# Causal Inference based on Undersmoothing the Highly Adaptive Lasso

[1] **Mark J. van der Laan,** [2]**David Benkeser,** [1]**Wilson Cai**
[1]Division of Biostatistics, UC Berkeley
2121 Berkeley Way, Room 5302
Berkeley, CA 94720-7360
[2]Division of Biostatistics, Emory University
1518 Clifton Road, N.E.
Atlanta , GA 30322

### Abstract

Nonparametric structural causal models provide statistical models for the data generating distribution, and allow the formal definition of causal impact of an intervention on an outcome of interest. Formal identification results establish non-testable assumptions that allows one to identify the causal quantity of interest as an estimand of the data distribution. Once we accept this estimand as a best or perfect approximation of the causal quantity of interest, we are left with a pure statistical estimation problem of learning the estimand based on knowing that the true data distribution falls in a specified infinite dimensional statistical model. Efficiency theory teaches us that the estimation of the data distribution or its relevant part requires machine learning at a rate of convergence faster than $n^{-1/4}$, combined with targeting the estimator so that it solves the critical efficient score equation for the target estimand. In this paper we discuss the previously introduced Highly Adaptive Lasso Minimum Loss Estimator (HAL-MLE) of the data distribution, which corresponds with minimizing an empirical risk over a linear span of tensor product of 0-order spline basis functions. It has been shown to converge at a rate faster than $n^{-1/4}$ as long as the true function is cadlag and has finite sectional variation norm. In this short paper we demonstrate that by selecting a large enough $L_1$-norm of the vector of coefficients associated with the collection of basis functions, the estimator preserves its $n^{-1/4}$-rate of convergence, while solving the efficient score equation for any desired pathwise differentiable target feature of the data distribution. As a consequence, an undersmoothed HAL-MLE results in an efficient plug-in estimator of the desired estimand, and moreoever, it will also be efficient for any other smooth estimand of the data distribution. We demonstrate this undersmoothed HAL-MLE for estimation of the average treatment effect.

**Key words:** Asymptotically efficient estimator, cadlag, canonical gradient, causal inference, cross-validation, efficient influence curve, Highly-Adaptive-Lasso MLE, loss-function, pathwise differentiable parameter, risk, sectional variation norm, structural causal model, undersmoothing.



## Introduction

Drawing causal inference from observational and experimental data requires a number of careful steps Petersen and van der Laan (2014); van der Laan and Rose (2011); Pearl (2009). To start with one defines a causal model, such as a structural causal model, that consists of a collection of functions that describe how endogenous nodes are generated as a function of parent endogenous nodes and exogenous errors Pearl (2009). These functions are realistically modeled, so that such structural causal models typically mostly incorporates time-ordering and exclusion restriction assumptions, while minimizing any other type of assumptions. Some of the endogenous nodes represent *intervention nodes*, while one of the final endogenous nodes represents the outcome of interest. An intervention-specific counterfactual outcome is then defined by replacing the intervention node equations by the desired intervention one wants to study. This modified system of structural equations now describes a post-intervention distribution for the endogenous nodes, and, in particular, for the outcome. One can now define a causal quantity of interest as the mean outcome under this post-intervention distribution, or a contrast involving two different interventions (e.g., an active and control treatment). One then establishes a link between the endogenous nodes and the observed data, and thereby a link between the observed-data distribution and the full-data distribution described by the structural causal model. In particular, the structural causal model implies the statistical model, the set of possible observed-data distributions. One then addresses whether the causal quantity can be described as a function of the observed-data distribution. This process of causal identification generally involves non-testable assumptions, such as conditional independence assumptions on the exogenous errors. If the assumptions are deemed plausible, then identifiability result provides a causal interpretation of an estimable parameter of the observed-data distribution.

The statistical estimation problem is now defined in terms of the statistical model and estimand, and stands apart from the previous causal modeling steps. At this point, we can draw on statistical theory to guide the construction of efficient estimators of the estimand of interest. This paper focuses on a general technique for efficient estimation in a realistic statistical model.

