TMLE, Superlearner, HAL and the roadmap towards a new Philosophy of Statistics

Dialogues concerning the historical and philosophical roots of the crisis in Statistics and Data Science

Mark J. van der Laan, Mathematician, Statistician and Biostatistician
University of California, Berkeley
laan@berkeley.edu

Richard J.C.M. Starmans, Philosopher and Statistician
Utrecht University, Tilburg University, the Netherlands

March 18, 2024

1 First dialogue: LASSO and beyond . . . Highly Adaptive LASSO

About regularization in traditional regression and why it neither helps to reconcile statistical learning and machine learning nor helps to overcome the current crisis in data analytical practice

Richard: Now, we are about to dive into the historical and philosophical roots of the crisis in statistics and data science. Before we start lamenting about flawed methodologies in machine learning and indulge in a litany of complaints about the Statistics Wars and persistent problems in data-analytic practice, let us start with a specific, more tangible example. I think that regularization techniques in statistical learning and machine learning would be a suitable starting point. LASSO (L_1) (Tibshirani, 1996), Ridge regression (L_2) and Elastic Net (which combines L_1 and L_2) are now well-established and frequently applied regularization techniques, aimed at improving on traditional ordinary least squares regression (OLS) and extensions (Hastie et al., 2001).

Mark: Yes, in many empirical disciplines and application domains the Generalized Linear Model (GLM) has become more or less exemplary. Especially in such fields as the social and behavioral sciences, but also in economics or epidemiology. They don't call regression analysis the "workhorse of statistics" for nothing.

Richard: As a result, the literature on the estimation of the regression function is large and there is a strong willingness to further develop statistics along these lines. That is understandable, and regularization methods like Ridge regression and LASSO fit nicely in this. They are not exactly a paradigm shift, I mean they neither do require a "Gestaltswitch" nor imply a need for a "Umkehrung aller Werte". They allow empirical scientists to continue working within the broad intellectual framework that they are familiar with, because they were raised within this framework, trained within this framework, think within this framework. They can continue using the same trusted and solid methodology with the same lingo, only adapted to changes in the environment. In a manner of speaking regularization techniques provide these necessary adjustments, allowing the prevailing paradigm of GLM to enter the era of Big Data and data science. This is exactly what the Least Absolute Shrinkage and Selection Operator (LASSO) promises. First, due to shrinkage of coefficients there will most likely be less overfitting, or -using traditional lingo- the external validity will be improved. Secondly, it offers a selection operator, that achieves variable selection in a better way than traditional methods such as subset regression, forward, backward or stepwise regression, etcetera. Already in the nineties these methods were used as tools for model building -to use the traditional lingo-, but at the same time these methods were deemed inappropriate or at least suspect by many social scientists; they present themselves as an open invitation to data dredging, chasing low p-values, torturing the data until they confess. So, in summary, rather than causing a paradigm shift, regularization techniques like LASSO are more an extension of the product lifecycle of the existing GLM-paradigm.

Mark: Well, I have at least two things to say. First, I would definitely comment on the way you introduce models in general and model building in particular, but I will postpone this for now. Secondly, this idea of paradigms and paradigm-shift needs some elucidation as well; the concepts are highly ambiguous and overused, but because I know that this is one of your hobbyhorses as well, let us postpone this for now and stick to the real topic. I agree that LASSO builds on the popular intellectual framework that GLM offers, and no doubt that is part of its success. Another thing is that all these regularization methods were made freely available to the scientific community and everybody who showed interest rather fast, through Python Libraries, R packages, with code made available at Github, et cetera.

Richard: I know that, and you could say that this is one of the blessings of open source and open science. But it does not impress me too much. In fact, I think that the full acceptance and adoption of these regularization methods in the sciences and in everyday data analytical practice actually went rather slowly. It came more or less in fits and starts, just like any new technology or any innovation aimed at achieving progress in science.

Mark: What do you mean by that?

Richard: Let me give you one example. In 2022 IBM released version 29 of IBM-SPSS. SPSS is arguably one of the eldest and most applied commercially available statistical software packages, raised in 1968 and bought by IBM in 2009. In its latest release the company proudly announced some brand-new features. In the Regression module, the menu was enhanced with three new tabs, for Ridge regression, the LASSO and Elastic net respectively. Now let us restrict ourselves to LASSO. The original paper was published by Tibshirani in 1996 and the technique was introduced already some years before, so it took nearly three decades before SPSS, the market leader in statistical software adopted it. Now the paper aims to improve on Ridge regression, which is an even older technique. So, regarding this commercial "acceptance" you could say that all this is not before time at all.

Mark: I agree that commercial adoption of a technique by such companies as IBM-SPSS is a token of recognition, but don't you exaggerate the significance of this a little?

Richard: Well, let me tell you this. I am fully aware that open source and open science are not the main interests of any commercial company. What's more, SPSS has never been famous for adopting new technologies very quickly; look for instance at the neural net module, that in no way keeps track with latest developments in deep learning. And that is no less than an understatement, a very polite way to put it. However, large segments in the social and behavioral sciences, the humanities, but especially application areas outside academia are now involved in statistics, quantitative data-analysis: in policy research or business analysis, law, finance, marketing, public health, education et cetera. In fact, all kinds of governmental and non-governmental organizations have entered the era of data science and big data. They often come from different intellectual communities with different research traditions, and they quite often rely on and are trained and educated with commercially available statistical packages, data mining suites, etc. SPSS is more or less part of the methodology, or at least the research environment and practice.

Mark: Indeed, and they have made great efforts over the last decades to make the package part of this environment and practice. Worldwide, hundreds of thousands of researchers and students, affiliated with universities and other educational or research institutions have got nearly free access to the software. I can get your point. So, you say that given the fact that the market leader knows how many people rely on the software, especially since it implements the basics of the GLM paradigm so well, and given the fact that LASSO extends this so smoothly, it is remarkable that they adopt the technique about three decades after its invention.

Richard: Exactly, and I know that the history of technology shows many divergent criteria that determine success or failure regarding the adoption of a technology; they could be rational or less rational, you may think of commercial interests, legal considerations or even ideological or political issues. Maybe not everybody was convinced about the necessity to implement these techniques, but let us leave this for now. Let me emphasize that compliance with the existing methodology, sticking to traditional scientific, customs, habits and habits is especially imperative in areas with formalized, protocolized methodologies, areas with high stakes (clinical research, public health, medicine, pharmaceutics) or in other disciplines with fixed well-established rules for doing research publication, such as in economics. Apart from the natural omnipresent reluctance to innovation in these fields adoption of technology can be even more complicated.

Mark: Are you actually saying that domains with formalized methodologies such as clinical or pharmaceutics are unwilling or reluctant to adopt new technologies or -more generally put- to innovate when it comes to research methods. Because as a biostatistician this certainly is not my experience.

Richard: No, I only say that open-source availability of code, et cetera for the research community is only a first step, getting people acquainted, it doesn't change anything. Changing methodology and research habits is always tricky, especially in highly regulated domains. You will have to convince different sub-disciplines within academia to change the methodologies, they have been using for decades, subsequently you have to convince professionals (based on or derived for these sub-disciplines) and you have to convince commercial companies offering statistical software, data mining suites, process mining tools, etc. They have to choose between hundreds if not thousands of methods, algorithms, proposed by the scientific community in the last decades, all promising to change the world, or at least make a difference. Sometimes these methods bring the inventor eternal fame, due to Turing awards or even Noble prices, but they remain laying on the shelf because despite their alleged universal applicability, because they are not acknowledged by the relevant disciplines and do not solve real problems, but only toy problems.

Mark: There is something to say for a commercial company to only adapt methods in their new version if they have gone through some serious testing period in the open source and academic community. A collaborator at the FDA told me that they are mostly interested in statistical approaches that have gone through decades of academic scrutiny and development, and that does not even get into the issue of robust and scalable implementations of these methods. They view the fact that Targeted Learning has been around since 2006 as a positive thing, and that, in spite of much and natural resistance, its adoption has grown a lot. However, I have also seen that the FDA is open to R and is willing to accept SAP's in terms of methods implemented in the open source R language, not necessarily SAS or SPSS. Of course, they will require simulations demonstrating the validity of the code and the claimed statistical properties.

Richard: Yes, in fact the history of statistics shows that many techniques became successful, developed by people who were working in a particular field, working with real life problems, no toy examples. Often this happened when the world was faced with a crisis, for example a war or health crisis. I always say that in order to understand the history of data science you need to understand the history of epidemiology, that shows the emergence of many pathbreaking new insights, methods, or techniques during crises. But let's leave this for now.

Mark: Right, so with respect to these regularization techniques you could say, better late than never. I guess that what you are saying exceeds LASSO. How about bootstrap methods, Monte Carlo simulation or Bayesian methods?

Richard: The same story. For example, they have all been integrated in SPSS as well by now and not before time. For example, since a few years SPSS offers Bayesian alternative to classical parametric T-test, ANOVA, correlation test, etc. This is significant not only in view of the philosophical differences between frequentists and Bayesians, but especially if you want to come up with a constructive methodology for data science. That being said, I am now mainly interested in / focusing on data scientists, who sometimes have no background in mathematical statistics, or even programming and who need a constructive methodology they can rely on without being involved, without discussions about frequentists versus Bayesians, statistical learning versus machine learning, potential outcomes versus DAG's, etc. We will discuss the implications for the Philosophy of Statistics later.

