BANK OF ENGLAND

Staff Working Paper No. 784
Parametric inference with universal function approximators

Andreas Joseph

July 2020
This is an updated version of the Staff Working Paper originally published on 8 March 2019

BANK OF ENGLAND

Staff Working Paper No. 784
Parametric inference with universal function approximators

Andreas Joseph[1]

Abstract

Universal function approximators, such as artificial neural networks, can learn a large variety of target functions arbitrarily well given sufficient training data. This flexibility comes at the cost of the ability to perform parametric inference. We address this gap by proposing a generic framework based on the Shapley-Taylor decomposition of a model. A surrogate parametric regression analysis is performed in the space spanned by the Shapley value expansion of a model. This allows for the testing of standard hypotheses of interest. At the same time, the proposed approach provides novel insights into statistical learning processes themselves derived from the consistency and bias properties of the nonparametric estimators. We apply the framework to the estimation of heterogeneous treatment effects in simulated and real-world randomised experiments. We introduce an explicit treatment function based on higher-order Shapley-Taylor indices. This can be used to identify potentially complex treatment channels and help the generalisation of findings from experimental settings. More generally, the presented approach allows for a standardised use and communication of results from machine learning models.

Key words: Machine learning, statistical inference, Shapley values, numerical simulations, macroeconomics, time series.

JEL classification: C45, C52, C71, E47.

[1] Bank of England and Data Analytics for Finance and Macro (DAFM) Research Centre, King's College London.
Email: andreas.joseph@bankofengland.co.uk

A previous version of this working paper was available under the title 'Shapley regressions: a framework for statistical inference on machine learning models'.

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. All errors are mine. Many thanks to David Bholat, David Bradnum, Marcus Buckmann, Mingli Chen, Victor Chernozhukov, Sinem Hacioglu, George Kapetanios, Evans Munro, Milan Nedeljkovic, Paul Robinson, Arthur Turrell and, the participants of '$(EC)^2$ conference: Big Data Econometrics with Applications' (Rome, December 2018), 'Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics' (Washington DC, October 2019), 'Economic Measurement' (London, May 2019), 'Machine Learning for Economics' (Summer Forum, Barcelona, May 2019) and the 'London Mathematical Finance Seminar Series' (London, January 2020), who provided crucial feedback and support at various stages of the project. The code and data used in this paper are available on https://github.com/bank-of-england/Shapley_regressions.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email enquiries@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1 Introduction

Model families from statistical learning,[1] like support vector machines, tree ensembles, and artificial neural networks, often excel in the accuracy of their predictions compared to more conventional approaches (Fernandez-Delgado (2014)). This is because these models are *universal function approximators*. Given sufficient training data, these models learn a large set of target functions arbitrarily well under suitable conditions.

However, this model flexibility comes at the cost of transparency due to their generally nonparametric structure.[2] Standard statistical inference, such as the testing of hypothesis and the construction of confidence intervals is mostly not possible for these models. Given that models like artificial neural networks form the backbone of recent advances in artificial intelligence or AI (Goodfellow et al. (2016)) and, thus, are at the forefront of fast technological change, this is a considerable concern not only from a technical, but also from ethical, legal or economic perspectives.

This paper aims to address this gap by combining insights from statistical learning, game theory and parametric inference. The property of universal function approximation, or (error) consistency, is the foundation we build on. Based on this, a central theme will be the conditions and by which means we achieve (estimation) consistency.

The idea presented in this paper is simple. Conventional inference is achieved by transforming a general model into something familiar, namely, parametric inference in a linear regression setting. The resulting problem of the right transformation is addressed by exploiting an established analogy between cooperative game theory and supervised problems in statistical learning (Strumbelj and Kononenko, 2010). The predictions of a model are decomposed into an exact Shapley-Taylor expansion (Agarwal et al. (2019)) making well-defined attribution to individual variables and higher-order interactions. The space of this decomposition is linear by construction, meaning we can use it as a transformed inputs, or generated regressors (Pagan (1984)), for a surrogate linear model.

The main questions are then the interpretation of such inference, the conditions under which it is valid, how to implement it in practice and its relations to previously proposed parameter estimation techniques in a machine learning context.

Previous work addressing the problem of inference using machine learning models mostly focused on estimating single quantities of interest in a high-dimensional setting, such as treatment effects in randomised experiments with potentially many covariates Athey and Imbens (2016); Chernozhukov et al. (2018). The advantage of using machine learning techniques in these settings is their effectiveness of dealing with high-dimensional nuisance parameters and unknown heterogeneity. Our approach generalises this work to the estimation of nonlinear effects of a single or multiple variables, e.g. *a priori* unknown higher-order terms and inter-

[1] Often referred to as "machine learning", which is the popular phrase nowadays. We will use the two terms exchangably.

[2] We recognise that models like artificial neural networks are parametric. However, these parameters do not have identifying properties, i.e. are not unique for describing a data generating process. Hence, we group them with other nonparametric function approximators, like tree models. See also below.

actions. Such effects are captured by the corresponding Shapley components. The standard way of communicating results, e.g. via regression coefficients breaks down in this context and we introduce summary statistics which allow for a similar presentation of results while also accounting for increased model complexity.

The current work is complemented by Chernozhukov et al. (2017) for cases where machine learning models may not converge due to known impossibilities in nonparametric inference (Stone, 1982). More generally, the presented Shapley regression approach aims at reconciling Breiman's 'two cultures' (Breiman, 2001) by allowing statistical inference without having to assume a stochastic model of the data.

The main concern for the validity of this inference form machine learning models is the often slow convergence of these models, leading to potentially inconsistent estimators. As we will show, this can be addressed by the use of cross-fitting and variational estimation and inference methods (VEIN, Chernozhukov et al. (2017)). The estimation of a machine learning model is based on a training sample, while *ex post* statistical inference is based on an independent test sample. Sub-optimal convergence rates in the first step can be countered by an appropriate ratio between training and test samples sizes with the former being proportionally larger than the latter. This procedure is well suited for cross-fitting, whereby the training and test samples are rotated across several folds, such that statistical inference can be performed on the full dataset. Quantities of interest, such as coefficients and confidence intervals are now random variables themselves with a sampling distribution across folds. To obtain obtain point estimates, the median of the respective quantity is taken according to VEIN while adjusting the confidence level accordingly.

To best of our knowledge, the presented approach is the first which allows to statistically assess the learning process of generally unknown data-generating processes going beyond empirical risk minimisation. By extracting variable Shapley attributions, we can evaluate learned functional forms of a model without having to specify them *ex ante*. For this, we introduce the concept of robust component estimation which quantifies the sample size dependent learning bias of single variables or functions thereof. These estimates can be used to extend the concept of estimation bias from individual coefficients or parameters to a wide range of functional forms. Hence, we can quantify the statistical learning process of a model in greater detail as previously possible. This may have important implications for the empirical and theoretical study of learning problems.

We test the proposed framework using numerical and empirical case studies of randomised experiments. We quantify the sample size dependent convergence of popular off-the-shelf machine learning models when estimating heterogeneous effects in the treated population. We derive several quantities of interest in this context. In particular, a generic treatment function allows to evaluate main treatment and higher-order interaction effects, both of which can be of unknown functional forms and do not need to be specified explicitly. This may be helpful to address challenges brought forward against randomised experiments and the generalisation of their findings (Deaton and Cartwright, 2018), e.g. by identifying and quantifying potentially complex treatment channels. The results from both applications show that universal approximators can learn complex treatment effects well. However, there are substantial differences between models in their ability to do so. Importantly, it mostly is not possible to assess these differences from first principles prior to an analysis. The proposed inference framework allows to quantify model differences to allow for robust inference and model selection.

The remainder of this paper is structured as follows. Section 2 introduces the notation, and main components from statistical learning and game theory needed. Section 3 states and discusses the main results. Section 4 discusses estimator properties based on universal approximators. Section 5 connects the Shapley regression framework to the estimation of treatment effects in randomised experiments and presents numerical and empirical case studies. Section 6 concludes. The literature is discussed together with the main results. An inference recipe, figures, tables and proofs are in the Appendix.

2 Methodological background

2.1 Notation and definitions

This paper considers the common case of an observed target

$$y \equiv f(x;\beta) + \sigma\,, \tag{1}$$

which we would like to model. Here, $f : D \subset \mathbb{R}^n \mapsto T \subset \mathbb{R}^q$ is the data generating process (DGP) of interest. It has domain D and maps into the finite target space T with probability law P, is p-times differentiable and piecewise continuous, and σ is an irreducible noise component with finite variance. We only consider the case $q = 1$, the extension to $q > 1$ is straightforward. The vector $\beta \in \mathbb{R}^{n'}$ describes the parameterisation of the DGP, which we are usually interested in studying.

The data $x \in \mathbb{R}^{m\times n}$ with n being the number of features or variables and m the number of observations. The data is assumed to consist of independent and identically distributed samples from the population generated by f. Note that we make no distributional assumption $P(x)$ or $P(y)$ apart from boundedness.

The nonparametric model $\hat{y} = \hat{f}(x;\theta) : D \subset \mathbb{R}^n \mapsto T \subset \mathbb{R}^q$ with $\theta \in \mathbb{R}^{q'}$, where $q' \to \infty$ as $m \to \infty$ is allowed. It represents our machine learning models of interest, such as artificial neural networks (ANN), support vector machines (SVM), or random forests (RF). The main difference between the sets of parameters β and θ is that the former identify the DGP, while the latter may be degenerate in the sense that different configurations of θ (number and values) can describe the same model.[3] The model parameters θ are also slightly different to their usage in semi-parametric statistics, where θ often describes a high-dimensional

[3] Strictly speaking, models like ANN and SVM are parametric opposed to, say, tree-based models. However, we refer to them as nonparametric too as single parameters do not have intrinsic meaning compared to the structural parameters β.

nuisance parameter, which may be present or not. The paper focuses on a cross-sectional setting. However, most concepts apply analogously to dynamic settings for which we will make references as needed.
The used index convention is that $i, j \in \{1, \ldots, m\}$ refer to individual observations (rows of x) and $k, l \in \{1, \ldots, n\}$ to variable dimensions (columns of x). No index refers to the whole dataset $x \in \mathbb{R}^{m \times n}$. Super-scripts S refer to "Shapley-related" quantities which will be clear from the context. Estimated quantities are hatted, except decompositions Φ/ϕ for simplicity. Primed inputs, e.g. x' refer to variable sets or elements thereof. We refrain from using set braces for a simplified notation, which will be clear from the context.

2.2 Learning theory

Statistical learning theory[4] addresses the general problem of approximating f from a finite sample x. This can be formalised by taking an appropriate loss $L\Big(y, \hat{f}(x,\theta)\Big)$ minimising the risk functional

$$R(\theta) = \int L\Big(y, \hat{f}(x,\theta)\Big)\, \mathrm{d}P(x,y)\,, \tag{2}$$

over the unknown probability measure $P(x,y) = P(x)P(y|x)$. This is achieved by minimising the expected empirical risk from the sample x,

$$R_e(\theta) = \frac{1}{m}\sum_{i=1}^{m} L\Big(y, \hat{f}(x,\theta)\Big)\,. \tag{3}$$

This setting encompasses many learning problems which are reflected in different choices of the loss L. The loss can be the squared or classification error, or a negative log density for regression, classification or density estimation problems, respectively. The main questions in learning theory are, under which conditions learning is possible, i.e. Eq. (3) convergences to some minimal possible value $R_e(\theta_0)$, what are the rates of convergence for different approaches, how can these rates be controlled and how can we construct efficient algorithms?

The *key theorem of learning theory* states that, if (2) is bounded, for the empirical risk (3) to be consistent, it is necessary and sufficient that $R_e(\theta)$ converges uniformly in probability to the actual risk $R(\theta)$ over the empirical loss $L(y, \hat{f}(x,\theta))$. That is, for any $\epsilon > 0$ and for any $\delta > 0$, there exists a number $m_0(\epsilon, \delta)$, such that for any $m > m_0$, $P(R(\theta_m) - R(\theta_0) < \epsilon) \geq 1 - \delta$ and

$$\lim_{m \to \infty} P\Big(\sup_{\theta}\big(R(\theta) - R_e(\theta)\big) > \epsilon\Big) = 0\,. \tag{4}$$

That is, uniform convergence of the empirical risk function R_e is the key requirement for a problem to be learnable.[5] The main assumption in this paper is that (4) holds, i.e. the

[4] This concise overview mostly follows Vapnik (1999).

[5] A slight generalisation of the key theorem of learning is provided in Shalev-Shwartz et al. (2010).

problems we consider are learnable. We denote the rate of error consistency $R_e \sim m^{-\xi_{ml}}$.[6] We focus on cross-sectional problems in this paper. However, the conditions for learning in dynamic settings, like mixing processes or prediction problems have also been investigated, see e.g. Yu (1994); Meir (2000); Adams and Nobel (2010); Mohri and Rostamizadeh (2010).

Making the assumption of learning, we can now contrast the current work with previous work on generic inference on machine learning models by Chernozhukov et al. (2017), which is complementary to the current paper. It covers the situation where machine learning models may fail to converge due to impossibilities in nonparametric inference, and efficient techniques are presented for the estimation of low dimensional quantities of interest. Particularly, the optimal rate of nonparemetric (error) convergence is given by $N^{p/(2p+n)}$ as $m \to \infty$ (Stone, 1982). If p is fixed and finite, no consistent nonparametric estimator exists if the data dimension n increases such that $n \geq m$. This is a practical possibility in situations with high-dimensional data, especially when the size of the available training sample is restricted relative to the (generally unknown) maximal dimension of the feature space. In contrast, when the feature dimension is finite or only slowly increasing with the sample size, such that convergence is possible, the techniques presented in this paper allow for what we call *full inference* on a machine learning model. That is, not only can we estimate and assess low dimensional coefficients but more general aspects like nonlinearities and interactions of the DGP without needing to specify them explicitly, as well as to quantify the state of learning, i.e. model convergence, itself.

