Chapter 27
Cross-Validated Targeted Minimum-Loss-Based Estimation

Wenjing Zheng, Mark J. van der Laan

In previous chapters, we introduced targeted maximum likelihood estimation 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 minimizing another loss-specific empirical risk, over a "clever" parametric working model through the initial estimator, treating the initial estimator as offset. This updating process may need to be iterated to convergence. The target parameter of the resulting updated estimator is then evaluated, and is called the targeted minimum-loss-based estimator (also 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.

However, we have seen in practice that the performance of the TMLE suffers when the initial estimator is too adaptive, leaving little signal in the data to fit the residual bias with respect to the initial estimator in the targeting step. Moreover, the use of adaptive estimators raises the question to what degree we can still rely on the central limit theorem for statistical inference. Our previous theorems (e.g., van der Laan and Robins 2003; van der Laan and Rubin 2006; van der Laan and Gruber 2010) 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 put some bounds on how adaptive the initial estimator can be.

We present a version of TMLE that uses V-fold sample splitting for the initial estimator in order to make the TMLE maximally robust in its bias reduction step. We refer to this estimator as the cross-validated targeted minimum-loss-based estimator (CV-TMLE). In a direct application, we formally establish its asymptotics under stated conditions that avoid such empirical process conditions.

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_27,

We refer to our accompanying technical report (Zheng and van der Laan 2010) for the generalization of the theorem presented in this chapter to arbitrary semiparametric models and pathwise differentiable parameters.

27.1 The CV-TMLE

Let $O \sim P_0$. 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 they are 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 one-dimensional submodels $\{P_h(\epsilon) : \epsilon\}$ 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 the notation by using interchangeably $\Psi(Q(P))$ and $\Psi(P)$. Let $g : \mathcal{M} \to \mathcal{G}$ be such 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 a 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, $P_0L(Q) = \int L(Q)(o)dP_0(o)$. 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" at $\epsilon = 0$ for which the linear combinations of its components generates $D^*(P)$:

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

Here we used the notation $\langle h \rangle$ for the linear span spanned by the components of $h = (h_1, \ldots, h_k)$.

We are now ready to define the CV-TMLE. 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 $P_n \to \hat{Q}_\epsilon(P_n)$ be a family of estimators indexed by ϵ chosen so that

$$\langle D^*(\hat{Q}(P_n), \hat{g}(P_n)) \rangle \subset \langle \frac{d}{d\epsilon} L(\hat{Q}_\epsilon(P_n)) \Big|_{\epsilon=0} \rangle. \tag{27.1}$$

One can think of $\{\hat{Q}_\epsilon(P_n) : \epsilon\} \subset \mathcal{M}$ as a submodel through $\hat{Q}(P_n)$ with parameter ϵ, chosen so that the derivative/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 $P^0_{B_n,n}$, $P^1_{B_n,n}$ be the empirical probability distributions of the training and validation samples, respectively. For a given cross-validation scheme $B_n \in \{0, 1\}^n$, we now define

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

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

It is important to point out that this cross-validated selector of ϵ equals the cross-validation selector among the library of candidate estimators $P_n \to \hat{Q}_\epsilon(P_n)$ 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

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

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

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

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:

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

Here c can be explicitly bounded by M_2 and an upper bound of L. This shows that under no conditions on the initial estimator does the selection of ϵ have good consistency properties.

