Explaining Predictive Uncertainty with Information Theoretic Shapley Values

David S. Watson[1*], Joshua O'Hara[1], Niek Tax[2], Richard Mudd[2], and Ido Guy[2]

[1]Department of Informatics, King's College London

[2]Meta Core Data Science

[*]Corresponding author: David S. Watson; david.watson@kcl.ac.uk

Abstract

Researchers in explainable artificial intelligence have developed numerous methods for helping users understand the predictions of complex supervised learning models. By contrast, explaining the *uncertainty* of model outputs has received relatively little attention. We adapt the popular Shapley value framework to explain various types of predictive uncertainty, quantifying each feature's contribution to the conditional entropy of individual model outputs. We consider games with modified characteristic functions and find deep connections between the resulting Shapley values and fundamental quantities from information theory and conditional independence testing. We outline inference procedures for finite sample error rate control with provable guarantees, and implement an efficient algorithm that performs well in a range of experiments on real and simulated data. Our method has applications to covariate shift detection, active learning, feature selection, and active feature-value acquisition.

1 Introduction

Machine learning (ML) algorithms can solve many prediction tasks with greater accuracy than classical methods. However, some of the most popular and successful algorithms, such as deep neural networks, often produce models with millions of parameters and complex nonlinearities. The resulting "black box" is essentially unintelligible to humans. Researchers in explainable artificial intelligence (XAI) have developed numerous methods to help users better understand the inner workings of such models (see Sect. 2).

Despite the rapid proliferation of XAI tools, the goals of the field have thus far been somewhat narrow. The vast majority of methods in use today aim to explain model *predictions*, i.e. point estimates. But these are not necessarily the only model output of interest. Predictive *uncertainty* can also vary widely across the feature space, in ways that may impact model performance and human decision making. Such variation makes it risky to rely on the advice of a black box, especially when generalizing to new environments. Discovering the source of uncertainty can be an important first step toward reducing it.

Quantifying predictive uncertainty has many applications in ML. For instance, it is an essential subroutine in any task that involves exploration, e.g. active learning [18, 34], multi-armed bandits [66, 37], and reinforcement learning more generally [50, 69]. Other applications of predictive uncertainty quantification include detecting covariate shift [70] and adversarial examples [64], as well as classification with reject option [23]. Our method aims to expand the scope of XAI to these varied domains by explaining predictive uncertainty via feature attributions.

Knowing the impact of individual features on local uncertainty can help drive data collection and model design. It can be used to detect the source of a suspected covariate shift, select informative features, and test for heteroskedasticity. Our attribution strategy makes use of the Shapley value framework for XAI [41, 67, 45, 9], a popular approach inspired by cooperative game theory, which we adapt by altering the characteristic function and augment with inference procedures for provable error rate control. The approach is fully model-agnostic and therefore not limited to any particular function class.

Our main contributions are threefold: (1) We describe modified variants of the Shapley value algorithm that can explain higher moments of the predictive distribution, thereby extending its explanatory utility beyond mere point estimates. We provide an information theoretic interpretation of the resulting measures and study their properties. (2) We introduce a split conformal inference procedure for Shapley variables with finite sample coverage guarantees. This allows users to test the extent to which attributions for a given feature are concentrated around zero with fixed type I error control. (3) We implement model-specific and model-agnostic variants of our method and illustrate their performance in a range of simulated and real-world experiments, with applications to feature selection, covariate shift detection, and active learning.

2 Related Work

XAI has become a major subfield of machine learning in recent years. The focus to date has overwhelmingly been on explaining predictions in supervised learning tasks, most prominently via feature attributions [58, 41, 68], rule lists [59, 36, 65], and counterfactuals [76, 46, 31]. Despite obvious differences between these methods, all arguably share the same goal of identifying minimal conditions sufficient to alter predictions in some pre-specified way [78].

Quantifying inductive uncertainty is a fundamental problem in probability theory and statistics, although machine learning poses new challenges and opportunities in this regard [28]. The classical literature on this topic comes primarily from Bayesian modeling [21] and information theory [13], which provide a range of methods for analyzing the distribution of random variables. More recent work on conformal inference [75, 38, 4] has expanded the toolkit for practitioners.

Important application areas for these methods include active learning (AL) and covariate shift detection. In AL, the goal is to selectively query labels for unlabeled instances aimed to maximize classifier improvement under a given query budget. Methods often select instances on which the model has high epistemic (as opposed to aleatoric) uncertainty [61], as for example in BatchBALD [34]. This is especially valuable when labels are sparse and costly to collect while unlabeled data is widely available. In covariate shift detection, the goal is to identify samples that are abnormal relative to the in-distribution observations that the classifier has seen during training. It is well-known that neural networks can be overconfident [22], yielding predictions with unjustifiably high levels of certainty on test samples. Addressing this issue is an active area of research, and a variety of articles take a perspective of quantifying epistemic uncertainty [28, 51]. In safety-critical applications, the degree of model uncertainty can be factored into the decision making, for example by abstaining from prediction altogether when confidence is sufficiently low [23].

Very little work has been done on explaining predictive uncertainty. A notable exception is the CLUE algorithm [3, 39], a model-specific method designed for Bayesian deep learning, which generates counterfactual samples that are maximally similar to some target observation but optimized for minimal conditional variance. This contrasts with our feature attribution approach, which is model-agnostic and thereby the first to make uncertainty explanations available to function classes beyond Bayesian deep learning models.

Predictive uncertainty is often correlated with prediction loss, and therefore explanations of model errors are close relatives of our method. LossSHAP [42] is an extension of Lundberg and Lee [41]'s SHAP algorithm designed to explain the pointwise loss of a supervised learner (e.g., squared error or cross entropy). Though this could plausibly help identify regions where the model is least certain about predictions, it requires a large labelled test dataset, which may not be available in practice. By contrast, our method only assumes access to some unlabelled dataset of test samples, which is especially valuable when labels are slow or expensive to collect. For instance, LossSHAP is little help in learning environments where covariate shift is detectable before labels are known [82]. This is common, for example, in online advertising [35], where an impression today may lead to a conversion next week but quick detection (and explanation) of covariate shift is vital.

Previous authors have explored information theoretic interpretations of variable importance measures; see [16, Sect. 8.3] for a summary. These methods often operate at global resolutions—e.g., Sobol' indices [52] and SAGE [15]—whereas our focus is on local explanations. Alternatives such as INVASE [79] must be trained alongside the supervised learner itself and are therefore not model-agnostic. L2X [10] and REAL-X [30] provide post-hoc local explanations, but they require surrogate models to approximate a joint distribution over the full feature space. Chen et al. [11] propose an information theoretic variant of Shapley values for graph-structured data, which we examine more closely in Sect. 4.

3 Background

Notation. We use uppercase letters to denote random variables (e.g., X) and lowercase for their values (e.g., x). Matrices and sets of random variables are denoted by uppercase boldface type (e.g., $\mathbf{X}$) and vectors by lowercase boldface (e.g., $\mathbf{x}$). We occasionally use superscripts to denote samples, e.g. $\mathbf{x}^{(i)}$ is the i^{th} row of $\mathbf{X}$. Subscripts index features or subsets thereof, e.g. $\mathbf{X}_S = \{X_j\}_{j \in S}$ and $\mathbf{x}_S^{(i)} = \{x_j^{(i)}\}_{j \in S}$, where $S \subseteq [d] = \{1, \ldots, d\}$. We define the complementary subset $\overline{S} = [d] \backslash S$.

