

Research Article

Zeyi Wang, Wenxin Zhang, and Mark van der Laan*

Super Ensemble Learning Using the Highly-Adaptive-Lasso

Abstract: We consider estimation of a functional parameter of a realistically modeled data distribution based on observing independent and identically distributed observations. Suppose that the true function is defined as the minimizer of the expectation of a specified loss function over its parameter space. We assume that we are provided J estimators of the true function, which in our proposed approach can also be viewed as a data adaptive coordinate-transformation for the true function. For any J -dimensional real valued cadlag function with finite sectional variation norm, we define a candidate ensemble estimator as the mapping from the data into the composition of the cadlag function and the J estimated functions. Using V -fold cross-validation, we define the cross-validated empirical risk of each cadlag function specific ensemble estimator. We then define the Meta Highly Adaptive Lasso Minimum Loss Estimator (M-HAL-MLE) as the cadlag function that minimizes this cross-validated empirical risk of the cadlag function specific ensemble over all cadlag functions with a uniform bound on the sectional variation norm (and respecting the parameter space of functional parameter), where the latter bound is a tuning parameter. For each of the V training samples, this yields an estimate of the true function defined as the composition of the M-HAL-MLE ensemble and the J estimated functions trained on the training sample. We can estimate the true function with the average of these V estimated functions, which we call the M-HAL super-learner, and we estimate a pathwise differentiable target feature of the true function with the corresponding plug-in estimator, or with an average of the V plug-in estimated functions.

Beyond generally defining this M-HAL super-learner, this article concerns the analysis of 1) the M-HAL super-learner as an estimator of the true function, and, of 2) a target feature of the M-HAL super-learner as an estimator of the target feature of the true function. We show that 1) the M-HAL super-learner converges to the oracle estimator at a rate $n^{-2/3}$ (up till $\log n$ -factor) w.r.t. excess risk, where the oracle estimator is defined as the ensemble of the J estimators that minimizes the excess risk among all considered ensembles; 2) by choosing the sectional variation norm large enough (undersmoothing) in the M-HAL-MLE, the target feature of the M-HAL super-learner is an asymptotically linear estimator of the target feature of the oracle estimator; 3) the excess risk of the oracle estimator and true function will generally be second order; 4) the difference of a target feature of the oracle estimator and target feature of true function will be of same order as this oracle excess risk. As a consequence, under weak conditions on the J candidate estimators, target features of the undersmoothed M-HAL super-learner are asymptotically linear estimators of the corresponding target features of true function, with influence curve either the efficient influence curve, or potentially, a super-efficient influence curve.

Keywords: Asymptotically linear estimator, canonical gradient, cross-validation, dimension reduction, efficient influence curve, highly adaptive lasso (HAL), influence curve, meta-learning, minimum loss estimation, pathwise differentiable target parameter, sectional variation norm, super-efficiency, super-learning, transformation of variables.

1 Introduction

We consider estimation of a functional parameter of a realistically modeled data distribution based on observing independent and identically distributed observations of a d -dimensional Euclidean valued random variable. Suppose that the true function is a k -variate real valued function defined as the minimizer over its parameter space of the expectation of a specified uniformly bounded loss function.

HAL-MLE: In our previous work [1–4] we showed that if the parameter space consists of k -variate real valued cadlag functions with a universal bound on the sectional variation norm [1, 5], then the MLE over this parameter space, selecting the variation norm bound with cross-validation, converges to the true function at a rate $n^{-2/3}(\log n)^d$ w.r.t. the loss-based dissimilarity, also called excess risk. Moreover, computation of this estimator corresponds with minimizing an empirical risk over a linear combination of spline-basis functions under the constraint that the L_1 -norm of the coefficient vector is bounded by this sectional variation norm bound, making it a high dimensional lasso estimation problem.

A general theory for undersmoothing sieve based estimators is developed in Shen [6, 7], and a powerful demonstration of such an estimator also presented in [8]. We have shown that an undersmoothed HAL-MLE that selects the L_1 -norm larger than the cross-validation selector (but still bounded) according to a specified global undersmoothing criterion is asymptotically efficient [9] for any pathwise differentiable target parameter, under weak regularity conditions [10]. Thus, the smooth target features of the HAL-MLE were shown to be asymptotically efficient for the target features of the true function, if the sectional variation norm satisfies a global undersmoothing criterion, even though this HAL-MLE is not targeted towards that target feature.

Super learner with small family of ensembles: In other past research, we have proposed super-learning as a general optimal approach to learn a true function [11–15]. The super-learner selects a library of estimators, defines a collection of ensembles such as all convex combinations, and chooses the ensemble that minimizes the cross-validated empirical risk of the ensemble specific candidate estimator. For simplicity, let's consider the case that we use V -fold sample splitting, so that the cross-validated empirical risk is defined as the average over the V sample splits in training and validation sample of the empirical mean over the validation sample of the loss function at the ensemble specific estimator trained on the training sample. In particular, the discrete super-learner (the collection of ensembles is trivially defined as the set of estimators in the library) simply selects the estimator in the library that minimizes the cross-validated empirical risk. Given the cross-validated selected ensemble, one could either rerun the selected ensemble of estimators based on the whole data set, or one can simply take the average over the V sample splits of the selected ensemble of the estimators based on the training sample only. The later is immediate available as a by product of the cross-validated empirical risk of the cross-validation selector of the ensemble. In this article we will represent the super-learner as this average across V sample splits.

Asymptotic equivalence of super-learner (with small family of ensembles) with oracle selected ensemble: Under some constraints on the size of the family of ensembles, the excess risk of the super-learner divided by the excess risk of the oracle selected ensemble converges to 1 as sample size increases (see references above): we say that the cross-validation selected ensemble (i.e., super-learner) is asymptotically equivalent with the oracle selected ensemble. The oracle ensemble minimizes the excess risk among all ensemble specific estimators applied to data sets of the size of the training sample. In particular, if one uses the discrete super-learner, then it is asymptotically equivalent with the oracle selected estimator that minimizes the excess risk, as long as the number of candidate estimators is polynomial in sample size. When the family of ensembles is defined as all convex combinations, the asymptotic equivalence will require the number of estimators in the library to grow slowly with sample size (i.e., $\log n$). By including the HAL-MLE in its library, it follows that the discrete super-learner converges at least as fast as $n^{-2/3}(\log n)^d$, while being adaptive to unknown structure of the true function if other candidate estimators are tailored to such structure. However, this gain of adaptivity of this super-learner relative to HAL-MLE comes also at a price in the sense that pathwise differentiable target features of this super-learner (even when it includes the

HAL-MLE in its library) are not asymptotically linear estimators of the true features, so that one cannot provide formal statistical inference.

Meta-HAL super-learner: In this article, we consider an aggressive super-learner that defines as family of ensembles a class of multivariate real valued cadlag functions with a bound on its sectional variation norm. We refer to this super-learner as the Meta-HAL super-learner since the computation of the cross-validation selector of the ensemble corresponds with applying the HAL-MLE at a meta-level data set in which each observation O_i is coupled with a cross-fitted realization of the library of estimators.

Size of family controlled by sectional variation norm bound: Due to the large size of this family of ensembles, we should not expect asymptotic equivalence of the cross-validation selector with the oracle selector anymore, so that our results are new relative to the current super-learner literature. In fact, if the J estimated functions represent a zero-loss transformation of the coordinate system for true function, then the oracle estimator equals the true function. The potential overfitting of the M-HAL-MLE in finite samples is controlled by the choice of sectional variation norm, thereby making it a practically powerful super-learner. In particular, we can select this sectional variation norm bound C with a discrete super-learner with library the C -specific M-HAL super-learners across a range of values C , thereby guaranteeing that our cross-validation selector $C_n = C_{n,cv}$ will be asymptotically equivalent with the oracle selector of C .

Due to large family of ensembles, candidate estimators can also represent a data adaptive coordinate-transformation: By selecting such a large family of ensembles, the library of estimators does not need to be restricted to good estimators of the functional parameter, but could include fixed functions, and simple parametric model based estimators, not necessarily approximating or even aiming to approximate the true function. That is, one could also view the library of estimators as a proposed data adaptive transformation of the coordinates $x \in \mathbb{R}^k$ for the true function. For example, one extreme choice of transformation would be to simply propose k fixed functions $f_1(x), \dots, f_k(x)$ of x so that the coordinate-transformation is invertible and thereby represents a zero-loss transformation. One of these fixed functions could even be a fit from a previous study. An interesting variation is that we include the fixed simple functions $f_j(x) = x_j$, $j = 1, \dots, k$, and add a super-learner as the $k + 1$ -th element in the library, thereby specifying an over parametrized set of functions of x , but one that potentially allows for a representation of the true function as a relatively simple (low sectional variation norm) cadlag function of $(f_1(x), \dots, f_{k+1}(x))$. Of course, one still has the option to define the library as k highly adaptive estimators targeting the true function. Therefore, we also refer more generally to the library of estimators as a data adaptive coordinate-transformation for the true function (so that it is clear they do not have to be estimators of the true function).

M-HAL super-learner behaves as an HAL-MLE for a transformed data problem with coordinates the J cross-fitted functions: Our results will demonstrate that the M-HAL super-learner of the true function will generally behave as an HAL-MLE of the J -dimensional oracle cadlag function for a transformed data problem in which the J coordinates for this oracle function are the cross-fitted J estimated functions applied to the original x . Due to the transformed coordinates, the true oracle cadlag function might be a simpler function of the J variables than the true function is of the original input variables, which is formally compared by contrasting the sectional variation norms of the two functions. This makes the M-HAL super-learner a potentially superior estimator.

Target features of M-HAL super-learner: In this article, analogue to our work on the HAL-MLE [10], we will also analyze the undersmoothed M-HAL super-learner, selecting the sectional variation norm bound in the meta-HAL-MLE larger than the value suggested by cross-validation. In this case, we establish that under the meta-level analogue of the global undersmoothing criterion for the HAL-MLE in [10], target features of the M-HAL-SL are either asymptotically efficient, or potentially super-efficient, depending on the coordinate-transformation implied by library of J estimators. Either way, we will establish asymptotic linearity with known influence curve implied by canonical gradient of the pathwise derivative of the target parameter, and thereby allowing for formal statistical inference.

Thus, contrary to the regular discrete or convex ensemble super-learner, this highly aggressive super-learner using an undersmoothed HAL-MLE in the meta-learning step results in asymptotically linear plug-in estimators of target features, just as we have shown for the regular plug-in HAL-MLE. Therefore, the

M-HAL super-learner is not only at least as powerful as a regular (small family of ensembles) super-learner, but, when undersmoothed, its smooth features are asymptotically linear, super-efficient or efficient.

Targeted HAL-MLE in meta-learning step: In an upcoming tech report [16] we also proposed a targeted-HAL-MLE (T-HAL-MLE) defined by enforcing in the definition of the HAL-MLE an additional constraint, beyond the sectional variation norm bound, defined as the empirical mean of the efficient influence curve equation for the target parameter being equal to zero (or its Euclidean norm small enough). We showed that this targeting preserves the rate of convergence of the HAL-MLE; that it is now asymptotically efficient for the target parameter without need for undersmoothing; and that it is still asymptotically efficient for other target parameters under the global undersmoothing condition. By using the analogue T-HAL-MLE of the true oracle cadlag function at the meta-level, the corresponding T-M-HAL super-learner of our true function has the same above mentioned asymptotic properties as the T-HAL-MLE. Such meta-level targeted MLE will be important extension of this work, achieving similar asymptotic performance without undersmoothing.

Using the M-HAL super-learner or undersmoothed M-HAL super-learner as initial estimator in TMLE: Of course, one could also decide to use the M-HAL super-learner as a highly powerful initial estimator of the functional parameter in the TMLE of a target parameter, in which case no undersmoothing will be needed [17–21].

When using as the initial estimator an undersmoothed M-HAL super-learner, we can also design this TMLE to preserve the score equations solved by the initial estimator. For example, one can choose to orthogonalize the efficient score with respect to the solved scores, or to jointly solve all the desired score equations with a one-step update along multivariate submodels [21]. Similar to an undersmoothed T-M-HAL super-learner, such score-preserving TMLE can also efficiently estimate other smooth features that it did not target. In other words, the targeting step of a score-preserving TMLE does not destroy the properties of the undersmoothed M-HAL super-learner for plug-in estimation of features that it did not target. Although not the topic of this article, we suggest that undersmoothing the M-HAL super-learner when using it as initial estimator in the TMLE (or undersmoothing the T-M-HAL super-learner) might also benefit the target feature (even though the estimator is already targeted towards this feature). That is, this undersmoothing will generally improve the second order behavior of the resulting plug-in estimator of the target parameter, by solving additional score equations that shrink the size of its exact second order remainder [22].

1.1 Organization of article

In Section 2 and Section 3 we formally define the statistical estimation problem, the relevant quantities, and the M-HAL super-learner. We also provide a transformation/reduction of the observed data implied by the library of J estimators, and corresponding statistical estimation problem addressed by the meta-learning step, where the latter treats the J -dimensional vector of cross-fitted estimates as a fixed J -dimensional coordinate-transformation. The latter reduced data statistical estimation problem will essentially make our study of the M-HAL-MLE cross-validation selector equivalent with our previous study of the HAL-MLE. In Section 4 we analyze the excess risk of the M-HAL super-learner. The excess risk will be decomposed as a sum of the excess risk of the M-HAL super-learner relative to the oracle estimator (i.e., the best possible ensemble/cadlag function of the J candidate estimators among all cadlag functions with sectional variation norm bounded by our bound), and the excess risk of the oracle estimator relative to the true function (which would be zero if the coordinate-transformation is a zero-loss transformation). In Section 5 we analyze a target feature of M-HAL super-learner as estimator of the target feature of true function. The difference of this plug-in estimator with the true target estimand is decomposed as the sum of 1) the difference of plug-in M-HAL super-learner with the plug-in oracle estimator, and 2) the difference of the plug-in of oracle estimator and the true target estimand. The latter is analyzed separately in Section 5.2, and is shown to be a second order difference, while the first is analyzed analogue to our previous analysis of the plug-in HAL-MLE [10]. In Section 6 we apply the formulations to an example, the nonparametric estimation of a

treatment specific mean. In this section we will also demonstrate our general results applied to this example, so that the reader obtains a comprehensive concrete understanding of the general results in this article through this example. We conclude with a discussion in Section 8.

Basic outline of proofs are presented in the main article, while more technical results are presented in a self-contained way in the Appendix. Appendix A and Appendix B presents a notation index that should help the reader, even though notation will be introduced in main article. Appendix F provides that undersmoothing makes the M-HAL super-learner solve the cross-validated empirical mean of the efficient influence curve equation, representing the only real challenge for establishing asymptotic linearity. In particular, we discuss in detail the two undersmoothing conditions (12) and (13). Appendix G generalizes the consistency and asymptotic linearity results to the targeted M-HAL SL. Finally, Appendix J provides deeper understanding of the undersmoothing condition (12), and that it can be easily achieved with bounded selectors of the sectional variation norm.

2 Statistical Model

Suppose we observe n independent and identically distributed copies O_1, \dots, O_n with probability measure P_0 that is known to be an element of a statistical model \mathcal{M} . We assume that $O \in \mathbb{R}^d$ is a d -variate bounded random variable. Let P_n be the empirical probability measure that puts mass $1/n$ on each O_i . We consider a functional parameter $Q : \mathcal{M} \rightarrow \mathcal{Q} \equiv \{Q(P) : P \in \mathcal{M}\}$, where the parameter space \mathcal{Q} represents a set of multivariate $[0, 1]$ -valued functions; that is, $Q(P) : \mathbb{R}^k \rightarrow [0, 1], \forall P \in \mathcal{M}$. In practice, the target function is often defined on a bounded Euclidean set; therefore, without loss of generality we assume that variables can be standardized so that $Q(P) : [0, 1]^k \rightarrow [0, 1], \forall P \in \mathcal{M}$. Let $L : \mathcal{Q} \rightarrow L^2(P_0)$ be a mapping from the parameter space \mathcal{Q} into a set of d -variate real valued functions of O satisfying $Q_0 = \arg \min_{Q \in \mathcal{Q}} P_0 L(Q)$. We refer to L as a loss function, where $L(Q)(o)$ evaluates a loss of candidate Q at observation o . Let $d_0(Q, Q_0) \equiv P_0 L(Q) - P_0 L(Q_0)$ be the loss-based dissimilarity, which denotes the excess risk of an estimator Q with respect to the class \mathcal{Q} . It is assumed that $M_1 = \sup_{Q \in \mathcal{Q}} |L(Q)(o)| < \infty$ and $M_{20} = \sup_{Q \in \mathcal{Q}} \frac{P_0(L(Q) - L(Q_0))^2}{d_0(Q, Q_0)} < \infty$, so that the cross-validation selector is well behaved and generally asymptotically equivalent with the oracle selector [11, 12].

We will consider a pathwise differentiable target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}$. It is assumed that it is pathwise differentiable at P with canonical gradient $D^*(P)$, and that $\Psi(P)$ only depends on P through $Q(P)$. Let $G : \mathcal{M} \rightarrow \mathcal{G}$ be a functional parameter so that $D^*(P)$ only depends on P through $Q(P)$ and $G(P)$. We will also denote $D^*(P)$ with $D^*(Q, G)$ and $\Psi(P)$ with $\Psi(Q)$. Let $R_{20}(Q, G, Q_0, G_0) \equiv \Psi(Q) - \Psi(Q_0) + P_0 D^*(Q, G)$ be the exact second order remainder implied by the canonical gradient. Let $L_1(G)$ be a loss function for $G_0 = \arg \min_{G \in \mathcal{G}} P_0 L_1(G)$, and $d_{01}(G, G_0) = P_0 L_1(G) - P_0 L_1(G_0)$.

Our goal is to construct an estimator of Q_0 , and to also use it to construct asymptotically linear estimators of $\Psi(Q_0)$, possibly for arbitrary Ψ in a large class of smooth features.

2.1 Data Adaptive Coordinate-Transformation with V -fold Cross-Validation

We split the n observations O_1, \dots, O_n in V blocks, and for each choice v of a block, let $P_{n,v}^1$ be the empirical measure of the observations O_i in that block, and let $P_{n,v}$ be the empirical measure of the observations O_i in the other $V - 1$ blocks. We refer to $P_{n,v}^1$ and $P_{n,v}$ as the empirical measures of the validation and training sample for the v -th sample split, respectively.

Let $\hat{\mathbf{Q}} = (\hat{Q}_j : j = 1, \dots, J)$ be a collection of J algorithms, possibly a library of estimators of Q_0 , but it is only required that \hat{Q}_j maps data into an element of the parameter space \mathcal{Q} . For an empirical measure $P_{n,v}$ of a training sample extracted from $\{O_1, \dots, O_n\}$, $\mathbf{Q}_{n,v} = \hat{\mathbf{Q}}(P_{n,v}) \in \mathcal{Q}^J$ represents its realization applied to the empirical probability measure $P_{n,v}$. Let \mathcal{M}_{np} denote the set of discrete empirical measures based on an arbitrary subset of $\{O_1, \dots, O_n\}$ so that $\hat{\mathbf{Q}} : \mathcal{M}_{np} \rightarrow \mathcal{Q}^J$ represents a vector of estimators that can be

applied to arbitrary training samples extracted from O_1, \dots, O_n . Let \mathbf{Q}_n be a function of (v, x) defined by $\mathbf{Q}_n(v, x) = \mathbf{Q}_{n,v}(x)$. Note that $\mathbf{Q}_{n,v} : [0, 1]^k \rightarrow [0, 1]^J$ constructs a data adaptive coordinate-transformation for Q_0 , which is a realization of the algorithm $\hat{\mathbf{Q}}$ applied to the v -th training sample.

For observation O_i , let v_i be the index of the block that contains the i -th observation O_i , $i = 1, \dots, n$. We represent our data set with $(v_1, O_1), \dots, (v_n, O_n)$. One can represent this sample as an i.i.d. sample from the true distribution $P_0^{\bar{V}}$ of a random variable (\bar{V}, O) , where the conditional distribution of O , given $\bar{V} = v$, equals P_0 , and $\bar{V} \sim U(1, \dots, V)$ is uniform on $\{1, \dots, V\}$. In this manner, for a function $f(v, o)$ of (v, o) , we can write $P_0^{\bar{V}} f = \frac{1}{V} \sum_{v=1}^V \int f(v, o) dP_0(o)$. Let $P_n^{\bar{V}}$ be the empirical measure putting mass $1/n$ on each (v_i, O_i) , $i = 1, \dots, n$.

In the next section, we will construct HAL-MLE with meta-level data, in which each observation O_i is coupled with the cross-fitted algorithm realizations \mathbf{Q}_{n,v_i} .

2.2 Family of Ensembles

We define the set of ensembles as a class of cadlag functions with a bound on the sectional variation norm. That is, consider a collection $\mathcal{Q}^r \subset \mathcal{D}_{C^u}[0, 1]^J$ of J -variate real valued cadlag functions $Q^r : [0, 1]^J \rightarrow [0, 1]$, with a uniform bound C^u on its sectional variation norm. We have $Q^r(x) = Q^r(0) + \sum_{s \subset \{1, \dots, J\}} \int \phi_{s, u_s}(x) dQ_s^r(u_s)$, where $\phi_{s, u_s}(x) = I(x_s \geq u_s)$, $Q_s^r(u_s) = Q^r(u_s, 0_{-s})$ is the s -specific section that sets the coordinates in s^c equal to zero [2]. The sectional variation norm of Q^r is defined as $\|Q^r\|_v^* = |Q^r(0)| + \sum_{s \subset \{1, \dots, J\}} \int |dQ_s^r(u_s)|$.

For any $Q^r \in \mathcal{Q}^r$ and $\mathbf{Q} \in \mathcal{Q}^J$, a Q^r -ensemble of \mathbf{Q} is given by $x \mapsto Q^r \circ \mathbf{Q}(x) = Q^r(\mathbf{Q}(x))$. We assume that any ensemble estimator constructed by $Q^r \in \mathcal{Q}^r$ respects the parameter space \mathcal{Q} . Specifically, for each $v = 1, \dots, V$,

$$\{Q^r \circ \mathbf{Q}_{n,v} : Q^r \in \mathcal{Q}^r\} \subset \mathcal{Q}. \quad (1)$$

In practice, one can consider additive HAL models with respect to a subset of the most nonparametric ensemble space, \mathcal{Q}^r , for potentially better finite sample performance. For example, a hyperparameter can restrict the size of the s section so that $\{Q^r \in \mathcal{Q}^r : Q^r(x) = Q^r(0) + \sum_{s \subset \{1, \dots, J\}, |s| \leq U} \int \phi_{s, u_s}(x) dQ_s^r(u_s)\} \subset \mathcal{Q}^r$ involves only up to U -th order interactions. Such hyperparameters, that further restrict the class of ensembles for the additive HAL models, can be decided jointly along with C^u by a cross-fitted discrete super learner similar to Section 3.4.

2.3 Data Reduction Implied by Cross-fitted Coordinate-Transformation

Note we can view $L(Q^r \circ \mathbf{Q}_n) : (v, O) \mapsto L(Q^r \circ \mathbf{Q}_{n,v})(O)$ as a function of (v, O) . In many settings X is a subvector of O , and $L(Q)(O)$ involves evaluating the function Q at X , and Q^r is only a function of original coordinates $X \subset O$ through transformations $\mathbf{Q}_{n,v}(X)$. In general, for any given L and \mathbf{Q}_n , there exists a data reduction of (\bar{V}, O) ,

$$O^r = O^r(\bar{V}, O),$$

such that $L(Q^r \circ \mathbf{Q}_n)(v, O)$ depends on (v, O) or $\mathbf{Q}_{n,v}(X)$ only through $(v, O^r(v, O))$. This allows us to define a loss for Q^r with the reduced data,

$$L^r(Q^r)(v, O^r(v, O)) \equiv L(Q^r \circ \mathbf{Q}_n)(v, O). \quad (2)$$

For example, if $L(Q)(X, Y) = (Y - Q(X))^2$, $O = (X, Y)$, then $L(Q^r \circ \mathbf{Q}_n)(v, O) = (Y - Q^r \circ \mathbf{Q}_{n,v}(X))^2$ depends only on (v, O) or $\mathbf{Q}_{n,v}(X)$ through $(v, O^r(v, O))$ with $O^r(v, O) = (X_v^r \equiv \mathbf{Q}_{n,v}(X), Y)$; so we can define $L^r(Q^r)(v, O^r) = (Y - Q^r(X_v^r))^2$.

Note that (\bar{V}, O^r) represents a resulting data reduction of (\bar{V}, O) implied by the cross-fitted coordinated transformations \mathbf{Q}_n . Let d^r be the dimension of O^r .

