HHS Public Access

Author manuscript

Int J Biostat. Author manuscript; available in PMC 2018 July 22.

Published in final edited form as:

Int J Biostat. ; 13(2): . doi:10.1515/ijb-2015-0097.

A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso

Mark van der Laan*

University of California, Berkeley, USA

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 of the data distribution based on a realistic statistical model. The only smoothness condition we will enforce on the statistical model is that the nuisance parameters of the data distribution that are needed to evaluate the canonical gradient of the pathwise derivative of the target parameter are multivariate real valued cadlag functions (right-continuous and left-hand limits, (G. Neuhaus. On weak convergence of stochastic processes with multidimensional time parameter. Ann Stat 1971;42:1285–1295.) and have a finite supremum and (sectional) variation norm. Each nuisance parameter is defined as a minimizer of the expectation of a loss function over over all functions it its parameter space. For each nuisance parameter, we propose a new minimum loss based estimator that minimizes the loss-specific empirical risk over the functions in its parameter space under the additional constraint that the variation norm of the function is bounded by a set constant. The constant is selected with cross-validation. We show such an MLE can be represented as the minimizer of the empirical risk over linear combinations of indicator basis functions under the constraint that the sum of the absolute value of the coefficients is bounded by the constant: i.e., the variation norm corresponds with this L_1-norm of the vector of coefficients. We will refer to this estimator as the highly adaptive Lasso (HAL)-estimator. We prove that for all models the HAL-estimator converges to the true nuisance parameter value at a rate that is faster than $n^{-1/4}$ w.r.t. square-root of the loss-based dissimilarity. We also show that if this HAL-estimator is included in the library of an ensemble super-learner, then the super-learner will at minimal achieve the rate of convergence of the HAL, but, by previous results, it will actually be asymptotically equivalent with the oracle (i.e., in some sense best) estimator in the library. Subsequently, we establish that a one-step TMLE using such a super-learner as initial estimator for each of the nuisance parameters is asymptotically efficient at any data generating distribution in the model, under weak structural conditions on the target parameter mapping and model and a strong positivity assumption (e.g., the canonical gradient is uniformly bounded). We demonstrate our general theorem by constructing such a one-step TMLE of the average causal effect in a nonparametric model, and establishing that it is asymptotically efficient.

*Corresponding author: Mark van der Laan, laan@berkeley.edu.

Keywords

asymptotic linear estimator; canonical gradient; cross-validated targeted minimum loss estimation (CV-TMLE); Donsker class; efficient influence curve; efficient estimator; empirical process; entropy; highly adaptive Lasso; influence curve; one-step TMLE; super-learning; targeted minimum loss estimation (TMLE)

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 asymptotically efficient substitution estimator of the target parameter. An estimator is asymptotically efficient if and only if it is asymptotically linear with influence curve equal to the canonical gradient (also called the efficient influence curve) of the pathwise derivative of the target parameter [1]. For realistic statistical models construction of efficient estimators requires using highly data adaptive estimators of the relevant parts of the data distribution the efficient influence curve depends upon. We will refer to these relevant parts of the data distribution as nuisance parameters.

One can construct an asymptotically efficient estimator with the following two general methods. 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 [1]. In the special case that the efficient influence curve can be represented as an estimating function, one can represent this methodology as the first step of the Newton-Raphson algorithm for solving the estimating equation defined by setting the empirical mean of the efficient influence curve equal to zero. Such general estimating equation methodology for construction of efficient estimators has been developed for censored and causal inference models in the literature (e.g., [2, 3]). 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 one-step TMLE of the target parameter is now the resulting plug-in estimator [4–6]. In this article we focus on the one-step TMLE since it is a more robust estimator by respecting the global constraints of the statistical model, which becomes evident when comparing the one-step estimator and TMLE 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., [7–9]). Nonetheless, the results in this article have immediate analogues for the one-step estimator and estimating equation method.

The asymptotic linearity and efficiency of the TMLE and one-step estimator relies on a second order remainder to be $o_P(n^{-1/2})$, which typically requires that the nuisance parameters are estimated at a rate faster than $n^{-1/4}$ w.r.t. an $L^2(P_0)$-norm (e.g., see our example in Section 7). To make the TMLE highly data adaptive and thereby efficient for large statistical models we have recommended to estimate the nuisance parameters with a super-learner based on a large library of candidate estimators [10–13]. Due to the oracle

inequality for the cross-validation selector, the super-learner will be asymptotically equivalent with the oracle selected estimator w.r.t. loss-based dissimilarity, even when the number of candidate estimators in the library grows polynomial in sample size. The loss-based dissimilarity (e.g., Kullback-Leibler divergence or loss-based dissimilarity for the squared error loss) behaves as a square of an $L^2(P_0)$-norm (see, for example Lemma 4 in our example). Therefore, in order to control the second order remainder, our goal should be to construct a candidate estimator in the library of the super-learner which will converge at a faster rate than $n^{-1/4}$ w.r.t. square-root of the loss-based dissimilarity.

In this article, for each nuisance parameter, we propose a new minimum loss based estimator that minim-izes the loss-specific empirical risk over its parameter space under the additional constraint that the variation norm is bounded by a set constant. The constant is selected with cross-validation. We show that these MLEs can be represented as the minimizer of the empirical risk over linear combinations of indicator basis functions under the constraint that the sum of the absolute value of the coefficients is bounded by the constant: i.e., the variation norm corresponds with this L_1-norm of the vector of coefficients. We will refer to this estimator as the highly adaptive Lasso (HAL)-estimator. We prove that the HAL-estimator converges at a rate that is for all models faster than $n^{-1/4}$ w.r.t. square-root of the loss-based dissimilarity. This even holds if the model only assumes that the true nuisance parameters have a finite variation norm. As a corollary of the general oracle inequality for cross-validation, we will then show that the super-learner including this HAL-estimator it its library is guaranteed to converge to its true counterparts at the same rate as this HAL-estimator (and thus faster than $n^{-1/4}$). By also including a large variety of other estimators in the library of the super-learner, the super-learner will also have excellent practical performance for finite samples relative to competing estimators [14]. Based on this fundamental result for the HAL-estimator and the super-learner, we proceed in this article with proving a general theorem for asymptotic efficiency of the one-step TMLE for arbitrary statistical models. In this article we will use a one-step cross-validated-TMLE (CV-TMLE), which avoids the Donsker-class entropy condition on the nuisance parameter space, in order to further minimize the conditions for asymptotic efficiency [5, 15]. In our accompanying technical report [16] we present the analogue results for the one-step TMLE. Beyond establishing these fundamental theoretical general results, we will also discuss the practical implementation of the HAL-estimator and corresponding TMLE.

2 Example: Treatment specific mean in nonparametric model

Before we start the main part of this article, in this section we will first introduce an example, and use this example to provide the reader with a guide through the different sections.

2.1 Defining the statistical estimation problem

Let $O = (W, A, Y) \sim P_0$ be a d-dimensional random variable consisting of a $(d-2)$-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 $\bar{Q}(P)(W) = E_P(Y|A = 1, W)$ and $\bar{G}(P)(W) = E_P(A|W)$. Let $Q_2(P)$ be the marginal cumulative

probability distribution of W, and $Q = (Q_1 = \bar{Q}, Q_2)$. 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 $P \in \mathcal{M}$, $W \mapsto \bar{Q}(P)(W)$ and $W \mapsto \bar{G}(P)(W)$ are cadlag functions in W on a set $[0, \tau_P] \subset \mathbb{R}^{d-2}$ [17], and that the variation norm of these functions $\bar{Q}(P)$ and $\bar{G}(P)$ are bounded. The definition of variation norm will be presented in the next section. Suppose that $\mathcal{G}$ assumes that $\bar{G}$ only depends on W through a subset of covariates of dimension $d_2 \leq d - 2$: if $d_2 = d - 2$, then this does not represent an assumption.

Our target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by $\Psi(P) = \int \bar{Q}(w) dQ_2(w) \equiv \Psi_1(Q_1 = \bar{Q}, Q_2)$. For notational convenience, we will use Ψ for both mappings Ψ and Ψ_1. It is well known that Ψ is pathwise differentiable so that for each 1-dimensional parametric submodel $\{P_\varepsilon : \varepsilon\} \subset \mathcal{M}$ through P with score S at $\varepsilon = 0$, we have

$$\frac{d}{d\varepsilon}\Psi(P_\varepsilon)\Big|_{\varepsilon = 0} = PD(P)S = \int_o D(P)(o)S(o)dP(o),$$

for some $D(P) \in L^2(P)$, where $L^2(P)$ is the Hilbert space of functions of O with mean zero endowed with inner product $\langle f, g \rangle_P = Pfg$. Here we use the notation $Pf \equiv \int f(o)dP(o)$. Such an object $D(P)$ is called a gradient at P of the pathwise derivative. The unique gradient that is also an element of the tangent space $T(P)$ is defined as the canonical gradient. The tangent space $T(P)$ at P is defined as the closure of the linear span of the set of scores of the class of 1-dimensional parametric submodels we consider. In this example the canonical gradient $D^*(P) = D^*(Q(P), G(P))$ at P is given by:

$$D^*(Q, G)(O) = \frac{A}{\bar{G}(W)}(Y - \bar{Q}(W)) + \bar{Q}(W) - \Psi(Q).$$

Let $D_1^*(Q, G) = A/\bar{G}(W)(Y - \bar{Q}(W))$ and $D_2^*(Q) = \bar{Q}(W) - \Psi(Q)$ and note that $D^*(Q, G) = D_1^*(Q, G) + D_2^*(Q)$.

An estimator ψ_n of $\psi_0 = \Psi(P_0)$ is asymptotically efficient (among the class of all regular estimators) if and only if it is asymptotically linear with influence curve equal to the canonical gradient $D^*(P_0)$ [1]:

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

where P_n is the empirical probability distribution of $O_1, \ldots, O_n$. Therefore, the canonical gradient is also called the efficient influence curve.

We have that

$$\Psi(P) - \Psi(P_0) = (P - P_0)D^*(Q, G) + R_{20}((\bar{Q}, \bar{G}), (\bar{Q}_0, \bar{G}_0)), \quad (1)$$

where $Q = Q(P)$, $G = G(P)$, and the second order remainder $R_{20}()$ is defined as follows:

$$R_{20}((\bar{Q}, \bar{G}), (\bar{Q}_0, \bar{G}_0)) \equiv \int \frac{\bar{G}(w) - \bar{G}_0(w)}{\bar{G}(w)} (\bar{Q}(w) - \bar{Q}_0(w)) dP_0(w).$$

Of course, $PD^*(Q, G) = 0$.

We define the following two log-likelihood loss functions for $\bar{Q}$, Q_2 and $\bar{G}$, respectively:

$$L_{11}(\bar{Q})(O) = -A\{Y \log \bar{Q}(W) + (1 - Y)\log(1 - \bar{Q}(W))\}; L_{12}(Q_2)(O) = -\log dQ_2(W); L_2(\bar{G})(O) = -\{A \log \bar{G}(W) + (1 - A)\log(1 - \bar{G}(W))\}.$$

We also define the corresponding Kullback-Leibler dissimilarities $d_{10,1}(\bar{Q}, \bar{Q}_0) = P_0\{L_{11}(\bar{Q}) - L_{11}(\bar{Q}_0)\}$, $d_{10,2}(Q_2, Q_{20}) = P_0\{L_{12}(Q_2) - L_{12}(Q_{20})\}$, and $d_{20}(\bar{G}, \bar{G}_0) = P_0\{L_2(\bar{G}) - L_2(\bar{G}_0)\}$. Here Q_2 represents an easy to estimate parameter which we will estimate with the empirical probability distribution $Q_{2n} = \hat{Q}_2(P_n)$ of $W_1, \ldots, W_n$.

Let the submodel $\mathcal{M}(\delta) \subset \mathcal{M}$ be defined by the extra restriction that $\delta < \bar{Q}(W) < 1 - \delta$ and $\bar{G}(W) > \delta$ P_0-a.e. If we would replace the log-likelihood loss $L_{11}(\bar{Q})$ (which becomes unbounded if $\bar{Q}$ approximates 0 or 1) by a squared error loss $(Y - \bar{Q}(W))^2 A$, then one can remove the restriction $\delta < \bar{Q}(W) < 1 - \delta$ in the definition of $\mathcal{M}(\delta)$. Given a sequence $\delta_n \to 0$ as $n \to \infty$, we can define a sequence of models $\mathcal{M}_n = \mathcal{M}(\delta_n)$ which grows from below to $\mathcal{M}$ as $n \to \infty$. By assumption, there exists an $N_0 = N(P_0) < \infty$ so that for $n > N_0$ we have $P_0 \in \mathcal{M}_n$.

Let $\mathcal{Q}_n = \mathcal{Q}_{1n} \times \mathcal{Q}_{2n}$ and $\mathcal{G}_n$ be the corresponding parameter spaces for $Q = (\bar{Q}, Q_2)$ and $\bar{G}$, respectively, and specifically, $\mathcal{Q}_{1n} = \{\bar{Q} : \delta_n < \bar{Q} < 1 - \delta_n\}$, while $\mathcal{Q}_{2n} = \mathcal{Q}_2$.

2.2 One step CV-TMLE

Let $\hat{\bar{Q}} : \mathcal{M}_{nonp} \to \mathcal{Q}_{1n}$ and $\hat{\bar{G}} : \mathcal{M}_{nonp} \to \mathcal{G}_n$ be initial estimators of $\bar{Q}_0$, $\bar{G}_0$, respectively, where $\mathcal{M}_{nonp}$ denotes a nonparametric model so that the estimator is defined for all realizations of the empirical probability distribution. Let $\hat{Q} : \mathcal{M}_{nonp} \to \mathcal{Q}_n$ be the estimator $\hat{Q}(P_n) = (\hat{\bar{Q}}(P_n), \hat{Q}_2(P_n))$ of $Q_0 = (\bar{Q}_0, Q_{20})$. For a given cross-validation scheme $B_n \in \{0, 1\}^n$, let P^1_{n,B_n}, P^0_{n,B_n} be the empirical probability distributions of the validation sample $\{O_i : B_n(i) = 1\}$ and training sample $\{O_i : B_n(i) = 0\}$, respectively. It is assumed that the proportion of observations in the validation sample (i.e., $\Sigma_i B_n(i)/n$) is between δ and $1-\delta$ for some $0 < \delta <$

1. Let $Q_{n,B_n} = (\bar{Q}_{n,B_n}, Q_{2n,B_n}) = \hat{Q}(P^0_{n,B_n})$ and $\bar{G}_{n,B_n} = \hat{\bar{G}}(P^0_{n,B_n})$ be the estimators applied to the training sample P^0_{n,B_n}. Given a $(\bar{Q}, \bar{G})$, consider the uniform least favorable submodel (van der Laan and Gruber, 2015)

$$\text{Logit}\bar{Q}_{\varepsilon_1} = \text{Logit}\bar{Q} + \varepsilon_1 H_{\bar{G}}$$

through $\bar{Q}$ at $\varepsilon_1 = 0$, where $H_{\bar{G}}(W) = 1/\bar{G}(W)$. We indeed have $\frac{d}{d\varepsilon_1} L_{11}(\bar{Q}_{\varepsilon_1}) = D^*_1(\bar{Q}_{\varepsilon_1}, \bar{G})$ for all ε_1. Given a $Q = (\bar{Q}, Q_2)$, consider also the local least favorable submodel

$$dQ^{lfm}_{2,\varepsilon_2}(W) = dQ_2(W)\left(1 + \varepsilon_2 D^*_2(Q)(W)\right)$$

through Q_2 at $\varepsilon_2 = 0$. Indeed, $\frac{d}{d\varepsilon_2} L_{12}(Q^{lfm}_{2,\varepsilon_2})\Big|_{\varepsilon_2 = 0} = D^*_2(\bar{Q}, Q_2)$. This local least favorable submodel implies the following uniform least favorable submodel (van der Laan and Gruber, 2015): for $\varepsilon_2 \geq 0$

$$dQ_{2,\varepsilon_2} = dQ_2 \exp\left(\int_0^{\varepsilon_2} D^*_2(\bar{Q}, Q_{2,x}) dx\right).$$

This universal least favorable submodel implies a recursive construction of $Q_{2,\varepsilon}$ for all ε-values, by starting at $\varepsilon = 0$ and moving upwards. For negative values of ε_2, we define $\int_0^{\varepsilon_2} = \int_{\varepsilon_2}^0$. For all ε_2, $\frac{d}{d\varepsilon_2} L_{12}(Q_{2,\varepsilon_2}) = D^*_2(\bar{Q}, Q_{2,\varepsilon_2})$, which shows that this is indeed a universal least favorable submodel for Q_2.