Information Theory. Let p, q be two probability distributions over the same σ-algebra of events. (In the continuous case, we additionally require that p, q be absolutely continuous with respect to Lebesgue measure.) We make use of several fundamental quantities from information theory [13], such as entropy $H(p)$, cross entropy $H(p, q)$, KL-divergence $D_{KL}(p \parallel q)$, and mutual information $I(X; Y)$ (all formally defined in Appx. B.1). We use shorthand for the conditional probability mass/density function of the random variable Y, e.g. $p_{Y|\mathbf{x}_S} := p(Y \mid \mathbf{X}_S = \mathbf{x}_S)$. We speak interchangeably of the entropy of a random variable and the entropy of the associated mass/density function: $H(Y \mid x) = H(p_{Y|x})$. We call this the *local* conditional entropy to distinguish it from its global counterpart, $H(Y \mid X)$, which requires marginalization over the joint space $\mathcal{X} \times \mathcal{Y}$.

Shapley Values. Consider a supervised learning model f trained on features $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ to predict outcomes $Y \in \mathcal{Y} \subseteq \mathbb{R}$. We assume that data are distributed according to some fixed but unknown distribution $\mathcal{D}$. Shapley values are a feature attribution method in which model predictions are decomposed as a sum: $f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi(j, \mathbf{x})$, where ϕ_0 is the baseline expectation (i.e., $\phi_0 = \mathbb{E}_{\mathcal{D}}[f(\mathbf{x})]$) and $\phi(j, \mathbf{x})$ denotes the Shapley value of feature j at point $\mathbf{x}$. To define this quantity, we require a value function $v : 2^{[d]} \times \mathbb{R}^d \mapsto \mathbb{R}$ that quantifies the payoff associated with subsets $S \subseteq [d]$ for a particular sample. This characterizes a cooperative game, in which each feature acts as a player. A common choice for defining payoffs in XAI is the following [9]:

$$v_0(S, \mathbf{x}) := \mathbb{E}_{\mathcal{D}}\big[f(\mathbf{x}) \mid \mathbf{X}_S = \mathbf{x}_S\big],$$

where we marginalize over the complementary features $\overline{S}$ in accordance with reference distribution $\mathcal{D}$. For any value function v, we may define the following random variable to represent j's marginal contribution to coalition S at point $\mathbf{x}$:

$$\Delta_v(S, j, \mathbf{x}) := v(S \cup \{j\}, \mathbf{x}) - v(S, \mathbf{x}).$$

Then j's Shapley value is just the weighted mean of this variable over all subsets:

$$\phi_v(j, \mathbf{x}) := \sum_{S \subseteq [d] \setminus \{j\}} \frac{|S|!\ (d - |S| - 1)!}{d!} \big[\Delta_v(S, j, \mathbf{x})\big]. \tag{1}$$

It is well known that Eq. 1 is the unique solution to the attribution problem that satisfies certain desirable properties, including efficiency, symmetry, sensitivity, and linearity [67] (for formal statements of these axioms, see Appx. B.2.)

4 Alternative Value Functions

To see how standard Shapley values can fall short, consider a simple data generating process with $X, Z \sim \mathcal{U}(0, 1)^2$ and $Y \sim \mathcal{N}(X, Z^2)$. Since the true conditional expectation of Y is X, this feature will get 100% of the attributions in a game with payoffs given by v_0. However, just because Z receives zero attribution does not mean that it adds no information to our predictions—on the contrary, we can use Z to infer the predictive variance of Y and calibrate confidence intervals accordingly. This sort of higher order information is lost in the vast majority of XAI methods.

We consider information theoretic games that assign nonzero attribution to Z in the example above, and study the properties of resulting Shapley values. We start in an idealized scenario in which we have: (i) oracle knowledge of the joint distribution $\mathcal{D}$; and (ii) unlimited computational budget, thereby allowing complete enumeration of all feature subsets.

INVASE [79] is a method for learning a relatively small but maximally informative subset of features $S \subset [d]$ using the loss function $D_{KL}(p_{Y|\mathbf{x}} \parallel p_{Y|\mathbf{x}_S}) + \lambda|S|$, where λ is a regularization penalty. While Yoon et al. [79] are not focused on Shapley-style explanations, we can take the first term of their loss function to define a new game:

$$v_{KL}(S, \mathbf{x}) := -D_{KL}(p_{Y|\mathbf{x}} \parallel p_{Y|\mathbf{x}_S}),$$

which can be interpreted as -1 times the excess number of bits one would need on average to describe samples from $Y \mid \mathbf{x}$ given code optimized for $Y \mid \mathbf{x}_S$.

Chen et al. [11] make a similar proposal, replacing KL-divergence with cross entropy:

$$v_{CE}(S, \mathbf{x}) := -H(p_{Y|\mathbf{x}}, p_{Y|\mathbf{x}_S}).$$

This value function is closely related to that of LossSHAP [42], which for likelihood-based loss functions can be written:

$$v_L(S, \mathbf{x}) := -\log p(Y = y \mid \mathbf{x}_S),$$

where y denotes the true value of Y at the point $\mathbf{x}$. As Covert et al. [16] point out, this is equivalent to the pointwise mutual information $I(y; \mathbf{x}_S)$, up to an additive constant. However, v_L requires true labels for Y, which may not be available when evaluating feature attributions on a test set. By contrast, v_{CE} averages over $\mathcal{Y}$, thereby avoiding this issue: $v_{CE}(S, \mathbf{x}) = -\mathbb{E}_{Y|\mathbf{x}}\big[v_L(S, \mathbf{x})\big]$. We reiterate that in all cases we condition on some fixed value of $\mathbf{x}$ and do not marginalize over the feature space $\mathcal{X}$. This contrasts with global feature attribution methods like SAGE [15], which can be characterized by averaging v_L over the complete joint distribution $p(\mathbf{X}, Y)$.

It is evident from the definitions that v_{KL} and v_{CE} are equivalent up to an additive constant not depending on S, namely $H(p_{Y|\mathbf{x}})$. This renders the resulting Shapley values from both games identical (all proofs in Appx. A.)

Proposition 4.1. *For all features $j \in [d]$, coalitions $S \subseteq [d] \backslash \{j\}$, and samples $\mathbf{x} \sim \mathcal{D}_X$:*

$$\begin{aligned}\Delta_{KL}(S, j, \mathbf{x}) &= \Delta_{CE}(S, j, \mathbf{x}) \\ &= \int_{\mathcal{Y}} p(y \mid \mathbf{x}) \log \frac{p(y \mid \mathbf{x}_S, x_j)}{p(y \mid \mathbf{x}_S)} \; dy.\end{aligned}$$

This quantity answers the question: if the target distribution were $p_{Y|\mathbf{x}}$, how many more bits of information would we get on average by adding x_j to the conditioning event $\mathbf{x}_S$? Resulting Shapley values summarize each feature contribution in bits to the distance between Y's fully specified local conditional distribution $p(Y \mid \mathbf{x})$ and the marginal $p(Y)$.

Proposition 4.2. *With $v \in \{v_{KL}, v_{CE}\}$, Shapley values satisfy $\sum_{j=1}^{d} \phi_v(j, \mathbf{x}) = D_{KL}(p_{Y|\mathbf{x}} \,\|\, p_Y)$.*

We introduce two novel information theoretic games, characterized by negative and positive local conditional entropies:

$$v_{IG}(S, \mathbf{x}) := -H(Y \mid \mathbf{x}_S), \quad v_H(S, \mathbf{x}) := H(Y \mid \mathbf{x}_S).$$

The former subscript stands for information gain; the latter for entropy. Much like v_{CE}, these value functions can be understood as weighted averages of LossSHAP payoffs over $\mathcal{Y}$, however this time with expectation over a slightly different distribution: $v_{IG}(S, \mathbf{x}) = -v_H(S, \mathbf{x}) = -\mathbb{E}_{Y|\mathbf{x}_S}\big[v_L(S, \mathbf{x})\big]$. The marginal contribution of feature j to coalition S is measured in bits of local conditional mutual information added or lost, respectively (note that $\Delta_{IG} = -\Delta_H$).