2.3 Shapley values

The linear model $\hat{f}(x_i) = x_i\hat{\beta} = \sum_{k=0}^{n} x_{ik}\hat{\beta}_k$, with $\hat{\beta}_0$ the intercept and $x_{i0} = 1$, is special in the sense that it provides local and global inference at the same time. The coefficients $\hat{\beta}$ describe *local* effects via the sum of the product of variable components and coefficients at point x_i. At the same time, the coefficient vector $\hat{\beta}$ determines the orientation of the *global* model plane with constant slope in each direction of the input space.
The linear model belongs to the class of additive variable attributions. For an observation $x_i \in \mathbb{R}^n$ we define the model decomposition Φ as

$$\Phi\Big[\hat{f}(x_i)\Big] \equiv \phi_0(\hat{f}) + \sum_{k=1}^{n} \phi_k(x_i;\hat{f}) \overset{lin.model}{=} \hat{\beta}_0 + \sum_{k=1}^{n} x_{ik}\hat{\beta}_k\,, \tag{5}$$

where $\phi_0 = \hat{\beta}_0$ is again the intercept. The standard approach to test for the importance of a certain variable is to test against the null hypothesis $\mathcal{H}_0^k : \{\beta_k = 0\}$. The aim of the current work is to derive similar tests valid for nonlinear models, i.e. where a decomposition (5) is not directly accessible. A general approach with desirable properties has been proposed

[6] Convergence properties are an active area of research. See for example Cybenko (1989); Geman et al. (1992); Farago and Lugosi (1993); Steinwart (2002); Steinwart and Scovel (2007); Christmann and Steinwart (2008); Biau (2012); Scornet et al. (2014); Andoni et al. (2014) and references therein.

by Strumbelj and Kononenko (2010). The authors make the analogy between variables in a model and players in a cooperative game. In both cases, it is often not clear which player (variable) contributes how much to a pay-off (prediction), because players (variables) can be compliments or substitutes. For example, adding a player (variable) k whose skills (statistical properties) are similar (correlated) with that of other players (variables) already in a coalition (set of variables) S is unlikely to add much to the coalition's pay-off (model prediction). That is, k is a substitute for some of the players in S. The opposite is true for compliments.

The situation of the game already has a general solution which is given by the *Shapley values* of players in a cooperative game (Shapley (1953)). Taking the analogy from Strumbelj and Kononenko (2010), we can rewrite (5) with

$$\hat{f}(x_i) \;=\; \phi_0^S + \sum_{k=1}^{n} \phi_k^S(x_i;\hat{f}) \;\equiv\; \Phi^S(x_i), \qquad \text{with} \tag{6}$$

$$\phi_k^S(x_i;\hat{f}) \;=\; \sum_{x' \subseteq \mathcal{C}(x)\setminus\{k\}} \frac{|x'|!(n-|x'|-1)!}{n!}\left[\hat{f}(x_i|x'\cup\{k\}) - \hat{f}(x_i|x')\right], \tag{7}$$

where $\mathcal{C}(x)\setminus\{k\}$ is the set of all possible variables combination (coalitions) of $n-1$ model variables when excluding the k^{th} variable. $|x'|$ denotes the number of included variables and $|x'|!(n-|x'|-1)!/n!$ is a combinatorial weighting factor that sums to one over $\mathcal{C}(x)$.
Eq. 7 is the weighted sum of marginal contributions of including variable k accounting for the number of possible coalitions.[7] Statistical models mostly do not allow for inputs x' with missing components. The calculation of model Shapley values is discussed in detail in Section 2.5.

Shapley values are the unique class of additive value attribution with the following properties (Shapley (1953); Young (1985); Strumbelj and Kononenko (2010)).

Property 1: Efficiency. The attribution model Φ^S matches the original model $\hat{f}$ at x_i,

$$\Phi^S(x_i) = \hat{f}(x_i)\,. \tag{8}$$

In a modelling context, this property is called *local accuracy.* The Shapley decomposition always sums to the predicted value at each point.

Property 2: Missingness (null player). If a variable is missing from a model, no attribution is given to it, i.e. $\phi_k^S = 0$ (dummy player).

Property 3: Symmetry. If k and k' are two variables, such that

$$\hat{f}(x'\cup k) = \hat{f}(x'\cup k') \tag{9}$$

[7] For example, assuming we have three players (variables) $\{A,B,C\}$, the Shapley value of player C would be $\phi_C^S(\hat{f}) = 1/3[\hat{f}(\{A,B,C\}) - \hat{f}(\{A,B\})] + 1/6[\hat{f}(\{A,C\}) - \hat{f}(\{A\})] + 1/6[\hat{f}(\{B,C\}) - \hat{f}(\{B\})] + 1/3[\hat{f}(\{C\}) - \hat{f}(\{\emptyset\})]$.

for all possible x' not containing k or k', then $\phi^S_k = \phi^S_{k'}$.

Property 4: Strong monotonicity. Variable attributions do not decrease if an input's contribution to a model increases or stays the same regardless of other variables in the model. That is, for two models $\hat{f}$ and $\hat{f}'$ on the same domain and a coalition $x' \setminus k$, if

$$\hat{f}(x' \cup k) - \hat{f}(x') \geq \hat{f}'(x' \cup k) - \hat{f}'(x') \quad \Rightarrow \quad \phi^S_k(f, x') \geq \phi^S_k(f', x')\,. \tag{10}$$

This property is also called *attribution consistency* in the context of variable attribution. Most alternative attributions are lacking this property, substantially reducing trust (Lundberg et al. (2018)). The strong monotonicity property can also be formulated using partial derivatives for models where they exist. The Shapley attribution of a variable for a model with the larger partial derivative across the domain ought not to be less than that for a model with a smaller partial derivative in the same domain.

Property 5: Linearity. For any two independent models $\hat{f}$ and $\hat{f}'$, i.e. where the outcome of the one does not depend on the inputs or outcome of the other, the joint Shapley decomposition for a variable k can be written as

$$\phi^S_k\big(a_1(\hat{f} + a_2\hat{f}')\big) = a_1\phi^S_k(\hat{f}) + a_1 a_2 \phi^S_k(\hat{f}') \tag{11}$$

for any real numbers a_1 and a_2. A consequence of these properties is the following proposition.[8]

Proposition 2.1. *The Shapley decomposition Φ^S of a model $\hat{f}$ linear in parameters $\hat{\beta}$, $\hat{f}(x) = x\hat{\beta}$, is the model itself. The proof is given in the Appendix.*

Hence, the Shapley decomposition of the linear model is already known and is the model itself.

2.4 The Shapley-Taylor decomposition

Nonlinearities can be multiplicative, polynomial, or both. That is, nonlinear terms can be approximated by two or more variables being multiplied, taken by some power, or a combination of both at each point of the input space.[9] Eq. 6–7 do not differentiate between these situations. It defines an additive *main effect* for each variable. However, the a more detailed functional form of two or more variables may be of interest in many situations, e.g. when they are complements or dependent.

This situation can be addressed by using the Shapley-Taylor index (Agarwal et al. (2019)) for interaction terms, resulting in the Shapley-Taylor expansion of a model. We define the

[8] This corresponds to linear Shap in Lundberg and Lee (2017).

[9] An approximation may be pointwise for discrete variables or valid within some region for continuous variables up to a certain precision.

discrete set derivative of model $\hat{f}$ at point x_i as

$$\delta_{x'}\hat{f}(x_i|x'') \equiv \sum_{x''' \subseteq x'} (-1)^{|x'''|-|x'|}\, \hat{f}(x''' \cup x'')\,, \tag{12}$$

with x', x'' and $x''' \subseteq \mathcal{C}(x)$. The case $|x'| = 1$ corresponds to (7). Let $h \leq n$ denote the maximal order of interaction terms we consider, then the Shapley-Taylor index for variable interactions up to order h at x_i is

$$\mathcal{T}_h^S\big(\hat{f}, x_i \,|\, x'\big) \;=\; \begin{cases} \delta_{x'}\hat{f}(x_i \,|\, \emptyset) & \text{if} \quad |x'| < h\,, \\ \frac{h}{n} \sum_{x'' \subseteq \mathcal{C}(x) \setminus x'} \frac{\delta_{x'}\hat{f}(x_i \,|\, x'')}{\binom{n-1}{|x''|}} & \text{if} \quad |x'| = h\,. \end{cases} \tag{13}$$

Shapley components for terms of order smaller than h are given by the set derivative (12) relative to the empty set with all variables missing. Terms of order h and higher are included in the terms of order $|x'| = h$. This means that terms up to order $k-1$ are unbiased with respect to higher-order interactions, while terms of order k include all higher order terms.[10] The efficiency statement (8) takes the form

$$\hat{f}(x_i) \;=\; \phi_0^S \quad + \sum_{x' \subseteq \mathcal{C}(x),\, |x'| \leq h} \mathcal{T}_h^S\big(\hat{f}, x_i \,|\, x'\big)\,. \tag{14}$$

That is, the Shapley-Taylor decomposition of a model, by construction, sums to the model prediction at each point of the input space (local accuracy). We again arrive at (7) for $h = 1$. The Shapley-Taylor decomposition also satisfies the other properties of Shapley values in addition to an interaction axiom which is less relevant in the current context (see Agarwal et al. (2019)).

2.5 Shapley value computation

The computation of both the simple Shapley value (6 & 7) and the more general Shapley-Taylor expansion (14) are of exponential complexity in the number of variables n. This means that an exact calculation is intractable for even moderate variable sets. Approximation techniques are sampling from $\mathcal{C}(x)$ (Strumbelj and Kononenko (2010)), efficient model-based algorithms (e.g. Lundberg et al. (2018) for tree models), or a compression of the input space. In many applications, the modeller has a prior about what variables are most important or of greatest interest, while still using all variables to fit a model. Variables not of interest can be grouped into a set O of size $\bar{n}$, reducing the dimension of the input space to $n' = n - \bar{n} + 1$.[11] This allows for an exact calculation of terms in $x \setminus O$. There will likely be a trade-off between n' and the order of the Shapley-Taylor decomposition h, with a larger h forcing a smaller n'. However, the set O is given the status of a single variable in this computation, which allows to assess the overall importance of variables in O jointly. For instance, the set O can

[10] In the analogy to the Taylor expansion of a function around some point x_i, we can say that the order-h term contains the remainder.

[11] The plus one stands for all variables in O.

be adjusted if the overall share given to it is deemed to large.
We have not yet clarified how a model value conditioned on a subset x' of variables, $\hat{f}(x_i|x')$, can be calculated. That is, how to account for missing variables in coalitions of variables. Sundararajan and Najmi (2019) provide an overview of different approaches with likely differences in preference depending on the modelling situation. In this paper, we take the stand to use *conditional expectations* together with the Shapley-Taylor decomposition to account for variables dependencies. That is,

$$\hat{f}(x_i|x') = \mathbb{E}_b\big[\hat{f}(x_i)\,|\,x'\big] = \int \hat{f}(x_i)\,\mathrm{d}b(\bar{x}') = \frac{1}{|b|}\sum_b \hat{f}(x_i|\bar{x}_i'), \tag{15}$$

Variables not included in x' are integrated out over a background dataset b. The set $\bar{x}'$ represents variables not included in x'. The components of observations $\{x_i|\bar{x}_i'\}$ take the values of x_{ik} if $k \in x'$ and values from b otherwise. The background provides an informative reference point by determining the intercept ϕ_0^S. Possible choices are the training dataset incorporating all information the model has learned from, or the subset of non-treated individuals in an experimental setting. The reference value ϕ_0^S for these choices is the expected value of the model and the expected value for the non-treated sub-population, respectively. A potential issue with the conditional expectation approach (15) is that it is possible to violate strong monotonicity (10) without an additional requirement.

Proposition 2.2. *For strong monotonicity (Property 4) to hold in the conditional expectation framework (15) for two functions f and f', it is sufficient that $\phi_0^S(f\,|\,b) = \phi_0^S(f'\,|\,b)$, i.e. that they have the same base value on the background b. The proof is given in the Appendix.*

Proposition 2.2 is not unreasonable in practical situations, where different models are compared on the same dataset and on the same loss function. However, caution in comparing decomposition from different models is warranted if their base values strongly differ.

The advantage of using conditional expectations as background compared to, say, a single reference point (e.g. BShap in Sundararajan and Najmi (2019)), is that it reflects more information by accounting for the global model structure and the distribution of the data. All Shapley attributions require the choice of a background, with different choices leading to different attributions all fulfilling the properties of Shapley values. It is therefore of tantamount importance to be consistent and transparent in the choice of background when comparing results across models or between different studies.

Finally, the conditional expectation approach (15) makes the implicit assumption of variables independence, i.e. column-wise independence in our setting. This may not be justified in many situations. This is addressed by the Shapley-Taylor expansion. Higher-order terms respect dependencies between variables, while lower order terms are the net of these dependencies. Only terms of order h do not respect dependencies of order $h+1$ and higher. Thus, the Shapley-Taylor expansion provides a practical tool to assess and account for variable dependence. For instance, unknown variable dependence is likely of minor concern regarding variable main effects if second order terms are comparable small. The same holds for assessing second order terms and so on. This makes the above presented Shapley value framework highly flexible by imposing minimal conditions on the DGP and the used models.[12]

3 The Shapley regression framework

3.1 Shapley regressions

We are know in a position to state our main ideas. Parametric inference on a universal approximator can be performed by estimating the *Shapley regression*

$$y_i \;=\; \phi_0^S + \sum_{k=1}^{n} \phi_k^S(\hat{f}, x_i)\hat{\beta}_k^S + \hat{\epsilon}_i \;\equiv\; \Phi^S(x_i)\hat{\beta}^S + \hat{\epsilon}_i \,. \tag{16}$$

The last expression uses an inner product with $\hat{\beta}_0^S = 1$, and $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.[13] The surrogate coefficients $\hat{\beta}_k^S$ are tested against the null hypothesis

$$\mathcal{H}_0^k(\Omega) \;:\; \{\beta_k^S \leq 0 \;\;\Omega\} \,. \tag{17}$$

Eq. 17 tests the alignment of components ϕ_k^S with the target dependent variable. If it can be rejected, ϕ_k^S provides statistically relevant information about y. If not, its contribution are noise-like. The working of (16) and (17) is illustrated in Fig. 1. Without loss of generality, two variables x_1 and x_2 contribute additively to a model $\hat{f}$ in describing the target y (LHS). These contributions can be nonlinear (as for x_1) or close to noise (as for x_2). We usually do not know this *a priori*. The Shapley values attributed to each variable and observation on the RHS provide a linear representation of the model by construction. We now can test the alignment of each component with the target y and test against $\mathcal{H}_0$ or other hypotheses of interest. We expect $\hat{\beta}_k^S \approx 1$, i.e. perfect alignment as for x_1, if a model has perfectly learned from a variable, or $\hat{\beta}_k^S \approx 0$, i.e. no alignment as for x_2. This will also be the asymptotic values of $\hat{\beta}^S$ (see Section 4).