Mark: But now, I would comment on another point you make, i.e. that we want robust and reliable methodology in the hands of the average practitioner and that goes far beyond some open-source code that has many options and often requires expertise in order to know how to properly apply it in a particular data application setting. This is where commercial companies can play an important role by tailoring robust implementations to very specific settings and providing a user-friendly interface obtaining the right input from the user.

Richard: Yes, we definitely should discuss this issue with respect to TMLE in the dialogue on Philosophical Roots of Statistics. But let's go back to the topic of Lasso. Let's discuss this idea of regularization a little further. They sometimes are called shrinking-techniques because what they actually do is shrinking the coefficients of the regression equation by adding a penalty term, as to avoid -or at least decrease- overfitting, thus improving the performance of the model on new, unseen data. Now, obviously that is what we usually want and as such these techniques are important from the perspective of achieving or improving a bias-variance trade-off. Both statisticians and people from machine learning agree that this bias-variance trade-off is of crucial importance in analytics, but unfortunately, they do not fully agree on what the concept exactly means, but let's postpone this issue a bit.

Mark: Yes, for our purposes it is important to understand that we deliberately introduce a little bit of bias in the model based on the training data, in order to have better prediction performance on test data representing the part of reality we are actually interested in. When we talk about bias and variance of an estimator such as Lasso regression, we have to first define the target. This target could be the prediction function in which case the LASSO regression represents an estimator $\hat{f}(X)$ of the conditional mean function $f(X) = E(Y|X)$. If prediction is the goal, then the bias variance trade-off is with respect to this optimal prediction function $f(X) = E(Y|X)$. The mean squared error could then be defined as the expected squared error $(\hat{f}(X) - f(X))^2$ averaged w.r.t. population distribution of X. So this is really an average over possible inputs x of the mean squared error of the estimator $\hat{f}(x)$ representing the average performance of the estimator as an estimator of the true curve at a randomly drawn X. We should probably refer to this as an average MSE.

But maybe one wants to be more specific and make a prediction for a particular input x so that the target is $E(Y|X = x)$. In that case, the MSE would be the expectation of squared error of the Lasso regression at this given x versus the true conditional mean of Y, given $X = x$. In either case, we can decompose MSE or average MSE of an estimator w.r.t. its target in terms of its (average) variance and (average) square of its bias w.r.t. target, so that minimizing MSE or average MSE comes down to balancing the variance of the estimator with the square of the bias of the estimator, either for a given x or averaged. By shrinking the coefficients Lasso induces some bias but reduces variance in such a way that the MSE is reduced relative to the MSE of the unpenalized least squares linear regression estimator. Anyway, I guess I went overboard here a little, so forgive me.

Richard: Yes, and no doubt ”less overfitting” is an important aspect of this better future prediction behavior, and in order to achieve this the penalty term should do the trick. Anyway, in L_1 and L_2 the optimal value of the penalty parameter lambda, -which of course is only a hyper-parameter, that does not refer to any characteristic of interest in a population and therefore needs no interpretation- can easily be found by cross-validation.

Mark: Yes, and as theoretically proven cross-validation will do a great job in optimizing the average MSE w.r.t. the conditional mean of Y, given X (van der Laan et al., 2007). So, it is safe to say that the LASSO will outperform the non-penalized least squares estimator for the sake of prediction.

Richard: What's more, LASSO, unlike Ridge regression is often credited for allowing feature selection in a swift way, because the shrinkage actually can go to zero, which means that the variables with zero coefficient can be ignored. This can all be achieved in a straightforward manner, more or less automatically. In case of high dimensional data or in case of multicollinearity this could be beneficial, at least according to some voices. But let's postpone this as well.

Mark: I can see that this is an attractive feature of Lasso regression. There might be other penalties that achieve the same, but the L_1-penalty indeed does that nicely. Either way, at this point, in case the user uses the LASSO fit to evaluate a treatment effect by its coefficient in the model fit, we have to start evaluating the LASSO w.r.t. such particular features of the true regression curve. In other words, as I mentioned above, we need to define the target first before we can evaluate the validity of a method and a coefficient represents a very specific feature of the regression curve.

Richard: Now, as a data scientist and applied statistician I welcome a new set of techniques one can rely on, to be added to the analyst's or practitioner's toolkit. As a philosopher, specialized in the history and foundations of statistics, I am not so very much pleased with a few aspects. I would like to share some thoughts with you on this matter and you would most certainly oblige me, by giving your comments. That is, if you have not been bored yet so far.

Mark: No, not at all, in my work I am always very concerned with the philosophy behind an approach, and that then also naturally sheds light on its scientific validity.

Richard: Let's restrict ourselves to the LASSO. Does it really add that much to regression, when it comes to the problems, we are facing in data analytic practice? Is it not true that it just inherits the benefits and drawbacks of linear regression? I mean, when you like OLS because it is so very much interpretable (whatever that may mean if you add higher order interaction terms), then of course LASSO-regression will do the trick as well, a clear benefit. However, if your traditional OLS is wrongly specified, then LASSO-regression will not help you very much. Is that not so?

Mark: You make an excellent point. Indeed, I view a lot of the literature on LASSO as a desperate effort to hang on the linear regression models in the context of high dimensional covariates, giving the user a sense that they are just ending up with a linear regression model so that they can employ their traditional training and practice. Of course, it is very attractive that they can even obtain such a linear regression model when having high dimensional covariates, even to the point that the number of covariates can exceed the number of observations. The problem is that these models are misspecified so it becomes unclear what these coefficients even mean, and this misspecification is not getting any better by having higher dimensional covariates. Many theoretical results in the LASSO literature concern statistical inference for a coefficient in the Lasso model fit, often referred to as post-model selection inference. They then typically assume that the Lasso is able to fit a correct model, so that they only have to correct for the data adaptivity of the selected model. They might even prove that if the true regression is contained in the starting model, then the Lasso is able to select the right non-zero coefficients with probability tending to 1. Their typical theoretical results assume sparsity. These assumptions are totally unrealistic.

Richard: What does sparsity mean here exactly, and why is this assumption unrealistic?

Mark: Their notion of sparsity means that among the many main terms in the user supplied regression model relatively few, like or the order of logarithm of sample size in some cases, have a true non-zero coefficient and then they prove that the LASSO is able to learn the true underlying parametric model as sample size grows: i.e. it is able to screen out the variables that have no effect on the outcome. Of course, the real world does not work that way, so that all these results have no link to reality. That is, true coefficients might be small but rarely will be exactly zero, and that is not taking into account yet that the model is misspecified making it even more unreasonable to assume that most coefficients are equal to zero. I often refer to these theorems as empty theorems since they make assumptions that are known to be wrong, and not just a little bit wrong. The problem in the practice of traditional statistics is precisely that one aims to interpret the coefficients in linear regression models and hanging on to them in the context of high dimensional data makes these interpretations even more biased and problematic. On the other hand, on the positive side, it would be perfectly appropriate to view LASSO as a rich set of candidate machine learning algorithms for estimating the conditional mean of the outcome as a function of the covariates, among other target functions depending on the chosen loss function. By varying the sets of covariate-terms in the linear model, including interactions and other possible transformations of the covariates, such as so called basis functions, one obtains a rich set of candidate regression algorithms.

Richard: Now these are also the days of ensemble learning and many data scientist distinguish between three categories: bagging, boosting and stacking. The latter is of particular importance, because it combines the best models generated by divergent analytical techniques. So you say that LASSO as such does not tackle the problems associated with wrongly specified parametric models, but from a practical point of view LASSO is a powerful extension of classical regression, aimed at prediction and interpretation according to the original paper. As such it has been added to the practitioner's toolkit.

Mark: Correct.

Richard: Considered as a rich set of candidate machine learning algorithms for estimating the conditional mean of the outcome as a function of the covariates, would you say that LASSO should be added to a stacking algorithm?

Mark: Yes, and in that case, it is viewed as a prediction algorithm, and we could include it as candidate algorithms in, for example, a super-learner that uses cross-validation to select the best performing algorithm, or the convex super-learner that was originally introduced as a stacking ensemble technique. In fact in our applications I love to add all kinds of Lasso's to the the collection of candidate machine learning algorithms in the Super-learner. So somehow, my experience is that one can get very far with LASSO-based algorithms. In fact, my work on highly adaptive lasso demonstrates that, but I am sure we will get to that in this discussion.

Richard: All right, so you introduced Highly adaptive LASSO or HAL some years ago (van der Laan, 2017). The fact that you deliberately choose a name derived from LASSO is significant. How does HAL extend the concept of LASSO, especially in view of my concerns and the drawbacks and limitations that you just mentioned? I mean, could you explain the theoretical, mathematical part of HAL, but also consider the practical implications.

