

Collaborative targeted inference from continuously indexed nuisance parameter estimators

Cheng Ju, Antoine Chambaz, Mark van Der Laan

▶ **To cite this version:**

Cheng Ju, Antoine Chambaz, Mark van Der Laan. Collaborative targeted inference from continuously indexed nuisance parameter estimators. 2018. hal-01759020

HAL Id: hal-01759020
https://hal.science/hal-01759020

Preprint submitted on 5 Apr 2018

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.

arXiv:1804.00102v1 [stat.ME] 31 Mar 2018

Collaborative targeted inference from continuously indexed nuisance parameter estimators

Cheng Ju
Division of Biostatistics, UC Berkeley
and
Antoine Chambaz
MAP5 (UMR 8145), Université Paris Descartes
and
Mark J. van der Laan
Division of Biostatistics, UC Berkeley

April 3, 2018

Abstract

Suppose that we wish to infer the value of a statistical parameter at a law from which we sample independent observations. Suppose that this parameter is smooth and that we can define two variation-independent, infinite-dimensional features of the law, its so called Q- and G-components (comp.), such that if we estimate them consistently at a fast enough product of rates, then we can build a confidence interval (CI) with a given asymptotic level based on a plain targeted minimum loss estimator (TMLE). The estimators of the Q- and G-comp. would typically be by products of machine learning algorithms. We focus on the case that the machine learning algorithm for the G-comp. is fine-tuned by a real-valued parameter h. Then, a plain TMLE with an h chosen by cross-validation would typically not lend itself to the construction of a CI, because the selection of h would trade-off its empirical bias with something akin to the empirical variance of the estimator of the G-comp. as opposed to that of the TMLE. A collaborative TMLE (C-TMLE) might, however, succeed in achieving the relevant trade-off. We prove that this is the case indeed.

We construct a C-TMLE and show that, under high-level empirical processes conditions, and if there exists an oracle h that makes a bulky remainder term asymptotically Gaussian, then the C-TMLE is asymptotically Gaussian hence amenable to building a CI provided that its asymptotic variance can be estimated too. The construction hinges on guaranteeing that an additional, well chosen estimating equation is solved on top of the estimating equation that a plain TMLE solves. The optimal h is chosen by cross-validating an empirical criterion that guarantees the wished trade-off between empirical bias and variance.

We illustrate the construction and main result with the inference of the so called average treatment effect, where the Q-comp. consists in a marginal law and a conditional expectation, and the G-comp. is a propensity score (a conditional probability). We also conduct a multi-faceted simulation study to investigate the empirical properties of the collaborative TMLE when the G-comp. is estimated by the LASSO. Here, h is the bound on the ℓ^1-norm of the candidate coefficients. The variety of scenarios shed light on small and moderate sample properties, in the face of low-, moderate- or high-dimensional baseline covariates, and possibly positivity violation.

Keywords: cross-validation, empirical process theory, semiparametric models

1 Introduction

We wish to infer the value of a statistical parameter at a law from which we sample independent observations. The parameter is a smooth function of the data distribution. We assume that we can define two variation-independent, infinite-dimensional features of the law, its so called Q- and G-components, such that if we estimate them consistently at a fast enough joint rate, then we can build a confidence interval (CI) with a given asymptotic level based on a plain targeted minimum loss estimator (TMLE) [30, 29]. Typically, the parameter depends on the law only through its Q-component, whereas its canonical gradient depends on the law through both its Q- and G-components. The estimators of the Q- and G-components would typically be by products of machine learning algorithms. We focus on the case that the machine learning algorithm for the G-component is fine-tuned by a real-valued parameter h. Is it possible to construct an estimator that will lend itself to the construction of a CI, by fine-tuning data-adaptively and in a targeted fashion both the algorithm for the estimation of the G-component and the resulting estimator of the parameter of interest?

Literature overview. The general problem that we address is often encountered in observational studies of the effect of an exposure, for instance when one wishes to infer the average effect of a two-level exposure. It is then necessary to account for the fact that the level of exposure is not fully randomized in the observed population. A pivotal object of interest in such studies, the so called exposure mechanism (that is, the conditional law of exposure given baseline covariates) is an example of what we generally call a G-component of the law of the experiment.

A wide range of estimators of the average effect of a two-level exposure require the estimation of the propensity score: Horvitz-Thompson estimators [9]; estimators based on propensity score matching [23, 8, 7] or stratification [1, 24]; any estimator relying on the efficient influence curve, among which double-robust inverse probability of exposure weighted estimators [20, 22, 18] or estimators built based on the targeted minimum loss estimation (TMLE) methodology [30, 29].

Common methods for the estimation of the propensity score are multivariate logistic regression [14], high-dimensional propensity score adjustment [25, 2], and a variety of machine learning algorithms [15, 6, 10]. Except in the so called *collaborative* variant of TMLE that we will discuss shortly, the estimators of the propensity score can be derived at a preliminary step, regardless essentially of why they are needed and how they are used at the subsequent step. This is problematic because optimality at the preliminary step has little if any relation to optimality at the subsequent step. For instance, the optimal estimator of the propensity score at the preliminary step might take values very close to zero, therefore disqualifying it as a viable estimator at the subsequent step, not to mention an optimal one. In a less dramatic scenario, using an instrumental variable (which only influences exposure but not the outcome) to estimate the propensity score could concomitantly yield a better estimator thereof and only increase the variance of the resulting estimator of the effect of exposure [31, 29].

This prompted the development of the so called *collaborative* version of the targeted minimum loss estimation methodology [31, 29], where the estimation of the G-component is not separated from that of the parameter of main interest anymore. More concretely, collaborative TMLE (C-TMLE) consists in building a sequence of estimators of the G-component and in selecting one of them by optimizing a criterion that targets the parameter of main interest. For instance, in the above less dramatic scenario, covariates that are strongly predictive of exposure but not of the outcome would be removed, resulting in less bias for the estimator of the parameter of main interest.

The C-TMLE methodology has been adapted to a wide range of fields, including genomics [4, 34],

survival analysis [27], and clinical studies[11]. Because the derivation of C-TMLE estimators is often computationally demanding, scalable versions have also been developed [11].

In [26], the authors propose a C-TMLE algorithm that uses regression shrinkage of the exposure model for the estimation of the propensity score. It sequentially reduces the parameter that determines the amount of penalty placed on the size of the coefficient values, and selects the appropriate parameter by cross-validation. The methodology for continuously fine-tuned, collaborative targeted learning that we develop in this article encompasses the algorithm of [26]. Its statistical analysis sheds light on why, and under which assumptions, it would provide valid statistical inference.

The present study builds upon [28]. The methodology is also studied in [12, 13], the latter an example of real-life application.

At this point in the introduction, we wish to formalize what is the problem at stake. What follows recasts the introductory paragraph in the theoretical framework that we adopt in the article.

Setting the scene. Let $O_1, \ldots, O_n$ be n independent draws from a law P_0 on a set $\mathcal{O}$. We view P_0 as an element of the statistical model $\mathcal{M}$, a collection of plausible laws for $O_1, \ldots, O_n$. The more we know about P_0, the smaller is $\mathcal{M}$. Our primary goal is to infer the value of parameter $\Psi : \mathcal{M} \to \mathbb{R}$ at P_0, namely, $\psi_0 \equiv \Psi(P_0)$. Our statistical analysis is asymptotic in the number of observations.

We consider the case that Ψ is pathwise differentiable at every $P \in \mathcal{M}$ with respect to (w.r.t.) a tangent set $\mathcal{S}_P \subset L_0^2(P)$: there exists $D^*(P) \in L_0^2(P)$ such that, for every $s \in S_P$, there exists a submodel $\{P_t : t \in \mathbb{R}, |t| < c\} \subset \mathcal{M}$ satisfying *(i)* $P_t|_{t=0} = P$, *(ii)* $P_t \ll P$ for all $t \in]-c, c[$, *(iii)*

$$\frac{d}{dt} \log \frac{dP_t}{dP}(O)\bigg|_{t=0} = s(O)$$

(the submodel's score function equals s), and *(iv)* the real valued mapping $t \mapsto \Psi(P_t)$ is differentiable at $t = 0$ with a derivative equal to $PD^*(P)s$, where Pf is a shorthand notation for $E_P(f(O))$ (any measurable f). It is assumed moreover that every $P \in \mathcal{M}$ is associated with two possibly infinite-dimensional features $Q \in \mathcal{Q}$ and $G \in \mathcal{G}$ such that *(i)* Q and G are unrelated (*i.e.*, variation independent: knowing anything about Q tells nothing about G and vice versa), *(ii)* $\Psi(P)$ depends on P only through Q, *(iii)* $D^*(P)$ depends on P only through Q and G, and *(iv)* G is a mapping from $\mathcal{O}$ to $\mathbb{R}$. At this early stage, we can introduce the pivotal

$$\mathrm{Rem}_{20}(Q, G) \equiv \Psi(P) - \Psi(P_0) + P_0 D^*(P)$$

for every $P \in \mathcal{M}$. The notation is justified *(i)* because we wish to think of the right-hand-side expression as a remainder term, and *(ii)* by the fact that $\Psi(P)$ and $D^*(P)$ depend on P only through Q and G. We consider the case that parameter Ψ is such that, for some pseudo-distances $d_{\mathcal{Q}}$ and $d_{\mathcal{G}}$ on $\mathcal{Q}$ and $\mathcal{G}$,

$$|\mathrm{Rem}_{20}(Q, G)| \lesssim d_{\mathcal{Q}}(Q, Q_0) \times d_{\mathcal{G}}(G, G_0), \tag{1}$$

where $a \lesssim b$ stand for "there exists a universal positive constant $c > 0$ such that $a \leq bc$". A remainder term satisfying (1) is said double-robust.

Let $\hat{Q}$ be an algorithm for the estimation of Q_0, the Q-component of the true law P_0. Likewise, let $\hat{G}_h$ ($h \in \mathcal{H}$, an open interval of $\mathbb{R}_+^*$ of which the closure contains 0) be an h-specific algorithm for the estimation of G_0, the G-component of P_0. Formally, we view $\hat{Q}$ and each $\hat{G}_h$ as mappings from

$$\bigcup_{N \geq 1} \left\{ N^{-1} \sum_{i=1}^{N} \mathrm{Dirac}(o_i) : o_1, \ldots, o_N \in \mathcal{O} \right\}$$

to $\mathcal{Q}$ and $\mathcal{G}$, respectively, that can "learn" from the empirical measure P_n some estimators $\hat{Q}(P_n)$ and $\hat{G}_h(P_n)$ of Q_0 and G_0. Set $Q_n^0 \equiv \hat{Q}(P_n)$ (the superscript 0 stands for "initial"), $G_{n,h} \equiv \hat{G}_h(P_n)$, and let $P_{n,h}^0 \in \mathcal{M}$ be any element of the model of which the Q- and G-components equal Q_n^0 and $G_{n,h}$. Derived by the mere substitution of P_n^0 for P_0 in $\Psi(P_0)$, $\Psi(P_{n,h}^0)$ is a natural estimator of $\Psi(P_0)$. It is *not* targeted toward the inference of $\Psi(P_0)$ in the sense that none of the known features of $P_{n,h}^0$ was derived specifically for the sake of ultimately estimating $\Psi(P_0)$.

It is well documented in the TMLE literature that one way to target $\Psi(P_{n,h}^0)$ toward $\Psi(P_0)$ is to build $P_{n,h}^* \in \mathcal{M}$ from $P_{n,h}^0$ in such a way that

$$P_n D^*(P_{n,h}^*) = o_P(1/\sqrt{n})$$

and to infer $\Psi(P_0)$ with $\Psi(P_{n,h}^*)$. This can be achieved, in such a way that $G_{n,h}$ is not modified, by "fluctuating" $P_{n,h}^0$, a procedure that we will develop in details in the specific example studied in the article. Then, by (1), the estimator satisfies the asymptotic expansion:

$$\Psi(P_{n,h}^*) - \Psi(P_0) = (P_n - P_0)D^*(P_{n,h}^*) + \mathrm{Rem}_{20}(Q_{n,h}^*, G_{n,h}) + o_P(1/\sqrt{n}). \tag{2}$$

By convention, we agree that small values of h correspond with less bias for $G_{n,h}$ as an estimator of G_0. Moreover, we assume that there exists $h_n \in \mathcal{H}$, $h_n = o(1)$, such that $d_{\mathcal{G}}(G_{n,h_n}, G_0) = o_P(\rho_{1,n})$ for some $\rho_{1,n} = o(1)$, *i.e.*, that G_{n,h_n} consistently estimates G_0 at rate $\rho_{1,n}$. If Q_{n,h_n}^* is also such that $d_{\mathcal{Q}}(Q_{n,h_n}^*, Q_0) = o_P(\rho_{2,n})$ for some $\rho_{2,n} = o(1)$, and if $\rho_{1,n}\rho_{2,n} = o(1/\sqrt{n})$, then (1) and (2) yield

$$\Psi(P_{n,h_n}^*) - \Psi(P_0) = (P_n - P_0)D^*(P_{n,h_n}^*) + o_P(1/\sqrt{n})$$

which may in turn imply the asymptotic linear expansion

$$\Psi(P_{n,h_n}^*) - \Psi(P_0) = (P_n - P_0)\mathrm{IF} + o_P(1/\sqrt{n}), \tag{3}$$

with influence function $\mathrm{IF} \equiv D^*(P_0)$, depending in particular on how data-adaptive are algorithms $\hat{Q}$ and $\hat{G}_h$ ($h \in \mathcal{H}$). By the central limit theorem, (3) guarantees that $\sqrt{n}(\Psi(P_{n,h_n}^*) - \Psi(P_0))$ is asymptotically Gaussian.

We focus on a more challenging situation, where $\rho_{1,n}\rho_{2,n}$ is not necessarily $o(1/\sqrt{n})$. We anticipate that our analysis is also very relevant at small and moderate sample sizes when $\rho_{1,n}\rho_{2,n} = o(1/\sqrt{n})$.

In order to derive an asymptotic linear expansion similar to (3) from (2) in this situation, we would have to derive an asymptotic expansion of $\mathrm{Rem}_{20}(Q_{n,h_n}^*, G_{n,h_n})$. Unfortunately, we have reasons to believe that this is not possible without targeting (their presentation in an example is deferred to Section 3.3).

Now, observe that the estimators $\Psi(P_{n,h}^*)$ ($h \in \mathcal{H}$) do not cooperate in the sense that, although $Q_{n,h}^*$ and $Q_{n,h'}^*$ (for any two $h, h' \in \mathcal{H}$, $h \neq h'$) share the same initial estimator Q_n^0, the construction of the latter does not capitalize on that of the former. In contrast, we propose to build collaboratively a continuum of estimators of the form $\Psi(P_{n,h}^*)$ ($h \in \mathcal{H}$) and to select data-adaptively one among them that will be asymptotically Gaussian, under conditions often encountered in empirical process theory.

Organization of the article. In Section 2, we lay out a high-level presentation of collaborative TMLE, and state a high-level result. In Sections 3, 4, 5 and 6, we consider a specific example. In Section 3, we particularize the theoretical construction and analysis. In Section 4, we describe two practical instantiations of the estimator developed in Section 3. In Sections 5 and 6, we carry out a mutli-faceted simulation study of their performances and comment upon its results. In Section 7, we summarize the content of the article. All the proofs are gathered in the appendix.

2 High-level presentation and result

We now state and prove a general result about continuously fine-tuned, collaborative targeted minimum loss estimation, a version of [Theorem 10.1 in 28]. Its high-level assumptions are clarified in the particular example that we study in the next sections.

From now on, we slightly abuse notation and denote $D^*(Q,G)$ instead of $D^*(P)$, where Q and G are the Q- and G-components of P. Let $G_\cdot \equiv \{G_t : t \in \mathcal{T}\} \subset \mathcal{G}$ be a (one-dimensional) subset of $\mathcal{G}$ (indexed by a real parameter ranging in an open subset $\mathcal{T}$ of $\mathcal{H}$) such that $t \mapsto D^*(Q,G_t)(O)$ is twice differentiable over $\mathcal{T}$ for all $Q \in \mathcal{Q}$ (P_0-almost surely). We characterize ∂D^* and $\partial^2 D^*$ by setting, for every $h \in \mathcal{T}$ and $Q \in \mathcal{Q}$,

$$\begin{aligned} \partial_h D^*(Q,G_\cdot)(O) &\equiv \frac{d}{dt} D^*(Q,G_t)(O)|_{t=h}, \\ \partial_h^2 D^*(Q,G_\cdot)(O) &\equiv \frac{d^2}{dt^2} D^*(Q,G_t)(O)|_{t=h}. \end{aligned} \tag{4}$$

Consider the following inter-dependent assumptions. The first one is indexed by $(Q,h,c) \in \mathcal{Q} \times \mathcal{H} \times \mathbb{R}_+^*$.

A1(Q,h,c) There exists an open neighborhood $\mathcal{T} \subset \mathcal{H}$ of $h \in \mathcal{H}$ for which the set $G_{n,\cdot} \equiv \{\hat{G}_h(P_n) \equiv G_{n,h} : h \in \mathcal{T}\} \subset \mathcal{G}$ is such that $t \mapsto D^*(Q,G_{n,\cdot})(O)$ is twice differentiable over $\mathcal{T}$ (P_0-almost surely). Moreover, P_0-almost surely,

$$\sup_{h \in \mathcal{T}} |\partial_h^2 D^*(Q,G_{n,\cdot})(O)| \leq c.$$

A2 For all $h \in \mathcal{H}$, we know how to build $P_{n,h}^* \in \mathcal{M}$, with Q- and G-components denoted by $Q_{n,h}^*$ and $G_{n,h}$, in such a way that $P_n D^*(Q_{n,h}^*, G_{n,h}) = o_P(1/\sqrt{n})$. Moreover, we know how to choose $h_n \in \mathcal{H}$ such that

$$P_n D^*(Q_{n,h_n}^*, G_{n,h_n}) = o_P(1/\sqrt{n}). \tag{5}$$

and, for some deterministic $c_2 > 0$, **A1**(Q_{n,h_n}^*, h_n, c_2) is met and

$$P_n \partial_{h_n} D^*(Q_{n,h_n}^*, G_{n,\cdot}) = o_P(1/n^{1/4}). \tag{6}$$

A3 It holds that $d_{\mathcal{G}}(G_{n,h_n}, G_0) = o_P(1)$, and there exists $Q_1 \in \mathcal{Q}$ such that $d_{\mathcal{Q}}(Q_{n,h_n}^*, Q_1) = o_P(1)$. In addition,

$$\begin{aligned} (P_n - P_0)\left(D^*(Q_{n,h_n}^*, G_{n,h_n}) - D^*(Q_1,G_0)\right) &= o_P(1/\sqrt{n}), & (7) \\ \mathrm{Rem}_{20}(Q_{n,h_n}^*, G_{n,h_n}) - \mathrm{Rem}_{20}(Q_1, G_{n,h_n}) &= o_P(1/\sqrt{n}). & (8) \end{aligned}$$

A4 Let $\Phi_0 : \mathcal{G} \to \mathbb{R}$ be given by $\Phi_0(G) \equiv P_0 D^*(Q_1, G)$. There exist $\tilde{h}_n \in \mathcal{H}$ and $\Delta(P_1) \in L_0^2(P_0)$ such that

$$\Phi_0(G_{n,\tilde{h}_n}) - \Phi_0(G_0) = (P_n - P_0)\Delta(P_1) + o_P(1/\sqrt{n}). \tag{9}$$

A5 It holds that $(h_n - \tilde{h}_n)^2 = o_P(1/\sqrt{n})$. Moreover, there exists a deterministic $c_5 > 0$ such that **A1**(Q_1, h_n, c_5) is met, and

$$\begin{aligned} (P_n - P_0)\left(D^*(Q_1, G_{n,h_n}) - D^*(Q_1, G_{n,\tilde{h}_n})\right) &= o_P(1/\sqrt{n}), & (10) \\ (h_n - \tilde{h}_n) \times P_0\left(\partial_{h_n} D^*(Q_{n,h_n}^*, G_{n,\cdot}) - \partial_{h_n} D^*(Q_1, G_{n,\cdot})\right) &= o_P(1/\sqrt{n}), & (11) \\ (P_n - P_0)\left(\partial_{h_n} D^*(Q_{n,h_n}^*, G_{n,\cdot}) - \partial_{h_n} D^*(Q_1, G_{n,\cdot})\right) &= o_P(1/\sqrt{n}). & (12) \end{aligned}$$

Now that we have introduced our high-level assumptions, we can state the corresponding high-level result that they entail. The proof is relegated to the appendix.

