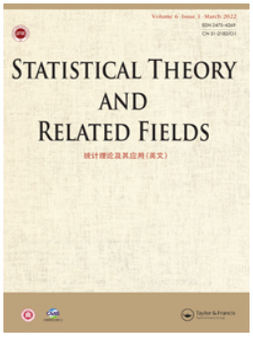

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Interpreting uninterpretable predictors: kernel methods, Shtarkov solutions, and random forests

T. M. Le & Bertrand Clarke

To cite this article: T. M. Le & Bertrand Clarke (2022) Interpreting uninterpretable predictors: kernel methods, Shtarkov solutions, and random forests, Statistical Theory and Related Fields, 6:1, 10-28, DOI: 10.1080/24754269.2021.1974157

To link to this article: https://doi.org/10.1080/24754269.2021.1974157

Published online: 08 Sep 2021.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 1, 10–28
https://doi.org/10.1080/24754269.2021.1974157

Taylor & Francis
Taylor & Francis Group

Interpreting uninterpretable predictors: kernel methods, Shtarkov solutions, and random forests

T. M. Le[a] and Bertrand Clarke[b]

[a]Department of Mathematics, Science, and Informatics, Mercer University, Atlanta, GA, USA; [b]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, USA

ABSTRACT
Many of the best predictors for complex problems are typically regarded as hard to interpret physically. These include kernel methods, Shtarkov solutions, and random forests. We show that, despite the inability to interpret these three predictors to infinite precision, they can be asymptotically approximated and admit conceptual interpretations in terms of their mathematical/statistical properties. The resulting expressions can be in terms of polynomials, basis elements, or other functions that an analyst may regard as interpretable.

ARTICLE HISTORY
Received 18 March 2020
Revised 20 April 2021
Accepted 8 August 2021

KEYWORDS
Bayes; boosting; kernel methods; random forest; Shtarkov predictor; stacking

1. Introduction

Fundamentally point prediction is an input–output relation. Given a pair of related sequences $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ where each y_i is an outcome of some Y_i, the predicted value for y_i is $\hat{y}_i = F_i(x_i)$ in which F_i is typically chosen using the earlier x_i's and y_i's, i.e., $x_{i-1}, \ldots, x_1$ and $y_{i-1}, \ldots, y_1$. An extra set of burn-in data may be used to choose F_i and there may be added complexities from side information. We may put many sorts of desiderata on the F_i's – low predictive error, simplicity, even insisting each $F_i \in \mathcal{F}_i$ for some set of functions $\mathcal{F}_i$. However, the point predictor, F_i, is a merely a function – a way to convert an input x to an output y.

Interval predictors are somewhat more complex: They give a prediction interval, say I_i with a pre-assigned probability that the event $\{Y_i \in I_i\}$ will occur. Thus they have the same input but a different output. Regardless of any further desiderata we might impose, such as optimality criteria, interval predictors (or their generalization to regional predictors) remain input–output relations. This may be regarded as an example of conformal prediction, see Vovk et al. (2005), as we discuss later in Section 5.

By contrast, modelling is a conceptually different process. In principle, a statistical modeler proposes a model, say $Y = F(x) + \epsilon$ for simplicity, to be true and has a collection of terms, say $t_j(x)$ for $j = 1, \ldots, m$, that may be part of an additive model. The modeler uses the data to choose terms and the end result is a model something like $Y(x) = \sum_{j=1}^{q} \hat{t}_j(x) + \epsilon$, where the $\hat{t}_j$'s are the same as the t_j's apart from estimating some coefficients. The modeler then asserts that the model reflects reality by ensuring that each component has a correlate in reality: Each t_j means something physical, there is a reason that the terms are added, and it has been verified that the error ϵ represents intrinsic variability rather than small terms that have been ignored i.e., bias. In this case, the model gives the point predictor $\hat{Y}(x) = \sum_{j=1}^{q} \hat{t}_j(x)$ and it is assumed that any estimates in $\hat{t}_j$ are satisfactorily close to their true values that the model is 'good' – not readily falsifiable. Note that even though $\hat{Y}$ is an estimator of F, we focus on how well it predicts Y.

What are the differences between these two approaches? First, a predictor is just a mathematical construct to match the output from a data generator (DG). It has no greater significance. All that matters about $\hat{Y}$ is how close $\hat{Y}(x)$ is to $Y = y$, i.e., how well using $\hat{Y}$ lets someone predict Y. The quality of the prediction is usually measured formally, e.g., by some cumulative error such as the prediction sum of squares

$$\text{PRESS}_n = \sum_{i=1}^{n} (y_i - \hat{Y}_{-i}(x_i))^2,$$

where the subscript $-i$ indicates that data point (x_i, y_i) was not used to form $\hat{Y}$. However, the point remains: $\hat{Y}$ predicts well, adequately, or poorly according to how we assess predictive performance. While there is nothing more that must necessarily be said, there is much that can be said – how the predictor was found, what its properties are, etc. – and some aspects of these will be discussed for the predictors here.

A clear statement of the basic problem studied here can be found in Geisser (1975) who focused on 'low structure' data and treated it strictly empirically.

CONTACT T. M. Le le_tm@mercer.edu

i.e., with minimal discussions about abstract constructs such as distributions. This is analogous to what we call complex data – data about which it is hard to make any strong assumptions. Geisser (1975)'s treatment was prescient in that he clearly expressed many of the ideas here in elementary settings such as ANOVA, regression, and posterior means. Our work goes beyond this by examining techniques that were not available in 1975 and we do so from a contemporary conceptual standpoint, not just empirically (although that is clearly no less important).

The notion of 'interpretability' used here is the same as used in Le and Clarke (2020). Briefly, a model M consists of K components, say $M = \{c_1, \ldots, c_K\}$. The c_k's are the components that go into the formulation of a model such as variables, parameters, and rules for how they are to be combined to produce a model. The model M is interpretable if and only if each c_k has a physical correlate, i.e., they correspond to some identifiable and measurable feature of the DG. We say a model is valid if and only if it is interpretable and correct at least to the degree that its predictions and future outcomes are sufficiently close.

In this setting, the key question in modelling is how well the terms in $\hat{Y}$ encapsulate the components of the DG. That is, what aspect (or c_k) of the DG does a specific t_j represent and how accurately? In short, the subcomponents of a model matter because it is hoped they have physical correlates. The model is meant to match reality in that its components can be interpreted physically in the context of the DG. Otherwise put, a model can be falsified by falsifying one of its components. Thus, it makes sense to ask if a model is 'true' – it being understood that 'true' may only be in a provisional sense. A better model may be found that discredits the earlier model and science proceeds by sequentially falsifying ever better models hopefully arriving at a model that is either not-falsifiable or so close to true (in the absolute sense) that it isn't worth the trouble to falsify. In either case, there is the idea of a model being true that has no genuine analog for predictors. The closest analog would be for a predictor to be optimal (within a class) but this is not part of measurable, objective reality.

The main link between predictors and model is that models regularly provide predictors while predictors do not in general lead to models – at least not directly. (Indeed, models that do not make measurable predictions are not valid models as they are not falsifiable.) Thus, we may speak of model-induced predictors and non-model induced predictors. Loosely, the class of predictors is very much larger than the class of models. So, one way to find good models is to find a good predictor and determine a model that performs almost as well in terms of prediction. That way, there is a physical interpretation even if some predictive power is lost.

Ideally, we want a model that is not falsified (or at least is very hard to falsify) and that gives an extremely good predictor. Sometimes this occurs but often it does not especially for complex problems. That is, a really good predictor often outperforms the predictor generated from a provisionally true model and does so by a substantial margin. In earlier work, we gave examples of this, see Le and Clarke (2020). That is, we showed how certain interpretable predictors, linear models in particular, could be modified to give improved predictions. The modification were chiefly to introduce non-interpretable features to the model induced predictor. The end result was a predictor that had some interpretable and some non-interpretable components, and gave demonstrably improved prediction over the model induced predictor. We called the difference between the error of using the partially interpretable predictors over the model induced predictors the cost of modelling. Intuitively, the flexibility from the loss of full interpretability enabled improved prediction.

Here we examine the same problem but from the reverse direction. That is, we start with uninterpretable predictors and find interpretable models that are close to them. The interpretable models do not in general perform as well as the uninterpretable predictors since the latter are the result of optimization. Thus, non-interpretable (but optimal) predictors can lead to approximate models that may be examined to see how they relate to physical components of a DG. This is another way to formalize the cost of modelling, or interpretability, in terms of the loss of predictive accuracy because the approximate model may still say something useful about the DG.

One implication of this reasoning is that the *principle of falsification* may have to be reconsidered. Falsifiability is the assertion that any conjectured model must be disprovable before it can become accepted. The problem is that if the predictions from essentially every model for a DG can be improved, then essentially every model is flawed and can be discredited. We are not actually uncovering truth. Accordingly, the *principle of falsification* may itself be 'false' in the sense that since every model is disprovable, disprovability can only be used to discriminate better models from worse ones not to arrive at a model that can be generally accepted – at least not often enough that it is useful as a foundational philosophical principle.

A further complication is the concept of $\mathcal{M}$-closed, -complete, and -open problems; see Bernardo and Smith (2000) for the original definitions. These are defined by the location of a class of proposed models or predictors relative to the DG. Here we say that an $\mathcal{M}$-closed problem is one in which the DG is exactly one of a finite list of explicit candidate models. The problem is therefore merely selecting the one that matches the data generator. In the Bayesian case, this also means that the prior is well-defined in the sense that the prior probability of a model represents the pre-experimental belief that the model is true.

An $\mathcal{M}$-complete problem is one in which the DG has a true model but it is inaccessible in some sense. For instance, it might be too complicated to formulate. There may not be any closed form expression for it, so it can only be approximated numerically. The true model might be so complicated that it is unrealistic to learn much about it from data that might realistically be gathered. Indeed, the true model may be so complex that no approximation to it is adequate even if a serviceable one can be found under restrictions. The main point is the model exists so that, for instance, expectations and convergences are well defined but any properties of it are problematic and uncertain. In this case, the prior probabilities are not that a model is true but rather that it is close to the true model given the model list. (Interpreting the prior as weights on actions in a decision theory problem is also possible; see Le and Clarke (2016b) for a brief discussion of this.)

The two problem classes contrast sharply with the $\mathcal{M}$-open problem class in which no model for the data generator can be assumed to exist. Hence, expectations and expressions related to the form of a model, e.g., modes of convergence, do not make sense and the meaning of a prior is unclear unless it is taken as a belief that a given predictor, possibly model-based, will perform better than another predictor within the class of predictors under study. In the $\mathcal{M}$-open problem class modelling makes little sense; we are essentially left only with predictors and their properties. Thus, a model for a DG in the class really means the predictor the model generates because the model has no necessary meaning. A predictor may be examined to learn something about the DG, such as the relevance or irrelevance of a variable, but detailed knowledge in the sense of a complete set of correlates for a DG is unobtainable.

A key point is that the existence of $\mathcal{M}$-open problems undermines the principle of falsifiability. If there is no true model and predictors can only be evaluated in terms of how well they perform on a relative basis, then the principle of falsifiability is irrelevant to many modern complex problems. That is, in many settings, all models are wrong and hence already falsified. They are not useful either except insofar as they give a good predictor that may or may not say anything about the DG. So, falsifiability per se is often merely a distraction from good prediction.

Indeed, many of the most important data sets currently being or recently gathered were not generated by a DG that admits a model, or, more precisely, were not generated by a mechanism that has anything stable or identifiable enough to model effectively. However, as long as there is something to measure we can make a guess as to its next value. Moreover, for these situations, predictors derived from model averages are often found to be better than their individual components, regardless of whether the models in the average are 'true' or merely regarded as the predictors they generate. There are other classes of predictors also do not generally admit physical interpretations yet also have clear predictive utility. Here, we study three classes of such predictors.

Our goal is to show that it is possible to provide interpretations for predictors that are generally uninterpretable. Specifically, kernel methods, the Bayes Shtarkov solution, and random forests are predictors for complex ($\mathcal{M}$-complete or $\mathcal{M}$-open) problems that are typically regarded as hard or impossible to interpret. We develop two types of interpretation for each of them. One is an interpretation in the usual sense of finding physical correlates, usually approximately, for components of the predictors. The other is a theoretical characterization of these predictors in terms of concepts to help guide their use.

