RandomHAL: Quickly approximating a learning algorithm while preserving its statistical rate

Salvador Balkus and Nima Hejazi

August 2, 2025

Abstract

In statistical learning, algorithms that fit a lasso over a set of basis functions can achieve desirable theoretical properties. For example, the Highly Adaptive Lasso (HAL) estimator applies the lasso to a very high-dimensional indicator or spline basis, attaining dimension-free rates of convergence across a large class of functions. However, the time complexity of such algorithms is often exponential in the number of features, meaning they are too computationally intensive for most practical data analysis problems. In this work, we show that a lasso-based empirical risk minimizer over a growing set of basis functions retains its asymptotic rate even if fit only on a random, relatively small subset of the basis. Applying this idea, we propose RandomHAL: a fast approximation to the HAL estimator that retains its desirable properties, yet can be fit on datasets with many more features. The empirical performance of RandomHAL is evaluated using simulation experiments.

1 Introduction

Across various fields of statistics, especially semiparametrics and causal inference, estimating a parameter often involves learning the form of one or more nuisance functions. For instance, in these settings, a common estimator involves computing

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i; f_n) \tag{1}$$

where X_i represents some data, ϕ is an influence function, and f_n is an estimate of a nuisance function f or collection thereof. See, for instance, [11, 14, 10, 21, 23, 22, 6].

In order to converge asymptotically to a normal distribution centered around the true parameter, estimators such as the influence function-based estimator in Equation 1 require f_n to converge to f "fast enough". Often, ϕ is "rate doubly-robust", meaning that $f = (f_1, f_2)$ and asymptotic normality is achieved if the product of both functions' rates is $o_P(n^{-1/2})$. In other words, in order for the estimator to be consistent and

asymptotically normal, both nuisance estimates must satisfy $\|f_{nk} - f_k\| = o_P(n^{-1/4})$ at the slowest.

Since the forms of nuisance functions are usually unknown, one would prefer to avoid imposing too many unnecessary assumptions when learning f_n. While parametric models converge at rate $n^{-1/2}$, they often require very strict assumptions about the form f. Hence, it is often preferable to assume that f lies in a broader function class, and learn it using a more flexible statistical learning algorithm that has been proven to converge in that class.

Unfortunately, the best known rates of convergence for many popular algorithms such as various Random Forests [4, 2, 24] or Bayesian Additive Regression Trees (BART) [7, 16] often suffer from the "asymptotic curse of dimensionality" – their proven rates depend heavily on the dimension of the data, meaning that, in the absence of sufficient smoothness, they will converge slower than the required $o_P(n^{-1/4})$ on datasets involving more than even a few covariates. These slow rates often arise due to an assumption commonly enforced when analzing these algorithms that f lies in a Lipschitz or Hölder class. Table 1 displays some common rates given for various function classes.

Function Class	Rough Optimal Rate	Example
Parametric	$n^{-1/2}$	Linear Regression
Additive	$n^{-1/3}$	GAM
Cadlag with finite sectional variation	$n^{-1/3}(\log n)^{3(d-1)/2}$	HAL
Lipschitz/Hölder	$n^{-\alpha/(d+\alpha)}$	Random Forest

Table 1: Rates for function classes used to analyze common regression models

How can an algorithm match the required $o_P(n^{-1/4})$ as closely as possible while imposing as few assumptions as possible? One solution is the Highly Adaptive Lasso (HAL) [1, 19, 20]. This algorithm assumes that f lies in a special function class: the collection of càdlàg (right-continuous with left limits) functions with bounded sectional variation norm. All functions in this class can be represented as a lasso model over a high-dimensional indicator basis; by solving this lasso problem, HAL converges at rate $o_P(n^{-1/4})$ (technically $o_P(n^{-1/3}\log^{d/n} n)$, which is slightly faster [3]) thereby breaking the curse of dimensionality. Independently, this rate was discovered to be minimax-optimal [8].

Clearly, HAL has many convenient properties for semiparametric inference. The fatal downside, however, is that enumerating the entire HAL basis requires a space complexity of $O(2^d n)$, where d is the number of covariates. This exponential dependence on the covariates becomes computationally intractable for problems with more than just a few dimensions, making the "vanilla" implement of HAL impractical to use in many real-world data analysis problems.

But is it really necessary to enumerate the entire HAL basis to fit a model that achieves the $o_P(n^{-1/4})$ rate? In this manuscript, we prove the contrary: that even a lasso model fit over a random, relatively small $o(n \log n)$ subset of the HAL basis can achieve the same $o_P(n^{-1/4})$ asymptotic rate. In fact, this result holds in general: given any empirical risk minimizer over a large basis, solving the same problem over

a sufficiently large random subset of the basis will suffice to preserve the statistical asymptotic rate of the original estimator. In addition, we implement this procedure over the HAL basis, which we term "RandomHAL", and study its finite-sample performance using numerical simulation.