Due to the curse of dimensionality Donoho (2000), maximum likelihood estimation is often ill defined, and a regularized MLE, for example, one based on a sieve, is generally too biased for the estimand, and therefore fails to produce statistical behavior typical of MLEs, namely asymptotic linearity and normality. Global-bias-reduced regularized MLE sometimes provides a path forward (e.g., undersmoothing as in Newey (2014); van der Vaart (1998); van der Laan (2006)), but the success depends on the particular sieve, and choices of key tuning parameters control the level of smoothing. General guidelines on how to undersmooth are often not available, which makes these techniques difficult to utilize in practice.

In response to this curse of dimensionality issue with the regularized MLE, the current literature has proposed three general methods for construction of efficient estimators. Each of these methods rely on the estimand being a pathwise differentiable functional of the data distribution, whose derivative is identified by the so-called canonical gradient. The canonical gradient represents the score of the submodel through the data distribution in which the target estimands locally changes maximally Bickel et al. (1997). By the convolution theorem, a regular estimator is efficient if and only if it is asymptotically linear with influence curve equal to the canonical gradient. The estimation error of such estimators behaves approximately (i.e., in first-order) as the empirical mean of the canonical gradient at the true data distribution.

Each of these three methods uses the canonical gradient as principal ingredient to target the estimator towards the estimand. The one-step estimator adds to an initial plugin estimator of the estimand the canonical gradient at the initial estimator Bickel et al. (1997). The estimating equation-based framework assume the canonical gradient can be expressed as an estimating function in the target estimand, possibly indexed by nuisance parameters. The estimator is the solution of the resulting estimating equation Robins and Rotnitzky (1992); van der Laan and Robins (2003). A targeted minimum loss estimation updates an initial estimator of the data distribution with a minimum loss estimator of the least favorable parametric submodel through the initial estimator, and estimates the estimand with the corresponding plug-in estimator van der Laan and Rubin (2006); van der Laan (2008); van der Laan and Gruber (2015).

Each estimator requires estimation of key nuisance parameters. For example, in causal inference, these are often objects like the conditional mean of the outcome given intervention nodes and other endogenous nodes. The three frameworks above accommodate the use of state-of-the-art machine learning techniques. For example, targeted minimum loss estimators are often studied in combination with a cross-validation-based ensemble machine learning technique termed super learning. Super learning builds an ensemble of candidate machine learning techniques. Important oracle properties have been established that demonstrate conditions whereby the ensemble converges essentially at the same rate as the theoretically optimal ensemble van der Laan and Dudoit (2003); van der Vaart et al. (2006); van der Laan et al. (2006). Super learning has been shown to perform well in a variety of settings van der Laan et al. (2007); Polley et al. (2011).

Recently, van der Laan (2015) and Benkeser and van der Laan (2016)van der Laan (2015); Benkeser and van der Laan (2016) introduced the Highly Adaptive Lasso MLE (HAL-MLE) machine learning algorithm. This technique yields learners whose error converges to the optimal error at rate faster than $n^{-1/4}$ under minimal conditions. Thus, including this learning as a candidate in a super learner guarantees this rate for the super learner as well. As a consequence, utilizing such learners to estimate key nuisance parameters in one of the three estimating frameworks described above yields efficient estimators of the estimand in great generalityvan der Laan (2015).

In this work, we revisit the undersmoothing paradigm in the contet of HAL-MLE. We argue that using a properly undersmoothed HAL-MLE of the data distribution (or the relevant nuisance parameters of the data distribution) results in a generally efficient plug-in estimator for pathwise differentiable target estimands. Here, we mainly discuss the formal results, provide intuition for the proof; the formal (quite involved) mathematical proofs will be presented elsewhere. We present an application to estimating the treatment-specific mean, a canonical problem in causal inference. We demonstrate through a simulation that indeed the theory works out as predicted. Due to this contribution, we can conclude that this undersmoothed HAL-MLE provides a fourth general method for constructing efficient estimators, beyond the three general targeted methods presented above.