In Section 5 we summarize out findings by stating what we call the *prediction principle* that we propose should be added to the *falsification principle*. This is separate from and in addition to the celebrated *prequential principle*, see Dawid (1984, 1992, 2010) and Dawid and Vovk (1999), that we regard as foundational.

The structure of this paper is as follows. In Section 2 we provide an interpretation for kernel methods such as relevance vector machines (RVM's) and support vector machines (SVM's) using the eigenfunctions of kernels with a consistency result so the interpretations will be valid. In Section 3 we present a Bayes version of the Shtarkov predictor and indicate how to interpret it as a mixing of Beta distributions or in terms of a Pearson distribution. In Section 4 we show that a random forest is asymptotically equivalent to boosting so random forest builds an additive logistic regression model. They can also be approximated in a regression sense. Some concluding remarks are made in Section 5. Longer technical proofs are in Appendix 1 and some further discussion of interpretability versus complexity can be found in Appendix 2.

2. RVM's and SVM's for regression

For a regression problem with a training data set $\mathcal{D} = \mathcal{D}_n = \{(y_i, x_i), i = 1, \ldots, n\}$, where y is the response variable and x is a covariate of dimension p, the goal is to find a function $f(x)$ to predict the responses y in a test set. This can be viewed as a regularization problem of the form

$$\min_{f \in \mathcal{H}_K} \left[\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \right], \qquad (1)$$

where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) with kernel K and norm $\|\cdot\|_{\mathcal{H}_K}$, L is a loss function, and $\lambda > 0$ is the smoothing parameter. It can be shown that (1) has a solution of the form

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x); \qquad (2)$$

see Kimeldorf and Wahba (1971). The solution $\hat{f}_\lambda$ is not a conventional model because the evaluations of K do not have any necessary physical correlates and in addition to its dependence on the parameters $\alpha_i, \hat{f}_\lambda$ depends explicitly on the x_i's to define the functions in the sum, the number of which depends on n. Indeed, the optimal values of the α_i's are $\alpha_i = \alpha_i(\mathcal{D})$. Estimating the α_i's means finding a data-driven approximation $\hat{\alpha}_i(\mathcal{D})$ even though the 'true' value depends on $\mathcal{D}$. That is, the data is used once to define the 'true' α_i's from the optimization (1) and then again to obtain good estimates for them. It is easiest to regard the latter estimates as converging in the sense of real numbers rather than stochastically.

Conditional on $\mathcal{D}$, the (uninterpretable) representer theorem predictor is

$$\hat{Y}_{\text{rep}}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x). \tag{3}$$

It is well known that RVM's and SVM's are of the form (3). Moreover, it is seen that (3) is the mode of a posterior where $L(y_i, f(x_i))$ is the exponent in an exponential family and $\lambda\|f\|^2_{\mathcal{H}_K}$ is the log of the prior on f. The representation (2) of f is of special interest when the number of covariates p is much larger than the sample size n and predictors such as $\hat{Y}_{\text{rep}}(\cdot)$ are often used in $\mathcal{M}$-complete (and $\mathcal{M}$-open) settings.

We start with a consistency result so the approximate interpretations we present will be asymptotically valid.

2.1. Consistency of the representer theorem predictor

Consider an $\mathcal{M}$-complete problem and let

$$\hat{Q}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\|f\|^2_{\mathcal{H}_K}. \tag{4}$$

The population version of (4) is

$$Q_0(f) = E_{(X,Y)} L(Y, f(X)) + \lambda\|f\|^2_{\mathcal{H}_K}. \tag{5}$$

We can assume that, for each i the $\hat{\alpha}_i = \hat{\alpha}_i(\mathcal{D})$'s are known from the empirical optimization of $\hat{Q}_n$ and that the true values for the α_i's given $\mathcal{D}$ are fixed with limits, assuming they exist, due to the optimization of Q_0.

Theorem 2.1: *Assume* (i) $Q_0(f)$ *is uniquely minimized at* f_0, (ii) $Q_0(f)$ *is continuous in* f, *and* (iii) $\hat{Q}_n(f)$ *converges uniformly in probability to* $Q_0(f)$ *i.e.,* $\sup_{f\in\mathcal{H}_K} |\hat{Q}_n(f) - Q_0(f)| \xrightarrow{P} 0$. *Then*

$$\hat{Y}_{\text{rep}} \xrightarrow{P} f_0,$$

where the convergence is over independent outcomes from the distribution of (X, Y).

Proof: This is a modification of Theorem 2.1 in Newey and McFadden (1994). Since $\hat{Q}_n(\hat{Y}_{\text{rep}}) \le \hat{Q}_n(f)$ for any f by the optimality of $\hat{Y}_{\text{rep}}$, we have that for any $\epsilon > 0$, $\hat{Q}_n(\hat{Y}_{\text{rep}}) \le \hat{Q}_n(f_0) + \epsilon/3$. Also, for any $\epsilon > 0$, Assumption (iii) gives

$$Q_0(\hat{Y}_{\text{rep}}) < \hat{Q}_n(\hat{Y}_{\text{rep}}) + \frac{\epsilon}{3},$$
$$\hat{Q}_n(f_0) < Q_0(f_0) + \frac{\epsilon}{3},$$

with probability approaching one (w.p.a.1), as $n \to \infty$. Therefore, w.p.a.1,

$$Q_0(\hat{Y}_{\text{rep}}) < \hat{Q}_n(\hat{Y}_{\text{rep}}) + \frac{\epsilon}{3} < \hat{Q}_n(f_0) + \frac{2\epsilon}{3} < Q_0(f_0) + \epsilon. \tag{6}$$

For any $\delta > 0$, let

$$B(f_0, \delta) = \{f \in \mathcal{H}_K : \|f - f_0\|_{\mathcal{H}_K} < \delta\}.$$

Since $B(f_0, \delta)^c$ is closed, Assumptions (i) and (ii) give

$$\inf_{f\in B(f_0,\delta)^c} Q_0(f) = Q_0(f^*) > Q_0(f_0)$$

for some $f^* \in B(f_0, \delta)^c$.

Choosing $\epsilon = \inf_{f\in B(f_0,\delta)^c} Q_0(f) - Q_0(f_0) = Q_0(f^*) - Q_0(f_0)$, expression (6) implies

$$Q_0(\hat{Y}_{\text{rep}}) < Q_0(f^*) = \inf_{f\in B(f_0,\delta)^c} Q_0(f) \text{ w.p.a.1},$$

and hence $\hat{Y}_{\text{rep}} \in B(f_0, \delta)$. Letting $\delta \to 0$ completes the proof. ■

Some discussion of what Theorem 2.1 means and does not mean is important here. The mode of convergence is stochastic, in the Hilbert space norm. The objects converging are functions of the form (3) in which the x_i's appear as arguments in the kernel evaluations and the (x_i, y_i)'s appear implicitly in the definition of the α_i's. The whole function (3) converges to the minimizer f_0. The function itself depends on $\mathcal{D}$ through the values of the x_i's and the α_i's. This means that the connection between (3) for one $\mathcal{D}$ and other data set $\mathcal{D}^*$ is unclear. The x_i's, y_i's, and the sample sizes, say n and n^*, may be different. So, there is no necessary relationship between $\alpha_i(\mathcal{D})$ and $\alpha_i(\mathcal{D}^*)$ even when $i \le \min(n, n^*)$ and the first $\min(n, n^*)$ pairs (x_i, y_i) are the same for $\mathcal{D}$ and $\mathcal{D}^*$. It may be easiest to regard increasing data sets as a sequence of problems corresponding to the accumulation of data and the convergence in the joint distribution of (X, Y) as summarizing the effect of replications over the entirety of all the countably infinite data sequences. There may be further structure in the convergence of the α_i's and their effect on the convergence of (2.1) to f_0, but we do not treat this here.

The convergence in Theorem 2.1 is in probability. The mode can be improved to L^2 with some extra hypotheses as seen in the following.

Corollary 2.1: *Assume the conditions in Theorem* 2.1 *hold. Assume*

(i) *Let X be a generic random variable representing any* X_i. *Then there is an* $\epsilon > 0$ *so that* $\forall x \in \overline{\text{supp}(X)}$

$$E[K^{2+\epsilon}(X,x)] < \infty;$$

(ii) *The sum of squares of the* α_i*'s is bounded with rate* $(1/n)$, *i.e.,* $\exists M$ *so that for any* $\mathcal{D}$, $\sum_{i=1}^{\infty} \alpha_i^2(\mathcal{D}) < M$ *and* $\exists N = N(n)$ *so that* $\sum_{i=N}^{n} \alpha_i^2 = o_P(1/(n-N))$.

Then, as $n \to \infty$, *for any x we have*

$$\hat{Y}_{\text{rep}}(x) \xrightarrow{L^2} f_0(x).$$

Proof: To establish the result, it is enough to show that as $n \to \infty$,

$$E[\hat{Y}_{\text{rep}}(x) - f_0(x)]^2 \to 0. \tag{7}$$

This is done in Appendix A. ■

Again, some discussion of this result is worthwhile. First, Assumption (i) is easy to verify for most kernels K. However, Assumption (ii) is asymptotic and therefore hard to verify. The boundedness clause, while intuitive, may have to be enforced by restricting (X, Y) to a compact set and renormalizing $P_{(X,Y)}$. The compact set can then be allowed to increase slowly while still preserving the result. The second clause of the assumption, the rate, is harder to deal with. Nevertheless, the rate assumption (and the boundedness assumption) are nearly always satisfied, at least approximately, in practice. While not a verification, the second clause can be checked by seeing how the α_j's perform using bootstrap samples from an $\mathcal{D}$. If the clause is satisfied for bootstrap samples and a range of finite n then it may be reasonable to take as true. In practice, with RVM's few of the α_i's are non-zero so as a practical matter, Assumption (ii) usually appears to be satisfied. Indeed, this can be seen in Tipping (2001).

A counterfactual may make Assumption (ii) less unpalatable. If the Representer Theorem solution were a Fourier expansion, the two clauses of Assumption (ii) would seem fairly reasonable. The first clause of Assumption (ii) would only mean that $\sum_i \alpha_i K(x_i, \cdot)$ is in the Hilbert space because Bessel's inequality gives that the sum of squared Fourier coefficients is less than the norm of the function. (This is true for any orthonormal basis.) Also, the rate $\sum_{i=N}^{n} \alpha_i^2 = o_P(1/(n-N))$ as n and $N(n)$ increase imposes a sparsity condition. It limits the collection of functions that can be well approximated because, as n increases, the last α_i's can't be too large. That is, the true function is only being approximated by $N(n)$ evaluations of the kernel. Otherwise put, this method is only effective for functions f_0 that are sufficiently sparse in terms of the $K(x_i, x)$'s required to express them. The rate clause in Assumption (ii) bounds how far f_0 can be from the approximations $\hat{Y}_{\text{rep}}$, in L^2, for consistency – as opposed to merely optimal approximation – to hold.

The mode of convergence in Corollary 2.1 is in L^2 pointwise in x. If a distribution P is assigned to X, then Egoroff's theorem can be applied. It strengthens the result by giving $\hat{Y}_{\text{rep}}(X) - f_0(X) \to 0$ in L^2 uniformly for $X \in A$ where A has arbitrarily large probability under P. That is, the convergence is almost uniform over most of the sample space of X.

To complete our treatment of the consistency of $\hat{Y}_{\text{rep}}$ we note that Assumption (iii) in Theorem 2.1 is hard to verify. So, we provide sufficient conditions for it.

Theorem 2.2: *Assume* (i) *the loss function L is continuous in f,* (ii) *there exists* $\delta > 0$ *so that for any* $f^* \in \mathcal{H}_K$, $E_{(X,Y)}[\sup_{f \in B(f^*,\delta)} L(Y, f(X))] < \infty$ *where* $B(f^*, \delta) = \{f \in \mathcal{H}_K : \|f - f^*\|_{\mathcal{H}_K} < \delta\}$, *and* (iii) *there exists an increasing sequence of compact subsets of* $\mathcal{H}_K$, $\{\mathcal{D}_j\}_{j=1}^{\infty}$, *converging to* $\mathcal{H}_K$ *such that, for each fixed n,* $\lim_{j\to\infty} \sup_{f\in\mathcal{D}_j} |\hat{Q}_n(f) - Q_0(f)| = \sup_{f\in\mathcal{H}_K} |\hat{Q}_n(f) - Q_0(f)|$. *Then*

$$\sup_{f \in \mathcal{H}_K} |\hat{Q}_n(f) - Q_0(f)| \xrightarrow{P} 0.$$

Proof: The proof is adapted from Theorem 6.10, the uniform weak law of large numbers, in Bierens (2005). The details are in Appendix A. ■

Remark: Uniform laws of large numbers emerge from empirical process theory, see Van de Geer (2000) Chapter 2 for instance. Often these results have weaker hypotheses that are harder to verify. We have used extensions of the classical law of large numbers since our goal is predictor interpretability not weakest conditions.