A key difference to the linear case is that tests against $\mathcal{H}_0^k$ depend on $\Omega \in \mathbb{R}^n$. That is, only *local* statements about the significance of variable components can be made due to the potential nonlinearity of the model. The model hyperplane may be curved in the input-target space compared to that of the linear model. Note that the Shapley decomposition absorbs the signs of each component, leading to the half-sided test in (17). Negative values may occur and indicate that a model as poorly learned from a variable. However, these are usually not significant even in a two-sided test.

Shapley regressions establish a further connection between parametric inference and statistical learning, namely the possibility to account for different error structures when testing

[12] It is possible to construct functions with high-order terms having an arbitrarily large importance. However, this is highly unlikely in real-word applications, in addition to such effects not being known a priori.

[13] Model Shapley values have been used in linear regression analysis before to address isues around collinearity (Lipovetsky and Conklin (2001)). I do not see scope for confusion with the current application in context of universal approximators.

hypotheses, such as heteroskedasticity. Eq. 16 also provides a direct connection to dynamic settings for statistical leaning without the need to adjust the learning framework itself. That is, most machine learning models treat the DGP as cross-sectional, while (16) allows to account for temporal dependences.

The following proposition provides further justification for the use of Shapley regressions (16) for inference with machine learning models.

Proposition 3.1. *The Shapley regression problem of Eq. 16 for a model $\hat{f}$ linear in parameters $\hat{\beta}$ is identical to the least-square problem related to $\hat{f}(x) = x\hat{\beta}$, i.e. $\hat{\beta}^S = 1$. The proof is given in the Appendix.*

That is, the Shapley regression from a linear model returns the model itself. The surrogate regression model (16) can, therefore, be seen as a natural generalisation of parametric inference to nonparametric models.
One caveat of inference using the auxiliary regression (16) is that the coefficients $\hat{\beta}^S$ are uninformative about the direction or magnitude of components ϕ_k^S. We therefore introduce generalised coefficients suitable for the standardised communication of results.

3.1.1 Shapley share and mean coefficients

The Shapley share coefficients (SSC) of variable x_k, or other Shapley-Taylor terms from (13), in the Shapley regression (16) is defined as

$$\begin{aligned} \Gamma_k^S(\hat{f},\Omega) \quad &\equiv \quad \left[sign\big(\hat{\beta}_k^{lin}\big) \left\langle \frac{|\phi_k^S(\hat{f})|}{\sum_{l=1}^n |\phi_l^S(\hat{f})|} \right\rangle_{\Omega_k} \right]^{(*)} \in [-1,1]\,, \qquad (18) \\ &\overset{\hat{f}(x)=x\hat{\beta}}{=} \quad \hat{\beta}_k^{(*)} \left\langle \frac{|(x_k - \langle x_k \rangle_b)|}{\sum_{l=1}^n |\hat{\beta}_k (x_l - \langle x_l \rangle_b)|} \right\rangle_{\Omega_k}, \qquad (19) \end{aligned}$$

where $\langle \cdot \rangle_{\Omega_k}$ stands for the average of x_k in $\Omega_k \in \mathbb{R}$. The SSC is a summary statistic for the contribution of x_k to the model within Ω. It consists of three parts. The first is the sign, which is the sign of the corresponding linear model. The motivation for this is to indicate the direction of alignment of a variable with the target y. The second part is coefficient size. It is defined as the fraction of absolute variable attribution allotted to x_k within Ω. The sum of absolute values is one by construction.[14] It measures how much of the model is explained by x_k. The last component $(*)$ is used to indicate the significance level of Shapley attributions from x_k against the null hypothesis (17).
Eq. 19 provides the explicit form for the linear model. The main difference to the conventional case is the normalising factor, which accounts for localised properties of non-linear models. Given the definition over a range of $x_k \in \Omega_k$, it is important to also interpret them in this

[14] The normalisation is not needed in binary classification problems where the model output is a probability. Here, the a Shapley contribution relative to a base rate can be interpreted as the expected change in probability due to that variable.

context. For example, contributions may vary over the input space such that $\hat{\beta}_k^S$ takes on difference values at different points or times.

A related summary statistic is the Shapley mean coefficient (SMC). It is defined as

$$\bar{\Gamma}_k^S(\hat{f}, \Omega) \quad \equiv \quad \Big[sign\big(\hat{\beta}_k^{lin}\big) \, \Big\langle \phi_k^S(\hat{f}) \Big\rangle_{\Omega_k} \Big]^{(*)}, \tag{20}$$

$$\overset{\hat{f}(x)=x\hat{\beta}}{=} \quad \hat{\beta}_k^{(*)} \quad \langle (x_k - \langle x_k \rangle_b) \rangle_{\Omega_k} . \tag{21}$$

The SMC highlights deviations from the background sample b, see (15), which is made explicit for the linear case in Eq. 21. The SMC can be used to measure differences between subgroups, such as between treated and untreated sub-populations in an experimental setting.

Under the conditions we required from $\hat{f}$, the classical central limit theorem applies to the sampling distribution of Shapley values $\phi_k^S(\hat{f})$, such that it tends to a multivariate normal distribution. This can be used to construct standard errors and confidence intervals for SSC or SMC. For instance, let $\mu_k = |\Gamma_k^S(\hat{f}, \Omega)| \in [0, 1]$ be the absolute value of the k-th SSC. The upper bounds on the variance of μ_k and its sampling standard error of the mean are given by[15]

$$var(\mu_k) \leq \mu_k(1-\mu_k) \leq \frac{1}{2} \quad \Rightarrow \quad \sigma_k^\phi \equiv se(\mu_k) \leq \frac{1}{\sqrt{2|\Omega|}} . \tag{22}$$

The sampling distribution of μ_k approaches a Gaussian $|\Omega| \to \infty$. Thus, σ^Φ provides a well-defined measure of the certainty we can attach to Γ^S within Ω.

3.2 Validity conditions for Shapley regressions

The Shapley regression (16) is an auxiliary model building on generated regressors (Pagan (1984)), minimising the log-likelihood

$$l\big(\beta^S, \hat{\theta}; y, x\big) \quad \sim \quad -\frac{1}{2\sigma_\epsilon^2} \Big[\big(y - \Phi^S(\hat{\theta})\beta^S\big)^T \big(y - \Phi^S(\hat{\theta})\beta^S\big) \Big] . \tag{23}$$

Inference with regard to β^S is valid under two conditions. First, the cross terms of the Fisher information must vanish, i.e. $\mathcal{I}(\beta^S, \hat{\theta}) = 0$.[16] This is achieved by a two-step approach commonly used in statistical learning, making the optimisation processes for θ and β^S independent from each other. The input data x are randomly split between a training sample x_{train} on which the model parameters θ are fixed. A hold-out sample x_{test} is then used to evaluate model performance, e.g. to test its out-of-sample prediction accuracy, and to perform inference estimating $\hat{\beta}^S$.

The second condition for valid inference is that the nonparametric part, Φ^S in our case, is

[15] One will generally be interested in the expected explanatory fraction μ_k of a variable, while the sign of the SSC is fixed. Accounting for the sign, the bound on the RHS of (22) needs to be multiplied by four.

[16] $\mathcal{I}(\eta, \eta) = -\mathbb{E}\Big[\frac{\partial^2 l}{\partial \eta \partial \eta'}\Big]$, with $\eta \in \{\theta, \beta^S\}$.

$\sqrt{m}$-consistent. The accuracy property of Shapley values (8) relates this to the error consistency of a model and its convergence rate $R_e \sim m^{-\xi_{ml}}$. However, nonparametric techniques, including machine learning models, often converge slower than $\sqrt{m}$, i.e. $\xi_{ml} < \frac{1}{2}$, which would lead to inconsistent estimation of $\hat{\beta}^S$. However, a low ξ_{ml} can be accounted for via appropriate sample splitting between the training and the test set. Specifically, the condition for the maximal size of the test is

$$m_{test} \quad \leq \quad m_{train}^{2\min_{k\in\{1,\ldots,n\}}\xi_k} \quad \overset{uniform}{=} \quad m_{train}^{2\xi_{ml}}\,, \tag{24}$$

$$R_e \quad \sim \quad \textstyle\sum_{k=1}^{n} |\Gamma_k^S(m)|\; m_{train}^{-\xi_k} \quad \overset{uniform}{=} \quad m_{train}^{-\xi_{ml}}. \tag{25}$$

The convergence rate of individual Shapley components ϕ_k^S are labelled ξ_k. If the ξ_k are different, the smallest ξ_k will dominate R_e at some point and lead to the most conservative condition for m_{test}^{max} determining ξ_{ml} beyond some value m_0. Rates of convergence can depend on the sample data, model and algorithm used, which also means that general rules and practically useful bounds are often not available. Nevertheless, the convergence properties of R_e can be determined empirically by fitting learning curves, i.e. the model error dependence on the sample size. This is also the practical approach we will be following in the applications in Section 6.

Condition (24) puts an upper bound on the number of observations we can use for inference. If $\xi_{ml} < \frac{1}{2}$, this reduction of available sample size may impose an undesirable trade-off. However, we can avoid this by the use of *K-fold cross-fitting* (Chernozhukov et al., 2018), which also is amenable to how machine learning models are evaluated in practice. Namely, we first set

$$K \;=\; \left\lceil \frac{m_{train}}{m_{test}} \right\rceil + 1 \;=\; \left\lceil m^{1-2\xi_{ml}} \right\rceil + 1\,, \tag{26}$$

i.e. the ceiling of the ratio, and then randomly split x in K approximately equally sized folds. For $l \in \{1, \ldots, K\}$, we next use the K_l^{th} fold for testing and inference, and the remaining $K-1$ folds for training.[17] The number of folds K can become large depending on the relation between m (high) and ξ_{ml} (low). We see this together with the challenge of calculating Shapley decompositions in high-dimensional settings as the two main drawbacks of using universal function approximators for statistical inference.

The above procedure leads to the the situation of *variational inference and estimation (VEIN)* methods (Chernozhukov et al., 2017), where we have two sources of sampling uncertainty: (i) the estimation uncertainty of quantities like $\hat{\Xi} \equiv (\hat{\beta}^S, \Gamma^S)$ and (ii) variation induced by cross-fitting. (i) is dealt with in the standard way resulting in point estimates and confidence intervals for $\hat{\Xi}$. However, (ii) leaves us with a set of K random realisation of Ξ and their confidence intervals, some of which may be favourable and some not. To arrive at robust adjusted point and interval estimators, we adapt the VEIN estimators from Chernozhukov et al. (2017):

[17] This procedure is slightly different from what in statistical learning is called K-fold cross-validation, which uses the above procedure for hyperparameter tuning. This still can be done using cross-validation on the training datasets.

$$\hat{\Xi}^V := Med[\hat{\Xi}|x] \qquad \text{(point estimate)} \qquad (27)$$

$$\left[\hat{\Xi}^V_{low},\, \hat{\Xi}^V_{up}\right] := \left[\underline{Med}[\hat{\Xi}_{low}|x],\, \overline{Med}[\hat{\Xi}_{up}|x]\right] \qquad \text{(confidence interval)} \qquad (28)$$

$$\alpha^V := 2\alpha \qquad \text{(confidence level)} \qquad (29)$$

That is, the main cost of data splitting is a discount in the confidence level by a factor of two. $\underline{Med}[X]$ is the usual median and $\overline{Med}[X]$ is the next distinct quantile of a random variable X. The two medians coincide for continuous random variable but can (slightly) differ for discrete ones.

4 Estimator properties of universal approximators

Without loss of generality we focus on regression problems,[18] where it is common to minimise the mean squared error (MSE) between a target y from an unknown DGP $f(\beta)$ and a model $\hat{f}(\theta)$ over the dataset x. The expected MSE can be decomposed in terms of the bias-variance trade-off

$$\mathbb{E}_x\left[\left(y - \hat{f}(\theta)\right)^2\right] = \underbrace{\left(f(\beta) - \mathbb{E}_x\left[\hat{f}(\theta)\right]\right)^2}_{\text{bias}^2} + \underbrace{\left(\hat{f}(\theta) - \mathbb{E}_x\left[\hat{f}(\theta)\right]\right)^2}_{\text{variance}} + \sigma^2\,, \qquad (30)$$

where σ^2 is the irreducible error or unlearnable component of the DGP corresponding to the variance of y setting effective bounds on the empirical risk R_e. The separation model parameters θ and structural parameters β in (30) is important, because machine learning models are often subject to regularisation as part of model calibration and training. This affects θ during optimisation. Thus, if $\hat{\beta}$ would explicitly be part of the training process, its values would be biased as was shown in Chernozhukov et al. (2018).
It is a fundamental problem in statistical learning of how to generalise to $f(\beta)$ from (y, x) by the means of $\hat{f}(\theta)$. While it may be intuitive that universal approximators generalise to $f(\beta)$ via the convergence of the empirical risk R_e, there are so far no results which make this explicit, and where the generalisation process can be evaluated going beyond R_e and towards the structural parameters β. This problem can be addressed using Shapley regressions which we discuss next.

4.1 Estimator consistency

Statistical inference on machine learning models requires two steps. First, the control of bias and variance according to (30) and, second, the extraction of and inference on $\hat{\beta}$. Regarding the former, the imposition of error consistency requires that the squared error tends towards

[18] A regression problem in statistical learning refers to fitting a continuous target or dependent variable, as opposed to a discrete classification problem. Results presented here can be adapted to the classification case (Domingos (2000)).