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

satisfying the derivative condition (27.1), and define

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

resulting in another update $\hat{Q}^1_{\epsilon_n^1}(P^0_{B_n,n})$ for each B_n. This process is iterated till $\epsilon_n^k = 0$ (or close enough to zero). The final update will be denoted by $\hat{Q}^*(P^0_{B_n,n})$ for each split B_n. The TMLE of ψ_0 is now defined as

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

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

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

Linear components. This TMLE can also be generalized to the case where only one component of Q should be estimated 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., a nonparametric maximum likelihood estimator). 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 mapping

$$P \to \Psi(Q_1(P), Q_2(P_0)) \tag{27.2}$$

at $P = P_0$. Assume also that $D_1^*(P)$ does not depend on $Q_2(P)$. Then we may estimate (27.2) at P_0, viewed as a function of $Q_1(P_0)$, as if $Q_2(P_0)$ were known, with the above-defined CV-TMLE. In this case, the parametric fluctuation model needs to satisfy

$$\langle D_1^*(\hat{Q}_1(P_n), \hat{g}(P_n))\rangle \subset \langle \frac{d}{d\epsilon} L(\hat{Q}_{1,\epsilon}(P_n))\Big|_{\epsilon=0} \rangle.$$

The optimal ϵ at each step is selected using cross-validation as described above but now with respect to a loss function $L(Q_1)$. The procedure ends when ϵ_n^k converges to 0. This yields a CV-TMLE $\hat{Q}_1^*(P^0_{B_n,n})$ for each sample split B_n. The resulting CV-TMLE of ψ_0 is given by

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

Note that we estimate Q_{20} on validation samples, which allows the asymptotics of the estimator to minimally depend on empirical process conditions, while the stated

linearity in Q_2 makes this estimator behave well (just like it is fine to estimate a mean with the average of subsample specific empirical means over the subsamples that partition the whole sample). We illustrate this estimator with an application to the additive causal effect of a binary treatment on a continuous or binary outcome.

27.2 The CV-TMLE for the Additive Causal Effect

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 $\Psi : \mathcal{M} \to \mathbb{R}$:

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

Here, $Q(P) = (\bar{Q}(P), Q_W(P))$, where $\bar{Q}(P)(A, W) \equiv E_P(Y \mid A, W)$ 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 $E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W)$. The distinctions will be clear from the arguments given to the function or from context. Let $g(P)(A \mid W) \equiv Pr_P(A \mid W)$. We also adopt the notations $\bar{Q}_0 = \bar{Q}(P_0)$ and $Q_{W,0} = Q_W(P_0)$.

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

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

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

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

where

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

For convenience, we will also use the notation

$$H^*_g(W) = H^*_g(1, W) - H^*_g(0, W).$$

Firstly, note that the map $Q_W \mapsto \Psi(\bar{Q}, Q_W)$ is linear. 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 what follows 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 $\hat{\bar{Q}}_\epsilon(P_n)$, 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 $\{\hat{\bar{Q}}_\epsilon(P_n) : \epsilon\}$ will be

selected such that

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

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

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

with the parametric working model

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

Quasi-log-likelihood loss function. The quasi-log-likelihood loss function is given by:

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

with the parametric working model

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

We note that we would use this loss function if Y were binary or Y were 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 a 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 samples, respectively. Given the parametric working model, the optimal ϵ_n is selected using cross-validation:

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

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

$$0 = E_{B_n} P^1_{B_n,n} D_Y^*(\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n}), \hat{g}(P^0_{B_n,n})). \tag{27.3}$$

The TMLE of ψ_0 is defined as

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

In the theorem and proof, at each sample split B_n, we define the TMLE of Q_0 at (P_n, B_n) as

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

27.3 Asymptotics of the CV-TMLE

We will now use the squared error loss example to illustrate the theoretical advantages of CV-TMLE and the use of data-adaptive estimators for the initial estimators. We will show that under a natural rate condition on the initial estimators $\hat{\bar{Q}}$ and $\hat{g}$, the resulting TMLE ψ_n^* is asymptotically linear, and when $\hat{g}$ and $\hat{\bar{Q}}$ are consistent, its influence curve is indeed the efficient influence curve. For a similar theorem for the CV-TMLE using the quasi-log-likelihood loss function and its proof, we refer to the accompanying technical report (Zheng and van der Laan 2010).