Let $\varepsilon_{1n} = \arg\min_{\varepsilon_1} E_{B_n} P^1_{n,B_n} L_{11}(\bar{Q}_{n,B_n,\varepsilon_1})$, and $\bar{Q}^*_{n,B_n} = \bar{Q}_{n,B_n,\varepsilon_{1n}}$. The score equation for ε_{1n} shows that $E_{B_n} P^1_{n,B_n} D^*_1(\bar{Q}^*_{n,B_n}, \bar{G}_{n,B_n}) = 0$. Let $\varepsilon_{2n} = \arg\min_{\varepsilon_2} E_{B_n} P^1_{n,B_n} L_{12}(Q_{2n,B_n,\varepsilon_2})$ and $Q^*_{2n,B_n} = Q_{2n,B_n\varepsilon_{2_n}}$. The score equation for ε_{2n} shows that $E_{B_n} P^1_{n,B_n} D^*_2(\bar{Q}^*_{n,B_n}, Q^*_{2n,B_n}) = 0$, which implies

$$E_{B_n} P^1_{n,B_n} \bar{Q}^*_{n,B_n} = E_{B_n} Q^*_{2n,B_n} \bar{Q}^*_{n,B_n}. \quad (2)$$

The CV-TMLE of $\Psi(Q_0)$ is defined as $\psi^*_n \equiv E_{B_n} \Psi(Q^*_{n,B_n})$, where $Q^*_{n,B_n} = (\bar{Q}^*_{n,B_n}, Q^*_{2n,B_n})$. By eq. (2) this implies that the CV-TMLE can also be represented as:

$$\psi_n^* = E_{B_n} P^1_{n,B_n} \bar{Q}^*_{n,B_n}. \quad (3)$$

Note that this latter representation proves that we never have to carry out the TMLE-update step for Q_{2n}, but that the CV-TMLE is a simple empirical mean of $\bar{Q}^*_{n,B_n}$ over the validation sample, averaged across the different splits B_n. We also conclude that this one-step CV-TMLE solves the crucial cross-validated efficient influence curve equation

$$E_{B_n} P^1_{n,B_n} D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) = 0. \quad (4)$$

2.3 Guide for article based on this example

Section 3: Formulation of general estimation problem—The goal of this article is far beyond establishing asymptotic efficiency of the CV-TMLE eq. (3) in this example. Therefore, we start in Section 3 by defining a general model and general target parameter, essentially generalizing the above notation for this example. Therefore, having read the above example, the presentation in Section 3 of a very general estimation problem will be easier to follow. Our subsequent definition and results for the HAL-estimator, the HAL-super-learner, and the CV-TMLE in the subsequent Sections 4-6 apply now to our general model and target parameter, thereby establishing asymptotic efficiency of the CV-TMLE for an enormous large class of semi-parametric statistical estimation problems, including our example as a special case.

Let's now return to our example to point out the specific tasks that are solved in each section of this article. By eqs (1) and (4), we have the following starting identity for the CV-TMLE:

$$E_{B_n}\Psi(Q^*_{n,B_n}) - \Psi(Q_0) = E_{B_n}(P^1_{n,B_n} - P_0)D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) \quad (5)$$
$$+ E_{B_n} R_{20}((\bar{Q}^*_{n,B_n}, \bar{G}_{n,B_n}), (\bar{Q}_0, \bar{G}_0)).$$

By the Cauchy-Schwarz inequality and bounding $1/\bar{G}_{n,B_n}$ by $1/\delta_n$, we can bound the second order remainder as follows:

$$|E_{B_n} R_{20}((\bar{Q}^*_{n,B_n}, \bar{G}_{n,B_n}), (\bar{Q}_0, \bar{G}_0))| \leq \frac{1}{\delta_n} E_{B_n} \|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0} \|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0}, \quad (6)$$

where $\|f\|_{P_0} \equiv (P_0 f^2)^{1/2}$. Suppose we can construct estimators $\hat{\bar{Q}}$ and $\hat{\bar{G}}$ of $\bar{Q}_0$ and $\bar{G}_0$ so that $\|\bar{Q}_n - \bar{Q}_0\|_{P_0} = O_P(n^{-1/4-\alpha_1})$ and $\|\bar{G}_n - \bar{G}_0\|_{P_0} = O_P(n^{-1/4-\alpha_2})$ for some $\alpha_1 > 0$, $\alpha_2 > 0$.

Since the training sample is proportional to sample size n, this immediately implies $\|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0} = O_P(n^{-1/4-\alpha_2})$ and $\|\bar{Q}_{n,B_n} - \bar{Q}_0\|_{P_0} = O_P(n^{-1/4-\alpha_1})$. In addition, it is easy to show (as we will formally establish in general) that the rate of convergence of the initial estimator $\bar{Q}_{n,B_n}$ carries over to its targeted version so that $\|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0} = O_P(n^{-1/4-\alpha_1})$.

Thus, with such initial estimators, we obtain

$$E_{B_n} R_{20}((\bar{Q}^*_{n,B_n}, \bar{G}_{n,B_n}), (\bar{Q}_0, \bar{G}_0)) = o_P(\delta_n^{-1} n^{-1/2-\alpha_1-\alpha_2}). \quad (7)$$

Thus, by selecting δ_n so that $\delta_n^{-1} n^{-\alpha_1-\alpha_2} \to 0$, we obtain $E_{B_n} R_{20}((\bar{Q}^*_{n,B_n}, \bar{G}_{n,B_n}), (\bar{Q}_0, \bar{G}_0)) = o_P(n^{-1/2})$.

Section 4: Construction and analysis of an *M*-specific HAL-estimator that converges at a rate faster than $n^{-1/4}$—This challenge of constructing such estimators $\hat{\bar{Q}}$ and $\hat{\bar{G}}$ is addressed in Section 4. In the context of our example, in Section 4 we define a minimum loss estimator (MLE) $\bar{Q}_{n,M} = \arg\min_{\|\bar{Q}\|_v < M} P_n L_{11}(\bar{Q})$ that minimizes the empirical risk over all cadlag functions with variation norm smaller than M. In Section 4 we then show that, if M is chosen larger than the variation norm of $\bar{Q}_0$, $d^{1/2}_{10,1}(\bar{Q}_{n,M}, \bar{Q}_0)$ converges to zero at a faster rate than $n^{-1/4-\alpha_1}$ for some $\alpha_1 = \alpha_1(d) > 0$ (for each dimension d). We provide an explicit representation eq. (17) of a cadlag function with finite variation norm M as an infinite linear combination of indicator functions for which the sum of the absolute value of the coefficients is bounded by M. As a consequence, it is shown in Appendix D that this M-specific minimum loss-based estimator can be approximated by (or can be exactly defined as) a Lasso-generalized linear regression problem in which the sum of the absolute values of the coefficients is bounded by M. Therefore, we will refer to $\bar{Q}_{n,M}$ as the M-specific HAL-estimator. Our proof of Lemma 1 in Section 4, which establishes the rate of convergence of the M-specific HAL-estimator, relies on an empirical process result by [18] that expresses the upper bound for this rate of convergence in terms of the entropy of the model space $\mathcal{Q}_1$ of $\bar{Q}$. The representation eq. (17) demonstrates that the set of cadlag functions that have variation norm smaller than a constant M is a difference of a"convex" hull of indicator functions, and, as a consequence of a general convex hull result in [19] this proves that it is a Donsker class with a specified upper bound on its entropy. In this way, we obtain an explicit entropy bound for our model space $\mathcal{Q}_1$. Given this explicit upper bound for the entropy, the result in [18] establishes a rate of convergence of the M-specific HAL-estimator faster than $n^{-1/4-\alpha_1}$ for a specified $\alpha_1 > 0$. By selecting M larger than the unknown variation norm of the true nuisance parameter value, we obtain an HAL-estimator that converges at a faster rate than $n^{-1/4}$.

Section 5: Construction and analysis of an HAL-super-learner—Instead of assuming that the the variation norm of $\bar{Q}_0$ is bounded by a known M and use the corresponding M-specific HAL-estimator, in Section 5 we define a 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 infinity. We then use cross-validation to data adaptively select M. We now show that the resulting cross-validated selected estimator of $\bar{Q}_0$ will be asymptotically equivalent with the oracle (i.e., best w.r.t. loss-based dissimilarity) 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 (e.g., [11, 13]). By using such a data adaptively selected M one obtains an estimator with better practical performance and it avoids having to know an upper bound M. As a consequence, our statistical model does not need to assume a universal bound M on the variation norm of the nuisance parameters, but it only needs to assume that each nuisance parameter value has a finite variation norm. For the sake of finite sample performance, we want to use a super-learner that uses cross-validation to select an estimator from a library of candidate estimators that includes these M-specific estimators as candidates, beyond other candidate estimators. In this way, the choice of estimator will be adapted to what works well for the actual data set. Therefore, in Section 5, we actually define such a general super-learner $\hat{\bar{Q}}$ and Theorem 2 states that it will converge at least as fast as the best choice in the library, and thus certainly as fast as the M-specific HAL-estimator using M equal to the true variation norm of $\bar{Q}_0$. We refer to a super-learner whose library includes this collection of M-specific HAL-estimators as an HAL-super-learner. We will use an analogue HAL-super-learner of $\bar{G}_0$ (Theorem 6).

The convergence results for this super-learner in terms of the Kullback-Leibler loss-based dissimilarities also imply corresponding results for $L^2(P_0)$-convergence as needed to control the second order remainder eq. (6): see Lemma 4.

Section 6: Construction and analysis of HAL-CV-TMLE—To control the remainder we need to understand the behavior of the updated initial estimator $\bar{Q}^*_{n,B_n}$ instead of the initial estimator $\bar{Q}_{n,B_n}$ itself. In our example, since the updated estimator only involves a single updating step of the initial estimator, using a cross-validated MLE selector of ε, we can easily show that $\bar{Q}^*_{n,B_n}$ converges at same rate to $\bar{Q}_0$ as the initial estimator $\bar{Q}_{n,B_n}$. In general, in Section 6 we define a one-step CV-TMLE for our general model and target parameter so that the targeted versions of the initial estimator of $\bar{Q}_0$ converges at the same rate as the initial HAL-super-learner estimator $\bar{Q}_n$. (Since the initial estimator is an HAL-super-learner, we refer to this type of CV-TMLE as an HAL-CV-TMLE.) This concerns a choice of least favorable submodel for which the CV-TMLE-step separately updates each of the components of the initial estimator $\hat{Q}$. We then show that with this choice of least favorable submodel the CV-TMLE-step preserves the convergence rate of the initial estimator (Lemma 3). We also establish in Appendix D that the one-step CV-TMLE already

solves the desired cross-validated efficient influence curve equation (4) up till an $o_P(n^{-1/2})$-term, so that an iterative CV-TMLE can be avoided (Lemma 13 and Lemma 14). At that point, we have shown that the generalized analogue of eq. (7) indeed holds with a specified $a_1 > 0$, $a_2 > 0$. In the final subsection of Section 6, Theorem 1 then establish the asymptotic efficiency of the HAL-CV-TMLE, which now also involves analyzing the cross-validated empirical process term, specifically, showing that

$$E_{B_n}(P^1_{n,B_n} - P_0)D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) = (P_n - P_0)D^*(Q_0, \bar{G}_0) + o_P(n^{-1/2}). \quad (8)$$

This will hold under weak conditions, given that we have estimators Q^*_{n,B_n}, G_{n,B_n} that converge at specified rates to their true counterparts and that, for each split B_n, conditional on the training sample, the empirical process is indexed by a finite dimensional (i.e., dimension of :) class of functions.

Section 7: Returning to our example—In Section 7 we return to our example to present a formal Theorem 2 with specified conditions, involving an application of our general efficiency Theorem 1 in Section 6.

Appendix: Various technical results are presented in the Appendix.

3 Statistical formulation of the estimation problem

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 $\Psi: \mathcal{M} \to \mathbb{R}$ 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 rich class of one-dimensional submodels $\{P_\varepsilon : \varepsilon \in (-\delta, \delta)\} \subset \mathcal{M}$ through P at $\varepsilon = 0$ and score $S = \frac{d}{d\varepsilon}\log dP_\varepsilon/dP\big|_{\varepsilon=0}$, we have

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

Our goal in this article is to construct a substitution estimator (i.e., a TMLE $\Psi(P^*_n)$ for a targeted estimator P^*_n of P_0) 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 $Q: \mathcal{M} \to \mathcal{Q}$. Suppose that $Q(P) = (Q_j(P) : j = 1, \ldots, k_1 + 1)$ has $k_1 + 1$ components, and $Q_j: \mathcal{M} \to \mathcal{Q}_j$ 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 $j = 1, \ldots, k_1 + 1$, $Q_j(P_0) = \arg\min_{Q_j \in \mathcal{Q}_j} P_0 L_{1j}(Q_j)$ for specified loss functions $(O, Q_j) \mapsto L_{1j}(Q_j)(O)$. 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_{k1+1} represents an easy to estimate parameter for which we have an empirical estimator $\hat{Q}_{k_1+1}$ available with negligible bias. In our treatment specific mean example above $Q = (Q_1 = \bar{Q}, Q_2)$, where the easy to estimate parameter Q_2 was the probability distribution of W which is naturally estimated with the empirical probability distribution. The parameter $\bar{Q}(P_0)$ will be estimated with our proposed loss-based HAL-super-learner. In the special case that each of the components of Q require a super-learner type-estimator, we define Q_{k_1+1} as empty (or equivalently, a known value), and in that case $Q = \bar{Q}$. We define corresponding loss-based dissimilarities $d_{10j}(Q_j, Q_{j0}) = P_0 L_{1j}(Q_j) - P_0 L_{1j}(Q_{j0})$, $j = 1, \ldots, k_1 + 1$. We assume that $d_{10(k_1+1)}(\hat{Q}_{k_1+1}(P_n), Q_{(k_1+1)0}) = O_P(r_{Q,k_1+1}(n))$ for a known rate of convergence $r_{Q,k_1+1}(n)$. Let

$$d_{10}(Q, Q_0) = (d_{10j}(Q_j, Q_{j0}) : j = 1, \ldots, k_1 + 1) \quad (9)$$

be the collection of these $k_1 + 1$ loss-based 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 vector of k_1 loss-based dissimilarities for $\bar{Q}$.

Suppose that $D^*(P)$ only depends on P through $Q(P)$ and an additional nuisance parameter $G(P)$. In the special case that $D^*(P)$ only depends on P through $Q(P)$, we define G as empty (or equivalently, as a known value). 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 $\mathcal{G}_j : \mathcal{M} \to \mathcal{G}_j$. Let $G_{j0} = \arg\min_{G \in \mathcal{G}_j} P_0 L_{2j}(G_j)$ for a loss function $(O, G_j) \mapsto L_{2j}(G_j)(O)$, and let $d_{2j0}(G_j, G_{j0}) = P_0 L_{2j}(G_j) - P_0 L_{2j}(G_{j0})$ 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. The parameter $\bar{G}(P_0)$ will be estimated with our proposed HAL-super-learner. We assume that $d_{20(k_2+1)}(\hat{G}_{k_2+1}(P_n), G_{(k_2+1)0}) = O_P(r_{G,k_2+1}(n))$ 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

$\bar{G} = (G_1, \ldots, G_{k_2})$. In the special case that each G_j requires a super-learner based estimator, then we define G_{k_2+1} as empty, and $G = \bar{G}$.

We also define

$$d_0((Q, G), (Q_0, G_0)) = (d_{10j_1}(Q_{j_1}, Q_{j_1 0}), d_{20j_2}(G_{j_2}, G_{j_2 0}) : j_1, j_2) \quad (10)$$

as the vector 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

$$L_1(Q) = (L_{1j}(Q_j) : j = 1, \ldots, k_1 + 1) \quad (11)$$

as the vector of $k_1 + 1$-loss functions for $Q = (Q_1, \ldots, Q_{k_1+1})$, and similarly we define

$$L_2(G) = (L_{2j}(G_j) : j = 1, \ldots, k_2 + 1). \quad (12)$$

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 $\bar{Q} \mapsto L_1(\bar{Q})$ is a convex function in the sense that, for any $\bar{Q}_1 = (Q_{j1} : j = 1, \ldots, k_1), \ldots, \bar{Q}_m = (Q_{jm} : j = 1, \ldots k_1)$, for each $j = 1, \ldots, k_1$

$$P_0 L_{1j}\left(\sum_{k=1}^{m} \alpha_k Q_{jk}\right) \leq \sum_{k=1}^{m} \alpha_k P_0 L_{1j}(Q_{jk}) \quad (13)$$

when $\Sigma_k \alpha_k = 1$ and $\min_k \alpha_k \geq 0$. Similarly, we assume $\bar{G} \mapsto L_2(\bar{G})$ is a convex 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: for example, if $O \in \mathbb{R}$, $\mathcal{M}$ is nonparametric, and $\Psi(P) = \int p(o)^2 do$ is the integral of the square of the Lebesgue density p of P, then the canonical gradient is given by $D^*(P) = 2p^2 - 2\Psi(P)$, so that one would define $Q(P) = p$, and there is no G.