By providing a fast approximation to HAL, our work makes it possible to fit nuisance models at the proper rate necessary for many estimators, especially in semiparametric statistics and causal inference, that otherwise might be computationally intractable. Our results extend to similar empirical risk minimization problems with other types of regularization, such as ridge, ElasticNet, or the adaptive lasso, as well as other large collections of basis functions besides the HAL basis. In the next section, we introduce the mathematical foundations of this fast rate-preserving approximation method.

2 Mathematical foundation

2.1 Set-up

Suppose our data consists of observations (X_i, Y_i) for $i = 1, \ldots, n$ where Y_i is an outcome of interest and $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})$ is a vector of covariates. In this manuscript we use the empirical process notation Pf and P_n to denote integration over the ground-truth probability measure P and empirical measure P_n, respectively. That is,

$$Pf = \int f(x) dP(x) \text{ and } P_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tag{2}$$

Let Lf denote some loss function of f; a common example is the squared loss $Lf = (y - f(x))^2$. Now consider learning a function $f^\star$ that minimizes the mean loss PL over a function class $\mathcal{F}$, possibly subject to some regularization norm:

$$f^\star = \arg\min_{f \in \mathcal{F}} PLf + \lambda \|f\| \tag{3}$$

The *empirical minimizer* f_n is an estimator minimizing the *empirical* mean of the loss $f_n = \arg\min_{f \in \mathcal{F}} P_n Lf + \lambda \|f\|$. Typically, one can show that the estimator f_n converges to $f^\star$ in some norm at a given rate: $\|f^\star - f_n\| = o_P(r_n)$, where r_n is the rate of convergence.

2.2 Early-stopping estimators

When a function f_n is an empirical minimizer (such as the lasso), in many cases, it may not be possible to fully realize or compute. For example, if f_n has no closed-form solution, it is typically fit using an iterative procedure for which some stopping point must be chosen. Or, f_n may be too computationally intensive to fit exactly, so some approximation may be chosen. In such cases, we consider an *early-stopping estimator* $f_n^{(m)}$, where the parameter m denotes the degree of approximation. By the triangle inequality, the error of this estimator is bounded by

$$\|f^\star - f_n^{(m)}\| \leq \underbrace{\|f^\star - f_n\|}_{\text{estimation error}} + \underbrace{\|f_n - f_n^{(m)}\|}_{\text{approximation error}} = o_P(r_n) + o_P(r_m) \tag{4}$$

where

- r_n is the *estimation rate* of f_n to $f^\star$
- r_m is the *approximation rate* of $f_n^{(m)}$ to f_n

This means $\|f^\star - f_n^{(m)}\| = o_P(\min(r_n, r_m))$; achieving the same asymptotic rate using an approximation requires m to be set such that the approximation error converges faster than the estimation error. Note that [17] showed that if f_n is in a Donsker class, then $f_n^{(m)}$ will be too. We will use this idea approximate a full Lasso model using a random subset of features – specifically, the HAL model.

2.3 RandomHAL Algorithm

As an example of early-stopping estimator, in this section we describe our RandomHAL algorithm, which approximates a full HAL model by sampling only a small subset of the full HAL basis. The HAL algorithm performs the following optimization:

$$\arg\min_{\beta} \left\{ L(\Phi\beta) + \lambda\|\beta\|_1 \right\} \tag{5}$$

where Φ is a K-order spline basis with all possible knots and interactions; that is,

$$\Phi(x)\beta = \sum_{k=1}^{K} \sum_{i=1}^{n} \underbrace{\sum_{S \subset \{1,\ldots,d\}}}_{\substack{\text{all possible} \\ \text{subsets of} \\ \text{covariates}}} \beta_{i,S} \cdot \prod_{j \in S} (x_j - X_{i,j})^k \tag{6}$$

The most common form of HAL is 0-order HAL, which in the rest of the article we refer to as just "HAL" for readability. This HAL basis takes the form

$$\Phi(x)\beta = \sum_{i=1}^{n} \sum_{S \subset \{1,\ldots,d\}} \beta_{i,S} \cdot 1(x_S < X_{i,S}) \tag{7}$$

RandomHAL is a procedure that fits a HAL model over a randomly selected subset of Φ. The algorithm proceeds as follows:

1. Sample m basis functions $\phi_{S,i}(x)$ from $\Phi(x)$
2. Fit lasso over sub-basis $\Phi_m(x)$ to solve for β_{lasso}
3. $f_n^{(m)} = \Phi_m(x)\beta_{\text{lasso}}$

While instantiating Φ in memory requires $O(2^d n)$ space complexity, RandomHAL only requires $O(nm)$. Clearly, RandomHAL can be much more efficient even when d is of moderate size, if m is small. But how can we prove that RandomHAL retains the same rate as HAL? And what should we set m? How many basis functions must be sampled to preserve the rate? In the next section, we provide a general theorem that can be used to show that even only sampling $m = n \log n$ basis functions is sufficient for the rate of RandomHAL to match that of HAL.