σ^2 as $m \to \infty$. An important consequence of this is that the expected bias of a model tends to zero in probability, that is

$$p \lim_{m\to\infty} \mathbb{E}\big[y - \hat{f}|x\big] = 0\,, \tag{31}$$

However, we are usually interested in *estimator consistency* in a statistical inference setting, i.e. if

$$p \lim_{m\to\infty} \mathbb{E}_{\hat{f}}\big[\beta_k - \hat{\beta}_k|x\big] = 0\,. \tag{32}$$

That is, if universal approximators learn the true functional form.

To connect these two concepts, we take a two-step approach. In many applications of interest, f can be locally approximated by a polynomial regression. For a polynomial DGP the following result holds.

Theorem 4.1. *(polynomial consistency of universal approximators): Let f be a DGP of the form $f(\beta, x) = \sum_{k=0}^{n} \beta_k p_k^d(x) \equiv P_d(x)$, where $p^d(x)$ is a polynomial of order d of the input features on a subspace $x \in \Omega \subseteq D \subset \mathbb{R}^n$. If for each $x' \subseteq \Omega$, a model $\hat{f}(\theta)$ is error consistent according to (31), then the estimator $\hat{\beta}(\theta)$ is consistent in the sense of (32). The proof is given in the Appendix.*

Theorem 4.1 can be used to make a more general statement.

Corollary 1 *(universal consistency of universal approximators)*: *Let $f(\beta)$ be a DGP on $\Omega \subseteq D \subset \mathbb{R}^n$ and $\hat{f}(\theta)$ a model not involving β. If f can be approximated by a polynomial $\hat{f}_p(\theta')$ arbitrarily close within Ω and $\hat{f}$ is error consistent in the sense of (31), then $\hat{f}$ is estimator consistent for any $f(\beta)$ within Ω in the sense of (32). The proof is given in the Appendix.*

Corollary 1 tells us that an error consistent model will learn most functional forms of interest and the true DGP up to a certain precision primarily depending on the quantity if available data. This result may also be called *implicit estimation consistency*, as the above statements and proofs to not provide practical means to infer the state of estimation. However, Shapley regressions can be used to explicitly assess the state of estimation.

4.2 Estimator bias

When has a model sufficiently converged for well informed inference, e.g. judged by its Shapley share coefficients (18)? Before addressing this question, let us extend Shapley decompositions to general finite decompositions.

Lemma 1 *(model decomposition)*: There exists a decomposition $\hat{\Psi}\big(\Phi^S(x)\big) \equiv \sum_{c=1}^{C} \hat{\psi}_c(x) = \hat{f}(x)$ if the equation $\hat{\Psi}(x) = \Phi^S$ is solvable for each x_k, with $k \in \{1, \ldots, n\}$. The proof is simple as $x(\Phi^S)$ can be used to construct $\hat{\Psi}$. □

A decomposition $\hat{\Psi}$ can be called an *additive functional form representation* of $\hat{f}$ with the

Shapley decomposition Φ^S being the trivial representation. That is, $\hat{\Psi}$ is a parameterisation of f for which the following results holds.

Theorem 4.2. *(composition bias): Let f be a DGP and $\Psi^*(x) \equiv \sum_{c=1}^{C} \psi_c^*(x) = f(x)$ the true decomposition of f at x. Let $\hat{f}$ be an error consistent model according to Theorem 4.1 with a local decomposition $\hat{\Psi}(x) \equiv \sum_{c=1}^{C} \hat{\psi}_c(x) = \hat{f}(x)$, e.g. its Shapley decompositions (6). Applying the Shapley regression (16), $\hat{\Psi}$ is unbiased with respect to Ψ^* if and only if $\hat{\beta}_c^S = 1$, $\forall c \in \{1, \dots, C\}$. Particularly, there exists a minimal m_0 for which $\hat{\beta}_c = 1$, $\forall c \in \{1, \dots, C\}$ at a chosen confidence level. The proof is given in the Appendix.*

Theorem 4.2 implies that $\hat{\beta}^S \to 1$, as $m \to \infty$ for either Φ^S or $\hat{\Psi}$. Having Φ^S, the mapping $\Phi^S \mapsto \hat{\Psi}$ can be used to test the functional form of f. Corollary 1 extends this to local approximations of any form, i.e. for those to which Lemma 1 does not apply but a local decomposition can be formulated approximatively. For example, universal approximators can learn (regression) discontinuities (Imbens and Lemieux (2008)) as a results of treatment when given enough data. The Shapley regression framework can then be used to construct approximate parametric functional forms around a discontinuity and test the limits of their validity.

For a linear model, $\hat{\beta}_c = 1$ is nothing else as the unbiasedness of coefficients. This can be seen from Proposition 3.1 and shows again that Shapley regressions reduce to the standard case in this situation.

For a general non-linear model, unbiasedness can only be assessed if $\hat{\beta}_c = 1$, $\forall c \in \{1, \dots, C\}$ due to the accuracy condition (8) required from each decomposition $\hat{\Psi}$. The Shapley regression (16) tests linear alignment of $\hat{\Psi}$ with the dependent variable, while the level of individual components $\hat{\psi}_c$ may shift until $\hat{\beta}_c^S = 1$, $\forall c \in \{1, \dots, C\}$ for sample sizes smaller than m_0. Consistency implies that such a shift happens towards the level ψ_c^*.

This leads to the definition of *robust component estimates*: $\hat{\psi}_c$ is said to provide a robust estimation of ψ_c^* within a region Ω, if $\mathcal{H}_0^c(\Omega) : \{\beta_c^S = 0 | \Omega\}$ can be rejected and

$$\mathcal{H}_1^c(\Omega) : \{\beta_c^S = 1 | \Omega\} \quad \text{is } \textit{not} \text{ rejected} \tag{33}$$

at a chosen confidence level. That is, if the chosen confidence bounds for $\hat{\beta}_c^S$ exclude zero but include one.Alternatively, one may define an acceptable range for $\hat{\beta}_c^S$ provided $\mathcal{H}_0^c$ can be rejected, e.g. $\hat{\beta}_c^S \in [0.9, 1.1]$ admitting a small amount of bias.

Both conditions are necessary to guarantee meaningful information content in $\hat{\psi}_c$. This can be seen by considering a linear model with a pure noise variable. The best least-squares fit will return $\hat{\beta}_c^S = 1$ by construction, but $\mathcal{H}_0^c$ is almost certain not to be rejected. The practicality of robust component estimates is that they can provide useful information despite a failing test for a model being unbiased according to Theorem 4.2.

Changes between different points in a region Ω are independent from the actual level of $\hat{\psi}_c$ if the model and target are well aligned, i.e. $\mathcal{H}_0^c$ can be rejected. For example, the change of ψ_c^* between two points $x_1, x_2 \in \Omega$ can be approximated by $\hat{\psi}_c = \psi_c^* + b_c$ with b_c the component bias, if $\hat{\beta}_c^S \approx 1$,

$$\begin{aligned}\Delta\psi_c^*(x_1,x_2) &\equiv \psi_c^*(x_2)-\psi_c^*(x_1)=\hat{\beta}_c^S\big(\hat{\psi}_c(x_2)+b_c\big)-\hat{\beta}_c^S\big(\hat{\psi}_c(x_1)+b_c\big) \\ &\approx \hat{\psi}_c(x_2)-\hat{\psi}_c(x_1)=\Delta\hat{\psi}_c(x_1,x_2)\,. \end{aligned} \tag{34}$$

Connecting this with the consistency of universal approximators, condition (24) affects the asymptotic behaviour of $\hat{\beta}^S$, $\sqrt{m}(\hat{\beta}^S-\beta^S)\rightarrow_p \mathcal{N}(0,\mathcal{I}^{-1}(\hat{\beta}^S,\hat{\beta}^S))$, as $m\rightarrow\infty$. If $\xi_{ml}<\frac{1}{2}$, this quantity diverges resulting in an asymptotically biased estimator. Practically this means that confidence intervals from $\mathcal{I}^{-1}$ do not overlap (or fail to do so beyond some values of m) with one if $\beta^S=1$. Thus, tests for the robustness of a component using $\mathcal{H}_1(\Omega)$ and model bias will fail despite the model being consistent. However, we do know $p\lim_{m\rightarrow\infty}\hat{\beta}^S=1$, meaning we can quantify the bias in $\hat{\beta}^S$ at any point, e.g. for deciding if a component estimate is sufficiently robust for practical purposes. Importantly, this does not impose restrictions on m_{test} for tests against $\mathcal{H}_0(\Omega)$. Asymptotic inference on $\mathcal{H}_0(\Omega)$ is still valid without sample splitting because $\mathcal{I}(\beta^S,\hat{\theta})=0$ if $\beta^S=0$ (Pagan (1984)), but not for other hypotheses.

We see from the above discussion that the only possible true values for β^S are $\beta^S\in\{0,1\}^n$. This can be understood as follows. If there is a relationship between the target y and x_k (or $\hat{\psi}_c$ more generally), then the true value $\beta_k^S=1$ otherwise it is $\beta_k^S=0$. Intuitively this means that either there is information in x_k to describe y or it is just noise.

5 Applications

5.1 Shapley decompositions for experimental settings

We demonstrate the working of the Shapley inference framework with universal function approximators for the estimation of (heterogeneous) treatment effects under unconfoundedness (Rubin, 1974). This has been a main focus for the use of statistical learning for the estimation of causal effects, see e.g. Athey and Imbens (2016); Chernozhukov et al. (2018, 2017); Wager and Athey (2018); Lechner (2019). We consider the potential outcomes framework where subjects either receive a treatment or not, i.e. $t_i\in\{0,1\}$, leading to either y_{i0} or y_{i1}, respectively. We assume unconfoundedness, i.e. that the treatment assignment t_i is independent of the potential outcomes y_{it} conditioned on observables z, $\{y_{i0},y_{i1}\}\perp t\,|\,z_i$. Setting $x=(t,z)$, the expected treatment effect can be written as

$$\tau(x)=\mathbb{E}\big[y_{i1}-y_{i0}\,|\,z\big]\,. \tag{35}$$

The fundamental dilemma in the potential outcomes framework is that we only observe either y_{i0} or y_{i1} but never both. Moreover, treatment effects may be heterogeneous. This can arise from various causes, such as a differentiated reaction by subgroups of the population or nonlinear reactions to the treatment, or some form of interaction with some of the covariates or unobservables. It is hard to account for all possible causes in practice, e.g. when setting up a pre-analysis plan. However, inference based on universal function approximators may be able to address some of these challenges. Provided sufficient data, these models mimic the true DGP.[19]

Shapley value decompositions and regressions can be used to make variable dependencies explicit and assess components of interest, respectively. The Shapley-Taylor expansion (14) can be chosen to particularly suit an experimental setting. Let $b = x_{train}|t = 0$ be the background dataset consisting of untreated subjects in the training data, and ϕ^S_{00} the corresponding intercept. Making use of the missingness property of Shapley values, the decomposition of model prediction $\hat{f}(x_i)$ can then be written as

$$\begin{aligned} \hat{f}(x_i) &= \phi^S_{00} + \phi^S_{i,t} + \sum_{k \neq t} \phi^S_{i,k} & (36) \\ &= \phi^S_{00} + \phi^S_{i,t} - \sum_{k \neq t} \phi_{i,kt} + 2 \sum_{k \neq t} \phi^S_{i,kt} + \sum_{k \neq t} (\phi^S_{i,k} - \phi^S_{i,kt}) & (37) \\ &= \phi^S_{00} + \overline{\phi}^S_{i,t} + \sum_{k \neq t} \overline{\phi}^S_{i,kt} + \sum_{k \neq t} \overline{\phi}^S_{i,k} \,, & (38) \end{aligned}$$

where $\overline{\phi}_{i,x}$ indicates quantities where the interaction effects $\phi_{i,kt},\ (k \neq t)$ with the treatment t are netted out. The comma separates row and column indices.

The motivation behind (38) is to separate treatment effects from confounding effects in the second sum. Assuming the latter is balanced between the treated and non-treated, the average expected treatment effect is given by

$$ATE = \mathbb{E}_x \left[\overline{\phi}^S_t + \sum_{k \neq t} \overline{\phi}^S_{kt} \;\; t = 1 \right] . \tag{39}$$

Eq. 38 additionally splits treatment effects into a "bare" component (second term) and treatment interactions (third term). The latter can be useful for identifying channels through which the treatment may work, while both terms are not limited to be linear but can have themselves any functional form covered by Corollary 1. That is, we arrive at a *treatment function* (TF) for estimating (35),

$$\hat{\tau}(x_i) = \phi^S_{00} + \overline{\phi}^S_t(x_i) + \sum_{k \neq t} \overline{\phi}^S_{kt}(x_i) + \overline{\phi}^S_z(x_i) \,. \tag{40}$$

The last term, $\overline{\phi}^S_z(x) = \sum_{k \neq t} \overline{\phi}^S_{i,k}$, collects terms unrelated to the treatment. It is expected to zero out for balanced samples when calculating the ATE. Keeping it allows to assess the effect of sample imbalances as well.

5.2 A numerical experiment

Let $x = (t, x_1, x_2)$ with $t \sim \mathcal{B}(0.5) \in \{0, 1\}$ a treatment drawn from a fair Bernoulli distribution and $x_k \sim \mathcal{N}(0, 1)$, $k \in \{1, 2\}$ covariates sampled form a standard normal distribution.

[19] We do not expect good fits in many empirical settings. However, lacking information only contributes to the irreducible component in (30), while (31) still allows to estimate f asymptotically.

We consider the DGP

$$y \;=\; f_t(x;\beta) \;=\; \beta_1 t + \beta_2 t x_1 + \beta_3 x_1 x_2 + \beta_4 + \sigma \;\equiv\; \Psi^*(x) + \sigma\,, \tag{41}$$

with $\sigma \sim \mathcal{N}\big(0, 0.1\,\sigma^2(\Psi^*)\big)$ an irreducible noise component drawn from a normal distribution centred at zero and a standard deviation of 10% of Ψ^*. The coefficients are set to $\beta = (1,1,1,0)$. Ψ^* has homogeneous (β_1) and heterogeneous (β_2) treatment components, with the ATE set by β_1. Without *a priori* knowledge of Ψ^*, the heterogeneous treatment component is challenging to estimate as x_1 does not only interact with the treatment but also with the covariate x_2.