Theorem 27.1. *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. We will index the V possible outcomes of B_n with $v = 1,\ldots,V$. Let $\hat{\bar{Q}}$ and $\hat{g}$ be initial estimators of $\bar{Q}_0$ and g_0. In what follows, $\hat{\bar{Q}}(P_0)$ and $\hat{g}(P_0)$ denote limits of these estimators, not necessarily equal to $\bar{Q}_0$ and g_0, respectively. The CV-TMLE satisfies*

$$\begin{aligned}\psi_n^* - \psi_0 = {} & E_{B_n}\left(P^1_{B_n,n} - P_0\right) D^*\left(\hat{Q}_{B_n}(P_n), \hat{g}(P^0_{B_n,n})\right) \\ & + E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{B_n,n})} \left(\bar{Q}_0 - \hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})\right)\left(g_0 - \hat{g}(P^0_{B_n,n})\right)\right\}. \quad (27.4)\end{aligned}$$

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

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

Suppose that these minima exist and satisfy the derivative equations

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

where

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

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 $K \subset \mathbb{R}^k$ 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. $$\int_W (\hat{g}(P_n)(1|w) - \hat{g}(P_0)(1|w))^2 \, dQ_{W,0}(w) \to 0 \quad \textit{in probability.}$$
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,

$$\begin{aligned}\psi_n^* - \psi_0 = (P_n - P_0)\left\{D_Y^*\left(\hat{\bar{Q}}_{\epsilon_0}(P_0), \hat{g}(P_0)\right) + \hat{\bar{Q}}_{\epsilon_0}(P_0)\right\} \\ + E_{B_n} P_0 \left\{\frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{B_n,n})}\left(\bar{Q}_0 - \hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})\right)\left(g_0 - \hat{g}(P^0_{B_n,n})\right)\right\} \\ + o_P(1/\sqrt{n}). \end{aligned} \tag{27.5}$$

Furthermore, if $\hat{g}(P_n) = g_0$, the TMLE estimator ψ_n^* is an asymptotically linear estimator of ψ_0:

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(\hat{Q}_{\epsilon_0}(P_0), g_0) + o_P(1/\sqrt{n}), \tag{27.6}$$

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

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

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(Q_0, g_0) + o_P(1/\sqrt{n}). \tag{27.7}$$

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

$$\sqrt{E_{B_n} P_0 \left(\frac{g_0 - \hat{g}(P^0_{B_n,n})}{g_0 \hat{g}(P^0_{B_n,n})}\right)^2} \sqrt{E_{B_n} P_0 \left(\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n}) - \bar{Q}_0\right)^2} = o_P(1/\sqrt{n}), \tag{27.8}$$

then ψ_n^* is an asymptotically efficient estimator of ψ_0:

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

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

$$\sqrt{E_{B_n} P_0 \left(\frac{g_0 - \hat{g}(P^0_{B_n,n})}{g_0 \hat{g}(P^0_{B_n,n})} \right)^2} \sqrt{E_{B_n} P_0 \left(\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n}) - \hat{\bar{Q}}_{\epsilon_0}(P_0) \right)^2} = o_P(1/\sqrt{n}), \quad (27.9)$$

and

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

is an asymptotically linear estimator of

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

with influence curve IC', *then* ψ^*_n *is an asymptotically linear estimator of* ψ_0:

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

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

Lemma 27.1. *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* $r \geq 1$.

Definition 27.1. An envelope of a class of functions $\mathcal{F}$ is a function F such that $|f| \leq F$ for all $f \in \mathcal{F}$.

Definition 27.2. For a class of functions $\mathcal{F}$ whose elements are functions f that map O into a real number, we define the entropy integral

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

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.

We refer to van der Vaart and Wellner (1996) for empirical process theory. Lemma 27.2 below is an application of Lemma 2.14.1 in van der Vaart and Wellner (1996).

