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, $\mathcal{Y} = \{y_1, \ldots, y_C\}$. 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.[1] 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.[2] 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:

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

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:

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

[1] 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.

[2] 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.

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

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.[6] 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.

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	τ	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

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