Let P_0^r be the distribution of (\bar{V}, O^r) implied by $P_0^{\bar{V}}$, treating \mathbf{Q}_n as a fixed function. Similarly, let $\mathcal{M}^r = \{P^r : P \in \mathcal{M}\}$ be the statistical model for the distribution P_0^r of (\bar{V}, O^r) implied by the statistical model \mathcal{M} for P_0 . Let P_n^r be the empirical probability measure of (v_i, O_i^r) , where $O_i^r = Q^r(v_i, O_i)$ is the data reduction of O_i implied by \mathbf{Q}_{n,v_i} . We also use notation $\mathcal{M}_v^r = \{P_v^r : P \in \mathcal{M}\}$, where P_v^r denotes the distribution of $O_v^r = O^r(v, O)$ implied by $O \sim P$. Let $Q^r : \mathcal{M}^r \rightarrow \mathcal{Q}$ be a parameter of $P^r \in \mathcal{M}^r$ such that $Q^r(P^r) = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r} P^r L^r(Q^r)$.

3 Meta-Level Learning Using Highly Adaptive Lasso

3.1 M-HAL-MLE

For a given ensemble $Q^r \in \mathcal{Q}^r$, define the cross-validated empirical risk of corresponding Q^r -specific candidate estimator $Q^r \circ \hat{\mathbf{Q}} : \mathcal{M}_{np} \rightarrow \mathcal{Q}$ of Q_0 by

$$P_n^{\bar{V}} L(Q^r \circ \mathbf{Q}_n) = \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 L(Q^r \circ \mathbf{Q}_{n,v}).$$

Define the C_n -specific meta-level HAL-MLE (M-HAL-MLE) as the cross-validation selector among all ensembles of J estimators,

$$Q_n^r = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^{\bar{V}} L(Q^r \circ \mathbf{Q}_n) = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 L(Q^r \circ \mathbf{Q}_{n,v}).$$

The oracle ensemble targeted by the M-HAL-MLE is given by

$$Q_{0,n}^r = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r} P_0^{\bar{V}} L(Q^r \circ \mathbf{Q}_n) = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r} \frac{1}{V} \sum_{v=1}^V P_0 L(Q^r \circ \mathbf{Q}_{n,v}).$$

Note that Q_n^r is the empirical estimator of $Q_{0,n}^r$ defined by replacing $P_0^{\bar{V}}$ by its empirical counterpart $P_n^{\bar{V}}$, so that Q_n^r is a regular HAL-MLE of $Q_{0,n}^r$.

3.2 M-HAL-SL

For each coordinate-transformation $\mathbf{Q}_{n,v}$, the M-HAL-MLE ensemble defines an estimator $Q_{n,v} \equiv Q_n^r \circ \mathbf{Q}_{n,v}$ of Q_0 , $v = 1, \dots, V$. We also use the notation $Q_n \equiv Q_n^r \circ \mathbf{Q}_n$ for the function $Q_n(v, x) \equiv Q_{n,v}(x)$ which codes each of these V estimates of Q_0 . The C_n -specific meta-level HAL super-learner (M-HAL-SL) of Q_0 refers to either Q_n or to its average across splits given by

$$\bar{Q}_n(x) \equiv P_n^{\bar{V}} Q_n(\bar{V}, x) = \frac{1}{V} \sum_{v=1}^V Q_n(v, x).$$

Sometimes, to emphasize its dependence on the selector C_n , we denote these estimators with $Q_n^{C_n}$ and $\bar{Q}_n^{C_n}$, respectively. Note that, if the parameter space \mathcal{Q} is convex, $\bar{Q}_n \in \mathcal{Q}$. We will see that for estimation of smooth features $\Psi(Q_0)$ of Q_0 , one can either use the smooth feature of the average \bar{Q}_n , $\Psi(\bar{Q}_n)$, or use the average across the sample splits of $Q_{n,v}$,

$$\Psi^{\bar{V}}(Q_n) \equiv \frac{1}{V} \sum_v \Psi(Q_{n,v}),$$

as the difference is second order.

Similarly, the oracle ensemble defines an oracle estimator, $Q_{0,n} \equiv Q_{0,n}^r \circ \mathbf{Q}_n$, so that $Q_{0,n}(v, x) = Q_{0,n,v}(x) \equiv Q_{0,n}^r \circ \mathbf{Q}_{n,v}(x)$, $v = 1, \dots, V$. We will view this function $Q_{0,n}$ as a parameter of the distribution $P_0^{\bar{V}}$ of (\bar{V}, O) . In addition, we use the notation $\bar{Q}_{0,n}(x) \equiv \frac{1}{V} \sum_{v=1}^V Q_{0,n}(v, x)$. We note that Q_n and \bar{Q}_n are estimators of $Q_{0,n}$ and $\bar{Q}_{0,n}$, respectively.

The excess risk of the M-HAL-SL Q_n is defined as

$$d_0^{\bar{V}}(Q_n, Q_0) = P_0^{\bar{V}} L(Q_n) - P_0^{\bar{V}} L(Q_0) = \frac{1}{V} \sum_{v=1}^V \{P_0 L(Q_{n,v}) - P_0 L(Q_0)\}.$$

The excess risk of the M-HAL-SL \bar{Q}_n is given by $d_0(\bar{Q}_n, Q_0) = P_0 L(\bar{Q}_n) - P_0 L(Q_0)$, and for convex loss functions $L(Q)$, we have

$$d_0(\bar{Q}_n, Q_0) \leq d_0^{\bar{V}}(Q_n, Q_0).$$

Therefore, it suffices to analyze the excess risk $d_0^{\bar{V}}(Q_n, Q_0)$ of the M-HAL-SL Q_n , which is decomposed as

$$\begin{aligned} d_0^{\bar{V}}(Q_n, Q_0) &= P_0^{\bar{V}} L(Q_n^r \circ \mathbf{Q}_n) - P_0^{\bar{V}} L(Q_{0,n}^r \circ \mathbf{Q}_n) \\ &\quad + P_0^{\bar{V}} L(Q_{0,n}^r \circ \mathbf{Q}_n) - P_0^{\bar{V}} L(Q_0) \\ &\equiv d_0^{\bar{V}}(Q_n, Q_{0,n}) + d_0^{\bar{V}}(Q_{0,n}, Q_0). \end{aligned}$$

The first term, $d_0^{\bar{V}}(Q_n, Q_{0,n})$, involves comparing M-HAL-MLE Q_n^r with the oracle ensemble $Q_{0,n}^r$. Viewing it as a function of a given set of cross-fitted coordinate-transformations, we also denote this as $d_0(Q_n^r, Q_{0,n}^r)$, the loss-based dissimilarity of an HAL-MLE Q_n^r of $Q_{0,n}^r$. The second term, $d_0^{\bar{V}}(Q_{0,n}, Q_0)$, represents the dissimilarity between the oracle selected ensemble of \mathbf{Q}_n and the true function Q_0 .

Just as our decomposition above, we will also decompose $\Psi^{\bar{V}}(Q_n) - \Psi(Q_0)$ as the sum of the difference of the target feature of Q_n and the oracle estimator $Q_{0,n}$, $\Psi^{\bar{V}}(Q_n) - \Psi^{\bar{V}}(Q_{0,n})$, and the difference of the target feature of the oracle estimator and true target estimand, $\Psi^{\bar{V}}(Q_{0,n}) - \Psi(Q_0)$. Similarly, $\Psi(\bar{Q}_n) - \Psi(Q_0) = \Psi(\bar{Q}_n) - \Psi(\bar{Q}_{0,n}) + \Psi(\bar{Q}_{0,n}) - \Psi(Q_0)$. Both terms will be analyzed separately, where the “bias” (conditional on the training sample, it is truly a bias) of the target feature of the oracle estimator, $\Psi(Q_{0,n}) - \Psi(Q_0)$, will be shown to be second order, or even zero when $\mathbf{Q}_{n,v}$ are zero-loss coordinate-transformations.

3.3 Equivalent Formulation of Statistical Parameters Using Reduced Data

Given a data reduction $(\bar{V}, O^r) \sim P_0^r$ of $(\bar{V}, O) \sim P_0^{\bar{V}}$ as specified in Section 2.3, M-HAL-MLE of oracle ensemble is regular HAL-MLE with the reduced data. Specifically, with the corresponding reduced data loss function $L^r : \mathcal{Q}^r \rightarrow L^2(P_0^r)$ that satisfies (2), we have that the oracle ensemble

$$Q_{0,n}^r = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r} P_0^r L^r(Q^r)$$

can be represented as a functional of the reduced data distribution P_0^r , and our M-HAL-MLE of $Q_{0,n}^r$,

$$Q_n^r = \operatorname{argmin}_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^r L^r(Q^r),$$

is given by the HAL-MLE fitted with the reduced data $(v_i, O_i^r) \sim P_n^r$. This demonstrates that Q_n^r can be implemented as a standard HAL-MLE based on the reduced data (v_i, O_i^r) , $i = 1, \dots, n$, and loss function $L^r(Q^r)$. Denote the loss-based dissimilarity with reduced data as $d_0^r(Q^r, Q_{0,n}^r) = P_0^r L^r(Q^r) - P_0^r L^r(Q_{0,n}^r)$.

In our model \mathcal{M}^r for reduced data (\bar{V}, O^r) implied by the cross-fitted coordinate-transformation \mathbf{Q}_n , $\Psi^{\bar{V}}(Q^r \circ \mathbf{Q}_n)$ can be viewed as a parameter $\Psi^r : \mathcal{Q}^r \rightarrow \mathbb{R}$ defined by

$$\Psi^r(Q^r) = \frac{1}{V} \sum_{v=1}^V \Psi(Q^r \circ \mathbf{Q}_{n,v})$$

evaluated at the M-HAL-MLE Q_n^r . So now we have both $d_0^{\bar{V}}(Q_n, Q_{0,n}) = d_0^r(Q_n^r, Q_{0,n}^r)$ and $\Psi^{\bar{V}}(Q_n) = \Psi^r(Q_n^r)$. Therefore, the performance of the averaged plug-in estimator $\Psi^{\bar{V}}(Q_n)$ is decided by Q_n^r as HAL-MLE of oracle ensemble $Q_{0,n}^r$, treating the coordinate-transformation \mathbf{Q}_n as fixed.

Let $D^r(P^r)(\bar{V}, O^r)$ be the canonical gradient of Ψ^r at P^r . Let $G^r(P^r) \in \mathcal{G}^r$ be a nuisance parameter such that $D^r(P^r) = D^r(Q^r, G^r)$. We assume that $G^r(\cdot)$ given $\bar{V} = v$ defines an element G_v^r in \mathcal{G} , the parameter space for G . For example, if G is a function of X and G^r is a function of $\mathbf{Q}_n(\bar{V}, X)$, then conditional on $\bar{V} = v$ we may define $G_v^r(x) = G^r(\mathbf{Q}_{n,v}(x)) \in \mathcal{G}$; although in practice the definition is flexible depending on G^r and G . We assume the following link between $D^r(P^r)$ and $D^*(P)$:

$$D^r(Q^r, G^r)(v, O^r(v, O)) = D^*(Q^r \circ \mathbf{Q}_{n,v}, G_v^r)(O). \quad (3)$$

Condition (3) essentially states that for a fixed v , the reduced data structure O_v^r has the same structure as O , and as a result P_v^r has same model structure as P , so that the pathwise derivative of $Q^r \rightarrow \Psi(Q^r \circ \mathbf{Q}_{n,v})$ has the same structure as $Q \rightarrow \Psi(Q)$. For example, the general formula for the canonical gradient of $EE(Y | A = 1, W)$ with a nonparametric model remains the same in terms of the (conditional) densities of Y , A , and W , regardless of the dimension or definition of W . Similarly, the canonical gradient of a treatment specific mean $EY_{\bar{a}}$ for general longitudinal data structure $O = (L(0), A(0), \dots, L(K), A(K), Y = L(K+1))$ has the same general form in terms of all the conditional densities regardless of the dimension of definition of $L(k)$.

Recall that the true $Q_{0,n}^r = Q^r(P_0^r)$ and $G_{0,n}^r \equiv G^r(P_0^r)$ are indexed by a subscript n to emphasize dependence on coordinate-transformation \mathbf{Q}_n . Let $L_1^r(G^r)$ be a loss function for $G_{0,n}^r = \arg \min_{G^r \in \mathcal{G}^r} P_0^r L_1^r(G^r)$ and let $d_{01}^r(G^r, G_{0,n}^r) = P_0^r L_1^r(G^r) - P_0^r L_1^r(G_{0,n}^r)$, where \mathcal{G}^r is the parameter space for G^r . Let $R_{20}^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) = \Psi^r(Q^r) - \Psi^r(Q_{0,n}^r) + P_0^r D^r(Q^r, G^r)$ be the exact second order remainder. By (3) it follows that

$$R_{20}^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) = \frac{1}{V} \sum_{v=1}^V R_{20}(Q^r \circ \mathbf{Q}_{n,v}, G_v^r, Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r).$$

3.4 Cross-Validation Selector of the Sectional Variation Norm

Note that M-HAL-MLE and M-HAL-SL are fitted as functions of P_n across a V -fold sample splitting $(P_{n,v}, P_{n,v}^1 : v = 1, \dots, V)$ of P_n , and are indexed by a hyperparameter C for the sectional variation norm bound enforced on the ensembles $Q^r \in \mathcal{Q}^r$. Therefore, we can denote the algorithms by $Q_n^{r,C} : \mathcal{M}_{np} \rightarrow \mathcal{Q}^r$ and $\bar{Q}_n^C : \mathcal{M}_{np} \rightarrow \mathcal{Q}$, such that

$$\bar{Q}_n^C(P_n) = \frac{1}{V} \sum_{v=1}^V Q_n^{r,C}(P_n) \circ \mathbf{Q}_n(P_{n,v}).$$

We can define the cross-validation selector of C by

$$C_{n,cv} = \arg \min_C \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 L(\bar{Q}_n^C(P_{n,v})).$$

Note that this involves double cross-validation similar to the cross-validated risk of a regular super-learner. This cross-validation selector $C_{n,cv}$ is asymptotically equivalent with the oracle selector of C optimizing excess risk.

If we fix the realized coordinate-transformations $\mathbf{Q}_{n,v}$, $v = 1, \dots, V$, then the M-HAL-SL can be defined as an algorithm using fixed functions, independent of what data is provided. This defines the approximation

$$\bar{Q}_n^{C,fast}(P_n) = \frac{1}{V} \sum_{v=1}^V Q_n^{r,C}(P_n) \circ \mathbf{Q}_{n,v}.$$

The cross-validation selector for this algorithm is then given by:

$$C_{n,cv}^{fast} = \operatorname{argmin}_C \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 L(\bar{Q}_n^{C,fast}(P_{n,v})).$$

This algorithm still requires V times running the HAL-MLE Q_n^r at the meta-level on the training samples $P_{n,v}$, but it does not require rerunning the estimators \hat{Q}_j , $j = 1, \dots, J$. This could be used as a fast approximation of the double cross-validation selector with the risk of slightly overfitting the univariate hyperparameter C . We believe this criterion will still provide a good ranking and thereby selector $C_{n,cv}^{fast}$, especially as an initial value of an undersmoothing selector for the purpose of plug-in estimations.

3.5 Summary of Assumptions

We assume the analogue of $M_{20} < \infty$ for loss $L(Q)$ for the reduced data loss $L^r(Q^r)$. For that purpose, define $Q_{0,Q}^r = \arg \min_{Q^r \in \mathcal{Q}^r} P_0 L(Q^r \circ Q)$, and let

$$M_2^r \equiv \sup_{Q \in \mathcal{Q}^J} \sup_{Q^r \in \mathcal{Q}^r} \frac{P_0^r \{L(Q^r \circ Q) - L(Q_{0,Q}^r \circ Q)\}^2}{P_0^r \{L(Q^r \circ Q) - L(Q_{0,Q}^r \circ Q)\}} < \infty. \quad (4)$$

We also need that various classes functions of O_v^r are contained in the class of cadlag functions with bound on sectional variation norm. Therefore, it is convenient to let $\mathcal{D}_{dr}[0, \tau^r]$ represent a class of d^r -variate real valued cadlag functions on a cube $[0, \tau^r]$ with a universal bound (also over realizations of $\{Q_{n,v} : v\}$) on the sectional variation norm. Let $d_0^r((Q^r, G^r), (Q_{0,n}^r, G_{0,n}^r)) \equiv d_0^r(Q^r, Q_{0,n}^r) + d_{01}^r(G^r, G_{0,n}^r)$ be the loss-based dissimilarity for the joint (Q^r, G^r) . We make the usual assumption that the exact second order remainders can be bounded in terms of this loss-based dissimilarity.

The summarize the key assumptions throughout this article as follows

$$\begin{aligned} D^r(Q^r, G^r)(v, O_v^r = O^r(v, O)) &= D^*(Q^r \circ Q_{n,v}, G_v^r)(O) \\ M_2^r &< \infty \\ \{L^r(Q^r) : Q^r \in \mathcal{Q}^r\} &\subset \mathcal{D}_{dr}[0, \tau^r] \\ \{D^r(Q^r, G_{0,n}^r) : Q^r \in \mathcal{Q}^r\} &\subset \mathcal{D}_{dr}[0, \tau^r] \\ R_{20}(Q, G, Q_0, G) &= O(d_0(Q, Q_0)) \\ P_0\{D^*(Q, G) - D^*(Q_0, G)\}^2 &= O(d_0(Q, Q_0)) \\ R_{20}^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) &= O(d_0^r((Q^r, G^r), (Q_{0,n}^r, G_{0,n}^r))) \\ P_0^r\{D^r(Q^r, G^r) - D^r(Q_{0,n}^r, G_{0,n}^r)\}^2 &= O(d_0^r((Q^r, G^r), (Q_{0,n}^r, G_{0,n}^r))) \\ \sup_{\{Q_{n,v}:v\}, v} P_0\{D^*(Q_{0,n,v}, G_{0,n,v}^r) - D^*(Q_0, \tilde{G}_0)\}^2 &\rightarrow_p 0 \end{aligned} \quad (5)$$

for a limit $\tilde{G}_0 \in \mathcal{G}$ of $G_{0,n,v}^r$, $v = 1, \dots, V$. The last assumption is a universal consistency condition for the asymptotic normality of $n^{1/2} P_n^r D^r(Q_n^r, G_{0,n}^r)$. The other assumptions in general hold true once we enforce strong positivity so that $D^*(P_0)$ is a uniformly bounded function on a support of O . We will refer to this whole set of assumptions as assumption (5).

4 Convergence Rate of M-HAL-SL

Consider $d_0^{\bar{V}}(Q_n, Q_0) = d_0^{\bar{V}}(Q_n, Q_{0,n}) + d_0^{\bar{V}}(Q_{0,n}, Q_0)$. By assuming that $L(Q)$ is a convex loss function so that $P_0 L(\sum_j \alpha_j Q_j) \leq \sum_j \alpha_j P_0 L(Q_j)$ for α -vectors with $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$, we have $d_0(\bar{Q}_n, Q_0) \leq d_0^{\bar{V}}(Q_n, Q_0)$, so that our results imply the same rate result for $d_0(\bar{Q}_n, Q_0) = P_0 L(\bar{Q}_n) - P_0 L(Q_0)$.

The following lemma establishes the rate of convergence result for $d_0^{\bar{V}}(Q_n, Q_{0,n})$ (proof in Appendix C).

Lemma 4.1. *Recall assumption (5). We have*

$$d_0^{\bar{V}}(Q_n, Q_{0,n}) = O_P(n^{-2/3}(\log n)^{d^r}).$$

This yields the following result for $d_0^{\bar{V}}(Q_n, Q_0)$ and thereby for $d_0(\bar{Q}_n, Q_0)$.

Theorem 4.2. *Recall assumption (5). We have*

$$\begin{aligned} d_0^{\bar{V}}(Q_n, Q_0) &= d_0^{\bar{V}}(Q_{0,n}, Q_0) + d_0^{\bar{V}}(Q_n, Q_{0,n}) \\ &= \min_{Q^r \in \mathcal{Q}^r} \frac{1}{V} \sum_{v=1}^V P_0\{L(Q^r \circ \mathbf{Q}_{n,v}) - L(Q_0)\} + O_P(n^{-2/3}(\log n)^{d^r}). \end{aligned}$$

If $\hat{\mathbf{Q}} = (\hat{Q}_j : j)$ includes an estimator \hat{Q}_j such that $d_0(\hat{Q}_j(P_{n,v}), Q_0) = O_P(n^{-2/3}(\log n)^d)$, then it follows that

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

The leading term in $d_0^{\bar{V}}(Q_n, Q_0)$ represents the dissimilarity between the oracle estimator $Q_{0,n}^r \circ \mathbf{Q}_{n,v}$ and Q_0 . Since \mathcal{Q}^r includes the functions $f(x) = x_j$, $j = 1, \dots, J$, this leading term can be bounded by

$$\min_j \frac{1}{V} \sum_{v=1}^V P_0\{L(\hat{Q}_j(P_{n,v})) - L(Q_0)\}.$$

However, note that the oracle estimator is generally a much better estimator than one of the candidates in the library of J estimators. In fact, $d_0^{\bar{V}}(Q_{0,n}, Q_0)$ even equals zero for many coordinate-transformations.

5 Asymptotic Linearity of Target Features of Undersmoothed M-HAL-MLE

We can estimate $\Psi(Q_0)$ with $\Psi(\bar{Q}_n)$ or $\frac{1}{V} \sum_{v=1}^V \Psi(Q_{n,v})$. Under regularity conditions, the Taylor expansion at \bar{Q}_n gives that the difference between these two plug-in estimators will generally be second order

$$\begin{aligned} \frac{1}{V} \sum_{v=1}^V \Psi(Q_{n,v}) - \Psi(\bar{Q}_n) &= \frac{1}{V} \sum_{v=1}^V d\Psi(\bar{Q}_n)(Q_{n,v} - \bar{Q}_n) + O_P(d_0^{\bar{V}}(Q_n, Q_0)) \\ &= 0 + O_P(d_0^{\bar{V}}(Q_n, Q_0)) = O_P(d_0^{\bar{V}}(Q_n, Q_0)), \end{aligned}$$

where $d\Psi(\bar{Q}_n)(h) = \frac{d}{d\epsilon} \Psi(\bar{Q}_n + \epsilon h)|_{\epsilon=0}$ is the directional derivative of Ψ at \bar{Q}_n in direction h , and $O_P(\|Q_{n,v} - \bar{Q}_n\|_{P_0}^2) = O_P(\|\bar{Q}_n - Q_0\|_{P_0}^2 + \|Q_{n,v} - Q_0\|_{P_0}^2) = O_P(d_0^{\bar{V}}(Q_n, Q_0))$ under mild assumptions (Section 4). Note, the first terms equals zero since $\bar{Q}_n = \frac{1}{V} \sum_v Q_{n,v}$. Theorem 4.2 establishes that $d_0^{\bar{V}}(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^d)$ under reasonable conditions, so that this will indeed be $o_P(n^{-1/2})$. Therefore, it suffices to analyze the target feature $\Psi^r(Q_n^r) = \frac{1}{V} \sum_{v=1}^V \Psi(Q_{n,v})$ of the undersmoothed M-HAL-MLE Q_n^r .

Furthermore, we have

$$\Psi^r(Q_n^r) - \Psi(Q_0) = \Psi^r(Q_n^r) - \Psi^r(Q_{0,n}^r) + (\Psi^r(Q_{0,n}^r) - \Psi(Q_0)).$$

The second term represents a bias term in our reduced data model that treats the cross-fitted transformation \mathbf{Q}_n as fixed. If the coordinate-transformation \mathbf{Q}_n is zero loss, then we would have that $Q_{0,n,v} = Q_0$ for each v , so that $\Psi^r(Q_{0,n}^r) = \Psi(Q_0)$. There also exist examples of reductions \mathbf{Q}_n for which $Q_{0,n,v} \neq Q_0$, but nonetheless $\Psi^r(Q_{0,n}^r) = \Psi(Q_0)$ (Section 6). What matters is if the \mathbf{Q}_n represents a zero loss coordinate-transformation w.r.t $\Psi(Q_0)$. In Section 5.2 we will generally establish that this bias term $\Psi^r(Q_{0,n}^r) - \Psi(Q_0)$ is second order, due to either the cross-fitted transformation \mathbf{Q}_n being zero-loss w.r.t. $\Psi(Q_0)$, or due to ensembles of \mathbf{Q}_n being $n^{-1/4}$ -consistent estimators of Q_0 . This condition will not require C_n to undersmooth. The first estimation term, $\Psi^r(Q_n^r) - \Psi^r(Q_{0,n}^r)$, will be analyzed in Section 5.1.