2.2. Theoretical interpretation of the representer theorem solution predictors

Under Mercer's conditions, see Mercer's Theorem in Scholkopf and Smola (2002), the kernel K can be decomposed as

$$K(x_i, x) = \sum_{m=1}^{\infty} \lambda_m g_m(x_i) g_m(x), \tag{8}$$

where $\{g_m \mid m = 1, 2, \ldots\}$ is an orthonormal set of eigenfunctions of K with $\int K(x,y) g_m(y)\, dy = \lambda_m g_m(x)$, $m = 1, 2, \ldots$. Thus, different kernels correspond to different orthonormal bases.

We can now use the g_m's in a nonparametric regression expansion for f_0. Write the projection of f_0 onto the

span of $\{g_1, \ldots, g_M\}$ as

$$r(x) = \sum_{j=1}^{M} \beta_m g_m(x) \tag{9}$$

so that estimating the β_m's is equivalent to estimating the projection. Since the g_m's are orthonormal, the optimal β_m's in an L^2 projection sense are $\langle g_m, f_0 \rangle$ and these can be estimated by $\hat{\beta}_m = (1/n) \sum_{i=1}^{n} y_i g_m(x_i)$ for $m = 1, \ldots, M$. For any reasonable choice of joint distribution for (Y, X), the central limit theorem gives that for each m, there is a σ_m so that

$$\hat{\beta}_m \sim N\left(\beta_m, \frac{\sigma_m^2}{n}\right), \tag{10}$$

asymptotically in n, assuming the second moments of the $\hat{\beta}_m$'s exist. The integrated squared bias of r as an estimator for f_0 is

$$B_M = B(f_0, r) = \int_{-\infty}^{\infty} (f_0(x) - r(x))^2 \, \mathrm{d}x = \sum_{m=M+1}^{\infty} \beta_m^2. \tag{11}$$

Since both r and f_0 are in a Hilbert space, B_M is finite for any M and for any latter m, $\beta_m \to 0$ as $M \to \infty$. Now, having controlled the variance of the $\hat{\beta}_m$'s and the bias of r we see that

$$\hat{r}_M(x) = \sum_{m=1}^{M} \hat{\beta}_m g_m(x) \tag{12}$$

converges to r as $n \to \infty$ and to f_0 with bias roughly $B_{M_n}(f_0, \hat{r})$. We can let $M_n \to \infty$ so slowly as $n \to \infty$ that B_M also goes to zero. Doing this, it is seen that we have $B_{M_n} \to 0$ as well, possibly slowly with n. Now, $\hat{r}$ converges to f_0 pointwise in x, i.e.,

$$\hat{r}_M(x) \to f_0(x) \quad \text{in } P \tag{13}$$

for any x. More explicit formal conditions for (13) and for versions of (13) in stronger modes of convergence can be given, but that is beyond our present scope. However, we note that this argument holds for any orthonormal basis but that the g_m's, being derived from K, are natural for this problem. Of course, if the basis elements in any orthonormal basis have a physical interpretation for a given K that is more compelling than the g_m, that basis would be preferred.

From (13) and Theorem 2.1, we have the following.

Theorem 2.3: *Assume* (13) *holds and that the hypotheses of Theorem 1 are satisfied. For $\mathcal{M}$-closed and $\mathcal{M}$-complete problems, for any x, the orthonormal basis predictor $\hat{r}_M$ in* (12) *is asymptotically equivalent to the representer theorem solution predictor $\hat{Y}_{\mathrm{rep}}(x)$, i.e., as $n \to \infty$ and consequently $M_n \to \infty$ at an appropriate rate*

$$\hat{r}_{M_n}(x) - \hat{Y}_{\mathrm{rep}}(x) \xrightarrow{P} 0.$$

From Theorem 2.3 we are justified in regarding $\hat{r}$ as an interpretation of $\hat{Y}_{\mathrm{rep}}(x)$ on the grounds that the g_m's (or other orthonormal basis) admit a physical interpretation relevant to the DG. The cost of this interpretation is asymptotically zero but for finite n depends on B_{M_n} in (11) and the rate for the $\hat{\beta}_m$'s in (10).

Here we give some examples of the eigenfunctions g_m for common choices of kernel to show the interpretability is non-trivial. (Other orthonormal bases may be easier.)

Example 2.1: Consider the kernel $K(x, y) = \mathrm{e}^{-xy}$ on $(0, \infty)$.

By the definition of the Gamma function, for $\alpha > 0$,

$$\int_0^{\infty} t^{\alpha-1} \mathrm{e}^{-xt} \, \mathrm{d}t = \Gamma(\alpha) x^{-\alpha}.$$

Changing α to $1 - \alpha$, for $\alpha < 1$,

$$\int_0^{\infty} t^{-\alpha} \mathrm{e}^{-xt} \, \mathrm{d}t = \Gamma(1-\alpha) x^{\alpha-1}.$$

So, for $0 < \alpha < 1$,

$$\begin{aligned} &\int_0^{\infty} \left(\frac{1}{\sqrt{\Gamma(\alpha)}} t^{\alpha-1} + \frac{1}{\sqrt{\Gamma(1-\alpha)}} t^{-\alpha} \right) \mathrm{e}^{-xt} \, \mathrm{d}t \\ &\quad = \sqrt{\Gamma(\alpha)\Gamma(1-\alpha)} \\ &\qquad \times \left(\frac{1}{\sqrt{\Gamma(\alpha)}} x^{\alpha-1} + \frac{1}{\sqrt{\Gamma(1-\alpha)}} x^{-\alpha} \right). \end{aligned}$$

Therefore, by definition, the eigenfunctions of this kernel are

$$g_\alpha(x) = \frac{1}{\sqrt{\Gamma(\alpha)}} x^{\alpha-1} + \frac{1}{\sqrt{\Gamma(1-\alpha)}} x^{-\alpha}, \quad \text{for } 0 < \alpha < 1.$$

Example 2.2 (Polynomial kernel): The non-homogeneous version of the polynomial kernel of degree d is defined by $K(x, y) = (c + \langle x, y \rangle)^d$ where c is a constant and $\langle \cdot, \cdot \rangle$ is the inner product.

For $p = 2$, say, let $x = (x_1, x_2)$, $c = 0$, $d = 3$, and suppose that $(x_1, x_2) \sim 0.5N((-3, 1), I_2) + 0.5N((2, -1), I_2)$, then the eigenfunctions of this kernel are, see Liyang and Lee (2013),

$$\begin{aligned} g_1(x) &= \frac{1}{1862.615}(0.848x_1^3 - 0.791x_1^2x_2 \\ &\quad + 0.437x_1x_2^2 - 0.097x_2^3), \\ g_2(x) &= \frac{1}{343.748}(-0.518x_1^3 - 1.073x_1^2x_2 \\ &\quad + 0.862x_1x_2^2 - 0.317x_2^3), \\ g_3(x) &= \frac{1}{59.266}(-0.112x_1^3 - 1.079x_1^2x_2 \\ &\quad - 0.929x_1x_2^2 + 0.559x_2^3), \\ g_4(x) &= \frac{1}{1862.615}(0.848x_1^3 - 0.791x_1^2x_2 \\ &\quad + 0.437x_1x_2^2 - 0.097x_2^3). \end{aligned}$$

Example 2.3 (Exponential kernel): Consider the exponential kernel $K(x,y) = \exp(-\frac{|x-y|}{w})$ for the uniform distribution on the interval $[-1,1]$. In Diaconis et al. (2008) it was shown that the eigenfunctions of this kernel can be written as $\cos(bx)$ or $\sin(bx)$ inside the interval $[-1,1]$ for appropriately chosen values of b and decay exponentially away from it.

Example 2.4 (Gaussian kernel): Consider the Gaussian kernel $K(x,y) = \exp(-\frac{(x-y)^2}{2w^2})$ for the normal distribution $N(\mu,\sigma^2)$. Let $\beta = 2\sigma^2/w^2$ and let $H_i(x)$ be the i th-order Hermite polynomial, Shi et al. (2008) provided the eigenfunctions of this kernel,

$$g_i(x) = \frac{(1+2\beta)^{1/8}}{\sqrt{2^{i-1}(i-1)!}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\cdot\frac{\sqrt{1+2\beta}-1}{2}\right) \times H_{i-1}\left(\left(\frac{1}{4}+\frac{\beta}{2}\right)^{1/4}\frac{x-\mu}{\sigma}\right),$$

for $i = 1,2,\ldots$ In particular, the first eigenfunction is

$$g_1(x) = \left(1+\frac{4\sigma^2}{w^2}\right)^{1/8} \times \exp\left(-\frac{(x-\mu)^2}{4\sigma^2}\left(\sqrt{1+\frac{4\sigma^2}{w^2}}-1\right)\right).$$

Other examples may be developed but the key points are (i) the bases used for the orthonormal basis predictor should be chosen in view of the DG to ensure interpretability, and (ii) for finite n, using an interpretation of $\hat{Y}_{\text{rep}}$ will not in general be as good a predictor (in say PRESS_n) and this is the cost of interpretability.

2.3. A more empirical interpretation

If g_m's are not interpretable, and no obvious orthonormal basis can be identified, we might be led to default to the coordinates of the x_i's since they, presumably, were the quantities measured. In the $\dim(x) = 1$ case we can write Taylor expansions

$$g_m(x) = \beta_0 + \beta_1 x + \cdots + \beta_k x^k.$$

If each Taylor expansion converges i.e., $g_m(x)$ analytic, then we get an analogous result for projections of the form of $r(x)$ in (9). For ease of exposition, each Taylor series can be represented as a finite sum of orthonormal polynomials. A common choice is the Hermite polynomials, often denoted $H_0, H_1, \ldots$. Then the difference between using kth-order Taylor expansions and expansions using the first k Hermite polynomial basis elements is simply identifying the linear transformation between bases.

Following Section 2.2, we can form r as in (9) and $\hat{r}$ as in (12) using Hermite polynomials, and obtain a variant on Theorem 2.3 so that the $\hat{r}$ provides a quantifiably good approximation to f_0. The result can be left in Hermite polynomials or converted back to the Taylor expansion of each g_m in r. The point of converting back to the polynomial basis used for Taylor expansions is that functions of the form x^j are usually easier to interpret physically in the context of a DG than Hermite polynomials are simply because it was x that was measured.

Thus, using linear models, which are commonly regarded as interpretable, to approximate the g_m's, is asymptotically equivalent to approximating K in (8) directly. In practice, we suggest it will be easier to use Hermite polynomials (or any other orthonormal basis) on the terms on the right side of (8), and convert them to the polynomials used in Taylor expansions than approximating K directly. Again, the finite sample discrepancy between the approximation we just described and (3) quantifies the cost of passing from an optimal predictor to an interpretable predictor. An entirely analogous argument holds when $\dim(x) \geq 2$ provided that orthonormal bases for polynomial spaces with $\dim(x)$ variables are used.

3. Bayes Shtarkov predictors in $\mathcal{M}$-open settings

Consider the online prediction of arbitrary sequences $y_1, y_2, \ldots$, drawn from a finite set $\mathcal{Y}$. In $\mathcal{M}$-open settings, interest focusses on the case that no probability distribution can be assumed for a sequence of length n, say $y^n = (y_1, y_2, \ldots, y_n)$. This is the paradigm $\mathcal{M}$-open statistical prediction problem for strings of values.

This problem can be regarded as a sequential game between Nature, N, and a Forecaster, F, permitting F to access a collection of experts indexed by $\theta \in \Theta \subset \mathbb{R}^k$ for some k. In the special case of log-loss, each round of the game proceeds as follows. Each expert announces a density say p_θ. Given this, F announces a density $q(\cdot)$ that will be used to predict the value N issues. Finally, N issues y and pays F $\log q(y)$. If this number is negative, it is the amount of money F pays N and this concludes the round. See Shtarkov (1987) and Cesa-Bnachi and Lugosi (2006) for details of this game and its properties.

