University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 273

Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation

Wenjing Zheng* Mark J. van der Laan†

*University of California, Berkeley, Division of Biostatistics, wenjing.zheng@ucsf.edu
†University of California, Berkeley, laan@berkeley.edu

Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation

Wenjing Zheng and Mark J. van der Laan

Abstract

We consider a targeted maximum likelihood estimator of a path-wise differentiable parameter of the data generating distribution in a semi-parametric model based on observing n independent and identically distributed observations. The targeted maximum likelihood estimator (TMLE) uses V-fold sample splitting for the initial estimator in order to make the TMLE maximally robust in its bias reduction step. We prove a general theorem that states asymptotic efficiency (and thereby regularity) of the targeted maximum likelihood estimator when the initial estimator is consistent and a second order term converges to zero in probability at a rate faster than the square root of the sample size, but no other meaningful conditions are needed. In particular, the conditions of this theorem allow the full utilization of loss based super learning to obtain the initial estimator.

In particular, the theorem proves that first order efficient and unbiased estimation is enhanced in an important way by using adaptive estimators such as an super learner, thereby formally dealing with the concern that adaptive estimation might make it harder to construct valid confidence intervals. On the contrary, the theorem teaches us that to achieve first order efficiency and regularity, it is crucial to estimate the relevant parts of the true data generating distribution as good as possible. The theorem is applied to prove asymptotic efficiency of the targeted maximum likelihood estimator of the additive causal effect of a binary treatment on an outcome in a randomized controlled trial and in an observational study. Excellent finite sample performance of this estimator has been demonstrated in past articles (e.g.van der Laan et al. (September, 2009), Gruber and van der Laan (2010), Stitelman and van der Laan (2010), Petersen et al. (2010).

1 Introduction.

Current practice in statistics often involves fitting parametric or stringent semi-parametric regression models and using statistical inference for the regression coefficients in these models. These models are always wrong, and as a consequence the point estimates and confidence intervals are biased. Large sample sizes are not reducing this bias, but enhances false rejections of null hypotheses. In addition, this parametric approach does not focus on carefully translating the scientific question of interest in terms of a target parameter of the probability distribution of the data.

In van der Laan and Rubin (2006) we introduced targeted maximum likelihood estimation (TMLE) in semiparametric models, which incorporates adaptive estimation (e.g., loss based super learning) of the relevant part of the data generating distribution, and subsequently carries out a targeted bias reduction by maximizing the log-likelihood (or other loss function for the relevant part) of a "clever" parametric working-model through the initial estimator, treating the initial estimator as off-set, and possibly iterates this targeted updating step till convergence. The target parameter of the resulting updated estimator is then evaluated, and is called the TMLE of the target parameter of the data generating distribution. This estimator is, by definition, a substitution estimator, and, under regularity conditions, is a double robust semiparametric efficient estimator. We refer the reader to van der Laan et al. (September, 2009) for applications of TMLE.

The use of adaptive estimators raises the question till what degree we can still rely on the central limit theorem for statistical inference. Our previous theorems show that under empirical process conditions and rate of convergence conditions, one can indeed still prove asymptotic linearity, and thereby obtain CLT-based inference. The empirical process conditions puts some bounds on how adaptive the initial estimator can be. Indeed, we have experienced that using as initial estimator an adaptive regression algorithm that overfits the data such as the machine learning algorithm Random Forest can negatively impact the bias reduction performance of the subsequent TMLE-step. In this paper we present a version of targeted MLE that uses V-fold sample splitting. We refer to this as the cross-validated targeted MLE (CV-TMLE). We formally establish its asymptotics under stated conditions that avoid such empirical process conditions. The implications of this theorem for the role of super learning (i.e., adaptive estimation) in construction of semiparametric efficient estimators of target parameters is discussed. We also present a direct application of this version of targeted MLE to the estimation of the additive causal effect of a binary treatment on an outcome.

We shall see that under mild conditions (e.g. initial estimators need not be consistent), the resulting estimator is of the form

$$\psi_n^* - \psi_0 = (P_n - P_0)IC(P_0) + R_n,$$

where the remainder is second order:

$$R_n = E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})} \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) \left(g_0 - \hat{g}(P^0_{n,B_n}) \right) \right\}.$$

The conditions for asymptotic linearity of ψ_n^* thus follow from the analysis of this second order term.

The organization of this article is as follows. In section 2 we formally present the TMLE using V-fold sample splitting for the initial estimator (CV-TMLE). In section 3 we focus on the one-step CV-TMLE and present a theorem establishing its asymptotics. The conditions and implications of the theorem are discussed. We also present an extension of the theorem with more practical implications. In section 4 the theorem is demonstrated for the cross-validated TMLE of the causal effect of a binary treatment on a continuous or binary outcome. We discuss the implications of the theorem in strategies for estimating the target parameter of the data generating distribution using data adaptive estimators combined with CV-TMLE. In section 5 we present a theorem for the general iterative CV-TMLE, and its conditions are discussed. We end this article with a discussion. Technical derivations are put in the Appendix.

2 The TMLE using V-fold sample splitting for initial estimator.

Let $O \sim P_0$ and the probability distribution P_0 is known to be an element of a statistical model $\mathcal{M}$. We observe n i.i.d. copies $O_1, \ldots, O_n$ of O and wish to estimate a particular multivariate target parameter $\Psi(P_0) \in \mathbb{R}^d$, where $\Psi : \mathcal{M} \to \mathbb{R}^d$ and d denotes the dimension of the parameter. Let P_n denote the empirical probability distribution of $O_1, \ldots, O_n$ so that estimators can be represented as mappings from an empirical distribution to the parameter space of the parameter it is estimating: for example, $P_n \to \hat{\Psi}(P_n)$ denotes an estimator of $\psi_0 = \Psi(P_0)$.

We assume that Ψ is pathwise differentiable at each $P \in \mathcal{M}$ along a class of 1-dimensional sub-models $\{P_h(\epsilon) : \epsilon\}$ indexed by a choice h in an index set

$\mathcal{H}$: i.e., there exists a fixed d-variate function $D(P) = (D_1(P), \ldots, D_d(P))$ so that for all $h \in \mathcal{H}$

$$\frac{d}{d\epsilon}\Psi(P_h(\epsilon))\Big|_{\epsilon=0} = PD(P)S(h),$$

where $S(h)$ is the score of $\{P_h(\epsilon) : \epsilon\}$ at $\epsilon = 0$. Here we used the notation $PS = \int S(o)dP(o)$ for the expectation of a function S of O.

We assume that a parameter $Q : \mathcal{M} \to \mathcal{Q}$ is chosen so that $\Psi(P_0) = \Psi^1(Q(P_0))$ for some mapping $\Psi^1 : \mathcal{Q} \to \mathbb{R}^d$. For convenience, we will refer to both mappings with Ψ, so we will abuse notation by using interchangeably $\Psi(Q(P))$ as well as $\Psi(P)$. Let $g : \mathcal{M} \to \mathcal{G}$ be so that for all $P \in \mathcal{M}$,

$$D^*(P) = D^*(Q(P), g(P)).$$

In other words, the canonical gradient only depends on P through a relevant part $Q(P)$ of P and a nuisance parameter $g(P)$ of P.

Let $\mathcal{L}^\infty(K)$ be the class of functions of O with bounded supremum norm over a set of K so that $P_0(O \in K) = 1$, endowed with the supremum norm. We assume there exists an uniformly bounded loss function $L : \mathcal{Q} \to \mathcal{L}^\infty(K)$ so that

$$Q(P_0) = \arg\min_{Q \in \mathcal{Q}} P_0 L(Q),$$

where, we remind the reader that $P_0L(Q) = \int L(Q)(o)dP_0(o)$. In addition, we assume that for each $P \in \mathcal{M}$, for a specified d-dimensional (hardest) parametric model $\{P(\epsilon) : \epsilon\} \subset \mathcal{M}$ through P at $\epsilon = 0$ and with score $D^*(P)$ at $\epsilon = 0$,

$$\langle \frac{d}{d\epsilon} L(Q(P(\epsilon)))\Big|_{\epsilon=0} \rangle \supset \langle D^*(P) \rangle.$$

We are now ready to define a targeted maximum likelihood estimator. Let $P_n \to \hat{Q}(P_n)$ be an initial estimator of $Q_0 = Q(P_0)$. Let $P_n \to \hat{g}(P_n)$ be an initial estimator of $g_0 = g(P_0)$. Given $\hat{Q}, \hat{g}$, let $P_n \to \hat{Q}(P_n)(\epsilon)$ be a family of estimators indexed by ϵ chosen so that

$$\langle \frac{d}{d\epsilon} L(\hat{Q}(P_n)(\epsilon))\Big|_{\epsilon=0} \rangle \supset \langle D^*(\hat{Q}(P_n), \hat{g}(P_n)) \rangle. \quad (1)$$

Here we used the notation $\langle h \rangle$ for the linear span spanned by the components of $h = (h_1, \ldots, h_k)$. One can think of $\{\hat{Q}(P_n)(\epsilon) : \epsilon\} \subset \mathcal{M}$ as a submodel through $\hat{Q}(P_n)$ with parameter ϵ, chosen so that the derivative(or score) at $\epsilon = 0$ yields a function that equals or spans the efficient influence curve at

the initial estimator $(\hat{Q}(P_n), \hat{g}(P_n))$. Note that this submodel for fluctuating $\hat{Q}(P_n)$ uses the estimator $\hat{g}(P_n)$ in its definition.

Let $B_n \in \{0,1\}^n$ be a random vector indicating a split of $\{1,\ldots,n\}$ into a training and validation sample: $\mathcal{T} = \{i : B_n(i) = 0\}$ and $\mathcal{V} = \{i : B_n(i) = 1\}$. Let P^0_{n,B_n}, P^1_{n,B_n} be the empirical probability distributions of the training and validation sample, respectively. For a given cross-validation scheme $B_n \in \{0,1\}^n$, we now define

$$\epsilon^0_n = \hat{\epsilon}(P_n) \equiv \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L(\hat{Q}(P^0_{n,B_n})(\epsilon)).$$

This now yields an update $\hat{Q}(P^0_{n,B_n})(\epsilon^0_n)$ of $\hat{Q}(P^0_{n,B_n})$ for each split B_n.

As a side-note, it is of interest to point out that this cross-validated selector of ϵ equals the cross-validation selector among the library of candidate estimators $P_n \to \hat{Q}(P_n)(\epsilon)$ of Q_0 indexed by ϵ. As a consequence, we can apply the results for the cross-validation selector that show that it is asymptotically equivalent with the so called oracle selector. Formally, consider the oracle selector

$$\tilde{\epsilon}^0_n \equiv \arg\min_{\epsilon} E_{B_n} P_0 L(\hat{Q}(P^0_{n,B_n})(\epsilon)).$$

If, in addition to uniform boundedness, we assume that the loss function also satisfies

$$M_2 = \sup_{Q \in \mathcal{Q}} \frac{\mathrm{VAR}\{L(Q) - L(Q_0)\}}{E_0\{L(Q) - L(Q_0)\}} < \infty,$$

then the results in van der Laan and Dudoit (2003) and van der Vaart et al. (2006) imply that we have the following finite sample inequality:

$$\begin{aligned} 0 &\leq E E_{B_n} P_0\{L(\hat{Q}(P^0_{n,B_n})(\epsilon^0_n)) - L(\hat{Q}(P^0_{n,B_n})(\tilde{\epsilon}^0_n))\} \\ &\leq 2\sqrt{c}\frac{1}{\sqrt{n}}\sqrt{E E_{B_n} P_0\{L(\hat{Q}(P^0_{n,B_n})(\tilde{\epsilon}^0_n)) - L(Q_0)\}}. \end{aligned}$$

Here c can be explicitly bounded by M_2 and an upper bound of L. This finite sample inequality gives us insight in the benefit of using cross-validation to select the amount of fluctuation ϵ, since it shows that ϵ^0_n will be close to the oracle selector $\tilde{\epsilon}^0_n$ for any choice of initial estimators (even if the initial estimator is extremely data adaptive).

One could now iterate this updating process of the training sample specific estimators: define $\hat{Q}^1(P^0_{n,B_n}) = \hat{Q}(P^0_{n,B_n})(\epsilon^0_n)$, define the family of fluctuations $P_n \to \hat{Q}^1(P_n)(\epsilon)$ satisfying the derivative condition (1), and set

$$\epsilon^1_n = \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L(\hat{Q}^1(P^0_{n,B_n})(\epsilon)),$$

resulting in another update $\hat{Q}^1(P^0_{n,B_n})(\epsilon^1_n)$ for each B_n. This process is iterated till $\epsilon^k_n = 0$ (or close enough to zero). The final update will be denoted with $\hat{Q}^*(P^0_{n,B_n})$ for each split B_n. The targeted MLE is now defined as

$$\hat{\Psi}(P_n) \equiv E_{B_n}\Psi(\hat{Q}^*(P^0_{n,B_n})).$$

We refer to this as the *cross-validated TMLE* (CV-TMLE).

In a variety of examples, the convergence occurs in one step (i.e., $\epsilon^1_n = 0$ already). In this case, we write $\epsilon_n \equiv \epsilon^0_n$ and

$$\hat{\Psi}(P_n) = E_{B_n}\Psi(\hat{Q}(P^0_{n,B_n})(\epsilon_n)).$$

2.1 Cross-validated TMLE when one of the components is linear in data generating distribution

The CV-TMLE presented above can be generalized to the case where only one component of the initial estimator $\hat{Q}(P_n)$ should be updated using a parametric working fluctuation model, while the other component can be estimated using a substitution estimator plugging in the empirical probability distribution function (i.e., an NPMLE). In this case, it is not necessary to target the second component since it is already an unbiased estimator. Formally, consider a decomposition of Q into (Q_1, Q_2), such that $Q_2 \to \Psi(Q_1, Q_2)$ is linear, and $Q_2(P)$ is linear in P itself so that it is sensible to estimate it with an empirical probability distribution. Suppose that the canonical gradient D^* can be decomposed as

$$D^*(P) = D^*_1(P) + D^*_2(P),$$

where $D^*_1(P_0)$ is the canonical gradient of the map

$$P \to \Psi(Q_1(P), Q_2(P_0))$$

at $P = P_0$. Assume also that $D^*_1(P)$ does not depend on $Q_2(P)$.

Under these assumptions we can apply the CV-TMLE algorithm to obtain a targeted estimator of $Q_1(P_0)$, while not updating the initial estimator of $Q_2(P_0)$. In this case, the parametric fluctuation model satisfies

$$\langle \frac{d}{d\epsilon} L(\hat{Q}_1(P_n)(\epsilon))\Big|_{\epsilon=0} \rangle \supset \langle D^*_1(\hat{Q}_1(P_n), \hat{g}(P_n))\rangle,$$

where $L()$ is now a loss function for $Q_1(P_0)$ only. For a given cross-validation scheme $B_n \in \{0,1\}^n$, we define

$$\epsilon^0_n = \hat{\epsilon}(P_n) \equiv \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L(\hat{Q}_1(P^0_{n,B_n})(\epsilon)).$$

This now yields an update $\hat{Q}_1(P^0_{n,B_n})(\epsilon^0_n)$ of $\hat{Q}_1(P^0_{n,B_n})$ for each split B_n. One could now iterate this updating process of the training sample specific estimators: define $\hat{Q}^1_1(P^0_{n,B_n}) = \hat{Q}_1(P^0_{n,B_n})(\epsilon^0_n)$, define the family of fluctuations $P_n \to \hat{Q}^1_1(P_n)(\epsilon)$ satisfying the derivative condition (1), and set

$$\epsilon^1_n = \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L(\hat{Q}^1_1(P^0_{n,B_n})(\epsilon)),$$

resulting in another update $\hat{Q}^1_1(P^0_{n,B_n})(\epsilon^1_n)$ for each B_n. This process is iterated till $\epsilon^k_n = 0$ (or close enough to zero). The final update will be denoted with $\hat{Q}^*_1(P^0_{n,B_n})$ for each split B_n. The resulting CV-TMLE of ψ_0 is given by

$$\hat{\Psi}(P_n) = E_{B_n} \Psi\left(\hat{Q}^*_1(P^0_{n,B_n}), \hat{Q}_2(P^1_{n,B_n})\right).$$

We will illustrate this estimator with an application to the additive causal effect of a binary treatment on a continuous or binary outcome in section 4.

3 Asymptotics for the one-step cross-validated TMLE

In this section we analyze the cross-validated targeted MLE that converge in one step. The theorem carries relevance in general since it establishes the theoretical behavior of the targeted MLE updating algorithm. For convenience, in this section and the next ϵ^0_n is simply denoted with ϵ_n. In the following theorem, convergence in probability always refers to convergence when n converges to infinity.

Definition 1. *For a class of functions, $\mathcal{F}$, whose elements are functions f that map O into a real number, we define the entropy integral*

$$\textit{Entro}(\mathcal{F}) \equiv \int_0^\infty \sqrt{\log \sup_Q N(\epsilon \parallel F \parallel_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon,$$

where $N(\epsilon, \mathcal{F}, L_2(Q))$ is the covering number, defined as the minimal number of balls of radius $\epsilon > 0$ needed to cover $\mathcal{F}$, using the $L_2(Q)$-norm when defining a ball of radius ϵ. In addition, F is defined as the envelope of $\mathcal{F}$ which is a function F so that $\mid f \mid \leq F$ for all $f \in \mathcal{F}$.

We refer to van der Vaart and Wellner (1996) for empirical process theory. We state the following lemma (Lemma 2.14.1 in van der Vaart and Wellner (1996)) for ease of reference.

Lemma 1. *Let $\mathcal{F}$ denote a class of measurable functions of O. Let $G_n = \sqrt{n}(P_n - P_0)$. Then*

$$E\left(sup_{f\in\mathcal{F}}|G_n f|\right) \leq Entro(\mathcal{F})\sqrt{P_0 F^2}.$$

The following result is an immediate application of lemma 1.

Lemma 2. *Suppose $\| \epsilon_n - \epsilon_0 \| \xrightarrow{P} 0$. For each sample split of B_n, we condition on P^0_{n,B_n} and consider a class of measurable functions of O:*

$$\mathcal{F}(P^0_{n,B_n}) \equiv \left\{ f_\epsilon(P^0_{n,B_n}) \equiv f(\epsilon, P^0_{n,B_n}) - f(\epsilon_0, P_0) : \epsilon \right\},$$

where the index set contains ϵ_n with probability tending to 1. For a deterministic sequence $\delta_n \to 0$, define the subclasses

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) \equiv \left\{ f_\epsilon \in \mathcal{F}(P^0_{n,B_n}) : \| \epsilon - \epsilon_0 \| < \delta_n \right\}.$$

If for deterministic sequence $\delta_n \to 0$, we have

$$E\left\{ Entro(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0 F(\delta_n, P^0_{n,B_n})^2} \right\} \to 0 \quad as\ n \to \infty,$$

where $F(\delta_n, P^0_{n,B_n})$ is the envelope of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$, then

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{ f(\epsilon_n, P^0_{n,B_n}) - f(\epsilon_0, P_0) \right\} = o_P(1).$$

Theorem 1. *Let $\hat{Q}(P_n)$, $\hat{g}(P_n)$ be an initial estimator of Q_0, g_0, respectively. In the following, $\hat{Q}(P_0)$ and $\hat{g}(P_0)$ denote the limits of these estimators, not necessarily equal to Q_0 and g_0, respectively.*
Uniformly bounded loss function: *We assume that $\{\hat{Q}(P_n)(\epsilon) : \epsilon\} \in \mathcal{Q}$ with probability 1, the loss function $L(Q)$ for Q_0 is uniformly bounded in $Q \in \mathcal{Q}$, and over a support of $O \sim P_0$:*

$$M_1 = \sup_Q \sup_O \mid L(Q)(O) \mid < \infty.$$

Let $B_n \in \{0,1\}^n$ be a random vector indicating a split of $\{1, \ldots, n\}$ into a training and validation sample. Suppose B_n is uniformly distributed over a finite support.

