The Highly Adaptive Lasso Estimator

David Benkeser and Mark van der Laan
Group in Biostatistics
University of California, Berkeley
Berkeley, California 94709
Email: benkeser@berkeley.edu

***Abstract*—Estimation of a regression functions is a common goal of statistical learning. We propose a novel nonparametric regression estimator that, in contrast to many existing methods, does not rely on local smoothness assumptions nor is it constructed using local smoothing techniques. Instead, our estimator respects global smoothness constraints by virtue of falling in a class of right-hand continuous functions with left-hand limits that have variation norm bounded by a constant. Using empirical process theory, we establish a fast minimal rate of convergence of our proposed estimator and illustrate how such an estimator can be constructed using standard software. In simulations, we show that the finite-sample performance of our estimator is competitive with other popular machine learning techniques across a variety of data generating mechanisms. We also illustrate competitive performance in real data examples using several publicly available data sets.**

I. Introduction

Estimation of the conditional mean of a random variable is necessary in many statistical applications. For example, we are often interested in estimating the causal effect of a binary intervention A, randomized based on covariates W, on an outcome Y. Under standard causal assumptions, efficient estimation in a nonparametric model of the causal effect requires estimation of two conditional means: the outcome regression $E_{P_0}(Y|A,W)$ and propensity regression $E_{P_0}(A|W)$, where P_0 denotes the true distribution of the observed data. Estimating conditional means is also important in settings where prediction is of interest. For example, we may be interested in predicting an outcome Y based on X, a set of features of data units. In such settings, the target parameter is the prediction function ψ_0 in a class of prediction functions $\mathbf{\Psi}$ that minimizes the average of a scientifically relevant loss function. That is, we would like to estimate $\Psi(P_0)(X) = \text{argmin}_{\psi\in\mathbf{\Psi}} E_{P_0}\{L(\psi)(X,Y)\}$, where $L(\psi)$ is a loss function, such as squared-error loss $L(\psi)(x,y) = \{y - \psi(x)\}^2$. The minimizer of the mean squared-error loss is $\psi_0(X) = E_{P_0}(Y|X)$, making estimating of the conditional again an important goal.

Parametric and semiparametric methods for estimating the conditional mean assume its form is known up to a finite number of parameters. For example, generalized linear models express the conditional mean as a transformation of a linear function of the conditioning variables. Parametric methods suffer from a large bias when the assumed functional form is different from the true conditional mean. In contrast, nonparametric methods, such as machine learning techniques, approximate ψ_0 using highly flexible functions. This makes nonparametric methods more appealing in many applications where little is known about the relationships between observed variables and outcome. Nonparametric methods make far fewer assumptions than parametric methods; however, nonparametric methods do typically require at least some assumptions to ensure statistical properties of estimators. For example, many methods assume ψ_0 has nearly constant, linear, or low-order polynomial behavior for all points sufficiently close to each other in a given metric.

Examples of methods assuming local smoothness include many popular learning techniques, such as histogram regression, tree-based methods, and generalized additive models. A central feature of these methods is the implicit or adaptive selection of the size of a neighborhood. The neighborhood size that optimizes the bias-variance trade-off can often be calculated explicitly. For example, a kernel regression estimator using kernels that are orthogonal to polynomials in x of degree k and bandwidth h assumes that ψ_0 is k-times continuously differentiable. The bias of the kernel regression estimator is $O(h^k)$ and variance is $O(1/(nhd))$, so that the optimal rate for the bandwidth can be calculated as $h = O(n^{-1/(2k+d)})$, resulting in a rate of convergence $O(n^{-k/(2k+d)})$. This example illustrates a key point about existing nonparametric regression methods in the literature: the rate of convergence is largely influenced by the dimension of the conditioning variables. The only way to achieve fast rates of convergence in high dimensions is by making strong smoothness assumptions and these assumptions may not be true in practice.

This discussion highlights the tradeoffs we are faced with when constructing nonparametric regression estimators: we must make some smoothness assumptions to guarantee reasonable convergence rates, but if we assume too much smoothness, the estimator will be ineffective. In this work, we outline a regression estimator that makes global, rather than local, smoothness assumptions on the true regression function. We assume that the true conditional mean function is right-hand continuous with left-hand limits (cadlag) and has variation norm smaller than a constant M. These are exceedingly mild assumptions that are expected to hold in almost every practical application. Nevertheless, these assumptions are sufficient to ensure a fast convergence rate regardless of the dimension of the regression function [1]. We propose a minimum loss-based estimator in this class of functions, which can be computed using L_1-penalized regression. This allows the estimator to be constructed using standard Lasso estimator software [2].

However, in stark contrast to the usual Lasso estimator, our implementation does not require a parametric specification of the relationship between predictors and outcome, but rather uses a special set of data-dependent basis functions. We call our estimator the highly adaptive lasso (HAL) estimator.

The remainder of the article is organized as follows. In Section 2, we review the theory developed in van der Laan (2016) that establishes the convergence rate of an MLE over the class of cadlag functions with finite variation norm. In Section 3, we propose a practical implementation of such an MLE using L_1-penalized regression. In Section 4, we illustrate the performance of the proposed estimator relative to competitors. In Section 5, we evaluate the performance of our estimator on real data sets. In Section 6, we highlight potential extensions to big data and high dimensional settings. We conclude with a short discussion.

II. Theoretical Framework

A. Cadlag functions with finite variation norm

Suppose the observed data consist of n i.i.d. copies of the random variable $O = (X, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a nonparametric model. Define the support of $P \in \mathcal{M}$ as a set $\mathcal{O}_P \subset \mathbb{R}^{d+1}$ such that $P(\mathcal{O}_P) = 1$. We assume that for each $P \in \mathcal{M}$, $\mathcal{O}_P \subset [0, \tau_P]$ for a finite $\tau_P \in \mathbb{R}^{d+1}_{>0}$ and define $\tau = \sup_{P \in \mathcal{M}} \tau_P$. Thus, τ is an upper bound of all the supports and for simplicity, we assume that τ is finite, though this need not be true for our main theorem to hold.