5.1 Asymptotic Linearity Theorem

Under an undersmoothing selector $C_n > C_{n,cv}$ chosen large enough [10], we have (see Appendix F)

$$P_n^r D^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2}). \quad (6)$$

Once this efficient score equation (6) is solved, then we obtain

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

This can be represented as

$$\begin{aligned} \Psi^r(Q_n^r) - \Psi^r(Q_{0,n}^r) &= \frac{1}{V} \sum_{v=1}^V (P_{n,v}^1 - P_0) D^*(Q_n^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r) \\ &\quad + \frac{1}{V} \sum_{v=1}^V R_{20}(Q_n^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r, Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r). \end{aligned}$$

By assumption (5), $d_0^V(Q_n, Q_{0,n}) = O_P(n^{-2/3}(\log n)^{d^r})$ implies that the second order remainder is $O_P(n^{-2/3}(\log n)^{d^r})$. The empirical process term will be controlled in the following asymptotic linearity theorem (proof in Appendix D).

Note that conditional on $P_{n,v}$ or for fixed $\mathbf{Q}_{n,v}$, the Donsker class condition over $\mathcal{D}^* = \{D^*(Q, G) : Q \in \mathcal{Q}, G \in \mathcal{G}\}$, typically required for the original data problem, is avoided. Instead, it suffices to assume a Donsker class condition driven by Q_n^r only, which is satisfied if the sectional variation norms in $\mathcal{Q}^r, \mathcal{G}^r$ are universally bounded with probability tending to 1. With the reduced data dimensions, this meta-level regularity condition is easier to hold, especially when the original data problem is complex and constructs highly varying initial estimators. Moreover, (6) may be satisfied for not only one specific target, in which case the following theorem applies for arbitrary Ψ in a large class of smooth features.

Theorem 5.1. *Recall assumption (5). Assume C_n is chosen large enough so that*

$$\frac{1}{V} \sum_{v=1}^V P_{n,v}^1 D^*(Q_n^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r) = o_P(n^{-1/2}).$$

Then,

$$\begin{aligned} \Psi^r(Q_n^r) - \Psi^r(Q_{0,n}^r) &= \frac{1}{V} \sum_{v=1}^V (P_{n,v}^1 - P_0) D^*(Q_{0,n,v}, G_{0,n,v}^r) + O_P(n^{-2/3}(\log n)^{d^r}) \\ &= P_n D^*(Q_0, \tilde{G}_0) + o_P(n^{-1/2}). \end{aligned}$$

Combined with Section 5.2, we conclude that the target feature of M-HAL-MLE, $\Psi^r(Q_n^r)$, is root- n -consistent for the true target feature $\Psi(Q_0)$, it has known influence curve conditional on training samples (so that variance estimation follows), and it is asymptotically normally distributed, without any Donsker class assumption on \mathcal{D}^* . In addition, the target feature of M-HAL-SL is an asymptotically linear estimator of $\Psi(Q_0)$ with influence curve $D^*(Q_0, \tilde{G}_0)$. For zero-loss transformations \mathbf{Q}_n , and certain types of reductions \mathbf{Q}_n under which $G_{0,n,v}^r$ converges to true G_0 , then we will have that $D^*(Q_0, \tilde{G}_0) = D^*(Q_0, G_0)$, in which case $\Psi^r(Q_n^r)$ behaves as an asymptotically efficient estimator of $\Psi(Q_0)$. If $\tilde{G}_0 \neq G_0$, then $\Psi^r(Q_n^r)$ will typically end up being super-efficient.

5.2 Difference Between Target Feature of Oracle Estimator and Target Estimand

To establish the asymptotic linearity for the fixed parameter $\Psi(Q_0)$, it remains to establish that $\frac{1}{V} \sum_v \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) = o_P(n^{-1/2})$. The following two theorems address the scenarios with or without the nuisance parameter (see proofs in Appendix E).

The first result applies to the case in which $D^*(P) = D^*(Q)$ so that there is no nuisance parameter G .

Theorem 5.2. Assume (5). Suppose that $D^*(P) = D^*(Q(P))$. We have

$$\Psi^r(Q_{0,n}) - \Psi(Q_0) = \frac{1}{V} \sum_{v=1}^V R_{20}(Q_{0,n,v}, Q_0).$$

By (5), the latter is bounded by $O(d_0(Q_{0,n}, Q_0))$. Thus, $\Psi^r(Q_{0,n}) - \Psi(Q_0) = O(d_0(Q_{0,n}, Q_0))$.

Even if there exists a nuisance parameter, one could redefine Q as a joint parameter $Q = (Q^s, G)$ including both the sufficient Q^s and the nuisance parameter G , such that $\Psi(Q) = \Psi(Q^s, G) = \Psi(Q^s)$ and $D^*(P) = D^*(Q)$. This strategy simplifies the conditions required for $G_{0,n}^r$ but involves M-HAL-SL of both Q_0^s and G_0 .

The following general theorem handles the nuisance parameter using an approximation of $G_{0,n}^r$. In practice, this approximation can be chosen such that the residual r_n is also bounded by $(d_0^V(Q_{0,n}, Q_0))^{1/2}$ (Section 6).

Theorem 5.3.

Definitions: Let $G_{0,n,v}^* \in \mathcal{G}$ be a functional that approximates $G_{0,n,v} \equiv G_{0,n,v}^r$ and for which

$$\frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0) = \frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^*),$$

or equivalently,

$$R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^*, Q_0, G_0) = R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0, Q_0, G_0).$$

Let

$$r_n \equiv \frac{1}{V} \sum_v \{R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}, Q_0, G_0) - R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^*, Q_0, G_0)\}.$$

We note that this represents a second order term that generally can be bounded in terms of squares or products of $d_0^V(Q_{0,n}, Q_0)^{1/2}$ and a norm $\|G_{0,n,v}^* - G_{0,n,v}\|$ (such as $L^2(P_0)$ -norm).

Conclusion: We have

$$\left\{ \frac{1}{V} \sum_v \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) \right\} = r_n + O(d_0^V(Q_{0,n}, Q_0)).$$

6 Treatment Specific Mean Example

In this section, we will go through the definitions and conditions of the theorems in the context of a concrete example of treatment specific means.

Let $O = (W, A, Y) \sim P_0 \in \mathcal{M}$ be a vector random variable in which W are baseline covariates, $A \in \{0, 1\}$ is a binary treatment, and $Y \in [0, 1]$ a bounded continuous outcome. Suppose that we observe n i.i.d. copies O_1, \dots, O_n of O . Let d be the dimension of O , and assume O is a bounded random variable.

For all $P \in \mathcal{M}$, let $Q(P) = E_P(Y \mid A = 1, W)$ be the functional parameter of interest, and let $G(P) = E_P(A \mid W)$. Let the statistical model be given by $\mathcal{M} = \{P : G(P) > \delta > 0 \text{ for some } \delta > 0\}$, thereby enforcing a positivity assumption. Then $Q : \mathcal{M} \rightarrow \mathcal{Q} = \{Q(P) : P \in \mathcal{Q}\}$, where \mathcal{Q} is its parameter space. Note that each realization is a $k = d - 2$ -dimensional real valued measurable function of W . For $Q \in \mathcal{Q}$, we can choose the squared error loss function $L(Q)(O) = A(Y - Q(W))^2$, so that $Q_0 = Q(P_0) = \arg \min_Q P_0 L(Q)$. Note that the loss-based dissimilarity is given by $d_0(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0) = P_0 G_0(Q - Q_0)^2$, and is thus a square of a weighted L^2 -norm. Since G is bounded away from zero, this loss-based dissimilarity is equivalent with $\|Q - Q_0\|_{P_0}^2$, where $\|Q - Q_0\|_{P_0} = \sqrt{P_0(Q - Q_0)^2}$.

We will define $\Psi(P) = P_0 Q(P)$ as the target parameter, so that the treatment specific mean is given by $P_0 Q_0 = E_0 E_0(Y \mid A = 1, W)$ at $P = P_0$. We also denote $\Psi(P)$ with $\Psi(Q)$. The canonical

gradient of $\Psi(P)$ at P is given by $D^*(G, Q) = A/G(W)(Y - Q(W))$ and the exact second order remainder $R_{20}(Q, G, Q_0, G_0) = \Psi(Q) - \Psi(Q_0) + P_0 D^*(G, Q)$ is given by $R_{20}(Q, G, Q_0, G_0) = P_0(G - G_0)/G(Q - Q_0)$ (e.g., [19]).

Let $\mathbf{Q}_{n,v} = \hat{\mathbf{Q}}(P_{n,v})$ be a collection of J estimators of Q_0 based on training sample $P_{n,v}$, $v = 1, \dots, V$, which can also be viewed more generally as a J -dimensional data adaptive transformation $\mathbf{Q}_{n,v}(W)$ of the k -dimensional W . Recall $\mathbf{Q}_n(v, W) = \mathbf{Q}_{n,v}(W)$. Let \mathcal{Q}^r be a class of J -variate real valued cadlag functions with a universal bound C^u on its sectional variation norm. For a given cadlag function (also called ensemble) $Q^r \in \mathcal{Q}^r$, we can define the composition $Q^r \circ \mathbf{Q}_n$ by $Q^r \circ \mathbf{Q}_n(v, W) = Q^r \circ \mathbf{Q}_{n,v}(W) = Q^r(\mathbf{Q}_{n,v}(W))$.

6.1 Reduced data estimation problem treating \mathbf{Q}_n as fixed

Treating \mathbf{Q}_n as fixed, we can reduce the observed data (\bar{V}, O) to $(\bar{V}, O^r = (W^r, A, Y))$, where $W^r \equiv \mathbf{Q}_{n,\bar{V}}(W) = \sum_{v=1}^V \mathbb{I}_{\{\bar{V}=v\}} \mathbf{Q}_{n,v}(W)$. Define $W_v^r \equiv \mathbf{Q}_{n,v}(W)$. Let $d^r = J + 2$ be the dimension of the reduced data O^r . Recall (\bar{V}, O^r) follows the joint distribution $P_0^r \in \mathcal{M}^r$, where \bar{V} is uniform $\{1, \dots, V\}$, and O^r given $\bar{V} = v$ follows the distribution of $(W_v^r, A, Y) \sim P_{0,v}^r$ which is implied by $(\mathbf{Q}_{n,v}(W), A, Y)$ under P_0 . For all $P^r \in \mathcal{M}^r$, if $(\bar{V}, O^r) \sim P^r$, then there exists $P \in \mathcal{M}$ such that $(W_v^r, A, Y) \sim P_v^r$ follows the distribution implied by $(\mathbf{Q}_{n,v}(W), A, Y)$ under $O \sim P$.

Define the reduced data loss as $L^r(Q^r)(\bar{V}, O^r) = A(Y - Q^r(W^r))^2$. This satisfies condition (2): $L(Q^r \circ \mathbf{Q}_n)(O) = A(Y - Q^r(\mathbf{Q}_{n,v}(W)))^2 = L^r(Q^r)(v, O^r(v, O))$, which depends on O only through $(\mathbf{Q}_{n,v}(W), A, Y) = O^r(v, O)$.

Let $\mathcal{M}^r = \{P^r(P) : P \in \mathcal{M}, G^r(P^r) > \delta > 0, W^r \mapsto E_{P^r(P)}(Y|A = 1, W^r) \in \mathcal{Q}^r\}$ be the model for the distribution P_0^r of (\bar{V}, O^r) implied by \mathcal{M} . Define the functional parameter $Q^r : \mathcal{M}^r \rightarrow \mathcal{Q}^r$ by $Q^r(P^r) \equiv \arg \min_{Q^r \in \mathcal{Q}^r} P^r L^r(Q^r)$ which equals $E_{P^r}(Y|A = 1, W^r)$ due to the assumption on \mathcal{M}^r . $Q_{0,n}^r = Q^r(P_0^r)$ represents the optimal ensemble. Let $d_0^r(Q^r, Q_{0,n}^r) = P_0^r L^r(Q^r) - P_0^r L^r(Q_{0,n}^r)$ be the loss-based dissimilarity. Let $G_{0,n}^r(W^r) = E_0(A | W^r)$. Note also that, since $G_0 > \delta > 0$, we also have $G_{0,n}^r > \delta > 0$. The loss-based dissimilarity is given by

$$\begin{aligned} d^r(Q^r, Q_{0,n}^r) &= P_0^r G_{0,n}^r(Q^r - Q_{0,n}^r)^2 = E_0 G_{0,n}^r(W^r)(Q^r - Q_{0,n}^r)^2(W^r) \\ &= \frac{1}{V} \sum_{v=1}^V E_0 G_{0,n}^r(W_v^r)(Q^r - Q_{0,n}^r)^2(W_v^r). \end{aligned}$$

We define $\Psi^r : \mathcal{M}^r \rightarrow \mathbb{R}$ by $\Psi^r(P^r) = \Psi^r(Q^r(P^r)) = E_{P_0^r} E_{P^r}(Y | A = 1, W^r)$. It can be verified that $\Psi^r(P^r) = \frac{1}{V} \sum_{v=1}^V E_{P_0^r}(Q^r(W^r)|\bar{V} = v) = \frac{1}{V} \sum_{v=1}^V E_{P_{0,v}^r}(Q^r(W_v^r)) = \frac{1}{V} \sum_{v=1}^V E_{P_0} Q^r(\mathbf{Q}_{n,v}(W)) = \frac{1}{V} \sum_{v=1}^V \Psi(Q^r \circ \mathbf{Q}_{n,v})$. Note that, as a special case, $\Psi^r(Q_{0,n}^r) = \Psi(Q_0)$ if, for each v , W_v^r is such that A , given W , only depends on $W_v^r = \mathbf{Q}_{n,v}(W)$. The canonical gradient of Ψ^r at P^r is given by:

$$D^r(P^r)(\bar{V}, O^r) = \frac{A}{E(A | W^r)}(Y - E_{P^r}(Y | A = 1, W^r)),$$

or

$$D^r(Q^r, G^r)(\bar{V}, O^r) = \frac{A}{G^r(W^r)}(Y - Q^r(W^r)).$$

Condition (3) holds since $D^r(Q^r, G^r)(v, O_v^r(o)) = D^*(Q^r(\mathbf{Q}_{n,v}), G_v^r)(o)$, where $G_v^r(W) \equiv G^r(\mathbf{Q}_{n,v}(W))$. We have

$$P_0^r D^r(Q^r, G^r) = \frac{1}{V} \sum_{v=1}^V P_0 D^*(Q^r(\mathbf{Q}_{n,v}), G_v^r);$$

and $R_2^r(P^r, P_0^r) = P_0^r(G^r - G_{0,n}^r)/G^r(Q^r - Q_{0,n}^r)$. It follows that

$$R_{20}^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) = \frac{1}{V} \sum_{v=1}^V P_0(G^r - G_{0,n}^r)/G^r(\mathbf{Q}_{n,v})(Q^r(\mathbf{Q}_{n,v}) - Q_{0,n}^r(\mathbf{Q}_{n,v})).$$

Thus,

$$R_{20}^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) = \frac{1}{V} \sum_{v=1}^V R_{20}(Q^r(\mathbf{Q}_{n,v}), G_v^r, Q_{0,n}^r(\mathbf{Q}_{n,v}), G_{0,n,v}^r).$$

We also have the equivalences $d_0^r(Q_n^r, Q_{0,n}^r) = d_0^{\tilde{V}}(Q_n, Q_{0,n})$.

6.2 Convergence of M-HAL-MLE

We have that the M-HAL-MLE of the oracle ensemble $Q_{0,n}^r$ is given by $Q_n^r = \arg \min_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^r L^r(Q^r)$. This is just a regular HAL-MLE of $E(Y | A = 1, W^r)$ based on the reduced data set $O_i^r = (W_i^r, A_i, Y_i)$, $i = 1, \dots, n$, where $W_i^r = \mathbf{Q}_{n,v_i}(W_i)$. It corresponds with a linear least squares regression under an L_1 -constraint $\|\beta^r\|_1 < C_n$, and it results in a fit $Q_n^r = \sum_{s,j} \beta_n^r(s,j) \phi_{s,j}$ for a rich collection of spline basis functions. Given Q_n^r , we can compute the M-HAL-SL Q_n (collection of V estimators) by $Q_n(v, \cdot) = Q_n^r \circ \mathbf{Q}_{n,v}$, and corresponding average $\bar{Q}_n(\cdot) = \frac{1}{V} \sum_{v=1}^V Q_n(v, \cdot)$. The target of Q_n is the oracle estimator defined by $Q_{0,n}(v, \cdot) = Q_{0,n}^r \circ \mathbf{Q}_{n,v}$, and the target of \bar{Q}_n is accordingly given by $\bar{Q}_{0,n} = \frac{1}{V} \sum_{v=1}^V Q_{0,n}^r \circ \mathbf{Q}_{n,v}$. Note that $Q_{0,n}^r \circ \mathbf{Q}_{n,v}(w) = E_0(Y | A = 1, W_v^r = \mathbf{Q}_{n,v}(w))$.

Due to O being bounded, and \mathcal{Q} being bounded functions, we have that $M_1 < \infty$ and $M_2^r < \infty$. By assumption, \mathcal{Q}^r consists of J -dimensional real valued cadlag functions on $[0, 1]^J$ with sectional variation norm bounded by a universal C^u . Let $d^r = J + 2$ be the dimension of $O^r = (W^r, A, Y)$. Therefore, it follows that $\{L(Q^r \circ \mathbf{Q}_{n,v}) : Q^r \in \mathcal{Q}^r\}$ represents a class of d^r -valued cadlag functions with a universal bound on its sectional variation norm. This verifies all conditions of Lemma 4.1 so that $d_0^{\tilde{V}}(Q_n, Q_{0,n}) = O_P(n^{-2/3}(\log n)^{d^r})$. In addition, we have $L(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) = L(Q_0)$ so long as A depends on W only through $W_v^r = \mathbf{Q}_{n,v}(W)$, in which case Theorem 4.2 implies $d_0^{\tilde{V}}(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^{d^r})$. In the case that one of the algorithms \hat{Q}_1 satisfies $d_0(\mathbf{Q}_{n,v,1}, Q_0) = O_P(n^{-2/3}(\log n)^d)$, it follows that $d_0^{\tilde{V}}(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^d)$ by Theorem 4.2.

6.3 Plug-in Estimation with Undersmoothed M-HAL-SL

To apply Theorem 5.1, we verify that with large enough $C_n > C_{n,cv}$ the undersmoothed M-HAL-SL solves efficient score equations for the target feature such that $P_n^r D^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$ (Appendix H). Thus,

$$\Psi^r(Q_n^r) - \Psi^r(Q_{0,n}^r) = P_n D^*(Q_0, \tilde{G}_0) + o_P(n^{-1/2}),$$

where \tilde{G}_0 is the limit of $G_{0,n,v}^r$ as defined in (5). To achieve the asymptotic linearity of $\Psi^r(Q_n^r)$ for $\Psi(Q_0)$, it is left to be verified the conditions required for $\Psi^r(Q_{0,n}^r) - \Psi(Q_0) = o_P(n^{-1/2})$ as in Section 5.2.

There are important cases in which $\Psi^r(Q_{0,n}^r) = \Psi(Q_0)$ exactly. For example, suppose that A , given W , equals A , given W_1 for a lower dimensional vector W_1 . In that case, we could define $(\mathbf{Q}_{n,v} : v)$ as a collection of estimators of $E_0(Y | A = 1, W)$, but augmented with the fixed function W_1 of W . Then, $\Psi(Q_0) = E_0 E_0(Y | A = 1, W) = E_0 E_0(Y | A = 1, W_1)$, and, similarly, $\Psi(Q_{0,n}^r) = E_0 E_0(Y | A = 1, W^r) = E_0 E_0(Y | A = 1, W_1)$. So in this case, we have $\Psi^r(Q_{0,n}^r) - \Psi(Q_0) = 0$.

This generalizes to any causal estimation problem in which the intervention mechanism is known to only be affected by a low dimensional summary of all measured (baseline and time-dependent) covariates, and these are included in \mathbf{Q}_n as fixed functions. (With longitudinal data and iterative conditional expectations, \mathbf{Q}_n can have a nested structure, sequentially defining a set of coordinate transformations for each intervention time-point.) In particular, this means that this type of M-HAL-SL yields efficient estimators of causal effects of single time point and multiple time point interventions based on (sequentially) randomized trials and well understood observational studies, in which one knows, at each intervention time-point, a low dimensional summary measure of the past that predicts the intervention. Here, the intervention can have both a treatment and a censoring component.

In general, the difference of the random and fixed parameter is given by:

$$\Psi^r(Q_{0,n}^r) - \Psi(Q_0) = E_0 E_0(Y | W^r, A = 1) - E_0 E_0(Y | W, A = 1),$$

which, under similar conditions as Theorem 5.3, can be shown to behave as $d_0(Q_n, Q_0)$ and is thus second order (Appendix I).

6.4 Double Robustness

Lemma 6.1. *If $\mathbf{Q}_{n,v}(W)$ includes the correct propensity score model $G_0(W)$ (or if G_0 depends on W only through $\mathbf{Q}_{n,v}(W)$) for all $v = 1, \dots, V$, or if $\mathbf{Q}_{n,v}(W)$ includes the correct outcome model $Q_0(W)$ for all $v = 1, \dots, V$, then we have $\Psi^r(Q_{0,n}^r) - \Psi(Q_0) = 0$. Moreover, if $\mathbf{Q}_{n,v}(W)$ includes a consistent estimator for either Q_0 or G_0 , then we have that $\Psi^r(Q_{0,n}^r)$ is consistent for $\Psi(Q_0)$.*

Proof. Define $G_{0,n,v}(W) = E_0(A | \mathbf{Q}_{n,v}(W))$. Note that $P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}) = 0$ by iterated conditional expectation. Utilizing the double robustness structure of R_{20} and the positivity assumption, we have

$$\begin{aligned} |\Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0)| &= |P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}) - P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0)| \\ &\leq |P_0(G_{0,n,v} - G_0)(Q_0 - Q_{0,n}^r \circ \mathbf{Q}_{n,v})| \\ &\leq \|G_{0,n,v} - G_0\|_{P_0} \|Q_{0,n}^r \circ \mathbf{Q}_{n,v} - Q_0\|_{P_0}. \end{aligned}$$

If $\mathbf{Q}_{n,v}(W)$ includes estimators $Q_1(W)$ and $G_1(W)$, then $Q_1(W)$ and $G_1(W)$ are measurable functions of $\mathbf{Q}_{n,v}(W)$. Note that by tower properties, $E_0\{G_0(W) | \mathbf{Q}_{n,v}(W)\} = E_0(A | \mathbf{Q}_{n,v}(W)) = G_{0,n,v}(W)$, and $E_0\{Q_0(W) | A = 1, \mathbf{Q}_{n,v}(W)\} = E_0(Y | A = 1, \mathbf{Q}_{n,v}(W)) = Q_{0,n}^r \circ \mathbf{Q}_{n,v}(W)$. By projection properties,

$$|\Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0)| \leq \|G_1 - G_0\|_{P_0} \|Q_1 - Q_0\|_{P_0}.$$

This proves the claims when $Q_1 = Q_0$, or $G_1 = G_0$, or either Q_1 or G_1 is a consistent estimator (A and Y are both bounded). Lastly, if G_0 depends on W only through $\mathbf{Q}_{n,v}(W)$, then $G_0(W) = G_{0,n,v}(W)$ directly. \square

Note that a similar result can be achieved following Lemma I.1.

When Lemma 6.1 holds, the target feature of a properly undersmoothed M-HAL-SL $\Psi^r(Q_n^r)$, that satisfies the undersmoothing condition of Theorem 5.1, is asymptotically linear for $\Psi(Q_0)$. Therefore, the typical double robustness property, that full-data TMLE is asymptotic linear with a correct model for either Q or G , can be preserved in meta-learning estimators for treatment specific means.

7 Numerical Experiments

7.1 Prediction Performance of M-HAL-SL

In this section, we evaluate the prediction performance of M-HAL-SL relative to regular super-learners. The tuning parameter λ in M-HAL-SL is selected by cross-validation with `glmnet`, using honest or fast selectors. Other super-learners for comparisons include: non-negative least square superlearner without normalization (NNLS-SL), convex super-learner which restricts the weights to be positive with sum equal to one (Convex-SL), and the simple ensemble that takes the average of base learners' predictions (Average). Each super-learner uses the following five base learner algorithms: intercept-only model (`Lrnr_mean`), simple linear regression (`Lrnr_glm`), Xgboost (`Lrnr_xgboost`), support vector machine (`Lrnr_svm`) and random forest (`Lrnr_rf`).