Consider the estimator defined above

$$\hat{\Psi}(P_n) = E_{B_n}\Psi(\hat{Q}(P^0_{n,B_n})(\epsilon_n)).$$

If the parameter $P \to \Psi(Q(P))$ satisfies

A1:

$$\Psi(Q(P)) - \Psi(Q_0) = -P_0 D^*(Q(P), g_0) + O_P(\| \Psi(Q(P)) - \Psi(Q_0) \|^2).$$

Then

$$\begin{aligned}
\hat{\Psi}(P_n) - \psi_0 &= E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \\
&+ E_{B_n} P_0 \left\{ D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon_n), g_0\right)\right\} \\
&- E_{B_n} P_0 \left\{ D^*\left(Q_0, \hat{g}(P^0_{n,B_n})\right) - D^*\left(Q_0, g_0\right)\right\} \\
&+ O_P(\| \hat{\Psi}(P_n) - \psi_0 \|^2). \qquad (2)
\end{aligned}$$

Consider $\epsilon_0 = \epsilon(P_0)$ *such that* $\| \epsilon_n - \epsilon_0 \| \xrightarrow{P} 0$. *Suppose the following assumption also holds:*

A2: (Given $\| \epsilon_n - \epsilon_0 \| \xrightarrow{P} 0$*)*

For each sample split B_n, *condition on* P^0_{n,B_n} *and define the class of functions*

$$\mathcal{F}(P^0_{n,B_n}) \equiv \{O \to D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)\right) : \epsilon\},$$

where the set over which ϵ *varies is chosen so that it contains* ϵ_n *with probability tending to 1. In addition, for a deterministic sequence* δ_n *converging to zero as* $n \to \infty$, *we also define the sequence of sub-classes*

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) \equiv \left\{ f_\epsilon \in \mathcal{F}(P^0_{n,B_n}) : \| \epsilon - \epsilon_0 \| < \delta_n \right\}.$$

Assume that for deterministic sequence δ_n *converging to* 0, *we have*

$$EEntro(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0 F^2(\delta_n, P^0_{n,B_n})} \to 0 \text{ as } n \to \infty,$$

where $F(\delta_n, P^0_{n,B_n})$ *is the envelope of* $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$.

Then we have:

$$\begin{aligned}
\hat{\Psi}(P_n) - \psi_0 &= (P_n - P_0) D^*\left(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + o_P(1/\sqrt{n}) \\
&+ E_{B_n} P_0 \left\{ D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon), g_0\right)\right\} \\
&- E_{B_n} P_0 \left\{ D^*\left(Q_0, \hat{g}(P^0_{n,B_n})\right) - D^*\left(Q_0, g_0\right)\right\} \\
&+ O_P(\| \hat{\Psi}(P_n) - \psi_0 \|^2). \qquad (3)
\end{aligned}$$

Furthermore, suppose $\hat{g}(P_n) = g_0$. *Then*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*\left(\hat{Q}(P_0)(\epsilon_0), g_0\right) + o_P(1/\sqrt{n}). \tag{4}$$

If, in addition to $\hat{g}(P_n) = g_0$, *we also have* $\hat{Q}(P_0)(\epsilon_0) = Q_0$, *then* $\hat{\Psi}(P_n)$ *is in fact asymptotically efficient:*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*\left(Q_0, g_0\right) + o_P(1/\sqrt{n}). \tag{5}$$

More generally, suppose $\hat{g}(P_0) = g_0$. *Let* $\tilde{Q}$ *denote the limit of* $\hat{Q}(P_n)(\epsilon_n)$ *which is not necessarily* Q_0. *Assume in addition*

A3:

$$\begin{aligned}
&E_{B_n} P_0 \left\{ D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P^0_{n,B_n})(\epsilon_n), g_0\right)\right\} \\
&- E_{B_n} P_0 \left\{ D^*\left(\tilde{Q}, \hat{g}(P^0_{n,B_n})\right) - D^*\left(\tilde{Q}, g_0\right)\right\} \\
&= o_P(1/\sqrt{n}).
\end{aligned}$$

A4: For some mean zero function $IC'(P_0) \in L^2_0(P_0)$, *we have*

$$\begin{aligned}
&E_{B_n} P_0 \left\{ D^*\left(\tilde{Q}, \hat{g}(P^0_{n,B_n})\right) - D^*\left(\tilde{Q}, g_0\right)\right\} \\
&- E_{B_n} P_0 \left\{ D^*\left(Q_0, \hat{g}(P^0_{n,B_n})\right) - D^*(Q_0, g_0)\right\} \\
&= (P_n - P_0)\, IC'(P_0) + o_P(1/\sqrt{n}).
\end{aligned}$$

NOTE: *If* $\hat{Q}(P_n)(\epsilon_n)$ *converges to* Q_0 *then A4 is automatically true with* $IC' \equiv 0$.

Then $\hat{\Psi}(P_n)$ *is asymptotically linear*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\left\{ D^*\left(\hat{Q}(P_0)(\epsilon_0), g_0\right) + IC'(P_0)\right\} + o_P(1/\sqrt{n}).$$

Proof of Theorem 1:

From definition of ϵ_n and the one-step convergence of $\hat{Q}(P)(\epsilon_n)$, we have that

$$E_{B_n} P^1_{n,B_n} D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) = 0.$$

Combining this result with A1 and the double robustness of D^*, which guarantees $P_0 D^*(Q_0, g) = 0$ for all g, we readily have (2):

$$\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 = E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) && (6)\\
&+ E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), g_0)\right\} && (7)\\
&- E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0)\right\} && (8)\\
&+ O_P(\parallel \hat{\Psi}(P_n) - \psi_0 \parallel^2).
\end{aligned}$$

We may rewrite (6) as

$$
\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))\\
&=E_{B_n}\left(P^1_{n,B_n}-P_0\right)\left\{D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^*(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0))\right\}\\
&+E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0)).
\end{aligned}
$$

An application of lemma 2 and A2 implies that for each sample split B_n,

$$
\begin{aligned}
&\left(P^1_{n,B_n}-P_0\right)\left\{D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^*(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0))\right\}\\
&=o_P(1/\sqrt{n}).
\end{aligned}
$$

Since B_n is uniformly distributed on a finite support, it now follows that indeed

$$
\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n}-P_0\right)\left\{D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^*(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0))\right\}\\
&=o_P(1/\sqrt{n}).
\end{aligned}
$$

In other words, the term (6) is given by

$$
\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))\\
&=E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0))+o_P(1/\sqrt{n}).
\end{aligned}
$$

This result and the established equality in (2) now prove (3).

Now, if $\hat{g}(P_n)=g_0$, then the (7) and (8) are exactly 0. Consequently, (3) becomes

$$
\begin{aligned}
\hat{\Psi}(P_n)-\psi_0 &= (P_n-P_0)\,D^*(\hat{Q}(P_0)(\epsilon_0),g_0)\\
&+ o_P(1/\sqrt{n})+O_P(\parallel\hat{\Psi}(P_n)-\psi_0\parallel^2).
\end{aligned}
$$

However, note that taking $\|\|$ on both sides of the equality above yields $\parallel\hat{\Psi}(P_n)-\psi_0\parallel=o_P(1/\sqrt{n})$. We thereby have asymptotically linearity of $\hat{\Psi}(P_n)$:

$$
\hat{\Psi}(P_n)-\psi_0=(P_n-P_0)\,D^*(\hat{Q}(P_0)(\epsilon_0),g_0)+o_P(1/\sqrt{n}).
$$

If, in addition, $\hat{Q}(P_0)(\epsilon_0)=Q_0$, then the influence curve is indeed the efficient influence curve $D^*(Q_0,g_0)$.

Next we consider a more general case where $\hat{g}(P_0)=g_0$. Let $\tilde{Q}$ be the limit of $\hat{Q}(P_n)(\epsilon_n)$. It is not necessarily the case that $\tilde{Q}=Q_0$. We now

rewrite the established equality (3) to account for $\tilde{Q}$:

$$\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*\left(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + o_P(1/\sqrt{n}) \\
&+E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), g_0) \right\} \\
&-E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&+E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&-E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\
&+O_P(\parallel \hat{\Psi}(P_n) - \psi_0 \parallel^2).
\end{aligned}$$

From A3, the term

$$\begin{aligned}
&E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\epsilon_n), g_0) \right\} \\
&-E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&= o_P(1/\sqrt{n}).
\end{aligned}$$

From A4, the term

$$\begin{aligned}
&E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&-E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\
&= (P_n - P_0) IC'(P_0) + o_P(1/\sqrt{n}).
\end{aligned}$$

Therefore (3) becomes

$$\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0) \left\{ D^*\left(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + IC'(P_0) \right\} + o_P(1/\sqrt{n}) \\
&+O_P(\parallel \hat{\Psi}(P_n) - \psi_0 \parallel^2).
\end{aligned}$$

Taking $\|\|$ on both sides again yields $\parallel \hat{\Psi}(P_n) - \psi_0 \parallel = o_P(1/\sqrt{n})$. We thereby have the desired result

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0) \left\{ D^*\left(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + IC'(P_0) \right\} + o_P(1/\sqrt{n}).$$

□

3.1 Remarks about conditions of Theorem 1

To understand assumption A1 we note the following. By general property of the efficient influence curve, we have

$$\Psi(P) - \Psi(P_0) = -P_0 D^*(P) + R(P, P_0),$$

where the specifics of the behavior of the remainder R as a function of P, P_0 depend on the particular data structure, semiparametric model, and target parameter. For example, for linear parameters on convex models we have $\Psi(P) - \Psi(P_0) = -P_0 D^*(P)$ exact, as shown in van der Laan (2006).

Under no conditions on the estimators, we determined an exact identity (2) for the cross-validated TMLE minus its target ψ_0, which already provides the main insights about the performance of this estimator. It shows that the analysis of the CV-TMLE involves a cross-validated empirical process term applied to the efficient influence curve, and a remainder term (In many examples we shall see that this remainder is second order). The cross-validated empirical process term is nice because it involves, for each sample split, an empirical mean over a validation sample of an estimated efficient influence curve that is largely estimated based on the training sample. Based on this, one would predict that one can establish a CLT for this cross-validated empirical process term without having to enforce restrictive entropy conditions on the support of (i.e., class of functions that contains) the estimated efficient influence curve (and thereby limit the adaptiveness of the initial estimators). This is formalized by A2 and our second result (3), which replaces the cross-validated empirical process term by an empirical mean of mean zero random variables $D^*(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0))$ plus a negligible $o_P(1/\sqrt{n})$-term. This result only requires the positivity assumption, and *that the estimators converge to a target*. That is, under essentially no conditions beyond the positivity assumption, the CV-TMLE minus the true ψ_0, behaves as an empirical mean of mean zero i.i.d. random variables (which thus converges to a normal distribution, by CLT), plus a specified remainder term. In particular, we control bias of the estimator by making this remainder term as small as possible.

Regarding assumption A2 we note the following. Combined with lemma 2, A2 implies that the cross-validated empirical process term minus an empirical mean of mean zero random variables converges to 0 at root-n rate. The entropy-term in A2 concerns the entropy of a class of functions that are indexed by a finite dimensional parameter. Such entropies are bounded under very weak conditions, mainly that the class of functions are uniformly bounded. As a consequence, to obtain the wished convergence, one first simply provides a bound on the entropy of $\mathcal{F}(P)$ for a fixed P uniformly in all P. In this way, it remains to show that

$$EP_0\mathbf{F}^2(\delta_n, P^0_{n,B_n}) \to 0 \text{ as } n \to \infty.$$

In other words, one shows that the $L^2(P_0)$-norm of the envelope converges to zero for $\delta_n \to 0$ and P^0_{n,B_n} converging to P_0. Again, this is mainly a

consistency condition on $\hat{Q}(P_n)(\epsilon_n)$ (with respect to its limit, which is not necessarily Q_0). More importantly, we do not require that the entropy of the space of initial estimator $\hat{Q}(P_n)(\epsilon)$, and thereby also the entropy of $\hat{g}(P_n)$, is controlled. The latter are typical conditions putting strong restrictions on how data adaptive the estimators $\hat{Q}$ and $\hat{g}$ can be, but these conditions are now completely avoided. This result allows us to fully utilize data adaptive estimators to make the remainder term negligible.

Moreover, in an RCT g_0 is known, and one might set $\hat{g}(P_n) = g_0$, so that the remainder term is exactly equal to zero, giving us the asymptotic linearity (4) of the CV-TMLE under no other conditions than the positivity assumption and convergence of $\hat{\bar{Q}}(P_n)$ to some fixed function. This teaches us the remarkable lesson that in an RCT, one can use very aggressive super learning without causing any violations of the conditions, but one will achieve asymptotic efficiency for smaller sample sizes. In particular, in an RCT in which we use a consistent estimator $\hat{\bar{Q}}$ the CV-TMLE is asymptotically efficient, as stated in (5). That is, in an RCT, this theorem teaches us that CV-TMLE with adaptive estimation of $\bar{Q}_0$ is the way to go.

In more general types of studies, when $\hat{g}(P_n) \neq g_0$, the remainder may not be exactly zero. But its form, as described in (3), will allow us to identify the necessary conditions and general strategies for estimation of Q_0 and g_0 to make this term negligible. We will illustrate this in our example with estimation of additive causal effect of binary treatment on an outcome.

Implication for the use of super learning The importance of using super learning for estimation of both Q_0 and g_0 is now clear. Super learning is essential to make the remainder as small as possible, for controlling bias. Interestingly, at least asymptotically, there seems to be no price for using super learning, but only benefits: one wants the remainder term in (3) to be small, and that requires approximating the true Q_0 and g_0 well, and simultaneously, the use of very data adaptive estimators did not affect the conditions required for the analysis of the asymptotically linear term in (3), due to the V-fold sample splitting. Therefore, to control the bias term asymptotically, the utilization of super learning is essential, while it also improves the efficiency of the first order term. Further investigation of the required conditions for the bias-term will have to teach us if there will be any trade-off between obtaining a good rate of convergence and the entropy of the estimators. We will return to this issue in our example and its following remarks.

3.2 Asymptotics for CV-TMLE when one of the components is linear in data generating distribution

We now study the asymptotics of the CV-TMLE described in section 2.1 when the algorithm converges in one step.

Consider a decomposition of Q into $Q = (Q_1, Q_2)$, such that $Q_2 \mapsto \Psi(Q_1, Q_2)$ is linear, and $Q_2(P)$ is linear in P itself. Suppose we can decompose the canonical gradient D^* as

$$\begin{aligned} D^*(Q_1(P), Q_2(P), g(P)) &= D_1^*(Q_1(P), g(P)) \\ &+ D_2^*(Q_1(P), g(P)) + D_3^*(Q_1(P), Q_2(P), g(P)), \end{aligned}$$

where $D_1^*(P_0)$ is the canonical gradient of the map

$$P \mapsto \Psi\left(Q_1(P), Q_2(P_0)\right)$$

at $P = P_0$. In our additive causal effect example in next section, Q_2 plays the role of the marginal distribution of the baseline covariates and $Q_1 = E(Y|1, W) - E(Q|0, W)$. Since $\Psi(Q_0)$ only involves taking an average w.r.t. the covariate distribution, $Q_{2,0}$ is naturally estimated with its empirical distribution. In our example, $D_2^*(Q_1, g) = Q_1$ and $D_3^*(Q_1, Q_2, g) = -\Psi(Q_1, Q_2)$.

Under certain conditions on D_2^* and D_3^*, the asymptotic results of previous theorem extend naturally to the CV-TMLE where $Q_{1,0}$ is estimated using a fluctuation model and $Q_{2,0}$ is estimated using a substitution estimator plugging in the empirical distribution.

Theorem 2. *Consider a decomposition of Q into $Q = (Q_1, Q_2)$, such that $Q_2 \mapsto \Psi(Q_1, Q_2)$ is linear and $Q_2(P)$ is linear in P.*

Suppose the canonical gradient D^* can be decomposed into

$$\begin{aligned} D^*(Q_1(P), Q_2(P), g(P)) &= D_1^*(Q_1(P), g(P)) \\ &+ D_2^*(Q_1(P), g(P)) + D_3^*(Q_1(P), Q_2(P), g(P)), \end{aligned}$$

where $D_1^*(P_0)$ is the canonical gradient of the map

$$P \mapsto \Psi\left(Q_1(P), Q_2(P_0)\right)$$

at $P = P_0$. Denote $D^{*\prime} \equiv (D_1^* + D_2^*)$

Let $\hat{Q}_1(P_n)$, $\hat{Q}_2(P_n)$, $\hat{g}(P_n)$ be estimators of $Q_{1,0}$, $Q_{2,0}$, g_0, respectively. We will denote their limits with $\hat{Q}_1(P_0)$,$\hat{Q}_2(P_0)$, and $\hat{g}(P_0)$, which are not necessarily equal to $Q_{1,0}$,$Q_{2,0}$ and g_0, respectively.

Uniformly bounded loss function: *We assume that* $\{\hat{Q}_1(P_n)(\epsilon) : \epsilon\} \in \mathcal{Q}$ *with probability 1, the loss function* $L(Q_1)$ *for* $Q_{1,0}$ *is uniformly bounded in* $Q_1 \in \mathcal{Q}$*, and over a support of* $O \sim P_0$*:*

$$M_1 = \sup_Q \sup_O \mid L(Q_1)(O) \mid < \infty.$$

Let $B_n \in \{0,1\}^n$ *be a random vector indicating a split of* $\{1,\ldots,n\}$ *into a training and validation sample. Suppose* B_n *is uniformly distributed over a finite support.*

Denote $\hat{Q}(P_n, B_n)(\epsilon_n) \equiv \left(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n), \hat{Q}_2(P^1_{n,B_n})\right)$*, and let*

$$\hat{\Psi}(P_n) \equiv E_{B_n} \Psi\left(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n), \hat{Q}_2(P^1_{n,B_n})\right).$$

If the parameter $P \to \Psi(Q(P))$ *satisfies*

A1:

$$\Psi(Q(P)) - \Psi(Q_0) = -P_0 D^*(Q(P), g_0) + O_P(\parallel \Psi(Q(P)) - \Psi(Q_0) \parallel^2),$$

and

A2:

$$\begin{aligned} & E_{B_n} P^1_{n,B_n} D_2^*(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) \\ & + E_{B_n} P^1_{n,B_n} D_3^*(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})) \\ & = 0. \end{aligned}$$

Then

$$\begin{aligned} \hat{\Psi}(P_n) - \psi_0 = & E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})) \\ & + E_{B_n} P_0 \left\{ D^*(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_n, B_n)(\epsilon_n), g_0) \right\} \\ & - E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\ & + O_P(\parallel \hat{\Psi}(P_n) - \psi_0 \parallel^2), \end{aligned} \tag{9}$$

Let $\epsilon_0 = \epsilon(P_0)$ *be such that* $\parallel \epsilon_n - \epsilon_0 \parallel \xrightarrow{P} 0$*.*

Suppose the following assumptions also hold

A3: For each sample split B_n

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{D^*_3\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) - D^*_3\left(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)\right)\right\} = o_P(1),$$

where $\hat{Q}(P_0)(\epsilon_0) = \left(\hat{Q}_1(P_0)(\epsilon_0), \hat{Q}_2(P_0)\right)$.