Second order remainder for target parameter

We define the second order remainder $R_2(P, P_0)$ as follows:

$$R_2(P, P_0) \equiv \Psi(P) - \Psi(P_0) + P_0 D^*(P). \quad (14)$$

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 the above 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 so 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

$$\tau = \sup_{P \in \mathcal{M}} \tau_P, \quad (15)$$

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 $\mathbb{D}[0, \tau]$ be the Banach space of d-variate real valued cadlag functions (right-continuous with left-hand limits) [17]. For a $f \in \mathbb{D}[0, \tau]$, let $\| f \|_\infty = \sup_{x \in [0, \tau]} | f(x) |$ be the supremum norm. For a $f \in \mathbb{D}[0, \tau]$, we define the variation norm of f [20] as

$$\| f \|_v = |f(0)| + \sum_{s \subset \{1, \ldots, d\}} \int_{(0_s, \tau_s)} \left| f(dx_s, 0_{-s}) \right|. \quad (16)$$

For a subset $s \subset \{1, \ldots, d\}$, $x_s = (x_j : j \in s)$, $x_{-s} = (x_j : j \notin s)$, and the Σ_s in the above definition of the variation norm is over all subsets of $\{1, \ldots, d\}$. In addition, $x_s \to f(x_s, 0_{-s})$ is the s-specific section of $x \to f(x)$ that sets the coordinates in the compliment of s equal to 0. Note that $\| f \|_v$ is the sum of variation norms of s-specific sections of f (including f itself). Therefore, one might refer to this norm as the sectional variation norm, but, for convenience, for the purpose of this article, we will just refer to it as variation norm. If $\| f \|_v < \infty$, then we can, in fact, represent f as follows [20]:

$$f(x) = f(0) + \sum_{s \subset \{1, \ldots, d\}} \int_{\{0_s, x_s\}} f(du_s, 0_{-s}), \quad (17)$$

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

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

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

Cartesian product of cadlag function spaces, 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 $f_j \in \mathbb{D}[0, \tau]$, $j = 1, \ldots, k$. 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 $\| f \|_{P0} = (\| f_j \|_{P0} : j = 1, \ldots, k)$ 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, $f \times g$, $\max(f, g)$ 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:

Assumption 1. (Smoothness Assumption)

For each $P \in \mathcal{M}$, $\bar{Q} = \bar{Q}(P) \in \mathbb{D}^{k_1}[0,\tau]$, $\bar{G} = \bar{G}(P) \in \mathbb{D}^{k_2}[0,\tau]$, $D^*(P) = D^*(Q,G) \in \mathbb{D}[0,\tau]$, $L_1(\bar{Q}) \in \mathbb{D}^{k_1}[0,\tau]$, $L_2(\bar{G}) \in \mathbb{D}^{k_2}[0,\tau]$, 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 $\mathcal{M}$. Our estimators will also allow for unbounded models by using a sieve of models for which its finite bounds slowly approximate the actual model bound as sample size converges to infinity. These bounds will be defined now:

$$\tau = \tau(\mathcal{M}) = \sup_{P \in \mathcal{M}} \tau(P), \tag{18}$$

$$M_{1Q} = M_{1Q}(\mathcal{M}) = \sup_{Q, Q_0 \in \mathcal{Q}} \|L_1(\bar{Q}) - L_1(\bar{Q}_0)\|_\infty,$$

$$M_{2Q} = M_{2Q}(\mathcal{M}) = \sup_{P, P_0 \in \mathcal{M}} \frac{\|L_1(\bar{Q}) - L_1(\bar{Q}_0)\|_{P_0}}{\{d_{10}(\bar{Q}, \bar{Q}_0)\}^{1/2}},$$

$$M_{1G} = M_{1G}(\mathcal{M}) = \sup_{G, G_0 \in \mathcal{G}} \|L_2(\bar{G}) - L_2(\bar{G}_0)\|_\infty,$$

$$M_{2G} = M_{2G}(\mathcal{M}) = \sup_{P, P_0 \in \mathcal{M}} \frac{\|L_2(\bar{G}) - L_2(\bar{G}_0)\|_{P_0}}{\{d_{20}(\bar{G}, \bar{G}_0)\}^{1/2}},$$

$$M_{D^*} = M_{D^*}(\mathcal{M}) = \sup_{P \in \mathcal{M}} \|D^*(P)\|_\infty.$$

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 $L_1(\bar{Q})$ (e.g., [11, 13]). 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 $L_2(\bar{G})$.

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), $M_{1Q}, M_{2Q}, M_{1G}, M_{2G}, M_{D^*}$ are finite. In words, in essence, a bounded model is a model for which the support and the supremum 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. The counterparts of the above defined universal bounds on $\mathcal{M}$ applied to $\mathcal{M}_n$ are denoted with $\tau_n, M_{1Q,n}, M_{2Q,n}, M_{1G,n}, M_{2G,n}, M_{D^*,n}$. The conditions of our general asymptotic efficiency Theorem 1 will enforce that these bounds converge slowly enough to infinity (in the case the corresponding true model bound is infinity). This model $\mathcal{M}_n$ could be defined as the largest subset of $\mathcal{M}$ for which these latter bounds apply. By Assumption 1, with this choice of definition of $\mathcal{M}_n$, for any $P_0 \in \mathcal{M}$, there exists an $N_0 = N(P_0)$, so that for $n > N_0$ $P_0 \in \mathcal{M}_n$. Either way, we assume that $\mathcal{M}_n$ is defined such that the latter is true.

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 $\mathcal{M}_n$:

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

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 $Q_{(k_1+1)0}$ and $G_{(k_2+1)0}$. 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).

4 Highly adaptive Lasso estimator of Nuisance parameters

Let $M_1 < \infty$ be given. Our M_1-specific HAL-estimator of $\bar{Q}_0$ is defined as the minimizer of the empirical risk $P_n L_1(\bar{Q})$ over $\bar{Q} \in \bar{\mathcal{Q}}_n$ for which $L_1(\bar{Q})$ has a variation norm bounded by M_1 (see eq. (21)). The rate of convergence of a minimum empirical risk estimator is driven by the rate of convergence of the covering number of the parameter space over which one minimizes (e.g., [19]). This explains why the rate of convergence of the covering number of this set of functions $L_1(\bar{Q})$ defines a minimal rate of convergence for this HAL-estimator (while M_1 will be selected with the cross-validation selector). Similarly, this applies to our HAL-estimator of $\bar{G}_0$. In the next subsection we define the relevant covering numbers and their rates a_1, a_2, and establish an upper bound on them. Subsequently, we establish in Lemma 1 the minimal rate of convergence of the HAL-estimator in terms of these rates a_1, a_2.

4.1 Upper bounding the entropy of the parameter space for the HAL-estimator

We remind the reader that a covering number $N(\varepsilon, \mathcal{F}, L^2(\Lambda))$ is defined as the minimal 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 $L^2(\Lambda)$. Let $\alpha_1 \in \mathbb{R}^{k_1}_{\geq 0}$ and $\alpha_2 \in \mathbb{R}^{k_2}_{\geq 0}$ be such that for fixed M_1, M_2

$$\begin{aligned} \sup_{\Lambda} \log^{1/2}(N(\varepsilon, L_1(\bar{\mathcal{Q}}_{n,M_1}), L^2(\Lambda)) &= O(\varepsilon^{-(1-\alpha_1)}) \\ \sup_{\Lambda} \log^{1/2}(N(\varepsilon, L_2(\bar{\mathcal{Q}}_{n,M_2}), L^2(\Lambda)) &= O(\varepsilon^{-(1-\alpha_2)}), \end{aligned} \quad (19)$$

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

$$\begin{aligned} \bar{\mathcal{Q}}_{n,M_1} &\equiv \{\bar{Q} \in \bar{\mathcal{Q}}_n : \|L_1(\bar{Q})\|_\nu < M_1\} \\ \bar{\mathcal{G}}_{n,M_2} &\equiv \{\bar{G} \in \bar{\mathcal{Q}}_n : \|L_2(\bar{G})\|_\nu < M_2\}. \end{aligned} \quad (20)$$

The minimal rates of convergence of our HAL-estimator of $\bar{Q}_0$ and $\bar{G}_0$ are defined in terms of a_1 and a_2, respectively.

By eq. (17) it follows that any cadlag function with finite variation norm can be represented as a difference of two bounded monotone increasing functions (i.e., cumulative distribution function). The class of d-variate monotone increasing/cumulative distribution functions is a convex hull of d-variate indicator functions, which is again concretely implied by the representation eq. (17) by noting that $\int_0^x df(u) = \int I(u \leq x) df(u)$ Thus, $\mathcal{F}_{\nu,M}$ consists of a difference of two convex hulls of d-variate indicator functions. By Theorem 2.6.9 in [19], which maps the covering number of a set of functions into a covering number of the convex

hull of these functions, for a fixed $M < \infty$, we have that the universal covering number of $\mathcal{F}_{\nu,M}$ is bounded as follows:

$$\sup_{\Lambda} \log^{1/2}(N(\varepsilon, \mathcal{F}_{\nu,M}, L^2(\Lambda)) = O(\varepsilon^{-(1-\alpha(d))}),$$

where $\alpha(d) = 2/(d+2)$. 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 $\bar{G} = (G_1, \ldots, G_{k_2})$. Since $L_1(\bar{\mathcal{Q}}_{n,M_1}) \subset \mathcal{F}_{\nu,M_1}$ with $d = d_1$, $L_2(\bar{\mathcal{G}}_{n,M_2}) \subset \mathcal{F}_{\nu,M_2}$ with $d = d_2$, we have that $\alpha_1 \geq \alpha(d_1)$ and $\alpha_2 \geq \alpha(d_2)$.

4.2 Minimal rate of convergence of the HAL-estimator

Lemma 1 below proves that 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 HAL-estimator 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:

$$r_{\bar{Q}}(n) = r_{Q,1:k_1}(n) = n^{-(1/2+\alpha_1/4)}$$
$$r_{\bar{G}}(n) = r_{G,1:k_2}(n) = n^{-(1/2+\alpha_2/4)}.$$

Let r_{Q,k_1+1} and r_{G,k_2+1} be the rates of the simple estimators $\hat{Q}_{k_1+1}$ and $\hat{G}_{k_2+1}$ of $Q_{(k_1+1)0}$ and $G_{(k_2+1)0}$, respectively. This defines $r_Q(n) \in \mathbb{R}^{k_1+1}$ and $r_G(n) \in \mathbb{R}^{k_2+1}$.

Lemma 1—For a given vector $M \in \mathbb{R}^{k_1}_{\geq 0}$ of constants, let $\bar{\mathcal{Q}}_{n,M} \subset \left\{\bar{Q} \in \bar{\mathcal{Q}}_n : \|L_1(\bar{Q})\|_\nu \leq M\right\} \subset \mathcal{F}_{\nu,M}$ be the set of 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 < ∞. (In this definition one can also incorporate some extra M-constraints, as long as $\bar{\mathcal{Q}}_{n,M=\infty} = \bar{\mathcal{Q}}_n$.) Let $\bar{Q}^M_{0n} \in \bar{\mathcal{Q}}_{n,M}$ be so that $P_0 L_1(\bar{Q}^M_{0n}) = \inf_{\bar{Q} \in \bar{\mathcal{Q}}_{n,M}} P_0 L_1(\bar{Q})$. Assume that for a fixed $M < \infty$,

$$M_{2Q,M} \equiv \lim_{n \to \infty} \sup_{\bar{Q} \in \bar{\mathcal{Q}}_{n,M}} \frac{\|L_1(\bar{Q}) - L_1(\bar{Q}^M_{0n})\|_{P_0}}{\{d_{10}(\bar{Q}, \bar{Q}^M_{0n})\}^{1/2}} < \infty.$$

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

$$P_n L_1(\bar{Q}_n^M) = \inf_{\bar{Q} \in \bar{\mathcal{Q}}_{n,M}} P_n L_1(\bar{Q}) + r_n, \quad (21)$$

where $r_n = o_P(n^{-1/2})$. then

$$0 \leq d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) \leq -(P_n - P_0)\{L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)\} + r_n, \quad (22)$$

and

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

Proof—We have

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

which proves eq. (22). Since $L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)$ falls in a P_0-Donsker class $\mathcal{F}_{\nu,M}$, it follows that the right-hand side is $O_P(n^{-1/2})$, 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 $\|L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)\|_{P_0}^2 = O_P(n^{-1/2})$. By empirical process theory we have that $n^{1/2}(P_n - P_0)f_n \to_p 0$ if f_n falls in a P_0-Donsker class with probability tending to 1, and $P_0 f_n^2 \to_p 0$ as $n \to \infty$. Applying this to $f_n = L_1(\bar{Q}_n^M) - L_1(\bar{Q}_{0n}^M)$ shows that $(P_n - P_0)(L_1(\bar{Q}_n^M) - L(\bar{Q}_{0n}^M)) = o_P(n^{-1/2})$, which proves $d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) = o_P(n^{-1/2})$.

We now apply Lemma 7 with $\mathcal{F}_n = \{L_1(\bar{Q}) - L_1(\bar{Q}_{0n}^M) : \bar{Q} \in \bar{\mathcal{Q}}_{n,M}\}$, $\alpha = \alpha_1$ (see eq. (19)), envelope bound $M_n = M$ and $r_0(n) = n^{-1/4}$, which proves that

$$\left| n^{1/2}(P_n - P_0) f_n \right| = O_P(n^{-\alpha_1/4}).$$

This proves $d_{01}(\bar{Q}_n^M, \bar{Q}_{0n}^M) = O_P(n^{-(1/2+\alpha_1/4)}) + r_n$ □

5 Super-learning: HAL-estimator tuning the variation norm of the fit with cross-validation

Defining the library of candidate estimators

For an $M \in \mathbb{R}^{k_1}_{>0}$, let $\hat{\bar{Q}}_M : \mathcal{M}_{nonp} \to \bar{\mathcal{Q}}_{n,M} \subset \mathcal{F}_{\nu,M}$ be the HAL-estimator eq. (21) and let $\bar{Q}_{n,M} = \hat{\bar{Q}}_M(P_n)$. By Lemma 1 we have $d_{01}(\bar{Q}_{n,M} = \hat{\bar{Q}}_M(P_n), \bar{Q}^M_{0n}) = O_P(r^2_{\bar{Q}}(n))$, assuming that the numerical approximation error r_n is of smaller order. Let $\mathcal{K}_{1,n,\nu}$ be an ordered collection $M^n_1 < M^n_2 < \ldots < M_{K_{1,n,\nu}}$ of k_1-dimensional constants, and consider the corresponding collection of $K_{1,n,\nu}$ candidate estimators $\hat{\bar{Q}}_M$ with $M \in \mathcal{K}_{1,n,\nu}$. We impose that this index set $\mathcal{K}_{1,n,\nu}$ is increasing in n such that $\limsup_{n\to\infty} M_{K_{1,n,\nu}}$ equals $\sup_{P\in\mathcal{M}} \|L_1(\bar{Q}(P))\|_\nu$, so that for any $P \in \mathcal{M}$, there exists an $N(P)$ so that for $n > N(P)$, we will have that $M_{K_{1,n,\nu}} > \|L_1(\bar{Q}(P))\|_\nu$. Note that for all $M \in \mathcal{K}_{1,n,\nu}$ with $M > \|L_1(\bar{Q}_0)\|_\nu$, we have that $d_{01}(\hat{\bar{Q}}_M(P_n), \bar{Q}_0) = O_P(r^2_{\bar{Q}}(n))$. In addition, let $\hat{\bar{Q}}_j : \mathcal{M}_{nonp} \to \mathcal{Q}_n, j \in \mathcal{K}_{1,n,a}$ 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 $\mathcal{K}_{1,n} = \mathcal{K}_{1,n,\nu} \cup \mathcal{K}_{1,n,a}$ representing a collection of $K_{1n} = K_{1,n,\nu} + K_{1,n,a}$ candidate estimators $\{\hat{\bar{Q}}_k : k \in \mathcal{K}_{1n}\}$.

Super Learner

Let $B_n \in \{0, 1\}^n$ 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. We impose throughout the article that $q < q_n \leq 1/2$ for some $q > 0$ and that this random vector B_n has a finite number V possible realizations for a fixed $V < \infty$. 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 $\hat{\bar{Q}} : \mathcal{M}_{nonp} \to \bar{\mathcal{Q}}_n$ of $\bar{Q}_0$ is defined as $E_{B_n} P^1_{n,B_n} L_1(\hat{\bar{Q}}(P^0_{n,B_n}))$.

