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

Year 2015 *Paper* 343

A Generally Efficient Targeted Minimum Loss Based Estimator

Mark J. van der Laan[*]

[*]University of California, Berkeley, Division of Biostatistics, laan@berkeley.edu

A Generally Efficient Targeted Minimum Loss Based Estimator

Mark J. van der Laan

Abstract

Suppose we observe n independent and identically distributed observations of a finite dimensional bounded random variable. This article is concerned with the construction of an efficient targeted minimum loss-based estimator (TMLE) of a pathwise differentiable target parameter based on a realistic statistical model.

The canonical gradient of the target parameter at a particular data distribution will depend on the data distribution through an infinite dimensional nuisance parameter which can be defined as the minimizer of the expectation of a loss function (e.g., log-likelihood loss). For many models and target parameters the nuisance parameter can be split up in two components, one required for evaluation of the target parameter and one real nuisance parameter. The only smoothness condition we will enforce on the statistical model is that these nuisance parameters are multivariate real valued cadlag functions and have a finite supremum and variation norm.

We propose a general one-step targeted minimum loss-based estimator (TMLE) based on an initial estimator of the nuisance parameters defined by a loss-based super-learner that uses cross-validation to combine a library of candidate estimators. We enforce this library to contain minimum loss based estimators minimizing the empirical risk over the parameter space under the additional constraint that the variation norm is bounded by a set constant, across a set of constants for which the maximal constant converges to infinity with sample size. We show that this super-learner is not only asymptotically equivalent with the best performing algorithm in the library, but also that it always converges to the true nuisance parameter values at a rate faster than $nˆ{-1/4}$. This minimal rate applies to each dimension of the data and even to nonparametric statistical models. We also demonstrate that

the implementation of these constant-specific minimum loss-based estimators can be carried out by minimizing the empirical risk over linear combinations of basis functions under the constraint that the sum of the absolute value of the coefficients is smaller than the constant (e.g., Lasso regression), making our proposed estimators practically feasible.

Based on this rate of the super-learner of the nuisance parameter, we can establish that this one-step TMLE is asymptotically efficient at any data generating distribution in the model, under very weak structural conditions on the target parameter mapping and model. We demonstrate our general theorems by constructing such a one-step TMLE of the average causal effect in a nonparametric model, and presenting the corresponding efficiency theorem.

1 Introduction

We consider the general statistical estimation problem defined by a statistical model for the data distribution, a Euclidean valued target parameter mapping defined on the statistical model, and observing n independent and identically distributed draws from the data distribution. Our goal is to construct a generally efficient substitution estimator of the target parameter. For realistic statistical models this requires a highly data adaptive estimator. The current wisdom is that due to the curse of dimensionality this will typically require assuming very strong smoothness assumptions (e.g., Robins and Ritov (1997)).

There are two general methods for constructing an asymptotically efficient estimator. Firstly, the one-step estimator is defined by adding to an initial plug-in estimator of the target parameter an empirical mean of an estimator of the efficient influence curve at this same initial estimator (Bickel et al., 1993). In the special case that the efficient influence curve can be represented as an estimating function, one can represent this methodology as an estimating equation methodology, as has been developed for censored and causal inference models in the literature (van der Laan and Robins, 2003; Robins and Rotnitzky, 1992). Secondly, the TMLE defines a least favorable parametric submodel through an initial estimator of the relevant parts (nuisance parameters) of the data distribution, and updates the initial estimator with the MLE over this least favorable parametric submodel. The TMLE of the target parameter is now the resulting plug-in estimator (van der Laan and Rubin, 2006; van der Laan, 2008; van der Laan and Rose, 2011). In this article we focus on the TMLE since it is a more robust estimator by respecting the global constraints of the statistical model, which becomes evident when comparing the two estimators in simulations for which the information is low for the target parameter (e.g., even resulting in one-step estimators of probabilities that are outside the (0,1) range) (e.g., (Porter et al., 2011; Sekhon et al., 2012; Gruber and van der Laan, 2010)). Nonetheless, the results in this article have immediate analogues for the one-step estimator.

To make the TMLE highly data adaptive and thereby efficient for large statistical models we have recommended to estimate the relevant parts of the data distribution with a super-learner based on a large library of candidate estimators (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006, 2007; Polley et al., 2012). Due to the oracle inequality for the cross-validation selector, the super-learner will be asymptotically equivalent with the oracle selected estimator even when the number of candidate estimators in the library grows polynomial in sample size. In this article we develop a specific super learner which adapts to the underlying variation norm of the relevant nuisance parameters of the data distribution. We show that this super learner is guaranteed to converge to its true counterparts at a rate faster than the critical rate $n^{-1/4}$, even when the model only assumes that the true nuisance parameters have a finite variation norm.

Based on this fundamental result, we can then prove a general theorem for asymptotic efficiency of the TMLE for arbitrary statistical models. We will also use a so called cross-validated TMLE in order to further minimize the conditions for asymptotic efficiency (Zheng and van der Laan, 2011; van der Laan and Rose, 2011). By also including a large variety of other estimators in the library of the super-learner the TMLE will also have excellent practical performance for finite samples relative to competing estimators (Polley et al., 2012). Beyond establishing these fundamental theoretical general results, we will also discuss the practical implementation of such a super-learner and TMLE.

1.1 Organization of the article

In Section 2 we define the general estimation problem in terms of a pathwise differentiable statistical target parameter and statistical model, and define all the key characteristics of the estimation problem that will play a role in the definition of the estimator and our analysis. To address the estimation problem we will have to define the canonical gradient of the pathwise derivative, and the nuisance parameters this canonical gradient depends upon. These nuisance parameters will have to be estimated as part of the TMLE. We will introduce loss functions and loss-based dissimilarities for these nuisance parameters and define the bounded variation norm assumption on the true nuisance parameter values. The analysis of the super-learner of these nuisance parameters and the corresponding one-step TMLE will involve controlling various universal bounds on the statistical model. For that purpose we will define these model bounds. Since we will allow that some of these bounds are infinite, we will also define a sequence of bounded statistical submodels that grows to the complete statistical model, and the corresponding bounds that converge to the actual (possibly infinite) model bounds as the sample size converges to infinity. The minimal rate (i.e., worst case rate) of convergence of our super-learner estimators of the nuisance parameter are driven by entropies for the relevant parameter spaces, and we also need to control the entropy of the corresponding plug-in estimator of the canonical gradient. We will define these entropy bounds and the corresponding worst case rates of convergence for the super-learners. These worst-case rates will always (even for nonparametric models) be faster than the critical rate $n^{-1/4}$.

In Section 3 we define and analyze our first super-learner of the nuisance parameters. This first super-learner incorporates candidate estimators that minimize the empirical risk of the loss functions over all parameters in the parameter space that have a variation norm smaller than a $M < \infty$, across a set of such M values. Therefore, Section 3 will start out with analyzing these M-specific minimum loss-based estimators, and then proceeds with analyzing the corresponding super-learner based on the oracle inequality of the cross-validation selector. In Section 4 we analyze our second super-learner that is similar to the first super-learner except that its candidate M-specific minimum loss-based estimators minimize over finite epsilon-nets of the parameter space and selects both M and the resolution ϵ with cross-validation. Due to previous results for such cross-validated epsilon-net estimators we obtain a finite sample inequality for the resulting super-learner and slightly more optimal worst-case rates of convergence.

In Section 5 we define the one-step TMLE and discuss the local least favorable submodel that is used to update the super-learner estimator and establish results that show, under regularity conditions, that the one-step TMLE already guarantees that the empirical mean of the canonical gradient at the TMLE equals zero up till an asymptotically negligible remainder. Subsequently, in Section 5 we present a formal theorem establishing asymptotic efficiency of the one-step TMLE under specified conditions. In Section 6 we define the one-step cross-validated TMLE and present a formal theorem establishing its asymptotic efficiency. The advantage of the cross-validated TMLE is that it is asymptotically efficient under even weaker conditions than required for efficiency of the one-step TMLE, and, in particular, it allows the model bounds for the sieve to grow to infinity at a faster rate with sample size than for the TMLE.

In Section 7 we discuss the practical implementation of the M-specific minimum loss-based estimator that minimizes the empirical risk over all parameters in the parameter space that have variation norm smaller than M. We show that this MLE can be approximated by minimizing over linear combinations of basis functions under the constraint that the sum of the absolute value of the coefficients is bounded by M. In particular, we demonstrate that for nonparametric models these estimators can be implemented with Lasso type regression algorithms. In Section 8 we apply our theorems to the estimation of the average causal effect of a single time point binary treatment. We conclude with a discussion in Section 9. Our appendix is split up in various sections establishing the required empirical process results, and proofs of the various lemmas the efficiency of the one-step TMLE and CV-TMLE rely upon.

1.2 General idea

In order to follow the logic of this article it might help to understand the main idea behind the proposed one-step TMLE. The TMLE relies on an initial estimator of the key nuisance parameters that are required to evaluate the efficient influence curve of the target parameter. It is well known that the asymptotic efficiency of the TMLE mostly relies on a second order remainder being $o_P(n^{-1/2})$. Therefore, one wants to construct an initial estimator of the nuisance parameters that converges w.r.t. a suitable dissimilarity at a rate faster than $n^{-1/4}$.

The most important observation is that a minimum loss-based estimator minimizing the empirical risk over all candidate nuisance parameter values that have a variation norm smaller than $M < \infty$ converges at a rate faster than $n^{-1/4}$ to its M-specific true counterpart. By using a recent empirical process result by (van der Vaart and Wellner, 2011) we can establish the precise minimax rate of convergence in terms of the entropy of the model space. So, by selecting M larger than the unknown variation norm of the true nuisance parameter value, we obtain an initial estimator that converges at a faster rate than $n^{-1/4}$.

The second important observation is that if we define a collection of such M-specific estimators for a set of M-values for which the maximum value converges to infinity as sample size converges to infinitiy, and use cross-validation to data adaptively select M, then the resulting cross-validated selected estimator will be asymptotically equivalent with the oracle choice. This follows from a previously established oracle inequality for the cross-validation selector, as long as the supremum norm bound on the loss-function at the candidate estimators does not grow too fast to infinity as a function of sample size. As a consequence, our statistical model does not need to assume a universal bound on the variation norm of the nuisance parameters, but it only needs to assume that each nuisance parameter value has a finite variation norm. In this manner, we can construct super-learners that have a worst case rate faster than the critical rate $n^{-1/4}$. We obtain a super-learner that also in finite samples outperforms any competing algorithm by simply including these competing algorithms in the library of the super-learner beyond all these M-specific minimum loss-based estimators.

The typical TMLE involves iteratively updating this initial estimator through a parametric local least favorable submodel through the initial estimator/current estimator, so that the efficient score /influence curve equation is solved exactly. For the analysis of the TMLE it is very helpful if this TMLE algorithm converges in a single or finite number of steps. In many problems this TMLE updating algorithm converges in one step, and we developed a so called universal least favorable submodel that guarantees this convergence in one step (van der Laan and Gruber, 2015). However, we want our theory to apply to general local least favorable submodels, which can be easier to implement than a universal least favorable model. In order to deal with this challenge, we observe that if the initial estimators of the nuisance parameters converge at a rate faster than $n^{-1/4}$, then in great generality we can show that the one-step TMLE (thus only updating the super-learner once) already solves the efficient score equation up till a remainder of size $o_P(n^{-1/2})$. This is important, since this makes it relatively straightforward to establish the minimal rate of convergence of the TMLE update of the super-learner, and typically this minimal rate will not be worse than the rate of the super-learner itself. In this manner, we establish that also the TMLE update of the super-learner achieves the desired minimal rate faster than $n^{-1/4}$.

Given the understood behavior of the super-learners and their TMLE update, we can now carry out the general proof for asymptotic efficiency of the TMLE as presented in various of our previous articles on TMLE. Some extra care is needed in our proof since we allow that our true statistical model is unbounded. We allow such an unbounded statistical model by approximating it by a sequence of bounded submodels that grow slowly enough (w.r.t. sample size) to the true statistical model, and by enforcing our super-learners to respect that sequence of models. Finally, by using the CV-TMLE we can further reduce the conditions for asymptotic efficiency.

2 Formulation of the estimation problem, and definitions

Let $O_1, \ldots, O_n$ be n independent and identically distributed copies of a d-dimensional random variable O with probability distribution P_0 that is known to be an element of a statistical model $\mathcal{M}$. Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be a one-dimensional target parameter, so that $\psi_0 = \Psi(P_0)$ is the estimand of interest we aim to learn from the n observations $o_1, \ldots, o_n$. We assume that Ψ is pathwise differentiable at any $P \in \mathcal{M}$ with canonical gradient $D^*(P)$: for a specified class of one-dimensional submodels $\{P_\epsilon : \epsilon \in (-\delta, \delta)\} \subset \mathcal{M}$ through P at $\epsilon = 0$ and score $S = \frac{d}{d\epsilon} \log dP_\epsilon/dP\Big|_{\epsilon=0}$, we have

$$\frac{d}{d\epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0} = PD^*(P)S \equiv \int_o D^*(P)(o)S(o)dP(o).$$

Here we used the notation $Pf \equiv \int f(o)dP(o)$ for the expectation operator under P. The closure of the linear span of all scores generated by this class of one-dimensional submodels in the Hilbert space $L_0^2(P)$ (endowed with the inner product $\langle f, g\rangle_P = Pfg$) is called the tangent space at P and will be denoted with $T(P) \subset L_0^2(P)$. For a $f \in L^2(P)$, we denote its norm with $\| f \|_P = \sqrt{Pf^2}$. The canonical gradient at P is the unique gradient at P that is also an element of the tangent space $T(P)$.

Let P_n be the empirical probability distribution of $O_1, \ldots, O_n$. We view an estimator $\hat{\Psi} : \mathcal{M}_{np} \rightarrow \mathbb{R}$ as a mapping from the nonparametric model $\mathcal{M}_{np}$ to the real line so that it is well defined for any realization of the empirical distribution P_n. We recall from efficiency theory that an estimator $\hat{\Psi}(P_n)$ of ψ_0 is asymptotically efficient at P_0 if and only if $\hat{\Psi}(P_n)$ is asymptotically linear at P_0 with influence curve equal to the canonical gradient $D^*(P_0)$:

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

Therefore the canonical gradient is also called the efficient influence curve. Our goal in this article is to construct a substitution estimator (i.e., a TMLE) that is asymptotically efficient under minimal conditions.

Relevant nuisance parameters Q, G and their loss functions: Let $Q(P)$ be a nuisance parameter of P so that $\Psi(P) = \Psi_1(Q(P))$ for some Ψ_1, so that $\Psi(P)$ only depends on P through $Q(P)$. Let $\mathcal{Q} = Q(\mathcal{M}) = \{Q(P) : P \in \mathcal{M}\}$ be the parameter space of this parameter $Q : \mathcal{M} \rightarrow \mathcal{Q}$. Suppose that $Q(P) = (Q_j(P) : j = 1, \ldots, k_1 + 1)$ has $k_1 + 1$-components, and $Q_j : \mathcal{M} \rightarrow \mathcal{Q}_j$ are variation independent parameters $j = 1, \ldots, k_1 + 1$. Let $\mathcal{Q}_j = Q_j(\mathcal{M})$ be the parameter space of Q_j. Thus, the parameter space of Q is a cartesian product $\mathcal{Q} = \prod_{j=1}^{k_1+1} \mathcal{Q}_j$. In addition, suppose that for $j = 1, \ldots, k_1 + 1$, $Q_j(P_0) = \arg\min_{Q_j \in \mathcal{Q}_j} P_0 L_j(Q_j)$ for specified loss functions $(O, Q_j) \rightarrow L_j(Q_j)(O)$. Let $\bar{Q} = (Q_1, \ldots, Q_{k_1})$ represent parameters that require data adaptive estimation trading off variance and bias (e.g., densities), while Q_{k_1+1} represents an easy to estimate parameter for which we have an empirical estimator $\hat{Q}_{k_1+1}$ available with negligible bias. The parameter $\bar{Q}(P_0)$ will be estimated with our proposed loss-based super-learner. We define corresponding loss-based dissimilarities $d_{10j}(Q_j, Q_{j0}) = P_0 L_{1j}(Q_j) - P_0 L_{1j}(Q_{j0})$, $j = 1, \ldots, k_1$, while $d_{10k_1+1}(Q_{k_1+1}, Q_{k_1+10})$ represents a norm (e.g., supremum norm) or dissimilarity for which we know that $d_{10k_1+1}(\hat{Q}_{k_1+1}(P_n), Q_{k_1+10}) = O_P(r_{Q,k_1+1}(n))$ for a known rate of convergence $r_{Q,k_1+1}(n)$. It could be that $d_{10k_1+1}(Q_{k_1+1},, Q_{k_1+10}) = P_0 L_{1k_1+1}(Q_{k_1+1}) - P_0 L_{1k_1+1}(Q_{k_1+10})$, but that is not necessarily the case. Let

$$d_{10}(Q, Q_0) = (d_{10j}(Q_j, Q_{j0}) : j = 1, \ldots, k_1 + 1)$$

be the collection of these k_1+1 dissimilarities. We use the notation $d_{10}(\bar{Q},\bar{Q}_0) = (d_{10j}(Q_j,Q_{j0}) : j=1,\ldots,k_1)$ for the loss-based dissimilarities for $\bar{Q}$.

Suppose that $D^*(P)$ only depends on P through $Q(P)$ and an additional nuisance parameter $G(P)$. Let $G=(G_1,\ldots,G_{k_2+1})$ be a collection of k_2+1-variation independent parameters of G for some integer $k_2+1\geq 1$. Thus the parameter space of G is a cartesian product $\mathcal{G}=\prod_{j=1}^{k_2+1}\mathcal{G}_j$, where $\mathcal{G}_j$ is the parameter space of $G_j : \mathcal{M}\rightarrow\mathcal{G}_j$. Let $G_{j0}=\arg\min_{G\in\mathcal{G}_j} P_0L_{2j}(G_j)$ for a loss function $(O,G_j)\rightarrow L_{2j}(G_j)(O)$, and let $d_{2j0}(G_j,G_{j0})=P_0L_{2j}(G_j)-P_0L_{2j}(G_{j0})$ be the corresponding loss-based dissimilarity, $j=1,\ldots,k_2+1$. Let G_{k_2+1} represents an easy to estimate parameter for which we have a well behaved and understood estimator $\hat{G}_{k_2+1}$ available. We define corresponding loss-based dissimilarities $d_{20j}(G_j,G_{j0})=P_0L_{2j}(G_j)-P_0L_{2j}(G_{j0})$, $j=1,\ldots,k_2$, while $d_{20k_2+1}(G_{k_2+1},G_{k_2+10})$ represents a norm or dissimilarity for which we know that $d_{20k_2+1}(\hat{G}_{k_2+1}(P_n),G_{k_2+10})=O_P(r_{G,k_2+1}(n))$ for a known rate of convergence $r_{G,k_2+1}(n)$. As above, let $d_{20}(G,G_0)=(d_{20j}(G_j,G_{j0}) : j=1,\ldots,k_2+1)$ be the collection of these loss-based dissimilarities, and let $d_{20}(\bar{G},\bar{G}_0)=(d_{20j}(G_j,G_{j0}) : j=1,\ldots,k_2)$, where $\bar{G}=(G_1,\ldots,G_{k_2})$.

We also define

$$d_0((Q,G),(Q_0,G_0)) = (d_{10j_1}(Q_j,Q_{j_10}), d_{20j_2}(G_{j_2},G_{j_20}) : j_1,j_2)$$

as the collection of k_1+k_2+2 loss based dissimilarities. We will also use the short-hand notation $d_0(P,P_0)$ for $d_0((Q,G),(Q_0,G_0))$.

We define $L_1(Q)=(L_{1j}(Q_j) : j=1,\ldots,k_1+1)$ as the vector of k_1+1-loss functions for $Q=(Q_1,\ldots,Q_{k_1+1})$, and similarly we define $L_2(G)=(L_{2j}(G_j) : j=1,\ldots,k_2+1)$. We will also use the notation $L_1(\bar{Q})=(L_1(Q_j) : j=1,\ldots,k_1)$ and $L_2(\bar{G})=(L_{2j}(G_j) : j=1,\ldots,k_2)$. We will assume that $L_1(\bar{Q})$ is a convex loss function in the sense that for each $j=1,\ldots,k_1$ $P_0L_{1j}(\sum_{k=1}^m \alpha_k Q_{jk}) \leq \sum_{k=1}^m \alpha_k P_0L_{1j}(Q_{jk})$ when $\sum_k \alpha_k=1$ and $\min_k \alpha_k\geq 0$. Similarly, we assume $L_2(\bar{G})$ is a convex loss function. Our results for the TMLE generalize to non convex loss functions, but the convexity of the loss functions allows a nicer representation for the super-learner oracle inequality, and in most applications a natural convex loss function is available.

We will abuse notation by also denoting $\Psi(P)$ and $D^*(P)$ with $\Psi(Q)$ and $D^*(Q,G)$, respectively. A special case is that $D^*(P)=D^*(Q(P))$ does not depend on an additional nuisance parameter G.

First order expansion of pathwise differentiable target parameter: We define the second order remainder $R_2(P,P_0)$ as follows:

$$\Psi(P)-\Psi(P_0)=(P-P_0)D^*(P)+R_2(P,P_0),$$

or equivalently,

$$R_2(P,P_0) \equiv \Psi(P) - \Psi(P_0) + P_0 D^*(P).$$

We will also denote $R_2(P,P_0)$ with $R_{20}(Q,G,Q_0,G_0)$ to indicate that it involves differences between Q and Q_0 and G and G_0, beyond possibly some additional dependence on P_0. In our experience, this remainder $R_2(P,P_0)$ can be represented as a sum of terms of the type $\int (H_1(P) - H_1(P_0))(H_2(P) - H_2(P_0))f(P,P_0)dP_0(o)$ for some functionals H_1, H_2 and f, where, typically, $H_1(P)$ and $H_2(P)$ represent functions of $Q(P)$ or $G(P)$. In certain classes of problems we have that $R_2(P,P_0)$ only involves cross-terms of the type $\int (H_1(Q)-H_1(Q_0))(H_2(G)-H_2(G_0))f(P,P_0)dP_0$, so that $R_{20}(Q,G,Q_0,G_0) = 0$ if either $Q = Q_0$ or $G = G_0$. In these cases, we say that the efficient influence curve is double robust w.r.t. misspecification of Q_0 and G_0:

$$P_0 D^*(P) = \Psi(P_0) - \Psi(P) \text{ if } G(P) = G(P_0) \text{ or } Q(P) = Q(P_0).$$

Given this latter double robustness property of the canonical gradient (i.e, of the target parameter), if P solves $P_0 D^*(P) = 0$, and either $G(P) = G_0$ or $Q(P) = Q_0$, then $\Psi(P) = \Psi(P_0)$. This allows for the construction of si called double robust estimators of ψ_0 that will be consistent if either the estimator of Q_0 is consistent or the estimator of G_0 is consistent.

Support of data distribution: The support of $P \in \mathcal{M}$ is defined as a set $\mathcal{O}_P \subset \mathbb{R}^d$ so that $P(\mathcal{O}_P) = 1$. It is assumed that for each $P \in \mathcal{M}$, $\mathcal{O}_P \subset [0,\tau_P]$ for some finite $\tau_P \in \mathbb{R}^d_{>0}$. We define $\tau = \sup_{P\in\mathcal{M}} \tau_P$, so that $[0,\tau_P] \subset [0,\tau]$ for all $P \in \mathcal{M}$, where $\tau = \infty$ is allowed, in which case $[0,\tau] \equiv \mathbb{R}^d_{\geq 0}$. That is, $[0,\tau]$ is an upper bound of all the supports, and the model $\mathcal{M}$ states that the support of the data structure O is known to be contained in $[0,\tau]$.

Cadlag functions on $[0,\tau]$, supremum norm and variation norm: Suppose τ is finite, and, in fact, if τ is not finite, then we will apply the definitions below to a $\tau = \tau_n$ that is finite and converges to τ. Let $D[0,\tau]$ be the Banach space of d-variate real valued cadlag functions (Neuhaus, 1971). For a $f \in D[0,\tau]$, let $\parallel f \parallel_\infty = \sup_{x\in[0,\tau]} \mid f(x) \mid$ be the supremum norm. For a $f \in D[0,\tau]$, we define the variation norm of f (Gill et al., 1995) as

$$\parallel f \parallel_v = \mid f(0) \mid + \sum_{s\subset\{1,\ldots,d\}} \int_{(0_s,\tau_s]} \mid f(dx_s, o_{-s}) \mid .$$

For a subset $s \subset \{1,\ldots,d\}$, $x_s = (x_j : j \in s)$, $x_{-s} = (x_j : j \notin s)$, and the $\sum_s$ in the above definition of the variation norm is over all subsets of $\{1,\ldots,d\}$.

If $\| f \|_v < \infty$, then we can, in fact, represent f as follows (Gill et al., 1995):

$$f(x) = f(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{(0_s, x_s]} f(du_s, o_{-s}),$$

where $f(du_s, 0_{-s})$ is the measure generated by the cadlag function $u_s \rightarrow f(u_s, 0_{-s})$. For a $M \in \mathbb{R}_{\geq 0}$, let

$$\mathcal{F}_{v,M} = \{ f \in D[0,\tau] : \| f \|_v < M \}$$

denote the set of cadlag functions $f : [0,\tau] \rightarrow \mathbb{R}$ with variation norm bounded by M.

Cartesian product of cadlag function space, and its component-wise operations: Let $D^k[0,\tau]$ be the product Banach space of k-dimensional $(f_1, \ldots, f_k)$ where each $f_j \in D[0,\tau]$, $j = 1, \ldots, k$. If $f \in D^k[0,\tau]$, then we define $\| f \|_\infty = (\| f_j \|_\infty : j = 1, \ldots, k)$ as a vector whose j-th component equals the supremum norm of the j-th component f_j of f. Similarly we define a variation norm of $f \in D^k[0,\tau]$ as a vector

$$\| f \|_v^* = (\| f_j \|_v : j = 1, \ldots, k)$$

of variation norms, If $f \in D^k[0,\tau]$, then $\| f \|_{P_0} = (\| f_j \|_{P_0} : j = 1, \ldots, k)$ is a vector whose components are the $L^2(P_0)$-norms of the components of f. Generally speaking, in this paper any operation on a function $f \in D^k[0,\tau]$, such as taking a norm $\| f \|_{P_0}$, an expectation $P_0 f$, operations on a pair of functions $f, g \in D^k[0,\tau]$, such as f/g, $f * g$, $\max(f,g)$ or an inequality $f < g$, is carried out component wise: for example, $\max(f,g) = (\max(f_j, g_j) : j = 1, \ldots, k)$ and $\inf_{Q \in \mathcal{Q}} P_0 L_1(Q) = (\inf_{Q_j \in \mathcal{Q}_j} P_0 L_{1j}(Q_j) : j = 1, \ldots, k_1 + 1)$. In a similar manner, for an $M \in \mathbb{R}^k_{>0}$, let $\mathcal{F}_{v,M} = \prod_{j=1}^k \mathcal{F}_{v,M_j}$ denote the cartesian product. This general notation allows us to present results with minimal notation, avoiding the need to continuously having to enumerate all the components.