We simulate multiple realisations of f_t for different sample sizes. On these groups of samples we then use off-the-shelf implementations of random forests (RF), support vector machines (SVM) and artificial neural networks (ANN).[20] All models only see the raw features $x_k, k \in \{t,1,2\}$, and do not have information on $\Psi^*(x)$. We draw independent samples x_{cv}, x_{train} and x_{test} for each sample size and realisations for cross-validation, training, and testing and inference, respectively. Sample sizes are equally spaced on a logarithmic scale between 10 and ten thousand, i.e. $m_q = 10^q$ for $q \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}$ with 25 realisations for each sample size.[21] The sample size dependent prediction performance of all models is evaluated out-of-sample by the root-mean-squared error on x_{test}. We decompose all model predictions on the test sets into the treatment function (40) and perform the Shapley value regression (16) using VEIN estimates (27)-(29). The treatment function takes the explicit form

$$TF_{f_t}(x_i) = \phi^S_{00} + \phi^S_t(x_i) + \phi^S_1(x_i) + \phi^S_2(x_i) + \phi^S_{t1}(x_i) + \phi^S_{t2}(x_i) + \phi^S_{12}(x_i)\,. \tag{42}$$

The components ϕ^S_1, ϕ^S_2 and ϕ^S_{t2} are spurious as they are not present in the DGP (41). Models should not learn these components, while they may show up in (42) due to imperfect learning in the presence of noise.

We next perform Shapley regressions on (42) for inference on the learned treatment functions. Instead of scaling the sample size for valid inference, we adjust the confidence intervals assuming a test sample size given by (24) using the rates of convergence of each model.[22]

The results for this exercise are summarised in Fig. 2 with the sample size on the horizontal axis in all panels. Each row corresponds to a summand in (42) . Columns correspond to models (RF, SVM and ANN). The LHS vertical axes refer to the surrogate coefficients $\hat{\beta}^S_{f_b}$ (red lines) using VEIN estimates (27)-(29). The red shaded areas are confidence intervals corresponding to $\alpha^V = 1\%$. Only three of six components of the treatment function (42) actually converge to their unbiased values $\hat{\beta}^S = 1$, namely ϕ^S_t, ϕ^S_{t1} and ϕ^S_{12}. These are precisely

[20] We use `RandomForestRegressor`, `SVR` and `MLPRegressor` from the `Python` package `scikit-learn` (Pedregosa, 2011), respectively.

[21] Cross-validation is limited to $m_{cv} \leq 10^{3.5}$ to reduce computational costs.

[22] For each m_q, we take the ratio $r_{CI} = \sqrt{df_0/df_{adj}} * [ppf_t(1-\alpha^V/2, df_{adj})/ppf_t(1-\alpha^V/2, df_0)]$, with ppf_t being the percentile function of the Student's t-distribution, df_0 the degrees of freedom of the unadjusted regression and $df_{adj} = m_q^{2\xi_{ml}} - 6$ the adjusted degrees of freedom. It is set to one if negative which can happen for small samples. The rates of convergence are $\xi_{RF} = 0.25$, $\xi_{SVM} = 0.33$ and $\xi_{NN} = 0.28$, i.e. all models converge substantially slower then $\sqrt{m}$. r_{CI} is dominated by the first factor for large samples.

the components actually present in Ψ^* (41). The sample dependence of the spurious components does not show patterns of convergence, particularly confidence intervals are largely overlapping with zero, i.e. there is no alignment with the target and, thus, no learning of these components.

However, given small sample sizes and finite noise levels, models may still attribute some fraction of output to these components. We investigate this by looking at the learned versus actual (zero for spurious components) fraction $|\Gamma^S|$ (see Eq. 18) of Shapley values attributed to each summand of the treatment function (RHS vertical axes of Fig. 2). Grey lines show the actual fractions, while cyan lines show the learned fractions for each model and summand. We observe convergence to the true predictive fraction in all cases. Particularly, the fractions attributed to spurious components tend to zero with increasing sample size.[23]

The above analysis looked at the average learning behaviour of each model. However, the treatment function (42) allows to make statements about each observation. Here it is of interest how much predicted values vary around true values despite an average alignment. Fig. 3 shows predicted versus true values for the heterogeneous treatment component ϕ^S_{t1} for all models (columns) and different sample sizes (rows). There are clear differences between models and between samples sizes. The RF seems to have difficulty of learning this term. This is not surprising, because tree-based models are not well suited for modelling smooth function which are mostly not aligned with the variable axes. On the contrary, the SVM learns an almost perfect representation of the heterogeneous treatment term despite a considerable amount of noise. One the one hand this is encouraging, as can have high fidelity in some models. One the other hand, this also warrants caution. It is mostly not clear *a priori* which model delivers the best results in a given situation.

5.3 A real-world experiment

We next look at a large-scale field experiment from Brazil (Bruhn et al., 2016).[24] The authors investigate the impact of high-school financial education on financial proficiency (FP), consumption and saving behaviour, as well as educational outcomes using a randomised control trial (RCT) involving about 900 schools and 25.000 students in six states spanning 17 months between August 2010 and December 2011. Treatment consisted of an adjusted curriculum intermeshed with concepts relevant for financial literacy in classes in selected schools. Results come from three rounds of surveys, a baseline before treatment and two follow-ups after treatment has been received. Students in treated schools showed significant higher FP, improved savings behaviour, educational outcomes and positive spillovers to their parents, but also more concerning short-term consumption behaviour through the use of expensive credit.

We would like to revisit some of the study's findings using universal function approximators

[23] ϕ^S_{12} is somewhat harder to learn across models, particularly the RF.

[24] We would like to thank the authors for making this rich dataset publicly available.

from statistical learning.[25] One strength of statistical learning techniques is their ability to cope well with high-dimensional settings in the absence of prior knowledge of the DGP. To assess this we create a baseline dataset from the raw data containing 41 potential regressors using the baseline survey. Only the treatment dummy and the FP score are taken from the first follow-up (we do not consider the second follow-up here). The FP score is our target for what follows.[26]

Fig. 4 shows averaged learning curves for the three statistical learning models and an analogue linear regression for ten iterations of cross-validation and testing for different sample sizes between ten and 10.000. Machine learning models have mostly considerably lower errors compared to the linear model, especially for small sample sizes. This can be an advantage in experimental settings which often have modest sample sizes and the collection of more observations is not possible. The loss of machine learning models also does not saturate for the RF and ANN.

We next move to inference using universal function approximators. We compute terms up to order $h = 3$ in the Shapley-Taylor expansion (14) for all models. To enable a sampling-free calculation, we take two different approaches focusing on a main set of variables $\mathcal{M}$ which explains the majority of model variation.[27]

In the first approach, the models with all predictors are used and variables not in $\mathcal{M}$ are treated as a single variable "others" substantially reducing the number of coalition sets to evaluate. In the second approach, the model is retrained on the same sample but only using predictors in $\mathcal{M}$. This substantially reduces noise from irrelevant covariates and reduces the number of coalition sets further. Learning rates of each model are taken from Fig. 4 with all covariates included for better comparison. These prescribe the number of folds needed according to (26). The background dataset is taken to be the sub-population of untreated students in the each training fold summarised using 25 k-means centroids, which considerably speeds up the computation.

[25] We again use off-the-shelf implementations of random forests (RF), support vector machines (SVM) and artificial neural networks (ANN) from the `scikit-learn` package (Pedregosa, 2011).

[26] We do not consider behavioural outcomes. The covariates are the FP base value, school dropout rate, school passing rate, education mother, education father, family receives social support (Bolsa Familia), gender (female), has PC at home, home has internet at home, has repeated school years, income, savings, has borrowed before, has outstanding debt, money managing style, makes expenditure list, does save for the future, does compare products, does negotiate price, does negotiate payment method, financial autonomy index, intention to save index, parent employed, is in arrears (bank, store, or kin), parent has savings, used credit card, used installments, used cash or debit card, log number teachers in school, log number students in school, log students per teacher and state dummies. Sortable categorical variables are given corresponding discrete values, while not sortable variables are encoded as dummies. All continuous variables are divided by their mean over the full dataset, i.e. they have a mean of one. This increses comparability and training stability. See Bruhn et al. (2016) for more details on the dataset.

[27] Treatment, FP base value, has repeated year and intention to save. The adjusted R^2 of the linear model drops by less than 1% using only these variables. This selection obviously includes some judgement. For instance, student gender was also at the boarder line to be considered an important explanatory variable and could lead to gender specific conclusions.

Tables 1 and 2 summarise Shapley regression inference results on the full treatment function (40, LHS) for approaches one and two, respectively, for all students where a full set of covariates is available.[28] The tables also show the convergence rates and the corresponding numbers of folds needed for valid inference. The middle parts refer to estimating β^S from (16) including 95% confidence intervals. The RHS parts show the components of Shapley share and mean coefficients. The α-levels refers to test against $\mathcal{H}_0$ (17) and robustness to $\mathcal{H}_1$ (33), i.e. if zero is excluded and one included in the confidence intervals.

We first consider the case of models with all covariates included in Tab. 1. We see that all main effects, i.e. those of single variables, are learned with high confidence with the exception of year-repetition by the SVM. The coefficients (β^S) for the main effects are generally in the vicinity of one indicating some alignment between model components and the FP scores. However, confidence intervals to mostly *not* include one, especially for the main treatment, such that robust learning has to be rejected. We also see that all models have considerable difficulty in learning interaction effects, especially the SVM with some very large coefficients suggesting estimation instability. These results indicate a lack of convergence, which we can assess in the Shapley regression framework. Insufficient convergence also explains the different magnitudes of Shapley share and mean coefficients. Overall, all models have difficulty to generalise to the full DGP. To remedy this, one can either include more observations or increase data quality, both difficult in experimental settings, or reduce the number of variables which requires more judgement.

The latter corresponds to our second approach, where we retrain all models on the same folds using the main variables in $\mathcal{M}$ only. Inference results are given in Tab. 2. This has the same structure but no "others" contributions and interaction terms up to order three are shown. The RF has now learned almost all components of the treatment function confidently and many also robustly. This means that its predictions are well aligned with the target and in expectation unbiased for these components. The ANN has learned most main effects robustly and some second order interaction effects. The results for the SVM again indicate numerical instability in the Shapley regression step, as almost all Shapley shares have been allocation the main effects with poor robustness results. This is contrary to the numerical simulation in the previous subsection, where the SVM performed best, highlighting the difficulty in *a priori* judging on the most suitable model.[29]

These differences demonstrate how the Shapley regression framework can be used to evaluate the learning process of individual models. It especially allows to assess individual parts of each model separately and if these can be trusted, e.g. to inform decisions if learned robustly. Comparing Shapley mean coefficients (SMC) $\bar{\Gamma}^S$ of the two most prominent pre-

[28] This selection criterion slightly increases the average FP base value. The patterns of sample differences between treated and untreated sub-populations of variables in $\mathcal{M}$ is the same between this sample and the full dataset.

[29] The SVM also has slightly different baseline values ϕ_0^S for both approaches. This can be explained by its different objective function, which is similar to a mean absolute error compared to mean squared errors for the RF and ANN.

dictors in Tab. 2, the FP base value and the treatment, we see that all models agree on their values as it should be for robust components.[30] The SMC is suitable for assessing if terms in the treatment function deviate from zero, i.e. if there is a treatment effect or not, because of its definition relative to the baseline value ϕ_0^S. Note that we do not necessarily expect comparable values across models from Shapley share coefficients (SSC) as long as not all (major) components have been learned robustly due to the normalisation in (18).

Finally, we are interested if we can extract interesting functional forms learned by different models, for instance telling us something about the mechanism how the treatment affects different sub-populations. The SSC of the RF and the ANN indicate that the interaction between the FP base value and the treatment is among the most important interactions.[31] We use degree-4 polynomials to approximate the functional forms learned by all models for this interaction and the FP base value. The results is shown in Fig. 5. We see a clear and comparable nonlinearity learned by both the RF and ANN, while the flat line for the SVM is in line with its zero SSC component in Table 2. The overall effect of this term for the RF and ANN is sizeable. It comprises about 50% from peak to trough relative to the average main treatment effect. Taking this at face value indicates that treatment was most effective for students with pre-existing financial knowledge and much less so for a priori less knowledgeable students. This may suggest different curricula based on existing knowledge for more effective treatment.

6 Conclusion

We proposed a generic statistical inference framework based on universal function approximators from statistical learning. It consists of a two-step approach. First, the decomposition of a model into its Shapley-Taylor expansion - a solution concept from cooperative game theory - is used to make well-defined attributions, e.g. for variable components or their interactions. In the second step, a linear regression of the target on these Shapley values - or, a Shapley regression - allows for standard testing of hypotheses.

The major appeal of the presented framework is its universality. It applies to most functions and dataset imposing only light conditions. It can be applied whenever a models is error consistent and sufficient data for inference are available. We presented practical approaches and solutions to challenges arising from the framework. For instance the often slow convergence of conventional machine learning models requires the use of cross-fitting for valid inference. However, this technique is amenable to standard procedures training and testing machine learning models. We developed a set of summary statistics, particularly Shapley share and mean coefficients, which allow for a similar communication of results based on universal approximators as for a conventional linear regression analysis.

[30] The SVM is again a slight outlier with the FP base value only being almost learned robustly judged as the confidence interval does not overlap with one.

[31] Note that the RF also robustly learned a set of treatment triple interactions.

The Shapley regression framework also provides insights quantifying statistical learning processes. For example the state of convergence and generalisation of a learning algorithm can be tested by assessing the robustness of individual Shapley component estimates. Such insights have the potential to assist in the practical and theoretical development of new algorithms. From a practical point of view, this allows to analyse and use "parts of a model" and to quantify differences in learning between different models.

The main difficulties for applying Shapley regressions can be computational. Both, the precise computation of Shapley decompositions for large variable sets as well as cross-fitting for slowly converging models can be challenging. We presented approaches of how to address some of these issues, while the use of approximations and high-performance computing may be necessary in some settings. We do not believe that these are insurmountable given the quick proliferation of accessible computing resources and the skills to operate them.