2.4 Preserving statistical rates using a random subset of a basis

The lasso and similar learning algorithms are typically fit using coordinate descent [9], an algorithm that iteratively optimizes over each individual feature $j = 1, \ldots, d$ holding the others fixed. To analyze the performance of the Lasso over a random subset of features, we will lean on analysis from several authors of *random coordinate descent* [18, 13, 15], which optimizes coordinates randomly one at a time. For this setting, we adapt the work of [15].

In our setting, we want to find a vector of d parameters $\beta = [\beta_j]_{j=1}^d$ that minimize the loss L. Consider an objective function consisting of a *loss* and a decomposable regularizer:

$$L(\beta) = \underbrace{\ell(\beta)}_{\text{smooth}} + \underbrace{\sum_{j=1}^{d} \psi_j(\beta_j)}_{\text{decomposable}} \tag{8}$$

with the following regularity conditions:

1. $\sum_{j=1}^d \psi_j$ is a convex regularizer.
2. ℓ is a differentiable function with a coordinate-wise Lipschitz gradient $\nabla \ell$; that is, $\nabla \ell$ satisfies $\|\nabla_j \ell(\beta + t_j) - \nabla_j \ell(\beta)\|_{(j)} \leq C_i \|t\|_{(j)}$ for t_j all zero except for in the jth coordinate, and ∇_j denoting the jth element of the gradient.
3. ℓ satisfies a *restricted strong convexity* condition: if $t \in \mathbb{C}$ must hold for some $\mathbb{C}$, then $\ell(\beta + t) - \ell(\beta) \geq \langle \nabla \ell(\beta), t \rangle + \frac{\mu}{2} \|t\|_2^2$ is true for some "strong convexity parameter" μ for all $t \in \mathbb{C}$

Restricted strong convexity is a central condition in the analysis of regularized M-estimators in high-dimensional statistics [12]. In lasso or compressed sensing problems, it is satisfied for design matrices that satisfy *restricted isometry* [5] or *restricted eigenvalue* [25] conditions. If ℓ is convex but not strongly convex, we can always add a shrinking ridge penalization $\frac{\lambda}{n} \|\beta\|_2^2$ to make it convex [26]; asymptotically, this penalty will converge to 0 and yield the same penalization as the original loss.

Theorem 1 (Error bound of Random Coordinate Descent with uneven sampling)**.** *Suppose each basis function j is sampled with probability p_j. After running random coordinate descent for m iterations, the average difference in loss satisfies*

$$E\Big(L(\beta^{(m)}) - L(\beta^\star)\Big) \leq \Big(1 - C_\mu \max p_j\Big)^m \Big(L(\beta^{(0)}) - L(\beta^\star)\Big) \tag{9}$$

where C_μ is a constant depending only on μ, the strong convexity parameter of the loss L.

Proof. This proof follows similarly to those of Theorem 6 and Lemma 5 of [15], but adapted to account for possible non-uniform sampling. To begin, we establish an initial bound. Consider a random vector U_j which has a j entry is equal to 1 and all other entries equal to 0. By the Lipschitz property,

$$L(\beta + U_j t) \leq \ell(\beta) + V_j(\beta, t) + \sum_{k \neq j} \psi_k(\beta_k) \tag{10}$$

where $V_j(\beta, t) = \langle \nabla_j \ell(\beta), t\rangle + \frac{L_i}{2}\|t\|_j^2 + \psi_j(\beta_j + t)$. Note that the update step of coordinate descent can be expressed in terms of V_j by writing $\beta_j \leftarrow T_j(\beta^{(m)})$ where

$$T_j(\beta^{(m)}) = \arg\min \Big\{ V_j(\beta^{(m)}, t) : t \in R_j \Big\} \tag{11}$$

Next, we bound the error of a single update step in the random coordinate descent algorithm. By the definition of expectation, plugging in the bound above we have

$$E\Big(L(\beta^{(m+1)}) \mid \beta^{(m)}\Big) = \sum_{j=1}^d p_j \Big(L(\beta^{(m)} + U_j T_j(\beta^{(m)}))\Big)$$

$$\leq \sum_{j=1}^d p_j \Big(\ell(\beta^{(m)}) + V_j(\beta^{(m)}, T_j(\beta^{(m)})) + \sum_{k \neq j} \psi_k(\beta_k)\Big)$$

