on the empirical coverage, empirical variance and mean of the estimated variance of the TMLE. Consider the table about the simulation under $P_{0,1}$ first. For $n \in \{10^3, 5\times10^3, 10^4\}$,

| | $P_{0,1}$, optimal $h_1$ | | | | | $P_{0,1}$, $h_0 \equiv 1$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | b. | $p$-val. | c. | v. | e. v. | b. | $p$-val. | c. | v. | e. v. |
| $1 \times 10^3$ | 0.024 | 0.018 | 0.957 | 0.946 | 1.149 | 0.025 | 0.499 | 0.963 | 1.010 | 1.219 |
| $5 \times 10^3$ | 0.011 | 0.858 | 0.971 | 0.972 | 1.199 | 0.011 | 0.320 | 0.968 | 0.981 | 1.265 |
| $1 \times 10^4$ | 0.008 | 0.948 | 0.970 | 0.961 | 1.210 | 0.008 | 0.215 | 0.964 | 1.060 | 1.277 |
| $5 \times 10^4$ | 0.004 | 0.441 | 0.920 | 1.334 | 1.213 | 0.004 | 0.253 | 0.916 | 1.282 | 1.283 |
| $1 \times 10^5$ | 0.004 | 0.858 | 0.861 | 1.601 | 1.214 | 0.004 | 0.750 | 0.874 | 1.664 | 1.284 |

| | $P_{0,2}$, optimal $h_2$ | | | | | $P_{0,2}$, $h_0 \equiv 1$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | b. | $p$-val. | c. | v. | e. v. | b. | $p$-val. | c. | v. | e. v. |
| $1 \times 10^3$ | 0.110 | 0.001 | 0.955 | 20.14 | 26.01 | 0.124 | 0.001 | 0.945 | 25.49 | 30.32 |
| $5 \times 10^3$ | 0.045 | 0.526 | 0.986 | 16.08 | 25.84 | 0.052 | 0.156 | 0.978 | 21.46 | 32.00 |
| $1 \times 10^4$ | 0.032 | 0.419 | 0.991 | 16.34 | 25.83 | 0.036 | 0.686 | 0.983 | 20.69 | 32.17 |
| $5 \times 10^4$ | 0.015 | 0.501 | 0.990 | 16.69 | 25.89 | 0.016 | 0.775 | 0.989 | 20.01 | 32.45 |
| $1 \times 10^5$ | 0.011 | 0.956 | 0.985 | 17.20 | 25.88 | 0.012 | 0.839 | 0.986 | 20.23 | 32.38 |

| | $P_{0,3}$, optimal $h_3$ | | | | | $P_{0,3}$, $h_0 \equiv 1$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | b. | $p$-val. | c. | v. | e. v. | b. | $p$-val. | c. | v. | e. v. |
| $1 \times 10^3$ | 0.229 | 0.001 | 0.987 | 86.85 | 184.2 | 0.532 | 0.001 | 0.910 | 518.5 | 549.6 |
| $5 \times 10^3$ | 0.093 | 0.242 | 0.994 | 70.15 | 175.7 | 0.181 | 0.001 | 0.994 | 264.6 | 627.7 |
| $1 \times 10^4$ | 0.069 | 0.268 | 0.997 | 73.32 | 174.5 | 0.127 | 0.022 | 0.995 | 253.8 | 629.4 |
| $5 \times 10^4$ | 0.029 | 0.085 | 1.000 | 65.54 | 174.0 | 0.055 | 0.459 | 0.999 | 228.3 | 642.7 |
| $1 \times 10^5$ | 0.022 | 0.584 | 0.998 | 73.98 | 174.2 | 0.040 | 0.054 | 1.000 | 242.7 | 644.5 |

Table 1: Summarizing the results of the simulation study. The top, middle and bottom tables correspond to simulations under $P_{0,1}$, $P_{0,2}$ and $P_{0,3}$. Each of them reports the empirical bias of the estimators (b., $B^{-1}\sum_{b=1}^{B}|\psi^*_{n,b} - \psi^c_0|$), $p$-value of a Shapiro-Wilk test of normality ($p$-val.), empirical coverage of the confidence intervals (c., $B^{-1}\sum_{b=1}^{B}\mathbf{1}\{\psi^c_0 \in I_{n,b}\}$), $n$ times the empirical variance of the estimators (v., $n[B^{-1}\sum_{b=1}^{B}\psi^{*2}_{n,b} - (B^{-1}\sum_{b=1}^{B}\psi^*_{n,b})^2]$) and empirical mean of $n$ times the estimated variance of the estimators (e. v., $B^{-1}\sum_{b=1}^{B}\Sigma_{n,b}$), for every sub-sample size $n$ and for both $h$ optimal and $h = h_0 \equiv 1$.

