Unfooling SHAP and SAGE: Knockoff Imputation for Shapley Values

Kristin Blesch[1,2](✉), Marvin N. Wright[1,2,3], and David Watson[4]

[1] Leibniz Institute for Prevention Research and Epidemiology – BIPS,
Bremen, Germany
blesch@leibniz-bips.de

[2] Faculty of Mathematics and Computer Science, University of Bremen,
Bremen, Germany

[3] Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[4] Department of Informatics, King's College London, London, UK

Abstract. Shapley values have achieved great popularity in explainable artificial intelligence. However, with standard sampling methods, resulting feature attributions are susceptible to adversarial attacks. This originates from target function evaluations at extrapolated data points, which are easily detectable and hence, enable models to behave accordingly. In this paper, we introduce a novel strategy for increased robustness against adversarial attacks of both local and global explanations: Knockoff imputed Shapley values. Our approach builds on the model-X knockoff methodology, which generates synthetic data that preserves statistical properties of the original samples. This enables researchers to flexibly choose an appropriate model to generate on-manifold data for the calculation of Shapley values upfront, instead of having to estimate a large number of conditional densities or make strong parametric assumptions. Through real and simulated data experiments, we demonstrate the effectiveness of knockoff imputation against adversarial attacks.

Keywords: XAI · Shapley Values · Adversarial Attacks · Knockoffs

1 Introduction

Explainable artificial intelligence (XAI) oftentimes strives to deliver insights on the underlying mechanisms of black-box machine learning models in order to generate trust in these algorithms. To do so, XAI methods themselves must be trustworthy.

Several popular XAI tools, such as SHAP [17] and LIME [19], are vulnerable to adversarial attacks [23]. The issue stems from how these approaches generate new data during the explanation process – typically by independently permuting feature values. Permute-and-predict methods force models to extrapolate beyond their training data, yielding off-manifold samples. This results in potentially misleading assessments [13] and enables adversaries to pass fairness audits

L. Longo (Ed.): xAI 2023, CCIS 1901, pp. 131–146, 2023.
https://doi.org/10.1007/978-3-031-44064-9_8

even with discriminatory models. For example, an algorithm could fool the XAI method by using a fair model for queries on synthetic, extrapolated data during XAI evaluation in order to suggest the model would be fair even though it may produce discriminatory outcomes for non-synthetic, i.e. real data [23].

Robustness against such adversarial attacks can be achieved by reducing extrapolation during data generation. Ideally, conditional sampling procedures should be used, which ensures that the generated data is indistinguishable from the original data. Figure 1 visualizes data points generated through marginal in contrast to a conditional sampling method.

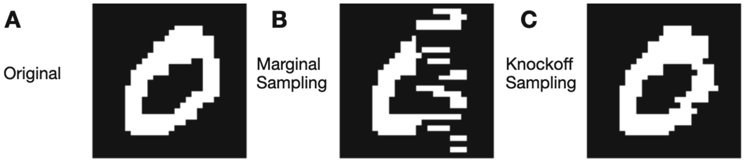

Fig. 1. Sampling of out-of-coalition features for a digit from $\{28 \times 28\}$ `mnist` data. The first 14 columns from the left are in-coalition, whereas the remaining 14 columns are sampled either from marginals (as in Kernel SHAP [17]) or deep knockoffs [21].

For Shapley values [22] – one of the most prominent XAI methods – conditional variants and their properties have been widely discussed in the literature [6,8,10,25,29]. Conditional Shapley values sample out-of-coalition features from a distribution conditioned on the in-coalition features. However, this requires knowledge about conditional distributions for all possible feature coalitions and, since estimating conditional distributions is generally challenging, there remains considerable room for improvement. However, to prevent adversarial attacks, calculating conditional Shapley values may be unnecessarily challenging. It suffices to minimize extrapolation, which is a strictly simpler task.

In that vein, we propose the model-X knockoff framework [5] in its full generality to sample out-of-coalition features for protection against adversarial attacks on Shapley value explanations. Knockoffs are characterized by two properties, formally defined below: (1) pairwise exchangeability with the original features; and (2) conditional independence of the response, given the true data. We argue that this makes them well-suited to serve as reference data in Shapley value pipelines. For example, property (1) allows us to estimate knockoffs upfront and use them to impute out-of-coalition features, which effectively avoids extrapolation and does not require the separate estimation of conditional distributions for any feature coalition. Knockoff imputed Shapley values balance on-manifold data sampling with maintaining utmost generality and flexibility in application.