Theorem 1 (Asymptotics of the collaborative TMLE – a high-level result)**.** *Under assumptions **A2** to **A5**, it holds that*

$$\Psi(P^*_{n,h_n}) - \Psi(P_0) = (P_n - P_0)\left(D^*(Q_1, G_0) + \Delta(P_1)\right) + o_P(1/\sqrt{n}). \tag{13}$$

Commenting on the high-level assumptions. Assumption $\mathbf{A1}(Q, h, c)$ concerns both D^* (specifically, how $D^*(Q, G)(O)$ depends on $G(O)$) and algorithms $\hat{G}_t$, $t \in \mathcal{H}$ (specifically, how smooth is $t \mapsto \hat{G}_t(P_n)(O)$ around h). In the particular example studied in the following sections, the counterpart **C1** of $\mathbf{A1}(Q, h, c)$ concerns only algorithms $\hat{G}_t$, $t \in \mathcal{H}$.

In the example, we show how P^*_{n,h_n} can be built collaboratively in such a way that **A2** is met, under a series of nested assumptions about the smoothness of data-dependent, real-valued functions over $\mathcal{H}$, the construction of which notably involve algorithms $\hat{G}_t$, $t \in \mathcal{H}$. To understand why achieving (6) is relevant, observe that the following oracle version of $P_n \partial_{h_n} D^*(Q^*_{n,h_n}, G_{n,.})$,

$$\lim_{\substack{t \to 0 \\ t \neq 0}} \frac{1}{t} P_0 \left(D^*(Q^*_{n,h_n}, G_{n,h_n+t}) - D^*(Q^*_{n,h_n}, G_{n,h_n}) \right),$$

can be rewritten as

$$\lim_{\substack{t \to 0 \\ t \neq 0}} \frac{1}{t} \left(\mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h_n+t}) - \mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h_n}) \right)$$

in view of (1). Thus, achieving (5) relates to finding critical points of $h \mapsto \mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h})$.

Assumption **A3** formalizes the convergence of G_{n,h_n} to its target G_0 w.r.t. $d_{\mathcal{G}}$, and that of Q^*_{n,h_n} to some limit $Q_1 \in \mathcal{Q}$ w.r.t. $d_{\mathcal{Q}}$. It does not require that Q_1 be equal to the target Q_0 of Q^*_{n,h_n}, but **A4** may be impossible to meet when $Q_1 \neq Q_0$ (see below). Condition (7) in **A3** is met for instance if the $L^2(P_0)$-norm of $D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q_1, G_0)$ goes to zero in probability and if the difference falls in a P_0-Donsker class with probability tending to one. As for (8), it typically holds whenever the product of the rates of convergence of Q^*_{n,h_n} and G_{n,h_n} to their limits is $o_P(1/\sqrt{n})$. The counterpart of **A3** in the example studied in the following sections is **C2**.

With **A4**, we assume the existence of an oracle $\tilde{h}_n$ that undersmoothes $G_{n,h}$ enough so that $\Phi_0(G_{n,\tilde{h}_n})$ is an asymptotically linear estimator of $\Phi_0(G_0)$, where we note that Φ_0 is pathwise differentiable in a similar way as Ψ. We say that $\tilde{h}_n$ is an oracle because the definition of Φ_0 involves P_0 and Q_1. It happens that

Lemma 2. *Under **A2** and **A3**, if $Q_1 = Q_0$, $d_{\mathcal{G}}(G_{n,h_n}, G_0) \times d_{\mathcal{Q}}(Q^*_{n,h_n}, Q_0) = o_P(1/\sqrt{n})$, and if* (3) *is met with* $\mathrm{IF} = D^*(P_0)$*, then **A4** holds with $h_n = \tilde{h}_n$ and $\Delta(P_1) = 0$.*

It is difficult to assess whether or not **A4** is a tall order when $d_{\mathcal{G}}(G_{n,h_n}, G_0) \times d_{\mathcal{Q}}(Q^*_{n,h_n}, Q_0)$ is not necessarily $o_P(1/\sqrt{n})$, or if $Q_1 \neq Q_0$.

Finally, **A5** states that the distance between $\tilde{h}_n$ and h_n, introduced in **A2**, is of order $o_P(1/n^{1/4})$ at most. Its conditions (10) and (12) are of similar nature as (7). As for (11), the Cauchy-Schwarz inequality reveals that it is met if the $L^2(P_0)$-norm of $\partial_{h_n} D^*(Q^*_{n,h_n}, G_{n,.}) - \partial_{h_n} D^*(Q_1, G_{n,.})$ is $o_P(1/n^{1/4})$.

3 Collaborative TMLE for continuous tuning when inferring the average treatment effect: presentation and analysis

In this section, we specialize the discussion to the inference of a specific statistical parameter, the so called average treatment effect. Section 3.1 introduces the parameter and recalls what are the corresponding D^* and Rem_{20} from Section 1. Section 3.2 describes the *un*cooperative construction of a continuum of uncooperative TMLEs. Section 3.3 argues why the selection of one of the uncooperative TMLEs is unlikely to yield a well behaved (*i.e.*, asymptotically Gaussian) estimator when the product of the rates of convergence of the estimators of Q_0 and G_0 to their limits is not fast enough (*i.e.*, $o(1/\sqrt{n})$). Then, Sections 3.4 and 3.5 present the collaborative construction of collaborative TMLEs and how to select one among them that is well behaved, under assumptions that are spelled out in Section 3.6, where the high-level Theorem 1 and its assumptions are specialized.

3.1 Preliminary

We observe n independent draws $O_1 \equiv (W_1, A_1, Y_1)$, ..., $O_n \equiv (W_n, A_n, Y_n)$ from P_0, the true law of $O \equiv (W, A, Y)$. It is known that Y takes its values in $[0, 1]$. We consider the statistical model $\mathcal{M}$ that leaves unspecified the law $Q_{W,0}$ of W and the conditional law of Y given (A, W), while we might know that the conditional expectation G_0 of A given W belongs to a set $\mathcal{G}$.

Introduce

$$\bar{Q}_0(A, W) \equiv E_{P_0}(Y|A, W), \quad G_0(W) \equiv P_0(A = 1|W).$$

The parameter of interest is the average treatment effect,

$$\psi_0 \equiv E_{Q_{W,0}}\left(\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\right).$$

We choose it because its study provides a wealth of information and paves the way for the analysis of a variety of other parameters often encountered in the statistical literature.

More generally, every $P \in \mathcal{M}$ gives rise to Q_W, $\bar{Q}(A, W)$, $G(W)$ and $Q \equiv (Q_W, \bar{Q})$, which are respectively the marginal law of W under P, the conditional expectation of Y given (A, W) under P, the conditional probability that $A = 1$ given W under P, and the couple consisting of Q_W and $\bar{Q}$. For each of them, the average treatment effect is $\Psi(P)$, where $\Psi : \mathcal{M} \to [0, 1]$ is given by

$$\Psi(P) \equiv E_{Q_W}\left(\bar{Q}(1, W) - \bar{Q}(0, W)\right).$$

For notational conciseness, we let ℓG be given by

$$\ell G(A, W) \equiv AG(W) + (1 - A)(1 - G(W)) \tag{14}$$

for every $G \in \mathcal{G}$. Note that $\ell G(A, W)$ is the conditional likelihood of A given W when A given W is drawn from the Bernoulli law with parameter $G(W)$, hence the "ℓ" in the notation. Parameter Ψ viewed as a real-valued mapping over $\mathcal{M}$ is pathwise differentiable at every $P \in \mathcal{M}$ w.r.t. the maximal tangent set $\mathcal{S}_P = L_0^2(P)$. The efficient influence curve $D^*(P)$ of Ψ at $P \in \mathcal{M}$ is given by

$$\begin{aligned} D^*(P)(O) &\equiv D_2^*(\bar{Q}, G)(O) + \left(\bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(P)\right) \quad \text{where} \\ D_2^*(\bar{Q}, G)(O) &\equiv \frac{2A - 1}{\ell G(A, W)}(Y - \bar{Q}(A, W)). \end{aligned} \tag{15}$$

Recall definition (1). It is easy to check that, for every $P \in \mathcal{M}$,

$$\mathrm{Rem}_{20}(\bar{Q}, G) = E_{P_0}\left[(2A - 1)\left(1 - \frac{\ell G_0(A, W)}{\ell G(A, W)}\right)\left(\bar{Q}(A, W) - \bar{Q}_0(A, W)\right)\right]. \tag{16}$$

Writing $\mathrm{Rem}_{20}(\bar{Q}, G)$ instead of $\mathrm{Rem}_{20}(Q, G)$ slightly abuses notation, but is justified because integrating out A in the RHS of (16) reveals that it only depends on P_0, $\bar{Q}$ and G. Furthermore, by the Cauchy-Schwartz inequality, it holds that

$$\mathrm{Rem}_{20}(\bar{Q}, G)^2 \leq P_0(\bar{Q} - \bar{Q}_0)^2 \times P_0\left(\frac{G - G_0}{\ell G}\right)^2. \tag{17}$$

3.2 Uncooperative construction of a continuum of uncooperative TMLEs

Prerequisites. Let $\bar{Q}_n^0 \equiv \hat{\bar{Q}}(P_n)$ be an initial estimator of $\bar{Q}_0$ and $\{G_{n,h} \equiv \hat{G}_h(P_n) : h \in \mathcal{H}\}$ be a continuum of candidate estimators of G_0 indexed by a real-valued tuning parameter $h \in \mathcal{H}$, an open interval of $\mathbb{R}_+^*$. By convention, we agree that small values of h correspond with less bias for $G_{n,h}$ as an estimator of G_0. Specifically, denoting L_1 the valid loss function for the estimation of G_0 given by

$$L_1(G)(A, W) \equiv -\log \ell G(A, W) = -A \log G(W) - (1 - A) \log(1 - G(W)), \tag{18}$$

for every $G \in \mathcal{G}$, where ℓG was defined in (14), we assume from now on that the empirical risk $h \mapsto P_n L_1(G_{n,h})$ increases.

For example, $\hat{G}_h$ could correspond to fitting a logistic linear regression maximizing the log-likelihood under the constraint that the sum of the absolute values of the coefficients is smaller than or equal to $1/h$ with $h \in \mathcal{H} \equiv \mathbb{R}_+^*$. We will refer to this algorithm as the LASSO logistic regression algorithm.

Uncooperative TMLEs. Let $Q_{W,n}$ be the empirical law of $\{W_1, \ldots, W_n\}$. Set arbitrarily $h \in \mathcal{H}$ and let $P_{n,h}^0 \in \mathcal{M}$ denote any element of $\mathcal{M}$ such that the marginal law of W under $P_{n,h}^0$ equals $Q_{W,n}$ and the conditional expectation of Y given (A, W) under P_n^0 is equal to $\bar{Q}_n^0$, hence $Q_n^0 = (Q_{W,n}, \bar{Q}_n^0)$ on the one hand; and the conditional expectation of A given W under $P_{n,h}^0$ coincide with $G_{n,h}$ on the other hand. Evaluating Ψ at $P_{n,h}^0$ yields an estimator of $\Psi(P_0)$,

$$\Psi(P_{n,h}^0) = \frac{1}{n} \sum_{i=1}^{n} \left(\bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i)\right),$$

which is not targeted toward the inference of $\Psi(P_0)$ in the sense that none of the known features of P_n^0 was derived specifically for the sake of ultimately estimating $\Psi(P_0)$.

One way to target $\Psi(P_{n,h}^0)$ toward $\Psi(P_0)$ is to build $P_{n,h}^* \in \mathcal{M}$ from $P_{n,h}^0$ in such a way that

$$P_n D^*(P_{n,h}^*) = o_P(1/\sqrt{n})$$

and to infer $\Psi(P_0)$ with $\Psi(P_{n,h}^*)$. This can be achieved by "fluctuating" $P_{n,h}^0$ in the following sense. For every $G \in \mathcal{G}$, introduce the so called "clever covariate" $\mathcal{C}(G)$ given by

$$\mathcal{C}(G)(A, W) \equiv \frac{2A - 1}{\ell G(A, W)}. \tag{19}$$

Now, for every $\varepsilon \in \mathbb{R}$, let $\bar{Q}_{n,h,\varepsilon}^0$ be characterized by

$$\mathrm{logit}\left(\bar{Q}_{n,h,\varepsilon}^0(A, W)\right) \equiv \mathrm{logit}\left(\bar{Q}_n^0(A, W)\right) + \varepsilon \mathcal{C}(G_{n,h})(A, W) \tag{20}$$

and $P^0_{n,h,\varepsilon} \in \mathcal{M}$ be defined like $P^0_{n,h}$ except that the conditional expectation of Y given (A,W) under $P^0_{n,h,\varepsilon}$ equals $\bar{Q}^0_{n,h,\varepsilon}$ (and not $\bar{Q}^0_n$). Clearly, $P^0_{n,h,\varepsilon} = P^0_{n,h}$ when $\varepsilon = 0$. Moreover, denoting L_2 the loss function given by

$$L_2(\bar{Q})(O) \equiv -Y \log \bar{Q}(A,W) - (1-Y) \log \big(1 - \bar{Q}(A,W)\big)$$

for every $\bar{Q}$ induced by a $P \in \mathcal{M}$, it holds that

$$\frac{d}{d\varepsilon} L_2(\bar{Q}^0_{n,h,\varepsilon})(O) = -D_2(\bar{Q}^0_{n,h,\varepsilon}, G_{n,h})(O),$$

a property that prompts us to say that the one-dimensional submodel $\{P^0_{n,h,\varepsilon} : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$ "fluctuates" $P^0_{n,h}$ "in the direction of" $D_2(\bar{Q}^0_n, G_{n,h})$.

The optimal fluctuation of $P^0_{n,h}$ along the above submodel is indexed by the minimizer of the empirical risk

$$\varepsilon_{n,h} \equiv \arg\min_{\varepsilon \in \mathbb{R}} P_n L_2(\bar{Q}^0_{n,h,\varepsilon}), \tag{21}$$

of which the existence is assumed (note that $\varepsilon \mapsto P_n L_2(\bar{Q}^0_{n,h,\varepsilon})$ is twice differentiable and strictly convex). We call $P^*_{n,h} \equiv P^0_{n,h,\varepsilon_{n,h}}$ the TMLE of P_0, and the resulting estimator

$$\psi^*_{n,h} \equiv \Psi(P^*_{n,h}) = \frac{1}{n} \sum_{i=1}^{n} \Big(\bar{Q}^0_{n,h,\varepsilon_{n,h}}(1, W_i) - \bar{Q}^0_{n,h,\varepsilon_{n,h}}(0, W_i) \Big) \tag{22}$$

the TMLE of $\Psi(P_0)$. It is readily seen that (22) is equivalent to

$$P_n \left(D^*(P^*_{n,h}) - D^*_2(\bar{Q}^*_{n,h}, G_{n,h}) \right) = 0$$

where $\bar{Q}^*_{n,h} \equiv \bar{Q}^0_{n,h,\varepsilon_{n,h}}$. Since $\varepsilon_{n,h}$ minimizes the differentiable mapping $\varepsilon \mapsto P_n L_2(Q^0_{n,h,\varepsilon})$, it holds moreover that

$$P_n D^*_2(\bar{Q}^*_{n,h}, G_{n,h}) = 0 \tag{23}$$

which, combined with the previous display, yields

$$P_n D^*(P^*_{n,h}) = 0; \tag{24}$$

in words, $\psi^*_{n,h}$ is targeted toward $\Psi(P_0)$ indeed. Furthermore, in view of (16) and (24), $\psi^*_{n,h}$ satisfies

$$\psi^*_{n,h} - \Psi(P_0) = (P_n - P_0) D^*(P^*_{n,h}) + \mathrm{Rem}_{20}(\bar{Q}^*_{n,h}, G_{n,h}). \tag{25}$$

Finally, the TMLEs $\psi^*_{n,h}$ $(h \in \mathcal{H})$ are said uncooperative because, although they share the same initial estimator $\bar{Q}^0_n$, for any two $h, h' \in \mathcal{H}$, $h \neq h'$, the construction of $\psi^*_{n,h}$ does not capitalize on that of $\psi^*_{n,h'}$.

3.3 Selecting one of the uncooperative TMLEs

At this stage of the procedure, a crucial question is to select one TMLE in the collection of uncooperative TMLEs, one that lends itself to the construction of a CI for $\Psi(P_0)$ with a given asymptotic level. Such a TMLE necessarily writes as ψ^*_{n,h_n} for some well chosen $h_n \in \mathcal{H}$. This could possibly be a deterministic (fixed in n) or a data-driven (random and n-dependent) element of $\mathcal{H}$.

The risk R_1 generated by L_1 (18) is given by

$$R_1(G, G_0) \equiv E_{Q_{0,W}} \left[\mathrm{KL}(G_0(W), G(W)) \right],$$

where $\mathrm{KL}(p,q)$ is the Kullback-Leibler divergence between the Bernoulli laws with parameters $p,q \in [0,1]$. By Pinsker's inequality, it holds that

$$0 \leq 2P_0 \left(G - G_0\right)^2 \leq R_1(G, G_0)$$

for all $G \in \mathcal{G}$. Therefore, if G is bounded away from zero and one, then (17) implies

$$\mathrm{Rem}_{20}(\bar{Q}, G)^2 \lesssim P_0(\bar{Q} - \bar{Q}_0)^2 \times R_1(G, G_0). \tag{26}$$

If the deterministic $h_n \in \mathcal{H}$ is such that *(i)* there exist two rates $\rho_{1,n} = o(1)$ and $\rho_{2,n} = o(1)$ such that $R_1(G_{n,h_n}, G_0) = o_P(\rho_{1,n}^2)$ and $P_0(\bar{Q}^*_{n,h_n} - \bar{Q}_0)^2 = o_P(\rho_{2,n}^2)$, *(ii)* $P_0 \left(D^*(P^*_{n,h_n}) - D^*(P_0)\right)^2 = o_P(1)$, *(iii)* $D^*(P^*_{n,h_n})$ falls in a P_0-Donsker class with P_0-probability tending to one, *(iv)* $o_P(\rho_{1,n}\rho_{2,n}) = o_P(1/\sqrt{n})$, then [32, Lemma 19.24], (25) and (26) guarantee that (3) is met (with $\mathrm{IF} = D^*(P_0)$). [This argument will be used repeatedly throughout the article.] Thus, by the central limit theorem, $\sqrt{n}(\psi^*_{n,h_n} - \Psi(P_0))$ converges in law to the centered Gaussian law with variance $\mathrm{Var}_{P_0}(D^*(P_0)(O))$. So, if the synergy between the convergences of $\bar{Q}^*_{n,h_n}$ and G_{n,h_n} to their respective limits $\bar{Q}_0$ and G_0 is sufficient, then the TMLE ψ^*_{n,h_n} can be used to build CIs.