Mark: Yes, it is ironic that a person (myself) that has always been critical of the lasso literature, with its sparsity assumptions and post-model selection, views the Highly Adaptive Lasso as one of his main contributions. Let me clarify how this happened and that will then also provide the explanation of this estimation approach I termed HAL. I asked myself the following question. Let's consider the problem of estimating the conditional mean of an outcome on covariates, precisely what linear least squares regression aims to address. Least squares linear regression corresponds with minimizing the empirical mean squared error (MSE) over all linear combinations of the d covariates $(X_1, .., X_d)$. Instead, I suggested let's minimize this MSE over all functions of these d covariates.

Richard: What is the difference "over all linear combinations of the d covariates" and "all functions of these d covariates"? Do you mean "linear functions" with "linear combinations"?

Mark: Yes, the class of functions linear regression considers are functions of form beta X, which are linear combinations of the covariates X_j, $j = 1, \ldots, d$. Our goal is to learn the true target function $E(Y|X)$: or, equivalently, given a d-dimensional covariate vector $X = x$, one wants to know the best possible prediction $E(Y|X = x)$ from a mean squared error perspective. Of course, the true target function will never be a linear combination of the main terms and one expects that these linear main term models are just too simplistic approximations of the true target function to be any good. When I say, all functions of these d covariates, I am looking literally at all functions that map X into a number/prediction, including the functions captured by these linear models, but also including linear combinations of highly complex functions of X. That is, just make up any function of X, and it will be included. What I really want is to define a class of functions that includes the true target function.

Richard: So you were proposing to literally minimize the empirical MSE over all functions. But that is crazy, it would totally overfit the data, and will result in terrible prediction function for future use, right?

Mark: Of course, that would not work since it would simply fit the data perfectly. So, we need a constraint on this class of functions. So, then the next question becomes, what kind of constraint shall we put on our class of functions, and preferably one that might hold for the true target function? I had experience through my Ph.D research with so called multivariate real valued cadlag functions with a finite sectional variation norm (van der Laan, 1996; Gill et al., 1995). These are functions of the d-covariates that are right-continuous, are allowed to have left-discontinuities, and the variation norm of the function restricted to a subset of one of more of the d covariates is finite, across all possible subsets of the d covariates.

Richard: Please explain this variation norm, is it bounded by some constant or what?

Mark: Yes, we would want these variation norms to be bounded by some constant, possibly a large one. And we can then use cross-validation to select the right bound on this variation norm.

Richard: Maybe it is helpful if you first explain variation norm for a univariate function.

Mark: Yes, we can get into a more detailed explanation of how this sectional variation norm of a multivariate function is defined, but for now just view it as a generalization of how we define the variation norm of a univariate function as the sum of the absolute value of all its changes up and down. To nail this concept for a univariate function, let's consider an example. Consider a univariate function on the unit interval that starts at $x = 0$ with value 1 and increases till value 5 at 0.5 (half-way) and then decreases till value 0 at $x = 1$. What would the variation norm of this function be?

Richard: I would add up the change of the function from $x = 0$ to $x = 1/2$, which equals 5-1, and the change of the function from $x = 1/2$ to $x = 1$, which equals 5. So the total absolute change of the function between 0 and 1 is 4+5=9. I would define the variation norm as 9. Did I get that right?

Mark: In fact, you got it exactly right with one little twist that is not very important. Note in my example the function started at value 1 at $x = 0$, which we view as a jump from 0 to 1 at the start and we would add that to the variation norm. So we end up with a variation norm of 10. In general, for a univariate function defined on $[0, 1]$, we first partition the interval $[0, 1]$ in many little intervals; we then compute the difference of the function over each interval, and sum the absolute value of all these differences; finally, we also add the absolute value of $f(0)$ itself. This would then be a numerical approximation of the variation norm of this univariate function. By letting the number of intervals grow, i.e. the width of the intervals start to approximate 0, we will actually obtain the precise variation norm of the function, even for functions that can be highly erratic in certain areas. But this definition of the variation norm as an integral of its infinitesimal differences is exactly the same as you would get by just tracking the ups and down of the function. However, this integral definition becomes important when we generalize this variation norm for univariate functions to multivariate functions. Before we go there, let's stick to univariate functions for now. In other words, let's act as if we care about a regression of a continuous outcome on a single continuous covariate X. Let's try to understand if it is reasonable to assume that this true target function $E(Y|X)$ has bounded variation norm. Can you think of a univariate function that would have infinite variation norm?

Richard: You tell me.

Mark: For example, $\sin(1/x)$ on $[0,1]$ would be such a function. As x approximates zero, this function just keeps going up and down from -1 to 1 forever.

Richard: Okay, in the real world I don't expect outcomes to be such a weird function of a univariate covariate X. For example, if anything relations between key variables measured on a subject, such as blood-pressure and age, are often known to be monotone, and at most I would think a typical univariate relation might be unimodal.

Mark: Exactly my point. For a univariate function the sectional variation norm equals the variation norm. Because of the same reason, I find it also very reasonable to assume that the true target function of a multivariate X has finite sectional variation norm. I think of the sectional variation norm of a function as a measure of its complexity. The more the function goes up and down the more complex it is. As you can imagine a function with very large variation norm is much harder to fit with finite data than a function that is almost constant. Therefore, I suggested that we use a bound on the sectional variation norm of the function as a constraint.

Richard: Okay, but nowhere in this definition for univariate function you talked about so called sections. Can you explain somewhat this notion of a sectional variation norm of a multivariate function?

Mark: Just for the sake of not making things too abstract, if you don't mind, I like to also explain how one computes this sectional variation norm for a function of two variables (X_1, X_2). Then we have an accurate picture of what a bound on the sectional variation norm means. Consider such a function on the unit square $[0,1]^2$. Firstly, consider the function only as a function of X_1 while setting the second coordinate X_2 equal to 0. This is just a univariate function and we compute its variation norm as we defined above. This univariate function is called a section of the function. We can do the same for the other section that views f as a function of X_2 and sets X_1 equal to zero. Note you can also view these univariate sections as simply restrictions of the function to the horizontal and vertical edges of the unit square that start at the origin. Finally, we need to compute the variation norm of the bivariate function on the unit square excluding these zero-edges. The variation norm of a bivariate function is computed by first partitioning the unit square in many small squares, left-open and right closed. We then compute the difference of difference of the function for this little square. Let's say the square has x_1 in $(a_1, b_1]$ and x_2 in $(a_2, b_2]$. If we set $x_2 = a_2$, then we can take the change of f as x_1 goes from a_1 to b_1, giving $f(b_1, a_2) - f(a_1, a_2)$. Similarly, we have the change of f at $x_2 = b_2$, giving $f(b_1, b_2) - f(a_1, b_2)$. So the difference of these two changes of f due to small change in x_1 at the neighboring values of x_1 is given by $f(b_1, b_2) - f(a_1, b_2) - f(b_1, a_2) + f(a_1, a_2)$. This is the second order change of f over this little square. We then take the absolute value of this, and we do this for each of the squares. The numerical approximation of the variation norm of the bivariate function f is then the sum over all squares of the absolute value of this second order change of f. The real variation norm is then obtained by letting the partitioning get finer and finer. This second order change of f over this little square is also called a generalized difference of f defined as this sum of positive and negative signs of the function at the corners of the square. This generalized difference of the function over this little square represents how one would calculate the probability that a bivariate random variable fall in the little square in terms of the bivariate cumulative distribution function of this bivariate random variable. Anyway, I think you can sense that this variation norm of the bivariate function quantifies second order interactions of this function, while the variation norm of its univariate sections quantifies the first order changes of this function. Finally, the sectional variation norm of this bivariate function is now defined as the sum of the variation norms of the two univariate sections and the bivariate function, and we also add the absolute value of f at the origin $(0, 0)$. So now I have told you the precise definition of the sectional variation norm of a bivariate function. I am very sorry for this little math journey, and hope you are not bored.

Richard: You should not and I'm not.

Mark: Nice, because I just want to give you a real sense that all the little variations this sectional variation norm is adding up is actually picking up first order and second order variations of the function. For example, if the bivariate variation norm of the function equals zero, then this implies that the function is an additive function $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$, while if the bivariate variation norm is substantial then such an additive model would be missing a lot of interactions between x_1 and x_2.

Richard: Fascinating, but why don't you go on and explain HAL, where we accept this sectional variation norm of a function as an interesting measure of the complexity of the function.

Mark: With this control on the complexity of the function, I then defined a nonparametric least squares estimator as the minimizer of the empirical MSE over all cadlag functions with a bound M on their sectional variation norm. Interestingly, this is an enormous model for the true underlying regression function, so assuming that the true regression function falls in this set for a large enough chosen M is not much of an assumption at all and can be expected to hold in most applications. In other words, we don't expect functions as $\sin(1/x)$ to describe the real world. Note that this model for the target function does not assume that the true function is very smooth or so. In fact, it allows that the true function has jumps in same way as a surface represented by stairs, where each step in the stair represents a sudden change in the function. Believe it or not but this is the most nonparametric HAL estimator for least squares regression, and the same principled definition can be applied to other definitions of empirical risk beyond MSE. For example, we can compute such HAL-estimators of the density of the data giving a nonparametric maximum likelihood estimator: in that case, we maximize the empirical log-likelihood over all densities, possibly parametrized through a link function, with a bound on the sectional variation norm. That is, we also have HAL estimators of densities, conditional densities, conditional means, conditional treatment effects and really essentially any functional parameter of the data distribution you might be interested in. HAL is just a minimum empirical risk minimizer and different definitions of empirical risk give HAL estimators of different target functions. Minimum empirical risk minimization over a function class is nothing new in the machine learning literature. What makes HAL unique is the particular class of functions we define, especially by using this sectional variation norm as a key constraint. Okay so at this point you wonder, how is this in anyway related to LASSO?