Our results will hold for general models and pathwise differentiable target parameters, as long as the statistical model satisfies the following key smoothness assumption:

Key Smoothness Assumption: For each $P \in \mathcal{M}$, $\bar{Q} = \bar{Q}(P) \in D^{k_1}[0,\tau]$, $\bar{G} = \bar{G}(P) \in D^{k_2}[0,\tau]$, $D^*(P) = D^*(Q,G) \in D[0,\tau]$, $L_1(\bar{Q}) \in D^{k_1}[0,\tau]$, $L_2(\bar{G}) \in D^{k_2}[0,\tau]$, and $\bar{Q}, \bar{G}, D^*(P), L_1(\bar{Q}), L_2(\bar{G})$ have a finite supremum and variation norm.

Definition of bounds on the statistical model: The properties of the super-learner and TMLE rely on bounds on the model. Our estimators will also allow for unbounded models by using a sieve of models for which its finite bounds approximate the actual model bound as sample size converges to infinity. These bounds will be defined now:

$$\begin{aligned}
\tau &= \tau(\mathcal{M}) = \sup_{P \in \mathcal{M}} \tau(P) \\
M_{1Q} &= M_{1Q}(\mathcal{M}) = \sup_{Q, Q_0 \in \mathcal{Q}} \parallel L_1(\bar{Q}) - L_1(\bar{Q}_0) \parallel_\infty \\
M_{2Q} &= M_{2Q}(\mathcal{M}) = \sup_{P, P_0 \in \mathcal{M}} \frac{\parallel L_1(\bar{Q}) - L_1(\bar{Q}_0) \parallel_{P_0}}{\{d_{10}(\bar{Q}, \bar{Q}_0)\}^{0.5}} \\
M_{1G} &= M_{1G}(\mathcal{M}) = \sup_{G, G_0 \in \mathcal{G}} \parallel L_2(\bar{G}) - L_2(\bar{G}_0) \parallel_\infty \\
M_{2G} &= M_{2G}(\mathcal{M}) = \sup_{P, P_0 \in \mathcal{M}} \frac{\parallel L_2(\bar{G}) - L_2(\bar{G}_0) \parallel_{P_0}}{\{d_{20}(\bar{G}, \bar{G}_0)\}^{0.5}} \\
M_{D^*} &= M_{D^*}(\mathcal{M}) = \sup_{P \in \mathcal{M}} \parallel D^*(P) \parallel_\infty
\end{aligned}$$

Note that $M_{1Q}, M_{2Q} \in \mathbb{R}^{k_1}_{\geq 0}$ and $M_{1G}, M_{2G} \in \mathbb{R}^{k_2}_{\geq 0}$ are defined as vectors of constants, a constant for each component of $\bar{Q}$ and $\bar{G}$, respectively. The bounds M_{1Q}, M_{2Q} guarantee excellent properties of the cross-validation selector based on the loss-function $L_1(\bar{Q})$. A bound on M_{2Q} shows that the loss-based dissimilarity $d_{01}(\bar{Q}, \bar{Q}_0)$ behaves as a square of a difference between $\bar{Q}$ and $\bar{Q}_0$. Similarly, the bounds M_{1G}, M_{2G} control the behavior of the cross-validation selector based on the loss function $L_2(\bar{G})$.

We also define the following universal variation norm bounds on the model $\mathcal{M}$:

$$\begin{aligned}
M_{Q,v} &= \sup_{P \in \mathcal{M}} \parallel \bar{Q}(P) \parallel_v \\
M_{G,v} &= \sup_{P \in \mathcal{M}} \parallel \bar{G}(P) \parallel_v \\
M_{D^*,v} &= \sup_{P \in \mathcal{M}} \parallel D^*(P) \parallel_v \\
M_{L_1(Q),v} &= \sup_{P \in \mathcal{M}} \parallel L_1(\bar{Q}) \parallel_v \\
M_{L_2(G),v} &= \sup_{P \in \mathcal{M}} \parallel L_2(\bar{G}) \parallel_v
\end{aligned}$$

Again, $M_{Q,v} \in \mathbb{R}^{k_1}_{\geq 0}, M_{L_1(Q),v} \in \mathbb{R}^{k_1}_{\geq 0}$ and $M_{G,v} \in \mathbb{R}^{k_2}_{\geq 0}, M_{L_2(G),v} \in \mathbb{R}^{k_2}_{\geq 0}$ are vectors of constants, one for each component of $\bar{Q}, L_1(\bar{Q}), \bar{G}, L_2(\bar{G})$, respectively.

Bounded and Unbounded Models: We will call the model $\mathcal{M}$ bounded if it is a model for which $\tau < \infty$ (i.e., universally bounded support), M_{1Q}, M_{2Q}, M_{1G}, M_{2G}, M_{D^*}, $M_{Q,v}$, $M_{G,v}$, $M_{D^*,v}$, $M_{L_1(Q),v}$ and $M_{L_2(G),v}$ are finite. In words, in essence, a bounded model is a model for which the supremum norm and variation norm of $\bar{Q}(P)$, $\bar{G}(P)$, $L_1(\bar{Q})$, $L_2(\bar{G})$ and $D^*(Q,G)$ are uniformly (over the model) bounded. Any model that is not bounded will be called an unbounded model.

Sequence of bounded submodels approximating the unbounded model: For an unbounded model $\mathcal{M}$, our initial estimators $(\bar{Q}_n, \bar{G}_n)$ of $(\bar{Q}_0, \bar{G}_0)$ are defined in terms of a sequence of bounded submodels $\mathcal{M}_n \subset \mathcal{M}$ that are increasing in n and approximate the actual model $\mathcal{M}$ as n converges to infinity. As a consequence, this sequence of models satisfies that for any $P_0 \in \mathcal{M}$, there exists an $N_0 = N(P_0)$, so that for $n > N_0$ $P_0 \in \mathcal{M}_n$. The counterparts of the above defined universal bounds on $\mathcal{M}$ applied to $\mathcal{M}_n$ are denoted with $\tau_n, M_{1Q,n}$, $M_{2Q,n}$, $M_{1G,n}$, $M_{2G,n}$, $M_{D^*,n}$, $M_{Q,v,n}$, $M_{G,v,n}$, $M_{D^*,v,n}$, $M_{L_1(Q),v,n}$ and $M_{L_2(G),v,n}$.

Let $\mathcal{Q}_n = Q(\mathcal{M}_n)$ and $\mathcal{G}_n = G(\mathcal{M}_n)$ be the parameter spaces of Q and G under model $\mathcal{M}_n$, and let $\bar{\mathcal{Q}}_n = \bar{Q}(\mathcal{M}_n)$ and $\bar{\mathcal{G}}_n = \bar{G}(\mathcal{M}_n)$ be the parameter spaces of $\bar{Q}$ and $\bar{G}$. We define the following true parameters corresponding with this model $\mathcal{M}_n$:

$$\begin{aligned}\bar{Q}_{0n} &= \arg\min_{\bar{Q}\in\bar{\mathcal{Q}}_n} P_0 L_1(\bar{Q}) \\ \bar{G}_{0n} &= \arg\min_{\bar{G}\in\bar{\mathcal{G}}_n} P_0 L_2(\bar{G}).\end{aligned}$$

We will assume that $\mathcal{M}_n$ is chosen so that $Q_{k_1+1}(P_{0n}) = Q_{k_1+1}(P_0)$ and $G_{k_2+1}(P_{0n}) = G_{k_2+1}(P_0)$, where $P_{0n} = \arg\max_{P\in\mathcal{M}_n} P_0 \log \frac{dP}{dP_0}$. That is, our sieve is not affecting the estimation of the easy nuisance parameters Q_{k_1+10} and G_{k_2+10}. Note that for $n > N_0$, we have $Q_{0n} = Q_0$ and $G_{0n} = G_0$.

In this paper our initial estimators of $\bar{Q}_0$ and $\bar{G}_0$ are always enforced to be in the parameter spaces of this sequence of models $\mathcal{M}_n$, but if the model $\mathcal{M}$ is already bounded, then one can set $\mathcal{M}_n = \mathcal{M}$ for all n. However, even for bounded models $\mathcal{M}$, the utilization of a sequence of submodels $\mathcal{M}_n$ with stronger universal bounds than $\mathcal{M}$ could result in finite sample improvements (e.g., if the universal bounds on $\mathcal{M}$ are very large relative to sample size and the dimension of the data).

Cross-validation: Our initial estimators rely on cross-validation. For that purpose, $B_n \in \{0,1\}^n$ will denote a random cross-validation scheme that randomly splits the sample $\{O_1, \ldots, O_n\}$ in a training sample $\{O_i : B_n(i) = 0\}$ and validation sample $\{O_i : B_n(i) = 1\}$. Let $q_n = \sum_{i=1}^n B_n(i)/n$ denote the proportion of observations in the validation sample, and we assume that $q < q_n \leq 0.5$ for some $q > 0$. We also assume that this random vector B_n has only V possible realizations for a $V < \infty$. In addition, P^1_{n,B_n}, P^0_{n,B_n} will denote the empirical probability distributions of the validation and training sample, respectively. Thus, the cross-validated risk of an estimator $\hat{\bar{Q}} : \mathcal{M}_{np} \rightarrow \bar{\mathcal{Q}}_n$ of $\bar{Q}_0$ is defined as $E_{B_n} P^1_{n,B_n} L_1(\hat{\bar{Q}}(P^0_{n,B_n}))$.

Entropy bounds for Super Learner II: For $M_1 \in \mathbb{R}^{k_1}$ and $M_2 \in \mathbb{R}^{k_2}$, we define

$$\begin{aligned} \bar{\mathcal{Q}}_{n,M_1} &\equiv \bar{\mathcal{Q}}_n \cap \mathcal{F}_{v,M_1} \\ \bar{\mathcal{G}}_{n,M_2} &\equiv \bar{\mathcal{G}}_n \cap \mathcal{F}_{v,M_2}, \end{aligned}$$

the sub-parameter spaces of $\bar{Q}, \bar{G}$ under model $\mathcal{M}_n$ obtained by only including the functions for which all its components have variation norm bounded by the corresponding constants in the vectors M_1, M_2. By our Key Assumption we have that if $n > N_0$, $M_1 > \| \bar{Q}_0 \|_v$ and $M_2 > \| \bar{G}_0 \|_v$, then $\bar{Q}_0 \in \bar{\mathcal{Q}}_{n,M_1}$ and $\bar{G}_0 \in \bar{\mathcal{G}}_{n,M_2}$. Let $\sup_\Lambda \log N(\epsilon M_1, \bar{\mathcal{Q}}_{n,M_1}, L^2(\Lambda))$ and $\sup_\Lambda \log N(\epsilon M_2, \bar{\mathcal{G}}_{n,M_2}, L^2(\Lambda))$ be the k_1 and k_2-dimensional universal log covering numbers as a function of $\epsilon \in (0,1)$ for $\bar{\mathcal{Q}}_{n,M_1}$ and $\bar{\mathcal{G}}_{n,M_2}$, respectively. We remind the reader that a covering number $N(\epsilon, \mathcal{F}, L^2(\Lambda))$ is defined as the number of balls of size ϵ w.r.t. $L^2(\Lambda)$-norm that are needed to cover the set $\mathcal{F}$ of functions embedded in $L^2(\Lambda)$.

The minimal rate of our second super learner II of $\bar{Q}_0, \bar{G}_0$ relies on the following entropy bounds. Let $\alpha_1 \in \mathbb{R}^{k_1}$ and $\alpha_2 \in \mathbb{R}^{k_2}$ be vectors satisfying: for some $C < \infty$ (allowed to depend on M_1, M_2, but not on ϵ)

$$\begin{aligned} \sup_\Lambda \log^{0.5} N(\epsilon, \bar{\mathcal{Q}}_{n,M_1}, L^2(\Lambda)) &\leq C\epsilon^{-(1-\alpha_1)} \\ \sup_\Lambda \log^{0.5} N(\epsilon, \bar{\mathcal{G}}_{n,M_2}, L^2(\Lambda)) &\leq C\epsilon^{-(1-\alpha_2)}. \end{aligned}$$

Entropy bounds for Super Learner I: The minimal rate of our first super learner I of $\bar{Q}_0, \bar{G}_0$ relies on the following entropy bounds α_1^*, α_2^* which are essentially the same as α_1, α_2. Let $\alpha_1^* \in \mathbb{R}^{k_1}_{\geq 0}$ and $\alpha_2^* \in \mathbb{R}^{k_2}_{\geq 0}$ be such that for some $C < \infty$

$$\begin{aligned} \sup_\Lambda \log^{0.5}(N(\epsilon, L_1(\bar{\mathcal{Q}}^*_{n,M_1}), L^2(\Lambda)) &< C\epsilon^{-(1-\alpha_1^*)} \\ \sup_\Lambda \log^{0.5}(N(\epsilon, L_2(\bar{\mathcal{G}}^*_{n,M_2}), L^2(\Lambda)) &< C\epsilon^{-(1-\alpha_2^*)}, \end{aligned}$$

where $L_1(\bar{\mathcal{Q}}^*_{n,M_1}) = \{L_1(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}^*_{n,M_1}\}$, $L_2(\bar{\mathcal{G}}^*_{n,M_2}) = \{L_2(\bar{G}) : \bar{G} \in \bar{\mathcal{G}}^*_{n,M_2}\}$, and

$$\begin{aligned}\bar{\mathcal{Q}}^*_{n,M_1} &\equiv \{\bar{Q} \in \bar{\mathcal{Q}}_n :\parallel L_1(\bar{Q}) \parallel_v < M_1\} \\ \bar{\mathcal{G}}^*_{n,M_2} &\equiv \{\bar{G} \in \bar{\mathcal{G}}_n :\parallel L_2(\bar{G}) \parallel_v < M_2\}.\end{aligned}$$

By Corollary 2.6.12 in van der Vaart, Wellner (1996), we have that the universal covering number of $\mathcal{F}_{v,M}$ is bounded as follows:

$$\sup_{\Lambda} \log^{0.5} N(\epsilon, \mathcal{F}_{v,M}, L^2(\Lambda)) \leq C\epsilon^{-(1-\alpha(d))},$$

where $\alpha(d) = 1/(d+1)$. Let $d_1 \in \mathbb{N}^{k_1}_{>0}$ be the vector of integers indicating the dimension of the domain of $\bar{Q} = (Q_1, \ldots, Q_{k_1})$, and similarly, let $d_2 \in \mathbb{R}^{k_2}_{>0}$ be the vector of integers indicating the dimension of the domain of $\bar{G} = (G_1, \ldots, G_{k_2})$. Thus, we have that $\max(\alpha_1, \alpha^*_1) \geq \alpha(d_1)$ and $\max(\alpha_2, \alpha^*_2) \geq \alpha(d_2)$.

Minimal rate for Super learner II: Let $C(m_1, m_2) = m_1 + m_2^2$. The minimal rates $r_{Q,1:k_1}(n) \in \mathbb{R}^{k_1}$ and $r_{G,1:k_2}(n) \in \mathbb{R}^{k_2}$ of our super-learner II of $\bar{Q}_0$ and $\bar{G}_0$ w.r.t. the loss-based dissimilarities $d_{01}(Q, Q_0)$ and $d_{02}(G, G_0)$ are given by:

$$\begin{aligned} r_{Q,1:k_1}(n) &= O\left(n^{-\frac{1}{4-2\alpha_1}} C(M_{1Q,n}, M_{2Q,n})^{\frac{1}{4-2\alpha_1}}\right) \\ r_{G,1:k_2}(n) &= O\left(n^{-\frac{1}{4-2\alpha_2}} C(M_{1G,n}, M_{2G,n})^{\frac{1}{4-2\alpha_2}}\right). \end{aligned}$$

We already defined $r_{Q,k_1+1}(n)$ and $r_{G,k_2+1}(n)$ as the rates of the estimators $\hat{Q}_{k_1+1}, \hat{G}_{k_2+1}$ of the easy parameters Q_{k_1+10}, G_{k_2+10}. This defines $r_Q(n) \in \mathbb{R}^{k_1+1}$ and $r_G(n) \in \mathbb{R}^{k_2+1}$.

Minimal rate for Super Learner I: The minimal rates $r_{Q,MLE,1:k_1}(n) \in \mathbb{R}^{k_1}$ and $r_{G,MLE,1:k_2}(n) \in \mathbb{R}^{k_2}$ of our super-learner I of $\bar{Q}_0$ and $\bar{G}_0$ w.r.t. the loss-based dissimilarities $d_{01}(Q, Q_0)$ and $d_{02}(G, G_0)$ are given by:

$$\begin{aligned} r_{Q,MLE,1:k_1}(n) &= n^{-(0.5+\alpha^*_1/4)} \\ r_{G,MLE,1:k_2}(n) &= n^{-(0.5+\alpha^*_2/4)}. \end{aligned}$$

Let $r_{Q,MLE,k_1+1} = r_{Q,k_1+1}$ and $r_{G,MLE,k_2+1} = r_{G,k_2+1}$ be the rates of the simple estimators $\hat{Q}_{k_1+1}$ and $\hat{G}_{k_2+1}$ of Q_{k_1+10} and G_{k_2+10}, respectively. This defines $r_{Q,MLE} \in \mathbb{R}^{k_1+1}$ and $r_{G,MLE} \in \mathbb{R}^{k_2+1}$.

Guaranteed minimal rate faster than $n^{-1/4}$: Since $\alpha_1, \alpha^*_1, \alpha_2, \alpha^*_2$ are all larger than $\alpha(d_1), \alpha(d_1), \alpha(d_2), \alpha(d_2)$, respectively, it follows that all four rates are faster than $n^{-1/4}$ if the model $\mathcal{M}_n$ grows at a slow enough rate to $\mathcal{M}$. Note also that for most sequence of models $\mathcal{M}_n$, the minimal rates $r_Q(n), r_G(n)$ for super-learner II are slightly better than the minimal rates $r_{Q,MLE}(n), r_{G,MLE}(n)$ for super-learner 1.

Entropy bound for estimated efficient influence curve: Let $\alpha^* \in \mathbb{R}_{>0}$ be chosen so that for $\mathcal{F}_n = \{D^*(Q,G) : Q \in \mathcal{Q}_n, G \in \mathcal{G}_n\}/M_{D^*v,n}$ with envelope $F_n < M_{D^*,n}/M_{D^*v,n}$ we have

$$\sup_{\Lambda} \sqrt{\log(1 + N(\epsilon \parallel F_n \parallel_{P_0}, \mathcal{F}_n, L^2(\Lambda)))} = O\left(\frac{1}{\epsilon^{1-\alpha^*}}\right).$$

By the same argument as above, we have $\alpha^* > 1/(d+1)$, where d is the dimension of O.

Entropy bound for controlling TMLE update: Suppose that $\tilde{\alpha}_1 \in \mathbb{R}^{k_1}_{>0}$ is chosen so that for $\mathcal{F}_{1n} = \{L_1(Q) : Q \in \mathcal{Q}_n\}/M_{L_1(Q)v,n}$ with envelope $F_{1n} < M_{1Q,n}/M_{L_1(Q)v,n}$ we have

$$\sup_{\Lambda} \sqrt{\log(1 + N(\epsilon \parallel F_{1n} \parallel_{P_0}, \mathcal{F}_{1n}, L^2(\Lambda)))} = O\left(\frac{1}{\epsilon^{1-\tilde{\alpha}_1}}\right).$$

Reduction to a single entropy bound for each nuisance parameter: Due to the very minor differences (if any) between α_1, α_1^*, $\tilde{\alpha}_1$, there is typically no loss to select all three equal to the an upper bound for all three so that $\alpha_1 = \alpha_1^* = \tilde{\alpha}_1$. Similarly, one would select α_2 and α_2^* equal to each other. Therefore, when reading this article, the reader can just replace $(\alpha_1, \alpha_1^*, \tilde{\alpha}_1)$ and (α_2, α_2^*) by a single $\alpha_1 \in \mathbb{R}^{k_1}$ and $\alpha_2 \in \mathbb{R}^{k_2}$, respectively.

3 Super Learner I

3.1 An MLE restricting the variation norm

Our goal is to construct an estimator $\hat{\bar{Q}} : \mathcal{M}_{np} \rightarrow \bar{\mathcal{Q}}_n$ of $\bar{Q}_0 = \bar{Q}(P_0) = \arg\min_{\bar{Q} \in \bar{\mathcal{Q}}} P_0 \bar{L}_1(\bar{Q})$ so that $d_{01}(\bar{Q}_n, \bar{Q}_0) = o_P(n^{-1/4})$. The following Lemma defines such an estimator.

Lemma 1 *For a given vector $M \in \mathbb{R}^{k_1}_{\geq 0}$ of constants, let $\bar{\mathcal{Q}}^*_{n,M} \subset \{\bar{Q} \in \bar{\mathcal{Q}}_n : \parallel L_1(\bar{Q}) \parallel_v \leq M\} \subset \mathcal{F}_{v,M}$ be all functions in the parameter space $\bar{\mathcal{Q}}_n$ for $\bar{Q}_{0n}$ for which the variation norm of its loss is smaller than $M < \infty$. (In this definition one can also incorporate some extra M-constraints, as long as $\bar{\mathcal{Q}}^*_{n,M=\infty} = \bar{\mathcal{Q}}_n$.) If $M > \max(M_{Q,v}, M_{L_1(Q),v})$, then $\bar{\mathcal{Q}}^*_{n,M} = \bar{\mathcal{Q}}_n$. Let $\bar{Q}^{M*}_{0n} \in \bar{\mathcal{Q}}^*_{n,M}$ be so that $P_0 \bar{L}_1(\bar{Q}^{M*}_{0n}) = \inf_{\bar{Q} \in \bar{\mathcal{Q}}^*_{n,M}} P_0 \bar{L}_1(\bar{Q})$. Assume that for a fixed $M < \infty$,*

$$M_{2Q,M} \equiv \limsup_{n \rightarrow \infty} \sup_{\bar{Q} \in \bar{\mathcal{Q}}^*_{n,M}} \frac{\parallel L_1(\bar{Q}) - L_1(\bar{Q}^{M*}_{0n}) \parallel_{P_0}}{\{d_{10}(\bar{Q}, \bar{Q}^{M*}_{0n})\}^{0.5}} < \infty.$$

Consider an estimator $\bar{Q}_n^M$ for which

$$P_n L_1(\bar{Q}_n^M) = \inf_{\bar{Q} \in \bar{\mathcal{Q}}^*_{n,M}} P_n L_1(\bar{Q}) + r_n,$$

where $r_n = o_P(1/\sqrt{n})$. Then

$$0 \leq d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^{M*}) \leq -(P_n - P_0)\{L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^{M*})\} + r_n, \qquad (1)$$

and

$$d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^{M*}) = O_P(r^2_{\bar{Q},MLE}(n)) + r_n.$$

Proof: In this proof we suppress the * in the notation. We have

$$\begin{array}{rcl} 0 & \leq & d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) = P_0\{L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)\} \\ & = & -(P_n - P_0)\{L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)\} + P_n\{L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)\} \\ & \leq & -(P_n - P_0)\{L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)\} + r_n, \end{array}$$

which proves (1). Since $L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)$ falls in a P_0-Donsker class $\mathcal{F}_{v,M}$, it follows that the right-hand side is $O_P(1/\sqrt{n})$, and thus $d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) = O_P(n^{-1/2})$. Since $M_{2,Q,M} < \infty$, this also implies that $\| L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M) \|^2_{P_0} = O_P(1/\sqrt{n})$. By empirical process theory we have that $\sqrt{n}(P_n - P_0)f_n \rightarrow_p 0$ if f_n falls in a P_0-Donsker class with probability tending to 1, and $P_0 f_n^2 \rightarrow_p 0$ as $n \rightarrow \infty$. Applying this to $f_n = L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)$ shows that $(P_n - P_0)(L_1(\bar{Q}_n^M) - L(\bar{Q}_{0n}^M)) = o_P(1/\sqrt{n})$, which proves $d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) = o_P(1/\sqrt{n})$.

We now apply Lemma 9 with $\mathcal{F}_n = \{L_1(\bar{Q}) - L_1(\bar{Q}_{0n}^M) : \bar{Q} \in \bar{\mathcal{Q}}_{n,M}\}$, $\alpha = \alpha_1^*$, envelope bound $M_n = M$ and $r_0(n) = n^{-1/4}$, which proves that

$$\mid \sqrt{n}(P_n - P_0)f_n \mid = O_P(n^{-\alpha_1/4}).$$

This proves $d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) = O_P(n^{-(0.5+\alpha_1/4)})$. □

3.2 Super-Learning: A cross-validated MLE tuning the variation norm of the fit.

Defining the library of candidate estimators: For an $M \in \mathbb{R}^{k_1}_{>0}$, let $\hat{\bar{Q}}^*_M : \mathcal{M}_{np} \rightarrow \bar{\mathcal{Q}}^*_{n,M} \subset \mathcal{F}_{v,M}$ be the above MLE satisfying $d_{01}(\bar{Q}_{n,M} = \hat{\bar{Q}}_M(P_n), \bar{Q}_{0n}^{M*}) = O_P(r^2_{\bar{Q},MLE}(n))$. Let $\mathcal{K}_{1,n,v}$ be an ordered collection $M_1^n < M_2^n < \ldots < M_{K_{1,n,v}}$ of k_1-dimensional constants, and consider the corresponding collection of $K_{1,n,v}$ candidate estimators $\hat{\bar{Q}}_M$ with $M \in \mathcal{K}_{1,n,v}$. We assume that this index set $\mathcal{K}_{1,n,v}$ is increasing in n and that $\limsup_{n\rightarrow\infty} M_{K_{1,n,v}} =$

$\max(M_{Q,v}, M_{L_1(Q),v})$. Note that for all $M \in \mathcal{K}_{1,n,v}$ with $M >\| L_1(\bar{Q}_0) \|_v$, we have that $d_{01}(\hat{\bar{Q}}_M(P_n), \bar{Q}_0) = O_P(r^2_{\bar{Q},MLE}(n))$. In addition, let $\hat{\bar{Q}}_j : \mathcal{M}_{np} \rightarrow \mathcal{Q}_n$, $j \in \mathcal{K}_{1,n,a}$ be an additional collection of $K_{1,n,a}$ estimators of $\bar{Q}_0$. For example, these candidate estimators could include a variety of parametric model as well as machine learning based estimators. This defines an index set $\mathcal{K}_{1,n} = \mathcal{K}_{1,n,v} \cup \mathcal{K}_{1,n,a}$ representing a collection of $K_{1n} = K_{1,n,v} + K_{1,n,a}$ candidate estimators $\{\hat{\bar{Q}}_k : k \in \mathcal{K}_{1n}\}$.

Super Learner I: We define the cross-validation selector as the index

$$k_{1n} = \hat{K}_1(P_n) = \arg\min_{k \in \mathcal{K}_{1n}} E_{B_n} P_n L_1(\hat{\bar{Q}}_k(P^0_{n,B_n}))$$

that minimizes the cross-validated risk $E_{B_n} P_n L_1(\hat{\bar{Q}}_k(P^0_{n,B_n}))$ over all choices k of candidate estimators. Our proposed super-learner is defined by

$$\bar{Q}_n = \hat{\bar{Q}}(P_n) = E_{B_n} \hat{\bar{Q}}_{k_{1n}}(P^0_{n,B_n}). \tag{2}$$

The following lemma proves that the super-learner $\hat{\bar{Q}}(P_n)$ converges to $\bar{Q}_0$ at least at the rate $r_{\bar{Q},MLE}(n)$: $d_{01}(\hat{\bar{Q}}(P_n), \bar{Q}_0) = O_P(r_{\bar{Q},MLE}(n))$. This lemma also shows that the super-learner is either asymptotically equivalent with the oracle selected candidate estimator, or achieves the parametric rate $1/n$ of a correctly specified parametric model.