We make two key smoothness assumptions about ψ_0. The first is that $\psi_0 \in D[0, \tau]$, the Banach space of d-variate cadlag functions [3]. The second is that the variation norm of ψ_0 is finite. The typical definition of the variation norm is $||\psi||_v = \int_{[0,\tau]} |\psi(dx)|$; however, we make use of an alternative formulation of the variation norm defined as the sum of the variation norm over sections of ψ_0. Specifically, we define $X_s = \{X_j : j \in s\}$ for a given subset $s \subset \{1, \dots, d\}$. We also define $X_{s,c} = \{X_j : j \in s^c\}$, where $s^c = \{j : j \notin s\} \subset \{1, \dots, d\}$ denotes the complementary set of indices of s. For example, if $X = (X_1, X_2)$, we have subsets $X_1 = \{X_1\}$, $X_{1,c} = \{X_2\}$, $X_2 = \{X_2\}, X_{2,c} = \{X_1\}, X_{1,2} = \{X_1, X_2\}, X_{1,2,c} = \emptyset$.

For any function $\psi \in D[0, \tau]$, we define the s-th section of ψ as $\psi_s(x) = \psi\big(x_1 I(1 \in s), \dots, x_d I(d \in s)\big)$. This is the function that varies along the variables in x_s according to ψ, but sets the variables in $x_{s,c}$ equal to zero. Using these definitions, we can re-express the variation norm as

$$||\psi||_v = \psi(0) + \sum_{s \subset \{1,\dots,d\}} \int_{0_s}^{\tau_s} \left|\psi_s(du)\right| \ ,$$

where the sum is taken over all subsets of $\{1, \dots, d\}$.

Consider the class of functions $\mathbf{\Psi}_M = \{\psi \in \mathbf{\Psi} : ||\psi||_v < M\}$, where M can be taken to be an arbitrarily large constant. We expect that the true conditional mean in almost every real data application will fall into this class and we therefore view this as an extremely mild assumption of the true conditional mean. Examples of functions with infinite variation norm tend to be pathological; for example, $\psi_0(x) = \cos(1/x)$ has infinite variation norm.

B. Minimum loss-based estimator in class of cadlag functions with finite variation norm

We now turn to what can be said about a minimum loss-based estimator in the class $\mathbf{\Psi}_M$ for a given M. We define the MLE in this class as

$$\psi_{n,M} = \arg\min_{\psi \in \mathbf{\Psi}_M} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \psi(X_i)\}^2 \ ,$$

and the true minimizer of the average loss as

$$\psi_{0,M} = \arg\min_{\psi \in \mathbf{\Psi}_M} E_{P_0}\{Y - \psi(X)\}^2 \ .$$

Note that if $M > ||\psi_0||_v$, then $\psi_0 \in \mathbf{\Psi}_M$ and $\psi_{0,M} = \psi_0$. The following theorem establishes the rate of convergence of the MLE with respect to $L_2(P_0)$ norm denoted $||\psi_n - \psi_0||_{P_0} = \{P_0(\psi_n - \psi_0)^2\}^{1/2}$, where we use the notation $Pf = E_P\{f(O)\}$.

Theorem 1: Let $L(\psi)$ be the squared-error loss function and $\mathbf{\Psi}_M = \{\psi \in \mathbf{\Psi} : ||\psi||_v < M\}$ be the set of cadlag functions with variation norm smaller than M, and $d_0(\psi, \psi_0) = P_0\{L(\psi) - L(\psi_0)\} = ||\psi_n - \psi_0||^2_{P_0}$ be the loss-based dissimilarity for squared-error loss. If

$$\text{(A1)} \sup_{\psi \in \mathbf{\Psi}_M} \frac{||L(\psi)||_v}{||\psi||_v} < \infty \ , \text{ and}$$

$$\text{(A2)} \sup_{\psi \in \mathbf{\Psi}_M} \frac{||L(\psi) - L(\psi_{0,M})||^2_{P_0}}{d_0(\psi_0, \psi_{0,M})} < \infty \ ,$$

then $||\psi_{n,M} - \psi_{0,M}||_{P_0} = O_P(n^{-1/4-\alpha(d)/8})$ where $\alpha(d) = 1/(d+1)$. Specifically, if $M > ||\psi_0||_v$, then $||\psi_{n,M} - \psi_0||_{P_0} = O_P(n^{-1/4-\alpha(d)/8})$.

Proof: We have

$$\begin{aligned} 0 \le d_0(\psi_{n,\lambda}, \psi_{0,\lambda}) &= P_0\{L(\psi_{n,M}) - L(\psi_{0,M})\} \\ &= -(P_n - P_0)\{L(\psi_{n,M}) - L(\psi_{0,M})\} \\ &\qquad + P_n\{L(\psi_{n,M}) - L(\psi_{0,M})\} \\ &\le -(P_n - P_0)\{L(\psi_{n,M}) - L(\psi_{0,M})\}. \end{aligned}$$