Now suppose n independent rounds of this game are to be played. Prior to the first round, each expert θ announces a density $p(\cdot \mid \theta)$ for y^n. F receives these p_θ's and chooses the density $q(y^n)$ by trying to match the performance of the best expert θ for predicting y^n. Then, N reveals y^n and incurs the loss (or gain) $\log q(y^n)$. The question remains how F should use the p_θ's to choose q. Obviously, the best expert will incur the loss $\min_\theta \log 1/p(y^n \mid \theta)$ to F where θ ranges over the experts.

3.1. The Bayesian version

In the Bayes version of the game, F has access to experts that are weighted by a prior $w(\theta)$. (If $w \equiv 1$, this reduces to the frequentist version.) In this case, F would want to choose q to minimize the maximum regret

$$\sup_{y^n}\left[\log\frac{1}{q(y^n)} - \inf_{\theta}\log\frac{1}{w(\theta)p(y^n \mid \theta)}\right] = \sup_{y^n}\left[\sup_{\theta}\log\frac{w(\theta)p(y^n \mid \theta)}{q(y^n)}\right]. \quad (14)$$

More formally, the solution q_{opt} to (14) that we henceforth call the Bayes Shtarkov predictor (for the discrete case) is, see Le and Clarke (2016a),

$$q_{\mathsf{opt}}(y^n) = \arg_q\left[\inf_{q\in\mathcal{P}}\left(\sup_{y^n}\sup_{\theta}\log\frac{w(\theta)p(y^n \mid \theta)}{q(y^n)}\right)\right] = \frac{w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))}{\sum_{y^n} w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))}, \quad (15)$$

where θ ranges over the 'parameter space' indexing the experts, $\mathcal{P}$ is the collection of all densities for y^n with respect to counting measure, and $\tilde\theta$ is the posterior mode. (Since q is in the denominator of (15), $q(y) \neq 0$ for any $y \in \mathcal{Y}$.)

In the continuous case, the sum becomes an integral over a subset of a real space, $\mathcal{P}$ becomes a class of densities with respect to Lebesgue measure, so (15) becomes

$$q_{\mathsf{opt}}(y^n) = \frac{w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))}{\int w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))\mathrm{d}y^n}; \quad (16)$$

see Clarke (2007, Sec. 5.2) for discussion of (15) and (16).

The solution $q_{\mathsf{opt}}(y^n)$ does not factor into a product of $q(y_i)$'s and so does not correspond to a stochastic process. Nevertheless, regardless of how $q_{\mathsf{opt}}(y^n)$ is computed, univariate Bayes Shtarkov densities predictors, when they exist, are of the form

$$q_{\mathsf{opt}}(y_{n+1} \mid y^n) = \frac{q_{\mathsf{opt}}(y^{n+1})}{q_{\mathsf{opt}}(y^n)}, \quad (17)$$

and can be used prequentially, i.e., to generate sequential predictions.

The foregoing generalizes directly to the case where side information, i.e., a value x_i is associated to each y_i. So, let us write the corresponding q_{opt} as $q_{\mathsf{shk}}(y \mid x^{n+1}, y^n)$, namely the Bayes Shtarkov predictive density for Y where $x^{n+1} = (x_1, \ldots, x_{n+1})$ and $y^n = (y_1, \ldots, y_n)$. Now, there are two ways to generate predictions from q_{shk}: (i) use q_{shk} as a density to generate interval predictors, and (ii) convert q_{shk} to a point predictor.

For the first, recall that under various regularity conditions, q_{shk} can be approximated by the Bayesian's marginal density for the data, i.e., $q_{\mathsf{shk}}(y^n) \approx m(y^n)$ in terms of regret. That is, (14) leads to

$$\sup_{y^n}\left[\log\frac{w(\tilde\theta)p(y^n \mid \tilde\theta)}{q_{\mathsf{shk}}(y^n)\hat I(\tilde\theta)}\right] = \sup_{y^n}\left[\log\frac{w(\tilde\theta)p(y^n \mid \tilde\theta)}{m(y^n)\hat I(\tilde\theta)}\right] = \frac{1}{2}\log\frac{n}{2\pi} + o(1), \quad (18)$$

where $\hat I(\cdot)$ is the empirical Fisher information from $p(y \mid \theta)$; see Xie and Barron (2000) and Clarke (2007). This means a stochastic process, the Bayesian's mixture, is approximating a density q_{shk} that does not correspond to a stochastic process. Hence, even in $\mathcal{M}$-open problems, there may be good – new – predictors that resemble familiar predictors. If computing q_{shk} is onerous, $m(\cdot)$ may be a good predictor. Likewise, conditionals from $m(\cdot)$ such as $m(y_{n+1} \mid y^n)$ may be a good approximation for (17).

More important for the present, if the Bayesian's marginal for the data is 'interpretable' – perhaps because the densities proposed by the experts are – (18) provides an asymptotic interpretation of q_{shk} in terms of $m(\cdot)$. The difference between $m(\cdot)$ and q_{shk} represents the degree to which $m(\cdot)$ is predictively suboptimal to q_{shk} – in regret under log-loss – and the degree of suboptimality decreases as $n \to \infty$. Thus, $(1-\alpha)$% predictive intervals under $m(\cdot)$ and q_{shk} are equivalent in the limit even though for all finite n q_{shk} is better in terms of regret.

By contrast, if a generic 'interpretable' predictor $r(y^n)$ is used, the regret usually becomes far worse. For instance,

$$\sup_{y^n}\left[\log\frac{w(\tilde\theta)p(y^n \mid \tilde\theta)}{r(y^n)\hat I(\tilde\theta)}\right] = \sup_{y^n}\left[\log\frac{w(\tilde\theta)p(y^n \mid \tilde\theta)}{q_{\mathsf{shk}}(y^n)\hat I(\tilde\theta)} + n\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{q(y_i \mid y^{i-1})}{r(y_i \mid y^{i-1})}\right)\right] \approx \frac{1}{2}\log\frac{n}{2\pi} + n\sup_{y^n} D(q_i \mid r_i) + o(1), \quad (19)$$

in which $D(q_i \mid r_i)$ is defined by (19) and typically dominates, resulting in a much larger loss for F. In part, this is an artifact of using the log-loss of a density ratio which is much more sensitive to tail behaviour than other functions. Statements analogous to (17), (18), and (19) hold for sequences of discrete outcomes y_i. In either case, this suggests that interpretability per se, without any direct relationship to the minimum regret in a logarithmic sense, gives much larger costs.

For the second, if we consider pointwise prediction, we want a point predictor from q_{shk} that represents

where the mass of q_{shk} is, say $E_{q_{\mathsf{shk}}}(Y)$. One such predictor is

$$\hat{Y}_{\mathsf{shk}}(x) = E_{q_{\mathsf{shk}}}(Y \mid x^{n+1}, y^n). \tag{20}$$

Others include $\mathsf{med}(Y \mid x^{n+1}, y^n)$, $\mathsf{mode}(Y \mid x^{n+1}, y^n)$ etc. In addition, because of (18) we are led to approximate each of these point predictors by expectations with respect to the conditional $m(\cdot \mid x^{n+1}, y^n)$ from the mixture.

Assume $p_j(y)$ is the proposed density for Y from expert j, $j = 1, \ldots, J$, i.e., we assume finitely many experts or that a continuum of experts can be approximated by a weighted sum of finitely many. Even though there is no distribution associated with Y because the problem is $\mathcal{M}$-open, we can still take expectations with respect to q_{shk} and the p_j's. We can write

$$\begin{aligned} &|Y - E_{\sum_{j=1}^J \gamma_j p_j}(Y)| \\ &\quad \le |Y - E_{q_{\mathsf{shk}}}(Y)| + |E_{q_{\mathsf{shk}}}(Y) - E_{\sum_{j=1}^J \gamma_j p_j}(Y)| \\ &\quad = |Y - \hat{Y}_{\mathsf{shk}}| + |\hat{Y}_{\mathsf{shk}} - E_{\sum_{j=1}^J \gamma_j p_j}(Y)|, \end{aligned} \tag{21}$$

where $\gamma_j > 0$, $\sum_{j=1}^J \gamma_j = 1$, and subscripts on E indicate density in which expectation is taken. Since the Bayes Shtarkov predictor is best in log-loss, the first term in (21) is likely to be small. Thus, we only need to find γ_j's such that the second term in (21) is as small as possible if not zero. If this is done, then the first term represents the minimal cost of prediction and the second term represents the additional cost of interpretation, if the p_j's are interpretable.

We can modify (21) by adding and subtracting expressions involving expectations with respect to $m(y_{n+1} \mid x^{n+1}, y^n)$ as a way to try to identify the components of the error meant by (21) in terms of the uninterpretable q_{shk}, its distance from a mixture (possibly regarded as interpretable) and sums of weighted densities of the experts (if they are interpretable), e.g., we can add and subtract $E_{m(\cdot|y^n)}Y$ in the last term of (21). In practice, the last terms would represent the cost of interpretability while the first term on the right in (21) represents the minimal cost of prediction.

3.2. A theoretical interpretation for Bayes Shtarkov predictors

Separate from approximating q_{shk} by mixture densities or other expressions, in some cases we can characterize q_{shk} as a mixture itself. The hypotheses are rather strong and the mixing density is somewhat artificial but in some examples this characterization may be useful. We have the following.

Theorem 3.1: *Assume q_{shk} is m-monotone over $(0, \infty)$ i.e., $(-1)^k q_{\mathsf{shk}}^{(k)}(|x|) \ge 0$ for $k = 0, \ldots, m-1$ where $q_{\mathsf{shk}}^{(k)}$ is the kth derivative of q_{shk} and $q_{\mathsf{shk}}^{(0)} = q_{\mathsf{shk}}$. Then q_{shk} can be represented as the following mixture for any integer k, $1 \le k \le m$,*

$$q_{\mathsf{shk}}(y) = \int_0^\infty \left[\frac{1}{s} k \left(1 - \frac{|y|}{s}\right)_+^{k-1}\right] g(s)\, \mathrm{d}s, \tag{22}$$

where $a_+ = \max\{a, 0\}$ and the mixing density $g(s)$ is

$$g(s) = \frac{1}{k} \sum_{j=0}^{k-1} \frac{(-1)^j}{j!} [j s^j q_{\mathsf{shk}}^{(j)}(s) + s^{j+1} q_{\mathsf{shk}}^{(j+1)}(s)].$$

Proof: The proof of Theorem 1 in Polson et al. (2014), based on Williamson (1956), holds for all positive values of y. For negative values of y, define $f(y)$ on $(0, \infty)$ by $f(y) = q_{\mathsf{shk}}(-y)$, then we have the same result for $f(y)$:

$$f(y) = \int_0^\infty \left[\frac{1}{s} k \left(1 - \frac{y}{s}\right)_+^{k-1}\right] g(s)\, \mathrm{d}s.$$

Hence, for the negative values of y,

$$q_{\mathsf{shk}}(y) = f(-y) = \int_0^\infty \left[\frac{1}{s} k \left(1 - \frac{-y}{s}\right)_+^{k-1}\right] g(s)\, \mathrm{d}s.$$

Thus, for all y, q_{shk} has the representation (22). ■

We do not have sufficient conditions for q_{shk} to be m-monotone. However, examples of q_{shk} in Le and Clarke (2016a) look like graphs of $1/y$ or e^{-y} which are m-monotone.

One of the implications of Theorem 3.1 is that if $X \sim \mathrm{Beta}(1, k)$ and $S \sim g(\cdot)$ then $Y = SX$ will have density (22). For instance,

(1) if $k = 1$, $g(s) = s q'_{\mathsf{shk}}(s)$ and $X \sim \mathrm{Beta}(1, 1) = \mathrm{Uniform}(0, 1)$,
(2) if $k = 2, g(s) = -\frac{s^2}{2} q''_{\mathsf{shk}}(s)$ and $X \sim \mathrm{Beta}(1, 2)$,
(3) if $k = 3, g(s) = \frac{s^3}{3} q'''_{\mathsf{shk}}(s)$ and $X \sim \mathrm{Beta}(1, 3)$.

Thus, while g remains uninterpretable, X is recognizable and Y is a product. A limitation of this result is that it is only for univariate y. However, as suggested by the relationship between $m(\cdot)$ and q_{shk} in (18), extensions to multivariate y may be possible under some $\mathcal{M}$-open analog of stationarity – provided there is the right sort of dependence so that the middle term of order n in (19) can be avoided.