Lemma 2 *Let λ_1 be chosen so that $r^2_{\bar{Q},MLE}(n) = O(n^{-\lambda_1})$. We have*

$$d_{01}(\bar{Q}_n, \bar{Q}_{0n}) = O_P(n^{-\lambda_1}) + O_P\left(C(M_{1Q,n}, M_{2Q,n}, \delta)\frac{\log K_{1n}}{n}\right), \tag{3}$$

where $C(M_1, M_2, \delta) = 2(1+\delta)^2(2M_1/3 + M_2^2/\delta)$. In addition, we have, for any $\delta > 0$,

$$\begin{aligned} d_{01}(\bar{Q}_n, \bar{Q}_{0n}) &\leq (1+2\delta)E_{B_n} \min_{k \in \mathcal{K}_{1n}} d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_{0n}) \\ &\quad + O_P\left(C(M_{1Q,n}, M_{2Q,n}, \delta)\frac{\log K_{1n}}{n}\right). \end{aligned}$$

If for a fixed $\delta > 0$, $C(M_{1Q,n}, M_{2Q,n}, \delta)\log K_{1n}/n$ divided by $E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_{0n})$ is $o_P(1)$, then

$$\frac{d_{01}(\hat{\bar{Q}}(P_n), \bar{Q}_{0n})}{E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_{0n})} - 1 = o_P(1).$$

If for a fixed $\delta > 0$, $E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_{0n}) = O_P(C(M_{1Q,n}, M_{2Q,n}, \delta) \log K_{1n}/n)$, *then*

$$d_{01}(\hat{\bar{Q}}(P_n), \bar{Q}_{0n}) = O_P\left(\frac{C(M_{1n}, M_{2n}, \delta) \log K_{1n}}{n}\right).$$

The proof of this lemma is a simple corollary of the finite sample oracle inequality for cross-validation (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006), also presented in Lemma 8. It uses the convexity of the loss function to bring the E_{B_n} inside the loss-based dissimilarity.

Super-Learner I of $\bar{G}_0$: Similarly, we can define such a super-learner of G_0. For an $M \in \mathbb{R}^{k_2}_{>0}$, let $\hat{\bar{G}}_M : \mathcal{M}_{np} \rightarrow \bar{\mathcal{G}}^*_{n,M} \subset \mathcal{F}_{v,M}$ be the MLE of the previous subsection for which $d_{02}(\bar{G}_{n,M} = \hat{\bar{G}}_M(P_n), \bar{G}^{M*}_{0n}) = O_P(r^2_{\bar{G},MLE}(n))$. Let $\mathcal{K}_{2,n,v}$ be an ordered collection of k_2-dimensional constants, and consider the corresponding collection of candidate estimators $\hat{\bar{G}}_M$ with $M \in \mathcal{K}_{2,n,v}$. We assume the index set $\mathcal{K}_{2,n,v}$ is increasing in n and that $\limsup_{n\rightarrow\infty} M_{K_{2,n,v}} = \max(M_{G,v}, M_{L_2(G),v})$. Note that for all $M \in \mathcal{K}_{2,n,v}$ with $M >\parallel L_2(\bar{G}_0) \parallel_v$, we have that $d_{02}(\hat{\bar{G}}_M(P_n), \bar{G}_0) = O_P(n^{-\lambda_2})$. In addition, let $\hat{\bar{G}}_j : \mathcal{M}_{np} \rightarrow \bar{\mathcal{G}}_n$, $j \in \mathcal{K}_{2,n,a}$, be an additional collection of $K_{2,n,a}$ estimators of G_0. This defines a collection of $K_{2n} = K_{2,n,v} + K_{2,n,a}$ candidate estimators $\{\hat{\bar{G}}_k : k \in \mathcal{K}_{2n}\}$ of $\bar{G}_0$.

We define the cross-validation selector as the index

$$k_{2n} = \hat{K}_2(P_n) = \arg\min_{k\in\mathcal{K}_{2n}} E_{B_n} P_n L_1(\hat{\bar{G}}_k(P^0_{n,B_n}))$$

that minimizes the cross-validated risk $E_{B_n} P_n L_2(\hat{\bar{G}}_k(P^0_{n,B_n}))$ over all choices k of candidate estimators. Our proposed super-learner of $\bar{G}_0$ is defined by

$$\bar{G}_n = \hat{\bar{G}}(P_n) = E_{B_n} \hat{\bar{G}}_{k_n}(P^0_{n,B_n}). \tag{4}$$

The same Lemma 2 applies to this estimator $\hat{\bar{G}}(P_n)$ of $\bar{G}_0$.

Lemma 3 *Let λ_2 be chosen so that $r^2_{\bar{G},MLE}(n) = O(n^{-\lambda_2})$. We have*

$$d_{02}(\hat{\bar{G}}(P_n), \bar{G}_{0n}) = O_P(n^{-\lambda_2}) + O_P\left(C(M_{1G,n}, M_{2G,n}, \delta)\frac{\log K_{2n}}{n}\right). \tag{5}$$

In addition, we have, for any $\delta > 0$,

$$\begin{aligned} d_{02}(\bar{G}_n, \bar{G}_{0n}) &\leq (1 + 2\delta) E_{B_n} \min_{k\in\mathcal{K}_{2n}} d_{02}(\hat{\bar{G}}_k(P^0_{n,B_n}), \bar{G}_{0n}) \\ &\quad + O_P\left(C(M_{1G,n}, M_{2G,n}, \delta)\frac{\log K_{2n}}{n}\right), \end{aligned}$$

for a constant $C(M_1, M_2, \delta) = 2(1+\delta)^2(2M_1/3 + M_2^2/\delta)$. *If for a fixed* $\delta > 0$, $C(M_{1G,n}, M_{2G,n}, \delta)\log K_{2n}/n$ *divided by* $E_{B_n}\min_k d_{02}(\hat{\bar{G}}_k(P^0_{n,B_n}), \bar{G}_{0n})$ *is* $o_P(1)$, *then*

$$\frac{d_{02}(\hat{\bar{G}}(P_n), \bar{G}_{0n})}{E_{B_n}\min_k d_{02}(\hat{\bar{G}}_k(P^0_{n,B_n}), \bar{G}_{0n})} - 1 = o_P(1).$$

If for a fixed $\delta > 0$, $E_{B_n}\min_k d_{02}(\hat{\bar{G}}_k(P^0_{n,B_n}), \bar{G}_{0n}) = O_P(C(M_{1G,n}, M_{2G,n}, \delta)\log K_{2n}/n)$, *then*

$$d_{02}(\hat{\bar{G}}(P_n), \bar{G}_{0n}) = O_P\left(\frac{C(M_{1G,n}, M_{2G,n}, \delta)\log K_{1n}}{n}\right).$$

4 Super Learner II

For an $M \in \mathbb{R}^{k_1}$ we define $\bar{\mathcal{Q}}_{n,M} = \bar{\mathcal{Q}}_n \cap \mathcal{F}_{v,M}$, but one could impose additional M-constraints as long as they disappear as $M \rightarrow \infty$. Accordingly, we define

$$\bar{Q}^{M_1}_{0n} \quad = \quad \arg\min_{\bar{Q}\in\bar{\mathcal{Q}}_{n,M_1}} P_0 L_1(\bar{Q}).$$

If $n > N_0$ and $M_1 > \parallel \bar{Q}_0 \parallel_v$, then $\bar{Q}^{M_1}_{0n} = \bar{Q}_0$.

ϵ-nets: For an $M \in \mathbb{R}^{k_1}$, let $\bar{\mathcal{Q}}_{n,M,\epsilon} \subset \bar{\mathcal{Q}}_{n,M}$ be an ϵ-net (i.e., a finite subset) of $\bar{\mathcal{Q}}_{n,M}$ in the sense that there exists a $\bar{Q}^M_{0,n,\epsilon} \in \bar{\mathcal{Q}}_{n,M,\epsilon}$ so that $\sqrt{d_{01}(\bar{Q}^M_{0,n,\epsilon}, \bar{Q}^M_{0,n})} \leq \epsilon$. Suppose that for a fixed M, uniformly in n and uniformly over $\bar{Q} \in \bar{\mathcal{Q}}_{n,M}$, the loss-based dissimilarity $d_{01}(\bar{Q}, \bar{Q}^M_{0,n}))$ is not a stronger norm than the $L^2(P_0)$-norm:

$$C(M) = \lim\sup_{n\rightarrow\infty}\sup_{\bar{Q}\in\bar{\mathcal{Q}}_{n,M}} \frac{d^{0.5}_{01}(\bar{Q}, \bar{Q}^M_{0n})}{\parallel \bar{Q} - \bar{Q}^M_{0n} \parallel_{P_0}} < \infty. \tag{6}$$

This allows us to conclude the following. Suppose that $N_{n,M}(\epsilon) = \sup_\Lambda N(\epsilon, \bar{\mathcal{Q}}_{n,M}, L^2(\Lambda))$ as a function in ϵ is given. Then it follows that there exists a finite subset $\bar{\mathcal{Q}}_{n,M,\epsilon}$ of $\bar{\mathcal{Q}}_{n,M}$ of size $N_{n,M}(\epsilon)$ so that w.r.t. all $L^2(\Lambda)$-norms, the distance between an element in $\bar{\mathcal{Q}}_{n,M}$ and the finite set $\bar{\mathcal{Q}}_{n,M,\epsilon}$ is smaller than ϵ. Then, by assumption (6), we know that this finite subset $\bar{\mathcal{Q}}_{n,M,\epsilon}$ of $\bar{\mathcal{Q}}_{n,M}$ also approximates $\bar{Q}^M_{0,n}$ within dissimilarity ϵ w.r.t. d_{01}:

$$\sup_{P_0\in\mathcal{M}}\min_{\bar{Q}\in\bar{\mathcal{Q}}_{n,M,\epsilon}} \sqrt{d_{01}(\bar{Q}, \bar{Q}^M_{0n})} \leq C(M)\epsilon.$$

This proves that we can guarantee the desired ϵ-approximation of $\bar{Q}^M_{0n}$ with a finite net of size bounded by $N_{n,M}(\epsilon)$, and in the following it is assumed that indeed the size of $\bar{\mathcal{Q}}_{n,M,\epsilon}$ is bounded by $N_{n,M}(\epsilon)$. By definition of α_1, we have $\sup_\Lambda \sqrt{\log N_{n,M}(\epsilon)} = O(\epsilon^{-(1-\alpha_1)})$, and we know that $\alpha_1 \leq \alpha(d_1) = 1/(d_1+1)$.

Candidate estimators: Let

$$\bar{Q}_{n,M,\epsilon} = \arg \min_{\bar{Q} \in \bar{\mathcal{Q}}_{n,M,\epsilon}} P_n L_1(\bar{Q}) \tag{7}$$

be the ϵ-MLE. Let $\mathcal{K}_{1n,v} \times \mathcal{E}_n$ be the cartesian product of a finite set $\mathcal{K}_{1n,v}$ of constants M and a finite set $\mathcal{E}_n$ of ϵ-values. Let $K_{1n,v}$ be the size of $\mathcal{K}_{1n,v}$, and E_n is the size of $\mathcal{E}_n$. We assume that $E_n = n^p$ for some $p > 0$, which allows us to approximate the minimum over all ϵ-values by a minimum over $\mathcal{E}_n$ without affecting the theoretical performance of the estimator. Let $M_{n,v} \equiv \max_{M \in \mathcal{K}_{1n,v}} M$ be the largest M value in $\mathcal{K}_{1n,v}$. It is assumed that $M_{n,v}$ equals or exceeds the upper bound $M_{Q,v,n}$ of the variation norm of a $\bar{Q} \in \bar{\mathcal{Q}}_n$. For each $(M, \epsilon) \in \mathcal{K}_{1n,v} \times \mathcal{E}_n$, this MLE (7) is a candidate estimator $\hat{\bar{Q}}_{M,\epsilon} : \mathcal{M}_{np} \rightarrow \bar{\mathcal{Q}}_{n,M}$. In addition, let $\hat{\bar{Q}}_j : \mathcal{M}_{np} \rightarrow \bar{\mathcal{Q}}_n$, $j \in \mathcal{J}_n$, be an additional set of candidate estimators of $\bar{Q}_{0n}$. Let $\mathcal{K}_{1n} = \mathcal{K}_{1n,v} \times \mathcal{E}_n \cup \mathcal{J}_n$ be the index set of the resulting total set of candidate estimators, and let $K_{1n} = K_{1n,v} + J_n$ be the total number of candidate estimators. This defines now our library of candidate estimators $\hat{\bar{Q}}_k$, $k \in \mathcal{K}_{1n}$.

Super-Learner II: Let $k_{1n} \in \mathcal{K}_{1n}$ be the cross-validation selector of k:

$$k_{1n} = \arg \min_{k \in \mathcal{K}_{1n}} E_{B_n} P^1_{n,B_n} L_1(\hat{\bar{Q}}_k(P^0_{n,B_n})).$$

The super-learner is now defined as:

$$\hat{\bar{Q}}(P_n) = E_{B_n} \hat{\bar{Q}}_{k_{1n}}(P^0_{n,B_n}). \tag{8}$$

For example, if $\mathcal{J}_n$ is empty, then $k_{1n} = (M_n, \epsilon_n)$ is the cross-validation selector of (M, ϵ):

$$(M_n, \epsilon_n) = \arg \min_{(M,\epsilon) \in \mathcal{K}_{1n,v}} E_{B_n} P^1_{n,B_n} L_1(\hat{\bar{Q}}_{n,M,\epsilon}(P^0_{n,B_n})),$$

and the above super-learner is given by

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

We will refer to this latter estimator as the cross-validated ϵ-net MLE.

The following lemma proves that the latter $\hat{\bar{Q}}_e : \mathcal{M}_{np} \rightarrow \bar{\mathcal{Q}}_n$ converges to $\bar{Q}_0$ w.r.t. d_{01} at the minimax rate $r_{\bar{Q}}(n)$:

Lemma 4 *We have the following finite sample inequality: for each $\delta > 0$, we have*

$$\begin{aligned} E_0 d_{01}(E_{B_n}\hat{\hat{\bar{Q}}}_{n,M_n,\epsilon_n}(P^0_{n,B_n}), \bar{Q}_{0n}) &\leq \\ (1+2\delta)\min_{M,\epsilon}\left\{(1+2\delta)\min_{\bar{Q}\in\bar{\mathcal{Q}}_{n,M,\epsilon}} d_{01}(\bar{Q},\bar{Q}_{0n}) + 2C(M_{1Q,n},M_{2Q,n},\delta)\frac{1+\log N_{n,M}(\epsilon)}{nq}\right\} \\ +2C(M_{1Q,n},M_{2Q,n},\delta)\frac{1+\log K_{1n,v}}{nq}. \end{aligned}$$

Let M_{0v} be the smallest $M \in \mathcal{K}_{1n,v}$ that is larger than $\parallel \bar{Q}_0 \parallel_v$. Assume an ϵ-net $\bar{\mathcal{Q}}_{n,M,\epsilon}$ whose size $N_{n,M}(\epsilon)$ is bounded by $\log^{0.5} N_{n,M}(\epsilon) = O(\epsilon^{-(1-\alpha_1)})$, whose existence is shown above. The above implies now

$$\begin{aligned} E_0 d_{01}(E_{B_n}\hat{\hat{\bar{Q}}}_{n,M_n,\epsilon_n}(P^0_{n,B_n}), \bar{Q}_{0n}) &\leq \\ (1+2\delta)\min_{\epsilon}\left\{(1+2\delta)\epsilon^2 + 2C(M_{1Q,n},M_{2Q,n},\delta)\frac{1+\log N_{n,M_{0,v}}(\epsilon)}{nq}\right\} \\ +2C(M_{1Q,n},M_{2Q,n},\delta)\frac{1+\log K_{1n,v}}{nq}. \end{aligned}$$

By definition of $r_Q(n)$, in particular, if

$$C(M_{1Q,n},M_{2Q,n},\delta)\frac{1+\log K_{1n,v}}{nq} = O(r^2_Q(n)),$$

then

$$E_0 d_{01}(\hat{\hat{\bar{Q}}}_e(P_n), \bar{Q}_0) = O(r^2_Q(n)).$$

Under the same condition, we also have

$$E_0 d_{01}(\hat{\hat{\bar{Q}}}(P_n), \bar{Q}_0) = O(r^2_Q(n)).$$

Analogue to above, we can also present the super-learner $\hat{\hat{\bar{G}}}$ of $\bar{G}_0$:

$$\hat{\hat{\bar{G}}}(P_n) = E_{B_n}\hat{\hat{\bar{G}}}_{k_n}(P^0_{n,B_n}). \qquad (9)$$

Of course, we can present the same result for this super-learner and the cross-validated ϵ-net MLE $\hat{\hat{\bar{G}}}_e$ of $\bar{G}_0$.

Lemma 5 *We have the following finite sample inequality: for each $\delta > 0$, we have*

$$\begin{aligned} E_0 d_{02}(E_{B_n}\hat{\hat{\bar{G}}}_{n,M_n,\epsilon_n}(P^0_{n,B_n}), \bar{G}_{0n}) &\leq \\ (1+2\delta)\min_{M,\epsilon}\left\{(1+2\delta)\min_{\bar{G}\in\bar{\mathcal{G}}_{n,M,\epsilon}} d_{02}(\bar{G},\bar{G}_{0n}) + 2C(M_{1G,n},M_{2G,n},\delta)\frac{1+\log N_{n,M}(\epsilon)}{nq}\right\} \\ +2C(M_{1G,n},M_{2G,n},\delta)\frac{1+\log K_{2n,v}}{nq}. \end{aligned}$$

Let M_{0v} be the smallest $M \in \mathcal{K}_{2n,v}$ that is larger than $\parallel \bar{G}_0 \parallel_v$ and assume that $\bar{\mathcal{G}}_{n,M,\epsilon}$ is an ϵ-net whose size $N_{n,M}(\epsilon)$ is bounded by $\log^{0.5} N_{n,M}(\epsilon) < C\epsilon^{-(1-\alpha_2)}$. The above implies:

$$\begin{aligned} E_0 d_{02}(E_{B_n}\hat{\bar{G}}_{n,M_n,\epsilon_n}(P^0_{n,B_n}), \bar{G}_{0n}) \quad \leq & \\ (1+2\delta)\min_{\epsilon}\left\{(1+2\delta)\epsilon^2 + 2C(M_{1G,n}, M_{2G,n}, \delta)\frac{1+\log N_{n,M_{0v}}(\epsilon)}{nq}\right\} & \\ +2C(M_{1G,n}, M_{2G,n}, \delta)\frac{1+\log K_{2n,v}}{nq}. & \end{aligned}$$

By the definition of $r_G(n)$, in particular, if

$$C(M_{1G,n}, M_{2G,n}, \delta)\frac{1+\log K_{2n,v}}{nq} = O(r^2_{\bar{G}}(n)),$$

then

$$E_0 d_{02}(\hat{\bar{G}}_e(P_n), \bar{G}_0) = O(r^2_{\bar{G}}(n)).$$

Under the same condition, we also have

$$E_0 d_{02}(\hat{\bar{G}}(P_n), \bar{G}_0) = O(r^2_{\bar{G}}(n)).$$

5 One-step TMLE

5.1 The one-step TMLE

We consider a one-step TMLE defined by an initial estimator $Q_n = \hat{Q}(P_n) \in \mathcal{Q}_n$, $\hat{G}(P_n) \in \mathcal{G}_n$ of Q_0, G_0, and a finite dimensional least favorable submodel $\{Q_{n,\epsilon} : \epsilon\} \subset \mathcal{Q}_n$ of $\mathcal{Q}_n$ through Q_n at $\epsilon = 0$. Specifically, it is assumed that the linear span of the components of its score

$$\left.\frac{d}{d\epsilon}\bar{L}_1(Q_{n,\epsilon})\right|_{\epsilon=0}$$

w.r.t. sum loss $\bar{L}_1(Q) = \sum_{j=1}^{k_1+1} L_{1j}(Q_j)$ contains the efficient influence curve $D^*(Q_n, G_n)$ at (Q_n, G_n).

Let $\epsilon_n = \arg\min_\epsilon P_n \bar{L}_1(Q^0_{n,\epsilon})$ be the MLE, and we define the one-step TMLE of Q_0 as $Q^1_n = Q_{n,\epsilon_n}$. One could iterate this process of updating to construct a final update $Q^*_n = Q^K_n$ for K large enough that solves $P_n D^*(Q^*_n, G_n) = 0$ exactly or numerically. In various examples this iterative TMLE converges in one step in which case $Q^*_n = Q^1_n$.

In this article, we study the one-step TMLE $Q_n^* = Q_n^1$, and our results also apply to the K-th step Q_n^K for a fixed integer $K > 1$ (where K does not depend on the data or on P_0). The one-step TMLE of ψ_0 is defined by the plug-in estimator $\psi_n^* = \Psi(Q_n^*)$.

We assume that

$$P_n D^*(Q_n^*, G_n) = o_P(1/\sqrt{n}). \tag{10}$$

That is, it is assumed that the one-step TMLE already solves the efficient influence curve equation up till an asymptotically negligible approximation error. In (van der Laan and Gruber, 2015) it is shown that one can always construct a so called universal least favorable submodel with a one dimensional ϵ so that $\frac{d}{d\epsilon}\bar{L}_1(Q_{n,\epsilon}^0) = D^*(Q_{n,\epsilon}^0, G_n)$ at each ϵ, so that indeed $P_n D^*(Q_{n,\epsilon_n}^0, G_n) = 0$ (exactly). In addition, as formalized by Lemma 16 in the Appendix, for our choice of initial estimators Q_n, G_n of Q_0, G_0 a one-step TMLE will satisfy (10) for one dimensional local least favorable submodels under weak regularity conditions.

We can establish an asymptotic efficiency theorem for our one-step TMLE for any finite dimensional local least favorable submodel, including a universal least favorable submodel. However, our theorem is developed for a local least favorable submodel of the type $Q_\epsilon = (Q_{1,\epsilon_1}, \ldots, Q_{k_1+1,\epsilon_{k_1+1}})$. By using such a submodel we have $Q_{jn}^* = Q_{jn,\epsilon_n(j)}$ and $\epsilon_n(j) = \arg\min_\epsilon P_n L_{1j}(Q_{jn,\epsilon})$. Thus in this case each Q_{jn} is updated with its own $\epsilon_n(j)$, $j = 1, \ldots, k_1 + 1$. The advantage of such a least favorable submodel is that the one-step update of $\bar{Q}_{jn}$ is not affected by the statistical behavior of the other estimators $\bar{Q}_{ln}$, $l \neq j$: e.g., if one uses a single ϵ, the MLE ϵ_n is very much driven by the worst performing estimator $\bar{Q}_{jn}$. By using such a submodel the rate of convergence of the initial estimator $\bar{Q}_{jn}$ is fully preserved by the TMLE-update step for bounded models, and still well controlled for unbounded models.

A general approach for constructing such a least favorable submodel is the following. Let $D_j^*(P)$ be the efficient influence curve at a P for the parameter $\Psi_{j,P} : \mathcal{M} \to \mathrm{I\!R}$ defined by $\Psi_j(P_1) = \Psi(Q_{-j}(P), Q_j(P_1))$ that sets all the other components of Q_l with $l \neq j$ equal to its true value under P, $j = 1, \ldots, k_1 + 1$. Then, it follows immediately from the definition of pathwise derivative that

$$D^*(P) = \sum_{j=1}^{k_1+1} D_j^*(P),$$

so that, $D^*(P)$ is an element of the linear span of $\{D_j^*(P) : j = 1, \ldots, k_1 + 1\}$. Let $\{Q_{jn,\epsilon(j)} : \epsilon(j)\} \subset \mathcal{Q}_{jn}$ be so that

$$\left.\frac{d}{d\epsilon(j)} L_{1j}(Q_{jn,\epsilon(j)})\right|_{\epsilon(j)=0} = D_j^*(Q_n, G_n),\ j = 1, \ldots, k_1 + 1.$$

That is, $\{Q_{jn,\epsilon(j)} : \epsilon(j)\}$ is a local least favorable submodel at (Q_n, G_n) for the parameter $\Psi_{j,Q_n} : \mathcal{M} \to \mathbb{R}$, $j = 1, \ldots, k_1 + 1$. Now, define $\{Q_{n,\epsilon} : \epsilon\} \subset \mathcal{Q}_n$ by $Q_{n,\epsilon} = (Q_{nj,\epsilon(j)} : j = 1, \ldots, k_1 + 1)$. Then, we have

$$\frac{d}{d\epsilon} \bar{L}(Q_{n,\epsilon}) \bigg|_{\epsilon=0} = (D_j^*(Q_n, G_n) : j = 1, \ldots, k_1 + 1)^\top,$$

so that the submodel is indeed a local least favorable submodel.

The only key ingredient of this type of submodel we rely upon is that the MLE $\epsilon_n \in \mathbb{R}^{k_1+1}$ is now defined by $\epsilon_n(j) = \arg\min_{\epsilon(j)} P_n L_{1j}(Q_{jn,\epsilon(j)})$, $j = 1, \ldots, k_1 + 1$, so that indeed the update for Q_{jn} is not affected by the other estimators Q_{ln}, $l \neq j$. Since $\hat{Q}_{k_1+1}(P_n)$ is typically an MLE of Q_{k_1+10}, we would typically have $\epsilon_n(k_1 + 1) = 0$. Lemma 17 provides a sufficient set of minor conditions under which this one-step TMLE will satisfy (10). We will not assume these conditions in our general efficiency theorem below, since there are many examples in which this TMLE solves the efficient influence curve equation exactly without any need to verify these conditions, as in our example.

5.2 Efficiency of the one-step TMLE

If we use super-learner II, then choose λ_1 and λ_2 so that $r_Q^2(n) = O(n^{-\lambda_1})$ and $r_G^2(n) = O(n^{-\lambda_2})$. Our initial super-learner II satisfies that $d_{01}(Q_n, Q_0) = O_P(r_Q^2(n))$ and $d_{01}(G_n, G_0) = O_P(r_G^2(n))$, where $r_Q(n) = (r_{\bar{Q}}(n), r_{Q,k_1+1}(n))$ and $r_G(n) = (r_{\bar{G}}(n), r_{G,k_2+1}(n))$. If we use super-learner I, then we choose λ_1 and λ_2 so that $r_{Q,MLE}^2(n) = O(n^{-\lambda_1})$ and $r_{G,MLE}^2(n) = O(n^{-\lambda_2})$. Our initial super-learner I satisfies that $d_{01}(Q_n, Q_0) = O_P(r_{Q,MLE}^2(n))$ and $d_{01}(G_n, G_0) = O_P(r_{G,MLE}^2(n))$, where again $r_{Q,MLE}(n) = (r_{\bar{Q},MLE}(n), r_{Q,k_1+1}(n))$ and $r_{G,MLE}(n) = (r_{\bar{G}}(n), r_{G,k_2+1}(n))$.

Let $\lambda_1^*(1 : k_1) = 0.5 + \tilde{\alpha}_1/4$ and $\lambda_1^*(k_1 + 1) = \lambda_1(k_1 + 1)$. Our update Q_n^* conservatively satisfies $d_{01}(\bar{Q}_n^*, \bar{Q}_0) = O_P(n^{-\lambda_1^*})$ (Lemma 13 and Corollary 1). For bounded models we could have set $\lambda_1^* = \lambda_1$ (Lemma 14). We assumed that $\hat{Q}_{k_1+1}$ is not updated by the one-step TMLE, or that its update is not affecting its initial rate. Let $r^2(n) \equiv (n^{-\lambda_1^*}, r_G^2(n)) \in \mathbb{R}^{k_1+k_2+2}$ so that $d_0(P_n^*, P_0) = O_P(r^2(n))$.