## Defining the HAL-MLE

### 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 $k$ with support $\mathcal{O}$ contained in $[0, \tau_o] \subset \mathbb{R}^k$. Let $Q : \mathcal{M} \rightarrow Q(\mathcal{M}) = \{Q(P) : P \in \mathcal{M}\}$ be a functional parameter of the data distribution. 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 $Q \rightarrow PL(Q)$ 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, which for most loss functions behaves as a square of an $L^2(P)$-type norm (e.g., Kullback-Leibler divergence for the log-likelihood loss). 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 $M_1 \equiv \sup_{o \in \mathcal{O}, P \in \mathcal{M}} \mid L(Q(P))(o) \mid < \infty$. These latter two assumptions are sufficient to guarantee good theoretical behavior of cross-validation-based estimator selection. In particular, these assumptions provide conditions whereby the a cross-validation-selected estimator is asymptotically equivalent with an oracle selector (see above super-learner references).

**Parameter space for functional parameter $Q$: Cadlag and uniform bound on sectional variation norm.** We assume that the parameter space $Q(\mathcal{M}) = \{Q(P) : P \in \mathcal{M}\}$

is a collection of multivariate real-valued cadlag functions on a cube $[0,\tau] \subset \mathbb{R}^k$ with finite sectional variation norm $\parallel Q(P) \parallel_v^* < C^u$ for some $C^u < \infty$ Gill et al. (1995); van der Laan (2006, 2015): 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 $\parallel Q(P) \parallel_v^* < C^u$, where the sectional variation norm is defined by

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

For a given subset $s \subset \{1,\ldots,k\}$, $Q_s : (0_s, \tau_s] \rightarrow \mathbb{R}$ 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 $x_s$.

Note also that $[0,\tau] = \{0\} \cup (\cup_s (0_s, \tau_s])$ 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. 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 to $x_s$ over the $s$-specific edges with respect to the measure generated by the corresponding $s$-specific section $Q_s$. Thus, we refer to $Q_s$ both as a cadlag function and 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 in this representation is precisely the sectional variation norm $\parallel Q \parallel_v^*$.

### Definition of the HAL-MLE

Let $\mathcal{Q}(C^u) = \{Q \in D[0,\tau] : \parallel Q \parallel_v^* < C^u\}$ be the class of cadlag functions with sectional variation norm bounded by $C^u$, which is thus the parameter space for $Q$. Let $C_0 \equiv \parallel Q_0 \parallel_v^*$ be the sectional variation norm of the true $Q_0$, and let $C^u$ be an upper bound guaranteeing that $C_0 < C^u$. For a data adaptive selector $C_n$, we define the HAL-MLE as

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

We will restrict the minimization to $Q$ for which for all subsets $s \subset \{1,\ldots,k\}$, $dQ_s(u_s)$ is a discrete measure with a finite support $\{z_{s,j} : j = 1,\ldots,n_s\}$, where this support is chosen fine enough so that its resulting bias is negligible. Typically, one can actually prove that the unrestricted HAL-MLE (1) is attained at a discrete $Q_n$. Generally, if $O$ includes observing $X$ where $L(Q)(O)$ depends on $Q$ through $Q(X)$, we recommend to select the support of $dQ_s$ as a subset (or whole set) of the observed data $X_i(s)$, $i = 1,\ldots,n$. The above representation for functions in $D[0,\tau]$ shows that all such discrete $Q$ are represented by a finite dimensional linear combination of basis functions indexed subset $s$ and knotpoint $z_{s,j}$. Therefore, in this case the HAL MLE can be represented as $Q_n = \sum_{s,j \in \mathcal{J}_n(s)} \beta_n(s,j)\phi_{s,j}$, where

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

and $\mathcal{J}_n(s)$ is the collection of support points of the $s$-specific section $Q_{n,s}$ of $Q_n$.