Proposition 4.3. *For all features $j \in [d]$, coalitions $S \subseteq [d] \backslash \{j\}$, and samples $\mathbf{x} \sim \mathcal{D}_X$:*

$$\begin{aligned}\Delta_{IG}(S, j, \mathbf{x}) &= I(Y; x_j \mid \mathbf{x}_S) \\ &= \int_{\mathcal{Y}} p(y \mid x_j, \mathbf{x}_S) \log \frac{p(y, x_j \mid \mathbf{x}_S)}{p(y \mid \mathbf{x}_S)\; p(x_j \mid \mathbf{x}_S)} \; dy.\end{aligned}$$

This represents the decrease in Y's uncertainty attributable to the conditioning event $X_j = x_j$ when we already know that $\mathbf{X}_S = \mathbf{x}_S$. This quantity is similar (but not quite equivalent) to the *information gain*, a common optimization objective in tree growing algorithms [55, 56]. The difference again lies in the fact that we do not marginalize over $\mathcal{X}$, but instead condition on a single instance. Resulting Shapley values summarize each feature's contribution in bits to the overall local information gain.

Proposition 4.4. *Under v_{IG}, Shapley values satisfy $\sum_{j=1}^{d} \phi_{IG}(j, \mathbf{x}) = I(Y; \mathbf{x})$.*

These games share an important and complex relationship to conditional independence structures. We distinguish here between global claims of conditional independence, e.g. $Y \perp\!\!\!\perp X \mid Z$, and local or context-specific indepencies (CSI), e.g. $Y \perp\!\!\!\perp x \mid z$. The latter occurs when $X = x$ adds no information about Y under the conditioning event $Z = z$ [7] (see Appx. B.1 for an example).

Theorem 4.5. *For value functions $v \in \{v_{KL}, v_{CE}, v_{IG}, v_H\}$, we have:*

(a) $Y \perp\!\!\!\perp X_j \mid \mathbf{X}_S \Leftrightarrow \sup_{\mathbf{x} \in \mathcal{X}} |\Delta_v(S, j, \mathbf{x})| = 0.$
(b) $Y \perp\!\!\!\perp x_j \mid \mathbf{x}_S \Rightarrow \Delta_v(S, j, \mathbf{x}) = 0.$

(c) The set of distributions such that $\Delta_v(S, j, \mathbf{x}) = 0 \wedge Y \not\perp\!\!\!\perp x_j \mid \mathbf{x}_S$ *is Lebesgue measure zero.*

Item (a) states that Y is conditionally independent of X_j given $\mathbf{X}_S$ if and only if j makes no contribution to S at any point $\mathbf{x}$. Item (b) states that the weaker condition of CSI is sufficient for zero marginal payout. However, while the converse does not hold in general, item (c) states that the set of counterexamples is *small* in a precise sense—namely, it has Lebesgue measure zero. Measure zero events are not necessarily harmless, especially when working with finite samples. Near violations may in fact be quite common due to statistical noise [73]. Together, these results establish a powerful, somewhat subtle link between conditional independencies and information theoretic Shapley values. Similar results are lacking for the standard value function v_0—with the notable exception that conditional independence implies zero marginal payout [43]—an inevitable byproduct of the failure to account for predictive uncertainty.

5 Method

The information theoretic quantities described in the previous section are often challenging to calculate, as they require extensive conditioning and marginalization. Computing some $\mathcal{O}(2^d)$ such quantities per Shapley value, as Eq. 1 requires, quickly becomes infeasible. (See [74] for an in-depth analysis of the time complexity of Shapley value algorithms.) Therefore, we make several simplifying assumptions that strike a balance between computational tractability and error rate control.

First, we require some uncertainty estimator $h : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$. In the previous section, we assumed access to the true data generating process, which greatly simplifies theoretical analysis. In practice, we must train this model from finite samples, often using outputs from the base model f. In the regression setting, this may be a conditional variance estimator, as in heteroskedastic error models [32]; in the classification setting, we assume that f outputs a pmf over class labels and write $f_y : \mathbb{R}^d \mapsto [0, 1]$ to denote the predicted probability of class $y \in \mathcal{Y}$. Then predictive entropy is estimated via the plug-in formula $h_t(\mathbf{x}) := -\sum_{y\in\mathcal{Y}} f_y(\mathbf{x}) \log f_y(\mathbf{x})$, where the subscript t stands for *total*.

In many applications, we must decompose total entropy into epistemic and aleatoric components—i.e., uncertainty arising from the model or the data, respectively. We achieve this via ensemble methods, using a set of B basis functions, $\{f^1, \ldots, f^B\}$. These may be decision trees, as in a random forest [62], or subsets of neural network nodes, as in Monte Carlo (MC) dropout [20]. Let $f_y^b(\mathbf{x})$ be the conditional probability estimate for class y given sample $\mathbf{x}$ for the b^{th} basis function. Then aleatoric uncertainty is given by $h_a(\mathbf{x}) := -\frac{1}{B}\sum_{b=1}^{B}\sum_{y\in\mathcal{Y}} f_y^b(\mathbf{x}) \log f_y^b(\mathbf{x})$. Epistemic uncertainty is simply the difference [27], $h_e(\mathbf{x}) := h_t(\mathbf{x}) - h_a(\mathbf{x})$. Alternative methods may be appropriate for specific function classes, e.g. Gaussian processes [57] or Bayesian deep learning models [47]. We leave the choice of which uncertainty measure to explain up to practitioners. In what follows, we use the generic $h(\mathbf{x})$ to signify whichever estimator is of relevance for a given application.

We are similarly ecumenical regarding reference distributions. This has been the subject of much debate in recent years, with authors variously arguing that $\mathcal{D}$ should be a simple product of marginals [29]; or that the joint distribution should be modeled for proper conditioning and marginalization [1]; or else that structural information should be encoded to quantify causal effects [24]. Each approach makes sense in certain settings [8, 77], so we leave it up to practitioners to decide which is most appropriate for their use case. We stress that information theoretic games inherit all the advantages and disadvantages of these samplers from the conventional XAI setting, and acknowledge that attributions should be interpreted with caution when models are forced to extrapolate to off-manifold data [25].

In our experiments below, we use the independent sampling approach with KernelSHAP and DeepSHAP, which is the default in popular XAI software [41]. That is, we create synthetic points $\tilde{\mathbf{x}}$ by combining the coordinates of $\mathbf{x}_S$ with those of all observed $\mathbf{x}_{\overline{S}}$, thereby ignoring correlations between in- and out-of-coalition features. By contrast, when our target model is tree-based, we approximate conditionals by exploiting the model structure itself using a method called TreeSHAP [42]. Solutions based on copula methods [1] or variational autoencoders [49] have also shown promise, although these may not be appropriate with all data types; see [9] for a discussion.

Finally, we adopt standard methods to efficiently sample candidate coalitions. Observe that the distribution on subsets implied by Eq. 1 induces a symmetric pmf on cardinalities $|S| \in \{0, \ldots, d-1\}$ that places exponentially greater weight at the tails than it does at the center. Thus while there are over 500 billion coalitions at $d = 40$, we can cover 50% of the total weight by sampling just over 0.1% of these subsets (i.e., those with cardinality ≤ 9 or ≥ 30). To reach 90% accuracy requires just over half of all coalitions. Shapley values can therefore be well approximated via Monte Carlo using just a small fraction

of all subsets. We also employ the paired sampling approach of Covert and Lee [14] to reduce variance and speed up convergence still further.

Some authors propose Bayesian methods for statistical inference with Shapley values [63]. While these methods help quantify the uncertainty of model explanations, we are interested instead in an explanation of model uncertainty. We therefore seek inference procedures not for individual Shapley values, per se—although existing methods could in principle be extended to do so for our revised games—but rather for the random variable $\phi(j, \mathbf{x})$, which varies across samples but will tend to concentrate around zero for uninformative j. We take a conformal approach [75, 38] that provides the following finite sample coverage guarantee.

