Targeted Highly Adaptive Lasso Minimum Loss Estimation

Mark J. van der Laan[1]
Division of Biostatistics, UC Berkeley
laan@berkeley.edu

December 24, 2019

Abstract

We consider estimation of a functional parameter of a realistically modeled data distribution based on observing independent and identically distributed observations. In van der Laan (2015) we proposed a Highly Adaptive Lasso Minimum Loss Estimator of a functional parameter that is defined by minimizing the empirical risk function over subset of multivariate real valued cadlag functions with finite sectional variation norm. In Bibaut and van der Laan (2019) it has been shown that this HAL-MLE converges to the true functional target parameter w.r.t. loss-based dissimilarity at a rate as fast as $n^{-2/3}$ up till $\log n$-factor. In recent work van der Laan et al. (2019) we showed that an undersmoothed version of this HAL-MLE that uses an L_1-bound large enough so that its fit selects sparsely supported basis functions will be asymptotically efficient for smooth functionals of the functional parameter. Even though the undersmoothed HAL-MLE is asymptotically efficient for any smooth functional, it is not targeted towards a specific target estimand as a TMLE. Therefore, in this article we define a targeted-HAL-MLE (targeting a particular estimand) that minimizes the empirical risk under the additional constraint that the Euclidean norm of the empirical mean of the efficient influence curve of the target estimand equals zero. We show that it preserves the rate of convergence of HAL-MLE that only enforces the sectional variation norm constraint; by being a TMLE it is efficient for the target estimand; and, in addition, that, if it satisfies the undersmoothing criterion, it is still asymptotically efficient for any other smooth functional as well. We demonstrate the practical performance of the T-HAL-MLE relative to HAL-MLE in a simulation study.

Key words: Asymptotically efficient estimator, cadlag, canonical gradient, cross-validation, efficient influence curve, Highly-Adaptive-Lasso MLE, loss-function, pathwise differentiable parameter, risk, sectional variation norm, targeted minimum loss estimation (TMLE), undersmoothing.

1 Introduction

We consider the problem of statistical estimation of a functional parameter of a data distribution for realistic statistical models, based on independent and identically distributed observations. The statistical model is defined as the set of possible data distributions. It is assumed that the statistical model implies a parameter space for the functional parameter that is contained in the set of cadlag functions with bounded sectional variation norm. It is assumed that the functional parameter is defined as the minimizer over the parameter space of the expectation of a loss function at a candidate function. Functions in this parameter space can be represented as infinitesimal linear combinations of tensor products of indicator basis functions (0-level splines), where the L_1-norm of the coefficients (represented by an integral) equals the sectional variation norm of the function (van der Laan, 2015). The HAL-MLE is defined by minimizing the empirical risk over all functions in the parameter space with a sectional variation norm bounded by a specific constant. It is implemented with Lasso type implementations that minimize the empirical risk over a finite linear combination of the indicator basis functions with knot points that can typically be a priori selected based on the observed sample, under the constraint that the L_1-norm of the coefficients is bounded by the specified constant. A natural selector of the L_1-norm is the cross-validation selector, which is asymptotically optimal w.r.t. minimizing the loss-based dissimilarity of estimator with true value of functional parameter. We have shown that this nonparametric HAL-MLE converges in loss-based dissimilarity as fast as $n^{-2/3}$ up till a $(\log n)^d$-factor (Bibaut and van der Laan, 2019). This corresponds with convergence in an L^2-norm at a rate as fast as $n^{-1/3}(\log n)^{d/2}$, and thus faster than the critical rate $n^{-1/4}$ needed to control second order terms in the analysis of plug-in estimators of target estimands based on this HAL-MLE.

One is often also interested in low dimensional target estimand of the functional parameter. The plug-in HAL-MLE of such a target estimand, when selecting the L_1-norm with cross-validation, will generally not be efficient. This is due to the cross-validation selector optimizing bias-variance trade-off for the functional itself instead of for the target parameter. A general theory for undersmoothing sieve based estimators is developed in Shen (1997, 2007),

and a powerful demonstration of such an estimator was earlier presented in (Newey, 2014). We have shown that an undersmoothed HAL-MLE that selects the L_1-norm significantly larger than the cross-validation selector (but still bounded) according to a specified criterion that measures the maximal sparsity of the selected basis functions is asymptotically efficient (Bickel et al., 1997) for any pathwise differentiable target parameter, under weak regularity conditions (van der Laan et al., 2019).

Alternatively, one could use a TMLE that uses the cross-validated HAL-MLE as initial estimator (van der Laan and Rubin, 2006; van der Laan, 2008; van der Laan and Rose, 2011; van der Laan and Gruber, 2015). This so called HAL-TMLE is asymptotically efficient for the target parameter under weak regularity conditions (van der Laan, 2015). This raises the issue of how to compare the undersmoothed HAL-MLE with a HAL-TMLE. Let's consider the case that the target parameter is well supported by the data in the sense that its efficient influence curve is uniformly bounded (by a reasonable value).

One might prefer the undersmoothed HAL-MLE by being asymptotically efficient for all target parameters, while the TMLE needs to be tailored to each specific target parameter, separately, or to a finite collection of target parameters. On the other hand, in many estimation problems the TMLE step relies on an estimator of nuisance parameter, and, if that nuisance parameter is well specified, it results in excellent bias reduction for the target parameter, thereby resulting in potentially superior finite sample performance of the TMLE relative to undersmoothed HAL-MLE. In particular, in double robust estimation problems, knowing this nuisance parameter, as in randomized trials, would make the TMLE asymptotically linear even when it uses an inconsistent initial estimator for the functional parameter. Therefore, in statistical models in which one has strong knowledge about this nuisance parameter, the HAL-TMLE appears to be the preferred estimator relative to an undersmoothed HAL-MLE, while, for models where there is no nuisance parameter or the nuisance parameter is not any easier to estimate, then the undersmoothed HAL-MLE is an interesting alternative. It is worthwhile to mention that a formal way to compare knowledge about a nuisance parameter is to compare the size of the parameter spaces w.r.t. its entropy. However, if the target parameter is weakly supported by the data, then the TMLE update step might be erratic, making the undersmoothed HAL-MLE potentially a more robust estimator.

In this article we want to marry the undersmoothed HAL-MLE with the HAL-TMLE by constructing a so called targeted HAL-MLE (T-HAL-MLE) whose definition is analogue to the HAL-MLE, but it now also enforces a constraint that guarantees solving the efficient influence curve equation for

the target estimand. As a consequence, this T-HAL-MLE can be analyzed as the HAL-TMLE, and is thus asymptotically efficient under weak regularity conditions. Thus, this T-HAL-MLE is able to utilize knowledge about the nuisance parameter, as much as the HAL-TMLE. It is efficient for the target parameter if either the L_1-norm is selected with cross-validation or if is chosen to to satisfy the undersmoothing condition. Moreover, the undersmoothed T-HAL-MLE also preserves the global efficiency across all target parameters of the undersmoothed HAL-MLE. Thus, we can then conclude that the T-HAL-MLE inherits both the properties of the HAL-TMLE as well as the undersmoothed HAL-MLE (by essentially being both a HAL-TMLE as well as an undersmoothed HAL-MLE).

The organization of this article is as follows. In Section 2 we formulate the statistical estimation problem and define the HAL-MLE. In Section 3 we define the targeted HAL-MLE. In Section 4 we present a general algorithm for computing an HAL-MLE and T-HAL-MLE. In Section 5 we establish its rate of convergence w.r.t. loss-based dissimilarity. In Section 6 we establish its asymptotic efficiency for the target parameter, and in Section 7 we show that when it is undersmoothed, then it is also efficient for all other target parameters, just as the undersmoothed HAL-MLE. In Section 8 we show a simulation study comparing the T-HAL-MLE with HAL-MLE. We conclude with a discussion in Section 9.

2 Defining the estimation problem and relevant quantities

2.1 Functional estimation problem

Suppose we observe $O_1, \ldots, O_n \sim_{iid} P_0 \in \mathcal{M}$, where O is a Euclidean random variable of dimension d with support contained in $[0, \tau_o] \subset \mathbb{R}^d$. 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 $Pf \equiv \int f(o) dP(o)$. 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.

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 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 $C^u < \infty$: i.e., 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 (Gill et al., 1995; van der Laan, 2015) is defined by

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

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. 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. We will also denote this parameter space $Q(\mathcal{M})$ with $\mathcal{Q}(C^u)$, and, more generally, $\mathcal{Q}(C)$ for $C < C^u$ is defined as the subset of all functions in $\mathcal{Q}(C^u)$ for which $\| Q \|_v^* < C$.

We assume that

$$\begin{aligned} M_{20} &\equiv \sup_{P \in \mathcal{M}} P_0\{L(Q(P)) - L(Q_0)\}^2 / d_0(Q(P), Q_0) < \infty \qquad (1) \\ M_1 &\equiv \sup_{o, P \in \mathcal{M}} | L(Q(P))(o) | < \infty \\ M_3 &\equiv \sup_{Q \in \mathcal{Q}(C^u)} \| L(Q) \|_v^* < \infty. \end{aligned}$$

The bounds M_1, M_{20} guarantee good behavior of the cross-validation selector (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006, 2007; Polley et al., 2011). The bound M_3 will provide us with the empirical process bounds in the analysis of the HAL-MLE. Thus, we assume $L(\mathcal{Q}(C^u)) = \{L(Q) : Q \in \mathcal{Q}(C^u)\}$ is contained in the class of d-variate real valued cadlag functions on a cube $[0,\tau_o] \subset \mathbb{R}^d_{\geq 0}$ with a sectional variation norm bounded by universal constant $M_3 < \infty$.

Note also that $[0,\tau] = \{0\} \cup (\cup_s (0_s, \tau_s])$ is partitioned in the singleton $\{0\}$, the s-specific $| s |$-dimensional 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. It is also important to note that any cadlag function Q with finite sectional variation norm can be represented as

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

That is, $Q(x)$ is a sum of integrals up till x_s over all the s-specific edges w.r.t. 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 $x \rightarrow \phi_{s,u_s}(x) \equiv I(x_s \geq u_s)$ indexed by knot-point u_s with coefficient $dQ_s(u_s)$:

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

Note that the L_1-norm of the coefficients $dQ_s(u_s)$, across subsets s and knot-points u_s, in this representation is precisely the sectional variation norm $\| Q \|_v^*$.

Let $Q_n^C = \arg\min_{Q \in \mathcal{Q}(C), Q \ll^* \mu_n} P_n L(Q)$ be the HAL-MLE restricted to a sectional variation norm bounded by C and that the support of the s-specific sections $Q_s(du_s)$ is restricted to a finite set of support points of a discrete measure $\mu_{n,s}$, $s \subset \{1,\ldots,k\}$. In (Bibaut and van der Laan, 2019) it is shown that if $M_1 < \infty$, $M_{20} < \infty$, and $M_3 < \infty$, then $d_0(Q_n^C, Q_0) = O_P(n^{-2/3}(\log n)^d)$ as long as $C > C_0^v \equiv \| Q_0 \|_v^*$. Similarly, this will hold at a random selector C_n so that $P(C_n \geq C_0^v) \rightarrow 1$. In (van der Laan and Bibaut, 2017) it is also shown that the HAL-MLE is uniformly consistent under a continuity assumption.