The data adaptive selector $C_n$ defining the $L_1$-norm restriction will be selected larger or equal than the cross-validation selector

$$C_{n,cv} = \arg \min_C \frac{1}{V} \sum_{v=1}^{V} P^1_{n,v} L(\hat{Q}_C(P_{n,v})) ,$$

where $P^1_{n,v}, P_{n,v}$ are the empirical distributions of the validation and training sample, respectively, corresponding with the $v$-th sample split in a typical $V$-fold cross-validation scheme. Here $\hat{Q}_C(P_{n,v})$ is the HAL-MLE applied to the training sample corresponding with the $v$-th sample split. For any selector $C_n \leq C^u < \infty$ for which $P(C_n > C_0) \rightarrow 1$, we have that $d_0(Q_n, Q_0) = o_P(n^{-1/2-\alpha(k)})$ for $\alpha(k) = 1/(2(k+2))$ van der Laan (2015). In particular, we have this rate of convergence for the cross-validation selector, which is optimal for estimation of $Q_0$ as a whole.

## Efficient estimation with the undersmoothed HAL-MLE

Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ represent the statistical target parameter of interest, so that $\Psi(P_0)$ is the estimand we aim to learn. We assume that $\Psi$ is pathwise differentiable at $P \in \mathcal{M}$ in the sense that $\frac{d}{d\epsilon}\Psi(P_\epsilon)\big|_{\epsilon=0} = PD(P)S$ for a rich collection of submodels $\{P_\epsilon : \epsilon\}$ through $P$ at $\epsilon = 0$ with score $S$. If the gradient $D(P)(O)$ is chosen to be a score itself (or an arbitrarily fine approximation of a score), then it is called the canonical gradient, which we denote by $D^*(P)$. As above, let $Q : \mathcal{M} \rightarrow Q(\mathcal{M}) = \{Q(P) : P \in \mathcal{M}\}$ be a functional parameter such that $\Psi(P) = \Psi_1(Q(P))$ for some $\Psi_1$: we will abuse notation, and simply use $\Psi(Q)$ and $\Psi(P)$ interchangeably. Let $G : \mathcal{M} \rightarrow \mathcal{G}$ be a functional nuisance parameter so that the canonical gradient $D^*(P)$ only depends on $P$ through $(Q(P), G(P))$. Let $R_2(P, P_0) = \Psi(P) - \Psi(P_0) + P_0 D^*(P)$ be the exact second-order remainder for the target parameter expansion. This remainder $R_2(P, P_0)$ only involves differences between $(Q, G)$ and $(Q_0, G_0)$ so that we will use notation $D^*(P) = D^*(Q(P), G(P))$ and $R_2(P, P_0) = R_2(Q, G, Q_0, G_0)$.

Consider that for a plug-in estimator $\Psi(Q_n)$ of $\Psi(Q_0)$,

$$\begin{aligned}&\Psi(Q_n)-\Psi(Q_0)\\&\quad=(P_n-P_0)D^*(Q_n,G_0)-P_nD^*(Q_n,G_0)\\&\qquad\qquad+R_2(Q_n,G_0,Q_0,G_0).\end{aligned}$$

Assuming that $\{D^*(Q,G):Q,G\}$ falls in a class of cadlag functions with a universal bound on the sectional variation norm (which is, importantly, a Donsker class), using empirical process theory we can establish a simple $L^2(P_0)$-consistency $P_0\{D^*(Q_n,G_0)-D^*(Q_0,G_0)\}^2\to_p 0$ implies $(P_n-P_0)D^*(Q_n,G_0)=(P_n-P_0)D^*(Q_0,G_0)+o_P(n^{-1/2})$ van der Vaart and Wellner (1996). In addition, the above stated convergence $d_0(Q_n,Q_0)=o_P(n^{-1/2})$ will generally imply (under a strong positivity assumption) that $R_2(Q_n,G_0,Q_0,G_0)=o_P(n^{-1/2})$. In so-called double-robust causal inference or censored data problems the second-order remainder only involves cross-terms like $(Q_n-Q_0)(G_n-G_0)$ so that we even have $R_2(Q_n,G_0,Q_0,G_0)=0$ van der Laan and Robins (2003). Thus,