Richard: Yes, for me the math is both profound and straightforward, I mean without these theoretical underpinnings you cannot actually do science, and what's more you cannot build a constructive methodology. But for now, you would oblige me by explaining me how this estimator relates to the LASSO.

Mark: It happens to be the case that any cadlag function with finite sectional variation norm can be written as an infinite linear combination of zero order splines, i.e. indicator functions of the form $I(X > u) = \prod_{k=1}^{d} I(X_k > u_k)$ indexed by a multivariate knot-point $u = (u_1, .., u_d)$, and that the L_1-norm of the coefficients in this representation is precisely the sectional variation norm of the function. So from a pure practical algorithmic point of view this estimator just requires running the LASSO on a very large number of zero order spline basis functions $I(X > u_j)$ across many knot-points u_j in the d-dimensional unit cube $[0,1]^d$. To give you some sense that this result is not that surprising, let's consider the case that the dimension $d = 1$. Note that I can write $f(x) = f(0) + \int_{(0,x]} df(u)$. That is, I can write $f(x)$ as its value at 0 plus the sum of all its little increments from 0 to x. However, we can write this as $f(x) = f(0) + \int_{(0,1]} I(x \geq u) df(u)$. That is, we can write f as an intercept plus a linear combination of basis functions $x \rightarrow I(x \geq u_j)$ across a grid of knot-points u_j, and moreover, the coefficient in front of these indicator basis functions is the actual increment $df(u_j) = f(u_{j+1}) - f(u_j)$ of the function. So the L_1-norm of the vector of coefficients is precisely the variation norm $\int_{(0,1]} |df(u)|$ of the function! So what I said above is that we have a generalization of this result showing that $f(x)$ can be expressed as $f(0)$ plus a sum over all its sections of all the generalized increments of the section from 0 to the corresponding sub-vector of x. Just to annoy you, for the bivariate case this representation is given by

$$f(x_1, x_2) = f(0,0) + \int I_{u \leq x_1} df_1(u) + \int I_{v \leq x_2} df_2(v) + \int I_{u \leq x_1} I_{v \leq x_2} df(u,v).$$

Richard: So, this is why you kept the reference to LASSO in HAL? In the sense that you can construct the estimator using standard software available for LASSO?

Mark: Indeed, the practical implementation of HAL is just a matter of running LASSO with a very large set of binary covariates $I(X_i \geq u_j)$ across many knot-points u_j. For example, for the case that $d = 2$, a bivariate function is a linear model in univariate indicators $I(X_1 \geq u_{1j})$ and $I(X_2 \geq u_{2k})$ and bivariate indicators $I(X_1 \geq u_{1j}, X_2 \geq u_{2k})$ across a large set of knot-points u_{1j} and u_{2k}. The variation norm of the first section is just the L_1 norm of the coefficients in front of $I(X_1 \geq u_{1j})$ and the variation norm of the second section is the L_1-norm of the coefficients in front of $I(X_2 \geq u_{2k})$ and finally the bivariate variation norm is the L_1-norm of the coefficients in front of $I(X_1 \geq u_{1j}, X_2 \geq u_{2k})$. So the sectional variation norm is the L_1-norm of all these coefficients combined and we also include the intercept representing $f(0,0)$. Moreover, we can let cross-validation figure out the right bound on the sectional variation norm. As a statistician, one then wonders if this non-parametric least square estimator is any good. For example, is it able to learn the true target function as sample size grows to infinity? Moreover, how fast does it approximate the true target function as a function of sample size? The answer is that HAL approximates the true target function at a rate $n^{-1/3}$ up till a $\log n$-factor $(\log n)^{d/2}$ (Bibaut and van der Laan, 2019). Notice that this rate does essentially not depend on the dimension d of the covariate vector $X = (X_1, .., X_d)$.

Richard: Now this is significant. All too often statisticians claim their estimators to have "desirable statistical properties", without any further clarification or even reference to practical implications for data scientists or empirical researchers working in a real-life setting. But establishing a fast minimal rate of convergence is extremely important in machine learning as well. The fact that it actually does not depend on the number of independent variables will also be music to the ears of any data scientist working with big data. How did you prove this all? Did you do simulations, use publicly available datasets of use empirical process theory or what?

Mark: We did all of these. But yes, the theoretical proof uses empirical process theory in the sense that it utilizes bounds on the maximal difference between an empirical mean of a function of the observed random variable and the true mean of this function, where the maximum is taken over all cadlag functions with a universal bound on the sectional variation norm. Such bounds rely on the so-called covering number or entropy of the class of functions. So, I could basically just refer to such existing very fundamental results and carry out a relatively standard proof that can be written out in one page. To appreciate this result, one has to understand that at that point no other machine learning algorithm in the very rich literature of nonparametric estimation and ML had been shown to converge that fast without making extreme smoothness assumptions and the degree of smoothness they have to assume heavily depends on the dimension d. The theoretical results we have been able to show about HAL go far beyond its remarkable dimension free rate of convergence, and are from a statistical point of view even more important. Establishing a rate of convergence is nice, but it does not teach us anything yet about how to construct confidence intervals. For that we want to understand the sampling distribution of the estimator. We can discuss that little later.

Richard: That's fine with me, but first these smoothness conditions. What are these all about, are we talking about local or global smoothing conditions and how is the variation norm condition different from what others in the ML literature typically assume?

Mark: That is an excellent question. A local smoothness condition on a function is a statement about how smooth it is locally around a given x. A typical definition of smoothness is the so called Holder differentiability that extends k-th order continuous differentiability to non-integer values k. For example, one might assume that a function is continuously differentiable at x, or that it is 10 times differentiable at x. If we make such assumptions on the function at each x in its domain, then we would still state that we made local smoothness conditions about the function. When we assume that a function has finite variation norm, we still allow the function to have jumps and thus be heavily non smooth but we are restricting its overall amount of movement. I would call that a global smoothness constraint. Interestingly, popular classes of functions in the ML literature make a lot of local smoothness assumptions but up till smoothness up till degree d these were still not enough to exclude that the function has a universal bound on the sectional variation norm. Due to this they ended up with rates of convergence heavily affected by the dimension d of the covariate vector. Our class of functions does include a function whose first order mixed derivatives are all uniformly bounded: for example, for a bivariate function f, one assumes that $d/dx_1 f$, $d/dx_2 f$, $d/dx_1 d/dx_2 f$ are all bounded. If the typical literature based on smoothness classes (e.g., k-th order continuously differentiable) would have realized that bounding these mixed derivatives up till degree 1 is enough instead of assuming that all higher order derivatives need to be uniformly bounded as well ($d^2/dx_1 f, d^2/dx_2 f, d/dx_1 d/dx_2 f$), then they might have come up with the same result. But even that would not get to the essence, which is that we need to bound the overall variation of the function, and that smoothness is not needed at all. Somehow it is the global variation norm constraint that gives these dimension free rates of convergence. It is actually quite intuitive why the global variation is the right constraint to avoid overfitting. Suppose one perfectly fits the data. Then the fitted function jumps to all the observed values Y_i across all X_i, which means that this would end up having a variation norm of order sample size n. So, by restricting the variation norm of the function by a bounded constant, it is not allowed to overfit, and, in particular, the amount of variation of the fitted function is a nice measure of how data adaptive the fit is. In a sense it is hard to come up with a better constraint when it comes to controlling overfitting. By finding the best fit of the data under a given bound on the variation norm, one uses the allowed variation where it matters most. Let me now make a few remarks about the practical impact of HAL. HAL just spits out a finite dimensional linear regression model, allowing easy inspection of the fitted curve. Therefore, HAL can be viewed as an important contribution to interpretable machine learning.

Richard: Wait a minute, now for many data scientists interpretability is a kind of holy grail. Why and how is HAL more or better interpretable than LASSO?