Typically, the unrestricted MLE $\arg\min_{Q \in \mathcal{Q}(C)} P_n L(Q)$ is attained at a discrete measure Q_n^C so that the support restriction $Q \ll^* \mu_n$ is not needed. Either way, we assume that the support of μ_n is chosen rich enough so that it has the same rate of convergence as the non-support restricted HAL-MLE. Note that in this case each Q can be represented as a finite dimensional linear combination of basis functions $\phi_{s,u_{s,j}}$ for a specified set of knot-points $u_{s,j}$ across subsets s and knot points $u_{s,j}$, $j = 1,\ldots$: $Q_\beta = \sum_{s,j} \beta(s,j) \phi_{s,u_{s,j}}$. As a consequence, computation of this HAL-MLE corresponds with minimizing a criterion over a vector β under an L_1-norm constraint. For a range of values of C, $Q_n^C = \hat{Q}^C(P_n)$ represents a collection of candidate estimators indexed by C.

Let $B_n \in \{0,1\}^n$ be a random split of the sample $\{O_i : i = 1,\ldots,n\}$ into a training sample $\{O_i : B_n(i) = 0\}$ and validation sample $\{O_i : B_n(i) = 1\}$. Let P_{n,B_n}^0 and P_{n,B_n}^1 be the empirical probability measures of the training and validation sample, respectively. The cross-validation selector of C is then given by $C_{n,cv} = \arg\min_C E_{B_n} P_{n,B_n}^1 L(\hat{Q}^C(P_{n,B_n}^0))$. By (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006, 2007; Polley et al., 2011) we have that $d_0(Q_n^{C_{n,cv}}, Q_0)$ is asymptotically equivalent with the oracle selector and thus as good as selecting C_0: therefore, it follows that $d_0(Q_n^{C_{n,cv}}, Q_0) = O_P(n^{-2/3}(\log n)^d)$ as well. Let C_n be the actual selector used, chosen to be a random choice in $[C_{n,cv}, C^u]$. In some cases we simply

use $C_n = C_{n,cv}$, but, in the case of undersmoothed HAL-MLE the selector satisfies $C_{n,cv} < C_n \leq C^u$. Since $C_n \geq C_{n,cv}$, we also have $d_0(Q_n^{C_n}, Q_0) = O_P(n^{-2/3}(\log n)^d)$.

m-th order spline HAL-MLE: We have generalized the HAL-MLE to a general m-th order Spline HAL-MLE that assumes that the true function is m-times differentiable with a finite sectional variation norm for the m-th order derivatives (van der Laan et al., 2019). Our representation theorem shows that this m-th order smoothness class is equivalent with assuming that the true function can be represented as an infinitesimal linear combination of $\leq m$-th order tensor product spline basis functions. In particular, this collection of m-th order Spline HAL-MLEs implies a smoothness adaptive Spline HAL-MLE by selecting m with cross-validation. This smoothness adaptive Spline HAL-MLE will achieve the rate of convergence of the m_0-th order Spline HAL-MLE with m_0 the largest m whose smoothness class still contains the true function. For notational convenience and simplicity, we will focus on the above non-smooth HAL-MLE, which corresponds with $m = 0$. However, the results equally apply to the m-th order Spline HAL-MLE if the true function falls in the m-th order smoothness class, and for the smoothness adaptive HAL-MLE (since the cross-validation selector m_n will equal m_0 with probability tending to 1, so that is just requires assuming the results for the m_0-th order Spline HAL-MLE).

Estimation of pathwise diffferentiable target parameters: Let $\Psi : \mathcal{M} \to \mathbb{R}^m$ be an m-dimensional pathwise differentiable target parameter, and let $D^*(P)$ be the canonical gradient of the pathwise derivative at $P \in \mathcal{M}$. We assume that $\Psi(P)$ only depends on P through $Q(P)$ so that we can also denote it with $\Psi : \mathcal{Q}(C^u) \to \mathbb{R}$. Let $G(P)$ be a nuisance parameter so that $D^*(P)$ only depends on P through $Q(P)$ and $G(P)$: $D^*(P) = D^*(Q(P), G(P))$. Let $R_2(P, P_0) \equiv \Psi(P) - \Psi(P_0) + P_0 D^*(P)$ be the exact second order remainder, and we will also denote it with $R_2(Q, G, Q_0, G_0)$ to emphasize that it involves a second order difference between (Q, G) and (Q_0, G_0). The T-HAL-MLE defined in the next section will be targeted towards $\Psi(Q_0)$.

We will also consider the behavior of the undersmoothed T-HAL-MLE when used to obtain plug-in estimators of other pathwise differentiable target parameters that depend on P through $Q(P)$ only. Let $\tilde{\Psi} : \mathcal{M} \to \mathbb{R}$ represent such a parameter, where $\tilde{\Psi}(P)$ will also be denoted with $\tilde{\Psi}(Q)$. Let $\tilde{D}^*(P) = \tilde{D}^*(Q(P), G(P))$ denotes its canonical gradient at P, where for simplicity G will be defined so that it includes the nuisance parameter for both parameters Ψ and $\tilde{\Psi}$. Let $\tilde{R}_2(P, P_0) \equiv \tilde{\Psi}(P) - \tilde{\Psi}(P_0) + P_0 \tilde{D}(P)$ be the exact second order remainder, which will also be denoted with $\tilde{R}_2(Q, G, Q_0, G_0)$.

2.2 An example: treatment specific mean in nonparametric model

In order to demonstrate the results for the T-HAL-MLE throughout this article we will use the following relatively simple example. Let $O = (W, A, Y) \sim P_0$, W covariate vector, $A \in \{0, 1\}$ binary subsequent treatment, and Y a final continuous outcome. Let the statistical model $\mathcal{M}$ for P_0 be nonparametric, beyond $\delta < P_0(A = 1 \mid W) < 1 - \delta$ for some $\delta > 0$, and sectional variation norm assumptions mentioned below. We assume that we observe n i.i.d. copies of O. We assume that $O \in [0, \tau_o] \subset \mathbb{R}^d$ under P_0. Suppose that we are concerned with estimation of the treatment specific mean $E_P E_P(Y \mid A = 1, W)$ (which equals EY_1 in a causal model under causal assumptions). This target estimand depends on the probability distribution Q_w of W and conditional mean $Q(P) \equiv E_P(Y \mid A, W)$. Since $Q_w(P)$ is estimated with the unbiased empirical probability measure of $W_1, \ldots, W_n$ (which is also an NPMLE and thus efficient), the only challenge is the asymptotic efficient plug-in estimation $\Psi(Q_n)$ of $\Psi(Q_0) \equiv E_{P_0} Q_0(1, W)$ treating the expectation over W as given, so that $1/n \sum_{i=1}^n Q_n(1, W_i)$ is an efficient estimator of $E_0 E_0(Y \mid A = 1, W)$. The canonical gradient of Ψ at P is given by $D^*(P) = A/g(1 \mid W)(Y - Q(A, W))$, where $g(1 \mid W) = P(A = 1 \mid W)$. Let $G(W) \equiv \text{Logit} g(1 \mid W)$. We can also denote $D^*(P)$ with $D^*(Q, G)$. The exact second order remainder $R_{20}(Q, G, Q_0, G_0) = \Psi(Q) - \Psi(Q_0) + P_0 D^*(Q, G)$ is given by $R_{20}(Q, G, Q_0, G_0) = E_{P_0}(Q - Q_0)(1, W)(G - G_0)/G(W)$. As loss function for Q we select the squared error loss $L(Q) = (Y - Q(A, W))^2$, while we could use as loss-function for G the log-likelihood loss $L_1(G) = -\{A \log g(1 \mid W) + (1 - A) \log(1 - g(1 \mid W)\}$, where $g(1 \mid W) = \text{expit} G(W)$ with $\text{expit}(x) = 1/(1 + \exp(-x))$. As part of the statistical model $\mathcal{M}$, we will also assume that Q_0 and G_0 are multivariate cadlag functions (on the cubes implied by $[0, \tau_o]$) with a sectional variation norm bounded by a universal constant C^u. Under the implied sup-norm bound on these functional parameters Q_0, G_0, the bounds M_1, M_{20} are bounded for loss-functions $L(Q)$ and $L_1(G)$, respectively. As a consequence, the HAL-MLEs $Q_n^{C^u} = \arg\min_{Q, \|Q\|_v^* < C^u} P_n L(Q)$ and $\arg\min_{G, \|G\|_v^* < C^u} P_n L_1(G)$ converge to Q_0 and G_0 at rate $n^{-2/3} (\log n)^d$ w.r.t. $d_0(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0)$ and $d_{01}(G, G_0) = P_0 L_1(G) - P_0 L_1(G_0)$, respectively. Moreover, it also follows that $M_3 < \infty$. This verifies (1).

Consider as an alternative parameter $E_P E_P(Y \mid A = 0, W)$. Again, due to the empirical mean over W being an unbiased efficient estimator, we only need to be concerned with asymptotic efficient plug-in estimation of $\tilde{\Psi}(Q_n)$ of $\tilde{\Psi}(Q_0) \equiv E_{P_0} Q_0(0, W)$, treating the expectation over W as given. The canonical gradient of $\tilde{\Psi}$ at P is given by $\tilde{D}(P) = (1 - A)/g(0 \mid W)(Y -$

$Q(A,W)$), while the corresponding exact second order remainder is given by $\tilde{R}_{20}(Q,G,Q_0,G_0) = E_{P_0}(Q-Q_0)(0,W)(g-g_0)/g(0 \mid W)$. We already note here that, by using the Cauchy-Schwarz inequality and strong positivity $\delta < g(1 \mid W) < 1-\delta$ for some $\delta > 0$, both $R_{20}(Q,G,Q_0,G_0)$ and $\tilde{R}_{20}(Q,G,Q_0,G_0)$ can be bounded in terms of $\parallel Q - Q_0 \parallel_{P_0} \parallel G - G_0 \parallel_{P_0}$, where $\parallel f \parallel_{P_0} = (P_0 f^2)^{1/2}$.

In this example, we will be constructing a T-HAL-MLE Q_n^*, targeted towards $\Psi(Q_0)$, so that $\Psi(Q_n^*)$ is asymptotically efficient estimator of $\Psi(Q_0)$, and, by selecting the sectional variation norm bound $C_{n,cv} < C_n \leq C^u$ large enough, this same T-HAL-MLE will also result in an efficient estimator $\tilde{\Psi}(Q_n^*)$ of the alternative parameter $\tilde{\Psi}(Q_0)$.