3.3. An empirical interpretation for Bayes Shtarkov predictors

We can approximate the univariate density $q_{\mathsf{shk}}(y)$ by finding the member of a large family of densities closest to it. For this we want a relatively large family of candidate densities that are parametrized in some way that reflects out understanding of the shapes of densities. One such family consists the Pearson distributions

first introduced by Pearson (1895). There are at least 7 useful subtypes within the Pearson family; Pearson himself ultimately identified 12. Overall, this family is characterized by five parameters: A location parameter a (often interpretable as a mode), a location parameter λ (often interpretable as a mean, μ_1), a variance μ_2, a skewness γ_1 (this enters as $\beta_1 = \gamma_1^2$), and a kurtosis β_2. While this family is only for univariate densities, there are proposed generalizations to the multivariate case, see Steyn (1960) amongst others, although they are not well developed and there is little recent work on them.

Formally, a Pearson density p is any solution to the differential equation,

$$\frac{p'(y)}{p(y)} + \frac{a + (y - \lambda)}{b_0 + b_1(y - \lambda) + b_2(y - \lambda)^2} = 0, \tag{23}$$

where

$$b_0 = \frac{4\beta_2 - 3\beta_1}{10\beta_2 - 12\beta_1 - 18}\mu_2,$$
$$b_1 = a = \sqrt{\mu_2}\sqrt{\beta_1}\frac{\beta_2 + 3}{10\beta_2 - 12\beta_1 - 18},$$
$$b_2 = \frac{2\beta_2 - 3\beta_1 - 6}{10\beta_2 - 12\beta_1 - 18}.$$

Now, in principle, values of $(a, \lambda, \mu_1, \beta_1, \beta_2)$ yielding the Pearson density closest to a given q_{shk} density can be found. This Pearson density may be regarded as an interpretation of q_{shk}, but it cannot be as good as q_{shk} in terms of regret, see (14). The increase in regret from using the Pearson density closest to q_{shk}, rather than q_{shk} itself, is the cost of interpretation.

The solution of (23) is the indefinite integral

$$p_{\mathrm{prs}}(y) \propto \exp\left(-\int \frac{y - a}{b_2 y^2 + b_1 y + b_0}\mathrm{d}y\right).$$

For the sake of completeness, we look at an example, namely the two cases of the Pearson type IV distributions based on whether $b_1^2 - 4b_0b_2$ is negative or non-negative. (The term discriminant arises from the use of the quadratic root formula.) They are:

Case I: if $b_1^2 - 4b_0b_2 < 0$, then

$$p_{\mathrm{prs}}(y) \propto \left[1 + \left(\frac{y - \lambda}{\alpha}\right)^2\right]^{-m} \times \exp\left[-\nu \arctan\left(\frac{y - \lambda}{\alpha}\right)\right], \tag{24}$$

where

$$\alpha = \frac{\sqrt{4b_0b_2 - b_1^2}}{2b_2},$$
$$\nu = -\frac{2b_2a + b_1}{2b_2^2\alpha},$$
$$m = \frac{1}{2b_2}.$$

Case II: if $b_1^2 - 4b_0b_2 \geq 0$, then

$$p_{\mathrm{prs}}(y) \propto \left(1 - \frac{y}{a_1}\right)^{-\nu(a_1 - a)}\left(1 - \frac{y}{a_2}\right)^{-\nu(a_2 - a)}, \tag{25}$$

where

$$a_1 = \frac{-b_1 - \sqrt{b_1^2 - 4b_0b_2}}{2b_2},$$
$$a_2 = \frac{-b_1 + \sqrt{b_1^2 - 4b_0b_2}}{2b_2},$$
$$\nu = \frac{1}{b_2(a_1 - a_2)}.$$

Let d be a distance between densities. We can find $\hat{\nu}_1 = (\lambda_{\min}, \alpha_{\min}, \nu_{\min}, m_{\min})$ and $\hat{\nu}_2 = (a_{\min}, a_{1,\min}, a_{2,\min}, \nu_{\min})$ that achieve

$$\arg(\min d(q_{\mathsf{shk}}, p_{\mathrm{prs}})), \tag{26}$$

where the minimum is over the parameters in (24) or (25), respectively. Naturally, we would choose the Pearson density corresponding to whichever of $\hat{\nu}_1$ and $\hat{\nu}_2$ gave a lower value of the minimum in (26). In principle, this can be done over the other types of Pearson densities (or any parametrized class of densities) so that the parameters giving the overall minimum for the distance in (26) can be found. The result is a Pearson distribution whose density approximates q_{shk} as closely as possible within the family and hence has an interpretation based on the shape of the approximating density. If the minimum is not small enough given the choice of d, we may be led to consider richer families of densities than Pearson.

As a final point for this section, recall it is well known that if all moments of a distribution exist then they characterize the distribution and that the more the moments of two distributions that are close, the closer the two distributions are. Thus, as long as the moments of the distribution are meaningful the distributions can be regarded as interpretable, but, as we have seen, interpretation has a cost.

4. Random forest predictors

Random forest (RF) predictors were introduced by Breiman (2001a). The main idea is to use bootstrap aggregation on trees with binary splits or, more formally, binary recursive partitioning models, and add one extra step to reduce the correlation between any pair of trees in the forest. The extra step is random selection from the explanatory variables. That is, when growing a tree on a bootstrapped sample, before each split, choose $m \leq p$ of the explanatory variables at random to be candidates for splitting. Values for m typically range from $\log_2 p$ to $\sqrt{p}$. Despite the de-correlation

step, RF's have many of the properties of bagging trees. Here we focus on RF's because they are one of the most successful model averaging methods for binary classification and often have benefits other methods, including boosting, do not.

Here we will relate RF's to boosting, see Schapire (1990), to show the main point of this paper – that interpretable methods have a performance cost over optimal predictive methods – holds for classification as well as regression. This is a different perspective from Wyner et al. (2017) who argue that RF's and boosting work well because both are interpolating and averaging. Our point has to do with interpreting classifiers not understanding why they work.

4.1. Interpreting RF's in terms of boosting

Our point in this subsection can be stated succinctly as follows. Let $RF(x)$ be the random forest classifier based on x. Consider a data set $\mathcal{D} = \{(y_i, x_i), i = 1, \ldots, n\}$ where the pairs (y_i, x_i) are independent over i, $y_i \in \mathcal{Y} = \{-1, 1\}$ and $x_i \in \mathbb{R}^p$. Write $RF(x) = RF_{B,\rho,\tau}(x)$ where ρ is a splitting rule, τ is a stop splitting rule and B is the number of trees used to form the random forest. Thus,

$$RF(x) = \left(\frac{1}{B}\right) \sum_{b=1}^{B} T_{b,\rho,\tau}(x)$$

where $T_b = T_{b,\rho,\tau}$ is the classification tree formed from the bth bootstrap sample and the decorrelation used to form T_b is suppressed in the notation. Let $BSTC_J(x) = BSTC(x)$ be the boosted classifier using J iterations of the boosting procedure starting with the initial tree-based classifier $C(x)$ assumed to be 'weak'. That is $P(Y \neq C(X)) < .5$, but not by much. Now we can write

$$RF(x) = (RF(x) - BSTC(x)) + BSTC(x). \tag{27}$$

The idea is that $RF(\cdot)$ is uninterpretable but $BSTC(\cdot)$ has a limited interpretation (due to Friedman et al. (2000), see below) so the term $(RF(x) - BSTC(x))$ represents the cost of that interpretation. We assume that RF's perform a little better than boosted trees because they generally do unless the boosting classifier is sufficiently well-calibrated and does not overfit. That is, RF's are nearly automatic and hence more robust. Moreover, a majority of successful classifiers are random forests or variants on random forests; see Caruana et al. (2008) for a definitive report emphasizing high dimensional problems. As a generality, boosting also does not generalize well beyond binary classification.

To make this more precise, recall that while there are a variety of boosting algorithms, and variants on boosting algorithms including gradient boosting, AdaBoost due to Freund and Schapire (1997) is arguably the most popular. The basic idea of boosting is to improve a weak classifier iteratively by averaging the reweighted improvements i.e., to help the weak classifier learn from its mistakes. To generate the improved classifiers, the boosting procedure re-uses the data like bagging or RF's but builds iterates rather than starting anew with each iteration.

Let $C(x)$ be a fixed initial classifier for Y and assume that, as a function, $C(x)$ is representable as a tree. Denote the iterates of C under the boosting algorithm by $\hat{C}_j, j = 1, \ldots, J$. The iterates are also classifiers. They are ensembled into a final classifier by weighted majority voting to yield

$$BSTC(x) = \text{sign}\left(\sum_{j=1}^{J} \beta_j \hat{C}_j(x)\right), \tag{28}$$

where the weights $\beta_1, \ldots, \beta_J$ are computed by the AdaBoost algorithm, see Freund and Schapire (1997) for details. The central intuition in Adaboost is that increasing the penalty for misclassified data points forces successive $\hat{C}_j$'s to make fewer errors. There is a performance criterion that is satisfied by most versions of boosting, see Freund and Schapire (1997) Theorem 6. This result leaves open the possibility that a boosted classifier could be perfect in a limiting sense but does not actually give convergence of $BSTC$ to a limit.

Even though Adaboost was a novel approach to classification by ensembles, Friedman et al. (2000) showed it was equivalent to to forward stagewise additive logistic regression under exponential loss. This provides an interpretation of boosting because the logistic regression gives an explicit expression for $P(Y = 1 \mid \mathcal{D})$ in terms of specifically constructed functions of x (the $\hat{C}_j$'s below), see Friedman et al. (2000, Sec. 3.3) and Hastie et al. (2009, Secs. 10.4 and 10.5) for details. Since the $\hat{C}_j$'s are individual trees they admit interpretations in terms of the explanatory variables. This result holds in a limiting sense as the terms in the logistic regression increase and as $n \to \infty$. So, for each finite step, the boosted classifier is suboptimal even though it is Bayes optimal in the limit; we use this below in the proof of Corollary 4.1.

One key step in the Adaboost procedure is choosing how the iterates $\hat{C}_j$ are to be generated. There are various choices; the most popular are probably naive Bayes classifiers or trees with a maximum number of splits. Here we use the latter. So, we start by taking C to be a tree classifier and want our iterates to be trees as well. Specifically, the criterion the iterates must satisfy is

$$\hat{C}_{j+1}(x) = \arg\min_{h \in G_j} \sum_{i=1}^{n} D_j(i) \mathbf{1}_{\{y_i \neq h(x_i)\}}(x), \tag{29}$$

where $D_j = (D_j(1), \ldots, D_j(n))$ for $j \geq 1$ is the 'empirical' distribution on the n data points given by

$$D_j(i) = \frac{D_{j-1}(i) e^{\beta_{j-1} \mathbf{1}_{\{y_i \neq \hat{C}_{j-1}(x_i)\}}}}{N_{j-1}} \tag{30}$$

in which $\hat{C}_0 = C$, N_{j-1} is a normalization constant, the β_{j-1}'s are found by an auxiliary procedure, and

the sequence of distributions D_j is initialized by $D_0 = (1/n. \ldots, 1/n)$.

In (29), the classifiers h at step j vary over a set of classifiers G_j. As noted, C is a tree and we want the iterates to be trees. So, each G_j must be a class of trees. If G_j is too small a class, e.g., trees with exactly one split (often called 'stumps'), then the range of boosted tree-based classifiers will be too small. For instance, if G_j only contained stumps, it would not include trees that allowed interactions between entries in x. On the other hand, if G_j is too big, $\hat{C}_j$ will fit the data perfectly even though such a classifier often has poor generalization error. So, we have to choose a reasonable value for the number of splits to allow in the classifiers in G_j.

Chapter 10, Sec. 11 in Hastie et al. (2009) recommends using trees with four to eight splits. Three splits allows for interactions between explanatory variables, but often not enough so starting with four splits and working up to eight often a good overall procedure. Hastie et al. (2009) comment that 10 or more splits are rarely required for good performance, e.g., low generalization error.

To see that under these circumstances (28) is a tree it is enough to show that, as functions of x, a linear combination of trees is a tree. First, a real constant times a tree function is a tree so it is enough to show that the sum of two trees is again a tree. Let T_1 and T_2 be trees. To see that $T(x) = T_1(x) + T_2(x)$ is also a tree, observe that if each leaf node of T_1 is taken as the root node of T_2 the result is a tree T of twice the depth for which each input vector of explanatory variables ends up in exactly one leaf. However, some of the nodes or leaves may be void. This does not make T invalid, just artificial, and it may be collapsible into a much smaller tree. Nevertheless, in a trivial sense, the linear combination of trees is tree.