Theorem 5.1 (Coverage). *We partition n training samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n \sim \mathcal{D}$ into two equal-sized subsets $\mathcal{I}_1, \mathcal{I}_2$ where $\mathcal{I}_1$ is used for model fitting and $\mathcal{I}_2$ for computing Shapley values. Fix a target level $\alpha \in (0, 1)$ and define the mean Shapley value $\mu_j := \frac{2}{n} \sum_{i:\mathbf{x}^{(i)} \in \mathcal{I}_2} \phi(j, \mathbf{x}^{(i)})$, with conformity score $\delta_j^{(i)} := |\phi(j, \mathbf{x}^{(i)}) - \mu_j|$. Let τ_j be the qth smallest value in δ_j, for $q = \lceil (n/2 + 1)(1 - \alpha) \rceil$. Then for any test sample $\mathbf{x}^{(n+1)} \sim \mathcal{D}$, we have:*

$$\mathbb{P}\big(\phi(j, \mathbf{x}^{(n+1)}) \in \mu_j \pm \tau_j\big) \geq 1 - \alpha.$$

Moreover, if conformity scores have a continuous joint distribution, then the upper bound on this probability is $1 - \alpha + 2/(n + 2)$.

Note that this is not a *conditional* coverage claim, insomuch as the margin τ_j is fixed for a given X_j and does not vary with other feature values. However, Thm. 5.1 provides a PAC-style guarantee that Shapley values do not exceed a given (absolute) threshold with high probability, or that zero falls within the $(1 - \alpha) \times 100\%$ confidence interval for a given Shapley value. These results can inform decisions about feature selection, since narrow intervals around zero are necessary (but not sufficient) evidence of uninformative predictors. This result is most relevant for tabular or text data, where features have some consistent meaning across samples; it is less applicable to image data, where individual pixels have no stable interpretation over images.

6 Experiments

Full details of all datasets and hyperparameters can be found in Appx. C, while code for all experiments and figures can be found online.[1] Since our goal is to explain predictive *entropy* rather than *information*, we use the value function v_H in our experiments, with plug-in estimators for total, epistemic, and/or aleatoric uncertainty.

6.1 Supervised Learning Examples

First, we perform a simple proof of concept experiment that illustrates the method's performance on image and text data.

Image Data. We examine binary classifiers on subsets of the MNIST dataset. Specifically, we train deep convolutional neural nets to distinguish 1 vs. 7, 3 vs. 8, and 4 vs. 9. These digit pairs tend to look similar in many people's handwriting and are often mistaken for one another. We therefore expect relatively high uncertainty in these examples, and use a variant of DeepSHAP to visualize the pixel-wise contributions to predictive entropy, as estimated via MC dropout. We compute attributions for epistemic and aleatoric uncertainty, visually confirming that the former identifies regions of the image that most increase or reduce uncertainty (see Fig. 1A).

Applying our method, we find that epistemic uncertainty is reduced by the upper loop of the 9, as well as by the downward hook on the 7. By contrast, uncertainty is increased by the odd angle of the 8 and its small bottom loop. Aleatoric uncertainty, by contrast, is more mixed across the pixels. We take total entropy to be the sum of these two terms.

[1] https://github.com/joshwa71/UncertaintyShap.

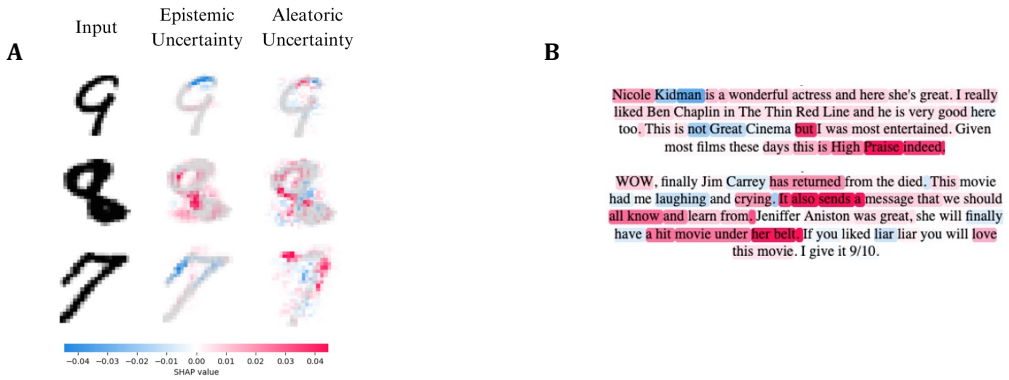

Figure 1: **A**. MNIST examples. We highlight pixels that increase (red) and decrease (blue) predictive uncertainty in digit classification tasks (1 vs. 7, 3 vs. 8, and 4 vs. 9). **B**. Example reviews from the IMDB dataset, with tokens colored by their relative contribution to the entropy of sentiment predictions.

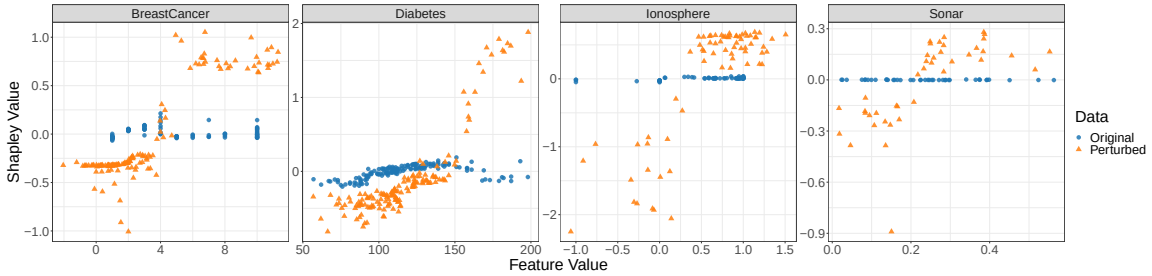

Figure 2: Information theoretic Shapley values explain the uncertainty of predictions on original and perturbed test sets. Our method correctly attributes the excess entropy to the perturbed features.

Text Data. We apply a transformer network to the IMDB dataset, which contains movie reviews for some 50,000 films. This is a sentiment analysis task, with the goal of identifying positive vs. negative reviews. We visualize the contribution of individual words to the uncertainty of particular predictions as calculated using the modified DeepSHAP pipeline, highlighting how some tokens tend to add or remove predictive information.

We report results for two high-entropy examples in Fig. 1B. In the first review, the model appears confused by the sentence "This is not Great Cinema but I was most entertained," which clearly conveys some ambiguity in the reviewer's sentiment. In the second example, the uncertainty comes from several sources including unexpected juxtapositions such as "laughing and crying", as well as "liar liar...you will love this movie."

6.2 Covariate Shift

To illustrate the utility of our method for explaining covariate shift, we consider several semi-synthetic experiments. We start with four binary classification datasets from the UCI machine learning repository [17]—`BreastCancer`, `Diabetes`, `Ionosphere`, and `Sonar`—and make a random 80/20 train/test split on each. We use an XGBoost model [12] with 50 trees to estimate conditional probabilities and the associated uncertainty. We then perturb a random feature from the test set, adding a small amount of Gaussian noise to alter its underlying distribution. Resulting predictions have a large degree of entropy, and would therefore be ranked highly by an AL acquisition function. We compute information theoretic Shapley values for original and perturbed test sets. Results are visualized in Fig. 2.

Our method clearly identifies the source of uncertainty in these datapoints, assigning large positive or negative attributions to perturbed features in the test environment. Note that the distribution shifts are fairly subtle in each case, rarely falling outside the support of training values for a given feature. Thus we find that information theoretic Shapley values can be used in conjunction with covariate shift detection algorithms to explain the source of the anomaly, or in conjunction with AL algorithms to explain the exploratory selection procedure.