3 The targeted HAL-MLE

Definition of Targeted HAL-MLE: Let G_n be an HAL-MLE of G_0, where we can use cross-validation to select is sectional variation norm bound. The T-HAL-MLE can be defined as the HAL-MLE enforcing as extra constraint $P_n D^*(Q,G_n) = 0$:

$$Q_n^* = \arg \min_{Q \in \mathcal{Q}(C_n), \parallel P_n D^*(Q,G_n) \parallel = 0} P_n L(Q).$$

The constraint $\parallel P_n D^*(Q,G_n) \parallel = 0$ could be weakened by restricting its norm to being smaller than r_n, where $n^{1/2} r_n \to 0$. In our example, we have

$$Q_n^* = \arg \min_{Q, \|Q\|_v^* < C_n, \frac{1}{n}\sum_{i=1}^n A_i/g_n(1|W_i)(Y_i - Q(A_i,W_i)) = 0} \frac{1}{n}\sum_{i=1}^n (Y_i - Q(A_i,W_i))^2,$$

where we know Q can be approximated by $Q_\beta = \sum_{s,j} \beta(s,j)\phi_{s,j}$, where, for each subset s, we choose as knot points $X_j(s)$, $j = 1,\ldots,n$, where $X_j = (A_j, W_j)$. So, with this approximation $\mathcal{Q}_n(C_n) = \{Q_\beta : \parallel \beta \parallel_1 < C_n\}$ of $\mathcal{Q}(C_n)$, we have $Q_n^* = Q_{\beta_n^*}$, where

$$\beta_n^* = \arg \min_{\{\beta : \|\beta\|_1 < C_n, P_n D^*(Q_\beta, g_n) = 0\}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \sum_{s,j} \beta(s,j)\phi_{s,j}(A_i,W_i) \right)^2.$$

Lagrange multiplier formulation of T-HAL-MLE: We could use the Lagrange multiplier formulation to obtain Q_n^* in two steps. For a given $\lambda \geq 0$, consider the λ-specific penalized HAL-MLE

$$Q_{n,\lambda} = \arg \min_{Q \in \mathcal{Q}(C_n)} P_n L(Q) + \lambda \parallel P_n D^*(Q,G_n) \parallel^2 .$$

Given this λ-specific HAL-MLE, we define λ_n so that $\| P_n D^*(Q_{n,\lambda_n}, G_n) \| = 0$. One could weaken the latter requirement by (for example) setting

$$\lambda_n = \min\{\lambda : \| P_n D^*(Q_{n,\lambda}, G_n) \| < n^{-1/2}\sigma_n / \max(10, \log n)\},$$

where $\sigma_n^2 = P_n\{D^*(Q_n, G_n)\}^2$ is an estimate of the variance of efficient influence curve based on a regular cross-validated HAL-MLE Q_n. This cut-off already makes sure that the value $P_n D^*(Q_{n,\lambda_n}, G_n)$ will be small enough to not meaningfully affect finite sample behavior of $\Psi(Q_{n,\lambda_n})$, and the coverage of the corresponding confidence intervals.

If we consider the case that $P_n D^*(Q_{n,\lambda_n}, G_n) = 0$, then this procedure corresponds with globally minimizing $(\lambda, Q) \to P_n L(Q) + \lambda \| P_n D^*(Q, G_n) \|^2$ using the profile method, so that, by the Lagrange multiplier theorem, it also corresponds with the constrained HAL-MLE $Q_n^* = \arg\min_{Q \in \mathcal{Q}(C_n), P_n D^*(Q,G_n)=0} P_n L(Q)$. So $Q_n^* = Q_{n,\lambda_n}$.

Example: The HAL-MLE G_n has been implemented with Lasso logistic regression (glmnet()) using $\phi_{s,j}$ with s-specific knot-points $W_j(s)$, $j = 1, \ldots, n$, where we use cross-validation to select the L_1-norm. Similarly, we can use e can write $Q_\beta = \sum_{s,j} \beta(s,j)\phi_{s,j}$ using as s-specific knot-points $X_j(s)$, $j = 1, \ldots, n$, where $X = (A, W)$. Let a selector $C_{n,cv} \leq C_n$ be given. Let $Q_{n,\lambda} = Q_{\beta_{n,\lambda}}$, where

$$\beta_{n,\lambda} = \arg \min_{\|\beta\|_1 < C_n} P_n L(Q_\beta) + \lambda \{P_n D^*(Q_\beta, G_n)\}^2;$$

$P_n L(Q_\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \sum_{s,j} \beta(s,j)\phi_{s,j}(A_i, W_i))^2$; and
$P_n D^*(Q_\beta, G_n) = \frac{1}{n}\sum_{i=1}^n \frac{A_i}{g_n(1|W_i)}(Y_i - \sum_{s,j} \beta(s,j)\phi_{s,j}(1, W_i))$.

3.1 Risk-based dissimilarity for lagrange-multiplier penalized HAL-MLE

The true counterpart of the empirical risk minimizes by $Q_{n,\lambda}$ is given by $R_\lambda(Q, P_0) = P_0 L(Q) + \lambda \| P_0 D^*(Q, G_0) \|^2$. Let $Q_{0,\lambda} = \arg\min_{Q \in \mathcal{Q}(C)} R_\lambda(Q, P_0)$ be the minimizer of this risk function. Note that $Q_{0,\lambda} = Q_0$ for all λ. Let $d_{0,\lambda}(Q, Q_0) \equiv R_\lambda(Q, P_0) - R_\lambda(Q_{0,\lambda}, P_0)$ be the risk based dissimilarity. Note that

$$d_{0,\lambda}(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0) + \lambda \| P_0 D^*(Q, G_0) \|^2 .$$

Since $-P_0 D^*(Q, G_0) = \Psi(Q) - \Psi(Q_0) - R_2(Q, G_0, Q_0, G_0)$, it follows that the penalty represents a first order approximation of $\| \Psi(Q) - \Psi(Q_0) \|^2$. This demonstrates that the λ-penalized empirical risk minimized by $Q_{n,\lambda}$ is also concerned with minimizing the distance $\Psi(Q) - \Psi(Q_0)$. For example, in problems where the exact second order remainder $R_{20}(Q, G, Q_0, G_0)$ of target

parameter Ψ has the double robustness structure we have $R_{20}(Q, G_0, Q_0, G_0) = 0$ so that $\| P_0 D^*(Q, G_0) \|^2 = \| \Psi(Q) - \Psi(Q_0) \|^2$. Thus, for these problems the risk-based dissimilarity is given by:

$$d_{0,\lambda}(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0) + \lambda \| \Psi(Q) - \Psi(Q_0) \|^2 .$$

In the appendix we show that $\lambda_n = O_P(1)$ for $C_n = C_{n,cv}$ and, if we undersmooth, then one expects $\lambda_n \to_p 0$ since $P_n D^*(Q_n, G_n) = o_P(n^{-1/2})$ already (where $Q_n = Q_{n,\lambda=0}$ using the same C_n).

Example: Notice that in our example, we have $d_{0,\lambda}(Q, Q_0) = P_0(Q - Q_0)^2 + \lambda(\Psi(Q) - \Psi(Q_0))^2$.

4 Algorithm for computing T-HAL-MLE

We consider two types of equivalent goals, each one inspiring its own algorithm. Let the bound C on the sectional variation nom be given. Firstly, for a given λ, we want to obtain the minimizer $Q_{n,\lambda} = \arg\min_{Q, \|Q\|_v^* < C} P_n L(Q) + \lambda \| P_n D^*(Q, G_n) \|^2$. Approximating the parameter space $\mathcal{Q}$ of cadlag functions with $\mathcal{Q}_n = \{Q_\beta = \sum_{s,j} \beta(s,j)\phi_{s,j} : \beta\}$, this corresponds, with finding

$$\beta_{n,\lambda} = \arg \min_{\beta, \|\beta\|_1 < C} f_n(\beta) + \lambda \Phi_n(\beta),$$

where $f_n(\beta) = P_n L(Q_\beta)$ and $\Phi_n(\beta) = \| P_n D^*(Q_\beta, G_n) \|^2$.

In the next subsection we provide an algorithm for minimizing any function of β under the L_1-constraint. After having computed $\beta_{n,\lambda}$ for all λ one determines the λ_n choice for which $\Phi_n(\beta_{n,\lambda_n}) = 0$ (or r_n), and $\beta_n = \beta_{n,\lambda_n}$ and Q_{β_n} is our desired solution (for a given C).

Alternatively, using the same finite dimensional approximation $\mathcal{Q}_n$, our goal is more directly

$$\beta_n = \arg \min_{\beta, \|\beta\|_1 < C, \Phi_n(\beta) = 0} f_n(\beta).$$

Above, we wrote out this equation for our example. In the second subsection we will generalize our algorithm to handle such an additional constraint without using the Lagrange multiplier formulation.

4.1 General algorithm for minimizing a function under L_1-norm constraint

Consider a function $\beta \to f(\beta)$ of a vector β. We consider β as a vector in the Hilbert space $\mathbb{R}^N$ with inner product $\langle x, y \rangle = \sum_j x(j)y(j)$. Let $\| x \| =$

$\sqrt{\langle x, x\rangle}$ be the usual Euclidean norm. We will also use notation xy for the vetor $(x(j)y(j) : j = 1, \ldots, N)$. Suppose that our goal is to minimize $f(\beta)$ over the convex set defined by all vector $\beta \in \mathbb{R}^N$ satisfying $\| \beta \|_1 = \sum_j | \beta(j) | \leq C$. Let $\beta^* = \arg\min_{\beta, \|\beta\|_1 \leq C} f(\beta)$.

Given a β satisfying $\| \beta \|_1 \leq C$, consider a collection of paths $\beta_{\epsilon,h} = (1 + \epsilon h)\beta$, where $\beta_{\epsilon,h}(j) = (1 + \epsilon h(j))\beta(j)$, and $\epsilon \in (-\delta, \delta)$ for small enough $\delta > 0$ so that $1 + \epsilon h(j) > 0$ for all j with $\beta(j) \neq 0$: for example, $\delta = 1/\max_j | h(j) |$. The only restriction on h is that $\langle h, | \beta | \rangle = \sum_j h(j) | \beta(j) | = 0$. We note that any such path satisfies $\| \beta_{\epsilon,h} \|_1 = \| \beta \|_1$ for $\epsilon \in (-\delta, \delta)$. Consider the pathwise derivative

$$\dot{f}_\beta(h) \equiv \frac{d}{d\epsilon} f(\beta_{\epsilon,h}) \Big|_{\epsilon=0},$$

which is a linear real valued operator on $\mathbb{R}^N$. We assume this linear operator is a bounded operator, so that, by the Riesz representation theorem

$$\dot{f}_\beta(h) = \langle D_\beta, h \rangle$$

for a vector D_β. This latter vector is called a gradient of this pathwise derivative. Consider the subspace $T(\beta) \equiv \{x : \langle x, | \beta | \rangle = 0\}$ spanned by all allowed vectors h. Let $D^*_\beta = \Pi(D_\beta \mid T(\beta))$ be the projection of D_β on this sub-Hilbert space. This projection is simply given by

$$\begin{aligned} D^*_\beta &= D_\beta - \frac{\langle D_\beta, | \beta | \rangle}{\langle | \beta |, | \beta | \rangle} | \beta | \\ &= D_\beta - \frac{\sum_j D_\beta(j) | \beta(j) |}{\sum_j \beta(j)^2} | \beta | . \end{aligned}$$

This is called the canonical gradient of the pathwise derivative of f at β.