$$\begin{aligned}&\Psi(Q_n)-\Psi(Q_0)\\&\quad=P_nD^*(Q_0,G_0)-P_nD^*(Q_n,G_0)+o_P(n^{-1/2}).\end{aligned}$$

The only remaining obstacle in proving efficiency of the HAL-MLE is that we need $P_nD^*(Q_n,G_0)=o_P(n^{-1/2})$. We can show that this can be proven under two fundamental conditions: 1) the loss function $L(Q)$ must generate the canonical gradient as a score; 2) $C_n$ must be selected "large enough". We now discuss these two conditions.

**Canonical gradient of target parameter in tangent space of loss function:** We assume that the 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 $h$,through $Q$ at $\epsilon=0$, so that for any $G\in\mathcal{G}$, one of these $h$-specific submodels generates a score that equals the canonical gradient $D^*(Q,G)$ at $(Q,G)$:

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

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 $Q$ so that $\Psi(P)$ depends on $P$ only through $Q$ there are typically more direct loss functions $L(Q)$, so that the loss-based dissimilarity $d_0(Q,Q_0)=P_0L(Q)-P_0L(Q_0)$ directly measures a dissimilarity between $Q$ and $Q_0$, for which this condition holds as well.

**Choosing $C_n$ large enough:** A key property of an MLE such as the HAL-MLE is that it solves many score equations of the form $0=P_nS_h(Q_n)$, where $S_h(Q_n)=\frac{d}{d\epsilon}L(Q^h_{n,\epsilon})\big|_{\epsilon=0}=0$, generated by paths $\{Q^h_{n,\epsilon}:\epsilon\}$, such that

$$\begin{aligned}&Q^h_{n,\epsilon}(x)=Q(0)(1+\epsilon h(0))\\&\quad+\sum_{s\subset\{1,\ldots,k\}}\int\phi_{s,u_s}(x)(1+\epsilon h(s,u_s))dQ_s(u_s),\end{aligned}\tag{2}$$

where $h$ is any uniformly bounded function such that $\parallel Q^h_{n,\epsilon}\parallel^*_v=\parallel Q_n\parallel^*_v$ for a small enough neighborhood $\epsilon\in(-\delta,\delta)$. The latter constraint translates into a linear constraint $r(h,Q_n)=0$, where

$$\begin{aligned}&r(h,Q_n)=\mid Q(0)\mid h(0)\\&\quad+\sum_{s\subset\{1,\ldots,k\}}\int\phi_{s,u_s}(x)h(s,u_s)\mid dQ_s(u_s)\mid.\end{aligned}\tag{3}$$

The canonical gradient $D^*(Q_n,G_0)$ is well-approximated by one of these $h$-specific scores $S_h(Q_n)$, but not necessarily by one that satisfies this linear constraint $r(h,Q_n)=0$. As the dimension of the fit of the HAL-MLE $Q_n$ grows, i.e., as more basis functions have a non-zero coefficient, so too does the dimension of the linear space spanned by these score equations $\{P_nS_h(Q_n):h,r_n(h,Q_n)=0\}$. At some large-enough dimension, this linear span of score equations, in spite of the constraint, will be rich enough so as to approximately solve the efficient score equation up to a term that is $o_P(n^{-1/2})$. Indeed, we can formally prove that the main condition for $P_nD^*(Q_n,G_0)=o_P(n^{-1/2})$ is

$$\min_{s,j\in\mathcal{J}_n(s),\beta_n(s,j)\neq0}\parallel P_n\frac{d}{dQ_n}L(Q_n)(\phi_{s,j})\parallel=o_P(n^{-1/2}).\tag{4}$$