Mark: Yes, thank you for that question since you raise an important issue. The goal of interpretable machine learning is to both approximate reality and be able to understand how the machine learning algorithms comes to its predictions and if the way it maps the input $(X_1, .., X_d)$ into a prediction of the outcome make sense. So yes, it is true that a main term linear model is easily inspected in the sense that one can check the variables it uses and if the effects of these variables are positive or negative and if there are interactions of interest and so on. However, if its predictions are poor, who cares about this? One only cares about the operating characteristics of a prediction algorithm if it is good. Indeed, people use as argument in favor of simple linear regression models that they are interpretable. But if these models are very wrong, the coefficients just represent projections of the true target function on this simplistic model making it almost impossible to interpret them as meaningful features of the true target function. So, you could say that it is fake to claim that such models are interpretable: any feature one would evaluate from this fitted function would be very biased for the actual feature of the target function. So, it would be brilliant if our fitted function is not only a linear model in a number of spline basis functions, but that it also a good or best possible approximation of reality. HAL is precisely that. Moreover, the HAL-fit is just a parametric model in indicators of the form $I(X_j \geq u)$, $j = 1, \ldots, d$, and its possible higher order interactions. Moreover, due to the terms in the HAL model being these indicators $I(X \geq u_j)$, the HAL fit is just the analogue of a "piecewise" constant function. That is, HAL partitions the covariate space in rectangular areas, or cubes really, and it assigns a value to each of them. Therefore, it is just as easy to interpret the HAL estimator as it is to interpret CART since they both represent a partitioning of the covariate space in regions defined by cubes or unions of cubes, or, equivalently, by logical statements such as X_1 smaller than u_1, X_2 larger than u_2 and X_3 larger than u_3, and so on. CART is a popular algorithm because of its easy interpretability but it has typically terrible practical performance relative to other algorithms. In fact, it gets never selected when one uses cross-validation to choose among a collection of machine learning algorithms, as in the discrete super-learner. On the other hand, HAL gives the same type of fit, but is actually good.

Richard: I have always liked CART because of its interpretability. So, if I understand you correctly HAL can give me a very similar interpretation, but it will actually be an algorithm competitive with state-of-the-art machine learning algorithms such as random forest, gradient boosting, Bayesian regression trees, among others.

Mark: Yes, I agree with that characterization.

Richard: It is also nice that HAL just gives an actual linear model we can report and easily inspect itself. However, it seems to me that I cannot read off easily from this what the effect is of a binary treatment, while I would be able to read that off immediately with a main term linear regression model as standard LASSO produces.

Mark: Yes, you are asking: What about learning effects of particular variables such as the effect of a binary treatment on the outcome? As you state, that is something a linear model with main terms easily spits out (although biased). Did we give that all up? For example, HAL might include indicators for treatment but also various interactions of that treatment with other covariates. In traditional statistics, practitioners typically avoid including interactions with the exposure/treatment of interest since they want a single coefficient that defines the treatment effect.

Richard: Wasn't that the beauty of the LASSO using main terms only?

Mark: My response is that we have to give up on thinking in terms of coefficients, but we have to start defining so called estimands as features of the regression curve. That is, we should ask the question: If I give you the true regression function, what feature of it would you like to calculate? These features could be defined so that they reduce to the coefficient in the linear regression model if it happens to be a correct linear model, but, much more importantly, it has a nonparametric interpretation as a target feature of interest of the true curve. One might think of it as a variable importance measure that one could calculate for any of the d covariates. For example, for measuring the importance of a binary treatment variable, one may define the target feature as a population average of the change in individual prediction when one flips the treatment from 0 to 1. In the sample, this means that for each individual in your sample you evaluate the prediction of the outcome under treatment and control, given its covariate vector, take the difference of these two predictions corresponding with the covariate vector of the individual, and average across all individuals in the sample. Clearly, we can then just apply that feature mapping to the HAL-curve to obtain a so-called plug-in estimate of this target feature. We refer to this particular estimand as an ATE estimand since under causal assumptions it actually equals the average treatment effect (ATE) of the binary treatment controlling for the baseline-confounders, where one then needs to assume that these covariates are indeed measured before the realization of treatment. This ATE estimand would be equal to the coefficient in front of treatment in a linear regression model if that model would be correct, so we did not have to give up anything by defining such an estimand, but we enriched its meaning to a meaningful quantity in the real world, not just in a toy world that does not exist.

Richard: Time for a little intermezzo. So far we have been discussing the regularization/shrinkage part of LASSO, now let's discuss the issue of automatic feature / variable selection, which often is mentioned as a benefit of LASSO. This looks a little beside the point, but you know that I like being beside the point. I think this property is not always a benefit and automatic feature selection is just part of the problem we are dealing with (i.e. a flawed methodology in machine learning and data-analytic practice) not part of the solution. We now experience a situation were students in data science, usually without proper statistical training, are encouraged to use automatic variable selection routinely. In the open-source language R there is even a package called "Feature TerminatoR", honoring the legacy of Arnold Schwarzenegger in a most peculiar way".

Mark: That looks a little drastic, sounds like a commercial for a new kind of pesticide.

Richard: What's more, we have had these techniques in regression for over 40 years (subset, stepwise, forward, backward, etc) so this is nothing new and even in those days, handbooks of methodology warned that this is not a proper way to do data-analysis, -not to mention inferential statistics- because coefficients, p-values, confidence bounds could be flawed / unreliable. The choice of variables should not be based on some automatic process that is not even really data-driven. Fancy nomenclature, buzzwords or lingo like ”feature extraction” (basically PCA or some other unsupervised learning technique for dimension reduction) or even worse, the more general term ”feature engineering” does not make all this legitimate. This criticism especially applies to feature engineering, which suggests that the choice of variables is just part of data preparation, getting the raw data in the ”right” shape for machine learning algorithms to deal with them, just like standardizing, normalizing or log-transforming data, imputation methods for missing values, cleaning data errors, et cetera. Wouldn't it be good to return to traditional ”Victorean” values, put on the breaks a bit and stick to sound, good old-fashioned empirical research. Such as formulating carefully a hypothesis, that gives a tentative answer to a research question, do your literature review, choose a conceptual model, collect data on variables and choose your variables on sound theoretical grounds, analyze the data with the appropriate analytical technique, interpret and visualize results. I mean, that sounds like a constructive methodology. Am I missing the point, is it too old-fashioned or what?

Mark: These are excellent questions and many practitioners will recognize themselves in these questions. You actually raise many points I cannot address in one response, but they are important. But one take on what you just said is that the practice of statistics is so unbelievable confusing if one is not clear about the goal and benchmarks. In one way, you raise the issue that we should have a question in mind before we start the process of learning from data and that this learning should be guided by the question. And that is absolutely correct, but, at the same time, if you return to Victorian values to use your bizarre vocabulary you might fall into the trap of holding on to a simple model as something important and that great art is involved by an expert and that one should not just throw the kitchen and sink at this.

Richard: How is that?

Mark: If you exaggerate this traditional notion of a model, like you did, then implicitly you are then making the wrong assumption that the answer to the question of interest cannot be learned in such a machine learning type fashion. In the end, if we agree on the target, we can benchmark and judge different approaches in terms of the bias and variance of the resulting estimator, beyond if it provides valid inference in terms of confidence intervals or testing the null hypothesis of interest.

Richard: Well, perhaps I committed the strawman fallacy a bit. Of course, it is about the level of guidance of your research: a specific question about the parameter of interest, a more vague concept of a hypothesis, or basically nothing. I mean since Leo Breimans ”Statistical Modelling (Breiman, 2001); the Two cultures” from 2001 these questions have become ”en vogue” again.

Mark: Yes, essentially you are referring to two camps of data scientists, the hypothesis driven statisticians that provide p-values and confidence intervals based on typically misspecified parametric models, and the machine learning community going for fancy prediction functions based on blackbox algorithms as ”crazy” as deep learning algorithms. In your question you seem to lean towards the traditional statistician as representative of a more scientific approach.

Richard: Let me make two remarks about this. Yes, from a more epistemic point of view, scientific reasoning is -at least since the probabilistic revolution that started in the 19th century- associated with statistical inference, bridging the gap between sample and population, hypothesis testing and parameter estimation as two sides of the same coin. Now statistical inference can be done in the classical, frequentist way with p-values, confidence intervals, H0 and H1 hypothesis, Type I and Type II error and -we cannot overemphasize this- with sampling distributions (obtained theoretically by CLT or empirically by bootstrap) representing the variability of the chosen statistics, but the parameter of interest viewed as a fixed unknown quantity, that as such is no random variable, has no distribution. It can also be done in the Bayesian way with bayes factors, priors and posteriors, credible intervals, model comparison and a parameter of interest that is no longer fixed, but is considered to be a random variable itself. This classical epistemic view is eroding due to successes of exploratory data analysis (EDA) (pioneered by Tukey (Tukey, 1978), since the early sixties of the previous century) and of course computational data analysis and machine learning since the eighties and nineties.

Mark: This all sounds as music to my ears.

Richard: Another intermezzo, I do have a little problem with your qualification of deep learning though.

Mark: Why is that? I mentioned that these are in some sense the opposite of interpretable machine learning by fitting the data with massive dimensional parametric models. They can still be judged by how good they are actually predicting. Anyway, love to hear your perspective.

Richard: The huge successes of AI over the last years are for a substantial part due to the rise of deep learning. In fact, we are talking about neural nets, which were at the heart of AI even before Turing wrote his paper in 1950. Although the roots of AI are clearly probabilistic, for a long time the sub-symbolic techniques were bashed by classical logic-based AI and were even defunded in the sixties and seventies, invoking an AI Winter and a Neural Net Winter. After great improvements in the nineties, the deep learning revolution emerged in the beginning of this century. Now these neural nets receive backlash again, partly because of the fact that all these millions of coefficients of parameters are not interpretable and the conclusions can not be explained. Now the field of Explainable and Responsible AI has been receiving much attention and funding. I am involved in all this myself. The threads however are often exaggerated or not fully understood, but let's leave this to a following discussion, we really should dedicate a separate dialogue to this. In my view, the fact that statistics is basically a disunited field with many unsolved debates and controversies, whereas textbooks suggest coherence, is far more problematic than deep learning, so let's close this intermezzo for now.