Example: In our example, we have $f(\beta) = f_n(\beta) + \lambda \Phi_n(\beta)$, where $f_n(\beta) = P_n L(Q_\beta)$. The gradient of $\Phi(\beta) = P_n^2 D^*(Q_\beta, g_n)$ is given by

$$D^\Phi_\beta(s, j) = 2\beta(s, j) P^*_n(Q_\beta, g_n) \frac{1}{n} \sum_{i=1}^n A_i / g_n(1 \mid W_i) \phi_{s,j}(1, W_i). \tag{2}$$

The gradient of $f_n(\beta)$ is given by

$$D^{f_n}_\beta(s, j) = 2\beta(s, j) \frac{1}{n} \sum_{i=1}^n \phi_{s,j}(A_i, W_i)(Y_i - \sum_{s,j} \beta(s, j) \phi_{s,j}(A_i, W_i)). \tag{3}$$

Thus, the gradient of $f(\beta)$ at β is given by $D_\beta = D_\beta^{f_n} + \lambda D_\beta^{\Phi}$. Thus, the canonical gradient D_β^* of $f(\beta)$ at β is defined by

$$D_\beta^*(s,j) = D_\beta(s,j) - \frac{\sum_{s_1,j_1} D_\beta(s_1,j_1) \mid \beta \mid (s_1,j_1)}{\sum_{s_1,j_1} \beta(s_1,j_1)^2} \mid \beta \mid (s,j). \tag{4}$$

Let $D_\beta^{\Phi,*}$ be the canonical gradient for $\Phi(\beta)$ which is obtained with same formula as in (4) but with D_β replaced by D_β^{Φ}. Let $D_\beta^{f_n,*}$ be the canonical gradient of $f_n(\beta)$ which is obtained with same formula as in (4) but with D_β replaced by $D_\beta^{f_n}$. Of course, we could also express $D_\beta^* = D_\beta^{f_n,*} + \lambda D_\beta^{\Phi,*}$.

Expressing the canonical gradient in terms of standard partial derivatives: Let's now relate this canonical gradient of $\dot{f}_\beta$ to the standard vector $df/d\beta = (d/d\beta_1 f(\beta), \ldots, d/d\beta_N f(\beta))$ of partial derivatives of f at β. Firstly, we note that $f(\beta_{\epsilon,h}) = f((1+\epsilon h)\beta) = f(\beta + \epsilon h\beta)$, so that $\frac{d}{d\epsilon} f(\beta_{\epsilon,h}) = \langle \frac{df}{d\beta}, h\beta \rangle$. This can be written as $\sum_j (df/d\beta_j \beta(j)) h(j)$, so that we could select as gradient $D_\beta = \frac{d}{d\beta} f(\beta)\beta$, i..e each partial derivative $\frac{d}{d\beta_j} f(\beta)$ is multiplied by $\beta(j)$. Thus, the canonical gradient D_β^* can be expressed as

$$D_\beta^* = \frac{d}{d\beta} f(\beta)\beta - \frac{\sum_j \frac{d}{d\beta(j)} f(\beta)\beta(j) \mid \beta(j) \mid}{\sum_j \beta(j)^2} \mid \beta \mid .$$

Specifically, the i-th component of D_β^* is defined as

$$D_\beta^*(i) = \frac{d}{d\beta(i)} f(\beta)\beta(i) - \frac{\sum_j \frac{d}{d\beta(j)} f(\beta)\beta(j) \mid \beta(j) \mid}{\sum_j \beta(j)^2} \mid \beta(i) \mid .$$

We note that the vector of partial derivatives is generally easily derived. In a typical application we have that $f(\beta) = g(\sum_j \beta(j)\phi_j)$ for a collection of basis functions ϕ_j and a real valued function $g : \mathbb{R} \to \mathbb{R}$. So in that case, defining δ_i as the unit vector with a 1 at the i-th component,

$$\begin{aligned}
\frac{d}{d\beta(i)} f(\beta) &= \frac{d}{d\epsilon} g\left(\sum_j (\beta(j) + \epsilon\delta_i(j))\phi_j \right)\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon} g\left(\sum_j \beta(j)\phi_j + \epsilon\phi_i \right)\Big|_{\epsilon=0} \\
&= \dot{g}\left(\sum_j \beta(j)\phi_j \right)(\phi_i),
\end{aligned}$$

where $\dot{g}(\sum_j \beta(j)\phi_j)(h)$ is the directional derivative of g at $\sum_j \beta(j)\phi_j$ in the direction h. To conclude, the standard gradient $d/d\beta f(\beta)$ is obtained by applying the directional derivative of g to directions $(\phi_1, \ldots, \phi_N)$.

Example: In our example, we have that the partial derivatives of the empirical risk are given by

$$\frac{d}{d\beta(s,j)} f(\beta) = \frac{1}{n}\sum_{i=1}^{n} 2\phi_{s,j}(A_i, W_i)(Y_i - \sum_{s_1,j_1} \beta(s_1,j_1)\phi_{s_1,j_1}(A_i, W_i)).$$

Steepest gradient algorithm for determining minimum β^*: We can now use the steepest gradient path $\beta_\epsilon^{sgp} \equiv \beta_{\epsilon, D^*_\beta}$ to optimize the function $f(\beta)$ and thereby find β^*. Specifically, we could use the following algorithm.

- Choose a starting value β^0 with $\parallel \beta^0 \parallel = C$.
- Let $\epsilon^0 = \arg\min_\epsilon f(\beta_\epsilon^{0,sgp})$, and define the update $\beta^1 = \beta_{\epsilon^0}^{0,sgp}$.
- Let $k = 1$, $\epsilon^k = \arg\min_\epsilon f(\beta^{k,sgp})$ and $\beta^{k+1} = \beta_{\epsilon^k}^{k,sgp}$, and iterate this till $\epsilon^k \approx 0$.
- Let β^* be the final limit of this iterative algorithm. We have $\frac{d}{d\epsilon} f(\beta_\epsilon^{*,sgp}) = 0$ at $\epsilon = 0$ and thus
$$\dot{f}_{\beta^*}(D^*_{\beta^*}) = \parallel D^*_{\beta^*} \parallel^2 = 0.$$

Specifically, for each i,

$$\frac{d}{d\beta(i)} f(\beta)\beta(i) - \frac{\sum_j \frac{d}{d\beta(j)} f(\beta)\beta(j) \mid \beta(j) \mid}{\sum_j \beta(j)^2} \mid \beta(i) \mid = 0.$$

If the function $f(\beta)$ is convex, then the obtained local minimum will also be a global minimum. In general, good starting values β^0 might be important. We could carry out this algorithm in parallel for a grid of values C and select C based on a valid criterion such as cross-validated risk.

Example: In our example, we have $\beta_\epsilon^{sgp} = (1 + \epsilon D^*_\beta)\beta$, where D^*_β is given by (4). The k-th MLE step for computing ϵ^k in the above algorithm is thus given by

$$\begin{aligned}
\epsilon^k &= \arg\min_\epsilon \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \sum_{s,j} (1 + \epsilon D^*_{\beta^k}(s,j))\beta^k(s,j)\phi_{s,j}(A_i, W_i) \right)^2 \\
&= \arg\min_\epsilon \frac{1}{n}\sum_{i=1}^{n} \left(Z_i - \epsilon \sum_{s,j} D^*_{\beta^k}(s,j)\beta^k(s,j)\phi_{s,j}(A_i, W_i) \right)^2,
\end{aligned}$$

where $Z_i = Y_i - \sum_{s,j} \beta^k(s,j)\phi_{s,j}(A_i, W_i)$. Thus, ϵ^k is obtained with univariate linear least squares regression with single covariate $\sum_{s,j} D^*_{\beta^k}(s,j)\beta^k(s,j)\phi_{s,j}(A,W)$, using as off-set the current Q_{β^k}. However, one should truncate the absolute value of ϵ^k at the δ-value so that for all (s,j), $(1 + \epsilon D^*_{\beta^k}(s,j)) \mid \beta(s,j) \mid \geq 0$.

Exponential paths: The above paths $\beta_{\epsilon,h}$ restrict $\epsilon \in (-\delta, \delta)$ with $\delta = 1/\max_j \mid h(j) \mid$. This means that in the iterative algorithm we can only make small moves. This itself is not a bad thing since it will make us move in the steepest direction mostly (the canonical gradient D^*_β is only the steepest path locally). However, here we consider an alternative class of paths that does not restrict ϵ. Define

$$\beta_{\epsilon,h}(j) = C(\epsilon, h, \beta) \exp(\epsilon h(j))\beta(j),$$

where $C(\epsilon, h, \beta)$ is chosen so that $C(\epsilon, h, \beta) \sum_j \exp(\epsilon h(j)) \mid \beta(j) \mid = C$, and thus

$$C(\epsilon, h, \beta) = \frac{C}{\sum_j \exp(\epsilon h(j)) \mid \beta(j) \mid}.$$

We now have that for all ϵ and any vector $h \parallel \beta_{\epsilon,h} \parallel_1 = C$. Note that at $\epsilon = 0$, we have $C(\epsilon, h, \beta) = 1$ and we also note that $\frac{d}{d\epsilon} C(\epsilon, h, \beta) = 0$ at $\epsilon = 0$ if $\langle h, \mid \beta \mid \rangle = 0$. As with our previous paths we will therefore restrict h to this condition $\langle h, \mid \beta \mid \rangle = 0$. In that case, we know that $\frac{d}{d\epsilon}\beta_{\epsilon,h}$ at $\epsilon = 0$ is equal to the derivative of the previously defined paths. As a consequence, we have that the pathwise derivative $\frac{d}{d\epsilon} f(\beta_{\epsilon,h})$ at $\epsilon = 0$ is identical to the pathwise derivative along our previous paths. Therefore, we can conclude that

$$\frac{d}{d\epsilon} f(\beta_{\epsilon,h}) = \langle D^*_\beta, h \rangle,$$

where the canonical gradient D^*_β is defined above. In other words, the pathwise derivative for these two classes of paths are identical and thereby also the canonical gradients are identical.

We can use the same iterative algorithm defined above with the same canonical gradient. The difference is that the step for determining $\epsilon^k = \arg\min_\epsilon f(\beta^{k,sgp}_\epsilon)$ can now optimize over an unrestricted ϵ. This will mean that we need fewer iterations of this iterative algorithm to achieve convergence, but the ϵ^k-computation step is also slightly more cumbersome due to the normalizing constant $C(\epsilon, h, \beta)$.

4.2 General algorithm for minimizing a function under L_1-norm constraint and equation-constraint

We want to determine $\beta^* = \arg\min_{\beta, \|\beta\|_1 \leq C, \Phi(\beta)=0} f(\beta)$. Let β be given. Use the above approach to determine the canonical gradient D^Φ_β of Φ at β, con-

straining the paths $\| \beta_{\epsilon,h} \|_1 = \| \beta \|_1$ (i.e., $\langle h, | \beta | \rangle = 0$). Let β^{Φ}_{ϵ} be the path with $h = D^{\Phi}_{\beta}$ so that it is tailored for maximally changing the constraint Φ at β. Let D^{f}_{β} be the canonical gradient of f at β constraining the paths to preserve the L_1-norm as above. The canonical gradient D^{*}_{β} of f under the jointly constrained parameter space $\| \beta \|_1 < C$ and $\Phi(\beta) = 0$ is given by $D^{*}_{\beta} = D^{f}_{\beta} - \frac{\langle D^{f}_{\beta}, D^{\Phi}_{\beta} \rangle}{\langle D^{\Phi}_{\beta}, D^{\Phi}_{\beta} \rangle} D^{\Phi}_{\beta}$.

Therefore we could employ the steepest gradient algorithm described above based on the path $\beta^{sgp}_{\epsilon} = \beta_{\epsilon, h = D^{*}_{\beta}}$, starting with a β^0 satisfying $\| \beta^0 \|_1 = C$ and $\Phi(\beta^0) = 0$. In first order, the change in the constraint Φ along a small move $\beta^{sgp}_{d\epsilon}$ along this path is zero, so that the path implied by iteratively moving along β^{sgp}_{ϵ} with infinitesimal small moves $d\epsilon$ would provide the desired minimum β^*. However, in practice, each move would result in a small deviation $\Phi(\beta^{sgp}_{d\epsilon}) \neq 0$. Therefore, we recommend the following iterative algorithm that also builds in the correction so that the constraint remains at zero.