Note that $L(\psi_{n,M}) - L(\psi_{0,M})$ falls in the P_0-Donsker class of all cadlag functions with variation norm smaller than a constant. By (A1) it follows that $P_0\{L(\psi_{n,M}) - L(\psi_{0,M})\} = O_P(n^{-1/2})$. As a consequence of (A2), we have $||L(\psi_{n,M}) - L(\psi_{0,M})||^2_{P_0} = O_P(n^{-1/2})$. By empirical process theory [4], we have that $\sqrt{n}(P_n - P_0)f_n \to 0$ in probability if f_n falls in a P_0-Donsker class with probability tending to 1, and $P_0 f_n^2 \to 0$ in probability as $n \to \infty$. Applying this to $f_n = L(\psi_{n,M}) - L(\psi_{0,M})$ shows that $(P_n - P_0)\{L(\psi_{n,M}) - L(\psi_{0,M})\} = o_P(n^{-1/2})$, which proves $d_0(\psi_{n,M}, \psi_{0,M}) = o_P(n^{-1/2})$ and that $||\psi_{n,M} - \psi_{0,M}||_{P_0} = o_P(n^{-1/4})$. However, it is possible to provide a more precise rate by utilizing the bound on the entropy of the Donsker class of cadlag functions with variation norm smaller than a constant established in van der Vaart and Wellner (2011) [5]. This result was used by van der Laan (2015) [1] to give the precise rate

$d_0(\psi_{n,M},\psi_{0,M}) = O_P(n^{-1/2-\alpha(d)/4})$ with $\alpha(d) = 1/(d+1)$. Thus, we have that $||\psi_{n,M} - \psi_{0,M}||_{P_0} = O_P(n^{-1/4-\alpha(d)/8})$. □

Theorem 1 establishes that MLE $\psi_{n,M}$ converges to its M-specific true counterpart in L_2-norm no slower than $n^{-1/4}$ regardless of the dimension of X. This is a remarkable result since this minimum rate does not depend on the underlying local smoothness of ψ_0. For example, ψ_0 could be a function that is non-differentiable at many points.

C. Cross-validated choice of bound on variation norm

The bound on the variation norm of ψ_0 is unlikely to be known in practice and will need to be chosen based on the data via cross validation. Consider a grid $M_1,\ldots,M_{K_n}$ of K_n potential values for the bound, the largest of which is larger than $||\psi_0||_v$. The latter can be achieved in practice by selecting M_{K_n} to be such that $\psi_{n,M_{K,n}}$ perfectly fits the data. For $k = 1,\ldots,K_n$, let $\psi_{n,M_k} = \hat{\Psi}_M(P_n)$ denote the MLE over the class of cadlag functions with variation norm smaller than M_k. Consider a V-fold cross-validation scheme and let $P^0_{n,v}$, $P^1_{n,v}$ be the empirical distribution of the training and validation sample corresponding with sample split v, $v = 1,\ldots,V$. The cross-validation selector M_n of M is the choice with the lowest estimated cross-validated risk,

$$M_n = M_{k'} : k' = \arg\min_k \frac{1}{V}\sum_{v=1}^{V} P^1_{n,v} L(\hat{\Psi}_{M_k}(P^0_{n,v})) \ .$$

The estimator of ψ_0 is given by $\psi_n = \psi_{n,M_n} = \hat{\Psi}_{M_n}(P_n)$. To examine how the cross-validation-based selection of M affects the statistical properties of ψ_n relative to the MLE presented in Theorem 1 (where the bound was known), we now assume that the loss function is uniformly bounded

(A3) $\sup_{\psi\in\mathbf{\Psi},o\in\mathcal{O}} \mid L(\psi)(o) \mid < \infty$.

The oracle selector is defined as the value of M that minimizes the true cross-validated risk

$$M^*_n = M_{\tilde{k}} : \tilde{k} = \arg\min_k \frac{1}{V}\sum_{v=1}^{V} P_0 L(\hat{\Psi}_{M_k}(P^0_{n,v})) \ .$$

We can compare the performance of ψ_n to the oracle choice $\psi^*_n = \psi_{n,M^*_n}$ using the finite-sample oracle inequality for cross validation [6]–[8], which establishes

$$\frac{d_0(\psi_n,\psi_0)}{d_0(\psi^*_n,\psi_0)} \to 1 \ ,$$

in probability. Thus, the cross-validated selector is asymptotically equivalent with the oracle choice in loss-based dissimilarity. We can further establish that

$$||\psi_{n,M} - \psi_{0,M}||_{P_0} = O_P(n^{-1/4-\alpha(d)/8}) + O_P(n^{-1/2}\log K_n) \ ,$$

so that by choosing K_n such that $n^{-1/2}\log K_n$ converges to zero as n goes to infinity, the minimal convergence rate established in Theorem 1 will be preserved. Note that this requirement is mild and in particular, admits schemes with $K_n = n^p$ for any $p > 0$.

We have now established that the statistical properties of the MLE presented in Theorem 1 are retained even when the true variation norm is not known and must be chosen via cross validation. We now turn to how such an MLE can be constructed in practice.

III. The highly adaptive lasso estimator

For any $\psi \in D[0,\tau]$ with $||\psi||_v < \infty$, we have the following representation of ψ [9]:

$$\psi(x) = \psi(0) + \sum_{s\subset\{1,\ldots,d\}} \int_{0_s}^{x_s} \psi_s(du) \ ,$$

which can also be written as

$$\psi(0) + \sum_{s\subset\{1,\ldots,d\}} \int_{0_s}^{\tau_s} I\big(u \leq x_s\big)\psi_s(du) \ . \tag{1}$$

Consider approximating the representation of ψ in equation (1) using a discrete measure ψ_m with m support points. For each subset s, define $\psi_{m,s}$ to be the discrete approximation of ψ_s with support points given by $(u_{s,j} : j)$. Let $d\psi_{m,s,j}$ be the pointmass assigned to point $u_{s,j}$ by ψ_m. An approximation of ψ may then be constructed as:

$$\psi_m(x) = \psi(0) + \sum_{s\subset\{1,\ldots,d\}} \sum_j I(u_{s,j} \leq x_s) d\psi_{m,s,j} \ . \tag{2}$$