A4: (Given $\| \epsilon_n - \epsilon_0 \| \xrightarrow{P} 0$*,)*

Conditional on each P^0_{n,B_n}*, define the class of functions*

$$\mathcal{F}(P^0_{n,B_n}) \equiv \{O \to D^{*\prime}(\hat{Q}_1(P^0_{n,B_n})(\epsilon), \hat{g}(P^0_{n,B_n})) - D^{*\prime}(\hat{Q}_1(P_0)(\epsilon_0), \hat{g}(P_0)) : \epsilon\},$$

where the set over which ϵ *varies is chosen so that it contains with probability tending to 1* ϵ_n*. In addition, for a deterministic sequence* δ_n *converging to zero as* $n \to \infty$*, we also define the sequence of subclasses*

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) \equiv \left\{f_\epsilon \in \mathcal{F}(P^0_{n,B_n}) : \| \epsilon - \epsilon_0 \| < \delta_n\right\}.$$

Assume that for deterministic sequence δ_n *converging to* 0*, we have*

$$E\,Entr(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0 F(\delta_n, P^0_{n,B_n})^2} \to 0 \text{ as } n \to \infty,$$

where $F(\delta_n, P^0_{n,B_n})$ *is the envelope of* $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$*.*

Then we have:

$$\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*(\hat{Q}_1(P_0)(\epsilon_0), \hat{Q}_2(P_0), \hat{g}(P_0)) + o_P(1/\sqrt{n}) \\
&+E_{B_n} P_0\left\{D^*(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_n, B_n)(\epsilon_n), g_0)\right\} \\
&-E_{B_n} P_0\left\{D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0)\right\} \\
&+O_P(\| \hat{\Psi}(P_n) - \psi_0 \|^2), \qquad (10)
\end{aligned}$$

Furthermore, suppose $\hat{g}(P_n) = g_0$*. Then*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*\left(\hat{Q}_1(P_0)(\epsilon_0), \hat{Q}_2(P_0), g_0\right) + o_P(1/\sqrt{n}).$$

If, in addition to $\hat{g}(P_n) = g_0$*, we also have* $\hat{Q}_1(P_0)(\epsilon_0) = Q_{1,0}$ *and* $\hat{Q}_2(P_0) = Q_{2,0}$*, then* $\hat{\Psi}(P_n)$ *is in fact asymptotically efficient*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*(Q_0, g_0) + o_P(1/\sqrt{n}).$$

More generally, suppose $\hat{g}(P_0) = g_0$*. Let* $\tilde{Q}_1$ *denote the limit of* $\hat{Q}_1(P_n)(\epsilon_n)$ *which is not necessarily* $Q_{1,0}$*. Assume in addition*

A5:

$$E_{B_n}P_0\left\{D^*(\hat{Q}(P_n,B_n)(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^*(\hat{Q}(P_n,B_n)(\epsilon_n),g_0)\right\}$$
$$-E_{B_n}P_0\left\{D^*(\tilde{Q}_1,\hat{Q}_2(P_0),\hat{g}(P^0_{n,B_n}))-D^*(\tilde{Q}_1,\hat{Q}_2(P_0),g_0)\right\}$$
$$=o_P(1/\sqrt{n}).$$

A6: For some mean zero function $IC'(P_0)\in L^2_0(P_0)$, we have

$$E_{B_n}P_0\left\{D^*(\tilde{Q}_1,\hat{Q}_2(P_0),\hat{g}(P^0_{n,B_n}))-D^*(\tilde{Q}_1,\hat{Q}_2(P_0),g_0)\right\}$$
$$-E_{B_n}P_0\left\{D^*(Q_0,\hat{g}(P^0_{n,B_n}))-D^*(Q_0,g_0)\right\}$$
$$=(P_n-P_0)\,IC'(P_0)+o_P(1/\sqrt{n}).$$

NOTE: *If $\hat{Q}_1(P_n)(\epsilon_n)$ converges to $Q_{1,0}$ and $\hat{Q}_2(P_n)$ converges to $Q_{2,0}$ then A6 is automatically true with $IC'\equiv 0$.*

Then $\hat{\Psi}(P_n)$ is asymptotically linear

$$\hat{\Psi}(P_n)-\psi_0=(P_n-P_0)\left\{D^*\left(\hat{Q}_1(P_0)(\epsilon_0),\hat{Q}_2(P_0),g_0\right)+IC'(P_0)\right\}+o_P(1/\sqrt{n}).$$

Proof of Theorem 2:

From definition of ϵ_n and one-step convergence of $\hat{Q}_1(P)(\epsilon_n)$, we have that

$$E_{B_n}P^1_{n,B_n}D^*_1(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))=0.$$

Combining this result with A1, A2 and the double robustness of D^*, which guarantees $P_0D^*(Q_0,g)=0$ for all g, we readily have (9):

$$\hat{\Psi}(P_n)-\psi_0=E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n})) \quad (11)$$
$$+E_{B_n}P_0\left\{D^*(\hat{Q}(P_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^*(\hat{Q}(P_{n,B_n})(\epsilon_n),g_0)\right\} \quad (12)$$
$$-E_{B_n}P_0\left\{D^*(Q_0,\hat{g}(P^0_{n,B_n}))-D^*(Q_0,g_0)\right\} \quad (13)$$
$$+O_P(\parallel\hat{\Psi}(P_n)-\psi_0\parallel^2),$$

On the other hand, we may rewrite (11) as

$$E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P_n,B_n)(\epsilon_n),\hat{g}(P^0_{n,B_n}))$$
$$=E_{B_n}\left(P^1_{n,B_n}-P_0\right)\left\{D^{*\prime}(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^{*\prime}(\hat{Q}_1(P_0)(\epsilon_0),\hat{g}(P_0))\right\}$$
$$+(P_n-P_0)\,D^{*\prime}(\hat{Q}_1(P_0)(\epsilon_0),\hat{g}(P_0))$$
$$+E_{B_n}(P^1_{n,B_n}-P_0)\left\{D^*_3(\hat{Q}(P_n,B_n)(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^*_3(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0))\right\}$$
$$+(P_n-P_0)\,D^*_3(\hat{Q}(P_0)(\epsilon_0),\hat{g}(P_0))$$

Applying the lemma 2 with A4 we have that

$$E_{B_n}\left(P^1_{n,B_n}-P_0\right)\left\{D^{*\prime}(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n),\hat{g}(P^0_{n,B_n}))-D^{*\prime}(\hat{Q}_1(P_0)(\epsilon_0),\hat{g}(P_0))\right\}$$
$$=o_P(1/\sqrt{n}).$$

It follows from this result and A3 that the term (11) becomes

$$\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}(P_n,B_n)(\epsilon_n),\hat{g}(P^0_{n,B_n}))\\
&=E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^{*\prime}(\hat{Q}_1(P_0)(\epsilon_0),\hat{g}(P_0))\\
&+(P_n-P_0)D^*_3(\hat{Q}_1(P_0)(\epsilon_0),\hat{Q}_2(P_0),\hat{g}(P_0))+o_P(1/\sqrt{n})\\
&=E_{B_n}\left(P^1_{n,B_n}-P_0\right)D^*(\hat{Q}_1(P_0)(\epsilon_0),\hat{Q}_2(P_0),\hat{g}(P_0))+o_P(1/\sqrt{n})
\end{aligned}$$

These results and the established equality in (9) now prove (10).
Similar steps as in the proof of theorem 1 now complete this proof. □

3.2.1 Remark on conditions of theorem 2

For some parameters, it is more efficacious to only target one component of Q_0 while estimating the other component using a substitution estimator plugging in the empirical distribution. Theorem 2 teaches us that the resulting CV-TMLE, under this partial-targeting scheme, has all the desired properties of its full-targeting counterpart. The analysis of the theoretical behavior of $\hat{\Psi}$ in theorem 1 can be extended natural to obtain the results in theorem 2 if D^*_2 and D^*_3 satisfy A2 and A3. These two conditions give us insight into when it is sensible to use this partial-targeting CV-TMLE for Q.

Condition A2 implies that one may still solve the estimation equation by only targeting Q_1 and estimating Q_2 using the validation set, i.e.

$$E_{B_n}P^1_{n,B_n}D^*\left(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n),\hat{Q}_2(P^1_{n,B_n}),\hat{g}(P^0_{n,B_n})\right)=0.$$

This suggests that it's sensible to employ this partial-targeting scheme only if the estimator $\left(\hat{Q}_1(P^0_{n,B_n})(\epsilon_n),\hat{Q}_2(P^1_{n,B_n})\right)$ will be as good as its full-targeting counterpart in terms of solving the cross validated estimating equation.

In our examples, $D^*_3(Q_1,Q_2,g)=-\Psi(Q_1,Q_2)$, in which case A3 is automatically true since $(P_n-P_0)D^*_3(Q_1,Q_2,g)=0$ for all Q_1,Q_2,g. In such instances, no requirements are imposed on the estimators, and thus the partial-targeting scheme is highly effective. However, when that is not the

case, A3 implies that one will need to control the entropy of the class of estimators $\hat{\bar{Q}}_2$, since they will be evaluated at the training set P^1_{n,B_n}. In these cases, the partial-targeting scheme may not be as effective as the full-targeting one.

4 Application of Theorem 2 to estimation of additive causal effect in nonparametric model

Let $O = (W, A, Y)$, W be a vector of baseline covariates, A a binary treatment variable, and Y an outcome of interest. Let $\mathcal{M}$ be the class of all probability distributions for O. We consider the parameter $\Psi : \mathcal{M} \to \mathbf{R}$

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

Several estimators, in addition to TMLE, have been proposed for the estimation of this parameter: the G-comp estimator (Robins (1986)), the IPTW estimator (Hernan et al. (2000); Robins (1999)), the DR-IPTW estimator (Robins and Rotnitzky (2001); Robins (2000); Robins et al. (2000)). We refer to van der Laan et al. (September, 2009), Gruber and van der Laan (2010), Stitelman and van der Laan (2010), and Petersen et al. (2010) for comparisons of performance between TMLE and these various estimators.

Let $Q(P) = (\bar{Q}(P), Q_W(P))$, where $\bar{Q}(P)(W, A) \equiv E_P(Y|W, A)$ and $Q_W(P)$ is the density of the marginal probability distribution of W. For convenience, we will use $\bar{Q}(P)(W)$ to denote $E_P(Y|W, A = 1) - E_P(Y|W, A = 0)$. The distinctions will be clear from the arguments given to the function or from context. Let $g(P)(A|W) \equiv Pr_P(A|W)$. We also adopt the notations $\bar{Q}_0 \equiv \bar{Q}(P_0)$, $Q_{W,0} \equiv Q_W(P_0)$, and $g_0 \equiv g(P_0)$.

Our parameter of interest is Ψ evaluated at the distribution $P_0 \in \mathcal{M}$ of the observed O:

$$\psi_0 \equiv \Psi(Q_0) = E_{W,0}\left[E_0(Y|W, A = 1) - E_0(Y|W, A = 0)\right].$$

The canonical gradient of Ψ at $P \in \mathcal{M}$ is

$$\begin{aligned} D^*(Q(P), g(P))(O) &= \left\{H^*_{g(P)}(A, W)\left(Y - \bar{Q}(P)(A, W)\right)\right\} \\ &+ \left\{\bar{Q}(P)(W) - Q_W(P)\bar{Q}(P)\right\} \\ &\equiv D^*_Y(\bar{Q}(P), g(P)) + D^*_W(\bar{Q}(P), Q_W(P)), \end{aligned}$$

where

$$H^*_g(A, W) = \left(\frac{A}{g(1|W)} - \frac{1 - A}{g(0|W)}\right).$$

For convenience, we will also use the notation

$$H_g^*(W) \equiv H_g^*(1, W) - H_g^*(0, W).$$

Firstly, note that the map $Q_W \mapsto \Psi(\bar{Q}, Q_W)$ is linear and $Q_W(P)$ is linear in P. Secondly, $D_Y^*(\bar{Q}_0, g_0)$ is the canonical gradient of the map $P \mapsto \Psi(\bar{Q}(P), Q_W(P_0))$ at $P = P_0$, and does not depend on $Q_W(P_0)$. In the following we present a TMLE of Q_0 where only the initial estimator $\hat{\bar{Q}}(P_n)$ of $\bar{Q}_0$ is updated using a parametric working model $\hat{\bar{Q}}(P_n)(\epsilon)$, while the marginal distribution of W is estimated with the empirical distribution which is not updated. Given an appropriate loss function $L(\bar{Q})$ and initial estimators $\hat{\bar{Q}}$ and $\hat{g}$ of $\bar{Q}_0$ and g_0, respectively, the parametric working model $\{\hat{\bar{Q}}(P_n)(\epsilon) : \epsilon\}$ will be selected such that

$$\frac{d}{d\epsilon} L(\hat{\bar{Q}}(P_n)(\epsilon))\Big|_{\epsilon=0} = D_Y^*(\hat{\bar{Q}}(P_n), \hat{g}(P_n)).$$

We consider here two possible loss functions for binary outcome or continuous outcomes $Y \in [0, 1]$.

Squared error loss function: The squared error loss function is given by

$$L(\bar{Q})(O) \equiv (Y - \bar{Q}(A, W))^2,$$

with the parametric working model

$$\hat{\bar{Q}}(P_n)(\epsilon) = \hat{\bar{Q}}(P_n) + \epsilon H_{\hat{g}(P_n)}^*.$$

Quassi-log-likelihood loss function: The quassi-log-likelihood loss function is given by

$$L(\bar{Q})(O) \equiv -\left(Y \log(\bar{Q}(W, A)) + (1 - Y) \log(1 - \bar{Q}(W, A))\right),$$

with parametric working model

$$\hat{\bar{Q}}(P_n)(\epsilon) = \frac{1}{1 + e^{-logit(\hat{\bar{Q}}(P_n)) - \epsilon H_{\hat{g}(P_n)}^*}}.$$

We note that we would use this loss function if Y is binary or Y is continuous with values in $(0, 1)$. If Y is a bounded continuous random variable with values in (a, b), then we can still use this loss function by using the transformed outcome $Y^* = (Y - a)/(b - a)$ and mapping the obtained TMLE of the additive treatment effect on Y^* (and confidence intervals) into a TMLE

of the additive treatment effect on Y (and confidence intervals).

It is important to point out that the TMLE of $\bar{Q}_0$ corresponding with both fluctuation models will converge in one step, since the clever covariate $H^*_{\hat{g}(P_n)}$ in the update of $\hat{\bar{Q}}$ does not involve $\hat{\bar{Q}}$.

Let $B_n \in \{0,1\}^n$ be a random vector indicating a split of $\{1,\ldots,n\}$ into a training and validation sample: $\mathcal{T} = \{i : B_n(i) = 0\}$ and $\mathcal{V} = \{i : B_n(i) = 1\}$. Let P^0_{n,B_n}, P^1_{n,B_n} be the empirical probability distributions of the training and validation sample, respectively. Given the parametric working model, the optimal ϵ_n is selected using cross validation:

$$\epsilon_n = \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)).$$

In particular, the one-step convergence implies that ϵ_n satisfies

$$0 = E_{B_n} P^1_{n,B_n} D^*_Y(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})). \tag{14}$$

At each sample split B_n, we define the TMLE of Q_0 at (P_n, B_n) as

$$\hat{Q}(P_n, B_n)(\epsilon_n) \equiv \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n})\right).$$

The TMLE of ψ_0 is defined as

$$\hat{\Psi}(P_n) \equiv E_{B_n}\Psi\left(\hat{Q}(P_n, B_n)(\epsilon_n)\right) = E_{B_n}\Psi\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n})\right).$$

Next, we illustrate the theoretical advantages of this estimator under both loss functions. We will show that under a natural rate condition on the initial estimators $\hat{\bar{Q}}$ and $\hat{g}$, the resulting TMLE $\hat{\Psi}(P_n)$ is asymptotically linear, and when $\hat{g}$ and $\hat{\bar{Q}}$ are consistent, its influence curve is indeed the efficient influence curve.

4.1 Squared error loss for $\bar{Q}$

Let the loss function for $\bar{Q}_0$ be:

$$L(\bar{Q})(O) \equiv (Y - \bar{Q}(A,W))^2,$$

and consider the parametric working model through $\bar{Q}(P)$ for any $P \in \mathcal{M}$:

$$\bar{Q}(P)(\epsilon) = \bar{Q}(P) + \epsilon H^*_{g(P)}.$$

Then, for given initial estimators $\hat{g}$ and $\hat{Q}$, we have

$$\hat{\bar{Q}}(P_n)(\epsilon) = \hat{\bar{Q}}(P_n) + \epsilon H^*_{\hat{g}(P_n)}. \tag{15}$$

The cross validation selector of ϵ in (15) is defined as

$$\begin{aligned}\epsilon_n &\equiv \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)\right) \\ &= \arg\min_{\epsilon} E_{B_n} \sum_{i,B_n(i)=1} \left(Y_i - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)(A_i,W_i)\right)^2.\end{aligned}$$

At each sample split B_n, we define the TMLE of Q_0 at (P_n, B_n) as

$$\hat{Q}(P_n,B_n)(\epsilon_n) \equiv \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n})\right),$$

where $\hat{Q}_W(P^1_{n,B_n})$ is the marginal empirical distribution of W in the validation set. The TMLE of ψ_0 is defined as

$$\hat{\Psi}(P_n) \equiv E_{B_n} \Psi\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n})\right).$$

We will now apply the main Theorem 2 to $\hat{\Psi}(P_n)$ which provides us with the following result.

Theorem 3. *Consider the setting above under the squared error loss function.*

Let $B_n \in \{0,1\}^n$ be a random vector indicating a split of $\{1,\ldots,n\}$ into a training and validation sample. Suppose B_n is uniformly distributed on a finite support.

Let $\hat{\bar{Q}}$ and $\hat{g}$ be initial estimators of $\bar{Q}_0$ and g_0. In the following, $\hat{\bar{Q}}(P_0)$ and $\hat{g}(P_0)$ denote limits of these estimators, not necessarily equal to $\bar{Q}_0$ and g_0, respectively.

The cross-validated TMLE satisfies

$$\begin{aligned}\hat{\Psi}(P_n) - \psi_0 &= E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n,B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \\ &+ E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})} \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\}. \quad (2)\end{aligned}$$

Suppose now that there exists a constant $L > 0$ such that $P_0(|Y| < L) = 1$.