the empirical coverage is satisfying when relying on both $h_1$ and $h_0$. At each of these sub-sample sizes, it does seem that we achieve the conservative estimation of $\sigma_1^2$. However, the empirical coverage deteriorates sharply for $n \in \{5 \times 10^4, 10^5\}$. It appears that, concomitantly, the empirical variance of the estimators increases strongly. This may be due to the fact that, here, $n$ is not that small compared to $N$, so that neglecting the second LHS term in (10) is inadequate, so that $\sigma_1^2$ is not the limiting variance. In conclusion, note that resorting to the optimal $h$ does not yield much gain in terms of empirical variance of the estimators.

We now turn to the two remaining tables. The first striking feature is that the empirical coverage exceeds largely the nominal coverage of 95%. The comparison of the empirical variance with the mean of the estimated variance reveals that we do achieve the conservative estimation of $\sigma_1^2$. The second striking feature is that the empirical variance stabilizes for $n$ larger than $10^3$, contrary to what happens under $P_{0,1}$. It still holds that $n$ may not be small compared to $N$. Perhaps this is counterbalanced by the fact that, by increasing starkly the conditional variance of $Y$ given $(A, W)$ under $P_{0,2}$ and $P_{0,3}$ relative to $P_{0,1}$, we make $P_0 f_1^2 h^{-1}$, the first LHS term in (10), much larger than the second LHS term $n(P_0 f_1)^2/N$. Finally, resorting to the optimal $h$ yields, both under $P_{0,2}$ and $P_{0,3}$, considerable gains in terms of empirical variance of the estimators and in terms of the width of the resulting confidence intervals.

**Acknowledgements.** The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-BS01-0005 (project SPADRO).

# A Proof of Theorem 1

Throughout the proofs, "$a \lesssim b$" means that there exists a universal constant $L > 0$ such that $a \leq Lb$.

We start with a central limit theorem for the empirical process $(\sqrt{n}(P_{R_N}^{\boldsymbol{p}} - P_0)f)_{f \in \mathcal{F}}$. Its proof is given at the end of this section. Recall that a random process $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$ is $\|\cdot\|_{2,P_0}$-equicontinuous if for each $\xi > 0$, there exists $\delta > 0$ such that, for all $f, f' \in \mathcal{F}$, $\|f - f'\|_{2,P_0} \leq \delta$ implies $P_0(|\mathbb{G}(f - f')|) \leq \xi$.

**Theorem 2.** *Under* **A1** *and* **A2** *there exists a $\|\cdot\|_{2,P_0}$-equicontinuous Gaussian process $\mathbb{G}^h \in \ell^\infty(\mathcal{F})$ with covariance operator $\Sigma$ such that $(\sqrt{n}(P_{R_N}^{\boldsymbol{p}} - P_0)f)_{f \in \mathcal{F}}$ converges weakly in $\ell^\infty(\mathcal{F})$ towards $\mathbb{G}^h$. The same result holds with $\mathcal{F}$ replaced by $\{f - f_1 : f \in \mathcal{F}\}$.*

We now turn to the proof of Theorem 1. Since $P_n^* D(P_n^*) = 0$ (by definition, the influence function $D(P_n^*)$ is centered under $P_n^*$), **A4** rewrites

$$\mathcal{R}(P_n^*, P_0) = \psi_0 - \psi_n^* - P_0 D(P_n^*) = -\gamma_n(\psi_n^* - \psi_0) + o_P(1/\sqrt{n}),$$

hence

$$(1 - \gamma_n)\sqrt{n}(\psi_n^* - \psi_0) = -\sqrt{n} P_0 D(P_n^*) + o_P(1).$$

Moreover, (3) implies that the above equality also yields

$$(1 - \gamma_n)\sqrt{n}(\psi_n^* - \psi_0) \quad = \quad \sqrt{n}(P_{R_N}^{\boldsymbol{p}} - P_0) D(P_n^*) + o_P(1)$$

$$= \sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)f_1 + \sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)(D(P^*_n) - f_1) + o_P(1).$$

Theorem 2 implies in particular that $\sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)f_1$ converges in law to the centered Gaussian distribution with variance $\Sigma(f_1, f_1)$.