Now, note that

$$\sum_{j=1}^d p_j \sum_{k \neq j} \psi_k(\beta_k) = \sum_{j=1}^d (1 - p_j)\psi_j(\beta_j) = \sum_{j=1}^d \psi_j(\beta) - \sum_{j=1}^d p_j \psi_j(\beta)$$

Substituting the above into the previous bound, we can note

$$E\Big(L(\beta^{(m+1)}) \mid \beta^{(m)}\Big) \leq \Big(\ell(\beta_k) + \sum_{i=1}^d \psi_j(\beta^{(m)})\Big) \tag{12}$$

$$+ \Big(\max_j p_j\Big) \sum_{j=1}^d \Big(V_j(\beta^{(m)}, T_j(\beta^{(m)}) - \psi_j(\beta^{(m)})\Big) \tag{13}$$

$$\tag{14}$$

The first term is just $L(\beta^{(m)})$, and the second can be bounded in terms of the distance from $\beta^\star$. Let $G(\beta, t) = \sum_{j=1}^d \Big(V_j(\beta, t) - \psi_j(\beta)\Big)$. This function has two convenient

properties. First, by the definition of inner product,

$$G(\beta, t) = \sum_{j=1}^{d} \Big(\langle \nabla_j \ell(\beta), t \rangle + \frac{L_i}{2} \|t\|_j^2 + \psi_j(\beta_j + t_j) - \psi_j(\beta) \Big)$$
$$= \langle \nabla \ell(\beta), t \rangle + \|t\|_L^2 + \sum_{j=1}^{d} \Big(\psi_j(\beta_j + t_j) - \psi_j(\beta_j) \Big)$$

Second, if $T = [T_1(\beta), \ldots, T_d(\beta)]$ then $G(\beta, T)$ has, by the definition of T, the minimizing property

$$G(\beta, T) = \min_{t \in \mathbb{C}} G(\beta, t)$$

Therefore, by restricted strong convexity,

$$\min_{t \in \mathbb{C}} G(\beta, t) = \min_{t \in \mathbb{C}} \langle \nabla \ell(\beta), t \rangle + \|t\|_L^2 + \sum_{j=1}^{d} \Big(\psi_j(\beta_j + t_j) - \psi_j(\beta_j) \Big)$$
$$\leq \min_{t \in \mathbb{C}} \ell(\beta + t) - \ell(\beta) + \frac{1 - \mu}{2} \|t\|_L^2 + \sum_{j=1}^{d} \Big(\psi_j(\beta_j + t_j) - \psi_j(\beta_j) \Big)$$
$$= \min_{t \in \mathbb{C}} L(\beta + t) - L(\beta) + \frac{1 - \mu}{2} \|t\|_L^2$$

Then, if $t = \tilde{\beta} - \beta$, we can re-express $\tilde{\beta}$ in terms of a mixture of $\beta^\star$ and β using the convexity of L:

$$\min_{t \in \mathbb{C}} L(\beta + t) - L(\beta) + \frac{1 - \mu}{2} \|t\|_L^2$$
$$\leq \min_{\alpha \in [0,1]} L(\alpha \beta^\star - (1 - \alpha)\beta) + \frac{1 - \mu}{2} \|\alpha \beta^\star + (1 - \alpha)\beta - \beta)\|_L^2$$
$$\leq \min_{\alpha \in [0,1]} \alpha (L(\beta) - L(\beta^\star)) + \frac{\alpha^2 (1 - \mu)}{2} \|\beta - \beta^\star\|_L^2$$

to which minimizing over α and applying strong convexity of L (which follows from the strong convexity of ℓ and convexity of all ψ_j) yields

$$\min_{\alpha \in [0,1]} \alpha (L(\beta) - L(\beta^\star)) + \frac{\alpha^2 (1 - \mu)}{2} \|\beta - \beta^\star\|_L^2 = -C_\mu (L(\beta) - L(\beta^\star))$$

where $C_\mu = 1 - \mu/4$ if $\mu > 2$ and $1/\mu$ otherwise [15]. Applying this bound to the previous expression,

$$E\Big(L(\beta^{(m+1)}) - L(\beta^\star) \mid \beta^{(m)} \Big) \leq \Big(1 - C_\mu \max_j p_j \Big) \Big(L(\beta^{(m)}) - L(\beta^\star) \Big)$$

Finally, applying this bound recursively over all m iterations of the algorithm, we find the expected result that

$$E\Big(L(\beta^{(m)}) - L(\beta^\star)\Big) \leq \Big(1 - C_\mu \max_j p_j\Big)^m \Big(L(\beta^{(0)}) - L(\beta^\star)\Big)$$