We showcased the proposed framework using simulated and real-world experimental settings with randomised controls. Particularly, the choice of a suitable background dataset for the calculation of Shapley value attributions - the sub-population of non-treated in the training data - allowed us to derive an explicit treatment function. This function can account for potentially nonlinear treatment effects and complex treatment interactions. That is, it allows a detailed assessment of the heterogeneity of treatment effects. Such insights could be used to help to generalise information derived from RCTs, e.g by identifying channels through which treatment acts potentially addressing some of the challenges faced by RCT (Deaton and Cartwright, 2018).

More generally, the Shapley value and regression framework provides a set of tools which builds a bridge between the two cultures of statistical modelling (Breiman, 2001). It brings the use of machine learning models in line with that of standard statistical techniques regularly used to inform decisions. The main trade-off faced by the modeller is between lighter assumptions on the DGP and mostly more accurate models on the one side, and a more involved two-step procedure for inference on the other. Provided the potential benefits of machine learning models evidenced by rapid advances in artificial intelligence, this trade-off is likely to be favourable in many, but not all, situations.

Appendix

Box 1: Statistical inference recipe for machine learning models

1. **Cross-validation, training and testing** of a model $\hat{f}$, e.g. RF, SVM, ANN, etc., with K-fold cross-fitting according to Eq. 26.

2. **Model decomposition**

   (a) Shapley value decomposition $\Phi^S(\hat{f})$ [Eq. 6] or Shapley-Taylor expansion [Eq. 14] on the test set with suitable background dataset b.

   (b) (if any) Test of assumptions and approximations made to calculate Φ^S

   (c) (optional) Mapping of Φ^S to parametric functional form $\hat{\Psi}\big(\Phi^S\big)$ [see Section 4.2]

3. **Model inference**

   (a) Shapley regression [Eq. 16] with appropriate standard errors

   $$y_i \;=\; \phi_0^S + \sum_{k=1}^{n} \phi_k^S(\hat{f}, x_i)\hat{\beta}_k^S + \hat{\epsilon}_i \;\equiv\; \Phi^S(x_i)\hat{\beta}^S + \hat{\epsilon}_i$$

   (if needed) Replace Φ^S with $\hat{\Psi}$ in case of 2(b); use VEINs (27)-(29) (Chernozhukov et al., 2017).

   (b) Assessment of model bias and component robustness based on $\hat{\beta}^S$ over a region Ω of the input space:

   Robustness (component): $\mathcal{H}_0^k : \{\beta_k^S = 0|\Omega\}$ rejected and
   $\mathcal{H}_1^k : \{\beta_k^S = 1|\Omega\}$ not rejected for single $k \in \{1, \ldots, n\}$

   Unbiasedness (model): $\mathcal{H}_k^1 : \{\beta_k^S = 1|\Omega\}$ not rejected $\forall\, k \in \{1, \ldots, n\}$

   (c) Calculate Shapley share or mean coefficients $\Gamma^S/\bar{\Gamma}^S(\hat{f}, \Omega)$ [Eq. 18 / 20] and appropriate measures of variation, e.g. Eq. 22.

Figures and Table

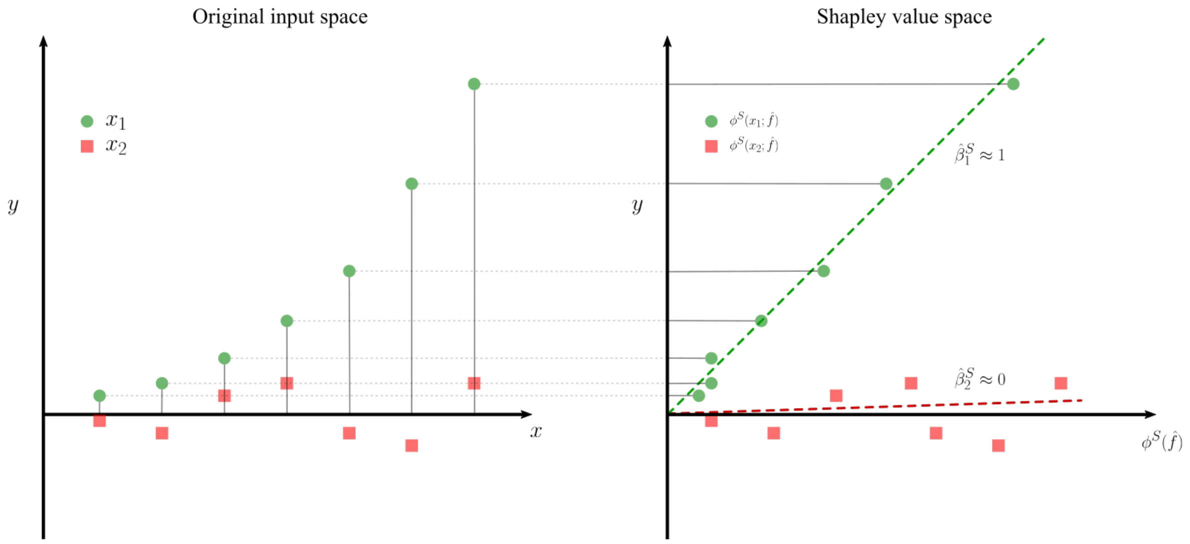

Figure 1: The principle behind Shapley regression (16). Shapley values project unknown but learned functional forms (LHS) into a linear space (RHS), where hypotheses on β^S can be tested. Variable x_1 contributes nonlinearly and significantly towards explaining the target y ($\hat{\beta}_1^S \approx 1$), while x_2 adds mostly noise ($\hat{\beta}_2^S \approx 0$).

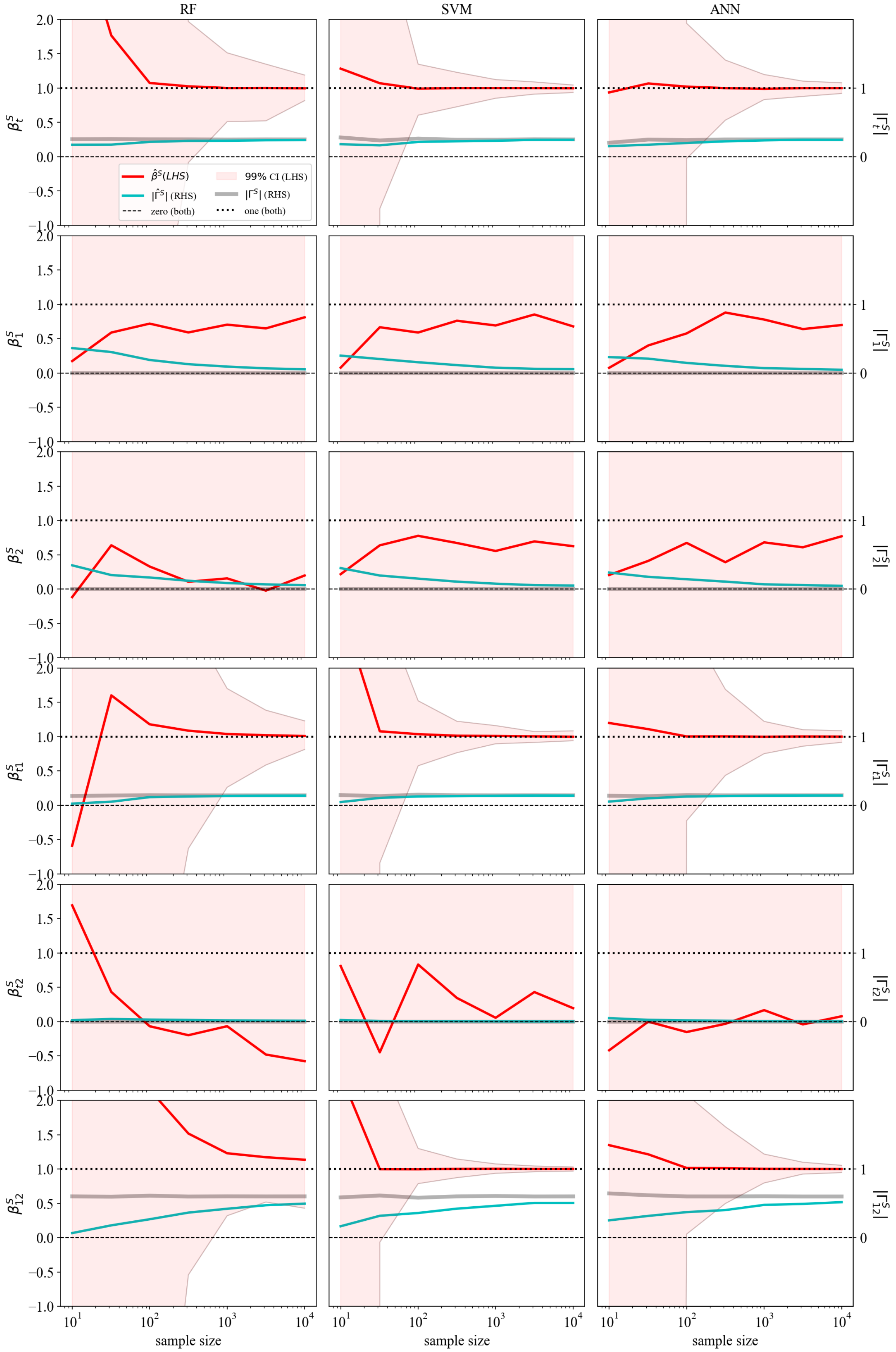

Figure 2: Inference analysis on simulated DGP (41) using RF, SVM and ANN (columns) for learning real and spurious components (row). LHS axes: Shapley regression coefficients β^S and 99% confidence intervals. RHS axes: Real (grey) and learned (cyan) Shapley shares $|\Gamma^S|$. Source: Author's calculation.

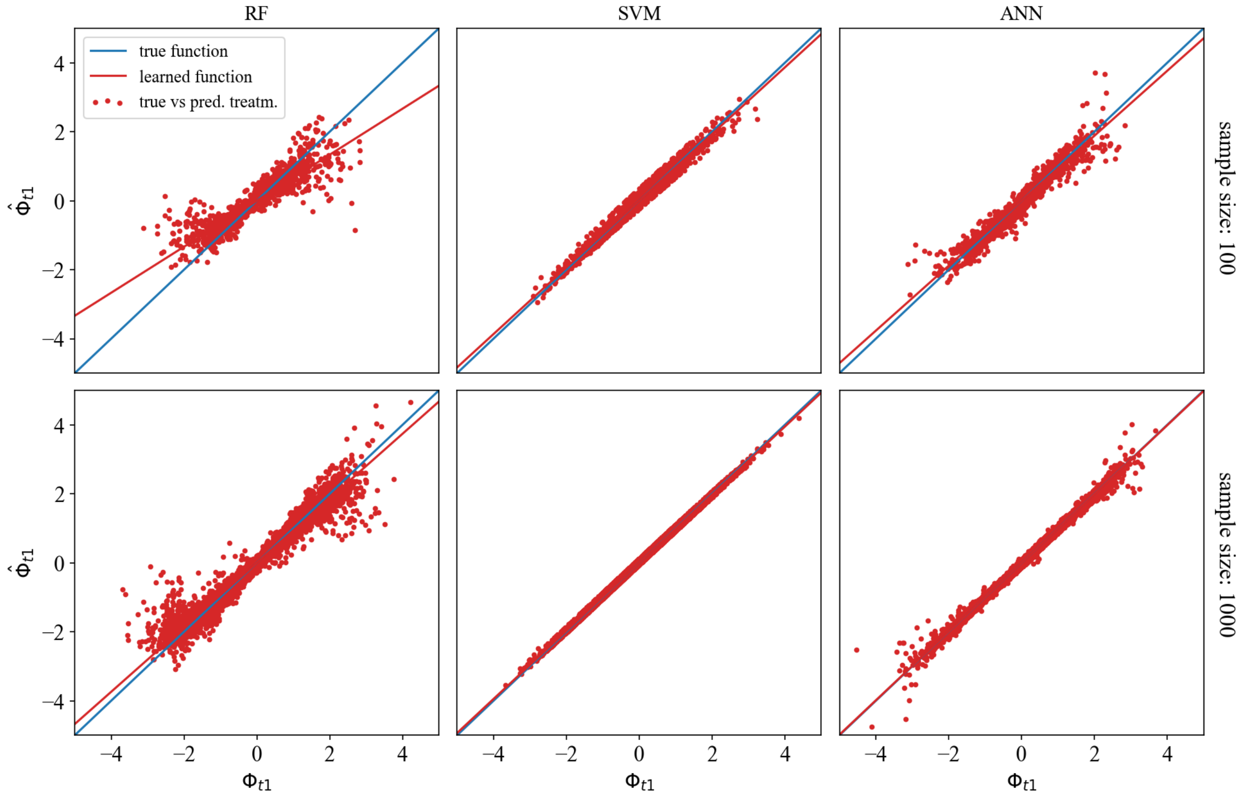

Figure 3: Actual versus learned treatment interaction effect ϕ_{t1} of (41) for RF, SVM and ANN (columns) and for different sample sizes: 100 (upper row), 1000 (lower row). Source: Author's calculation.

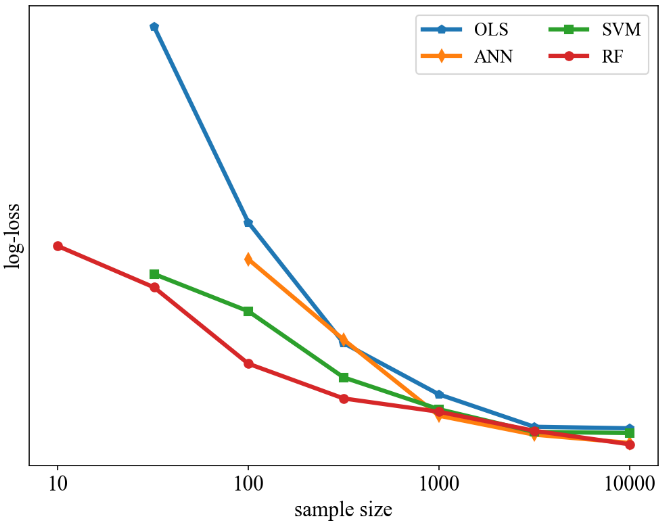

Figure 4: Learning curves for different models. Log loss is the logarithm of the root mean squared error. Only sections of decreasing test error are shown. Source: Bruhn et al. (2016) and author's calculation.