Let us prove now that $\sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)(D(P^*_n) - f_1) = o_P(1)$. This is a consequence of Theorem 2 and the concentration inequality of [25, Corollary 2.2.8].

Let $\|\cdot\|_{2,\Sigma}$ be the norm on $\mathcal{F}$ given by $\|f\|^2_{2,\Sigma} \equiv \Sigma(f, f)$. For every $\delta > 0$, introduce

$$\mathcal{F}_\delta \equiv \{f \in \mathcal{F} : P_0(f - f_1)^2 \leq \delta^2\} \subset \mathcal{F}.$$

The diameter of $\mathcal{F}_\delta$ wrt $\|\cdot\|_{2,\Sigma}$ is at most $\delta/\sqrt{c(h)}$. By [25, Corollary 2.2.8],

$$\begin{aligned} E_0\left[\sup_{f\in\mathcal{F}_\delta} \mathbb{G}^h(f - f_1)\right] &\lesssim \int_0^{\delta/\sqrt{c(h)}} \sqrt{\log N(\epsilon, \mathcal{F}_\delta, \|\cdot\|_{2,\Sigma})}d\epsilon \\ &\lesssim \int_0^{\delta/\sqrt{c(h)}} \sqrt{\log N(\epsilon, \mathcal{F}, \|\cdot\|_{2,\Sigma})}d\epsilon. \end{aligned} \tag{26}$$

Set arbitrarily $\alpha, \beta > 0$, and choose $\delta > 0$ in such a way that

$$\int_0^{\delta/\sqrt{c(h)}} \sqrt{\log N(\epsilon, \mathcal{F}, \|\cdot\|_{2,\Sigma})}d\epsilon \leq \alpha\beta.$$

By Markov's inequality, (26) and choice of $\delta$, it holds that

$$P_0\left(\sup_{f\in\mathcal{F}_\delta} \mathbb{G}^h(f - f_1) \geq \alpha\right) \leq \alpha^{-1} E_0\left[\sup_{f\in\mathcal{F}_\delta} \mathbb{G}^h(f - f_1)\right] \lesssim \beta.$$

Hence, Theorem 2 implies that, for $n$ large enough,

$$P_0\left(\sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)(D(P^*_n)) - f_1) \geq \alpha\right) \leq 2\beta. \tag{27}$$

Furthermore, by **A3**, $P_0(D(P^*_n) \notin \mathcal{F}_\delta) \leq \beta$ for $n$ large enough. Combining this inequality and (27) finally yields

$$\begin{aligned} P_0\left(\sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)(D(P^*_n)) - f_1) \geq \alpha\right) &\leq P_0\left(\sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)(D(P^*_n)) - f_1) \geq \alpha \ , \ D(P^*_n)) \in \mathcal{F}_\delta\right) \\ &\quad + P_0\left(D(P^*_n)) \notin \mathcal{F}_\delta\right) \\ &\leq P_0\left(\sup_{f\in\mathcal{F}_\delta} \sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)(f - f_1) \geq \alpha\right) \\ &\quad + P_0\left(D(P^*_n)) \notin \mathcal{F}_\delta\right) \leq 3\beta \end{aligned}$$

for $n$ large enough.

Consequently, $(1-\gamma_n)(\psi^*_n - \psi_0)$ converges in law to the centered Gaussian distribution with variance $\Sigma(f_1, f_1)$. Applying Slutsky's lemma completes the proof.

**Proof of Theorem 2.**

The proof relies on results from [14, 1]. For each $f \in \mathcal{F}$, define

$$Z_N(f) \equiv P^{\boldsymbol{p}}_{R_N} f$$

and

$$\mathbb{G}^h_n(f) \equiv \sqrt{n}(P^{\boldsymbol{p}}_{R_N} - P_0)f = \sqrt{n}(Z_N(f) - P_0 f).$$

We first state and prove the following lemma, by using [14, Lemma 4.3 and Theorem 7.1]:

**Lemma 2.** *For every (measurable) real-valued function $f$ on $\mathcal{O}$ such that $P_0 f^2/h$ is finite, $\mathbb{G}^h_n(f)$ converges in law to the centered Gaussian distribution with variance $\sigma^2(f) \equiv E_{P_0}\left[f^2(O)h(V)^{-1}\right]$.*

*Proof of Lemma 2.* This is a three-step proof.