The paper is structured as follows. First, we present the relevant background on Shapley values and model-X knockoffs in Sect. 2. We combine these approaches and study the theoretical properties of the resulting algorithm in Sect. 3. In Sect. 4, we present a series of experiments to demonstrate the effectiveness of our approach against adversarial attacks. We present a comprehensive discussion and directions for future research in Sects. 5 and 6, respectively.

2 Background and Related Work

2.1 Shapley Values

Originating from cooperative game theory, Shapley values [22] aim to attribute payouts fairly amongst a game's players. The basic idea is to evaluate the average change in output when a player is added to a coalition.

In XAI, we can think of the features $\mathbf{X} = \{X_1, \ldots, X_d\}$, where each X_j denotes a random variable, as a set of players $\mathcal{D} = \{1, \ldots, d\}$ who may or may not participate in a coalition of players $\mathcal{S} \subseteq \mathcal{D}$, i.e. $\mathcal{S}$ is a subset of $\mathcal{D}$. The value function v assigns a real-valued payout to each possible coalition, i.e. to every element of the power set of $\mathcal{D}$, which consists of $2^{|\mathcal{D}|} = 2^d$ elements, to a real value. This may be the expected output of a machine learning model f [17], or other quantities related to the model's prediction, such as the expected loss [8]. To compute the Shapley value ϕ_j for player j, we take a weighted average of j's marginal contributions to all subsets that exclude it:

$$\phi_j = \sum_{\mathcal{S} \subseteq \mathcal{D} \setminus \{j\}} \frac{|\mathcal{S}|!(|\mathcal{D}| - |\mathcal{S}| - 1)!}{|\mathcal{D}|!} \Big(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S}) \Big). \tag{1}$$

It is not immediately obvious how to evaluate v on strict subsets of $\mathcal{D}$, since f requires d-dimensional input. One common solution is to use an expectation with respect to some reference distribution $\mathcal{R}$:

$$v(\mathcal{S}) = \mathbb{E}_{\mathcal{R}} \big[f(\mathbf{x}_{\mathcal{S}}, \mathbf{X}_{\bar{\mathcal{S}}}) \big]. \tag{2}$$

In other words, for the random variables $\mathbf{X}_{\mathcal{S}}$, which are the in-coalition features, we take the realized values $\mathbf{x}_{\mathcal{S}}$ as fixed and sample values for out-of-coalition features $\mathbf{X}_{\bar{\mathcal{S}}}$ according to $\mathcal{R}$. Common choices for $\mathcal{R}$ include the marginal distribution $P(\mathbf{X}_{\bar{\mathcal{S}}})$ and the conditional distribution $P(\mathbf{X}_{\bar{\mathcal{S}}} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$.

Adversarial Attack Vulnerability. Taking the marginal distribution $\mathcal{R} = P(\mathbf{X}_{\bar{\mathcal{S}}})$ typically serves as an approximation to the conditional distribution $P(\mathbf{X}_{\bar{\mathcal{S}}} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$ in order to facilitate computation, e.g. as in KernelSHAP [17]. However, marginal and conditional distributions only coincide when features are jointly independent, which is scarcely ever the case in empirical applications. A consequence from a violation of feature independence is that sampling a set of $\mathbf{x}'_{\bar{\mathcal{S}}}$ from marginals instead of conditional distributions will lead to generated instances $\mathbf{x}' = (\mathbf{x}_{\mathcal{S}}, \mathbf{x}'_{\bar{\mathcal{S}}})$ that are off the data manifold of original, i.e. real data observations $\mathbf{x} = (\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}})$. In such cases, it is possible to train a prediction model ω that successfully distinguishes real from generated data. In adversarial explanations, e.g. the strategy described by [23], such an out-of-distribution (OOD) detector ω that exposes synthetic data is the primary workhorse. If the data is synthetic, the adversary deploys a different model as with real data, which effectively fools the explanation.

We want to highlight that even though this fooling strategy was introduced and is typically discussed for local Shapley values [23,28], it can also be applied to global aggregates such as Shapley additive global importance (SAGE) [8].

Achieving Adversarial Attack Robustness. Avoiding the generation of extrapolated data protects against adversarial attacks by preventing ω from distinguishing real from generated data during Shapley value calculation.