We first generate data with simple distributions of five covariates X_1, \dots, X_5 and a continuous outcome Y . The distribution of variables are as follows:

$$\begin{aligned} X_1 &\sim U(-4, 4), X_2 \sim U(-4, 4), X_3 \sim \text{Bernoulli}(0.5), \\ X_4 &\sim N(0, 1), X_5 \sim \text{Gamma}(2, 1). \end{aligned} \quad (7)$$

The outcome Y is generated by the “jump” regression function [3]:

$$\begin{aligned} \psi_0(x) &= -I(x_1 < -3)x_3 + 0.5I(x_1 > -2) - I(x_1 > 0) + 2I(x_1 > 2)x_3 \\ &\quad - 3I(x_1 > 3) + 1.5I(x_2 > -1) - 5I(x_2 > 1)x_3 + 2I(x_2 > 3) \\ &\quad + 2I(x_4 < 0) - I(x_5 > 5) - I(x_4 < 0)I(x_1 < 0) + 2x_3, \\ Y &= \psi_0(X) + \epsilon, \epsilon \sim N(0, 1). \end{aligned}$$

In each iteration, a dataset is generated with sample sizes $n = 200, 500, 1000$, and 2000 , respectively. The measure of performance is the mean of squared error (MSE) on an external test dataset of 5000 samples generated from the same distribution. We also provide their relative MSEs using Convex-SL as the baseline. In this scenario, M-HAL-SL performs as good as other super-learners (Table 1).

In the second scenario, we generate data with more complicated distributions of 20 covariates (X_1, \dots, X_{20}). These covariates are divided in four equal sized groups, and the conditional mean of Y is an additive model of four functions of the corresponding clusters of covariates. Each function involves two-way intersections between the five covariates. The data-generating distribution is as follows:

$$\begin{aligned} X_1 &\sim U(-4, 4), X_2 \sim U(-4, 4), X_3 \sim \text{Bernoulli}(0.5), X_4 \sim N(0, 1), \\ X_5 &\sim \text{Gamma}(2, 1), X_6 \sim \text{Pois}(2), X_7 \sim \text{Exp}(3), X_8 \sim \text{Beta}(1, 1), \\ X_9 &\sim \chi^2(2), X_{10} \sim \text{Geom}(0.6), X_{11} \sim U(-4, 4), X_{12} \sim U(-4, 4), \\ X_{13} &\sim \text{Bernoulli}(0.5), X_{14} \sim N(0, 1), X_{15} \sim \text{Gamma}(2, 1), \\ X_{16} &\sim \text{Pois}(1), X_{17} \sim \text{Exp}(1), X_{18} \sim \text{Beta}(2, 1), \\ X_{19} &\sim \chi^2(1), X_{20} \sim \text{Geom}(0.8). \end{aligned}$$

$$\begin{aligned} G_1(x) &= -I(x_1 < -3)x_3 + 0.5I(x_1 > -2) - I(x_1 > 0) + 2I(x_1 > 2)x_3 \\ &\quad - 3I(x_1 > 3) + 1.5I(x_2 > -1) - 5I(x_2 > 1)x_3 + 2I(x_2 > 3) \\ &\quad + 2I(x_4 < 0)x_2 - I(x_5 > 5)x_1 - I(x_4 < 0)I(x_1 < 0) + 2x_3 \\ G_2(x) &= -I(x_{10} > 3)x_8 + 0.5I(x_{10} > 2) - I(x_{10} > 0) + 2I(x_{10} > 2)x_8 \\ &\quad - 3I(x_{10} > 3) + 1.5I(x_9 > 5) - 5I(x_9 > 1)x_8 + 2I(x_9 > 3) \\ &\quad + I(x_7 > 4)x_9 - I(x_6 > 5)x_{10} - I(x_7 > 1)I(x_{10} < 2) + 2x_8 \\ G_3(x) &= -I(x_{11} < -3)x_{13} + 0.5I(x_{11} > -2) - I(x_{11} > 0) + 2I(x_{11} > 2)x_{13} \\ &\quad - 3I(x_{11} > 3) + 1.5I(x_{12} > -1) - 5I(x_{12} > 1)x_{13} + 2I(x_{12} > 3) \\ &\quad + I(x_{14} < 0)x_{12} - I(x_{15} > 5)x_{11} - I(x_{14} < 0)I(x_{11} < 0) + 2x_{13} \\ G_4(x) &= -I(x_{19} > 3)x_{17} + 0.5I(x_{19} > 2) - I(x_{19} > 0) + 2I(x_{19} > 2)x_{18} \\ &\quad - 3I(x_{19} > 3) + 1.5I(x_{16} > 5) - 5I(x_{16} > 1)x_{18} + 2I(x_{16} > 3) \\ &\quad + I(x_{16} > 4)x_{19} - I(x_{16} > 5)x_{20} - I(x_{17} > 1)I(x_{20} < 2) + 2x_{18} \\ \psi_0(x) &= G_1(x) + G_2(x) + G_3(x) + G_4(x) \\ Y &= \psi_0(X) + \epsilon, \epsilon \sim N(0, 1). \end{aligned} \quad (8)$$

In the second scenario with more complicated distributions, M-HAL-SL performs slightly better than other super-learners (Table 2).

metalearner	n = 200		n = 500		n = 1000		n = 2000	
	mse	relative_mse	mse	relative_mse	mse	relative_mse	mse	relative_mse
Meta-HAL-d2 (Honest CV)	2.15	1.01	1.67	0.98	1.37	0.98	1.26	1.01
Meta-HAL-d2 (Fast CV)	2.10	0.99	1.67	0.99	1.37	0.98	1.26	1.01
Meta-HAL-d1 (Honest CV)	2.01	0.95	1.66	0.98	1.37	0.98	1.25	1.00
Meta-HAL-d1 (Fast CV)	2.28	1.08	1.66	0.98	1.37	0.98	1.24	1.00
Convex	2.12	1.00	1.69	1.00	1.40	1.00	1.25	1.00
NNLS	2.09	0.99	1.71	1.01	1.39	0.99	1.24	1.00
Discrete SL	2.30	1.09	1.71	1.01	1.47	1.05	1.30	1.05
Average	2.81	1.33	2.52	1.49	2.28	1.63	2.10	1.69
Lrrn_mean	5.09	2.40	5.06	2.99	5.04	3.60	5.04	4.05
Lrrn_glm	4.68	2.21	4.64	2.74	4.63	3.31	4.59	3.68
Lrrn_xgboost	2.30	1.09	1.71	1.01	1.47	1.05	1.26	1.01
Lrrn_svm	3.16	1.49	2.94	1.74	2.73	1.95	2.50	2.01
Lrrn_rf	2.24	1.06	1.82	1.07	1.46	1.04	1.31	1.05

Tab. 1: Prediction performance of super-learners with simple distributions (7) in Section 7.1.

metalearner	n = 200		n = 500		n = 1000		n = 2000	
	mse	relative_mse	mse	relative_mse	mse	relative_mse	mse	relative_mse
Meta-HAL-d2 (Honest CV)	13.35	0.95	9.60	0.89	7.29	0.84	5.26	0.86
Meta-HAL-d2 (Fast CV)	13.33	0.95	9.63	0.89	7.21	0.83	5.23	0.86
Meta-HAL-d1 (Honest CV)	13.20	0.94	9.46	0.87	7.06	0.81	5.22	0.86
Meta-HAL-d1 (Fast CV)	13.20	0.94	9.46	0.87	7.04	0.81	5.23	0.86
Convex	14.08	1.00	10.83	1.00	8.70	1.00	6.11	1.00
NNLS	13.47	0.96	9.90	0.91	7.35	0.84	5.44	0.89
Discrete SL	14.50	1.03	11.06	1.02	8.30	0.95	6.05	0.99
Average	14.66	1.04	12.55	1.16	11.66	1.34	10.22	1.67
Lrrn_mean	20.32	1.44	19.99	1.85	19.98	2.30	19.98	3.27
Lrrn_glm	17.27	1.23	16.20	1.50	15.60	1.79	15.58	2.55
Lrrn_xgboost	16.57	1.18	11.06	1.02	8.30	0.95	6.05	0.99
Lrrn_svm	15.14	1.08	12.76	1.18	11.80	1.36	10.55	1.73
Lrrn_rf	14.47	1.03	12.61	1.16	11.50	1.32	10.11	1.66

Tab. 2: Prediction performance of super-learners with more complicated distributions (8) in Section 7.1.

Note that the asymptotic performance of M-HAL-SL relies on the convergence rate of the HAL algorithm, which allows M-HAL-SL to approximate more complex functions of the base learners. In comparison, the learner library of other super-learners are limited to simple linear combinations, and their asymptotic performances rely on oracle inequality and therefore the best estimator in the learner library. Indeed, at the largest sample size ($n = 2000$), both Convex-SL and NNLS-SL perform similarly as the best candidate algorithm Lrrn_xgboost under either simple or more complicated distributions, but M-HAL-SL gains additional precision when the data generating distribution is more complex and a more flexible combination of base learners may jointly assist prediction. Our results show that M-HAL-SL is a valid alternative super-learner, and its finite-sample performance can be similar or slightly better than Convex-SL and NNLS-SL when base learners operate on all variables, depending on sample sizes and complexity of data generating distributions.

metalearner	n = 200		n = 500		n = 1000		n = 2000	
	mse	relative_mse	mse	relative_mse	mse	relative_mse	mse	relative_mse
Meta-HAL-d2 (Honest CV)	4.80	1.01	2.48	0.53	3.40	0.72	2.11	0.45
Meta-HAL-d2 (Fast CV)	4.78	1.01	2.62	0.55	3.46	0.73	2.12	0.45
Meta-HAL-d1 (Honest CV)	4.69	0.99	3.85	0.81	4.50	0.96	3.63	0.77
Meta-HAL-d1 (Fast CV)	4.68	0.99	3.87	0.82	4.40	0.93	3.70	0.79
Convex	4.74	1.00	4.72	1.00	4.72	1.00	4.70	1.00
NNLS	4.74	1.00	4.73	1.00	4.72	1.00	4.70	1.00
Discrete SL	4.74	1.00	4.73	1.00	4.72	1.00	4.71	1.00
Average	4.93	1.04	4.90	1.04	4.88	1.03	4.87	1.04
x1	5.06	1.07	5.02	1.06	4.99	1.06	4.99	1.06
x2	5.09	1.07	5.08	1.08	5.04	1.07	5.04	1.07
x3	5.08	1.07	4.99	1.06	5.00	1.06	4.97	1.06
x4	4.74	1.00	4.73	1.00	4.72	1.00	4.71	1.00
x5	5.09	1.07	5.07	1.07	5.04	1.07	5.04	1.07

Tab. 3: Ensemble performance of super-learners when each base learner is a univariate linear regression on each single covariate X_i (Section 7.2) with simple distributions (7).

7.2 Performance of M-HAL-SL as Ensemble Method

When each of the base learners utilizes only part of the variables and captures partial information, M-HAL-SL can illustrate advantages as a more flexible ensemble method. In those scenarios, the unknown and possibly non-linear relationship between the dependent variables and base learners can still be approximated by M-HAL-SL so long as the variation norm is bounded on the meta-level (which usually has much lower dimensions depending on the number of base learner algorithms), whereas Convex-SL and NNLS-SL may be capped by the performance of each individual learner.

We consider the following two scenarios:

1. data is generated with the simple distribution, and base learner algorithms only include univariate linear regression estimators on single covariate X_i ,
2. data is generated with the more complicated distribution, and base learner algorithms are univariate linear regression estimators on single covariate X_i .

In both scenarios, the simulation results show expected performance. As base learners only learn from a subset of covariates, the prediction accuracy of M-HAL-SL significantly increases with larger sample sizes relative to other super-learners and base learners (Table 3 and Table 4). These results demonstrate the potential application of M-HAL-SL as a better model fusion method that can achieve higher precision than each individual model and regular super-learners, especially when each base learner contains only partial information. For example, this is applicable to the analysis of multi-modal data, where each modality requires separate training of complex models with tailored structures.

7.3 Undersmoothed M-HAL-MLE as Plug-in Estimator of Target Feature

This section evaluates the performance of (undersmoothed) M-HAL-SL plug-in estimation, following the treatment specific mean example (Section 6). We first verify asymptotic linearity properties with low-dimensional ($p \ll n$) covariates. In addition, we simulate high-dimensional covariates with small n large p , which is common in electronic health records, genomics, and brain imaging data. The higher dimensional and highly varying models present specific challenges, particularly with the application of machine learning and deep learning algorithms.

We simulate 4 clinical covariates along with 4 or $4n$ additional covariates for lower or higher dimensional settings at sample size $n = 200$. Only 4 additional covariates remain active in the propensity score model

metalearner	n = 200		n = 500		n = 1000		n = 2000	
	mse	relative_mse	mse	relative_mse	mse	relative_mse	mse	relative_mse
Meta-HAL-d2 (Honest CV)	15.03	0.83	10.69	0.56	7.57	0.41	5.81	0.31
Meta-HAL-d2 (Fast CV)	15.07	0.83	10.81	0.57	7.73	0.42	6.22	0.33
Meta-HAL-d1 (Honest CV)	15.96	0.88	12.34	0.65	10.89	0.59	10.70	0.57
Meta-HAL-d1 (Fast CV)	16.07	0.89	12.40	0.65	11.09	0.60	11.06	0.58
Convex	18.15	1.00	19.02	1.00	18.35	1.00	18.91	1.00
NNLS	17.61	0.97	16.10	0.85	16.89	0.92	15.84	0.84
Discrete SL	19.04	1.05	18.90	0.99	18.73	1.02	18.74	0.99
Average	19.83	1.09	19.54	1.03	19.54	1.06	19.52	1.03
x1	20.45	1.13	19.90	1.05	19.89	1.08	19.88	1.05
x2	19.04	1.05	18.83	0.99	18.80	1.02	18.80	0.99
x3	20.31	1.12	19.97	1.05	19.98	1.09	19.97	1.06
x4	20.30	1.12	19.96	1.05	19.95	1.09	19.96	1.06
x5	20.36	1.12	20.09	1.06	20.03	1.09	19.98	1.06
x6	20.35	1.12	20.00	1.05	19.98	1.09	19.98	1.06
x7	20.25	1.12	19.93	1.05	19.95	1.09	19.97	1.06
x8	20.21	1.11	19.88	1.05	19.87	1.08	19.89	1.05
x9	20.32	1.12	19.97	1.05	19.97	1.09	19.96	1.06
x10	20.40	1.12	19.82	1.04	19.82	1.08	19.81	1.05
x11	20.28	1.12	19.94	1.05	19.95	1.09	19.96	1.06
x12	19.07	1.05	18.86	0.99	18.73	1.02	18.74	0.99
x13	20.38	1.12	19.97	1.05	19.95	1.09	19.92	1.05
x14	20.28	1.12	19.98	1.05	19.97	1.09	19.94	1.05
x15	20.33	1.12	19.99	1.05	19.98	1.09	20.01	1.06
x16	19.14	1.05	18.90	0.99	18.87	1.03	18.88	1.00
x17	20.12	1.11	19.84	1.04	19.80	1.08	19.85	1.05
x18	20.27	1.12	19.96	1.05	19.92	1.09	19.92	1.05
x19	20.43	1.13	19.81	1.04	19.81	1.08	19.82	1.05
x20	20.65	1.14	19.98	1.05	19.98	1.09	19.99	1.06

Tab. 4: Ensemble performance of super-learners when each base learner is a univariate linear regression on each single covariate X_i (Section 7.2) with more complicated distributions (8).

and the outcome model; choosing a larger L1 norm bound than the cross-validated choice for the Lasso algorithm emulates an overfitted estimator with inflated sectional variation. The clinical covariates are always part of the true data generating process not subject to regularization. Specifically, we have the following additive model,

$$\begin{aligned} W &= (W_c, W_h) \\ W_h &= (W_a, W_n) \\ G_0(W) &= \beta_{0A} + \beta_A \cdot \mathbf{1}^\top(W_c, W_a) \\ Y &= \beta_Y \cdot \mathbf{1}^\top(W_c, W_a) + \psi_0 A + \epsilon \\ \epsilon &\sim N(0, 1). \end{aligned}$$

W_c is the vector of clinical covariates with $|W_c| = 4$. $W_h = (W_a, W_n)$ is the vector of additional covariates where only W_a is active with $|W_a| = 4$; we set the noise vector $W_n = \emptyset$ for low-dimensional cases and $|W_h| = 4n$ for high-dimensional (sparse) cases. We set $\beta_{0A} = -|W_a|\beta_A/2$ to avoid violating positivity assumptions. $\beta_A = 0.2$. $\beta_Y = 0.6$. $\psi_0 = 1$. Note that $\psi_0 = E_{P_0}(Y|A = 1, W) - E_{P_0}(Y|A = 0, W)$ is the target parameter, which is the difference between the treatment specific means as defined in Section 6.

The following estimators are evaluated:

- $\hat{\psi}_{\text{noadj}}$: the difference between the observed group means without covariate adjustment;
- $\hat{\psi}_{\text{tmle}}$: TMLE adjusting for all covariates, W ;
- $\hat{\psi}_{\text{meta}}^{\text{init}}$: M-HAL-SL plug-in, where the learner library, $\hat{\mathbf{Q}}$, consists of four univariate identity maps for the clinical covariate, the two group mean estimations using the same initial learner for Q_0 in $\hat{\psi}_{\text{tmle}}$, and the initial learner for G_0 in $\hat{\psi}_{\text{tmle}}$; the learner library without an estimator for G_0 is also evaluated;
- $\hat{\psi}_{\text{meta}}^{\text{us}}(\hat{G})$: Undersmoothed M-HAL-SL plug-in. Note that $\hat{\psi}_{\text{meta}}^{\text{us}}$ is a function of a separate G estimator that converges to \tilde{G}_0 in (5), which may lead to super-efficiency when $\tilde{G}_0 \neq G_0$. We consider \hat{G} to be the same as the initial learner for G_0 in $\hat{\psi}_{\text{tmle}}$, or a simplified version using no additional covariate but only the 4 clinical covariates (`us_simpleG`).

For low-dimensional settings, the initial estimators of G_0 and Q_0 for $\hat{\psi}_{\text{tmle}}$ are main-term logistic and linear regressions. For high-dimensional settings with additional covariates W_h , we use either regularized (generalized) linear regression or neural networks with one hidden layer as the initial estimators in $\hat{\psi}_{\text{tmle}}$. For main-term lasso, we used an additional loss term $\lambda_1 \|\hat{\beta}_h\|_1$ for regularization, where $\hat{\beta}_h$ is the estimated coefficients for high-dimensional covariates W_h ; we then reduce λ_1 to 0.1% of the cross-validated choice λ_{cv} to simulate over-fitted estimators with increasing total variations.

In low-dimensional settings, undersmoothed M-HAL-SL plug-in achieves asymptotic linearity with a reasonable finite-sample bias/SE ratio at a small cost in MSE compared to regular TMLE (Figure 1, Table 5). Note that undersmoothing reduces bias and improves bias-variance trade-off of the M-HAL-SL plug-in. In high-dimensional settings with LASSO initial learners, undersmoothed M-HAL-SL plug-in achieves reasonable bias/SE ratios in all scenarios, illustrating stable performance given overfitted learners and increased variation (Figure 2, Table 6). By removing the overfitted G estimator in $\hat{\psi}_{\text{meta}}^{\text{us}}$ we achieved better MSE performance without losing asymptotic linearity; this suggests that undersmoothing may flexibly span a larger space of scores and solve the same estimating equations when a G estimator is not involved. Strategies without estimating G such as plateau undersmoothing algorithms may be desirable when model knowledge on G is lacking and getting reasonable estimation is unlikely; for example, when the baseline scan images are strongly predictive for the outcomes but not drug prescriptions. In the last setting, the LASSO learners are replaced with neural networks with one hidden layer and two hidden nodes. Such networks each consists of more than 1,600 weights which is especially challenging given the sample size $n = 200$; this mimics carelessly tuned neural networks that are heavily overfitted but may still be informative. Only (undersmoothed) M-HAL-SL plug-in estimators perform stably and improve upon the raw sample mean difference without covariate adjustment. Note that in this specific example, undersmoothing is not reducing bias further compared to the initial M-HAL-SL; a super-efficient estimator (`us_simpleG`), which uses a misspecified estimator for G_0 in undersmoothing criteria, outperforms the undersmoothed M-HAL-SL using

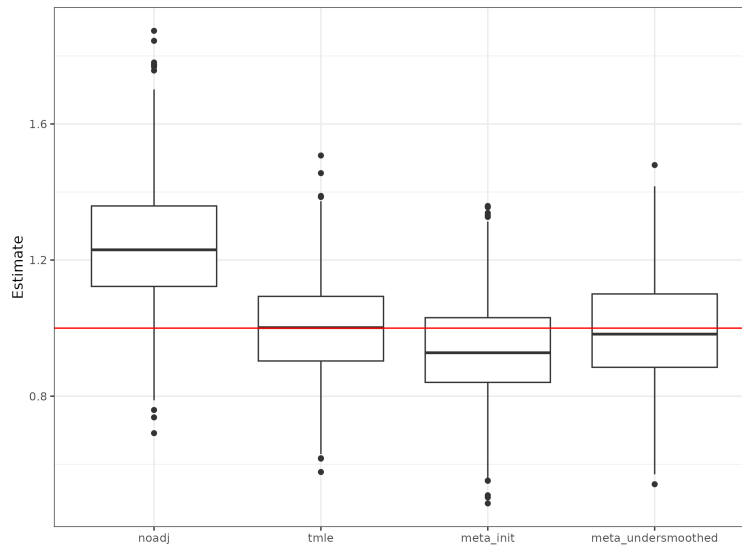


Fig. 1: Performance of (undersmoothed) M-HAL-SL plug-in estimators compared with regular TMLE or no adjustment in Section 7.3. Low-dimensional settings; $n = 200$ with 4 clinical covariates and 4 additional covariates.

the overfitted initial estimators. This suggests that model selection for G should consider the performance of final plug-in estimations, where collaborative adjustment similar to [23] may be applicable.

	MSE	Bias	SD	Ratio
noadj	0.095	0.242	0.192	1.262
tmle	0.022	0.003	0.149	0.019
meta_init	0.029	-0.067	0.158	-0.423
meta_undersmoothed	0.026	-0.006	0.161	-0.040

Tab. 5: Performance of (undersmoothed) M-HAL-SL plug-in estimators compared with regular TMLE or no adjustment in Section 7.3. Low-dimensional settings; $n = 200$ with 4 clinical covariates and 4 additional covariates.

8 Discussion

We proposed a super-learner that uses HAL-MLE to select a best ensemble among all cadlag functions of the J estimators in its library with a bound on its sectional variation norm. The oracle ensemble estimated by this HAL-MLE results in an oracle estimator that truly represents a powerful estimator. This is also reflected by the fact that if one selects J equal to the dimension k of the input of the true function, and the realized J estimates represent an invertible transformation of the input, then the realized oracle estimate actually equals the true function. Even when the library of estimators is very small, but includes one good estimator, then the oracle ensemble will improve upon this estimator by having enormous flexibility to correct its errors.

Therefore, in great generality, our results demonstrate that the performance of the M-HAL super-learner w.r.t. true function is all about how well the M-HAL-MLE estimates the oracle ensemble, where the J cross-fitted estimates are treated as fixed and represent the coordinates of the oracle ensemble. In particular, we show that the same applies for any pathwise differentiable target feature: the performance of the target feature of the M-HAL super-learner w.r.t. target estimand is all about how well the corresponding target

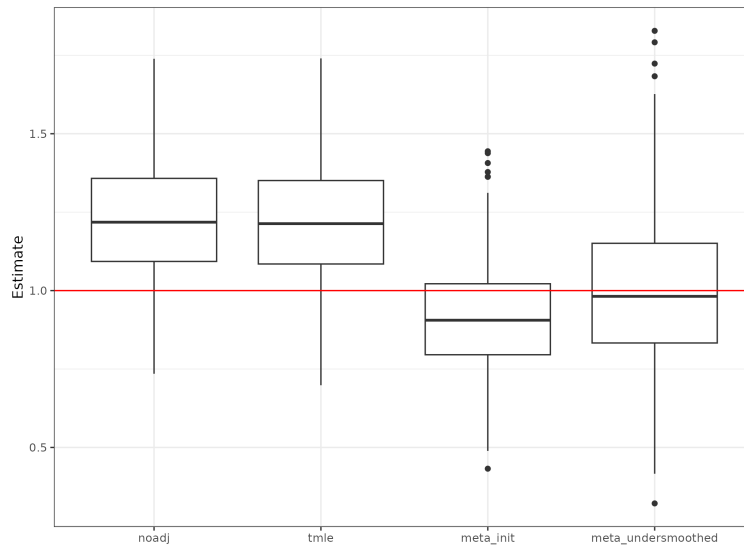


Fig. 2: Performance of (undersmoothed) M-HAL-SL plug-in estimators compared with regular TMLE or no adjustment in Section 7.3. High-dimensional settings; $n = 200$ with 4 clinical covariates and 800 additional covariates; initial learners use the LASSO algorithm and 0.1% of the cross-validated choice of λ_1 .