Lemma 27.2. *Conditional on* $P^0_{B_n,n}$, *let* $\mathcal{F}(P^0_{B_n,n})$ *denote a class of measurable functions of* O. *Suppose that the entropy integral of this class is bounded and there is an envelope function* $\mathbf{F}(P^0_{B_n,n})$ *of* $\mathcal{F}(P^0_{B_n,n})$ *such that* $EP_0\mathbf{F}(P^0_{B_n,n})^2 \to 0$. *Then for any* $\delta > 0$

$$EP \left(\sup_{f \in \mathcal{F}(P^0_{B_n,n})} \left| \sqrt{n} \left(P^1_{B_n,n} - P_0 \right) f \right| > \delta \middle| P^0_{B_n,n} \right) \to 0.$$

Lemma 27.3. *Suppose $\hat{g}$ is such that for some $\delta > 0$, $P(1 - \delta > \hat{g}(P_n)(1 \mid W) > \delta) = 1$. If $\hat{g}$ satisfies $P_{W,0}\left(\hat{g}(P_n) - \hat{g}(P_0)\right)^2 \xrightarrow{P} 0$, then we have that $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)^2$, $P_0\left(H^*_{\hat{g}(P_n)} - H^*_{\hat{g}(P_0)}\right)$, and $P_0\left((H^*_{\hat{g}(P_n)})^2 - (H^*_{\hat{g}(P_0)})^2\right)$ also converge to zero in probability.*

Lemma 27.4. *Suppose $\hat{g}$ and $\hat{\bar{Q}}$ satisfy conditions 2–5 in Theorem 27.1. Then, for any $r \geq 1$:*

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

We are now ready to prove Theorem 27.1.

Proof. Firstly, we wish to establish that

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

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

Note that

$$\begin{aligned}-P_0D^*(Q(P), g_0) &\equiv -P_0\left\{\left(Y - \bar{Q}(P)\right)H^*_{g_0} + \bar{Q}(P) - Q_W(P)\bar{Q}(P)\right\} \\ &= -\left\{P_0YH^*_{g_0} - P_0\bar{Q}(P)H^*_{g_0} + P_{W,0}\bar{Q}(P) - Q_W(P)\bar{Q}(P)\right\} \\ &= Q_W(P)\bar{Q}(P) - P_0YH^*_{g_0} \\ &= \Psi(Q(P)) - \Psi(Q_0).\end{aligned}$$

Applying this result to each sample split B_n and averaging over its support, it follows that

$$\psi^*_n - \psi_0 \equiv E_{B_n}\Psi\left(\hat{Q}_{B_n}(P_n)\right) - \Psi(Q(P_0)) = -E_{B_n}P_0D^*\left(\hat{Q}_{B_n}(P_n), g_0\right). \tag{27.10}$$

On the other hand,

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

Moreover, it follows from the definition of ϵ_n and the one-step convergence of the chosen fluctuation model that $\left(\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n}), \hat{g}(P^0_{B_n,n})\right)$ satisfies (27.3). Therefore, we have

$$\begin{aligned} & E_{B_n} P^1_{B_n,n} D^* \left(\hat{Q}_{B_n}(P_n), \hat{g}(P^0_{B_n,n})\right) \\ & \equiv E_{B_n} P^1_{B_n,n} D^*_Y \left(\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n}), \hat{g}(P^0_{B_n,n})\right) + E_{B_n} P^1_{B_n,n} D^*_W \left(\hat{Q}_W(P^1_{B_n,n}), \hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})\right) \\ & = 0. \end{aligned} \tag{27.11}$$

Combining (27.10), (27.11), and the robustness of D^*, $P_0 D^*(Q_0, g) = 0$ for all g, we may now rewrite $\psi^*_n - \psi_0$ as

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

The last two summands in this equality can be combined as

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

Therefore, we indeed have the desired expression (27.4):

$$\psi^*_n - \psi_0 = E_{B_n}\left(P^1_{B_n,n} - P_0\right) D^*\left(\hat{Q}_{B_n}(P_n), \hat{g}(P^0_{B_n,n})\right) \tag{27.12}$$

$$+ E_{B_n} P_0 \left\{\frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{B_n,n})}\left(\bar{Q}_0 - \hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})\right)\left(g_0 - \hat{g}(P^0_{B_n,n})\right)\right\}. \tag{27.13}$$