We define the cross-validation selector as the index

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

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 $k \in \mathcal{K}_{1n}$ of candidate estimators. Our proposed super-learner is defined by

$$\bar{Q}_n = \hat{\bar{Q}}(P_n) \equiv E_{B_n} \hat{\bar{Q}}_{k_{1n}}(P^0_{n,B_n}). \quad (23)$$

The following lemma proves that the super-learner $\hat{\bar{Q}}(P_n)$ converges to $\bar{Q}_0$ at least at the rate $r_{\bar{Q}}(n)$ the HAL-estimator converges to $\bar{Q}_0$: $d_{01}(\hat{\bar{Q}}(P_n), \bar{Q}_0) = O_P(r_{\bar{Q}}(n))$. 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

Recall the definition of the model bounds $M_{1Q,n}$, $M_{2Q,n}$ eq. (18), and let $C(M_1, M_2, \delta) \equiv 2(1+\delta)^2(2M_1/3 + M_2^2/\delta)$.

For any fixed $\delta > 0$,

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

If for each 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 each 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

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

Suppose that for each finite M, the conditions of Lemma 1 hold with negligible numerical approximation error r_n, so that $d_{01}(\bar{Q}_{n,M} = \hat{\bar{Q}}_M(P_n), \bar{Q}^M_{0n}) = O_P(r^2_{\bar{Q}}(n))$. Let $\lambda_1 \in \mathbb{R}^{k_1}_{>0}$ be chosen so that $r^2_{\bar{Q}}(n) = O(n^{-\lambda_1})$. For each fixed $\delta > 0$, we have

$$d_{01}(\bar{Q}_n, \bar{Q}_{0n}) = O_P(n^{-\lambda_1}) + O_P\left(C(M_{1Q,n}, M_{2Q,n}, \delta)\frac{\log K_{1n}}{n}\right). \quad (24)$$

The proof of this lemma is a simple corollary of the finite sample oracle inequality for cross-validation [11, 13, 21, 33, 34], also presented in Lemma 5 in Section A of the Appendix. It uses the convexity of the loss function to bring the E_{B_n} inside the loss-based dissimilarity.

In the Appendix we present the analogue super-learner eq. (37) of G_0 and its corresponding Lemma 6.

6 One-step CV-HAL-TMLE

Cross-validated TMLE (CV-TMLE) robustifies the bias-reduction of the TMLE-step by selecting : based on the cross-validated risk [5, 15]. In the next subsection we define the CV-TMLE. In this subsection we propose a particular type of local least favorable submodel that separately updates the initial estimator of Q_{j0} for each $j = 1, \ldots, k_1$. Due to this choice, in subsection 2 we now easily establish that the CV-TMLE of $\bar{Q}_0$ converges at the same rate to $\bar{Q}_0$ as the initial estimator, which is important for control of the second order remainder in the asymptotic efficiency proof of the CV-TMLE. In subsection 3 we establish the asymptotic efficiency of the CV-TMLE.

6.1 The CV-HAL-TMLE

Definition of one-step CV-HAL-TMLE for general local least favorable submodel—Let $\bar{L}_1(Q) \equiv \sum_{j=1}^{k_1+1} L_{1j}(Q_j)$ be the sum loss-function. For a given (Q, G), let $\{Q_\varepsilon : \varepsilon\} \subset \mathcal{Q}_n \subset \mathcal{Q}$ be a parametric submodel through Q at $\varepsilon = 0$ such that the linear span of $\frac{d}{d\varepsilon}\bar{L}_1(Q_\varepsilon)$ at $\varepsilon = 0$ includes the canonical gradient $D^*(Q, G)$. Let $\hat{Q}: \mathcal{M}_{nonp} \to \mathcal{Q}_n$ and $\hat{G}: \mathcal{M}_{nonp} \to \mathcal{G}_n$ be our initial estimators of $Q_0 = (\bar{Q}_0, Q_{0,k_1+1})$ and $G_0 = (\bar{G}_0, G_{0,k_2+1}$. We recommend defining the initial estimators $\hat{\bar{Q}}$ and $\hat{\bar{G}}$ of $\bar{Q}_0$ and $\bar{G}_0$ to be HAL-super-learners as defined by eqs (23) and (37), so that $d_{10}(\hat{Q}(P_n), Q_{0n}) = O_P(r^2_{\bar{Q}}(n))$ and $d_{20}(\hat{G}(P_n), G_{0n}) = O_P(r^2_{\bar{G}}(n))$. 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 estimator $\hat{Q}$ applied to the training sample P^0_{n,B_n}. Similarly, let $G_{n,B_n} = \hat{G}(P^0_{n,B_n})$. Let $\{Q_{n,B_n,\varepsilon} : \varepsilon\}$ be the above submodel with $(Q, G) = (Q_{n,B_n}, G_{n,B_n})$ through Q_{n,B_n} at $\varepsilon = 0$. Let

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

be the MLE of ε minimizing the cross-validated empirical risk. This defines $Q^*_{n,B_n} = Q_{n,B_n,\varepsilon_n}$ as the B_n-specific targeted fit of Q_0. The one-step CV-TMLE of ψ_0 is defined as

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

One-step CV-HAL-TMLE solves cross-validated efficient score equation—Our efficiency Theorem 1 assumes that

$$E_{B_n} P_{n,B_n}^1 D^*(Q_{n,B_n}^*, G_{n,B_n}) = o_P(n^{-1/2}). \quad (25)$$

That is, it is assumed that the one-step CV-TMLE already solves the cross-validated efficient influence curve equation up till an asymptotically negligible approximation error. By definition of ε_n we have that it solves its score equation $E_{B_n} P_{n,B_n}^1 \frac{d}{d\varepsilon_n} \bar{L}(Q_{n,B_n,\varepsilon_n}) = 0$, which provides a basis for verifying eq. (25). As formalized by Lemma 13 in the Appendix D, for our choice of $n^{-(1/4+)}$-consistent initial estimators Q_n, G_n of Q_0, G_0, a one-step CV-TMLE will satisfy eq. (25) for one-dimensional local least favorable submodels under weak regularity conditions. We believe that such a result can be proved in great generality for arbitrary (also multivariate) local least favorable submodels. Instead, below we propose a particular class of multivariate local least favorable submodels eq. (26) for which we establish eq. (25) under regularity conditions. In (van der Laan and Gruber, 2015) it is shown that one can always construct a so called universal least favorable submodel through Q with a one dimensional ε so that $\frac{d}{d\varepsilon} \bar{L}_1(Q_\varepsilon) = D^*(Q_\varepsilon, G)$ at each ε so that $E_{B_n} P_{n,B_n}^1 D^*(Q_{n,B_n,\varepsilon_n}^*, G_{n,B_n}) = 0$(exactly), independent of the properties of the initial estimator (Q_n, G_n).

One-step CV-HAL-TMLE preserves fast rate of convergence of initial estimator —Our efficiency Theorem 1 also assumes that the updated estimator Q_{n,B_n}^* satisfies for each split $B_n d_{01}(Q_{n,B_n}^*, Q_0) = o_P(n^{-1/2})$. This is generally a very reasonable condition given that $d_{01}(Q_{n,B_n}, Q_0) = O_P(n^{-\lambda_1})$ for a specified $\lambda_1 > 1/2$. Our proposed class of local least favorable submodels eq. (26) below guarantees that the rate of convergence of the initial estimator Q_{n,B_n} is completely preserved by Q_{n,B_n}^*, so that this condition is automatically guaranteed to hold.

A class of multivariate local least favorable submodels that separately updates each nuisance parameter component—One way to guarantee that $d_{01}(Q_{n,B_n}^*, Q_0) = o_P(n^{-1/2})$ is to make sure that the updated estimator Q_{n,B_n}^* converges as fast to Q_0 as the initial estimator Q_{n,B_n}. For that purpose we propose a $k_1 + 1$-dimensional local least favorable submodel of the type

$$Q_\varepsilon = (Q_{1,\varepsilon_1}, \ldots, Q_{k_1+1,\varepsilon_{k_1+1}}) \text{such that} \frac{d}{d\varepsilon_j} L_{1j}(Q_{j,\varepsilon_j})\Big|_{\varepsilon_j = 0} = D_j^*(Q,G), \quad (26)$$

for $j = 1, \ldots, k_1 + 1$, and where $D^*(Q,G) = \sum_{j=1}^{k_1+1} D_j^*(Q,G)$. By using such a submodel we have $Q^*_{j,n,B_n} = Q_{j,n,B_n,\varepsilon_n(j)}$ and $\varepsilon_n(j) = \arg\min_\varepsilon E_{B_n} P^1_{n,B_n} L_{1j}(Q_{j,n,B_n,\varepsilon})$. Thus, in this case Q_{j,n,B_n} is updated with its own $\varepsilon_n(j)$, $j = 1, \ldots, k_1 + 1$. The advantage of such a least favorable submodel is that the one-step update of $\bar{Q}_{j,n,B_n}$ is not affected by the statistical behavior of the other estimators $\bar{Q}_{l,n,B_n}, l \neq j$. On the other hand, if one uses a local least favorable submodel with a single ε, the MLE ε_n is very much driven by the worst performing estimator $\bar{Q}_{j,n,B_n}$. Lemma 3 shows that, by using such a $k_1 + 1$-variate local least favorable submodel satisfying eq. (26), the rate of convergence of the initial estimator $\bar{Q}_{j,n}$ is fully preserved by the TMLE-update $\bar{Q}^*_{j,n,B_n}$ (see Lemma 3 below).

How to construct a local least favorable submodel of type eq. (26)—A general approach for constructing such a $k_1 + 1$-variate least favorable submodel is the following. Let $D_j^*(P)$ be the efficient influence curve at a P for the parameter $\Psi_{j,P}: \mathcal{M} \to \mathbb{R}$ defined by $\Psi_{j,P}(P_1) = \Psi(Q_{-j}(P), Q_j(P_1))$ 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 $\{Q_{j,\varepsilon(j)}: \varepsilon(j)\} \subset \mathcal{Q}_{jn}$ be a one-dimensional submodel through Q_j so that

$$\frac{d}{d\varepsilon(j)} L_{1j}(Q_{j,\varepsilon(j)})\Big|_{\varepsilon(j)=0} = D_j^*(Q,G), j = 1, \ldots, k_1 + 1.$$

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

$$\frac{d}{d\varepsilon} \bar{L}(Q\varepsilon)\Big|_{\varepsilon=0} = (D_j^*(Q,G): j = 1, \ldots, k_1 + 1)^\top,$$

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

Lemma 14 provides a sufficient set of minor conditions under which the one-step-HAL-CV-TMLE using a local least favorable submodel of the type eq. (26) will satisfy eq. (25). Therefore, the class of local least favorable submodels eq. (26) yields both crucial conditions for the HAL-CV-TMLE: it solves eq. (25) and it preserve the rate of convergence of the initial estimator.

6.2 Preservation of the rate of initial estimator for the one-step CV-HAL-TMLE using eq. (26)

Consider the submodel $\{Q_\varepsilon : \varepsilon\}$ of the type eq. (26) presented above. Given an initial estimator $\hat{Q}: \mathcal{M}_{nonp} \to \mathcal{Q}_n$, recall the definition $Q_{n,B_n,\varepsilon} = \hat{Q}_\varepsilon(P^0_{n,B_n})$ as the fluctuated version of the initial estimator applied to the training sample, and $\varepsilon_n = \arg\min_\varepsilon E_{B_n} P^1_{n,B_n} L_1(Q_{n,B_n,\varepsilon})$. We want to show that Q_{n,B_n,ε_n} converges to Q_0 at the same rate as the initial estimator Q_{n,B_n} (and thus also $\hat{Q}(P_n)$). The following lemma establishes this result and it 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 $P_n \to Q_{jn,\varepsilon(j)} = \hat{Q}_{j,\varepsilon(j)}(P_n)$ indexed by $\varepsilon(j)$, for each $j = 1, \ldots, k_1 + 1$.

Lemma 3—Let $\varepsilon_n = \arg\min_\varepsilon E_{B_n} P^1_{n,B_n} L_1(Q_{n,B_n,\varepsilon})$. We have

$$E_{B_n} d_{01}(\hat{Q}_{\varepsilon_n}(P^0_{n,B_n}), Q_{0n}) \leq (1+2\delta) \min E_{B_n} d_{01}(\hat{Q}_\varepsilon(P^0_{n,B_n}), Q_{0n}) + O_P\left(\frac{C(M_{1Q,n}, M_{2Q,n}, \delta)\log K_{1n}}{nq}\right).$$

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

$$d_{01}(E_{B_n} \hat{Q}_{\varepsilon_n}(P^0_{n,B_n}), O_{0n}) \leq (1+2\delta) \min E_{B_n} d_{01}(\hat{Q}_\varepsilon(P^0_{n,B_n}), Q_{0n}) + O_P\left(\frac{C(M_{1Q,n}, M_{2Q,n}, \delta)\log K_{1n}}{nq}\right).$$

We have

$$\min_\varepsilon E_{B_n} d_{01}(\hat{Q}_\varepsilon(P^0_{n,B_n}), Q_{0n}) \leq E_{B_n} d_{01}(\hat{Q}(P^0_{n,B_n}), Q_{0n}).$$

Thus, if for some $\lambda_1 > 0$ $C(M_{1Q,n}, M_{2Q,n}, \delta)\log K_{1n}/(nq) = O(n^{-\lambda_1})$ and for each B_n $d_{01}(\hat{Q}(P^0_{n,B_n}), Q_{0n}) = O_P(n^{-\lambda_1})$, then

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

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

6.3 Efficiency of the one-step CV-HAL-TMLE

We have the following theorem.

Theorem 1—Consider the above defined corresponding one-step CV-TMLE $\psi^*_n = E_{B_n}\Psi(Q_{n,B_n,\varepsilon_n})$ of $\Psi(Q_0)$.

Initial estimator conditions: Consider the HAL-super-learners $\hat{\bar{Q}}(P_n)$ and $\hat{\bar{G}}(P_n)$ defined by eqs (23) and (37), respectively, and, recall that we are given simple estimators $\hat{Q}_{k_1+1}$ and $\hat{G}_{k_2+1}$ of Q_{0,k_1+1} and G_{0,k_2+1}. Let λ_1 and λ_2 be chosen so that $r_{\bar{Q}}(n) = O(n^{-\lambda_1})$ and $r_{\bar{G}}(n) = O(n^{-\lambda_2})$. Assume the conditions of Theorem 2 and Theorem 6 so that we have

$$d_{01}(\hat{\bar{Q}}(P_n), \bar{Q}_0) = O_P(n^{-\lambda_1(1:k_1)}) + O_P(C(M_{1Q,n}, M_{2Q,n}, \delta)\log K_{1n}/n)$$
$$d_{02}(\hat{\bar{G}}(P_n), \bar{G}_0) = O_P(n^{-\lambda_2(1:k_2)}) + O_P(C(M_{1G,n}, M_{2G,n}, \delta)\log K_{2n}/n),$$

where $\lambda_1(1:k_1) > 1/2$ and $\lambda_2(1:k_2) > 1/2$. Let $\hat{Q} = (\hat{\bar{Q}}, \hat{Q}_{k_1+1})$ and $\hat{G} = (\hat{\bar{G}}, \hat{G}_{k_2+1})$ be the corresponding estimators of Q_0 and G_0, respectively.

"Preserve rate of convergence of initial estimator"-condition: In addition, assume that either (Case A) the CV-TMLE uses a local least favorable submodel of the type eq. (26) so that Lemma 3 applies, or (Case B) assume that for each split B_n $d_{01}(Q^*_{n,B_n}, Q_0) = O_P(n^{-\lambda_1^*})$ for some $\lambda_1^* > 1/2$.

Efficient influence curve score equation condition and second order remainder condition: Define $f_{n,\varepsilon} = D^*(\hat{Q}_\varepsilon(P^0_{n,B_n}), G_{n,B_n}) - D^*(Q_0, G_0)$ and the class of functions $\mathcal{F}_n = \{f_{n,\varepsilon}: \varepsilon\}$. Assume

$$E_{B_n} P^1_{n,B_n} D^*(Q_{n,B_n,\varepsilon_n}, G_{n,B_n}) = o_P(n^{-1/2}), \quad (27)$$

$$\|D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0)\|_{P_0} = o_P(r_{D^*,n}) \text{ for } r_{D^*,n} = o(1), \quad (28)$$

$$E_{B_n} R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)) = o_P(n^{-1/2}), \quad (29)$$

$$\frac{\max(M_{1Q,n}, M^2_{2Q,n})\log K_{1n}}{n} = O(n^{-\lambda_1}), \quad (30)$$

$$\frac{\max(M_{1G,n}, M^2_{2G,n})\log K_{2n}}{n} = O(n^{-\lambda_1}), \quad (31)$$

$$\sup_{\Lambda} N(\varepsilon M_{D^*,n}, \mathcal{F}_n, L^2(\Lambda)) < K\varepsilon^{-p} \text{ for a } K < \infty, p < \infty. \quad (32)$$

In Case A, for verification of assumption eq. (27) one could apply Lemma 14.