The right-hand side can typically be bounded in terms of $\min_{s,j,\beta_n(s,j)\neq0}\mid P_n\phi_{s,j}\mid$, so that a sufficient condition for (4) is that the HAL-MLE fit selects sparse enough basis functions. In particular, a sufficient condition is that $\min_{s,j\in\mathcal{J}_n(s),\beta_n(s,j)\neq0}\mid P_n\phi_{s,j}\mid=O_P(n^{-1/2})$, but, this rate can be lowered by utilizing that $Q_n$ converges to $Q_0$, $P_n$ approximates $P_0$, and $P_0\frac{d}{dQ_0}L(Q_0)(\phi_{s,j})=0$ for all $(s,j)$ (since $Q_0$ minimizes $P_0L(Q)$). For example, using the known $L^2(P_0)$-rate of convergence of $Q_n$, this rate for the support of the most sparse basis function in $Q_n$ can be lowered to $O_P(n^{-1/4+\alpha(k)/2})$, and, if one is able to prove $\parallel Q_n-Q_0\parallel_\infty=o_P(n^{-1/4})$, then even a rate of $O_P(n^{-1/4})$ would be sufficient.

## Application of undersmoothed HAL-MLE to provide causal inference for the ATE

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 in $[0,\tau]\in\mathbb{R}^k$, where various of its components are discrete and thereby supported on a finite grid within $[0,\tau]$. Let $\bar{G}(W)=E_P(A\mid W)$ and $\bar{Q}(A,W)=E_P(Y\mid A,W)$. Assume the positivity assumption $\bar{G}_0(W)>\delta>0$ for some $\delta>0$; $\bar{Q}_0$ and $\bar{G}_0$ are cadlag functions with $\parallel\bar{Q}_0\parallel^*_v\leq C^u$ and $\parallel\bar{G}_0\parallel^*_v\leq C^u_2$ for some finite constants $C^u,C^u_2$; $\delta<\bar{Q}_0<1-\delta$ for some $\delta>0$. This defines the statistical model $\mathcal{M}$ for $P_0$.

Let $\Psi:\mathcal{M}\to\mathbb{R}$ be defined by $\Psi(P)=E_PE_P(Y\mid W,A=1)$. For simplicity, we focus on estimation of this treatment specific mean, but the presentation trivially generalizes to the average treatment effect (ATE) $\Psi(P)=E_PE_P(Y\mid W,A=1)-E_PE_P(Y\mid A=0,W)$. 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 $\Psi$ is pathwise differentiable at $P$ with canonical gradient given by $D^*(\tilde{Q}, G) = A/\bar{G}(W)(Y - \bar{Q}(A,W)) + \bar{Q}(1,W) - \Psi(\tilde{Q})$. Let $L(\bar{Q})(O) = -\left\{Y \log \bar{Q}(A,W) + (1-Y)\log(1-\bar{Q}(A,W))\right\}$ 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 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

$$R_2(\bar{Q}, \bar{G}, \bar{Q}_0, \bar{Q}_0) = P_0 \frac{\bar{G} - \bar{G}_0}{\bar{G}}(\bar{Q} - \bar{Q}_0).$$

We have $\sup_{P \in \mathcal{M}} \parallel D^*(\tilde{Q}(P), G(P)) \parallel_v^* < C(C^u, C_2^u)$ for some finite constant $C$ implied by the universal bounds $(C^u, C_2^u)$ on the sectional variation norm of $\bar{Q}, \bar{G}$. We also note that, using Cauchy-Schwarz inequality, $R_{20}(\bar{Q}, \bar{G}, \bar{Q}_0, \bar{G}_0) \leq \frac{1}{\delta} \parallel \bar{Q} - \bar{Q}_0 \parallel_{P_0} \parallel \bar{G} - \bar{G}_0 \parallel_{P_0}$, where $\parallel f \parallel_{P_0}^2 = \int f^2(o) dP_0(o)$.