Steepest gradient algorithm for determining minimum β^* under L_1 and zero-constraint: Consider the steepest gradient paths $\beta^{sgp}_{\epsilon} \equiv \beta_{\epsilon, D^{*}_{\beta}}$ for f and $\beta^{\Phi}_{\epsilon} = \beta_{\epsilon, D^{\Phi}_{\beta}}$ for the constraint Φ.

- Choose a starting value β^0 with $\| \beta^0 \| = C$ and $\Phi(\beta^0) = 0$.
- Let $\epsilon^{0a} = \arg\min_{\epsilon} f(\beta^{0,sgp}_{\epsilon})$, and define the update $\beta^{1a} = \beta^{0,sgp}_{\epsilon^{0a}}$. One may restrict the optimization over a small interval $(-\delta, \delta)$ of ϵ-values instead so that the violation of the constraint $\Phi(\beta^{1a}) = 0$ as corrected in next step is small.
- Construct the path $\beta^{1a,\Phi}_{\epsilon}$ through β^{1a}. Determine ϵ^0 so that $\Phi(\beta^{1a,\Phi}_{\epsilon}) = 0$. Let $\beta^1 = \beta^{1,\Phi}_{\epsilon^0}$.
- Let $k = 1$; $\epsilon^{ka} = \arg\min_{\epsilon} f(\beta^{k,sgp}_{\epsilon})$; $\beta^{k+1,a} = \beta^{k,sgp}_{\epsilon^{ka}}$; Construct $\beta^{k+1,a,\Phi}_{\epsilon}$ through $\beta^{k+1,a}$; determine ϵ^k so that $\Phi(\beta^{k+1,a,\Phi}_{\epsilon}) = 0$; set $\beta^{k+1} = \beta^{k+1,a,\Phi}_{\epsilon^k}$.
- Set $k \leftarrow k + 1$ and iterate this in $k = 1, \ldots, K$ till $\epsilon^K \approx 0$.
- Let β^* be the final limit β^K of this iterative algorithm.
- We have $\frac{d}{d\epsilon} f(\beta^{*,sgp}_{\epsilon}) = 0$ at $\epsilon = 0$, and thus

$$\dot{f}_{\beta^*}(D^{*}_{\beta^*}) = \| D^{*}_{\beta^*} \|^2 = 0.$$

Example: In our example, we have that the canonical gradient of $\Phi(\beta) = P^2_n D^*(Q_\beta, g_n)$, under L_1-norm constraint only, is given by $D^{\Phi,*}_{\beta}$ defined above

under (4). In addition, the canonical gradient of $f(\beta) = P_n L(Q_\beta)$, under L_1-norm constraint only, is given by $D_\beta^{f_n,*}$ defined above under (4). Thus, the canonical gradient of $f(\beta)$ under the joint L_1-constraint and $\Phi(\beta) = 0$ is given by

$$D_\beta^* = D_\beta^{f_n,*} - \frac{\langle D_\beta^{f_n,*}, D_\beta^{\Phi,*} \rangle}{\langle D_\beta^{\Phi,*}, D_\beta^{\Phi,*} \rangle} D_\beta^{\Phi,*}.$$

This allows us now to explicitly write down the two paths β_ϵ^{sgp} and β_ϵ^{Φ} corresponding with $\beta_{\epsilon,h} = (1 + \epsilon h)\beta$ with $h = D_\beta^*$ and $D_\beta^{\Phi,*}$, respectively. As we showed in previous subsection, the MLE-steps for $\epsilon^{k,a}$ correspond with univariate linear regression using an off-set. The subsequent correction step of determining ϵ^k based on $\beta_\epsilon^{k,\Phi}$ corresponds with solving $P_n D^*(Q_{\beta_\epsilon^{k,\Phi}}, g_n) = 0$. If we use the linear (in ϵ) paths, then this equation is linear in ϵ so that also this step has a simple closed form solution. One needs to truncate ϵ^k by δ. This then fully describes the above iterative algorithm for optimizing $f_n(\beta)$ under the joint constraint $\| \beta \|_1 < C$ and $\Phi(\beta) = 0$. Similarly, we can describe this algorithm for the exponential paths.

5 Rate of convergence of the targeted HAL-MLE

The following theorem shows that the T-HAL-MLE achieves the same rate of convergence as the regular HAL-MLE.

Theorem 1 *Let $\mathcal{Q}(C) = \{Q : \| Q \|_v^* < C\}$ be the sub-parameter space of $\mathcal{Q} = Q(\mathcal{M})$ consisting of k-variate cadlag functions with uniform sectional variation norm smaller than C. Let $r_n = o(n^{-1/2})$ be a sequence such as $r_n = \sigma_n/(n^{1/2} \log n)$. Let $C_0^v \equiv \| Q_0 \|_v^*$ and $C > C_0^v$.*

Let $\mathcal{Q}^{LFM}(Q_0) \equiv \{Q_{0,\epsilon,G_n} : \epsilon\} \subset \mathcal{Q}(C)$, with $\epsilon \in (-\delta, \delta)$ for some arbitrary small $\delta > 0$, be a local least favorable submodel through Q_0 at $\epsilon = 0$ so that $\frac{d}{d\epsilon} L(Q_{0,\epsilon,G_n}) = D^*(Q_0, G_n)$ at $\epsilon = 0$. Note that this parametric model $\mathcal{Q}^{LFM}(Q_0)$ with parameter ϵ is correctly specified and the true parameter $\epsilon_0 = 0$. Let $\epsilon_n = \arg\min_\epsilon P_n L(Q_{0,\epsilon,G_n})$ be the MLE of ϵ_0, where ϵ may vary over larger set than $(-\delta, \delta)$.

T-HAL-MLE: *Let $C_n(Q) \equiv \| P_n D^*(Q, G_n) \|$, and consider the T-HAL MLE*

$$Q_n^* = \arg \min_{\|Q\|_v^* < C, C_n(Q) \leq r_n} P_n L(Q).$$

Assumptions:

- $M_1 < \infty$; $M_{20} < \infty$ *and* $M_3 < \infty$.
- *Assume regularity conditions, so that the MLE* $\epsilon_n = O_P(n^{-1/2})$, *and thereby* $d_0(Q_{0,\epsilon_n,G_n}, Q_0) = O_P(n^{-1})$, *and* $\| P_n D^*(Q_{0,\epsilon_n,G_n}, G_n) \| \leq r_n$ *with probability tending to 1.*

Conclusion: *We have*

$$d_0(Q_n^*, Q_0) = O_P(n^{-2/3}(\log n)^d).$$

This remains true for data adaptive selectors $C = C_{n,cv}$ *or* $C = C_n$ *with a* $C_{n,cv} \leq C_n < C^u$.

Proof: Let $\mathcal{Q}_n = \{Q :\| Q \|_v^* < C, C_n(Q) \leq r_n\}$, while $\mathcal{Q} = \{Q :\| Q \|_v^* < C\}$. Let $Q_{0,n} = \arg\min_{Q \in \mathcal{Q}_n} P_0 L(Q)$. By usual HAL-MLE proof, using that $P_n L(Q_n^*) \leq P_n L(Q_{0,n})$, it follows that

$$d_0(Q_n^*, Q_{0,n}) \leq -(P_n - P_0)\{L(Q_n^*) - L(Q_{0,n})\}.$$

The typical HAL-MLE proof now proceeds with the following ingredients: 1) $\mathcal{F} = L(\mathcal{Q}(C^u))$ is a Donsker class with bracketing entropy number $\log N_{[]}(\epsilon, L(\mathcal{F}(C^u)), L^2(P)) \lesssim \epsilon^{-1}(\log \epsilon)^{-d}$ (Bibaut, van der Laan, 2019, Proposition 2); 2) $P_0\{L(Q) - L(Q_{0,n})\}^2 < M_2 d_0(Q, Q_{0,n})$ for some $M_2 < \infty$; 3) the bracketing entropy integral can be bounded by $J_{[]}(\delta, \mathcal{F}, L^2(P)) \lesssim \delta^{1/2}(\log \delta)^{-d/2}$; the modulus of discontinuity for the empirical process can be bounded accordingly as

$$\sup_{f, \|f\| < \delta} | G_n(f) | \lesssim J_{[]}(\delta, \mathcal{F}, L^2(P)) \left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L_2(P))}{\delta^2 n^{1/2}} M\right)$$

(van der Vaart and Wellner (2011) and Lemma 3.4.2 in van der Vaart and Wellner (1996)). This allows us to apply the iterative HAL-MLE proof in Appendix of (van der Laan, 2015) or direct proof (Bibaut and van der Laan, 2019) to establish that $d_0(Q_n^*, Q_{0,n}) = O_P(n^{-2/3}(\log n)^d)$. For example, (Bibaut and van der Laan, 2019) gives this result as a corollary.

Therefore, it remains to analyze $d_0(Q_{0,n}, Q_0)$. Let $\{Q_{0,\epsilon,G_n} : \epsilon\} \subset \mathcal{Q}$, with $\epsilon \in (-\delta, \delta)$ for some $\delta > 0$, be a local least favorable submodel through Q_0 at $\epsilon = 0$ so that $\frac{d}{d\epsilon} L(Q_{0,\epsilon,G_n}) = D^*(Q_0, G_n)$ at $\epsilon = 0$. Let $\epsilon_n = \arg\min_\epsilon P_n L(Q_{0,\epsilon,G_n})$ be the MLE, and consider the resulting first-step TMLE Q_{0,ϵ_n,G_n}). By assumption, $\epsilon_n = o_P(1)$, so that $\epsilon_n \in (-\delta, \delta)$ with probability tending to 1. This shows that $Q_{0,\epsilon_n,G_n} \in \mathcal{Q}(C)$ with probability tending to 1. Since Q_{0,ϵ_n,G_n} is a first-step TMLE that uses as initial estimator the true Q_0, under weak regularity conditions, we also have that $P_n D^*(Q_{0,\epsilon_n,G_n}, G_n) = o_P(r_n)$: in particular, if

we use a universal least favorable model (van der Laan and Gruber, 2015), then $P_n D^*(Q_{0,\epsilon_n,G_n}, G_n) = 0$. We conclude that with probability tending to 1, we have $Q_{0,\epsilon_n,G_n} \in \mathcal{Q}_n$ (i.e., it satisfies both the sectional variation norm constraint and $C_n(Q) \leq r_n$ constraint). Thus, we know that with probability tending to 1,

$$d_0(Q_{0n}, Q_0) \leq d_0(Q_{0,\epsilon_n,G_n}, Q_0).$$

Note that ϵ_n is an MLE in a correctly specified parametric model with true value $\epsilon_0 = 0$, so that under weak regularity conditions $\epsilon_n = O_P(n^{-1/2})$ (assumed in theorem). Therefore, this will generally imply $d_0(Q_{0,\epsilon_n,G_n}, Q_0) = O_P(n^{-1})$ (assumed in theorem). This proves that $d_0(Q_n^*, Q_0) = O_P(n^{-2/3}(\log n)^d)$. □

Example: We already verified that M_1, M_2, M_3 are all finite. We can select as least favorable submodel through Q_0 $Q_{0,\epsilon,G} = Q(A,W) + \epsilon A/g(1 \mid W)$. Note that indeed $\frac{d}{d\epsilon} L(Q_{0,\epsilon,G}) = D^*(Q_{0,\epsilon,G}, G)$ for all ϵ. It trivially follows that the least squares estimator $\epsilon_n = O_P(n^{-1/2})$, and thus $d_0(Q_{0,\epsilon_n,G_n}, Q_0) = P_0(Q_{0,\epsilon_n,G_n} - Q_0)^2 = O_P(n^{-1})$. Since this is a universal least favorable submodel it also follows $P_n D^*(Q_{0,\epsilon_n,G_n}, G_n) = 0$ exactly. Finally, assuming that $\| Q_0 \|_v^* < C$, for ϵ_n small enough (i.e., n large enough) we have that $\| Q_{0,\epsilon_n,G_n} \|_v^* < C$, so that is satisfies both constraints $\| Q \|_v^* < C$ and $P_n D^*(Q, G_n) = 0$. This verifies all conditions in Theorem 1 and thus proves $d_0(Q_n^*, Q_0) = O_P(n^{-2/3}(\log n)^d)$ for the T-HAL-MLE Q_n^*.