In Case A, for verification of the two assumptions eqs (28) and (29) 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 $d_{02}(G_{n,B_n}, G_0) = O_P(n^{-\lambda_2})$.

In Case B, for verification of the latter two assumptions eqs (28) and (29) 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, $\psi^*_n = E_{B_n}\Psi(Q_{n,B_n,\varepsilon_n})$ is asymptotically efficient:

$$\psi^*_n - \psi_0 = (P_n - P_0)D^*(Q_0, G_0) + o_P(n^{-1/2}). \quad (33)$$

Condition eq. (32) will practically always trivially hold for $p = k_1 + 1$ equal to the dimension of ε: note that this is even true for unbounded models due to the normalizing constant $M_{D^*,n}$. We already discussed the crucial condition eq. (27) in our subsection defining the CV-TMLE. Conditions eqs (30) and (31) are easily satisfied by controlling the speed at which the model bounds $M_{1Q,n}$, $M_{2Q,n}$, $M_{1G,n}$, $M_{2G,n}$ can converge to infinity, and are always true for bounded models (as long as the size of the library of the super-learner behaves as a polynomial power of sample size). For bounded models $\mathcal{M}$, condition eq. (28) will typically hold with $r_{D^*,n} = n^{-\lambda}$ and λ equal to the minimum of the components of $\lambda_1/2$ and $\lambda_2/2$: i.e., the efficient influence curve estimator will converge to its true counterpart as fast as the slowest converging nuisance parameter estimator. If the model $\mathcal{M}$ is unbounded so

that the model bounds of the sieve $\mathcal{M}_n$ will converge to infinity, then eq. (28) will hold with $r_{D^*,n} = n^{-\lambda} M_n$ for some M_n converging to infinity (e.g., $M_n = M_{D^*,n}$). So, in the latter case one has to control the rate at which the model bounds of the sieve $\mathcal{M}_n$, such as the supremum norm bound $M_{D^*,n}$ for the efficient influence curve, converge to infinity. Finally, the crucial condition eq. (29) will easily hold for bounded models $\mathcal{M}$ if this slowest rate λ is larger than 1/4, which we know to be true for the HAL-estimator and its super-learner. For unbounded models, this condition eq. (29) will put a serious brake on the speed as which the model bounds of $\mathcal{M}_n$ can converge to infinity.

Proof—By assumptions eqs (30) and (31), we have

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

Consider Case A. Lemma 3 proves that under these same assumptions eqs (30), (31), we also have, for each B_n, $d_{01}(Q_{n,B_n\varepsilon_n}, Q_{0n}) = O_P(n^{-\lambda_1})$. This proves that for each B_n, $d_0((Q^*_{n,B_n} = Q_{n,B_n,\varepsilon_n}, G_{n,B_n}), (Q_0, G_0)) = O_P(n^{-\lambda_1, n^{-\lambda_2}})$. For Case B, we replace in latter expression λ_1 by λ_1^*. 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 eq. (27) yields the following identity:

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

By assumption eq. (29) we have that $E_{B_n}R_{20}((Q^*_{n,B_n}, G_{n,B_n}), (Q_0, G_0)) = o_P(n^{-1/2})$. Thus, we have shown

$$\Psi(Q^*_n) - \Psi(Q_0) = E_{B_n}(P^1_{n,B_n} - P_0)D^*(Q^*_{n,B_n}, G_{n,B_n}) + o_P(n^{-1/2}).$$

We now note

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

Thus, it remains to prove that $E_{B_n}(P^1_{n,B_n} - P_0)\{D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0)\} = o_P(n^{-1/2})$. For this we apply Lemma 10 with $f_{n,\varepsilon} = D^*(\hat{Q}_\varepsilon(P^0_{n,B_n}), G_{n,B_n}) - D^*(Q_0, G_0)$, conditional on P^0_{n,B_n}, and $\mathcal{F}_n = \{f_{n,\varepsilon}: \varepsilon\}$. By assumption eq. (28), there exists a rate $r_{D^*,n} = o(1)$ so that $\|f_{n,\varepsilon_n}\|_{P_0} = O_P(r_{D^*,n})$, where (e.g., for Case A) this rate will be determined based upon $d_0((Q^*_{n,B_n}, G_{n,B_n})(Q_0, G_0)) = O_P(n^{-\lambda_1, n^{-\lambda_2}})$. Note also that the envelope of $\mathcal{F}_n$ satisfies $\|F_n\|_\Lambda \leq M_{D^*,n}$ for any measure Λ (see eq. (18)). Since ε is p-dimensional for some integer p, the entropy of $\mathcal{F}_n$ easily satisfies $\sup_\Lambda M(\varepsilon \| F_n \|_\Lambda, \mathcal{F}_n, L^2(\Lambda)) = O(\varepsilon^{-p})$, which is assumed to hold by condition eq. (32). Application of Lemma 10 proves now that, if $r_{D^*,n} = o(1)$, then, given P^0_{n,B_n},

$$(P^1_{n,B_n} - P_0)f_{n,\varepsilon_n} = o_P(n^{-1/2}).$$

This proves also that $E_{B_n}(P^1_{n,B_n} - P_0)f_{n,\varepsilon_n} = o_P(n^{-1/2})$. This completes the proof. □

7 Example: Treatment specific mean

We will now apply Theorem 1 to the example introduced in Section 2. We have the following sieve model bounds (van der Laan et al., 2004): $M_{1Q,n} = O(\log \delta_n^{-1})$; $M_{2Q,n} = O(1/\delta_n)$; $M_{1G,n} = O(\log \delta_n^{-1})$; $M_{2G,n} = O(1/\delta_n)$; $M_{D^*,n} = O(1/\delta_n)$.

Since the parameter space $\mathcal{Q}_{1n}$ consists of the cadlag functions with bounded variation norms, without any further restrictions beyond the global bound δ_n, we can select the entropy quantities for $\mathcal{Q}_1$ as follows: $a_1 = a(d_1) = 2/(d_1 + 2)$, where $d_1 = d-2$ 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, then we can select the entropy quantities for $\mathcal{G}_n$ as $a_2 = a(d_2) = 2/(d_2 + 2)$. 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 a_2 can be replaced by a sharper upper bound than $a(d_2)$. We already established that condition eq. (27) in Theorem 1 holds exactly. Condition eq. (32) trivially holds.

Verification of eqs (30) and (31)

Let $\bar{Q}_n \in \mathcal{Q}_{1n}$ be a super-learner of $\bar{Q}_0$ of the type presented in eq. (23). Similarly, let $\bar{G}_n \in \mathcal{G}_n$ be such a super-learner of $\bar{G}_0$ as presented in eq. (37). Suppose that $\max(M_{1Q,n}M_{2Q,n}^2)\log K_{1n}/n = O(n^{-\lambda(d_1)})$ and $\max(M_{1G,n}M_{2G,n}^2)\log K_{2n}/n = O(n^{-\lambda(d_2)})$, where $\lambda(d) = 1/2 + \alpha(d)/4$. Then, by Lemma 2 and Lemma 6, we have $d_{10,1}(\bar{Q}_n, \bar{Q}_0) = O_P(n^{-\lambda(d_1)})$ and $d_{02}(\bar{G}_n, \bar{G}_0) = O_P(n^{-\lambda(d_2)})$. 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 $\delta_n^{-1} = O(n^{1/2 - 1/2\lambda(d_1)}(\max(\log K_{1n}, \log K_{2n}))^{-1/2})$. (Improvements can be obtained by selecting a separate δ_{1n} for truncating $\bar{Q}$ and δ_{2n} for truncating $\bar{G}$.) Let $K_n = \max(K_{1n}, K_{2n})$ and impose that $\log K_n = O(n^{1/2 - \alpha(d_1)/2})$. Then, it follows that this bound for δ_n^{-1} is larger than $n^{\alpha(d_1)/6}$, so that this constraint on δ_n is dominated by our later constraint given below $\delta_n^{-1} = o(n^{\alpha(d_1)/6})$.

Above we showed that if $\delta_n^{-1} = O(n^{1/2 - 1/2\lambda(d_1)}(\max(\log K_{1n}, \log K_{2n}))^{-1/2})$, then the two super-learners $\bar{Q}_{n,B_n}$ and $\bar{G}_{n,B_n}$ of $\bar{Q}_0$ and $\bar{G}_0$ based on the training sample P^0_{n,B_n} converge at the rate $n^{-\lambda(d_1)}$ and $n^{-\lambda(d_2)}$ w.r.t the loss-based dissimilarities $d_{10,1}$ and d_{02}, respectively. By Lemma 3, under the same conditions stated above for $d_{01}(\bar{Q}_n, \bar{Q}_0) = O_P(n^{-\lambda(d_1)})$, the TMLE update $\bar{Q}^*_{n,B_n}$ converges at this same rate: for each split B_n, we have $d_{01}(\bar{Q}^*_{n,B_n}, \bar{Q}_0) = O_P(n^{-\lambda(d_1)})$.

Verification of eq. (28)

Using straightforward algebra and using the triangle inequality for a norm, we obtain

$$\|D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) - D^*(Q_0, \bar{G}_0)\|_{P_0} \leq \left\|A\frac{\bar{G}_{n,B_n} - \bar{G}_0}{\bar{G}_{n,B_n}\bar{G}_0}(Y - \bar{Q}_0)\right\|_{P_0}$$

$$+ \left\|\frac{A}{\bar{G}_{n,B_b}}(\bar{Q}^*_{n,B_n} - \bar{Q}_0)\right\|_{P_0} + \|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0} + \left|E_{B_n}\Psi(Q^*_{n,B_n}) - \Psi(Q_0)\right|.$$

Using that $\min(\bar{G}_{n,B_n}\bar{G}_0) > \delta_n$ and $|Y - \bar{Q}_0| < 1$ it follows that the first term is bounded by $\delta^{-3/2}\|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0}$. Using that $\bar{G}_{n,B_n} > \delta_n$, it follows that the second term is bounded by $\delta_n^{-1}\|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0}$. So, we have

$$\|D^*(Q^*_{n,B_n}, G_{n,B_n}) - D^*(Q_0, G_0)\|_{P_0} \leq \delta_n^{-3/2}\|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0} + 2\delta_n^{-1}\|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0} + \left|E_{B_n}\Psi(Q^*_{n,B_n}) - \Psi(Q_0)\right|.$$

We bound the last term as follows:

$$\begin{aligned} &E_{B_n}\Psi(Q^*_{n,B_n}) - \Psi(Q_0) = E_{B_n}Q^1_{2n,B_n}\bar{Q}^*_{n,B_n} - Q_{20}\bar{Q}_0 \\ &= E_{B_n}(Q^1_{2n,B_n} - Q_{20})\bar{Q}_0 + E_{B_n}Q^1_{2n,B_n}(Q^*_{n,B_n} - \bar{Q}_0) \\ &= O_P(n^{-1/2}) + E_{B_n}(Q^1_{2n,B_n} - Q_{20})(\bar{Q}^*_{n,B_n} - \bar{Q}_0) + E_{B_n}Q_{20}(\bar{Q}^*_{n,B_n} - \bar{Q}_0) \\ &= O_P(n^{-1/2}) + E_{B_n}(Q^1_{2n,B_n} - Q_{20})(\bar{Q}^*_{n,B_n} - \bar{Q}_0) + O_P(E_{B_n}d^{1/2}_{10,1}(\bar{Q}^*_{n,B_n}, \bar{Q}_0)), \end{aligned}$$

where we used at the third equality that for each split $B_n(O^1_{2n,B_n} - Q_{20})\bar{Q}_0 = O_P(n^{-1/2})$, by the standard central limit theorem.

In order to bound the second empirical process term we apply Lemma 10 to the term $n^{1/2}(Q^1_{2n,B_n} - Q_{20})(\bar{Q}^*_{n,B_n} - \bar{Q}_0)$. Lemma 4 below shows that $\|\bar{Q}_{n,B_n} - \bar{Q}_0\|_{P_0} = O_P(n^{-\lambda(d_1)/2}\delta_n^{-1/2})$. Therefore, we can apply Lemma 10 with $r_{D^*,n}$ equal to this latter rate. This yields the following bound:

$$E_{B_n}(Q^1_{2n,B_n} - Q_{20})(\bar{Q}^*_{n,B_n} - \bar{Q}_0) = O_P(n^{-\lambda(d_1)/2}\delta_n^{-1/2}(1 + \log n + \log\delta_n)).$$

Thus, we have shown

$$\|D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) - D^*(Q_0, \bar{G}_0)\|_{P_0} = O_P(n^{-\lambda(d_1)/2}\delta_n^{-1/2}(1 + \log n + \log\delta_n)) + O_P(\delta_n^{-1}\|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0}) + O_P(\delta_n^{-3/2}\|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0}).$$

We have $d_{10,1}(\bar{Q}^*_{n,B_n}, \bar{Q}_0) = O_P(n^{-\lambda(d_1)})$ and $d_{02}(\bar{G}_{n,B_n}, \bar{G}_0) = O_P(n^{-\lambda(d_2)})$. These rates first need to be translated in terms of $L^2(P_0)$-norms in order to utilize the above bound. Lemma 4 below shows that $\|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0} = O_P(n^{-\lambda(d_1)/2}\delta_n^{-1/2})$ and $\|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0} = O_P(n^{-\lambda(d_2)})$. So we obtain the following bound:

$$\|D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) - D^*(Q_0, \bar{G}_0)\|_{P_0} = O_P(n^{-\lambda(d_1)/2}\delta_n^{-1/2}(1 + \log n + \log\delta_n))$$
$$+ O_P(\delta_n^{-3/2}n^{-\lambda(d_1)/2}) + O_P(\delta_n^{-3/2}n^{-\lambda(d_2)/2}).$$

We can conservatively bound this as follows:

$$\|D^*(Q^*_{n,B_n}, \bar{G}_{n,B_n}) - D^*(Q_0, \bar{G}_0)\|_{P_0} = O_P(\delta_n^{-3/2}n^{-\lambda(d_1)/2}\log n),$$

where we used conservative bounding by not utilizing that d_2 could be significantly smaller than d_1. We conclude that we can set $r_{D^*,n} = \delta_n^{-3/2}n^{-\lambda(d_1)/2}\log n$. We need that $r_{D^*,n} = o(1)$ and thus that $\delta_n^{-3/2} = o(n^{\lambda(d_1)/2}\log n)$, or $\delta_n^{-1} = o(n^{1/6 + \alpha(d_1)/6}\log n)$ The latter condition is dominated by the condition $\delta_n^{-1} = o(n^{\alpha(d_1)/6})$ we need in the analysis below of the second order remainder.

Verification of eq. (29)

By eq. (6), we can bound the second order remainder as follows:

$$R_{20}(P^*_{n,B_n}, P_0) \leq \delta_n^{-1}\|\bar{G}_{n,B_n} - \bar{G}_0\|_{P_0}\|\bar{Q}^*_{n,B_n} - \bar{Q}_0\|_{P_0}$$
$$= O_P(\delta_n^{-3/2}n^{-\lambda(d_1)/2 - \lambda(d_2)/2}).$$

Thus, it suffices to assume that $\delta_n^{-3/2}n^{-\lambda(d_1)} = o(n^{-1/2})$, and thus $\delta_n^{-1} = o(n^{\alpha(d_1)/6})$.

We verified the conditions of Theorem 1. Application of Theorem 1 yields the following result.

Theorem 2—Consider the nonparametric statistical model $\mathcal{M}$ for P_0 of the d-dimensional $O = (W, A, Y) \sim P_0 \in \mathcal{M}$ and target parameter $\Psi: \mathcal{M} \to \mathbb{R}$ defined by $\Psi(P) = E_P E_P(Y \mid A = 1, W)$. In this nonparametric model we only assume that for each $P \in \mathcal{M}$,

$\bar{Q}(P) = E_P(Y|A=1,W)$ and $\bar{G}(P) = E_P(A|W)$ are cadlag functions on $\left[0,\tau\right] \subset \mathbb{R}^{d-2}_{\geq 0}$ for some finite τ with finite variation norm.

Consider the above defined one-step CV-TMLE $\psi^*_n = E_{B_n}\Psi(Q^*_{n,B_n})$ of $\Psi(Q_0)$ based on the HAL-super-learner $\bar{Q}_n$ and $\bar{G}_n$ of type eqs (23) and (37), where $\bar{Q}_n$ and $\bar{G}_n$ are enforced to be contained in interval $(\delta_n, 1-\delta_n)$. Let $d_1 = d-2$. Let $\alpha(d_1) = 2/(d_1+2)$, $\lambda(d_1) = 1/2 + \alpha(\mathrm{d}_1)/4$, and $K_n = \max(K_{1n}, K_{2n})$.