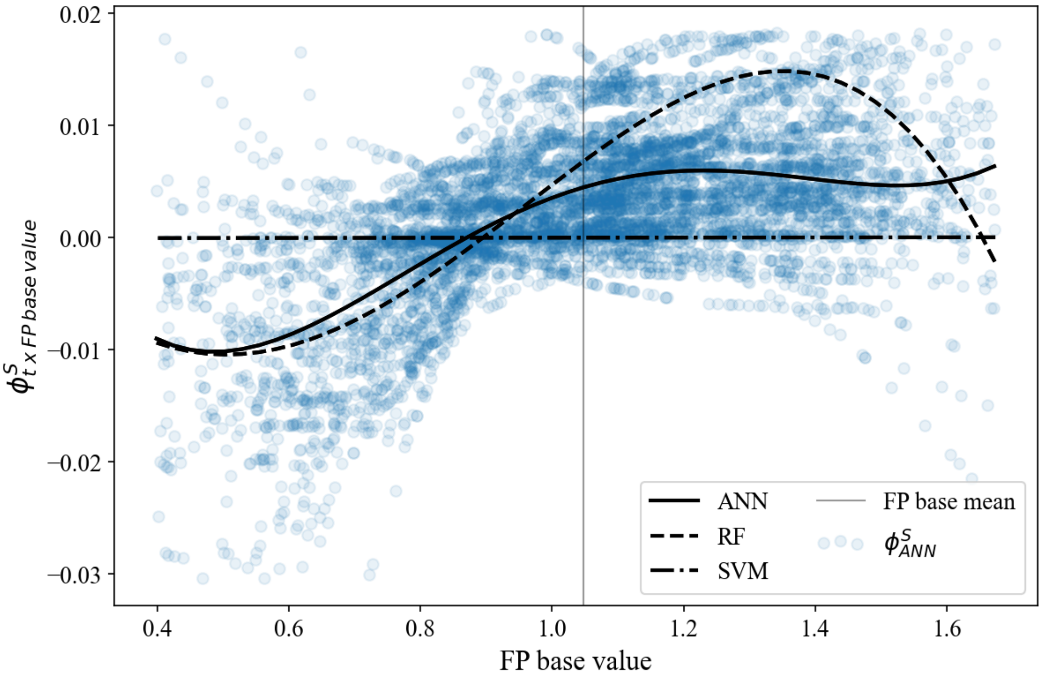

Figure 5: Estimated treatment function learned by different models for the interaction between the treatment and FP base value plotted against the base value. Curves are best-fit order-4 polynomials for the treated. Blue circles show individual Shapley-Talor index values for that term for the ANN (inner 98% of vertical span shown). Source: Bruhn et al. (2016) and author's calculation.

RF

	$\hat{\beta}^S$	$se(\hat{\beta}^S)$	$p_{\mathcal{H}_0}$	$\hat{\beta}^S_{2.5\%}$	$\hat{\beta}^S_{97.5\%}$	sign	$\lvert\Gamma^S\rvert$	$\bar{\Gamma}^S$	α-level	robust
treatment	1.388	0.146	0.000	1.102	1.675	+	0.051	0.016	***	
FP base value	1.054	0.018	0.000	1.019	1.09	+	0.381	0.013	***	
has repeated year	1.687	0.292	0.000	1.112	2.262	-	0.019	-0.005	***	
intention to save	1.127	0.174	0.000	0.786	1.468	+	0.043	-0.007	***	x
others	0.722	0.080	0.000	0.565	0.88	n.a.	0.123	-0.025	***	
treatment - FP base value	0.441	0.089	0.000	0.265	0.616	+	0.037	0.010	***	
treatment - has rep. year	-0.375	0.459	0.586	-1.277	0.527	-	0.005	0.002		
treatment - int. to save	0.672	0.175	0.000	0.328	1.016	+	0.022	0.007	***	x
treatment - others	0.905	0.122	0.000	0.666	1.145	n.a.	0.038	0.010	***	x
FP base value - has rep. year	0.595	0.245	0.016	0.113	1.077	-	0.011	0.000	**	x
FP base value - int. to save	0.072	0.122	0.555	-0.168	0.312	+	0.035	0.004		
FP base value - others	0.496	0.066	0.000	0.366	0.627	n.a.	0.090	0.01	***	
has rep. year - int. to save	-0.974	1.077	0.633	-3.091	1.144	-	0.002	-0.000		
has rep. year - others	-0.679	0.566	0.769	-1.791	0.433	n.a.	0.006	0.001		
int. to save - others	-0.442	0.306	0.851	-1.044	0.160	n.a.	0.024	0.002		
nbr. obs.: 13.874	model FE :	yes	ξ :	0.187		K :	393	ϕ^S_{00} :	0.994	

SVM

	$\hat{\beta}^S$	$se(\hat{\beta}^S)$	$p_{\mathcal{H}_0}$	$\hat{\beta}^S_{2.5\%}$	$\hat{\beta}^S_{97.5\%}$	sign	$\lvert\Gamma^S\rvert$	$\bar{\Gamma}^S$	α-level	robust
treatment	1.206	0.063	0.000	1.083	1.329	+	0.156	0.030	***	
FP base value	1.008	0.019	0.000	0.970	1.046	+	0.535	0.008	***	x
has repeated year	-6.979	1.184	1.000	-9.307	-4.650	-	0.008	0.001		
intention to save	2.160	0.287	0.000	1.596	2.723	+	0.051	-0.000	***	
others	0.132	0.033	0.000	0.066	0.197	n.a.	0.241	0.005	***	
treatment - FP base value	-966	-966	0.723	-966	-966	+	0.000	-0.000		
treatment - has rep. year	-25.169	28.113	0.629	-80.456	30.118	-	0.000	0.000		
treatment - int. to save	-229	-229.4	0.514	-229	-229	+	0.000	-0.000		
treatment - others	5.924	2.335	0.012	1.333	10.515	n.a.	0.002	-0.000	**	
FP base value - has rep. year	69.515	36.964	0.061	-3.177	70	-	0.000	0.000	*	
FP base value - int. to save	511	511.4	0.017	93.159	511	+	0.000	0.000	**	
FP base value - others	-0.418	2.606	0.873	-5.544	4.707	n.a.	0.002	0.000		
has rep. year - int. to save	65.738	81.717	0.422	-94.965	66	-	0.000	0.000		
has rep. year - others	-0.406	0.988	0.681	-2.348	1.536	n.a.	0.004	-0.000		
int. to save - others	-2.703	10.212	0.791	-22.786	17.38	n.a.	0.001	0.000		
nbr. obs.: 13.874	model FE :	yes	ξ :	0.191		K :	361	ϕ^S_{00} :	1.023	

ANN

	$\hat{\beta}^S$	$se(\hat{\beta}^S)$	$p_{\mathcal{H}_0}$	$\hat{\beta}^S_{2.5\%}$	$\hat{\beta}^S_{97.5\%}$	sign	$\lvert\Gamma^S\rvert$	$\bar{\Gamma}^S$	α-level	robust
treatment	0.883	0.055	0.000	0.774	0.992	+	0.152	0.037	***	
FP base value	1.124	0.019	0.000	1.087	1.161	+	0.410	0.008	***	
has repeated year	1.912	0.361	0.000	1.195	2.629	-	0.014	-0.002	***	
intention to save	0.842	0.187	0.000	0.471	1.213	+	0.059	-0.000	***	x
others	0.043	0.020	0.035	0.003	0.083	n.a.	0.306	-0.005	**	
treatment - FP base value	-0.450	0.279	0.890	-1.004	0.104	+	0.003	-0.001		
treatment - has rep. year	1.276	2.836	0.654	-4.352	6.904	-	0.001	0.000		
treatment - int. to save	0.543	1.565	0.729	-2.563	3.650	+	0.001	-0.000		
treatment - others	-0.174	0.158	0.727	-0.487	0.139	n.a.	0.010	-0.000		
FP base value - has rep. year	0.484	0.967	0.618	-1.434	2.402	-	0.001	0.000		
FP base value - int. to save	0.540	0.515	0.296	-0.481	1.562	+	0.002	-0.000		
FP base value - others	-0.002	0.149	0.988	-0.297	0.292	n.a.	0.017	-0.000		
has rep. year - int. to save	0.068	2.569	0.979	-5.031	5.166	-	0.000	-0.000		
has rep. year - others	0.012	0.273	0.965	-0.530	0.554	n.a.	0.006	0.000		
int. to save - others	-1.090	0.405	0.992	-1.894	-0.287	n.a.	0.007	0.000		
nbr. obs.: 13.874	model FE :	yes	ξ :	0.260		K :	98	ϕ^S_{00} :	1.008	

Table 1: Summary of Shapley regression for modelling students' financial proficiency score (FP) from first follow-up study using all variables. Only terms for the main variables and interaction terms up to order two are shown for RF (upper part), SVM (middle part) and ANN (lower part). *Others* groups all remaining variables in one group for which no clear sign is available. The number of folds K for valid inference is calculated from (26) with convergence rates ξ taken from Fig. 4. ϕ^S_{00} is the model mean predicted value for the background dataset. Significance levels: ***: 0.01, **: 0.05, **: 0.1. Source: Bruhn et al. (2016) and author's calculation.

RF

	$\hat{\beta}^S$	$se(\hat{\beta}^S)$	$p_{\mathcal{H}_0}$	$\hat{\beta}^S_{2.5\%}$	$\hat{\beta}^S_{97.5\%}$	sign	$\lvert\Gamma^S\rvert$	$\bar{\Gamma}^S$	α-level	robust
treatment	1.018	0.052	0.000	0.916	1.120	+	0.106	0.032	***	x
FP base value	0.993	0.013	0.000	0.967	1.020	+	0.415	0.035	***	x
has repeated year	1.256	0.118	0.000	1.024	1.487	-	0.087	0.009	***	
intention to save	0.929	0.087	0.000	0.757	1.101	+	0.072	0.002	***	x
treatment - FP base value	0.797	0.071	0.000	0.658	0.936	+	0.034	0.003	***	
treatment - has rep. year	0.749	0.358	0.037	0.046	1.453	-	0.008	-0.000	**	x
treatment - int. to save	0.220	0.236	0.353	-0.245	0.684	+	0.013	-0.000		
FP base value - has rep. year	0.510	0.169	0.003	0.179	0.842	-	0.045	-0.001	***	
FP base value - int. to save	0.864	0.071	0.000	0.725	1.003	+	0.089	-0.001	***	x
has rep. year - int. to save	0.906	0.334	0.007	0.249	1.562	-	0.021	-0.000	***	x
treatment-FP base value - has rep. year	0.618	0.140	0.000	0.343	0.893	-	0.024	-0.000	***	
treatment-FP base value - int. to save	0.936	0.077	0.000	0.784	1.088	+	0.039	0.001	***	x
treatment-has rep. year - int. to save	1.038	0.333	0.002	0.384	1.692	-	0.010	0.001	***	x
FP base value - has rep. year - int. to save	0.798	0.168	0.000	0.467	1.128	-	0.036	0.000	***	x
nbr. obs.: 13.874	model FE :	yes	ξ :	0.187		K :	393	ϕ^S_{00} :	0.946	

SVM

	$\hat{\beta}^S$	$se(\hat{\beta}^S)$	$p_{\mathcal{H}_0}$	$\hat{\beta}^S_{2.5\%}$	$\hat{\beta}^S_{97.5\%}$	sign	$\lvert\Gamma^S\rvert$	$\bar{\Gamma}^S$	α-level	robust
treatment	1.069	0.075	0.000	0.922	1.216	+	0.162	0.031	***	x
FP base value	0.894	0.021	0.000	0.852	0.936	+	0.591	0.036	***	
has repeated year	0.677	0.130	0.000	0.422	0.933	-	0.144	0.012	***	
intention to save	0.573	0.151	0.000	0.276	0.871	+	0.102	0.002	***	
treatment - FP base value	-2.557	-2.6	0.992	-3	-3	+	0.000	0.000		
treatment - has rep. year	-717	-717.1	1.000	-717	-717	-	0.000	-0.000		
treatment - int. to save	-1753	-1753	1.000	-1753	-1753	+	0.000	-0.000		
FP base value - has rep. year	57.141	41.619	0.171	-24.705	57	-	0.000	0.000		
FP base value - int. to save	335	334.7	0.010	79.448	335	+	0.000	0.000	**	
has rep. year - int. to save	173	74.417	0.020	27.131	173	-	0.000	0.000	**	
treatment - FP base value - has rep. year	702800	702800	0.728	702800	702800	-	0.000	0.000		
treatment - FP base value - int. to save	3.243e+06	3.243e+06	0.424	3.243e+06	3.243e+06	+	0.000	-0.000		
treatment - has rep. year - int. to save	1.139e+06	1.139e+06	0.002	1.139e+06	1.139e+06	-	0.000	-0.000	***	
FP base value - has rep. year - int. to save	45560	45560	0.952	45560	45560	-	0.000	0.000		
nbr. obs.: 13.874	model FE :	yes	ξ :	0.191		K :	361	ϕ^S_{00} :	0.960	

ANN

	$\hat{\beta}^S$	$se(\hat{\beta}^S)$	$p_{\mathcal{H}_0}$	$\hat{\beta}^S_{2.5\%}$	$\hat{\beta}^S_{97.5\%}$	sign	$\lvert\Gamma^S\rvert$	$\bar{\Gamma}^S$	α-level	robust
treatment	0.996	0.050	0.000	0.898	1.095	+	0.161	0.035	***	x
FP base value	1.009	0.014	0.000	0.981	1.037	+	0.516	0.031	***	x
has repeated year	1.005	0.092	0.000	0.822	1.187	-	0.137	0.010	***	x
intention to save	1.048	0.114	0.000	0.821	1.275	+	0.091	0.001	***	x
treatment - FP base value	0.915	0.294	0.002	0.332	1.498	+	0.013	0.001	***	x
treatment - has rep. year	-0.068	0.337	0.842	-0.737	0.602	-	0.009	0.000		
treatment - int. to save	1.369	0.712	0.057	-0.044	2.782	+	0.004	0.000	*	
FP base value - has rep. year	0.941	0.125	0.000	0.693	1.188	-	0.027	0.003	***	x
FP base value - int. to save	-0.250	0.581	0.668	-1.403	0.904	+	0.010	0.001		
has rep. year - int. to save	-0.019	0.395	0.961	-0.804	0.765	-	0.012	0.001		
treatment - FP base value - has rep. year	0.424	0.440	0.338	-0.450	1.297	-	0.006	-0.000		
treatment - FP base value - int. to save	-0.888	0.732	0.772	-2.341	0.564	+	0.004	-0.000		
treatment - has rep. year - int. to save	0.274	1.091	0.802	-1.891	2.439	-	0.004	-0.000		
FP base value - has rep. year - int. to save	-0.696	0.725	0.660	-2.136	0.744	-	0.007	-0.000		
nbr. obs.: 13.874	model FE :	yes	ξ :	0.260		K :	98	ϕ^S_{00} :	0.944	

Table 2: Summary of Shapley regression for modelling students' financial proficiency score (FP) from first follow-up study using only the main variables. Interaction terms up to order three are shown for RF (upper part), SVM (middle part) and ANN (lower part). The number of folds K for valid inference is calculated from (26) with convergence rates ξ taken from Fig. 4. ϕ^S_{00} is the model mean predicted value for the background dataset. Significance levels: ***: 0.01, **: 0.05, **: 0.1. Source: Bruhn et al. (2016) and author's calculation.