Less artificially, write

$$T_k(x) = \sum_{\ell=1}^{u_k} \alpha_\ell I_{R_\ell^{(k)}}(x) \tag{31}$$

for $k = 1, 2$, where the $R_\ell^{(k)}$ are disjoint and exhaustive regions in $\mathbb{R}^p$ assumed to have edges parallel to the axes of the real space. In the special case of $p = 1$, $\mathbb{R}$ has an ordering so it is easy to see that assigning the right constant to each intersection $R_\ell^{(1)} \cap R_{\ell'}^{(2)}$ gives a function that can be expressed as constants times indicator functions for intervals in $\mathbb{R}$. Since intervals can be defined by splits on the single real variable, the sum $T(x) = T_1(x) + T_2(x)$ has the same form as (31) and arises from a tree structure using binary splits on the explanatory variables.

The same sort of argument applies to $\mathbb{R}^2$. It is easy to see that in the two dimensional case $T(x) = T_1(x) + T_2(x)$ can be written in the form (31). What harder is to see that the regions defined by the $R_\ell^{(1)} \cap R_{\ell'}^{(2)}$ arise from binary splits on the entries of x. While harder, this is not hard: In the real plane, label the coordinates as x_1 and x_2 and let the tree structure of T_1 and T_2 be represented as partitions of $\mathbb{R}^2$. Pick the root note of, say, T_1. Without loss of generality, assume it is a split of the form $x_1 < c_1$ versus $x_1 \geq c_1$. Consider the left branch. The region $x_1 < c_1$ will be partitioned by horizontal lines, i.e., by ranges of x_2. Choose the largest cutoff for x_2, say c_2. Thus, splitting on $x_2 < c_2$ versus $x_2 \geq c_2$ will give us two daughter nodes on the left branch. We can then repeat this for each cutpoint on the left branch splitting at c_3, c_4 etc. The issue arises when the horizontal band represented by the split is itself split by some value of x_1. If this happens at, say c_4 then this simply adds another split that has to be carried over to the other splits on the left, i.e., for $x_2 < c_5$ versus $x_2 \geq c_5$, $x_2 < c_6$ versus $x_2 \geq c_6$ etc. as far down the left side of the tree as there are splits on x_2. If there are further splits, they can be accommodated in the same way and the argument can be applied analogously to the right branch of the tree. The same argument can be applied in three or more dimensions; it is simply a matter of considering splits on each dimension in turn and all the splits that may be performed on it using the other explanatory variables, in all possible sequences.

Now, if $C(\cdot)$ is a tree then its iterates $\hat{C}_1, \hat{C}_2, \ldots$ can be assumed to be trees as can the final output $BSTC(x)$. It is reasonable to expect the boosted classifier to be Bayes optimal. Indeed, Theorem 6 in Freund and Schapire (1997) gives that the probability of misclassification by a J-step boosted classifier $P(Y \neq C_{\text{Boost},J}(X))$ can only decrease as J increases. Separately, Biau et al. (2008) gives conditions under which some randomized *RF*-like and majority vote averaging classifiers achieve the minimal Bayes risk. (These are not pure *RF*'s because they ignore the decorrelation and choose split points randomly.) Moreover, since we are using trees here, and trees are a very rich class of nonparametric function estimators (in this case classification functions), it is safe to assume that there are trees that are arbitrarily close to the Bayes classifier even if they are not the same as studied in Biau et al. (2008).

Suppose a Bayes optimal classifier $C_B(x)$ exists, i.e., there is a classifier C_B that achieves the minimal misclassification error, $\arg\min_{C \in S} P(Y \neq C(X))$ where S is the set of essentially all classifiers. Then,

$$C_B(x) = \arg \max_{r \in \{-1,1\}} P(Y = r \mid X = x)$$

for each x. We begin with a result that initiates a boosting procedure with a Bayes optimal classifier. Of course, we find that the boosting procedure is unable to improve an optimal classifier. This is no surprise. Indeed, it is an artificial hypothesis – why boost an optimal classifier? However, we will use this result in Corollary 4.1 below and remove the hypothesis.

Theorem 4.1: *Suppose a Bayes classifier is used as* C_0 *in a boosting procedure. Then, on average, in the limit of*

increasing n, the weights β_j in (30) *are identical positive constants, for appropriately chosen G_j's.*

Proof: It is easy to see that the first step of the boosting procedure gives

$$\begin{aligned}\text{err}_1 &= \frac{1}{n}\sum_{i=1}^{n} I(Y \neq \hat{C}_1) \to E_{(X,Y)}(I(Y \neq \hat{C}_1(X))) \\ &\to P(Y \neq C_B(X))\end{aligned}$$

because of the minimum in (29), provided $n \to \infty$ and $G_1 = G_{1,n}$ is increasing by invoking Theorem 6 Freund and Schapire (1997). So, as $n \to \infty$,

$$\beta_1 = \log\left(\frac{1-\text{err}_1}{\text{err}_1}\right) \to \log\left(\frac{1-P(Y \neq C_B(X))}{P(Y \neq C_B(X))}\right).$$

Now $w_i \propto e^{\beta_1}$ if $y_i \neq \hat{C}_1(x_i)$ and $w_i \propto 1$ if $y_i = \hat{C}_1(x_i)$ for $i = 1, \ldots, n$. Also, $P(Y \neq C_B(X)) < 1/2$, we have $e^{\beta_1} > 1$. So, the first iteration w_i's are derived from the β_1's and the number and indices of the misclassifications. Out of n data points there will be asymptotically $nP(Y \neq C_B(X))$ misclassifications and they will occur randomly over the n data points.

Thus, from examining (30), any instance of the distribution D_1 randomly permutes the locations of the e^{β_1} and '1' entries, but the number of each type will be asymptotically constant. So, the first step optimization will again, on average, lead to the Bayes classifier: The misclassifications of the Bayes classifier are randomly located and spread uniformly over the occurrences of e^{β_1} and '1' entries. Therefore the output $\hat{C}_1$ is on average the same as the initial classifier $C_0(x) = C_B(x)$. The only way another classifier could improve on C_B would be to have fewer misclassifications on the indices i that had e^{β_1} rather than one which is impossible (on average) because the locations are random.

The same reasoning applies to step, $j = 2$. We get that the same proportion of observations have the weights $w_i \propto e^{\beta_1}$ and '1'. Hence, at the end of this step we still get that as $n \to \infty$,

$$\beta_2 = \log\left(\frac{1-\text{err}_2}{\text{err}_2}\right) \to \log\left(\frac{1-P[Y \neq C_B(X)]}{P[Y \neq C_B(X)]}\right)$$

on average, $w_i \propto e^{\beta_2} = e^{\beta_1}$ if $y_i \neq \hat{C}_2(x_i)$, and $w_i \propto 1$ if $y_i = \hat{C}_2(x_i)$ for $i = 1, \ldots, n$. Again, $e^{\beta_2} > 1$ with the number of misclassifications the same as before and the locations of the misclassifications permuted randomly. Hence D_2 is unchanged on average and $\hat{C}_2$ is essentially C_B, as before.

If we continue this process for steps $j = 3, \ldots, J$, we have in the limit

$$\beta_1, \beta_2, \ldots, \beta_J \to \log\left(\frac{1-P(Y \neq C_B(X))}{P(Y \neq C_B(X))}\right) > 0,$$

as $n \to \infty$, on average in $P_{X,Y}$. ■

Now, we get a corollary by initializing a boosting procedure with a 'weak' classifier.

Corollary 4.1: *Let $C_0(\cdot)$ be weak classifier with $P(Y \neq C_0(X)) < 1/2$ (i.e., as a classifier, $C_0(X)$ is better than a coin toss). Then as $n \to \infty$, the output of the AdaBoost algorithm is a majority vote of a sum of trees, i.e., a 'forest'.*

Remark: The output of Adaboost is only a forest, a collection of trees, not a *random* forest. In fact, the output of boosting is the sign of a sum of a weighted sequence of trees. As seen above, this is a tree. That is, as a weighted sum of functions, the majority vote of the individual trees from boosting is the same as the output of a single tree. In the same spirit, a random forest is a weighted sum of trees and therefore can be represented as a single tree if desired. Since Adaboost and *RF*'s are good – essentially Bayes classifiers – we expect the two should be close to each other as functions of their inputs. Even though the tree from boosting is a *RF*-like classifier, not actually a *RF*, we can still compare it to an actual *RF* classifier as in (27).

Proof: Recall that the boosting classifier is asymptotically Bayes optimal because it is a greedy approximation to the relative classification rate, see Le and Clarke (2018) Theorem 3.5, cf. Theorem 6 in Freund and Schapire (1997). So, for J large enough the iterates from the boosting procedure are Bayes optimal asymptotically. Thus, by Theorem 4.1 the β_j's converge to the same positive constant and for large enough J, the latter terms in *BSTC* will dominate to give the limit. That is,

$$\begin{aligned}BSTC(x) &= \text{sign}\left(\sum_{j=1}^{J} \beta_j \hat{C}_j(x)\right) \\ &\approx \text{sign}\left(\sum_{j=1}^{J} \hat{C}_j(x)\right) \\ &= \text{majority vote of } \{\hat{C}_j(x)\}|_{j=1}^{J} = RF^*(x),\end{aligned}$$

where RF^* is a sum of trees and the approximation improves as $n \to \infty$. ■

In view of (27), the Corollary means that if *BSTC* is a good approximation for *RF*'s asymptotically then *RF*'s will be well-approximated by an additive logistic regression model under the exponential loss. It agrees with the result in Le and Clarke (2018) showing that the risks of *RF*'s and boosting converge to the minimal Bayes's risk. Furthermore, we agree with Mease and Wuner (2008) and offer an argument supporting their claim that boosting does not overfit since *RF*'s are asymptotically equivalent to boosting, at least in a misclassification sense, and they do not overfit, see

Breiman (2001b). The new part here is using *BSTC* as an interpretation for *RF*'s and noting the cost of interpretation.

4.2. A more empirical interpretation for RF's

By construction, *RF*'s give a function $f_{\mathrm{RF}}(x_1, \ldots, x_p)$. So, suppose we also have a collection of functions $f_1, \ldots, f_K$ that we want to use as a way to 'interpret' f_{RF} by regression. Conditional on the data and f_k's, solving

$$\hat{w} = \arg \min_{w=(w_1,\ldots,w_K)} \sum_{i=1}^{n} \left(f_{\mathrm{RF}}(x_i) - \sum_{k=1}^{K} w_k f_k(x_i) \right)^2 \tag{32}$$

gives an approximation of $f_{\mathrm{RF}}(x)$,

$$\hat{Y}_{RF} = \sum_{k=1}^{K} \hat{w}_k f_k(x). \tag{33}$$

The $\hat{w}_k$'s may be found by using a standard least squares approach treating $f_{\mathrm{RF}}(x_i)$ as the Y_i's and the $f_k(x_i)$'s as K explanatory variables. More generally, since $x_i = (x_{i1}, \ldots, x_{ip})$ the f_k's can be regarded as the leading terms in a basis expansion, e.g., a Taylor expansion in p dimensions, cf. Section 2.2. It is easy to see that replacing *BSTC* in (27) by the right side of (33) gives the cost of interpreting f_{RF} in terms of its regression function (using the f_k's) by the residuals from (33). Whether the residuals are satisfactorily small can be assessed by a variety of established techniques.

Expression (32) can also be phrased in terms of logistic regression, which may be more appropriate for a classifier; just replace the inner sum in (32) by the corresponding expression from a logit in a selection of variables such as the f_i's and again minimize the error over choices of the parameters. This is not hard, but we have not done it for of ease of exposition.

5. Discussion

In this paper we have examined three classes of predictors – kernel methods, Shtarkov solutions, and random forests – and shown that, despite the inability to interpret them, they can be asymptotically approximated both theoretically and pragmatically by interpretable expressions. In each case we have given an expression that quantifies the cost of approximating the ideal but uninterpretable predictor by its interpretable expression. Consequently, up to approximation error, the hitherto uninterpretable expressions that for the most part did not permit physical inference have been manipulated into forms from which physical inference may be possible.

For the sake of completeness, it is important to discuss conformal prediction; see Shafer and Vovk (2008) for a comprehensive overview based on Vovk et al. (2005). Applications to exponential families and generalized linear models are given in Eck and Crawford (2019) and recent computational progress is given in Vovk et al. (2020). Essentially, conformal prediction assumes sequential data and that future data will resemble past data, i.e., the DG has enough stability that prediction is feasible.