Assume that $\log K_n = O(n^{1/2-\alpha(d_1)/2})$, and that d_n^{-1} converges slowly enough to ∞ so that $\delta_n^{-1} = o(n^{\alpha(d_1)/6})$ 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 for large dimension d, δ_n^{-1} is only allowed to converge to infinity at a very slow rate. Note that δ_n^{-1} immediately implies a bound on the efficient influence curve and such bounds are naturally very crucial.

Above we used the following lemma.

Lemma 4—We have

$$\|\bar{Q}-\bar{Q}_0\|^2_{P_0} \leq 4\delta_n^{-1} d_{01}(\bar{Q},\bar{Q}_0)\,. \quad (34)$$

We also have

$$\|\bar{G}-\bar{G}_0\|^2_{P_0} \leq 4 d_{02}(\bar{G},\bar{G}_0)\,. \quad (35)$$

Proof—We first prove eq. (34). Let

$$KL(\bar{Q}(W),\bar{Q}_0(W)) = \bar{Q}_0(W)\log\frac{\bar{Q}_0(W)}{\bar{Q}(W)} + (1-\bar{Q}_0(W))\log\frac{1-\bar{Q}_0(W)}{1-\bar{Q}(W)}$$

be the Kullback-Leibler divergence between the Bernoulli laws with probabilities $\bar{Q}(W)$ and $\bar{Q}_0(W)$. Then,

$$d_{01}(\bar{Q},\bar{Q}_0) = E_{P_0}\bar{G}_0(W)KL(\bar{Q}(W),\bar{Q}_0(W))\,.$$

In van der Vaart (1998, page 62) it is shown that for two densities p, p_0, we have $\|p^{1/2} - p_0^{1/2}\|_{P_0}^2 \leq -\int \log(p/p_0) dP_0$. Applying this inequality to Bernoulli laws with probabilities $\bar{Q}(W)$ and $\bar{Q}_0(W)$ yields:

$$KL(\bar{Q}(\cdot)) \geq \bar{Q}_0(\bar{Q}^{1/2} - \bar{Q}_0^{1/2})^2 + (1 - \bar{Q}_0)((1 - \bar{Q})^{1/2} - (1 - \bar{Q}_0)^{1/2})^2.$$

Applying the inequality $(a - b)^2 \leq 4(a^{1/2} - b^{1/2})^2$ (for $a, b \in [0, 1]$) to the square terms on the right-hand side now yields:

$$KL(\bar{Q}(\cdot), \bar{Q}_0(\cdot)) \geq 4^{-1}(\bar{Q} - \bar{Q}_0)^2. \quad (36)$$

Now, note that $d_{01}(\bar{Q}, \bar{Q}_0) = E_{P_0} \bar{G}_0(W) KL(\bar{Q}(W), \bar{Q}_0(W))$. We can use that $\bar{G}_0 > \delta_n$, which provides us with the following bound:

$$\begin{aligned} d_{01}(\bar{Q}, \bar{Q}_0) &\geq \delta_n E_{P_0} KL(\bar{Q}(W), \bar{Q}_0(W)) \\ &\geq \delta_n 4^{-1} E_{P_0} (\bar{Q} - \bar{Q}_0)^2(W) = \delta_n 4^{-1} \|\bar{Q} - \bar{Q}_0\|_{P_0}^2. \end{aligned}$$

This completes the proof of eq. (34). We have

$$d_{02}(\bar{G}, \bar{G}_0) = E_{P_0} KL(\bar{G}(W), \bar{G}_0(W)).$$

Completely analogue to the derivation above of eq. (36) we obtain

$$KL(\bar{G}(\cdot), \bar{G}_0(\cdot)) \geq 4^{-1}(\bar{G} - \bar{G}_0)^2,$$

and thus

$$d_{02}(\bar{G}, \bar{G}_0) \geq 4^{-1} \|\bar{G} - \bar{G}_0\|_{P_0}^2.$$

This proves eq. (35). □

8 Discussion

In this article we established that a 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 L^1-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, which is typically equivalent with the $L^2(P_0)$-norm. 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 focused 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 theorem applies 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 focused 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.

By now, we also have implemented the HAL-estimator for nonparametric regression and dimensions $d \leq 10$, and established that its practical performance appears to be very good [22]. In addition, we also implemented the HAL-TMLE for the ATE (i.e., our example) for such low dimensions and the coverage of the confidence intervals has been remarkable good for normal sample sizes, suggesting that the asymptotics of the HAL-TMLE kicks in at earlier sample sizes then theory would predict. We suspect that part of the reason for the excellent practical performance is the double robust nature of the second order remainder, which suggest more finite sample bias cancelation than an actual square of a difference. The practical implementation and evaluation of the HAL-estimator and HAL-TMLE across a diversity of problems remains an area of future research.

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 [4, 23–26].

This article focused on a CV-TMLE that represents the statistical target parameter $\Psi(P)$ as a function $\Psi(Q_1(P), \ldots, Q_{k_1+1}(P))$ of variation independent nuisance parameters $(Q_1, \ldots, Q_{k_1+1})$. 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 [27, 28] relies on a sequential regression representation of the target parameter [29]. 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 [11, 30, 31] 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 $\hat{Q}_{k_1,j}$ 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., [32]). For example, in our example we could have defined the target parameter $EY_{d_1} - EY_{d_0}$, where $d_1(W) = I(\bar{G}_n(W) > \delta)\pi$ and $d_0(W) = 1 - I((1 - \bar{G}_n(W) > \delta)$, and $\bar{G}_n$ 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 data dependent 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. $M_{D^*} < \infty$.

Acknowledgments

This research is funded by NIH-grant 5R01AI074345-07. The author thanks Marco Carone, Antoine Chambaz, and Alex Luedtke for stimulating discussions, and the reviewers for their very helpful comments.

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 [11, 13], combined with exploiting the convexity of the loss function allowing us to bring the E_{B_n} inside the loss-based dissimilarity.

Lemma 5

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

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

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 Super learner of G_0

Completely analogue to the super-learner eq. (23), we can define such a super-learner of G_0, which we will do here. For an $M \in \mathbb{R}^{k_2}_{>0}$, let $\hat{\bar{G}}_M : \mathcal{M}_{nonp} \to \overline{\mathcal{G}}_{n,M} \subset \mathcal{F}_{\nu,M}$ be the MLE for which $d_{02}(\bar{G}_{n,M} = \hat{\bar{G}}_M(P_n), \bar{G}^M_{0n}) = O_P(r^2_{\bar{G}}(n))$. Let $\mathcal{K}_{2,n,\nu}$ be an ordered collection of k_2-dimensional constants, and consider the corresponding collection of candidate estimators $\hat{\bar{G}}_M$ with $M \in \mathcal{K}_{2,n,\nu}$. We assume the index set $\mathcal{K}_{2,n,\nu}$ is increasing in n and that $\text{limsup}_{n\to\infty} M_{K_{2,n,\nu}} = \max(M_{G,\nu}, M_{L_2(G),\nu})$. Note that for all $M \in \mathcal{K}_{2,n,\nu}$ with $M > \|L_2(\bar{G}_0)\|_\nu$, we have that $d_{02}(\hat{\bar{G}}_M(P_n), \bar{G}_0) = O_P(n^{-\lambda_2})$. In addition, let $\hat{\bar{G}}_j : \mathcal{M}_{nonp} \to \overline{\mathcal{G}}_n$, $j \in \mathcal{K}_{2,n,a}$, be an additional collection of $K_{2,n,a}$ estimators of G_0. This defines a collection of $K_{2n} = K_{2,n,\nu} + K_{2,n,a}$ candidate estimators $\{\hat{\bar{G}}_k : k \in \mathcal{K}_{2n}\}$ of $\bar{G}_0$.

We define the cross-validation selector as the index

$$k_{2n} = \hat{K}_2(P_n) = \arg \min_{k \in \mathcal{K}_{2n}} E_{B_n} P^1_{n,B_n} L_1(\hat{\bar{G}}_k(P^0_{n,B_n}))$$

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

$$\bar{G}_n = \hat{\bar{G}}(P_n) = E_{B_n} \hat{\bar{G}}_{k_n}(P^0_{n,B_n}). \quad (37)$$

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

Lemma 6

Recall the definition of the model bounds $M_{1G,n}$, $M_{2G,n}$ eq. (18), and let $C(M_1, M_2, \delta) \equiv 2(1+\delta)^2(2M_1/3 + M_2^2/\delta)$. For any fixed $\delta > 0$,

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

If for each 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

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

Suppose that for each fixed M the conditions of Lemma 1 hold with negligible numerical approximation error r_n, so that $d_{02}(\bar{G}_{n,M}, \bar{G}^M_{0n}) = O_P(r^2_{\bar{G}}(n))$. Let λ_2 be chosen so that $r^2_{\bar{G}}(n) = O(n^{-\lambda_2})$. For each fixed $\delta > 0$, we have

$$d_{02}(\hat{\bar{G}}(P_n), \bar{G}_{0n}) = O_P(n^{-\lambda_2}) + O_P\left(C(M_{1G,n}, M_{2G,n}, \delta)\frac{\log K_{2n}}{n}\right). \quad (38)$$

C Empirical process results

Theorem 2.1 in [18] establishes the following result for a Donsker class $\mathcal{F}_n$ with uniformly bounded envelope F_n and for which for each $f \in \mathcal{F}_n P_0 f^2 \leq \delta^2 P F_n^2$:

$$E\|G_n\|_{\mathcal{F}_n} \lesssim J(\delta, \mathcal{F}_n)\left(1 + \frac{J(\delta, \mathcal{F}_n)}{\delta^2 n^{1/2}\|F_n\|_{P_0}}\right)\|F_n\|_{P_0},$$

where $G_n(f) = n^{1/2}(P_n - P_0)f$ and

$$J(\delta, \mathcal{F}_n) \equiv \sup_{\Lambda} \int_0^{\delta} \log^{1/2}(1 + N(\varepsilon \|F_n\|_{\Lambda}, \mathcal{F}_n, L^2(\Lambda)) d\varepsilon$$

is the entropy integral from 0 to δ. This definition of the entropy integral is slightly different from a common definition in which the supremum over P is taken within the integral.

Suppose we want a bound on $\sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < \delta} |G_n(f)|$. Of course, $\|f\|_{P_0} < \delta$ is equivalent with $\|f\|_{P_0} < \delta_1 \|F_n\|_{P_0}$, where $\delta_1 = \delta / \|F_n\|_{P_0}$. Application of the above result with this choice of $\delta = \delta_1$ yields:

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < \delta} |G_n(f)| \lesssim J(\delta / \|F_n\|_{P_0}, \mathcal{F}_n) \left(1 + \frac{J(\delta / \|F_n\|_{P_0}, \mathcal{F}_n) \|F_n\|_{P_0}}{\delta^2 n^{1/2}}\right) \|F_n\|_{P_0}. \quad (39)$$

Suppose that $\sup_{\Lambda} \log^{1/2}(1 + N(\varepsilon \|F_n\|_{\Lambda}, \mathcal{F}_n, L^2(\Lambda))) = O(\varepsilon^{-(1-\alpha)})$ for some $\alpha \in (0, 1)$. Then,

$$J(\delta / \|F_n\|_{P_0}, \mathcal{F}_n) = O(\delta^{\alpha} \|F_n\|_{P_0}^{-\alpha}).$$

Thus, we have

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < \delta} |G_n(f)| \lesssim \delta^{\alpha} \|F_n\|_{P_0}^{1-\alpha} + \delta^{2\alpha - 2} n^{-1/2} \|F_n\|_{P_0}^{2-2\alpha}.$$

Note that this is a decreasing function in $\|F_n\|_{P_0}$. Given a bound M_n so that $\|F_n\|_{P_0} < M_n$, a conservative bound is obtained by replacing $\|F_n\|_{P_0}$ by M_n.

This proves the following lemma.

Lemma 7

Consider $\mathcal{F}_n$ with $\|F_n\|_{P_0} < M_n$ and $\sup_{\Lambda} \log^{1/2}(1 + N(\varepsilon \|F_n\|_{\Lambda}, \mathcal{F}_n, L^2(\Lambda))) = O(\varepsilon^{-(1-\alpha)})$ for some $\alpha \in (0, 1)$. Then,

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < r_0(n)} |G_n(f)| \lesssim \{r_0(n)/M_n\}^{\alpha} M_n + \{r_0(n)/M_n\}^{2\alpha - 2} n^{-1/2}.$$

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:

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < r_0(n)} |G_n(f)| \lesssim \{n^{-1/4}/M_n\}^{\alpha} M_n + \{M_n\}^{2-2\alpha} n^{-\alpha/2}.$$

Consider eq. (39) again, but suppose now that $\sup_\Lambda N(\varepsilon \|F_n\|_\Lambda, \mathcal{F}_n, L^2(\Lambda)) = O(\varepsilon^{-p})$ for some $p > 0$. Then,

$$J(\delta/\|F_n\|_{P_0}, \mathcal{F}_n) = p^{1/2} \int_0^{\delta/\|F_n\|_{P_0}} \log^{1/2} \varepsilon^{-1} d\varepsilon.$$

We can conservatively bound $\log^{1/2} \varepsilon^{-1}$ by $\log \varepsilon^{-1}$ for ε small enough, and then note $\int_0^x \log \varepsilon d\varepsilon = x(1 - \log x)$. Thus, we have the bound

$$J(\delta/\|F_n\|_{P_0}, \mathcal{F}_n) = O(\delta \|F_n\|_{P_0}^{-1} (1 - \log(\delta/\|F_n\|_{P_0}))).$$

By plugging this latter bound into eq. (39) we obtain

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < \delta} |G_n(f)| \lesssim \delta(1 - \log(\delta/\|F_n\|_{P_0})) + (1 - \log(\delta/\|F_n\|_{P_0}))^2 n^{-1/2}.$$

Note that the right-hand side is increasing in $\|F_n\|_{P_0}$. So if we know that $\|F_n\|_{P_0} \leq M_n$ for some M_n, we obtain the bound

$$E \sup_{f \in \mathcal{F}_n, \|f\|_{P_0} < \delta} |G_n(f)| \lesssim \delta(1 - \log(\delta/M_n)) + (1 - \log(\delta/M_n))^2 n^{-1/2}.$$

Lemma 8

Consider $\mathcal{F}_n$ with $\|F_n\|_{P_0} < M_n$ and $\sup_\Lambda N(\varepsilon \|F_n\|_\Lambda, \mathcal{F}_n, L^2(\Lambda))) = O(\varepsilon^{-p})$ for some $p > 0$. Then,

$$E \sup_{f \in \mathcal{F}_n, \|F_n\|_{P_0} < r_0(n)} \|G_n(f)\| \lesssim r_0(n) \left(1 - \log \frac{r_0(n)}{M_n}\right) + \left(1 - \log \frac{r_0(n)}{M_n}\right)^2 n^{-1/2}. \quad (40)$$

The following lemma is proved by first applying the Lemma 7 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 new initial rate $r_0(n)$.