Consider the following definition:

$$\epsilon_0 \equiv \arg\min_{\epsilon} P_0 L(\hat{\bar{Q}}(P_0)(\epsilon)).$$

Suppose that this minimum exists and satisfies the derivative equation

$$0 = P_0 D_Y(P_0, \epsilon_0),$$

where

$$\begin{aligned} D_Y(P,\epsilon) &\equiv \frac{d}{d\epsilon} L(\hat{\bar{Q}}(P)(\epsilon))(O) \\ &= \left(Y - \hat{\bar{Q}}(P)(A,W) - \epsilon H^*_{\hat{g}(P)}(A,W)\right) H^*_{\hat{g}(P)}(A,W) \\ &= D^*_Y\left(\hat{\bar{Q}}(P)(\epsilon), \hat{g}(P)\right). \end{aligned}$$

If there are multiple minima, then it is assumed that the argmin is uniquely defined and selects one of these minima.

Suppose that $\hat{\bar{Q}}$ and $\hat{g}$ satisfy the following conditions:

1. *There exists a closed bounded set $K \subset \mathbf{R}^k$ containing ϵ_0 such that ϵ_n belongs to K with probability 1;*
2. *For some $\delta > 0$, $P(1 - \delta > \hat{g}(P_n)(1 \mid W) > \delta) = 1$;*
3. *For some $K > 0$, $P(|\hat{\bar{Q}}(P_n)(A,W)| < K) = 1$;*
4. $$\int_W \left(\hat{g}(P_n)(1|W) - \hat{g}(P_0)(1|W)\right)^2 dQ_{W,0}(w) \to 0 \quad \textit{in probability;}$$
5. *For $a = 0, 1$,* $$\int_W \left(\hat{\bar{Q}}(P_n)(a,w) - \hat{\bar{Q}}(P_0)(a,w)\right)^2 dQ_{W,0}(w) \to 0 \quad \textit{in probability.}$$

Then,

$$\begin{aligned} \hat{\Psi}(P_n) - \psi_0 &= (P_n - P_0)\left\{D^*_Y\left(\hat{\bar{Q}}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + \hat{\bar{Q}}(P_0)(\epsilon_0)\right\} \\ &+ E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})} \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\} \quad (3) \\ &+ o_P(1/\sqrt{n}). \end{aligned}$$

Furthermore, If $\hat{g}(P_n) = g_0$, the TMLE estimator $\hat{\Psi}(P_n)$ is asymptotically linear estimator of ψ_0:

$$\hat{\Psi}(P_n) - \psi_0 \quad = \quad (P_n - P_0)\, D^*(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)) + o_P(1/\sqrt{n}), \quad (4)$$

where $\hat{Q}(P_0)(\epsilon_0) = (\hat{\bar{Q}}(P_0)(\epsilon_0), Q_{W,0})$.

If, in addition to $\hat{g}(P_n) = g_0$, $\hat{\bar{Q}}(P_0) = \bar{Q}_0$, which implies that $\hat{\bar{Q}}(P_0)(\epsilon_0) = \bar{Q}_0$, then $\hat{\Psi}(P_n)$ is an asymptotically efficient estimator of ψ_0:

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*(Q_0, g_0) + o_P(1/\sqrt{n}). \quad (5)$$

More generally, if the limits satisfy $\hat{g}(P_0) = g_0$ and $\hat{\bar{Q}}(P_0) = \bar{Q}_0$, and if the convergence satisfies

$$\sqrt{E_{B_n} P_0 \left(\frac{g_0 - \hat{g}(P^0_{n,B_n})}{g_0 \hat{g}(P^0_{n,B_n})} \right)^2} \sqrt{E_{B_n} P_0 \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \bar{Q}_0 \right)^2} = o_P(1/\sqrt{n}), \quad (16)$$

then $\hat{\Psi}(P_n)$ is an asymptotically efficient estimator of ψ_0:

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*(Q_0, g_0) + o_P(1/\sqrt{n}).$$

Consider now the case that $\hat{g}(P_0) = g_0$, but $\hat{\bar{Q}}(P_0) \neq \bar{Q}_0$. If the convergence satisfies

$$\sqrt{E_{B_n} P_0 \left(\frac{g_0 - \hat{g}(P^0_{n,B_n})}{g_0 \hat{g}(P^0_{n,B_n})} \right)^2} \sqrt{E_{B_n} P_0 \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(P_0)(\epsilon_0) \right)^2} = o_P(1/\sqrt{n}), \quad (17)$$

and $P_0 \left\{ H^*_{\hat{g}(P_n)} \left(\hat{\bar{Q}}(P_0)(\epsilon_0) - \bar{Q}_0 \right) \right\}$ is an asymptotically linear estimator of $P_0 \left\{ H^*_{\hat{g}(P_0)} \left(\hat{\bar{Q}}(P_0)(\epsilon_0) - \bar{Q}_0 \right) \right\}$ with influence curve IC', then $\hat{\Psi}(P_n)$ is an asymptotically linear estimator of ψ_0:

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0) \left\{ D^*(\hat{Q}(P_0)(\epsilon_0), g_0) + IC' \right\} + o_P(1/\sqrt{n}).$$

For convenience of reference, we state several simple but useful results in the proof of the theorem.

Lemma 3. *If X_n converges to X in probability, and there exists $A > 0$ such that $P(|X_n| < A) = 1$, then $E|X_n - X|^r \to 0$ for $r \geq 1$.*

Lemma 4. *Suppose $\hat{g}$ is such that for some $\delta > 0$, $P(1-\delta > \hat{g}(P_n)(1 \mid W) > \delta) = 1$. If for $a = 0, 1$, $\hat{g}$ satisfies $P_{W,0}\left(\hat{g}(P_n) - \hat{g}(P_0)\right)^2 \xrightarrow{P} 0$, then we have that $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^4$, $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^2$, $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)$ and $P_0\left({H^*_{\hat{g}(P_n)}}^2 - {H^*_{\hat{g}(P_0)}}^2\right)$ also converge to zero in probability.*

Lemma 5. *Suppose $\hat{g}$ and $\hat{\bar{Q}}$ satisfy the conditions 2-5 in Theorem 3. Then, for each split B_n, for any $r \geq 1$,*

1. $EP_0\left(\hat{\bar{Q}}(P^0_{n,B_n})H^*_{\hat{g}(P^0_{n,B_n})} - \hat{\bar{Q}}(P_0)H^*_{\hat{g}(P_0)}\right)^r \to 0;$
2. $EP_0\left((Y - \hat{\bar{Q}}(P^0_{n,B_n}))H^*_{\hat{g}(P^0_{n,B_n})} - (Y - \hat{\bar{Q}}(P_0))H^*_{\hat{g}(P_0)}\right)^r \to 0;$
3. $EP_0\left({H^*_{\hat{g}(P^0_{n,B_n})}}^2 - {H^*_{\hat{g}(P_0)}}^2\right)^r \to 0;$
4. $EP_0\left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right)^r \to 0.$

We are now ready to prove theorem 3.

Proof. Firstly, we wish to establish that

$$\hat{\Psi}(P_n) - \psi_0 = E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right)$$
$$+E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\},$$

where $\hat{Q}(P_n, B_n)(\epsilon_n) = \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n})\right)$.

Note that

$$\begin{aligned}
&-P_0 D^*(Q(P), g_0) \\
&\equiv -P_0\left\{\left(Y - \bar{Q}(P)\right)H^*_{g_0} + \bar{Q}(P) - P_W(P)\bar{Q}(P)\right\} \\
&= -\left\{P_0 Y H^*_{g_0} - P_0\bar{Q}(P)H^*_{g_0} + P_{W,0}\bar{Q}(P) - P_W(P)\bar{Q}(P)\right\} \\
&= P_W(P)\bar{Q}(P) - P_0 Y H^*_{g_0} \\
&= \Psi(Q(P)) - \Psi(Q_0).
\end{aligned}$$

Applying this result to each sample split of B_n and averaging, it follows that

$$\hat{\Psi}(P_n) - \psi_0 \equiv E_{B_n}\Psi\left(\hat{Q}(P_n, B_n)(\epsilon_n)\right) - \Psi(Q(P_0)) = -E_{B_n}P_0 D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), g_0\right). \tag{18}$$

On the other hand,

$$
\begin{aligned}
&E_{B_n} P^1_{n,B_n} D^*_W \left(\hat{Q}_W(P^1_{n,B_n}), \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) \\
&\equiv E_{B_n} P^1_{n,B_n} \left\{ \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - P_W(P^1_{n,B_n}) \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right\} \\
&= E_{B_n} \left\{ P_W(P^1_{n,B_n}) \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - P_W(P^1_{n,B_n}) \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right\} = 0.
\end{aligned}
$$

Moreover, it follows from the definition of ϵ_n and the one-step convergence of the chosen fluctuation model that $\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n}) \right)$ satisfies (14). Therefore, we have

$$
\begin{aligned}
&E_{B_n} P^1_{n,B_n} D^* \left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n}) \right) \\
&\equiv E_{B_n} P^1_{n,B_n} D^*_Y \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n}) \right) \\
&+ E_{B_n} P^1_{n,B_n} D^*_W \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n}) \right) \\
&= 0. \qquad (19)
\end{aligned}
$$

Combining (18), (19) and robustness of D^*, $P_0 D^*(Q_0, g) = 0$ for all g, we may now rewrite $\hat{\Psi}(P_n) - \psi_0$ as

$$
\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 = E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D^* \left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n}) \right) \\
&+ E_{B_n} P_0 \left\{ D^* \left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n}) \right) - D^* \left(\hat{Q}(P_n, B_n)(\epsilon_n), g_0 \right) \right\} \\
&- E_{B_n} P_0 \left\{ D^* \left(Q_0, \hat{g}(P^0_{n,B_n}) \right) - D^* \left(Q_0, g_0 \right) \right\}.
\end{aligned}
$$

The last two summands in this equality can be combined as

$$
\begin{aligned}
&E_{B_n} P_0 \left\{ D^* \left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n}) \right) - D^* \left(\hat{Q}(P_n, B_n)(\epsilon_n), g_0 \right) \right\} \\
&- E_{B_n} P_0 \left\{ D^* \left(Q_0, \hat{g}(P^0_{n,B_n}) \right) - D^* \left(Q_0, g_0 \right) \right\} \\
&\equiv E_{B_n} P_0 \left\{ D^*_Y(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})) + D^*_W(\hat{Q}(P_n, B_n)(\epsilon_n)) \right\} \\
&- E_{B_n} P_0 \left\{ D^*_Y(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), g_0) + D^*_W(\hat{Q}(P_n, B_n)(\epsilon_n)) \right\} \\
&- E_{B_n} P_0 \left\{ D^*_Y(\bar{Q}_0, \hat{g}(P^0_{n,B_n})) + D^*_W(Q_0) \right\} \\
&+ E_{B_n} P_0 \left\{ D^*_Y(\bar{Q}_0, g_0) + D^*_W(Q_0) \right\} \\
&= E_{B_n} P_0 \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{g_0} \right) \\
&= E_{B_n} P_0 \left\{ \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) (-1)^{1+A} \frac{\left(g_0 - \hat{g}(P^0_{n,B_n}) \right)}{g_0 \hat{g}(P^0_{n,B_n})} \right\}.
\end{aligned}
$$

Therefore, we indeed have the desired expression (2):

$$\hat{\Psi}(P_n) - \psi_0 = E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \quad (20)$$

$$+E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\}. \quad (21)$$

We now study each term separately. For convenience, we use the notation $D_Y(P, \epsilon) \equiv D^*_Y(\hat{\bar{Q}}(P)(\epsilon), \hat{g}(P))$.

The term (20) can be written as

$$\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \\
&= E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*_Y\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \\
&+E_{B_n}\left(P^1_{n,B_n} - P_0\right)\left\{\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - P_W(P^1_{n,B_n})\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right\} \\
&= E_{B_n}\left(P^1_{n,B_n} - P_0\right)\left\{D_Y\left(P^0_{n,B_n}, \epsilon_n\right) - D_Y\left(P_0, \epsilon_0\right)\right\} \quad (22) \\
&+ \left(P_n - P_0\right) D_Y\left(P_0, \epsilon_0\right) \\
&+E_{B_n}\left(P^1_{n,B_n} - P_0\right)\left\{\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(P_0)(\epsilon_0)\right\} \quad (23) \\
&+ \left(P_n - P_0\right)\hat{\bar{Q}}(P_0)(\epsilon_0).
\end{aligned}$$

It follows from the following lemma that ϵ_n converges to ϵ_0 in probability.

Lemma 6. *Let ϵ_n and ϵ_0 be defined as in theorem 3 and suppose they solve the derivative equations as stated in the theorem. If $\hat{g}$ and $\hat{\bar{Q}}$ satisfy the conditions 1-5 in theorem 3, then ϵ_n converges to ϵ_0 in probability.*

Now consider the following lemmas:

Lemma 7. *If the initial estimators $\hat{\bar{Q}}$ and $\hat{g}$ satisfy the conditions 1-5 in the theorem, then, on a sample split of B_n,*

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{D_Y\left(P^0_{n,B_n}, \epsilon_n\right) - D_Y\left(P_0, \epsilon_0\right)\right\} = o_P(1).$$

Lemma 8. *If $\hat{\bar{Q}}$ and $\hat{g}$ satisfy conditions 1-5 of the theorem, then, on a sample split of B_n,*

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(P_0)(\epsilon_0)\right) = o_P(1).$$

Note that lemmas 6, 7 and 8 follow from lemmas 2, 4 and 5.

Lemmas 7 and 8 imply that (22) and (23) are $o_P(1/\sqrt{n})$. We thus have established that (20) is given by

$$E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right)$$
$$= (P_n - P_0)\left\{D^*_Y\left(\hat{\bar{Q}}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + \hat{\bar{Q}}(P_0)(\epsilon_0)\right\} + o_P(1/\sqrt{n}).$$

Combining this result with (21), we have proved (3):

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\left\{D^*_Y\left(\hat{\bar{Q}}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + \hat{\bar{Q}}(P_0)(\epsilon_0)\right\}$$
$$+E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\}$$
$$+o_P(1/\sqrt{n}).$$

Note that up till this point we have only used convergence of $\hat{\bar{Q}}(P_n)$ and $\hat{g}(P_n)$ to some limits, but we assumed neither consistency to the true Q_0, g_0, nor a rate of convergence for these initial estimators to such limits.

Finally, we study the remainder term (21):

$$E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\}.$$

We consider several cases. Firstly, consider the case, $\hat{g}(P_n) = g_0$. In this case, term (21) is exactly 0. Therefore, (3) now implies that $\hat{\Psi}(P_n)$ is asymptotically linear with influence curve $D^*(\hat{Q}(P_0)(\epsilon_0), g_0)$. If in addition, the initial estimator $\hat{\bar{Q}}$ is consistent for $\bar{Q}_0$, i.e. $\hat{\bar{Q}}(P_0) = \bar{Q}_0$, then

$$\begin{aligned}\epsilon_0 &\equiv \arg\min_\epsilon P_0(Y - \hat{\bar{Q}}(P_0) - \epsilon H^*_{\hat{g}(P_0)})^2 \\ &= \arg\min_\epsilon P_0(Y - Q_0 - \epsilon H^*_{\hat{g}(P_0)})^2 = 0.\end{aligned}$$

This implies that $\hat{\bar{Q}}(P_0)(\epsilon_0)$ is simply Q_0. Consequently, $\hat{\Psi}(P_n)$ is asymptotically linear with influence curve $D^*(Q_0, g_0)$, and is thereby asymptotically efficient.

Let's now consider the case that $\hat{g}(P_0) = g_0$ and $\hat{\bar{Q}}(P_0) = \bar{Q}_0$. These imply that (21) converges to 0. However, for $\hat{\Psi}(P_n)$ to be asymptotically linear, it is necessary that the convergence of this second order term occurs

at a $\sqrt{n}$ rate, i.e.

$$E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(g_0-\hat{g}(P^0_{n,B_n})\right)\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\bar{Q}_0\right)\right\}$$
$$=o_P(1/\sqrt{n}).$$

Applying Cauchy-Schwartz inequality, it follows that if

$$\sqrt{E_{B_n}P_0\left(\frac{g_0-\hat{g}(P^0_{n,B_n})}{g_0\hat{g}(P^0_{n,B_n})}\right)^2}\sqrt{E_{B_n}P_0\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\bar{Q}_0\right)^2}$$
$$=o_P(1/\sqrt{n}),$$

then $\hat{\Psi}(P_n)$ will be asymptotically efficient.

