Mark J. van der Laan*, David Benkeser and Weixin Cai

Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso

https://doi.org/10.1515/ijb-2019-0092

Received August 15, 2019; accepted May 9, 2022; published online July 15, 2022

Abstract: We consider estimation of a functional parameter of a realistically modeled data distribution based on observing independent and identically distributed observations. The highly adaptive lasso estimator of the functional parameter is defined as the minimizer of the empirical risk over a class of cadlag functions with finite sectional variation norm, where the functional parameter is parametrized in terms of such a class of functions. In this article we establish that this HAL estimator yields an asymptotically efficient estimator of any smooth feature of the functional parameter under a global undersmoothing condition. It is formally shown that the L_1-restriction in HAL does not obstruct it from solving the score equations along paths that do not enforce this condition. Therefore, from an asymptotic point of view, the only reason for undersmoothing is that the true target function might not be complex so that the HAL-fit leaves out key basis functions that are needed to span the desired efficient influence curve of the smooth target parameter. Nonetheless, in practice undersmoothing appears to be beneficial and a simple targeted method is proposed and practically verified to perform well. We demonstrate our general result HAL-estimator of a treatment-specific mean and of the integrated square density. We also present simulations for these two examples confirming the theory.

Keywords: asymptotically efficient estimator; canonical gradient; cross-validation; highly adaptive lasso; sectional variation norm; undersmoothing.

1 Introduction

We consider the estimation problem in which we observe n independent and identically distributed copies of a random variable with probability distribution known to be an element of an infinite-dimensional statistical model, while the goal is to estimate a particular smooth functional of the data distribution. It is assumed that the target parameter is a pathwise differentiable functional of the data distribution so that its derivative is characterized by its canonical gradient.

A regular asymptotically linear estimator is asymptotically efficient if and only if it is asymptotically linear with influence curve the canonical gradient [1] and a number of general methods for efficient estimation have been developed in the literature. If the model is not too large, then a regularized or sieve maximum likelihood estimator or minimum loss estimator (MLE) generally results in an efficient substitution estimator [2–4]. For a general theory on sieve estimation that also demonstrates sieve-based maximum likelihood estimators that are asymptotically efficient in large models, we refer to [5, 6]. These results generally require a sieve-based

***Corresponding author: Mark J. van der Laan**, Division of Biostatistics, University of California, Berkeley, USA, E-mail: laan@berkeley.edu

David Benkeser, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, USA. https://orcid.org/0000-0002-1019-8343

Weixin Cai, Division of Biostatistics, University of California, Berkeley, USA. https://orcid.org/0000-0003-2680-3066

MLE that overfits the data (or equivalently, *undersmooths* the estimated functional parameter) and are only applicable for certain type of sieves [7–10].

An alternative to undersmoothing is to use targeted estimator based on the canonical gradient, such as: the one-step estimator, which adds to an initial plug-in estimator the empirical mean of the canonical gradient at the estimated data distribution [1]; an estimating equations-based estimator, which defines the estimator of the target parameter as the solution of an estimating equation with the estimated canonical gradient as estimating function [11, 12]; and targeted minimum loss-estimation, which updates an initial estimator of the data distribution with an MLE of a least favorable parametric submodel through the initial estimator [13–16]. By using an initial estimator of the relevant parts of the data distribution that converges with respect to L^2-type norm to the truth at a rate faster than $n^{-1/4}$, such as achieved with the HAL-estimator [17, 18], these three procedures will generally result in an efficient estimator.

In this article we focus on a HAL-MLE, a particular sieve MLE described in [17, 18]. The HAL-MLE is defined as the minimizer of an empirical mean of the loss function (e.g., log-likelihood loss) over a particular class of functions. As such, such estimators could also be referred to as empirical risk minimizers. The particular class over which HAL-MLE minimizes risk are functions that can be arbitrarily well approximated by linear combinations of tensor products of univariate zero order spline-basis functions, but where the L_1-norm of the coefficient vector is constrained. The L_1-norm of the coefficients equals the sectional variation norm of the function [18, 19] so that the HAL-MLE corresponds with minimizing the empirical risk over all cadlag functions with a bound on their sectional variation norm.

The class of k_1-dimensional real-valued cadlag functions with finite sectional variation norm differs from typical smoothness classes that assume pointwise derivatives (e.g., Hölder classes) by assuming a global rather than local constraint. This finite sectional variation norm constraint allows for functions that are discontinuous, but puts a bound on the total variation of the measures generated by the section of the function that sets some of the coordinates equal to the left-origin of its support (a cube). In spite of the constraint, the class of functions with finite section variation norm is reasonably large, including for example any function whose first-order cross derivatives are uniformly bounded. In spite of its size, this class turns out to be a uniform Donsker class with a well-behaved entropy integral. In turn, this Donkser property affords appealing properties of the estimator, such as $n^{-1/3}(\log\ n)^{k_1/2}$-rate of convergence in loss-based dissimilarity (i.e., L^2-norm), as well as control over certain key empirical process conditions that are useful for proving asymptotic efficiency.

The target parameter is defined as a particular smooth real- or Euclidean-valued function of the functional parameter estimated by HAL-MLE, so that the HAL-MLE results in a plug-in estimator of the target parameter. In this case the sieve is indexed by a bound on the L_1-norm. By increasing this bound up to a large, finite value, the sieve includes the total parameter space for the true functional parameter. If the goal is to estimate the functional itself, then the constraint on the L_1-norm is optimally chosen with cross-validation.

In this article we investigate whether and how an appropriately undersmoothed HAL-MLE can be used to produce an efficient plug-in estimator of smooth functions of the functional parameter. There are essentially three key ingredients to establishing efficiency of a plug-in estimator:

(i) negligibility of the empirical mean of the canonical gradient;
(ii) control of the second-order remainder; and
(iii) asymptotic equicontinuity.

For (i), we argue that since the canonical gradient is a score, we essentially require that HAL-MLE solves a particular score equation. Because HAL-MLE is an MLE, it solves a large class of score equations, and we investigate whether these score equations might also approximate the particular score equation implied by the canonical gradient of the smooth target feature. We find that the larger the L_1-norm of the HAL-MLE, the more such score equations are solved by the HAL-MLE. We also find that the HAL-MLE solves the score equations of paths that ignore the L_1-norm constraint at rate $O_P(n^{-2/3})$, thereby better as the desired $o_P(n^{-1/2})$. Nonetheless, one might need to select a larger L_1-norm than the cross-validation selector to make sure that the basis functions selected by HAL generate enough scores to approximate the desired canonical gradient

at the desired precision for the given sample. Either way, by increasing the L_1-norm of the HAL-MLE, the linear span of equations solved by the HAL-MLE will approximate any canonical gradient score equation at the desired precision.

However, in order to satisfy (ii), we must preserve the $n^{-1/4}$-rate of convergence of achieved by the HAL-MLE, which is naturally achieved when the L_1-norm is selected with cross-validation. Fortunately, the rate of the HAL-MLE is not affected by the size of the L_1-norm as long as it remains bounded and, for n large enough, exceeds the sectional variation norm of the true function. Similarly, the asymptotic equicontinuity condition (iii) will also be satisfied for any bounded L_1-norm, since the class of cadlag functions with a finite sectional variation norm is a Donsker class. In fact, one can prove that this L_1-norm is allowed to slowly converge to infinity as sample size increases without affecting the asymptotic equicontinuity condition and the $n^{-1/4}$-rate of convergence of the HAL-MLE.

Taken together, our analysis highlights that when selecting the level of undersmoothing of a HAL-MLE, one wants to undersmooth enough to solve the efficient score equation up to an appropriate level of approximation, but in order to reasonable finite-sample performance one should not undersmooth beyond that level. This discussion highlights the need to establish empirical criterion by which the level of undersmoothing may be chosen to appropriately satisfy the conditions required of an efficient plug-in estimator. For that purpose we propose to simply select the L_1-norm till the empirical mean of the canonical gradient is solved at the desired level.

This article is organized as follows. In the next Section 2 we define the HAL-MLE. In Section 3 we establish our main theorem providing the undersmoothing conditions under which the HAL-MLE is asymptotically efficient for any pathwise differentiable parameter. In Section 4 we apply our theorem to the treatment-specific mean example providing a theorem for this particular nonparametric estimation problem. In Section 5 we apply our theorem to a nonparametric estimation problem with target parameter the integrated square of the data density. In Section 6 we demonstrate a simulation study for both examples, providing a practical verification of our theoretical results. We conclude with a discussion in Section 7. Some of the proofs are presented in the Appendices A and B.

2 Defining the functional estimation problem and HAL-MLE

2.1 Functional estimation problem

Suppose we observe $O_1, \ldots, O_n \sim_{\text{iid}} P_0 \in \mathcal{M}$, where O is a Euclidean random variable of dimension k_1 with support contained in $[0, \tau_o] \subset \mathbb{R}^{k_1}$. Let $Q: \mathcal{M} \to Q(\mathcal{M}) = \{Q(P): P \in \mathcal{M}\}$ be a functional parameter. It is assumed that there exists a loss function $L(Q)$ so that $P_0 L(Q(P_0)) = \min_{P \in \mathcal{M}} P_0 L(Q(P))$, where we use the notation $Pf \equiv \int f(o) \mathrm{d}P(o)$. Thus, $Q(P)$ can be defined as the minimizer of the risk function $Q \to PL(Q)$ over all Q in the parameter space. Let $d_0(Q, Q_0) \equiv P_0 L(Q) - P_0 L(Q_0)$ be the loss-based dissimilarity. We assume that $M_{20} \equiv \sup_{P \in \mathcal{M}} P_0 \{L(Q(P)) - L(Q_0)\}^2 / d_0(Q(P), Q_0) < \infty$, and $M_1 \equiv \sup_{0, P \in \mathcal{M}} \mid L(Q(P))(o) \mid < \infty$, thereby guaranteeing good behavior of the cross-validation selector [20–24].

Parameter space for functional parameter Q: Cadlag and uniform bound on sectional variation norm. We assume that the parameter space $Q(\mathcal{M})$ is a collection of multivariate real valued cadlag functions on a cube $[0, \tau] \subset \mathbb{R}^k$ with finite sectional variation norm $\| Q(P) \|_v^* < C^u$ for some $C^u < \infty$. That is, for all P, $Q(P)$ is a k-variate real valued cadlag function on $[0, \tau] \subset \mathbb{R}^k_{\geq 0}$ with $\| Q(P) \|_v^* < C^u$, where the sectional variation norm is defined by

$$\| Q \|_v^* \equiv Q(0) + \sum_{s \subset \{1, \ldots, k\}} \int_{[0_s, \tau_s]} \mid \mathrm{d}Q_s(u_s) \mid .$$

For a given subset $s \subset \{1, \ldots, k\}$, $Q_s: (0_s, \tau_s] \to \mathbb{R}$ is defined by $Q_s(x_s) = Q(x_s, 0_{-s})$. That is, Q_s is the s-specific section of Q which sets the coordinates in the complement of subset $s \subset \{1, \ldots, k\}$ equal to 0. Since Q_s is right-continuous with left-hand limits and has a finite variation norm over $(0_s, \tau_s]$, it generates a finite

measure, so that the integrals with respect to Q_s are indeed well defined. For a given vector $x \in [0, \tau]$, we define $x_s = (x(j): j \in s)$. Sometimes, we will also use the notation $x(s)$ for x_s.

Note also that $[0, \tau] = \{0\} \cup \left(\cup_s (0_s, \tau_s]\right)$ is partitioned in the singleton $\{0\}$, the s-specific left-edges $(0_s, \tau_s] \times \{0_{-s}\}$ of cube $[0, \tau]$, and, in particular, the full-dimensional inner set $(0, \tau]$ (corresponding with $s = \{1, \ldots, k\}$). Therefore, the above sectional variation norm equals the sum over all subsets s of the variation norm of the s-specific section over its s-specific edge. An important result is that any cadlag function Q with finite sectional variation norm can be represented as

$$Q(x) = Q(0) + \sum_{s \subset \{1,\ldots,k\}} \int_{(0_s, x_s]} \mathrm{d}Q_s(u_s).$$

That is, $Q(x)$ is a sum of integrals up to x_s over all the s-specific edges with respect to the measure generated by the corresponding s-specific section Q_s. We will refer to Q_s as a cadlag function as well as a measure. We note that this representation represents Q as an infinitesimal linear combination of indicator basis functions $x \to \phi_{s,u_s}(x) \equiv I(x_s \geq u_s)$ indexed by knot-point u_s with coefficient $\mathrm{d}Q_s(u_s)$:

$$Q(x) = Q(0) + \sum_{s \subset \{1,\ldots,k\}} \int \phi_{s,u_s}(x) \mathrm{d}Q_s(u_s).$$

Note that these basis functions are tensor products over the coordinates $j \in s$ of univariate indicator basis functions $I(x(j) \geq u(j))$, which are also known as zero-order splines. Note that the L_1-norm of the coefficients in this representation is precisely the sectional variation norm $\| Q \|_v^*$.

2.2 Definition of HAL-MLE

Recall $\mathcal{Q}(C^u) = \left\{Q \in D[0, \tau]: \| Q \|_v^* < C^u\right\}$ be the class of cadlag functions which with sectional variation norm bounded by C^u. Let $C_0 \equiv \| Q_0 \|_v^*$ be the sectional variation norm of Q_0, and let C^u be an upper bound guaranteeing that $C_0 < C^u$. For a constant $C < C^u$, consider the class $\mathcal{Q}(C) \equiv \left\{Q \in D[0, \tau]: \| Q \|_v^* < C^u\right\} \subset \mathcal{Q}(C^u)$. For a data adaptive selector C_n, we define

$$Q_n \equiv \arg \min_{Q \in \mathcal{Q}(C_n)} P_n L(Q) \tag{1}$$

be the HAL-MLE. We will restrict the minimization to Q for which, for all subsets s, $\mathrm{d}Q_s(u)$ is a discrete measure with a finite support $\{z_{s,j}: j = 1, \ldots, n_s\}$. That is, for each s, the $\mathrm{d}Q_s$ is absolutely continuous with respect to a discrete counting measure $\mu_{n,s}$. We will denote this form of absolute continuity with $Q \ll^* \mu_n$. In that case, the HAL-MLE is supported by $\mathcal{J}(\mu_n) = \{z_{s,j}: s \subset \{1, \ldots, d\}, j = 1, \ldots n_s\}$. Thus, the HAL-MLE then becomes

$$Q_n \equiv \arg \min_{Q \in \mathcal{Q}(C_n), Q \ll^* \mu_n} P_n L(Q).$$

In this case the HAL MLE can be represented as $Q_n = \sum_{j \in \mathcal{J}(\mu_n)} \beta_n(j) \phi_j$, where

$$\beta_n \equiv \arg \min_{\beta, \|\beta\|_1 \leq C_n} L\left(\sum_{j \in \mathcal{J}(\mu_n)} \beta(j) \phi_j\right),$$

and ϕ_j corresponds with one of the indicator basis functions $I(X_s \geq u_{s,j_1})$ indexed by subset $s \subset \{1, \ldots, d\}$ and knot-point u_{s,j_1} for some s and j_1. Note that $Q_n = \hat{Q}(P_n)$ is the realization of a mapping from the empirical probability measure to the parameter space.

As noted earlier, the data adaptive selector C_n might be selected larger or equal than cross-validation selector $C_{n,cv} = \arg\min_C E_{B_n} P^1_{n,B_n} L(\hat{Q}\left(P^0_{n,B_n}\right))$, where $B_n \in \{0,1\}^n$ represents a random sample split (e.g., V-fold cross-validation) into a training sample $\{i: B_n(i) = 0\}$ and validation sample $\{i: B_n(i) = 1\}$, while P^0_{n,B_n}

and P^1_{n,B_n} are the corresponding empirical probability measures. One wants that $C_n \geq C_0$ for n large enough, so that $Q_0 \in \mathcal{Q}(C_n)$.

Typically, one is able to prove that the unrestricted MLE (1) will be discrete on a support in which case our μ_n-discretization does not restrict the definition of the HAL-MLE. Generally, if O includes observing X where $L(Q)(0)$ depends on Q through $Q(X)$, we recommend to select the support of $\mathrm{d}Q_s$ as a subset (or whole set) of the observed data $X_i(s)$, $i = 1, \ldots, n$.