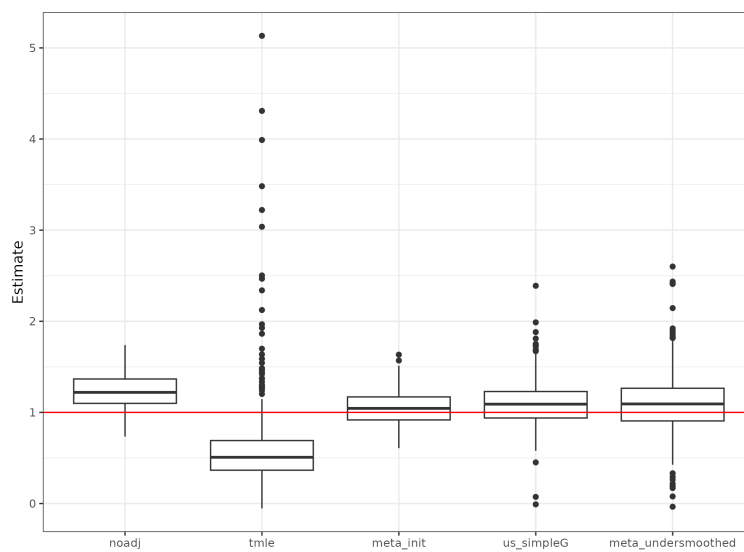


Fig. 3: Performance of (undersmoothed) M-HAL-SL plug-in estimators compared with regular TMLE or no adjustment in Section 7.3. High-dimensional settings; $n = 200$ with 4 clinical covariates and 800 additional covariates; initial learners use neural networks with one hidden layer.

	MSE	Bias	SD	Ratio
noadj	0.091	0.233	0.192	1.217
λ_{cv}	MSE	Bias	SD	Ratio
tmle	0.031	0.068	0.162	0.421
meta_init	0.035	-0.077	0.172	-0.445
meta_undersmoothed	0.055	0.014	0.235	0.058
$0.1\%\lambda_{cv}$	MSE	Bias	SD	Ratio
tmle	0.088	0.225	0.193	1.165
meta_init	0.037	-0.086	0.173	-0.496
meta_undersmoothed	0.058	-0.002	0.240	-0.007
$\lambda_{cv}, \hat{\psi}_{meta}^{init}$ without \hat{G}	MSE	Bias	SD	Ratio
tmle	0.031	0.068	0.162	0.421
meta_init	0.031	-0.065	0.164	-0.396
meta_undersmoothed	0.041	0.018	0.202	0.089
$0.1\%\lambda_{cv}, \hat{\psi}_{meta}^{init}$ without \hat{G}	MSE	Bias	SD	Ratio
tmle	0.088	0.225	0.193	1.165
meta_init	0.029	-0.041	0.166	-0.249
meta_undersmoothed	0.042	0.010	0.205	0.048

Tab. 6: Performance of (undersmoothed) M-HAL-SL plug-in estimators compared with regular TMLE or no adjustment in Section 7.3. High-dimensional settings; $n = 200$ with 4 clinical covariates and 800 additional covariates; initial learners use the LASSO algorithm. 100% or 0.1% of the cross-validated choice of λ_1 is used; learner libraries with or without a G estimator are compared.

feature (i.e., w.r.t. J coordinates) of the meta HAL-MLE estimates the target feature of the oracle ensemble, where the J cross-fitted estimates are treated as fixed coordinates. In this manner, we were able to show that 1) the M-HAL super-learner converges at a rate $n^{-2/3}(\log n)^{d^*}$ to the true function w.r.t. excess risk; and 2) when the sectional variation norm is chosen to satisfy a global undersmoothing criterion, then target features of the M-HAL super-learner will be asymptotically linear (efficient or super-efficient) estimators of the target features of the true function. These two properties equally apply to a targeted M-HAL super-learner in which we use targeted HAL-MLE to estimate the oracle ensemble, thereby enforcing it to solve the efficient influence curve equation for a user supplied set of target features.

A remaining question is then what we gained relative to using a regular HAL-MLE (or T-HAL-MLE). However, the transformed coordinates implied by the J cross-fitted functions will typically allow the oracle ensemble to be a cadlag function with significantly smaller sectional variation norm than the true function as a function of the original coordinates. This is obvious if we select $J = 1$ and choose a single super-learner as the "library" of estimators, in which case the oracle ensemble is a 1-dimensional function. However, even when we select J larger, if there are good estimators in the library, then, even using the best estimator among the J estimators as intercept would already mean that the sectional variation norm only needs to suffice to fit the residual bias. In addition, even when none of the J estimators are good, but they represent an effective coordinate-transformation, gains could be expected.

Another interesting feature of the M-HAL super-learner is that we did not have to make assumptions on the parameter space of the true function, beyond that the oracle estimator needs to do a good job in approximating the true function. Assuming the latter, this means that we allow the true function to be non-cadlag and/or have infinite sectional variation norm. So, in essence, the only smoothness condition we enforced on the parameter space of the true function is that an oracle ensemble of J candidate estimators has an excess risk that converges at a good rate (e.g, $n^{-1/4}$) to the true function.

The M-HAL super-learner is highly user friendly by not requiring the user to choose a meta-learning step. The only tuning parameter for the meta-learning is the sectional variation norm (i.e, L_1 -norm in HAL), and that one can be optimally selected with the cross-validation selector for the sake of the function as a whole, and we could target this selector to a collection of target features as well with relatively straightforward undersmoothing criteria.

An interesting question is if an HAL-MLE using as coordinates the J fitted functions is relatively interpretable if the J estimators are themselves interpretable estimators [24, 25]. The basis functions in HAL are just indicators, suggesting that these indicators evaluated at the cross-fitted functions represent an interpretable basis function. Since the HAL-MLE is just a linear regression model in these basis functions, this suggests that the M-HAL super-learner fit is represented by a linear combination of interpretable basis functions, making it interpretable itself. In this manner, M-HAL super-learner is able to map a set of interpretable algorithms into a very powerful algorithm that is still interpretable. Another feature of HAL-MLE and thereby the M-HAL super-learner is that the HAL-fits have at most $n - 1$ non-zero coefficients, again simplifying its interpretation (and fast evaluation of the fitted function).

The application of this M-HAL super-learner goes beyond the treatment specific mean examples, suitable for causal estimation problems with longitudinal interventions [26–28] and mediation problems with (static or stochastic) interventions across multiple variables [29, 30]. We can design coordinate transformations of baseline and time-varying covariates such that the loss function of M-HAL-SL only depends on the reduced data as (2) and the EIC of the reduced data problem links to the EIC of the original data problem as (3). The transformed coordinates lead to ensembles with generally smaller sectional variation norms and thereby potentially more reliable asymptotic linearity, as well as reduced computational costs for other applicable meta-level estimators such as collaborative TMLE [23], targeted HAL-MLE [16], and higher-order spline-HAL [10].

An alternative and potentially more flexible approach is to define submodels where the conditional data likelihoods depend on full data only through summary covariates. For example, one can data-adaptively define summary covariates with predictors of conditional densities based on conditional hazards with exponential link functions. The pre-determined submodels can be viewed as a particular class of model constraints through dimension reduction. The plug-in at the projection of the true data distribution onto this submodel can be estimated as a projection target parameter with adaptive TMLE [31], where the oracle bias of the projection parameter, similar to that of the plug-in at the oracle ensemble of M-HAL-MLE, can be reasonably controlled. This approach relaxes the conditions that the EICs need to be strictly linked before and after the coordinate transformation, and constructs a richer class of asymptotically linear and possibly super-efficient estimators. Future work following this direction is applicable to the analysis of network and single time-series data [32, 33] or dimension reduction of other high-dimensional data such as electronic health records (EHR), imaging, and genomics.

Acknowledgments

This research is funded by NIH-grant R01AI074345-10A1.

Supplementary Materials

A Notation for Meta-HAL super-learner of functional parameter and its target features

O : Unit data structure/random variable

P : Possible probability distribution of O

$Pf = \int f(o)dP(o)$: Expectation operator w.r.t. P

P_0 : True probability distribution of O

O_1, \dots, O_n : n i.i.d. copies of $O \sim P_0$

P_n : Empirical measure of O_1, \dots, O_n

$P_nf = \int f(o)dP_n(o) = 1/n \sum_{i=1}^n f(O_i)$: Empirical mean operator

\mathcal{M} : Statistical model for P_0 , set of possible probability distributions including P_0

$Q : \mathcal{M} \rightarrow \mathcal{Q}$: Functional parameter of interest, where $Q(P) : \mathbb{R}^k \rightarrow [0, 1]$ is a k -variate $[0, 1]$ -valued function

$\mathcal{Q} = \{Q(P) : P \in \mathcal{M}\}$: Parameter space of Q consisting of k -variate real valued functions

$\Psi : \mathcal{M} \rightarrow \mathbb{R}^k$: Euclidean valued target parameter mapping P into $\Psi(P)$, chosen so that $\Psi(P)$ only depends on P through $Q(P)$

$\Psi(Q)$: Alternative notation for $\Psi(P)$, represents a target feature of Q

$D^*(P)$: Canonical gradient of pathwise derivative $\left. \frac{d}{d\epsilon} \Psi(P_\epsilon) \right|_{\epsilon=0} = PD^*(P)S$ of Ψ at P w.r.t. class of paths $\{P_\epsilon : \epsilon \in (-\delta, \delta)\}$ through P with score S

$G : \mathcal{M} \rightarrow \mathcal{G}$: Functional nuisance parameter $G(P)$ so that $D^*(P)$ only depends on P through $Q(P)$ and $G(P)$. If $D^*(P)$ only depends on $Q(P)$, then $G(P)$ is empty and can be ignored

$D^*(Q, G)$: Alternative notation for $D^*(P)$

\mathcal{G} : Parameter space of G defined as $\mathcal{G} = \{G(P) : P \in \mathcal{M}\}$

$R_2(P, P_0)$: Notation for exact second order remainder $R_2(P, P_0) \equiv \Psi(P) - \Psi(P_0) + P_0 D^*(P)$ for $\Psi(P) - \Psi(P_0)$

$R_{20}(Q, G, Q_0, G_0)$: Alternative notation for $R_2(P, P_0)$

$L(Q)(o)$: Loss function for Q so that $Q_0 = Q(P_0) = \arg \min_{Q \in \mathcal{Q}} P_0 L(Q)$

$d_0(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0)$: Excess risk of Q , loss-based dissimilarity

$L_1(G)(o)$: Loss function for G so that $G_0 = G(P_0) = \arg \min_{G \in \mathcal{G}} P_0 L_1(G)$

$d_{01}(G, G_0)$: Excess risk of G , loss-based dissimilarity

\mathcal{M}_{np} : Set of all possible empirical probability measures that can occur as a realization of P_n for any sample size n

$\hat{Q}_j : \mathcal{M}_{np} \rightarrow \mathcal{Q}$: An estimator that maps an empirical measure P_n (e.g., of training sample) into an element of \mathcal{Q} , $j = 1, \dots, J$

$\hat{\mathbf{Q}} = (\hat{Q}_j : j = 1, \dots, J)$: Collection of J estimators, in context of super-learner it is called the library of J estimators of Q_0

(v_i, O_i) : Using V -fold sample splitting, each observation O_i gets assigned a value $v_i \in \{1, \dots, V\}$. For each v , it defines a v -th sample split in validation sample $\{O_i : v_i = v\}$ and training sample $\{O_i : v_i \neq v\}$

$P_{n,v}^1$: Empirical measure of validation sample $\{O_i : v_i = v\}$ (approximately n/V observations) for the v -th sample split

$P_{n,v}$: Empirical measure of training sample $\{O_i : v_i \neq v\}$ (approximately $n - n/V$ observations)

$\mathbf{Q}_{n,v} = \hat{\mathbf{Q}}(P_{n,v})$: Vector of J estimates based on training sample $P_{n,v}$, $v = 1, \dots, V$

$\mathbf{Q}_n(v, x) = \mathbf{Q}_{n,v}(x)$: representing the collection of V J -dimensional vector of estimates based on training sample $P_{n,v}$, $v = 1, \dots, V$

$P^{\bar{V}}$: Probability measure of (\bar{V}, O) implied by P defined by $\bar{V} \sim U\{1, \dots, V\}$ and conditional probability measure of O , given $\bar{V} = v$, equals P

$P^{\bar{V}} f = \frac{1}{V} \sum_{v=1}^V \int f(v, o) dP(o)$

$P_n^{\bar{V}}$: Empirical measure of (v_i, O_i) , $i = 1, \dots, n$

$P_n^{\bar{V}} f = \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 f(v, \cdot)$

- \mathcal{Q}^r : Collection of real valued cadlag functions $Q^r : [0, 1]^J \rightarrow \mathbb{R}$ with sectional variation norm $\|Q^r\|_v^*$ bounded by some $C^u < \infty$. We also refer to this as collection of candidate ensembles of J estimators
- $x_s = (x(j) : j \in s)$: subvector of $x \in [0, 1]^J$ defined by subset $s \subset \{1, \dots, J\}$
- $x_{-s} = (x(j) : j \notin s)$
- $y = (x_s, 0_{-s})$: vector defined by $y(j) = x(j)$ if $j \in s$ and $y(j) = 0$ if $j \notin s$
- $Q_s^r(x_s) = Q^r(x_s, 0_{-s})$: s -specific section of Q^r that sets coordinates in complement of s equal to 0 and is viewed as function on $|s|$ dimensional s -specific edge $E_s \equiv \{x \in [0, 1]^J : x_{-s} = 0\}$ of $[0, 1]^J$. Note $[0, 1]^J = \cup_{s \subset \{1, \dots, J\}} E_s$
- $\|Q^r\|_v^* = |Q^r(0)| + \sum_{s \subset \{1, \dots, J\}} \int_{(0_s, 1_s]} |dQ_s^r(u)|$: sectional variation norm of function $Q^r : [0, 1]^J \rightarrow \mathbb{R}$
- $Q^r(x) = \sum_{s \subset \{1, \dots, J\}} \int \phi_{s, x_s}(u) dQ_s^r(u)$: representation of cadlag function Q^r as infinitesimal linear combination of tensor products of zero-spline basis functions, $x \rightarrow \phi_{s, x_s}(u) = I(u \leq x_s)$, with knot point u . By convention, this sum includes intercept $Q^r(0)$ (corresponding with empty set s)
- Q^r for discrete measures: Note that if Q_s^r generates discrete measure dQ_s^r , then $Q^r(x) = \sum_{(s, j)} \beta^r(s, j) \phi_{s, j}(x)$ where $\beta^r(s, j) = dQ_s^r(u_{s, j})$ at support point $u_{s, j}$ of dQ_s^r
- $Q^r \circ \mathbf{Q}_n(v, x) = Q^r(\mathbf{Q}_n(v, x))$: composition of ensemble Q^r with J -dimensional vector $\mathbf{Q}_{n, v}$ of estimated functions based on $P_{n, v}$
- $Q_{0, n}^r = \arg \min_{Q^r \in \mathcal{Q}^r} P_0^{\bar{V}} L(Q^r \circ \mathbf{Q}_n)$: oracle ensemble that minimizes the conditional risk $1/V \sum_v P_0 L(Q^r \circ \mathbf{Q}_{n, v})$
- $Q_{0, n}(v, x) = Q_{0, n}^r \circ \mathbf{Q}_n(v, x)$: v -specific oracle estimator defined by applying the oracle ensemble $Q_{0, n}^r$ to the J estimates $\mathbf{Q}_{n, v}$, $v = 1, \dots, V$
- $\bar{Q}_{0, n}(x) = \frac{1}{V} \sum_{v=1}^V Q_{0, n}(v, x)$: Single oracle estimator obtained from $Q_{0, n}$ by averaging the v -specific oracle estimates $Q_{0, n, v}$ across the V sample splits
- $Q_n^r = \arg \min_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^{\bar{V}} L(Q^r \circ \mathbf{Q}_n)$: M-HAL-MLE of oracle ensemble $Q_{0, n}^r$ using bound C_n for sectional variation norm, minimizing the cross-validated risk $1/V \sum_v P_{n, v}^1 L(Q^r \circ \mathbf{Q}_{n, v})$ over all ensemble specific estimators $Q^r \circ \hat{\mathbf{Q}} : \mathcal{M}^{np} \rightarrow \mathcal{Q}$. It is thus the cross-validation selector for this class of ensemble specific estimators
- Q_n^{rC} Q_n^r using C as bound on sectional variation norm
- $Q_n^r = \sum_{(s, j)} \beta_n^r(s, j) \phi_{s, j}$: finite dimensional representation of Q_n^r due to the unrestricted M-HAL-MLE Q_n^r being discrete, or due to choosing it to be discrete on a user rich set of knot-points
- C_n : bound on sectional variation norm enforced in Q_n^r . $C_n \geq C_{n, cv}$, where $C_{n, cv} = \arg \min_C 1/V \sum_{v=1}^V P_{n, v}^1 L(Q_n^{r, C})$ is the cross-validation selector
- $\mathcal{J}_n(C_n)$: Set of coefficient-indices (s, j) with $\beta_n^r(s, j) \neq 0$
- $C_0^v \equiv \|Q_{0, n}^r\|_v^*$: sectional variation norm of oracle selector $Q_{0, n}^r$
- M-HAL-MLE: Meta Highly Adaptive Lasso Minimum Loss Estimator Q_n^r
- $Q_{n, v}(x) = Q_n(v, x) = Q_n^r \circ \mathbf{Q}_n(v, x)$: v -specific M-HAL super-learner (of oracle estimator $Q_{0, n, v} = Q_{0, n}(v, \cdot)$) defined by applying the M-HAL-MLE Q_n^r to the J -dimensional vector $\mathbf{Q}_{n, v}$ of estimates, $v = 1, \dots, V$
- $d_0^{\bar{V}}(Q_n, Q_{0, n}) = P_0^{\bar{V}} L(Q_n) - P_0^{\bar{V}} L(Q_{0, n})$: excess risk or M-HAL SL relative to oracle estimator $Q_{0, n}$, which is equal to $1/V \sum_v P_0 \{L(Q_{n, v}) - L(Q_{0, n, v})\}$
- $\bar{Q}_n(x) = \frac{1}{V} \sum_{v=1}^V Q_n(v, x)$: single M-HAL super-learner defined by averaging the v -specific M-HAL super-learners $Q_{n, v} = Q_n(v, \cdot)$ across the V sample splits
- $d_0(\bar{Q}_n, Q_0)$: excess risk of M-HAL SL
- M-HAL SL: Meta Highly Adaptive Lasso Super Learner defined by Q_n or by \bar{Q}_n
- $P_n^{\bar{V}} D^*(Q_n, G_{0, n}^r) = o_P(n^{-1/2})$: the cross-validated efficient influence curve equation $1/V \sum_v P_{n, v}^1 D^*(Q_{n, v}, G_{0, n, v}^r) = o_P(n^{-1/2})$. If this equation holds for the M-HAL SL Q_n by selecting C_n large enough in definition of HAL-MLE Q_n^r , then a standard analysis shows that $\Psi^{\bar{V}}(Q_n)$ is asymptotically linear

B Notation for equivalent estimation problem implied by treating \mathbf{Q}_n as a fixed coordinate transformation

B.1 Explanation

The most important task is the analysis of $Q_n = Q_n^r \circ \mathbf{Q}_n$ as an estimator of $Q_{0,n} = Q_{0,n}^r \circ \mathbf{Q}_n$ and its target features $\Psi^{\bar{V}}(Q_{0,n}) = 1/V \sum_{v=1}^V \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v})$. The remaining bias term $d_0(Q_{0,n}, Q_0)$ is typically of significantly smaller order (or even 0). If we treat \mathbf{Q}_n as fixed, and view $Q_n(v, x) = Q_n^r \circ \mathbf{Q}_{n,v}(x)$ and $Q_{0,n}(v, x) = Q_{0,n}^r \circ \mathbf{Q}_{n,v}(x)$ as a function in the new J coordinates $\mathbf{Q}_n(v, x)$ instead of (v, x) , then Q_n becomes $Q_n^r(\cdot)$ and $Q_{0,n}$ becomes $Q_{0,n}^r(\cdot)$. In particular, $\Psi^{\bar{V}}(Q_{0,n}) = \frac{1}{V} \sum_{v=1}^V \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v})$ is now only a target feature of $Q_{0,n}^r$, and $\Psi(Q_n)$ is now a target feature of Q_n^r . As a consequence, Q_n^r is just a regular HAL-MLE of $Q_{0,n}^r$ with this new coordinate transformation $(v, x) \rightarrow \mathbf{Q}_n(v, x)$ based on data set (v_i, O_i) , $i = 1, \dots, n$. In addition, the data (v_i, O_i) can be recoded in terms of the new coordinates \mathbf{Q}_n resulting in a reduction (v_i, O_i^r) . This then teaches us that if \mathbf{Q}_n would truly be fixed, all our previous results for standard HAL-MLE based on i.i.d. data, including efficient plug-in estimation of target features, can be applied to this Q_n^r as estimator of $Q_{0,n}^r$. The reason that the data dependence of \mathbf{Q}_n does not cause issues is due to the loss at (v_i, O_i) using the transformation \mathbf{Q}_{n,v_i} based on the training sample excluding O_i , allowing conditioning on $\mathbf{Q}_{n,v}$ whenever dealing with an empirical process w.r.t. $P_{n,v}^1$.

B.2 Notation for equivalent estimation problem treating \mathbf{Q}_n as fixed coordinate transformation

$O_v^r = O^r(v, O)$: reduction of O chosen so that the loss $L(Q^r \circ \mathbf{Q}_{n,v})(o)$ of candidate $Q^r \circ \mathbf{Q}_{n,v} \in \mathcal{Q}$ only depends on o through $o^r(v, o)$

$L^r(Q^r)(v, o^r)$: reduced data loss defined by $L^r(Q^r)(v, O^r(v, o)) = L(Q^r \circ \mathbf{Q}_{n,v})(o)$, $v = 1, \dots, V$

$O^r = O^r(\bar{V}, O)$: viewed as a random variable implied by distribution of $(\bar{V}, O) \sim P_0^{\bar{V}}$ treating \mathbf{Q}_n as fixed (i.e., non random fixed functions)

P^r : probability distribution of (\bar{V}, O^r) implied by $P^{\bar{V}}$

P_0^r : true probability distribution of (\bar{V}, O^r) implied by $P_0^{\bar{V}}$

$d_0^r(Q^r, Q_{0,n}^r) = P_0^r L^r(Q^r) - P_0^r L^r(Q_{0,n}^r)$

$\mathcal{M}^r = \{P^r : P \in \mathcal{M}\}$: statistical model for $(\bar{V}, O^r) \sim P_0^r$, again, treating \mathbf{Q}_n as fixed

P_n^r : empirical measure of $(v_i, O_i^r = O^r(v_i, O_i))$, $i = 1, \dots, n$

$Q^r : \mathcal{M}^r \rightarrow \mathcal{Q}^r$: $Q^r(P^r) = \arg \min_{Q^r \in \mathcal{Q}^r} P^r L^r(Q^r)$

$Q_0^r = Q^r(P_0^r)$

$\Psi^r : \mathcal{M}^r \rightarrow \mathbb{R}$: defined by $\Psi^r(P^r) = \frac{1}{V} \sum_{v=1}^V \Psi(Q^r \circ \mathbf{Q}_{n,v})$

$\Psi^r(Q^r)$: alternative notation for $\Psi^r(P^r)$ to emphasize it only depends on P^r through Q^r

$D^r(P^r)(\bar{V}, O^r)$: canonical gradient of Ψ^r at P^r

$G^r : \mathcal{M}^r \rightarrow \mathcal{G}^r$: nuisance parameter so that $D^r(P^r)$ only depends on P^r through Q^r and G^r . $G_{0,n}^r = G^r(P_0^r)$ true nuisance parameter. It is chosen so that $G_v^r : \mathcal{M}_v^r \rightarrow \mathcal{G}$ maps distribution P_v^r of O_v^r into parameter space \mathcal{G} of G

$D^r(Q^r, G^r)$: alternative notation for canonical gradient $D^r(P^r)$

$R_{20}(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) \equiv \Psi^r(Q^r) - \Psi^r(Q_{0,n}^r) + P_0^r D^r(Q^r, G^r)$: exact second order remainder for Ψ^r

$P_n^r D^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$: efficient influence curve equation that would be solved by HAL-MLE Q_n^r when selecting C_n large enough