The argument falls apart if $o_P(\rho_{1,n}\rho_{2,n})$ is not $o_P(1/\sqrt{n})$ (or, worse, if the $L^2(P_0)$-limit $\bar{Q}_1$ of $\bar{Q}^*_{n,h_n}$ is not $\bar{Q}_0$, because we do not expect that $R_1(G_{n,h_n}, G_0) = o_P(1/n)$). In that case, whether or not it is possible to derive a useful asymptotic linear expansion of a TMLE ψ^*_{n,h_n} similar to (3) will depend on whether or not we can derive an asymptotic linear expansion for $\sqrt{n}\,\mathrm{Rem}_{20}(\bar{Q}^*_{n,h_n}, G_{n,h_n})$.

If G_{n,h_n} was derived by maximizing the likelihood over a correctly specified, finite-dimensional and fine-tune-parameter-free parametric model, then $\sqrt{n}\,\mathrm{Rem}_{20}(\bar{Q}^*_{n,h_n}, G_{n,h_n})$ would be asymptotically linear. Because of how we estimate G_0, we now argue that there is little chance that we can select $h_n \in \mathcal{H}$ such that the remainder term $\sqrt{n}\,\mathrm{Rem}_{20}(\bar{Q}^*_{n,h_n}, G_{n,h_n})$ is asymptotically linear. A natural choice would be to use the likelihood-based cross-validation selector $h_{n,\mathrm{CV}}$. Let us recall how it is derived and explain why we do not believe it will solve our problem.

Let $B_n \in \{0,1\}^n$ be a cross-validation scheme. For instance, B_n could be a V-fold cross-validation scheme, *i.e.*, a random vector taking V different values $b_1, \ldots, b_V \in \{0,1\}^n$, each with probability $1/V$, such that *(i)* the proportion $n^{-1}\sum_{i=1}^n b_v(i)$ of ones among the coordinates of each b_v is close to $1/V$, and *(ii)* $\sum_{v=1}^V b_v(i) = 1$ for all $1 \leq i \leq n$. Let P^0_{n,B_n} be the empirical probability law of the training subsample $\{O_i : B_n(i) = 0, 1 \leq i \leq n\}$ and P^1_{n,B_n} be the empirical probability law of the validation subsample $\{O_i : B_n(i) = 1, 1 \leq i \leq n\}$. The likelihood-based cross-validation selector $h_{n,\mathrm{CV}}$ of $h \in \mathcal{H}$ is given by

$$h_{n,\mathrm{CV}} \equiv \arg\min_{h \in \mathcal{H}} E_{B_n}\left[P^1_{n,B_n} L_1(\hat{G}_h(P^0_{n,B_n}))\right]. \tag{27}$$

Unfortunately, we do not expect that $\sqrt{n}\,\mathrm{Rem}_{20}(\bar{Q}^*_{n,h}, G_{n,h})$ is asymptotically linear. Heuristically, $h_{n,\mathrm{CV}}$ trades off the bias and variance of $G_{n,h}$ as an estimator of G_0, whereas we wish to trade off this bias with the variance of $\psi^*_{n,h}$. Clearly, the variance of the estimator $\psi^*_{n,h} = \Psi(P^*_{n,h})$, where Ψ is a smooth functional, is significantly smaller than that of the infinite-dimensional object $G_{n,h}$.

3.4 Collaborative construction of finitely many collaborative TMLEs

The take-home message of Sections 3.2 and 3.3 is that the *uncooperative* construction of a continuum of standard TMLEs will typically fail to produce one asymptotically linear TMLE if the product of the rates of convergence of the estimators of $\bar{Q}_0$ and G_0 to their limits is not fast enough (*i.e.*,

$o(1/\sqrt{n})$). In Sections 3.4 and 3.5, we demonstrate how a *collaborative* construction of a continuum of standard TMLEs can produce one asymptotically linear TMLE in this challenging situation, under appropriate assumptions.

Recursive construction. We now present the collaborative construction of *finitely many* TMLEs. In the forthcoming theoretical presentation, we make on the fly a series of assumptions. The most important ones will be emphasized.

We argued that the cross-validated selector $h_{n,\mathrm{CV}}$ (27) does not sufficiently undersmooth $G_{n,h}$ to make of $\sqrt{n}\,\mathrm{Rem}_{20}(\bar{Q}^*_{n,h}, G_{n,h})$ an asymptotically linear term. Since we have assumed that $h \mapsto P_n L_1(G_{n,h})$ increases, we can focus on those tuning parameters h in $\mathcal{H}\cap]0, h_{n,\mathrm{CV}}]$, a set assumed non-empty from now (an assumption that we call $\mathbf{B1}(P_n, 1)$).

The construction is recursive. It unfolds as follows.

Initialization. We begin as in Section 3.2: for every $h \in \mathcal{H}\cap]0, h_{n,\mathrm{CV}}]$, we build $\bar{Q}^{(*)}_{n,h}$ and $P^{(*)}_{n,h}$ using $\bar{Q}^0_n$ as an initial estimator of $\bar{Q}_0$ and $G_{n,h}$ as the estimator of G_0. Note that placing the star symbol between parentheses suggests that $\bar{Q}^{(*)}_{n,h}$ and $P^{(*)}_{n,h}$ are the *tentative* h-specific estimator of $\bar{Q}_0$ and TMLE. Specifically, for every $h \in \mathcal{H}\cap]0, h_{n,\mathrm{CV}}]$, we define $\bar{Q}^0_{n,h,\varepsilon}$ as in (20), $\varepsilon_{n,h,1}$ as in (21), assuming that it exists (an assumption that we call $\mathbf{B2}(P_n, 1)$), then set $\bar{Q}^{(*)}_{n,h} \equiv \bar{Q}^0_{n,h,\varepsilon_{n,h,1}}$ and find $P^{(*)}_{n,h} \in \mathcal{M}$ such that the marginal law of W under $P^{(*)}_{n,h}$ is the empirical law $Q_{W,n}$ of $\{W_1, \ldots, W_n\}$ and the conditional expectation of Y given (A, W) under $P^{(*)}_{n,h}$ equals $\bar{Q}^{(*)}_{n,h}$, hence $Q^{(*)}_{n,h} = (Q_{W,n}, \bar{Q}^{(*)}_{n,h})$ on the one hand; and the conditional expectation of A given W under $P^{(*)}_{n,h}$ coincides with $G_{n,h}$ on the other hand.

We assume that $h \mapsto P_n L_2(Q^{(*)}_{n,h})$ is minimized globally at $h_{n,1}$ in the interior of $\mathcal{H}\cap]0, h_{n,\mathrm{CV}}]$ (an assumption that we call $\mathbf{B3}(P_n, 1)$). If there are several minimizers, then $h_{n,1}$ is the largest of them by choice. Observe that, for every $h \in \mathcal{H}\cap]0, h_{n,\mathrm{CV}}]$,

$$P_n L_2(\bar{Q}^{(*)}_{n,h_{n,1}}) \leq P_n L_2(\bar{Q}^{(*)}_{n,h}) \leq P_n L_2(\bar{Q}^0_{n,h})$$

and, in particular,

$$P_n L_2(\bar{Q}^{(*)}_{n,h_{n,1}}) < P_n L_2(\bar{Q}^*_{n,h_{n,\mathrm{CV}}}) \leq P_n L_2(\bar{Q}^0_{n,h_{n,\mathrm{CV}}}).$$

Let us assume now that, in addition, $h \mapsto \varepsilon_{n,h,1}$, $h \mapsto 1/G_{n,h}(W_i)$ and $h \mapsto 1/(1 - G_{n,h}(W_i))$ (all $1 \leq i \leq n$) are differentiable in an open neighborhood of $h_{n,1}$ (an assumption that we call $\mathbf{B4}(P_n, 1)$). Consequently, *(i)* $\partial_{h_{n,1}} D^*(\bar{Q}^{(*)}_{n,h_{n,1}}, G_{n,\cdot})(O_i)$ is well defined for each $1 \leq i \leq n$ (see (4)), and *(ii)* $h \mapsto P_n L_2(\bar{Q}^{(*)}_{n,h})$ is differentiable in that neighborhood. Moreover, since $h_{n,1}$ minimizes the previous mapping, we have

$$\begin{aligned}
0 &= -\frac{d}{dt} P_n L_2(\bar{Q}^{(*)}_{n,t})\Big|_{t=h_{n,1}} \\
&= \left(\frac{d}{dt}\varepsilon_{n,t,1}\Big|_{t=h_{n,1}}\right) \times P_n D_2^*(\bar{Q}^{(*)}_{n,h_{n,1}}, G_{n,h_{n,1}}) + \varepsilon_{n,h,1} \times P_n \partial_{h_{n,1}} D^*(\bar{Q}^{(*)}_{n,h_{n,1}}, G_{n,\cdot}) \\
&= \varepsilon_{n,h,1} \times P_n \partial_{h_{n,1}} D^*(\bar{Q}^{(*)}_{n,h_{n,1}}, G_{n,\cdot}),
\end{aligned}$$

where the third equality holds because

$$P_n D_2^*(\bar{Q}^{(*)}_{n,h_{n,1}}, G_{n,h_{n,1}}) = P_n D_2^*(\bar{Q}^0_{n,h,\varepsilon_{n,h,1}}, G_{n,h_{n,1}}) = 0$$

in light of (23). If $\varepsilon_{n,h,1} \neq 0$ (an assumption that we call $\mathbf{B5}(P_n, 1)$), then we thus have proven that the following equation is solved

$$P_n \partial_{h_{n,1}} D^*(\bar{Q}^{(*)}_{n,h_{n,1}}, G_{n,\cdot}) = 0.$$

To complete the initialization, we define $h_{n,0} \equiv h_{n,\mathrm{CV}}$, $\bar{Q}^*_{n,h_{n,1}} \equiv \bar{Q}^{(*)}_{n,h_{n,1}}$, $Q^*_{n,h_{n,1}} \equiv Q^{(*)}_{n,h_{n,1}}$, $P^*_{n,h_{n,1}} \equiv P^{(*)}_{n,h_{n,1}}$, $\psi^*_{n,h_{n,1}} \equiv \Psi(P^*_{n,h_{n,1}})$, and note that they satisfy

$$P_n \partial_{h_{n,1}} D^*(\bar{Q}^*_{n,h_{n,1}}, G_{n,\cdot}) = P_n D^*(P^*_{n,h_{n,1}}) = 0 \qquad \text{and}$$
$$P_n L_2(\bar{Q}^*_{n,h_{n,1}}) < P_n L_2(\bar{Q}^*_{n,h_{n,0}})$$

(recall how (23) implied (24) earlier).

Recursion. Let $k \geq 2$ be arbitrarily chosen. Suppose that, for all $1 \leq \ell < k$, we have already built the 5-tuples $(h_{n,\ell}, \bar{Q}^*_{n,h_{n,\ell}}, Q^*_{n,h_{n,\ell}}, P^*_{n,h_{n,\ell}}, \psi^*_{n,h_{n,\ell}})$ under assumptions $\mathbf{B1}(P_n, \ell)$ to $\mathbf{B5}(P_n, \ell)$, and also that $\mathcal{H} \cap]0, h_{n,k-1}] \neq \emptyset$ (an assumption that we call $\mathbf{B1}(P_n, k)$). Let us now present the construction of $(h_{n,k}, \bar{Q}^*_{n,h_{n,k}}, Q^*_{n,h_{n,k}}, P^*_{n,h_{n,k}})$ under assumptions $\mathbf{B1}(P_n, k)$ to $\mathbf{B5}(P_n, k)$. Because the presentation is very similar to that of the initialization, it is more laid out more directly.

For every $h \in \mathcal{H} \cap]0, h_{n,k-1}]$, we build *again* $\bar{Q}^{(*)}_{n,h}$ and $P^{(*)}_{n,h}$ *but* using $\bar{Q}^*_{n,h_{n,k-1}}$ as an initial estimator of $\bar{Q}_0$ and $G_{n,h}$ as the estimator of G_0. Specifically, for every $h \in \mathcal{H} \cap]0, h_{n,k-1}]$, we define $\bar{Q}^{k-1}_{n,h,\varepsilon}$ as in (20) with $\bar{Q}^*_{n,h_{n,k-1}}$ substituted for $\bar{Q}^0_n$, $\varepsilon_{n,h,k}$ as in (21) with $\bar{Q}^{k-1}_{n,h,\varepsilon}$ substituted for $\bar{Q}^0_{n,h,\varepsilon}$ ($\mathbf{B2}(P_n, k)$ assumes the existence of $\varepsilon_{n,h,k}$), then set $\bar{Q}^{(*)}_{n,h} \equiv \bar{Q}^k_{n,h,\varepsilon_{n,h,k}}$ and find $P^{(*)}_{n,h} \in \mathcal{M}$ such that the marginal law of W under $P^{(*)}_{n,h}$ is the empirical law $Q_{W,n}$ of $\{W_1, \ldots, W_n\}$ and the conditional expectation of Y given (A, W) under $P^{(*)}_{n,h}$ equals $\bar{Q}^{(*)}_{n,h}$, hence $Q^{(*)}_{n,h} = (Q_{W,n}, \bar{Q}^{(*)}_{n,h})$ on the one hand; and the conditional expectation of A given W under $P^{(*)}_{n,h}$ coincides with $G_{n,h}$ on the other hand.

We assume that $h \mapsto P_n L_2(Q^{(*)}_{n,h})$ is minimized globally at $h_{n,k}$ in the interior of $\mathcal{H} \cap]0, h_{n,k-1}]$ (an assumption that we call $\mathbf{B3}(P_n, k)$). If there are several minimizers, then $h_{n,k}$ is the largest of them by choice. Moreover, we also assume that $h \mapsto \varepsilon_{n,h,k}$, $h \mapsto 1/G_{n,h}(W_i)$ and $h \mapsto 1/(1 - G_{n,h}(W_i))$ (all $1 \leq i \leq n$) are differentiable in an open neighborhood of $h_{n,k}$ (an assumption that we call $\mathbf{B4}(P_n, k)$). Consequently, $\partial_{h_{n,k}} D^*(\bar{Q}^{(*)}_{n,h_{n,k}}, G_{n,\cdot})(O_i)$ is well defined for each $1 \leq i \leq n$ (see (4)), $h \mapsto P_n L_2(\bar{Q}^{(*)}_{n,h})$ is differentiable in that neighborhood and, since $h_{n,k}$ minimizes the previous mapping,

$$\varepsilon_{n,h,k} \times P_n \partial_{h_{n,k}} D^*(\bar{Q}^{(*)}_{n,h_{n,k}}, G_{n,\cdot}) = 0.$$

If $\varepsilon_{n,h,k} \neq 0$ (an assumption that we call $\mathbf{B5}(P_n, k)$), then it holds that

$$P_n \partial_{h_{n,k}} D^*(\bar{Q}^{(*)}_{n,h_{n,k}}, G_{n,\cdot}) = 0.$$

To complete the presentation and the recursion, we define $\bar{Q}^*_{n,h_{n,k}} \equiv \bar{Q}^{(*)}_{n,h_{n,k}}$, $Q^*_{n,h_{n,k}} \equiv Q^{(*)}_{n,h_{n,k}}$, $P^*_{n,h_{n,k}} \equiv P^{(*)}_{n,h_{n,k}}$, $\psi^*_{n,h_{n,k}} \equiv \Psi(P^*_{n,h_{n,k}})$, and note that they satisfy

$$P_n \partial_{h_{n,k}} D^*(\bar{Q}^*_{n,h_{n,k}}, G_{n,\cdot}) = P_n D^*(P^*_{n,h_{n,k}}) = 0 \qquad \text{and} \tag{28}$$

$$P_n L_2(\bar{Q}^*_{n,h_{n,\ell}}) < P_n L_2(\bar{Q}^*_{n,h_{n,\ell-1}})$$

for all $1 \leq \ell \leq k$.

We discuss when to stop the loop in the next paragraph. The collection $\{P^*_{n,h_{n,k}} : 0 \leq k \leq K_n\}$ of TMLEs is arguably built collaboratively, as the derivation of every $P^*_{n,h_{n,\ell}}$ heavily depends on $P^*_{n,h_{n,\ell-1}}$.

The loop is iterated until a stopping criterion is met. The instantiations of the collaborative TMLE laid out in Section 4 rely on the LASSO logistic regression algorithm. It is thus possible to pre-specify an upper bound on K_n. In general, we may decide to stop the recursive construction whenever a maximal number K of iterations has been reached, or $h_{n,k} \leq \hbar$, or M successive TMLEs $\psi^*_{n,h_{n,k+m}}$ $(0 \leq m < M)$ all belong to an interval of length smaller than $\eta_{n,k}$, for some user-supplied integers $K_{\max}$, M and small positive numbers $h_{\min}$ and $\eta_{n,k}$, the former chosen such that $\mathcal{H} \cap]0, h_{\min}[$ is non-empty and the latter possibly sample-size- and data-driven. The choice of $K_{\max}$ would typically be driven by considerations about the computational time. The choice of $h_{\min}$ would typically depend on the collection $\{\hat{G}_h : h \in \mathcal{H}\}$ of h-specific algorithms, $h \leq h_{\min}$ meaning that too much undersmoothing is certainly at play when using $\hat{G}_h$. We would suggest choosing $M \equiv 3$ and characterizing $\eta_{n,k}$ by $\eta^2_{n,k} \equiv \Upsilon_{P_n}(P^*_{n,h_{n,k}})/10n$ with $\Upsilon_{P_n} : \mathcal{M} \to \mathbb{R}^*_+$ given by

$$\Upsilon_{P_n}(P) \equiv E_{P_n}\left[D^*(P)(O)^2\right] = \frac{1}{n}\sum_{i=1}^{n} D^*(P)(O_i)^2.$$

The definition of Υ_{P_n} is justified by the fact that $\Upsilon_{P_n}(D^*(P^*_{n,h_n}))$ estimates the asymptotic variance of the TMLE $\Psi(P^*_{n,h_n})$ in the context where we prove (3) (with IF $= D^*(P_0)$) in Section 3.3.

3.5 Selecting one of the finitely many collaborative TMLEs

It remains to determine which TMLE to select among the collection of collaborative TMLEs that we constructed in Section 3.4. Again, the selection hinges on the cross-validation principle.