3 Efficiency of the HAL MLE for pathwise differentiable target parameters

3.1 Defining the efficient estimation problem and plug-in HAL-MLE

Let $\Psi: \mathcal{M} \to \mathbb{R}^d$ be the d-dimensional statistical target parameter of interest of the data distribution. We assume that Ψ is pathwise differentiable at any $P \in \mathcal{M}$ with canonical gradient $D^*(P)$. That is, for a class of paths $\mathcal{P}_h = \{P^h_\epsilon : \epsilon \in (-\delta, \delta)\}$ through P at $\epsilon = \epsilon_0 \equiv 0$ with score $h \in L^2_0(P)$, the pathwise derivative $\frac{\mathrm{d}}{\mathrm{d}\epsilon_0}\Psi\left(P^h_{\epsilon_0}\right)$ is a bounded linear operator on the tangent space $T_P \subset L^2_0(P)$ spanned by all the scores h. As a consequence, the pathwise derivative can be represented as an inner product $PD^*(P)h$ for an element $D^*(P)$ in the tangent space T_P which is called the canonical gradient. From efficiency theory we know that an estimator ψ_n is asymptotically efficient among the class of all regular estimators if and only if $\psi_n - \Psi(P_0) = P_n D^*(P_0) + o_P(n^{-1/2})$. For a pair $P, P_0 \in \mathcal{M}$, we define the exact second order remainder by

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

Relevant functional parameter and its loss function: Let $Q: \mathcal{M} \to Q(\mathcal{M}) = \{Q(P): P \in \mathcal{M}\}$ be a functional parameter such that $\Psi(P) = \Psi_1(Q(P))$ for some Ψ_1. It is assumed that Q is a functional parameter with parameter space $Q(\mathcal{M}) \subset \mathcal{Q}(C^u) = D_{C^u}[0, \tau]$ as defined above in Section 2, so that the model $\mathcal{M}$ does not make any smoothness assumptions on Q beyond that it is a cadlag function with sectional variation norm bounded by C^u. In particular, we have the HAL-MLE has a rate of convergence $d_0(Q_n, Q_0) = O_P(n^{-1/3}(\log n)^{d/2})$ [25].

Nuisance parameter for canonical gradient: Let $G: \mathcal{M} \to \mathcal{G}$ be a functional nuisance parameter so that $D^*(P)$ only depends on P through $(Q(P), G(P))$, and the remainder $R_2(P, P_0)$ only involves differences between (Q, G) and (Q_0, G_0):

$$D^*(P) = D^*(Q(P), G(P)), \text{ while } R_2(P, P_0) = R_{20}((Q, G), (Q_0, G_0)).$$

Here R_{20} could have some remaining dependence on P_0 and P, and $\mathcal{G} = G(\mathcal{M})$ is the parameter space for G.

Canonical gradient of target parameter in tangent space of loss function: We also assume that this loss function $L(Q)$ is such that there exists a class of submodels $\{Q^h_\epsilon : \epsilon\} \subset Q(\mathcal{M})$ indexed by a choice $h \in \mathcal{H}^1$, through Q at $\epsilon = 0$, so that for any $G \in \mathcal{G}$, one of these directions h generates a score that equals the canonical gradient $D^*(Q, G)$ at (Q, G):

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} L\left(Q^h_\epsilon\right)\Big|_{\epsilon=0} = D^*(Q, G).$$

Since the canonical gradient is an element of the tangent space and thereby typically a score of a submodel, this generally holds for Q defined as the density of P and the log-likelihood loss $L(Q) = -\log Q$. However, for any Q there are typically more direct loss functions $L(Q)$, so that the loss-based dissimilarity $d_0(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0)$ directly measures a dissimilarity between Q and Q_0, for which this condition holds as well.

Plug-in HAL-MLE: In this section, we are concerned with analyzing the plug-in estimator $\Psi(Q_n)$ of $\Psi(Q_0)$, where Q_n is the C_n-tuned HAL-MLE $\hat{Q}^m(P_n) = \hat{Q}_{C_n}(P_n)$, which minimizes the empirical risk over $\mathcal{Q}(C_n)$. We assume that $\mathcal{Q}$ is defined such that Q_n is in the interior of the model based parameter space $\mathcal{Q}$ (so that there are submodels through Q_n that generate the tangent space and the canonical gradient), even though Q_n

is typically on the edge of the parameter subspace $\mathcal{Q}(C_n) \subset \mathcal{Q} = Q(\mathcal{M})$ over which the estimator is minimizing the empirical risk. It is understood that verification of our conditions might require using a C_n different from the cross-validation selector.

Remark: Target parameter could be component of real target parameter. In many situations the real target parameter is a $P \to \Psi(Q_1(P), Q_2(P))$ for two (or more) functional parameters Q_1 and Q_2. One could apply our efficiency theorem below to the target parameter $\Psi_{Q_{10}}(Q_2) = \Psi(Q_{10}, Q_2)$ and $\Psi_{Q_{20}}(Q_1) = \Psi(Q_1, Q_{20})$ treating the indices Q_{10} and Q_{20} as known, and HAL-MLEs Q_{1n} and Q_{2n} of Q_{10} and Q_{20}, respectively. Application of our theorem to these two cases then proves that $\Psi(Q_{10}, Q_{2n})$ and $\Psi(Q_{1n}, Q_{20})$ are both asymptotically efficient, if both HAL-MLEs are appropriately tuned with respect to sectional variation norm bound. Since

$$\Psi(Q_{1n}, Q_{2n}) - \Psi(Q_{10}, Q_{20}) = \Psi(Q_{1n}, Q_{2n}) - \Psi(Q_{10}, Q_{2n}) + \Psi(Q_{10}, Q_{2n}) - \Psi(Q_{10}, Q_{20}),$$

this then also establishes asymptotic efficiency of $\Psi(Q_{1n}, Q_{2n})$ as estimator of $\Psi(Q_{10}, Q_{20})$, under the condition that

$$\Psi(Q_{1n}, Q_{2n}) - \Psi(Q_{10}, Q_{2n}) - (\Psi(Q_{1n}, Q_{20}) - \Psi(Q_{10}, Q_{20})) = o_P(n^{-1/2}).$$

This latter term can be viewed as a second order difference of (Q_{1n}, Q_{2n}) and (Q_{10}, Q_{20}) so that the latter condition will generally hold by using the already established rates of convergence $O_P(n^{-2/3}(\log n)^{k_1})$ with respect to risk based dissimilarity for Q_{1n} and Q_{2n}. The above immediately generalizes to the case that the target parameter is a function of more than two Q-components.

3.2 The HAL MLE solves the unconstrained score-approximation of the efficient influence curve equation by including sparse basis functions

Let $Q_n = \arg\min_{Q \in \mathcal{Q}(C_n)} P_n L(Q)$ be the HAL-MLE. Theorem 2 establishes that $\Psi(Q_n)$ is asymptotically efficient for $\Psi(Q_0)$ for large enough C_n, and some weak conditions specific towards the target parameter. The key property is $P_n D^*(Q_n, G_0) = o_P(n^{-1/2})$. This is addressed in two steps we describe now.

Main idea of result: The HAL-MLE minimizes over a class of functions so that is solves a class of score equations $P_n S_h(Q_n) = 0$ corresponding with paths $\left\{ Q_{n,\epsilon}^h : \epsilon \right\} \subset \mathcal{Q}(C_n)$ through the HAL-MLE Q_n that keep the L_1-norm constant which happens to be arranged by a simple linear real valued constraint $r(h, Q_n) = 0$. The directions h will be vectors with one component $h(j)$ for each coefficient $\beta_n(j)$ in the representation $Q_n = \sum_j \beta_n(j)\phi_j$ as a linear combination of spline-basis functions, while the paths through β_n are of form $(1 + \epsilon h(j))\beta_n(j)$ with $r(h, \beta_n) = 0$, which implies the path through Q_n. The canonical gradient $D^*(Q_n, G_0)$ can be well approximated by the class of all scores $\{S_h(Q_n) : h\}$ that ignore the L_1-norm constraint. We will refer to this best approximation with the linear span of such scores with $D_n^*(Q_n, G_0) = S_{h^*(Q_n, G_0)}(Q_n)$. Indeed, the first key condition is that $P_0\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\} = o_P(n^{-1/2})$, or, equivalently, the second order difference $P_0\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\} - P_0\{D_n^*(Q_0, G_0) - D^*(Q_0, G_0)\} = o_P(n^{-1/2})$, which behaves as a product of $d_0^{1/2}(Q_n, Q_0) = O_P(n^{-1/3}(\log n)^{k_1/2})$ times L^2-norm of difference of $\{D^*(Q_n, G_0) - D^*(Q_0, G_0)\}/d_0^{1/2}(Q_n, Q_0)$ (normalized to not converge to zero) minus its projection on the linear span of scores $\{S_h(Q_n) : h\}$. This itself corresponds with an undersmoothing condition since the more one undersmooths the better the approximation by the linear span of scores will be. This latter condition will be captured by Theorem 2 in next subsection.

It then remains to approximate the latter $D_n^*(Q_n, G_0)$ with the scores $\{S_h(Q_n) : h, r(h, Q_n) = 0\}$ of the paths that enforce the L_1-norm constraint. This requires approximating $h^*(Q_n, G_0)$ by a choice h with $r(h, Q_n) = 0$. For that purpose, we select h equal to $h^*(Q_n, G_0)$ for all its components, except one j^*. The key is then to select that choice j^* so that it minimizes the difference $P_n S_{h^*(Q_n, G_0)}(Q_n) - P_n S_h(Q_n)$ over all choices h that are equal to $h^*(Q_n, G_0)$ except at j^* (which equals $P_n S_{h^*(Q_n, G_0)}$ since $P_n S_h(Q_n) = 0$). We then want this minimizer to be $o_P(n^{-1/2})$ so that it establishes the desired $P_n S_{h^*(Q_n, G_0)} = o_P(n^{-1/2})$. This will correspond with having a basis function j^* with non-zero coefficient $\beta_n(j^*)$ for which its score equation is small enough, and correspondingly this is implied by $P_n \phi_{j^*}$ being small enough. As a result, our condition will correspond with undersmoothing enough to including a sparsely supported basis function. This result is addressed by Theorem 1 below.

Both of these conditions needed for $P_n D^*(Q_n, G_0) = o_P(n^{-1/2})$ might easily (asymptotically) hold for without any undersmoothing, but either way both will be guaranteed by enough undersmoothing. In practice we find that undersmoothing is important.

The statement of Theorem 1 relies on the following definitions that also provide the basis of the proof of the theorem as outlined above.

Definitions:

- Recall we can represent $Q_n = \arg\min_{Q \in \mathcal{Q}(C_n)} P_n L(Q)$ as follows:

$$Q_n(x) = Q_n(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{(0_s, x_s]} \mathrm{d}Q_{n,s}(u_s).$$

  For notational convenience, we define the extended measure $\mathrm{d}Q_n(u) = \sum_{s \subset \{1,\ldots,d\}} I_{E_s}(u) \mathrm{d}Q_{n,s}(u)$ onto the full cube $[0, \tau]$, not just $(0, \tau]$, where $[0, \tau] = \cup_s E_s$, $E_\emptyset = \{0\}$, $E_s = (0_s, \tau_s] \times \{0_{-s}\}$ is the s-specific left-edge for subsets $s \subset \{1, \ldots, d\}$, and $dQ_{n,s}(u)$ is the measure on E_s defined by the section $Q_{n,s}$ of Q. Note that E_s is defined by having coordinates in the complement of s being equal to zero. In this manner, we can use the compact representation:

$$Q_n(x) = \int_{[0,\tau]} \phi_x(u) \mathrm{d}Q_n(u),$$

  where we note that $\phi_x(u) \equiv I(x \geq u)$ reduces to $I(x_s \geq u_s)$ when u is on the edge E_s of $[0, \tau]$.
- Consider the family of paths $\left\{ Q^h_{n,\epsilon} : \epsilon \in (-\delta, \delta) \right\}$ through Q_n at $\epsilon = 0$ for arbitrarily small $\delta > 0$, indexed by any uniformly bounded $h \in D[0, \tau]$, defined by

$$Q^h_{n,\epsilon}(x) = \int_{[0,\tau]} \phi_x(u)(1 + \epsilon h(u)) \mathrm{d}Q_n(u), \tag{2}$$

- Let

$$r(h, Q_n) \equiv \int_{[0,\tau]} \phi_x(u) h(u) \mid \mathrm{d}Q_n(u) \mid,$$

- For any uniformly bounded h with $r(h, Q_n) = 0$ we have that for a small enough $\delta > 0$ $\left\{ Q^h_{n,\epsilon} : \epsilon \in (-\delta, \delta) \right\} \subset \mathcal{Q}(C_n)$.
- Let $S_h(Q_n) = \frac{\mathrm{d}}{\mathrm{d}\epsilon} L\left(Q^h_{n,\epsilon}\right)\Big|_{\epsilon=0}$ be the score of this h-specific submodel.
- Consider the set of scores

$$\mathcal{S}(Q_n) = \left\{ S_h(Q_n) = \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(f(h, Q_n)) : \| h \|_\infty < \infty \right\}, \tag{3}$$

  where

$$\begin{aligned} f(h, Q_n)(x) &\equiv \frac{\mathrm{d}}{\mathrm{d}\epsilon} Q^h_{n,\epsilon}\Big|_{\epsilon=0}(x) \\ &= \int_{[0,\tau]} \phi_x(u) h(u) \mathrm{d}Q_n(u). \end{aligned}$$

  This is the set of scores generated by the above class of paths if we do not enforce constraint $r(h, Q_n) = 0$.
- We have that Q_n solves the score equations $P_n S_h(Q_n) = 0$ for any uniformly bounded h satisfying $r(h, Q_n) = 0$.
- Let $D^*_n(Q_n, G_0) \in \mathcal{S}(Q_n)$ be an approximation of $D^*(Q_n, G_0)$ that is contained in this set of scores $\mathcal{S}(Q_n)$.
- We also consider a special case in which $D^*_n(Q_n, G_0) = D^*(Q_n, G_{0n})$ for an approximation $G_{0n} \in \mathcal{G}$ of G_0. Let

$$\mathcal{G}_n = \{ G \in \mathcal{G} : D^*(Q_n, G) \in \mathcal{S}(Q_n) \}$$

be the set of G's for which $D^*(Q_n, G)$ equals a score $S_h(Q_n)$ for some uniformly bounded h. One can then define $G_{0n} \in \mathcal{G}_n$ as an approximation of G_0.

– Let $h^*(Q_n, G_0)$ be the index so that $D_n^*(Q_n, G_0) = S_{h^*(Q_n, G_0)}(Q_n)$.

Remark: Understanding $\mathcal{G}_n$. It might seem that the class of paths $\left\{ Q_{n,\epsilon}^h : \epsilon \right\}$ for any bounded h above is rich enough to generate the full tangent space at Q_n and thereby $D^*(Q_n, G_0)$, even for finite n. However, a special property of this class of paths is that it is contained in the linear span of (order n) the basis functions ϕ_j that have non-zero coefficients $\beta_n(j)$ in Q_n. On the other hand, if n increases, and thereby the number of basis functions converges to infinity, this class of paths will indeed be able to approximate any function in the tangent space. Since the true G_0 or the relevant function of G_0 is generally not contained in this linear span of basis functions that make up Q_n, $D^*(Q_n, G_0) \notin \mathcal{S}(Q_n)$ is not contained in its set $\mathcal{S}(Q_n)$ of scores. For example, in the treatment-specific mean example, we would need that $1/\bar{G}_0(W)$ is approximated by this linear span of spline basis functions that are present in the fit Q_n. Therefore, indeed, there will be $G \in \mathcal{G}$ whose shape is such that $1/G(W)$ is in the linear span, which can then be used to define a G_{0n} so that $D^*(Q_n, G_{0n}) \in \mathcal{S}(Q_n)$. Alternatively, one directly approximates $1/\bar{G}_0(W)$ with a linear span, without being concerned if it results in a representation $1/\bar{G}_{0n}$, thereby determining an approximation $D_n^*(Q_n, G_0)$. Since in this example $\bar{G}_0$ can be any function of W with values in $(0, 1)$, in this example, both methods are equivalent: i.e., if $1/\bar{G}_0$ is approximated by $\sum_j \alpha_j \phi_j$, then we can solve for $\bar{G}_{0n}$ by setting $1/\bar{G}_{0n} = \sum_j \alpha_j \phi_j$, giving $G_{0n} = 1/\sum_j \alpha_j \phi_j$. This explains that indeed this set $\mathcal{G}_n$ will approximate $\mathcal{G}$ as n converges to infinity, so that G_{0n} will approximate G_0, typically as fast as Q_n approximates Q_0 (although that will depend on the undersmoothing of Q_n as well in the case that G_0 requires basis functions that are not needed for approximating Q_0). By increasing C_n, the number of selected basis functions in Q_n with non-zero coefficient will increase, thereby making the approximation G_{0n} better and better.