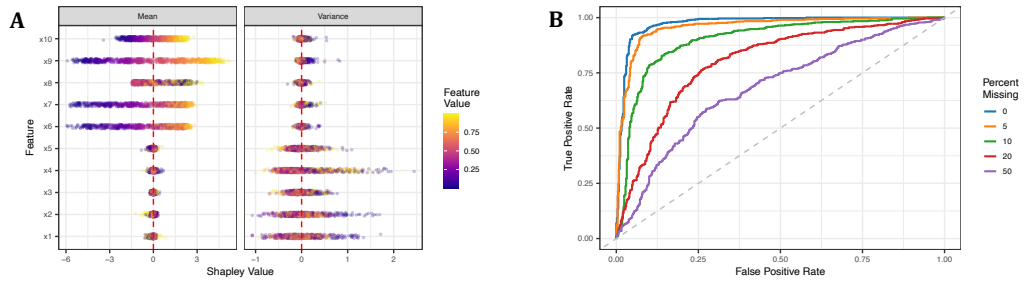

Figure 3: **A**. Results for the modified Friedman benchmark experiment. The conditional mean depends on $\{X_6, \ldots, X_{10}\}$, while the conditional variance relies on $\{X_1, \ldots, X_5\}$. **B**. ROC curves for a feature ranking task with variable levels of missingness. The proposed value function gives informative results for feature-value acquisition.

6.3 Feature Selection

Another application of the method is as a feature selection tool when heteroskedasticity is driven by some but not all variables. For this experiment, we modify the classic Friedman benchmark [19], which was originally proposed to test the performance of nonlinear regression methods under signal sparsity. Outcomes are generated according to:

$$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon_y,$$

with input features $\mathbf{X} \sim \mathcal{U}(0,1)^{10}$ and standard normal residuals $\epsilon_y \sim \mathcal{N}(0, 1^2)$. To adapt this DGP to our setting, we scale Y to the unit interval and define:

$$Z = 10 \sin(\pi X_6 X_7) + 20(X_8 - 0.5)^2 + 10X_9 + 5X_{10} + \epsilon_z,$$

with $\epsilon_z \sim \mathcal{N}(0, \tilde{Y}^2)$, where $\tilde{Y}$ denotes the rescaled version of Y. Note that Z's conditional variance depends exclusively on the first five features, while its conditional mean depends only on the second five. Thus with $f(\mathbf{x}) = \mathbb{E}[Z \mid \mathbf{x}]$ and $h(\mathbf{x}) = \mathbb{V}[Z \mid \mathbf{x}]$, we should expect Shapley values for f to concentrate around zero for $\{X_6, \ldots, X_{10}\}$, while Shapley values for h should do the same for $\{X_1, \ldots, X_5\}$.

We draw 2000 training samples and fit f using XGBoost with 100 trees. This provides estimates of both the conditional mean (via predictions) and the conditional variance (via observed residuals $\hat{\epsilon}_y$). We fit a second XGBoost model h with the same hyperparameters to predict $\hat{\epsilon}_y^2$. Results are reported on a test set of size 1000. We compute feature attributions using TreeSHAP [42] and visualize results in Fig. 3A. We clearly see that Shapley values are clustered around zero for unimportant features in each model, demonstrating the method's promise for discriminating between different modes of predictive information. In a supplemental experiment, we empirically evaluate our conformal coverage guarantee on this same task, achieving nominal coverage at $\alpha = 0.1$ for all features in the experiment (see Appx. C.3).

As an active feature-value acquisition example, we use the same modified Friedman benchmark, but this time increase the training sample size to 5000 and randomly delete some proportion of cells in the design matrix for $\mathbf{X}$. This simulates the effect of missing data, which may arise due to entry errors or high collection costs. XGBoost has native methods for handling missing data at training and test time, although resulting Shapley values are inevitably noisy. We refit the conditional variance estimator h and record feature rankings with variable missingness.

The goal in active feature-value acquisition is to prioritize the variables whose values will best inform future predictions subject to budgetary constraints. Fig. 3B shows receiver operating characteristic (ROC) curves for a feature importance ranking task as the frequency of missing data increases from zero to 50%. Importance is estimated via absolute Shapley values. Though performance degrades with increased missing data, as expected, we find that our method reliably ranks important features above unimportant ones in all trials. Even with fully half the data missing, we find an AUC of 0.682, substantially better than random.

7 Discussion

Critics have long complained that Shapley values (using the conventional payoff function v_0) are difficult to interpret in XAI. It is not always clear what it even means to delete features [40, 2], and large/small

attributions are neither necessary nor sufficient for important/unimportant predictors, respectively [5]. In an effort to ground these methods in classical statistical notions, several authors have analyzed Shapley values in the context of ANOVA decompositions [6] or conditional independence tests [71], with mixed results. Our information theoretic approach provides another window into this debate. With modified value functions, we show that marginal payoffs $\Delta_v(S, j, \mathbf{x})$ have an unambiguous interpretation as a local dependence measure. Still, Shapley values muddy the waters somewhat by averaging these payoffs over coalitions.

There has been a great deal of interest in recent years on *functional data analysis*, where the goal is to model not just the conditional mean of the response variable $\mathbb{E}[Y \mid \mathbf{x}]$, but rather the entire distribution $P(Y \mid \mathbf{x})$, including higher moments. Distributional regression techniques have been developed for additive models [60], gradient boosting machines [72], random forests [26], and neural density estimators [53]. Few if any XAI methods have been specifically designed to explain such models, perhaps because attributions would be heavily weighted toward features with a significant impact on the conditional expectation, thereby simply reducing to classic measures. Our method provides one possible way to disentangle those attributions and focus attention on higher moments. Future work will explore more explicit connections to the domain of functional data.

One advantage of our approach is its modularity. We consider a range of different information theoretic games, each characterized by a unique value function. We are agnostic about how to estimate the relevant uncertainty measures, fix reference distributions, or sample candidate coalitions. These are all active areas of research in their own right, and practitioners should choose whichever combination of tools works best for their purpose.

However, this flexibility does not come for free. Computing Shapley values can be #P-hard, depending on the function class [74]. Imputing values for out-of-coalition features is a statistical challenge that requires extensive marginalization. Some speedups can be achieved by making convenient assumptions, but these may incur substantial errors in practice. These are familiar problems in feature attribution tasks. Our method inherits the same benefits and drawbacks.

8 Conclusion

We introduced a range of methods to explain conditional entropy in ML models, bringing together existing work on uncertainty quantification and feature attributions. We studied the information theoretic properties of several games, and implemented our approach in model-specific and model-agnostic algorithms with numerous applications. Future work will continue to examine how XAI can go beyond its origins in prediction to inform decision making in areas requiring an exploration-exploitation trade-off, such as bandits and reinforcement learning.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.*, 298:103502, 2021.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[3] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*, 2021.

[4] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178, 2023.

[5] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *arXiv* preprint, 2212.11870, 2022.

[6] Sebastian Bordt and Ulrike von Luxburg. From Shapley values to generalized additive models and back. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 709–745, 2023.

[7] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in Bayesian networks. In *Proceedings of The 12th Conference on Uncertainty in Artificial Intelligence*, 1996.

[8] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv* preprint, 2006.16234, 2020.

[9] Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. Algorithms to estimate Shapley value feature attributions. *arXiv* preprint, 2207.07605, 2022.

[10] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892, 2018.

[11] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-Shapley and C-Shapley: Efficient model interpretation for structured data. In *International Conference on Learning Representations*, 2019.

[12] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016.

[13] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, Hoboken, NJ, second edition, 2006.

[14] Ian Covert and Su-In Lee. Improving kernelSHAP: Practical Shapley value estimation using linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 3457–3465, 2021.

[15] Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223, 2020.

[16] Ian Covert, Scott M. Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.*, 22(209):1–90, 2021.