□

2.5 Choosing m to match a convergence rate

Now let us use the previous result to approximate HAL and other models that optimize a loss over some basis growing in n to achieve a fast convergence rate. If we choose our sampling scheme so $p_j = w_j/n$, then

$$E\Big(L(\beta^{(m)}) - L(\beta^\star)\Big) \leq \Big(1 - \frac{C_\mu \max_j w_j}{n}\Big)^m \Big(L(\beta^{(0)}) - L(\beta^\star)\Big) \tag{15}$$

Hence, if we choose $m = \omega(n\log(n))$ (that is, m is bounded below by $n\log(n)$),

$$\begin{aligned} E\Big(L(\beta^{(m)}) - L(\beta^\star)\Big) &\leq \Big(1 - \frac{C}{n}\Big)^{n\omega(\log n)} \Big(L(\beta^{(0)}) - L(\beta^\star)\Big) \\ &\approx \exp(C)^{-\omega(\log n)} \Big(L(\beta^{(0)}) - L(\beta^\star)\Big) = o\Big(\frac{1}{n}\Big) \end{aligned}$$

Taking the norm of the loss difference, we have that $\|E(L(\beta^{(m)}) - L(\beta^\star))\| = o(n^{-1/2})$, clearly sufficient to ensure the approximation rate r_m of $f_n^{(m)}$ converges at faster than even the typical parameter rate. For HAL, this is clearly faster than the required $r_m = o(n^{-1/4})$, and so RandomHAL using an approximation basis with roughly $m = n\log n$ should preserve its rate of convergence.

To summarize, the theory above tells us that a random coordinate descent over greater than $n\log(n)$ basis functions will be sufficient to preserve the asymptotic convergence rate of HAL – or any lasso estimator over a basis that grow proportionally to n. The error attained by choosing $n\log(n)$ basis functions to start and then fitting coordinate descent over them must be bounded above by the error of taking $n\log(n)$ sequential steps. As a consequence, fitting HAL over a random (and possibly non-uniformly sampled) subset of $n\log(n)$ basis functions achieves the same asymptotic rate as fitting HAL over the full basis.

2.6 A note on simulation strategy

Sampling with uneven probability is necessary for RandomHAL because many higher-order interaction terms will be mostly composed of 0 entries, and therefore not be selected in the Lasso procedure. For example, if $d = 15$, sampling uniformly will result in a basis subsample of almost all 5-10th-order interactions, which are unlikely to be useful for estimation compared to main terms or 1st-order interactions. Hence, a sampling scheme that prioritizes lower-order interactions is necessary. While our result holds asymptotically, intelligent choices of sampling probability will most likely be

necessary to overcome the constant C in finite samples. Selecting each p_j to minimize the finite-sample error bound above may be an interesting strategy to pursue in future research.

3 Numerical Simulations

Despite being asymptotically correct, does RandomHAL perform well in finite samples? In this section, we compare HAL and RandomHAL with uniform sampling on a simulation setup with the DGP

$$\begin{aligned} X1 \sim \text{Beta}(2,2), X2 \sim \text{Beta}(2,4), X3 \sim \text{Beta}(4,2), \\ X4 \sim \text{Beta}(4,4), X5 \sim \text{Bernoulli}(0.5) \\ A \sim \text{Bernoulli}(\text{logistic}(L_1 L_5 + L_1 + L_1^2 + L_2 - L_2^2 + L_3 + L3^2 + L_4 - L_4^2 - 3.5)) \\ Y \sim \text{N}(A + L_5 A + 2 L_3 L_4 + (L_1 - L_1^{3/2} + L_2 - L_2^{3/2} + L_3 - L_3^{3/2} + L_4 - L_4^{3/2}, 0.3) \end{aligned}$$

Figure 1 displays the out-of-sample MSE attained by predicting from the outcome model using RandomHAL versus HAL. Figure 2 shows the convergence results of both algorithms for estimating the ATE of the above DGP using HAL to learn both the outcome model and the propensity score, with all values averaged over 200 iterations.

From these simulations, we can see that for prediction, Figure 1 demonstrates that RandomHAL achieves almost identical MSE to HAL. Furthermore, Figure 2 shows that one-step ATE estimates using RandomHAL achieve low bias, that their bias and MSE converge at the rates predicted by theory, and that nominal coverage is nearly achieved at large samples. In fact, on all metrics, RandomHAL actually exhibits an improvement over HAL. Such results demonstrate that random basis approximations can work well in practice.