Lemma 9

Consider the following setting:

$$f_n \in \mathscr{F}_n, \|F_n\|_{P_0} \leq M_n,$$
$$\sup_\Lambda \log^{1/2}(1 + N(\varepsilon\|F_n\|_\Lambda, \mathscr{F}_n, L^2(\Lambda))) = O(\varepsilon^{-(1-\alpha)}), \alpha \in (0,1),$$
$$d_0(Q_n, Q_0) \leq \left|(P_n - P_0)f_n\right|,$$
$$\|f_n\|_{P_0} \leq M_{2n}\{d_0(Q_n, Q_0)\}^{1/2}$$
$$1 < M_n \lesssim n^{1/(4(1-\alpha))}.$$

Then

$$d_0(Q_n, Q_0) \lesssim n^{-1/2}n^{-\alpha/4}C(M_n, M_{2n}, \alpha),$$

where

$$C(M_n, M_{2n}, \alpha) = M_{2n}^{\alpha}M_n^{1-\alpha/2-\alpha^2/2} + n^{-\alpha/4}M_{2n}^{2\alpha-1}M_n^{1-\alpha^2}.$$

Proof

We have $d_0(Q_n, Q_0) \leq |(P_n - P_0)f_n|$. We apply Lemma 7 to the right-hand side with $r_0(n) = 1$. This yields

$$E\left|(P_n - P_0)f_n\right| \lesssim n^{-1/2}M_n^{1-\alpha} + M_n^{2-2\alpha}n^{-1}.$$

This shows $d_0(Q_n, Q_0) \lesssim n^{-1/2}M_n^{(1-\alpha)} + M_n^{2-2\alpha}n^{-1}$. Using that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, this implies $d_0(Q_n, Q_0)^{1/2} \lesssim n^{-1/4}M_n^{(1-\alpha)/2} + M_n^{1-\alpha}n^{-1/2}$. By assumption, this implies

$$\|f_n\|_{P_0} \lesssim n^{-1/4}M_{2n}M_n^{(1-\alpha)/2} + n^{-1/2}M_{2n}M_n^{1-\alpha}.$$

The right-hand side is of order $n^{-1/4}M_{2n}M_n^{(1-\alpha)/2}$ if $M_n \lesssim n^{1/(4(1-\alpha))}$, which holds by assumption. Let $r_0(n) = n^{-1/4}M_{2n}M_n^{(1-\alpha)/2}$. We now apply Lemma 7 to $(P_n - P_0)f_n$ with this choice of $r_0(n)$. Note $r_0(n)$ converges to zero at slower rate (or equal than) $n^{-1/4}$. Thus, application of Lemma 7 gives the following bound:

$$E\left|(P_n - P_0)f_n\right| \lesssim n^{-1/2} r_{0(n)}{}^{\alpha} M_n^{1-\alpha} + r_{0(n)}{}^{2\alpha-2} M_n^{2-2\alpha_n - 1}$$
$$\lesssim n^{-1/2} n^{-\alpha/4} M_{2n}^{\alpha} M_n^{1-\alpha/2-\alpha^2/2} + n^{-1/2(1+\alpha)} M_{2n}^{2\alpha-2} M_n^{1-\alpha^2}.$$

We can factor out $n^{-1/2} n^{-\alpha/4}$, giving the bound

$$\lesssim n^{-1/2} n^{-\alpha/4} \{ M_{2n}^{\alpha} M_n^{1-\alpha/2-\alpha^2/2} + n^{-\alpha/4} M_{2n}^{2\alpha-2} M_n^{1-\alpha^2} \}.$$

This completes the proof of the lemma. □

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

Lemma 10

Let $f_{n,\varepsilon_n} \in \mathscr{F}_n = \{f_{n,\varepsilon} : \varepsilon\}$ where ε varies over a bounded set in $\mathbb{R}^p$ and $f_{n,\varepsilon}$ is a non-random function (i.e., not based on data $O_1, \ldots, O_n$). Let F_n be the envelope of $\mathscr{F}_n$ and let M_n be such that $\|F_n\| < M_{D^*,n}$. Assume that $\sup_{\Lambda} N(\varepsilon \|F_n\|_{\Lambda}, \mathscr{F}_n, L^2(\Lambda)) = O(\varepsilon^{-p})$. Suppose that $\|f_{n,\varepsilon_n}\|_{P_0} = o_P(r_{D^*}(n))$ for a rate $r_{D^*}(n) \to 0$. Then, $G_n(f_{n,\varepsilon_n}) = G_n(\bar{f}_{n,\varepsilon_n}) + E_n$, where

$$E_0\left|G_n(\tilde{f}_{n,\varepsilon_n})\right| = O(r_{D^*}(n)(1 - \log(r_{D^*}(n)/M_{D^*,n}))),$$

and E_n equals 0 with probability tending to 1. Thus, if $r_{D^*}(n)\log(M_{D^*,n}/r_{D^*}(n)) = o(1)$, then $G_n(f_{n,\varepsilon_n}) = o_P(1)$.

Proof

For notational convenience, let's denote f_{n,ε_n} with f_n. We have that with probability tending to 1 $\|f_n\|_{P_0} < r_{D^*}(n)$. We have $f_n = f_n I(\|f_n\|_{P_0} < r_{D^*}(n)) + f_n I(\|f_n\|_{P_0} > r_{D^*}(n))$. Denote the first term with $\tilde{f}_n$ and note that the second term equals zero with probability tending to 1. This shows that $G_n(f_n) = G_n(\tilde{f}_n) + E_n$ where E_n equals zero with probability tending to 1 while $\|\tilde{f}_n\|_{P_0} < r_{D^*}(n)$ with probability 1. Application of Lemma 8 shows that

$$E\left|G_n(\tilde{f}_n)\right| \lesssim r_{D^*}(n)\log(M_{D^*,n}/r_{D^*}(n)).$$

This completes the proof. □

D Implementing the HAL-estimator

For notational convenience, consider the case that $\mathcal{Q}_n = \mathcal{Q}$. The M-specific HAL-estimator is defined for a given $M < \infty$ vector, by minimizing $P_n L_1(\bar{Q})$ over all $\bar{Q} \in \bar{\mathcal{Q}}$ for which the variation norm of $L_1(\bar{Q})$ is bounded by this M. We need to calculate this estimator for a series of M-vectors ranging from 0 to infinity, and we will then select M with cross-validation (see next section). Suppose that, for a fixed n, there exists an $M_{n,v} \in \mathbb{R}^{k_1}$ so that for all $\bar{Q} \in \bar{\mathcal{Q}}$, $\|L_1(\bar{Q})\|_v \leq M_{n,v}\|\bar{Q}\|_v$. This is typically an assumption that is trivially satisfied. Then, calculating this collection of M-specific HAL-estimators across a set of M-vectors can also be achieved by computing an MLE of $\bar{Q} \to P_n L_1(\bar{Q})$ over all $\bar{Q} \in \bar{\mathcal{Q}}$ with $\|\bar{Q}\|_v < M$, for a series of M-vectors. Therefore we rephrase our goal as to compute a $\bar{Q}_{n,M}$ so that

$$P_n L_1(\bar{Q}_{n,M}) = \min_{\bar{Q} \in \bar{\mathcal{Q}}_M} P_n L_1(\bar{Q}) + r_n, \quad (41)$$

where in this section we redefine $\bar{\mathcal{Q}}_M = \{\bar{Q} \in \bar{\mathcal{Q}} : \|\bar{Q}\|_v < M\}$, and r_n is a controlled small number. We will now address a strategy for implementation of this MLE $\bar{Q}_{n,M}$.

D.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 $f \in \mathbb{D}[0,\tau]$ 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 $|s|$, consider a partitioning of $(0_s, \tau_s]$ in $|s|$-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

$$f_m(x) = f_m(f)(x) = f(0) + \sum_{s \subset \{1,\ldots,p\}} \sum_{j \in \mathcal{R}_{h_m}(s)} \phi^s_{h_m,j}(x)\beta^s_{h_m,j},$$

$\beta^s_{h_m,j} = f(R_{h_m}(j,s))$ 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 $\| f_m(f) - f \|_\Lambda \to 0$ for any $L^2(D)$-norm. Moreover, the variation norm of f is approximated by the sum of the absolute values of all the coefficients $\beta^s_{h_m,j}$:

$$\|f\|_v = \lim_{h_m \to 0} f(0) + \sum_{s \subset \{1,\ldots,p\}} \sum_{j \in \mathcal{R}_{h_m}(s)} \left| \beta^s_{h_m,j} \right|.$$

Let β_0 denote the intercept $f(0)$. 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 indicator basis functions $\phi^s_{h_m,j}$ plus an intercept β_0 for which the L^1-norm of its coefficient vector $(\beta_0, (\beta^s_{h_m,j} : j, s))$ approximates the variation norm of f. The support points $m_{h_m}(j,s)$ could also be selected based on the data support $\{O_1, \ldots, O_n\}$. Such a strategy is presented and implemented for the HAL-estimator of a nonparametric regression in [22]. In the latter paper we select n support points for each s-specific measure, possibly resulting in as many as $n * 2^d$-number of basis functions.

D.2 An approximation of the MLE over functions of bounded variation using L^1-penalization

For an $M \in \mathbb{R}_{>0}$, let's define

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

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 $\bar{\mathcal{Q}}_j$ for $\bar{Q}_j(P)$, $j \in \{1, \ldots, k_1\}$ is nonparametric, so that the MLE over $\bar{\mathcal{Q}}_{j,M} = \mathcal{F}_{v,M}$ of $\bar{Q}_{j0}$ would correspond with minimizing over $\mathcal{F}_{v,M}$. 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 $\bar{Q}_j$ of dimension $d_1(j)$, $j = 1, \ldots, k_1$, and unspecified functions $\bar{G}_j$ of dimension $d_2(j)$, $j = 1, \ldots, k_2$.

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

Lemma 11—Let $M \in \mathbb{R}_{>0}$ be given. Consider $f_0 \in \mathcal{F}_{v,M} \subset \mathbb{D}[0,\tau]$ so that for a loss function $(O, f) \to L(f)(O)$, we have $P_0 L(f_0) = \min_{f \in \mathcal{F}_{v,M}} P_0 L(f)$. 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 a 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 $P_0 L(f_{0,m}) = \min_{f \in \mathcal{F}^m_{v,M}} P_0 L(f)$. We have $P_0(L(f_{0,m}) - L(f_0)) \to 0$ as $h_m \to 0$.

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

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

$$0 \le P_0 L(f_{0,m}) - P_0 L(f_0) \le P_0 L(f^*_{0,m}) - P_0 L(f_0) \to 0,$$

which proves that $P_0 L(f_{0,m}) - P_0 L(f_0) \to 0$, as $h_m \to 0$.

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

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

$$0 \le P_n L(f_{n,m}) - P_n L(f_n) \le P_n L(f^*_{n,m}) - P_n L(f_n) \to 0,$$

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

D.3 An approximation of the MLE over the subspace $\overline{\mathcal{Q}}_M$ by an MLE over an L_1-constrained linear model

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 $\overline{\mathcal{Q}}_M = \overline{\mathcal{Q}} \cap \mathcal{F}_{v,M}$ with the MLE over $\overline{\mathcal{Q}}^m_M = \{\overline{Q}_m(\overline{Q}) : \overline{Q} \in \overline{\mathcal{Q}}_M\}$, 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 $\overline{\mathcal{Q}}^m_M$ is typically not a submodel of $\overline{\mathcal{Q}}_M$, but it is obtained by replacing each element $\overline{Q}$ in $\overline{\mathcal{Q}}_M$ with its approximation $\overline{Q}_m(\overline{Q})$.

Lemma 12—Assume that if $\bar{Q}_m \in \mathcal{F}_{v,M}$ converges pointwise to a $\bar{Q} \in \mathcal{F}_{v,M}$ on $[0,\tau]^{k_1}$, then $L_1(\bar{Q}_m)$ converges pointwise to $L_1(\bar{Q})$ on a support of P_0, including the support of the empirical distribution P_n. For an $M \in \mathbb{R}^{k_1}$, let $\bar{\mathcal{Q}}_M = \bar{\mathcal{Q}} \cap \mathcal{F}_{v,M}^{k+1} = \{\bar{Q}(P): P \in \mathcal{M}, \bar{Q}(P) \in \mathcal{F}_{v,M}\}$ 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 $\{\phi_{h_m,j}^s: j,s\}$ with coefficient vector $\{\beta_{h_m,j}^s(\bar{Q}): j,s\}$.

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 $P_0 L_1(\bar{Q}_{0,M}^m) = \min_{\bar{Q} \in \bar{\mathcal{Q}}_M^m} P_0 L_1(\bar{Q})$ be such that $P_0(L_1(\bar{Q}_{0,M}^m) - L_1(\bar{Q}_{0,M})) \to 0$ as $h_m \to 0$.

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 $\bar{Q}_{n,M}^m \in \bar{\mathcal{Q}}_M^m$ be such that $P_n L_1(\bar{Q}_{n,M}^m) = \min_{\bar{Q} \in \bar{\mathcal{Q}}_M^m} P_n L_1(\bar{Q})$. Then, $P_n(L_1(\bar{Q}_{n,M}^m - L_1(\bar{Q}_{n,M})) \to 0$ as $h_m \to 0$.

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

$$0 \le P_0 L_1(\bar{Q}_{0,M}^m) - P_0 L_1(\bar{Q}_{0,M}) \le P_0 L_1(\bar{Q}_{0,M}^{m,*}) - P_0 L_1(\bar{Q}_{0,M}) \to 0,$$

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

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

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

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

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

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

In this section we focus on the one-step TMLE, but the results can be straightforwardly generalized to the one-step CV-TMLE.

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 $P_n D^*(Q_{n,\varepsilon_n}, G_n) = o_P(n^{-1/2})$ under some regularity conditions.

Lemma 13

$\Psi: \mathcal{M} \to \mathbf{R}$ 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 $G \in \mathcal{G}$, $\{Q_\varepsilon: \varepsilon\} \subset \mathcal{Q}$ is a one dimensional parametric submodel through Q with $\frac{d}{d\varepsilon} L(Q_\varepsilon)\big|_{\varepsilon = 0} = D^*(Q, G)$ be an initial estimator of (Q_0, G_0), and consider the one-step TMLE $\Psi(Q_n, \varepsilon_n)$ with $\varepsilon_n = \arg\min_\varepsilon P_n L(Q_n, \varepsilon)$.

Let $f_n(\varepsilon) = P_n D^*(Q_n, \varepsilon, G_n)$ and $g_n(\varepsilon) = \frac{d}{d\varepsilon} P_n L(Q_{n,\varepsilon})$. Let $f'_n(\varepsilon) = \frac{d}{d\varepsilon} f_n(\varepsilon)$ and $g'_n(\varepsilon) = \frac{d}{d\varepsilon} g_n(\varepsilon)$. Let $\varepsilon_0 = 0$. Assume

- $f_n(\varepsilon_n) = f_n(0) + f'_n(0)\varepsilon_n + O_P(\varepsilon_n^2)$ and $g_n(\varepsilon_n) = g_n(0) + g'_n(0)\varepsilon_n + O_P(\varepsilon_n^2)$;
- $\varepsilon_n^2 = o_P(n^{-1/2})$;
- $\{\frac{d}{d\varepsilon_n} D^*(Q_{n,\varepsilon_n}, G_n) - \frac{d^2}{d\varepsilon_n^2} L(Q_{n,\varepsilon_n})\}/n^{1/4}$ falls in a P_0-Donsker class with probability tending to 1;
- $$P_0\left\{\frac{d}{d\varepsilon_0} D^*(Q_{n,\varepsilon_0}, G_n) - \frac{d}{d\varepsilon_0} D^*(Q_{0,\varepsilon_0}, , G_0)\right\} = O_P(n^{-1/4}) \quad (42)$$
$$P_0\left\{\frac{d^2}{d\varepsilon_0^2} L(Q_{n,\varepsilon_0}) - \frac{d^2}{d\varepsilon_0^2} L(Q_{0,\varepsilon_0})\right\} = O_P(n^{-1/4});$$

- $$P_0 \frac{d^2}{d\varepsilon_0^2} L(Q_{0,\varepsilon_0}) = -P_0 D^*(P_0)\{D^*(P_0)\}^\top. \quad (43)$$

  If $L(Q(P)) = -\log p_{Q(P),\eta(P)}$ for some density parameterization $(Q, \eta) \to p_{Q,\eta}$, *then* (43) *holds*;

- $\frac{d}{d\varepsilon_0} R_{20}((Q_{0,\varepsilon_0}, G_0), (Q_0, G_0)) = 0.$

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

The first bullet point condition only assumes that the chosen least favorable submodel is smooth in ε. The second bullet point condition will be satisfied if the initial estimators Q_n, G_n converge to the true Q_0, G_0 at a rate faster than $n^{-1/4}$. The third bullet condition will hold without $n^{-1/4}$-scalar if the estimators Q_n, G_n have uniformly bounded variation norm. Due to the scaling $n^{-1/4}$, it could even allow that the variation norm grows with sample size, again showing that this is a very weak condition. Conditions eq. (42) are expected to hold if Q_n, G_n converge to Q_0, G_0 at a rate $n^{-1/4}$. Condition eq. (43) is a condition that holds for loss-functions that can be represented as log-likelihood loss function, and is therefore again a natural condition for a local least favorable submodel w.r.t. loss function L. Finally, consider the last bullet point condition. If this remainder has a double robust form $R_{20}((Q, G), (Q_0, G_0)) = \int (H_1(Q) - H_1(Q_0))(H_2(G) - H_2(G_0))dP_0$ for some functionals H_1, H_2, then this condition holds. If the remainder is of the form $R_{20}((Q, G), (Q_0, G_0)) = \int (H(Q) - H(Q_0))^2 dP_0$, then again this condition trivially holds. This shows that also the latter condition is a weak regularity condition.