[17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[18] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.

[19] Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1 – 67, 1991.

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, 2016.

[21] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald Rubin. *Bayesian data analysis*. Chapman & Hall, New York, 1995.

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[23] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

[24] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems*, 2020.

[25] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82, 2021.

[26] Torsten Hothorn and Achim Zeileis. Predictive distribution modeling using transformation forests. *J. Comput. Graph. Stat.*, 30(4):1181–1196, 2021.

[27] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv* preprint, 1112.5745, 2011.

[28] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110:457–506, 2021.

[29] Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. Feature relevance quantification in explainable AI: A causal problem. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 2907–2916, Online, 2020.

[30] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1459–1467, 2021.

[31] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 895–905, 2020.

[32] Robert Kaufman. *Heteroskedasticity in regression.* SAGE, London, 2013.

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The 3rd International Conference for Learning Representations*, 2015.

[34] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

[35] Sofia Ira Ktena, Alykhan Tejani, Lucas Theis, Pranay Kumar Myana, Deepak Dilipkumar, Ferenc Huszár, Steven Yoo, and Wenzhe Shi. Addressing delayed feedback for continuous training with neural networks in CTR prediction. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 187–195, 2019.

[36] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019.

[37] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, Cambridge, 2020.

[38] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, jul 2018.

[39] Dan Ley, Umang Bhatt, and Adrian Weller. Diverse, global and amortised counterfactual explanations for uncertainty estimates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7390–7398, 2022.

[40] Zachary Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, 2018.

[41] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774. 2017.

[42] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1):56–67, 2020.

[43] Christoph Luther, Gunnar König, and Moritz Grosse-Wentrup. Efficient SAGE estimation via causal structure learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[44] Christopher Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of The 11th Conference on Uncertainty in Artificial Intelligence*, 1995.

[45] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *CD-MAKE*, pages 17–38. Springer, 2020.

[46] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[47] Kevin Murphy. *Probabilistic Machine Learning: An Introduction.* The MIT Press, Cambridge, MA, 2022.

[48] Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763 – 765, 1973.

[49] Lars H. B. Olsen, Ingrid K. Glad, Martin Jullum, and Kjersti Aas. Using shapley values and variational autoencoders to explain predictive models with dependent mixed features. *Journal of Machine Learning Research*, 23(213), 2022.

[50] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29, 2016.

[51] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[52] Art B. Owen. Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1): 245–251, 2014.

[53] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Stiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019.

[55] John R. Quinlan. Induction of decision trees. *Mach. Learn*, 1(1):81–106, 1986.

[56] John R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[57] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, 2006.

[58] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. 32(1):1527–1535, 2018.

[60] Robert A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

[61] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison, 2009.

[62] Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In Michael R. Berthold, Ad Feelders, and Georg Krempl, editors, *Advances in Intelligent Data Analysis XVIII*, pages 444–456, Cham, 2020. Springer International Publishing.

[63] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Advances in Neural Information Processing Systems*, volume 34, pages 9391–9404, 2021.

[64] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, 2018.

[65] Kacper Sokol and Peter Flach. LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. *arXiv* preprint, 2005.01427, 2020.

[66] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[67] Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[68] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 3319–3328, 2017.

[69] Richard Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. The MIT Press, Cambridge, MA, 2nd edition, 2018.

[70] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[71] Jacopo Teneggi, Beepul Bharti, Yaniv Romano, and Jeremias Sulam. From Shapley back to Pearson: Hypothesis testing via the Shapley value. *arXiv* preprint, 2207.07038, 2022.

[72] Janek Thomas, Andreas Mayr, Bernd Bischl, Matthias Schmid, Adam Smith, and Benjamin Hofner. Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28(3):673–687, 2018.

[73] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436 – 463, 2013.

[74] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. On the tractability of SHAP explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021.

[75] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[76] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.*, 31(2):841–887, 2018.

[77] David S. Watson. Conceptual challenges for interpretable machine learning. *Synthese*, 200(2):65, 2022.

[78] David S. Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. *Minds Mach.*, 32(1):185–218, 2022.

[79] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019.

[80] Jiji Zhang and Peter L. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of The 19th Conference on Uncertainty in Artificial Intelligence*, 2003.

[81] Jiji Zhang and Peter L. Spirtes. The three faces of faithfulness. *Synthese*, 193(4):1011–1027, 2016.

[82] Indre Žliobaite. Change with delayed labeling: When is it detectable? In *2010 IEEE International Conference on Data Mining Workshops*, pages 843–850. IEEE, 2010.

A Proofs

Proof of Prop. 4.1. Substituting v_{KL} into the definition of $\Delta_v(S, j, \mathbf{x})$ gives:

$$\Delta_{KL}(S, j, \mathbf{x}) = -D_{KL}(p_{Y|\mathbf{x}} \parallel p_{Y|\mathbf{x}_S, x_j}) + D_{KL}(p_{Y|\mathbf{x}} \parallel p_{Y|\mathbf{x}_S}).$$

Rearranging and using the definition of KL-divergence, we have:

$$\Delta_{KL}(S, j, \mathbf{x}) = \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[\log p(y \mid \mathbf{x}) - \log p(y \mid \mathbf{x}_S)\big] - \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[\log p(y \mid \mathbf{x}) - \log p(y \mid \mathbf{x}_S, x_j)\big].$$

Cleaning up in steps:

$$\begin{aligned}
\Delta_{KL}(S, j, \mathbf{x}) &= \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[\log p(y \mid \mathbf{x}) - \log p(y \mid \mathbf{x}_S) - \log p(y \mid \mathbf{x}) + \log p(y \mid \mathbf{x}_S, x_j)\big] \\
&= \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[\log p(y \mid \mathbf{x}_S, x_j) - \log p(y \mid \mathbf{x}_S)\big] \\
&= \int_{\mathcal{Y}} p(y \mid \mathbf{x}) \ \log \frac{p(y \mid \mathbf{x}_S, x_j)}{p(y \mid \mathbf{x}_S)} \ dy.
\end{aligned}$$

Substituting v_{CE} into the definition of $\Delta_v(S, j, \mathbf{x})$ gives:

$$\Delta_{CE}(S, j, \mathbf{x}) = -H(p_{Y|\mathbf{x}}, p_{Y|\mathbf{x}_S, x_j}) + H(p_{Y|\mathbf{x}}, p_{Y|\mathbf{x}_S}).$$

Rearranging and using the definition of cross entropy, we have:

$$\begin{aligned}
\Delta_{CE}(S, j, \mathbf{x}) &= H(p_{Y|\mathbf{x}}, p_{Y|\mathbf{x}_S}) - H(p_{Y|\mathbf{x}}, p_{Y|\mathbf{x}_{S\cup\{j\}}}) \\
&= \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[-\log p(y \mid \mathbf{x}_S)\big] - \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[-\log p(y \mid \mathbf{x}_S, x_j)\big] \\
&= \mathop{\mathbb{E}}_{Y|\mathbf{x}}\big[\log p(y \mid \mathbf{x}_S, x_j) - \log p(y \mid \mathbf{x}_S)\big] \\
&= \int_{\mathcal{Y}} p(y \mid \mathbf{x}) \ \log \frac{p(y \mid \mathbf{x}_S, x_j)}{p(y \mid \mathbf{x}_S)} \ dy.
\end{aligned}$$

Proof of Prop. 4.2. Since the Shapley value $\phi_v(j, \mathbf{x})$ is just the expectation of $\Delta_v(S, j, \mathbf{x})$ under a certain distribution on coalitions $S \subseteq [d]\backslash\{j\}$ (see Eq. 1), it follows from Prop. 4.1 that feature attributions will be identical under v_{KL} and v_{CE}. To show that resulting Shapley values sum to the KL-divergence between $p(Y \mid \mathbf{x})$ and $p(Y)$, we exploit the efficiency property:

$$\begin{aligned}
\sum_{j=1}^{d} \phi_{KL}(j, \mathbf{x}) &= v_{KL}([d], \mathbf{x}) - v_{KL}(\emptyset, \mathbf{x}) \\
&= -D_{KL}(p_{Y|\mathbf{x}} \parallel p_{Y|\mathbf{x}}) + D_{KL}(p_{Y|\mathbf{x}} \parallel p_Y) \\
&= D_{KL}(p_{Y|\mathbf{x}} \parallel p_Y).
\end{aligned}$$

The last step exploits Gibbs's inequality, according to which $D_{KL}(p \parallel q) \geq 0$, with $D_{KL}(p \parallel q) = 0$ iff $p = q$.