6 Efficiency of T-HAL-MLE for target parameter

The following theorem establishes the efficiency of the targeted HAL-MLE for the parameter Ψ it targeted. In the definition of the T-HAL-MLE we would select an estimator G_n so that $d_{02}(G_n, G_0) = o_P(n^{-2/3}(\log n)^d)$ such as an HAL-MLE. Assuming a universal bound $\sup_{P \in \mathcal{M}} \| D^*(P) \|_\infty$ (positivity assumption), one will generally be able to bound $R_{20}(Q, G, Q_0, G_0)$ by a constant times $d_0(Q, Q_0)^{1/2} d_{02}(G, G_0)^{1/2}$ so that $R_{2n} \equiv R_{20}(Q_n^*, G_n, Q_0, G_0) = O_P(n^{-2.3}(\log n)^d)$. This thus implies that the main assumption $R_{2n} = o_P(n^{-1/2})$ in the next theorem holds.

Theorem 2 *Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be pathwise differentiable with canonical gradient at P given by $D^*(Q(P), G(P))$. Suppose $\Psi(P) = \Psi(Q(P))$, where $Q_0 = \arg\min_{Q \in \mathcal{Q}} P_0 L(Q)$. For a given G_n, and $r_n = o(n^{-1/2})$, consider the targeted HAL-MLE $Q_n^* = \arg\min_{Q \in \mathcal{Q}(C_n), C_n(Q) \leq r_n} P_n L(Q)$. For example, Q_n^**

is defined by $Q_{n,\lambda} = \arg\min_{Q \in \mathcal{Q}(C)} P_n L(Q) + \lambda \parallel P_n D^*(Q, G_n) \parallel^2$; $Q_n^* = Q_{n,\lambda_n}$ *with* λ_n *so that* $P_n D^*(Q_{n,\lambda_n}, G_n) \leq r_n$ *(e.g, 0).*

Assume the conditions of Theorem 1. Then, $d_0(Q_n^*, Q_0) = O_P(n^{-2/3}(\log n)^d)$. *In addition, assume* $P_0\{D^*(Q_n^*, G_n) - D^*(Q_0, G_0)\}^2 \to_p 0$ *and* $R_{20}(Q_n^*, G_n, Q_0, G_0) = o_P(n^{-1/2})$.

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

As remarked above, the conditions stated beyond the conditions of Theorem 1 generally hold given that we already achieved the rate-consistency for $d_0(Q_n^*, Q_0)$; we already had to assume G_n is a good estimator of G_0; and we bounded the supremum norm of $D^*(Q, G)$ uniformly in (Q, G) so that $R_2()$ can be bounded by Cauchy-Schwarz in terms of L^2-norms.

Example: We already established $d_{01}(G_n, G_0) = O_P(n^{-2/3}(\log n)^d)$ by G_n being an HAL-MLE. Due to the density g_n, g_0 being uniformly bounded away from zero, this also implies $\parallel g_n - g_0 \parallel_{P_0}^2 = O_P(n^{-2/3})$. We have that $R_{20}(Q_n^*, G_n, Q_0, G_0)$ can be bounded by $\parallel Q_n^* - Q_0 \parallel_{P_0} \parallel g_n - g_0 \parallel_{P_0}$, so that the established consistency of Q_n^* and g_n w.r.t. the loss-based dissimilarities implies $R_{20}(Q_n^*, G_n, Q_0, G_0) = O_P(n^{-2/3}(\log n)^d)$. We already verified the conditions of Theorem 1, so that all conditions of Theorem **??** are now verified. This proves the following corollary.

Corollary 1 *Recall the description of the statistical model involving* $\min_{a,w} g_0(a \mid W) > \delta > 0$ *for some* $\delta > 0$, $O \in [0, \tau_o]$ *under* P_0, g_0 *and* Q_0 *being multivariate real valued cadlag functions with a universal bound on their sectional variation norm. Consider the definition of the T-HAL-MLE* $Q_n^* = \arg\min_{Q, \|Q\|_v^* < C_n, \|P_n D^*(Q,G_n)\| \leq r_n} P_n L(Q)$ *with* $r_n = o(n^{-1/2})$ *and* $P(C_0^v \leq C_n \leq C^u) \to 1$ *(e.g.,* $C_n = C_{n,cv}$*). We have* $d_0(Q_n^*, Q_0) = O_P(n^{-2/3}(\log n)^d)$ *and* $\Psi(Q_n^*)$ *is asymptotically efficient estimator of* $\Psi(Q_0)$.

7 Efficiency of targeted HAL-MLE for other parameters

The targeted HAL-MLE $Q_n^* = Q_{n,\lambda_n}$ satisfies

$$Q_n^* = \arg \min_{Q \in \mathcal{Q}(C_n)} P_n L(Q) + \lambda_n \parallel P_n D^*(Q, G_n) \parallel^2,$$

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

In this section we will require that λ_n is chosen so that $\lambda_n \parallel P_n D^*(Q_{n,\lambda_n}, G_n) \parallel = o_P(n^{-1/2})$. As we argued earlier, we expect $\lambda_n = O_P(1)$, or even converge to zero. In particular, we might select λ_n so that $P_n D^*(Q_{n,\lambda_n}, G_n) = 0$.

Consider a path $\{Q_{n,\epsilon}^{*,h} : \epsilon\}$, indexed by a function h, defined by

$$Q_{n,\epsilon}^{*h} = (1 + \epsilon h(0))Q_n^*(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{(0_s, x_s]} (1 + \epsilon h(s, u_s)) dQ_{n,s}^*(u_s), \quad (5)$$

where h has to satisfy the restriction $r_v(h, Q_n^*) = 0$ defined by

$$r_v(h, Q_n^*) \equiv h(0) \mid Q_n^*(0) \mid + \sum_{s \subset \{1,\ldots,d\}} \int_{(0_s, \tau_s]} (1 + \epsilon h(s, u_s)) \mid dQ_{n,s}^*(u_s) \mid .$$

Due to the constraint $r(h, Q_n^*)$, it follows that for any uniformly bounded function h with $r(h, Q_n^*)$, $\{Q_{n,\epsilon}^{*,h} : \epsilon\} \subset \mathcal{Q}(C_n)$. Specifically, the sectional variation norm of $Q_{n,\epsilon}^{*,h}$ does not change as ϵ moves away from zero locally. Consider the score equation for Q_n^* for the penalized risk it minimizes:

$$S_h(Q_n^*) \equiv \frac{d}{d\epsilon} \left\{ P_n L(Q_{n,\epsilon}^{*,h}) + \lambda_n \parallel P_n D^*(Q_{n,\epsilon}^{*,h}, G_n) \parallel^2 \right\}\Big|_{\epsilon=0}.$$

Since Q_n^* minimizes this penalized risk over all $Q \in \mathcal{Q}(C_n)$, we know that $P_n S_h(Q_n^*) = 0$ for all uniformly bounded h with $r(h, Q_n^*) = 0$. Suppose first that λ_n is chosen so that exactly $P_n D^*(Q_n^*, G_n) = 0$. In that case, the $\frac{d}{d\epsilon}$ of the penalty term equals zero, so that

$$S_h(Q_n^*) = \frac{d}{d\epsilon} P_n L(Q_{n,\epsilon}^{*,h}\Big|_{\epsilon=0}.$$

Similarly, if $\lambda_n P_n D^*(Q_n^*, G_n) = o_P(n^{-1/2})$, then it follows that $P_n S_h(Q_n^*) = o_P(n^{-1/2})$, uniformly in all bounded h with $r(h, Q_n^*) = 0$. Thus, we conclude that

$$\begin{aligned} &\tfrac{d}{d\epsilon} \left\{ P_n L(Q_{n,\epsilon}^{*h}) + \lambda_n \parallel P_n D^*(Q_{n,\epsilon}^{*h}, G_n^*) \parallel^2 \right\}\Big|_{\epsilon=0} \\ &= \tfrac{d}{d\epsilon} P_n L(Q_{n,\epsilon}^{*h})\Big|_{\epsilon=0} + o_P(n^{-1/2}) = P_n S_h(Q_n^*) + o_P(n^{-1/2}). \end{aligned}$$

Thus, we have $P_n S_h(Q_n^*) = o_P(n^{-1/2})$ for all bounded h with $r(h, Q_n^*) = 0$, and, if λ_n is chosen so that $P_n D^*(Q_n^*, G_n) = 0$ exactly, then $P_n S_h(Q_n^*) = 0$ for all such h. Therefore, it solves (either exact or up till approximation $o_P(n^{-1/2})$) the same score equations as the regular HAL MLE Q_n, which solves $P_n S_h(Q_n) = 0$ for all bounded h with $r(h, Q_n) = 0$. Let $\mathcal{S}(Q_n^*) = \{S_h(Q_n^*) : h\}$ be the linear span of all these score functions $S_h(Q_n^*)$ indexed by any bounded function h.

Since Q_n^* solves the same score equations as a regular HAL-MLE Q_n we can apply Theorems in (van der Laan et al., 2019) to establish that for large enough C_n, the score equations $P_n S_h(Q_n)$ span the efficient influence curve

equation for a given target parameter and thereby that it can be shown to be asymptotically efficient for that target parameter under usual conditions (as used for TMLE or any other efficient estimator). This results in the next two theorems.

The following theorem provides an undersmoothing condition for the sectional variation norm bound C_n under which the efficient influence curve equation $P_n\tilde{D}(Q_n^*, G_0) = o_P(n^{-1/2})$, so that under the usual conditions it follows also that $\tilde{\Psi}(Q_n^*)$ is asymptotically efficient.