Note that this approximation consists of a linear combination of basis functions $x \to \phi_{s,j}(x) = I(x_s \geq u_{s,j})$ with corresponding coefficients $d\psi_{m,s,j}$ summed over s and j. Additionally note that the sum of the absolute values of these coefficients gives the variation norm of ψ_m:

$$||\psi_m||_v = \psi(0) + \sum_{s\subset\{1,\ldots,d\}} \sum_j |d\psi_{m,s,j}| \ .$$

In words, we have illustrated that any function of finite variation norm admits a representation as the sum over subsets of an integral with respect to a subset-specific measure. Each subset-specific measure can be approximated using a discrete measure with a given set of support points. In simpler terms, each subset-specific function ψ_s can be approximated by a step function with jumps at a given set of points. At a given point, the sum of the subset-specific step functions at that point gives an approximation of the function ψ at that point. Each subset-specific step function can be represented as a linear combination of indicator basis functions for a given set points where the step function jumps. Furthermore, the variation norm of the approximation of ψ is given by the sum over all approximating step functions of the absolute value of the jumps made by each function.

The question natrually arises as to where the support points of the discrete measure should be placed; that is, where should the step functions jump? We argue that the minimizer of the empirical risk over all measures (continuous and discrete) is equivalent with minimizing the empirical risk over all discrete measures with support defined by the actual n observations. The empirical risk $P_nL(\psi)$ only depends on ψ through $\{\psi(\tilde{x}_i) : i = 1,\ldots,n\}$, where we use $\tilde{x}_i$ to denote the observed values of X, $i = 1, \dots, n$. This suggests that the infinite-dimensional minimization over $\mathbf{\Psi}_M$ can be replaced by a finite dimensional minimization problem. Specifically, for each subset $s \subset \{1, \dots, d\}$, let $\tilde{x}_{i,s}$ be the subvector $\{\tilde{x}_{i,k} : k \in s\}, i = 1, \dots, n$. We now apply (2) with support points of ψ_s given by $\{\tilde{x}_{i,s} : i = 1, \dots, n\}$ for each s:

$$\psi_m(x) = \psi(0) + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} I(\tilde{x}_{i,s} \leq x_s) d\psi_{m,s,i} \ . \quad (3)$$

We can consider minimization only over linear combinations of basis functions $x \to \phi_{s,i}(x) = I\big(x_s \geq \tilde{x}_{s,i}\big)$ and corresponding coefficients $d\psi_{m,s,i}$ summed over subsets $s \subset \{1, \dots, d\}$ and for $i = 1, \dots, n$. For the purposes of carrying out the minimization, we define

$$\psi_\beta = \beta_0 + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} \beta_{s,i} \phi_{s,i} \ ,$$

and a corresponding subspace

$$\mathbf{\Psi}_{n,M} = \left\{ \psi_\beta : \beta,\ \beta_0 + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} |\beta_{s,i}| < M \right\} .$$

The minimizer $\psi_{n,M}$ will be equivalent with ψ_{β_n}, the minimizer in β over this subspace:

$$\beta_n = \arg \min_{\beta, \beta_0 + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} |\beta_{s,i}| < M} P_n L(\psi_\beta) \ .$$

Note that this minimization corresponds exactly to Lasso linear regression [2] with covariates $\phi_{s,i}$ across all subsets $s \subset \{1, \dots, d\}$ and for $i = 1, \dots, n$. Thus, computation of $\psi_{n,M}$ requires minimizing over the vector β under the constraint that its L_1-norm is bounded by M. If all d components of X are continuously distributed, then the dimension of the vector β will be at most $n(2^d - 1)$. If X has discrete components the representation of ψ given above generalizes. For example, a function on $\{0, 1\}$ is not cadlag, but may be extended to a cadlag function by defining it as constant on $(0, 1)$ with right-continuous jumps at 0 and 1. Note that when X has discrete components the number of unique basis functions required for the approximation (3) will be far fewer than $n(2^d - 1)$.

IV. Illustration in low dimensions

In this section, we illustrate our estimator in the simple situations with $d = 1$ and $d = 2$. For the univariate setting, we drew $n = 500$ independent copies of X from a Uniform(-4,4) and of ϵ from a Normal(0,1) distribution and let $Y = 2\text{sin}(\pi/2|X|) + \epsilon$, so that $||\psi_0||_v = 16$. The basis functions used by the HAL estimator consist of n indicators at the observed data points: $\phi_j(x) = I(x \geq \tilde{x}_j)$ for $j = 1, \dots, n$. To select the bound on the variation norm, we used ten-fold cross validation to select from 100 possible bounds ranging from 0 to about 350. We illustrate the fit from three of these choices in Figure 1. The solid line is the HAL estimator, which uses the cross-validation-selected value $M_n = 13.9$. The dashed and dotted lines represent choices that are smaller and larger than the true variation norm, respectively. The ticks at the bottom of the figure are placed at the 46 support points of ψ_n with a non-zero coefficient. The choice of 4.8 as bound on the variation norm (dashed line) visibly over-smooths the data, while the bound of 35.2 appears to provide a reasonable approximation and is similar with the prediction from the HAL estimator. However, the larger bound does appear to produce more noise near the edges of the support. Theory dictates that any choice of bound larger than the true norm will yield an estimator with the properties established in Theorem 1. Nevertheless, we expect that the HAL estimator will exhibit superior performance in finite samples by allowing for selection of a bound smaller than the true norm. The oracle inequality guarantees that so long as at least one bound larger than the true norm is considered as a candidate bound, then we will eventually select a bound that is larger than the true variation norm. Indeed, when the sample size was increased to 5,000 (data not shown), the cross-validation-selected bound was selected as 16.8, which is close to the true value.