Let $r_{D^*,n}$ be such that $\| D^*(Q_n^*, G_n) - D^*(Q_0, G_0) \|_{P_0} = O_P(r_{D^*,n})$, where this rate will be based on knowing $d_0(P_n^*, P_0) = O_P(r^2(n))$, and bounds of the model $\mathcal{M}_n$ that are enforced on our super-learners. Finally, we define the following sequences of constants that control how fast we can let grow $\mathcal{M}_n$

converge to a possibly unbounded model $\mathcal{M}$:

$$\begin{aligned} C_{1n} &= r^{\tilde{\alpha}_1}_{D^*,n,1} M^{1-\tilde{\alpha}_1}_{D^*,n} + r^{2\tilde{\alpha}_1-2}_{D^*,n,1} M^{2-2\tilde{\alpha}_1}_{D^*,n} n^{-0.5} M_{D^*v,n} \qquad (11) \\ C_{2n} &= M^{\tilde{\alpha}_1+\tilde{\alpha}_1^2}_{L_1(Q)v,n} M^{\tilde{\alpha}_1}_{2Q,n} M^{1-\tilde{\alpha}_1^2}_{1Q,n} + n^{-\tilde{\alpha}_1/2} M^{-1+2\tilde{\alpha}_1^2}_{L_1(Q)v,n} M^{2\tilde{\alpha}_1(1-\tilde{\alpha}_1)}_{1Q,n} M^{-2+2\tilde{\alpha}_1}_{2Q,n}. (12) \end{aligned}$$

Theorem 1 *Consider the super-learners $\hat{\hat{Q}}(P_n)$ and $\hat{\hat{G}}(P_n)$ defined by (2) and (4), respectively, or by (8) and (9), respectively. This defines the initial estimators $\hat{Q}(P_n) = (\hat{\hat{Q}}(P_n), \hat{Q}_{k_1+1}(P_n))$ and $\hat{G}(P_n) = (\hat{\hat{G}}(P_n), \hat{G}_{k_2+1}(P_n))$. Consider also the above defined corresponding one-step TMLE Q_{n,ϵ_n} of Q_0, and resulting one-step TMLE $\Psi(Q_n^*)$ of $\Psi(Q_0)$. Let $r^2(n)$ be the above defined vector of rates so that under the assumptions of this theorem we have $d_0(P_n^*, P_0) = O_P(r^2(n))$.*

Assume that

$$\begin{aligned} P_n D^*(Q_n^*, G_n) &= o_P(n^{-0.5}) \qquad (13) \\ \frac{\max(M_{1Q,n}, M_{2Q,n}^2)\log K_{1n}}{n} &= O(n^{-\lambda_1}) \qquad (14) \\ \frac{\max(M_{1G_n}, M_{2G_n}^2)\log K_{2n}}{n} &= O(n^{-\lambda_2}) \qquad (15) \\ n^{-\tilde{\alpha}_1/4} C_{2n} &= O(1) \qquad (16) \\ C_{1n} &= o(1) \qquad (17) \\ R_{20}((Q_n^*, G_n), (Q_0, G_0)) &= o_P(n^{-1/2}), \qquad (18) \end{aligned}$$

where we can use that $d_0(P_n^*, P_0) = O_P(r^2(n))$.

Then, $\Psi(Q_n^*)$ is asymptotically efficient:

$$\Psi(Q_n^*) - \Psi(Q_0) = (P_n - P_0) D^*(Q_0, G_0) + o_P(n^{-0.5}).$$

Proof: Consider the case that our initial estimators are based on Super Learner II. Combining $P_n D^*(Q_n^*, G_n) = o_P(1/\sqrt{n})$ (13) with $\Psi(Q_n^*) - \Psi(Q_0) = -P_0 D^*(Q_n^*, G_n) + R_{20}(Q_n^*, G_n, Q_0, G_0)$ yields the identity:

$$\Psi(Q_n^*) - \Psi(Q_0) = (P_n - P_0) D^*(Q_n^*, G_n) + R_{20}(Q_n^*, G_n, Q_0, G_0) + o_P(1/\sqrt{n}).$$

Under assumptions (14) and (15), by Lemmas 4 and 5, we have $d_0((Q_n, G_n), (Q_0, G_0)) = O_P(r_Q^2(n), r_G^2(n))$. Under the additional assumption (16), Corollary 1 shows that this implies $d_{01}(Q_n^*, Q_0) = O_P(n^{-\lambda_1^*})$. Thus, $d_0(P_n^*, P_0) = O_P(n^{-\lambda_1^*}, n^{-\lambda_2}) = O_P(r^2(n))$. Using this, by assumption (18), we have $R_{20}(P_n^*, P_0) = o_P(n^{-0.5})$. It remains to analyze the empirical process term $(P_n - P_0) D^*(Q_n^*, G_n)$. We have

$$(P_n - P_0) D^*(Q_n^*, G_n) = (P_n - P_0)\{D^*(Q_n^*, G_n) - D^*(Q_0, G_0)\} + (P_n - P_0) D^*(Q_0, G_0).$$

We apply Lemma 11 to the first term on the right-hand side, which proves that the expectation of the absolute value of this first term is bounded by $O_P(C_{1n}/n^{0.5})$, which is thus $o_P(n^{-1/2})$ under assumption (17). This proves $\Psi(Q_n^*)-\Psi(Q_0) = (P_n-P_0)D^*(Q_0,G_0)+o_P(n^{-0.5})$, and thereby the asymptotic efficiency of the one-step TMLE. The proof is the same for the case that our initial estimators are based on Super Learner I, but now we should apply Lemmas 2 and **??** to obtain $d_0((Q_n,G_n),(Q_0,G_0)) = O_P(r_{Q,MLE}(n), r_{G,MLE}(n))$.
□

6 Efficiency of the one-step CV-TMLE

Cross-validated TMLE (CV-TMLE) robustifies the bias-reduction of the TMLE-step by selecting ϵ based on the cross-validated risk (Zheng and van der Laan, 2011; van der Laan and Rose, 2011).

6.1 The CV-TMLE

For a given Q,G, let $\{Q_\epsilon : \epsilon\} \subset \mathcal{Q}_n \subset \mathcal{Q}$ be the same k_1+1-dimensional submodel through Q at $\epsilon=0$ such that the linear span of $\frac{d}{d\epsilon}\bar{L}_1(Q_\epsilon)$ at $\epsilon=0$ includes $D^*(Q,G)$, as presented in the previous section. Let $\hat{Q}:\mathcal{M}_{np}\rightarrow\mathcal{Q}_n$ and $\hat{G}:\mathcal{M}_{np}\rightarrow\mathcal{G}_n$ be our initial estimators. Given a cross-validation scheme $B_n\in\{0,1\}^n$, let $Q_{n,B_n}=\hat{Q}(P^0_{n,B_n})\in\mathcal{Q}_n$ be the super-learner applied to the training sample P^0_{n,B_n}. Let

$$\epsilon_n = \arg\min_\epsilon E_{B_n} P^1_{n,B_n}\bar{L}_1(Q_{n,B_n,\epsilon}),$$

where the submodel $\{Q_{n,B_n,\epsilon} : \epsilon\}$ is the submodel through Q_{n,B_n} at $\epsilon=0$. This submodel uses $\hat{G}(P^0_{n,B_n})$ as an estimator of G_0. Let $Q^*_{n,B_n}=Q_{n,B_n,\epsilon_n}$ be the B_n-specific targeted fit of Q_0. The one-step CV-TMLE of ψ_0 is defined as

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

As with the TMLE in the previous section, we only assume that $E_{B_n}P^1_{n,B_n}D^*(Q^*_{n,B_n},G_{n,B_n}) = o_P(n^{-1/2})$. By Lemma 17 in the Appendix this will hold in great generality for local least favorable submodels, if $d_0((Q_n,G_n),(Q_0,G_0)) = o_P(n^{-0.5})$.

6.2 Efficiency of the one-step CV-TMLE

Let λ_1 and λ_2 be defined as above, so that $d_{01}(\hat{Q}(P_n),Q_0) = O_P(n^{-\lambda_1})$ and $d_{02}(\hat{G}(P_n),G_0) = O_P(n^{-\lambda_2})$. Let $r_{D^*,n}$ be a rate such that for each B_n $\|$

$D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0) \|_{P_0} = o_P(r_{D^*,n})$, where one should use that we already know that $d_{01}(Q^*_{n,B_n}, Q_0) = O_P(n^{-\lambda_1})$ and $d_{02}(G_n, G_0) = O_P(n^{-\lambda_2})$.

Theorem 2 *Consider the super-learners $\hat{Q}(P_n)$ and $\hat{G}(P_n)$ defined by (2) and (4), respectively, or by (8) and (9), respectively. Consider the above defined corresponding one-step CV-TMLE $\psi_n^* = E_{B_n}\Psi(Q_{n,B_n,\epsilon_n})$ of $\Psi(Q_0)$.*

Assume

$$\begin{aligned}
E_{B_n} P^1_{n,B_n} D^*(Q_{n,B_n,\epsilon_n}, G_{n,B_n}) &= o_P(n^{-0.5}) & (19)\\
\frac{\max(M_{1Q,n}, M^2_{2Q,n}) \log K_{1n}}{n} &= O(n^{-\lambda_1}) & (20)\\
\frac{\max(M_{1G_n}, M^2_{2G_n}) \log K_{2n}}{n} &= O(n^{-\lambda_2}) & (21)\\
\| D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0) \|_{P_0} &= o_P(r_{D^*,n}) \text{ for a } r_{D^*,n} = o(1) & (22)\\
E_{B_n} R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)) &= o_P(n^{-1/2}), & (23)
\end{aligned}$$

where for the latter two assumptions (22) and (23) one can use that for each of the V realizations of B_n, $d_0(Q^*_{n,B_n}, Q_0) = O_P(n^{-\lambda_1})$ and $d_{02}(G_{n,B_n}, G_0) = O_P(n^{-\lambda_2})$.

Then, $\psi_n^* = E_{B_n}\Psi(Q_{n,B_n,\epsilon_n})$ is asymptotically efficient:

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(Q_0, G_0) + o_P(n^{-1/2}).$$

Proof: By assumptions (20) and (21), we have

$$d_0((\hat{Q}(P^0_{n,B_n}), \hat{G}(P^0_{n,B_n}), (Q_0, G_0)) = O_P(r^2_Q(n), r^2_G(n)) = O_P(n^{-\lambda_1}, n^{-\lambda_2}).$$

Lemma 15 proves that under these same assumptions (20), (21), we also have, for each B_n, $d_{01}(\hat{Q}_{n,B_n,\epsilon_n}, Q_{0n}) = O_P(n^{-\lambda_1})$. This proves that for each B_n, $d_0((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)) = O_P(n^{-\lambda_1}, n^{-\lambda_2})$. Suppose $n > N_0$ so that $Q_{0n} = Q_0$ and $G_{0n} = G_0$. By the identity $\Psi(Q^*_{n,B_n}) - \Psi(Q_0) = -P_0 D^*(Q^*_{n,B_n}, G_{n,B_n}) + R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0))$, we have

$$E_{B_n}\Psi(Q^*_{n,B_n}) - \Psi(Q_0) = -E_{B_n} P_0 D^*(Q^*_{n,B_n}, G_{n,B_n}) + E_{B_n} R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)).$$

Combining this with (19) yields the following identity:

$$\begin{aligned}
\psi_n^* - \Psi(Q_0) &= E_{B_n}(P^1_{n,B_n} - P_0) D^*(Q^*_{n,B_n}, G_{n,B_n})\\
&\quad + E_{B_n} R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)) + o_P(n^{-1/2}).
\end{aligned}$$

By assumption (23) we have that $E_{B_n} R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)) = o_P(n^{-0.5})$. Thus, we have shown

$$\Psi(Q^*_n) - \Psi(Q_0) = E_{B_n}(P^1_{n,B_n} - P_0) D^*(Q^*_{n,B_n}, G_{n,B_n}) + o_P(n^{-0.5}).$$

We now note

$$\begin{aligned} E_{B_n}(P^1_{n,B_n} - P_0) D^*(Q^*_{n,B_n}, G_{n,B_n}) &= E_{B_n}(P^1_{n,B_n} - P_0) D^*(Q_0, G_0) \\ &\quad + E_{B_n}(P^1_{n,B_n} - P_0)\{D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0)\} \\ &= (P_n - P_0) D^*(Q_0, G_0) + E_{B_n}(P^1_{n,B_n} - P_0)\{D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0)\}. \end{aligned}$$

Thus, it remains to prove that $E_{B_n}(P^1_{n,B_n} - P_0)\{D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0)\} = o_P(n^{-0.5})$. For this we apply Lemma 12 with $f_{n,\epsilon} = D^*(\hat{Q}_\epsilon(P^0_{n,B_n}), G_{n,B_n}) - D^*(Q_0, G_0)$, conditional on P^0_{n,B_n}, and $\mathcal{F}_n = \{f_{n,\epsilon} : \epsilon\}$. By assumption (22), there exists a rate $r_{D^*,n} = o(1)$ so that $\parallel f_{n,\epsilon_n} \parallel_{P_0} = O_P(r_{D^*,n})$, where this rate will be determined based upon $d_0(P^*_{n,B_n}, P_0) = O_P(r^2(n))$ with $r^2(n) = O((n^{-\lambda_1}, n^{-\lambda_2})$.

Note also that the envelope of $\mathcal{F}_n$ satisfies $\parallel F_n \parallel_{P_0} \leq M_{D^*,n}$. Since ϵ is p-dimensional for some integer p, the entropy of $\mathcal{F}_n$ satisfies $\sup_Q N(\epsilon \parallel F_n \parallel, \mathcal{F}_n, L^2(Q)) < \epsilon^{-p}$. Application of Lemma 12 proves now that, if $r_{D^*,n} = o(1)$, then, given P^0_{n,B_n},

$$(P^1_{n,B_n} - P_0) f_{n,\epsilon_n} = o_P(n^{-0.5}).$$

This proves also that $E_{B_n}(P^1_{n,B_n} - P_0) f_{n,\epsilon_n} = o_P(n^{-0.5})$. This completes the proof. □

7 Implementing an MLE over a class of functions with variation norm bounded by a specific constant.

Our super-learner I relies on an estimator defined by minimizing $P_n L_1(\bar{Q})$ over all $\bar{Q} \in \bar{\mathcal{Q}}_n$ for which the variation norm of $L_1(\bar{Q})$ is bounded by some $M < \infty$ for an ordered set of M-vectors. If for a fixed n, there exists a $M_{n,v} \in \mathbb{R}^{k_1}$ so that for all $\bar{Q} \in \bar{\mathcal{Q}}_n$, $\parallel L_1(\bar{Q}) \parallel_v \leq M_{n,v} \parallel \bar{Q} \parallel_v$, then we can achieve this as well by defining the MLE of $\bar{Q} \rightarrow P_n L_1(\bar{Q})$ over all $\bar{Q} \in \bar{\mathcal{Q}}$ with $\parallel \bar{Q} \parallel_v < M$, for a series of M-vectors. Therefore we rephrase our goal as to compute a $\bar{Q}_{n,M}$ so that $P_n L_1(\bar{Q}_{n,M}) = \min_{\bar{Q} \in \bar{\mathcal{Q}}_{n,M}} P_n L_1(\bar{Q}) + r_n$, where r_n is a controlled small number. In this section, we address a concrete strategy for implementation of this MLE.

7.1 Approximating a function with variation norm M by a linear combination of indicator basis functions with L^1-norm of the coefficient vector equal to M.

Any cadlag function $f \in D[0,\tau]$ with finite variation norm can be represented as follows:

$$f(x) = f(0) + \sum_{s \subset \{1,\ldots,p\}} \int_{(0_s, x_s]} f(du_s, 0_{-s}).$$

For each subset s of size $\mid s \mid$, consider a partitioning of $(0_s, \tau_s]$ in $\mid s \mid$-dimensional cubes with width h_m. Let's denote these cubes with $R_{h_m}(j,s)$, where j is the index of the j-th cube and j runs over $O(1/h_m^{|s|})$ cubes. Let $\mathcal{R}_{h_m}(s)$ be the index set, so that we can write $(0_s,\tau_s] = \cup_{j \in \mathcal{R}_{h_m}(s)} R_{h_m}(j,s)$. By definition of an integral, we have $f(x) = \lim_{h_m \to 0} f_m(x)$, where

$$f_m(x) = \sum_{s \subset \{1,\ldots,p\}} \sum_{j \in \mathcal{R}_{h_m}(s)} \phi^s_{h_m,j}(x) \beta^s_{h_m,j},$$

$\beta^s_{h_m,j} = f(R_{h_m}(j,s))$ is the measure f assigns to the cube $R_{h_m}(j,s)$, and $\phi^s_{h_m,j}(x) = I(m_{h_m}(j,s) \leq x_s)$ is the indicator that the midpoint $m_{h_m}(j,s)$ of the cube $R_{h_m}(j,s)$ is smaller or equal than x_s. By the dominated convergence theorem, it also follows that $\parallel f_m(f) - f \parallel_\Lambda \to 0$ for any $L^2(\Lambda)$-norm. Moreover, the variation norm of f is approximated by the sum of the absolute value of all the coefficients $\beta^s_{h_m,j}$:

$$\parallel f \parallel_v = \lim_{h_m \to 0} \sum_{s \subset \{1,\ldots,p\}} \sum_{j \in \mathcal{R}_{h_m}(s)} \mid \beta^s_{h_m,j} \mid .$$

Thus, we conclude that given a function $f \in \mathcal{F}_{v,M}$, we can approximate it with a finite linear combination $f_m(f)$ of basis functions $\phi^s_{h_m,j}$ for which the L^1-norm of its coefficient vector $\{\beta^s_{h_m,j} : j, s\}$ approximates the variation norm of f.

7.2 An approximation of the MLE over functions of bounded variation using L_1-penalization.

Let's define

$$\mathcal{F}^m_{v,M} = \left\{ \sum_{s \subset \{1,\ldots,p\}} \sum_{j \in \mathcal{R}_{h_m}(s)} \phi^s_{h_m,j}(x) \beta^s_{h_m,j} : \sum_{s,j} \mid \beta^s_{h_m,j} \mid \leq M \right\}$$

as the collection of all these finite linear combinations of this collection of basis functions under the constraint that its L^1-norm is bounded by M. Consider the case that the parameter space $\bar{\mathcal{Q}}_j$ is nonparametric, so that the MLE over $\bar{\mathcal{Q}}_{j,n,M_j}$ of $\bar{Q}_{j0}$ would correspond with minimizing over $\mathcal{F}_{v,M_j}$. Note that this does not imply that the model $\mathcal{M}$ is nonparametric: for example, the data distribution could be parameterized in terms of unspecified functions Q_j of dimension $d_1(j)$, $j = 1, \ldots, k_1 + 1$, and unspecified functions G_j of dimension $d_2(j)$, $j = 1, \ldots, k_2 + 1$.

The next lemma proves that we can approximate such an MLE over $\mathcal{F}_{v,M_j}$ for a loss function L_{1j} by an MLE over $\mathcal{F}^m_{v,M_j}$ by selecting m large enough.

Lemma 6 *Let $M \in \mathbb{R}_{\geq 0}$ be given. Consider $f_0 \in \mathcal{F}_{v,M} \subset D[0,\tau]$ so that for a loss function $(O,f) \rightarrow L(f)(O)$, we have $P_0L(f_0) = \min_{f\in\mathcal{F}_{v,M}} P_0L(f)$. Assume that if $f_m \in \mathcal{F}_{v,M}$ converges pointwise to a $f \in \mathcal{F}_{v,M}$ on $[0,\tau]$, then $L(f_m)$ converges pointwise to $L(f)$ on support of P_0, including the support of the empirical distribution P_n. Let $f_{0,m} \in \mathcal{F}^m_{v,M}$ be such that $P_0L(f_{0,m}) = \min_{f\in\mathcal{F}^m_{v,M}} P_0L(f)$. We have $P_0(L(f_{0,m}) - L(f_0)) \rightarrow 0$ as $h_m \rightarrow 0$.*

Consider now an $f_n \in \mathcal{F}_{v,M}$ so that $P_nL(f_n) = \min_{f\in\mathcal{F}_{v,M}} P_nL(f)$, and let $f_{n,m} \in \mathcal{F}^m_{v,M}$ be such that $P_nL(f_{n,m}) = \min_{f\in\mathcal{F}^m_{v,M}} P_nL(f)$. We have $P_n(L(f_{n,m}) - L(f_n)) \rightarrow 0$ as $h_m \rightarrow 0$.

Proof: We want to show that $P_0(L(f_{0,m}) - L(f_0)) \rightarrow 0$. By the approximation presented in the previous section, since $f_0 \in \mathcal{F}_{v,M}$, we can find a sequence $f^*_{0,m} \in \mathcal{F}^m_{v,M}$ so that $f^*_{0,m} \rightarrow f_0$ as $h_m \rightarrow 0$, pointwise and in $L^2(P_0)$ norm. By assumption and the dominated convergence theorem, this implies $P_0L(f^*_{0,m}) - P_0L(f_0)$ also converges to zero as $h_m \rightarrow 0$. But, since $f_{0,m}$ minimizes $P_0L(f)$ over all $f \in \mathcal{F}^m_{v,M}$, we have

$$0 \leq P_0L(f_{0,m}) - P_0L(f_0) \leq P_0L(f^*_{0,m}) - P_0L(f_0) \rightarrow 0,$$

which proves that $P_0L(f_{0,m}) - P_0L(f_0) \rightarrow 0$, as $h_m \rightarrow 0$.

We now want to show that $P_n(L(f_{n,m}) - L(f_n)) \rightarrow 0$ as $h_m \rightarrow 0$. Since $f_n \in \mathcal{F}_{v,M}$, we can find a sequence $f^*_{n,m} \in \mathcal{F}^m_{v,M}$ so that $f^*_{n,m} \rightarrow f_n$ as $h_m \rightarrow 0$, pointwise and in $L^2(P_n)$-norm.

Then, by assumption and the dominated convergence theorem, $P_nL(f^*_{n,m}) - P_nL(f_n)$ also converges to zero as $h_m \rightarrow 0$. But, since $f_{n,m}$ minimizes $P_nL(f)$ over all $f \in \mathcal{F}^m_{v,M}$, we have

$$0 \leq P_nL(f_{n,m}) - P_nL(f_n) \leq P_nL(f^*_{n,m}) - P_nL(f_n) \rightarrow 0,$$

which proves that $P_nL(f_{n,m}) - P_nL(f_n) \rightarrow 0$, as $h_m \rightarrow 0$. □

7.3 An approximation of the MLE over the subspace $\bar{\mathcal{Q}}_{n,M}$ by an MLE over a constrained linear model

For notational convenience, consider the case that $\mathcal{Q}_n = \mathcal{Q}$. Above we defined a mapping from a function $f \in \mathcal{F}_{v,M}$ into a linear combination $f_m(f) \in \mathcal{F}^m_{v,M}$ of basis functions for which the norm of the coefficient vector approximates the variation norm of f. The following lemma proves in general that we can compute the MLE over $\bar{\mathcal{Q}}_M = \bar{\mathcal{Q}} \cap \mathcal{F}^{k_1}_{v,M}$ with the MLE over $\bar{\mathcal{Q}}^m_M = \{\bar{Q}_m(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}_M\}$, which is a collection of these linear combinations of the basis functions for which the L^1-norm of the coefficient vector is bounded by M. Note that $\bar{\mathcal{Q}}^m_M$ is typically not a submodel of $\bar{\mathcal{Q}}_M$, but it is obtained by replacing each element $\bar{Q}$ in $\bar{\mathcal{Q}}_M$ with its approximation $\bar{Q}_m(\bar{Q})$.

Lemma 7 *Assume that the loss function $L_1(\bar{Q})$ satisfies the pointwise continuity condition of the previous lemma.*

For an $M \in \mathbb{R}^{k_1}$, let $\bar{\mathcal{Q}}_M = \bar{\mathcal{Q}} \cap \mathcal{F}_{v,M} = \{\bar{Q}(P) : P \in \mathcal{M}, \bar{Q}(P) \in \mathcal{F}_{v,M}\}$ be all functions in the parameter space for $\bar{Q}_0$ that have a variation norm smaller than $M < \infty$. Let $\bar{\mathcal{Q}}^m_M = \{\bar{Q}_m(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}_M\}$, where $\bar{Q}_m(\bar{Q})$ is defined above as the finite dimensional linear combination of the basis functions $\{\phi^s_{h_m,j} : j, s\}$ with coefficient vector $\{\beta^s_{h_m,j}(\bar{Q}) : j, s\}$.

Consider a $\bar{Q}_{0,M} \in \bar{\mathcal{Q}}_M$ so that $P_0 L_1(\bar{Q}_{0,M}) = \min_{\bar{Q} \in \bar{\mathcal{Q}}_M} P_0 L_1(\bar{Q})$, and let $\bar{Q}^m_{0,M} \in \bar{\mathcal{Q}}^m_M$ be such that $P_0 L_1(\bar{Q}^m_{0,M}) = \min_{\bar{Q} \in \bar{\mathcal{Q}}^m_M} P_0 L_1(\bar{Q})$. Then, $P_0(L_1(\bar{Q}^m_{0,M}) - L_1(\bar{Q}_{0,M})) \rightarrow 0$ as $h_m \rightarrow 0$.

Similarly, consider a $\bar{Q}_{n,M} \in \bar{\mathcal{Q}}_M$ so that $P_n L_1(\bar{Q}_{n,M}) = \min_{\bar{Q} \in \bar{\mathcal{Q}}_M} P_n L_1(\bar{Q})$, and let $\bar{Q}^m_{n,M} \in \bar{\mathcal{Q}}^m_M$ be such that $P_n L_1(\bar{Q}^m_{n,M}) = \min_{\bar{Q} \in \bar{\mathcal{Q}}^m_M} P_n L_1(\bar{Q})$. Then, $P_n(L_1(\bar{Q}^m_{n,M} - L_1(\bar{Q}_{n,M})) \rightarrow 0$ as $h_m \rightarrow 0$.

Proof: We want to show that $P_0(L_1(\bar{Q}^m_{0,M}) - L(\bar{Q}_{0,M})) \rightarrow 0$. By the approximation presented in the previous section, since $\bar{Q}_{0,M} \in \mathcal{F}_{v,M}$, we can find a sequence $\bar{Q}^{m,*}_{0,M} \in \mathcal{F}^m_{v,M}$ so that $\bar{Q}^{m,*}_{0,M} \rightarrow \bar{Q}_{0,M}$ as $h_m \rightarrow 0$, pointwise and in $L^2(P_0)$ norm. By assumption and the dominated convergence theorem, this implies $P_0 L_1(\bar{Q}^{m,*}_{0,M}) - P_0 L_1(\bar{Q}_{0,M})$ also converges to zero as $h_m \rightarrow 0$. But, since $\bar{Q}^m_{0,M}$ minimizes $P_0 L_1(\bar{Q})$ over all $\bar{Q} \in \bar{\mathcal{Q}}^m_M$, we have

$$0 \leq P_0 L_1(\bar{Q}^m_{0,M}) - P_0 L_1(\bar{Q}_{0,M}) \leq P_0 L_1(\bar{Q}^{m,*}_{0,M}) - P_0 L_1(\bar{Q}_{0,M}) \rightarrow 0,$$

which proves that $P_0 L_1(\bar{Q}^m_{0,M}) - P_0 L_1(\bar{Q}_{0,M}) \rightarrow 0$, as $h_m \rightarrow 0$.