Mark: Agree completely.

Richard: I guess that our aim to reconcile statistical learning and machine learning, and related to this solve the crisis in data analytic practice by devising a constructive methodology is not really affected by that. The fact that statisticians do not seem to agree on the foundations of their field, makes it all that easy for ML-people and data scientists to ignore it and say, let the data do their work, it is all just data management, storage and retrieval, just displaying and visualizing the results.

Mark: I agree, this immense incoherence about our field is the very reason why in our research early on we defined a roadmap for learning from data, also called the roadmap for causal inference and targeted learning (Pet; van der Laan and Rose, 2011). This roadmap unifies our field and sets clear benchmarks. Once you own this roadmap you don't get confused anymore and every contribution can be put in context and be evaluated accordingly. We should probably discuss the steps in some detail but for now let me just state that it involves defining the experiment that generated the data, specifying knowledge about the experiment and thereby the probability distribution behind the observed data; specify the statistical estimand one aims to learn from the data; specify (a priori) an estimator with inference; and sensitivity analysis acknowledging a potential gap between the statical estimand and the quantity of interest due to violation of non-testable assumption. Each step deserves its own discussion, but for now I just want to point to the key step that requires the definition of a statistical estimand, which requires defining a target feature mapping that maps the true data distribution into a number. This target feature mapping defines the statistical estimand as a well-defined feature of the data distribution without relying on unrealistic assumptions such as that the true regression function is described by a linear main term model. The estimand represents the answer to the question of interest that one wants to address with the data analysis.

Richard: A data scientist following the roadmap has to commit to a choice of statistical estimand, thereby commits to the goal to estimate this estimand from the data and possibly provide inference for it as well.

Mark: Exactly, in particular, we should then judge the quality of the procedure used by bias, variance, MSE, and coverage of proposed confidence intervals, all with respect to this estimand. One possible estimand is the whole regression function, while another one might be the ATE estimand we defined above for a binary treatment controlling for a set of confounders. In particular, if the goal is prediction (i.e., the whole regression function), then one can understand that feature engineering and highly flexible machine learning algorithms might represent a sound approach. Still it also points out that there are so many variations available to the user that one wonders how to choose among all these options. This then gets into the Super Learner. Either way, one could also understand that a pure machine learning approach might yield a poor estimate of the ATE, precisely the very point you raised earlier. This gets then into TMLE and targeted learning. You can tell you have opened a whole bunch of concepts we might want to discuss in a more modulated fashion.

Richard: No doubt, we will address this in a upcoming dialogue. So obviously HAL is unthinkable without the paradigm of TMLE, is that not so?

Mark: Yes, HAL is an intrinsic part of the TMLE literature (e.g, (van der Laan and Rubin, 2006; van der Laan and Rose, 2011))and also represents an important part of my journey towards the most robust and powerful methods for learning from high dimensional data. Of course, HAL can just be viewed as another machine learning algorithm in the prediction or function estimation literature, an enrichment of the super-learner, but it is much more than that. It also represents a very fundamental enrichment of TMLE, and HAL itself provides theoretically grounded plug-in estimator of any target feature of the target function one might be interested in, completely analogue to parametric MLE (van der Laan et al.). All of this really represents my overall journey from Ph.D till now.

Richard: Now let us consider the concepts of parametric, semi-parametric and non-parametric. Even in the statistical literature they are not always fully consistently defined. They do not always look mutually exclusive or totally exhaustive, in fact these concepts seem to evolve. Now you once mentioned that HAL is the first general non-parametric MLE.

Mark: Yes, if we apply HAL to maximize the log-likelihood over this class of functions constrained by the sectional variation norm, then that is an MLE over such a rich statistical model that it contains the true data distribution. So that is the analogue of non-parametric MLE but one that applies to realistic statistical models. Contrary to an NPMLE that often ends up with an empirical probability distribution, HAL is an MLE over a class of densities that actually ends up with a fit that is a density itself. For example, the NPMLE over all univariate densities of a continuous random variable ends up being the empirical probability distribution that puts mass 1/n on each observation. Similarly, the Kaplan-Meier estimator of a survival function based on right-censored failure time data is an NPMLE, and one that is again a discrete distribution with mass at the observed failure times. The MLE over the Cox-proportional hazards model also ends up being a discrete conditional hazard even when one knows that the true failure time is continuous. In these problems this NPMLE is a bad estimator of the actual density of the data but it still yields efficient estimators of smooth features of the density. In particular, in spite of the fact that the Nelson-Aalen estimator of the hazard is a terrible estimator of the real underlying hazard, it still yields an excellent estimator of the survival function, or of the regression coefficients in the Cox-proportional hazards regression model.

Richard: But the traditional non-parametric Kaplan-Meier curves and the parametric Cox-proportional hazards method were especially effective long before the era of big data emerged.

Mark: Yes, in all these examples where NPMLE still yields good estimates of smooth target features, the dimension of the data is usually low. For actual realistic data structures and corresponding realistic statistical models these NPMLE are generally not well defined and either way represent bad estimators even for smooth features of the data distribution. For example, suppose one wants to estimate the conditional mean of an outcome given a few continuous covariates. Then an NPMLE wants to estimate this conditional mean at a particular covariate x with the empirical mean of the outcome among all observations with $X_i = x$. However, there are either none or one subjects with $X_i = x$. If there are none, then it is not defined, and if there is one, then it just estimates it with the outcome Y_i that observation, which is a total overfit of the regression curve at that x. Similarly, if one wants to compute an NPMLE of the ATE estimand, one needs within each treatment group, for each covariate configuration in the sample, that there are one or multiple observations, but again, there will be none or one. Therefore, the moment our data takes into account baseline covariates or even time-dependent covariates, estimation of target features will require machine learning. Therefore, there was a real need to have an MLE that actually really estimates the data density which can then also be used to obtain plug-in estimators of target features.

Richard: It is in some way ironic that MLE has -from a historical point of view- long been identified or associated with Fisher and the "grandeur and misery" of parametric statistics, deemed less relevant in the Big Data era due to the "curse of dimensionality"; and now there is fully nonparametric MLE.

Mark: It is and to appreciate this, one should realize that in the recent literature, I was myself involved in as well, assumed that MLE was just not the right approach for estimation in realistic models due to curse of dimensionality. That is how the field concerned with semiparametric efficient estimators moved from MLE to the one-step estimator and estimating equation approach and eventually TMLE. HAL turns that on its head. HAL created a lot of clarity for me that I lacked about these movements in the literature. Anyway, this is one reason why the focus on such concepts as parametric, semi-parametric and non-parametric do evolve like you mentioned already.

Richard: Very interesting, we should continue our discussion in an upcoming dialogue. You mentioned that HAL enriches the super-learner, and I can understand that it does due to its excellent rate of convergence, and presumably excellent practical performance. However, you also mentioned that it resulted in a real advance for TMLE of target features of the data distribution such as the ATE estimand, and, more generally, of causal estimands under identification assumptions. I realize that we did not discuss the TMLE in detail yet, nor what TMLE really is, but for now, let's just understand that TMLE is a plug-in estimator of the statistical estimand, obtained by replacing in the estimand the true data density by a targeted estimator of the density. The targeted estimator of the data density is obtained through a two-stage procedure that first obtains an initial density estimator, generally speaking using the state of art in machine learning, and then carries out a parametric MLE update along a parametric model that has as off-set the initial estimator (van der Laan and Rubin, 2006). I have provided a philosophical and historical perspective on models, inference and truth and how TMLE takes these concepts at hard (**?**).

Can you tell us in what sense HAL enriched this general TMLE procedure?

Mark: Firstly, we can now use as initial estimator of the data density or the relevant parts thereof a discrete super learner that includes various HAL-algorithms that vary the function space beyond the sectional variation norm. This gives us now an initial estimator that is guaranteed to converge at a rate $n^{-1/3}$ to the true density (van der Vaart et al., 2006; van der Laan and Dudoit, 2003). The asymptotic normality of TMLE, and thereby asymptotic valid inference, relies on a second order remainder being negligible after scaling it by $n^{1/2}$. Therefore, the asymptotic analysis of TMLE relies on the initial density estimator to converge at a faster rate than $n^{-1/4}$. Due to HAL, for the first time we could now state that TMLE using HAL is guaranteed to be asymptotically efficient under absolutely realistic assumptions about the data density (i.e., finite sectional variation norm). For the first time in the literature we had shown that there is no real asymptotic curse of dimensionality, and that asymptotic efficient estimation is actually possible, even in nonparametric models. This point was made in my article that introduced HAL (van der Laan, 2017). In fact, this second order remainder converges at rate $n^{-2/3}$ clearly exceeding the required $n^{-1/2}$. The other key advantage of HAL relative to other machine learning algorithms is that it is an MLE and thereby solves a lot of score equations. These score equations can be shown to make this second order remainder even smaller: that, is the MLE is essentially knocking out the exact remainder by solving enough score equations. This insight was the basis of our work on higher order TMLE (van der Laan et al., 2021). So, we can say that HAL-TMLE is not just first order efficient but that HAL also really helps the finite sample performance. The key to remember is that an NPMLE, if it would exist, would set this second order remainder equal to zero in all these causal and censored data estimation problems, something I proved in my Ph.D thesis (van der Laan, 1996). The HAL is approximating the NPMLE and thereby also has a superior second order remainder relative to TMLE using general machine learning algorithm that lack this key MLE property.