As is evident from Theorem 2 below, this approximation G_{0n} should aim to approximate G_0 in the sense that $R_{20}(Q_n, G_{0n}, Q_0, G_0) = o_P(n^{-1/2})$ while also arranging $P_0\{D^*(Q_n, G_{0n}) - D^*(Q_0, G_0)\}^2 \to_p 0$ (the latter being trivial by not requiring any rate).

Convenient notation for finite dimensional spline-representation of Q_n: Due to finite support condition $Q \ll^* \mu_n$ in the definition of the HAL-MLE, we have

$$Q_n(x) = \sum_{j \in \mathcal{J}(\mu_n)} \beta_n(j) \phi_j(x), \tag{4}$$

where $\phi_j(x) = I(x \geq u_j)$ for the set of indices of all the knot-points $\{u_j : j\} \subset [0, \tau]$, varying over the s-specific edges E_s of $[0, \tau]$ and across the different subsets $s \subset \{1, \ldots, d\}$. Note $\beta_n(j) = dQ_n(u_j)$, $j \in \mathcal{J}(\mu_n)$. Let $\mathcal{J}(Q_n) = \{j : \beta_n(j) \neq 0\} \subset \mathcal{J}(\mu_n)$ be the indices for the basis functions that have non-zero coefficient.

The following theorem establishes an undersmoothing condition (5) on C_n that guarantees $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$. We remind the reader of the definition of a directional derivative $\frac{\mathrm{d}}{\mathrm{d}Q} L(Q)(h) \equiv \frac{\mathrm{d}}{\mathrm{d}\delta_0} L(Q + \delta_0 h)$ in the direction h, where $\delta_0 = 0$.

Theorem 1. *Consider an approximation $D_n^*(Q_n, G_0) \in \mathcal{S}(Q_n)$ (i.e., scores of submodels not enforcing L_1-norm constant of HAL-MLE) of $D^*(Q_n, G_0)$ as defined above, and let h_n^* be so that $D_n^*(Q_n, G_0) = S_{h_n^*}(Q_n)$. Consider the representation (4) of Q_n. Note that β_n minimizes $\beta \to P_n L\left(\sum_{j \in \mathcal{J}(\mu_n)} \beta_n(j) \phi_j\right)$ over all $\beta = (\beta(j) : j \in \mathcal{J}(\mu_n))$ with $\sum_{j \in \mathcal{J}(\mu_n)} | \beta(j) | \leq C_n$. This theorem applies to any $Q_n = \sum_{j \in \mathcal{J}(\mu_n)} \beta_n(j) \phi_j$ with β_n a minimizer of the latter empirical risk. Let $S_j(Q) = \frac{d}{dQ} L(Q)(\phi_j)$.*

Assume $\| h_n^* \|_\infty = O_P(1)$, and

$$\min_{j \in \mathcal{J}(Q_n)} \| P_n \frac{d}{dQ_n} L(Q_n)(\phi_j) \| = o_P(n^{-1/2}). \tag{5}$$

Then,

$$P_n D_n^*(Q_n, G_0) = o_P(n^{-1.2}).$$

Let $j^* = \arg\min_{j \in \mathcal{J}_n(Q_n)} P_0 \phi_j$. *We can replace* (5) *by the following:* $P_0 S_{j^*}(Q_n)^2 \to_p 0$ *(which will generally hold whenever* $P_0 \phi_{j^*} = o_P(1)$*);* $\{S_j(Q) : Q \in \mathcal{Q}, j \in \mathcal{J}(\mu_n)\}$ *is contained in a Donsker class (e.g., the class of cadlag functions with uniformly bounded sectional variation norm);*

$$\| P_0 \left\{ \frac{d}{dQ_n} L(Q_n)(\phi_{j^*}) - \frac{d}{dQ_0} L(Q_0)(\phi_{j^*}) \right\} \| = o_P(n^{-1/2}), \tag{6}$$

and $P_0 \left\{ \frac{d}{dQ_n} L(Q_n)(\phi_{j^*}) \right\}^2 \to_p 0$.

Regarding (6)*, if we have*

$$\| P_0 \left\{ \frac{d}{dQ_n} L(Q_n)(\phi_{j^*}) - \frac{d}{dQ_0} L(Q_0)(\phi_{j^*}) \right\} = O_P \left(P_0^{1/2} \phi_{j^*} d_0^{1/2}(Q_n, Q_0) \right);$$

and $d_0(Q_n, Q_0) = O_P(n^{-2/3} (\log n)^{k_1})$ *(as we showed for HAL-MLE), then* (5) *is implied by*

$$\min_{j \in \mathcal{J}(Q_n)} P_0 \phi_j = o_P(n^{-1/3} (\log n)^{-k_1}). \tag{7}$$

Condition (5) is directly verifiable on the data and can thus be used to select the sectional variation norm bound C_n for the HAL-MLE. For example, one could select a constant a and set C to the smallest value (larger than the cross-validation selector) for which the left-hand side is smaller than $a/(\sqrt{n} \log n)$ for some constant a. The sufficient assumption (6) provides understanding of what it requires in terms of Q_n and P_0. We note that $P_0 \left\{ \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{s^*,j^*}) \right\}^2 \to_p 0$ is a relatively weak condition and is generally implied by the support of ϕ_{s^*,j^*} converging to zero, and is thereby a non-condition, given our undersmoothing condition (6). In the folllowing lemma we consider a common structure on the loss function and demonstrate that if we know that $Q_n - Q_0$ converges to zero in supremum norm at rate close to $n^{-1/3} (\log n)^{k_1/2}$, then condition (7) be significantly weakened.

Lemma 1. *Consider the special case that* $O = (Z, X)$, $L(Q)(O)$ *depends on* Q *through* $Q(X)$ *only, and* $\frac{d}{dQ} L(Q)(\phi) = \frac{d}{dQ} L(Q) \times \phi$*, i.e., the directional derivative* $\frac{d}{d\epsilon} L(Q + \epsilon\phi)\Big|_{\epsilon=0}$ *of* $L()$ *at* Q *in the direction* ϕ *is just multiplication of a function* $\frac{d}{dQ} L(Q)$ *of* O *with* $\phi(X)$*. Assume* $\limsup_n \| \frac{d}{dQ_n} L(Q_n) \|_\infty < \infty$*. Let* $j^* = \arg\min_{j \in \mathcal{J}(Q_n)} P_0 \phi_j$*. Assume* $P_0 \phi_{j^*} = o_P(1)$*. Then, a sufficient condition for* $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$ *is given by* (6)*.*

Assume

$$\| \frac{d}{dQ_n} L(Q_n) - \frac{d}{dQ_0} L(Q_0) \|_\infty = O(\| Q_n - Q_0 \|_\infty).$$

Then, $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$ *if*

$$\| Q_n - Q_0 \|_\infty \min_{j \in \mathcal{J}(Q_n)} P_0 \phi_j = o_P(n^{-1/2}). \tag{8}$$

The condition (8) *can be replaced by*

$$\| Q_n - Q_0 \|_\infty \min_{j \in \mathcal{J}(Q_n)} P_n \phi_j = o_P(n^{-1/2}).$$

Here $P_0 \phi_j$ *and* $P_n \phi_j$ *can be bounded by* $P_0(X \geq u_j)$ *and* $P_n(X \geq u_j)$*, respectively.*

Alternatively, we apply Theorem 1 above so that (6) *holds if* $\min_{j \in \mathcal{J}(Q_n)} P_n \phi_j = o_P(n^{-1/3} (\log n)^{-k_1})$.

In [26]; we proved that $\| Q_n - Q_0 \|_\infty \to_p 0$ under an absolute continuity condition. However, we expect that the rate of convergence with respect to supremum norm to be close to its rate $n^{-1/3} (\log n)^{k_1/2}$ with respect to $d_0^{1/2}(Q_n, Q_0)$, in which case this would only require that $\min_{j \in \mathcal{J}(Q_n)} P_n \phi_j = o_P(n^{-1/6})$.

3.3 Condition for solving the unconstrained score approximation of the efficient influence curve in terms of number of non-zero coefficients in HAL-MLE fit

Let Q_n be the HAL-MLE. It solves scores along paths $Q^h_{n,\epsilon}(x) = \int \phi_u(x)(1+\epsilon h)(u)\mathrm{d}Q_n(u)$ with $r(h,Q_n) = \int h \mid \mathrm{d}Q_n(u) \mid = 0$. This correspond with paths $Q_n(x) + \epsilon \int \phi_u(x)h(u)\mathrm{d}Q_n(u)$. Let $\mathrm{d}Z_{Q_n}(u) = h(u)\mathrm{d}Q_n(u)$. Note that the constraint

$$r(h,Q_n) = \int h \mid \mathrm{d}Q_n \mid /\mathrm{d}Q_n \mathrm{d}Q_n = \int \mid \mathrm{d}Q_n \mid /\mathrm{d}Q_n \mathrm{d}Z_{Q_n}(u).$$

Let $\ell_{Q_n} = \mid \mathrm{d}Q_n \mid /\mathrm{d}Q_n$, which is a vector with elements in $\{-1,1\}$ representing the sign of $\mathrm{d}Q_n(u)$. So in terms of Z_{Q_n} the constraint $r(h,Q_n)=0$ corresponds with $\int \ell_{Q_n}\mathrm{d}Z_{Q_n} = 0$. This shows that we can view the paths as $Q_n + \epsilon Z$ for any $Z \ll^* Q_n$ and $\int \ell_{Q_n}\mathrm{d}Z = 0$. Since $\mathrm{d}Q_n$ is discrete we can use notation $\beta_n(u) = \mathrm{d}Q_n(u)$. Then the paths correspond with $\beta^z_{n,\epsilon} = \beta_n + \epsilon z$ with z any vector with $z/\beta_n < \infty$ so that $\langle z, \ell_{\beta_n}\rangle = \sum_{j\in S_n} z(j)\ell_{\beta_n}(j) = 0$. The HAL-MLE Q_n satisfies for any $z \perp \ell_{\beta_n}$ with $z/\beta_n < \infty$:

$$0 = P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)\left(\sum_{j\in\mathcal{J}(Q_n)} z(j)\phi_j\right). \tag{9}$$

The following lemma establishes a bound for the latter score equation for z without the orthogonality constraint $z \perp \ell_{\beta_n}$, where this bound is in terms of the number J_n of knot-points in the fit Q_n with non-zero coefficient. Specifically, the bound is given by $J_n^{-1} d_0(Q_n,Q_0)^{1/2}$. One then wonders what the approximate rate is for J_n, and specifically when using the cross-validation selector. For this purpose, we note that Q_n is also an MLE for the parametric model $\sum_{j\in\mathcal{J}(Q_n)}\beta(j)\phi_j$ and one can show that the rate of convergence of Q_n to the best approximation $Q_{0,n}$ in this J_n-dimensional parametric model is $O_P((J_n/n)^{1/2})$ (to be addressed in detail in future research). Given that we know that the rate of convergence of Q_n to Q_0, using the cross-validation selector $C_{n,cv}$, is given by $n^{-1/3}(\log n)^{k_1/2}$, this suggest that $J_n \sim n^{1/3}(\log n)^{k_1/2}$, which then implies the rate $O_P(n^{-2/3})$ for the score Eq. (9) without the constraint $z \perp \ell_{\beta_n}$. Therefore, this result appears to formally establish that even without undersmoothing the score equation $P_n D^*_n(Q_n,G_0) = O_P(n^{-2/3})$ is already solved at the desired error (asymptotically). However, undersmoothing might still be needed, even asymptotically, for achieving the desired approximation of $D^*(Q_n,G_0)$ by an element $D^*_n(Q_n,G_0)$ in the linear span of the scores $S_h(Q_n)$ for uniformly bounded h (i.e., the selected basis functions in Q_n, even though sufficient to fit Q_0 at a good rate, might not generate enough scores to approximate the possibly more complex $D^*(Q_n,G_0)$, due to complexity of G_0).

Lemma 2. *Let J_n be the number of elements in support $\mathcal{J}(Q_n)$ of Q_n. Assume*

$$P_0\left(\frac{d}{dQ_n}L(Q_n) - \frac{d}{dQ_0}L(Q_0)\right)\left(\sum_{j\in\mathcal{J}(Q_n)} z(j)\phi_j\right) = O_P(\| z\|_1 d_0^{1/2}(Q_n,Q_0));$$

and that, uniformly in z with $\| z\|_1 < M$ for some $M<\infty$, the random function $\frac{d}{dQ_n}L(Q_n)\left(\sum_{j\in\mathcal{J}(Q_n)}z(j)\phi_j\right)$ falls in a fixed P_0-Donsker class (e.g., cadlag functions with universal bound on sectional variation norm).

We have that for any $z/\beta_n < \infty$,

$$P_n\frac{d}{dQ_n}L(Q_n)\left(\sum_{j\in\mathcal{J}(Q_n)} z(j)\phi_j\right) = O_P\left(\| z\|_1 J_n^{-1}n^{-1/2} + \| z\|_1 J_n^{-1} d_0^{1/2}(Q_n,Q_0)\right).$$

Clearly, the first term is $o_P(n^{-1/2})$ as long as $J_n \to \infty$. For example, if $J_n = n^{1/3}(\log n)^{k_1/2}$ and $d_0^{1/2}(Q_n,Q_0) = O_P(n^{-1/3}(\log n)^{k_1/2})$, then this becomes

$$P_n\frac{d}{dQ_n}L(Q_n)\left(\sum_{j\in\mathcal{J}(Q_n)} z(j)\phi_j\right) = O_P\left(\| z\|_1 n^{-2/3}\right).$$

Moreover, then

$$\sup_{\|z\|_1<M, z/\beta_n<\infty} \left| P_n \frac{d}{dQ_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} z(j) \phi_j \right) \right| = O_P(M n^{-2/3}).$$

Finally, if for an $M < \infty$, $D_n^*(Q_n, G_0) = \frac{d}{dQ_n} L(Q_n)(\sum_{j \in \mathcal{J}(Q_n)} z(j)\phi_j)$ *for a* $z = z(Q_n, G_0)$, *and* $\| z(Q_n, G_0) \|_1 < M$ *with probability tending to 1, then this implies* $P_n D_n^*(Q_n, G_0) = O_P(J_n^{-1} n^{-1/3} (\log n)^{k_1/2})$ *(and thus* $O_P(n^{-2/3})$ *if* $J_n = n^{1/3} (\log n)^{k_1/2}$*).*

Proof. Let z with $\| z \|_1 < M$ and $z/\beta_n < \infty$ be given. We define $\tilde{z} = z - \Pi(z \mid \ell_{\beta_n})$. Note that $\langle \ell_{\beta_n}, \ell_{\beta_n} \rangle = J_n^2$. So

$$\tilde{z} = z - \frac{\sum_{j \in \mathcal{J}(Q_n)} z(j) \ell_{\beta_n}(j)}{J_n^2} \ell_{\beta_n}.$$

In short notation we write $\tilde{z} = z - \pi_n(z)$ with $\pi_n(z) = \Pi(z \mid \ell_{\beta_n})$. Above shows that

$$\pi_n(z) = \frac{1}{J_n} \left(\sum_{j \in \mathcal{J}(Q_n)} z(j) \ell_{\beta_n}(j) \right) \ell_{\beta_n} / J_n \equiv J_n^{-1} \pi_n^*(z),$$

where

$$\| \pi_n^*(z) \|_1 = \left| \sum_{j \in \mathcal{J}(Q_n)} z(j) \ell_{\beta_n}(j) \right| \leq \| z \|_1 \leq M.$$

We have

$$\begin{aligned} 0 &= P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} \tilde{z}(j) \phi_j \right) \\ &= P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} z(j) \phi_j \right) - P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} \pi_n(z)(j) \phi_j \right) \end{aligned}$$

Therefore, using that $\pi_n(z) = J_n^{-1} \pi_n^*(z)$ with $\| \pi_n^*(z) \|_1 < M$, the Donker class and bounding condition of the Lemma, it follows that

$$\begin{aligned} -P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} z(j) \phi_j \right) &= P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} \pi_n(z)(j) \phi_j \right) \\ &= J_n^{-1} (P_n - P_0) \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} \pi_n^*(z)(j) \phi_j \right) \\ &\quad + J_n^{-1} P_0 \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_{j \in \mathcal{J}(Q_n)} \pi_n^*(z)(j) \phi_j \right) \\ &= O_P\left(J_n^{-1} n^{-1/2} \right) + J_n^{-1} P_0 \left(\frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) - \frac{\mathrm{d}}{\mathrm{d}Q_0} L(Q_0) \right) \\ &\quad \times \left(\sum_{j \in \mathcal{J}(Q_n)} \pi_n^*(z)(j) \phi_j \right) \\ &= O_P\left(J_n^{-1} n^{-1/2} + J_n^{-1} M d_0^{1/2}(Q_n, Q_0) \right). \end{aligned}$$

This completes the proof. □

3.4 Efficiency of the plug-in HAL MLE

The typical general efficiency proof used to analyze the TMLE (e.g., [18]) can be easily generalized to the condition that $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$ for some approximation $D_n^*(Q, G)$ of the actual canonical gradient $D^*(Q, G_0)$. This results in the following theorem.