Some approaches naturally circumvent the task of generating synthetic data altogether, for example by using surrogate models [10], retraining the model such that it adapts to missing features [8] or fitting a separate model for each coalition [25,29]. However, these approaches come at a high computational costs, since repeated model refitting is required.

Another approach is to calculate conditional Shapley values, for which we will give a brief overview of methods in the following paragraph. Working with conditional Shapley values, i.e. using $\mathcal{R} = P(\mathbf{X}_{\bar{\mathcal{S}}} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$, is clearly the most rigorous way of enforcing on-manifold sampling of synthetic data, even though prior literature merely acknowledges the potential for preventing adversarial attacks. Several conditional Shapley value estimation procedures have been proposed, yet conditional feature sampling remains a challenging task and improvements are highly desirable.

A straightforward, empirical approach is to simply use the observed data that naturally satisfies the conditioning on the selected in-coalition features by using data points in close proximity to the instance to be explained [1,11]. For example, in Fig. 1, one could also sample the out-of-coalition features from other observations of digit zero in the data set. This approach, however, has the downside that the number of observations fulfilling the conditioning event might be very small, leading to only very few or even no appropriate observations available. Another approach to calculating conditional Shapley values is assuming a specific data distribution, e.g. a Gaussian distribution [1,7], for which conditional distributions are easy to derive, but this approach has the drawback of strong assumptions on the data generating process. Further, conditional generative models might be used [10,20], however, these models might be challenging to train and it is unclear whether they truly approximate the data well. In sum, conditional Shapley values are challenging to access and hence have limited applicability.

For the goal of preventing adversarial attacks, conditional Shapley values are sufficient but not necessary, since any method that avoids extrapolation will prevent the attack and hence related, but less strict frameworks provide another suite of promising methods. Such an idea is pursued by [28], where generative models use 'focused sampling' of new instances that are close to the original observations. However, this approach lacks theoretical guarantees and may fail depending on the fit of the generative models. We acknowledge that [28] investigate Gaussian knockoffs in conjunction with the so-called Interactions-based Method for Explanation (IME, [24]). However, the authors do not use model-X knockoffs for imputation in full generality, nor do they apply the strategy to SHAP or SAGE values directly. The authors even mention that the knockoff imputation idea merits further investigation as an approach, which is what the present paper contributes to.

2.2 Model-X Knockoffs

The model-X knockoff framework [5] is a theoretically sound concept to characterize synthetic variables with specific statistical properties. Intuitively speaking, knockoffs are synthetic variables that aim to copy the statistical properties of a given set of original variables, e.g. the covariance structure, such that they are indistinguishable from the original variables when the target variable Y is not looked at. Crucially, valid knockoffs ensure that original variables can be swapped with their knockoff counterparts without affecting the joint distribution.

Formally, in order for $\tilde{\mathbf{X}}$ to be a valid knockoff matrix for $\mathbf{X}$, two conditions have to be met:

1. Pairwise exchangeablility: For any proper subset $\mathcal{S} \subset \{1, \ldots, d\}$:

$$(\mathbf{X}, \tilde{\mathbf{X}})_{swap(\mathcal{S})} \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}}), \tag{3}$$

   where $\overset{d}{=}$ represents equality in distribution and *swap($\mathcal{S}$)* indicates swapping the variables in $\mathcal{S}$ with their knockoff counterparts.
2. Conditional independence:

$$\tilde{\mathbf{X}} \perp\!\!\!\perp Y \mid \mathbf{X}. \tag{4}$$

Generating valid knockoffs is an active field of research and various sampling algorithms have been proposed, which ensures that practitioners can flexibly choose appropriate algorithms. For example, there are algorithms based on distributional assumptions [3,5,21], Bayesian statistics [12] or deep learning [14,16,18,26].

3 Combining Model-X Knockoffs with Shapley Values

This paper proposes to impute out-of-coalition features with model-X knockoffs for the calculation of Shapley value based quantities. Knockoffs come with strong theoretical guarantees that ensure avoiding extrapolation. Moreover, they provide a major computational boost, since knockoffs can be sampled upfront for the full data matrix instead of requiring separate models for each possible coalition. Since many methods are available for knockoff generation—including some that are essentially tuning-free—practitioners have a large collection of tools available for valid, flexible and convenient sampling of the out-of-coalition space that ensures robustness against adversarial attacks.