We now want to show that $P_n(L_1(\bar{Q}^m_{n,M}) - L_1(\bar{Q}_{n,M})) \rightarrow 0$ as $h_m \rightarrow 0$. Since $\bar{Q}_{n,M} \in \mathcal{F}_{v,M}$, we can find a sequence $\bar{Q}^{m,*}_{n,M} \in \mathcal{F}^m_{v,M}$ so that $\bar{Q}^{m,*}_{n,M} \rightarrow \bar{Q}_{n,M}$ as $h_m \rightarrow 0$, pointwise and in $L^2(P_n)$-norm.

Then, by assumption and the dominated convergence theorem, $P_n L_1(\bar{Q}^{m,*}_{n,M}) - P_n L_1(\bar{Q}_{n,M})$ also converges to zero as $h_m \to 0$. But, since $\bar{Q}^m_{n,M}$ minimizes $P_n L_1(\bar{Q})$ over all $\bar{Q} \in \bar{\mathcal{Q}}^m_{n,M}$, we have

$$0 \leq P_n L_1(\bar{Q}^m_{n,M}) - P_n L_1(\bar{Q}_{n,M}) \leq P_n L_1(\bar{Q}^{m,*}_{n,M}) - P_n L_1(\bar{Q}_{n,M}) \to 0,$$

which proves that $P_n L_1(\bar{Q}^m_{n,M}) - P_n L_1(\bar{Q}_{n,M}) \to 0$, as $h_m \to 0$. □

7.4 What to do with too many basis functions?

If the dimension $d_1(j)$ is large, then for h_m small the linear approximations above are expressed in terms of too many basis functions to store in memory and to computationally handle. In the final subsection of the next section, we suggest that one might want to randomly sample the indicator basis functions $I(\cdot > x)$ by sampling the "midpoint" x from the data itself (i.e.., include the observed values), and variations of the data till the resulting MLE (or TMLE) is not changing anymore by more than a statistically negligible margin. We refer to this subsection for a more detailed discussion.

8 Example: Treatment specific mean

Let $O = (W, A, Y) \sim P_0$ be a random variable consisting of a d-dimensional vector of baseline covariates W, binary treatment $A \in \{0,1\}$ and binary outcome $Y \in \{0,1\}$. We observe n i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P_0$. Let $G(P)$ be the conditional probability distribution of A, given W, and let $Q_1(P)$ be the conditional distribution of Y, given A, W, under P. Let $Q_2(P)$ be the marginal cumulative probability distribution of W, and $Q = (Q_1, Q_2)$. Let the statistical model be of the form $\mathcal{M} = \{P : G(P) \in \mathcal{G}, Q(P) \in \mathcal{Q}\}$, where $\mathcal{G}$ is a possibly restricted set, and $\mathcal{Q}$ is nonparametric. The only key assumption we will enforce on $\mathcal{Q}$ and $\mathcal{G}$ is that for each $P \in \mathcal{M}$, $\bar{Q}_1(P)(a, W) = E_P(Y \mid A = a, W)$ and $\bar{G}(P)(W) = E_P(A \mid W)$ is a Cadlag function in W on a set $[0, \tau_P] \subset \mathbb{R}^d$, and that the variation norm of $\bar{Q}(P)$ and $\bar{G}(P)$ is bounded. Let $g(P)(a \mid W) = P(A = a|W)$ be the conditional probability density. Suppose that $\bar{G}$ only depends on W through a subset of covariates of dimension $d_2 \leq d$. Our target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by $\Psi(P) = \int \{\bar{Q}_1(1, w) - \bar{Q}_1(0, w)\} dQ_2(w) \equiv \Psi_1(Q_1, Q_2)$. For notational convenience, we will use Ψ for both mappings Ψ and Ψ_1. The efficient influence curve $D^*(P) = D^*(Q, G)$ at P is given by:

$$D^*(Q, G)(O) = \frac{2A - 1}{g(A|W)}(Y - \bar{Q}_1(A, W)) + \bar{Q}_1(1, W) - \bar{Q}_1(0, W) - \Psi(Q).$$

We have that $\Psi(P) - \Psi(P_0) = (P - P_0)D^*(Q, G) + R_{20}((\bar{Q}, \bar{G}), (\bar{Q}_0, \bar{G}_0))$, where the second order remainder $R_{20}()$ is defined as follows:

$$\begin{aligned}
R_{20}(P, P_0) &= R_{20,1}(P, P_0) - R_{20,0}(P, P_0) \\
R_{20,a}(P, P_0) &= \int \frac{g(a \mid w) - g_0(a \mid w)}{g(a \mid w)} (\bar{Q}_1(a, w) - \bar{Q}_{10}(a, w)) dP_0(w) \\
& \quad a \in \{0, 1\}.
\end{aligned}$$

We define the following two log-likelihood loss functions for $\bar{Q}_1$ and $\bar{G}$, respectively:

$$\begin{aligned}
L_1(\bar{Q}_1) &= -\left\{ Y \log \bar{Q}_1(A, W) + (1 - Y) \log(1 - \bar{Q}_1(A, W)) \right\} \\
L_2(\bar{G}) &= -\left\{ A \log \bar{G}(W) + (1 - A) \log(1 - \bar{G}(W)) \right\}.
\end{aligned}$$

We also define the corresponding two loss-based dissimilarities $d_{01,1}(\bar{Q}_1, \bar{Q}_{10}) = P_0\{L_1(\bar{Q}_1) - L_1(\bar{Q}_{10})\}$ and $d_{02}(\bar{G}, \bar{G}_0) = P_0\{L_2(\bar{G}) - L_2(\bar{G}_0)\}$. Here Q_2 represents the easy to estimate parameter which we will estimate with the empirical cumulative probability distribution Q_{2n} of $W_1, \ldots, W_n$. We know from empirical process theory that the supremum norm of the difference of the cumulative distribution functions Q_{2n} and Q_{20} converges to zero at rate $1/\sqrt{n}$. Therefore, we define $d_{01,2}(Q_2, Q_{20}) = \parallel Q_2 - Q_{20} \parallel_\infty$.

Let the submodel $\mathcal{M}(\delta, C) \subset \mathcal{M}$ be defined by the extra restriction that $\bar{Q}_1 > \delta$, $\min(g(0 \mid W), g(1 \mid W)) > \delta$, $\parallel \bar{Q}_1 \parallel_v < C$ and $\parallel \bar{G} \parallel_v < C$. Given a sequence $(\delta_n, C_n))$ for which $\delta_n \to 0$ and $C_n \to \infty$ as $n \to \infty$, we can define a sequence of models $\mathcal{M}_n = \mathcal{M}(\delta_n, C_n)$ which grows from below to $\mathcal{M}$ as $n \to \infty$.

Let $\mathcal{Q}_n = \mathcal{Q}_{1n} \times \mathcal{Q}_{2n}$, $\mathcal{G}_n$ be the corresponding parameter spaces for $Q = (Q_1, Q_2)$ and G, and specifically, $\mathcal{Q}_{1n} = \{\bar{Q}_1 : \parallel \bar{Q}_1 \parallel_v < C_n, \bar{Q}_1 > \delta_n\}$, while $\mathcal{Q}_{2n} = \mathcal{Q}_2$. We have the following sieve model bounds for $M_{1Q,n}, M_{2Q,n}, M_{1G,n}, M_{2G,n}$ (van der Laan et al., 2004) and for the supremum and variation norm of

$Q, G, L_1(Q), L_2(G)$:

$$\begin{aligned}
M_{1Q,n} &= O(\log \delta_n) \\
M_{2Q,n} &= O(1/\delta_n) \\
M_{1G,n} &= O(\log \delta_n) \\
M_{2G,n} &= O(1/\delta_n) \\
M_{D^*,n} &= O(1/\delta_n) \\
M_{Qv,n} &= O(C_n) \\
M_{Gv,n} &= O(C_n) \\
M_{L_1(Q)v,n} &= O(C_n\delta_n^{-1}) \\
M_{L_2(G)v,n} &= O(C_n\delta_n^{-1}) \\
M_{D^*v,n} &= O(C_n\delta_n^{-2}).
\end{aligned}$$

Regarding the variation norm bounds, it is easy to see that the variation norm of $\log f$ for an $f > \delta_n$ is bounded by $\parallel f \parallel_v / \delta_n$, and the variation norm of $\parallel \bar{Q}_1/g \parallel_v$ involves an integral with $1/g^2$, which explains the δ_n^{-2} factor.

Since the parameter space $\mathcal{Q}_{1n}$ consists of the cadlag functions with bounded variation norms, without any further restrictions beyond the global bounds δ_n, C_n, we can select the entropy quantities for $\mathcal{Q}_1$ as follows: $\alpha_1 = \alpha_1^* = \tilde{\alpha}_1 = \alpha(d) = 1/(d+1)$, where d is the dimension of W. Similarly, if $\mathcal{G}_n$ consists of all cadlag functions of dimension d_2, without further meaningful restrictions beyond δ_n, C_n, then we can select the entropy quantities for $\mathcal{G}_n$ as $\alpha_2 = \alpha_2^* = \alpha(d_2) = 1/(d_2+1)$. If the model $\mathcal{G}$ enforces more meaningful restrictions than that A only depends on W through a subset of W of dimension d_2, then $\alpha_2 = \alpha_2^*$ can be replaced by a sharper upper bound α_2 than $\alpha(d_2)$. The entropy bound $\tilde{\alpha}$ for the parameter space of $D^*(Q,G)$ can be set equal to $\alpha(d) = 1/(d+1)$.

Let $\bar{Q}_{1n} \in \mathcal{Q}_{1n}$ be a super-learner I of $\bar{Q}_{10}$ of the type presented in (2). Similarly, let $\bar{G}_n \in \mathcal{G}_n$ be such a super-learner I of $\bar{G}_0$ as presented in (4) . Suppose that $\max(M_{1Q,n}, M_{2Q,n}^2)\log K_{1n}/n = O(n^{-\lambda(d)})$ and $\max(M_{1G,n}, M_{2G,n}^2)\log K_{2n}/n = O(n^{-\lambda(d_2)})$, where $\lambda(d) = 0.5 + \alpha(d)/4 = 0.5 + 0.25(d+1)^{-1}$. Then, by Lemma 2 and Lemma 3, we have $d_{01,1}(Q_{1n}, Q_{10}) = O_P(n^{-\lambda(d)})$ and $d_{02}(\bar{G}_n, \bar{G}_0) = O_P(n^{-\lambda(d_2)})$, while $d_{01,2}(Q_{2n}, Q_{20}) = \parallel Q_{2n} - Q_{20} \parallel_\infty = O_P(n^{-1/2})$. Similar results apply to our Super Learner II, but we focus here on Super Learner I.

Plugging in the above bounds for $M_{1Q,n}, M_{2Q,n}, M_{1G,n}, M_{2G,n}$, it follows that it suffices to select δ_n so that $\delta_n^{-1} = O(n^{0.5-0.5\lambda(d)}(\max(\log K_{1n}, \log K_{2n}))^{-0.5})$. (Improvements can be obtained by selecting a separate δ_{1n} for truncating Q_1 and δ_{2n} for truncating G.) Let $K_n = \max(K_{1n}, K_{2n})$ and suppose that $\log K_n = O(n^{0.5-\alpha(d)/2})$. Then, it follows that this bound for δ_n^{-1} is larger than

$n^{\alpha(d)/8}$, so that this constraint on δ_n is dominated by our later constraint given below $\delta_n^{-1} = o(n^{\alpha(d)/8})$.

8.1 One-step TMLE

Consider the submodel:

$$\text{Logit}\bar{Q}_{1n,\epsilon_1} = \text{Logit}\bar{Q}_{1n} + \epsilon_1 H_{g_n},$$

where $H_{g_n}(A, W) = (2A-1)/g_n(A \mid W)$. Let $\epsilon_{1n} = \arg\min_{\epsilon_1} P_n L_1(\bar{Q}_{1n,\epsilon_1})$, and $\bar{Q}^*_{1n} = \bar{Q}_{1n,\epsilon_{1n}}$. The TMLE of $\Psi(Q_0)$ is given by $\Psi(Q^*_n)$, where $Q^*_n = (\bar{Q}^*_{1n}, Q_{2n})$. The second step TMLE would result in $\epsilon_{1n} = 0$ so that it follows that $P_n D^*(Q^*_n, G_n) = 0$. We will now verify the other conditions of Theorem 1.

Preservation of rate of convergence of TMLE update: By Lemma 13 and its corresponding corollary 1, if $C_{2n} n^{-\tilde{\alpha}_1/4} = o(1)$, then we also have $d_{01,1}(Q^*_{1n}, Q_{10}) = O_P(n^{-\lambda(d)})$, where

$$C_{2n} = M_{L_1(Q)v,n}^{\tilde{\alpha}_1+\tilde{\alpha}_1^2} M_{2Q,n}^{\tilde{\alpha}_1} M_{1Q,n}^{1-\tilde{\alpha}_1^2} + n^{-\tilde{\alpha}_1/2} M_{L_1(Q)v,n}^{-1+2\tilde{\alpha}_1^2} M_{1Q,n}^{2\tilde{\alpha}_1(1-\tilde{\alpha}_1)} M_{2Q,n}^{-2+2\tilde{\alpha}_1}.$$

The rate at which C_{2n} can converge to infinity is dominated by the first term $C_{2n,a}$ on the right-hand side. We have

$$C_{2n,a} = O\left(C_n^{\alpha(d)+\alpha(d)^2} \delta_n^{-2\alpha(d)-\alpha(d)^2} (\log \delta_n)^{1-\alpha(d)^2}\right).$$

So the condition is that

$$n^{-\alpha(d)/4} C_n^{\alpha(d)+\alpha(d)^2} \delta_n^{-2\alpha(d)-\alpha(d)^2} (\log \delta_n)^{1-\alpha(d)^2} = o(1).$$

Using that we will enforce $\delta_n^{-1} = o(n^{\alpha(d)/8})$, this condition will hold if

$$C_n^{\alpha(d)+\alpha^2(d)} n^{-\alpha(d)/4+\alpha^2(d)/4+\alpha^3(d)/8} (\log n)^{1-\alpha^2(d)} = o(1).$$

Rate of convergence $r_{D^*,n}$ of estimated efficient influence curve: We also note that

$$\begin{aligned} \parallel D^*(Q^*_n, G_n) - D^*(Q_0, G_0) \parallel_{P_0} \leq \tfrac{1}{\delta_n^{3/2}} \parallel g_n - g_0 \parallel_{P_0} + \tfrac{\|\bar{Q}^*_{1n} - \bar{Q}_{10}\|_{P_0}}{\delta_n} \\ + \mid \Psi(Q^*_n) - \Psi(Q_0) \mid . \end{aligned}$$

Let $\bar{Q}^b_1(W) = \bar{Q}_1(1, W) - \bar{Q}_1(0, W)$. Then,

$$\begin{aligned} &\Psi(Q^*_n) - \Psi(Q_0) = Q_{2n}\bar{Q}^{b*}_{1n} - Q_{20}\bar{Q}^b_{10} \\ &= (Q_{2n} - Q_{20})\bar{Q}^b_{10} + Q_{2n}(\bar{Q}^{b*}_{1n} - \bar{Q}^b_{10}) \\ &= O_P(n^{-1/2}) + (Q_{2n} - Q_{20})(\bar{Q}^{b*}_{1n} - \bar{Q}^b_{10}) + Q_{20}(\bar{Q}^{b*}_{1n} - \bar{Q}^b_{10}) \\ &= O_P(n^{-1/2}) + (Q_{2n} - Q_{20})(\bar{Q}^{b*}_{1n} - \bar{Q}^b_{10}) + O_P(d^{0.5}_{01,1}(\bar{Q}^*_{1n}, \bar{Q}_{10})). \end{aligned}$$

In order to bound the second empirical process term we apply Lemma 9 to the term $n^{0.5}(Q_{2n}-Q_{20})(\bar{Q}^b_{1n}-\bar{Q}^b_{10})/C_n$ with $r_0(n)=n^{-1/4}/C_n\leq n^{-1/4}$ (since $\parallel \bar{Q}^b_{1n}-\bar{Q}^b_{10} \parallel_{P_0}= O_P(n^{-\lambda(d)/2})=o_P(n^{-1/4})$) and $M_n=1/C_n$ and $\alpha=\alpha(d)$. This yield the following bound:

$$(Q_{2n}-Q_{20})(\bar{Q}^{b*}_{1n}-\bar{Q}^b_{10})=O_P(n^{-(0.5+\alpha(d)/4}C_n^{\alpha(d)})=O_P(n^{-\lambda(d)}C_n^{\alpha(d)}).$$

Thus, we have shown

$$\begin{aligned}\parallel D^*(Q^*_n,G_n)-D^*(Q_0,G_0)\parallel_{P_0} &= O_P\left(n^{-\lambda(d)}C_n^{\alpha(d)}+\delta_n^{-1}\parallel \bar{Q}^*_{1n}-\bar{Q}_{10}\parallel_{P_0}\right)\\ &\quad +O_P\left(\delta_n^{-3/2}\parallel g_n-g_0\parallel_{P_0}\right)\end{aligned}$$

We have $d_{01,1}(\bar{Q}^*_{1n},\bar{Q}_{10})=O_P(n^{-\lambda(d)})$ and $d_{02}(\bar{G}_n,\bar{G}_0)=O_P(n^{-\lambda(d_2)})$. These rates first need to be translated in terms of $L^2(P_0)$-norms in order to utilize the above bound. In van der Vaart (1998, page 62) it is shown that for two densities p,p_0, we have $\parallel \sqrt{p}-\sqrt{p_0}\parallel^2_{P_0}\leq -\int\log(p/p_0)dP_0$. By noting that $\sqrt{x}-\sqrt{x_0}=0.5(\xi(x,x_0))^{-0.5}(x-x_0)$ for some $\xi(x,x_0)\in(\min(x,x_0),\max(x,x_0))$, and that $\max(1/\bar{Q}_{1n},1/\bar{G}_n)\leq 1/\delta_n$, it follows that $\parallel \bar{Q}_{1n}-\bar{Q}_{10}\parallel_{P_0}=O_P(n^{-\lambda(d)/2}\delta_n^{-0.5})$ and $\parallel \bar{G}_n-\bar{G}_0\parallel_{P_0}=O_P(n^{-\lambda(d_2)}\delta_n^{-0.5})$. So we obtain the following bound:

$$\begin{aligned}\parallel D^*(Q_n,G_n)-D^*(Q_0,G_0)\parallel_{P_0} &= O_P(n^{-\lambda(d)}C_n^{\alpha(d)}+\delta_n^{-1.5}n^{-\lambda(d)/2}+\delta_n^{-2}n^{-\lambda(d_2)/2})\\ &= O_P(n^{-\lambda(d)}C_n^{\alpha(d)}+\delta_n^{-2}n^{-\lambda(d)/2}),\end{aligned}$$

where we used conservative bounding by not utilizing that d_2 could be significantly smaller than d. By assuming that $C_n<n^{0.5\lambda(d)/\alpha(d)}$ (a condition dominated by our other constraints), it follows that the latter term is $O_P(\delta_n^{-2}n^{-\lambda(d)/2})$ so that we can define $r_{D^*,n}=\delta_n^{-2}n^{-\lambda(d)/2}$.

We have the following upper bound for C_{1n} (11):

$$C_{1n}=o\left(r^{\alpha(d)}_{D^*,n,1}\delta_n^{\alpha(d)-1}+r^{2\alpha(d)-2}_{D^*,n,1}\delta_n^{2\alpha(d)-4}C_nn^{-0.5}\right),$$

where $r_{D^*,n,1}=\max(n^{-1/4},r_{D^*,n})$. Define

$$\begin{aligned}C_{1na} &= n^{-\alpha(d)/4}\delta^{\alpha(d)-1}+n^{-\alpha(d)/2}\delta_n^{2\alpha(d)-4}C_n\\ C_{1nb} &= \delta_n^{-(1+\alpha(d))}n^{-\alpha(d)/4-\alpha^2(d)/8}+\delta_n^{-2\alpha(d)}n^{-\alpha(d)/4-\alpha^2(d)/4}C_n.\end{aligned}$$

Then $C_{1n}=o(\max(C_{1na},C_{1nb}))$. We need $C_{1n}=o(1)$. The first term of C_{1nb} and the first term of C_{1na} are both $o(n^{-\alpha(d)/8})$ by using that $\delta_n^{-1}=o(n^{\alpha(d)/8})$. Thus, our condition is that

$$C_n\left(n^{-\alpha(d)/2}\delta_n^{2\alpha(d)-4}+\delta_n^{-2\alpha(d)}n^{-\alpha(d)/4-\alpha^2(d)/4}\right)=o(1).$$

Analysis of second order remainder $R_{20}(P_n^*, P_0)$: Consider the second order term $R_{20,a}(P_n^*, P_0)$ for $a \in \{0,1\}$. By the Cauchy-Schwarz inequality obtain the following bound

$$\begin{aligned} R_{20,a}(P_n^*, P_0) &\leq \delta_n^{-1} \parallel g_n - g_0 \parallel_{P_0} \parallel \bar{Q}_{1n}^* - \bar{Q}_{10} \parallel_{P_0} \\ &= O_P(\delta_n^{-2} n^{-\lambda(d)}), \end{aligned}$$

where we used that the $L^2(P_0)$ norms of $g_n - g_0$ and $\bar{Q}_{1n}^* - \bar{Q}_{10}$ are $O_P(\delta_n^{-0.5} n^{-\lambda(d)/2})$, as shown above. Thus, we need that $\delta_n^{-2} n^{-\lambda(d)} = o(n^{-0.5})$, and thus $\delta_n^{-1} = o(n^{\alpha(d)/8})$.

We have verified all conditions of Theorem 1 for the one-step TMLE. Application of Theorem 1 yields the following result.

Theorem 3 *Consider the one-step TMLE $\Psi(Q_{1n}^*, Q_{2n})$ of $\Psi(P_0) = \Psi(Q_{10}, Q_{20})$ based on Super Learner I defined above, where the super-learner I of $\bar{Q}_{10}$ and $\bar{G}_0$ is enforced to be contained in interval $(\delta_n, 1 - \delta_n)$ and its variation norm is enforced to be smaller than C_n. Let $\alpha(d) = 1/(d+1)$, $\lambda(d) = 0.5 + \alpha(d)/4$, $r_{1n}(\delta_n, C_n) = \max(n^{-1/4}, n^{-\lambda(d)} C_n^{\alpha(d)} + \delta_n^{-2} n^{-\lambda(d)/2})$.*

Assume that $\log K_n = O(n^{0.5-\alpha(d)/2})$, and that δ_n^{-1}, C_n are converging slowly enough to ∞ so that the following holds:

$$\begin{aligned} &\delta_n^{-1} = o(n^{\alpha(d)/8}) \\ &C_n^{\alpha(d)+\alpha^2(d)} n^{-\alpha(d)/4+\alpha^2(d)/4+\alpha^3(d)/8} (\log n)^{1-\alpha^2(d)} = o(1) \\ &C_n \left(n^{-\alpha(d)/2} \delta_n^{2\alpha(d)-4} + \delta_n^{-2\alpha(d)} n^{-\alpha(d)/4-\alpha^2(d)/4} \right) = o(1). \end{aligned}$$

Then ψ_n^* is a regular asymptotically linear estimator with influence curve equal to the efficient influence curve $D^*(P_0)$, and is thus asymptotically efficient.

The condition $\delta_n^{-1} = o(n^{\alpha(d)/8})$ makes clear that for large dimensions d, we only allow δ_n to converge to zero at a very slow rate. The second condition is for most d implied by the first condition, and the third condition requires $C_n = O(n^{\alpha(d)/2})$. Given that $\alpha(d) = 1/(d+1)$, we can conclude that both $\delta_n^{-1} < n^{\alpha(d)/8}$ and $C_n < n^{\alpha(d)/2}$ can only converge to infinity at a very slow rate when the dimension d is large.

8.2 One step CV-TMLE

For a given cross-validation scheme $B_n \in \{0,1\}^n$, let Q_{n,B_n} and G_{n,B_n} be the super-learner I (2) and (4) applied to the training sample P_{n,B_n}^0, respectively. Let $\epsilon_{1n} = \arg\min_{\epsilon_1} E_{B_n} P_{n,B_n}^1 L_1(\bar{Q}_{1n,B_n,\epsilon_1})$, and $\bar{Q}_{1n,B_n}^* = \bar{Q}_{1n,B_n,\epsilon_{1n}}$. The CV-TMLE of $\Psi(Q_0)$ is given by $E_{B_n} \Psi(Q_{n,B_n}^*)$, where $Q_{n,B_n}^* = (\bar{Q}_{1n,B_n}^*, Q_{2n,B_n})$.

Just as the TMLE, the CV-TMLE iterative updating algorithm converges in one-step so that we have $E_{B_n} P^1_{n,B_n} D^*(Q^*_{n,B_n}, G_{n,B_n}) = 0$.

Above we showed that if $\delta_n^{-1} = O(n^{0.5-0.5\lambda(d)}(\max(\log K_{1n}, \log K_{2n}))^{-0.5})$, then the two super-learners Q_{1n,B_n}, G_{n,B_n} of $\bar{Q}_{10}$ and $\bar{G}_0$ based on the training sample P^0_{n,B_n} converge at the rate $n^{-\lambda(d)}$ w.r.t the loss-based dissimilarities $d_{01,1}$ and d_{02}. In addition, by Lemma 15, under these same conditions, the TMLE update Q^*_{1n,B_n} converges at this same rate. We will assume that $\log K_n = O(n^{0.5-\alpha(d)/2})$ so that this constraint on δ_n^{-1} is dominated by our constraint $\delta_n^{-1} = o(n^{\alpha(d)/8})$ below.

Above, we also showed that for each of the V splits B_n, we have

$$\begin{aligned} \parallel D^*(Q^*_{n,B_n}, G_n) - D^*(Q_0, G_0) \parallel_{P_0} &= O_P(n^{-\lambda(d)} C_n^{\alpha(d)} + \delta_n^{-2} n^{-\lambda(d)/2}) \\ &\equiv O_P(r_{D^*,n}). \end{aligned}$$

We need that $r_{D^*,n} = o(1)$ and thus that $C_n = o(n^{0.5/\alpha(d)+0.25}) = o(n^{0.5(d+1)+0.25})$ and $\delta_n^{-2} = o(n^{\lambda(d)/2})$. For all $d \geq 1$, the latter condition $\delta_n^{-1} = o(n^{\lambda(d)/4})$ will be dominated by the condition $\delta_n^{-1} = o(n^{-\alpha(d)/8})$ below. Note that C_n can grow to infinity very fast, so that for all practical purposes there is no constraint on C_n. In fact, by using a slightly different definition $\psi^*_n = E_{B_n} \Psi(Q^*_{1n,B_n}, \hat{Q}_2(P^1_{n,B_n}))$ it follows that there is no constraint on C_n, since this constraint only appears as part of having to bound $\psi^*_n - \psi_0$ above.

Above, we also showed that if $\delta_n^{-1} = o(n^{\alpha(d)/8})$, then $R_2(P^*_{n,B_n}, P_0) = o_P(n^{-1/2})$. Application of Theorem 2 yields the following result.