*Step 1: preliminary.* Set arbitrarily a measurable function $f : \mathcal{O} \to \mathbb{R}$ such that $P_0 f^2/h$ is finite and define

$$T_N(f) \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{f(O_i)\varepsilon_i}{p_i}.$$

The only difference between $T_N(f)$ and $Z_N(f)$ is the substitution of $(\varepsilon_1, \ldots, \varepsilon_N)$ for $(\eta_1, \ldots, \eta_N)$. Since $(O_1, \varepsilon_1), \ldots, (O_N, \varepsilon_N)$ are independently sampled (from $P_0^{\varepsilon}$), it holds that $E_{P_0^{\varepsilon}}[T_N(f)] = P_0 f$ and

$$\begin{aligned}
\mathrm{Var}_{P_0^{\varepsilon}}(T_N(f)) &= \frac{1}{N}\mathrm{Var}_{P_0^{\varepsilon}}\left(\frac{f(O_1)\varepsilon_1}{p_1}\right) \\
&= \frac{1}{n}E_{P_0}\left[\frac{f^2(O)}{h(V)}\right] - \frac{1}{N}(P_0 f)^2 \\
&= \frac{\sigma^2(f)}{n} + o(1/n). \qquad (28)
\end{aligned}$$

Thus, $\sqrt{n}(T_N(f) - P_0 f)$ converges in law to the centered Gaussian distribution with variance $\sigma^2(f)$. The challenge is now to derive another central limit theorem for $Z_N(f)$ from this convergence in law.

*Step 2: coupling.* The rest of the proof mainly hinges on coupling. We may assume without loss of generality that there exist $U_1, \ldots, U_N$ independently drawn from the uniform distribution on $[0, 1]$ and independent of $(O_1, \ldots, O_N)$ such that, for each $1 \leq i \leq N$, $\varepsilon_i = \mathbf{1}\{U_i \leq p_i\}$. We now define $\ell_N \equiv n/\sum_{i=1}^N p_i$ and, for each $1 \leq i \leq N$, $\varepsilon_i(\ell_N) = \mathbf{1}\{U_i \leq \ell_N p_i\}$. This is the first coupling used in the proof.

The second coupling is more elaborate. Due to Hajek, it gives rise to two random subsets $s_K$ and $s_n$ of $\{1, \ldots, N\}$ that we characterize now, in three successive steps. In the rest of this step of the proof, we work conditionally on $O_1, \ldots, O_N$.

1. Drawing $s_n \subset \{1, \ldots, N\}$:

   (a) sample $(\eta'_1, \ldots, \eta'_N)$ from the conditional distribution of $(\varepsilon'_1, \ldots, \varepsilon'_N)$ given $\sum_{i=1}^N \varepsilon'_i = n$ when $\varepsilon'_1, \ldots, \varepsilon'_N$ are independently drawn from the Bernoulli distributions with parameters $\ell_N p_1, \ldots, \ell_N p_N$, respectively;

(b) define $s_n = \{1 \leq i \leq N : \eta'_i = 1\}$ and $D_n \equiv \sum_{i \in s_n}(1 - \ell_N p_i)$ for future use.

We say simply that $s_n$ is drawn from the rejective sampling scheme on $\{1, \ldots, N\}$ with parameter $(\ell_N p_i : i \in \{1, \ldots, N\})$ (see Section 2).

2. Drawing $K \in \{1, \ldots, N\}$:

   (a) sample $\varepsilon''_1, \ldots, \varepsilon''_N$ independently from the Bernoulli distributions with parameters $\ell_N p_1, \ldots, \ell_N p_N$, respectively;

   (b) define $K \equiv \sum_{i=1}^N \varepsilon''_i$.

3. Drawing $s_K$:

   (a) if $K = n$, then set $s_K \equiv s_n$;

   (b) if $K > n$, then draw $s_{K-n}$ from the rejective sampling scheme on $\{1, \ldots, N\} \setminus s_n$ with parameter $((K-n)\ell_N p_i / D_n : i \in \{1, \ldots, N\} \setminus s_n)$ and set $s_K \equiv s_n \cup s_{K-n}$;

   (c) if $K < n$, then draw $s_{n-K}$ from the rejective sampling scheme on $s_n$ with parameter $((K-n)\ell_N p_i / D_n : i \in s_n)$ and set $s_K \equiv s_n \setminus s_{n-K}$.