We now study each term separately. For convenience, we use the notation $D_Y(P, \epsilon) \equiv D^*_Y(\hat{\bar{Q}}_\epsilon(P), \hat{g}(P))$. Term (27.12) can be written as

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

$$
\begin{aligned}
&+E_{B_n}\left(P^1_{B_n,n}-P_0\right)\left\{\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})-Q_W(P^1_{B_n,n})\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})\right\}\\
&=E_{B_n}\left(P^1_{B_n,n}-P_0\right)\left\{D_Y\left(P^0_{B_n,n},\epsilon_n\right)-D_Y\left(P_0,\epsilon_0\right)\right\} \qquad (27.14)\\
&+\left(P_n-P_0\right)D_Y\left(P_0,\epsilon_0\right)\\
&+E_{B_n}\left(P^1_{B_n,n}-P_0\right)\left\{\hat{\bar{Q}}_{\epsilon_n}(P^0_{B_n,n})-\hat{\bar{Q}}_{\epsilon_0}(P_0)\right\} \qquad (27.15)\\
&+\left(P_n-P_0\right)\hat{\bar{Q}}_{\epsilon_0}(P_0).
\end{aligned}
$$

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

Lemma 27.5. *Let ϵ_n and ϵ_0 be defined as in Theorem 27.1 and suppose they solve the derivative equations as stated in the theorem. If $\hat{g}$ and $\hat{\bar{Q}}$ satisfy conditions 1–5 in Theorem 27.1, then ϵ_n converges to ϵ_0 in probability.*

Now consider the following lemmas

Lemma 27.6. *If the initial estimators $\hat{\bar{Q}}$ and $\hat{g}$ satisfy conditions 1–5 in the theorem, then, conditional on a sample split B_n,*

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

Lemma 27.7. *If $\hat{\bar{Q}}$ and $\hat{g}$ satisfy conditions 1–5 of the theorem, then, conditional on a sample split B_n,*

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

Note that Lemmas 27.5–27.7 follow from Lemmas 27.2–27.4.

Lemmas 27.6 and 27.7 imply that (27.14) and (27.15) are $o_P(1/\sqrt{n})$. We thus have established that (27.12) is given by

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

Combining this result with (27.13), we have proved (27.5):

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

Note that up to this point we have only used the 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 (27.13):

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

We consider several cases. Firstly, consider the case $\hat{g}(P_n) = g_0$. In this case, term (27.13) is exactly 0. Therefore, (27.5) now implies that ψ_n^* is asymptotically linear with influence curve $D^*(\hat{Q}_{\epsilon_0}(P_0), g_0)$. 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 $\hat{\hat{\bar{Q}}}_{\epsilon_0}(P_0)$ is simply Q_0. Consequently, ψ_n^* is asymptotically linear with influence curve $D^*(Q_0, g_0)$ and is thereby asymptotically efficient.

Let's now consider the case where $\hat{g}(P_0) = g_0$ and $\hat{\bar{Q}}(P_0) = \bar{Q}_0$. In this case, $\hat{\hat{\bar{Q}}}_{\epsilon_n}(P^0_{B_n,n})$ converges to $\bar{Q}_0$ and $\hat{g}(P^0_{B_n,n})$ converges to g_0. In particular, these imply that (27.13) converges to 0. However, for ψ_n^* to be asymptotically linear, it is necessary that the convergence of this second order term occurs at a $\sqrt{n}$ rate, i.e.,

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

Applying the Cauchy–Schwartz inequality, it follows that if

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

then ψ_n^* will be asymptotically efficient.