Theorem 2. *Assume $M_1, M_{20} < \infty$. We have $d_0(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^{k_1})$. Assume condition (5) or conditions of Lemma 2 so that $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$.*

If $D_n^*(Q_n, G_0) = D^*(Q_n, G_{0n})$, then we assume

- *$R_2((Q_n, G_{0n}), (Q_0, G_0)) = o_P(n^{-1/2})$ and $P_0 \{D^*(Q_n, G_{0n}) - D^*(Q_0, G_0)\}^2 \to_p 0$.*
- *$\{D^*(Q, G): Q \in \mathcal{Q}, G \in \mathcal{G}\}$ is contained in the class of k_1-variate cadlag functions on a cube $[0, \tau_o] \subset \mathbb{R}^{k_1}$ in a Euclidean space and that $\sup_{Q \in \mathcal{Q}, G \in \mathcal{G}} \| D^*(Q, G)\|_v^* < \infty$.*

Otherwise, we assume

- *$R_2((Q_n, G_0), (Q_0, G_0)) = o_P(n^{-1/2})$, $P_0 \left\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\right\} = o_P(n^{-1/2})$, and $P_0 \left\{D_n^*(Q_n, G_0) - D^*(Q_0, G_0)\right\}^2 \to_p 0$.*
- *$\left\{D_n^*(Q, G_0), D^*(Q, G_0): Q \in \mathcal{Q}\right\}$ is contained in the class of k_1-variate cadlag functions on a cube $[0, \tau_o] \subset \mathbb{R}^{k_1}$ in a Euclidean space and that $\sup_{Q \in \mathcal{Q}} \max(\| D^*(Q, G_0)\|_v^*, \| D_n^*(Q, G_0)\|_v^*) < \infty$.*

Then, $\Psi(Q_n)$ is asymptotically efficient.

The proof is straightforward, analogue to typical efficiency proof for TMLE, and is presented in the Appendix. Regarding the condition, $P_0 \left\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\right\} = o_P(n^{-1/2})$, we note the following. For a typical choice $D_n^*(Q_n, G_0)$ in the set of scores $S(Q_n)$, we have $P_0 D_n^*(Q_0, G_0) = 0$, so that

$$-P_0 \left\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\right\} = P_0 \left\{D^*(Q_n, G_0) - D^*(Q_0, G_0)\right\} - P_0 \left\{D_n^*(Q_n, G_0) - D_n^*(Q_0, G_0)\right\}.$$

However, this equals the P_0-mean of the function $D^*(Q_n, G_0) - D^*(Q_0, G_0) = O_P(n^{-1/3}(\log n)^{k_1/2})$ minus its projection onto the linear span of scores $\{S_h(Q_n): h\}$. This will generally behave as a second order remainder involving a product of differences $Q_n - Q_0$ and $D_n^* - D^*$. The latter is addressed in detail in our integrated square density example.

Remark about general impact of undersmoothing on behavior of plug-in HAL-MLE. Though undersmoothing is beneficial for controlling $P_n D^*(Q_n, G_0) = o_P(n^{-1/2})$, it might harm the degree to which the Donsker class condition holds. This puts a clear restriction on the amount of undersmoothing allowed for asymptotic efficiency. The impact of undersmoothing on the second order remainder $R_2((Q_n, G_{0n}), (Q_0, G_0))$ is beneficial with respect to the approximation error of $G_0 - G_{0,n}$ but might make Q_n a poor approximation of Q_0. However, in many estimation problems, it appears that undersmoothing reduces the second order remainder as well in the sense that the second order remainder $R_2(Q_n, G_n, Q_0, G_0)$ can itself be represented as a score $P_0 S_h(Q_n)$ for some h, so that undersmoothing reduces the size of the second order remainder. Therefore, undersmoothing may be generally beneficial for the behavior of the estimator as long as its variation norm stays bounded by a universal constant (or a slowly converging constant) as sample size increases.

3.5 Inference for the plug-in undersmoothed HAL-MLE

The undersmoothed HAL-MLE $\Psi(Q_n)$ is asymptotically linear with influence curve $D^*(Q_0, G_0)$ so that it is approximately $N(\Psi(Q_0), \sigma_0^2 = P_0 \{D^*(Q_0, G_0)\}^2)$. There are various possible methods for estimating this normal limit distribution with corresponding confidence intervals. Let σ_n^2 be an estimator of this asymptotic variance. Then, an asymptotic 0.95-confidence interval is given by $\Psi(Q_n) \pm 1.96\sigma_n/n^{1/2}$. Let $\hat{G}: \mathcal{M}_{np} \to \mathcal{G}$

be an estimator of G_0. Then we can estimate σ_0^2 with $\sigma_n^2 = P_n \{D^*(Q_n, G_n)\}^2$, or a cross-validated estimator $\sigma_{n,cv}^2 = E_{B_n} P_{n,B_n}^1 \left\{ D^*(\hat{Q}\left(P_{n,B_n}^0\right), \hat{G}\left(P_{n,B_n}^0\right)) \right\}^2$, based on cross-validation scheme $B_n \in \{0,1\}^n$. The cross-validated estimator is generally more accurate by not suffering from overfitting, just as a cross-validated MSE is a better estimator than the empirical plug-in MSE. In both of these plug-in estimators σ_n^2 and $\sigma_{n,cv}^2$ there is no argument for preferring an undersmoothed Q_n over the HAL-MLE based on the cross-validation selector of the L_1-norm. Therefore, we recommend the latter.

This approach for obtaining inference would require the construction of an estimator of G_0, even though the HAL-MLE Q_n does not require this. To avoid such reliance on an estimator of G_0 and to improve finite sample coverage, we can use the non-parametric bootstrap in which the sampling distribution of $n^{1/2}(\Psi(Q_n) - \Psi(Q_0))$ is estimated with the distribution of $n^{1/2}(\Psi\left(Q_n^{\#}\right) - \Psi(Q_n))$, conditional on P_n, where $Q_n^{\#}$ is the undersmoothed HAL-MLE based on an i.i.d. sample from P_n. This method was proposed and analyzed in [27]; showing that the nonparametric bootstrap is a valid method for estimating the limit distribution of the plug-in HAL-MLE. In this bootstrap one can fix the L_1-norm of the HAL-MLE at the L_1-norm selected by the undersmoothed HAL-MLE, thereby making it computationally feasible. In [27] we also proposed a more conservative version of this bootstrap method by carrying out the bootstrap distribution for each L_1-norm and selecting the L_1-norm at which the width of the bootstrap confidence intervals reaches a plateau. In this manner, we guarantee that we sample from a maximal complex bootstrap distribution which was shown to yield robust finite sample coverage.

4 Example: HAL-MLE of treatment-specific mean

4.1 Formulation and relevant quantities for statistical estimation problem

Data and statistical model: Let $O = (W, A, Y) \sim P_0$, where $Y \in \{0,1\}$ and $A \in \{0,1\}$ are binary random variables. Let (A, W) have support $[0, \tau] \in \mathbb{R}^k$, where $A \in [0,1]$ with only support on the edges $\{0,1\}$. Similarly, certain components of W might be discrete so that it only has a finite set of support points in its interval. Note $O \in [0, \tau_o] = [0, \tau] \times [0,1]$, where $[0, \tau_o]$ is a cube in Euclidean space of same dimension as (W, A, Y). Let $\bar{G}(W) = E_P(A \mid W)$ and $\bar{Q}(W) = E_P(Y \mid A = 1, W)$. Assume the positivity assumption $\bar{G}_0(W) > \delta > 0$ for some $\delta > 0$; $\bar{Q}_0$, $\bar{G}_0$ are cadlag functions with $\| \bar{Q}_0 \|_v^* \leq C^u$ and $\| \bar{G}_0 \|_v^* \leq C_2^u$ for some finite constants C^u, C_2^u; $\delta < \bar{Q}_0 < 1 - \delta$ for some $\delta > 0$. This defines the statistical model $\mathcal{M}$ for P_0.

Target parameter, canonical gradient and exact second order remainder: Let $\Psi: \mathcal{M} \to \mathbb{R}$ denote the treatment-specific mean, defined by $\Psi(P) = E_P E_P(Y \mid W, A = 1)$. If an alternative quantity, such as the average treatment effect, $E_P\{E_P(Y \mid W, A = 1) - E_P(Y \mid W, A = 0)\}$, is of interest, the following strategy could be employed separately in each treatment group. Let $\tilde{Q} = (Q_W, \bar{Q})$, where Q_W is the probability distribution of W. Note that $\Psi(P) = \Psi(\tilde{Q}) = Q_W \bar{Q}(\cdot, 1)$. We have that Ψ is pathwise differentiable at P with canonical gradient given by $D^*(\tilde{Q}, G) = A/\bar{G}(W)(Y - \bar{Q}(W, A)) + \bar{Q}(1, W) - \Psi(\tilde{Q})$. Let $L(\bar{Q})(O) = -\left\{ Y \log \bar{Q}(W, A) + (1 - Y) \log(1 - \bar{Q}(W, A)) \right\}$ be the log-likelihood loss for $\bar{Q}$, and note that by the above bounding assumptions on $\bar{Q}$, we have that this loss function has finite universal bounds $M_1 < \infty$ and $M_{20} < \infty$. Let $D_1^*(\bar{Q}, \bar{G}) = A/\bar{G}(Y - \bar{Q})$ be the $\bar{Q}$-component of the canonical gradient, $D_2^*(\tilde{Q}) = \bar{Q}(1, W) - \Psi(Q)$ the Q_W-component, and note that $D^*(\tilde{Q}, G) = D_1^*(\bar{Q}, G) + D_2^*(\tilde{Q})$. We have $\Psi(\tilde{Q}) - \Psi(\tilde{Q}_0) = -P_0 D^*(\tilde{Q}, G) + R_{20}(\bar{Q}, \bar{G}, \bar{Q}_0, \bar{G}_0)$, where

$$R_{20}(\bar{Q}, \bar{G}, \bar{Q}_0, \bar{Q}_0) = P_0 \frac{\bar{G} - \bar{G}_0}{\bar{G}} (\bar{Q} - \bar{Q}_0).$$

Bounds on sectional variation norm and exact second order remainder: We have $\sup_{P \in \mathcal{M}} \| D^*(\tilde{Q}(P), G(P)) \|_v^* < C\left(C^u, C_2^u\right)$ for some finite constant C implied by the universal bounds $\left(C^u, C_2^u\right)$ on the sectional variation norm of $\bar{Q}, \bar{G}$. We also note that, using Cauchy–Schwarz inequality, $R_{20}(\bar{Q}, \bar{G}, \bar{Q}_0, \bar{G}_0) \leq \frac{1}{\delta} \| \bar{Q} - \bar{Q}_0 \|_{P_0} \| \bar{G} - \bar{G}_0 \|_{P_0}$, where $\| f \|_{P_0}^2 = \int f^2(o) \mathrm{d}P_0(o)$.

4.2 HAL-MLE

Let $Q = \text{Logit}\bar{Q}$ and let's $L(Q)(O) = -A\{Y \log \bar{Q}(W) + (1-Y)\log(1-\bar{Q}(W))\}$ be the log-likelihood loss restricted to the observations with $A = 1$. Let $Q_{C,n} = \arg\min_{Q,\|Q\|_v^* < C} P_n L(Q)$ be the C-specific 0-order Spline-HAL-MLE for a given bound C on the sectional variation norm. Let $C_n \leq C^u$ be a data adaptive selector that is larger or equal than the cross-validation selector, so that $P\left(C_{n,CV} \leq C_n \leq C^u\right) = 1$. Let $Q_n = Q_{C_n,n}$, and $Q_{W,n}$ be the empirical probability measure of $W_1, \ldots, W_n$. We can represent $Q_n = \sum_{j \in \mathcal{J}(\mu_n)} \beta_n(j)\phi_j$, where $\phi_j = I(W \geq w_j)$ for a knot point w_j over all observations $\{(W_{s,i}, 0_{-s}): i = 1, \ldots, n\}$ across all subsets $s \subset \{1, \ldots, k_1\}$. By our rate of convergence results on the HAL-MLE we have that $\| Q_n - Q_0 \|_{P_0} = O_P(n^{-1/3}(\log n)^{k_1/2})$. The HAL-MLE of $\Psi(\tilde{Q}_0)$ is the plug-in estimator $\Psi(\tilde{Q}_n) = Q_{W,n}\bar{Q}_n$. Note that $P_n D_2^*(\tilde{Q}_n) = 0$ for any Q_n. Thus, we are only concerned with showing that $P_n D_1^*(Q_n, G_0) = o_P(n^{-1/2})$.

Class of paths absolute continuous with respect to Q_n: Consider the following class of paths

$$Q_{n,\epsilon}^h(x) = \int_{[0,\tau]} \phi_x(u)(1 + \epsilon h(u))\mathrm{d}Q_n(u),$$

where the right-hand side can also be written as $Q_n(x) + \epsilon f(h, Q_n)(x)$, where

$$f(h, Q_n)(x) = h(0)Q_n(0) + \sum_{s \subset \{1,\ldots,m\}} \int_{0_s, x_s]} h(s, u_s))\, Q_{n,s}(\mathrm{d}u_s).$$

This defines a path $\left\{ Q_{n,\epsilon}^h : \epsilon \in (-\delta, \delta) \right\}$ for each uniformly bounded function h, as in our general representation.

Set of scores generated by class of paths: The scores generated by this family of paths are given by:

$$S_h(Q_n) \equiv \frac{\mathrm{d}}{\mathrm{d}\epsilon} L\left(Q_{n,\epsilon}^h\right)\Big|_{\epsilon=0} = A f(h, Q_n)(Y - Q_n(W)).$$

This defines a set of scores $\mathcal{S}(Q_n) = \{S_h(\bar{Q}_n): \| h\|_\infty < \infty\}$. Note that in order to solve for an h so that $S_h(\bar{Q}_n) = D_1^*(\bar{Q}_n, \bar{G}_0)$ would require $f(h, \bar{Q}_n)(W) = 1/\bar{G}_0(W)$. However, since $\bar{G}_0$ is not sectional absolute continuous with respect to Q_n (i.e., $Q_{n,s}$ is discrete for all subsets s, while $\bar{G}_{0,s}$ is (say) continuous), there does not exist a h for which $f(h, Q_n) = 1/\bar{G}_0$. Thus, $D^*(Q_n, \bar{G}_0) \notin \{S_h(Q_n): \| h\|_\infty < \infty\}$.