We denote by $\mathbb{S}$ the joint law of $(s_K, s_n)$. Obviously, $\mathbb{S}$ is such that $s_K \subset s_n$ or $s_n \subset s_K$ $\mathbb{S}$-almost surely. We denote by $\mathbb{P}$ the law of the Poisson sampling scheme, *i.e.*, the law of $\{1 \leq i \leq N : \varepsilon'_i = 1\}$ from the description of how $s_n$ is drawn. Law $\mathbb{S}$ is a coupling of the rejective sampling scheme and an approximation to the Poisson sampling scheme $\mathbb{P}$ in the sense of the following corollary of [14, Lemma 4.3].

**Proposition 1** (Hajek)**.** *If $d_N \equiv \sum_{i=1}^N p_i(1-p_i)$ goes to infinity as $N$ goes to infinity, then the marginal distribution of $s_K$ when $(s_K, s_n)$ is drawn from $\mathbb{S}$ converges to $\mathbb{P}$ in total variation.*

The condition on $d_N$ is met for our choice of $(p_1, \ldots, p_N)$.

*Step 3: concluding.* Introduce

$$\begin{aligned} T_N^{\ell_N}(f) &\equiv \frac{1}{N}\sum_{i=1}^N \frac{f(O_i)\varepsilon_i(\ell_N)}{\ell_N p_i}, \\ T_N^{s_K}(f) &\equiv \sum_{i \in s_K} \frac{f(O_i)}{p_i}, \\ T_N^{s_n}(f) &\equiv \sum_{i \in s_n} \frac{f(O_i)}{p_i}. \end{aligned}$$

The random variables $Z_N(f)$, $T_N^{\ell_N}(f)$, $T_N^{s_K}(f)$ and $T_N^{s_n}(f)$ satisfy the following properties.

- $Z_N(f)$ and $T_N^{s_n}(f)$ share a common law.
  This is a straightforward consequence of Proposition 1.

- $\sqrt{n}(T_N^{s_n}(f) - T_N^{s_K}(f)) = o_P(1)$.
  Indeed, it is shown in the proof of [14, Theorem 7.1] that the convergence of $d_N$ (defined in Proposition 1) to infinity implies, conditionally on $O_1, \ldots, O_N$, $\sqrt{n}(T_N^{s_K}(f) - T_N^{s_n}(f)) = o_P(1)$. The unconditional result readily follows.

- $T_N^{s_K}(f)$ and $T_N^{\ell_N}(f)$ have asymptotically the same law, in the sense that the total variation distance between their laws goes to 0 as $N$ goes to infinity.
  This is a consequence of Proposition 1.

- $\sqrt{n}(T_N^{\ell_N}(f) - T_N(f)) = o_P(1)$.
  It suffices to show that $E_{P_0^\varepsilon}\left[(T_N(f) - T_N^{\ell_N}(f))^2\right] = o(1/n)$. Observe now that

$$\begin{aligned} nE_{P_0^\varepsilon}\left[\left(T_N^{\ell_N}(f) - T_N(f)\right)^2\right] &= \frac{n}{N}E_{P_0^\varepsilon}\left[\left(\frac{\varepsilon_1(\ell_N)}{\ell_N} - \varepsilon_1\right)^2 \frac{f(O_1)^2}{p_1^2}\right] \\ &= E_{P_0^\varepsilon}\left[\left(\frac{1}{\ell_N} - 2\frac{\min(1,\ell_N)}{\ell_N} + 1\right)\frac{f(O_1)^2}{h(V_1)}\right]. \end{aligned}$$

  The strong law of large numbers yields that $\ell_N$ converges to 1 almost surely, which implies $1/\ell_N - 2\min(1,\ell_N)/\ell_N + 1$ converges to 0 almost surely hence the result by the dominated convergence theorem.