Conformal prediction can also be regarded as an extension of Geisser (1975) by including the assumptions necessary for future data to look like past data and taking a probabilistic or stochastic processes approach to data analysis. At root is a non-conformity measure used to assess how close data points are; different non-conformity measures lead to different prediction regions. Thus, this framework has much in common with calibration and prequentialism, see Dawid (1984), and is more general than time series (Box–Jenkins or state space models). On the other hand, much contemporary sequential data will not fit into this framework. In any event, our work here is broadly consistent with conformal prediction even though our emphasis is on interpretability more than the stochastic properties of the DG.

Recall that our starting points was interpretability and in Section 1 we proposed a definition for interpretability. We did not distinguish between constructing an interpretable model pre-data or deriving an interpretable model post-data although it is clear the latter will generally be better justified if it is derived from an input–output relation that predicts well. Then, we took linear combinations of variables as the paradigm case for interpretable quantities. We did this in Sections 2.3 and 4.2. In Section 3.3 we allowed a broader interpretation – the interaction of variables and parameters were not linear but closely related enough that the effects could be readily seen. That is, in all three cases, we implicitly identified the c_k's with variables and parameters whose combination was explicit and properties could be queried.

An anonymous referee asked (i) if deep neural networks (DNN's) are interpretable and if so could they be used in place of the combinations of variables and parameters in Sections 2.3 and 4.2 and (ii) if reinforcement learning might provide an alternative interpretation to that in Section 3.3. for the Shtarkov solution. The referee also observed that there are cases where DNN's may perform better than kernel methods and RF's, and implicitly recognized that the Shtarkov solution, possibly with side information, was similar to reinforcement learning. Thus, if DNN's and reinforcement learning are interpretable why not simply start with the them and ignore these other methods?

First, the interpretability of DNN's is problematic. Defining a clear physical correlate for nodes, layers of nodes, types of layers of nodes, and connectivity will often be elusive because DNN's are usually overparametrized and may be mathematically distinct

even as they have very similar numerical properties as input–output relations. Some theorists have suggested that a DNN can be partitioned so that modules within it may have physical meaning or that reducing the number of neurons layer by layer might be akin to finding summary statistics. However, these suggestions remain conjectures. In short, it is not clear that DNN's are interpretable according to the definition used here even if sensitivity analyses can be used to understand the effect of parameters and variables on each other – a difficult task in practice for all but the smallest neural networks. If this holds then the kind of analysis done earlier, for kernel methods, say, to derive an interpretable approximation would have to be done for DNN's in order to assess what the DNN might have to say about a DG. That is, we are led to infer that the better performance of DNN's over other methods could be a result of their flexibility and associated non-interpretability, at least partially. Often, as a model becomes more complex or more general it becomes less interpretable. This does not contradict our basic assertion that interpretation has a cost in terms of prediction.

As to reinforcement learning, this is usually regarded as a sequential decision process in the context of discrete time, discrete space Markov processes. While transitions and actions may be interpretable, the analogy between reinforcement learning and the Shtarkov solution is not tight: The Shtarkov solution does not assume any distribution on the sequence of data points and arises from a minimization of regret while reinforcement learning finds an optimal action for each transition. It is reasonable to conjecture that some version of reinforcement learning will provide an approximation to Shtarkov in some settings, but investigating this is beyong our present scope.

Finally, we draw three implications from our results. First, from a pragmatic standpoint, we are arguing that, as a generality, models are at best only approximately true and the degree of approximation is usually unassessable. Proposed models can routinely be discredited by searching a more general class of predictors to get measurably better prediction. How can one assert a model is 'true' if its predictions can be improved? The consequence is that the other inferences from models taken as 'true' must be seen as unreliable absent further validation and assessment of their degree of mis-specification. In particular, physical interpretations are tentative at best.

Model mis-specification is an extensively studied topic, see Walker (2013) especially Sections 5 and 6, and the ensuing discussion. These authors generally focus on what we have called $\mathcal{M}$-complete problems and take this as being the typical setting for modelling and analysis. Accordingly, these authors try to characterize the inferential difference between whatever model used and the unknown but true model. Two recurring questions are: (1) Given that the true model class is unobtainable, what is it that we are making inferences about? and (2) Since the degree of mis-specification is important, how should we compare proposed models? These are addressed in a variety of ways by O'Hagan (2013), Hoff and Wakefield (2013) and De Blasi (2013).

Second, here we offer answers, perhaps unsatisfying, to these two questions. We argue that inferences should be about the next outcome, i.e., prediction, and that we should compare proposed models by how close the predictors they give are to the best predictor we can find. That is, prediction is the paradigm statistical problem, not estimation, interpretability, or other goals. Unless we have achieved good prediction there is no particular reason to trust other inferences. After all, from the falsification principle, it is unclear how to discredit estimates or the result of tests except by repeating an experiment, which is rarely done. The current term for this longstanding problem is the 'replicability crisis'. At root, requiring optimal prediction is a solution to problems with replicability: No predictive validation implies no valid inferences of any other sort.

Finally, we suggest as a default that experimenters achieve good prediction via optimal uninterpretable methods and then adapt the results (as much as they dare) to make them interpretable. The loss of predictive power as a consequence of constructing a physical interpretation can then be quantified and its cost assessed. In this way a simplified, and possibly interpretable, model may be validated up to a measurable degree of predictive loss. The cost of interpretability is then a limit on the validity of the model much as a standard error is a limit on the certainty of an estimate. We conclude with what might be called the *prediction principle*: The degree of predictive success a method has determines the reliability of any interpretation that rests on it.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Bernardo, J., & Smith, A. F. M. (2000). *Bayesian theory*. John Wiley & Sons.

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, *9*(66), 2015–2033. https://doi.org/10.1145/1390681.1442799

Bierens, H. (2005). *Introduction to the mathematical and statistical foundations of econometrics*. Cambridge University Press.

Billingsley, P. (2012). *Probability and measure*. Wiley.

Breiman, L. (2001a). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/BF00117832

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

Cesa-Bnachi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Clarke, B. (2007). Information optimality and Bayesian modelling. *Journal of Econometrics*, *138*(2), 405–429. https://doi.org/10.1016/j.jeconom.2006.05.003

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society, Series A 147*(2), 278–292. https://doi.org/10.2307/2981683

Dawid, A. P. (1992). Prequential data analysis. In: M. Ghosh & P. K. Pathak (Eds.), *Current issues in statistical inference: Essays in Honor of D. Basu* (pp. 113–126). IMS Lecture Notes Monograph Ser. 17. Institute of Mathematical Statistics.

Dawid, A. P. (2010). Fundamentals of prequential analysis. http://www3.stat.sinica.edu.tw/2013frontiers/presentation/29.pdf

Dawid, A. P., & Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli*, *5*(1), 125–162. https://doi.org/10.2307/3318616

De Blasi, P. (2013). Discussion on article 'Bayesian inference with misspecified models' by Stephen G. Walker. *Journal of Statistical Planning and Inference*, *143*(10), 1634–1637. https://doi.org/10.1016/j.jspi.2013.05.015

Diaconis, P., Goel, S., & Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, *2*(3), 777–807. https://doi.org/10.1214/08-AOAS165

Eck, D., & Crawford, F. (2019). Efficient and minimal length parametric conformal prediction regions. https://arxiv.org/pdf/1905.03657.pdf

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 1(119), 139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J., Hastie, T., & Tibshirani, R. (2000). A statistical view of boosting. *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of statistical learning* (2nd ed.). Springer.

Hoff, P., & Wakefield, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters: A discussion of 'Bayesian inference with misspecified models' by Stephen Walker. *Journal of Statistical Planning and Inference*, *143*(10), 1638–1642. https://doi.org/10.1016/j.jspi.2013.05.014

Kimeldorf, G., & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*(1), 82–95. https://doi.org/10.1016/0022-247X(71)90184-3

Le, T. M., & Clarke, B. (2016a). Using the Bayesian Shtarkov solution for predictions. *Computational Statistics & Data Analysis*, *104*(9), 183–196. https://doi.org/10.1016/j.csda.2016.06.018

Le, T. M., & Clarke, B. (2016b). A Bayes interpretation of stacking for $\mathcal{M}$-complete and $\mathcal{M}$-open settings. *Bayesian Analysis*, *12*(3), 807–829.https://doi.org/10.1214/16-BA1023

Le, T. M., & Clarke, B. (2018). On the interpretation of ensemble classifiers in terms of Bayes classifiers. *Journal of Classification*, *35*(2), 198–229. https://doi.org/10.1007/s00357-018-9257-y

Le, T. M., & Clarke, B. (2020). In praise of partially interpretable predictors. *Statistical Analysis and Data Mining*, *13*(2), 113–133. https://doi.org/10.1002/sam.v13.2

Liyang, Z., & Lee, Y. (2013). Eigen-analysis of nonlinear PCA with polynomial kernels. *Statistical Analysis and Data Mining*, *6*(6), 529–544. https://doi.org/10.1002/sam.11211

Mease, D., & Wuner, A. (2008). Evidence contrary to the statistical view of boosting. *The Journal of Machine Learning Research*, *9*(6), 131–156.https://doi.org/10.1145/1390681.1390687

Newey, W., & McFadden, D. (1994). *Large sample estimation and hypothesis testing*. Elsevier Science.

O'Hagan, A. (2013). Bayesian inference with misspecified models: Inference about what? *Journal of Statistical Planning and Inference*, *143*(10), 1643–1648. https://doi.org/10.1016/j.jspi.2013.05.016

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A*, (186), 343–414. https://doi.org/10.1098/rsta.1895.0010

Polson, N. G., Scott, J. G., & Windle, J. (2014). The Bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(4), 713–733. https://doi.org/10.1111/rssb.2014.76.issue-4

Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, *5*(June 1990), 197–227. https://doi.org/10.1007/BF00116037

Scholkopf, B., & Smola, A. (2002). *Learning with kernels*. MIT Press.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *The Journal of Machine Learning Research*, *9*(12), 371–421.

Shi, T., Belkin, M., & Yu, B. (2008). Data spectroscopy: Learning mixture models using eigenspaces of convolution operators. In Andrew McCallum & Sam Roweis (Eds.), *Proceedings of the 25th Annual International Conference on Machine Learning* (pp. 936–953). University of Cambridge.

Shtarkov, Y. (1987). Universal sequential coding of single messages. *Problems in Information Transmission*, *23*(3), 3–17.

Steyn, H. (1960). On regression properties of multivariate probability functions of Pearso's types. *Proceedings of the Royal Academy of Sciences*, *63*, 302–311. https://doi.org/10.1016/S1385-7258(60)50038-2

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*(June 2001), 211–244. https://doi.org/10.1162/15324430152748236.

Van de Geer, S. (2000). *Applications of empirical process theory*. Cambridge University Press.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Vovk, V., Petej, I., Nouretdinov, I., Manokhin, V., & Gammerman, A. (2020). Computationally efficient versions of conformal predictive distributions. *Neurocomputing*, *397*(July 2020), 292–308. https://doi.org/10.1016/j.neucom.2019.10.110

Walker, S. G. (2013). Bayesian inference with misspecified models, with discussion and rejoinder. *Journal of Statistical Planning and Inference*, *143*(10), 1621–1633. https://doi.org/10.1016/j.jspi.2013.05.013

Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Mathematical Journal*, *23*(2), 189–207. https://doi.org/10.1215/S0012-7094-56-02317-1

Wyner, A., Olsen, M., & Bleich, J. (2017). Explaining the success of AdaBoost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, *18*(May 2017), 1–33.