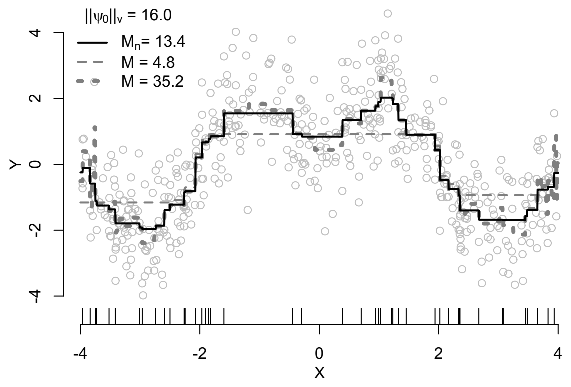

Fig. 1. The highly adaptive lasso in the univariate setting.

We now illustrate the bivariate setting where X has a discrete component. We again drew X_1 from a Uniform(-4,4) distribution and also drew X_2 independently from a Bernoulli(0.5) distribution. We let $Y = -0.5 * X_1 + X_2 X_1^2/2.75 + X_2 + \epsilon$ where ϵ was drawn from a Normal(0,1) distribution. To construct the HAL estimator in this setting, we first created n basis functions corresponding with indicators at the observed values of X_1: $\phi_{1,i}(x) = I(x_1 \geq \tilde{x}_{1,i})$ for $i = 1, \dots, n$. Next, we added basis functions for the subset consisting only of X_2: $\phi_{2,i}(x) = I(x_2 \geq \tilde{x}_{2,i})$ for $i = 1, \dots, n$. Note that because X_2 is binary, there was only be a single unique basis function to be added, $\phi_2(x) = I(x_2 \geq 1)$. Finally, we created bivariate basis functions of the form $\phi_{12,i}(x) = I(x_1 \geq \tilde{x}_{1,i}, x_2 \geq \tilde{x}_{2,i})$ for $i = 1, \dots, n$. These basis functioned number fewer than n due to binary X_2. In particular, it was unnecessary to add basis functions $\phi_{12,i}(x)$ for any i for which $\tilde{x}_{2,i} = 0$ due to the fact that for any such i we had already placed support on this zero-edge by including $\psi_{i,1}$. This illustrates the point made at the end of the previous section that the number of basis functions in a given sample will be at most $n(2^d - 1)$, while in practice the number may

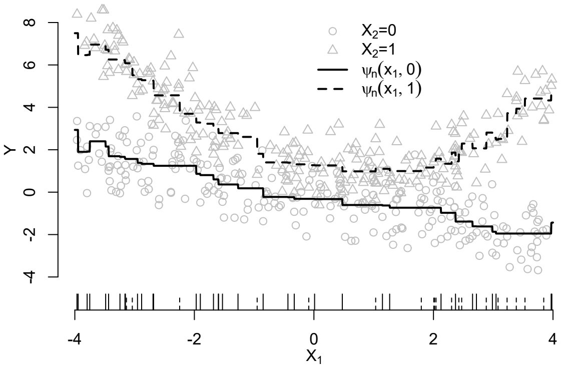

Fig. 2. The highly adaptive lasso in the bivariate setting.

be far fewer depending on the number of unique observations in a given data set.

Figure 2 illustrates a random draw of size $n = 500$ from this data generating mechanism. Two lines are shown corresponding with the estimate of ψ_0 when $X_2 = 1$ (upper dashed line) and when $X_2 = 0$ (lower solid line). The solid tick marks across the bottom of the figure indicate the univariate basis functions with a non-zero coefficient in ψ_n. Accordingly, these marks corresponding with jumps in both $\psi_n(\cdot, 0)$ and $\psi_n(\cdot, 1)$. The dashed tick marks indicate the bivariate basis functions with non-zero coefficients and thus correspond with values of a jump in $\psi_n(\cdot, 1)$, but not $\psi_n(\cdot, 0)$. Notice that, as expected these ticks occur most frequently when $X_1 > 2$, corresponding with the values for which $\psi_0(x_1, 0)$ is decreasing in x_1, while $\psi_1(x_1, 1)$ is increasing.

V. Simulation

We evaluated the finite-sample performance of the HAL estimator relative to other nonparametric algorithms: regression trees [10], random forests [11], gradient boosted machines (GBM) [12], kernel regression [13], [14], support vector machines (SVM) [15], and polynomial multivariate adaptive regressions splines (Polynomial MARS) [16]. We considered three types of data generating mechanisms, which we call smooth, jumps, and sinusoidal. For each type of data generating mechanism, we varied the dimension of X and considered $d \in \{1, 3, 5\}$ and sample sizes $n \in \{500, 1000, 2000\}$. Performance was judged based on R^2, which was calculated on an independent test set of size $N = 1e4$, where for a given estimator $\hat{\psi}$, we define

$$R^2 = 1 - \frac{\sum_{i=1}^N \{Y_i - \hat{\psi}(X_i)\}^2}{\sum_{i=1}^N \{Y_i - \bar{Y}_N\}^2} \ .$$

Each setting was designed so that the optimal R^2 value was $R^2_{opt} = 0.80$, where

$$R^2_{opt} = 1 - \frac{E_{P_0}\{Y - \psi_0(X)\}^2}{\text{Var}_0(Y)}$$

is the value of R^2 obtained when using the true regression function ψ_0. This value can be viewed as an upper bound on the performance of any estimator.

The distribution of X was as follows:

$$\begin{aligned} &X_1 \sim \text{Uniform}(-4, 4) &&; \ X_2 \sim \text{Uniform}(-4, 4) \\ &X_3 \sim \text{Bernoulli}(0.5) &&; \ X_4 \sim \text{Normal}(0, 1) \\ &X_5 \sim \text{Gamma}(2, 1) \ . \end{aligned}$$

For dimension d, call the target parameter $\psi_0^d(X)$ and let $X = (X_j : j = 1, \ldots, d\})$. We define $Y = \psi_0^d(X) + \epsilon$ where $\epsilon \sim \text{Normal}(0, 1)$.