We also evaluate RandomHAL on a DGP with too many covariates for a full HAL model to be fit in a timely manner – fifteen in total, evaluated over 400 simulations of:

$$\begin{aligned} X_1, \ldots, X_5 &\sim \text{MVN}(j, 0.25(I_5 + 1_{5\times 5})) \\ X_6, \ldots X_{10} &\sim \text{Multinomial}(10, (j-6)/15) \\ p &\sim \text{Beta}(2,2), \\ X_{11}, \ldots X_{15} &\sim \text{Bernoulli}(p) \\ A &\sim \text{Ber}(\text{expit}(\mu)/10) \\ Y &\sim \text{Normal}(\mu, 1.0) \end{aligned}$$

where $\mu = \sum_{j=1}^{5} \sqrt{|X_j|} + \sum_{j=1}^{5} X_{j+5} X_{j+10} - 5$. The MSE of fitting an outcome model using RandomHAL to learn Y from this data and predicting on unseen samples is shown in Figure 3. The operating characteristics of the one-step estimator using an outcome regression and propensity score learned by RandomHAL on this data are shown in Figure 4. In these simulations, we employ an uneven sampling scheme, where lower-order interactions are more likely to be sampled than higher-order ones.

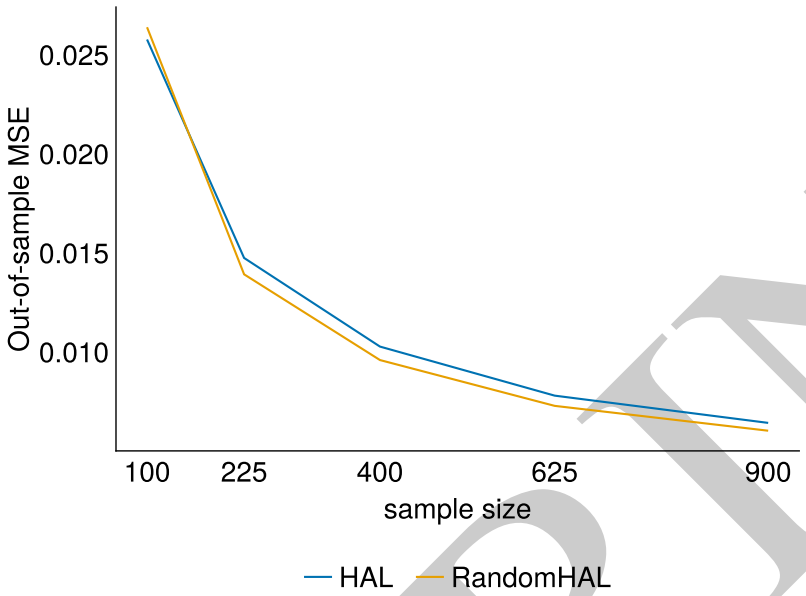

Figure 1: Out-of-sample MSE for HAL vs RandomHAL outcome regression.

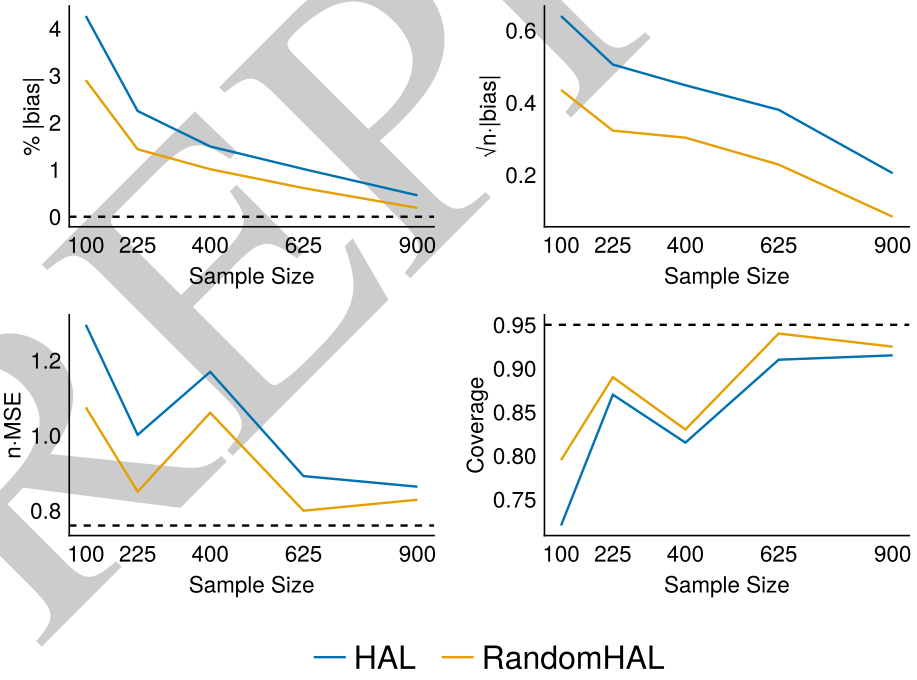

Figure 2: One-Step ATE estimates using HAL and RandomHAL to estimate outcome regression and propensity score nuisance parameters.