Proof of Prop. 4.3. Substituting v_{IG} into the definition of $\Delta_v(S, j, \mathbf{x})$ gives:

$$\begin{aligned}
\Delta_{IG}(S, j, \mathbf{x}) &= -H(Y \mid \mathbf{x}_S, x_j) + H(Y \mid \mathbf{x}_S) \\
&= H(Y \mid \mathbf{x}_S) - H(Y \mid \mathbf{x}_S, x_j) \\
&= I(Y; x_j \mid \mathbf{x}_S) \\
&= \int_{\mathcal{Y}} p(y, x_j \mid \mathbf{x}_S) \log \frac{p(y, x_j \mid \mathbf{x}_S)}{p(y \mid \mathbf{x}_S) \ p(x_j \mid \mathbf{x}_S)} dy.
\end{aligned}$$

In the penultimate line, we exploit the equality $I(Y; X) = H(Y) - H(Y \mid X)$, by which we define mutual information (see Appx. B.1).

Proof of Prop. 4.4. We once again rely on efficiency and the definition of mutual information in terms of marginal and conditional entropy:

$$\begin{aligned}\sum_{j=1}^{d} \phi_{IG}(j, \mathbf{x}) &= v_{IG}([d], \mathbf{x}) - v_{IG}(\emptyset, \mathbf{x}) \\ &= -H(Y \mid \mathbf{x}) + H(Y) \\ &= H(Y) - H(Y \mid \mathbf{x}) \\ &= I(Y; \mathbf{x}).\end{aligned}$$

Proof of Thm. 4.5. Begin with item (a). Note that the conditional independence statement $Y \perp\!\!\!\perp X_j \mid \mathbf{X}_S$ holds iff, for all points $(\mathbf{x}, y) \sim \mathcal{D}$, we have:

$$p(y \mid \mathbf{x}_S, x_j) = p(y \mid \mathbf{x}_S) \quad \text{and} \quad p(y, x_j \mid \mathbf{x}_S) = p(y \mid \mathbf{x}_S)\ p(x_j \mid \mathbf{x}_S).$$

The former guarantees that marginal payouts evaluate to zero for $v \in \{v_{KL}, v_{CE}\}$; the latter does the same for $v \in \{v_{IG}, v_H\}$. This follows because the log ratio in each formula evaluates to zero when numerator and denominator are equal.

Of course, conditional independence is also sufficient for zero marginal payout with more familiar value functions such as v_0. But item (a) makes an additional claim—that the *converse* holds as well, i.e. that conditional independence is *necessary* for zero marginal payout across all $\mathbf{x}$. This follows from the definitions of the value functions themselves. Observe:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_X}\left[\Delta_{KL}(S, j, \mathbf{x})\right] &= \mathop{\mathbb{E}}_{(\mathbf{x}, y) \sim \mathcal{D}}\left[\log \frac{p(y \mid \mathbf{x}_s, x_j)}{p(y \mid \mathbf{x}_S)}\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}_X}\left[\mathop{\mathbb{E}}_{Y \mid \mathbf{x}_s, x_j}\left[\log \frac{p(y \mid \mathbf{x}_s, x_j)}{p(y \mid \mathbf{x}_S)}\right]\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}_X}\left[D_{KL}(p_{Y \mid \mathbf{x}_S, x_j} \parallel p_{Y \mid \mathbf{x}_S})\right]\end{aligned}$$

By Gibbs's inequality, the KL-divergence between two distributions is zero iff they are equal, so setting this value to zero for all $\mathbf{x}$ satisfies the first definition of conditional independence above. For the latter, we simply point out that:

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_X}\left[\Delta_{IG}(S, j, \mathbf{x})\right] = I(Y; X_j \mid \mathbf{X}_S).$$

Since conditional mutual information equals zero iff the relevant variables are conditionally independent, this satisfies the second definition above.

Item (b) states that CSI, which is strictly weaker than standard conditional independence, is also sufficient for zero marginal payout at a given point $\mathbf{x}$. This follows directly from the sufficiency argument above.

The converse relationship is more complex, however. Call a distribution *conspiratorial* if there exists some $S, j, \mathbf{x}$ such that $\Delta_v(S, j, \mathbf{x}) = 0 \wedge Y \not\perp\!\!\!\perp x_j \mid \mathbf{x}_S$ for some $v \in \{v_{KL}, v_{CE}, v_{IG}, v_H\}$. Such distributions are so named because the relevant probabilities must coordinate in a very specific way to guarantee summation to zero as we marginalize over $\mathcal{Y}$. As a concrete example, consider the following data generating process:

$$X \sim \text{Bern}(0.5), \quad Z \sim \text{Bern}(0.5), \quad Y \sim \text{Bern}(0.3 + 0.4X - 0.2Z).$$

What is the contribution of X to coalition $S = \emptyset$ when $X = 1$ and $Z = 1$? In this case, we have neither global nor context-specific independence, i.e. $Y \not\perp\!\!\!\perp x$. Yet, evaluating the payoffs in a KL-divergence game, we have:

$$\begin{aligned}\Delta_{KL}(S, j, \mathbf{x}) &= \sum_y P(y \mid X = 1, Z = 1) \log \frac{P(y \mid X = 1)}{P(y)} \\ &= 0.5 \log \frac{0.4}{0.6} + 0.5 \log \frac{0.6}{0.4} \\ &= 0.\end{aligned}$$

In this case, we find that negative and positive values of the log ratio cancel out exactly as we marginalize over $\mathcal{Y}$. (Similar examples can be constructed for v_{IG} and v_H.) This shows that CSI is sufficient but not necessary for $\Delta_v(S, j, \mathbf{x}) = 0$.

However, just because conspiratorial distributions are possible does not mean that they are common. Item (c) states that the set of all such distributions has Lebesgue measure zero. Our proof strategy here follows that of Meek [44], who demonstrates a similar result in the case of *unfaithful* distributions, i.e. those whose (conditional) independencies are not entailed by the data's underlying graphical structure. This is an important topic in the causal discovery literature (see, e.g., [80, 81]).

For simplicity, assume a discrete state space $\mathcal{X} \times \mathcal{Y}$. Fix some S, j such that $Y \not\perp\!\!\!\perp x_j \mid \mathbf{x_S}$. Let C be the number of possible outcomes, $\mathcal{Y} = \{y_1, \dots, y_C\}$. Define vectors $\mathbf{p}, \mathbf{r}$ of length C such that, for each $c \in [C]$:

$$p_c = p(y_c \mid \mathbf{x}), \quad r_c = \log \frac{p(y_c \mid \mathbf{x}_S, x_j)}{p(y_c \mid \mathbf{x}_S)}.$$

(Technically, we only require $C - 1$ entries to fully describe these conditional distributions, but there is no penalty for overparametrization here.) By the assumption of local conditional dependence, we know that $\|\mathbf{r}\|_0 > 0$. Yet for our conspiracy to obtain, the inner product of these vectors must satisfy $\mathbf{p} \cdot \mathbf{r} = 0$. A well-known algebraic lemma of Okamoto [48] states that if a polynomial constraint is non-trivial (i.e., if there exists some $\mathbf{p}, \mathbf{r}$ for which it does not hold), then the subset of parameters for which it does hold has Lebesgue measure zero. Since the conspiracy requires nontrivial constraints that are linear in the parameters $\mathbf{p}, \mathbf{r}$, we conclude that the set of conspiratorial distributions has Lebesgue measure zero.