The recursive construction described in Section 3.4 can be applied to the empirical measure $\mathbb{P}_n$ of any subset of the complete data set. Starting from $h_{n,\mathrm{CV}}$ (as defined in (27) even when $\mathbb{P}_n$ differs from P_n), let the 5-tuple $(\mathbb{H}_{n,1}, \bar{\mathbb{Q}}^*_{n,\mathbb{H}_{n,1}}, \mathbb{Q}^*_{n,\mathbb{H}_{n,1}}, \mathbb{P}^*_{n,\mathbb{H}_{n,1}}, \Psi(\mathbb{P}^*_{n,\mathbb{H}_{n,1}}))$ be defined like the 5-tuple $(h_{n,1}, \bar{Q}^*_{n,h_{n,1}}, Q^*_{n,h_{n,1}}, P^*_{n,h_{n,1}}, \psi^*_{n,h_{n,1}})$ with $\mathbb{P}_n$ substituted for P_n, under assumptions $\mathbf{B1}(\mathbb{P}_n, 1)$ to $\mathbf{B5}(\mathbb{P}_n, 1)$. Then, recursively, let $(\mathbb{H}_{n,k}, \bar{\mathbb{Q}}^*_{n,\mathbb{H}_{n,k}}, \mathbb{Q}^*_{n,\mathbb{H}_{n,k}}, \mathbb{P}^*_{n,\mathbb{H}_{n,k}}, \Psi(\mathbb{P}^*_{n,\mathbb{H}_{n,k}}))$ be defined like $(h_{n,k}, \bar{Q}^*_{n,h_{n,k}}, Q^*_{n,h_{n,k}}, P^*_{n,h_{n,k}}, \psi^*_{n,h_{n,k}})$ with $\mathbb{P}_n$ substituted for P_n, under assumptions $\mathbf{B1}(\mathbb{P}_n, k)$ to $\mathbf{B5}(\mathbb{P}_n, k)$. The recursive construction is stopped when $\mathbb{K}_n$ 5-tuples have been derived, where $\mathbb{K}_n$ is defined like K_n with $\mathbb{P}_n$ substituted for P_n.

The collection

$$\left\{(\mathbb{H}_{n,k}, \bar{\mathbb{Q}}^*_{n,\mathbb{H}_{n,k}}, \mathbb{Q}^*_{n,\mathbb{H}_{n,k}}, \mathbb{P}^*_{n,\mathbb{H}_{n,k}}, \Psi(\mathbb{P}^*_{n,\mathbb{H}_{n,k}})) : 1 \leq k \leq \mathbb{K}_n\right\} \tag{29}$$

of $\mathbb{K}_n$ collaborative TMLEs is used to define a continuum of collaborative TMLEs in the following straightforward way. The challenge is to associate a 4-tuple $(\bar{\mathbb{Q}}^*_{n,h}, \mathbb{Q}^*_{n,h}, \mathbb{P}^*_{n,h}, \Psi(\mathbb{P}^*_{n,h}))$ to any $h \in \mathcal{H} \cap]0, h_{n,\mathrm{CV}}]$. To do so, we simply let $\mathbb{H}_n(h)$ be the element of $\{\mathbb{H}_{n,k} : 1 \leq k \leq \mathbb{K}_n\}$ that is closest to h (with a preference for the larger of the two closer ones when h is right in the middle), that is, formally, we set

$$\mathbb{H}_n(h) \equiv \max\left\{\mathbb{H}_{n,k} : |h - \mathbb{H}_{n,k}| = \min\{|h - \mathbb{H}_{n,\ell}| : 1 \leq \ell \leq \mathbb{K}_n\}\right\} \tag{30}$$

and associate to h the corresponding 4-tuple

$$(\bar{\mathbb{Q}}^*_{n,\mathbb{H}_n(h)}, \mathbb{Q}^*_{n,\mathbb{H}_n(h)}, \mathbb{P}^*_{n,\mathbb{H}_n(h)}, \Psi(\mathbb{P}^*_{n,\mathbb{H}_n(h)})).$$

Let B_n be the cross-validation scheme introduced in Section 3.3. By convention, let the max of the empty set be 0. The collaborative TMLE that we select is

$$(\bar{Q}^*_{n,h_{n,\kappa_n}}, Q^*_{n,h_{n,\kappa_n}}, P^*_{n,h_{n,\kappa_n}}, \psi^*_{n,h_{n,\kappa_n}}) \tag{31}$$

where κ_n is given by

$$\kappa_n \equiv 1 \vee \max \left\{ 1 \leq k \leq K_n : h_{n,k} \geq \underset{h \in \mathcal{H} \cap]0, h_{n,\mathrm{CV}}]}{\arg\min} E_{B_n} \left[P^1_{n,B_n} L_2(\bar{\mathbb{Q}}^*_{n,\mathbb{H}_n(h)}\big|_{\mathbb{P}_n = P^0_{n,B_n}})) \right] \right\}. \tag{32}$$

In words, κ_n is the unique element of $\{1, \ldots, K_n\}$ such that h_{n,κ_n} is the smallest element of $\{h_{n,1}, \ldots, h_{n,K_n}\}$ that is larger than the minimizer of the cross-validated L_2-risk of the collaborative TMLE, if there exists such an element, and 1 otherwise. In (32), $\bar{\mathbb{Q}}^*_{n,\mathbb{H}_n(h)}\big|_{\mathbb{P}_n = P^0_{n,B_n}}$ equals $\bar{\mathbb{Q}}^*_{n,\mathbb{H}_n(h)}$ when $\mathbb{P}_n = P^0_{n,B_n}$.

The contrast between h_{n,κ_n} and $h_{n,\mathrm{CV}}$ is stark. At first glance, the main difference is that the role play by cross-validated L_1-risks of algorithms to estimate G_0 in (27) is played by cross-validated L_2-risks of algorithms to estimate $\bar{Q}_0$ in (32). A closer examination reveals that the difference is deeper. Replacing $L_1(\hat{G}_h(P^0_{n,B_n}))$ by $L_2(\bar{Q}^{0*}_{n,B_n h})$ (with $\bar{Q}^{0*}_{n,B_n,h}$ defined like $\bar{Q}^*_{n,h}$ in Section 3.2 but based on P^0_{n,B_n} instead of P_n) would not make of the resulting alternative cross-validated selector of h a good candidate: because of the inherent lack of cooperation between the uncooperative TMLEs $\psi^*_{n,h}$, the resulting estimator of G_0 would not even be consistent. This fact motivates the general C-TMLE methodology, of which the present instantiation includes a twist consisting in solving two critical equations, see (28).

3.6 Asymptotics

The study of the asymptotic properties of the collaborative TMLE $\psi^*_{n,h_{n,\kappa_n}}$ hinges on Theorem 1. We first specify two pseudo-distances $d_{\mathcal{G}}$ and $d_{\mathcal{Q}}$ on $\mathcal{G}$ and $\mathcal{Q}$ in light of requirement (1). On the one hand, because we will eventually assume that G_0 is bounded away from zero and one, (16) yields that we can choose $d_{\mathcal{G}}$ such that, for each $G_1, G_2 \in \mathcal{G}$,

$$d_{\mathcal{G}}(G_1, G_2)^2 \equiv P_0(G_1 - G_2)^2.$$

On the other hand, note that any data-dependent $Q_n \equiv (Q_{W,n}, \bar{Q}_n) \in \mathcal{Q}$ naturally gives rise to a substitution estimator ψ_n of ψ_0:

$$\psi_n \equiv E_{Q_{W,n}} \left(\bar{Q}_n(1, W) - \bar{Q}_n(0, W) \right) = \frac{1}{n} \sum_{i=1}^{n} \left(\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \right).$$

It is easy to check (see the appendix) that the following result holds.

Lemma 3. *Assume that G_0 is bounded away from zero and one and that $\bar{Q}_n(1, \cdot) - \bar{Q}_n(0, \cdot)$ falls in a P_0-Donsker class with P_0-probability tending to one. If $P_0(\bar{Q}_n - \bar{Q}_1)^2 = o_P(1)$ for some $\bar{Q}_1$, then $\psi_n = P_0(\bar{Q}_1(1, \cdot) - \bar{Q}_1(0, \cdot)) + o_P(1)$.*

Since we always estimate the marginal law of W under P_0, $Q_{W,0}$, with its empirical counterpart $Q_{W,n}$, we can thus define the pseudo-distance $d_{\mathcal{Q}}$ by setting, for each $Q_1, Q_2 \in \mathcal{Q}$,

$$d_{\mathcal{Q}}(Q_1, Q_2)^2 \equiv P_0(\bar{Q}_1 - \bar{Q}_2)^2,$$

an expression that does not depend on the first components of Q_1 and Q_2.

Consider the following inter-dependent assumptions. The first one is related to $\mathbf{A1}(Q, h, c)$ and completes $\mathbf{B5}(P_n, h_{n,\kappa_n})$.

C1 There exists a universal constant $C_1 \in]0, 1/2[$ such that G_0 and any by-product $G_{n,h}$ of algorithm $\hat{G}_h$ (any $h \in \mathcal{H}$) trained on the empirical measure P_n take their values in $[C_1, 1 - C_1]$. Moreover, there exists an open neighborhood $\mathcal{T} \subset \mathcal{H}$ of h_{n,κ_n} and a universal constant $C_2 > 0$ such that $t \mapsto G_{n,t}(W)$ is twice differentiable over $\mathcal{T}$ (P_0-almost surely) and, P_0-almost surely,

$$\sup_{h \in \mathcal{T}} \left| \frac{d}{dt} G_{n,t}(W)|_{t=h} \right| \vee \sup_{h \in \mathcal{T}} \left| \frac{d^2}{dt^2} G_{n,t}(W)|_{t=h} \right| \leq C_2.$$

When **C1** is met, we denote $G'_{n,h}(W)$ the first derivative of $t \mapsto G_{n,t}(W)$ at $h \in \mathcal{H}$.

C2 Both $G_{n,h_{n,\kappa_n}}$ and $\bar{Q}^*_{n,h_{n,\kappa_n}}$ converge in $L^2(P_0)$, to G_0 and $\bar{Q}_1$ respectively. Moreover, it holds that $P_0(G_{n,h_{n,\kappa_n}} - G_0)^2 = o_P(1/\sqrt{n})$ and $P_0(G_{n,h_{n,\kappa_n}} - G_0)^2 \times P_0(\bar{Q}^*_{n,h_{n,\kappa_n}} - \bar{Q}_1)^2 = o_P(1/n)$.

C3 Assumption **A4** is met, $(h_{n,\kappa_n} - \tilde{h}_n)^2 = o_P(1/\sqrt{n})$ and $(h_{n,\kappa_n} - \tilde{h}_n)^2 \times P_0(\bar{Q}^*_{n,h_{n,\kappa_n}} - \bar{Q}_1)^2 = o_P(1/n)$.

C4 With P_0-probability tending to one, $\bar{Q}^*_{n,h_{n,\kappa_n}}$, $G_{n,h_{n,\kappa_n}}$, $G_{n,\tilde{h}_n}$ and $G'_{n,h_{n,\kappa_n}}$ fall in P_0-Donsker classes.

We are now in a position to state the corollary of Theorem 1 that describes the asymptotic properties of the collaborative TMLE targeting the average treatment effect.

Corollary 4 (Asymptotics of the collaborative TMLE – targeting the average treatment effect)**.** *Suppose that assumptions **B1**$(\cdot,\cdot)$ to **B5**$(\cdot,\cdot)$ that we made in Sections 3.4 and 3.5 when constructing the collaborative TMLE given in* (31) *are met. In addition, suppose that **C1** to **C4** are satisfied. Then*

$$\psi^*_{n,h_{n,\kappa_n}} - \Psi(P_0) = (P_n - P_0)\left(D^*(Q_1, G_0) + \Delta(P_1)\right) + o_P(1/\sqrt{n}).$$

By the central limit theorem, the corrolary implies that $\sqrt{n}(\psi^*_{n,h_{n,\kappa_n}} - \Psi(P_0))$ converges in law to the centered Gaussian law with a variance $\sigma^2 \equiv P_0(D^*(Q_1, G_0) + \Delta(P_1))^2$. Therefore, provided that we can estimate σ^2 consistently (or conservatively), we can build CIs for $\Psi(P_0)$ with a given asymptotic level. Sections 4, 5 and 6 investigate the practical implementation of the collaborative TMLE and its performances in a simulation study.

4 Collaborative TMLE for continuous tuning when inferring the average treatment effect: practical implementation

In this section, we describe the practical implementation of the two instantiations of the collaborative TMLE algorithm presented and studied in Section 3. In both of them, the collection $\{\hat{G}_h : h \in \mathcal{H}\}$ is embodied in `R` by the `glmnet` algorithm [3]. The nature of algorithm $\hat{\bar{Q}}$ is left unspecified. As for $\bar{Q}^0_n$, it is obtained once and for all by training $\hat{\bar{Q}}$ on P_n at the beginining of the procedure. More specifically, we never evaluate $\hat{\bar{Q}}(\mathbb{P}_n)$ for $\mathbb{P}_n \neq P_n$.

4.1 LASSO-C-TMLE

We now describe our LASSO-C-TMLE algorithm. Recall that $\mathbb{P}_n$ denotes the empirical measure of a generic subset of the complete data set. The following algorithm implements the theoretical procedure laid out in Sections 3.4 and 3.5.

1. Build a sequence $\{G_{n,h} \equiv \hat{G}_h(P_n) : h \in \mathcal{H}_{100}\}$ by computing a discretized version of the path of the LASSO logistic regression of A on W with a regularization parameter h ranging in the set $\mathcal{H}_{100}$ provided by `cv.glmnet` with options `nlambda=100` (hence $\mathrm{card}(\mathcal{H}_{100}) = 100$) and `nfolds=10`. Set $h_{\min} \equiv \min \mathcal{H}_{100}$ and let $h_{n,\mathrm{CV}}$ be equal to `lambda.min`.

2. Build a sequence $\{\mathbb{G}_{n,h} \equiv \hat{G}_h(\mathbb{P}_n) : h \in \mathcal{H}_{100} \cap [h_{\min}, h_{n,\mathrm{CV}})\}$ by computing a discretized version of the path of the LASSO logistic regression of A on W with a regularization parameter h ranging in $\mathcal{H}_{100} \cap [h_{\min}, h_{n,\mathrm{CV}})$ using `glmnet` with a `lambda` set to $\mathcal{H}_{100} \cap [h_{\min}, h_{n,\mathrm{CV}})$ from step 1.

   Set $k \equiv 1$ and $\mathbb{H}_{n,k-1} \equiv h_{n,\mathrm{CV}}$.

3. For every $h \in \mathcal{H}_{100} \cap [h_{\min}, \mathbb{H}_{n,k-1})$, determine $\bar{\mathbb{Q}}^k_{n,h}$ by fluctuating $\bar{\mathbb{Q}}^{k-1}_n$ based on $\mathbb{G}_{n,h}$ (and $\mathbb{P}_n$) as Section 3.2.

4. Identify the minimizer $\mathbb{H}_{n,k}$ of $h \mapsto \mathbb{P}_n L_2(\bar{\mathbb{Q}}^k_{n,h})$ over $\mathcal{H}_{100} \cap [h_{\min}, \mathbb{H}_{n,k-1})$, define and store $\bar{\mathbb{Q}}^*_{n,h} \equiv \bar{\mathbb{Q}}^k_{n,h}$ for every $h \in \mathcal{H}_{100} \cap [\mathbb{H}_{n,k}, \mathbb{H}_{n,k-1})$, and finally define $\bar{\mathbb{Q}}^k_n \equiv \bar{\mathbb{Q}}^k_{n,h_{n,k}}$.

5. As long as $\mathbb{H}_{n,k} > h_{\min}$, set $k \leftarrow k+1$ and repeat steps 3 and 4 recursively.

The algorithm necessarily converges in a finite number of repetitions of step 4. Let $\mathbb{K}_n$ be the number of repetitions. For every $1 \leq k \leq \mathbb{K}_n$, set $\mathbb{Q}^*_{n,\mathbb{H}_{n,k}} \equiv (\mathbb{Q}_{W,n}, \bar{\mathbb{Q}}^*_{n,\mathbb{H}_{n,k}}) \in \mathcal{Q}$ (its first component is the empirical law of W under $\mathbb{P}_n$) and let $\mathbb{P}^*_{n,\mathbb{H}_{n,k}} \in \mathcal{M}$ be any element of model $\mathcal{M}$ of which the Q-component equals $\mathbb{Q}^*_{n,\mathbb{H}_{n,k}}$. The collection (29) of $\mathbb{K}_n$ collaborative TMLEs and mapping $h \mapsto \mathbb{H}_n(h)$ over $\mathcal{H}_{100}[h_{\min}, h_{n,\mathrm{CV}})$ as in (30) are thus now well defined.

Recall the definition of the cross-validation scheme B_n introduced in Section 3.3. Set

$$\hbar(P_n) \equiv \underset{h \in \mathcal{H}_{100} \cap [h_{\min}, h_{n,\mathrm{CV}})}{\arg\min} E_{B_n}\left[P^1_{n,B_n} L_2(\bar{\mathbb{Q}}_{n,\mathbb{H}_n(h)}\big|_{\mathbb{P}_n = P^0_{n,B_n}})\right]$$

and run once steps 1 to 5 with $\mathbb{P}_n \equiv P_n$, hence the collection

$$\left\{(h_{n,k}, \bar{Q}^*_{n,h_{n,k}}, Q^*_{n,h_{n,k}}, P^*_{n,h_{n,k}}, \Psi(P^*_{n,h_{n,k}})) : 1 \leq k \leq K_n\right\}$$

of collaborative TMLEs. Finally, set

$$\kappa_n \equiv 1 \vee \max\left\{1 \leq k \leq K_n : h_{n,k} \geq \hbar(P_n)\right\}.$$

The collaborative TMLE that we select, our LASSO-C-TMLE estimator, is $\Psi(P^*_{n,h_{n,\kappa_n}})$, as in (31).

4.2 LASSO-PSEUDO-C-TMLE

The LASSO-C-TMLE procedure described in Section 4.1 is quite demanding computationally. It is thus tempting to try and develop an alternative algorithm that would mimick LASSO-C-TMLE but be simpler.