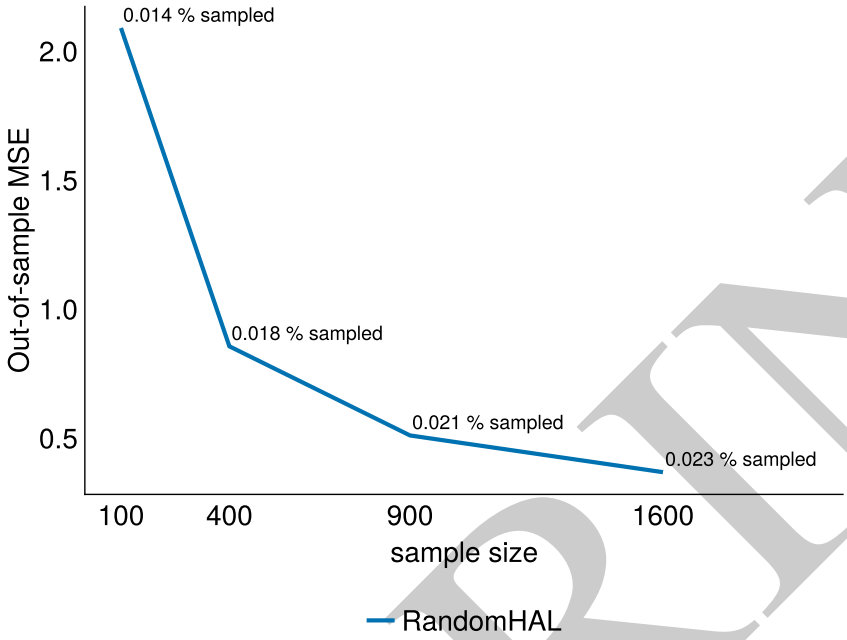

Figure 3: Out-of-sample MSE for RandomHAL outcome regression in a DGP that would be computationally intractable for HAL.

Similar to those results shown previously, we can see that the theory for RandomHAL also holds in these high-dimensional simulations. At just 100 samples, fitting a full HAL model would require instantiating over 3 million basis functions – but just from sampling roughly 0.02% of the possible basis functions, the out-of-sample MSE of the RandomHAL outcome regression decreases rapidly with sample size. Furthermore, the RandomHAL-based ATE estimator also achieves low bias, its bias and MSE converge at the expected rate, and close to nominal coverage is achieved. Even when the number of covariates is too large to fit a full HAL model, RandomHAL can be used to achieve the necessary nuisance estimation rates for valid causal inference.

4 Conclusion

In this work, we showed that when a regularized minimum loss-based estimator attains a desirable statistical rate, but has too many parameters to be computationally tractable, it is possible to construct an approximation that preserves its statistical rate while performing minimization over only a small subset of those parameters. We apply this idea to the Highly Adaptive Lasso (HAL), which performs l_1-regularized loss minimization over a set of basis functions of size $O(n \cdot 2^d)$. Our approach, RandomHAL, is able to preserve HAL's asymptotic $o_P(n^{-1/4})$ rate by fitting the lasso over a random subset of just over $n \log n$ components of the basis. Despite relying on asymptotic results, our numerical experiments demonstrate that this approach works well even in finite

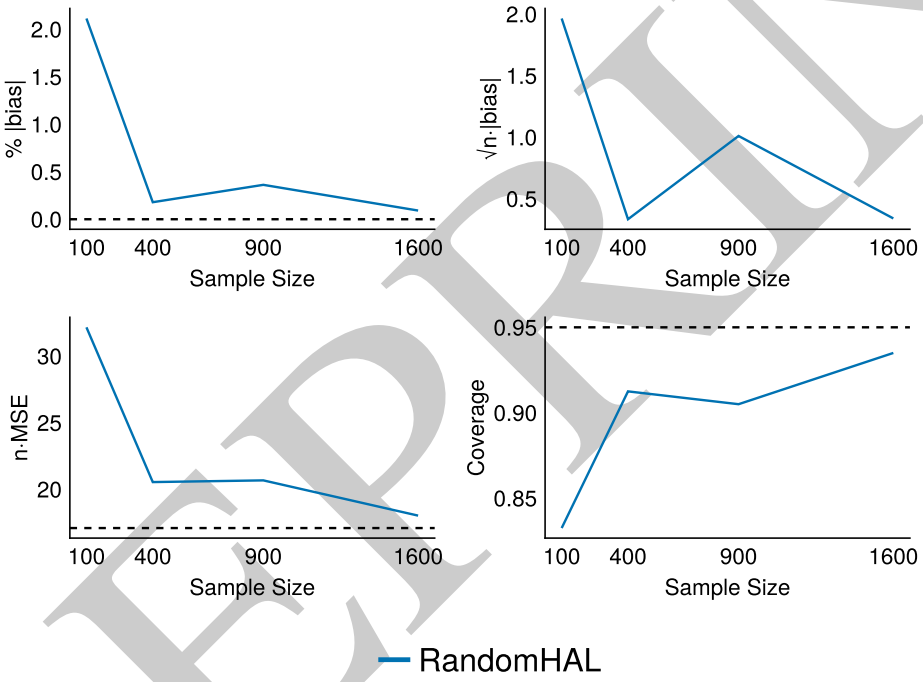

Figure 4: One-Step ATE estimates using RandomHAL to estimate outcome regression and propensity score nuisance parameters in a DGP that would be computationally intractable for HAL.