Score equations solved by HAL-MLE:

Let

$$r(h, Q_n) = \int_{[0,\tau]} h(u) \mid \mathrm{d}Q_n(u) \mid,$$

which can also be written as

$$r(h, Q_n) \equiv h(0)) \mid Q_n(0) \mid + \sum_{s \subset \{1,\ldots,m\}} \int_{(0_s, x_s]} h(s, u_s)) \mid \mathrm{d}Q_{n,s}(u_s) \mid .$$

The HAL-MLE solves

$$P_n S_h(Q_n) = 0 \text{ for all } h \text{ with } r(h, Q_n) = 0.$$

4.3 Defining approximation G_{0n}

We define

$$\mathcal{G}_n \equiv \left\{ \bar{G} \in \mathcal{G}: \bar{G} \ll^* \bar{Q}_n \right\}.$$

We note that if $\bar{G}_s \ll \bar{Q}_{n,s}$, then we also have $1/\bar{G}_s \ll \bar{Q}_{n,s}$ as well. Here we use that if $g(x) = 1/f(x)$, then $g_s(\mathrm{d}x_s) = -1/f_s^2(x_s) f_s(\mathrm{d}x_s)$. Therefore, if $\bar{G} \ll^* \bar{Q}_n$, then we can find a h so that $f(h, Q_n)(A, W) = A/\bar{G}(W)$, and thereby that $D_1^*(Q_n, \bar{G}) = S_h(Q_n)$.

Let

$$G_{0n} = \arg\min_{\bar{G}\in\mathcal{G}_n} \| \bar{G} - \bar{G}_0\|_{P_0},$$

where $\| \bar{G} - \bar{G}_0\|_{P_0}$ is the $L^2(P_0)$-norm of $\bar{G} - \bar{G}_0$. Then, $D_1^*(Q_n, \bar{G}_{0n}) \in \{S_h(Q_n): h\}$ so that we can find a $h^*(Q_n, \bar{G}_0)$ so that

$$D_1^*(Q_n, \bar{G}_{0n}) = S_{h^*(Q_n,\bar{G}_0)}(Q_n).$$

4.4 Application of Theorem 2

We need to assume $R_2((Q_n, G_{0n}), (Q_0, G_0)) = o_P(n^{-1/2})$ and $P_0 \{D^*(Q_n, G_{0n}) - D^*(Q_0, G_0)\}^2 \to_p 0$. The latter already holds if $\| \bar{G}_{0n} - \bar{G}_0\|_{P_0} \to_p 0$. However, the first condition relies on a rate of convergence. Given the rate of convergence for the HAL-MLE Q_n, it thus suffices that $\| \bar{G}_{0n} - \bar{G}_0\|_{P_0} = o_P(n^{-1/6}(\log n)^{k_1/2})$. This appears to be a reasonable condition, since $\bar{G}_{0n}$ is the $L^2(P_0)$-projection of $\bar{G}_0$ onto $\mathcal{G}_n$, so that the only concern would be that the set $\mathcal{G}_n$ does not approximate fast enough $\mathcal{G}$ as n converges to infinity. However, if the set of basis functions is rich enough for $\bar{Q}_n$ to converge at a rate than $n^{-1/3}(\log n)^{k_1/2}$ to $\bar{Q}_0$ (not allowing to choose the coefficients based on P_0), then the resulting linear combination of indicator basis functions should generally also be rich enough for approximating the true G_0 with a rate $n^{-1/6}(\log n)^{k_1/2}$ (now allowing to select the coefficients of the basis functions in terms of G_0).

Verification of Assumption 5 of Theorem 2: Assumption (5) is stating that

$$\min_{j\in\mathcal{J}(Q_n)} P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_j) = 2\frac{1}{n}\sum_i \phi_j(1, W_i) I(A_i = 1) I(W_i > w_j)(Y_i - \bar{Q}_n(1, W_i))$$

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

We apply the last part of Theorem 1. Since $\frac{\mathrm{d}}{\mathrm{d}Q} L(Q)(\phi) = \phi(A, W)(Y - \bar{Q}(A, W))$, it follows that have

$$\| \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) - \frac{\mathrm{d}}{\mathrm{d}Q_0} L(Q_0)\|_{P_0} = O(\| Q_n - Q_0\|_{P_0}). \tag{10}$$

Given that we have $d_0(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^{k_1})$, it follows that the remaining condition is (7), or, equivalently,

$$\min_{j\in\mathcal{J}(Q_n)} P_n\phi_j = o_P\left(n^{-1/3}(\log n)^{-k_1}\right).$$

This reduces to the assumption that $O\left(\min_{j\in\mathcal{J}(Q_n)} P_n(W \geq w_j)\right) = o_P(n^{-1/3}(\log n)^{-k_1})$. We arrange this assumption to hold by selecting C_n accordingly.

Verification of assumptions of Lemma 2: Given our assumptions, it is straightforward to verify the conditions of Lemma 2. This lemma provides the bound $J_n^{-1} d_0^{1/2}(Q_n, Q_0)$ for $P_n D_n^*(Q_n, G_0)$. This provides then the alternative condition for choosing C_n (and thereby J_n) to establish $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$.

This proves the following efficiency theorem for the HAL-MLE in this particular estimation problem.

Theorem 3. *Consider the formulation above of the statistical estimation problem. Let*

$$\mathcal{G}_n = \left\{\bar{G} \in \mathcal{G}: \bar{G} \ll^* \bar{Q}_n\right\},$$

and

$$\bar{G}_{0n} = \arg\min_{\bar{G}\in\mathcal{G}_n} \| G - G_0\|_{P_0}.$$

Assumptions:

- $\| \bar{G}_{0n} - \bar{G}_0\|_{P_0} = o_P(n^{-1/6}(\log n)^{-k_1/2})$, *where we can use that* $\| Q_n - Q_0\|_{P_0} = o_P(n^{-1/3}(\log n)^{k_1/2})$.
- *Given the fit* $Q_n = \sum_{j\in\mathcal{J}(Q_n)} \beta_n(j)\phi_j$ *with knot-points the observations* $\{W_j(s): j = 1, \ldots, n, s\}$ *and indicator basis functions* $\phi_j(W) = I(W > W_j)$, *we assume that* $C_n < C^u$ *for some finite* C^u *is chosen so that*

$$\min_{j\in\mathcal{J}(Q_n)} P_n\phi_j = o_P(n^{-1/3}(\log n)^{-k_1}).$$

Alternatively, select the number of knot-points J_n (i.e., C_n) so that $J_n^{-1} d_0^{1/2}(Q_n, Q_0) = o_P(n^{-1/2})$ (e.g. $J_n = n^{1/3}(\log n)^{k_1/2}$).

Then, $\Psi(Q_n)$ is an asymptotically efficient estimator of $\Psi(Q_0)$.

5 Example: HAL-MLE for the integrated square of the data density

Let $O \sim P_0$ be a k_1-variate random variable with Lebesgue density p_0 that is assumed to be bounded from below by a $\delta > 0$ and from above by an $M < \infty$. Let $\{P_Q : Q \in \mathcal{Q}\}$ be a parametrization of the probability measure of O in terms of a functional parameter Q that varies over a class of multivariate real valued cadlag functions on $[0, \tau]$ with finite sectional variation norm. Below we will focus on the particular parameterization given by $p_Q = c(Q)\{\delta + (M - \delta)\text{expit}(Q)\}$, where $\text{expit}(x) = 1/(1 + \exp(-x))$, and $c(Q)$ is the normalizing constant defined by $\int p_Q \mathrm{d}o = 1$. Note that in this parameterization Q can be any cadlag function with finite sectional variation norm, thereby allowing that the densities p_Q are discontinuous (but cadlag). Another possible parametrization is obtained through the following steps: (1) modeling the density $p(x)$ as a product $\prod_{j=1}^{k} p_j(x_j \mid \bar{x}(j-1))$ of univariate conditional densities of x_j, given $\bar{x}(j-1)$; (2) modeling each univariate conditional density p_j in terms of its univariate conditional hazard λ_j; (3) modeling this hazard as $\lambda(x_j \mid \bar{x}(j-1)) = \exp(Q_j(x_j, \bar{x}(j-1)))$ (or discretizing it and modeling it with a logistic function in Q_j), and (4) setting $Q = (Q_1, \ldots, Q_{k_1})$. With this latter parametrization each Q_j varies over a parameter space of cadlag functions with finite sectional variation norm.

Let the statistical model $\mathcal{M} = \{P_Q : Q \in \mathcal{Q}(C^u)\}$ for P_0 be nonparametric beyond that each probability distribution is dominated by the Lebesgue measure, Q varies over cadlag functions with sectional variation norm bounded by C^u. The statistical target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by $\Psi(P) = \int p^2(o)\mathrm{d}o$. The canonical gradient of Ψ at P is given by $D^*(P)(O) = 2(p(O) - \Psi(P))$, and, the exact second order remainder $R_2(P, P_0) = \Psi(P) - \Psi(P_0) + P_0 D^*(P)$ is given by $R_2(P, P_0) = -\int (p - p_0)^2(o)\mathrm{d}o$.

Let $L(Q) = -\log p_Q$ be the log-likelihood loss function for Q. Let Q_n be an HAL-MLE bounding the sectional variation norm by a $C_n < C^u$. We wish to establish conditions on C_n so that $\Psi(Q_n) = \int p_{Q_n}^2 \mathrm{d}o$ is an asymptotically efficient estimator of $\Psi(Q_0) = \int p_{Q_0}^2 \mathrm{d}o$. We assume this HAL-MLE is discrete so that we can use the finite dimensional representation $Q_n = \sum_{j \in J(Q_n)} \beta_n(j)\phi_j$ with $\| \beta_n \|_{L_1} \leq C_n$, as in our general presentation. Let $Q_{n,\epsilon}^h(x) = \int_{[0,\tau]} \phi_x(u)(1 + \epsilon h(u))\mathrm{d}Q_n(u)$, indexed by any bounded function h, be the paths as defined in our general presentation (and previous section). Let $S_h(Q_n) = \frac{\mathrm{d}}{\mathrm{d}\epsilon} L\left(Q_{n,\epsilon}^h\right)\Big|_{\epsilon=0}$ be score of this path under the log-likelihood loss. These scores are given by

$$S_h(Q_n) = \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(f(h, Q_n)),$$

where $f(h, Q_n) = \int_{[0,\tau]} \phi_x(u)h(u)\mathrm{d}Q_n(u)$, which also equals $Q_n(0)h(0) + \sum_s \int_{(0_s, x_s]} h(s, u_s)\mathrm{d}Q_{n,s}(u_s)$. Let $S(Q_n) = \{S_h(Q_n) : h\}$ be the collection of scores. In order to apply Theorem 2 we need to determine an approximation $D_n(Q_n) \in S(Q_n)$ of the canonical gradient $D^*(Q_n) = 2(p_{Q_n} - \Psi(Q_n))$. We have

$$S_h(Q) = A(f(h, Q))/C(Q) - M\frac{\exp(Q)}{(1 + \exp(Q))(\delta + \delta\ \exp(Q) + M)} f(h, Q),$$

where

$$A(f) = \frac{\int \exp(Q)/(1 + \exp(Q))^2 f \mathrm{d}o}{\left(\int (\delta + M/(1 + \exp(Q)))\mathrm{d}o\right)^2}.$$

Let $G(Q) = -M\frac{\exp(Q)}{(1+\exp(Q))(\delta+\delta\ \exp(Q)+M)}$, so that the equation $S_h(Q) = D^*(Q)$ corresponds with $G(Q)f(h, Q) + C(Q)^{-1}A(f(h, Q)) = D^*(Q)$, which can be rewritten as $f(h, Q) + G_1(Q)A(f(h, Q)) = D^*(Q)/G(Q)$, and $G_1(Q) = 1/(C(Q)G(Q))$. Let $D_1(Q) = D^*(Q)/G(Q)$, so that the equation becomes $f + G_1(Q)A(f) = D_1(Q)$. Once we have solved for f, whose solution we will denote with $f(Q)$, then it remains to solve for h in $f(h, Q) = f(Q)$ or find a

closest solution. It is important to note the $f \to A(f)$ is a linear real valued operator. Applying this operator to both sides yields $A(f) + A(f)A(G_1(Q)) = A(D_1(Q))$, so that we obtain the solution

$$A(f) = \frac{A(D_1(Q))}{1 + A(G_1(Q))}.$$

Plugging this back into the equation, we obtain $f(Q) \equiv D_1(Q) - \frac{G_1(Q)A(D_1(Q))}{1+A(G_1(Q))}$. Thus, we have shown that if we can set $f(h, Q_n) = f(Q_n)$, then we have $S_h(Q_n) = D^*(Q_n)$. It remains to determine a choice $h(Q_n)$ so that $f(h, Q_n) \approx f(Q_n)$. The space $\{f(h, Q_n): h\}$ equals $\{\sum_{j\in\mathcal{J}(Q_n)}\alpha(j)\phi_j: \alpha\}$ the linear span of the basis functions $\{\phi_j: j \in \mathcal{J}(Q_n)\}$. Let $f_n(Q_n)$ be the projection of $f(Q_n)$ onto this linear space, for example, with respect to $L^2(P_0)$-norm. Let $h_n(Q_n)$ be the solution of $f(h, Q_n) = f_n(Q_n)$, and let $D_n^*(Q_n) = S_{h_n(Q_n)}(Q_n)$ be our desired approximation of $D^*(Q_n)$ which is an element of the set of scores $\{S_h(Q_n): h\}$. We note that

$$\begin{aligned} D_n^*(Q_n) - D^*(Q_n) = S_{h_n(Q_n)}(Q_n) - D^*(Q_n) &= G(Q_n)f_n(Q_n) + C(Q_n)^{-1}A(f_n(Q_n)) - D^*(Q_n) \\ &= G(Q_n)f_n(Q_n) + C(Q_n)^{-1}A(f_n(Q_n)) \\ &\quad - G(Q_n)f(Q_n) - C(Q_n)^{-1}A(f(Q_n)) \\ &= G(Q_n)(f_n(Q_n) - f(Q_n)) + C(Q_n)^{-1}A(f_n(Q_n) - f(Q_n)). \end{aligned}$$

We will assume that $\| f_n(Q_n) - f(Q_n)\|_{P_0} = o_P(n^{-1/4})$. The main condition beyond (5) of Theorem 2 is that $P_0\left\{D_n^*(Q_n) - D^*(Q_n)\right\} = o_P(n^{-1/2})$. Note that $P_0D_n^*(Q_0) = 0 = P_0D^*(Q_0)$. Therefore,

$$\begin{aligned} P_0\left\{D_n^*(Q_n) - D^*(Q_n)\right\} &= P_0\left\{D_n^*(Q_n) - D_n^*(Q_0)\right\} - P_0\left\{D^*(Q_n) - D^*(Q_0)\right\} \\ &= P_0\{G(Q_n)(f_n(Q_n) - f(Q_n))\} + P_0\left\{C(Q_n)^{-1}A(f_n(Q_n) - f(Q_n))\right\} \\ &\quad - P_0\left\{G(Q_0)(f_n(Q_0) - f(Q_0)) - C(Q_0)^{-1}A(f_n(Q_0) - f(Q_0))\right\}. \end{aligned}$$

Let Π_n be the projection operator on the linear span generated by the basis function of Q_n, which is of the same dimension as the number of basis functions. The latter difference can also be represented as

$$P_0\left\{D^*(Q_n) - D^*(Q_0) - \Pi_n(D^*(Q_n) - D^*(Q_0))\right\},$$

or, if we define $\Pi_n^\perp = (I - \Pi_n)$ as the projection operator onto the orthogonal complement of the linear space spanned by the basis functions in Q_n, then this term can be denoted as

$$P_0\left\{\Pi_n^\perp(D^*(Q_n) - D^*(Q_0))\right\}, \tag{11}$$

which can, in particular, be bounded by the operator norm $\| \Pi_n^\perp \|$ of $\Pi_n^\perp$ times the $L^2(P_0)$-norm of $D^*(Q_n) - D^*(Q_0)$. Thus, if we assume that $\| \Pi_n^\perp \| = o_P(n^{-1/6}(\log n)^{-k_1/2})$, then it follows that this term is $o_P(n^{-1/2})$. We will simply assume (11) to be $o_P(n^{-1/2})$. The other conditions, beyond (5) of Theorem 2 hold by the fact that $\| Q_n - Q_0\|_{P_0} = O_P(n^{-1/3}(\log n)^{k_1/2})$ and that $D^*(Q_n), D_n^*(Q_n)$ fall in a P_0-Donsker class of cadlag functions with universal bound on sectional variation norm.