Theorem 4 *Consider the one-step CV-TMLE $\psi^*_n = E_{B_n} \Psi(Q^*_{n,B_n})$ of $\Psi(Q_0)$ based on Super Learner I defined above, where the super-learner I of $\bar{Q}_{10}$ and $\bar{G}_0$ is enforced to be contained in interval $(\delta_n, 1-\delta_n)$ and its variation norm is enforced to be smaller than C_n. Let $\alpha(d) = 1/(d+1)$, $\lambda(d) = 0.5 + \alpha(d)/4$, and $K_n = \max(K_{1n}, K_{2n})$.*

Assume that $\log K_n = O(n^{0.5-\alpha(d)/2})$, and that δ_n^{-1}, C_n are converging slowly enough to ∞ so that the following holds:

$$\begin{array}{l} \delta_n^{-1} = o(n^{\alpha(d)/8}) \\ C_n = o(n^{0.5(d+1)+0.25}). \end{array}$$

Then ψ^*_n is a regular asymptotically linear estimator with influence curve equal to the efficient influence curve$D^*(P_0)$, and is thus asymptotically efficient.

Thus, again, just as with the one-step TMLE, for large dimension d, δ_n is only allowed to converge to infinity at a very slow rate. Contrary to the condition on C_n in the previous theorem for the one-step TMLE, the condition on C_n

can be ignored for all practical purposes. The fact that the CV-TMLE allows C_n to be unbounded demonstrates the important gain of CV-TMLE relative to the TMLE.

8.3 Practical implementation of MLE over functions with bounded variation norm smaller than M

Consider a logistic linear regression model in which Logit$\bar{Q}_1$ is approximated by $\sum_j \beta_j \Phi_j$ in $\{\Phi_j : j\}$ where $\Phi_j(W) = I(W \geq w_j)$ for values w_j, $j = 1, \ldots, J_n$, that correspond with midpoints chosen of cubes in a fine partitioning of $[0, \tau] \subset \mathbb{R}^d$ and its sections $[0^s, \tau^s]$ for all subsets $s \subset \{1, \ldots, d\}$. Let $\bar{Q}^M_{1n}$ be the MLE over this linear logistic regression model under the constraint that $\sum_j \mid \beta_j \mid< M$. In the previous section we showed that for a fine enough partitioning this approximates the MLE over all functions $\bar{Q}_1$ for which its logit has a variation norm smaller than $M < \infty$. By including these M-specific MLEs in the library of the super-learner for a range of M-values, the resulting super-learner satisfies the conditions of the above two theorems (i.e., it is the type of super-learner defined by (2)). However, suppose that the dimension of d is reasonably large. Then the number of basis functions is too large to store into memory. In practice, we suggest the following practical approximation. For simplicity, let's consider the case that the true regression $\bar{Q}_0(w) = 0$ for any w for which one or more of its components equals zero. In that case, we can ignore the partitioning of $[0_s, \tau_s]$ for $s \subset \{1, \ldots, d\}$. Firstly, we select the n basis functions corresponding with $w_j \in \{W_1, \ldots, W_n\}$. In that manner, we are already guaranteed that for large M the MLE is able to perfectly fit the data. In addition, we could select another $O(n)$ basis functions by taking a random sample of points in $[0, \tau]$. We suggest to keep adding such randomly sampled basis functions till the resulting MLE is not changing anymore w.r.t $L^2(P_n)$-norm by more than $1/\sqrt{n}$. Our hope would be that this approximation procedure of the actual desired MLE will very quickly converge and will not require more than $O(n)$ basis functions. Regarding selecting a random sampling procedure, one might decide to sample from the uniform distribution on the cartesian product over $l = 1, \ldots, d$, of the sets $\{W_i(l) : i = 1, \ldots, n\}$.

9 Discussion

In this article we established that a one-step TMLE or one-step CV-TMLE, using a super learner with a library that includes L_1-penalized MLEs that minimize the empirical risk over high dimensional linear combinations of indicator basis functions under a series of $L1$-constraints, will be asymptotically efficient. This was shown to hold under remarkable weak conditions and for an arbitrary dimension of the data structure O.

This remarkable fact is heavily driven by the fact that this super-learner will always converge at a rate faster than $n^{-1/4}$ w.r.t. the loss-based dissimilarity. This holds for every dimension of the data and any underlying smoothness of the true nuisance parameter values, as long as these true nuisance parameter values have a finite variation norm. Since the second order remainder $R_2(P_n^*, P_0)$ of the first order expansion for the TMLE can be bounded in terms of these loss-based dissimilarities between the super-learner and its true counterpart, this rate of convergence is fast enough to make the second order remainder asymptotically negligible. As a consequence, the first order empirical mean of the canonical gradient/efficient influence curve drives the asymptotics of the TMLE.

In order to prove our theorems it was also important to establish that a one-step TMLE already approximately solves the efficient influence curve equation, under very general reasonable conditions. In this article we focussed on a one-step TMLE that updates each nuisance parameter with its own one-dimensional MLE update step. This choice of local least favorable submodel guarantees that the one-step TMLE update of the super-learner of the nuisance parameters is not driven by the nuisance parameter component that is hardest to estimate, which might have finite sample advantages. Nonetheless, our asymptotic efficiency results naturally extend to any local least favorable submodel.

The fact that a one-step TMLE already solves the efficient influence curve equation is particularly important in problems in which the TMLE update step is very demanding due to a high complexity of the efficient influence curve. In addition, a one-step TMLE has a more predictable robust behavior than a limit of an iterative algorithm. We could have focussed on the universal least favorable submodels so that the TMLE is always a one-step TMLE, but in various problems local least favorable submodels are easier to fit and can thus have practical advantages.

Even though we did not implement this new super-learner yet, we discussed practical tools for this implementation by relating it to minimizing the empirical risk over L_1-constrained linear model. In a future article we will implement this one-step TMLE and CV-TMLE in order to practically demonstrate these theoretical results and to provide a powerful TMLE for data analyses.

In this article we assumed independent and identically distributed observations. Nonetheless, this type of super learner and the resulting asymptotic efficiency of the one-step TMLE will be generalizable to a variety of dependent data structures such as data generated by a statistical graph that assumes sufficient conditional independencies so that the desired central limit theorems can still be established (van der Laan, 2008; Chambaz and van der Laan, 2011a,b; van der Laan et al., 2013; van der Laan, 2012).

This article focused on variation independent nuisance parameters. However, there are key examples in which representing $\Psi(P)$ in terms of recursively defined nuisance parameters has key advantages. For example, the longitudinal one-step TMLE of causal effects of multiple time point interventions in (Gruber and van der Laan, 2012; Petersen et al., 2013) relies on a sequential regression representation of the target parameter (Bang and Robins, 2005). In this case, the next regression is defined as the regression of the previous regression on a shrinking history, across a number of regressions, one for each time point at which an intervention takes place. In this case, a super-learner of nuisance parameter Q_k is based on a loss function $L_{1,k,Q_{k+1}}(Q_k)$ that depends on a next nuisance parameter Q_{k+1} (representing the outcome for the regression defining Q_k), $k = 1, \ldots, k_1 + 1$.. One would now start with obtaining the desired result for the super-learner of Q_{k_1+1} whose loss function does not depend on other nuisance parameters. For the second super-learner of Q_{k_1} based on candidate estimators $\hat{Q}_{k_1,j}$, $j = 1, \ldots, J$, we would use as cross-validated risk $E_{B_n} P^1_{n,B_n} L_{1,k_1,\hat{Q}_{k_1+1}(P^0_{n,B_n})}(\hat{Q}_{k_1,j})$. In other words, one estimates the nuisance parameter of the loss-function based on the training sample. In (van der Laan and Dudoit, 2003; van der Laan and Petersen, 2012; Díaz and van der Laan, 2013, In press) we establish oracle inequalities for the cross-validation selector based on loss-functions indexed by an unknown nuisance parameter, which now also rely on a remainder concerning the rate at which $\hat{Q}_{k_1+1}(P_n)$ converges to $Q_{k_1+1,0}$. In this manner, one can establish that the super-learner of $Q_{k_1,0}$ will converge at the same or better rate than the super-learner of $Q_{k_1+1,0}$. This process can be iterated to establish convergence of all the super-learners at the same or better rate than the initial super-learner of $Q_{k_1+1,0}$. Our asymptotic efficiency results for the one-step TMLE and one-step CV-TMLE can now be generalized to one-step TMLE and CV-TMLE that rely on sequential targeted learning. The disadvantage of sequential learning is that the behavior of previous super-learners affects the behavior of the next super-learners in the sequence, but the practical implementation of a sequential super-learner can be significantly easier.

Our general theorems and specifically the theorems for our example demonstrate that the model bound on the variance of the efficient influence curve heavily affects the stability of the TMLE, and that we can only let this bound converge to infinity at a slow rate when the dimension of the data is large. Therefore, knowing this bound instead of enforcing it in a data adaptive manner is crucial for good behavior of these efficient estimators. This is also evident from the well known finite sample behavior of various efficient estimators in causal inference and censored data models that almost always rely on using truncation of the treatment and/or censoring mechanism. If one uses highly data adaptive estimators, even when the censoring or treatment mechanism is bounded away from zero, the estimators of these nuisance parameters could easily get very close to zero, so that truncation is crucial. Careful data adaptive selection of this truncation level is therefore an important component in the definition of these efficient estimators.

Alternatively, one can define target parameters in such a way that their variance of the efficient influence curve is uniformly bounded over the model (e.g., van der Laan and Petersen (2007)). For example, in our example we could have defined the target parameter $EY_{d_1} - EY_{d_0}$, where $d_1(W) = I(\bar{G}_n(W) > \delta)$ and $d_0(W) = 1 - I((1 - \bar{G}_n(W) > \delta)$, where $\bar{G}_n$ is the super-learner of $\bar{G}_0 = E_0(A \mid W)$ and $\delta > 0$ is a user supplied constant. In this case, the static interventions have been replaced by realistic dynamic interventions that approximate the static interventions but are guaranteed to only carry out the intervention when there is enough support in the data. Due to the fact that such parameters have a guaranteed amount of support in the data, the variance of the efficient influence curve is uniformly bounded over the model: i.e. $M_{D^*} < \infty$.

Acknowledgement

This grant is funded by NIH-grant 5R01AI074345-07. The author also thank Marco Carone and Alex Luedtke for stimulating discussions.

References

H. Bang and J.M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

P.J. Bickel, C.A. Klassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation of semiparametric models.* Johns Hopkins University Press, Baltimore, MD, 1993.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, theoretical study. *Int J Biostat*, 7(1):1–32, 2011a. Working paper 258, www.bepress.com/ucbbiostat.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, simulation study. *Int J Biostat*, 7(1):33–, 2011b. Working paper 258,www.bepress.com/ucbbiostat.

Iván Díaz and Mark J van der Laan. Targeted data adaptive estimation of the causal dose response curve. *Journal of Causal Inference*, 2013, In press.

R.D. Gill, M.J. van der Laan, and J.A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré*, 31:545–597, 1995.

S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*, 6(1), 2010.

S. Gruber and M.J. van der Laan. Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics*, 8(1):Article 11, doi: 10.1515/1557–4679.1332, 2012.

G. Neuhaus. On weak convergence of stochastic processes with multidimensional time parameter. *Annals of Statistics*, 42:1285–1295, 1971.

M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M.J. van der Laan. Targeted minimum loss based estimation of marginal structural working models. *Journal of Causal Inference*, submitted, technical report http://biostats.bepress.com/ucbbiostat/paper312/, 2013.

E.C. Polley, Sherri Rose, and M.J. van der Laan. Super learning. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2012.

Kristin E. Porter, Susan Gruber, Mark J. van der Laan, and Jasjeet S. Sekhon. The relative performance of targeted maximum likelihood estimators. *Int J Biostat.*, Jan 1, 2011; 7(1): Article 31., 2011. Published online Aug 17, 2011. doi: 10.2202/1557-4679.1308. Also available at: U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 279, http://www.bepress.com/ucbbiostat/paper279.

J.M. Robins and Y. Ritov. Towards a curse of dimensionality appropriate asymptotic theory for semi parametric models. *Statistics in Medicine*, 16: 285–319, 1997.

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser, 1992.

J.S. Sekhon, S. Gruber, K.E. Porter, and M.J. van der Laan. Propensity score-based estimators and c-tmle. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2012.

M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

M.J. van der Laan. Causal inference for networks. Technical Report 300, UC Berkeley, 2012. http://biostats.bepress.com/ucbbiostat/paper300,to appear in Journal of Causal Inference.

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. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *to appear in International Journal of Biostatistics*, 2015.

M.J. van der Laan and M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *International Journal of Biostatistics*, 3(1), 2007.

M.J. van der Laan and M.L. Petersen. Targeted learning. In *Ensemble Machine Learning*, chapter pages 117–156, ISBN 978-1-4419-9326-7. Springer, New York, 2012.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

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

M.J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

M.J. van der Laan, L.B. Balzer, and M.L. Petersen. Adaptive matching in randomized trials and observational studies. *Journal of Statistical Research*, 46(2):113–156, 2013.

A.W. van der Vaart and J.A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. ISSN: 1935-7524, DOI: 10.1214/11-EJS605.

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.

W. Zheng and M.J. van der Laan. Cross-validated targeted minimum loss based estimation. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Studies*. Springer, New York, 2011.

Appendix

A Oracle inequality for the cross-validation selector.

Lemma 2 is a simple corollary of the following finite sample oracle inequality for cross-validation (van der Laan and Dudoit, 2003), combined with exploiting the convexity of the loss function allowing us to bring the E_{B_n} inside the loss-based dissimilarity.

Lemma 8 *For any $\delta > 0$, there exists a constant $C(M_{1Q,n}, M_{2Q,n}, \delta) = 2(1+\delta)^2(2M_{1Q,n}/3 + M_{2Q,n}^2/\delta)$ such that*

$$\begin{aligned} E_0\{E_{B_n} d_{01}(\hat{\bar{Q}}_{k_{1n}}(P^0_{n,B_n}), \bar{Q}_0)\} &\leq (1+2\delta)E_0\{E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_0)\} \\ &\quad + 2C(M_{1Q,n}, M_{2Q,n}, \delta)\frac{\log K_{1n}}{n\bar{B}_n}. \end{aligned}$$

Similarly, for any $\delta > 0$,

$$E_{B_n} d_{01}(\hat{\bar{Q}}_{k_{1n}}(P^0_{n,B_n}, \bar{Q}_0) \leq (1+2\delta)E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_0)\} + R_n,$$

where $ER_n \leq 2C(M_{1Q,n}, M_{2Q,n}, \delta)\frac{\log K_{1n}}{n\bar{B}_n}$.

If $\log K_{1n}/n$ divided by $E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_0)\}$ converges to zero in probability, then we also have

$$\frac{E_{B_n} d_{01}(\hat{\bar{Q}}_{k_n}(P^0_{n,B_n}, \bar{Q}_0)}{E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}, \bar{Q}_0)} \to_p 1.$$

Similarly, if $\log K_{1n}/n$ divided by $E_0 E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}), \bar{Q}_0)\}$ converges to zero, then we also have

$$\frac{E_0 E_{B_n} d_{01}(\hat{\bar{Q}}_{k_n}(P^0_{n,B_n}, \bar{Q}_0)}{E_0 E_{B_n} \min_k d_{01}(\hat{\bar{Q}}_k(P^0_{n,B_n}, \bar{Q}_0)} \to 1.$$

B Empirical process results

A theorem in (van der Vaart and Wellner, 2011) establishes the following result for a Donsker class $\mathcal{F}_n$ with envelope F_n: If $Pf^2 \leq \delta^2 PF^2$, then

$$E \parallel G_n \parallel_{\mathcal{F}_n} \leq J(\delta, \mathcal{F}_n, L^2)\left(1 + \frac{J(\delta, \mathcal{F}_n, L_2)}{\delta^2\sqrt{n} \parallel F_n \parallel_{P_0}}\right) \parallel F_n \parallel_{P_0},$$

where

$$J(\delta, \mathcal{F}_n, L^2) = \sup_{\Lambda} \int_0^{\delta} \left(\log(1 + N(\epsilon \parallel F_n \parallel_{P_0}, \mathcal{F}_n, L^2(\Lambda)))\right)^{0.5} d\epsilon$$

is the entropy integral from 0 to δ. A simple corollary of this theorem is the following lemma.

Lemma 9 *Consider* $\mathcal{F}_n$ *with* $\| F_n \|_{P_0} < M_n$ *and* $\sup_\Lambda \sqrt{\log(1+N(\epsilon \| F_n \|_{P_0}, \mathcal{F}_n, L^2(\Lambda)))} < 1/\epsilon^{1-\alpha}$. *Then,*

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < r_0(n)} \mid G_n(f) \mid \leq \{r_0(n)/M_n\}^{\alpha} M_n + \{r_0(n)/M_n\}^{2\alpha-2} n^{-0.5}.$$

If $r_0(n) < n^{-1/4}$, *one should select* $r_0(n) = n^{-1/4}$ *in the above right hand side, giving the bound:*

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < r_0(n)} \mid G_n(f) \mid \leq \{n^{-0.25}/M_n\}^{\alpha} M_n + \{M_n\}^{2-2\alpha} n^{-\alpha/2}.$$

The following lemma is proved by first applying the above lemma to $(P_n - P_0)f_n$ with $r_0(n) = 1$ to obtain an initial rate $r_0(n)$, and then applying the above lemma again with this initial rate $r_0(n)$.

Lemma 10 *Consider the following setting:*

$$\begin{array}{l} d_0(Q_n, Q_0) \leq (P_n - P_0) f_n \\ f_n \in \mathcal{F}_n, \ \| F_n \|_{P_0} \leq M_n \\ \sup_\Lambda \sqrt{\log(1+N(\epsilon \| F_n \|_{P_0}, \mathcal{F}_n, L^2(\Lambda)))} < 1/\epsilon^{1-\alpha} \\ \| f_n \|_{P_0} \leq M_{2n} \{d_0(Q_n^*, Q_0)\}^{0.5}. \end{array}$$

Then

$$d_0(Q_n, Q_0) \leq n^{-1/2} n^{-\alpha/2} C(M_n, M_{2n}, \alpha),$$

where

$$C(M_n, M_{2n}, \alpha) = M_2(n)^{\alpha} M(n)^{1-\alpha^2} + M(n)^{2\alpha(1-\alpha)} M_2(n)^{-2(1-\alpha)} n^{-\alpha/2}.$$

Lemma 11 *Let* $f_n = D^*(Q_n^*, G_n) - D^*(Q_0, G_0)$. *Assume* $\| f_n \|_{P_0} = O_P(r_{D^*,n})$. *Assume*

$$\sup_\Lambda \sqrt{\log(1+N(\epsilon \| F_{1n} \|_{P_0}, \mathcal{F}_{1n}, L^2(Q)))} = O(\epsilon^{-(1-\alpha)}),$$

where $\mathcal{F}_{1n} = \{D^*(P) : P \in \mathcal{M}_n\}/M_{D^*,v}$. *We can always select* $\alpha = 1/(d+1)$ *where* d *is the dimension of* O.

Let $r_{D^*,n,1} = \max(n^{-1/4}, r_{D^*,n})$. *Then,*

$$E_0 \mid \sqrt{n}(P_n - P_0) f_n \mid \leq \left(r_{D^*,n,1}/M_{D^*,n}\right)^{\alpha} M_{D^*,n} + \left(r_{D^*,n,1}/M_{D^*,n}\right)^{2\alpha-2} n^{-0.5} M_{D^* v,n}.$$

Proof: Let $\mathcal{F}_n = \{D^*(P) : P \in \mathcal{M}_n\} \subset \{f :\parallel f \parallel_v \leq M_{D^*v,n}, \parallel f \parallel_\infty < M_{D^*,n}\}$, and note that its envelope F_n satisfies $\parallel F_n \parallel_{P_0} \leq M_{D^*,n}$. Note that $f_n \in \mathcal{F}_n$. Let $\mathcal{F}_{1n} = \mathcal{F}_n / M_{D^*v,n}$, and note that $(P_n - P_0) f_n = M_{D^*v,n}(P_n - P_0) f_{1n}$ where $f_{1n} = f_n / M_{D^*v,n} \in \mathcal{F}_{1n}$. We have $\parallel F_{1n} \parallel_{P_0} \leq M_{D^*,n}/M_{D^*v,n}$ and, by assumption, $\sup_\Lambda \sqrt{\log(1 + N(\epsilon \parallel F_{1n} \parallel_{P_0}, \mathcal{F}_{1n}, L^2(\Lambda)))} = O(\epsilon^{-(1-\alpha)})$. We can now apply Lemma 9 with $\mathcal{F}_n = \mathcal{F}_{1n}$, $M_n = M_{D^*,n}/M_{D^*v,n}$, α, $r_0(n) = r_{D^*,n}/M_{D^*v,n}$. Application of Lemma 9 yields

$$E_0 G_n(f_{1n}) \leq (r_{D^*,n,1}/M_{D^*,n})^{\alpha} M_{D^*,n}/M_{D^*v,n} + (r_{D^*,n,1}/M_{D^*,n})^{2\alpha-2} n^{-0.5}.$$

The desired bound for $E_0 \mid G_n(f_n) \mid$ is the right-hand side multiplied with $M_{D^*v,n}$. $\Box$

The following lemma is needed in the analysis of the CV-TMLE, where $f_{n,\epsilon} = D^*(Q_{n,B_n,\epsilon}, G_{n,B_n}) - D^*(Q_0, G_0)$.

Lemma 12 *Let $f_{n,\epsilon_n} \in \mathcal{F}_n = \{f_{n,\epsilon} : \epsilon\}$ where ϵ varies over a bounded set in $\mathbb{R}^p$ and $f_{n,\epsilon}$ is a non-random function (i.e., not based on data $O_1, \ldots, O_n$). Suppose that $\parallel f_{n,\epsilon_n} \parallel_{P_0} = o_P(r_{D^*,n})$ for a rate $r_{D^*,n}$ satisfying $r_{D^*,n} \log r_{D^*,n}^{-1} \rightarrow 0$, and $r_{D^*,n} n^{0.5} \rightarrow \infty$. Suppose that the envelope F_n of $\mathcal{F}_n$ satisfies $\parallel F_n \parallel_{P_0} \leq M_{D^*,n}$. We have $\sup_\Lambda N(\epsilon \parallel F_n \parallel, \mathcal{F}_n, L^2(\Lambda)) = O(\epsilon^{-p})$. Then,*

$$E_0 \mid G_n(f_{n,\epsilon_n}) \mid = O\left(r_{D^*,n}(1 + \log r_{D^*,n}^{-1})\right).$$

Thus, if $r_{D^*,n} = o(1)$, then $G_n(f_{n,\epsilon_n}) = o_P(1)$.

Proof: For notational convenience, let's denote f_{n,ϵ_n} with f_n. We apply the Theorem in van der Vaart, Wellner providing us with

$$E_0 \mid G_n(f_n) \mid \leq J(\delta_n, \mathcal{F}_n) \left(1 + \frac{J(\delta_n, \mathcal{F}_n)}{\delta_n^2 n^{0.5} \parallel F_n \parallel_{P_0}}\right) \parallel F_n \parallel_{P_0}, \tag{24}$$

where we can select $\delta_n = r_{D^*,n}$. Using the bound ϵ^{-p} on the uniform covering number, it follows that $J(\delta_n, \mathcal{F}_n) = -p^{0.5} \int_0^{\delta_n} (\log \epsilon)^{0.5} d\epsilon$. We can conservatively bound $(\log \epsilon)^{0.5}$ by $\log \epsilon$, and use that $\int_0^{\delta_n} \log \epsilon d\epsilon = \delta_n - \delta \log \delta_n = \delta(1 + \log(\delta_n^{-1}))$. This shows that $J(\delta, \mathcal{F}_n) \leq \delta_n(1 + \log \delta_n^{-1})$. If $J(\delta_n, \mathcal{F}_n) = O(\delta_n^2 n^{0.5})$, then the leading term in (24) is given by $J(\delta_n, \mathcal{F}_n) \parallel F_n \parallel_{P_0}$. Using the above bound for $J(\delta_n, \mathcal{F}_n)$, it follows that this holds if $\delta_n(1 + \log \delta_n^{-1}) = O(\delta_n^2 n^{0.5})$, or equivalently, $\delta_n(1 + \log \delta_n^{-1}) = O(\delta_n n^{0.5})$. By assumption we have $\delta_n n^{0.5} \rightarrow \infty$ and $\delta_n \log \delta_n^{-1} \rightarrow 0$, so that this always holds. This results in the following bound:

$$E_0 \mid G_n(f_n) \mid = O(r_{D^*,n}(1 + \log r_{D^*,n}^{-1})),$$

which equals the stated bound. $\Box$

C Constraining the variation norm of the data distribution does not affect the canonical gradient of the model

In the above formulation of the one-step TMLE, we assumed that both $\hat{Q}(P_n)$ and its update Q_n^* are elements of $\mathcal{Q}_n$. By construction this holds for our initial estimator $\hat{Q}(P_n)$. The update will naturally be an element of $\mathcal{Q}_n$ by our assumption that the least favorable submodel $\{Q_{n,\epsilon} : \epsilon\} \subset \mathcal{Q}_n$ is a submodel of $\mathcal{Q}_n$. However, one might wonder why a least favorable submodel through Q_n with score equal to the efficient influence curve at (Q_n, G_n) in the actual model $\mathcal{M}$ would also be a submodel of $\mathcal{M}_n$? The key to the answer is that the universal bound assumptions enforced by $\mathcal{M}_n$ are not affecting the tangent space at a $P \in \mathcal{M}_n$: i.e. the tangent space at P in model $\mathcal{M}$ is identical to the tangent space at P in the model $\mathcal{M}_n$. This means that the efficient influence at a $P \in \mathcal{M}_n$ for the model $\mathcal{M}_n$ is the same as the efficient influence curve at this P for the model $\mathcal{M}$. This strongly suggests that a naturally selected least favorable submodel for model $\mathcal{M}$ through a $Q \in \mathcal{M}_n$ will also be a submodel of $\mathcal{M}_n$ for a small enough range of ϵ values. Therefore, we expect this to be a non-issue, as also demonstrated in our example.

The fact that supremum norm and $L^2(P_0)$ bounds do not change the tangent space at a $P \in \mathcal{M}_n$ is easily understood, but one might wonder if a strict bound on the variation norm could restrict the class of possible submodels through P enough so that the closure of the linear span of its scores is a strict subset of the tangent space at P under model $\mathcal{M}$. To obtain some insight in this we consider a particular example.

Let O be a univariate random variable and let $\mathcal{M}$ be a nonparametric model dominated by the Lebesgue measure μ that assumes that all densities are differentiable. The tangent space $T(P)$ at $P \in \mathcal{M}$ for this model is saturated and thus equals $L_0^2(P)$. For a $P \in \mathcal{M}$, let $p = dP/d\mu$ be its density and let p' be its derivative Let $\mathcal{M}_n$ be the submodel that enforces that $C = \sup_{P \in \mathcal{M}_n} \parallel p' \parallel_\infty < \infty$. Let $T_n(P)$ be the tangent space at $P \in \mathcal{M}_n$ for model $\mathcal{M}_n$. Suppose that P has compact support $\mathcal{O}(P) \subset \mathbb{R}$. Consider a submodel $\{p_\epsilon = (1 + \epsilon S)p : \epsilon \in (-\delta, \delta)\}$ for some $\delta > 0$ with score $S \in L_0^2(P)$ at a $P \in \mathcal{M}_n$ so that $\parallel p' \parallel_\infty < C < \infty$, where the supremum is over the support $\mathcal{O}(P)$. We have $p'_\epsilon = (1 + \epsilon S')p'$. This will satisfy $\parallel p'_\epsilon \parallel_\infty < C$ if

$$\parallel S' \parallel_\infty < \frac{C - \parallel p' \parallel_\infty}{\delta \parallel p' \parallel_\infty}.$$

We can select $\delta > 0$ equal to an arbitrarily small number larger than 0. This shows that any $S \in L^2_0(P)$ for which its derivative S' has a bounded supremum norm is an element of the tangent space $T_n(P)$. However, any function in $L^2_0(P)$ can be arbitrarily well approximated in $L^2_0(P)$ by functions that have a uniformly bounded derivative (where this bound can be arbitrarily large). This proves that $T_n(P) = L^2_0(P)$ and thus that putting bounds on the derivative do not affect the tangent space of the model.