The "smooth" regression functions for $d = 1, 3, 5$ respectively were defined as

$$\begin{aligned} \psi_0^1(x) &= 0.05x_1 + 0.42x_1^2 \ ; \\ \psi_0^3(x) &= 0.07x_1 - 0.28x_1^2 + 0.5x_2 + 0.25x_2x_3 \ ; \\ \psi_0^5(x) &= 0.1x_1 - 0.3x_1^2 + 0.25x_2 + 0.5x_3x_2 - 0.5x_4 + \\ &\quad 0.04x_5^2 - 0.1x_5 \ . \end{aligned}$$

The "jump" regression functions were defined as

$$\begin{aligned} \psi_0^1(x) &= -2.7I(x_1 < -3) + 2.5I(x_1 > -2) - 2I(x_1 > 0) + \\ &\quad 4I(x_1 > 2) - 3I(x_1 > 3) \ ; \\ \psi_0^3(x) &= -2I(x_1 < -3)x_3 + 2.5I(x_1 > -2) - 2I(x_1 > 0) + \\ &\quad 2.5I(x_1 > 2)x_3 - 2.5I(x_1 > 3) + I(x_2 > -1) - \\ &\quad 4I(x_2 > 1)x_3 + 2I(x_2 > 3) \ ; \\ \psi_0^5(x) &= -I(x_1 < -3)x_3 + 0.5I(x_1 > -2) - I(x_1 > 0) + \\ &\quad 2I(x_1 > 2)x_3 - 3I(x_1 > 3) + 1.5I(x_2 > -1) - \\ &\quad 5I(x_2 > 1)x_3 + 2I(x_2 > 3) + 2I(x_4 < 0) - \\ &\quad I(x_5 > 5) - I(x_4 < 0)I(x_1 < 0) + 2x_3 \ . \end{aligned}$$

The "sinusoidal" regression functions were defined as

$$\begin{aligned} \psi_0^1(x) &= 2\text{sin}(0.5\pi|x_1|) + 2\text{cos}(0.5\pi|x_1|) \ ; \\ \psi_0^3(x) &= 4x_3I(x_2 < 0)\text{sin}(0.5\pi|x_1|) + \\ &\quad 4.1I(x_2 \geq 0)\text{cos}(0.5\pi|x_1|) \ ; \\ \psi_0^5(x) &= 3.8x_3I(x_2 < 0)\text{sin}(0.5\pi|x_1|) + \\ &\quad 4I(x_2 > 0)\text{cos}(\pi|x_1|/2) + \\ &\quad 0.1x_5\text{sin}(\pi x_4) + x_3\text{cos}(|x_4 - x_5|) \ . \end{aligned}$$

Figure 3 displays the results of the simulation study with rows representing the different data generating mechanisms and columns representing the different dimensions of X. The margins of the figure show the results aggregated across data generating mechanisms of a particular dimension (bottom margin) and aggregated across different dimensions of a particular data generating mechanism (right margin). In each plot, the algorithms have been sorted by their average R^2 value across the three sample sizes with the highest R^2 at the top of the figure and the lowest R^2 at the bottom.

Beginning with the top row corresponding to the "smooth" data generating mechanisms, we find that all algorithms other than random forests perform well when $d = 1$, with kernel regression performing the best in this case. However, as