Proof of Thm. 5.1. Our proof is an application of the split conformal method (see [38, Thm. 2.2]). Whereas that method was designed to bound the distance between predicted and observed outcomes for a regression task, we effectively treat the mean Shapley value as a constant outcome to measure the concentration of feature attributions. To achieve this, we replace out-of-sample absolute residuals with out-of-sample Shapley values and labels with the mean Shapley value. With these substitutions in place, the result follows immediately from the symmetry of $\phi(j, \mathbf{x}^{(i+1)})$ and $\phi(j, \mathbf{x}^{(i)}), i \in \mathcal{I}_2$, which is itself a direct implication of the i.i.d. assumption.[2] Since the margin is calculated so as to cover $(1 - \alpha) \times 100\%$ of the distribution, it is unlikely that new samples will fall outside this region. Specifically, such exceptions occur with probability at most α. This amounts to a sort of PAC guarantee, i.e. that Shapley values will be within radius τ_j of their mean μ_j with probability at least $1 - \alpha$.

B Addenda

This section includes extra background material on information theory and Shapley values.

B.1 Information Theory

Let p, q be two probability distributions over the same σ-algebra of events. (In the continuous case, we additionally require that p, q be absolutely continuous with respect to Lebesgue measure.) The *entropy* of p is defined as $H(p) := \mathbb{E}_p[-\log p]$, i.e. the expected number of bits required to encode the distribution.[3] The *cross entropy* of p and q is defined as $H(p, q) := \mathbb{E}_p[-\log q]$, i.e. the expected number of bits required to encode samples from p using code optimized for q. The *KL-divergence* between p and q is defined as $D_{KL}(p \parallel q) := \mathbb{E}_p[\log p/q]$, i.e. the cost in bits of modeling p with q. These three quantities are related by the formula $D_{KL}(p \parallel q) = H(p, q) - H(p)$. The reduction in Y's uncertainty attributable to X is also called the *mutual information*, $I(Y; X) := H(Y) - H(Y \mid X)$. This quantity is nonnegative, with $I(Y; X) = 0$ if and only if the variables are independent.

However, conditioning on a specific value of X may increase uncertainty in Y, in which case the local conditional entropy exceeds the marginal. Thus it is possible that $H(Y \mid x) > H(Y)$ for some $x \in \mathcal{X}$. For example, consider the following data generating process:

$$X \sim \text{Bern}(0.8), \quad Y \sim \text{Bern}(0.5 + 0.25X).$$

[2] Note that conformal inference relies on the weaker assumption of exchangeability. However, since we operate in the standard i.i.d. setting of statistical learning theory (see Sect. 3), exchangeability naturally follows.

[3] Though the term "bit" is technically reserved for units of information measured with logarithmic base 2, we use the word somewhat more loosely to refer to any unit of information.

In this case, we have $P(Y=1)=0.7, P(Y=1 \mid X=0)=0.5$, and $P(Y=1 \mid X=1)=0.75$. It is easy to see that even though the marginal entropy $H(Y)$ exceeds the global conditional entropy $H(Y \mid X)$, the local entropy at $X=0$ is larger than either quantity, $H(Y \mid X=0) > H(Y) > H(Y \mid X)$. In other words, conditioning on the event $X=0$ increases our uncertainty about Y.

Similarly, there may be cases in which $I(Y;X \mid Z) > 0$, but $I(Y;X \mid z) = 0$. This is what Boutilier et al. [7] call *context-specific independence* (CSI). For instance, if $X, Z \in \{0,1\}^2$ and $Y := X \vee Z$, then we have $Y \not\perp\!\!\!\perp X \mid Z$, but $Y \perp\!\!\!\perp X \mid (Z=1)$ since Y's value is determined as soon as we know that either parent is 1.

B.2 The Shapley Axioms

For completeness, we here list the Shapley axioms.

Efficiency. Shapley values sum to the difference in payoff between complete and null coalitions:

$$\sum_{j=1}^{d} \phi(j, \mathbf{x}) = v([d], \mathbf{x}) - v(\emptyset, \mathbf{x}).$$

Symmetry. If two players make identical contributions to all coalitions, then their Shapley values are equal:

$$\forall S \subseteq [d] \backslash \{i,j\} : v(S \cup \{i\}, \mathbf{x}) = v(S \cup \{j\}, \mathbf{x}) \Rightarrow \phi(i, \mathbf{x}) = \phi(j, \mathbf{x}).$$

Sensitivity. If a player makes zero contribution to all coalitions, then its Shapley value is zero:

$$\forall S \subseteq [d] \backslash \{j\} : v(S \cup \{j\}, \mathbf{x}) = v(S, \mathbf{x}) \Rightarrow \phi(j, \mathbf{x}) = 0.$$

Linearity. The Shapley value for a convex combination of games can be decomposed into a convex combination of Shapley values. For any $a, b \in \mathbb{R}$ and value functions v_1, v_2, we have:

$$\phi_{a \cdot v_1 + b \cdot v_2}(j, \mathbf{x}) = a\phi_{v_1}(j, \mathbf{x}) + b\phi_{v_2}(j, \mathbf{x}).$$

C Experiments

C.1 Datasets.

The MNIST dataset is available online.[4] The IMDB dataset is available on Kaggle.[5] The `BreastCancer`, `Diabetes`, `Ionosphere`, and `Sonar` datasets are all distributed in the `mlbench` package, which is available on `CRAN`.[6]

C.2 Models.

All neural network training was conducted in PyTorch [54]. We use a standard convolutional neural network for the MNIST experiment, including convolutions, max pooling, and batch norm. We use ReLU activations, cross entropy loss, and optimize with Adam [33]. For the IMDB experiment, we use a pre-trained BERT model from the Hugging Face transformers library.[7] All hyperparameters are set to their default values. All XGBoost models are trained with the default hyperparameters, with the number of training rounds cited in the text.

[4] http://yann.lecun.com/exdb/mnist/.
[5] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.
[6] https://cran.r-project.org/web/packages/mlbench/index.html.
[7] https://huggingface.co/docs/transformers/model_doc/bert.

C.3 Coverage

To empirically test our conformal coverage guarantee, we compute means and margins on out-of-sample Shapley values for the modified Friedman benchmark. Results for conditional expectation and conditional variance are reported in Table 1, with target level $\alpha = 0.1$. Note that what constitutes a "small" or "large" margin is context dependent. The conditional variance model is fit to ϵ_y^2, which has a tighter range than Z, leading to smaller Shapley values on average. However, nominal coverage is very close to the target 90% throughout, illustrating how the conformal method can be used for feature selection and outlier detection.

Table 1: Means, margins, and nominal coverage at $\alpha = 0.1$ for Shapley values from the conditional mean and conditional variance models. Results are averaged over 50 replicates.

	Mean			Variance		
Feature	μ	τ	Coverage	μ	r	Coverage
X_1	-0.002	0.066	0.899	-0.009	0.505	0.898
X_2	0.008	0.141	0.898	-0.001	0.435	0.900
X_3	0.002	0.084	0.899	0.001	0.278	0.898
X_4	-0.004	0.098	0.901	-0.006	0.727	0.900
X_5	-0.004	0.092	0.905	0.020	0.333	0.902
X_6	-0.162	3.637	0.903	-0.001	0.060	0.900
X_7	-0.032	3.555	0.901	0.003	0.049	0.899
X_8	-0.027	1.981	0.898	0.001	0.055	0.900
X_9	0.190	4.114	0.898	-0.002	0.053	0.899
X_{10}	-0.044	1.952	0.903	-0.001	0.053	0.900