Consider now a submodel $\mathcal{M}_n$ of $\mathcal{M}$ that enforces $C = \sup_{P \in \mathcal{M}_n} \parallel p \parallel_v < \infty$. As above, let's consider again the submodel $\{p_\epsilon = (1 + \epsilon S)p : \epsilon \in (-\delta, \delta)\}$ for some $\delta > 0$ with score $S \in L^2_0(P)$ at a $P \in \mathcal{M}_n$, so that we know $\parallel p \parallel_v < C$. Let $\mathcal{O}(P)$ be a compact subset of $\mathbb{R}$. We have

$$\parallel p_\epsilon \parallel_v \leq \int_{\mathcal{O}(P)} \mid p(dx) \mid + \epsilon \int_{\mathcal{O}(P)} p(x) \mid S(dx) \mid + \mid S(x)p(dx) \mid .$$

Since $\parallel p \parallel_\infty \leq C$, we can further conservatively bound the right-hand side by $\parallel p \parallel_v + \epsilon C(\parallel S \parallel_v + \parallel S \parallel_\infty)$. Thus, if $\parallel S \parallel_v < (C - \parallel p \parallel_v)/(2C\delta)$, then $\{p_\epsilon : \epsilon \in (-\delta, \delta)\} \subset \mathcal{M}_n$. We can select $\delta > 0$ arbitrarily small. This proves that any $S \in L^2_0(P)$ for which $\parallel S \parallel_v < \infty$ is a score of a parametric submodel of $\mathcal{M}_n$. The closure of the linear span of this set in $L^2_0(P)$ equals $L^2_0(P)$ again, so that $T_n(P) = T(P) = L^2_0(P)$.

Based on these two examples, it follows that indeed additional universal variation norm bounds on the model $\mathcal{M}$ are truly global constraints which do thus not affect the tangent space and accordingly the efficient influence curve of a pathwise differentiable target parameter. As a consequence, at any $P \in \mathcal{M}_n$, a least favorable submodel of $\mathcal{M}_n$ through P for model $\mathcal{M}_n$ will also be a least favorable submodel through P for model $\mathcal{M}$.

D Preservation of the desired rate for the one-step TMLE

In this article we constructed an initial super-learner $\hat{\bar{Q}}$ of $\bar{Q}_0$ satisfying $d_{01}(\hat{\bar{Q}}(P_n), \bar{Q}_0) = O_P(n^{-\lambda_1})$ where $\lambda_1 \in \mathbb{R}^{k_1}_{>0}$. For example, if we select Super Learner II thenλ_1 can be chosen so that $r^2_{\bar{Q}}(n) = O(n^{-\lambda_1})$. However, our one-step TMLE relies on its targeted version $\bar{Q}^*_n = \bar{Q}_{n,\epsilon_n}$, so that we still need to establish a rate for this targeted version, where we can use that we already have a rate for $\bar{Q}_n$. We assume that $\hat{Q}_{k_1+1}(P_n)$ is not updated by the TMLE-update step, by already being an MLE type estimator. Due to the fact that our model $\mathcal{M}$ is allowed to be unbounded, so that the global bounds on $\mathcal{M}_n$ can converge to infinity, we might worsen the rate of the initial estimator $\bar{Q}_n$. Thus the preservation of the desired rate $n^{-1/4}$ might require controlling the rate at which the bounds of $\mathcal{M}_n$ converge to the corresponding infinite bounds for $\mathcal{M}$. This is formalized by the following lemma. We remind the reader that our submodel is such that $\epsilon_n(j) = \arg\min_{\epsilon(j)} P_n L_{1j}(Q_{jn,\epsilon(j)})$, or, in our short-hand notation $\epsilon_n = \arg\min_\epsilon P_n L_1(Q_{n,\epsilon_n})$.

Lemma 13 *Suppose $d_{01}(\bar{Q}_n, \bar{Q}_{0n}) = O_P(n^{-\lambda_1})$ for a $\lambda_1 \in \mathbb{R}^{k_1}_{>0}$ with $\lambda_1 > 0.5$. Suppose also that $\alpha_1 \in \mathbb{R}^{k_1}_{>0}$ is chosen so that for $\mathcal{F}_{1n} = \{L_1(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}_n\}/M_{L_1(Q)v,n}$ with envelope $F_{1n} < M_{1Q,n}/M_{L_1(Q)v,n}$ we have*

$$\sup_\Lambda \sqrt{\log(1 + N(\epsilon \parallel F_{1n} \parallel_{P_0}, \mathcal{F}_{1n}, L^2(\Lambda)))} = O(\epsilon^{-(1-\alpha_1)}).$$

Let $d_1 \in \mathbb{N}^{k_1}_{>0}$ be the vector of integers indicating the dimensions of the domains of the components of $\bar{Q} = (\bar{Q}_1, \ldots, \bar{Q}_{k_1})$. In a nonparametric model $\mathcal{M}_n$ we have that this holds for $\alpha_1 = 1/(d_1+1)$, so that we will always have $\alpha_1 \geq 1/(d_1+1)$. Let $\epsilon_n \in \mathbb{R}^{k_1}$ be defined by $\epsilon_n(j) = \arg\min_{\epsilon(j)} P_n L_{1j}(\bar{Q}_{jn,\epsilon_j})$, $j = 1, \ldots, k_1$, and let $\bar{Q}^*_n = \bar{Q}_{n,\epsilon_n}$. Then,

$$E_0 d_{01}(\bar{Q}^*_n, \bar{Q}_{0n}) \;\leq\; n^{-1/2} n^{-\alpha_1/2} C(M_{L_1(Q)v,n}, M_{2Q,n}, M_{1Q,n}, \alpha_1) + O(n^{-\lambda_1}),$$

where

$$\begin{aligned}
&C(M_{L_1(Q)v,n}, M_{2Q,n}, M_{1Q,n}, \alpha_1) = \\
&M_{L_1(Q)v,n}\left\{M_{2Q,n}M_{L_1(Q)v,n}\right\}^{\alpha_1}\left\{M_{1Q,n}/M_{L_1(Q)v,n}\right\}^{1-\alpha_1^2} \\
&+M_{L_1(Q)v,n}\left\{\{M_{1Q,n}/M_{L_1(Q)v,n}\}^{2\alpha_1(1-\alpha_1)}\{M_{2Q,n}M_{L_1(Q)v,n}\}^{-2(1-\alpha_1)}n^{-\alpha_1/2}\right\}.
\end{aligned}$$

Proof: We have that $P_n L_1(\bar{Q}_{n,\epsilon_n}) = \min_\epsilon P_n L_1(\bar{Q}_{n,\epsilon})$. This yields the following inequality:

$$\begin{aligned}
0 &\leq d_{01}(\bar{Q}^*_n, \bar{Q}_{0n}) = P_0 L_1(\bar{Q}_{n,\epsilon_n}) - P_0 L_1(\bar{Q}_{0n}) \\
&= P_0 L_1(\bar{Q}_{n,\epsilon_n}) - P_0 L_1(\bar{Q}_n) + P_0 L_1(\bar{Q}_n) - P_0 L_1(\bar{Q}_{0n}) \\
&= P_0 L_1(\bar{Q}_{n,\epsilon_n}) - P_0 L_1(\bar{Q}_n) + d_{01}(\bar{Q}_n, \bar{Q}_{0n}) \\
&= P_0\{L_1(\bar{Q}_{n,\epsilon_n}) - L_1(\bar{Q}_n)\} + O_P(n^{-\lambda_1}) \\
&= -(P_n - P_0)\{L_1(\bar{Q}_{n,\epsilon_n}) - L_1(\bar{Q}_n)\} + P_n\{L_1(\bar{Q}_{n,\epsilon_n}) - L_1(\bar{Q}_n)\} + O_P(n^{-\lambda}) \\
&\leq -(P_n - P_0)\{L_1(\bar{Q}_{n,\epsilon_n}) - L_1(\bar{Q}_n)\} + O_P(n^{-\lambda}).
\end{aligned}$$

We have that the variation norm and supremum norm of $L_1(\bar{Q}^*_n) - L_1(\bar{Q}_n)$ are bounded by $M_{L_1(Q)v,n}$ and $M_{1Q,n}$, respectively. Let $\mathcal{F}_n = \{L_1(\bar{Q}_\epsilon) - L_1(\bar{Q}) :$

$\bar{Q} \in \bar{\mathcal{Q}}_n, \epsilon\}$ and note that it is a subset of $\mathcal{F}_n^+ = \{f \in D^{k_1}[0,\tau] :\| f \|_v \leq M_{L_1(Q)v,n}, \| f \|_\infty < M_{1Q,n}\}$. We also note that the envelope F_n of $\mathcal{F}_n$ satisfies $\| F_n \|_{P_0} \leq M_{1Q,n}$. Then, the right-hand side of the above inequality can be represented as $(P_n - P_0)f_n + O_P(n^{-\lambda})$ for an $f_n \in \mathcal{F}_n$. Let $\mathcal{F}_{1n} = \mathcal{F}_n / M_{L_1(Q)v,n}$, and let F_{1n} be its envelope. We have $\| F_{1n} \|_{P_0} \leq M_{1Q,n}/M_{L_1(Q)v,n}$. By assumption we have $\sup_\Lambda \sqrt{\log(1 + N(\epsilon \| F_{1n} \|_{P_0}, \mathcal{F}_{1n}, L^2(\Lambda))} = O(\epsilon^{-(1-\alpha_1)})$. We have $\mathcal{F}_n \subset \mathcal{F}_n^+$, and the latter set has a covering number bounded by the choice $\alpha(d) = 1/(d+1)$ so that we know that $\alpha_1 \leq 1/(d+1)$. Define $d_{01}^n(\bar{Q}_n^*, \bar{Q}_{0n}) = d_{01}(\bar{Q}_n^*, \bar{Q}_{0n})/M_{L_1(Q)v,n}$. We have

$$d_{01}^n(\bar{Q}_n^*, \bar{Q}_{0n}) \leq (P_n - P_0)f_{1n} + O_P(n^{-\lambda_1}/M_{L_1(Q)v,n}),$$

where $f_{1n} \in \mathcal{F}_{1n}$, $\| F_{1n} \|_{P_0} \leq M_{1Q,n}/M_{L_1(Q)v,n}$. We also want to bound $\| f_{1n} \|_{P_0}$ in terms of $d_{01}^n(\bar{Q}_n^*, \bar{Q}_0)$. For this purpose we note

$$\begin{aligned}
\| L_1(\bar{Q}_n^*) - L_1\bar{Q}_n) \|_{P_0} &\leq \| L_1(\bar{Q}_n^*) - L_1(\bar{Q}_{0n}) \|_{P_0} + \| L_1(\bar{Q}_n) - L_1(\bar{Q}_{0n}) \|_{P_0} \\
&\leq M_{2Q,n}\{d_{01}(\bar{Q}_n^*, \bar{Q}_{0n})\}^{0.5} + O_P(n^{-\lambda_1/2}) \\
&= M_{2Q,n}M_{L_1(Q)v,n}\{d_{01}^n(\bar{Q}_n^*, \bar{Q}_{0n})\}^{0.5} + O_P(n^{-\lambda_1/2}).
\end{aligned}$$

So we have the following setting for the analysis of $d_{01}^n(\bar{Q}_n^*, \bar{Q}_0)$:

$$\begin{array}{l}
d_{01}^n(\bar{Q}_n^*, \bar{Q}_{0n}) \leq (P_n - P_0)f_{1n} + O_P(n^{-\lambda_1}/M_{L_1(Q)v,n}) \\
f_{1n} \in \mathcal{F}_{1n}, \| F_{1n} \| < M_{1Q,n}/M_{L_1(Q)v,n} \\
\sup_Q \sqrt{\log(1 + N(\epsilon \| F_{1n} \|_{P_0}, \mathcal{F}_{1n}, L^2(Q)))} = O(\epsilon^{-(1-\alpha_1)}) \\
\| f_{1n} \|_{P_0} \leq M_{2Q,n}M_{L_1(Q)v,n}\{d_{01}^n(\bar{Q}_n^*, \bar{Q}_{0n})\}^{0.5}.
\end{array}$$

We now apply Lemma 10 with $d_0 = d_{01}^n$, $Q_n = \bar{Q}_n^*$, $Q_0 = \bar{Q}_{0n}$, $f_n = f_{1n}$, $\mathcal{F}_n = \mathcal{F}_{1n}$, $\alpha = \alpha_1$, $M_n = M_{1Q,n}/M_{L_1(Q)v,n}$ and $M_{2n} = M_{2Q,n}M_{L_1(Q)v,n}$. This proves

$$d_{01}^n(\bar{Q}_n^*, \bar{Q}_{0n}) \leq n^{-1/2}n^{-\alpha_1/2}\frac{C(M_{L_1(Q)v,n}, M_{2Q,n}, M_{1Q,n}, \alpha_1)}{M_{L_1(Q)v,n}}.$$

Thus,

$$d_{01}(\bar{Q}_n^*, \bar{Q}_{0n}) \leq n^{-1/2}n^{-\alpha_1/2}C(M_{L_1(Q)v,n}, M_{2Q,n}, M_{1Q,n}, \alpha_1).$$

This completes the proof. □

An immediate corollary of this lemma is given below. Note that the rate $n^{-\lambda_1^*}$ guaranteed by this corollary is the same as the rate $r_{Q,MLE}(n)$. Thus, if one uses Super-Learner I, then the conditions stated guarantee that the rate of the initial estimator is preserved, even though the model is allowed to be unbounded.

Corollary 1 *Consider the super-learner $\hat{\bar{Q}}(P_n)$ defined in (2) or (8).*

Suppose $d_{01}(\bar{Q}_n, \bar{Q}_{0n}) = O_P(n^{-\lambda_1})$ for a $\lambda_1 \in \mathbb{R}^{k_1}_{>0}$ with $\lambda_1 > 0.5$. We can set λ_1 so that $n^{-\lambda_1} = O(r_Q^2(n))$ or $n^{-\lambda_1} = O(r_{Q,MLE}(n))$.

Suppose also that $\alpha_1 \in \mathbb{R}^{k_1}_{>0}$ is chosen so that for $\mathcal{F}_{1n} = \{L_1(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}_n\}/M_{L_1(Q)v,n}$ with envelope $F_{1n} < M_{1Q,n}/M_{L_1(Q)v,n}$ we have

$$\sup_{\Lambda} \sqrt{\log(1 + N(\epsilon \parallel F_{1n} \parallel_{P_0}, \mathcal{F}_{1n}, L^2(\Lambda)))} = O(\epsilon^{-(1-\alpha_1)}).$$

Let $d_1 \in \mathbb{N}^{k_1}_{>0}$ be the vector of integers indicating the dimension of the domain of $\bar{Q} = (\bar{Q}_1, \ldots, \bar{Q}_{k_1})$. In a nonparametric model $\mathcal{M}_n$ we have that this holds for $\alpha_1 = 1/(d_1+1)$, so that we will always have $\alpha_1 \geq 1/(d_1+1)$.

Let $\lambda_1^* = 0.5 + \alpha_1/4$. Assume that

$$\begin{aligned} \frac{\max(M_{1Q,n}, M_{2Q,n}^2)\log K_{1n}}{n} &= O(n^{-\lambda_1}) \\ \frac{\max(M_{1G_n}, M_{2G_n}^2)\log K_{2n}}{n} &= O(n^{-\lambda_2}) \\ n^{-\alpha_1/4} C(M_{L_1(Q)v,n}, M_{2Q,n}, M_{1Q,n}, \alpha_1) &= O(1). \end{aligned}$$

Then, $d_{01}((\bar{Q}_n^*, \bar{Q}_0) = O_P(n^{-\lambda_1^*})$.

Suppose that the model $\mathcal{M}$ is bounded and we select $\mathcal{M}_n = \mathcal{M}$. Then we can not only show that we preserve a rate $n^{-\lambda_1^*}$, as in the previous corollary, but we can now guarantee preservation of the rate of the initial estimator $\bar{Q}_n$. This result is stated in the following lemma.

Lemma 14 *Assume the model $\mathcal{M}$ is bounded. Assume $d_{01}(Q_n, Q_0) = O_P(n^{-\lambda_1})$ and $d_{02}(G_n, G_0) = O_P(n^{-\lambda_2})$. For a $\lambda_1 \in \mathbb{R}^{k_1+1}_{>0}$ and $\lambda_2 \in \mathbb{R}^{k_2+1}_{>0}$, we define $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^{k_1+k_2+2}_{>0}$.*

Let $S(Q, G, \epsilon) = \frac{d}{d\epsilon} L_1(Q_\epsilon)$. Note, $S(Q, G, \epsilon) = (S_j(Q, G, \epsilon(j)) : j = 1, \ldots, k_1{+}1)$. Let $\epsilon_n = \arg\min_\epsilon P_n L_1(Q_{n,\epsilon})$. Assume it solves $P_n S(Q_n, G_n, \epsilon_n) = 0$. Assume the following analytic properties of the least favorable submodel:

- *$d_{01}(Q_{n,\epsilon_n}, Q_0) \to 0$, $d_{01}(Q_n, Q_0) \to 0$ and $d_{02}(G_n, G_0) \to 0$ imply that $\epsilon_n \to 0$;*

- *Along a sequence (Q_n, G_n) with $d_0((Q_n, G_n), (Q_0, G_0)) = O(n^{-\lambda})$, we have*

$$\begin{aligned} P_0\{S(Q_n, G_n\epsilon_n) - S(Q_n, G_n, \epsilon_{0n})\} &= \frac{d}{d\epsilon_{0n}} P_0 S(Q_n, G_n, \epsilon_{0n})(\epsilon_n - \epsilon_{0n}) \\ &\quad + o(\mid \epsilon_n - \epsilon_{0n} \mid); \end{aligned}$$

- *For a sequence $\epsilon_n \rightarrow 0$ and (Q_n, G_n) with $d_0((Q_n, G_n), (Q_0, G_0)) = O(n^{-\lambda})$, we have $\frac{d}{d\epsilon_n} P_0 S(Q_n, G_n, \epsilon_n)$ converges to $c_0 = -\frac{d}{d\epsilon} P_0 S(Q_0, G_0, \epsilon)\Big|_{\epsilon=0} > 0$ with $c_0 > 0$;*
- *For some $\delta > 0$, $\sup_{|\epsilon|<\delta, P\in\mathcal{M}} \parallel S(Q, G, \epsilon) \parallel_v < \infty$.*
- *If $d_0((Q_n, G_n), (Q_0, G_0)) = O(n^{-\lambda})$ and $\epsilon_n \rightarrow 0$, then $P_0\{S(Q_n, G_n, \epsilon_n) - S(Q_0, G_0, \epsilon_n\} = O(n^{-\lambda/2})$;*
- *If $\epsilon_n = O(n^{-\lambda_1/2})$ and $d_{01}(Q_n, Q_0) = O(n^{-\lambda_1})$, then $d_{01}(Q_{n,\epsilon_n}, Q_0) = O(n^{-\lambda_1})$.*

Then,

$$\epsilon_n = O_P(n^{-\lambda_1/2}) \text{ and } d_{01}(Q_{n,\epsilon_n}, Q_0) = O_P(n^{-\lambda_1}).$$

Proof: Define $\epsilon_{0n} = \arg\min_\epsilon P_0 L_1(Q_{n,\epsilon})$, which solves $P_0 S(Q_n, G_n, \epsilon_{0n}) = 0$. We also define $\epsilon_0 = 0$ which solves $P_0 S(Q_0, G_0, \epsilon_0) = 0$. It follows, analogue to the proof of previous Lemma 13, that $d_{01}(Q_{n,\epsilon_n}, Q_{0n}) = o_P(1)$. By the first assumption, $d_{01}(Q_{n,\epsilon_n}, Q_{0n}) = o_P(1)$ implies $\epsilon_n = o_P(1)$, which also implies $\epsilon_{0n} = o_P(1)$. Thus, we have $\epsilon_n = o_P(1)$ and $\epsilon_{0n} = o_P(1)$. Given that we know that $\epsilon_n = o_P(1)$, we want to prove that $\epsilon_n = O_P(n^{-\lambda_1/2})$. We will prove this by proving $\epsilon_n - \epsilon_{0n} = O_P(n^{-0.5})$ and $\epsilon_{0n} = O_P(n^{-\lambda_1/2})$. Note

$$P_0\{S(Q_n, G_n, \epsilon_n) - S(Q_n G_n, , \epsilon_{0n})\} = -(P_n - P_0) S(Q_n, G_n, \epsilon_n).$$

We assumed differentiability in ϵ uniformly in the sequence Q_n with $d_{01}(Q_n, Q_0) = O_P(n^{-\lambda_1})$ in the sense that:

$$P_0\{S(Q_n, G_n, \epsilon_n) - S(Q_n, G_n, \epsilon_{0n})\} = \frac{d}{d\epsilon_{0n}} P_0 S(Q_n, G_n, \epsilon_{0n})(\epsilon_n - \epsilon_{0n}) + o(\mid \epsilon_n - \epsilon_{0n} \mid).$$

We also assumed that $-\frac{d}{d\epsilon_{0n}} P_0 S(Q_n, G_n, \epsilon_{0n})$ converges to $c_0 = -\frac{d}{d\epsilon_0} P_0 S(Q_0, G_0, \epsilon_0) > 0$. Then, it follows that

$$\epsilon_n - \epsilon_{0n} = c_0^{-1}(P_n - P_0) S(Q_n, G_n, \epsilon_n) + o(\mid \epsilon_n - \epsilon_{0n} \mid).$$

By assumption, $\sup_{Q\in\mathcal{Q}, G\in\mathcal{G}, |\epsilon|<\delta} \parallel S(Q, G, \epsilon) \parallel_v < C < \infty$, so that it follows that $\epsilon_n - \epsilon_{0n} = O_P(n^{-1/2})$.

Consider now the equations $P_0 S(Q_n, G_n, \epsilon_{0n}) = P_0 S(Q_0, G_0, \epsilon_0) = 0$. We have

$$P_0\{S(Q_0, G_0, \epsilon_{0n}) - S(Q_0, G_0, \epsilon_0)\} = -P_0\{S(Q_n, G_n, \epsilon_{0n}) - S(Q_0, G_0, \epsilon_{0n})\}.$$

Thus,

$$\epsilon_{0n} - \epsilon_0 = c_0^{-1} P_0\{S(Q_n, G_n, \epsilon_{0n}) - S(Q_0, G_0, \epsilon_{0n})\} + o(\mid \epsilon_{0n} - \epsilon_0 \mid).$$

We assumed that $P_0\{S(Q_n, G_n, \epsilon_{0n}) - S(Q_0, G_0, \epsilon_{0n}\} = O_P(n^{-\lambda_1/2})$ if $d_{01}(Q_n, Q_0) = O_P(n^{-\lambda_1})$ and $\epsilon_{0n} = o_P(1)$. This proves that $\epsilon_{0n} - \epsilon_0 = O_P(n^{-\lambda_1/2}) + o(\mid \epsilon_{0n} - \epsilon_0 \mid)$, and thus $\epsilon_{0n} = O_P(n^{-\lambda_1/2})$, which proves the first statement of the lemma. By the final assumption, we have that the latter, combined with $d_{01}(Q_n, Q_0) = O_P(n^{-\lambda_1})$ implies $d_{01}(Q_{n,\epsilon_n}, Q_0) = O_P(n^{-\lambda_1})$. □

E Preservation of the rate of initial estimator for the one-step CV-TMLE

Consider the submodel $\{Q_\epsilon : \epsilon\}$ of the type presented in main article and previous section. The following lemma is an immediate consequence of the oracle inequality of the cross-validation selector for the loss function L_{1j}, applied to the set of candidate estimators $P_n \rightarrow Q_{jn,\epsilon(j)} = \hat{Q}_{j,\epsilon(j)}(P_n)$ indexed by $\epsilon(j)$, for each $j = 1, \ldots, k_1 + 1$.

Lemma 15 *Let* $\epsilon_n = \arg\min_\epsilon E_{B_n} P^1_{n,B_n} L_1(Q_{n,B_n,\epsilon})$, *and* $\tilde{\epsilon}_n = \arg\min_\epsilon E_{B_n} P_0 L_1(\hat{Q}_\epsilon(P^0_{n,B_n}))$. *Assume* $d_{01}(\hat{Q}(P^0_{n,B_n}), Q_{0n}) = O_P(n^{-\lambda_1})$. *We have*

$$\begin{aligned} E_{B_n} d_{01}(\hat{Q}_{\epsilon_n}(P^0_{n,B_n}), Q_{0n}) \quad &\leq \quad (1+2\delta) \min_\epsilon E_{B_n} d_{01}(\hat{Q}_\epsilon(P^0_{n,B_n}), Q_{0n}) \\ &\quad + \frac{C(M_{1Q,n}, M_{2Q,n}, \delta) \log K_{1n}}{nq}. \end{aligned}$$

By convexity of the loss function $L_1(Q)$, *this implies*

$$\begin{aligned} d_{01}(E_{B_n} \hat{Q}_{\epsilon_n}(P^0_{n,B_n}), Q_{0n}) \quad &\leq \quad (1+2\delta) \min_\epsilon E_{B_n} d_{01}(\hat{Q}_\epsilon(P^0_{n,B_n}), Q_{0n}) \\ &\quad + \frac{C(M_{1Q,n}, M_{2Q,n}, \delta) \log K_{1n}}{nq}. \end{aligned}$$

We have

$$\min_\epsilon E_{B_n} d_{01}(\hat{Q}_\epsilon(P^0_{n,B_n}), Q_{0n}) \leq E_{B_n} d_{01}(\hat{Q}(P^0_{n,B_n}), Q_{0n}) = O_P(n^{-\lambda_1}).$$

Thus, if $C(M_{1Q,n}, M_{2Q,n}, \delta) \log K_{1n}/(nq) = O(n^{-\lambda_1})$, *then*

$$d_{01}(E_{B_n} Q_{n,B_n,\epsilon_n}, Q_{0n}) = O_P(n^{-\lambda_1}).$$

It also follows that for each B_n, $d_{01}(\hat{Q}_{\epsilon_n}(P^0_{n,B_n}), Q_{0n}) = O_P(n^{-\lambda_1})$.

F General bounding of second order remainder $R_2()$.