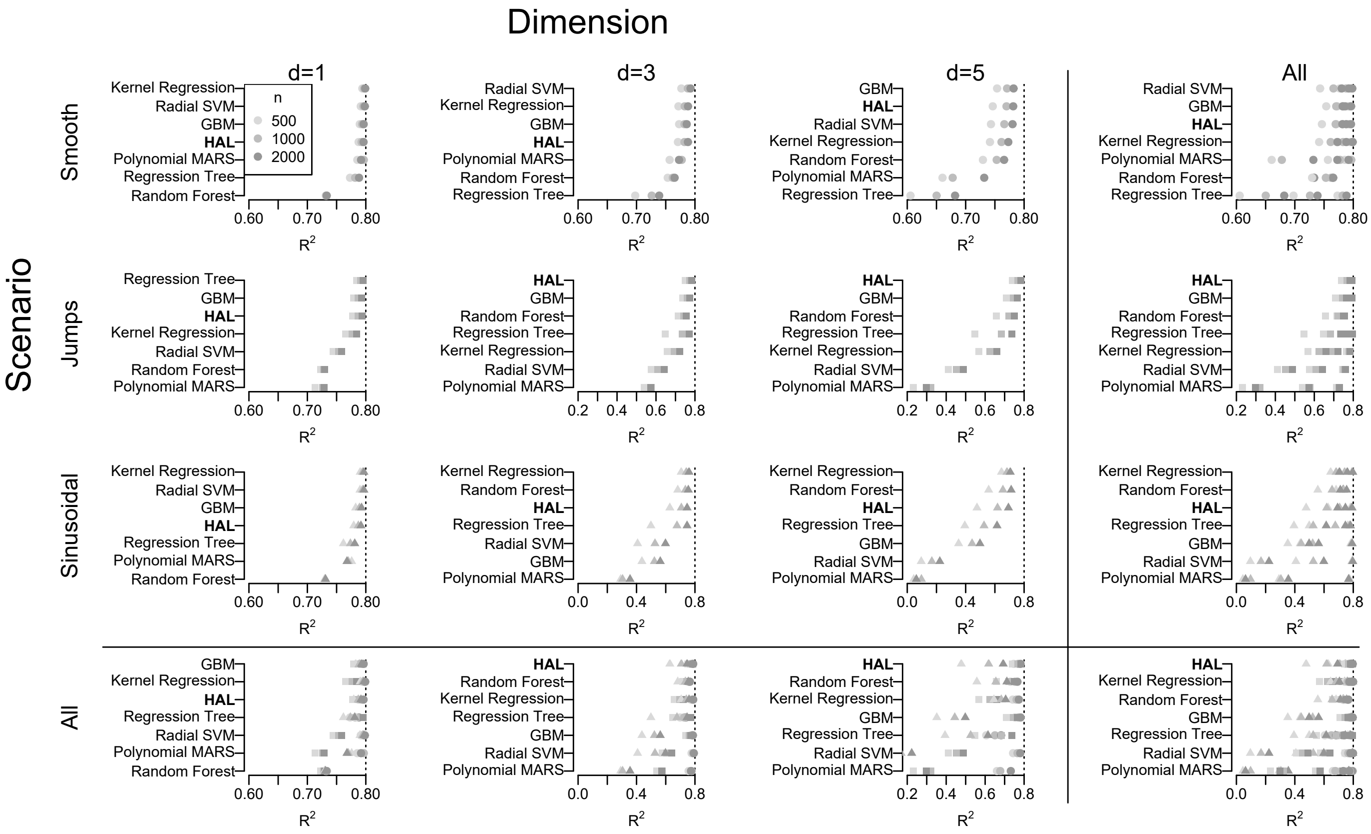

Fig. 3. Simulation study results

the dimension increases, the relative performance of kernel regression decreases, while the relative performance of HAL increases. Across all dimensions the SVM had the best overall performance; however, the performance of the GBM and HAL were comparable. In the second row corresponding with the "jumps" scenario, we see that the HAL performs extremely well, nearly achieving the optimal R^2 when $n = 2000$ for all dimensions. In the third row corresponding with the "sinusoidal" scenario, we find that somewhat surprisingly the kernel regression performs the best across all dimensions. This appears to be due in part to superior performance relative to other estimators when $n = 500$. For the larger sample sizes, the R^2 achieved by kernel regression, random forests, and HAL are similarly high. The far bottom right plot shows the results over all simulations with algorithms sorted by average R^2 and we see that HAL had the highest average R^2 followed by kernel regression and random forests. Overall, HAL performed well relative to competitors in all scenarios and particularly well in the jump setting, where local smoothness assumptions fail. Though the estimator was not ranked highest for the smooth and sinusoidal data generating mechanisms, its performance was comparable to the best-performing machine learning algorithms, which are generally considered to be state-of-the-art.

VI. Data Analysis

We separately analyzed five publicly available data sets listed with citation in Table I. Sample sizes for the data sets ranged from 201 to 654 and d ranged from four to eleven. In addition to the nonparametric methods evaluated in simulations, we considered estimation of ψ_0 with several parametric methods as well. These included a main terms generalized linear model (GLM), a stepwise GLM based on AIC including two-way interactions, and a generalized additive model (GAM) with the degree of splines determined via ten-fold cross-validation.

In practice, the individual best method for estimating a conditional mean will not be known a-priori. The performance of the various methods will be determined by both the sample size and the underlying true data distribution. Therefore, in a given data example, we may wish to consider many algorithms for the purpose of estimating ψ_0. The Super Learner is an ensemble learning method that, based on the methods' cross-validated risk, selects either the single best method or the best weighted combination of methods from a library of candidate methods. The former is often referred to as the discrete Super Learner, while the latter is referred to as either the continuous Super Learner or simply the Super Learner [17]. The same oracle inequality used to establish the performance of the HAL estimator in Section II-C may be used to establish that the performance of the Super Learner will be asymptotically

TABLE I
DATA SETS ANALYSED USING THE HAL ESTIMATOR AND COMPETITORS

Name	n	d
cpu [18]	209	6
laheart [19]	201	11
oecdpanel [20]	616	6
pima [21]	392	7
fev [22]	654	4

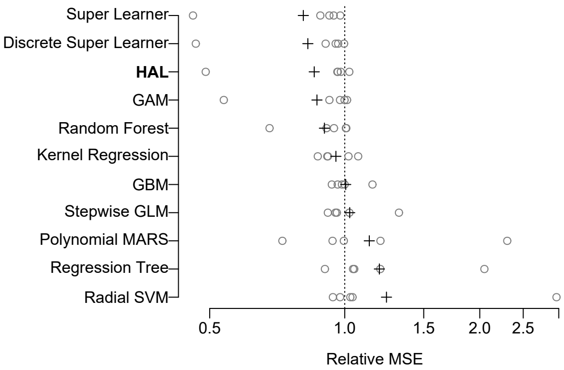

Fig. 4. Relative cross-validated mean squared error of methods in five real data sets. circle = result on a single data set, cross = geometric mean over five data sets.

equivalent with the performance of the best method considered. Therefore, by including the HAL estimator in the Super Learner, under mild conditions we are guaranteed to perform at least as well as the HAL estimator, but possibly much better.

In order to compare the performance of the various methods across different data sets with different outcomes, we studied the ten-fold cross-validated mean squared-error of each method relative to that of the main terms GLM. Values greater than one correspond to better performance of the GLM. The results of each of the data analyses are shown in Figure 4. The gray dots corresponds to the relative MSE in a particular data set, while the black cross corresponds to the geometric mean across all five studies. The Super Learner and discrete Super Learner perform best, followed by the HAL estimator. The HAL estimator performed particularly well on the cpu dataset, where its cross-validated MSE was nearly half that of the main terms GLM.