B.3 Equivalences

We now have the following equivalences between our estimation problem and corresponding fixed \mathbf{Q}_n -formulation of the estimation problem defined above in model \mathcal{M}^r :

$$\begin{aligned}
 Q_{0,n}^r &= \arg \min_{Q^r \in \mathcal{Q}^r} P_0^r L^r(Q^r) = \arg \min_{Q^r \in \mathcal{Q}^r} P_0^{\bar{V}} L(Q^r \circ \mathbf{Q}_n) \\
 Q_n^r &= \arg \min_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^r L^r(Q^r) \\
 &= \arg \min_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^{\bar{V}} L(Q^r \circ \mathbf{Q}_{n,\bar{V}}) \\
 \Psi^{\bar{V}}(Q_{0,n}) &= \Psi^r(Q_{0,n}^r) \\
 \Psi^{\bar{V}}(Q_n) &= \Psi^r(Q_n^r) \\
 d_0^{\bar{V}}(Q_n, Q_{0,n}) &= d_0^r(Q_n^r, Q_{0,n}^r) \\
 D^r(Q^r, G^r)(\bar{V}, O^r) &= D^*(Q^r \circ \mathbf{Q}_{n,\bar{V}}, G_{\bar{V}}^r)(O) \text{ by assumption (5)} \\
 R_{20}^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) &= \frac{1}{V} \sum_{v=1}^V R_{20}(Q^r \circ \mathbf{Q}_{n,v}, G_v^r, Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r) \\
 P_n^r D^r(Q_n^r, G_{0,n}^r) &= P_n^{\bar{V}} D^*(Q_n, G_{0,n}^r)
 \end{aligned}$$

C Proofs for Convergence Rate of M-HAL-SL

Proof of Lemma 4.1: We have

$$\begin{aligned}
 0 &\leq d_0^{\bar{V}}(Q_n, Q_{0,n}) \\
 &= \frac{1}{V} \sum_{v=1}^V P_0 \{L(Q_{n,v}) - L(Q_{0,n,v})\} \\
 &= -\frac{1}{V} \sum_{v=1}^V (P_{n,v}^1 - P_0) \{L(Q_{n,v}) - L(Q_{0,n,v})\} \\
 &\quad + \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 \{L(Q_{n,v}) - L(Q_{0,n,v})\} \\
 &\leq -\frac{1}{V} \sum_{v=1}^V (P_{n,v}^1 - P_0) \{L(Q_{n,v}) - L(Q_{0,n,v})\},
 \end{aligned}$$

where the last inequality is by definition of Q_n^r and thus Q_n . For a given v , conditional on the training sample so that $\mathbf{Q}_{n,v}$ is fixed, we have

$$\begin{aligned}
 &| (P_{n,v}^1 - P_0) \{L(Q_{n,v}) - L(Q_{0,n,v})\} | \\
 &\leq \sup_{Q^r, Q_1^r \in \mathcal{Q}^r} | (P_{n,v}^1 - P_0) \{L(Q^r \circ \mathbf{Q}_{n,v}) - L(Q_1^r \circ \mathbf{Q}_{n,v})\} | \\
 &= \sup_{Q^r, Q_1^r \in \mathcal{Q}^r} | (P_{n,v}^{r,1} - P_{0,v}^r) \{L^r(Q^r) - L^r(Q_1^r)\} |
 \end{aligned}$$

For a given v , $\mathcal{F}_1^r \equiv \{L^r(Q^r) - L^r(Q_1^r) : Q^r, Q_1^r \in \mathcal{Q}^r\}$ are cadlag functions of $O^r(v, O)$ with a universal bound on its sectional variation norm. Let d^r be the dimension of O^r . The typical HAL-MLE proof now proceeds with the following ingredients: 1) \mathcal{F}_1^r is a Donsker class with bracketing entropy number $\log N_{[]}(\epsilon, \mathcal{F}_1^r, L^2(P^r)) \lesssim \epsilon^{-1}(\log \epsilon)^{-d}$ (Proposition 2 in [4]); 2) for each v , $P_{0,v}^r \{L^r(Q^r) - L(Q_1^r)\}^2 \leq M_2^r d_0^r(Q^r, Q_1^r)$; 3) the bracketing entropy integral is bounded as follows $J_{[]}(\delta, \mathcal{F}_1^r, L^2(P^r)) \lesssim \delta^{1/2}(\log \delta)^{-d/2}$ [4]; the modulus of discontinuity for the empirical process can be bounded accordingly as

$$\sup_{f, \|f\| < \delta} |G_n(f)| \lesssim J_{[]}(\delta, \mathcal{F}_1^r, L^2(P^r)) \left(1 + \frac{J_{[]}(\delta, \mathcal{F}_1^r, L_2(P^r))}{\delta^2 n^{1/2}} M \right)$$

(van der Vaart and Wellner [34] and Lemma 3.4.2 in van der Vaart and Wellner [35]). This allows us to apply the iterative HAL-MLE proof in Appendix of [1] or direct proof [4] to establish that $d_0(Q_n, Q_{0,n}) = O_P(n^{-2/3}(\log n)^{d^*})$. For example, [4] gives this result as a corollary.

□

D Proofs for Asymptotic Linearity Theorem

Let $J(\delta, \mathcal{F}, L_2)$ denote the uniform entropy integral for a class \mathcal{F} with respect to the $L^2(P_0)$ norm. Define $\mathbb{G}_n f = \sqrt{n}(P_n - P_0)f$ so that $\mathbb{G}_{\frac{n}{V}, v} f = \sqrt{n/V}(P_{n,v}^1 - P_0)f$. Define

$$Q_0^r = \arg \min_{Q^r} \frac{1}{V} \sum_{v=1}^V P_0 L(Q^r \circ Q^*)$$

for some Q^* such that $\|Q_{n,v} - Q^*\|_{P_0}^2 \xrightarrow{P} 0$ for all v . Assume that G^r depends on O and $Q_n(\bar{V}, X)$ through a transformation of $O^r(\bar{V}, O)$ so that $G^r \in \mathcal{G}^r$ and $Q_{n,v}(X)$ define $G_v^r \in \mathcal{G}$. Use notation $G^r(Q_{n,v}) \equiv G_v^r$ to highlight the definition of G_v^r using G^r and $Q_{n,v}(X)$. For each $G^r \in \mathcal{G}^r$, replace $Q_{n,v}(X)$ with $Q^*(X)$ in the definition of G_v^r to define $G_*^r(Q^*) \in \mathcal{G}$. Also assume there exists $G_0^r \in \mathcal{G}^r$ such that $\|G_{0,n,v}^r - G_{0,*}^r(Q^*)\|_{P_0}^2 \xrightarrow{P} 0$ for all v .

Lemma D.1. *Define*

$$\begin{aligned} f(Q^r, G^r, Q) &= D^*(Q^r \circ Q, G^r(Q)) - D^*(Q_0^r \circ Q, G_{0,*}^r(Q)) \\ f_{n,v} &= f(Q_n^r, G_{0,n}^r, Q_{n,v}) = D^*(Q_n^r \circ Q_{n,v}, G_{0,n,v}^r) - D^*(Q_0^r \circ Q_{n,v}, G_{0,v}^r) \\ \mathcal{F}_{nv} &= \{f(Q^r, G^r, Q_{n,v}) : Q^r \in \mathcal{Q}^r, G^r \in \mathcal{G}^r\} \\ \mathcal{F} &= \cup \{\mathcal{F}_{nv} : n = 1, 2, \dots; v = 1, \dots, V\}. \end{aligned}$$

Assume

1. $P_0 f_{n,v}^2 \xrightarrow{P} 0$ uniformly for all v ,
2. \mathcal{F} is a P_0 -measurable class with envelope function F such that $|f| < F < c < \infty$ for some constant c and for all $f \in \mathcal{F}$,
3. there exists a positive sequence $1 > \delta_n \rightarrow 0$ at a slow rate such that $\frac{1}{\log(n)\delta_n^2} \rightarrow 0$ and $\frac{P_0 f_{n,v}^2}{\delta_n^2} \xrightarrow{P} 0$ (uniformly over v), and $\sum_{v=1}^V J(\delta_n, \mathcal{F}_{nv}, L_2) \xrightarrow{P} 0$.

Then, we have that for fixed V ,

$$\begin{aligned} & \frac{1}{V} \sum_{v=1}^V (P_{n,v}^1 - P_0) \left\{ D^*(Q_n^r \circ Q_{n,v}, G_{0,n,v}^r) - D^*(Q_0^r \circ Q_{n,v}, G_{0,v}^r) \right\} \\ & \equiv \frac{1}{\sqrt{nV}} \sum_{v=1}^V \mathbb{G}_{\frac{n}{V}, v} f_{n,v} = o_P(n^{-1/2}). \end{aligned}$$

Proof. Define $f_{n,v}^* = f_{n,v} \mathbb{I}_{\{f_{n,v}^2 \leq \delta_n^2 P_0 F^2\}}$. Then F is also an envelope for $\mathcal{F}_{nv}^* \equiv \{f \mathbb{I}_{\{f^2 \leq \delta_n^2 P_0 F^2\}} : f \in \mathcal{F}_{nv}\}$, and $J(\delta, \mathcal{F}_{nv}^*, L_2) \leq J(\delta, \mathcal{F}_{nv}, L_2)$. Using conditions 2 and 3, by Theorem 2.1 of [36], we have

$$\begin{aligned} E_{P_0} \|\mathbb{G}_{\frac{n}{V}, v} f_{n,v}\|_{\mathcal{F}_{nv}^*} & \lesssim J(\delta_n, \mathcal{F}_{nv}, L_2) \left(1 + \frac{J(\delta_n (\log(1/\delta_n))^{1/p}, \mathcal{F}_{nv}, L_2)}{\delta_n^2 \sqrt{n/V} \|F\|_{P_0}} \right) \|F\|_{P_0} \\ & = O_P(J(\delta_n, \mathcal{F}_{nv}, L_2)). \end{aligned}$$

Then,

$$\begin{aligned} |\mathbb{G}_{\frac{n}{V}, v} f_{n,v}| & \leq |\mathbb{G}_{\frac{n}{V}, v} f_{n,v} \mathbb{I}_{\{f_{n,v}^2 \leq \delta_n^2 P_0 F^2\}}| + |\mathbb{G}_{\frac{n}{V}, v} f_{n,v} \mathbb{I}_{\{f_{n,v}^2 > \delta_n^2 P_0 F^2\}}| \\ & \leq E_{P_0} \|\mathbb{G}_{\frac{n}{V}, v} f_{n,v}\|_{\mathcal{F}_{nv}^*} = O_P(J(\delta_n, \mathcal{F}_{nv}, L_2)), \end{aligned}$$

noting that $\mathbb{I}_{\{f_{n,v}^2 > \delta_n^2 P_0 F^2\}} \stackrel{P}{=} 0$ with large enough n . Thus,

$$\left| \frac{1}{\sqrt{V}} \sum_{v=1}^V \mathbb{G}_{\frac{n}{V},v} f_{n,v} \right| \leq O_P\left(\sum_{v=1}^V J(\delta_n, \mathcal{F}_{n,v}, L_2)\right) = o_P(1)$$

by using condition 3. \square

Remark 1. Condition 3 is a much weaker assumption than a Donsker class condition over $\{D^*(Q, G) : Q \in \mathcal{Q}, G \in \mathcal{G}\}$ for the original data problem. For example, if for fixed $\mathbf{Q}_{n,v}$, $\{f(Q^r, G^r, \mathbf{Q}_{n,v}) : Q^r \in \mathcal{Q}^r, G^r \in \mathcal{G}^r\}$ is a class of cadlag functions with a sectional variation norm bound not depending on $\mathbf{Q}_{n,v}$ — which is a reasonable assumption due to the bounded sectional variation norms of $\mathcal{Q}^r, \mathcal{G}^r$ — then $J(\delta_n, \mathcal{F}, L_2) = O_P(\sqrt{\delta_n}) = o_P(1)$ uniformly across all possible $\mathbf{Q}_{n,v}$. Such (uniform) Donsker class conditions defined on the meta level may be easier to achieve, and have an advantage when the original data problem is complex and all reasonable initial estimators are highly variable.

Lemma D.2. Suppose that

1. there exist uniform convergence limits $\tilde{Q}_0 \in \mathcal{Q}$ and $\tilde{G}_0 \in \mathcal{G}$: for all $\epsilon > 0$, there exists N such that for all $n \geq N$, $v = 1, \dots, V$, and all realizations of $\mathbf{Q}_{n,v}$,

$$P_0\{D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r) - D^*(\tilde{Q}_0, \tilde{G}_0)\}^2 < \epsilon;$$

denote $h(\mathbf{Q}) = D^*(Q_{0,n}^r \circ \mathbf{Q}, G_{0,n}^r) - D^*(\tilde{Q}_0, \tilde{G}_0)$,

2. $\sup_{\mathbf{Q}_{n,v}} \|h(\mathbf{Q}_{n,v})\|_\infty \leq M_n$ for some $M_n < \infty$, and $\lim_{n \rightarrow \infty} M_n < \infty$ or $M_n \rightarrow \infty$ at a rate slower than \sqrt{n} so that $M_n/\sqrt{n} \rightarrow 0$.

Then we have

$$\frac{1}{V} \sum_{v=1}^V (P_{n,v}^1 - P_0) \left\{ D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r) - D^*(\tilde{Q}_0, \tilde{G}_0) \right\} = \frac{1}{\sqrt{nV}} \sum_{v=1}^V \mathbb{G}_{\frac{n}{V},v} h(\mathbf{Q}_{n,v}) = o_P(n^{-1/2}).$$

Proof. By Bernstein's Inequality, for all $a > 0$,

$$E_{P_0}(\mathbb{I}_{\{|\mathbb{G}_{n/V,v} h(\mathbf{Q}_{n,v})| > a\}} | \mathbf{Q}_{n,v}) \leq 2 \exp\left\{-\frac{1}{4} \frac{a^2}{E_{P_0}[h(\mathbf{Q}_{n,v})^2 | \mathbf{Q}_{n,v}] + aM/\sqrt{n/V}}\right\}$$

Due to the uniform bound on $h(\mathbf{Q}_{n,v})$ across all realizations of $\mathbf{Q}_{n,v}$, by double expectations, for all small enough $\epsilon > 0$ and $\lambda = -\frac{a^2}{4 \log(\epsilon/2)} > 0$, there exists N such that for all $n \geq N$, $aM_n/\sqrt{n/V} \leq \lambda/2$, and

$$P_0(|\mathbb{G}_{n/V,v} h(\mathbf{Q}_{n,v})| > a) \leq E_{P_0} 2 \exp\left\{-\frac{1}{4} \frac{a^2}{\frac{\lambda}{2} + \frac{\lambda}{2}}\right\} \leq \epsilon.$$

\square

We note that, so far, we have not assumed any Donsker class assumption on \mathbf{Q}_n (we only relied on the oracle selected ensemble of \mathbf{Q}_n to be a good estimator of $Q_{0,n}$). This again demonstrates that the asymptotic properties of M-HAL-MLE features do not rely on a Donsker class assumption on $\mathcal{D}^* \equiv \{D^*(Q, G) : Q \in \mathcal{Q}, G \in \mathcal{G}\}$, but on a Donsker class condition on the meta level with fixed $\mathbf{Q}_{n,v}$.

$$\Psi^r(Q_{0,n}^r) - \Psi^r(Q_{0,n}^r) = P_n D^*(Q_0, \tilde{G}_0) + o_P(n^{-1/2}).$$

E Proofs for Difference of Target Feature of Oracle Estimator and Target Estimand

Proof of Theorem 5.2: Recall that $D^r(P^r) = D^r(Q^r)$ is the canonical gradient of $\Psi^r : \mathcal{M}^r \rightarrow \mathbb{R}$. Thus, $P_0^r D^r(Q_{0,n}^r) = 0$. We also have that $D^r(Q_{0,n}^r)(v, O^r(v, o)) = D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v})(o)$, so that $P_0^r D^r(Q_{0,n}^r) = 0$

implies $\frac{1}{V} \sum_{v=1}^V P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) = 0$. Now, use that, by definition of R_{20} ,

$$\Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) = -P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) + R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, Q_0).$$

Recall notation $Q_{0,n,v} = Q_{0,n}^r \circ \mathbf{Q}_{n,v}$. Thus, this proves

$$\begin{aligned} \left\{ \frac{1}{V} \sum_v \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) \right\} &= -\frac{1}{V} \sum_{v=1}^V P_0 D^*(Q_{0,n,v}) \\ &\quad + \frac{1}{V} \sum_{v=1}^V R_{20}(Q_{0,n,v}, Q_0) = \frac{1}{V} \sum_{v=1}^V R_{20}(Q_{0,n,v}, Q_0). \end{aligned}$$

By assumption (5), we have $\frac{1}{V} \sum_v R_{20}(Q_{0,n,v}, Q_0) = O(d_0(Q_{0,n}, Q_0))$. \square

In our treatment specific mean example we could have defined $G_{0,n,v}^*$ as the conditional mean of A , given $\mathbf{Q}_{n,v}(W)$, $\mathbf{Q}_0(W)$, where \mathbf{Q}_0 represents the limit of \mathbf{Q}_n , while $G_{0,n,v}^r = E_0(A \mid \mathbf{Q}_{n,v}(W))$. As in our example, we then note that $G_{0,n,v}^r$ and $G_{0,n,v}^*$ are conditional expectations in which $G_{0,n,v}^*$ conditions on an extra $\mathbf{Q}_0(W)$ coming from the limit of $\mathbf{Q}_{n,v}$. As in our example, we can then utilize that conditional expectations are projection operators to bound the L^2 -norm of the difference in terms of $\mathbf{Q}_{n,v}$ and \mathbf{Q}_0 . We improved on this approach by selecting $G_{0,n,v}^*$ as the conditional mean of A , given $Q_{0,n,v}(W)$, $Q_0(W)$, while still following the same subsequent steps. This allowed us to bound the L^2 -norm of the difference of $G_{0,n,v}^* - G_{0,n,v}^r$ in terms of the difference of the oracle ensemble $Q_{0,n}^r \circ \mathbf{Q}_{n,v}$ of $\mathbf{Q}_{n,v}$ and Q_0 , instead of a difference of the J -dimensional $\mathbf{Q}_{n,v}$ versus \mathbf{Q}_0 . In this manner we obtained a natural bound $d_0^{\tilde{V},1/2}(Q_{0,n}, Q_0)$ for the L^2 -norm of $G_{0,n,v}^* - G_{0,n,v}^r$. This insight clearly suggests that one should aim to define $G_{0,n,v}^*$ with minimal conditioning, even though conditioning on $\mathbf{Q}_{n,v}$ and \mathbf{Q}_0 would suffice.

Proof of Theorem 5.3: We have

$$\begin{aligned} \left\{ \frac{1}{V} \sum_v \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) \right\} &= -\frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0) \\ &\quad + \frac{1}{V} \sum_v R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0, Q_0, G_0). \end{aligned}$$

By assumption (5), $\frac{1}{V} \sum_v R_{20}(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0, Q_0, G_0) = O_P(d_0^{\tilde{V}}(Q_{0,n}, Q_0))$. It remains to analyze $\frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0)$. By definition of $G_{0,n,v}^*$ we have

$$\frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_0) = \frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^*).$$

So now it remains to analyze $\frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^*)$.

We have $0 = P_0^r D^r(Q_{0,n}^r, G_{0,n}^r)$, which equals $\frac{1}{V} \sum_{v=1}^V P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r)$. Thus, it follows that $\frac{1}{V} \sum_v P_0 D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r) = 0$. Subtracting this from our expression yields the following expression for $\Psi^r(Q_{0,n}^r) - \Psi(Q_0)$

$$\begin{aligned} &\frac{1}{V} \sum_v \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) \\ &= \frac{1}{V} \sum_v P_0 \{ D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}) - D^*(Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^*) \} + O_P(d_0^{\tilde{V}}(Q_{0,n}, Q_0)) \end{aligned}$$

Recall $Q_{0,n,v} = Q_{0,n}^r \circ \mathbf{Q}_{n,v}$. Using $\Psi(Q_{0,n,v}) - \Psi(Q_0) = -P_0 D^*(Q_{0,n,v}, G_{0,n,v}) + R_2(Q_{0,n,v}, G_{0,n,v}, Q_0, G_0)$, and $\Psi(Q_{0,n,v}) - \Psi(Q_0) = -P_0 D^*(Q_{0,n,v}, G_{0,n,v}^*) + R_2(Q_{0,n,v}, G_{0,n,v}^*, Q_0, G_0)$, it follows that the leading term on right-hand side above equals r_n . \square

F Undersmoothing Conditions

Here we will prove a formal Theorem F.1 establishing sufficient conditions for $P_n^r D^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$, involving an undersmoothing condition (11) that will hold by selecting C_n large enough, and a condition

(13). The latter is shown to be about the linear span of the basis functions in Q_n^r with non-zero coefficients approximating a function having to do with $G_{0,n}^r$.

Recall $Q_n^r = \arg \min_{Q^r \in \mathcal{Q}^r, \|Q^r\|_v^* < C_n} P_n^r L^r(Q^r)$. Consider a path $\{Q_{n,\epsilon}^{r,h} : \epsilon\}$, indexed by a function h , defined by

$$Q_{n,\epsilon}^{r,h} = (1 + \epsilon h(0))Q_n^r(0) + \sum_{s \subset \{1, \dots, d\}_{(0_s, x_s)}} \int (1 + \epsilon h(s, u_s)) dQ_{n,s}^r(u_s), \quad (9)$$

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

$$r(h, Q_n^r) \equiv h(0) | Q_n^r(0) | + \sum_{s \subset \{1, \dots, d\}_{(0_s, \tau_s)}} \int (1 + \epsilon h(s, u_s)) | dQ_{n,s}^r(u_s) |.$$

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

$$S_h(Q_n^r) \equiv \frac{d}{d\epsilon} L^r(Q_{n,\epsilon}^{r,h})|_{\epsilon=0}. \quad (10)$$

By (5) (on loss L^r), it also follows that $\{S_h(Q^r) : Q^r \in \mathcal{Q}^r\} \subset \mathcal{D}_{d^r, v}[0, \tau^r]$.

Since Q_n^r minimizes this empirical risk over all $Q^r \in \mathcal{Q}^r(C_n)$, we know that $P_n^r S_h(Q_n^r) = 0$ for all uniformly bounded h with $r(h, Q_n^r) = 0$. Thus, we have $P_n^r S_h(Q_n^r) = 0$ for all bounded h with $r(h, Q_n^r) = 0$. Let $\mathcal{S}(Q_n^r) = \{S_h(Q_n^r) : h\}$ be the linear span of all these score functions $S_h(Q_n^r)$ indexed by any bounded function h .

Analogue to the proof of the theorems in [10] we can now establish that for large enough C_n , the linear span of the score equations $P_n^r S_h(Q_n^r)$ with $r(h, Q_n^r) = 0$ approximate the efficient influence curve equation $P_n^r D^r(Q_n^r, G_{0,n}^r)$. This works as follows. Let $D_n^r(Q_n^r, G_{0,n}^r)(v, o^r) = D_n^*(Q^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r)(o)$ be an approximation of $D^r(Q_n^r, G_{0,n}^r)$ that is contained in $\mathcal{S}(Q_n^r) = \{S_h(Q_n^r) : h\}$, without the restriction $r(h, Q_n^r) = 0$. Let $h^*(Q_n^r, G_{0,n}^r)$ be the corresponding index so that $D_n^r(Q_n^r, G_{0,n}^r) = S_{h^*(Q_n^r, G_{0,n}^r)}(Q_n^r)$. For notational convenience, in this proof let's denote it with h^* . Recall $Q_n^r = \sum_{(s,j) \in \mathcal{J}_n(C_n)} \beta_n^r(s,j) \phi_{s,j}$. Let $\tilde{h}(s,j) = h^*(s,j)$ except at one $(s^*, j^*) \in \mathcal{J}_n(C_n)$ and defined such that $r(\tilde{h}, Q_n^r) = \sum_{(s,j) \in \mathcal{J}_n(C_n)} \tilde{h}(s,j) | \beta_n^r(s,j) | = 0$. Thus, $\tilde{h}(s^*, j^*) = - \frac{\sum_{(s,j) \neq (s^*, j^*)} h^*(s,j) | \beta_n^r(s,j) |}{| \beta_n^r(s^*, j^*) |}$. Then $P_n^r S_{\tilde{h}}(Q_n^r) = 0$. We now want to choose (s^*, j^*) such that $P_n^r (S_{h^*}(Q_n^r) - S_{\tilde{h}}(Q_n^r))$ minimal, so that subsequently setting it smaller than $o(n^{-1/2})$ yields the global undersmoothing criterion.