In Section 3.5, we emphasized (see comment before statement of theorem) that one of the keys of LASSO-C-TMLE is to ensure the existence of $h_n \in \mathcal{H}$ and $\bar{Q}^*_{n,h_n}$ such that

$$P_n \partial_{h_n} D^*(\bar{Q}^*_{n,h_n}, G_{n,\cdot}) = 0.$$

If we knew how to compute the derivative $G'_{n,h}(W)$ of $t \mapsto G_{n,t}(W)$ at $t = h$, then this could be easily achieved by enriching the fluctuation of the initial $\bar{Q}^0_n \equiv \hat{\bar{Q}}(P_n)$. Specifically, in light of (19) and (20), given any $h \in \mathcal{H}$, we would define

$$\mathcal{C}^+_h(G_{n,\cdot})(A,W) \equiv \mathcal{C}(G_{n,h})(A,W)\left(1, G'_{n,h}(W)\right), \tag{33}$$

introduce $\bar{Q}^0_{n,h,\varepsilon^+}$ characterized for any $\varepsilon^+ \in \mathbb{R}^2$ by

$$\text{logit}\left(\bar{Q}^0_{n,h,\varepsilon^+}(A,W)\right) \equiv \text{logit}\left(\bar{Q}^0_n(A,W)\right) + \mathcal{C}^+_h(G_{n,\cdot})(A,W)\varepsilon^+$$

and $P^0_{n,h,\varepsilon^+} \in \mathcal{M}$ defined like $P^0_{n,h,\varepsilon}$ except that the conditional expectation of Y given (A,W) under P^0_{n,h,ε^+} equals $\bar{Q}^0_{n,h,\varepsilon^+}$ (and not $\bar{Q}^0_{n,h,\varepsilon}$). Then, the optimal fluctuation would be indexed by the minimizer of the empirical risk

$$\varepsilon^+_{n,h} \equiv \underset{\varepsilon^+ \in \mathbb{R}^2}{\arg\min}\, P_n L_2(\bar{Q}^0_{n,h,\varepsilon^+}).$$

It would result in $\bar{Q}^{*+}_{n,h} \equiv \bar{Q}^0_{n,h,\varepsilon^+}$, $P^{*+}_{n,h} \equiv P^0_{n,h,\varepsilon^+_{n,h}}$ and the TMLE

$$\psi^{*+}_{n,h} \equiv \Psi(P^{*+}_{n,h}) = \frac{1}{n}\sum_{i=1}^n \left(\bar{Q}^{*+}_{n,h}(1,W_i) - \bar{Q}^{*+}_{n,h}(0,W_i)\right)$$

where, by construction, we would have

$$P_n D^*(\bar{Q}^{*+}_{n,h}, G_{n,h}) = P_n \partial_h D^*(\bar{Q}^{*+}_{n,h}, G_{n,h}) = 0.$$

The LASSO-PSEUDO-C-TMLE algorithm that we now describe adapts the above procedure. It unfolds as follows.

1. Build a sequence $\{G_{n,h} \equiv \hat{G}_h(P_n) : h \in \mathcal{H}_{100}\}$ by computing a discretized version of the path of the LASSO logistic regression of A on W with a regularization parameter h ranging in the set $\mathcal{H}_{100}$ provided by `cv.glmnet` with option `nlambda=100` (hence $\text{card}(\mathcal{H}_{100}) = 100$). Let h_n be equal to `lambda.min`. Our estimator of G_0 is G_{n,h_n}.

2. Choose arbitrarily $h^+_n \in \arg\min\{|h - h_n| : h \in \mathcal{H}_{100}, h \neq h_n\}$ and, for every $1 \leq i \leq n$, define

$$G'^+_{n,h_n}(W_i) \equiv \frac{G_{n,h^+_n}(W_i) - G_{n,h_n}(W_i)}{h^+_n - h_n},$$

   a rudimentary numerical approximation of the derivative $G'_{n,h_n}(W_i)$ of $t \mapsto G_{n,t}(W_i)$ at $t = h_n$.

3. Determine $\bar{Q}^{*+}_{n,h_n}$ and P^{*+}_{n,h_n} as described above, with $h = h_n$ and G'^+_{n,h_n} substituted for G'_{n,h_n} in (33).

The LASSO-PSEUDO-C-TMLE estimator is ψ^{*+}_{n,h_n}.

5 Main simulation study

In this section, we present the results of a multi-faceted simulation study of the behaviors and performances of the two instantiations of the collaborative TMLE described in Section 4. Section 5.1 specifies the synthetic data-generating distribution P_0 that we use, Section 5.2 introduces the competing estimators, Section 5.3 outlines the structure of the simulation study, and Section 5.4 gathers its results. Written in `R` [16], our code makes extensive use of the `C-TMLE` package [5].

5.1 Synthetic data-generating distribution

Our synthetic data-generating distribution $P_0 = \Pi_{0,p,\delta}$ depends on two fine-tune parameters: the dimension p of the baseline covariate W and a nonnegative constant $\delta \geq 0$. Sampling $O \equiv (W, A, Y)$ under $\Pi_{0,p,\delta}$ unfolds sequentially along the following steps.

1. Sample $\tilde{W}$ from the centered Gaussian law on $\mathbb{R}^M$, $M = \lceil p/10 \rceil$, of which the covariance matrix Σ is the block-diagonal matrix $(A_{kl})_{1 \leq k,l \leq M}$ where: A_{11} is the 10×10 identity matrix; each A_{kk} for $1 < k \leq M$ is the block-diagonal matrix $(B_{k,st})_{1 \leq s,t \leq 4}$ with

$$B_{k,11} = \begin{pmatrix} 1 & 0 & .25 \\ 0 & 1 & .25 \\ .25 & .25 & 1 \end{pmatrix}, \quad B_{k,22} = B_{k,33} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}, \quad B_{k,44} = \begin{pmatrix} 1 & .5 & 0 \\ .5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

   and $B_{k,st}$ is a zero matrix for $1 \leq s \neq t \leq 4$; each A_{kl} for $1 \leq k \neq l \leq M$ is a zero matrix. If $p = 10$, then $\Sigma = A_{11}$ and we set $W \equiv \tilde{W}$. If $M > 10p$, then we set $W \equiv (\tilde{W}_1, \ldots, \tilde{W}_p)^\top$.

2. Sample A conditionally on W from the Bernoulli law with paramater

$$G_0(W) \equiv \text{expit}\left(\delta + \sum_{k=1}^{p} \beta_k W_k\right),$$

   where $(\beta_1, \cdots, \beta_p) = (1, 1, 3/(p-2), \ldots, 3/(p-2))$.

3. Sample $\tilde{Y}$ conditionally on (A, W) from the Gaussian law with (conditional) variance $1/25$ and expectation

$$f_0(A, W) \equiv \frac{2}{5}(1 + W_1 + W_2 + W_5 + W_6 + W_8 + A),$$

   then define $Y \equiv \text{expit}(\tilde{Y})$.

The covariance matrix Σ induces a loose dependence structure. The components of $\tilde{W}$ can be gathered in $1 + 4 \times (M-1)$ independent groups, one group consisting of $10 + (M-1)$ independent random variables, and the other groups consisting of either two or three mildly dependent random variables (with correlations equal to either 0.25 or 0.5). Neither

$$\bar{Q}_0(A, W) \equiv \int_{[0,1]} \frac{e^{-[\text{logit}(u) - f_0(A,W)]^2/50}}{10\sqrt{\pi}(1-u)} du$$

nor $\Psi(\Pi_{0,p,\delta})$ has a closed form expression. Independently of p and δ, $\Psi(\Pi_{0,p,\delta}) \approx 0.0799$.

5.2 Competing estimators

Let $O_1, \ldots, O_n$ be independent draws from P_0. Recall the characterization of $\hat{G}_h$ ($h \in \mathcal{H}_{100}$) and definition of $h_{n,\mathrm{CV}}$ given in Section 4.1, step 1. Let the algorithm $\hat{\bar{Q}}$ for the estimation of $\bar{Q}_0$ consist in fitting the working model $\{\bar{Q}_\theta : \theta = (\theta_0, \theta_1), \theta_0, \theta_1 \in \mathbb{R}^8\}$ where $\bar{Q}_\theta$ is given by

$$\bar{Q}_\theta(A, W) \equiv \Phi\left((A\theta_1^\top + (1-A)\theta_0^\top))(W_3, \ldots, W_{10})^\top\right)$$

with Φ the distribution function of the standard normal law. Note that the working model is necessarily mis-specified, notably because of the absence of W_1 and W_2 in the above definition. Recall that $\bar{Q}_n^0$ is obtained by training $\hat{\bar{Q}}$ on the whole data set once and for all. To emphasize, $\hat{\bar{Q}}$ is never re-trained during the cross-validation procedure. This is consistent with implementation the original instantiation of the C-TMLE algorithm and of its scalable instantiations.

We compare the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators of $\Psi(\Pi_{0,p,\delta})$ from Section 4 with the following commonly used competitors:

- the unadjusted estimator:

$$\psi_n^{\mathrm{unadj}} \equiv \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1-A_i) Y_i}{\sum_{i=1}^n (1-A_i)};$$

- the so called G-comp estimator [17]:

$$\psi_n^{\text{G-comp}} \equiv \frac{1}{n} \sum_{i=1}^n \left(\bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i)\right);$$

- the so called IPTW estimator [9, 21]:

$$\psi_n^{\mathrm{IPTW}} \equiv \frac{1}{n} \sum_{i=1}^n \frac{(2A_i - 1) Y_i}{G_{n,h_{\mathrm{CV}}}(A_i, W_i)};$$

- the so-called A-IPTW estimator [19]:

$$\psi_n^{\text{A-IPTW}} \equiv \frac{1}{n} \sum_{i=1}^n \frac{(2A_i - 1)}{G_{n,h_{\mathrm{CV}}}(A_i, W_i)} \left(Y_i - \bar{Q}_n^0(W_i, A_i)\right) + \frac{1}{n} \sum_{i=1}^n \left(Q_n^0(1, W_i) - Q_n^0(0, W_i)\right);$$

- and the plain TMLE estimator $\psi^*_{n,h_{n,\mathrm{CV}}}$, see (22).

5.3 Outline of the structure of the simulation study

We consider six different scenarios. In each of them, we repeat independently $B = 200$ times the following steps: for each (n, p, δ) in a collection of scenario-specific triplets,

1. simulate a data set of n independent observations drawn from $\Pi_{0,p,\delta}$;
2. derive the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators of Sections 4.1 and 4.2 as well as the competing estimators presented in Section 5.2;
3. for the double-robust estimators only, *i.e.*, $\psi_n^{\text{A-IPTW}}$, $\psi^*_{n,h_{n,\mathrm{CV}}}$ and our two collaborative TMLEs, construct 95% CIs and check whether or not each of them contains $\Psi(\Pi_{0,p,\delta})$.

Building confidence intervals based on the collaborative TMLEs. By Corollary 4, the asymptotic variances of our collaborative TMLEs both write as

$$\mathrm{Var}_{P_0}\left[\left(D^*(Q_1, G_0) + \Delta(P_1)\right)(O)\right]. \tag{34}$$

Because $\Delta(P_1)$ is difficult to estimate, we estimate (34) with the empirical variance of $D^*(P^*_{n,h_{n,\kappa_n}})$, *i.e.*, with

$$P_n D^*(P^*_{n,h_{n,\kappa_n}})^2 = \frac{1}{n}\sum_{i=1}^{n} D^*(P^*_{n,h_{n,\kappa_n}})(O_i)^2$$

(recall that $P_n D^*(P^*_{n,h_{n,\kappa_n}}) = 0$ by construction). Therefore, the 95% CIs based on our collaborative TMLEs take the form

$$\psi^*_{n,h_{n,\kappa_n}} \pm 1.96\sqrt{P_n D^*(P^*_{n,h_{n,\kappa_n}})^2/n}.$$

We anticipate that the asymptotic variances are over-estimated, resulting in CIs that are too wide. However, we also anticipate that the omitted correction term is of second order relative to main term, or, put in other words, that the difference between (34) and $\mathrm{Var}_{P_0}(D^*(Q_1, G_0)(O))$ is small.

Six scenarios. The three first scenarios investigate what happens when $\delta = 0$ and the number of covariates p increases as a function of sample size n. In scenario 1, $p = 0.2 \times n$ and we increase n. In scenario 2, $p = \lfloor 2.83 \times \sqrt{n} \rfloor$ and we increase n. In scenario 3, $p = \lfloor 7.6 \times \log n \rfloor$ and we increase n. The values of the pairs (p, n) used in these scenarios are presented in Table 0. The constants 0.2, 2.83 and 7.6 are chosen so that $p = 40$ at sample size $n = 200$ in the three scenarios.

Table 0: Values of p and n in scenarios 1, 2 and 3.

n	200	400	600	800	1000	1200	1400	1600	1800	2000
p in scenario 1	40	80	120	160	200	240	280	320	360	400
p in scenario 2	40	56	69	80	89	98	105	113	120	126
p in scenario 3	40	45	48	50	52	53	55	56	56	57

In scenarios 4 and 5, we still set $\delta = 0$ and either keep p fixed and increase n (scenario 4) or keep n fixed and increase p (scenario 5). Finally, in scenario 6, we keep n and p fixed and challenge the positivity assumptions that G_0 is bounded away from 0 and 1 by progressively increasing δ.

In each scenario and for all estimators, we report in a table the average bias (bias, multiplied by 10), standard error (SE, multiplied by 10) and mean squared error (MSE, multiplied by 100) across the $B = 200$ repetitions. Specifically, if $\{\phi_n^{(b)} : 1 \leq b \leq B\}$ are the B realizations of an estimator of $\psi_0 = \Psi(\Pi_{0,p,\delta})$ based on n independent draws from $\Pi_{0,p,\delta}$, then we call $\overline{\phi}_n^{1:B} \equiv B^{-1}\sum_{b=1}^{B}(\phi_n^{(b)} - \psi_0)$ the average bias, $(B^{-1}\sum_{b=1}^{B}(\phi_n^{(b)} - \overline{\phi}_n^{1:B})^2)^{1/2}$ the standard error, and $(\overline{\phi}_n^{1:B})^2 + B^{-1}\sum_{b=1}^{B}(\phi_n^{(b)} - \overline{\phi}_n^{1:B})^2$ the mean squared error.

We also represent in a series of figures how MSE, the empirical coverage of the 95% CIs and their widths evolve as the sample size (scenarios 1 to 4) or number of covariates (scenario 5) or parameter δ (scenario 6) increase. To ease comparisons, all similar figures share the same x- and y-axes.

5.4 Results

Scenarios 1, 2, and 3: increasing n and setting $p = 0.2n, \lfloor 2.83\sqrt{n} \rfloor, \lfloor 7.6 \log n \rfloor$. The results of the three simulation studies under scenarios 1, 2 and 3 are best presented and commented

upon altogether. Figure 1 and Table 1 summarize the numerical findings under scenario 1; Figure 2 and Table 2 summarize the numerical findings under scenario 2; Figure 3 and Table 3 summarize the numerical findings under scenario 3.

Figures 1a, 2a and 3a reveal a general trend: MSE decreases as sample size n increases, despite the fact that the number of covariates p also increases (at different n-rates in each scenario). Overall, LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE perform similarly and better than TMLE; TMLE outperforms IPTW, and IPTW outperforms A-IPTW. Moreover, the gap between LASSO-C-TMLE, LASSO-PSEUDO-C-TMLE on the one hand and TMLE on the other hand *(i)* reduces as sample size n increases (in each scenario), and *(ii)* reduces as p decreases (for each sample size n, across scenarios).

Judging by Tables 1, 2 and 3, the unadjusted, G-comp, IPTW and A-IPTW estimators are strongly biased. The TMLE estimator is strongly biased too, even for large sample size n, when the number of covariates p is not sufficiently small. Note however that the bias of TMLE vanishes at sample size $n = 2000$ in scenario 2 (then, $p = 126$) and at sample size $n \in \{1000, 2000\}$ (then, $p \in \{52, 57\}$). Double-robustness is in action. In contrast, the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators are both essentially unbiased in all configurations.

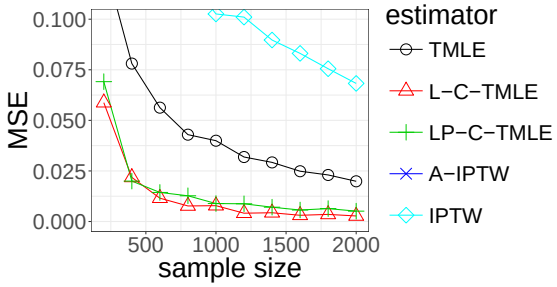
Tables 1, 2 and 3 also reveal that the variance of the TMLE estimator tends to be smaller than those of the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, those last two variances being very similar. Moreover, the gap between them tends to diminish as sample size n increases, in all scenarios.

We now turn to Figures 1b, 2b and 3b. The LASSO-C-TMLE estimator performs best in terms of empirical coverage, followed by the LASSO-PSEUDO-C-TMLE, TMLE and A-IPTW estimators, in that order. At moderate sample size, the superiority of LASSO-C-TMLE-based CIs over the others is striking. However, even they fail to provide the wished coverage except when sample size n is large (say, larger than 1500).

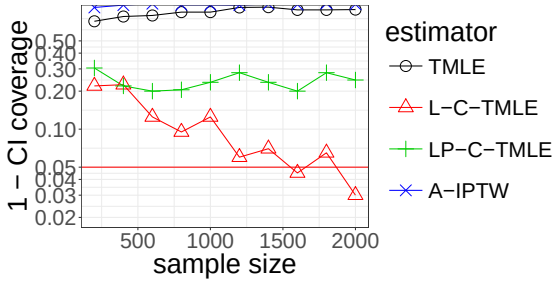
As a side note, we recall that if S is drawn from the Binomial law with parameter $(B, q) = (200, q)$, then $S \leq 185$ with probability approximately 8% for $q = 95\%$, 22% for $q = 94\%$ and 43% for $q = 93\%$. In this light, an empirical coverage of 7.5% is not that abnormal for $B = 200$ independent CIs of exact coverage $q = 93\%$, and even $q = 94\%$. Moreover, we anticipated to get conservative CIs because of how we estimate the asymptotic variance of the LASSO-C-TMLE estimator, see Section 5.3. The fact that the "ratio" entries of Table 3, scenario 3, are that close to one for the LASSO-C-TMLE estimator at sample size $n \in \{1000, 2000\}$ (not to mention at sample size $n = 2000$ in Table 1, scenario 1) reveals that little over-estimation of the asymptotic variance is at play. Finally, we see in Figures 1c, 2c and 3c that the CIs based on the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators are systematically slightly wider and slightly narrower than those based on the TMLE estimator, all much narrower than those based on the A-IPTW estimator.

Scenario 4: keeping p fixed and increasing n. Figure 4 and Table 4 summarize the numerical findings under scenario 4, where the number of covariates p is set to 40 and sample size n goes from 200 to 2000 by steps of 200.

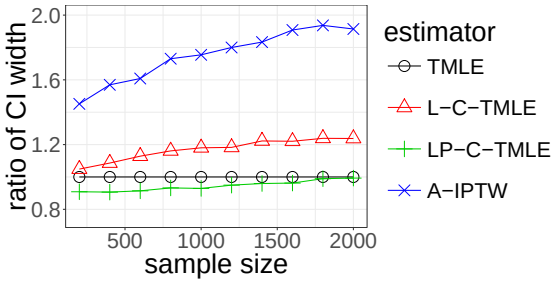
We observe the same trend in Figure 4a as in Figures 1a, 2a and 3a: MSE decreases as sample size n increases; overall, LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE perform similarly and better than TMLE; TMLE outperforms IPTW, and IPTW outperforms A-IPTW. Moreover, the gap between LASSO-C-TMLE, LASSO-PSEUDO-C-TMLE on the one hand and TMLE on the other hand vanishes completely as sample size increases, whereas it only got smaller in scenarios 1, 2, 3.



(a) MSE for five of the seven estimators. The MSE of the A-IPTW estimator is so large that it does not fit in the picture. *MSE is multiplied by 100.*



(b) Coverage of 95% CIs based on the double-robust estimators.