Theorem 3 Definitions: *Consider the targeted HAL MLE* $Q_n^* = Q_{n,\lambda_n} = \arg\min_{Q\in\mathcal{Q}(C_n),Q\ll^*\mu_n} P_nL(Q)+\lambda_n \parallel P_nD^*(Q,G_n)\parallel^2$ *for some selector* C_n *with* $C_{n,cv} \leq C_n \leq C^u < \infty$ *with probability tending to 1, where* $C_{n,cv}$ *is the cross-validation selector of* C*;* λ_n *chosen so that* $\lambda_n \parallel P_nD^*(Q_{n,\lambda_n}, G_n)\parallel = o_P(n^{-1/2})$*. Recall that* $Q_n^* = \sum_{(s,j)\in\mathcal{J}_n(C_n)} \beta_n^*(s,j)\phi_{s,j}$*, where* $\mathcal{J}_n(C_n)$ *provide the indices of the basis functions with non-zero coefficients and* β_n^* *denotes the corresponding coefficients. Here we emphasize that* $\mathcal{J}_n(C_n)$ *is implied by the* L_1*-norm bound* C_n *in definition of* Q_n^**.*

Consider a different (from the Ψ *targeted by* Q_n^**) pathwise differentiable target parameter* $\tilde{\Psi} : Q(\mathcal{M}) \to \mathbb{R}^d$ *with canonical gradient* $\tilde{D}^*(Q(P), G(P))$ *at* $P \in \mathcal{M}$ *and exact second order remainder* $\tilde{R}_{20}((Q,G),(Q_0,G_0)) = \tilde{\Psi}(Q) - \tilde{\Psi}(Q_0) + P_0\tilde{D}^*(Q,G)$*. Let* $\tilde{D}_n^*(Q_n^*, G_0)$ *be an approximation of* $\tilde{D}^*(Q_n^*, G_0)$ *that is contained in* $\mathcal{S}(Q_n^*) = \{S_h(Q_n^*) : h\}$ *(without the restriction* $r(h, Q_n^*) = 0$*). Let* $\tilde{h}^*(Q_n^*, G_0)$ *be the corresponding index so that* $\tilde{D}_n^*(Q_n^*, G_0) = S_{\tilde{h}^*(Q_n^*,G_0)}(Q_n^*)$*.*

Assumptions: *Assume* $\parallel \tilde{h}^*(Q_n^*, G_0)\parallel_\infty = O_P(1)$*, and*

$$\min_{(s,j)\in\mathcal{J}_n(C_n)} \parallel P_n\frac{d}{dQ_n^*}L(Q_n^*)(\phi_{s,j})\parallel = o_P(n^{-1/2}). \tag{6}$$

Conclusion: *Then,*

$$P_n\tilde{D}_n^*(Q_n^*, G_0) = o_P(n^{-1.2}).$$

We can also replace (6) by

$$\min_{(s,j)\in\mathcal{J}_n(C_n)} \parallel P_0\left\{\frac{d}{dQ_n^*}L(Q_n^*)(\phi_{s,j}) - \frac{d}{dQ_0}L(Q_0)(\phi_{s,j})\right\}\parallel = o_P(n^{-1/2}), \tag{7}$$

and, for the choice (s^*, j^*) *that minimizes the latter, we have* $P_0\{\frac{d}{dQ_n^*}L(Q_n^*)(\phi_{s^*,j^*})\}^2 \to_p 0$*.*

Finally, we have

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

if

$$\{\tilde{D}^*(Q,G_0), \tilde{D}^*_n(Q,G_0) : Q \in \mathcal{Q}(C^u)\} \text{ is a } P_0\text{-Donsker class}, \tag{8}$$

(like d-variate cadlag functions with universal bound on sectional variation norm), and

$$P_0\{\tilde{D}^*(Q^*_n,G_0) - \tilde{D}^*_n(Q^*_n,G_0)\}^2 \to_p 0. \tag{9}$$

Thus, if also

$$P_0\{\tilde{D}^*(Q^*_n,G_0) - \tilde{D}^*_n(Q^*_n,G_0)\} = o_P(n^{-1/2}), \tag{10}$$

then we have

$$P_n\tilde{D}^*(Q^*_n,G_0) = o_P(n^{-1/2}).$$

Given that $P_n\tilde{D}^*(Q^*_n,G_0) = o_P(n^{-1/2})$, it follows that

$$\tilde{\Psi}(Q^*_n) - \tilde{\Psi}(Q_0) = (P_n - P_0)\tilde{D}^*(Q^*_n,G_0) + R_{20}(Q^*_n,G_0,Q_0,G_0) + o_P(n^{-1/2}).$$

For so called double robust problems we would have $R_{20}(Q^*_n,G_0,Q_0,G_0) = 0$, and, in general, we can bound this term in terms of $d_0(Q^*_n,Q_0)$, which then finalizes the efficiency proof by applying empirical process theory. Thus, the following theorem follows.

Theorem 4 *Assume conditions 6, 8,9,10 in Theorem 3 so that* $P_n\tilde{D}^*(Q^*_n,G_0) = o_P(n^{-1/2})$.

In addition, assume $M_1, M_{20}, M_3 < \infty$. *We have* $d_0(Q^*_n,Q_0) = O_P(n^{-2/3}(\log n)^d)$. *Assume also* $\tilde{R}_2(Q^*_n,G_0,Q_0,G_0) = o_P(n^{-1/2})$; $\{\tilde{D}^*(Q,G_0) : Q \in \mathcal{Q}(C^u)\}$ *is contained in the class of d-variate cadlag functions on a cube* $[0,\tau_o] \subset \mathbb{R}^d$ *in a Euclidean space and that* $\sup_{Q\in\mathcal{Q}} \parallel \tilde{D}^*(Q,G_0) \parallel^*_v < \infty$; $P_0\{\tilde{D}^*(Q^*_n,G_0) - \tilde{D}^*(Q_0,G_0)\}^2 \to_p 0$.

Then, $\tilde{\Psi}(Q^*_n)$ *is asymptotically efficient.*

7.1 Understanding assumption (10).

Before, we apply this to our running example, we first want to discuss the key condition (10), beyond the undersmoothing condition (6). For this purpose, we reparametrize the paths $Q^{*,h}_{n,\epsilon}$ as follows:

$$Q^{*,f(h,Q^*_n)}_{n,\epsilon}(x) = Q^*_n(x) + \epsilon f(h,Q^*_n)(x),,$$

where

$$f(h,Q^*_n)(x) = h(0)Q^*_n(0) + \sum_{s\subset\{1,\ldots,d\}} \int_{(0_s,x_s]} h(s,u_s)dQ^*_{n,s}(u_s).$$

Therefore, we could also define the class of paths $\{Q^{*,h}_{n,\epsilon} :\| h \|_\infty < \infty\}$ as $\{Q^{*,f}_{n,\epsilon} : f \in \mathcal{F}(Q^*_n)\}$, where the index set is given by $\mathcal{F}(Q^*_n) = \{f(h, Q^*_n) :\| h \|_\infty < \infty\}$. The set $\mathcal{F}(Q^*_n)$ is restricted since it consists of the linear span of $\{\phi_{s,j} : (s,j) \in \mathcal{J}_n(C_n)\}$, that is, the linear span of all basis functions $\phi_{s,u_{s,j}}$ with non-zero coefficient $\beta^*_n(s, u_{s,j})$ in the fit $Q^*_n = \sum_{(s,j)} \beta^*_n(s,j)\phi_{s,j}$. The scores $S_h(Q^*_n)$ are linear in $f(h, Q^*_n)$ and the set of scores $\{S_h(Q^*_n) \parallel h \parallel_\infty < \infty\}$ can be parametrized accordingly as $\{S_f(Q^*_n) : f \in \mathcal{F}(Q^*_n)\}$. We will typically have that $\tilde{D}^*(Q^*_n, G_0) = \frac{d}{d\epsilon} L(Q^{*,f_{0n}}_{n,\epsilon})\big|_{\epsilon=0}$ for a choice $f_{0n} = f_0(Q^*_n, G_0)$, generally not an element of $\mathcal{F}(Q^*_n)$. Let $\mathcal{F}^+(Q^*_n)$ be this richer set so that $f_{0n} \in \mathcal{F}^+(Q^*_n)$ and $\{S_f(Q^*_n) : f \in \mathcal{F}^+(Q^*_n)\}$ is an augmented set of scores satisfying $P_0\{S_f(Q_0) = 0$ for all $f \in \mathcal{F}^*(Q^*_n)$. So let's make this assumption. Then, we can write $\tilde{D}^*(Q^*_n, G_0) = S_{f_{0n}}(Q^*_n)$. We can define $\tilde{D}^*_n(Q^*_n, G_0)$ as the projection of $\tilde{D}^*(Q^*_n, G_0)$ onto the finite dimensional linear span $\{S_f(Q^*_n) : f \in \mathcal{F}(Q^*_n)\}$. Thus, $\tilde{D}^*_n(Q^*_n, G_0) = S_{f_n}(Q^*_n)$ for a $f_n \in \mathcal{F}(Q^*_n)$. Somewhat conservatively, we could define f_n as the projection of f_{0n} onto the finite dimensional space $\{\sum_{(s,j)\in\mathcal{J}_n(C_n)} \alpha(s,j)\phi_{s,j} : \alpha\}$ spanned by the basis functions $\phi_{s,j}$ with a non-zero coefficient $\beta^*_n(s,j)$ in Q^*_n. Let $\parallel f_{0n} - f_n \parallel_0$ be the chosen Hilbert space norm so that $f_n = \arg\min_{f\in\mathcal{F}(Q^*_n)} \parallel f_{0n} - f \parallel_0$. Since the set of basis functions $\{\phi_{s,j} : (s,j) \in \mathcal{J}(C_n)\}$ is rich enough (even when we select $C_n = C_{n,cv}$) to approximate Q_0 w.r.t. $d_0(Q, Q_0)$ at a rate $n^{-2/3}(\log n)^d$, one generally expects that $\parallel f_n - f_{0n} \parallel_0$ will also be $O_P(n^{-1/3}(\log n)^{d/2})$. So

$$\begin{aligned} P_0\{\tilde{D}^*_n(Q^*_n, G_0) - \tilde{D}^*(Q^*_n, G_0)\} &= P_0 S_{f_n - f_{0n}}(Q^*_n) \\ &= P_0\{S_{f_n-f_{0n}}(Q^*_n) - S_{f_n-f_{0n}}(Q_0)\}, \end{aligned}$$

since $P_0 S_f(Q_0) = 0$ for all f. This now proves that the left-hand difference is indeed a second order term that can typically be bounded by $d_0(Q^*_n, Q_0)^{1/2} \parallel f_n - f_{0n} \parallel_0$, so that it will be $O_P(n^{-2/3}(\log n)^d)$.