Verification of Assumption 5 of Theorem 2: Assumption (5) is stating that

$$\min_{j\in\mathcal{J}(Q_n)} P_n\frac{\mathrm{d}}{\mathrm{d}Q_n}L(Q_n)(\phi_j) = o_P(n^{-1/2}).$$

We apply the last part of Theorem 1. We have

$$\| \frac{\mathrm{d}}{\mathrm{d}Q_n}L(Q_n) - \frac{\mathrm{d}}{\mathrm{d}Q_0}L(Q_0)\|_{P_0} = O(\| Q_n - Q_0\|_{P_0}). \tag{12}$$

Given that we have $d_0(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^{k_1})$, it follows that the remaining condition is (7), or, equivalently,

$$\min_{j\in\mathcal{J}(Q_n)} P_n\phi_j = O_P(n^{-1/3}(\log n)^{-k_1}).$$

This reduces to the assumption that $O\left(\min_{j\in\mathcal{J}(Q_n)}P_n(O\geq u_j)\right)=O_P(n^{-1/3}(\log n)^{-k_1})$, where u_j are the knot-points making up the support of μ_n. We arrange this assumption to hold by selecting C_n accordingly.

Alternatively, as in previous example, by applying Lemma 2, we select the number J_n of knot-points with non-zero coefficients so that $J_n d_0^{1/2}(Q_n,Q_0)=o_P(n^{-1/2})$.

This proves the following efficiency theorem for the HAL-MLE in this particular estimation problem.

Theorem 4. *Let $O\sim P_0$ be a k_1-variate random variable with Lebesgue density p_0 that is assumed to be bounded from below by a $\delta>0$ and from above by an $M<\infty$. Let $p_Q=c(Q)\{\delta+(M-\delta)expit(Q)\}$, where $expit(x)=1/(1+\exp(-x))$, and $c(Q)$ is the normalizing constant defined by $\int p_Q\mathrm{d}o=1$, where $Q\in\mathcal{Q}(C^u)$ can be any cadlag function with finite sectional variation norm bounded by C^u. Let the statistical model $\mathcal{M}=\left\{P_Q\colon Q\in\mathcal{Q}(C^u)\right\}$ for P_0 be nonparametric beyond that each probability distribution is dominated by the Lebesgue measure, Q varies over cadlag functions with sectional variation norm bounded by C^u. The statistical target parameter $\Psi\colon\mathcal{M}\to\mathbb{R}$ is defined by $\Psi(P)=\int p^2(o)\mathrm{d}o$, which we also denote with $\Psi(Q)$. The canonical gradient of Ψ at P is given by $D^*(P)(O)=2(p(O)-\Psi(P))$, and, the exact second order remainder $R_2(P,P_0)=\Psi(P)-\Psi(P_0)+P_0D^*(P)$ is given by $R_2(P,P_0)=-\int(p-p_0)^2do$.*

Consider the formulation above of the statistical estimation problem. We have $\|Q_n-Q_0\|_{P_0}=O_P(n^{-1/3}(\log n)^{k_1})$.

Assumptions*:*

- *Given the fit $Q_n=\sum_{j\in\mathcal{J}(\mu_n)}\beta_n(j)\phi_j$ with support points the observations $\mathcal{J}(\mu_n)=\{(O_j(s),0(-s))\colon j=1,\ldots,n,s\}$ and indicator basis functions $\phi_j(W)=I(O>u_j)$ with $u_j\in\mathcal{J}(\mu_n)$, we assume that $C_n<C^u$ for some finite C^u is chosen so that*

$$\min_{j\in\mathcal{J}(Q_n)}P_n\phi_j=O_P(n^{-1/3}(\log n)^{-k_1}).$$

  Alternatively, select the number J_n of knot-points with non-zero coefficients so that $J_n d_0^{1/2}(Q_n,Q_0)=o_P(n^{-1/2})$.
- *Let $\Pi_n^{\perp}$ be the projection operator in $L^2(P_0)$ onto the orthogonal complement of the linear span of the basis functions $\{\phi_j\colon j\in\mathcal{J}(Q_n)\}$ in the fit of Q_n. Assume*

$$P_0\left\{\Pi_n^{\perp}(D^*(Q_n)-D^*(Q_0))\right\}=o_P(n^{-1/2}). \tag{13}$$

  A sufficient condition is that the operator norm $\|\Pi_n^{\perp}\|$ of $\Pi_n^{\perp}$ is $o_P(n^{-1/6}(\log n)^{-k_1/2})$.

Then, $\Psi(Q_n)$ is an asymptotically efficient estimator of $\Psi(Q_0)$.

6 Simulation study

Our global undersmoothing conditions only specify a sufficient rate at which the sparsest selected basis function should converge to zero, or at which the number of basis functions selected will converge to infinity, but it does not provide a constant in front of this rate. Thus, it does not immediately yield a practical method for tuning the level of undersmoothing. In our simulation studies, we investigate the targeted L_1-norm selector that is chosen so that the empirical mean of the canonical gradient at the HAL-MLE (indexed by L_1-norm) and possibly a HAL-MLE of the nuisance parameter in the canonical gradient is $o_P(n^{-1/2})$. In extensive simulations, this method appears to give better practical results than several direct implementations of our global undersmoothing criterion (i.e., the choice of constant matters for practical performance). More research will be needed to investigate if one can construct a global undersmoothing selector (according to our theorem) that would result in well behaved efficient plug-in estimators across a large class of target parameters. Our simulations also demonstrate that our targeted selection method for undersmoothing controls the sectional variation norm of the fit, which is a crucial part of the Donsker class or asymptotic equicontinuity condition.

6.1 Simulations for the treatment-specific mean

We simulated a vector $W = (W_1, W_2)$, with W_1 created by drawing $Z \sim \text{Beta}(0.85, 0.85)$ and setting $W_1 = 4Z - 2$. W_2 was drawn independently from a Bernoulli(0.5) distribution. Given $W = w$, a binary random variable A was drawn with probability $A = 1$ equal to $\bar{G}_0(w) = \text{logit}^{-1}\{w_1 - 2w_1w_2\}$. Thus, the required positivity condition holds by design with $P_0\{0.119 < \bar{G}_0(W) < 0.881\} = 1$. Given $W = w$, we set $Y = \bar{Q}_0(w) + \epsilon$, where $\bar{Q}_0(w) = \text{logit}^{-1}\{w_1 - 2w_1w_2\}$ and $\epsilon \sim \text{Normal}(0, 0.25)$. The true value of the treatment-specific mean is $\Psi(P_0) = E_{P_0}E_{P_0}(Y \mid W, A = 1) = 0.5$. We refer readers back to Section 4 for the form of the canonical gradient.

We built our undersmoothed estimator of $\Psi(P_0)$ as follows. We estimate $\bar{Q}_0$ using a HAL regression estimator and select the regularization of the estimator by choosing the smallest value of C for L_1-norm such that

$$P_nD^*(Q_{C,n}, \bar{G}_n) < \frac{P_n^{1/2}\{D^*(Q_{C,n}, \bar{G}_n)^2\}}{\log(n)n^{1/2}},$$

where $\bar{G}_n$ is the HAL-MLE estimate of $\bar{G}_0$ (i.e., a HAL regression that uses cross-validated choice for C). We then computed the plug-in estimator as described in Section 4.

We generated 3000 data sets in this way and computed the undersmoothed HAL estimate. We report the estimator's bias (scaled by $n^{1/2}$), Monte Carlo variance (scaled by n), mean squared error (by n), and the sampling distribution of $n^{1/2}\{\Psi(\tilde{Q}_n) - \Psi(P_0)\}$. We additionally report on the behavior of $n^{1/2}P_nD^*(Q_{C_n,n}, \bar{G}_0)$ and

$$n^{1/2}\left\{\min_{s,j\in J_n(s),\beta_n(s,j)\neq 0} \| P_n\frac{\mathrm{d}}{\mathrm{d}Q_n}L(Q_n)(\phi_{s,j}) \|\right\}.$$

As predicted by theory, the bias of the estimator diminishes faster than $n^{-1/2}$ and the variance of the estimator approaches the efficiency bound in larger samples (Figure 1 and 2). The empirical average of the canonical gradient is appropriately controlled (top right) and our selection criteria for the HAL tuning parameter appears to also satisfy the global criteria stipulated by Eq. (5). At all sample sizes, the sampling distribution of the scaled and centered estimator is well-approximated by the efficient asymptotic distribution.

6.2 Simulations for the integral of the square of the density

We simulated a univariate variable $O \sim N(-4, 5/3)$ and evaluated the performance of undersmoothed HAL for estimating the integral of the square of the density of O (Section 5). We implemented a HAL-based estimator of the density using an approach similar to the one described in [28]. This approach entails estimating a discrete hazard function using HAL using a pre-specified binning of the real line. For this simulation, we used 320 equidistant bins, and note that the HAL density estimator is robust to this choice, so long as a large enough value is chosen. We sample 1000 data sets for each of several sample sizes ranging from $n = 100$ to 100,000. We compare the results for undersmoothed HAL to those obtained by using a typical implementation of HAL that selects the level of smoothing based on cross-validation. We compared these estimators on the same criterion described in the previous subsection.

The simulations results reflect what is expected based on theory. In particular, the undersmoothed HAL achieves the efficiency bound in large samples and the scaled-centered sampling distribution of the estimator is well approximated by the efficient asymptotic distribution. We found that our selection criterion for the level of undersmoothing based on the EIF led to control of the variation norm of the resultant fit. On the other hand, results for the HAL estimator with level of smoothing selected via cross-validation demonstrated that this estimator does not have bias that is decreasing faster than $n^{-1/2}$. Thus, this estimator performs worse in terms of all criteria that we considered. The estimated cross-validated and undersmoothed function paths as well as the true function are illustrated in Figure 3.

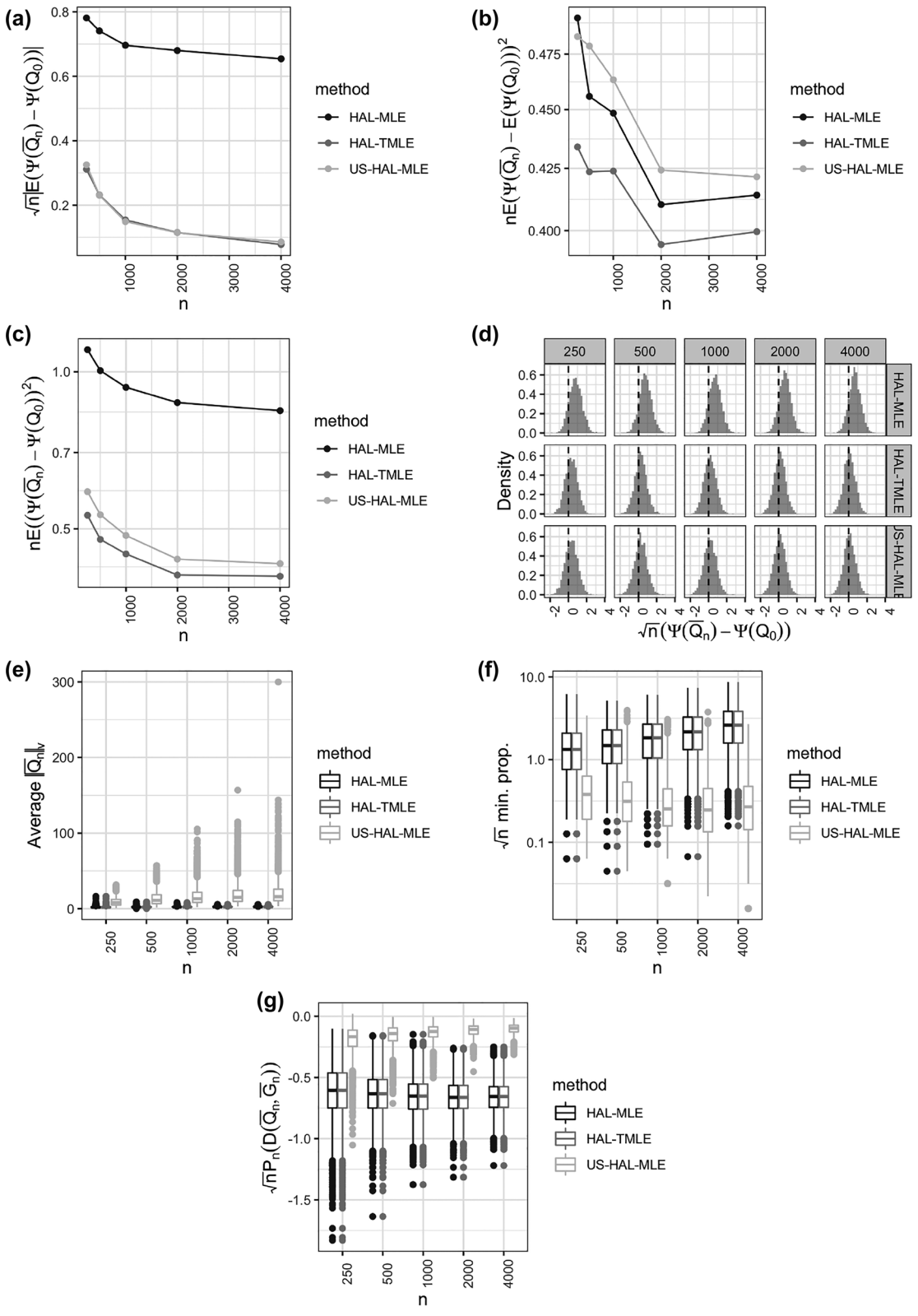

Figure 1: Simulation results for the treatment-specific mean parameter: (a) bias in absolute value, (b) variance, (c) mean-squared error (all scaled by $n^{1/2}$), (d) Sampling distribution of scaled and centered estimator, (e) Sectional variation norm of the nuisance parameter, (f) empirical average of quantity given in Eq. (5), (g) sample average of the efficient influence function, evaluated at the sample estimate.

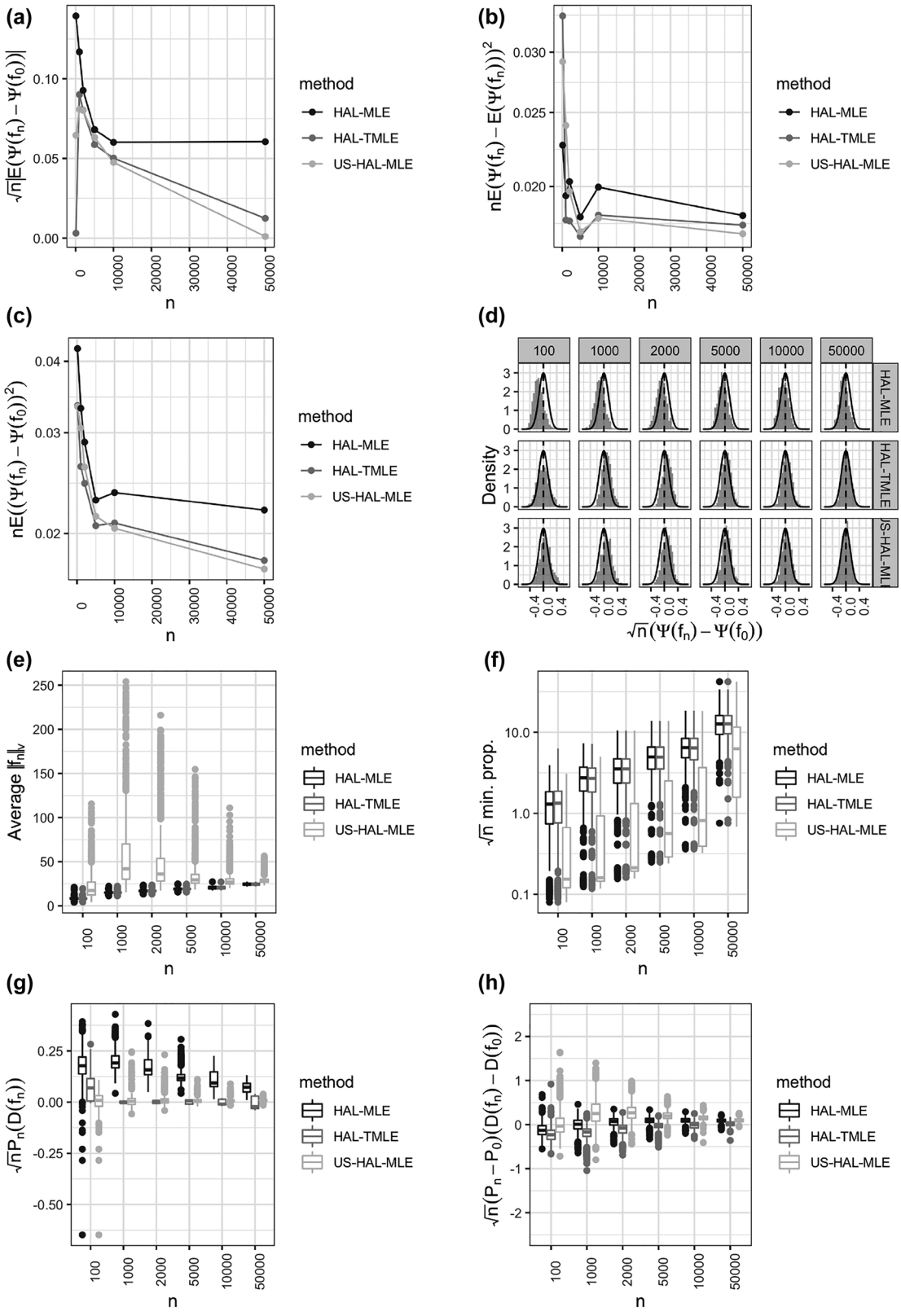

Figure 2: Simulation results for the average density value parameter: (a) bias in absolute value (b) variance (c) mean-squared error (all scaled by $n^{1/2}$); (d) Sampling distribution of scaled and centered estimator, (e) Sectional variation norm of the nuisance parameter (f) empirical average of quantity given in Eq. (5), (g) sample average of the efficient influence function, evaluated at the sample estimate, (h) $\sqrt{n}(P_n - P_0)(D_n - D_0)$. The dashed lines in the mean-squared error plots denote the efficiency bound. The reference sampling distribution for the estimators is a mean-zero normal distribution with this variance.