Note that

$$P_n^r (S_{\tilde{h}}(Q_n^r) - S_{h^*}(Q_n^r)) = P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r) \left(\sum_{(s,j)} (\tilde{h} - h^*)(s,j) \beta_n(s,j) \phi_{s,j} \right).$$

We have $\sum_{(s,j)} (\tilde{h} - h^*)(s,j) \beta_n^r(s,j) \phi_{s,j} = c_n(s^*, j^*) \phi_{s^*, j^*}$ with

$$c_n(s^*, j^*) = -\beta_n^r(s^*, j^*) \left\{ \frac{\sum_{(s,j) \neq (s^*, j^*)} h^*(s,j) | \beta_n^r(s,j) |}{| \beta_n^r(s^*, j^*) |} + h^*(s^*, j^*) \right\}.$$

Note that $c_n(s^*, j^*)$ is bounded by $\sum_{(s,j)} | h^*(s,j) | | \beta_n^r(s,j) |$ which is thus bounded by $\| h^* \|_\infty C_n$ (using $\sum_{(s,j)} | \beta_n^r(s,j) | = C_n$). Thus, under this trivial assumption we have $c_n(s^*, j^*) = O_P(1)$. So we have obtained:

$$\begin{aligned} P_n^r (S_{\tilde{h}}(Q_n^r) - S_{h^*}(Q_n^r)) &= c_n(s^*, j^*) P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r) (\phi_{s^*, j^*}) \\ &= O_P \left(C_n P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r) (\phi_{s^*, j^*}) \right). \end{aligned}$$

Therefore, our undersmoothing condition can be chosen to be $C_n \min_{(s,j) \in \mathcal{J}_n(C_n)} |P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s^*,j^*})| = o(n^{-1/2})$, which then implies $P_n^r D_n^r(Q_n^r, G_{0,n}^r) = o(n^{-1/2})$. This completes the proof of the first part of the next theorem, while the remaining part provides the extra condition (13) that makes $P_n^r(D_n^r - D^r)(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$, so that we also obtain $P_n^r D^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$ as well.

Theorem F.1.

Definitions:

Consider HAL-MLE ensemble $Q_n^r = \arg \min_{Q^r \in \mathcal{Q}^r(C_n), f \ll^* \mu_n} P_n^r L^r(Q^r)$ for some selector C_n with $C_{n,cv} \leq C_n \leq C^u < \infty$ with probability tending to 1, where $C_{n,cv}$ is the cross-validation selector of C . Recall that $L^r(Q^r)(v, o^r) = L(Q^r \circ \mathbf{Q}_{n,v})(o)$; $P_n^r L^r(Q^r) = \frac{1}{V} \sum_{v=1}^V P_{n,v}^1 L(Q^r \circ \mathbf{Q}_{n,v})$; $Q_n^r = \sum_{(s,j) \in \mathcal{J}_n(C_n)} \beta_n(s, j) \phi_{s,j}$, where $\mathcal{J}_n(C_n)$ provide the indices of the basis functions with non-zero coefficients and β_n denotes the corresponding coefficients. Here we emphasize that $\mathcal{J}_n(C_n)$ is implied by the L_1 -norm bound C_n in definition of Q_n^r . We have that Q_n^r solves the score equations $P_n^r S_h(Q_n^r) = 0$ for all bounded h with $r(h, Q_n^r) = 0$, and $S_h(Q_n^r)$ defined by (10).

Given a realization of \mathbf{Q}_n (i.e., treating it as fixed), consider the target parameter $\Psi^r : \mathcal{Q}^r \rightarrow \mathbb{R}$ defined by $\Psi^r(Q^r) = \frac{1}{V} \sum_{v=1}^V \Psi(Q^r \circ \mathbf{Q}_{n,v})$. Assume its canonical gradient $D^r(Q^r, G^r)(v, o^r) = D^*(Q^r \circ \mathbf{Q}_{n,v}, G_v^r)(o)$ at $P^r \in \mathcal{M}^r$, and exact second order remainder $R_2^r(Q^r, G^r, Q_{0,n}^r, G_{0,n}^r) = \Psi^r(Q^r) - \Psi^r(Q_{0,n}^r) + P_0 D^r(Q^r, G^r)$ given by $\frac{1}{V} \sum_v R_{20}(Q^r \circ \mathbf{Q}_{n,v}, G_v^r, Q_{0,n}^r \circ \mathbf{Q}_{n,v}, G_{0,n}^r)$. Let $D_n^r(Q_n^r, G_{0,n}^r)(v, o^r) = D_n^*(Q^r \circ \mathbf{Q}_{n,v}, G_{0,n,v}^r)(o)$ be an approximation of $D^r(Q_n^r, G_{0,n}^r)$ that is contained in $\mathcal{S}(Q_n^r) = \{S_h(Q_n^r) : h\}$, without the restriction $r(h, Q_n^r) = 0$. Let $h^*(Q_n^r, G_{0,n}^r)$ be the corresponding index so that $D_n^r(Q_n^r, G_{0,n}^r) = S_{h^*(Q_n^r, G_{0,n}^r)}(Q_n^r)$.

Assumptions: Assume $\|h^*(Q_n^r, G_{0,n}^r)\|_\infty = O_P(1)$, and the global undersmoothing criterion

$$C_n \min_{(s,j) \in \mathcal{J}_n(C_n)} \|P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j})\| = o(n^{-1/2}). \quad (11)$$

Conclusion: Then,

$$P_n^r D_n^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2}).$$

We can also replace (11) by

$$\min_{(s,j) \in \mathcal{J}_n(C_n)} \|P_0^r \left\{ \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j}) - \frac{d}{dQ_{0,n}^r} L^r(Q_{0,n}^r)(\phi_{s,j}) \right\}\| = o_P(n^{-1/2}), \quad (12)$$

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

If also

$$P_0^r \{D^r(Q_n^r, G_{0,n}^r) - D_n^r(Q_n^r, G_{0,n}^r)\} = o_P(n^{-1/2}), \quad (13)$$

then we have

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

Remark regarding setting cut-off for undersmoothing condition Our proof shows that $P_n^r D_n^r(Q_n^r, G_{0,n}^r) \approx \max_{(s,j) \in \mathcal{J}_n(C_n)} |h^*(s, j)| C_n \min_{(s,j) \in \mathcal{J}_n(C_n)} \|P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j})\|$, where h^* is so that $D_n^r(Q_n^r, G_{0,n}^r) = S_{h^*}(Q_n^r)$. A sensible bound for $P_n^r D_n^r(Q_n^r, G_{0,n}^r)$ is $\sigma_n/(n^{1/2} \log n)$. Thus, we would want to select $C_n > C_{n,cv}$ so that

$$\max_{(s,j) \in \mathcal{J}_n(C_n)} |h^*(s, j)| C_n \min_{(s,j) \in \mathcal{J}_n(C_n)} \|P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j})\| \approx \sigma_n/(n^{1/2} \log n).$$

If one knows the canonical gradient $D^r(Q_n^r, G_n^r)$ for a given estimator G_n^r , then one can determine the corresponding $h^*(Q_n^r, G_n^r)$ so that $D^r(Q_n^r, G_n^r) = S_{h^*(Q_n^r, G_n^r)}(Q_n^r)$, and use the max-norm of $h^*(Q_n^r, G_n^r)$. Therefore, a recommended concrete criterion is given by

$$\max_{(s,j) \in \mathcal{J}_n(C_n)} |h^*(s, j)| C_n \min_{(s,j) \in \mathcal{J}_n(C_n)} \|P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j})\| \approx \sigma_n/(n^{1/2} \log n).$$

However, in this case we are aiming to make the undersmoothing criterion tailored for the particular target parameter Ψ . Of course, one might as well select C_n so that $P_n^r D_n^r(Q_n^r, G_n^r) \approx \sigma_n/(n^{1/2} \log n)$, since one

already used G_n^r and even aimed to estimate the max-norm of h^* . However, we also see that if we set $\frac{\sigma_n}{\|h^*\|_\infty}$ to some constant K (e.g, $K = 1$), then we obtain a global undersmoothing criterion

$$\min_{(s,j) \in \mathcal{J}_n(C_n)} \|P_n^r \frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j})\| \approx KC_n^{-1}/(n^{1/2} \log n).$$

In many censored or causal inference problem, both this max-norm of h^* , which is aligned with the sup-norm of $D^r(Q_n^r, G_{0,n}^r)$, and the standard error σ_n of $D^r(Q_n^r, G_{0,n}^r)$ are driven by a positivity assumption and increase as the support for the target parameter decreases. This suggest that it might be quite reasonable to assume that across a large class of target parameters K is uniformly bounded away from 0, so that the above global undersmoothing condition will work well across a large family of target parameters.

Proof Theorem F.1: We already showed above that $P_n^r D_n^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$ by the undersmoothing condition. Now we note that

$$\begin{aligned} P_n^r D^r(Q_n^r, G_{0,n}^r) &= P_n^r \{D^r(Q_n^r, G_{0,n}^r) - D_n^r(Q_n^r, G_{0,n}^r)\} + o_P(n^{-1/2}) \\ &= P_0^r \{D^r(Q_n^r, G_{0,n}^r) - D_n^r(Q_n^r, G_{0,n}^r)\} + o_P(n^{-1/2}), \end{aligned}$$

if, for each v , conditional on the training sample (and thus, fixed $\mathbf{Q}_{n,v}$)

$$\{D^r(Q^r, G_{0,n}^r), D_n^r(Q^r, G_{0,n}^r) : Q^r \in \mathcal{Q}^r(C^u)\} \text{ is a } P_0^r\text{-Donsker class,}$$

and

$$P_0^r \{D^r(Q_n^r, G_{0,n}^r) - D_n^r(Q_n^r, G_{0,n}^r)\}^2 \rightarrow_p 0.$$

The Donsker assumption holds since, by (5) (and remark above showing it also applies to D_n^r), it consists of d^r -variate cadlag functions with universal bound on sectional variation norm. The consistency condition is implied by (13). \square

F.1 Understanding assumption (13).

Here we discuss the key condition (13), beyond the undersmoothing condition (11). For this purpose, we reparametrize the paths $Q_{n,\epsilon}^{r,h}$ as follows:

$$Q_{n,\epsilon}^{r,l(h,Q_n^r)}(x) = Q_n^r(x) + \epsilon l(h, Q_n^r)(x),$$

where

$$l(h, Q_n^r)(x) = h(0)Q_n^r(0) + \sum_{s \in \{1, \dots, J\}_{(0_s, x_s]}} \int h(s, u_s) dQ_{n,s}^r(u_s).$$

Therefore, we could also define the class of paths $\{Q_{n,\epsilon}^{r,h} : \|h\|_\infty < \infty\}$ as $\{Q_{n,\epsilon}^{r,l} : l \in \mathcal{F}(Q_n^r)\}$, where the index set is given by $\mathcal{F}(Q_n^r) = \{l(h, Q_n^r) : \|h\|_\infty < \infty\}$. The set $\mathcal{F}(Q_n^r)$ is restricted since it consists of the linear span of $\{\phi_{s,j} : (s,j) \in \mathcal{J}_n(C_n)\}$, that is, the linear span of all basis functions $\phi_{s,u_{s,j}}$ with non-zero coefficient $\beta_n(s, u_{s,j})$ in the fit $Q_n^r = \sum_{(s,j)} \beta_n(s, j) \phi_{s,j}$. The scores $S_h(Q_n^r)$ are linear in $l(h, Q_n^r)$ and the set of scores $\{S_h(Q_n^r) : \|h\|_\infty < \infty\}$ can be parametrized accordingly as $\{S_l(Q_n^r) : l \in \mathcal{F}(Q_n^r)\}$. We will typically have that the canonical gradient $D^r(Q_n^r, G_{0,n}^r) = \frac{d}{d\epsilon} L(Q_{n,\epsilon}^{r,l_0,n}) \Big|_{\epsilon=0}$ for a choice $l_0,n = l_0(Q_n^r, G_{0,n}^r)$, generally not an element of $\mathcal{F}(Q_n^r)$. Let $\mathcal{F}^+(Q_n^r)$ be this richer set so that $l_0,n \in \mathcal{F}^+(Q_n^r)$ and $\{S_l(Q_n^r) : l \in \mathcal{F}^+(Q_n^r)\}$ is an augmented set of scores satisfying $P_0^r S_l(Q_{0,n}^r) = 0$ for all $l \in \mathcal{F}^+(Q_{0,n}^r)$. So let's make this assumption. Then, we can write $D^r(Q_n^r, G_{0,n}^r) = S_{l_0,n}(Q_n^r)$. We can define $D_n^r(Q_n^r, G_{0,n}^r)$ as the projection of $D^r(Q_n^r, G_{0,n}^r)$ onto the finite dimensional linear span $\{S_l(Q_n^r) : l \in \mathcal{F}(Q_n^r)\}$: this actually equals the linear span of $\frac{d}{dQ_n^r} L^r(Q_n^r)(\phi_{s,j})$ across (s,j) with $\beta_n^r(s, j) \neq 0$. Thus, $D_n^r(Q_n^r, G_{0,n}^r) = S_{l_n}(Q_n^r)$ for a $l_n \in \mathcal{F}(Q_n^r)$. Somewhat conservatively, we could define l_n as the projection of l_0,n onto the finite dimensional space $\{\sum_{(s,j) \in \mathcal{J}_n(C_n)} \alpha(s, j) \phi_{s,j} : \alpha\}$ spanned by the basis functions $\phi_{s,j}$ with a non-zero coefficient $\beta_n(s, j)$ in Q_n^r . Let $\|l_0,n - l_n\|_0$ be the chosen Hilbert space norm so that $l_n = \arg \min_{l \in \mathcal{F}(Q_n^r)} \|l_0,n - l\|_0$. Since

the set of basis functions $\{\phi_{s,j} : (s,j) \in \mathcal{J}(C_n)\}$ is rich enough (even when we select $C_n = C_{n,cv}$) to approximate $Q_{0,n}^r$ w.r.t. $d_0^{1/2}(f, Q_{0,n}^r)$ at a rate $n^{-1/3}(\log n)^{d^r/2}$, one generally expects that $\|l_n - l_{0,n}\|_0$ will also be $O_P(n^{-1/3}(\log n)^{d^r/2})$. However, as argued in main section, if (e.g.) due to $G_{0,n}^r$ being more complex, $l_{0,n}$ is spanned by basis functions that are not needed for $Q_{0,n}^r$, then it might require $C_n > C_{n,cv}$ to obtain this rate of convergence. Finally, we note that

$$\begin{aligned} P_0^r \{D_n^r(Q_n^r, G_{0,n}^r) - D^r(Q_n^r, G_{0,n}^r)\} &= P_0^r S_{l_n - l_{0,n}}(Q_n^r) \\ &= P_0^r \{S_{l_n - l_{0,n}}(Q_n^r) - S_{l_n - l_{0,n}}(Q_{0,n}^r)\}, \end{aligned}$$

since $P_0^r S_l(Q_{0,n}^r) = 0$ for all $l \in \mathcal{F}^+(Q_{0,n}^r)$. This now proves that the left-hand difference is indeed a second order term that can typically be bounded by $d_0(Q_n^r, Q_{0,n}^r)^{1/2} \|l_n - l_{0,n}\|_0$, so that it will be $O_P(n^{-1/3}(\log n)^{d^r/2}) \|l_n - l_{0,n}\|_0$. Therefore, a sufficient assumption for (13) is that $\|l_n - l_{0,n}\|_0 = O_P(n^{-1/6-\delta})$ for some $\delta > 0$.

G Analysis of the Targeted HAL super-learner

G.1 Rate of convergence of T-M-HAL-SL

Analogue to [16], we obtain the same rate of convergence for $d_0(Q_n^{r,*}, Q_{0,n}^r)$ as for $d_0(Q_n^r, Q_{0,n}^r)$. Firstly, we can copy the proof of Lemma 4.2 by defining $\mathcal{Q}_n^r \subset \mathcal{Q}^r$ as the subset of functions $Q^r \in \mathcal{Q}^r$ for which $\|P_n^r D^r(Q^r, G_n^r)\| < r_n$, and $Q_{0,n}^{r,*}$ as the corresponding oracle ensemble. This then proves $d_0(Q_n^{r,*}, Q_{0,n}^{r,*}) = O_P(n^{-2/3}(\log n)^{d^r})$. It then remains to show that $d_0(Q_{0,n}^{r,*}, Q_{0,n}^r) = O_P(n^{-2/3}(\log n)^{d^r})$. We then construct a local least favorable submodel through $Q_{0,n}^r$ and define a corresponding TMLE update which maps $Q_{0,n}^r$ into a targeted version $\tilde{Q}_{0,n}^r$ that is an element of \mathcal{Q}_n^r . However, in this case this LFM is centered at the true $Q_{0,n}^r$ so that the MLE $\epsilon_n = O_P(n^{-1/2})$, thereby showing that $d_0(Q_{0,n}^r, \tilde{Q}_{0,n}^r) = O_P(n^{-1})$. This then shows that $d_0(Q_{0,n}^{r,*}, Q_{0,n}^r) \leq d_0(\tilde{Q}_{0,n}^r, Q_{0,n}^r) = O_P(n^{-1})$. Therefore, one can conclude that $Q_{0,n}^r$ and $Q_{0,n}^{r,*}$ only differ by a negligible amount so that we indeed have $d_0(Q_n^r, Q_{0,n}^r) = O_P(n^{-2/3}(\log n)^{d^r})$. This results in the following analogue of Lemma 4.1 for this T-HAL-MLE ensemble selector $Q_n^{r,*}$.

Lemma G.1.

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

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

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

Assumptions: Assume (5); regularity conditions on the least favorable submodel $\mathcal{Q}^{r,LFM}(Q_{0,n}^r)$, so that the MLE $\epsilon_n = O_P(n^{-1/2})$, thereby $d_0(Q_{0,n,\epsilon_n,G_n^r}^r, Q_{0,n}^r) = O_P(n^{-1})$, and $\|P_n^r D^r(Q_{0,n,\epsilon_n,G_n^r}^r, G_n^r)\| \leq r_n$ with probability tending to 1.

Conclusion: We have

$$d_0^{\bar{V}}(Q_n^{r,*}, Q_{0,n}^r) = d_0^r(Q_n^{r,*}, Q_{0,n}^r) = O_P(n^{-2/3}(\log n)^{d^r}).$$

Since $d_0^{\bar{V}}(Q_{0,n}, Q_0)$ is not affected by the targeting, Theorem 4.2 applies also to $Q_n^{r,*}$.

H Treatment Specific Mean Example: Solve Score Equations by Undersmoothing

Since, Q_n^r is an MLE it solves a large class of score equations $P_n^r S_h(Q_n^r) = 0$ defined by $S_h(Q_n^r) = \frac{d}{d\epsilon} L^r \left(\sum_{s,j} (1 + \epsilon h(s,j)) \beta_n^r(s,j) \phi_{s,j} \right) \Big|_{\epsilon=0}$ and h any bounded function satisfying that $r(h, Q_n^r) = 0$. The constraint is defined as $r(h, Q_n^r) \equiv \sum_{s,j} h(s,j) |\beta_n^r(s,j)| = 0$.

Note that $S_h(Q_n^r)(O^r) = A \sum_{s,j} h(s,j) \beta_n^r(s,j) \phi_{s,j}(W^r)(Y - Q_n^r(W^r))$. Thus, this class of scores $\{S_h(Q_n^r) : r(h, Q_n^r) = 0\}$ across all bounded h with $r(h, Q_n^r) = 0$ equals the dimension of the number of non-zero coefficients $\beta_n^r(s,j)$ in its representation $Q_n^r = Q_{\beta_n^r} = \sum_{s,j} \beta_n^r(s,j) \phi_{s,j}$, minus 1 due to the constraint $r(h, Q_n^r) = 0$. Therefore, as the L_1 -norm $C_n = \|\beta_n^r\|_1$ increases, the number of non-zero coefficients grows so that this linear span of these score equations grows accordingly. This is the intuitive argument why choosing C_n large enough should give us $P_n^r D^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2})$:

$$\frac{1}{n} \sum_{i=1}^n \frac{A_i}{G_{0,n}^r(W_i^r)} (Y_i - Q_n^r(W_i^r)) = o_P(n^{-1/2}).$$

Formally, we apply Theorem F.1. Let $\mathcal{S} = \{S_h(Q_n^r) : h\}$ the class of scores not enforcing the constraint $r(h, Q_n^r) = 0$. We first need to define an approximation $D_n^r(Q_n^r, G_{0,n}^r) \in \mathcal{S}$. By selecting C_n large enough we will have that we can find an h that makes $A \sum_{s,j} h(s,j) \beta_n^r(s,j) \phi_{s,j}(W^r)$ approximate $A/G_{0,n}^r(W^r)$. Let $h_n^* = h^*(Q_n^r, G_{0,n}^r) \equiv \arg \min_h P_0^r(\sum_{s,j} \beta_n^r(s,j) h(s,j) \phi_{s,j}(W^r) - 1/G_{0,n}^r(W^r))^2$ be defined as this $L^2(P_0^r)$ -projection of the desired $1/G_{0,n}^r$ onto this linear span, where the argmin includes any bounded h (not restricting to $r(h, Q_n^r) = 0$). We can then define the approximation $D_n^r(Q_n^r, G_{0,n}^r) \equiv A \sum_{s,j} h_n^*(s,j) \beta_n^r(s,j) (Y - Q_n^r(W^r)) \in \mathcal{S}$ of $D^r(Q_n^r, G_{0,n}^r)$.

Since $E(A | W) > \delta > 0$, by iterative conditional expectation, it follows that $G_{0,n}^r = E(A | W^r) > \delta > 0$ for some $\delta > 0$. Then, it follows that the sup-norm of h_n^* is $O_P(1)$, which verified the first condition of Theorem F.1. Consider now the global undersmoothing criterion (11) of Theorem F.1, and note that it is given by: select C_n large enough so that the fit Q_n^r includes a sparsely supported basis function with $\min_{s,j} |P_n \phi_{s,j}|$ small enough in the sense that

$$\min_{(s,j), \beta_n^r(s,j) \neq 0} \left| \frac{1}{n} \sum_{i=1}^n \phi_{s,j}(W_i^r) (Y_i - Q_n^r(W_i^r)) \right| = o_P(n^{-1/2}). \quad (14)$$

Application of Theorem F.1 proves now that

$$P_n^r D_n^r(Q_n^r, G_{0,n}^r) = o_P(n^{-1/2}).$$

We assume that $G_{0,n}^r$ is cadlag and has a uniformly bounded sectional variation norm. Since Q_n^r is by definition cadlag with finite sectional variation, this shows that $\mathcal{D}^r = \{D^r(Q^r, G_{0,n}^r) : Q^r \in \mathcal{Q}^r\}$ is a class of d^r -dimensional real valued cadlag functions with a uniformly bounded sectional variation norm. This verifies the Donsker class conditions in (5). Condition (13) of Theorem F.1 states

$$E_{P_0^r} A \left\{ Q_{\beta_n^r h_n^*}^r - 1/G_{0,n}^r \right\} (Y - Q_n^r(W^r)) = o_P(n^{-1/2}).$$

So this requires our approximation $Q_{\beta_n^r h_n^*}^r$ of $1/G_{0,n}^r$ to converge fast enough. By Cauchy-Schwarz inequality, the left-hand side can be bounded by $\|Q_n^r - Q_{0,n}^r\|_{P_0^r} \|Q_{\beta_n^r h_n^*}^r - 1/G_{0,n}^r\|_{P_0^r}$. Given our rate $\|Q_n^r - Q_{0,n}^r\|_{P_0} = n^{-1/3}(\log n)^{d^r/2}$ it follows that it suffices that

$$\inf_{\beta} \left\| \sum_{s,j, \beta_n^r(s,j) \neq 0} \beta(s,j) \phi_{s,j} - \frac{1}{G_{0,n}^r} \right\|_{P_0^r} = O_P(n^{-1/6-\delta}), \quad (15)$$

for some $\delta > 0$. We will assume this to hold. Since we know that the left-hand side with $1/G_{0,n}^r$ replaced by $Q_{0,n}^r$ would be $O_P(n^{-1/3}(\log n)^{d^r/2})$, even without undersmoothing (i.e., setting $C_n = C_{n,cv}$) this

might already hold. On the other hand, if the true $Q_{0,n}^r$ is a relatively simple function spanned by a subset of all possible spline basis functions, while approximating $1/G_{0,n}^r$ requires these basis functions, then undersmoothing will be needed. This verifies all conditions of Theorem F.1 and thus proves the following result.

Lemma H.1. Assume (15); and undersmoothing condition (14).

Then, $P_n^r D^r(Q_n^r, G_{0,n}^r) = r_1(n)$ with $r_1(n) = o_P(n^{-1/2})$.

I Treatment Specific Mean Example: Difference Between Targets

The following lemma establishes that $\Psi^r(Q_{0,n}^r) - \Psi(Q_0)$ behaves as $d_0(Q_n, Q_0)$ and is thus second order.

Lemma I.1. Definitions: Recall $Q_{0,n,v}(W) = Q_{0,n}^r \circ \mathbf{Q}_{n,v}(W)$. Define

$$\begin{aligned}\tilde{G}_{0,n,v}^*(x, y) &\equiv E_0(A \mid Q_{0,n,v}(W) = x, Q_0(W) = y) \\ \tilde{G}_{0,n,v}(x) &\equiv E_0(A \mid Q_{0,n,v}(W) = x).\end{aligned}$$

Due to $G_0 > \delta > 0$, these two functions are also bounded away from this δ . These two functions (where chosen to) satisfy $P_0 D^*(Q_{0,n,v}, G_0) = P_0 D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}^*(Q_{0,n,v}, Q_0))$, and $P_0 D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}(Q_{0,n,v})) = 0$.