Richard: Would you say that the fact that HAL-TMLE is guaranteed asymptotically efficient, has practical implications, like the convergence at fast rate, or is it just a nice theoretical property for statisticians?

Mark: In general, my take is that I want an estimator that is asymptotically optimal, efficient say, but, among all such estimators, one wants one that works best in the type of sample you are dealing with in your data analysis. It was frustrating to me that before HAL I had to rely on the super-learner to include a good enough machine learning algorithm to obtain this rate $n^{-1/4}$ even though we had no theoretical guarantees. So, I find it very comforting that we can now state that an HAL-TMLE is known to be asymptotically optimal for any smooth feature of the data under any realistic statistical model. Of course, asymptotics does not always kick in at the given sample size, so that this kind of result might not be that relevant. That is why it is also so important that HAL is an MLE and a theoretical result shows that if the HAL solves the right score equation then the exact second order remainder gets knocked out completely and, by approximately solving this right score equation it becomes a third or fourth order (etc.) difference. Since HAL solves a growing collection of score equations it also solves any score equation in the linear span of these score equations, and the latter starts approximating the desired score equation that would knock out the exact second order remainder in the expansion of the HAL-TMLE. In that sense by using HAL in the HAL-TMLE we are improving its finite sample behavior in a practically important way. Another nice feature about solving score equations is that the density estimator starts becoming asymptotically normally distributed as an estimator of the density itself (van der Laan, 2023) and, in fact, the plug-in HAL-estimator of smooth target features of the data density are already asymptotically efficient. These are theoretical results, but they are also very telling about the finite sample robustness of the whole procedure. Interestingly, due to using HAL as initial estimator, even without the TMLE it already yields an asymptotically efficient plug-in estimator.

Richard: So if HAL already provides us with an asymptotically efficient plugin estimator, what is then the point of TMLE?

Mark: The key is that the TMLE is still essential since in finite samples there is a real curse of dimensionality, and the targeting of the TMLE step will be crucial. So, the TMLE says, let's not just solve all these non-targeted score equations that HAL solves, but let's make sure we also solve some key score equations that are directly relevant for our target feature. In my more recent research, we keep seeing that undersmoothed HAL does not get us far enough and that the targeting step is absolutely making a big difference in finite samples. So, the TMLE not only survives HAL, but is enriched by HAL.

Richard: Let's elaborate al little on the significance of nonparametric techniques. Already in 1995 computer scientist Paul Cohen stated in his "Empirical Methods for Artificial intelligence" that statistical data analysis provides us with a general methodology for AI. One of the reasons that since the nineties statistical data analysis became popular in AI was the rise of computer intensive statistical techniques. We should first mention permutation tests or randomization tests, which are similar to the classical exact tests, anticipated and developed by Fisher 100 years ago. In fact, his famous experiment, conducted in a tearoom -about a lady who claimed being able to taste whether milk was added to the tea or just the other way round- was a small data exact test. Then of course, Monte Carlo methods, which help us solving intractable problems, which are essentially deterministic in nature in a computer intensive stochastic way. And last but not least, bootstrap methods, that give us "empirical" sampling distributions for virtually any statistic of interest and because they are fully non-parametric, they can be used when we cannot rely on CLT. Still, it would appear that for many machine learning algorithms the application of nonparametric bootstrap is known to be problematic.

Mark: You could say that again. Often these applications of the bootstrap in machine learning are proven to be invalid. It is therefore significant that the nonparametric bootstrap is a valid method for HAL-TMLE or HAL itself, which also this demonstrates the remarkable robustness of HAL (Cai and van der Laan, 2020). For example, the kernel density estimation literature clarifies that one should bootstrap from a fitted density and not from the empirical distribution. This makes sense since these kernel density estimators rely on underlying smoothness so they might have a very different behavior when applied to a sample from a pure discrete distribution that does not have any smoothness. However, one can use the nonparametric bootstrap to estimate the sampling distribution of HAL! Probably this is due to HAL not relying on any smoothness conditions about the underlying density, thereby making its sampling distribution not dramatically sensitive to the underlying smoothness of the data density.

Richard: How about recent extensions or applications of HAL, such as Multi-task HAL(Malenica et al., 2023), meta-HAL (Wang et al., 2023) and outcome adaptive HAL-TMLE (Ju et al., 2019)? To start with the first, I have been working with many people from the field of AI and cognitive science and multi-task learning obviously is something that has gained their attention because in a manner of speaking it refers to the way human agents learn and solve problems. Sometimes complex tasks can be divided into smaller subtasks which can be solved more efficiently in case of a common latent structure. What's more the idea that multiple tasks can be learned simultaneously by a shared model, is now a kind of subfield or paradigm in machine learning. How does Multi-task HAL fit into this?

Mark: Well, suppose we have 5 studies concerning patient data on patients that suffer from a certain disease, and in each data set we collect a number of observations on a covariate vector and an outcome. You could then define the collection of tasks as predicting the outcome from the covariates for each of the 5 studies. Maybe for some of the studies the outcome is an indicator of death and for others it is a biomarker measuring progression of disease. Or, each study uses a different instrument for measuring depression. The studies could have sampled from different populations, but they could also have overlapping populations. In particular, one can imagine a situation in which we have one sample from a population of patients diagnosed with some progressive disease, but we have multivariate outcome on each subject measuring different facets of the progression of this disease. One might also have the same outcome in each study but each study sampled a different population. In the latter case, it appears particularly clear that there is a lot of common structure across the five studies and it would be inefficient to not utilize that. Given we would like to use HAL or a discrete super learner with HAL candidates, one would naturally simply run such an HAL for each data set separately. However, as you state, if there is some common latent structure in each of these prediction problems one wonders if one cannot borrow from each other. For example, could it be that covariates predictive of a death outcome are also predictive of the biomarker outcome and maybe affect these outcomes in a similar way (e.g. same interactions). Or could it be that a prediction function optimized for predicting the biomarker outcome is actually a great dimension reduction for the prediction of the death outcome?

Richard: So, the hard work used to fit one of the prediction problems could be a great input for addressing the other prediction problem.

Mark: Exactly, and a very simple way to map this multi-task problem in a single task problem is to view the combined data set as a single data set in which each subject also has a study indicator, indicating that a subject belongs to one of the five studies. The combined prediction problem is then nothing else than predicting the outcome from the covariate vector and the study indiator, a categorical variable with 5 possible values. Stratifying on this categorical variable, would make it a separate prediction problem for each study. However, we could now fit a single HAL, where the loss for subject could be a different loss depending on what study it contributed to, based on the combined data set with this categorical study indicator as an extra covariate. One could now minimize the empirical risk for the overall combined data set over all cadlag functions, or subsets of these, with finite sectional variation norm. Such a HAL-model would now include indicators of study as well as interactions of the study indicator with the zero-order spline basis functions in the actual covariates. If there is now a lot of common structure the HAL-fit would now utilize this structure to reserve some of the L_1-norm for finding key interactions with study indicators instead of fitting a separate model for each study. For example, the HAL-fit might contain an indicator of belonging to study 1 or 2 times a linear combination of spline basis functions, thereby clearly expressing a common structure across the two studies. One might also utilize a group-lasso optimization in which we assign a separate L_1-norm to each study and thereby constrain the overall prediction function with 5 L_1-norms that can be chosen with cross-validation. In that way, one might control that the HAL-fit spends all its L_1-norm on one prediction problem.

Richard: Then let us elaborate a little on meta-reasoning, which obviously has always been playing a role in the sciences particularly in logic, cognitive science and AI and of course in the philosophy of science. In what way does HAL or meta-HAL to be precise, proceed in this tradition?

Mark: Well, the reality is that HAL has turned around my world and has opened up so many theoretically grounded advances not available before HAL. Meta learning in the context of prediction corresponds with creating a meta level data set in which one column is the outcome and the other columns are cross-fitted prediction functions based on a collection of machine learning algorithms. So, in essence the meta-level data set represents a coordinate transformation of the original covariate X into a new set of coordinates, possibly of much lower dimension. If one applies convex linear regression, one obtains the convex super-learner or the stacking algorithm. How about we replace convex linear least squares regression by HAL? We call that the meta-HAL super-learner. This estimator is just an HAL but applied to a different set of covariates, so that all the results on HAL carry over to the meta-HAL. When applying HAL at the meta-level one has to carefully tune the L_1 norm as well as the possibly additive model choice. Therefore, we really think of this as a discrete super learner in which we use a number of meta-HAL super-learners as candidates, beyond the convex super-learner, and beyond the base algorithms used in these super-learners. The discrete super-learner will then optimally tune the amount of fitting one should do at the meta-level. Meta-HAL is particularly attractive when the original covariates are extremely high dimensional such as the case that includes a brain image. One could then use a deep learning algorithm to create the cross-fitted predictions of the outcome from the brain image as one of covariates at the meta-level. In this manner, we can switch from a original data sets with diverse sources such as brain-image, NLP from text, claims, etc into a meta level data sets that utilizes the state of the art algorithms that were specifically developed for reducing the dimension of such objects.