Proofs

Proof of Proposition 2.1

Without loss of generality, we can write $\hat{f}$ in terms of linear and non-linear components, $\hat{f}(x) = \hat{f}_l(x) + \hat{f}_{nl}(x)$, e.g. using a Taylor expansion. The Shapley decomposition can then be written as

$$\Phi^S(x) = \sum_{k=0}^{n} \phi_k^S = f(x) = \hat{f}_l(x) + \hat{f}_{nl}(x) = \hat{f}_l(x) = x\hat{\beta}. \tag{43}$$

The first step follows from local accuracy and the third from the assumption of linearity. Properties (1)-(5) can be easily verified. □

Proof of Proposition 2.2

Without loss of generality, we can consider a function of either only one variable or only variable coalitions of one variable k, i.e. $x' = \emptyset$ on the LHS of (10). Then,

$$f - f' = \phi_0^S(f) + \phi_k^S(f) - \phi_0^S(f') - \phi_k^S(f) \overset{!}{\geq} 0\,. \tag{44}$$

We need $\phi_0^S(f) \geq \phi_0^S(f')$ for this inequality to hold. Since must be true for the negative of both functions and changing their order for strong monotonicity to hold, we arrive at $\phi_0^S(f) = \phi_0^S(f')$. □

Proof of Proposition 3.1

Without loss of generality, we can again write $\hat{f}$ in terms of a linear and a non-linear component, $\hat{f}(x) = \hat{f}_l(x) + \hat{f}_{nl}(x)$. The Shapley regression can then be written as

$$\Phi^S(x)\hat{\beta}^S = \sum_{k=0}^{n} \phi_k^S \hat{\beta}_k^S = \sum_{k=0}^{n} (\phi_{k,l}^S + \phi_{k,nl}^S)\hat{\beta}_k^S = \phi_l^S(x)\hat{\beta}^S = \hat{f}_l(x)\hat{\beta}^S = x\,diag(\hat{\beta})\hat{\beta}^S = x\hat{\beta}. \tag{45}$$

The last two steps follows from Proposition 1 and the uniqueness of the coefficients $\hat{\beta}$ as solution to the convex least-squared problem. This can be made explicit for the OLS estimator. By setting $x \to x\,diag(\hat{\beta}) \equiv xD_{\hat{\beta}}$, one obtains

$$\hat{\beta}^S = \frac{xD_{\hat{\beta}}y}{(xD_{\hat{\beta}})^T(xD_{\hat{\beta}})} = \frac{D_{\hat{\beta}}}{D_{\hat{\beta}}^2}\frac{Xy}{x^Tx} = D_{\hat{\beta}}^{-1}\hat{\beta} = 1_{n+1}\,. \tag{46}$$

□

Proof of Theorem 4.1

I provide proofs for analytic and non-analytic models, reflecting prominent model types from machine learning. Analytic models $\hat{f}(x, \theta)$ are differentiable almost everywhere (ANN and SVM in our case).

Proof. (*analytic models)*: Let $\hat{f}(\theta, x)$ be a function of inputs $x \in \mathbb{R}^n$ and parameters $\theta \in \mathbb{R}^q$, which is $(d' + 1)$ times differentiable, where d' is the degree of the highest polynomial $p^{d'}(x)$

of the DGP $f(\beta, x)$, such that the Taylor expansion of $\hat{f}$ exists. Then, there exists an open interval $\Omega \subset \mathbb{R}^n$ where the difference between f and $\hat{f}$ is error consistent for each $x' \in \Omega$ around a. Namely,

$$f - \hat{f}\;_{\Omega}(x') = \sum_{k=0}^{n} \big(\beta_k - \hat{\beta}_k\big) p_k^d(x' - a) + R\big(\hat{f}^{(d'+1)}(c), (x' - a)^{(n-d)}\big)\,. \tag{47}$$

That is, the polynomial expansion of $\hat{f}$ around a will be functionally identical to f up to a residual R with c between x' and a. By assumption, (47) vanishes with increasing sample size, from which follows $(\beta - \hat{\beta}, R) \to 0$, as $m \to \infty$. □

Second, non-analytic models are tree-based models with $\hat{f}(x) \equiv T_{x_\theta}(x) = \langle x_\theta \rangle_x$, with T_{x_θ} describing the set leaf nodes of the model from training. Usually, $|T| \to \infty$, as $|x_{train}| \to \infty$. Examples are classification trees, random forests or extreme trees (Geurts et al. (2006)). The main difference to analytic models is that tree-based models are not differentiable. However, many tree-based models are based on bagging (e.g. forests) which smoothens model output (Bühlmann and Yu (2002)).

Proof. (*non-analytic tree-based models)*: Let $x' \in \Omega \subseteq D$, where D is the domain of f and Ω is the leaf node region of x', with $|\Omega|$ being the number of x_θ in this region. The difference between f and $\hat{f}$ can then be written as

$$\begin{aligned} f - \hat{f}\;_{\Omega}(x') &= \sum_{k=1}^{n} \beta_k p_k^d(x') - \frac{1}{|\Omega|} \sum_{j=1}^{|\Omega|} \hat{\beta}_k p_k^d(x_j) \\ &= \frac{1}{|\Omega|} \sum_{j=1}^{|\Omega|} \Big(\sum_{k=1}^{n} \big(|\Omega| \beta_k p_k^d(x') - \hat{\beta}_k p_k^d(x_j) \big) \Big) \\ &= \sum_{k=1}^{n} p_k^d(x') \big(\beta_k - \hat{\beta}_k\big)\,. \end{aligned} \tag{48}$$

We used the model optimising condition that values x' fall into leave nodes with the same expected value, i.e. $\langle x_j \rangle_\Omega = x'$ in the limit $m \to \infty$. The above expression can then only vanish if $\beta - \hat{\beta} \to 0$ as $m \to \infty$. □

Proof of Corollary 4.1

For each $\epsilon > 0$ there is a neighbourhood $\mathcal{B}_\delta$ of radius $\delta > 0$ around every $x' \in \Omega$, such that $|f - \hat{f}|_{x'} < \epsilon$. For each $\mathcal{B}_\delta$ and ϵ, there will be an large enough n' such that there exist $\delta' \leq \delta$ and ϵ' and ϵ'' with $\epsilon' + \epsilon'' < \epsilon$ for which $|f - \hat{f}|_{x'} < \epsilon'$ and $|\hat{f} - \hat{f}_p|_{x'} < \epsilon''$. The conclusion follows form the assumption of error consistency. □

Proof of Theorem 4.2

The second part is a consequence of error consistency. For the first part, it is enough to show that the difference between $\hat{\Psi}$ and $\Psi^{\star}$ vanishes beyond n_u. Here,

$$0 = \Psi^{*} - \hat{\beta}^{S}\hat{\Psi} = \sum_{c=1}^{C} \psi_c^{*} - \sum_{c=1}^{C} \hat{\beta}_c^{S}\hat{\psi}_c = \sum_{c=1}^{C} \psi_c^{*} - \hat{\psi}_c \,. \tag{49}$$

The RHS of Eq. 49 only vanishes if $\psi_c^{*} = \hat{\psi}_c$, $\forall c \in \{1, \ldots, C\}$. □

References

Adams, Terrence, and Andrew Nobel. 2010. “Uniform convergence of Vapnik–Chervonenkis classes under ergodic sampling.” *Annals of Probability*, 38(4): 1345–1367.

Agarwal, Ashish, Kedar Dhamdhere, and Mukund Sundararajan. 2019. “A New Interaction Index inspired by the Taylor Series.” *arXiv e-prints*, 1902.05622.

Andoni, Alexandr, Rina Panigrahy, Gregory Valiant, and Li Zhang. 2014. “Learning Polynomials with Neural Networks.” Vol. 32 of *Proceedings of Machine Learning Research*, 1908–1916.

Athey, Susan, and Guido Imbens. 2016. “Recursive partitioning for heterogeneous causal effects.” *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.

Biau, Gérard. 2012. “Analysis of a Random Forests Model.” *Journal of Machine Learning Research*, 13(1): 1063–1095.

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author).” *Statistical Science*, 16(3): 199–231.

Bruhn, Miriam, Luciana de Souza Leão, Arianna Legovini, Rogelio Marchetti, and Bilal Zia. 2016. “The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil.” *American Economic Journal: Applied Economics*, 8(4): 256–95.

Bühlmann, Peter, and Bin Yu. 2002. “Analyzing bagging.” *The Annals of Statistics*, 30(4): 927–961.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. “Double/debiased machine learning for treatment and structural parameters.” *The Econometrics Journal*, 21(1): C1–C68.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2017. “Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments.” *arXiv e-prints*, 1712.04802.

Christmann, Andreas, and Ingo Steinwart. 2008. *Support Vector Machines.* Srpinger.

Cybenko, George. 1989. “Approximation by Superpositions of a Sigmoidal Function.” *Mathematics of Control, Signals, and Systems*, 2: 303–314.

Deaton, Angus, and Nancy Cartwright. 2018. “Understanding and misunderstanding randomized controlled trials.” *Social Science & Medicine*, 210: 2–21.

Domingos, Pedro. 2000. "A Unified Bias-Variance Decomposition and its Applications." *In Proc. 17th International Conf. on Machine Learning*, 231–238.

Farago, Andras, and Gabor Lugosi. 1993. "Strong universal consistency of neural network classifiers." *IEEE Transactions on Information Theory*, 39(4): 1146–1151.

Fernandez-Delgado, et al. 2014. "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research*, 15: 3133–3181.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. "Neural Networks and the Bias/Variance Dilemma." *Neural Computation*, 4(1): 1–58.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. "Extremely Randomized Trees." *Machine Learning*, 63(1): 3–42.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press.

Imbens, Guido, and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics*, 142(2): 615–635.

Lechner, Michael. 2019. "Modified Causal Forests for Estimating Heterogeneous Causal Effects." CEPR Discussion Paper 13430.

Lipovetsky, Stan, and Michael Conklin. 2001. "Analysis of regression in game theory approach." *Applied Stochastic Models in Business and Industry*, 17(4): 319–330.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30.* 4765–4774.

Lundberg, Scott, Gabriel Erion, and Su-In Lee. 2018. "Consistent Individualized Feature Attribution for Tree Ensembles." *ArXiv e-prints*, 1802.03888.

Meir, Ron. 2000. "Nonparametric Time Series Prediction Through Adaptive Model Selection." *Machine Learning*, 39(1): 5–34.

Mohri, Mehryar, and Afshin Rostamizadeh. 2010. "Stability Bounds for Stationary phi-mixing and beta-mixing Processes." *Journal of Machine Learning Research*, 11: 789–814.

Pagan, Adrian. 1984. "Econometric Issues in the Analysis of Regressions with Generated Regressors." *International Economic Review*, 25(1): 221–47.

Pedregosa, F. et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12: 2825–2830.

Rubin, Donald. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5): 688–701.

Scornet, Erwan, Gérard Biau, and Jean-Philippe Vert. 2014. "Consistency of random forests." *ArXiv e-prints*, 1405.2881.

Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2010. "Learnability, Stability and Uniform Convergence." *Journal of Machine Learning Reseach*, 11: 2635–2670.

Shapley, Lloyd. 1953. "A value for n-person games." *Contributions to the Theory of Games*, 2: 307–317.

Steinwart, Ingo. 2002. "Support Vector Machines are Universally Consistent." *Journal of Complexity*, 18(3): 768 – 791.

Steinwart, Ingo, and Clint Scovel. 2007. "Fast rates for support vector machines using Gaussian kernels." *The Annals of Statistics*, 35(2): 575–607.

Stone, Charles. 1982. "Optimal Global Rates of Convergence for Nonparametric Regression." *The Annals of Statistics*, 10(4): 1040–1053.

Strumbelj, Erik, and Igor Kononenko. 2010. "An Efficient Explanation of Individual Classifications Using Game Theory." *Journal of Machine Learning Research*, 11: 1–18.

Sundararajan, Mukund, and Amir Najmi. 2019. "The many Shapley values for model explanation."

Vapnik, V. N. 1999. "An overview of statistical learning theory." *IEEE Transactions on Neural Networks*, 10(5): 988–999.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Young, Peyton. 1985. "Monotonic solutions of cooperative games." *International Journal of Game Theory*, 14: 65–72.

Yu, Bin. 1994. "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences." *Annals of Probability*, 22(1): 94–116.