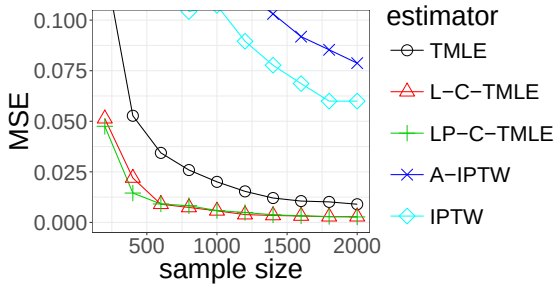


(c) Relative width of 95% CIs based on the double-robust estimators w.r.t. that of the plain TMLE, $\psi^*_{n,h_{n,\mathrm{CV}}}$.

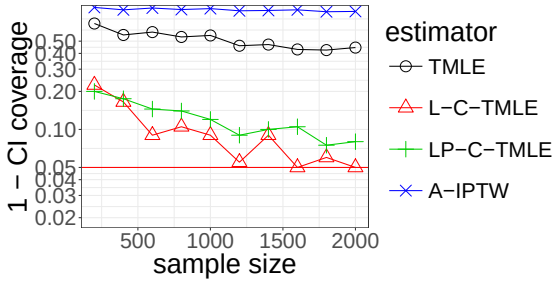
Figure 1: **Scenario 1.** We fix the ratio $p/n = 0.2$, and increase the sample size n from 200 to 2000.

n		ψ_n^{unadj}	$\psi_n^{\text{G-comp}}$	ψ_n^{IPTW}	$\psi_n^{\text{A-IPTW}}$	$\psi^*_{n,h_{n,\text{CV}}}$	L-C-TMLE	LP-C-TMLE
200	bias	1.259	1.212	0.435	0.632	0.327	-0.007	0.039
	SE	0.236	0.152	0.320	0.137	0.151	0.242	0.260
	MSE	1.641	1.491	0.291	0.418	0.130	0.059	0.069
	ratio				1.414	0.882	0.577	0.466
1000	bias	1.171	1.206	0.279	0.407	0.189	0.032	0.026
	SE	0.110	0.066	0.157	0.061	0.064	0.083	0.091
	MSE	1.382	1.459	0.103	0.169	0.040	0.008	0.009
	ratio				1.765	0.969	0.882	0.632
2000	bias	1.175	1.217	0.232	0.339	0.134	0.014	0.020
	SE	0.076	0.047	0.120	0.050	0.045	0.050	0.069
	MSE	1.386	1.483	0.068	0.118	0.020	0.003	0.005
	ratio				1.666	0.959	1.062	0.626

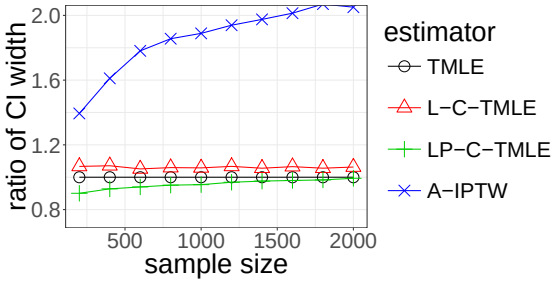
Table 1: **Scenario 1.** The performance of each estimator at sample size $n \in \{200, 1000, 2000\}$, with ratio $p/n = 0.2$. The columns named L-C-TMLE and LP-C-TMLE correspond to the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, respectively. Rows *ratio* report the ratios of the average of the SE estimates across the B repetitions to the empirical SE. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*



(a) MSE for five of the seven estimators.



(b) Coverage of 95% CIs based on the double-robust estimators. *MSE is multiplied by 100.*

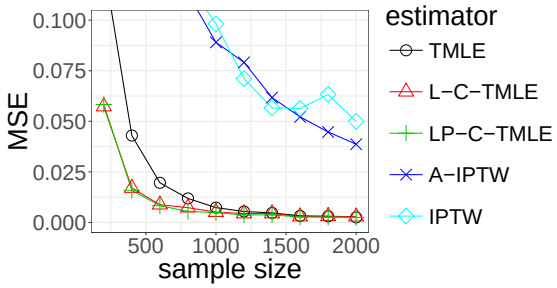


(c) Relative width of 95% CIs based on the double-robust estimators w.r.t. that of the plain TMLE, $\psi^*_{n,h_{n,\mathrm{CV}}}$.

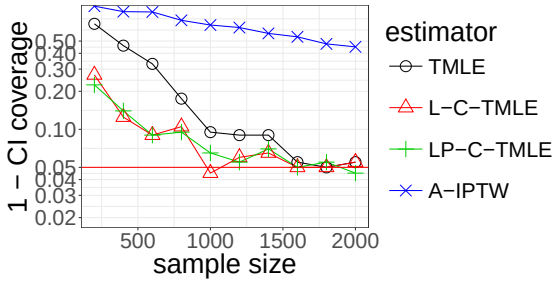
Figure 2: **Scenario 2.** We increase the sample size n from 200 to 2000 and set $p = \lfloor 2.83\sqrt{n} \rfloor$.

n		ψ_n^{unadj}	$\psi_n^{\text{G-comp}}$	ψ_n^{IPTW}	$\psi_n^{\text{A-IPTW}}$	$\psi^*_{n,h_{n,\text{CV}}}$	L-C-TMLE	LP-C-TMLE
200	bias	1.242	1.173	0.469	0.614	0.322	0.014	0.018
	SE	0.221	0.157	0.278	0.135	0.156	0.226	0.217
	MSE	1.592	1.401	0.297	0.395	0.128	0.051	0.048
	ratio				1.377	0.857	0.630	0.553
1000	bias	1.216	1.214	0.271	0.361	0.126	0.003	0.019
	SE	0.104	0.068	0.184	0.077	0.064	0.076	0.074
	MSE	1.489	1.479	0.107	0.136	0.020	0.006	0.006
	ratio				1.505	0.965	0.862	0.790
2000	bias	1.192	1.214	0.201	0.274	0.080	0.007	0.018
	SE	0.075	0.051	0.140	0.061	0.051	0.053	0.049
	MSE	1.426	1.477	0.060	0.079	0.009	0.003	0.003
	ratio				1.488	0.870	0.898	0.904

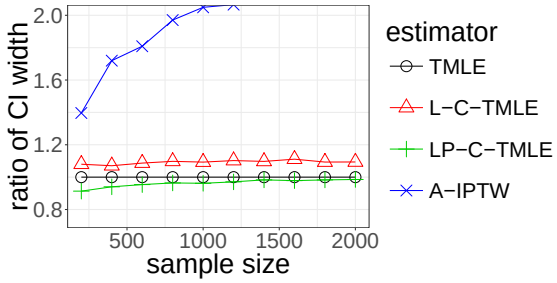
Table 2: **Scenario 2.** The performance of each estimator at sample size $n \in \{200, 1000, 2000\}$, with ratio $p = \lfloor 2.83\sqrt{n} \rfloor$. The columns named L-C-TMLE and LP-C-TMLE correspond to the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, respectively. Rows *ratio* report the ratios of the average of the SE estimates across the B repetitions to the empirical SE. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*



(a) MSE for five of the seven estimators. *MSE is multiplied by 100.*



(b) Coverage of 95% CIs based on the double-robust estimators.



(c) Relative width of 95% CIs based on the double-robust estimators w.r.t. that of the plain TMLE, $\psi^*_{n,h_{n,\mathrm{CV}}}$.

Figure 3: **Scenario 3.** We increase the sample size n from 200 to 2000 and keep $p = \lfloor 7.6 \log(n) \rfloor$.

n		ψ_n^{unadj}	$\psi_n^{\text{G-comp}}$	ψ_n^{IPTW}	$\psi_n^{\text{A-IPTW}}$	$\psi^*_{n,h_{n,\text{CV}}}$	L-C-TMLE	LP-C-TMLE
200	bias	1.244	1.190	0.427	0.634	0.314	-0.009	0.014
	SE	0.228	0.148	0.316	0.136	0.170	0.239	0.241
	MSE	1.600	1.439	0.282	0.420	0.128	0.057	0.058
	ratio				1.363	0.777	0.598	0.502
1000	bias	1.242	1.208	0.224	0.286	0.056	-0.008	-0.006
	SE	0.114	0.077	0.219	0.086	0.065	0.071	0.069
	MSE	1.555	1.465	0.098	0.089	0.007	0.005	0.005
	ratio				1.544	0.995	0.992	0.894
2000	bias	1.227	1.205	0.159	0.185	0.019	-0.006	-0.003
	SE	0.074	0.050	0.157	0.066	0.046	0.053	0.050
	MSE	1.511	1.453	0.050	0.039	0.003	0.003	0.002
	ratio				1.552	1.023	0.978	0.942

Table 3: **Scenario 3.** The performance of each estimator at sample size $n \in \{200, 1000, 2000\}$, with $p = \lfloor 7.6 \log(n) \rfloor$. The columns named L-C-TMLE and LP-C-TMLE correspond to the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, respectively. Rows *ratio* report the ratios of the average of the SE estimates across the B repetitions to the empirical SE. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*

Judging by Table 4, the unadjusted, G-comp, IPTW and A-IPTW estimators are strongly biased whereas the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators are both essentially unbiased even at small sample size $n = 200$. The TMLE estimator is strongly biased too at sample size $n = 200$, but much less so as n increases, with no bias at all at $n = 2000$. Again, double-robustness is in action. Moreover, there is little if any difference between the LASSO-C-TMLE, LASSO-PSEUDO-C-TMLE and TMLE estimators in terms of bias, SE and MSE for sample size $n \in \{1000, 2000\}$.

Figure 4b reveals that, at sample sizes $n \geq 1000$, the empirical coverage of the CIs based on the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators is satisfactory, and that CIs based on the TMLE estimator may provide more coverage than wished. By Table 4 (*ratio* rows), the estimation of the actual variance of the LASSO-C-TMLE and TMLE estimator is quite good at sample size $n \in \{1000, 2000\}$. Apparently, the variance of the LASSO-PSEUDO-C-TMLE estimator is under-estimated at sample size $n = 1000$, but much better estimated at sample size $n = 2000$.

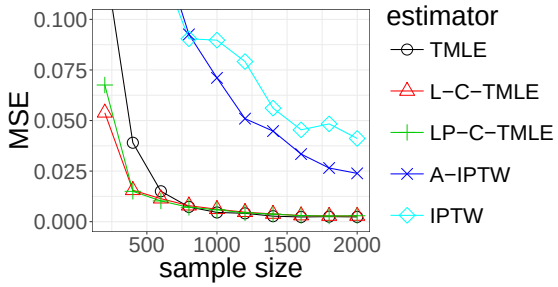
Scenario 5: keeping n fixed and increasing p. Figure 5 and Table 5 summarize the numerical findings under scenario 5, where the sample size n is set to 1000 and the number of covariates p ranges over $\{50, 75, 100, 150, 200\}$.

The take home message of Figure 5a is that, in terms of MSE, the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators outperform the TMLE estimator, which outperforms the IPTW and A-IPTW estimators. Figure 5b further shows that the above message is also valid when considering the empirical coverage of the CIs based on the different estimators. As the number of covariates p increases, all the empirical coverage degrade. However, the CIs based on the LASSO-C-TMLE estimator behave remarkably better than those based on the LASSO-PSEUDO-C-TMLE estimator, which are themselves superior to those based on the TMLE estimator.

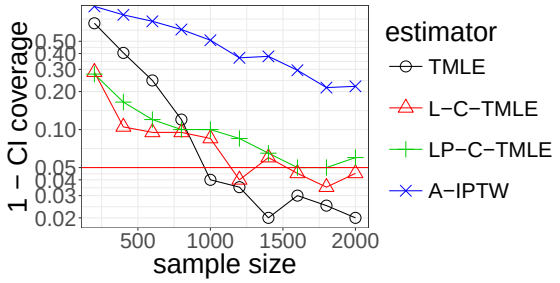
Examining Table 5 helps to better understand the general pattern. The unadjusted, G-comp, IPTW and A-IPTW estimators are too strongly biased to compete. The TMLE estimator performs rather well when the number of covariates p equals 50, like the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators. However, when $p \in \{100, 200\}$, then the TMLE estimator is too biased to compete too – even double-robustness does not help *yet* at the moderate sample size of $n = 1000$. In contrast, the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators are essentially unbiased, and exhibit relatively small variances (compared to all the variances reported in Tables 1, 2, 3, 4. Finally, let us note that the estimation of the variance of the LASSO-C-TMLE estimator is rather good (see the *ratio* rows of Table 5), as opposed to that of the variance of the LASSO-PSEUDO-C-TMLE estimator.

Scenario 6: keeping n and p fixed and challenging the positivity assumption. In this scenario, we study how the level of posivitity violation influences the performance of the estimators, at small sample size $n = 100$ and with $p = 50$ covariates, by progressively increasing $\delta \in \{0.5+k/10 : 0 \leq k \leq 15\}$. Figure 6a illustrates how the positivity violation is challenged. We recover the fact that $\delta \mapsto \Pi_{0,50,\delta}(A = 1|W)$ is increasing. When $\delta = 2$, the law is highly skewed to 1, and the positivity assumption is practically violated. Figures 6b, 6c, 6d and Table 6 summarize the numerical findings under scenario 6.

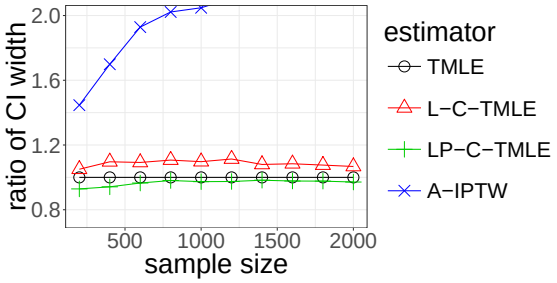
We see in Figure 6 that, overall, the TMLE estimator is much more affected than the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators by the near violation of the positivity assumption at sample size $n = 500$, and that the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators behave similarly in terms of MSE and empirical coverage. Judging by Table 6, The unadjusted, G-comp, IPTW, A-IPTW and TMLE estimators are too strongly biased to compete with



(a) MSE for five of the seven estimators. *MSE is multiplied by 100.*



(b) Coverage of 95% CIs based on the double-robust estimators.

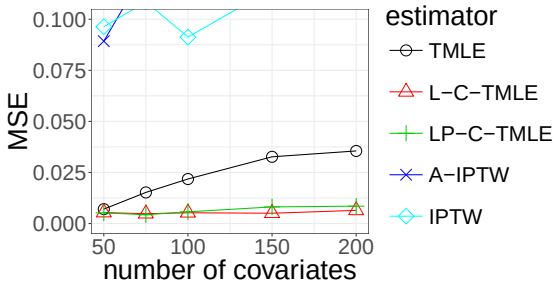


(c) Relative width of 95% CIs based on the double-robust estimators w.r.t. that of the plain TMLE, $\psi^*_{n,h_{n,\mathrm{CV}}}$.

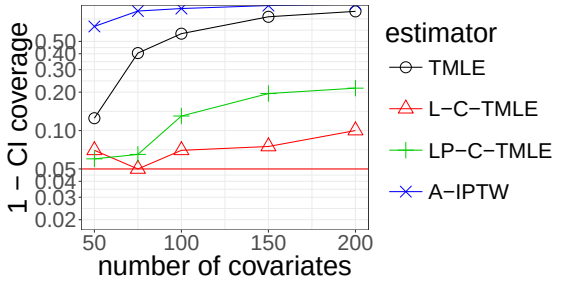
Figure 4: **Scenario 4.** We fix the number of covariates $p = 40$, and increase the sample size n from 200 to 2000.

n		ψ_n^{unadj}	$\psi_n^{\text{G-comp}}$	ψ_n^{IPTW}	$\psi_n^{\text{A-IPTW}}$	$\psi^*_{n,h_{n,\text{CV}}}$	L-C-TMLE	LP-C-TMLE	
200	bias	1.286	1.215	0.485	0.649		0.317	0.020	0.020
	SE	0.226	0.159	0.345	0.147		0.172	0.231	0.259
	MSE	1.705	1.501	0.354	0.443		0.130	0.054	0.068
	ratio				1.283		0.761	0.594	0.469
1000	bias	1.259	1.191	0.213	0.251		0.028	-0.020	-0.021
	SE	0.109	0.073	0.211	0.090		0.062	0.076	0.074
	MSE	1.597	1.425	0.090	0.071		0.005	0.006	0.006
	ratio				1.490		1.062	0.947	0.866
2000	bias	1.260	1.202	0.117	0.140		-0.002	-0.001	-0.002
	SE	0.080	0.049	0.165	0.066		0.046	0.049	0.048
	MSE	1.595	1.448	0.041	0.024		0.002	0.002	0.002
	ratio				1.673		1.062	1.014	0.974

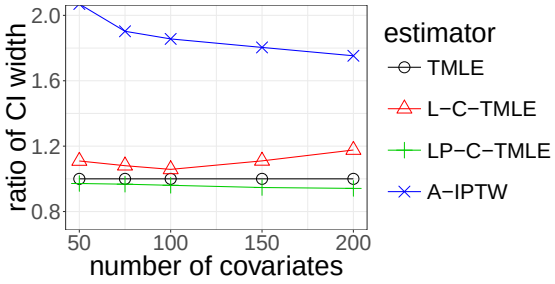
Table 4: **Scenario 4.** The performance of each estimator at sample size $n \in \{200, 1000, 2000\}$, with $p = 40$. The columns named L-C-TMLE and LP-C-TMLE correspond to the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, respectively. Rows *ratio* report the ratios of the average of the SE estimates across the B repetitions to the empirical SE. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*



(a) MSE for five of the seven estimators. *MSE is multiplied by 100.*



(b) Coverage of 95% CIs based on the double-robust estimators.



(c) Relative width of 95% CIs based on the double-robust estimators w.r.t. that of the plain TMLE, $\psi^*_{n,h_{n,\mathrm{CV}}}$.

Figure 5: **Scenario 5.** We fix the sample size $n = 1000$, and increase the number of covariates p from 50 to 200.

p		ψ_n^{unadj}	$\psi_n^{\text{G-comp}}$	ψ_n^{IPTW}	$\psi_n^{\text{A-IPTW}}$	$\psi^*_{n,h_{n,\text{CV}}}$	L-C-TMLE	LP-C-TMLE
50	bias	1.237	1.205	0.228	0.285	0.053	-0.017	-0.008
	SE	0.107	0.072	0.211	0.090	0.065	0.071	0.072
	MSE	1.542	1.457	0.096	0.089	0.007	0.005	0.005
	ratio				1.472	0.983	0.998	0.864
100	bias	1.179	1.199	0.252	0.357	0.130	0.005	0.013
	SE	0.102	0.069	0.167	0.073	0.071	0.072	0.075
	MSE	1.402	1.443	0.091	0.133	0.022	0.005	0.006
	ratio				1.561	0.867	0.896	0.791
200	bias	1.190	1.221	0.297	0.417	0.179	0.024	0.024
	SE	0.107	0.067	0.159	0.067	0.060	0.077	0.089
	MSE	1.428	1.494	0.114	0.178	0.036	0.006	0.009
	ratio				1.591	1.009	0.933	0.645

Table 5: **Scenario 5.** The performance of each estimator at sample size $n = 1000$, with $p \in \{50, 100, 200\}$. The columns named L-C-TMLE and LP-C-TMLE correspond to the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, respectively. Rows *ratio* report the ratios of the average of the SE estimates across the B repetitions to the empirical SE. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*

the nearly unbiased LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators. The rather poor performance in terms of empirical coverage of the CIs based on the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators may be due to the apparent failure in estimating well their variance (see the *ratio* rows of Table 6).