Consequently, $\mathbb{G}_n^h(f) \equiv \sqrt{n}(Z_N(f) - P_0 f)$ and $\sqrt{n}(T_N(f) - P_0 f)$ have asymptotically the same law. The same arguments are valid when $\{f - f_1 : f \in \mathcal{F}\}$ is substituted for $\mathcal{F}$. Thus, the proof is complete. □

We can now prove Theorem 2. We first note that Lemma 2 implies the asymptotic tightness of the real-valued random variable $\mathbb{G}_n^h(f)$ for all $f \in \mathcal{F}$. Moreover, Lemma 2 and the Cramér-Wold device yield the convergence in law of $(\mathbb{G}_n^h f_1, \ldots, \mathbb{G}_n^h f_M)$ to $(\mathbb{G}^h f_1, \ldots, \mathbb{G}^h f_M)$ for all $(f_1, \ldots, f_M) \in \mathcal{F}^M$. Indeed, for each $(f_1, \ldots, f_M) \in \mathcal{F}^M$ and any $(\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$, $\bar{f} \equiv \sum_{m=1}^M \lambda_m f_m$ is measurable and $P_0\bar{f}^2/h$ is finite hence, by Lemma 2, $\sum_{m=1}^M \lambda_m \mathbb{G}_n^h f_m = \mathbb{G}_n^h(\bar{f})$ converges in law to $\mathbb{G}^h(\bar{f}) = \sum_{m=1}^M \lambda_m \mathbb{G}_n^h f_m$. In addition, **A1** implies that the diameter of $\mathcal{F}$ wrt $\|\cdot\|_{2,P_0}$ is finite. Therefore, by [25, Theorems 1.5.4 and 1.5.7], if for all $\alpha, \beta > 0$, there exists $\delta > 0$ such that

$$\limsup_{N\to\infty} P\left(\sup_{f,f':\|f-f'\|_{2,P_0}<\delta} \left|\mathbb{G}_n^h f - \mathbb{G}_n^h f'\right| > \alpha\right) \leq \beta, \tag{29}$$

then Theorem 2 is valid.

Set arbitrarily $\alpha, \beta, \delta > 0$ and introduce $\mathcal{F}_\delta \equiv \{f - f' : f, f' \in \mathcal{F}, \|f - f'\|_{2,P_0} \leq \delta\}$. It is shown in [16] (see also [1]) that $\eta_1, \ldots, \eta_N$ are *negatively associated* in the following sense. For each $A_1, A_2 \subset \{1, \ldots, N\}$ with $A_1 \cap A_2 = \emptyset$ and all (measurable) $f : \mathbb{R}^{d_1} \to \mathbb{R}$ and $g : \mathbb{R}^{d_2} \to \mathbb{R}$ ($d_1 \equiv \text{card}(A_1)$ and $d_2 \equiv \text{card}(A_2)$), if $f$ and $g$ are increasing in every coordinate, then

$$\text{cov}\left(f(\eta_i : i \in A_1), g(\eta_i : i \in A_2)\right) \leq 0.$$

Hoeffding's inequality for negatively associated bounded random variables in [3, Theorem S1.2] guarantees that, conditionally on $O_1, \ldots, O_N$, for all $t > 0$,

$$P\left(|\mathbb{G}_n^h(f)| > t \middle| O_1, \ldots, O_N\right) \leq \exp\left(-\frac{2t^2}{\rho_N^2(f)}\right).$$

Therefore, a classical chaining argument [25, Corollary 2.2.8, for instance]) yields

$$E\left[\sup_{f,f' \in \mathcal{F}_\delta} |\mathbb{G}_n^h(f) - \mathbb{G}_n^h(f')| \middle| O_1, \ldots, O_N\right] \lesssim \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}_\delta, \rho_N)} d\epsilon. \tag{30}$$

By **A2**, there exists a deterministic sequence $\{a_N\}_{N \geq 1}$ tending to 0 such that, for all $f, g \in \mathcal{F}$, $\rho_N(f,g) \leq (1 + a_N)\rho(f,g)$ $P_0$-almost surely. Consequently, for every $\epsilon > 0$, it holds $P_0$-almost surely that

$$N(\epsilon, \mathcal{F}_\delta, \rho_N) \leq N(\epsilon/(1 + a_N), \mathcal{F}_\delta, \|\cdot\|_{2,P_0}).$$

Plugging the previous upper-bound in (30), taking the expectation, using Markov's inequality and letting $N$ go to infinity then give

$$\begin{aligned} \limsup_{N \to \infty} P\left(\sup_{f,f' \in \mathcal{F}_\delta} \left|\mathbb{G}_n^h(f) - \mathbb{G}_n^h(f')\right| > \alpha\right) &\lesssim \alpha^{-1} \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}_\delta, \|\cdot\|_{2,P_0})} d\epsilon \\ &\lesssim \alpha^{-1} J(\delta, \mathcal{F}, \|\cdot\|_{2,P_0}). \end{aligned}$$

By **A1**, it is possible to choose $\delta > 0$ small enough to ensure that the above RHS expression is smaller then $\beta$, hence (29) holds.