Finally, consider the case where $\hat{g}(P_0) = g_0$, but $\hat{\hat{\bar{Q}}}(P_0) \neq \bar{Q}_0$. We reconsider expression (27.13) to account for the limit $\hat{\hat{\bar{Q}}}_{\epsilon_0}(P_0)$ of $\hat{\hat{\bar{Q}}}_{\epsilon_n}(P^0_{B_n,n})$, which does not equal $\bar{Q}_0$:

$$\begin{aligned} & E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{B_n,n})} \left(g_0 - \hat{g}(P^0_{B_n,n}) \right) \left(\hat{\hat{\bar{Q}}}_{\epsilon_n}(P^0_{B_n,n}) - \bar{Q}_0 \right) \right\} \\ &= E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{B_n,n})} \left(g_0 - \hat{g}(P^0_{B_n,n}) \right) \left(\hat{\hat{\bar{Q}}}_{\epsilon_n}(P^0_{B_n,n}) - \hat{\hat{\bar{Q}}}_{\epsilon_0}(P_0) \right) \right\} \quad (27.16) \\ &\quad + E_{B_n} P_0 \left\{ \frac{(-1)^{1+A}}{g_0 \hat{g}(P^0_{B_n,n})} \left(g_0 - \hat{g}(P^0_{B_n,n}) \right) \left(\hat{\hat{\bar{Q}}}_{\epsilon_0}(P_0) - \bar{Q}_0 \right) \right\}. \quad (27.17) \end{aligned}$$

Firstly, we require again that the rate of convergence for the second order term in (27.16) be $\sqrt{n}$, that is,

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

Applying the Cauchy–Schwartz inequality, it suffices that

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

For (27.17) 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

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

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

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

If $\hat{g}$ is such that $\Phi(P_n)-\Phi(P_0)$ is asymptotically linear (with some influnce curve IC'), then (27.17) becomes

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

Therefore, if $\hat{g}$ and $\hat{\bar{Q}}$ satisfy the convergence speed condition and $\Phi(P_n)-\Phi(P_0)$ is asymptotically linear, then it follows from (27.16) and (27.17) that the remainder (27.13) becomes

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

This completes the proof.

27.4 Discussion of Conditions of the Theorem

Under no conditions we determined an exact identity (27.4) for the CV-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 second-order remainder term. The cross-validated empirical process term is convenient because it involves, for each sample split, an empirical mean over a validation sample of an estimated efficient influence curve that is largely (up till a finite dimensional ϵ) estimated based on the training sample. Based on this, one would predict that one could 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 our second result (27.5), which replaces the cross-validated empirical process term by an empirical mean of mean zero random variables $D^*(\hat{Q}_{\epsilon_0}(P_0), \hat{g}(P_0))$ plus a negligible $o_P(1/\sqrt{n})$-term. This result only requires the positivity assumption and *that the estimators converge to a limit*. That is, under essentially no conditions beyond the positivity assumption does the CV-TMLE minus the true ψ_0 behave 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 for 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 to be $o(1/\sqrt{n})$. In an RCT, g_0 is known, and one might set $\hat{g}(P_n) = g_0$, so that the second-order remainder term is exactly equal to zero, giving us the asymptotic linearity (27.6) 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 (27.7). 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 (27.8) 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 (27.9) 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 at being consistent for g_0, one will need to uti-

lize 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 (27.8) 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 (27.9) holds, then another condition is required. Namely, we now need for $\hat{g}$ to be 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 $\bar{Q}^* = \hat{\bar{Q}}_{\epsilon_0}(P_0)$. 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 toward 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.

The goal of this chapter was to present a TMLE that allows one to learn the truth ψ_0, while also providing statistical inference based on a CLT, *under as large a statistical model as possible*. For that purpose, adaptive estimation (super learning), targeted minimum-loss-based estimation, and cross-validated selection of the fluctuation parameter in the TMLE are *all* essential tools to achieve this goal. The CV-TMLE combines these tools in one machine that is able to utilize all the state-of-the-art algorithms in machine learning and still provide proper inference in terms of confidence intervals and type I error control for testing null hypotheses, under minimal conditions.