Finally, consider the case that $\hat{g}(P_0)=g_0$, but $\hat{\bar{Q}}(P_0)\neq\bar{Q}_0$. We reconsider the expression (21) to account for the limit $\hat{\bar{Q}}(P_0)(\epsilon_0)$ of $\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)$ which does not equal $\bar{Q}_0$:

$$E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(g_0-\hat{g}(P^0_{n,B_n})\right)\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\bar{Q}_0\right)\right\}$$
$$=E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(g_0-\hat{g}(P^0_{n,B_n})\right)\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\hat{\bar{Q}}(P_0)(\epsilon_0)\right)\right\} \tag{24}$$
$$+E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(g_0-\hat{g}(P^0_{n,B_n})\right)\left(\hat{\bar{Q}}(P_0)(\epsilon_0)-\bar{Q}_0\right)\right\}. \tag{25}$$

Firstly, we require again that the rate of convergence for the second order term in (24) be $\sqrt{n}$, that is,

$$E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(g_0-\hat{g}(P^0_{n,B_n})\right)\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\hat{\bar{Q}}(P_0)(\epsilon_0)\right)\right\}$$
$$=o_P(1/\sqrt{n}).$$

Applying Cauchy-Schwartz inequality, it suffices that

$$\sqrt{E_{B_n}P_0\left(\frac{g_0-\hat{g}(P^0_{n,B_n})}{g_0\hat{g}(P^0_{n,B_n})}\right)^2}\sqrt{E_{B_n}P_0\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\hat{\bar{Q}}(P_0)(\epsilon_0)\right)^2}$$
$$=o_P(1/\sqrt{n}).$$

For (25) to be asymptotically linear, stronger requirements on the performance of $\hat{g}$ are needed in order to address the inconsistency of $\hat{\bar{Q}}$. For convenience of notation, we recall that

$$E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})} \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) \left(g_0 - \hat{g}(P^0_{n,B_n}) \right) \right\}$$
$$= E_{B_n} P_0 \left\{ \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{g_0} \right) \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) \right\}.$$

Now, for the given initial estimator $\hat{\bar{Q}}$ and $\hat{g}$, let

$$\Phi(P) \equiv P_0 \left\{ H^*_{\hat{g}(P)} \left(\hat{\bar{Q}}(P_0)(\epsilon_0) - \bar{Q}_0 \right) \right\}.$$

If $\hat{g}$ is such that $\Phi(P_n) - \Phi(P_0)$ is asymptotically linear (with some influence curve IC'), then (25) becomes

$$\begin{aligned} & E_{B_n} P_0 \left\{ \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{g_0} \right) \left(\hat{\bar{Q}}(P_0)(\epsilon_0) - \bar{Q}_0 \right) \right\} \\ & \equiv E_{B_n} \left(\Phi(P^0_{n,B_n}) - \Phi(P_0) \right) \\ & = E_{B_n} \left(P^0_{n,B_n} - P_0 \right) IC' + o_P(1/\sqrt{n}) \\ & = (P_n - P_0) IC' + o_P(1/\sqrt{n}). \end{aligned}$$

Therefore, if $\hat{g}$ and $\hat{\bar{Q}}$ satisfy the convergence speed condition and $\Phi(P_n) - \Phi(P_0)$ asymptotically linear, then it follows from (24) and (25) that the remainder (21) becomes

$$\begin{aligned} & E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})} \left(g_0 - \hat{g}(P^0_{n,B_n}) \right) \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \bar{Q}_0 \right) \right\} \\ & = (P_n - P_0) \, IC' + o_P(1/\sqrt{n}). \end{aligned}$$

This completes the proof. □

4.2 Quassi-log-likelihood loss for $\bar{Q}$

Suppose now that the outcome Y has support in $\mathbf{R}$ and is naturally bounded. After a linear transformation, we may assume without loss of generality that Y has support in $(0, 1)$. Let the loss function be

$$L(\bar{Q})(O) \equiv - \left(Y \log(\bar{Q}) + (1 - Y) \log(1 - \bar{Q}) \right),$$

and consider the parametric working model through $\bar{Q}(P)$ for any $P \in \mathcal{M}$:

$$\bar{Q}(P)(\epsilon) = \frac{1}{1 + e^{-logit(\bar{Q}(P)) - \epsilon H^*_{g(P)}}}.$$

Then, for the given initial estimators $\hat{g}$ and $\hat{Q}$, we obtain the following parametric working model:

$$\hat{\bar{Q}}(P_n)(\epsilon) = \frac{1}{1 + e^{-logit(\hat{\bar{Q}}(P_n)) - \epsilon H^*_{\hat{g}(P_n)}}}. \tag{26}$$

The cross validation selector of ϵ in (26) is defined as

$$\begin{aligned}
\epsilon_n &\equiv \arg\min_{\epsilon} E_{B_n} P^1_{n,B_n} L(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)) \\
&= \arg\min_{\epsilon} -E_{B_n} \sum_{i, B_n(i)=1} \Big\{ Y_i \log(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)(A_i, W_i)) \\
&\qquad + (1 - Y) \log(1 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)(A_i, W_i)) \Big\}.
\end{aligned}$$

At each sample split B_n, we define the TMLE of Q_0 at (P_n, B_n) as

$$\hat{Q}(P_n, B_n)(\epsilon_n) \equiv \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n}) \right),$$

where $\hat{Q}_W(P^1_{n,B_n})$ is the marginal empirical distribution of W in the validation set.

The TMLE of ψ_0 is defined as

$$\hat{\Psi}(P_n) \equiv E_{B_n} \Psi(\hat{Q}(P_n, B_n)(\epsilon_n) = E_{B_n} \Psi \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n), \hat{Q}_W(P^1_{n,B_n}) \right).$$

Asymptotic results for CV-TMLE under the quassi-log-likelihood loss parallel those for the squared error loss function.

Theorem 4. *Consider the setting defined above.*

Suppose that $P_0(|Y| < 1) = 1$.

Let $B_n \in \{0, 1\}^n$ *be a random vector indicating a split of* $\{1, \ldots, n\}$ *into a training and validation sample. Suppose* B_n *is uniformly distributed on a finite support.*

Let $\hat{\bar{Q}}$ *and* $\hat{g}$ *be initial estimators of* $\bar{Q}_0$ *and* g_0. *In the following,* $\hat{\bar{Q}}(P_0)$ *and* $\hat{g}(P_0)$ *denote limits of these estimators, not necessarily equal to* $\bar{Q}_0$ *and* g_0, *respectively.*

The cross validated TMLE satisfies

$$\hat{\Psi}(P_n) - \psi_0 = E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right)$$
$$+E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})} \left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\}. \quad (2)$$

Consider the following definition:

$$\epsilon_0 \equiv \arg\min_{\epsilon} P_0 L(\hat{\bar{Q}}(P_0)(\epsilon)).$$

Suppose that this minimum exists and satisfies the derivative equation

$$0 = P_0 D_Y(P_0, \epsilon_0),$$

where

$$\begin{aligned} D_Y(P,\epsilon) &\equiv -\frac{d}{d\epsilon} L(O, \hat{\bar{Q}}(P)(\epsilon)) \\ &= \left(Y - \hat{\bar{Q}}(P)(\epsilon)\right) H^*_{\hat{g}(P)} \\ &= D^*_Y\left(\hat{\bar{Q}}(P)(\epsilon), \hat{g}(P)\right). \end{aligned}$$

If there are multiple minima, then it is assumed that the argmin is uniquely defined and selects one of these minima.

Suppose that $\hat{\bar{Q}}$ and $\hat{g}$ satisfy the following conditions:

1. *There exists a closed bounded set $K \subset \mathbf{R}^k$ containing ϵ_0 such that ϵ_n belongs to K with probability 1;*
2. *For some $\delta > 0$, $P(1 - \delta > \hat{g}(P_n)(1 \mid W) > \delta) = 1$;*
3. *For some $\gamma > 0$, $P(1 - \gamma|\hat{\bar{Q}}(P_n)(A, W)| < \gamma) = 1$;*
4. $$\int_W \left(\hat{g}(P_n)(1|w) - \hat{g}(P_0)(1|w)\right)^2 dQ_{W,0}(w) \to 0 \quad \textit{in probability;}$$
5. *For $a = 0, 1$,*
$$\int_W \left(\hat{\bar{Q}}(P_n)(a,w) - \hat{\bar{Q}}(P_0)(a,w)\right)^2 dQ_{W,0}(w) \to 0 \quad \textit{in probability.}$$

Then,

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\left\{D_Y^*\left(\hat{\bar{Q}}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + \hat{\bar{Q}}(P_0)(\epsilon_0)\right\}$$
$$+E_{B_n}P_0\left\{\frac{(-1)^{1+A}}{g_0\hat{g}(P^0_{n,B_n})}\left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\} \quad (3)$$
$$+o_P(1/\sqrt{n}).$$

Furthermore, If $\hat{g}(P_n) = g_0$*, the TMLE estimator* $\hat{\Psi}(P_n)$ *is asymptotically linear estimator of* ψ_0*:*

$$\hat{\Psi}(P_n) - \psi_0 \quad = \quad (P_n - P_0)\, D^*(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0)) + o_P(1/\sqrt{n}), \quad (4)$$

where $\hat{Q}(P_0)(\epsilon_0) = (\hat{\bar{Q}}(P_0)(\epsilon_0), Q_{W,0})$*.*

If, in addition to $\hat{g}(P_n) = g_0$*,* $\hat{\bar{Q}}(P_0) = \bar{Q}_0$*, which implies that* $\hat{\bar{Q}}(P_0)(\epsilon_0) = \bar{Q}_0$*, then* $\hat{\Psi}(P_n)$ *is an asymptotically efficient estimator of* ψ_0*:*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*(Q_0, g_0) + o_P(1/\sqrt{n}). \quad (5)$$

More generally, if the limits satisfy $\hat{g}(P_0) = g_0$ *and* $\hat{\bar{Q}}(P_0) = \bar{Q}_0$*, and if the convergence satisfies*

$$\sqrt{E_{B_n}P_0\left(\frac{g_0 - \hat{g}(P^0_{n,B_n})}{g_0\hat{g}(P^0_{n,B_n})}\right)^2}\sqrt{E_{B_n}P_0\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \bar{Q}_0\right)^2}$$
$$= o_P(1/\sqrt{n}), \qquad (16)$$

then $\hat{\Psi}(P_n)$ *is an asymptotically efficient estimator of* ψ_0*:*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\, D^*(Q_0, g_0) + o_P(1/\sqrt{n}).$$

Consider now the case that $\hat{g}(P_0) = g_0$*, but* $\hat{\bar{Q}}(P_0) \neq \bar{Q}_0$*. If the convergence satisfies*

$$\sqrt{E_{B_n}P_0\left(\frac{g_0 - \hat{g}(P^0_{n,B_n})}{g_0\hat{g}(P^0_{n,B_n})}\right)^2}\sqrt{E_{B_n}P_0\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(P_0)(\epsilon_0)\right)^2}$$
$$= o_P(1/\sqrt{n}), \qquad (17)$$

and $P_0\left\{H^*_{\hat{g}(P_n)}\left(\hat{\bar{Q}}(P_0)(\epsilon_0) - \bar{Q}_0\right)\right\}$ *is an asymptotically linear estimator of* $P_0\left\{H^*_{\hat{g}(P_0)}\left(\hat{\bar{Q}}(P_0)(\epsilon_0) - \bar{Q}_0\right)\right\}$ *with influence curve* IC'*, then* $\hat{\Psi}(P_n)$ *is an asymptotically linear estimator of* ψ_0*:*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)\left\{D^*(\hat{Q}(P_0)(\epsilon_0), g_0) + IC'\right\} + o_P(1/\sqrt{n}).$$

The proof of this theorem follows the same steps as that of theorem 3. The two only differ in the proofs of some of the auxiliary lemmas. We state the following useful results. For convenience, we adopt the notation $C_P \equiv \frac{1-\hat{\bar{Q}}(P)}{\hat{\bar{Q}}(P)}$.

Lemma 9. *Suppose* $\hat{\bar{Q}}$ *is such that for some* $1 > \gamma > 0$, $P(1-\gamma > \hat{\bar{Q}}(P_n)(A,W) > \gamma) = 1$. *If for* $a = 0, 1$, $\hat{g}$ *satisfy*

$$\int_W \left(\hat{\bar{Q}}(a,w) - \hat{\bar{Q}}(P_0)(a,w)\right)^2 dQ_{W,0}(w) \xrightarrow{P} 0,$$

then

$$P_0 \left(C_{P_n} - C_{P_0}\right)^4 \xrightarrow{P} 0. \tag{27}$$

Lemma 10. *Suppose* $\hat{g}$ *and* $\hat{\bar{Q}}$ *satisfy the conditions 2-5 in theorem 4. Then, on each split of* B_n, *for any* $r \geq 1$,

1. $EP_0 \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right)^r \to 0;$
2. $EP_0 \left(C_{P^0_{n,B_n}} - C_{P_0}\right)^r \to 0.$

Proof of theorem 4.

The identity in (2) is a result of the properties of $\Psi(P)$, its canonical gradient, the definition of ϵ_n and the one-step convergence of $\hat{\bar{Q}}(P)(\epsilon_n)$. Therefore, identical arguments as in the proof of theorem 3 yield (2).

We may express the first summand of (2) as

$$\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \\
&= E_{B_n}\left(P^1_{n,B_n} - P_0\right)\left\{D_Y\left(P^0_{n,B_n}, \epsilon_n\right) - D_Y\left(P_0, \epsilon_0\right)\right\} \qquad (28) \\
&+ \left(P_n - P_0\right) D_Y\left(P_0, \epsilon_0\right) \\
&+ E_{B_n}\left(P^1_{n,B_n} - P_0\right)\left\{\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(P_0)(\epsilon_0)\right\} \qquad (29) \\
&+ \left(P_n - P_0\right) \hat{\bar{Q}}(P_0)(\epsilon_0).
\end{aligned}$$

It follows from the following lemma that ϵ_n converges to ϵ_0 in probability.

Lemma 11. *Let* ϵ_n *and* ϵ_0 *be defined as in theorem 4 and suppose they solve the derivative equations as stated in the theorem. If* $\hat{g}$ *and* $\hat{\bar{Q}}$ *satisfy the conditions 1-5 in theorem 4, then* ϵ_n *converges to* ϵ_0 *in probability.*

The following lemmas 12 and 13 now prove that (28) and (29) are $o_P(1/\sqrt{n})$.

Lemma 12. *If the initial estimators $\hat{\bar{Q}}$ and $\hat{g}$ satisfy conditions 1-5 in the theorem, then, on a sample split of B_n,*

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{D_Y\left(P^0_{n,B_n}, \epsilon_n\right) - D_Y\left(P_0, \epsilon_0\right)\right\} = o_P(1).$$

Lemma 13. *If $\hat{\bar{Q}}$ and $\hat{g}$ satisfy conditions 1-5 of the theorem, then, on each sample split,*

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(P_0)(\epsilon_0)\right) = o_P(1).$$

Lemmas 7 and 8 imply that (22) and (23) are $o_P(1/\sqrt{n})$.

We thus have established that

$$\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}(P_n, B_n)(\epsilon_n), \hat{g}(P^0_{n,B_n})\right) \\
&= (P_n - P_0)\left\{D^*_Y\left(\hat{\bar{Q}}(P_0)(\epsilon_0), \hat{g}(P_0)\right) + \hat{\bar{Q}}(P_0)(\epsilon_0)\right\} + o_P(1/\sqrt{n}).
\end{aligned}$$

Combining this result with the (2), we have (3).

Finally, we study the remainder term:

$$E_{B_n} P_0\left\{\frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{n,B_n})}\left(\bar{Q}_0 - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)\right)\left(g_0 - \hat{g}(P^0_{n,B_n})\right)\right\}.$$

Firstly note that if the initial estimator $\hat{\bar{Q}}$ is consistent for $\bar{Q}_0$, i.e. $\hat{\bar{Q}}(P_0) = \bar{Q}_0$, then $\epsilon = 0$ is a solution to the derivative equation $P_0 D_Y(\epsilon, P_0) = 0$. On the other hand, we have seen in the proof of lemma 11 that the derivative function is monotonic in ϵ. Hence, we have $\epsilon_0 = 0$ and $\hat{\bar{Q}}(P_0)(\epsilon_0)$ is simply Q_0. Now, identical arguments in the proof of theorem 3 complete the proof. □.

4.3 Discussion of conditions of Theorems 3 and 4.

Under no conditions, we determined an exact identity (2), which shows that the analysis of the CV-TMLE involves a cross-validated empirical process term applied to the efficient influence curve, and a second order remainder term. Our second result (3) replaces the cross-validated empirical process term by an empirical mean of mean zero random variables

$D^*(\hat{Q}(P_0)(\epsilon_0), \hat{g}(P_0))$ plus a negligible $o_P(1/\sqrt{n})$-term. That is, under essentially no conditions beyond the positivity assumption, the CV-TMLE minus the true ψ_0, behaves as an empirical mean of mean zero i.i.d. random variables (which thus converges to a normal distribution, by CLT), plus a specified second order remainder term.

The second order remainder term predicts immediately that to make it negligible we will need that the product of the rates of convergence for $\hat{\bar{Q}}(P_n)$ and $\hat{g}(P_n)$ to their targets $\bar{Q}_0$ and g_0 is $o(1/\sqrt{n})$. As mentioned before, in an RCT g_0 is known, so that one might set $\hat{g}(P_n) = g_0$, in which case the second order remainder term is exactly equal to zero, giving us the asymptotic linearity (4) of the CV-TMLE under no other conditions than the positivity assumption and convergence of $\hat{\bar{Q}}(P_n)$ to some fixed function. This teaches us in particular that in an RCT in which we use a consistent estimator $\hat{\bar{Q}}$ the CV-TMLE is asymptotically efficient, as stated in (5). That is, in an RCT, this theorem teaches us that CV-TMLE with adaptive estimation of $\bar{Q}_0$ is the way to go.

Let's now consider a study in which g_0 is not known, but one has available a correctly specified parametric model: for example, one knows that A is only a function of a discrete variable, and one uses a saturated model. If the initial estimator $\hat{\bar{Q}}$ is consistent for $\bar{Q}_0$, then the rate condition (16) holds, so that it follows that the CV-TMLE is asymptotically efficient. That is, in this scenario there is only benefit in using an adaptive estimator of $\bar{Q}_0$. If, by chance, the estimator $\hat{\bar{Q}}$ is actually inconsistent for $\bar{Q}_0$, then the rate condition (17) still holds, and the asymptotic linearity condition on $\hat{g}$ will also hold under minimal conditions, so that we still have that the CV-TMLE is asymptotically linear.

Finally, let's consider a case in which the assumed model for g_0 is a large semiparametric model. To have a chance of being consistent for g_0, one will need to utilize adaptive estimation to estimate g_0 such as a maximum likelihood based super learner respecting the semiparametric model. There are now two scenarios possible. Firstly, suppose that $\hat{\bar{Q}}$ converges to $\bar{Q}_0$ fast enough so that (16) holds. Then the CV-TMLE is asymptotically efficient. If, on the other hand, $\hat{\bar{Q}}$ converges fast enough to a misspecified $\bar{Q}$ so that (17) holds, then another condition is required. Namely, we now need that $\hat{g}$ is such that the smooth functional $\Phi_{P_0}(\hat{g})$, indexed by P_0, is an asymptotically linear estimator of its limit $\Phi_{P_0}(g_0)$. This smooth functional can be represented as $\Phi_{P_0}(g) = P_0 H_g^*(\bar{Q}^* - Y)$, where $\bar{Q}^* = \hat{\bar{Q}}(P_0)(\epsilon_0)$. A data adaptive estimator $\hat{g}$ of g_0, only tailored to fit g_0 as a whole, may be too biased for this smooth functional (the whole motivation of TMLE!).

Therefore, we suggest that the estimator $\hat{g}$ should be targeted towards this smooth functional. That is, one might want to work out a TMLE $\hat{g}^*$ that aims to target this parameter $\Phi_{P_0}(g_0)$. We leave this for future research.

5 The iterative targeted MLE using V-fold sample splitting.

For a given cross-validation scheme $B_n \in \{0,1\}^n$, we defined

$$\epsilon_n^0 = \hat{\epsilon}(P_n) = \arg\min_{\epsilon} E_{B_n} P_{n,B_n}^1 L(\hat{Q}(P_{n,B_n}^0)(\epsilon)).$$

This now yields an update $\hat{Q}(P_{n,B_n}^0)(\epsilon_n^0)$ of $\hat{Q}(P_{n,B_n}^0)$ for each split of B_n. One could now iterate this updating process of the training sample specific estimators: define $\hat{Q}^1(P_{n,B_n}^0) = \hat{Q}(P_{n,B_n}^0)(\epsilon_n^0)$,

$$\epsilon_n^1 = \arg\min_{\epsilon} E_{B_n} P_{n,B_n}^1 L(\hat{Q}^1(P_{n,B_n}^0)(\epsilon)),$$

resulting in another update $\hat{Q}^1(P_{n,B_n}^0)(\epsilon_n^1)$ for each B_n. This process is iterated till $\epsilon_n^K = 0$ (or close enough to zero). We denote the k-step estimator $\hat{Q}^{k-1}(P)(\epsilon_n^{k-1})$ as $\hat{Q}(P)(\vec{\epsilon_n}^k)$ to reminds us that it is a function of the initial estimators $\hat{Q}$, $\hat{g}$ and the fluctuation vector $\vec{\epsilon_n}^k \equiv \left(\epsilon_n^0, \ldots, \epsilon_n^{k-1}\right)$. We denote the k-step TMLE of ψ_0 as

$$\hat{\Psi}^k(P_n) \equiv E_{B_n} \Psi(\hat{Q}(P_{n,B_n}^0)(\vec{\epsilon_n}^k)).$$

The final update will be denoted with $\hat{Q}(P_{n,B_n}^0)(\vec{\epsilon_n}^*)$ for each split B_n. The targeted MLE is now defined as $\hat{\Psi}^*(P_n) = E_{B_n} \Psi(\hat{Q}(P_{n,B_n}^0)(\vec{\epsilon_n}^*))$. We assume that, due to the derivative condition, $\frac{d}{d\epsilon} L(\hat{Q}(P_n)(\epsilon))\Big|_{\epsilon=0} = D^*(\hat{Q}(P_n), \hat{g}(P_n))$, we have

$$0 = E_{B_n} P_{n,B_n}^1 D^*(\hat{Q}(P_{n,B_n}^0)(\vec{\epsilon_n}^*), \hat{g}(P_{n,B_n}^0)).$$

We note that $\hat{Q}(P)(\vec{\epsilon_n}^*)$ is itself dependent on the data through the iterative sequence of selected ϵ's: $\epsilon_n^0, \ldots, \epsilon_n^K$.