Richard: So, as I see it, this is all about switching between reasoning in the object-language (that speaks and makes claims about the world) and reasoning in the meta-language (that speaks and makes claims about the object language itself). That is essentially an old philosophical distinction: reasoning and meta-reasoning, ethics and meta-ethics.

Mark: Yes, you could say that; one switches to the meta level, for example if the candidate base algorithms are too biased, because then the meta-HAL will go for an aggressive HAL to compensate for this, while if the candidate base algorithm are doing an excellent job themselves, then cross-validation might select a simple meta-learning algorithm such as the convex super-learner, which will do the trick. Either way this discrete super-learner is now not only improving on all the base-algorithms, but it also improves on the various meta super-learners. One can now use this meta-HAL super learner as a powerful prediction algorithm, but it could also play the role of the initial estimator in the TMLE or we could even apply the HAL-TMLE at the meta level. The meta-HAL super-learner also inherits the plug-in properties of the HAL itself, thereby resulting in efficient plug-in estimators of target features of the target function.

Richard: Finally, let us review the main features of outcome adaptive HAL-

TMLE? So far, we have only been dealing with this relation rather informally. Now, talking about desired statistical properties, one of these is no doubt the double robustness of an estimator. Especially from a practical perspective. You will remember that some time ago one of my students wrote his thesis on causal inference on observational data, estimating the ATE using TMLE. The aim was to explore the applicability of TMLE in data science and -among other things- the thesis involved making a comparison with other methods and the way they adjust for confounding. When it comes to dealing with confounding G-computation typically depends on the outcome mechanism, $E(Y \mid A, X)$, the conditional expectation of the outcome given the intervention A and covariates X. Propensity scores (more specifically inverse probability weighting) compute $P(A = 1|X)$, i.e. the conditional probability of the exposure given the observed covariates. TMLE is doubly robust by combining both in such a way that it gives unbiased estimates if either $E(Y|A, X)$ or $P(A = 1|X)$ is consistently estimated and this "two for the price of one" strategy proved its value on both simulation data and a small famous dataset. How about double robustness and HAL?

Mark: Yes, I remember you sent me the thesis. Indeed, a TMLE of, for example, the ATE based on observing n observations on (X, A, Y), X baseline covariates, A subsequent binary treatment, and Y a final outcome, relies on an estimator of both the outcome regression $E(Y \mid A, X)$ and the propensity score $P(A = 1|X)$. The regular HAL-TMLE would estimate both of these nuisance functions with HAL or a discrete super-learner with HAL-candidates. However, suppose that there are some covariates that are highly predictive of the treatment decision but are not very predictive at all of the outcome. These are like instrumental variables. The asymptotic variance of the TMLE of the ATE is very sensitive to the propensity score approximating zero or 1 for certain strata, and, similarly the finite sample performance of the TMLE update might be erratic due to such so called positivity violations. Therefore, for the sake of finite sample performance (such as MSE) of the TMLE it would be good that the propensity score gets only fitted on the important predictors of the outcome. For this purpose, (Shortreed and Ertefaie, 2017) introduced the outcome adaptive lasso. HAL allows us generalize their idea that relied on a typical main term lasso model as follows. We first fit the outcome regression with HAL. This HAL fit selects a bunch of zero-order spline basis functions, the indicators we talked about. We then fit the propensity score with HAL where we force these basis functions in the fit. That is, when we run the LASSO to fit the propensity score, we force in these variables that the outcome regression HAL selected, thereby not penalizing their coefficient in the

L_1-norm of the LASSO. We can then still add a lot of other basis functions in this initial HAL-model on top of that and let cross-validation tune the LASSO as usual.

Richard: So, in this manner the fit of the propensity score is prioritizing the variables that were shown to be predictive of the outcome, the most important confounders.

Mark: Yes, while it uses the L_1-regularization to still pick up the important features among all other variables. We have shown that this type of outcome adaptive HAL-TMLE remains a well-behaved estimator even under violations of the positivity assumption and just enhances the finite sample performance. The double robustness of the TMLE as you mentioned is preserved but the algorithm is just helped to focus its estimation of the key propensity score by giving it the right set of variables to work with and not waste a real effort on bias reduction by variables that are not real confounders (especially given the finite sample size, even though they might kick in at much larger sample sizes as still being relevant). A nice demonstration of this outcome adaptive HAL-TMLE is shown in an FDA funded demonstration project for a SENTINEL safety analysis aiming to estimate the causal effect of opioids on kidney injury (Wys). In that analysis we had to adjust for over 20,000 claim codes beyond a collection of 80 clinical variables. This approach for fitting the PS combined with targeted selection of the L_1-norm in the propensity score, through a technique Collaborative TMLE (CTMLE), resulted in a superior version of TMLE relative to a more standard TMLE that just fits the PS as one normally does. Anyway, HAL gave us a natural way to obtain targeted fits of the propensity score for the sake of TMLE.

Richard: Well, the proof of the pudding is always in the eating. I know TMLE, Superlearner, HAL have been applied in -and developed in cooperation with- the life sciences, pharmaceutical research and other areas where a lot is at stake. As I said before, the history of statistics has shown many examples that progress is fueled by working on real-life problems, rather than dealing with or relying on toy-examples. Or, by complacently retreating into the ivory tower and showing a monomaniacal focus on small technical results, postulated desirable properties with no practical (clinical, societal, commercial) importance whatsoever. So, we should deal with these issues in our next dialogue.

Mark: We will definitely do so.

Richard: Time for a little wrapping up. We have now discussed to some extent why regularization is an important extension of the still enormously influential statistical paradigm of GLM in the era of big data, why it doesn't sufficiently deal with important problems in data analytic practice, such as reliance on wrongly specified (parametric) models and -most importantly- how HAL goes far beyond this idea of shrinkage and feature selection of the traditional LASSO. We have then examined the practical significance of the method HAL, (in terms of "desirable" statistical properties), how it enhances ensemble learning and the TMLE. More specifically, we outlined MT-HAL, meta-HAL and HAL-TMLE. You will not be surprised that still a lot of methodological issues, not to mention philosophical problems in the field, have so far been unrevealed. Let us take up a few of these in our next dialogue.

Mark: We will do that all right!

References

A. Bibaut and M.J. van der Laan. Fast rates for empirical risk minimization over cadlag functions with bounded sectional variation norm. Technical Report https://arxiv.org/abs/1907.0924, Division of Biostatistics, University of California, Berkeley, 2019.

L. Breiman. Statistical modelling: the two cultures. *Statistical Science*, 16: 199–231, 2001.

W. Cai and M.J. van der Laan. Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (lasso) estimator. *The International Journal of Biostatistics*, 16 (2):20170070. https://doi.org/10.1515/ijb–2017–0070, 2020.

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.

T.J. Hastie, R.J. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer, Berlin Heidelberg New York, 2001.

C. Ju, D. Benkeser, and M.J. van der Laan. Robust inference on the average treatment effect using the outcome adaptive lasso. *Biometrics*, page https://doi.org/10.1111/biom.13121, 2019.

Ivana Malenica, Rachael V. Phillips, Daniel Lazzareschi, Jeremy R. Coyle, Romain Pirracchio, and Mark J. van der Laan. Multi-task highly adaptive lasso. Technical report, arxiv 2301.12029, 2023.

S.M. Shortreed and A. Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, doi: 10.1111/biom.12679, 2017.

R.J.C.M. Starmans. Models, inference and truth: Probabilistic reasoning in the information era. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Studies*, pages 1–20. Springer, New York, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (Vol. 58, No. 1): 267–288, 1996.

J.W. Tukey. *Exploratory data analysis.* Addison Wesley, 1978.

Mark van der Laan. Higher order spline highly adaptive lasso estimators of functional parameters: Pointwise asymptotic normality and uniform convergence rates. Technical report arxiv 2301.13354, 2023.

Mark van der Laan, Zeyi Wang, and Lars van der Laan. Higher order targeted maximum likelihood estimation. Technical report arxiv 2101.06290, 2021.

M.J. van der Laan. *Efficient and Inefficient Estimation in Semiparametric Models.* Centre of Computer Science and Mathematics, Amsterdam, cwi tract 114 edition, 1996.

M.J. van der Laan. A generally efficient targeted minimum loss based estimator. *International Journal of Biostatistics*, pages 1106–1118, 2017.

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

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

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

M.J. van der Laan, D. Benkeser, and W. Cai. *The International Journal of Biostatistics*, 19 (1):261–289.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

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

Zeyi Wang, Wenxin Zhang, and Mark van der Laan. Super ensemble learning using the highly-adaptive-lasso. Technical report arxiv 2312.16953, 2023.