Assumptions: Assume

- \hat{Q}_1 is an HAL-MLE so that, by Theorem 4.2 we have $d_0(Q_{0,n}, Q_0) = O_P(n^{-2/3}(\log n)^d)$, and, thus also, for each v , $d_0(Q_{n,v}, Q_0) = O_P(n^{-2/3}(\log n)^d)$.
- $\limsup_n \sup_{x,y} \left| \frac{d}{dy} \tilde{G}_{0,n,v}^*(x, y) \right| < \infty$, where the supremum over $(x, y) \in \mathbb{R}^2$ is over a support of $(Q_{0,n}(v, W), Q_0(W))$.

Then, we have that $|\Psi(Q_{0,n,v}) - \Psi(Q_0)|$ is bounded by $\delta^{-2} \|(\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v})(Q_{0,n,v}, Q_0)\|_{P_0} \|Q_{0,n,v} - Q_0\|_{P_0}$. In addition, we have, by Lemma I.2 below that $\|(\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v})(Q_{0,n,v}, Q_0)\|_{P_0} = O(\|Q_{0,n,v} - Q_0\|_{P_0})$. This proves

$$\left\{ \frac{1}{V} \sum_v \Psi(Q_{0,n}^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0) \right\} = O(d_0(Q_{0,n}, Q_0)).$$

Thus, by first bullet point assumption, we have that this is $O_P(n^{-2/3}(\log n)^d)$.

The differentiability condition on $G_{0,n,v}^*$ is not a bad assumption since it only concerns the dependence of the conditional expectation of A , given $(Q_{0,n,v}(W), Q_0(W))$, on the fixed random variable $Q_0(W)$.

In this example, $d_0(Q_{n,v}, Q_0) = O_P(n^{-2/3}(\log n)^d)$ can also be shown directly, instead of as an application of Lemma 4.1. We have $Q_{0,n,v}(W) = E_0(Y \mid \mathbf{Q}_{n,v}(W) = \mathbf{Q}_{n,v}(W))$. Therefore, $Q_{0,n,v} - Q_0$ requires analyzing $E_0(Y \mid \mathbf{Q}_{n,v}(W)) - Q_0(W)$. We have that $E_0(Y \mid \mathbf{Q}_{n,v}(W), A = 1)$ is the projection of $Q_0(W) = E_0(Y \mid W, A = 1)$ onto the set of functions of $\mathbf{Q}_{n,v}(W)$ (in the Hilbert space $L^2(P_{0|A=1})$). One such candidate function for the projection is given by $\mathbf{Q}_{n,v,j=1}(W)$, showing that the $L^2(P_{0|A=1})$ -norm of $Q_0(W) - E_0(Y \mid \mathbf{Q}_{n,v}(W), A = 1)$ is smaller than the $L^2(P_0)$ -norm of $Q_0 - \mathbf{Q}_{n,v,1}$, but the latter is $O_P(n^{-1/3}(\log n)^{d/2})$, by assumption.

Proof of Lemma I.1: We have

$$\Psi(Q_{0,n,v}) - \Psi(Q_0) = -P_0 D^*(Q_{0,n,v}, G_0), \quad (16)$$

since the second order remainder $R_{20}(Q, G_0, Q_0, G_0) = 0$ (due to its double robust structure). Note now that $P_0 D^*(Q_{0,n,v}, G_0) = P_0 D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}^*(Q_{0,n,v}, Q_0))$. This follows since

$$\begin{aligned}E_0 A / G_0 (Y - Q_{0,n,v}(W)) &= E_0 A / G_0 (Q_0 - Q_{0,n,v})(W) \\ &= E_0 (Q_0 - Q_{0,n,v})(W) = E_0 A / \tilde{G}_{0,n,v}^*(Q_{0,n,v}(W), Q_0(W)) (Q_0 - Q_{0,n,v})(W).\end{aligned}$$

So we can replace the right-hand side in (16) by $-P_0 D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}^*(Q_{0,n,v}, Q_0))$:

$$\Psi(Q_{0,n,v}) - \Psi(Q_0) = -P_0 D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}^*(Q_{0,n,v}, Q_0)). \quad (17)$$

We now note that $P_0 D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}(Q_{0,n,v})) = 0$: due to $E_0(Y \mid A = 1, Q_{0,n,v}(W)) = Q_{0,n,v}(W)$ it follows that

$$E_0 \frac{A}{E_0(A \mid Q_{0,n,v}(W))} (Y - Q_{0,n,v}(W)) = 0.$$

Thus, we have

$$\begin{aligned} \Psi(Q_{0,n,v}) - \Psi(Q_0) &= P_0 \{D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}(Q_{0,n,v})) - D^*(Q_{0,n,v}, \tilde{G}_{0,n,v}^*(Q_{0,n,v}, Q_0))\} \\ &= P_0 \frac{\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v}}{\tilde{G}_{0,n,v}^* \tilde{G}_{0,n,v}} (Q_{0,n,v}, Q_0) (Q_0 - Q_{0,n,v}). \end{aligned}$$

By Cauchy-Schwarz inequality, we can bound the latter term by $\|(\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v})(Q_{0,n,v}, Q_0)\|_{P_0} \|Q_{0,n,v} - Q_0\|_{P_0}$. This shows that it remains to show $\|(\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v})(Q_{0,n,v}, Q_0)\|_{P_0} = O_P(n^{-1/6-\delta})$ for some $\delta > 0$. For that we apply Lemma I.2. This lemma shows that we can bound $E_0(A \mid Q_{0,n,v}(W), Q_0(W)) - E_0(A \mid Q_{0,n,v}(W))$ by the L^2 -norm of $Q_{0,n,v} - Q_0$. This completes the proof. \square

Lemma I.2. For notational convenience, let $X_n(w) = Q_{0,n,v}(w)$, $X(w) = Q_0(w)$, $\tilde{G}_{0,n,v}^*(X_n(w), X(w)) = E_0(A \mid X_n(W) = X_n(w), X(W) = X(w))$, and $\tilde{G}_{0,n,v}(X_n(w)) = E_0(A \mid X_n(W) = X_n(w))$. Let $X_n = X_n(W)$ and $X = X(W)$, and let $L^2(P_{X_n, X})$ be the corresponding Hilbert space of functions of (X_n, X) with covariance inner product. Assume that $\limsup_n \sup_{x,y} \|\frac{d}{dy} \tilde{G}_{0,n,v}^*(x, y)\| < \infty$, where the supremum over (x, y) is over a support of $(X_n(W), X(W))$. We note that $\tilde{G}_{0,n,v}$ is the projection of $\tilde{G}_{0,n,v}^*$ onto the subspace $L^2(P_{X_n})$ of functions of X_n only in the Hilbert space $L^2(P_{X_n, X})$.

We have

$$\|\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v}\|_{P_{X_n, X}} = O(\|X_n - X\|_{P_0}).$$

Proof of Lemma I.2: We have that $\tilde{G}_{0,n,v}$ is the projection of $\tilde{G}_{0,n,v}^*$ (a function of X_n, X) on the subspace $L^2(P_{X_n})$ of all functions that only depend on X_n , a subspace of $L^2(P_{X_n, X})$, endowed with the usual covariance as inner product. Thus,

$$\begin{aligned} \|\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v}\|_{P_{X_n, X}}^2 &= \inf_{f \in L^2(X_n)} \|\tilde{G}_{0,n,v}^* - f\|_0^2 \\ &= \inf_{f \in L^2(X_n)} \int \{\tilde{G}_{0,n,v}^*(X_n(w), X(w)) - f(X_n(w))\}^2 dP_0(w) \\ &= \inf_{f \in L^2(X_n)} \int \{\tilde{G}_{0,n,v}^*(X_n(w), X(w)) - \tilde{G}_{0,n,v}^*(X_n(w), X_n(w)) \\ &\quad + \tilde{G}_{0,n,v}^*(X_n(w), X_n(w)) - f(X_n(w))\}^2 dP_0(w) \\ &= \int \{\tilde{G}_{0,n,v}^*(X_n(w), X(w)) - \tilde{G}_{0,n,v}^*(X_n(w), X_n(w))\}^2 dP_0(w) \\ &\quad + \inf_{f \in L^2(X_n)} \int \{\tilde{G}_{0,n,v}^*(X_n(w), X_n(w)) - f(X_n(w))\}^2 dP_0(w) \\ &\quad + 2 \int (\tilde{G}_{0,n,v}^*(X_n, X) - \tilde{G}_{0,n,v}^*(X_n, X_n))(\tilde{G}_{0,n,v}^*(X_n, X_n) - f(X_n)) dP_{X_n}. \end{aligned}$$

The latter infimum over all functions f of $X_n(w)$ is attained at $f = \tilde{G}_{0,n,v}^*(X_n(w), X_n(w))$, so we obtain

$$\|\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v}\|_{P_{X_n, X}}^2 = \int \{\tilde{G}_{0,n,v}^*(X_n(w), X(w)) - \tilde{G}_{0,n,v}^*(X_n(w), X_n(w))\}^2 dP_0(w).$$

By the assumed differentiability of $\tilde{G}_{0,n,v}^*$ in its second coordinate we have

$$\begin{aligned} \tilde{G}_{0,n,v}^*(X_n(w), X(w)) &= \tilde{G}_{0,n,v}^*(X_n(w), X_n(w)) \\ &\quad + \left. \frac{d}{dy} \tilde{G}_{0,n,v}^*(X_n(w), y) \right|_{y=\xi(X_n(w), X(w))} (X(w) - X_n(w)), \end{aligned}$$

for an intermediate point $\xi(X_n(w), X(w))$ in between $X_n(w)$ and $X(w)$. By the assumed uniform bound on the derivative we have $|\tilde{G}_{0,n,v}^*(X_n(w), X(w)) - \tilde{G}_{0,n,v}^*(X_n(w), X_n(w))| < C |X_n(w) - X(w)|$ for some $C < \infty$, so that we have

$$\|\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v}\|_{P_{X_n, X}}^2 \leq C \int (X_n(w) - X(w))^2 dP_0(w) = C \|Q_{0,n,v} - Q_0\|_0^2.$$

So this proves that $\|\tilde{G}_{0,n,v}^* - \tilde{G}_{0,n,v}\|_{P_0} = O_P(n^{-1/3}(\log n)^{d/2})$. This proves the lemma. \square

J Relation between undersmoothing criterion (12) and bound C_n on sectional variation norm

Consider assumption (12). In our treatment specific mean example, this states

$$\min_{s,j,\beta_n^r(s,j) \neq 0} P_0^r \phi_{s,j}(Q_n^r - Q_{0,n}^r) = o_P(n^{-1/2}). \quad (18)$$

The next theorem denotes the left-hand side with R_n and shows that $R_n = o_P(n^{-1/2})$ is generally expected to hold for a selector C_n under which $d_0^r(Q_n^r, Q_{0,n}^r) = O_P(n^{-2/3}(\log n)^{d^r})$.

Theorem J.1.

Definitions: For a given s , let $j^* = j_s^* = \arg \min_j P_0^r \phi_{s,j}$, and let u_{s,j^*} be the corresponding knot point. Let $\bar{P}_0^r(s) \equiv \min_j P_0^r \phi_{s,j^*}$ be the probability that $O_s^r \geq u_{s,j^*}$ under P_0^r . Let $P_{0,s}^r$ represent the probability distribution $O_s^r = (O^r(j) : j \in s)$. For a cadlag function Q , define $\tilde{Q}_s(x_s) = \int_{(0,x_s]} dQ_s(u)$, so that $Q = \sum_s \tilde{Q}_s$. Thus, $Q_n^r = \sum_s \tilde{Q}_{n,s}^r$ and $Q_{0,n}^r = \sum_s \tilde{Q}_{0,n,s}^r$. Let $R_n(s) \equiv |P_0^r \phi_{s,j^*}(Q_n^r - Q_{0,n}^r)|$ and $R_n \equiv \min_{s,j,\beta_n^r(s,j) \neq 0} |P_0^r \phi_{s,j}(Q_n^r - Q_{0,n}^r)|$. Let $r(n) = n^{-1/3}(\log n)^{d^r/2} \approx n^{-1/3}$.

Assumptions:

- $d_0(Q_n^r, Q_{0,n}^r) = O_P(r(n))$.
- The loss-based dissimilarity is equivalent with a square of the $L^2(P_0^r)$ -norm: $d_0^r(Q_n^r, Q_{0,n}^r) \sim \|Q_n^r - Q_{0,n}^r\|_{P_0^r}^2$.
- For at least one subset s , we have that $\|\tilde{Q}_{n,s}^r - \tilde{Q}_{0,n,s}^r\|_{P_{0,s}^r} = O_P(r(n))$; that there exists a $\delta > 0$ so that

$$\frac{\int_{u \geq u_{s,j^*}} |Q_{n,s}^r(u) - Q_{0,n,s}^r(u)| dP_{0,s}^r(u)}{\{\bar{P}_0^r(s)\}^{(|s|+1)/|s|}} > \delta > 0 \quad (19)$$

with probability tending to 1; and, for some $0 \leq \alpha \leq 1/2$,

$$\begin{aligned} \|\phi_{s,j^*}(Q_{n,s}^r - Q_{0,n,s}^r)\|_{1,P_0^r} &\leq (\bar{P}_0^r(s))^\alpha \|Q_{n,s}^r - Q_{0,n,s}^r\|_{P_0} \\ \|\phi_{s,j^*}(Q_n^r - Q_{0,n}^r)\|_{1,P_0^r} &\leq (\bar{P}_0^r(s))^\alpha \|Q_n^r - Q_{0,n}^r\|_{P_0^r}. \end{aligned}$$

Note, we always can select $\alpha = 1/2$ (conservatively), and, if $\|Q_n^r - Q_{0,n}^r\|_\infty = O_P(r(n))$, then we can set $\alpha = 0$. Let \mathcal{S}_1 be the collection of subsets s for which these two conditions hold.

Conclusion: We have for each subset $s \in \mathcal{S}_1$:

$$\begin{aligned} \bar{P}_0^r(s) &= O_P\left((r(n))^{\frac{|s|}{\alpha|s|+1}}\right) \\ &\approx O_P\left(n^{-\frac{|s|}{3+3\alpha|s|}}\right). \end{aligned}$$

We have

$$R_n(s) = O_P\left(r(n)^{\frac{1+|s|}{1+\alpha|s|}}\right).$$

This implies

$$R_n = O_P\left(\min_{s \in \mathcal{S}_1} r(n)^{\frac{1+|s|}{1+\alpha|s|}}\right).$$

If \mathcal{S}_1 includes a set s with $|s| \geq 3$, then, even for $\alpha = 1/2$, we have $R_n = o_P(n^{-1/2})$.

How does the bound on R_n improve if we would have supnorm convergence: Suppose that $\|Q_n^r - Q_{0,n}^r\|_\infty = O_P(r(n))$ as well. Then we can select $\alpha = 1$, so that for we obtain $R_n(s) = O_P(r(n)^{1+|s|}) \approx n^{-2/3}$ (even for $|s| = 1$).

Bounding assumption: In this theorem we assumed that for some s $\|\tilde{Q}_{n,s}^r - \tilde{Q}_{0,n,s}^r\|_{P_0^r}$ can be bounded by $\|Q_n^r - Q_{0,n}^r\|_{P_0^r}$. P_0^r describes a random variable on a cube $[0, \tau^r] \subset \mathbb{R}^{d^r}$. In our example, this would be the

distribution of W^r . In many applications, one might artificially truncate the covariate space from below and above so that its values are in a cube $[0, \tau^r]$. In that case, the s -specific edges $E_s = [0_s, \tau_s^r] \times \{0_{-s}\}$ of $[0, \tau^r]$ would have positive mass under P_0^r . Then, $\|Q_n^r - Q_{0,n}^r\|_{P_0^r} = \sum_s \int_{E_s} (Q_n^r - Q_{0,n}^r)^2(u_s, 0_{-s}) dP_0^r(u_s, 0_{-s})$. So, in that case, $\|Q_n^r - Q_{0,n}^r\|_{P_0^r} = O_P(r(n))$ would imply that the $L^2(P_0^r)$ -norm of the difference of the s -specific sections, $Q_{n,s}^r - Q_{0,n,s}^r$, converge at same rate. This would naturally imply the same rate for the s -specific generalized differences $\tilde{Q}_{n,s}^r - \tilde{Q}_{0,n,s}^r$ of the sections, and thereby verify this bounding condition. We suspect that this bounding assumption will apply to continuous P_0^r as well (i.e., no mass on the edges E_s). Our reasoning is based on $Q_n^r - Q_{0,n}^r = \sum_s (\tilde{Q}_{n,s}^r - \tilde{Q}_{0,n,s}^r)$, and that for $s \neq s_1$, the two sets of basis functions in Q_{0,n,s_1}^r and Q_{0,n,s_2}^r , respectively, are largely independent. That is, $Q_{0,n}^r$ represents an additive model (e.g., GAM), where each component $Q_{0,n,s}^r$ is identifiable from the total sum function (by our definition of \tilde{Q}_s^r). It is true that, for example, basis functions $I(X_1 > c_1)I(X_2 > c_2)$ across knot points (c_1, c_2) (i.e., $s_2 = \{1, 2\}$) can approximate $I(X_1 > c_1)$ (i.e., $s_1 = \{1\}$) by letting $c_2 \approx 0$, but the L^1 -norm (i.e., contribution to the variation norm of $\tilde{Q}_{0,n,s}^r$) of this small vector of coefficients represent a negligible proportion of the full L^1 -norm (i.e., full variation norm of $\tilde{Q}_{0,n,s}^r$) of the coefficients making up $Q_{0,n,s}^r$.

Condition (19): Consider one of the subsets $s \in \mathcal{S}_1$. Note that the numerator in (19) is the $L^1(P_0^r)$ -norm $\|\phi_{s,j^*}(Q_{n,s}^r - Q_{0,n,s}^r)\|_{1,P_0^r}$. This follows since $\phi_{s,j^*}(u) = I(u \geq u_{s,j^*})$, and since $Q_{n,s}^r - Q_{0,n,s}^r$ is only a function of $W^r(s)$, the expectation w.r.t. P_0^r becomes an expectation w.r.t. its marginal $P_{0,s}^r$. One expects that $\phi_{s,j^*}(u) = 1$ for all $u \geq u_{s,j^*}$ for most of the basis functions with $\beta_n^r(s, j) \neq 0$. So only a few basis functions will have some variation over $u > u_{s,j^*}$. For example, if s is a singleton, then, for all $u \geq u_{s,j^*}$, we have $Q_{n,s}^r(u) = Q_{n,s}^r(u_{s,j^*})$ is constant in u , due to all basis functions in $Q_{n,s}^r$ being 1 at such a u (i.e., there are no basis functions $\phi_{s,j}$ in $Q_{n,s}^r$ with knot points larger than u_{s,j^*}). So in that case $(Q_{n,s}^r - Q_{0,n,s}^r)(u) = (Q_{n,s}^r(u_{s,j^*}) - Q_{0,n,s}^r(u))$ for all $u \geq u_{s,j^*}$. Over a cube A in an $|s|$ -dimensional space, the variation in each of the $|s|$ coordinates is $A^{1/|s|}$, assuming that the sides of the cube are proportional to each other. So, $\max_{k \in s} (\tau_s(k) - u_{s,j^*}(k))$ behaves as $(\bar{P}_0^r(s))^{1/|s|}$. Thus, the integral of $(Q_{0,n,s}^r(u_{s,j^*}) - Q_{0,n,s}^r(u))$ behaves as $\bar{P}_0^r(s)^{1+1/|s|} = (\bar{P}_0^r(s))^{|s|+1/|s|}$.

Proof of Theorem J.1: Consider one of the sets $s \in \mathcal{S}_1$. By assumption we have $\|\phi_{s,j^*}(Q_{n,s}^r - Q_{0,n,s}^r)\|_{1,P_0^r} \geq \{\bar{P}_0^r(s)\}^{\frac{|s|+1}{|s|}}$. Now, use that $\|\phi_{s,j^*}(Q_{n,s}^r - Q_{0,n,s}^r)\|_{P_0^r} \leq \bar{P}_0^r(s)^{1-\alpha} \|Q_{n,s}^r - Q_{0,n,s}^r\|_{P_0^r}$. This gives then

$$\|Q_{n,s}^r - Q_{0,n,s}^r\|_{P_0^r} \geq \bar{P}_0^r(s)^{\frac{\alpha|s|+1}{|s|}}$$

Since the left-hand side is $O_P(r(n))$, this then shows that

$$\bar{P}_0^r(s) = O_P\left(r(n)^{\frac{|s|}{\alpha|s|+1}}\right).$$

Consider now the term $R_n(s)$ and note we can bound this by $\bar{P}_0^r(s)^{1-\alpha} \|Q_n^r - Q_{0,n}^r\|_{P_0^r}$. Combining the bound on $\bar{P}_0^r(s)$ above and $\|Q_n^r - Q_{0,n}^r\|_{P_0^r} = O_P(r(n))$, gives then

$$R_n(s) = O_P\left(r(n)^{-\frac{1+|s|}{1+\alpha|s|}}\right).$$

This implies the bound for R_n by minimizing the latter over all $s \in \mathcal{S}_1$. \square

References

- [1] 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.
- [2] M.J. van der Laan. A generally efficient targeted minimum loss estimator based on the highly adaptive lasso. *International Journal of Biostatistics*, 2017.
- [3] 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.
- [4] A. Bibaut and M.J. van der Laan. Fast rates for empirical risk minimization over cadlag functions with bounded sectional variation norm. Technical report, Division of Biostatistics, University of California, Berkeley, 2019.

- [5] 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.
- [6] X. Shen. On methods of sieves and penalization. *Annals of Statistics*, 25(6):2555–2591, 1997.
- [7] X. Shen. Large sample sieve estimation of semiparametric models. *Chapter in Handbook of Econometrics*, 76(00):0000, 2007.
- [8] W. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 2014.
- [9] 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.
- [10] M.J. van der Laan, D. Benkeser, and W. Cai. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Technical report, Division of Biostatistics, University of California, Berkeley, 2019.
- [11] 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, Division of Biostatistics, University of California, Berkeley, November 2003.
- [12] 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.
- [13] M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.
- [14] M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.
- [15] 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.
- [16] M.J. van der Laan and W. Cai. Targeted highly adaptive minimum loss estimation. Technical Report ?, Division of Biostatistics, University of California, Berkeley, 2020.
- [17] M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.
- [18] M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, page <http://www.bepress.com/ijb/vol4/iss1/17/>, 2008.
- [19] M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin Heidelberg New York, 2011.
- [20] M.J. van der Laan and S. Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer, Berlin Heidelberg New York, 2018.
- [21] Mark van der Laan and Susan Gruber. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The international journal of biostatistics*, 12(1):351–378, 2016.
- [22] Mark van der Laan, Zeyi Wang, and Lars van der Laan. Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*, 2021.
- [23] S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*, 6(1), 2010. PMCID: PMC3126668.
- [24] Gilmer Valdes, Yannet Interian, Efsthios Gennatas, and Mark Van der Laan. The conditional super learner. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10236–10243, 2021.
- [25] Ana Barragán-Montero, Adrien Bibal, Margerie Huet Dastarac, Camille Dragnet, Gilmer Valdes, Dan Nguyen, Siri Willems, Liesbeth Vandewinckele, Mats Holmström, Fredrik Löfman, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. *Physics in Medicine & Biology*, 67(11):11TR01, 2022.
- [26] M.J. van der Laan. Targeted maximum likelihood based causal inference: Part I. *Int J Biostat*, 6(2):Article 2, 2010.
- [27] M.J. van der Laan. Targeted maximum likelihood based causal inference: Part II. *Int J Biostat*, 6(2):Article 3, 2010.
- [28] Maya Petersen, Joshua Schwab, Susan Gruber, Nello Blaser, Michael Schomaker, and Mark van der Laan. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference*, 2(2):147–185, 2014.
- [29] Wenjing Zheng and Mark van der Laan. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of causal inference*, 5(2):20160006, 2017.
- [30] Zeyi Wang, Lars van der Laan, Maya Petersen, Thomas Gerds, Kajsa Kvist, and Mark van der Laan. Targeted maximum likelihood based estimation for longitudinal mediation analysis. *arXiv preprint arXiv:2304.04904*, 2023.
- [31] Lars van der Laan, Marco Carone, Alex Luedtke, and Mark van der Laan. Adaptive debiased machine learning using data-driven model selection techniques. *arXiv preprint arXiv:2307.12544*, 2023.
- [32] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, pages 1–15, 2022.
- [33] Ivana Malenica, Aurelien Bibaut, and Mark J van der Laan. Adaptive sequential design for a single time-series. *arXiv preprint arXiv:2102.00102*, 2021.

- [34] A.W. van der Vaart and J.A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. ISSN: 1935-7524, DOI: 10.1214/11-EJS605.
- [35] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, Berlin Heidelberg New York, 1996.
- [36] Aad Van Der Vaart and Jon A Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.