samples, and actually can improve over a full HAL fit.

That said, this method of rate preservation for random basis approximation is limited in several ways. When the constant $C_\mu \max_j w_j$ is very small, the error bound will remain large even at "galactically huge" finite sample sizes. Future work should attempt to achieve tighter bounds on the finite sample approximation error for specific problems such as the classical lasso. Relatedly, it may be that many basis functions are irrelevant to the estimation problem. Developing more optimal sampling schemes to select basis functions that are more likely to be relevant, perhaps based on Markov chains that update probabilities based on previous samples, would improve the performance of RandomHAL and other random basis approximations.

Finally, despite being faster than HAL, the time complexity of RandomHAL is still $O(n^2 \log n)$ – much slower than the $O(n \log n)$ time complexity of a decision tree or MARS algorithm. Finding algorithmic improvements to speed up computation in n would go a long way to making HAL an even more viable option for statistical learning. Such improvements would enable more theoretically-justified semiparametric statistics and causal machine learning.

References

[1] D. Benkeser and M. van der Laan. The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, page 689–696. IEEE, Oct. 2016. doi: 10.1109/dsaa.2016.93.

[2] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

[3] A. F. Bibaut and M. J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv*, 2019. doi: 10.48550/ARXIV.1907.09244.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/a:1010933404324.

[5] E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, Dec. 2005. ISSN 0018-9448. doi: 10.1109/tit.2005.858979. URL http://dx.doi.org/10.1109/TIT.2005.858979.

[6] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, Jan. 2018. ISSN 1368-423X. doi: 10.1111/ectj.12097. URL http://dx.doi.org/10.1111/ectj.12097.

[7] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), Mar. 2010. ISSN 1932-6157. doi: 10.1214/09-aoas285. URL http://dx.doi.org/10.1214/09-AOAS285.

[8] B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *The Annals of Statistics*, 49(2), Apr. 2021. ISSN 0090-5364. doi: 10.1214/20-aos1977.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL http://dx.doi.org/10.18637/jss.v033.i01.

[10] C. A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562, 1987.

[11] Y. A. Koshevnik and B. Y. Levit. On a non-parametric analogue of the information matrix. *Theory of Probability & Its Applications*, 21(4):738–753, 1977.

[12] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4), Nov. 2012. ISSN 0883-4237. doi: 10.1214/12-sts400. URL http://dx.doi.org/10.1214/12-STS400.

[13] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, Jan. 2012. ISSN 1095-7189. doi: 10.1137/100802001. URL http://dx.doi.org/10.1137/100802001.

[14] J. Pfanzagl and W. Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

[15] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1–2):1–38, Dec. 2012. ISSN 1436-4646. doi: 10.1007/s10107-012-0614-z. URL http://dx.doi.org/10.1007/s10107-012-0614-z.

[16] V. Ročková and S. van der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48 (4), Aug. 2020. ISSN 0090-5364. doi: 10.1214/19-aos1879. URL http://dx.doi.org/10.1214/19-AOS1879.

[17] A. Schuler, Y. Li, and M. van der Laan. Lassoed tree boosting. 2022. doi: 10.48550/ARXIV.2205.10697. URL https://arxiv.org/abs/2205.10697.

[18] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(52):1865–1892, 2011. URL http://jmlr.org/papers/v12/shalev-shwartz11a.html.

[19] M. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics*, 13 (2), Oct. 2017. ISSN 1557-4679. doi: 10.1515/ijb-2015-0097.

[20] M. van der Laan. Higher order spline highly adaptive lasso estimators of functional parameters: Pointwise asymptotic normality and uniform convergence rates. *arXiv*, 2023. doi: 10.48550/ARXIV.2301.13354.

[21] M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003. doi: 10.1007/978-0-387-21700-0.

[22] M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*, volume 4. Springer, 2011.

[23] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[24] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, June 2018. ISSN 1537-274X. doi: 10.1080/01621459.2017.1319839.

[25] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[26] X. Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. 2018. doi: 10.48550/ARXIV.1803.06573. URL `https://arxiv.org/abs/1803.06573`.