We are now ready to present the asymptotics of the k-step cross validated TMLE.

Theorem 5. *Let $\hat{Q}(P_n)$, $\hat{g}(P_n)$ be initial estimators of Q_0, g_0, respectively, and we will denote their limits with $\hat{Q}(P_0)$ and $\hat{g}(P_0)$, which are not necessarily Q_0 and g_0, respectively.*

Uniformly bounded loss function: *We assume that* $\{\hat{Q}(P_n)(\epsilon) : \epsilon\} \in \mathcal{Q}$ *with probability 1, and that the loss function* $L(Q)$ *for* Q_0 *is uniformly bounded in* $Q \in \mathcal{Q}$*, and over a support of* $O \sim P_0$*:*

$$M_1 = \sup_Q \sup_O \mid L(Q)(O) \mid < \infty.$$

Let $B_n \in \{0,1\}^n$ *be a random vector indicating a split of* $\{1,\ldots,n\}$ *into a training and validation sample. Suppose* B_n *is uniformly distributed over a finite support.*

Suppose there exists $k_n = \hat{k}(P_n) > 0$ *such that* $P(\hat{k}(P_n) \leq k_0) \to 1$ *for some* $k_0 \equiv k(P_0)$ *and*

$$E_{B_n} P^1_{n,B_n} D^* \left(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n}) \right) = o_P(1/\sqrt{n}). \tag{30}$$

Consider a k_0*-dimensional random vector* $\vec{\epsilon_n}^{k_0} \equiv \left(\vec{\epsilon_n}^{k_n}, a_0, \ldots, a_0 \right)$*, where* a_0 *is a constant that depends on the choice of the parametric working model such that* $\hat{Q}(P)(\vec{\epsilon_n}^{k_0}) = \hat{Q}(P)(\vec{\epsilon_n}^{k_n})$*. (e.g.* $a_0 = 0$ *in most cases) Note that* $\vec{\epsilon_n}^{k_n}$ *is a projection of* $\vec{\epsilon_n}^{k_0}$ *onto its first* k_n *coordinates.*

If parameter $P \to \Psi(Q(P))$ *satisfies*

A1:

$$\Psi(Q(P)) - \Psi(Q_0) = -P_0 D^*(Q(P), g_0) + O_P(\parallel \Psi(Q(P)) - \Psi(Q_0) \parallel^2).$$

Then

$$\begin{aligned} \hat{\Psi}^{k_n}(P_n) - \psi_0 &= E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) \\ &+ E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), g_0) \right\} \\ &- E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\ &+ o_P(1/\sqrt{n}) + O_P(\parallel \hat{\Psi}^{k_n}(P_n) - \psi_0 \parallel^2). \end{aligned} \tag{31}$$

Let $\vec{\epsilon_0}^{k_0}$ *denote the limit of* $\vec{\epsilon_n}^{k_0}$ *as* $n \to \infty$*, that is,* $\parallel \vec{\epsilon_n}^{k_0} - \vec{\epsilon_0}^{k_0} \parallel \xrightarrow{P} 0$*. Suppose the following assumption also holds*

A2: (Given $\parallel \vec{\epsilon_n}^{k_0} - \vec{\epsilon_0}^{k_0} \parallel \xrightarrow{P} 0$*,)*

Define the class of functions

$$\mathcal{F}(P^0_{n,v}) \equiv \{O \to D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0)) : \vec{\epsilon}\},$$

where the set over which $\vec{\epsilon}$ varies is chosen so that it is a subset of $\mathbf{R}^{k_0}$ and contains $\vec{\epsilon_n}^{k_0}$ with probability tending to 1. In addition, for a deterministic sequence δ_n converging to zero as $n \to \infty$, we also define the sequence of sub-classes

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) \equiv \left\{ f_\epsilon \in \mathcal{F}(P^0_{n,B_n}) : \| \vec{\epsilon} - \vec{\epsilon_0}^{k_0} \| < \delta_n \right\}.$$

Assume that for deterministic sequence δ_n converging to 0, we have

$$EEntro(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0 F^2(\delta_n, P^0_{n,B_n})} \to 0 \text{ as } n \to \infty,$$

where $F(\delta_n, P^0_{n,B_n})$ is the envelope of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$.

Then we can write $\hat{\Psi}^{k_n}(P_n) - \psi_0$ as:

$$\begin{aligned}
\hat{\Psi}^{k_n}(P_n) - \psi_0 &= (P_n - P_0) D^* \left(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0) \right) + o_P(1/\sqrt{n}) \\
&+ E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), g_0) \right\} \\
&- E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\
&+ O_P(\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \|^2). \qquad (32)
\end{aligned}$$

Furthermore, suppose $\hat{g}(P_n) = g_0$. Then

$$\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0) D^* \left(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), g_0 \right) + o_P(1/\sqrt{n}).$$

If, in addition to $\hat{g}(P_n) = g_0$, we also have $\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}) = Q_0$, then $\hat{\Psi}^{k_n}(P_n)$ is in fact asymptotically efficient

$$\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0) D^*(Q_0, g_0) + o_P(1/\sqrt{n}).$$

More generally, suppose $\hat{g}(P_0) = g_0$. Let $\tilde{Q}$ denote the limit of $\hat{Q}(P_n)(\vec{\epsilon_n}^{k_n})$ which is not necessarily Q_0. Assume in addition

A3:

$$\begin{aligned}
&E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), g_0) \right\} \\
&- E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&= o_P(1/\sqrt{n}).
\end{aligned}$$

A4: For for some mean zero function $IC'(P_0) \in L_0^2(P_0)$, we have

$$\begin{aligned}
&E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&- E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\
&= (P_n - P_0)\, IC'(P_0) + o_P(1/\sqrt{n}).
\end{aligned}$$

NOTE: *If $\hat{Q}(P_n)(\vec{\epsilon_n}^{k_n})$ converges to Q_0 then A5 is automatically true with $IC' \equiv 0$.*

Then $\hat{\Psi}^{k_n}(P_n)$ is asymptotically linear

$$\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0) \left\{ D^* \left(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), g_0 \right) + IC'(P_0) \right\} + o_P(1/\sqrt{n}).$$

Proof of Theorem 5:
From definition of k_n, we have that

$$E_{B_n} P^1_{n,B_n} D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) = o_P(1/\sqrt{n}).$$

Combining this result with A1 and the double robustness of D^*, which guarantees $P_0 D^*(Q_0, g) = 0$ for all g, we readily have (31):

$$\begin{aligned}
&\hat{\Psi}^{k_n}(P_n) - \psi_0 = E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) && (33) \\
&+ E_{B_n} P_0 \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), g_0) \right\} && (34) \\
&- E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} && (35) \\
&+ o_P(1/\sqrt{n}) + O_P(\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \|^2).
\end{aligned}$$

We may rewrite (33) as

$$\begin{aligned}
&E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D^*(\hat{Q}(\vec{\epsilon_n}^{k_n}) P^0_{n,B_n}), \hat{g}(P^0_{n,B_n})) \\
&= E_{B_n} \left(P^1_{n,B_n} - P_0 \right) \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_0}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0)) \right\} \\
&+ E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0))
\end{aligned}$$

An application of A2 and lemma 2, combined with the fact that B_n is uniformly distributed over a finite support, we have

$$\begin{aligned}
&E_{B_n} \left(P^1_{n,B_n} - P_0 \right) \left\{ D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_0}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0)) \right\} \\
&= o_P(1/\sqrt{n}).
\end{aligned}$$

In other words, the term (33) is given by

$$E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n}))$$
$$= E_{B_n}\left(P^1_{n,B_n} - P_0\right) D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0)) + o_P(1/\sqrt{n}).$$

This result and the established equality in (31) now prove (32).

Now, if $\hat{g}(P_n) = g_0$, then the (34) and (35) are exactly 0. Consequently, (32) becomes

$$\begin{aligned}\hat{\Psi}^{k_n}(P_n) - \psi_0 &= (P_n - P_0)\, D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), g_0) \\ &+ o_P(1/\sqrt{n}) + O_P(\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \|^2).\end{aligned}$$

However, note that taking $\|\|$ on both sides of the equality above yields $\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \| = o_P(1/\sqrt{n})$. We thereby have asymptotically linearity of $\hat{\Psi}^{k_n}(P_n)$:

$$\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0)\, D^*(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), g_0) + o_P(1/\sqrt{n}).$$

If, in addition, $\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}) = Q_0$, then the influence curve is indeed the efficient influence curve $D^*(Q_0, g_0)$.

Next we consider a more general case where $\hat{g}(P_0) = g_0$. Let $\tilde{Q}$ be the limit of $\hat{Q}(P_n)(\vec{\epsilon_n}^{k_n})$. It is not necessarily the case that $\tilde{Q} = Q_0$. We now rewrite the established equality (32) to account for $\tilde{Q}$:

$$\begin{aligned}&\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0)\, D^*\left(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0)\right) + o_P(1/\sqrt{n}) \\ &+E_{B_n}P_0\left\{D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), g_0)\right\} \\ &-E_{B_n}P_0\left\{D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0)\right\} \\ &+E_{B_n}P_0\left\{D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0)\right\} \\ &-E_{B_n}P_0\left\{D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0)\right\} \\ &+O_P(\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \|^2).\end{aligned}$$

From A3, the term

$$\begin{aligned}&E_{B_n}P_0\left\{D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon_n}^{k_n}), g_0)\right\} \\ &-E_{B_n}P_0\left\{D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0)\right\} \\ &= o_P(1/\sqrt{n}).\end{aligned}$$

From A4, the term

$$
\begin{aligned}
&E_{B_n} P_0 \left\{ D^*(\tilde{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\tilde{Q}, g_0) \right\} \\
&- E_{B_n} P_0 \left\{ D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0) \right\} \\
&= (P_n - P_0) IC'(P_0) + o_P(1/\sqrt{n}).
\end{aligned}
$$

Therefore (3) becomes

$$
\begin{aligned}
&\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0) \left\{ D^* \left(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0) \right) + IC'(P_0) \right\} + o_P(1/\sqrt{n}) \\
&+ O_P(\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \|^2).
\end{aligned}
$$

Taking $\|\|$ on both sides again yields $\| \hat{\Psi}^{k_n}(P_n) - \psi_0 \| = o_P(1/\sqrt{n})$. We thereby have the desired result

$$
\hat{\Psi}^{k_n}(P_n) - \psi_0 = (P_n - P_0) \left\{ D^* \left(\hat{Q}(P_0)(\vec{\epsilon_0}^{k_0}), \hat{g}(P_0) \right) + IC'(P_0) \right\} + o_P(1/\sqrt{n}).
$$

□

6 Concluding remarks.

We presented a TMLE that allows to learn the truth ψ_0, while also providing statistical inference based on an CLT, under an as large statistical model as possible. For that purpose, the combination of adaptive estimation (super learning), targeted maximum likelihood estimation, and cross-validated selection of the fluctuation parameter in the TMLE, are essential tools to achieve this goal.

In future work we wish to investigate the extension of CV-TMLE to collaborative targeted maximum likelihood estimation, as in van der Laan and Gruber (2010), and the incorporation of targeted estimators of g_0 to enhance the asymptotic linearity of the CV-TMLE of ψ_0 for the case that the initial estimator of Q_0 is inconsistent.

7 Appendix

Proof of lemma 2: Let $G^1_{n,B_n} = \sqrt{n}(P^1_{n,B_n} - P_0)$. For any $\delta > 0$.

$$
\begin{aligned}
&P\left(|G^1_{n,B_n} f_{\epsilon_n}(P^0_{n,B_n})| > \delta\right) = EP\left(|G^1_{n,B_n} f_{\epsilon_n}(P^0_{n,B_n})| > \delta \middle| P^0_{n,B_n}\right) \\
&= EP\left(\left|G^1_{n,B_n} f_{\epsilon_n}(P^0_{n,B_n}) I(\parallel \epsilon_n - \epsilon_0 \parallel < \delta_n)\right| > \delta \middle| P^0_{n,B_n}\right) \\
&+EP\left(\left|G^1_{n,B_n} f_{\epsilon_n}(P^0_{n,B_n}) I(\parallel \epsilon_n - \epsilon_0 \parallel \geq \delta_n)\right| > \delta \middle| P^0_{n,B_n}\right) \\
&\leq EP\left(sup_{f \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})} \left|G^1_{n,B_n} f\right| > \delta \middle| P^0_{n,B_n}\right) \\
&+EP\left(\parallel \epsilon_n - \epsilon_0 \parallel \geq \delta_n \middle| P^0_{n,B_n}\right) \\
&= EP\left(sup_{f \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})} \left|G^1_{n,B_n} f\right| > \delta \middle| P^0_{n,B_n}\right) \\
&+P\left(\parallel \epsilon_n - \epsilon_0 \parallel \geq \delta_n\right).
\end{aligned}
$$

By our assumption, $P\left(\parallel \epsilon_n - \epsilon_0 \parallel \geq \delta_n\right) \to 0$. On the other hand, by Chebysev inequality, lemma 1 and Cauchy-Schwartz inequality

$$
\begin{aligned}
&EP\left(sup_{f \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})} \left|G^1_{n,B_n} f\right| > \delta \middle| P^0_{n,B_n}\right) \\
&\leq \frac{1}{\delta} EE\left(sup_{f \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})} \left|G^1_{n,B_n} f\right|\right) \\
&\leq \frac{1}{\delta} EEntro\left(\mathcal{F}_{\delta_n}(P^0_{n,B_n})\right) \sqrt{P_0 F(\delta_n, P^0_{n,B_n})^2}.
\end{aligned}
$$

Therefore, $G^1_{n,B_n} f_{\epsilon_n}(P^0_{n,B_n}) \xrightarrow{P} 0$ by our assumption. □

Proof of lemma 3:

By our assumption, $P(|X_n - X| < 2A) = 1$. Then for any $\delta > 0$,

$$
\begin{aligned}
E|X_n - X|^r &= E\left\{|X_n - X|^r I_{|X_n - X| \leq \delta}\right\} + E\left\{|X_n - X|^r I_{|X_n - X| > \delta}\right\} \\
&\leq \delta^r P(|X_n - X| \leq \delta) + (2A)^r P(|X_n - X| > \delta) \\
&= \delta^r \cdot 1 + \left((2A)^r - \delta^r\right) P(|X_n - X| > \delta).
\end{aligned}
$$

We assumed that $P(|X_n - X| > \delta) \to 0$. Hence the last equality converges to δ^r. This holds for all $\delta > 0$. Thus we must have $E|X_n - X|^r \to 0$. □

Proof of lemma 4:

First note that

$$H^*_{\hat{g}(P_n)}(A,W) - H^*_{\hat{g}(P_0)}(A,W) = (-1)^{A+1} \frac{\left(\hat{g}(P_0)(A|W) - \hat{g}(P^0_{n,B_n})(A|W)\right)}{\hat{g}(P_0)(A|W)\hat{g}(P^0_{n,B_n})(A|W)}. \tag{36}$$

The expression $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^4$ can be expanded into

$$P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^4 = \sum_{a=0,1} E_{W,0}\left(H^*_{\hat{g}(P_n)}(a,W) - H^*_{\hat{g}(P_0)}(a,W)\right)^4 g_0(a|W).$$

Applying Cauchy-Schwartz and (36), each summand can be bounded as follows:

$$\begin{aligned}
&E_{W,0}\left(H^*_{\hat{g}(P_n)}(a,W) - H^*_{\hat{g}(P_0)}(a,W)\right)^4 g_0(a|W) \\
&\leq \sqrt{E_{W,0}\left(\frac{g_0(a|W)}{\left(\hat{g}(P_0)\hat{g}(P_n)(a|W)\right)^4}\right)^2} \sqrt{E_{W,0}\left\{\hat{g}(P_0)(a|W) - \hat{g}(P_n)(a|W)\right\}^8} \\
&\leq \sqrt{E_{W,0}\left(\frac{g_0(a|W)}{\left(\hat{g}(P_0)\hat{g}(P_n)(a|W)\right)^4}\right)^2} \sqrt{E_{W,0}\left\{\hat{g}(P_0)(a|W) - \hat{g}(P_n)(a|W)\right\}^2}.
\end{aligned}$$

Since $E_{W,0}\left(\frac{g_0(a|W)}{\left(\hat{g}(P_0)\hat{g}(P_n)(a|W)\right)^4}\right)^2$ is bounded and, by assumption,

$$E_{W,0}\left(\hat{g}(P_0)(a|W) - \hat{g}(P_n)(a|W)\right)^2 \xrightarrow{P} 0$$

this inequality implies that

$$P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^4 \xrightarrow{P} 0. \tag{37}$$

To prove $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^2 \xrightarrow{P} 0$, we use a simple application of Cauchy-Schwartz inequality and (37). Similarly for $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right) \xrightarrow{P} 0$.