## HAL-MLE

Let $Q = \text{Logit}\bar{Q}$ and denote $L(\bar{Q})$ by $L(Q)$. Let $Q_{C,n} = \arg\min_{Q, \|Q\|_v^* < C} P_n L(Q)$ be the $C$-specific 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 $P(C_{n,cv} \leq C_n \leq C^u) = 1$. 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 $Q_n = \sum_{s,j \in \in \mathcal{J}_n(s)} \beta_n(s,j)\phi_{s,j}$, where $\phi_{s,j} = I(W(s) \geq w_{s,j})$ for a knot point $w_{s,j}$. By our rate of convergence results on the HAL-MLE we have that $\parallel Q_n - Q_0 \parallel_{P_0} = O_P(n^{-1/4 - \alpha(k)})$. The HAL-MLE of $\Psi(\tilde{Q}_0)$ is the plug-in estimator $\Psi(\tilde{Q}_n) = \frac{1}{n}\sum_{i=1}^n 1/(1 + \exp(-Q_n(W_i)))$. Note that $P_n D_2^*(\tilde{Q}_n) = 0$ for any $Q_n$. Thus, for showing that $P_n D^*(\tilde{Q}_n, G_0) = o_P(n^{-1/2})$, we only need $P_n D_1^*(Q_n, G_0) = o_P(n^{-1/2})$.

According to our theory, selecting $C_n$ as the smallest constant larger than $C_{n,cv}$ for which there is a selected basis function that has support smaller than a constant times $n^{-1/4+\alpha(k)/2}$ would make sure that $P_n D_1^*(Q_n, G_0) = o_P(n^{-1/2})$ and thereby, assuming this $C_n$ exists while being smaller than some finite constant $C^u$, that $\Psi(\tilde{Q}_n)$ is asymptotically efficient. Unfortunately, this global (i.e., not parameter specific) undersmoothing condition is not helpful in practice since we do not have a criterion for selecting the constant in front of the rate. Therefore, we implemented the following selector $C_n$ instead. Let $\bar{G}_n$ be an HAL-MLE of $\bar{G}_0$ using cross-validation for selecting the $L_1$-norm, and $\sigma_n^2 = P_n D_1^*(Q_{C_{n,cv},n}, G_n)^2$ be the resulting estimator of the sample variance of the canonical gradient. We select $C_n$ as the smallest constant $C$ larger than $C_{n,cv}$ for which $\mid P_n D_1^*(Q_{C,n}, G_n) \mid \leq \sigma_n/(n^{1/2}\log n)$. In this manner, this selector $C_n$ guarantees that indeed $P_n D_1^*(Q_n, G_n) = o_P(n^{-1/2})$, so that the efficiency $\Psi(\tilde{Q}_n)$ follows.

## Simulation results for undersmoothed HAL-MLE of treatment specific mean

We evaluated the proposed estimators via simulation. We drew 1000 samples of size $n \in \{250, 500, 1000, 2000, 4000\}$ from the following data distribution. We let $W_1 = 4Z - 2$, where $Z$ was drawn from a Beta(0.85,0.85) distribution. $W_2$ was independently drawn from a Bernoulli(0.5) distribution. Given $W = (w_1, w_2)$ we drew $A$ from a Bernoulli distribution with the probability $A = 1$ equal to $\bar{G}_0(w_1, w_2) = \text{expit}(w_1 - 2w_1w_2)$. Given $A = a$, and $W = (w_1, w_2)$ we drew $Y$ from a Normal($\bar{Q}_0(w_1, w_2)$, $0.33^2$) distribution with $\bar{Q}_0(w_1, w_2) = \text{expit}(w_1 - 2w_1w_2)$. As predicted by theory, the bias of the estimator is appropriately controlled and the variance of the estimator approaches the efficiency bound in larger samples (Figure 1). 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 equation (4). At all sample sizes, the sampling distribution of the scaled and centered estimators are well-approximated by the asymptotic distribution well.