6 Secondary simulation study: LASSO-C-TMLE as a fine-tuning procedure

In this shorter section, we describe a second, less ambitious simulation study. Its aim is to evaluate the interest in using the LASSO-C-TMLE procedure as a fine-tuning procedure. Specifically we wish to investigate, in the same context as in Section 5, how the rivals of the LASSO-C-TMLE estimator that also rely on the estimation of G_0 (*i.e.*, the IPTW, A-IPTW, TMLE and LASSO-PSEUDO-C-TMLE estimators) perform when they are provided with the estimator $G_{n,h_{n,\kappa_n}}$ indexed by the data-adaptive, targeted, fine-tune parameter h_{n,κ_n}.

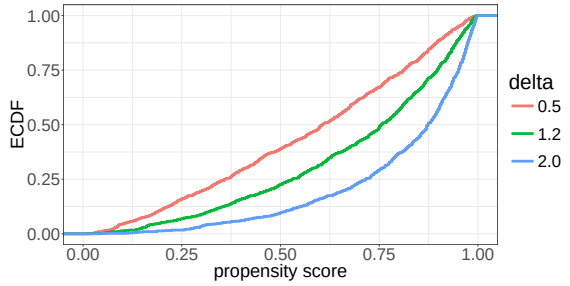
We thus choose to repeat independently $B = 200$ times the following steps: for each number of covariates $p \in \{100, 200\}$,

1. simulate a data set of $n = 1000$ independent observations drawn from $\Pi_{0,p,0}$;
2. derive the LASSO-C-TMLE estimator of Sections 4.1
3. derive the LASSO-PSEUDO-C-TMLE estimator of Section 4.2 as well as the competing IPTW, A-IPTW and TMLE estimators *exactly* as presented in Section 5.2, *and also using* $G_{n,h_{n,\kappa_n}}$ *in place of* G_{n,h_n} *(LASSO-PSEUDO-C-TMLE) and* $G_{n,h_{n,\mathrm{CV}}}$ *(the others).*

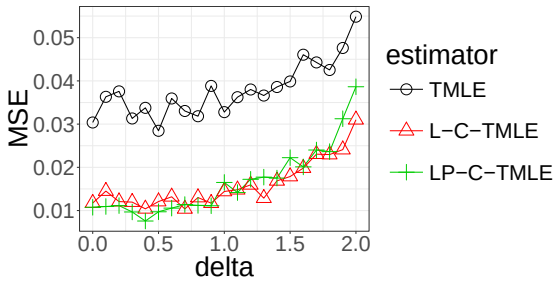
The results are reported in Table 7. A clear pattern emerges from Table 7: the bias is systematically reduced when using $G_{n,h_{n,\kappa_n}}$ in place of $G_{n,h_{n,\mathrm{CV}}}$ or G_{n,h_n}. Nevertheless, the MSE of the IPTW estimator increases, with a two-fold increase when the number of covariates $p = 200$. In contrast, the A-IPTW estimator benefits more from the substitution, with a stark decrease of the MSE on top of that of the bias, the latter being still far too large. This makes only more remarkable the fact that the TMLE estimator greatly benefits from the substitution on all fronts, bias and MSE. On the contrary, the benefit for the LASSO-PSEUDO-C-TMLE estimator is not convincing. In summary, $G_{n,h_{n,\kappa_n}}$ is targeted even "out of context", *i.e.*, even when it is used to build a plain TMLE estimator as opposed to the full-fledged C-TMLE estimator.

7 Discussion

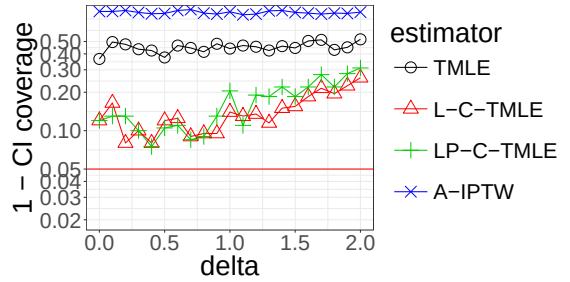
We study the inference of the value of a smooth statistical parameter at a law P_0 from which we sample n independent observations, in situations where *(i)* we rely on a machine learning algorithm fine-tuned by a real-valued parameter h to estimate the G-component G_0 of P_0, possibly consistently, and *(ii)* the product of the rates of convergence of the estimators of the Q- and G-components of P_0 to their targets may be slower than the convenient $o(1/\sqrt{n})$. A plain TMLE with an h chosen by cross-validation would typically not lend itself to the construction of a CI, because the selection of h would trade-off its empirical bias with something akin to the empirical variance of the estimator of G_0 as opposed to that of the TMLE. We develop a collaborative TMLE procedure that succeeds in achieving the relevant trade-off: under high-level empirical processes conditions, and if there exists an oracle h that makes a bulky remainder term asymptotically Gaussian, then the C-TMLE is asymptotically Gaussian hence amenable to building a CI provided that its asymptotic variance can be estimated too.



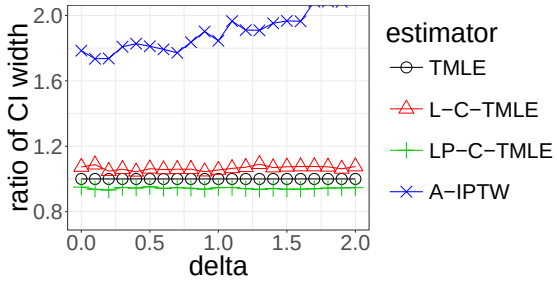
(a) For every $\delta \in \{0.5, 1.2, 2\}$, we simulate $n = 1000$ observations $(W_1, A_1, Y_1), \ldots, (W_n, A_n, Y_n)$ from $\Pi_{0,50,\delta}$, compute $\{\Pi_{0,50,\delta}(A = 1|W = W_i) : 1 \leq i \leq n\}$, and finally plot the corresponding empirical cumulative distribution.



(b) MSE for three of the seven estimators. *MSE is multiplied by 100.*



(c) Coverage of 95% CIs based on the double-robust estimators.



(d) Relative width of 95% CIs based on the double-robust estimators w.r.t. that of the plain TMLE, $\psi^*_{n,h_{n,\mathrm{CV}}}$.

Figure 6: **Scenario 6.** We fix $n = 500, p = 50$, and vary $\delta \in \{0.5 + k/10 : 0 \leq k \leq 15\}$. As the MSEs for IPTW and A-IPTW are too large, we only plot the MSEs of the plain TMLE $\psi^*_{n,h_{n,\mathrm{CV}}}$ and the two collaborative TMLEs to ease comparisons.

δ		ψ_n^{unadj}	$\psi_n^{\text{G-comp}}$	ψ_n^{IPTW}	$\psi_n^{\text{A-IPTW}}$	$\psi^*_{n,h_{n,\mathrm{CV}}}$	L-C-TMLE	LP-C-TMLE
1.0	bias	1.283	1.252	0.995	0.484	0.145	0.006	0.009
	SE	0.170	0.105	0.319	0.117	0.108	0.120	0.128
	MSE	1.675	1.579	1.092	0.248	0.033	0.014	0.017
	ratio				1.385	0.816	0.774	0.650
2.0	bias	1.391	1.340	1.620	0.625	0.185	0.053	0.063
	SE	0.223	0.142	0.409	0.154	0.143	0.168	0.186
	MSE	1.984	1.817	2.791	0.414	0.055	0.031	0.039
	ratio				1.283	0.650	0.597	0.474

Table 6: **Scenario 6.** The performance of each estimator at sample size $n = 500$, with $p = 50$ and $\delta \in \{1, 2\}$. The columns named L-C-TMLE and LP-C-TMLE correspond to the LASSO-C-TMLE and LASSO-PSEUDO-C-TMLE estimators, respectively. Rows *ratio* report the ratios of the average of the SE estimates across the B repetitions to the empirical SE. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*

The construction of the C-TMLE and the main result about its empirical behavior are illustrated with the inference of the average treatment effect, both theoretically and numerically. In the simulation study, the G-component is estimated by the LASSO, and h is the bound on the ℓ^1-norm of the candidate coefficients. Overall, the resulting LASSO-C-TMLE estimator is superior to all its competitors, including a plain TMLE estimator. Evaluated in terms of empirical bias, standard error, mean squared error and coverage of CIs, the superiority is striking in small and moderate sample sizes. It is also strong when the number of covariates increases, or when the positivity assumption is increasingly challenged, thus making the inference task progressively even more delicate.

The simulation study suggests that the CIs based on the C-TMLE do not provide the wished coverage, especially in small sample sizes. Obviously, this may be explained by the need for the C-TMLE estimator to reach its asymptotic regime. More subtly, this may also be related to high-level assumption **A4**, that states the existence of an oracle h making a bulky remainder term asymptotically Gaussian. The assumption may fail to hold in practice. We will devote future research to understanding better **A4** and finding strategies to avoir relying on it.

In conclusion, we believe that the present study further demonstrates the high versatility and potential of the collaborative targeted minimum loss estimation methodology. For (relative) simplicity, we focused on the inference of a smooth, real-valued statistical parameter from independent and identically distributed observations, assuming that the machine learning algorithm is fine-tuned by a real-valued parameter. Our instantiation of the collaborative targeted minimum loss estimation methodology can be extended to other statistical parameters, sampling schemes, and fine-tuning of machine learning algorithms.

A Proofs

Sections A.1, A.2, A.3 and A.4 respectively prove Lemma 2, Theorem 1, Lemma 3 and Corollary 4.

A.1 Proof of Lemma 2

Proof. By (1),

$$\begin{aligned}\Psi(P^*_{n,h_n}) - \Psi(P_0) + P_0 D^*(Q^*_{n,h_n}, G_{n,h_n}) &= \mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h_n}),\\ \Psi(P_0) - \Psi(P_0) + P_0 D^*(Q_1, G_{n,h_n}) &= \mathrm{Rem}_{20}(Q_1, G_{n,h_n}),\end{aligned}$$

hence, by (3) and (8),

$$\begin{aligned}&P_0\big(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q_1, G_{n,h_n})\big)\\ &\qquad = \mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h_n}) - \mathrm{Rem}_{20}(Q_1, G_{n,h_n}) - \big(\Psi(P^*_{n,h_n}) - \Psi(P_0)\big)\\ &\qquad = -(P_n - P_0)D^*(P_0) + o_P(1/\sqrt{n}).\end{aligned}$$

But (5) and $\Phi_0(G_0) = P_0 D^*(Q_1, G_0) = 0$ also imply that

$$\begin{aligned}-\Phi_0(G_{n,h_n}) + \Phi_0(G_0) &= P_0\left(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q_1, G_{n,h_n})\right)\\ &\quad + (P_n - P_0)\left(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q_1, G_0)\right)\\ &\quad + (P_n - P_0)D^*(Q_1, G_0) + o_P(1/\sqrt{n})\end{aligned}$$

which, combined with the previous display and (7), yield

$$\Phi_0(G_{n,h_n}) - \Phi_0(G_0) = (P_n - P_0)\left(D^*(P_0) - D^*(Q_1, G_0)\right) + o_P(1/\sqrt{n}) = o_P(1/\sqrt{n})$$

showing that (9) is satisfied with $\tilde{h}_n = h_n$ and $\Delta(P_1) \equiv 0$. □

A.2 Proof of Theorem 1

Proof. The proof unfolds in two parts.

Step one: extracting the would-be first order term. Equality (5) in **A2** rewrites as

$$\begin{aligned} o_P(1/\sqrt{n}) &= P_n D^*(Q^*_{n,h_n}, G_{n,h_n}) \\ &= (P_n - P_0)D^*(Q^*_{n,h_n}, G_{n,h_n}) + P_0 D^*(Q^*_{n,h_n}, G_{n,h_n}) \\ &= \left[(P_n - P_0)D^*(Q_1, G_0) + (P_n - P_0)\left(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q_1, G_0)\right)\right] \\ &\quad + \left[P_0 D^*(Q^*_{n,h_n}, G_0) + P_0\left(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q^*_{n,h_n}, G_0)\right)\right]. \end{aligned}$$

The second term of the sum between the first pair of brackets is $o_P(1/\sqrt{n})$ by (7) in **A3**. As for the first term between the second pair of brackets, (1) and (1) entail that it satisfies

$$P_0 D^*(Q^*_{n,h_n}, G_0) = \Psi(P_0) - \Psi(P^*_{n,h_n}) + \mathrm{Rem}_{20}(Q^*_{n,h_n}, G_0) = \Psi(P_0) - \Psi(P^*_{n,h_n}),$$

where we also use the fact that $\Psi(P^*_{n,h_n})$ depends on P^*_{n,h_n} only through Q^*_{n,h_n}. Thus, it holds that

$$\begin{aligned} \Psi(P^*_{n,h_n}) - \Psi(P_0) - (P_n - P_0)D^*(Q_1, G_0) + o_P(1/\sqrt{n}) \\ = P_0\left(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q^*_{n,h_n}, G_0)\right) \equiv T_{1,n}. \end{aligned} \quad (35)$$

Let us now study $T_{1,n}$, the right-hand side expression in (35). It rewrites as

$$\begin{aligned} T_{1,n} = P_0\left(D^*(Q_1, G_{n,h_n}) - D^*(Q_1, G_0)\right) \\ + P_0\left(D^*(Q^*_{n,h_n}, G_{n,h_n}) - D^*(Q^*_{n,h_n}, G_0)\right) - P_0\left(D^*(Q_1, G_{n,h_n}) - D^*(Q_1, G_0)\right). \end{aligned}$$

Consider the three terms in the right-hand side of the above equation. Combining (1), (1) and the fact that $\Psi(P) = \Psi(P')$ whenever P and P' have the same Q-component reveals that the first and third terms equal both $\Phi_0(G_{n,h_n}) - \Phi_0(G_0)$ and $\mathrm{Rem}_{20}(Q_1, G_{n,h_n})$. For similar reasons, the second term equals $\mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h_n})$. Therefore, by using successively (8) in **A3** then (9) from **A4**, we obtain that

$$\begin{aligned} T_{1,n} &= \Phi_0(G_{n,h_n}) - \Phi_0(G_0) + \left(\mathrm{Rem}_{20}(Q^*_{n,h_n}, G_{n,h_n}) - \mathrm{Rem}_{20}(Q_1, G_{n,h_n})\right) \\ &= \left[\Phi_0(G_{n,\tilde{h}_n}) - \Phi_0(G_0)\right] + \left[\Phi_0(G_{n,h_n}) - \Phi_0(G_{n,\tilde{h}_n})\right] + o_P(1/\sqrt{n}) \\ &= (P_n - P_0)\Delta(P_1) + \left[\Phi_0(G_{n,h_n}) - \Phi_0(G_{n,\tilde{h}_n})\right] + o_P(1/\sqrt{n}). \end{aligned} \quad (36)$$

Step two: showing that the would-be first order term is complete. The rest of the proof consists in showing that the term between brackets in (36), say $T_{2,n}$, is $o_P(1/\sqrt{n})$. The inequality below follows from the definition of Φ_0 and the triangle inequality, and the equality from (10) in **A5**:

$$\begin{aligned} |T_{2,n}| &\leq \left|P_n\left(D^*(Q_1, G_{n,\tilde{h}_n}) - D^*(Q_1, G_{n,h_n})\right)\right| \\ &\quad + \left|(P_n - P_0)\left(D^*(Q_1, G_{n,h_n}) - D^*(Q_1, G_{n,\tilde{h}_n})\right)\right| \\ &= \left|P_n\left(D^*(Q_1, G_{n,\tilde{h}_n}) - D^*(Q_1, G_{n,h_n})\right)\right| + o_P(1/\sqrt{n}). \end{aligned}$$

Therefore, it suffices to prove that the absolute value in the above right-hand side expression is $o_P(1/\sqrt{n})$. Under $\mathbf{A1}(Q_1, h_n, c_5)$ (guaranteed by **A5**), for every $1 \leq i \leq n$, the Taylor-Lagrange inequality yields

$$\left|\left(D^*(Q_1, G_{n,\tilde{h}_n}) - D^*(Q_1, G_{n,h_n}) - (\tilde{h}_n - h_n) \times \partial_{h_n} D^*(Q_1, G_{n,.})\right)(O_i)\right| \lesssim (\tilde{h}_n - h_n)^2$$

hence, by convexity,

$$\left|P_n\left(D^*(Q_1, G_{n,\tilde{h}_n}) - D^*(Q_1, G_{n,h_n}) - (\tilde{h}_n - h_n) \times \partial_{h_n} D^*(Q_1, G_{n,.})\right)\right| \lesssim (\tilde{h}_n - h_n)^2.$$

Since $P_n \partial_{h_n} D^*(Q^*_{n,h_n}, G_{n,.})$ and $(\tilde{h}_n - h_n)$ are both $o_P(1/n^{1/4})$ by (6) in **A2** and **A5**, we get

$$\begin{aligned}
& P_n\left(D^*(Q_1, G_{n,\tilde{h}_n}) - D^*(Q_1, G_{n,h_n})\right) \\
= \; & (\tilde{h}_n - h_n) \times P_n\left(\partial_{h_n} D^*(Q^*_{n,h_n}, G_{n,.}) - \partial_{h_n} D^*(Q_1, G_{n,.})\right) + o_P(1/\sqrt{n}) \\
= \; & (\tilde{h}_n - h_n) \times (P_n - P_0)\left(\partial_{h_n} D^*(Q^*_{n,h_n}, G_{n,.}) - \partial_{h_n} D^*(Q_1, G_{n,.})\right) \\
& + (\tilde{h}_n - h_n) \times P_0\left(\partial_{h_n} D^*(Q^*_{n,h_n}, G_{n,.}) - \partial_{h_n} D^*(Q_1, G_{n,.})\right) + o_P(1/\sqrt{n}).
\end{aligned}$$

In light of (11) and (12) in **A5**, the righ-hand side expression is $o_P(1/\sqrt{n})$. This completes the proof: $T_{2,n} = o_P(1/\sqrt{n})$, hence (36) rewrites as

$$T_{1,n} = (P_n - P_0)\Delta(P_1) + o_P(1/\sqrt{n}),$$

and (13) finally follows from (35). □

A.3 Proof of Lemma 3

Proof. Set $\bar{q}_n \equiv \bar{Q}_n(1,\cdot) - \bar{Q}_n(0,\cdot)$, $\bar{q}_1 \equiv \bar{Q}_1(1,\cdot) - \bar{Q}_1(0,\cdot)$, and $\psi_1 \equiv P_0\bar{q}_1$. Using inequality $(a+b)^2 \leq 2(a^2+b^2)$ (valid for all real numbers a, b), we first remark that

$$\begin{aligned}
P_0(\bar{q}_n - \bar{q}_1)^2 &\leq 2P_0[\bar{Q}_n(1,\cdot) - \bar{Q}_1(1,\cdot)]^2 + 2P_0[\bar{Q}_n(0,\cdot) - \bar{Q}_1(0,\cdot)]^2 \\
&= 2P_0(\bar{Q}_n - \bar{Q}_0)^2/\ell G_0 \lesssim P_0(\bar{Q}_n - \bar{Q}_0)^2.
\end{aligned}$$

Therefore, it also holds that $P_0(\bar{q}_n - \bar{q}_1)^2 = o_P(1)$. Second, we decompose the difference $\psi_n - \psi_1$ as

$$\psi_n - \psi_1 = P_n\bar{q}_n - P_0\bar{q}_1 = (P_n - P_0)(\bar{q}_n - \bar{q}_1) + (P_n - P_0)\bar{q}_1 + P_0(\bar{q}_n - \bar{q}_1).$$

Lemma 19.24 in [32] guarantees that the first term in the above RHS expression is $o_P(1/\sqrt{n})$. Because $\bar{q}_1$ is uniformly bounded, the standard central limit theorem (for sequences of independent and identically distributed, real-valued random variables with finite variance) implies that the second term is $O_P(1/\sqrt{n})$. Finally, the third term is $o_P(1)$ by the Cauchy-Schwarz inequality and the remark we previously made. In summary, $\psi_n - \psi_1 = o_P(1)$, as stated. □

A.4 Proof of Corollary 4

The proof of Corollary 4 uses repeatedly the fact that some specific random functions fall in P_0-Donsker classes with P_0-probability tending to one. Specifically, the proof will refer several times to the following lemma (its proof is deferred to the end of this section).