Proof of Lemma

Firstly, by the fact that $Q_{n,\varepsilon}$ has score $D^*(Q_n, G_n)$ at $\varepsilon = 0$, it follows that $f_n(0) = g_n(0)$. We also know that $g_n(\varepsilon_n) = 0$, and we want to show that $f_n(\varepsilon_n) = o_P(n^{-1/2})$. Let $\varepsilon_0 = 0$. By the second order Tailor expansion assumption for f_n, g_n at $\varepsilon = 0$, we have

$$\begin{aligned} f_n(\varepsilon_n) &= f_n(\varepsilon_n) - g_n(\varepsilon_n) \\ &= f_n(0) - g_n(0) + \varepsilon_n(f'_n - g'_n)(0) + O(\varepsilon_n^2) \\ &= \varepsilon_n \left\{ \frac{d}{d\varepsilon_0} P_n D^*(Q_{n,\varepsilon_0}, G_n) - \frac{d^2}{d\varepsilon_0^2} P_n L(Q_{n,\varepsilon_0}) \right\} + O(\varepsilon_n^2). \end{aligned}$$

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

$$P_n \frac{d}{d\varepsilon_0} D^*(Q_{n,\varepsilon_0}, G_n) - P_n \frac{d^2}{d\varepsilon_0^2} L(Q_{n,\varepsilon_0}) = O_P(n^{-1/4}).$$

By our Donsker class assumption, we have

$$(P_n - P_0)\left\{\frac{d}{d\varepsilon_0} D^*(Q_{n,\varepsilon_0}, G_n) - \frac{d^2}{d\varepsilon_0^2} L(Q_{n,\varepsilon_0})\right\}/n^{1/4} = O_P(n^{-1/2}).$$

Thus, it remains to show

$$\frac{d}{d\varepsilon_0} P_0 D^*(Q_{n,\varepsilon_0}, G_n) - P_0 \frac{d^2}{d\varepsilon_0^2} L(Q_{n,\varepsilon_0}) = O_P(n^{-1/4})$$

By assumptions eq. (42), we have that the left-hand side of last expression equals

$$\frac{d}{d\varepsilon_0} P_0 D^*(Q_{0,\varepsilon_0}, G_0) - P_0 \frac{d^2}{d\varepsilon_0^2} L(Q_{0,\varepsilon_0}) + O_P(n^{-1/4}),$$

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

$$\frac{d}{d\varepsilon_0} P_0 D^*(Q_{0,\varepsilon_0}, G_0) = -\frac{d}{d\varepsilon_0} \Psi(Q_{0,\varepsilon_0}) + \frac{d}{d\varepsilon_0} R_2((Q_{0,\varepsilon_0}, G_0), (Q_0, G_0)).$$

By assumption we have $\frac{d}{d\varepsilon_0} R_2((Q_{0,\varepsilon_0}, G_0), (Q_0, G_0)) = 0$. By definition of the pathwise derivative at P_0, we have that the derivative $\Psi(Q_{0,\varepsilon})$ = $\Psi(P_{0,\varepsilon})$ at $\varepsilon = 0$ equals $P_0 D^*(P_0)\{D^*(P_0)\}^\top$. Thus, we have shown

$$\frac{d}{d\varepsilon_0} P_0 D^*(Q_{0,\varepsilon_0}, G_0) = -P_0 D^*(P_0)\{D^*(P_0)\}^\top.$$

Thus, it remains to show eq. (43), which thus holds by assumption. Suppose that $L(Q(P)) = -\log p_{Q(P),\eta(P)}$ for some density parameterization $(Q, \eta) \to p_{Q,\eta}$. Then $L(Q_{0,\varepsilon}) = -\log p_{Q_{0,\varepsilon},\eta_0}$. Since $\{p_{Q_{0,\varepsilon},\eta_0} : \varepsilon\}$ is a correctly specified parametric model, we have that the second derivative of $-P_0 \log p_{Q_{0,\varepsilon},\eta_0}$ at $\varepsilon = 0$ equals its information matrix (i.e., covariance matrix of its score) $P_0 \frac{d}{d\varepsilon} \log p_{Q_{0,\varepsilon},\eta_0} \{\frac{d}{d\varepsilon} \log p_{Q_{0,\varepsilon},\eta_0}\}^\top$ at $\varepsilon = 0$. However, the latter equals $-P_0 D^*(P_0)\{D^*(P_0)\}^\top$, which proves eq. (43). This completes the proof of $f_n(\varepsilon_n) = o_P(n^{-1/2})$.

In the main article we have not proposed a 1-dimensional local least favorable submodel as in Lemma 13, 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 $\varepsilon(j)$ for updating Q_{jn} for each $j = 1, \ldots, k_1 + 1$. We will now state the desired lemma for the one-step TMLE for such a submodel by application of the above lemma across all j.

Lemma 14

Let $\Psi: \mathcal{M} \to \mathbb{R}$ 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 $\Psi_{Q,j}: \mathcal{M} \to \mathbb{R}$ 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 $\{Q_{jn,\varepsilon(j)} : \varepsilon(j)\} \subset \mathcal{Q}_{jn}$ be a submodel through Q_{jn} at $\varepsilon(j) = 0$ and satisfying $\frac{d}{d\varepsilon(j)} L_{1,j}(Q_{jn,\varepsilon(j)})\Big|_{\varepsilon(j)=0} = D^*_{Q_n,j}(Q_n, G_n)$, $j = 1, \ldots, k_1 + 1$. Let $\{Q_{n,\varepsilon} : \varepsilon\} \subset \mathcal{Q}_n$ be defined by $Q_{n,\varepsilon} = (Q_{jn,\varepsilon(j)} : j = 1, \ldots, k_1+1)$. Let $\varepsilon_n = \arg\min_\varepsilon P_n L_1(Q_{n,\varepsilon})$, where $P_n L_1(Q_{n,\varepsilon}) = (P_n L_{1j}(Q_{jn,\varepsilon(j)}) : j = 1, \ldots, k_1+1)$. Let $Q^*_n = Q_{n,\varepsilon_n}$.

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

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

For each $j = 1, \ldots, k_1 + 1$, assume the following conditions:

1. Suppose that by application of the previous lemma to $\Psi_{Q_n,j}: \mathcal{M} \to \mathbf{R}$, submodel $\{Q_{jn,\varepsilon(j)} : \varepsilon(j)\}$, loss function $L_{1j}(Q_j)$, $\varepsilon_n(j) = \arg\min_{\varepsilon(j)} P_n L_{1j}(Q_{jn,\varepsilon(j)})$, and one-step TMLE $Q_{jn,\varepsilon n(j)}$, we establish its conclusion $P_n D^*_{Q_n,j}(Q_{jn,\varepsilon_n(j)}, Q_{-jn}, G_n) = o_P(n^{-1/2})$. For completeness, Lemma 15 below explicitly states these j specific conditions of the previous lemma, which are sufficient for this conclusion.

2. 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 $(P_n - P_0)f_{nj} = o_P(n^{-1/2})$. For this to hold if suffices to assume that $P_0 f^2_{nj} \to_p 0$ and $\limsup_{n\to\infty} \|f_{nj}\|_v < M$ a.e.

3. Let $f_{nj,1} = D^*_{Q_n,j}(Q^*_n, G_n) - D^*_{Q^*_n,j}(Q^*_n, G_n)$, and assume $(P_n - P_0)f_{nj} = o_P(n^{-1/2})$. For this to hold if suffices to assume that $P_0 f^2_{nj} \to_p 0$ and $\limsup_{n\to\infty} \|f_{nj,1}\|_v < M$ a.e.

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

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

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

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

Lemma 15

Let $f_{nj}(\varepsilon(j)) = P_n D^*_{Q_n,j}(Q_{jn,\varepsilon(j)}, Q_{-jn}, G_n)$ and $g_{nj}(\varepsilon(j)) = \frac{d}{d\varepsilon(j)} P_n L_{1j}(Q_{jn,\varepsilon(j)})$. Let $f'_{nj}(\varepsilon(j)) = \frac{d}{d\varepsilon(j)} f_{nj}(\varepsilon(j))$ and $g'_{nj}(\varepsilon(j)) = \frac{d}{d\varepsilon(j)} g_{nj}(\varepsilon(j))$. Let : $\varepsilon_0(j) = 0$.

Assume the following conditions:

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

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

3. $\{\frac{d}{d\varepsilon_n(j)} D^*_{Q_n,j}(Q_{jn,\varepsilon_n(j)}, Q_{-jn} G_n) - \frac{d^2}{d\varepsilon_n(j)^2} L_{1j}(Q_{jn,\varepsilon_n(j)})\}/n^{1/4}$ falls in a P_0-Donsker class with probability tending to 1;

4. $$\frac{d}{d\varepsilon_0(j)} P_0 \left\{D^*_{Q_n,j}(Q_{jn,\varepsilon_0(j)}, Q_{-jn} G_n) - D^*_{Q_n,j}(Q_{j0,\varepsilon_0(j)}, Q_{-j0} G_0)\right\} = O_P(n^{-1/4})$$
$$\frac{d^2}{d\varepsilon_0(j)^2} P_0\{L_{1j}(Q_{jn,\varepsilon_0(j)}) - L_{1j}(Q_{j0,\varepsilon_0(j)})\} = O_P(n^{-1/4});$$

5. $$P_0 \frac{d^2}{d\varepsilon_0(j)^2} L_{1j}(Q_{j0,\varepsilon_0(j)}) = P_0 D^*_{Q_0,j}(P_0)\{D^*_{Q_0,j}(P_0)\}^\top. \quad (44)$$

If $L_{1j}(Q_j(P)) = -\log p_{Q_j(P),\eta(P)}$ for some density parameterization $(Q_j, \eta) \to p_{Q_j,\eta}$, then eq. (44) holds;

6. $\frac{d}{d\varepsilon_0(j)} R_{2,Q_0,j}((Q_{j_0,\varepsilon_0(j)}, Q_{-j0}, G_0), (Q_0, G_0)) = 0.$

Then, $P_n D^*_{Q_n,j}(Q_{jn,\varepsilon_n(j)}, Q_{-jn}, G_n) = o_P(n^{-1/2})$.

Proof—This is an immediate application of Lemma 13. □

Proof of Lemma 14

Consider a 1-dimensional submodel $\{P_\varepsilon : \varepsilon\} \subset \mathcal{M}$ with score S. We have

$$\frac{d}{d\varepsilon}\Psi(P_\varepsilon) = \frac{d}{d\varepsilon}\Psi(Q_\varepsilon)$$
$$= \frac{d}{d\varepsilon}\Psi(Q_{1\varepsilon}, \ldots, Q_{k_1+1\varepsilon})$$
$$\sum_{j=1}^{k_1+1} \frac{d}{d\varepsilon}\Psi(Q_{-j}, Q_{j\varepsilon}).$$

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 15.

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

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

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

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 $P_nD^*_{Q^*_n,j}(Q^*_n,G_n)\}=o_P(n^{-1/2})$. 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_nf_n. We have $P_nf_n=(P_n-P_0)f_n+P_0f_n$. By assumption 3., we have $(P_n-P_0)f_n=o_P(n^{-1/2})$. We now have to consider

$$
\begin{aligned}
&P_0\{D^*_{Q_n,j}(Q^*_n,G_n)-D^*_{Q^*_n,j}(Q^*_n,G_n)\}\\
&=\Psi_{Q^*_n,j}(Q^*_{jn})-\Psi_{Q^*_n,j}(Q_{j0})+R_{2,Q^*_n,j}((Q^*_n,G_n),(Q_0,G_0))\\
&-\Psi_{Q_n,j}(Q^*_{jn})+\Psi_{Q_n,J}(Q_{J0})-R_{2,Q_n,j}((Q^*_n,G_n),(Q_0,G_0))).
\end{aligned}
$$

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. □

References

1Bickel PJ, , Klaassen CAJ, , Ritov Y, , Wellner J. Efficient and adaptive estimation for semiparametric models Berlin/Heidelberg/New York: Springer; 1997

2Robins JM, , Rotnitzky A. AIDS epidemiology Basel: Birkhauser; 1992 Recovery of information and adjustment for dependent censoring using surrogate markers; 297331

3van der Laan MJ, , Robins JM. Unified methods for censored longitudinal data and causality New York: Springer; 2003

4van der Laan MJ. Estimation based on case-control designs with known prevalance probability. Int J Biostat. 2008; 4(1) Article 17.

5van der Laan MJ, , Rose S. Targeted learning: Causal inference for observational and experimental data Berlin/Heidelberg/New York: Springer; 2011

6van der Laan MJ, Rubin Daniel B. Targeted maximum likelihood learning. Int J Biostat. 2006; 2(1) Article 11.

7Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. Int J Biostat. 2010; 6(1)

8Porter KE, Gruber S, van der Laan MJ, Sekhon JS. The relative performance of targeted maximum likelihood estimators. Int J Biostat. Jan 1.2011 7(1) Article 31., 2011. Published online Aug 17,

2011. Also available at: U.C. Berkeley Division of Biostatistics. http://www.bepress.com/ucbbiostat/paper279. doi: 10.2202/1557-4679.1308Working Paper Series. Working Paper 279

9Sekhon JS, , Gruber S, , Porter KE, , van der Laan MJ. Propensity scorebased estimators and c-tmle. In: van der Laan MJ, , Rose S, editorsTargeted learning: Causal inference for observational and experimental data New York/Dordrecht/Heidelberg/London: Springer; 2012

10Polley EC, , Rose S, , van der Laan MJ. Super learner. In: van der Laan MJ, , Rose S, editorsTargeted learning: Causal inference for observational and experimental data New York/Dordrecht/Heidelberg/London: Springer; 2011

11van der Laan MJ, Gruber S. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. Int J Biostat. 2016; 12(1):351–378. DOI: 10.1515/ijb-2015-0054 [PubMed: 27227728]

12van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol. 2007; 6(1) Article 25.

13van der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross-validation. Stat Decis. 2006; 24(3):351–371.

14Polley EC, , Sherri Rose, van der Laan MJ. Super learning. In: van der Laan MJ, , Rose S, editorsTargeted learning: Causal inference for observational and experimental data New York/Dordrecht/Heidelberg/London: Springer; 2012

15Zheng W, , van der Laan MJ. Cross-validated targeted minimum loss based estimation. In: van der Laan MJ, , Rose S, editorsTargeted learning: Causal inference for observational and experimental studies New York: Springer; 2011

16van der Laan MJ. Technical Report 300 UC Berkeley: 2015 A generally efficient targeted minimum lossbased estimator. http://biostats.bepress.com/ucbbiostat/paper343

17Neuhaus G. On weak convergence of stochastic processes with multidimensional time parameter. Ann Stat. 1971; 42:1285–1295.

18van der Vaart AW, Wellner JA. A local maximal inequality under uniform entropy. Electr J Stat. 2011; 5:192–203. DOI: 10.1214/11-EJS605

19van der Vaart AW, , Wellner JA. Weak convergence and empirical processes Berlin/Heidelberg/New York: Springer; 1996

20Gill RD, van der Laan MJ, Wellner JA. Inefficient estimators of the bivariate survival function for three models. Annales de l'Institut Henri Poincaré. 1995; 31:545–597.

21van der Laan MJ, Dudoit S, van der Vaart AW. The cross-validated adaptive epsilon-net estimator. Stat Decis. 2006; 24(3):373–395.

22Benkeser D, van der Laan MJ. The highly adaptive lasso estimator. Proceedings of the IEEE Conference on Data Science and Advanced Analytics. 2016 To appear.

23Chambaz A, van der Laan MJ. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, theoretical study. Int J Biostat. 2011a; 7(1):1–32. www.bepress.com/ucbbiostat. Working paper 258

24Chambaz A, van der Laan MJ. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, simulation study. Int J Biostat. 2011b; 7(1):33. www.bepress.com/ucbbiostat. Working paper 258

25van der Laan MJ. Causal inference for networks UC Berkeley: 2012 Technical Report 300http://biostats.bepress.com/ucbbiostat/paper300, to appear in Journal of Causal Inference

26van der Laan MJ, Balzer LB, Petersen ML. Adaptive matching in randomized trials and observational studies. J Stat Res. 2013; 46(2):113–156.

27Gruber S, , van der Laan MJ. Targeted maximum likelihood estimation, R package version 1.2.0-1 2012Available at http://cran.rproject.org/web/packages/tmle/tmle.pdf

28Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan MJ. Targeted maximum likelihood estimation of dynamic and static marginal structural working models. J Causal Inf. 2013; 2:147–185.

29Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–972. [PubMed: 16401269]

30Iván Díaz, van der Laan MJ. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. Int J Biostat. In press.

31van der Laan MJ, , Petersen ML. Targeted learning. In: Zhang C, , Ma Y, editorsEnsemble machine learning: methods and applications New York: Springer; 2012

32van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. Int J Biostat. 2007; 3(1) Article 3.

33van der Laan MJ, , Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples Division of Biostatistics, University of California; Berkeley: 2003Technical Report 130

34van der Laan MJ, Dudoit S, Keles S. Asymptotic optimality of likelihood-based cross-validation. Stat Appl Genet Mol. 2004; 3(1) Article 4.