Recall the expansion $\Psi(Q_n^*)-\Psi(Q_0)=(P_n-P_0)D^*(Q_n^*,G_n)+R_{20}(Q_n^*,Q_0,G_n,G_0)$ for the TMLE Q_n^*. For asymptotic efficiency, we need that $R_{20}((Q_n^*,G_n),(Q_0,G_0))=o_P(n^{-1/2})$. Therefore, given $d_0(P_n^*,P_0)$, we want to bound $R_{20}(P_n^*,P_0)$ in terms of $d_0(P_n^*,P_0)$. Typically, our loss functions are log-likelihood or squared error loss functions in which case $d_{01j}(Q_j,Q_{j0})$ equals or is equivalent with $\parallel Q_j-Q_{j0}\parallel_{P_0}^2$, and similarly, $d_{02j}(G_j,G_{j0})$ equals or is equivalent with $\parallel G_j-G_{0j}\parallel_{P_0}$: here we use that the log likelihood based dissimilarity is equivalent with the square of the $L^2(P_0)$-norm if the log-likelihood loss is uniformly bounded on the model (our $M_1<\infty$). In all our applications, by using Cauchy-Schwarz inequality $\int fgdP_0\leq\parallel f\parallel_{P_0}\parallel g\parallel_{P_0}$, we have been able to naturally bound $R_{20}((Q_n^*,G_n),(Q_0,G_0))$ in terms of $L^2(P_0)$-norms of $Q_{jn}^*-Q_{j0}$ with $j\in\{1,\ldots,k_1+1\}$, and $L^2(P_0)$-norms of $G_{jn}-G_{j0}$ with $j\in\{1,\ldots,k_2+1\}$. Plugging in the rates $o_P(n^{-1/4})$ for each of these $L^2(P_0)$ norms, then proves that $R_2((Q_n^*,G_n),(Q_0,G_0))=o_P(1/\sqrt{n})$. So one is typically able to bound $R_2()$ as follows:

$$\begin{aligned}R_2((Q_n^*,G_n),(Q_0,G_0)) \leq & \sum_{j_1=1}^{k_1+1}\sum_{j_1'}^{k_1+1}c_{Qn}(j_1,j_1')\parallel Q_{j_1n}^*-Q_{j_10}\parallel_{P_0}\parallel Q_{j_1'n}^*-Q_{j_1'0}\parallel_{P_0}\\ &+\sum_{j_2=1}^{k_2+1}\sum_{j_2'}^{k_2+1}c_{Gn}(j_2,j_2')\parallel G_{j_2n}-G_{j_20}\parallel_{P_0}\parallel G_{j_2'n}-G_{j_2'0}\parallel_{P_0}\\ &+\sum_{j_1=1}^{k_1+1}\sum_{j_2=1}^{k_2+1}c_{QGn}(j_1,j_2)\parallel Q_{j_1n}^*-Q_{j_10}\parallel_{P_0}\parallel G_{j_2n}-G_{j_20}\parallel_{P_0},\end{aligned}$$

for certain matrices $c_{Qn}(),c_{Gn}(),c_{QGn}()$ of coefficients. In other words, the second order term $R_2()$ will be a sum of second order integral terms where each integral has a second order integrand, either a square difference or a cross-product of one difference with another difference, among the k_1+k_2+2 possible differences $Q_{j_1n}^*-Q_{j_10}$, $G_{j_2n}-G_{j_20}$. Of course, typically not all $(k_1+k_2+2)^2$ possible second order terms will appear in R_2. Each of these second order integrals can be bounded with Cauchy-Schwarz inequality in terms of a corresponding product of the norms of the two differences, resulting in the above type of bound. Since the second order integrands will typically also involve other functions depending on P_0 and (Q_n^*,G_n), these bounds will also involve bounds on these functions (e.g. supremum norm), resulting in coefficients that depend on n.

For bounded models, these coefficients will be uniformly bounded by fixed constants, while for sequence of models $\mathcal{M}_n$ converging to an unbounded model $\mathcal{M}$, we will have that $R_2((Q_n^*, G_n), (Q_{0n}, G_{0n}))$ is bounded as above but with coefficient matrices that are upper bounded by constants depending on the bounds of model $\mathcal{M}_n$. In the latter case, we will have that $R_2((Q_n^*, G_n), (Q_{0n}, G_{0n}))$ is bounded by some constant M_{R_2n} times the above second order polynomial sum in $L^2(P_0)$-norms, where M_{R_2n} will now possibly converge to infinity. That is, we obtain a bound of the form $M_{R_2n} f_{R_2}(d_0(P_n^*, P_{0n}))$, and by letting M_{R_2n} converge slowly enough to infinity, this will still be $o_P(n^{-1/2})$.

G A single updating step in TMLE suffices for approximately solving the efficient influence curve equation

The following lemma proves that for a local least favorable submodel with a 1-dimensional ϵ and $n^{-1/4+}$-consistent initial estimators, the one-step TMLE already solves $P_n D^*(Q_{n,\epsilon_n}, G_n) = o_P(n^{-1/2})$ under some regularity conditions.

Lemma 16 $\Psi : \mathcal{M} \to \mathbb{R}$ *is a pathwise differentiable parameter at* P *with canonical gradient* $D^*(P)$*, and assume* $\Psi(P) = \Psi(Q(P))$ *and* $D^*(P) = D^*(Q(P), G(P))$ *for parameters* $Q : \mathcal{M} \to \mathcal{Q} = \{Q(P) : P \in \mathcal{M}\}$ *and* $G : \mathcal{M} \to \mathcal{G} = \{G(P) : P \in \mathcal{M}\}$*. Let* $R_2()$ *be defined by* $\Psi(P) - \Psi(P_0) = (P - P_0)D^*(P) + R_2(P, P_0)$*, and let* $R_2(P, P_0) = R_{20}((Q, G), (Q_0, G_0))$*. Suppose* $Q_0 = \arg\min_Q P_0 L(Q)$ *for some loss function* $L(Q)$ *and that, for any* $Q \in \mathcal{Q}$ *and* $G \in \mathcal{G}$*,* $\{Q_\epsilon : \epsilon\} \subset \mathcal{Q}$ *is a one dimensional parametric submodel through* Q *with* $\frac{d}{d\epsilon} L(Q_\epsilon)\Big|_{\epsilon=0} = D^*(Q, G)$*. Let* (Q_n, G_n) *be an initial estimator of* (Q_0, G_0)*, and consider the one-step TMLE* $\Psi(Q_{n,\epsilon_n})$ *with* $\epsilon_n = \arg\min_\epsilon P_n L(Q_{n,\epsilon})$*.*

Let $f_n(\epsilon) = P_n D^*(Q_{n,\epsilon}, G_n)$ *and* $g_n(\epsilon) = \frac{d}{d\epsilon} P_n L(Q_{n,\epsilon})$*. Let* $f_n'(\epsilon) = \frac{d}{d\epsilon} f_n(\epsilon)$ *and* $g_n'(\epsilon) = \frac{d}{d\epsilon} g_n(\epsilon)$*. Let* $\epsilon_0 = 0$*.*

Assume

- $f_n(\epsilon_n) = f_n(0) + f_n'(0)\epsilon_n + O_P(\epsilon_n^2)$ *and* $g_n(\epsilon_n) = g_n(0) + g_n'(0)\epsilon_n + O_P(\epsilon_n^2)$*;*
- $\epsilon_n^2 = o_P(n^{-1/2})$*;*
- $\{\frac{d}{d\epsilon_n} D^*(Q_{n,\epsilon_n}, G_n) - \frac{d^2}{d\epsilon_n^2} L(Q_{n,\epsilon_n})\}/n^{0.25}$ *falls in a* P_0*-Donsker class with probability tending to 1;*

- $$\begin{aligned} P_0\left\{\frac{d}{d\epsilon_0}D^*(Q_{n,\epsilon_0},G_n)-\frac{d}{d\epsilon_0}D^*(Q_{0,\epsilon_0},,G_0)\right\} &= O_P(n^{-1/4}) \quad (25)\\ P_0\left\{\frac{d^2}{d\epsilon_0^2}L(Q_{n,\epsilon_0})-\frac{d^2}{d\epsilon_0^2}L(Q_{0,\epsilon_0})\right\} &= O_P(n^{-1/4}); \quad (26)\end{aligned}$$

- $$P_0\frac{d^2}{d\epsilon_0^2}L(Q_{0,\epsilon_0}) = -P_0D^*(P_0)\{D^*(P_0)\}^\top. \quad (27)$$

  If $L(Q(P)) = -\log p_{Q(P),\eta(P)}$ for some density parameterization $(Q,\eta)\to p_{Q,\eta}$, then (27) holds;

- $\frac{d}{d\epsilon_0}R_{20}((Q_{0,\epsilon_0},G_0),(Q_0,G_0))=0.$

Then, $P_nD^*(Q_{n,\epsilon_n},G_n)=o_P(1/\sqrt{n})$.

Proof of Lemma: Firstly, by the fact that $Q_{n,\epsilon}$ has score $D^*(Q_n,G_n)$ at $\epsilon=0$, it follows that $f_n(0)=g_n(0)$. We also know that $g_n(\epsilon_n)=0$, and we want to show that $f_n(\epsilon_n)=o_P(n^{-1/2})$. Let $\epsilon_0=0$. By the second order Tailor expansion assumption for f_n,g_n at $\epsilon=0$, we have

$$\begin{aligned} f_n(\epsilon_n) &= f_n(\epsilon_n)-g_n(\epsilon_n)\\ &= f_n(0)-g_n(0)+\epsilon_n(f_n'-g_n')(0)+O(\epsilon_n^2)\\ &= \epsilon_n\left\{\frac{d}{d\epsilon_0}P_nD^*(Q_{n,\epsilon_0},G_n)-\frac{d^2}{d\epsilon_0^2}P_nL(Q_{n,\epsilon_0})\right\}+O(\epsilon_n^2).\end{aligned}$$

By assumption, $\epsilon_n^2=o_P(n^{-1/2})$, so that $O(\epsilon_n^2)=o_P(n^{-1/2})$. Thus, it remains to show

$$P_n\frac{d}{d\epsilon_0}D^*(Q_{n,\epsilon_0},G_n)-P_n\frac{d^2}{d\epsilon_0^2}L(Q_{n,\epsilon_0})=O_P(n^{-1/4}).$$

By our Donsker class assumption, we have

$$(P_n-P_0)\left\{\frac{d}{d\epsilon_0}D^*(Q_{n,\epsilon_0},G_n)-\frac{d^2}{d\epsilon_0^2}L(Q_{n,\epsilon_0})\right\}/n^{1/4}=O_P(n^{-1/2}).$$

Thus, it remains to show

$$\frac{d}{d\epsilon_0}P_0D^*(Q_{n,\epsilon_0},G_n)-P_0\frac{d^2}{d\epsilon_0^2}L(Q_{n,\epsilon_0})=O_P(n^{-1/4}).$$

By assumptions (25), we have that the left-hand side of last expression equals

$$\frac{d}{d\epsilon_0} P_0 D^*(Q_{0,\epsilon_0}, G_0) - P_0 \frac{d^2}{d\epsilon_0^2} L(Q_{0,\epsilon_0}) + O_P(n^{-1/4}),$$

so that it remains to show that the first term equals zero. By $-P_0 D^*(P) = \Psi(P) - \Psi(P_0) - R_2(P, P_0)$, it follows that

$$\frac{d}{d\epsilon_0} P_0 D^*(Q_{0,\epsilon_0}, G_0) = -\frac{d}{d\epsilon_0} \Psi(Q_{0,\epsilon_0}) + \frac{d}{d\epsilon_0} R_2((Q_{0,\epsilon_0}, G_0), (Q_0, G_0)).$$

By assumption we have $\frac{d}{d\epsilon_0} R_2((Q_{0,\epsilon_0}, G_0), (Q_0, G_0)) = 0$. By definition of the pathwise derivative at P_0, we have that the derivative $\Psi(Q_{0,\epsilon}) = \Psi(P_{0,\epsilon})$ at $\epsilon = 0$ equals $P_0 D^*(P_0)\{D^*(P_0)\}^\top$. Thus, we have shown

$$\frac{d}{d\epsilon_0} P_0 D^*(Q_{0,\epsilon_0}, G_0) = -P_0 D^*(P_0)\{D^*(P_0)\}^\top.$$

Thus, it remains to show (27), which thus holds by assumption. Suppose that $L(Q(P)) = -\log p_{Q(P),\eta(P)}$ for some density parameterization $(Q, \eta) \rightarrow p_{Q,\eta}$. Then $L(Q_{0,\epsilon}) = -\log p_{Q_{0,\epsilon},\eta_0}$. Since $\{p_{Q_{0,\epsilon},\eta_0} : \epsilon\}$ is a correctly specified parametric model, we have that the second derivative of $-P_0 \log p_{Q_{0,\epsilon},\eta_0}$ at $\epsilon = 0$ equals its information matrix (i.e., covariance matrix of its score) $P_0 \frac{d}{d\epsilon} \log p_{Q_{0,\epsilon},\eta_0} \{\frac{d}{d\epsilon} \log p_{Q_{0,\epsilon},\eta_0}\}^\top$ at $\epsilon = 0$. However, the latter equals $-P_0 D^*(P_0)\{D^*(P_0)\}^\top$, which proves (27). This completes the proof of $f_n(\epsilon_n) = o_P(n^{-1/2})$. □

In the main article we have not proposed a 1-dimensional local least favorable submodel as in Lemma 16, even though our results are straightforwardly generalized to that case. Instead we proposed a $k_1 + 1$-dimensional least favorable submodel that uses a 1-dimensional $\epsilon(j)$ for updating Q_{jn} for each $j = 1, \ldots, k_1 + 1$. We will now prove the desired lemma for such a submodel by application of the above lemma across all j.

Lemma 17 *Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be pathwise differentiable with canonical gradient $D^*(P) = D^*(Q, G)$ and let $\Psi(P) = \Psi(Q(P))$ for $Q(P) = (Q_1(P), \ldots, Q_{k_1+1}(P))$. For a given Q, we define $\Psi_{Q,j} : \mathcal{M} \rightarrow \mathbb{R}$ by $\Psi_{Q,j}(P) = \Psi(Q_{-j}, Q_j(P))$, $j = 1, \ldots, k_1 + 1$. Let $D^*_{Q,j}(P) = D^*_{Q,j}(Q_j(P), Q_{-j}(P), G(P))$ be the efficient influence curve of $\Psi_{Q,j}$ at P, and define $R_{2,Q,j}(P, P_0) = R_{2,Q,j}((Q(P), G(P)), (Q_0, G_0))$ by $\Psi_{Q,j}(P) - \Psi_{Q,j}(P_0) = (P - P_0) D^*_{Q,j}(P) + R_{2,Q,j}(P, P_0)$, $j = 1, \ldots, k_1 + 1$. Here $Q_{-j} = (Q_l : l \neq j, l \in \{1, \ldots, k_1 + 1\})$. We have $D^*(P) = \sum_{j=1}^{k_1+1} D^*_{Q(P),j}(P)$.*

Let $Q_n \in \mathcal{Q}_n, G_n \in \mathcal{G}_n$ be a given initial estimator. Let $\{Q_{jn,\epsilon(j)} : \epsilon(j)\} \subset \mathcal{Q}_{jn}$ be a submodel through Q_{jn} at $\epsilon(j) = 0$ and satisfying $\frac{d}{d\epsilon(j)} L_{1,j}(Q_{jn,\epsilon(j)})\Big|_{\epsilon(j)=0} =$

$D^*_{Q_n,j}(Q_n,G_n)$, $j=1,\ldots,k_1+1$. Let $\{Q_{n,\epsilon}:\epsilon\}\subset\mathcal{Q}_n$ be defined by $Q_{n,\epsilon}=(Q_{jn,\epsilon(j)}:j=1,\ldots,k_1+1)$. Let $\epsilon_n=\arg\min_\epsilon P_nL_1(Q_{n,\epsilon})$, where $P_nL_1(Q_{n,\epsilon})=(P_nL_{1j}(Q_{jn,\epsilon(j)}):j=1,\ldots,k_1+1)$.

Suppose that by application of the previous lemma to $\Psi_{Q_n,j}:\mathcal{M}\rightarrow\mathbb{R}$, submodel $\{Q_{jn,\epsilon(j)}:\epsilon(j)\}$, loss function $L_{1j}(Q_j)$, $\epsilon_n(j)=\arg\min_{\epsilon(j)}P_nL_{1j}(Q_{jn,\epsilon(j)})$, and one-step TMLE $Q_{jn,\epsilon_n(j)}$, we establish its conclusion $P_nD^*_{Q_n,j}(Q_{jn,\epsilon_n(j)},Q_{-jn},G_n)=o_P(n^{-1/2})$. For completeness, Lemma 18 explicitly states these j specific conditions of the previous lemma, which are sufficient for this conclusion.

We wish to establish that $P_nD^*(Q_{n,\epsilon_n},G_n)=o_P(n^{-1/2})$, where

$$P_nD^*(Q_{n,\epsilon_n},G_n)=\sum_{j=1}^{k_1+1}P_nD^*_{Q_{n,\epsilon_n},j}(Q_{jn,\epsilon_n(j)},Q_{-jn,\epsilon_n},G_n).$$

For each $j=1,\ldots,k_1+1$, assume

1. *Let $f_{nj}=D^*_{Q_n,j}(Q^*_{jn},Q_{-jn},G_n)-D^*_{Q_n,j}(Q^*_{jn},Q^*_{-jn},G_n)$, and assume $(P_n-P_0)f_{nj}=o_P(n^{-1/2})$.*

   We can achieve this by applying Lemma 9: Assume $\sup_{Q\in\mathcal{Q}_n,G\in\mathcal{G}_n}\parallel D^*_{Q,j}(Q,G)\parallel_v\leq M_{D^*,v,n}$. Let $\mathcal{F}_{jn}=\{D^*_{Q,j}(Q,G):Q\in\mathcal{Q}_n,G\in\mathcal{G}_n\}/M_{D^*,v,n}$ with envelope bounded (up till constant) by $F_{jn}=M_{D^*,n}/M_{D^*,v,n}$, and let α^*_j (which can always be chosen to be smaller than $1/(d+1)$) be such that $\sup_\Lambda\sqrt{\log(1+N(\epsilon\mid F_{jn}\mid,\mathcal{F}_{jn},L^2(\Lambda))}<K\epsilon^{\alpha^*_j-1}$ for some $K<\infty$. Let $r_{0j}(n)$ be such that $\parallel f_{nj}\parallel_{P_0}<r_{0j}(n)$ with probability tending to 1. Define $r^*_{0j}(n)=\max(r_{0j}(n),n^{-1/4})$. Then, by Lemma 9, we have

   $$E\mid\sqrt{n}(P_n-P_0)f_{nj}\mid\leq\left\{r^*_{0j}(n)/M_{D^*,v,n}\right\}^{\alpha^*_j}M_{D^*,v,n}+\left\{r^*_{0j}(n)/M_{D^*,v,n}\right\}^{2\alpha^*_j-2}n^{-0.5}.$$

   Assume $M_{D^*,v,n}$ converges slowly enough to infinity so that the right-hand side is $o(1)$. Then, $(P_n-P_0)f_{nj}=o_P(n^{-1/2})$;

2. *$R_{2,Q_n,j}(((Q^*_{jn},Q^*_{-jn}),G_n),(Q_0,G_0))-R_{2,Q_n,j}(((Q^*_{jn},Q_{-jn}),G_n),(Q_0,G_0))=o_P(n^{-1/2})$;*

3. *Let $f_{nj,1}=D^*_{Q_n,j}(Q^*_n,G_n)-D^*_{Q^*_n,j}(Q^*_n,G_n)$, and assume $(P_n-P_0)f_{nj,1}=o_P(n^{-1/2})$. As above, we can achieve this by applying Lemma 9;*

4. *$R_{2,Q^*_n,j}((Q^*_n,G_n),(Q_0,G_0))-R_{2,Q_n,j}((Q^*_n,G_n),(Q_0,G_0))=o_P(n^{-1/2})$;*

5. *$\Psi_{Q^*_n,j}(Q^*_{jn})-\Psi_{Q^*_n,j}(Q_{j0})-\left\{\Psi_{Q_n,j}(Q^*_{jn})-\Psi_{Q_n,j}(Q_{j0})\right\}=o_P(n^{-1/2})$.*

Then, $P_n D^*(Q_{n,\epsilon_n}, G_n) = o_P(n^{-1/2})$.

Lemma 18 *Let* $f_{nj}(\epsilon(j)) = P_n D^*_{Q_n,j}(Q_{jn,\epsilon(j)}, Q_{-jn}, G_n)$ *and* $g_{nj}(\epsilon(j)) = \frac{d}{d\epsilon(j)} P_n L_{1j}(Q_{jn,\epsilon(j)})$. *Let* $f'_{nj}(\epsilon(j)) = \frac{d}{d\epsilon(j)} f_{nj}(\epsilon(j))$ *and* $g'_{nj}(\epsilon(j)) = \frac{d}{d\epsilon(j)} g_{nj}(\epsilon(j))$. *Let* $\epsilon_0(j) = 0$.

Assume

1. *Assume that* $f_{nj}(\epsilon_n(j)) = f_{nj}(0) + f'_{nj}(0)\epsilon_n(j) + O_P(\epsilon_n(j)^2)$ *and* $g_{nj}(\epsilon_n(j)) = g_{nj}(0) + g'_{nj}(0)\epsilon_n(j) + O_P(\epsilon_n^2(j))$;

2. $\epsilon_n^2(j) = o_P(n^{-1/2})$;

3. $\{\frac{d}{d\epsilon_n(j)} D^*_{Q_n,j}(Q_{jn,\epsilon_n(j)}, Q_{-jn}G_n) - \frac{d^2}{d\epsilon_n(j)^2} L_{1j}(Q_{jn,\epsilon_n(j)})\}/n^{0.25}$ *falls in a* P_0*-Donsker class with probability tending to 1;*

4. $$\begin{aligned} \frac{d}{d\epsilon_0(j)} P_0 \left\{ D^*_{Q_n,j}(Q_{jn,\epsilon_0(j)}, Q_{-jn}, G_n) - D^*_{Q_n,j}(Q_{j0,\epsilon_0(j)}, Q_{-j0}, G_0) \right\} &= O_P(n^{-1/4}) \\ \frac{d^2}{d\epsilon_0(j)^2} P_0 \left\{ L_{1j}(Q_{jn,\epsilon_0(j)}) - L_{1j}(Q_{j0,\epsilon_0(j)}) \right\} &= O_P(n^{-1/4}); \end{aligned}$$

5. $$P_0 \frac{d^2}{d\epsilon_0(j)^2} L_{1j}(Q_{j0,\epsilon_0(j)}) = -P_0 D^*_{Q_0,j}(P_0)\{D^*_{Q_0,j}(P_0)\}^\top. \qquad (28)$$
   If $L_{1j}(Q_j(P)) = -\log p_{Q_j(P),\eta(P)}$ *for some density parameterization* $(Q_j, \eta) \rightarrow p_{Q_j,\eta}$, *then (28) holds;*

6. $\frac{d}{d\epsilon_0(j)} R_{2,Q_0,j}((Q_{j0,\epsilon_0(j)}, Q_{-j0}, G_0), (Q_0, G_0)) = 0$.

Then, $P_n D^*_{Q_n,j}(Q_{jn,\epsilon_n(j)}, Q_{-jn}, G_n) = o_P(1/\sqrt{n})$.

Proof: This is an immediate application of Lemma 16. □

Proof of Lemma 17: Consider a 1-dimensional submodel $\{P_\epsilon : \epsilon\} \subset \mathcal{M}$ with score S. We have

$$\begin{aligned} \tfrac{d}{d\epsilon}\Psi(P_\epsilon) &= \tfrac{d}{d\epsilon}\Psi(Q_\epsilon) \\ &= \tfrac{d}{d\epsilon}\Psi(Q_{1\epsilon}, \ldots, Q_{k_1+1\epsilon}) \\ &= \textstyle\sum_{j=1}^{k_1+1} \tfrac{d}{d\epsilon}\Psi(Q_{-j}, Q_{j\epsilon}). \end{aligned}$$

By pathwise differentiability of Ψ at P the left-hand side equals $PD^*(P)S$, while, by pathwise differentiability of $\Psi_{Q,j}$ at P, each j-specific term on the right-hand side equals $PD^*_{Q,j}(P)S$. This proves that

$$PD^*(P)S = \sum_{j=1}^{k_1+1} PD^*_{Q,j}(P)S = P\left\{\sum_{j=1}^{k_1+1} D^*_{Q,j}(P)\right\}S.$$

Since this holds for each $S \in T(P)$ and $D^*_{Q,j}(P) \in T(P)$ for all j, this implies $D^*(P) = \sum_{j=1}^{k_1+1} D^*_{Q,j}(P)$. This proves the first statement of the lemma. This shows also that $P_n D^*(Q^*_n, G_n) = \sum_{j=1}^{k_1+1} P_n D^*_{Q^*_n,j}(Q^*_n, G_n)$, so it suffices to prove that $P_n D^*_{Q^*_n,j}(Q^*_n, G_n) = o_P(n^{-1/2})$ for each j. In the lemma we assumed that we already established $P_n D^*_{Q_n,j}(Q^*_{jn}, Q_{-jn}, G_n) = o_P(n^{-1/2})$, by application of Lemma 18.

Firstly, we want to prove that $P_n\{D^*_{Q_n,j}(Q^*_{jn}, Q_{-jn}, G_n) - D^*_{Q_n,j}(Q^*_{jn}, Q^*_{-jn}, G_n)\} = o_P(n^{-1/2})$, which then shows that $P_n D^*_{Q_n,j}(Q^*_n, G_n) = o_P(n^{-1/2})$. This term can be represented as $P_n f_n$. We can write $P_n f_n = (P_n - P_0) f_n + P_0 f_n$. By our first assumption, we have $(P_n - P_0) f_n = o_P(1)$. So we now have to consider

$$\begin{array}{l} P_0\{D^*_{Q_n,j}(Q^*_{jn}, Q_{-jn}, G_n) - D^*_{Q_n,j}(Q^*_{jn}, Q^*_{-jn}, G_n)\} \\ = \Psi_{Q_n,j}(Q^*_{jn}) - \Psi_{Q_n,j}(Q_{j0}) + R_{2,Q_n,j}(((Q^*_{jn}, Q^*_{-jn}), G_n), (Q_0, G_0)) \\ -\Psi_{Q_n,j}(Q^*_{jn}) + \Psi_{Q_n,j}(Q_{j0}) - R_{2,Q_n,j}(((Q^*_{jn}, Q_{-jn}), G_n), (Q_0, G_0)) \\ = R_{2,Q_n,j}(((Q^*_{jn}, Q^*_{-jn}), G_n), (Q_0, G_0)) - R_{2,Q_n,j}(((Q^*_{jn}, Q_{-jn}), G_n), (Q_0, G_0)). \end{array}$$

By assumption 2., the latter is $o_P(n^{-1/2})$.This proves now that $P_n D^*_{Q_n,j}(Q^*_n, G_n) = o_P(n^{-1/2})$.

We now want to prove that $P_n\{D^*_{Q_n,j}(Q^*_n, G_n) - D^*_{Q^*_n,j}(Q^*_n, G_n)\} = o_P(n^{-1/2})$, so that we can conclude $P_n D^*_{Q^*_n,j}(Q^*_n, G_n) = o_P(n^{-1/2})$. Let $f_n = \{D^*_{Q_n,j}(Q^*_n, G_n) - D^*_{Q^*_n,j}(Q^*_n, G_n)\}$, so that this term can be represented as $P_n f_n$. We have $P_n f_n = (P_n - P_0) f_n + P_0 f_n$. By assumption 3., we have $(P_n - P_0) f_n = o_P(n^{-1/2})$. We now have to consider

$$\begin{array}{l} P_0\{D^*_{Q_n,j}(Q^*_n, G_n) - D^*_{Q^*_n,j}(Q^*_n, G_n)\} \\ = \Psi_{Q^*_n,j}(Q^*_{jn}) - \Psi_{Q^*_n,j}(Q_{j0}) + R_{2,Q^*_n,j}((Q^*_n, G_n), (Q_0, G_0)) \\ -\Psi_{Q_n,j}(Q^*_{jn}) + \Psi_{Q_n,j}(Q_{j0}) - R_{2,Q_n,j}((Q^*_n, G_n), (Q_0, G_0)). \end{array}$$

By assumption 4., we have $R_{2,Q^*_n,j}() - R_{2,Q_n,j}() = o_P(n^{-1/2})$. By assumption 5, the "second order Ψ-difference" is $o_P(n^{-1/2})$ as well. □