Xie, Q., & Barron, A. R. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, *46*(2), 431–445. https://doi.org/10.1109/18.825803

Appendices

Appendix 1. Details of some proofs

Proof: To see (7), recall Theorem 2.1 gives

$$[\hat{Y}_{\text{rep}}(x) - f_0(x)]^2 \xrightarrow{P} 0 \quad \text{as } n \to \infty. \tag{A1}$$

More explicitly, given $\mathcal{D}$, using (3) the left-hand side of (A1) is

$$\left[\sum_{i=1}^{n} \alpha_i K(X_i, x) - f_0(x)\right]^2, \tag{A2}$$

where $\alpha_i = \alpha_i(\mathcal{D})$. Cauchy-Schwarz gives that (A2) is bounded by

$$\begin{aligned}
&2\left[\sum_{i=1}^{n} \alpha_i K(X_i, x)\right]^2 + 2f_0^2(x) \\
&\quad \le 4\left[\sum_{i=N}^{n} \alpha_i K(X_i, x)\right]^2 + 4\left[\sum_{i=1}^{N-1} \alpha_i K(X_i, x)\right]^2 + 2f_0^2(x) \\
&\quad \le 4(n-N+1)\left(\sum_{i=N}^{n} \alpha_i^2\right)\left[\frac{1}{n-N+1}\sum_{i=N}^{n} K^2(X_i, x)\right] \\
&\qquad + 4\left[\sum_{i=1}^{N-1} \alpha_i K(X_i, x)\right]^2 + 2f_0^2(x).
\end{aligned} \tag{A3}$$

Next, we show that the term $1/(n-N+1)\sum_{i=N}^{n} K^2(x_i, x)$ on the right-hand side of (A3) is uniformly integrable. Begin by noting that Jensen's inequality gives

$$\left(\sum_{i=N}^{n} a_i\right)^{1+\epsilon} \le (n-N+1)^{\epsilon}\left(\sum_{i=N}^{n} a_i^{1+\epsilon}\right), \tag{A4}$$

for any $\epsilon > 0$ and $a_i \ge 0$, $i = N, \ldots, n$. (Set $\varphi(x) = x^{1+\epsilon}$.) Now, by (A4)

$$\begin{aligned}
&\sup_n E\left[\frac{1}{n-N+1}\sum_{i=N}^{n} K^2(X_i, x)\right]^{1+\epsilon} \\
&\quad \le \sup_n \frac{1}{(n-N+1)^{1+\epsilon}} \\
&\qquad \times E\left[(n-N+1)^{\epsilon}\sum_{i=N}^{n} K^{2(1+\epsilon)}(X_i, x)\right] \\
&\quad \le E[K^{2(1+\epsilon)}(X, x)] < \infty,
\end{aligned}$$

by Assumption (i). Thus, $1/(n-N+1)\sum_{i=N}^{n} K^2(x_i, x)$ is uniformly integrable.

Assumption (ii) gives, as $n \to \infty$, $(n-N+1) \times \sum_{i=N}^{n} \alpha_i^2 = o_P(1)$ and is bounded. So, the right-hand side of (A3) is uniformly integrable in $P_{(X,Y)}$. This implies

$$[\hat{Y}_{\text{rep}}(x) - f_0(x)]^2$$

is uniformly integrable for any x and by (A1) has limit zero in probability. Therefore,

$$E[\hat{Y}_{\text{rep}}(x) - f_0(x)]^2 \to 0,$$

as $n \to \infty$ by the Theorem 25.12 in Billingsley (2012), concluding the proof. ■

Proof: From (4), (5), and using the fact that

$$\begin{aligned}
\sup_x |f(x)| &\le \max\left\{\left|\sup_x f(x)\right|, \left|\inf_x f(x)\right|\right\} \\
&\le \left|\sup_x f(x)\right| + \left|\inf_x f(x)\right|,
\end{aligned}$$

we have

$$\begin{aligned}
&\sup_{f \in B(f^*,\delta)} |\hat{Q}_n(f) - Q_0(f)| \\
&\quad = \sup_{f \in B(f^*,\delta)} \left|\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) - E_{(X,Y)} L(Y, f(X))\right| \\
&\quad \le \left|\sup_{f \in B(f^*,\delta)} \left\{\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) - E_{(X,Y)} L(Y, f(X))\right\}\right| \\
&\qquad + \left|\inf_{f \in B(f^*,\delta)} \left\{\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) - E_{(X,Y)} L(Y, f(X))\right\}\right|.
\end{aligned} \tag{A5}$$

To bound the terms on the right-hand side of (A5), since $E_{(X,Y)} \sup_{f \in B(f^*,\delta)} L(Y, f(X)) < \infty$, note that

$$\begin{aligned}
&\sup_{f \in B(f^*,\delta)} \left\{\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) - E_{(X,Y)} L(Y, f(X))\right\} \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} \sup_{f \in B(f^*,\delta)} L(y_i, f(x_i)) - \inf_{f \in B(f^*,\delta)} E_{(X,Y)} L(Y, f(X)) \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} \sup_{f \in B(f^*,\delta)} L(y_i, f(x_i)) - E_{(X,Y)} \inf_{f \in B(f^*,\delta)} L(Y, f(X)) \\
&\quad \le \left|\frac{1}{n}\sum_{i=1}^{n} \sup_{f \in B(f^*,\delta)} L(y_i, f(x_i)) - E_{(X,Y)} \sup_{f \in B(f^*,\delta)} L(Y, f(X))\right| \\
&\qquad + E_{(X,Y)} \sup_{f \in B(f^*,\delta)} L(Y, f(X)) - E_{(X,Y)} \inf_{f \in B(f^*,\delta)} L(Y, f(X)) \\
&\quad \le \left|\frac{1}{n}\sum_{i=1}^{n} \sup_{f \in B(f^*,\delta)} L(y_i, f(x_i)) - E_{(X,Y)} \sup_{f \in B(f^*,\delta)} L(Y, f(X))\right| \\
&\qquad + \left|\frac{1}{n}\sum_{i=1}^{n} \inf_{f \in B(f^*,\delta)} L(y_i, f(x_i)) \right. \\
&\qquad \left. - E_{(X,Y)} \inf_{f \in B(f^*,\delta)} L(Y, f(X))\right| \\
&\qquad + E_{(X,Y)} \sup_{f \in B(f^*,\delta)} L(Y, f(X)) \\
&\qquad - E_{(X,Y)} \inf_{f \in B(f^*,\delta)} L(Y, f(X)).
\end{aligned} \tag{A6}$$

Similarly, we have

$$\begin{aligned}
&\inf_{f\in B(f^*,\delta)}\left\{\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))-E_{(X,Y)}L(Y,f(X))\right\}\\
&\quad\geq\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))-\sup_{f\in B(f^*,\delta)}E_{(X,Y)}L(Y,f(X))\\
&\quad\geq\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\\
&\quad\geq-\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\\
&\quad\geq-\left|\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad-\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X)).
\end{aligned}\tag{A7}$$

Therefore, from (A6) and (A7), the first term on the RHS of (A5) is bounded by

$$\begin{aligned}
&\left|\sup_{f\in B(f^*,\delta)}\left\{\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))-E_{(X,Y)}L(Y,f(X))\right\}\right|\\
&\quad\leq\left|\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X)),
\end{aligned}\tag{A8}$$

$$\begin{aligned}
&\left|\inf_{f\in B(f^*,\delta)}\left\{\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))-E_{(X,Y)}L(Y,f(X))\right\}\right|\\
&\quad\leq\left|\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X)).
\end{aligned}\tag{A9}$$

Combining (A5), (A8), and (A9), we get

$$\begin{aligned}
&\sup_{f\in B(f^*,\delta)}\left|\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))-E_{(X,Y)}L(Y,f(X))\right|\\
&\quad\leq 2\left|\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in B(f^*,\delta)}L(y_i,f(x_i))\right.\\
&\qquad\left.-E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+2\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f^*,\delta)}L(y_i,f(x_i))\right.\\
&\qquad\left.-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right|\\
&\qquad+2\left[E_{(X,Y)}\sup_{f\in B(f^*,\delta)}L(Y,f(X))\right.\\
&\qquad\left.-E_{(X,Y)}\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right].
\end{aligned}\tag{A10}$$

It follows from the continuity of L in f and the dominated convergence theorem that the third term on the right-hand side of (A9) satisfies

$$\begin{aligned}
&\lim_{\delta\to 0}\sup_{f^*\in\mathcal{H}_K}E_{(X,Y)}\left[\sup_{f\in B(f^*,\delta)}L(Y,f(X))-\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right]\\
&\quad\leq\lim_{\delta\to 0}E_{(X,Y)}\sup_{f^*\in\mathcal{H}_K}\left[\sup_{f\in B(f^*,\delta)}L(Y,f(X))\right.\\
&\qquad\left.-\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right]=0,
\end{aligned}$$

and hence we can choose δ so small that

$$\sup_{f^*\in\mathcal{H}_K}E_{(X,Y)}\left[\sup_{f\in B(f^*,\delta)}L(Y,f(X))-\inf_{f\in B(f^*,\delta)}L(Y,f(X))\right]<\epsilon.\tag{A11}$$

Furthermore, by the compactness of $\mathcal{D}_j$, there exist finitely many of f^*'s, say $f_1,\ldots,f_{N(\delta)}$, such that $\mathcal{D}_j\subset\bigcup_{i=1}^{N(\delta)}B(f_i,\delta)$. Hence, by the union of events bound,

$$\begin{aligned}
&P\left(\sup_{f\in\mathcal{D}_j}\left|\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))-E_{(X,Y)}L(Y,f(X))\right|>\epsilon\right)\\
&\quad\leq P\left(\max_{1\leq i\leq N(\delta)}\sup_{f\in B(f_i,\delta)}\left|\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))\right.\right.\\
&\qquad\left.\left.-\ E_{(X,Y)}L(Y,f(X))\right|>\epsilon\right)\\
&\quad\leq\sum_{i=1}^{N(\delta)}P\left(\sup_{f\in B(f_i,\delta)}\left|\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))\right.\right.\\
&\qquad\left.\left.-\ E_{(X,Y)}L(Y,f(X))\right|>\epsilon\right).
\end{aligned}$$

Using (A10) and (A11) this expression is bounded by

$$\begin{aligned}
&\sum_{i=1}^{N(\delta)}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in B(f_i,\delta)}L(y_i,f(x_i))-E_{(X,Y)}\sup_{f\in B(f_i,\delta)}L(Y,f(X))\right|\right.\\
&\qquad+\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f_i,\delta)}L(y_i,f(x_i))\right.\\
&\qquad\left.\left.-\ E_{(X,Y)}\inf_{f\in B(f_i,\delta)}L(Y,f(X))\right|>\frac{\epsilon}{2}\right)
\end{aligned}$$

$$\leq \sum_{i=1}^{N(\delta)} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in B(f_i,\delta)} L(y_i,f(x_i)) - E_{(X,Y)}\sup_{f\in B(f_i,\delta)} L(Y,f(X))\right| > \frac{\epsilon}{4}\right) + \sum_{i=1}^{N(\delta)} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\inf_{f\in B(f_i,\delta)} L(y_i,f(x_i)) - E_{(X,Y)}\inf_{f\in B(f_i,\delta)} L(Y,f(X))\right| > \frac{\epsilon}{4}\right),$$

which goes to 0 as $n \to \infty$ by the weak law of large numbers, concluding the proof by Assumption (iii). ■

Appendix 2. Interpretability versus complexity

Interpretability is a different concept from complexity. We have implicitly assumed that the best predictors (and models) are highly complex, but we regard this as the most common case in current practice, not a priori true. In point of fact, an interpretable model may be simple or complex and an uninterpretable model may be simple or complex. Otherwise put, all pairs of (interpretability, complexity) can occur. As a generality, $\mathcal{M}$-closed problems are less complex than $\mathcal{M}$-complete problems and they in turn are less complex than $\mathcal{M}$-open problems. However, this ordering does not in general preclude the existence of an interpretable predictor or model for any problem.

Here, we use the notion of interpretability in Le and Clarke (2020). On the other hand, here complexity refers to how many components are required for good prediction or modelling. This is different from other notions of complexity such as VC dimension or code length since these do not require any components of a predictor or model to have any physical correlates.

Nevertheless, as an empirical observation we have noted that the more interpretability one demands of a model, the more complex it will typically be and, often, the more complex the true model is, the higher the model misspecification will be. Correspondingly, the less interpretability one requires, the smaller the error of the predictor can be but it is not clear what effect this has on complexity. Empirically, the predictions from an interpretable model are likely to be worse than the prediction from a well chosen non-interpretable model or predictor. This arises for the intuitive reason that restricting model classes to only those that are interpretable is likely to increase bias because real world phenomena are rarely captured to infinite precision by what we think are physically meaningful models. Because interpretable models may be simplifications of the real phenomena we expect some of the bias to be reflected in increased variance as well. The exception is when the model actually is valid to infinite precision or at least to a precision higher than that achieved by other models; this can happen but is atypical. These observations are neither new nor surprising. The issue is to quantify them to ascertain how much of an interpretation one can derive from an uninterpretable model without losing too much accuracy of prediction.