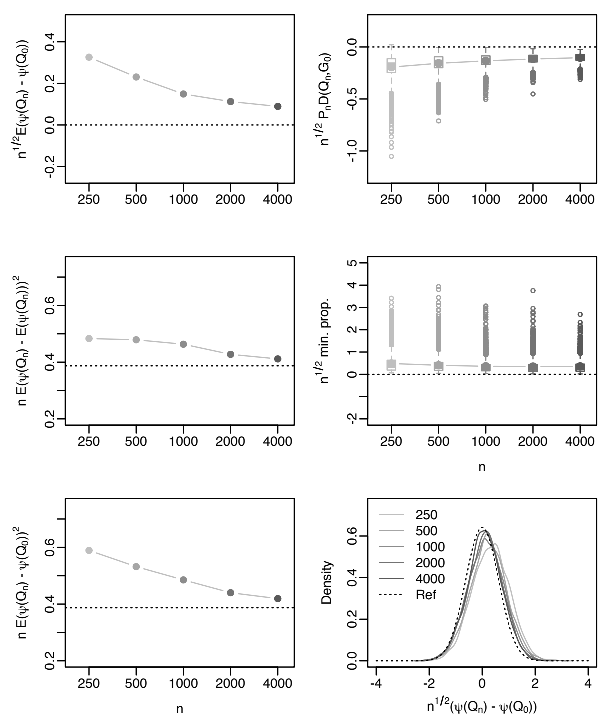


Figure 1: Left column top to bottom: bias, variance, and mean squared-error (all scaled by $n^{1/2}$) of undersmoothed HAL-MLE. Right column top to bottom: scaled empirical average of canonical gradient, empirical average of quantity given in equation (4), sampling distribution of scaled and centered estimator. The dashed lines in the variance and mean-squared error plots denote the efficiency bound. The reference sampling distribution for the estimators is a mean-zero Normal distribution with this variance.

## Discussion

Amongst the three standard frameworks for efficient estimation (estimating equations, one-step estimation, and TMLE), TMLE is often seen to be the most robust. This robustness may be attributed to its construction as a substitution estimator, which ensures it always respects global constraints on the target parameter and model. The comparison between TMLE and the undersmoothed HAL-MLE is less clear since both are substitution estimators. In causal inference and missing data settings, TMLE may behave erratically since it relies on extra model fitting that involves inverse probabilities of treatment and/or censoring. When these inverse weights are large, fitting may become unstable. Thus, the undersmoothed HAL-MLE may be more robust for weakly identifiable estimands. Nevertheless, in causal inference problems there might be substantial knowledge about the treatment and censoring mechanism, and the TMLE incorporates this knowledge to remove bias with respect to the target estimand. We expect that TMLE will be superior in such settings. However, in complex observational studies, where such knowledge on treatment and censoring mechanisms is lacking, and weak identifiability is a potential issue, the undersmoothed HAL-MLE might be the preferred procedure. Therefore, in future work we hope to establish a marriage between these two general methods that inherits the favorable properties of both procedures.

## References


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

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.

David L Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1(2000):32, 2000.

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

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

J. Pearl. *Causality: models, reasoning, and inference*. Cambridge, New York, 2nd edition, 2009.

M. Petersen and M.J. van der Laan. Causal models and learning from data: Integrating causal modeling and statistical estimation in the practice of epidemiology. *Epidemiology*, 25(3):418–426, 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.

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

M.J. van der Laan. Causal effect models for intention to treat and realistic individualized treatment rules. Technical report 203, Division of Biostatistics, University of California, Berkeley, 2006.

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 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 J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, Berlin Heidelberg New York, 2003.

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.

A.W. van der Vaart. *Asymptotic statistics*. Cambridge, New York, 1998.

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