It only remains to determine the covariance of $\mathbb{G}^h$. By adapting the proof of Lemma 2, it appears that $\mathrm{cov}(\mathbb{G}^h(f), \mathbb{G}^h(f')) = P_0 f f'/h = \Sigma(f, f')$ for all $f, f' \in \mathcal{F}$.

# B Tailoring the main theorem in the setting of Section 3.1

Let us show that $\mathbf{A1}^b$ and $\mathbf{A2}^b$ imply **A1**–**A4** in the setting of Section 3.1. Since $\mathcal{Q}^w$ and $\mathcal{G}^w$ are uniformly bounded away from 0 and 1, $t_n$ (19) necessarily belongs to a deterministic, compact subset $\mathcal{T}$ of $\mathbb{R}$. Define

$$\widetilde{\mathcal{Q}}^w \equiv \left\{\mathrm{expit}\left(\mathrm{logit}\, Q + t\frac{2A-1}{g(A|W)}\right) : Q \in \mathcal{Q}^w, g \in \mathcal{G}^w, t \in \mathcal{T}\right\}$$

then

$$\mathcal{F} \equiv \{D(P) : P \in \mathcal{M} \text{ s.t. } Q_P \in \widetilde{\mathcal{Q}}^w, g_P \in \mathcal{G}^w\}.$$

Obviously, $D(P_n^*) \in \mathcal{F}$ and $\sup_{f \in \mathcal{F}} \|f\|_\infty$ is finite. Furthermore, because expit is a 1-Lipschitz and logit is Lipschitz on any compact subset of $(0,1)$, it holds that $\widetilde{Q}, \widetilde{Q}' \in \widetilde{\mathcal{Q}}^w$ respectively parametrized by $(Q, g, t)$ and $(Q', g', t')$ satisfy

$$\|\widetilde{Q} - \widetilde{Q}'\|_{2,P_0} \lesssim \|Q - Q'\|_{2,P_0} + \|g - g'\|_{2,P_0} + |t - t'|.$$

Therefore, the finiteness of $J(1, \mathcal{Q}^w, \|\cdot\|_{2,P_0})$, $J(1, \mathcal{G}^w, \|\cdot\|_{2,P_0})$ and $J(1, \mathcal{T}, |\cdot|)$ implies the finiteness of $J(1, \widetilde{\mathcal{Q}}^w, \|\cdot\|_{2,P_0})$. Moreover, the separability of $\mathcal{Q}^w$ and $\mathcal{G}^w$ yields that $\widetilde{\mathcal{Q}}^w$ is also separable.

Furthermore, for every $P, P' \in \mathcal{M}$ such that $D^b(P), D^b(P') \in \mathcal{F}$, it holds that

$$\|D^b(P) - D^b(P')\|_{2,P_0} \lesssim \|Q_P - Q_{P'}\|_{2,P_0} + \|g_P - g_{P'}\|_{2,P_0} + |\Psi^b(P) - \Psi^b(P')|. \tag{31}$$

We will prove this at the end of the section. By (31), the separability of $\widetilde{\mathcal{Q}}^w$ and $\mathcal{G}^w$ implies that of $\mathcal{F}$. In addition, the finiteness of $J(1, \widetilde{\mathcal{Q}}^w, \|\cdot\|_{2,P_0})$, $J(1, \mathcal{G}^w, \|\cdot\|_{2,P_0})$, $J(1, [0,1], |\cdot|)$ and (31) imply that $J(1, \mathcal{F}, \|\cdot\|_{2,P_0})$ is finite. We prove likewise based on (31) that $\mathcal{F}$ has a finite uniform entropy integral because $\mathcal{Q}^w$ and $\mathcal{G}^w$ do. Finally, (15) and $\mathbf{A2}^b$ imply **A3** (by Cauchy-Schwarz's inequality) and **A4**.