7.2 Example.

Let $Q^*_n = \sum_{(s,j)\in\mathcal{J}_n(C_n)} \beta^*_n(s,j)\phi_{s,j}$, where $\mathcal{J}_n(C_n)$ is the set of non-zero coefficients in the fit of Q^*_n. We have $\tilde{D}^*(Q^*_n, G_0) = (1-A)f_0(W)(Y - Q^*_n(W))$ with $f_0(W) = 1/g_0(0 \mid W)$, while the linear span of $\{S_h(Q^*_n) : h\}$ equals the linear span of $\phi_{s,j}(A,W)(Y - Q^*_n(A,W))$ with $(s,j) \in \mathcal{J}_n(C_n)$. Therefore, we can define $\tilde{D}^*_n(Q^*_n, G_0)$ as the projection of $\tilde{D}^*(Q^*_n, G_0) = (1-A)f_0(W)(Y - Q^*_n(A,W))$ onto the linear span of $\{\phi_{s,j}(A,W)(Y - Q^*_n) : (s,j) \in \mathcal{J}_n(C_n)\}$ in $L^2(P_0)$. Or, slightly more conservative, we can define $(1-A)f_n(W)$ as the projection of $(1-A)f_0(W)$ onto the linear span of $\{(1-A)\phi_{s,j}(A,W) : (s,j) \in \mathcal{J}_n(C_n)\}$, and $\tilde{D}^*_n(Q^*_n, G_0) = (1-A)f_n(W)(Y - Q^*_n(A,W))$. Thus, we have

then

$$\mid P_0\{\tilde{D}_n^*(Q_n^*, G_0) - \tilde{D}^*(Q_n^*, G_0)\} \mid \leq P_0(1-A)(f_n - f_0)(W)(Q_0 - Q_n^*).$$

We can bound the latter by

$$\parallel f_n - f_0 \parallel_{P_0} \parallel Q_n^* - Q_0 \parallel_{P_0} = O_P(n^{-1/3}(\log n)^{d/2} \parallel f_n - f_0 \parallel_{P_0}).$$

From $d_0(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^d)$ it follows that the linear span of basis functions $(1-A)\phi_{s,j}$, $(s,j) \in \mathcal{J}_n(C_{n,cv})$, selected by the cross-validated HAL-MLE is sufficient to approximate the true $Q_0(0,W)$ at rate $n^{-1/3}(\log n)^{d/2}$. Therefore, one would expect that this same linear span will also allow approximation of $f_0 = 1/g_0(0 \mid W)$ at that rate. This approximation error $(1-A)(f_n - f_0)$ is even smaller due to the set of basis functions $\{\phi_{s,j} : (s,j) \in \mathcal{J}_n(C_n)\}$ for the undersmoothed Q_n^* represents a larger set of basis functions than $\{\phi_{s,j} : (s,j) \in \mathcal{J}_n(C_{n,cv})\}$. Thus, one expects $P_0(f_n - f_0)^2 = O_P(n^{-2/3}(\log n)^d)$. Therefore, in this example, we expect to have

$$\mid P_0\{\tilde{D}_n^*(Q_n^*, G_0) - \tilde{D}^*(Q_n^*, G_0)\} \mid \leq O_P(n^{-2/3}(\log n)^d).$$

Either way, to establish (10) we would only need to assume that $\parallel f_n - f_0 \parallel_{P_0} = O_P(n^{-1/6+\alpha})$ for an arbitrarily small $\alpha > 0$.

Corollary 2 *Recall the description of the statistical model involving* $\min_{a,w} g_0(a \mid W) > \delta > 0$ *for some* $\delta > 0$, $O \in [0, \tau_o]$ *under* P_0, g_0 *and* Q_0 *being multivariate real valued cadlag functions with a universal bound on their sectional variation norm. Consider the definition of the T-HAL-MLE* $Q_n^* = \arg\min_{Q, \|Q\|_v^* < C_n, \|P_n D^*(Q,G_n)\| \leq r_n} P_n L(Q)$ *with* $r_n = o(n^{-1/2})$, *targeted towards* $\Psi(P) = E_0 E_P(Y \mid A=1, W)$, *and* G_n *a log-likelihood based HAL-MLE using cross-validation to select the* L_1*-norm. We have* $Q_n^* = Q_{\beta_n^*} = \sum_{s,j} \beta_n^*(s,j)\phi_{s,j}$ *for the specified collection of basis functions, where*

$$\beta_n^* = \arg \min_{\beta, \|\beta\|_1 < C_n, |P_n D^*(Q_\beta, G_n)| \leq r_n} P_n L(Q_\beta).$$

Let $\mathcal{J}_n(C_n) = \{(s,j) : \beta_n^*(s,j) \neq 0\}$. *Let* $\tilde{\Psi}(P) = E_0 E_P(Y \mid A=0, W)$. *Let* $f_0(W) = 1/g_0(0 \mid W)$. *Let* $(1-A)f_n(W)$ *be the projection of* $(1-A)f_0(W)$ *onto the linear span of* $\{(1-A)\phi_{s,j}(A,W) : (s,j) \in \mathcal{J}_n(C_n)\}$ *in* $L^2(P_0)$, *and define* $\tilde{D}_n^*(Q_n^*, G_0) = (1-A)f_n(W)(Y - Q_n^*(A,W))$. *Recall* $C_0^v = \parallel Q_0 \parallel_v^*$.

Assumptions: *Assume* $P(C_0^v \leq C_n \leq C^u) \to 1$*;* C_n *satisfies the global undersmoothing condition (6);* $\parallel f_n - f_0 \parallel_{P_0} = O_P(n^{-1/6+\alpha})$ *for an arbitrarily small* $\alpha > 0$.

Conclusion: *We have* $d_0(Q_n^*, Q_0) = O_P(n^{-2/3}(\log n)^d)$; $\Psi(Q_n^*)$ *is asymptotically efficient estimator of* $\Psi(Q_0)$; *and* $\tilde{\Psi}(Q_n^*)$ *is also asymptotically efficient estimator of* $\tilde{\Psi}(Q_0)$.

8 Simulation study

9 Discussion

We defined C-specific targeted HAL-MLE as a minimizer of an empirical risk over a class of functions with bounded sectional variation norm bounded by C, and under the additional constraint that the Euclidean norm of an empirical mean of (target parameter specific) efficient influence curve equation equals zero. By selecting the sectional variation norm bound C with cross-validation selector $C_{n,cv}$, the plug-in estimator of the target parameter based on the T-HAL-MLE is asymptotically efficient. By selecting $C > C_{n,cv}$ so that our global undersmoothing criterion (6) is satisfied, the plug-in T-HAL-MLE will also be asymptotically efficient for a large class of other pathwise differentiable target parameters: i.e., it will be a globally efficient plug-in estimator. The T-HAL-MLE allows the user to employ targeting of a user supplied set of target parameters in which one believes that additional knowledge (e.g., about nuisance parameter, or strong positivity is known to hold) makes the targeting step particularly robust and powerful, while still preserving global efficiency for any other pathwise differentiable target estimand. Our results apply to an undersmoothed higher order spline HAL-MLE as well, and when one selects the order of the spline based on cross-validation. In this manner, the T-HAL-MLE is adaptive to underlying smoothness, guaranteed to converge to true Q_0 at rate $n^{-1/3}(\log n)^{d/2}$), while being a globally efficient plug-in estimator that is targeted towards a user supplied set of target estimands.

References

A. Bibaut and M.J. van der Laan. Fast rates for empirical risk minimization over cadlag functions with bounded sectional variation norm. Technical report, Division of Biostatistics, University of California, Berkeley, 2019.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, Berlin Heidelberg New York, 1997.

R.D. Gill, M.J. van der Laan, and J.A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré*, 31(3):545–597, 1995.

W. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 2014.

E.C. Polley, S. Rose, and M.J. van der Laan. Super Learner. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2011.

X. Shen. On methods of sieves and penalization. *Annals of Statitics*, 252(6): 2555–2591, 1997.

X. Shen. Large sample sieve estimation of semiparametric models. *Chapter in Handbook of Econometrics*, 76(00):0000, 2007.

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

M.J. van der Laan. A generally efficient targeted minimum loss-based estimator. Technical Report 300, UC Berkeley, 2015. http://biostats.bepress.com/ucbbiostat/paper343, to appear in IJB, 2017.

M.J. van der Laan and A. Bibaut. Uniform consistency of the highly adaptive lasso of infinite dimensional parameters. Technical Report arXiv:1709.06256, Division of Biostatistics, University of California, Berkeley, 2017.

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

M.J. van der Laan and S. Gruber. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *to appear in International Journal of Biostatistics*, 2015.

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

M.J. van der Laan and Daniel B. Rubin. Targeted maximum likelihood learning. *Int J Biostat*, 2(1):Article 11, 2006.

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

M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Stat Appl Genet Mol*, 6(1):Article 25, 2007.

M.J. van der Laan, D. Benkeser, and W. Cai. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Technical report, Division of Biostatistics, University of California, Berkeley, 2019.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, Berlin Heidelberg New York, 1996.

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

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

Appendix

A Behavior of λ_n.

Given that we now understand the risk-based dissimilarity $d_{0,\lambda}(Q, Q_0)$ the λ-specific T-HAL-MLE is aiming to minimize, it is of interest to understand the behavior of the weight λ_n the dissimilarity assigns to the Euclidean norm of $\Psi(Q) - \Psi(Q_0)$ (or first-order approximation). Recall that $Q_n = Q_{n,\lambda=0}$ is the regular HAL-MLE which corresponds with $\lambda = 0$. Our theorem 1 shows that $d_0(Q_{n,\lambda_n}, Q_0) = O_P(n^{-2/3}(\log n)^d)$, just as $d_0(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^d)$. This generally also shows that, w.r.t. a norm implied by the loss-based dissimilarity, $\| Q_{n,\lambda_n} - Q_n \| = O_P(n^{-1/3}(\log n)^{d/2})$. Recall the class of score equations $U_h(Q_{n,\lambda}, \lambda) = S_{1h}(Q_{n,\lambda}) + \lambda \| P_n D^*(Q_{n,\lambda}, G_n) \| f_h(Q_{n,\lambda})$ corresponding with a class of paths through $Q_{n,\lambda}$ indexed by direction $h \in \mathcal{H}$, where

$$\begin{array}{l} S_{1h}(Q_{n,\lambda}) = \frac{d}{d\epsilon} P_n L(Q^h_{n,\lambda,\epsilon})\big|_{\epsilon=0} \\ f_h(Q_{n,\lambda}) = 2 \frac{d}{d\epsilon} \| P_n D^*(Q^{*,h}_{n,\epsilon}, G_n) \| \big|_{\epsilon=0}. \end{array}$$

Note that $U(Q_{n,\lambda}, \lambda) = 0$ and $U(Q_n, 0) = 0$, where $U = (U_h : h \in \mathcal{H})$. We can write $U(Q_n, \lambda) - U(Q_n, 0) = -\{U(Q_{n,\lambda}, \lambda) - U(Q_n, \lambda)\}$. The left-hand side equals equals $\lambda \| P_n D^*(Q_n, G_n) \| f_h(Q_n)$. Taking the sup-norm over h on both sides yields $\lambda \| P_n D^*(Q_n, G_n) \| \sup_h \mid f_h \mid = \sup_h \mid U_h(Q_{n,\lambda}, \lambda) -$

$U_h(Q_n, \lambda)$ |. The right-hand side represents a norm $\| Q_{n,\lambda} - Q_n \|_{1,\lambda}$, so that this shows that $\lambda_n \| P_n D^*(Q_n, G_n) \| = O(\| Q_{n,\lambda_n} - Q_n \|_{1,\lambda})$. One expects that $\| Q_{n,\lambda_n} - Q_n \|_{1,\lambda_n} = O_P(n^{-1/3}(\log n)^d)$ So we have that $\lambda_n \| P_n D^*(Q_n, G_n) \| = O_P(n^{-1/3}(\log n)^{d/2})$.

It is straightforward to show that $P_n D^*(Q_n, G_n)$ behaves as $\Psi(Q_n) - \Psi(Q_0)$ in first order. If the bound C_n is selected with cross-validation, the latter behaves as $d_0^{1/2}(Q_n, Q_0)$. In that case, $\lambda_n = O_P(1)$.

Now we note that if the non-penalized Q_n already solves $P_n D^*(Q_n, G_n) = o_P(n^{-1/2})$, then little penalization should be needed (or none if one is not aiming to solve the constraint exactly), so that the resulting λ_n will be significantly smaller. For example, if C_n undersmooths so that $P_n D^*(Q_n, G_n) = o_P(n^{-1/2})$ is already close to zero, then we expect λ_n to even converge to zero. To conclude, we suggest that $\lambda_n = O_P(1)$ in general, and, under undersmoothing λ_n will approximate zero as sample size increases.