VII. CONSIDERATIONS FOR INCREASING DIMENSIONS

A limitation of the proposed HAL estimator is that as the dimension of X increases storage of the $n(2^d - 1)$ basis functions becomes infeasible. However, because the empirical risk $P_n L(\psi)$ only depends on ψ through n values $\{\psi(X_i) : i = 1, \ldots, n\}$, it may be possible to further reduce the number of basis functions while still attaining the minimum of the empirical risk. Our theorem proves that any $\psi_{n,M}$ attaining the minimum will converge to $\psi_{0,M}$ at the desired rate, while in fact, it suffices to achieve the minimum up to an approximation error smaller than this rate. This suggests that for finite samples it might suffice to work with a smaller subset of basis functions even though, in theory, all of these types of basis functions would be included as sample size increases so that any function can be arbitrarily well approximated.

One such strategy could consist of ordering basis functions from one-way, two-way, to d-way, while also ordering within each set of k-way basis functions $k = 1, \ldots, d$. An example of the latter ordering could be choosing basis functions based on p evenly spaced quantiles of the observed X, letting p go from one to n. The modified HAL estimator would determine the cross-validated risk of a HAL estimator based on $n_\ell \ll n(2^d - 1)$ basis functions and subsequently add the ordered basis functions until the cross-validated risk is no longer improved or increases by a fixed amount. Under reasonable assumptions, we expect such a scheme should perform well in finite samples, while ensuring that as sample size increases, there will be enough basis functions added to adequately approximate ψ_0.

Computational problems may also arise as the number of observations n grows even for small dimensions. In these cases, one may not be able to compute the empirical minimizer $\psi_{n,M}$ using standard software. However, it may be possible to adopt an online minimization approach, such as stochastic gradient descent [23]. The bound on the variation norm can also be selected in an online manner by minimizing online cross-validated risk. Developing computationally feasible algorithms which approximate the desired $\psi_{n,M}$ will be an important area of future research.

VIII. CONCLUSION

In this paper we defined a new nonparametric regression estimator, which we call the highly adaptive lasso estimator. A remarkable feature of this estimator is that it will converge to the truth at a rate faster than $n^{-1/4}$ regardless of the dimension of X. At first glance, this seems to contradict the well known minimax convergence rates from the nonparametric estimation literature. However, such minimax rates are with respect to estimation of ψ_0 at a single point x, while our rates are with respect to $L^2(P_0)$ norm. Another fascinating consequence of the minimal convergence rate achieved by HAL is its relation to construction of efficient estimators of pathwise differentiable parameters in infinite-dimensional models. These estimators typically require as an intermediate step estimation of the nuisance parameters that index the target parameter's efficient influence curve. For example, as mentioned in the introduction, the statistical parameter identifying the average causal effect of a treatment requires estimation of the outcome regression and propensity regression. These estimators can be used to construct a cross-validated targeted minimum loss-based estimator (CV-TMLE) [24] of the average causal effect. If every strata of covariates has a positive probability of receiving treatment, then a sufficient condition to prove asymptotic efficiency of the CV-TMLE estimator is that the estimated regressions converge to their true value in $L_2(P_0)$ norm faster than $n^{-1/4}$. The HAL estimator appears to be the first estimator that can guarantee asymptotically efficient estimation of the average causal effect without enforcing strong smoothness conditions [1].

We have made available the code used to execute the simulations and data analyses in a GitHub repository at https://github.com/benkeser/hal.

Acknowledgment

This work was supported by NIH grant 5R01AI074345-07 and Bill and Melinda Gates Foundation grant OPP1147962. The authors would like to thank Oleg Sofrygin for his input on computational considerations in constructing the HAL estimator.

References

[1] M. J. van der Laan, "A generally efficient targeted minimum loss based estimator," *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 343.*, 2015.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[3] G. Neuhaus, "On weak convergence of stochastic processes with multidimensional time parameter," *The Annals of Mathematical Statistics*, vol. 42, no. 4, pp. 1285–1295, 1971.

[4] A. W. van der Vaart and J. A. Wellner, *Weak Convergence*. Springer, 1996.

[5] A. van der Vaart and J. A. Wellner, "A local maximal inequality under uniform entropy," *Electronic Journal of Statistics*, vol. 5, no. 2011, p. 192, 2011.

[6] 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," *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130.*, 2003.

[7] A. W. van der Vaart, S. Dudoit, and M. J. van der Laan, "Oracle inequalities for multi-fold cross validation," *Statistics & Decisions*, vol. 24, no. 3, pp. 351–371, 2006.

[8] M. J. van der Laan, S. Dudoit, and A. W. van der Vaart, "The cross-validated adaptive epsilon-net estimator," *Statistics & Decisions*, vol. 24, no. 3, pp. 373–395, 2006.

[9] R. D. Gill, M. J. Laan, and J. A. Wellner, "Inefficient estimators of the bivariate survival function for three models," in *Annales de l'IHP Probabilités et statistiques*, vol. 31, no. 3, 1995, pp. 545–597.

[10] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[13] E. A. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[14] G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

[15] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[16] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, pp. 1–67, 1991.

[17] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, 2007.

[18] D. Kibler, D. W. Aha, and M. K. Albert, "Instance-based prediction of real-valued attributes," *Computational Intelligence*, vol. 5, no. 2, pp. 51–57, 1989.

[19] A. Afifi and S. Azen, "Statistical analysis, a computer oriented approach," 1979.

[20] Z. Liu, T. Stengos *et al.*, "Non-linearities in cross-country growth regressions: a semiparametric approach," *Journal of Applied Econometrics*, vol. 14, no. 5, pp. 527–538, 1999.

[21] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1988, p. 261.

[22] B. Rosner, *Fundamentals of Biostatistics*, 5th ed. Pacific Grove, 1999.

[23] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT 2010*. Springer, 2010, pp. 177–186.

[24] W. Zheng and M. J. van der Laan, "Cross-validated targeted minimum-loss-based estimation," in *Targeted Learning*. Springer, 2011, pp. 459–474.