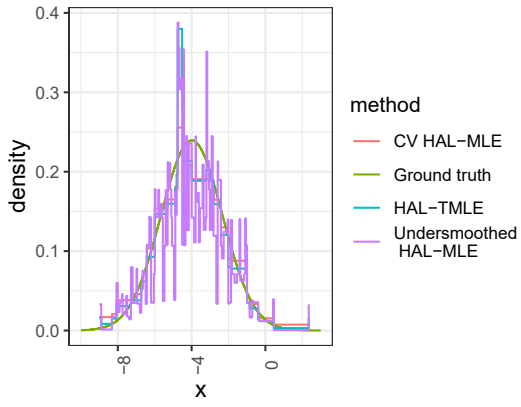

Figure 3: A random realization of the Simulation in Section 7.2 (n = 500).

7 Discussion

In this article we established that for realistic and nonparametric statistical models an overfitted zero-order spline HAL-MLE of a functional parameter of the data distribution results in efficient plug-in estimators of pathwise differentiable functionals of this functional parameter. The statistical model can be any model for which the parameter space of the functional parameter is a (cartesian product of a) subset of the set of multivariate cadlag functions with a universal bound on the sectional variation norm. The undersmoothing condition involves two purposes. Firstly, one wants to undersmooth so that solving the L_1-constrained scores $P_n S_h(Q_n)$ with $r(h, Q_n) = 0$ implies solving $P_n S_h(Q_n) = o_P(n^{-1/2})$, and thereby $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$ for the best score approximation $D_n^*(Q_n, G_0)$ of $D^*(Q_n, G_0)$. For that purpose we showed that it suffices to select the L_1-norm in the HAL-MLE large enough so that the basis functions with non-zero coefficients includes "sparse enough" basis functions, where "sparse enough" corresponds with assuming that the proportion of non-zero elements (among n observations of this basis function) in the basis function converges to zero at a rate faster than $n^{-1/3}(\log n)^{k_1}$. Alternatively, one controls the number J_n of non-zero coefficients so that $J_n d_0(Q_n, Q_0) = o_P(n^{-1/2})$. The latter establishes that, from an asymptotic perspective, this condition will even be satisfied for the cross-validation selector. The second purpose is to undersmooth enough so that the approximation $D_n^*(Q_n, G_0)$ becomes a good enough approximation of $D^*(Q_n, G_0)$. If there is no nuisance parameter G_0, then one generally expects this to hold for the cross-validation selector. However, if there is a nuisance parameter G_0, then undersmoothing might be needed since the dependence on G_0 might be more complex than it is to fit Q_0 itself, so that the fit Q_n needs to select extra basis functions beyond the ones needed to approximate Q_0. This shows that from an asymptotic point of view the need for undersmoothing appears minimal, but in practice (where the constant matters), we have observed that it is important.

The undersmoothing conditions presented are not parameter specific, so that such an undersmoothed HAL-MLE will be efficient for any of its smooth functionals. In addition, the undersmoothing of the HAL-MLE does not change its rate of convergence relative to the HAL-MLE optimally tuned with cross-validation, as long as the selected L_1-norm remains uniformly bounded, suggesting that it is still a good estimator of the true functional parameter.

On the other hand, a typical TMLE targeting one particular target parameter will generally only be asymptotically efficient for that particular target parameter, and not even asymtotically linear for other smooth functionals, even if it uses as initial estimator the HAL-MLE tuned with cross-validation. Therefore it appears to be an interesting topic to better understand the sampling distribution of the undersmoothed HAL-MLE in an asymptotic sense and in relation to a sampling distribution of a TMLE using an optimally smoothed (i.e., cross-validation) HAL-MLE as initial estimator. Note, however, that if the TMLE uses an undersmoothed HAL-MLE as initial estimator, than the TMLE step should result in small changes, thereby mostly preserving the behavior of the undersmoothed HAL-MLE. Therefore, the latter type of TMLE might be recommended,

inheriting the good global behavior of the HAL-MLE, also in light of the recent work on higher order TMLE [29].

It is also of interest to observe that the second order remainder of the HAL-MLE for a pathwise differentiable functional appears to either be driven by the square of the $L^2(P_0)$-norm of the HAL-MLE itself with respect to the functional parameter, or, in the case that the efficient influence curve has a nuisance parameter G, a second order remainder might also (or only) involve a product of differences of the HAL-MLE Q_n with respect to its true counterpart Q_0 and the difference of a projection $G_{0,n}$ of the true G_0 with respect to the linear space of basis functions selected by the undersmoothed HAL-MLE Q_n. Since $G_{0,n}$ is a type of oracle estimator of G_0, this suggest that in a model in which our knowledge on G_0 is not any better than our knowledge on Q_0, this HAL-MLE has a good second order remainder that might generally be smaller than what it would be for a TMLE that estimates G_0 with an actual estimator such as the HAL-MLE.

On the other hand, if the statistical model involves particularly strong knowledge on the nuisance parameter G_0, then a TMLE can fully utilize this model on G_0 and thereby obtain a better behaved second order remainder than the one for the overfitted HAL-MLE. One also suspects that a TMLE will be more sensitive to lack of positivity for the target parameter than the undersmoothed HAL-MLE. Therefore, we conjecture that an undersmoothed HAL-MLE might be the preferred estimator in models in which case the estimation of G_0 is as hard as estimation of Q_0, and when lack of positivity is a serious issue, while an HAL-TMLE might be the preferred estimator when estimation of G_0 is easier than estimation of Q_0. These are not formal statements, but indicate a qualitative comparison between the undersmoothed HAL-MLE and a HAL-TMLE using an estimator (HAL-MLE) G_n of G_0.

However, this above comparison has an additional twist of interest in favor of the HAL-MLE. That is, if G_0 happens to be a function with relative small variation norm, unknown to the analyst, then we will have much faster convergence of $G_{0,n}$ to G_0 than if the true G_0 is very complex. As such the undersmoothed HAL-MLE will have a remainder involving a very fast converging $G_{0,n}$, possibly faster than the estimator G_n used by the TMLE utilizing this simple model. Thus, the HAL-MLE is able to adapt to underlying (unknown) smoothness of G_0, making it even less obvious that TMLE utilizing knowledge on G_0 will do any better. All of this strongly suggest that the TMLE should use an undersmoothed HAL-MLE as initial estimator and make sure that the targeting step does not destroy the score equations already solved by the HAL-MLE. We will address the latter in a future article.

In future research we will address the comparison between undersmoothed HAL-MLE and HAL-TMLE in realistic simulations and by formal comparison by their second order remainders (some of it already shown by [29]). Specifically, in a subsequent article we will marry the TMLE with the HAL-MLE by defining a targeted HAL-MLE that minimizes the empirical risk over the linear span of basis functions (approximating the true cadlag function with finite sectional variation norm) under the L_1-constraint *and* under the constraint that the Euclidean norm of the empirical mean of the efficient influence curve at the HAL-MLE (as well as at an estimator G_n) is $o_P(n^{-1/2})$. We will show that undersmoothing this targeted HAL-MLE results in an estimator that is still efficient across all smooth functionals, while it is able to fully exploit all knowledge on G_0 for the sake of the specific target parameter. Moving forward it will also be critically important to compare approaches for building confidence intervals and performing hypothesis tests.

A key advantage of a TMLE is that it can utilize any super-learner so that its library can include many other powerful machine learning algorithms, including many variations of the HAL-MLE. In this manner a TMLE using a powerful super-learner might compensate for the favorable property of an undersmoothed HAL-MLE with respect to size of the second order remainder. In another future article we will provide a method that marries a powerful super-learner with HAL-MLE, by using the super-learner as a dimension reduction, and applying HAL-MLE as the meta learning step in an ensemble learner. We will show that an undersmoothed HAL-MLE in this metalearning step will result again in an estimator that is efficient, and possibly super-efficient, for any of its smooth functionals. By actually using a targeted HAL-MLE as meta learning step, we might end up with an estimator that is able to still fully exploit the strengths super-learning, undersmoothed HAL-MLE, and TMLE using a good estimator of G_0, combined in one method.

Undersmoothing of HAL-MLE can also be applied to nuisance parameters such as an HAL-MLE of the censoring and treatment mechanism in an inverse probability of treatment and censoring weighted (IPTCW) estimator. By undersmoothing the HAL-MLE G_n of the censoring and treatment mechanism G_0, smooth functionals of G_n become asymptotically efficient just as shown in this article for undersmoothed Q_n. An analysis of an IPTCW estimator precisely relies on showing that a smooth functional of G_n is asymptotically linear. Therefore, in this manner we can show that an IPTCW-estimator that uses an undersmoothed HAL-MLE for estimation of the censoring and treatment mechanism is regular and asymptotically linear and even efficient if the full data model is saturated [30].

Finally, we refer to our accompanying technical report [31] that presents a generalization of this highly adaptive lasso estimator to minimizers of empirical risk over smoothness classes that are spanned by the higher order spline basis functions. The current HAL-MLE corresponds with zero-order spline basis functions. The order of the spline can be selected with cross-validation resulting in an HAL-MLE that also adapts to underlying smoothness. We plan to publish this part in a later article.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This work was supported by the National Institute of Allergy and Infectious Diseases (grant number 5R01AI074345-09).
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendix A: Proof of Theorem 1 and Lemma 1

The HAL-MLE has the form $Q_n = \sum_{j\in\mathcal{J}(\mu_n)}\beta_n(j)\phi_j$ for a finite collection of basis functions. A basis function is of form $\phi_j(X) = I(X \geq x_j)$ for a knot point $x_j \in [0,\tau]$, and if the components of x_j in the complement of a subset $s \subset \{1,\ldots,k\}$ are equal to zero, then this indicator reduces to an indicator $I(X(s) \geq x_j(s))$. We also know that $\sum_j \mid \beta_n(j) \mid \leq C_n$ for the selected L_1-bound C_n (typically the L^1-norm will be equal to C_n). We have that

$$\beta_n = \arg\min_{\beta,\sum_j|\beta(j)|\leq C_n} P_n L\left(\sum_{j\in\mathcal{J}(\mu_n)}\beta(j)\phi_j\right).$$

Consider paths $(1+\epsilon h(j))\beta_n(j)$ for a bounded vector h, which yields a collection of scores

$$S_h(Q_n) = \frac{\mathrm{d}}{\mathrm{d}Q_n}L(Q_n)\left(\sum_{j\in\mathcal{J}(\mu_n)}h(j)\beta_n(j)\phi_j\right).$$

Let $r(h,Q_n) = \sum_{j\in\mathcal{J}(\mu_n)}h(j) \mid \beta_n(j) \mid$. If $r(h,Q_n) = 0$, then for ϵ small enough,

$$\begin{aligned}\sum_{j\in\mathcal{J}(\mu_n)} \mid (1+\epsilon h(j))\beta_n(j) \mid &= \sum_{j\in\mathcal{J}(\mu_n)}(1+\epsilon h(j)) \mid \beta_n(j) \mid \\ &= \sum_{j\in\mathcal{J}(\mu_n)} \mid \beta_n(j) \mid +\epsilon r(h,Q_n) \\ &= \sum_{j\in\mathcal{J}(\mu_n)} \mid \beta_n(j) \mid .\end{aligned}$$

Thus, by β_n being an MLE, $P_n S_h(Q_n) = 0$ for any h satisfying $r(h,Q_n) = 0$. Let $h^* = h_n^*$ be chosen so that $P_n S_{h_n^*}(Q_n) = P_n D_n^*(Q_n,G_0)$ for the approximation $D_n^*(Q_n,G_0)$ of $D^*(Q_n,G_0)$ specified in the theorem. We want to show that $P_n D_n^*(Q_n,G_0) = o_P(n^{-1/2})$, i.e. $P_n S_{h_n^*}(Q_n) = o_P(n^{-1/2})$. Let j^* be a particular choice in our finite index set $\mathcal{J}_n$ satisfying $\beta_n(j^*) \neq 0$, which we can specify later to minimize the bound. Let $\tilde{h}$ be defined by $\tilde{h}(j) = h^*(j)$

for $j \neq j^*$, and $\tilde{h}(j^*)$ is defined by $r(\tilde{h}, Q_n) = \sum_{j \in \mathcal{J}(\mu_n)} \tilde{h}(j) \mid \beta_n(j) \mid = 0$, so that we know $P_n S_{\tilde{h}}(Q_n) = 0$. Thus,

$$\sum_{j \neq j^*} h^*(j) \mid \beta_n(j) \mid + \tilde{h}(j^*) \mid \beta_n(j^*) \mid = 0.$$

This gives

$$\tilde{h}(j^*) = -\frac{\sum_{j \neq j^*} h^*(j) \mid \beta_n(j) \mid}{\mid \beta_n(j^*) \mid}.$$

So

$$\sum_j (\tilde{h} - h^*)(j)\beta_n(j)\phi_j = (\tilde{h} - h^*)(j^*)\beta_n(j^*)\phi_{j^*} = \left(-\frac{\sum_{j \neq j^*} h^*(j) \mid \beta_n(j) \mid}{\mid \beta_n(j^*) \mid} \beta_n(j^*) - h^*(j^*)\beta_n(j^*) \right) \phi_{j^*}$$

$$\equiv c_n(j^*)\phi_{j^*},$$

where

$$c_n(j^*) = -\frac{\sum_{(j) \neq (j^*)} h^*(j) \mid \beta_n(j) \mid}{\mid \beta_n(j^*) \mid} \beta_n(j^*) - h^*(j^*)\beta_n(j^*).$$

We note that $c_n(j^*)$ is bounded by $\sum_j \mid h^*(j) \| \beta_n(j) \mid$. So we can bound this by $\| h^* \|_\infty C_n$. Thus under the assumption that $\| h_n^* \|_\infty = O_P(1)$, we have that $c_n(j^*) = O_P(1)$.