Finally, to verify $P_0\left({H^*_{\hat{g}(P_n)}}^2 - {H^*_{\hat{g}(P_0)}}^2\right) \xrightarrow{P} 0$, we first bound the expectation using Cauchy-Schwartz inequality:

$$\begin{aligned}
&\left|P_0\left({H^*_{\hat{g}(P_n)}}^2 - {H^*_{\hat{g}(P_0)}}^2\right)\right| \\
&\leq \sqrt{P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^2} \sqrt{P_0\left(H^*_{\hat{g}(P_n)} + H^*_{\hat{g}(P_0)}\right)^2}.
\end{aligned}$$

By assumption, $P_0\left(H^*_{\hat{g}(P_n)} + H^*_{\hat{g}(P_0)}\right)^2$ is bounded. We established above that $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^2 \xrightarrow{P} 0$. Thus, the above inequality implies that $P_0\left({H^*_{\hat{g}(P_n)}}^2 - {H^*_{\hat{g}(P_0)}}^2\right) \xrightarrow{P} 0$. □

Proof of lemma 5:

1. Firstly, note that

$$\begin{aligned}
&E_{B_n}P_0\left(\hat{\bar{Q}}(P^0_{n,B_n})H^*_{\hat{g}(P^0_{n,B_n})} - \hat{\bar{Q}}(P_0)H^*_{\hat{g}(P_0)}\right)\\
&= E_{B_n}P_0\hat{\bar{Q}}(P^0_{n,B_n})\left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right)\\
&+E_{B_n}P_0H^*_{\hat{g}(P_0)}\left(\hat{\bar{Q}}(P^0_{n,B_n}) - \hat{\bar{Q}}(P_0)\right).
\end{aligned}$$

   By our assumptions, $P_0\hat{\bar{Q}}(P_n)^2$ is bounded with probability 1. Hence, it follows from Cauchy-Schwartz inequality and lemma 4 that the first summand converges to 0 in probability. On the other hand, we assumed that $P_0{H^*_{\hat{g}(P_0)}}^2$ is bounded and $P_0\left(\hat{\bar{Q}}(P_n) - \hat{\bar{Q}}(P_0)\right)^2 \xrightarrow{P} 0$. Therefore, it follows from an application of Cauchy-Schwartz inequality that the second summand also converge to 0. From these two results it follows that

$$E_{B_n}P_0\left\{\left(\hat{\bar{Q}}(P^0_{n,B_n})H^*_{\hat{g}(P^0_{n,B_n})} - \hat{\bar{Q}}(P_0)H^*_{\hat{g}(P_0)}\right)\right\} \xrightarrow{P} 0. \tag{38}$$

   Secondly, note also that our assumptions imply that $E_{B_n}P_0\hat{\bar{Q}}(P^0_{n,B_n})H^*_{\hat{g}(P^0_{n,B_n})}$ is bounded with probability 1. Hence, by lemma 3 we have obtain the desired result.

2. Firstly, note that

$$\begin{aligned}
&E_{B_n}P_0\left((Y - \hat{\bar{Q}}(P^0_{n,B_n}))H^*_{\hat{g}(P^0_{n,B_n})} - (Y - \hat{\bar{Q}}(P_0))H^*_{\hat{g}(P_0)}\right)\\
&= E_{B_n}P_0\left\{Y\left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right)\right\}\\
&-E_{B_n}P_0\left\{\left(\hat{\bar{Q}}(P^0_{n,B_n})H^*_{\hat{g}(P^0_{n,B_n})} - \hat{\bar{Q}}(P_0)H^*_{\hat{g}(P_0)}\right)\right\}.
\end{aligned}$$

   By our assumption, P_0Y^2 is bounded. Hence, it follows from Cauchy-Schwartz inequality and lemma 4 that $E_{B_n}P_0\left\{Y\left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right)\right\}$ converges to 0 in probability. Combining this result and (38), we have

$$E_{B_n}P_0\left((Y - \hat{\bar{Q}}(P^0_{n,B_n}))H^*_{\hat{g}(P^0_{n,B_n})} - (Y - \hat{\bar{Q}}(P_0))H^*_{\hat{g}(P_0)}\right) \xrightarrow{P} 0.$$

On the other hand, by our assumption, $E_{B_n} P_0 (Y - \hat{\bar{Q}}(P^0_{n,B_n})) H^*_{\hat{g}(P^0_{n,B_n})}$ is bounded with probability 1. Hence, an application of lemma 3 yields the desired result.

3. By our assumption, $E_{B_n} P_0 {H^*_{\hat{g}(P^0_{n,B_n})}}^2$ is bounded with probability 1. Hence, by lemma 4 and lemma 3, we have $EP_0 \left({H^*_{\hat{g}(P^0_{n,B_n})}}^2 - {H^*_{\hat{g}(P_0)}}^2 \right)^r \to 0$ for any $r \geq 1$.

4. Similarly, by our assumption, $E_{B_n} P_0 H^*_{\hat{g}(P^0_{n,B_n})}$ is bounded with probability 1. Hence by lemma 4 and lemma 3, we have

$$EP_0 \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)} \right)^r \to 0$$

for any $r \geq 1$.

□

Proof of lemma 6:

By our definition of ϵ_n and the one-step convergence of the fluctuation model,

$$E_{B_n} P^1_{n,B_n} D_Y(P^0_{n,B_n}, \epsilon_n) = 0.$$

This implies that

$$\begin{aligned} -P_0 D_Y(P_0, \epsilon_n) &= E_{B_n} \left(P^1_{n,B_n} - P_0 \right) \left\{ D_Y(P^0_{n,B_n}, \epsilon_n) - D_Y(P_0, \epsilon_n) \right\} && (39) \\ &+ E_{B_n} P_0 \left\{ D_Y(P^0_{n,B_n}, \epsilon_n) - D_Y(P_0, \epsilon_n) \right\} && (40) \\ &+ E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D_Y(P_0, \epsilon_n) && (41) \end{aligned}$$

The term (40) can be expanded into

$$\begin{aligned} & E_{B_n} P_0 \left\{ D_Y(P^0_{n,B_n}, \epsilon_n) - D_Y(P_0, \epsilon_n) \right\} \\ &= E_{B_n} P_0 \left\{ Y \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)} \right) \right\} \\ &- E_{B_n} P_0 \left\{ \left(\hat{\bar{Q}}(P^0_{n,B_n}) H^*_{\hat{g}(P^0_{n,B_n})} - \hat{\bar{Q}}(P_0) H^*_{\hat{g}(P_0)} \right) \right\} \\ &- \epsilon_n E_{B_n} P_0 \left\{ \left({H^*_{\hat{g}(P^0_{n,B_n})}}^2 - {H^*_{\hat{g}(P_0)}}^2 \right) \right\}. \end{aligned}$$

Note that ϵ_n is bounded with probability 1. Therefore, applying the arguments in the proof of lemma 5 to the corresponding summands, we have that (40) converges to 0 in probability.

The term (41) can be written as

$$\begin{aligned}
&E_{B_n}\left(P^1_{n,B_n}-P_0\right)D_Y(P_0,\epsilon_n)\\
&\equiv E_{B_n}\left(P^1_{n,B_n}-P_0\right)(Y-\hat{\bar{Q}}(P_0)-\epsilon_n H^*_{\hat{g}(P_0)})H^*_{\hat{g}(P_0)}\\
&=E_{B_n}\left(P^1_{n,B_n}-P_0\right)YH^*_{\hat{g}(P_0)}-E_{B_n}\left(P^1_{n,B_n}-P_0\right)\hat{\bar{Q}}(P_0)H^*_{\hat{g}(P_0)}\\
&-\epsilon_n E_{B_n}\left(P^1_{n,B_n}-P_0\right){H^*_{\hat{g}(P_0)}}^2.
\end{aligned}$$

All the empirical differences in the last equality are asymptotically normal with mean 0, and ϵ_n is bounded with probability 1. Therefore, we have that (41) converges to 0 in probability.

It remains to show that (39) converges to 0 in probability. By our assumption, there exists constant $M>0$ such that $P(|\epsilon_n|<M)=1$. Conditional on P^0_{n,B_n}, consider the class

$$\mathcal{F}(P^0_{n,B_n})=\left\{f_\epsilon(P^0_{n,B_n})=D_Y(P^0_{n,B_n},\epsilon)-D_Y(P_0,\epsilon):|\epsilon|<M\right\}.$$

Lemma 1 implies that

$$\sqrt{n}E\left(sup_{f\in\mathcal{F}(P^0_{n,B_n})}|(P^1_{n,B_n}-P_0)f|\right)\leq Entro(\mathcal{F}(P^0_{n,B_n}))\sqrt{P_0\mathbf{F}(P^0_{n,B_n})^2},$$

where $\mathbf{F}(P^0_{n,B_n})$ is an envelope of $\mathcal{F}(P^0_{n,B_n})$. Therefore, after an application of Chebysev inequality we may write

$$\begin{aligned}
&P\left(|(P^1_{n,B_n}-P_0)f_{\epsilon_n}(P^0_{n,B_n})|>\delta\right)\\
&\leq EP\left(sup_{f\in\mathcal{F}(P^0_{n,B_n})}|(P^1_{n,B_n}-P_0)f|>\delta\middle|P^0_{n,B_n}\right)\\
&\leq\frac{1}{\delta}EE\left(sup_{f\in\mathcal{F}(P^0_{n,B_n})}|(P^1_{n,B_n}-P_0)f|\right)\\
&\leq\frac{1}{\sqrt{n}}\frac{1}{\delta}EEntro(\mathcal{F}(P^0_{n,B_n}))\sqrt{P_0\mathbf{F}(P^0_{n,B_n})^2}.
\end{aligned}$$

Firstly note that f_ϵ is bounded per our assumptions. Hence $\sqrt{P_0\mathbf{F}(P^0_{n,B_n})^2}$ is bounded. On the other hand, the entropy of the class is also bounded. Therefore, we indeed have $P\left(|(P^1_{n,B_n}-P_0)f(\epsilon_n)(P_0)|>\delta\right)$ converges to 0 as $n\to\infty$. Consequently, (39) converges to 0 in probability.

Since K is compact, there is a subsequence ϵ_{nk} such that $\epsilon_{nk}\xrightarrow{P}\epsilon^*$ for some $\epsilon^*\in K$. This implies that for

$$\begin{aligned}
g(\epsilon)&\equiv P_0D_Y(P_0,\epsilon)\\
&=P_0YH^*_{\hat{g}(P_0)}-P_0\hat{\bar{Q}}(P_0)H^*_{\hat{g}(P_0)}-\epsilon P_0{H^*_{\hat{g}(P_0)}}^2,
\end{aligned}$$

which is continuous over K, we must have $g(\epsilon_{nk}) \xrightarrow{P} g(\epsilon^*)$.

We determined in above that $g(\epsilon_n) \xrightarrow{P} 0$, therefore it follows that $g(\epsilon^*) = 0$. On the other hand, by definition of ϵ_0 we have that $g(\epsilon_0) = 0$. Since $g(\epsilon)$ is a linear function in ϵ, it has unique solution at ϵ_0, therefore we indeed have $\epsilon^* = \epsilon_0$. This implies that all convergent subsequences of ϵ_n converge to ϵ_0 in probability. Since K is compact, it now implies that ϵ_n converge to ϵ_0 in probability. □

Proof of lemma 7: Conditional on P^0_{n,B_n}, for a deterministic sequence δ_n converging to 0, consider the class

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) \equiv \left\{ D_Y\left(P^0_{n,B_n}, \epsilon\right) - D_Y\left(P_0, \epsilon_0\right) : \| \epsilon - \epsilon_0 \| < \delta_n \right\}.$$

From lemma 6, we know that $\| \epsilon_n - \epsilon_0 \| \xrightarrow{P} 0$. To obtain the proposed result, we will show that this class satisfies the conditions of lemma 2.

For convenience, let $A_{P^0_{n,B_n}}(O) \equiv \left(Y - \hat{\bar{Q}}(P^0_{n,B_n})(A,W)\right) H^*_{\hat{g}(P^0_{n,B_n})}(A,W)$, $H_{P^0_{n,B_n}}(O)^2 \equiv {H^*_{\hat{g}(P^0_{n,B_n})}}^2$, and $A_{P_0}, H^2_{P_0}$ denote the analogous functions trained at P_0. Then, we can find an envelope for this class of functions as follows:

$$\begin{aligned}
&\left| D_Y\left(P^0_{n,B_n}, \epsilon\right) - D_Y\left(P_0, \epsilon_0\right) \right| \equiv \left| (A_{P^0_{n,B_n}} - \epsilon H^2_{P^0_{n,B_n}}) - (A_{P_0} - \epsilon_0 H^2_{P_0}) \right| \\
&= \left| (A_{P^0_{n,B_n}} - A_{P_0}) - H^2_{P^0_{n,B_n}}(\epsilon - \epsilon_0) - \epsilon_0 (H^2_{P^0_{n,B_n}} - H^2_{P_0}) \right| \\
&\leq |A_{P^0_{n,B_n}} - A_{P_0}| + |H^2_{P^0_{n,B_n}}| \delta_n + \epsilon_0 |H^2_{P^0_{n,B_n}} - H^2_{P_0}| \\
&\equiv \mathbf{F}_n.
\end{aligned}$$

Now, we study the convergence of $EP_0(\mathbf{F}_n)^2$. From the proposed conditions and lemma 5, we readily have that:

$$EP_0(A_{P^0_{n,B_n}} - A_{P_0})^2 \to 0,$$

and

$$EP_0\left\{ \epsilon_0^2 (H^2_{P^0_{n,B_n}} - H^2_{P_0})^2 \right\} = \epsilon_0^2 EP_0 (H^2_{P^0_{n,B_n}} - H^2_{P_0})^2 \to 0.$$

On the other hand, the boundedness conditions for $\hat{g}$ imply that $EP_0 H^4_{P^0_{n,B_n}}$ is bounded. Since δ_n converges to 0, this now implies that

$$EP_0\left\{ (H_{P^0_{n,B_n}})^4 \delta_n^2 \right\} \to 0.$$

Thus, all the square terms of $EP_0(\mathbf{F}_n)^2$ converge to 0 as $n \to \infty$. Applying Cauchy-Schwartz inequality and lemma 5 in a similar manner to the cross terms of $EP_0(\mathbf{F}_n)^2$ will show that they also converge to 0.

Moreover, this class has bounded entropy since the functions are linear in ϵ. Therefore, lemma 2 implies that we indeed have the desired result:

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{D_Y\left(P^0_{n,B_n}, \epsilon_n\right) - D_Y\left(P_0, \epsilon_0\right)\right\} = o_P(1).$$

□

Proof of lemma 8:

This result can be proved in a similar manner as lemma 7 by making use lemma 2 and the conditions of the theorem. □

Proof of lemma 9: Firstly, rewrite

$$C_{P_n} - C_{P_0} \equiv \frac{\hat{\bar{Q}}(P_0) - \hat{\bar{Q}}(P_n)}{\hat{\bar{Q}}(P_n)\hat{\bar{Q}}(P_0)}.$$

Then

$$\begin{aligned} P_0\left(C_{P_n} - C_{P_0}\right)^4 &= P_0 \frac{\left(\hat{\bar{Q}}(P_0) - \hat{\bar{Q}}(P_n)\right)^4}{(\hat{\bar{Q}}(P_n)\hat{\bar{Q}}(P_0))^4} \\ &\leq \sqrt{P_0 \frac{1}{(\hat{\bar{Q}}(P_n)\hat{\bar{Q}}(P_0))^8}} \sqrt{P_0\left(\hat{\bar{Q}}(P_0) - \hat{\bar{Q}}(P_n)\right)^8} \\ &\leq \sqrt{P_0 \frac{1}{(\hat{\bar{Q}}(P_n)\hat{\bar{Q}}(P_0))^8}} \sqrt{P_0\left(\hat{\bar{Q}}(P_0) - \hat{\bar{Q}}(P_n)\right)^2}, \end{aligned}$$

where the last inequality follows from the assumption that $\hat{\bar{Q}}(P)$ are bounded between 0 and 1 with probability 1. This last expression converges to 0 in probability by our assumption. □.

Proof of lemma 10:

1. By our assumption, $E_{B_n} P_0 H^*_{\hat{g}(P^0_{n,B_n})}$ is bounded with probability 1. An application of Cauchy-Schwartz inequality and lemma 4 implies that $E_{B_n} P_0\left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right) \xrightarrow{P} 0$. It now follows from lemma 3 that
$$EP_0\left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)}\right)^r \to 0$$
for any $r \geq 1$.

2. Similarly, By our assumption, $E_{B_n} P_0 C_{P^0_{n,B_n}}$ is bounded with probability 1. An application of Cauchy-Schwartz inequality and lemma 9 implies that $E_{B_n} P_0 (C_{P^0_{n,B_n}} - C_{P_0}) \xrightarrow{P} 0$. Hence, lemma 3 yields

$$E P_0 \left(C_{P^0_{n,B_n}} - C_{P_0} \right)^r \to 0$$

for any $r \geq 1$.

□

Proof of Lemma 11: By our definition,

$$E_{B_n} P^1_{n,B_n} D_Y(P^0_{n,B_n}, \epsilon_n) = 0.$$

This implies that

$$\begin{aligned} -P_0 D_Y(P_0, \epsilon_n) &= E_{B_n} \left(P^1_{n,B_n} - P_0 \right) \left\{ D_Y(P^0_{n,B_n}, \epsilon_n) - D_Y(P_0, \epsilon_n) \right\} && (42) \\ &+ E_{B_n} P_0 \left\{ D_Y(P^0_{n,B_n}, \epsilon_n) - D_Y(P_0, \epsilon_n) \right\} && (43) \\ &+ E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D_Y(P_0, \epsilon_n). && (44) \end{aligned}$$

The term (43) can be expanded into

$$\begin{aligned} & E_{B_n} P_0 \left\{ D_Y(P^0_{n,B_n}, \epsilon_n) - D_Y(P_0, \epsilon_n) \right\} \\ = \; & E_{B_n} P_0 \left\{ \left(Y - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) H^*_{\hat{g}(P^0_{n,B_n})} - \left(Y - \hat{\bar{Q}}(\epsilon_n)(P_0) \right) H^*_{\hat{g}(P_0)} \right\} \\ = \; & E_{B_n} P_0 \left\{ \left(Y - \hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) \right) \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)} \right) \right\} \\ & - E_{B_n} P_0 \left\{ H^*_{\hat{g}(P_0)} \left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n) - \hat{\bar{Q}}(\epsilon_n)(P_0) \right) \right\}. \end{aligned}$$

By lemma 4, $E_{B_n} P_0 \left(H^*_{\hat{g}(P^0_{n,B_n})} - H^*_{\hat{g}(P_0)} \right)^2 \xrightarrow{P} 0$. Moreover, Y is bounded by assumption and $\hat{\bar{Q}}(P)(\epsilon)$ is bounded by construction. Hence, an application of Cauchy-Schwartz imply that the first summand of the last equality converges to zero in probability. On the other hand the second summand

can be bounded by

$$\begin{aligned}
&\left|E_{B_n}P_0\left\{H^*_{\hat{g}(P_0)}\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\hat{\bar{Q}}(\epsilon_n)(P_0)\right)\right\}\right| \\
&\leq \sqrt{E_{B_n}P_0{H^*_{\hat{g}(P_0)}}^2}\sqrt{E_{B_n}P_0\left(\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon_n)-\hat{\bar{Q}}(\epsilon_n)(P_0)\right)^2} \\
&= \sqrt{E_{B_n}P_0{H^*_{\hat{g}(P_0)}}^2}\sqrt{E_{B_n}P_0\frac{\left(C_{P_0}e^{-\epsilon_n H^*_{\hat{g}(P_0)}}-C_{P^0_{n,B_n}}e^{-\epsilon_n H^*_{\hat{g}(P^0_{n,B_n})}}\right)^2}{(1+C_{P_0}e^{-\epsilon_n H^*_{\hat{g}(P_0)}})^2(1+C_{P^0_{n,B_n}}e^{-\epsilon_n H^*_{\hat{g}(P^0_{n,B_n})}})^2}} \\
&\leq \sqrt{E_{B_n}P_0{H^*_{\hat{g}(P_0)}}^2}\sqrt{E_{B_n}P_0\left(C_{P_0}e^{-\epsilon_n H^*_{\hat{g}(P_0)}}-C_{P^0_{n,B_n}}e^{-\epsilon_n H^*_{\hat{g}(P^0_{n,B_n})}}\right)^2}.
\end{aligned}$$