Lemma 5. *Suppose that the assumptions of Corollary 4 are met. Then, with P_0-probability tending to one, $\bar{Q}^*_{n,h_{n,\kappa_n}}(1,\cdot) - \bar{Q}^*_{n,h_{n,\kappa_n}}(0,\cdot)$, $D^*(P^*_{n,h_{n,\kappa_n}})$, $D^*(Q_1, G_{n,h_{n,\kappa_n}}) - D^*(Q_1, G_{n,\tilde{h}_n})$ and $\partial_{h_{n,\kappa_n}} D^*(Q^*_{n,h_{n,\kappa_n}}, G_{n,.}) - \partial_{h_{n,\kappa_n}} D^*(Q^*_{n,h_{n,\kappa_n}}, G_{n,.})$ also fall in P_0-Donsker classes.*

We can now present the proof of Corollary 4.

Proof of Corollary 4. There is no obvious counterpart in **C1**, **C2** and **C4** to (7) and (8) appearing within **A3**. Yet, under **C2** and **C4**, $G_{n,h_{n,\kappa_n}}$ consistently estimates G_0 and $\bar{Q}^*_{n,h_{n,\kappa_n}}$ converges to a limit $\bar{Q}_1$ that may differ from $\bar{Q}_0$. Moreover, since $G_{n,h_{n,\kappa_n}}$ and G_0 are bounded away from zero by **C1**, we have

$$P_0(D^*(P^*_{n,h_{n,\kappa_n}}) - D^*(P_1))^2 = o_P(1), \tag{37}$$

where $P_1 \in \mathcal{M}$ is any element of model $\mathcal{M}$ of which the Q- and G-components equal $Q_1 \equiv (Q_{W,0}, \bar{Q}_1)$ and G_0 (see proof below). By Lemma 5, $D^*(P^*_{n,h_{n,\kappa_n}})$ falls in a P_0-Donsker class with P_0-probability tending to one. It thus holds that

$$(P_n - P_0)(D^*(P^*_{n,h_{n,\kappa_n}}) - D^*(P_1)) = o_P(1/\sqrt{n}),$$

as requested in (7) of **A3**. In addition, the following convergence also occurs,

$$\mathrm{Rem}_{20}(\bar{Q}^*_{n,h_{n,\kappa_n}}, G_{n,h_{n,\kappa_n}}) - \mathrm{Rem}_{20}(\bar{Q}_1, G_{n,h_{n,\kappa_n}}) = o_P(1/\sqrt{n}), \tag{38}$$

as requested in (8) of **A3** (see proof below). Consequently, **C1**, **C2** and **C4** imply **A3**.

Let us now turn to assumption **A5**. To alleviate notation, let $G''_{n,h}(W)$ be the second order derivative of $t \mapsto G_{n,t}(W)$ at $h \in \mathcal{T}$ under **C1**. Given the definition of $D^*(Q_1, G_{n,t})(O)$, see (15), $t \mapsto D^*(Q_1, G_{n,t})(O)$ is twice differentiable on $\mathcal{T}$ and, for each $h \in \mathcal{T}$,

$$\partial_h^2 D^*(Q_1, G_{n,.})(O) = (Y - \bar{Q}_1(A,W)) \times \left(\frac{G''_{n,h}(W)}{\ell G_{n,h}(A,W)^2} - 2(2A-1)\frac{G'_{n,h}(W)^2}{\ell G_{n,h}(A,W)^3} \right).$$

Obviously, under **C1**, there exists a universal constant $C_3 > 0$ such that the supremum in $h \in \mathcal{T}$ of $\partial_h^2 D^*(Q_1, G_{n,.})(O)$ is P_0-almost surely smaller than C_3. Consequently, assumption $\mathbf{A1}(\bar{Q}_1, h_{n,\kappa_n}, C_3)$ is met. In addition, we show below that (10), (11) and (12) hold true whenever **C1** to **C4** are met.

In summary, **A2** is satisfied by construction of $\psi^*_{n,h_{n,\kappa_n}}$, see (28); **A4** is assumed to hold true; **A3** and **A5** are met. Thus, Theorem 1 applies and implies the result stated in Corollary 4. This completes the proof. □

Proof of (37). Suppose that the assumptions of Corollary 4 are met and recall decomposition (15). To alleviate notation, introduce $\bar{Q}_n \equiv \bar{Q}^*_{n,h_{n,\kappa_n}}$, $\bar{q}_n \equiv \bar{Q}^*_{n,h_{n,\kappa_n}}(1,\cdot) - \bar{Q}^*_{n,h_{n,\kappa_n}}(0,\cdot)$ and $G_n \equiv G_{n,h_{n,\kappa_n}}$. By Lemma 5, $\bar{q}_n$ falls in a P_0-Donsker class with P_0-probability tending to one. Using repeatedly inequality $(a+b)^2 \leq 2(a^2+b^2)$, we obtain

$$\begin{aligned}
P_0(D^*(P^*_{n,h_{n,\kappa_n}}) - D^*(P_1))^2 &\lesssim P_0(\ell G_n - \ell G_0)^2 + P_0(\bar{Q}_n \ell G_0 - \bar{Q}_1 \ell G_n)^2 \\
&\quad + P_0(\bar{q}_n - \bar{q}_1)^2 + (\psi^*_{n,h_{n,\kappa_n}} - \psi_0)^2 \\
&\lesssim P_0(G_n - G_0)^2 + P_0(\bar{Q}_n - \bar{Q}_1)^2 \\
&\quad + P_0(\bar{q}_n - \bar{q}_1)^2 + (\psi^*_{n,h_{n,\kappa_n}} - \psi_0)^2 \\
&= P_0(\bar{q}_n - \bar{q}_1)^2 + (\psi^*_{n,h_{n,\kappa_n}} - \psi_0)^2 + o_P(1).
\end{aligned}$$

The assumptions of Lemma 3 are met too. Therefore, we can retrieve the bound

$$P_0(\bar{q}_n - \bar{q}_1)^2 \lesssim P_0(\bar{Q}_n - \bar{Q}_1)^2 = o_P(1)$$

from its proof and assert that its conclusion holds: $(\psi^*_{n,h_{n,\kappa_n}} - \psi_0) = o_P(1)$. This completes the proof of (37). □

Proof of (38). Suppose that the assumptions of Corollary 4 are met. In view of (16), we have

$$\begin{aligned} T_n &\equiv |\mathrm{Rem}_{20}(\bar{Q}^*_{n,h_{n,\kappa_n}}, G_{n,h_{n,\kappa_n}}) - \mathrm{Rem}_{20}(\bar{Q}_1, G_{n,h_{n,\kappa_n}})| \\ &= \left| E_{P_0}\left[(2A-1)\left(1 - \frac{\ell G_0(A,W)}{\ell G_{n,h_{n,\kappa_n}}(A,W)}\right)(\bar{Q}^*_{n,h_{n,\kappa_n}}(A,W) - \bar{Q}_1(A,W))\right]\right|. \end{aligned}$$

Therefore, the Cauchy-Schwarz inequality and equality $(\ell G_{n,h_{n,\kappa_n}} - \ell G_0)^2 = (G_{n,h_{n,\kappa_n}} - G_0)^2$ yield

$$T_n^2 \lesssim P_0(G_{n,h_{n,\kappa_n}} - G_0)^2 \times P_0(\bar{Q}^*_{n,h_{n,\kappa_n}} - \bar{Q}_1)^2 = o_P(1/n),$$

which completes the proof of (38). □

Proof of (10) *in the context of Section 3.* Suppose that the assumptions of Corollary 4 are met. Using **C1** and the Taylor-Lagrange inequality yields

$$|G_{n,\tilde{h}_n}(W) - G_{n,h_{n,\kappa_n}}(W)| \lesssim |\tilde{h}_n - h_{n,\kappa_n}|$$

hence $P_0(G_{n,\tilde{h}_n} - G_{n,h_{n,\kappa_n}})^2 \lesssim (\tilde{h}_n - h_{n,\kappa_n})^2 = o_P(1)$ (with much to spare). Now, observe that

$$\begin{aligned} \left(D^*(Q_1, G_{n,h_{n,\kappa_n}}) - D^*(Q_1, G_{n,\tilde{h}_n})\right)(O) &= (Y - \bar{Q}_1(A,W))(2A-1) \\ &\quad \times \left(\frac{1}{\ell G_{n,h_{n,\kappa_n}}(A,W)} - \frac{1}{\ell G_{n,\tilde{h}_n}(A,W)}\right), \end{aligned}$$

which evidently implies the upper bound

$$\begin{aligned} \left|\left(D^*(Q_1, G_{n,h_{n,\kappa_n}}) - D^*(Q_1, G_{n,\tilde{h}_n})\right)(O)\right| &\lesssim |\ell G_{n,h_{n,\kappa_n}}(A,W) - \ell G_{n,\tilde{h}_n}(A,W)| \\ &= |G_{n,h_{n,\kappa_n}}(W) - G_{n,\tilde{h}_n}(W)|. \end{aligned}$$

Therefore, $P_0(D^*(Q_1, G_{n,h_{n,\kappa_n}}) - D^*(Q_1, G_{n,\tilde{h}_n}))^2 = o_P(1)$. Furthermore, Lemma 5 guarantees that $D^*(Q_1, G_{n,h_{n,\kappa_n}}) - D^*(Q_1, G_{n,\tilde{h}_n})$ falls in a P_0-Donsker class with P_0-probability tending to one. The same argument as the one that lead to (3) in Section 3.3 thus completes the proof of (10). □

Proof of (11) *in the context of Section 3.* Suppose that the assumptions of Corollary 4 are met. In view of (15), we have

$$\partial_{h_{n,\kappa_n}} D^*(Q^*_{n,h_{n,\kappa_n}}, G_{n,\cdot})(O) = \frac{2A-1}{\ell G_{n,h_{n,\kappa_n}}(A,W)} G'_{n,h_{n,\kappa_n}}(W)(Y - \bar{Q}^*_{n,h_{n,\kappa_n}}(A,W)), \quad (39)$$

$$\partial_{h_{n,\kappa_n}} D^*(Q_1, G_{n,\cdot})(O) = \frac{2A-1}{\ell G_{n,h_{n,\kappa_n}}(A,W)} G'_{n,h_{n,\kappa_n}}(W)(Y - \bar{Q}_1(A,W)), \quad (40)$$

hence

$$\left|\left(\partial_{h_{n,\kappa_n}} D^*(Q_{n,h_{n,\kappa_n}}, G_{n,\cdot}) - \partial_{h_{n,\kappa_n}} D^*(Q_1, G_{n,\cdot})\right)\right| \lesssim |\bar{Q}_1 - \bar{Q}^*_{n,h_{n,\kappa_n}}|. \quad (41)$$

Therefore, the Cauchy-Schwarz inequality implies the bound

$$\begin{aligned} &\left((h_{n,\kappa_n} - \tilde{h}_n) \times P_0\left(\partial_{h_{n,\kappa_n}} D^*(Q_{n,h_{n,\kappa_n}}, G_{n,\cdot}) - \partial_{h_{n,\kappa_n}} D^*(Q_1, G_{n,\cdot})\right)\right)^2 \\ &\qquad \lesssim (h_{n,\kappa_n} - \tilde{h}_n)^2 \times P_0(\bar{Q}^*_{n,h_{n,\kappa_n}} - \bar{Q}_1)^2 = o_P(1/n), \end{aligned}$$

thus completing the proof of (11). □

Proof of (12) *in the context of Section 3.* Suppose that the assumptions of Corollary 4 are met. By Lemma 5, $\partial_{h_{n,\kappa_n}} D^*(Q^*_{n,h_{n,\kappa_n}}, G_{n,\cdot}) - \partial_{h_{n,\kappa_n}} D^*(Q_1, G_{n,\cdot})$ falls in a P_0-Donsker class with P_0-probability tending to one. In view of (41), it holds that

$$P_0 \left(\partial_{h_{n,\kappa_n}} D^*(Q_{n,h_{n,\kappa_n}}, G_{n,\cdot}) - \partial_{h_{n,\kappa_n}} D^*(Q_1, G_{n,\cdot})\right)^2 = o_P(1).$$

The same argument as the one that lead to (3) in Section 3.3 thus completes the proof of (12). □

Proof of Lemma 5. We proceed by order of appearance in the statement of the lemma. First, note that

$$\bar{Q}^*_{n,h_{n,\kappa_n}}(1, W) - \bar{Q}^*_{n,h_{n,\kappa_n}}(0, W) = (2A - 1)\bar{Q}^*_{n,h_{n,\kappa_n}}(A, W).$$

Second, derive from (15) the explicit forms of $D^*(P^*_{n,h_{n,\kappa_n}})$, $D^*(Q_1, G_{n,h_{n,\kappa_n}})$, $D^*(Q_1, G_{n,\tilde{h}_n})$, and of the difference of the two last ones. Third, recall the explicit forms of $\partial_{h_{n,\kappa_n}} D^*(Q^*_{n,h_{n,\kappa_n}}, G_{n,\cdot})$ and $D^*(Q_1, G_{n,h_{n,\kappa_n}})$ given in (39) and (40), and derive from them that of their difference. Thanks to **C4** and the above explicit forms, straightforward applications of [33, Theorem 2.10.6] yield the result. □

We conclude this article on a final remark about $\mathbf{B4}(P_n, k)$. Suppose that $\mathbf{B2}(P_n, k)$ is met. If, in light of **C1**, we also assume that $t \mapsto G_{n,t}(W)$ is twice differentiable in a neighborhood of $h_{n,k}$, then the assumptions of the implicit function theorem are satisfied and $h \mapsto \varepsilon_{n,h,k}$ is differentiable around $h_{n,k}$.

References

[1] W. G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313, 1968.

[2] J. M. Franklin, W. Eddings, R. J. Glynn, and S. Schneeweiss. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *American journal of epidemiology*, 187(7):651–659, 2015.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

[4] S. Gruber and M. J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6(1): Article 18, 2010.

[5] S. Gruber and M. J. van der Laan. ctmle: an r package for collaborative targeted maximum likelihood estimation. Technical report, University of California, Berkeley, 2010.

[6] S. Gruber, R. W. Logan, I. Jarrín, S. Monge, and M. A. Hernán. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in medicine*, 34(1):106–117, 2015.

[7] D. Ho, K. Imai, G. King, and E. Stuart. Matchit: Matchit: Nonparametric preprocessing for parametric casual inference. *R package version*, pages 2–2, 2006.

[8] D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

[9] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[10] C. Ju, M. Combs, S. D. Lendle, J. M. Franklin, R. Wyss, S. Schneeweiss, and M. J. van der Laan. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *arXiv preprint arXiv:1703.02236*, 2017.

[11] C. Ju, S. Gruber, S. D. Lendle, A. Chambaz, J. M. Franklin, R. Wyss, S. Schneeweiss, and M. J. van der Laan. Scalable collaborative targeted learning for high-dimensional data. *arXiv preprint arXiv:1703.02237*, 2017.

[12] C. Ju, J. Schwab, and M. J. van der Laan. On adaptive propensity score truncation in causal inference. *arXiv preprint arXiv:1707.05861*, 2017.

[13] C. Ju, R. Wyss, J. M. Franklin, S. Schneeweiss, J. Häggström, and M. J. van der Laan. Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Statistical methods in medical research*, 2017.

[14] T. Kurth, A. M. Walker, R. J. Glynn, K. A. Chan, J. M. Gaziano, K. Berger, and J. M. Robins. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163(3):262–270, 2006.

[15] B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[17] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9-12):1393–1512, 1986.

[18] J. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association: Section on Bayesian Statistical Science*, pages 6–10, 2000.

[19] J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[20] J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica*, 11(4):920–936, 2001.

[21] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[22] J. M. Robins, A. Rotnitzky, and M. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

[23] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55, 1983.

[24] P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

[25] S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4):512, 2009.

[26] M. E. Schnitzer and M. Cefalu. Collaborative targeted learning using regression shrinkage. *Statistics in Medicine*, 2017.

[27] O. M. Stitelman and M. J. van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1):Article 21, 2010.

[28] M. van der Laan, A. Chambaz, and C. Ju. *Targeted Learning in Data Science. Causal Inference for Complex Longitudinal Studies*, chapter CTMLE for continuous tuning. Springer Series in Statistics. Springer, 2018.

[29] M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Science & Business Media, 2011.

[30] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[31] M. J. van der Laan, S. Gruber, et al. Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1):Article 17, 2010.

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

[33] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[34] H. Wang, S. Rose, and M. J. van der Laan. Finding quantitative trait loci genes with collaborative targeted maximum likelihood learning. *Statistics & Probability Letters*, 81(7):792–796, 2011.

p		ψ_n^{IPTW}	$(\psi_n^{\text{IPTW}})'$	$\psi_n^{\text{A-IPTW}}$	$(\psi_n^{\text{A-IPTW}})'$	
100	bias	0.252	0.078	0.357	0.151	
	SE	0.167	0.325	0.073	0.167	
	MSE	0.091	0.112	0.133	0.050	
200	bias	0.297	0.106	0.417	0.152	
	SE	0.159	0.480	0.067	0.150	
	MSE	0.114	0.242	0.178	0.045	
p		$\psi^*_{n,h_{n,\text{CV}}}$	$\psi^*_{n,h_{n,\kappa_n}}$	L-C-TMLE	LP-C-TMLE	LP-C-TMLE′
100	bias	0.130	0.017	0.005	0.013	-0.000
	SE	0.071	0.101	0.072	0.075	0.094
	MSE	0.022	0.011	0.005	0.006	0.009
200	bias	0.179	-0.037	0.024	0.024	-0.072
	SE	0.060	0.140	0.077	0.089	0.148
	MSE	0.036	0.021	0.006	0.009	0.027

Table 7: **Using LASSO-C-TMLE as a fine-tuning procedure.** The performance of each estimator at sample size $n = 1000$ with $p \in \{100, 200\}$. The prime symbol indicates the use of $G_{n,h_{n,\kappa_n}}$ as an estimator of G_0 in place of $G_{n,h_{n,\text{CV}}}$ or G_{n,h_n}. *Bias and SE are multiplied by 10, and MSE is multiplied by 100.*