For this choice $\tilde{h}$, let's compute $P_n S_{\tilde{h}}(Q_n) - P_n S_{h^*}(Q_n)$ (which equals $P_n S_{h^*}(Q_n)$):

$$\begin{aligned} P_n S_{\tilde{h}}(Q_n) - P_n S_{h^*}(Q_n) &= P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) \left(\sum_j (\tilde{h} - h^*)(j)\beta_n(j)\phi_j \right) \\ &= P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(c_n(j^*)\phi_{j^*}) \\ &= c_n(j^*) P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*}) \\ &= O_P\left(P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*}) \right). \end{aligned}$$

Therefore, our undersmoothing condition is that

$$\min_{j \in \mathcal{J}(\mu_n), \beta_n(j) \neq 0} \| P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_j) \| = o_P(n^{-1/2}). \tag{14}$$

Under this condition we have $P_n S_{\tilde{h}}(Q_n) - P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$, but, since $P_n S_{\tilde{h}}(Q_n) = 0$, this implies the desired conclusion $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$. This proves the first statement of Theorem 1.

Let now $j^* = \arg\min_{j \in \mathcal{J}(\mu_n)} P_0 \phi_j$. To understand, $P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*})$ we can proceed as follows.

$$P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*}) = (P_n - P_0) \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*}) + P_0 \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*}).$$

Let $S_j(Q_n) \equiv \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_j)$. Suppose that $P_0 S_{j^*}(Q_n)^2 \to_p 0$, which will generally hold whenever $P_0 \phi_{j^*} = o_P(1)$. We also have that $\{S_j(Q) : Q \in \mathcal{Q}, (j)\}$ is contained in the class of cadlag functions with uniformly bounded sectional variation norm, which is a Donsker class. Thereby, by asymptotic equicontinuity of the empirical process indexed by a Donsker class, we have $(P_n - P_0) S_{j^*}(Q_n) = o_P(n^{-1/2})$. Thus, it remains to show that $P_0 S_{j^*}(Q_n) = o_P(n^{-1/2})$. We now note that

$$P_0 S_{j^*}(Q_n) = P_0 \left\{ S_{j^*}(Q_n) - S_{j^*}(Q_0) \right\} + P_0 S_{j^*}(Q_0),$$

but $P_0 S_j(Q_0) = 0$ for all j, since $Q_0 = \arg\min_Q P_0 L(Q)$. Therefore, $P_n \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)(\phi_{j^*}) = o_P(n^{-1/2})$ if

$$P_0 \left\{ S_{j^*}(Q_n) - S_{j^*}(Q_0) \right\} = o_P(n^{-1/2}). \tag{15}$$

This proves the second statement of Theorem 1. The third statement is a trivial implication, which completes the proof of Theorem 1. □

Proof of Lemma 1. Consider the special case that $O = (Z, X)$, $L(Q)(O)$ depends on Q through $Q(X)$ only, and $\frac{\mathrm{d}}{\mathrm{d}Q} L(Q)(\phi) = \frac{d}{dQ} L(Q) \times \phi$, i.e., the directional derivative $\frac{\mathrm{d}}{\mathrm{d}\epsilon} L(Q + \epsilon\phi)\Big|_{\epsilon=0}$ of $L()$ at Q in the direction ϕ is just multiplication of a function $\frac{\mathrm{d}}{\mathrm{d}Q} L(Q)$ of O with $\phi(X)$. In that case, we have that (15) reduces to

$$P_0 \left\{ \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) - \frac{\mathrm{d}}{\mathrm{d}Q_0} L(Q_0) \right\} \phi_{j^*} = o_P(n^{-1/2}). \tag{16}$$

We assume

$$\| \frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n) - \frac{\mathrm{d}}{\mathrm{d}Q_0} L(Q_0) \|_\infty = O(\| Q_n - Q_0 \|_\infty).$$

Then, (16) reduces to

$$\| Q_n - Q_0 \|_\infty P_0 \phi_{j^*} = o_P(n^{-1/2}).$$

This teaches us that the critical condition (14) holds if

$$\min_{j \in \mathcal{J}(\mu_n), \beta_n(j) \neq 0} P_0 \phi_j = O_P(n^{-1/2}).$$

and that for this choice j^* we have $P_0 \left\{ S_{j^*}(Q_n) \right\}^2 \to_p 0$. The latter holds if $\min_j P_0 \phi_j = o_P(1)$, since $\frac{\mathrm{d}}{\mathrm{d}Q_n} L(Q_n)$ is uniformly bounded. Finally, since $P_0 \phi_j = O(P_0(X \geq x_j))$, $\sup_j \mid (P_n - P)\phi_j \mid = O_P(n^{-1/2})$, we can replace $P_0 \phi_j$ by $P_n \phi_j$ in the condition. This proves Lemma 1.

Appendix B: Proof of Theorem 2

Let G_{0n} be an approximation of G_0, and let $D^*(Q_n, G_{0n})$ be the approximation of $D^*(Q_n, G_0)$ in the space of scores $\mathcal{S}(Q_n)$. We have the following general theorem which proves the first part of Theorem 2.

Theorem 5. *Consider the HAL-MLE Q_n with $C = C_u$ or $C = C_n$. Assume $M_1, M_{20} < \infty$. We have $d_0(Q_n, Q_0) = O_P(n^{-1/2-\alpha(k_1)})$. Assume also that for a given approximation $G_{0n} \in \mathcal{G}$ of G_0 which satisfies*

$$P_n D^*(Q_n, G_{0n}) = o_P(n^{-1/2}). \tag{17}$$

- $R_2((Q_n, G_{0n}), (Q_0, G_0)) = o_P(n^{-1/2})$ *and* $P_0 \left\{ D^*(Q_n, G_{0n}) - D^*(Q_0, G_0) \right\}^2 \to_p 0$.
- $\{D^*(Q, G) : Q \in \mathcal{Q}, G \in \mathcal{G}\}$ *is contained in the class of k_1-variate cadlag functions on a cube $[0, \tau_o] \subset \mathbb{R}^{k_1}$ in a Euclidean space and that* $\sup_{Q \in \mathcal{Q}, G \in \mathcal{G}} \| D^*(Q, G) \|_v^* < \infty$.

Then $\Psi(Q_n)$ is asymptotically efficient at P_0.

Proof. The exact second order expansion at G_{0n} of the target parameter Ψ yields

$$\Psi(Q_n) - \Psi(Q_0) = (P_n - P_0) D^*(Q_n, G_{0n}) - P_n D^*(Q_n, G_{0n}) + R_2((Q_n, G_{0n}), (Q_0, G_0)).$$

Given that $d_0(Q_n, Q_0) = O_P(n^{-1/2-\alpha(k_1)})$, and that G_{0n} is presumably at least as good of an approximation of G_0, it is a reasonable assumption to assume $R_2((Q_n, G_{0n}), (Q_0, G_0)) = o_P(n^{-1/2})$ and $P_0 \left\{ D^*(Q_n, G_{0n}) - D^*(Q_0, G_0) \right\}^2 \to_p 0$. We also assume that $\{D^*(Q, G) : Q \in \mathcal{Q}, G \in \mathcal{G}\}$ is contained in the class of

k_1-variate cadlag functions on a cube $[0, \tau_o] \subset \mathbb{R}^{k_1}$ in a Euclidean space and that $\sup_{Q\in\mathcal{Q},G\in\mathcal{G}} \| D^*(Q,G)\|_v^* < \infty$. This essentially states that the sectional variation norm of $D^*(Q,G)$ can be bounded in terms of the sectional variation norm of Q and G, which will naturally hold under a strong positivity assumption that bounds denominators away from zero. Since the class of cadlag functions on $[0, \tau_o]$ with sectional variation norm bounded by a universal constant is a Donsker class, empirical process theory yields:

$$\Psi(Q_n) - \Psi(Q_0) = (P_n - P_0)D^*(Q_0, G_0) - P_n D^*(Q_n, G_{0n}) + o_P(n^{-1/2}).$$

□

This theorem can be easily generalized to a more general approximation $D_n^*(Q_n, G_0) \in S(Q_n)$ of $D^*(Q_n, G_0)$ (not necessarily of form $D_n^*(Q_n, G_0) = D^*(Q_n, G_{0n})$ for some G_{0n}).

Theorem 6. *Consider the HAL-MLE Q_n with $C = C_u$ or $C = C_n$. Assume $M_1, M_{20} < \infty$. We have $d_0(Q_n, Q_0) = O_P(n^{-1/2-\alpha(k_1)})$. Assume also that for a given approximation $D_n^*(Q_n, G_0)$ we have $P_n D^*(Q_n, G_{0n}) = o_P(n^{-1/2})$. In addition, assume*

- $R_2((Q_n, G_0), (Q_0, G_0)) = o_P(n^{-1/2})$, $P_0\left\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\right\} = o_P(n^{-1/2})$, *and* $P_0\left\{D_n^*(Q_n, G_0) - D^*(Q_0, G_0)\right\}^2 \to_p 0$.
- $\left\{D_n^*(Q, G_0), D^*(Q, G_0) : Q \in \mathcal{Q}\right\}$ *is contained in the class of k_1-variate cadlag functions on a cube $[0, \tau_o] \subset \mathbb{R}^{k_1}$ in a Euclidean space and that* $\sup_{Q\in\mathcal{Q}} \max(\| D^*(Q, G_0)\|_v^*, \| D_n^*(Q, G_0)\|_v^*) < \infty$.

Then $\Psi(Q_n)$ is asymptotically efficient at P_0.

Therefore, in order to prove Theorem 2, it remains to establish the condition under which (17) holds, which was proven in the previous Appendix A.

B.1 General proof of efficient score equation condition at G_0

This subsection can be skipped for the purpose of proving Theorem 2, but the following result fits here.

Lemma 3. *Under the conditions of Theorem 5, if $P_n D^*(Q_n, G_{0n}) = o_P(n^{-1/2})$, then also $P_n D^*(Q_n, G_0) = o_P(n^{-1/2})$. Under the conditions of Theorem 6, if $P_n D_n^*(Q_n, G_0) = o_P(n^{-1/2})$, then also $P_n D^*(Q_n, G_0) = o_P(n^{-1/2})$.*

Proof. Firstly, we have

$$\begin{aligned} P_n D^*(Q_n, G_0) &= P_n D_n^*(Q_n, G_0) + P_n\left\{D^*(Q_n, G_0) - D_n^*(Q_n, G_0)\right\} \\ &= P_n\left\{D^*(Q_n, G_0) - D_n^*(Q_n, G_0)\right\} + o_P(n^{-1/2}). \end{aligned}$$

In addition, we have

$$\begin{aligned} P_n\left\{D^*(Q_n, G_0) - D_n^*(Q_n, G_0)\right\} &= (P_n - P_0)\left\{D^*(Q_n, G_0) - D_n^*(Q_n, G_0)\right\} \\ &\quad + P_0\left\{D^*(Q_n, G_0) - D_n^*(Q_n, G_0)\right\} \\ &= o_P(n^{-1/2}) + P_0\left\{D^*(Q_n, G_0) - D_n^*(Q_n, G_0)\right\}, \end{aligned}$$

since $\sup_{Q\in\mathcal{Q}(\mathcal{M})} \max(\| D^*(Q, G_0)\|_v^*, \| D_n^*(Q, G_0)\|_v^*) < \infty$, and $P_0\left\{D_n^*(Q_n, G_0) - D^*(Q_n, G_0)\right\}^2 \to_p 0$. If $D_n^*(Q_n, G_0) = D^*(Q_n, G_{0n})$, then the first assumption holds if $\sup_{P\in\mathcal{M}} \| D^*(P)\|_v^* < \infty$.

To understand the last term, consider the case that $D_n^*(Q_n, G_0) = D^*(Q_n, G_{0n})$. By the exact second order expansion $\Psi(Q_n) - \Psi(Q_0) = -P_0 D^*(Q_n, G) + R_{20}(Q_n, G, Q_0, G_0)$ for all G, we have

$$P_0\left\{D^*(Q_n, G_0) - D^*(Q_n, G_{0n})\right\} = R_{20}(Q_n, G_0, Q_0, G_0) - R_{20}(Q_n, G_{0n}, Q_0, G_0).$$

In our general Theorem 5 we assumed $R_{20}(Q_n, G_{0n}, Q_0, G_0) = o_P(n^{-1/2})$, which certainly implies $R_{20}(Q_n, G_0, Q_0, G_0)$ (which actually equals zero in double robust problems). This then establishes that

$$P_n D^*(Q_n, G_0) = o_P(n^{-1/2}).$$

For general $D_n^*(Q_n, G_0)$, Theorem 6 simply assumed $P_0 \left\{ D^*(Q_n, G_0) - D_n^*(Q_n, G_0) \right\} = o_P(n^{-1/2})$. □

References

1. Bickel PJ, Klaassen CAJ, Ritov Y, Wellner J. Efficient and adaptive estimation for semiparametric models. Berlin, Heidelberg, New York: Springer; 1997.
2. Newey W. The asymptotic variance of semiparametric estimators. Econometrica 2014;62:1349–82.
3. van der Laan MJ. Causal effect models for intention to treat and realistic individualized treatment rules. In: Technical report 203. Berkeley: Division of Biostatistics, University of California; 2006.
4. van der Vaart AW. Asymptotic statistics. Cambridge, New York; 1998.
5. Shen X. On methods of sieves and penalization. Annals of Statitics 1997;252:2555–91.
6. Shen X. Large sample sieve estimation of semiparametric models. In: Chapter in handbook of econometrics, vol 76; 2007.
7. Giné E, Nickl R. A simple adaptive estimator of the integrated square of a density. Bernoulli 2008;14:47–61.
8. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 1998;2:315–31.
9. Newey WK, Robins JR. Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint arXiv:1801.09138 2018.
10. Newey WK, Hsieh F, Robins JM. Twicing kernels and a small bias property of semiparametric estimators. Econometrica 2004;72:947–62.
11. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: AIDS epidemiology. Basel: Birkhäuser; 1992.
12. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. Berlin, Heidelberg, New York: Springer; 2003.
13. van der Laan MJ. Estimation based on case-control designs with known prevalance probability. Int J Biostat 2008;4:17. https://doi.org/10.2202/1557-4679.1114.
14. van der Laan MJ, Gruber S. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. Int J Biostat 2016;12:351–78.
15. van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. Berlin, Heidelberg, New York: Springer; 2011.
16. van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. Int J Biostat 2006;2:11.
17. Benkeser D, van der Laan MJ. The highly adaptive lasso estimator. In: 2016 IEEE international conference on data science and advanced analytics (DSAA). Montreal, QC, Canada: IEEE; 2016:689–96 pp.
18. van der Laan MJ. A generally efficient targeted minimum loss-based estimator. In: Technical report 300. UC Berkeley; 2015. to appear in IJB, 2017 http://biostats.bepress.com/ucbbiostat/paper343.
19. Gill RD, van der Laan MJ, Wellner JA. Inefficient estimators of the bivariate survival function for three models. Ann Inst Henri Poincaré 1995;31:545–97.
20. Polley EC, Rose S, van der Laan MJ. Super learner. In: van der Laan MJ, Rose S, editors. Targeted learning: causal inference for observational and experimental data. New York, Dordrecht, Heidelberg, London: Springer; 2011.
21. van 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. In: Technical report 130. Berkeley: Division of Biostatistics, University of California; 2003.
22. van der Laan MJ, Dudoit S, van der Vaart AW. The cross-validated adaptive epsilon-net estimator. Stat Decis 2006;24:373–95.
23. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol 2007;6:25.
24. van der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross-validation. Stat Decis 2006;24:351–71.
25. Bibaut A, van der Laan MJ. Fast rates for empirical risk minimization over cadlag functions with bounded sectional variation norm. In: Technical report. Berkeley: Division of Biostatistics, University of California; 2019.
26. van der Laan MJ, Bibaut A. Uniform consistency of the highly adaptive lasso of infinite dimensional parameters. In: Technical report arXiv:1709.06256. Berkeley: Division of Biostatistics, University of California; 2017.

27. Cai W, van der Laan MJ. Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (lasso) estimator. Int J Biostat 2020;16:20170070. https://doi.org/10.1515/ijb-2017-0070.
28. Diaz Munoz I, van der Laan MJ. Super learner based conditional density estimation with application to marginal structural models. Int J Biostat 2011;7:1–20.
29. van der Laan MJ, Wang Z, van der Laan LWP. Higher order targeted maximum likelihood estimation; 2021.
30. Ertefaie A, Hejazi NS, van der Laan MJ. Nonparametric inverse probability weighted estimators based on the highly adaptive lasso; 2020.
31. van der Laan MJ, Benkeser D, Cai W. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso; 2019.