*Proof of (31).* For any $P \in \mathcal{M}$, denote $q_P(W) = Q_P(1, W) - Q_P(0, W)$. Set $P, P' \in \mathcal{M}^b$. It holds that

$$\|D_1^b(P) - D_1^b(P')\|_{2,P_0} \leq \|q_P - q_{P'}\|_{2,P_0} + |\Psi^b(P) - \Psi^b(P')|.$$

Moreover,

$$\begin{aligned}
\|D_2^b(P) - D_2^b(P')\|_{2,P_0} &= \left\| (Y - q_P(W)) \frac{2A-1}{g_P} - (Y - q_{P'}(W)) \frac{2A-1}{g_{P'}} \right\|_{2,P_0} \\
&\leq \left\| (Y - q_P(W))(2A-1) \left( \frac{1}{g_P} - \frac{1}{g_{P'}} \right) \right\|_{2,P_0} + \left\| (q_P - q_{P'}) \frac{2A-1}{g_{P'}(W)} \right\|_{2,P_0} \\
&\lesssim \|g_P - g_{P'}\|_{2,P_0} + \|q_P - q_{P'}\|_{2,P_0},
\end{aligned}$$

where the last inequality relies on the uniform boundedness of $(Y - q_P(W))(2A-1)$ and $g_P^{-1}$. The result follows since $\|q_P - q_{P'}\|_{2,P_0} \leq 2\|Q - Q'\|_{2,P_0}$. □

The same kind of arguments allow to verify that $\mathbf{A1}^c$ and $\mathbf{A2}^c$ also imply **A1**–**A4**.

# References


[1] A. D. Barbour. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5(4): 425–427, 1990.

[2] Y. Berger. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67(2):209–226, 1998.

[3] P. Bertail, E. Chautru, and S. Clémençon. Empirical processes in survey sampling. *Scandinavian Journal of Statistics*, October 2016. To appear.

[4] L. Bondesson, I. Traat, and A. Lundqvist. Pareto sampling versus Sampford and conditional Poisson sampling. *Scandinavian Journal of Statistics. Theory and Applications*, 33(4):699–720, 2006.

[5] N. E. Breslow and J. A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34(1): 86–102, 2007.

[6] N. E. Breslow and J. A. Wellner. 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". *Scandinavian Journal of Statistics*, 35, 2008.

[7] K. R. W. Brewer and M. E. Donadio. The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29(2):189–196, 2003.

[8] H. Cardot, D. Degras, and E. Josserand. Confidence bands for Horvitz–Thompson estimators using sampled noisy functional data. *Bernoulli*, 19(5A):2067–2097, 2013.

[9] H. Cardot, A. Dessertaine, C. Goga, E. Josserand, and P. Lardin. Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. *Survey Methodology/Techniques d'enquêtes*, 39:283–301, 2013.

[10] A. Chambaz and P. Neuvial. tmle.npvi: targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation. *Bioinformatics*, 31 (18):3054–3056, 2015.

[11] A. Chambaz and P. Neuvial. *Targeted Learning of a Non-Parametric Variable Importance Measure of a Continuous Exposure*, 2016. URL `http://CRAN.R-project.org/package=tmle.npvi`. R package version 0.10.0.

[12] A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6:1059–1099, 2012.

[13] A. Grafström. Entropy of unequal probability sampling designs. *Statistical Methodology*, 7(2): 84–97, 2010.

[14] J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 12 1964.

[15] M. Hanif and K. R. W. Brewer. Sampling with unequal probabilities without replacement: a review. *International Statistical Review/Revue Internationale de Statistique*, pages 317–335, 1980.

[16] K. Joag-Dev and F. Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 03 1983.

[17] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge University Press, 2000.

[18] M. R. Sampford. On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54(3-4):499–513, 1967.

[19] M. J. van der Laan. Statistical inference for variable importance. *International Journal of Biostatistics*, 2, 2006.

[20] M. J. van der Laan. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *International Journal of Biostatistics*, 2016. To appear.

[21] M. J. van der Laan and S. D. Lendle. Online targeted learning. Technical Report 330, U.C. Berkeley Division of Biostatistics, 2014. URL `http://biostats.bepress.com/ucbbiostat/paper330`.

[22] M. J. van der Laan and S. Rose. *Targeted learning*. Springer, 2011. ISBN 978-1-4419-9781-4.

[23] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2:Art. 11, 40, 2006. doi: 10.2202/1557-4679.1043.

[24] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[25] A. W. van Der Vaart and J. A. Wellner. *Weak Convergence and empirical processes*. Springer, 1996.

[26] J. C. Wang. Sample distribution function based goodness-of-fit test for complex surveys. *Computational Statistics & Data Analysis*, 56(3):664–679, 2012.