By our assumption $E_{B_n}P_0{H^*_{\hat{g}(P_0)}}^2$ is bounded. We now wish to show

$$E_{B_n}P_0\left(C_{P^0_{n,B_n}}e^{-\epsilon_n H^*_{\hat{g}(P^0_{n,B_n})}}-C_{P_0}e^{-\epsilon_n H^*_{\hat{g}(P_0)}}\right)^2\xrightarrow{P}0.$$

Let $H_{P^0_{n,B_n}}\equiv H^*_{\hat{g}(P^0_{n,B_n})}$, $H_{P_0}\equiv H^*_{\hat{g}(P_0)}$. Firstly, note that by property of the exponential function for every (a,w) in the support, there is $Y_{P^0_{n,B_n}}(a,w)$ between $\epsilon_n H_{P^0_{n,B_n}}(a,w)$ and $\epsilon_n H_{P_0}(a,w)$ such that

$$\begin{aligned}
e^{\epsilon_n H_{P^0_{n,B_n}}(a,w)}-e^{\epsilon_n H_{P_0}(a,w)} &= e^{-\epsilon_n H_{P_0}(a,w)}\epsilon_n(H_{P^0_{n,B_n}}-H_{P_0})(a,w) \\
&+ \frac{e^{Y_{P^0_{n,B_n}}(a,w)}}{2}\epsilon_n^2(H_{P^0_{n,B_n}}-H_{P_0})^2(a,w).
\end{aligned}$$

Boundedness of ϵ_n and $H_{P^0_{n,B_n}}$ implies that $Y_{P^0_{n,B_n}}$ is also bounded with

probability 1 over the support. Therefore, we have:

$$
\begin{aligned}
&E_{B_n}P_0\left(C_{P^0_{n,B_n}}e^{-\epsilon_n H_{P^0_{n,B_n}}}-C_{P_0}e^{-\epsilon_n H_{P_0}}\right)^2\\
&=E_{B_n}P_0\left\{C_{P^0_{n,B_n}}(e^{-\epsilon_n H_{P^0_{n,B_n}}}-e^{-\epsilon_n H_{P_0}})+e^{-\epsilon_n H_{P_0}}(C_{P^0_{n,B_n}}-C_{P_0})\right\}^2\\
&=E_{B_n}P_0\Bigg\{C_{P^0_{n,B_n}}\left(e^{-\epsilon_n H_{P_0}}\epsilon_n(H_{P^0_{n,B_n}}-H_{P_0})+\frac{e^{Y_{P^0_{n,B_n}}}}{2}\epsilon_n^2(H_{P^0_{n,B_n}}-H_{P_0})^2\right)\\
&\quad+e^{-\epsilon_n H_{P_0}}(C_{P^0_{n,B_n}}-C_{P_0})\Bigg\}^2\\
&=E_{B_n}P_0C^2_{P^0_{n,B_n}}\left(e^{-\epsilon_n H_{P_0}}\epsilon_n(H_{P^0_{n,B_n}}-H_{P_0})+\frac{e^{Y_{P^0_{n,B_n}}}}{2}\epsilon_n^2(H_{P^0_{n,B_n}}-H_{P_0})^2\right)^2\\
&+2E_{B_n}P_0\Bigg\{C_{P^0_{n,B_n}}\left(e^{-\epsilon_n H_{P_0}}\epsilon_n(H_{P^0_{n,B_n}}-H_{P_0})+\frac{e^{Y_{P^0_{n,B_n}}}}{2}\epsilon_n^2(H_{P^0_{n,B_n}}-H_{P_0})^2\right)\times\\
&\quad e^{-\epsilon_n H_{P_0}}(C_{P^0_{n,B_n}}-C_{P_0})\Bigg\}\\
&+E_{B_n}P_0\left\{e^{-2\epsilon_n H_{P_0}}(C_{P^0_{n,B_n}}-C_{P_0})^2\right\}\\
&=E_{B_n}P_0C^2_{P^0_{n,B_n}}\left(e^{-2\epsilon_n H_{P_0}}\epsilon_n^2(H_{P^0_{n,B_n}}-H_{P_0})^2\right)+E_{B_n}P_0\frac{e^{2Y_{P^0_{n,B_n}}}}{4}\epsilon_n^4(H_{P^0_{n,B_n}}-H_{P_0})^4\\
&+2E_{B_n}P_0\left\{e^{-\epsilon_n H_{P_0}}\frac{e^{Y_{P^0_{n,B_n}}}}{2}\epsilon_n^3(H_{P^0_{n,B_n}}-H_{P_0})^3\right\}\\
&+2E_{B_n}P_0\Bigg\{C_{P^0_{n,B_n}}\left(e^{-\epsilon_n H_{P_0}}\epsilon_n(H_{P^0_{n,B_n}}-H_{P_0})+\frac{e^{Y_{P^0_{n,B_n}}}}{2}\epsilon_n^2(H_{P^0_{n,B_n}}-H_{P_0})^2\right)\times\\
&\quad e^{-\epsilon_n H_{P_0}}(C_{P^0_{n,B_n}}-C_{P_0})\Bigg\}\\
&+E_{B_n}P_0\left\{e^{-2\epsilon_n H_{P_0}}(C_{P^0_{n,B_n}}-C_{P_0})^2\right\}.
\end{aligned}
$$

After repeated applications of Cauchy-Schwartz inequality to the summands,

the boundedness assumptions and lemmas 4 and 9 imply that indeed

$$E_{B_n} P_0 \left(C_{P^0_{n,B_n}} e^{-\epsilon_n H_{P^0_{n,B_n}}} - C_{P_0} e^{-\epsilon_n H_{P_0}} \right)^2 \xrightarrow{P} 0.$$

Hence (43) converges to 0 in probability.

The term (44) can be written as

$$E_{B_n} \left(P^1_{n,B_n} - P_0 \right) D_Y(P_0, \epsilon_n) \equiv E_{B_n} \left(P^1_{n,B_n} - P_0 \right) (Y - \hat{\bar{Q}}(\epsilon_n)(P_0)) H^*_{\hat{g}(P_0)}$$
$$= E_{B_n} \left(P^1_{n,B_n} - P_0 \right) Y H^*_{\hat{g}(P_0)} - E_{B_n} \left(P^1_{n,B_n} - P_0 \right) \hat{\bar{Q}}(\epsilon_n)(P_0) H^*_{\hat{g}(P_0)}$$

The first summand in the last equality is an empirical difference that is asymptotically normal with mean zero. In particular, it converges to zero in probability. The second summand also converges to 0 in probability. To see that, let $\mathcal{F}(P_0) = \{f_\epsilon = \hat{\bar{Q}}(P_0)(\epsilon) H^*_{\hat{g}(P_0)} : \epsilon\}$, where ϵ ranges over K. On a sample split of B_n, lemma 1 implies that

$$\sqrt{n} E \left(sup_{f \in \mathcal{F}} |(P^1_{n,B_n} - P_0) f| \right) \leq Entro(\mathcal{F}) \sqrt{P_0 \mathbf{F}^2},$$

where $\mathbf{F}$ is an envelope of $\mathcal{F}$. Therefore, we may write

$$P \left(|(P^1_{n,B_n} - P_0) f_{\epsilon_n}(P_0)| > \delta \right) \leq E P \left(sup_{f \in \mathcal{F}} |(P^1_{n,B_n} - P_0) f| > \delta \right)$$
$$\leq \frac{1}{\delta} E E \left(sup_{f \in \mathcal{F}} |(P^1_{n,B_n} - P_0) f| \right) \leq \frac{1}{\sqrt{n}} \frac{1}{\delta} E Entro(\mathcal{F}) \sqrt{P_0 \mathbf{F}^2}.$$

The entropy of this class is bounded. From the boundedness assumptions of $\hat{g}(P_0)$ and the definition of $\hat{\bar{Q}}(\epsilon)$, we see that all the functions the $\mathcal{F}$ are also bounded, hence $\sqrt{P_0 \mathbf{F}^2}$ is bounded. Therefore, the RHS of the last inequality converges to 0 in probability as $n \to \infty$. This result combined with the fact that B_n is uniformly distributed over a finite support now imply that $E_{B_n} \left(P^1_{n,B_n} - P_0 \right) \hat{\bar{Q}}(\epsilon_n)(P_0) H^*_{\hat{g}(P_0)}$ indeed converge to 0 in probability.

It remains to show that (42) converges to 0 in probability. By our assumption, there exists constant $M > 0$ such that $P(|\epsilon_n| < M) = 1$. Conditional on P^0_{n,B_n}, consider the class

$$\mathcal{F}(P^0_{n,B_n}) = \left\{ f_\epsilon(P^0_{n,B_n}) = D_Y(P^0_{n,B_n}, \epsilon) - D_Y(P_0, \epsilon) : |\epsilon| < M \right\}.$$

Lemma 1 implies that

$$\sqrt{n} E \left(sup_{f \in \mathcal{F}(P^0_{n,B_n})} |(P^1_{n,B_n} - P_0) f| \right) \leq Entro(\mathcal{F}(P^0_{n,B_n})) \sqrt{P_0 \mathbf{F}(P^0_{n,B_n})^2},$$

where $\mathbf{F}(P^0_{n,B_n})$ is an envelope of $\mathcal{F}(P^0_{n,B_n})$. Therefore, we may write

$$\begin{aligned}
&P\left(|(P^1_{n,B_n}-P_0)f_{\epsilon_n}(P^0_{n,B_n})|>\delta\right)\\
&\leq EP\left(sup_{f\in\mathcal{F}(P^0_{n,B_n})}|(P^1_{n,B_n}-P_0)f|>\delta\middle|P^0_{n,B_n}\right)\\
&\leq \frac{1}{\delta}EE\left(sup_{f\in\mathcal{F}(P^0_{n,B_n})}|(P^1_{n,B_n}-P_0)f|\right)\\
&\leq \frac{1}{\sqrt{n}}\frac{1}{\delta}EEntro(\mathcal{F}(P^0_{n,B_n}))\sqrt{P_0\mathbf{F}(P^0_{n,B_n})^2}.
\end{aligned}$$

Firstly note that f_ϵ is bounded per our assumptions and construction of $\hat{\bar{Q}}(P)(\epsilon)$. Hence $\sqrt{P_0\mathbf{F}(P^0_{n,B_n})^2}$ is bounded. On the other hand, the entropy of the class is also bounded. Therefore, we indeed have $P\left(|(P^1_{n,B_n}-P_0)f_{\epsilon_n}(P_0)|>\delta\right)$ converges to 0 as $n\to\infty$. Consequently, (42) converges to 0 in probability. We have thus shown that $P_0D_Y(P_0,\epsilon_n)\xrightarrow{P}0$.

Since K is compact, there is a subsequence ϵ_{nk} such that $\epsilon_{nk}\xrightarrow{P}\epsilon^*$ for some $\epsilon^*\in K$. This implies that for

$$\begin{aligned}
g(\epsilon) &\equiv P_0D_Y(P_0,\epsilon)\\
&= P_0YH^*_{\hat{g}(P_0)}-P_0\frac{H^*_{\hat{g}(P_0)}}{1+e^{-logit(\hat{\bar{Q}}(P_0))-\epsilon H^*_{\hat{g}(P_0)}}},
\end{aligned}$$

which is continuous over K, we must have $g(\epsilon_{nk})\xrightarrow{P}g(\epsilon^*)$.

Since $g(\epsilon_n)\xrightarrow{P}0$, as determined above, it follows that $g(\epsilon^*)=0$. On the other hand, by definition of ϵ_0 we have that $g(\epsilon_0)=0$. Note that $g'(\epsilon)<0$, hence it's monotonic in ϵ. Therefore we indeed have $\epsilon^*=\epsilon_0$. This implies that all convergent subsequences of ϵ_n converge to ϵ_0 in probability. Since K is compact, it now implies that ϵ_n converge to ϵ_0 in probability. □

Proof of Lemma 12:

Conditional on P^0_{n,B_n}, for a deterministic sequence δ_n converging to 0, consider the class

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n})\equiv\left\{D_Y(P^0_{n,B_n},\epsilon)-D_Y(P_0,\epsilon):\|\epsilon-\epsilon_0\|<\delta_n\right\},$$

where

$$\begin{aligned}
&D_Y(P^0_{n,B_n},\epsilon)-D_Y(P_0,\epsilon)\\
&=\left(Y-\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon)\right)H^*_{\hat{g}(P^0_{n,B_n})}-\left(Y-\hat{\bar{Q}}(P_0)(\epsilon_0)\right)H^*_{\hat{g}(P_0)}.
\end{aligned}$$

From lemma 11, we readily have $\| \epsilon_n - \epsilon_0 \| \xrightarrow{P} 0$. To obtain the desired result, it remains to show that $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$ satisfies the conditions of lemma 2.

For convenience, let $H_{P^0_{n,B_n}}(O) \equiv H^*_{\hat{g}(P^0_{n,B_n})}(A,W)$, and H_{P_0} its counterpart at P_0. Then, we can find an envelope for this class of functions as follows:

$$
\begin{aligned}
&\left|D_Y(P^0_{n,B_n},\epsilon) - D_Y(P_0,\epsilon)\right| \\
&\leq |Y||H_{P^0_{n,B_n}} - H_{P_0}| + |\hat{\bar{Q}}(P_0)(\epsilon_0)|H_{P^0_{n,B_n}} - H_{P_0}| + |H_{P^0_{n,B_n}}||\hat{\bar{Q}}(P^0_{n,B_n})(\epsilon) - \hat{\bar{Q}}(P_0)(\epsilon_0)| \\
&\leq |Y + \hat{\bar{Q}}(P_0)(\epsilon_0)||H_{P^0_{n,B_n}} - H_{P_0}| \\
&+|H_{P^0_{n,B_n}}|\left|\frac{C_{P^0_{n,B_n}} e^{-\epsilon H_{P^0_{n,B_n}}} - C_{P_0} e^{-\epsilon_0 H_{P_0}}}{(1 + C_{P^0_{n,B_n}} e^{-\epsilon H_{P^0_{n,B_n}}})(1 + C_{P^0_{n,B_n}} e^{-\epsilon H_{P^0_{n,B_n}}})}\right| \\
&\leq |Y + \hat{\bar{Q}}(P_0)(\epsilon_0)||H_{P^0_{n,B_n}} - H_{P_0}| + |H_{P^0_{n,B_n}}|e^{-\epsilon_0 H_{P_0}}|C_{P^0_{n,B_n}} - C_{P_0}| \\
&+|H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}||e^{-\epsilon H_{P^0_{n,B_n}}} - e^{-\epsilon_0 H_{P_0}}| \\
&\leq |Y + \hat{\bar{Q}}(P_0)(\epsilon_0)||H_{P^0_{n,B_n}} - H_{P_0}| + |H_{P^0_{n,B_n}}|e^{-\epsilon_0 H_{P_0}}|C_{P^0_{n,B_n}} - C_{P_0}| \\
&+|H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}||e^{-\epsilon_0 H_{P_0}}|\epsilon H_{P^0_{n,B_n}} - \epsilon_0 H_{P_0}| \\
&+|H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}||\frac{e^{M''}}{2}|\epsilon H_{P^0_{n,B_n}} - \epsilon_0 H_{P_0}|^2 \\
&\leq |Y + \hat{\bar{Q}}(P_0)(\epsilon_0)||H_{P^0_{n,B_n}} - H_{P_0}| + |H_{P^0_{n,B_n}}||e^{-\epsilon_0 H_{P_0}}||C_{P^0_{n,B_n}} - C_{P_0}| \\
&+|H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}|e^{-\epsilon_0 H_{P_0}}\epsilon_0|H_{P^0_{n,B_n}} - H_{P_0}| \\
&+|H_{P^0_{n,B_n}}|^2|C_{P^0_{n,B_n}}|e^{-\epsilon_0 H_{P_0}}\delta_n + |H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}|\frac{e^{M''}}{2}\epsilon_0^2|H_{P^0_{n,B_n}} - H_{P_0}|^2 \\
&+|H_{P^0_{n,B_n}}|^3|C_{P^0_{n,B_n}}|\frac{e^{M''}}{2}\delta_n^2 \\
&+2|H_{P^0_{n,B_n}}|^2|C_{P^0_{n,B_n}}|\frac{e^{M''}}{2}\epsilon_0|H_{P^0_{n,B_n}} - H_{P_0}|\delta_n
\end{aligned}
$$

$$
\begin{aligned}
&= \left(Y + \hat{\bar{Q}}(P_0)(\epsilon_0) + |H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}|e^{-\epsilon_0 H_{P_0}}\epsilon_0\right)|H_{P^0_{n,B_n}} - H_{P_0}| \\
&+|H_{P^0_{n,B_n}}||e^{-\epsilon_0 H_{P_0}}||C_{P^0_{n,B_n}} - C_{P_0}| \\
&+|H_{P^0_{n,B_n}}|^2|C_{P^0_{n,B_n}}|e^{-\epsilon_0 H_{P_0}}\delta_n + |H_{P^0_{n,B_n}}||C_{P^0_{n,B_n}}|\frac{e^{M''}}{2}\epsilon_0^2|H_{P^0_{n,B_n}} - H_{P_0}|^2 \\
&+|H_{P^0_{n,B_n}}|^3|C_{P^0_{n,B_n}}|\frac{e^{M''}}{2}\delta_n^2 \\
&+2|H_{P^0_{n,B_n}}|^2|C_{P^0_{n,B_n}}|\frac{e^{M''}}{2}\epsilon_0|H_{P^0_{n,B_n}} - H_{P_0}|\delta_n \\
&\equiv \mathbf{F}_n.
\end{aligned}
$$

Applying Cauchy-Schwartz inequality in combination with lemma 10 and boundedness assumptions, we thereby have that $EP_0(\mathbf{F}_n)^2 \to 0$. Furthermore, the entropy of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$ is bounded. Therefore, from lemma 2 it follows that

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{D_Y(P^0_{n,B_n}, \epsilon) - D_Y(P_0, \epsilon)\right\} = o_P(1).$$

□.

Proof of lemma 13: This is proved analogue to the proof of lemma 12. □.

References

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Technical report 265, Division of Biostatistics, University of California, Berkeley, May 2010.

M.A. Hernan, B. Brumback, and J.M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000.

M. Petersen, K. Porter, S.Gruber, Y. Wang, and M.J. van der Laan. Diagnosing and responding to violations in the positivity assumption. Technical report 269, Division of Biostatistics, University of California, Berkeley, 2010. URL `http://www.bepress.com/ucbbiostat/paper269`.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology: the environment and clinical trials*, pages 95–134. Springer-Verlag, 1999.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.

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.

J.M. Robins, A. Rotnitzky, and M.J. 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.

O.M. Stitelman and M.J. van der Laan. Collaborative targeted maximum likelihood for time-to-event data. Technical Report 260, Division of Biostatistics, University of California, Berkeley, 2010.

M.J. van der Laan. Causal effect models for intention to treat and realistic individualized treatment rules. Technical report 203, Division of Biostatistics, University of California, Berkeley, 2006.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003.

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1), 2010.

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

M.J. van der Laan, S. Rose, and S. Gruber. Readings on targeted maximum likelihood estimation. *Technical report, working paper series http://www.bepress.com/ucbbiostat/paper254*, September, 2009.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.