In detail, we propose Algorithm 1 to impute out-of-coalition features with knockoffs for Shapley values and Algorithm 2 (see Appendix A) for knockoff imputation with SAGE [8] values. In brief, the algorithms use N_{ko} knockoffs as the background distribution in the calculation of Shapley values. Note that for $N_{ko} = 1$, the Shapley values are with respect to a single knockoff baseline value, while for larger values of N_{ko}, Shapley values explain the difference between the selected instance and the expected value of the knockoff distribution.

Algorithm 1. Knockoff Imputed Shapley Values

```
Input: data matrix $(\mathbf{X}, Y)$, knockoff sampler $ko()$, model $f$, explanation instance
    $\mathbf{x}^0 = \{x^0_1, \dots, x^0_d\}$, number of knockoffs $N_{ko}$, power set $\pi$ of $\mathcal{D} \setminus \{j\}$
train knockoff sampler $ko(\mathbf{X})$
for feature $j$ in $\mathcal{D}$ do
    initialize $\phi_j = 0$
    for $i$ in $N_{ko}$ do
        draw $\tilde{\mathbf{x}}^i = \{\tilde{x}^i_1, \dots, \tilde{x}^i_d\}$ from $ko(\mathbf{X})$
        initialize $\Delta^i_j = 0$
        for $\mathcal{S}$ in $\pi$ do
            out-of-coalition set $\bar{\mathcal{S}} = \mathcal{D} \setminus \mathcal{S}$
            $v(\mathcal{S}) = f(\mathbf{x}^0_{\mathcal{S}}, \tilde{\mathbf{x}}^i_{\bar{\mathcal{S}}})$
            $\mathcal{S}' = \mathcal{S} \cup \{j\}$
            $\bar{\mathcal{S}}' = \bar{\mathcal{S}} \setminus \{j\}$
            $v(\mathcal{S}') = f(\mathbf{x}^0_{\mathcal{S}'}, \tilde{\mathbf{x}}^i_{\bar{\mathcal{S}}'})$
            $\Delta^i_j = \Delta^i_j + \frac{|\mathcal{S}|!(|\mathcal{D}|-|\mathcal{S}|-1)!}{|\mathcal{D}|!} \cdot \big(v(\mathcal{S}') - v(\mathcal{S})\big)$
        end for
    end for
    $\phi_j = \frac{1}{N_{ko}} \sum_{i=1}^{N_{ko}} \Delta^i_j$
end for
return Shapley values $\boldsymbol{\phi} = \{\phi_1, \dots, \phi_d\}$
```

To understand the advantages of knockoff imputed Shapley values on a theoretical level, let us investigate the implications of the exchangeability property (Eq. 3) in more depth. This property ensures that we can swap *any* set $\mathcal{S} \subseteq \mathcal{D}$ of original variables $\mathbf{X}$ with knockoffs $\tilde{\mathbf{X}}$, while maintaining the same joint distribution. The same joint distribution guarantees that any generated data is indeed on the same data manifold, so for the prevention of adversarial attacks, it is both necessary and sufficient that $\mathbf{x'}_{\bar{\mathcal{S}}}$ is generated such that $P(\mathbf{X}_{\mathcal{S}}, \mathbf{X}_{\bar{\mathcal{S}}}) = P(\mathbf{X}_{\mathcal{S}}, \mathbf{X'}_{\bar{\mathcal{S}}})$. Conditional Shapley values ensure this by sampling $\mathbf{x'}_{\bar{\mathcal{S}}}$ from $P(\mathbf{X}_{\bar{\mathcal{S}}}|\mathbf{X}_{\mathcal{S}})$. Doing so, the original joint distribution is maintained by factorizing through $P(\mathbf{X}_{\mathcal{S}}) \cdot P(\mathbf{X}_{\bar{\mathcal{S}}}|\mathbf{X}_{\mathcal{S}}) = P(\mathbf{X}_{\mathcal{S}}, \mathbf{X}_{\bar{\mathcal{S}}})$, whereas knockoffs directly guarantee $P(\mathbf{X}_{\mathcal{S}}, \mathbf{X}_{\bar{\mathcal{S}}}) = P(\mathbf{X}_{\mathcal{S}}, \tilde{\mathbf{X}}_{\bar{\mathcal{S}}})$ by exchangebility.

That said, it becomes obvious that we can generate knockoff copies for $\mathbf{X}$ upfront and then swap in knockoffs for the out-of-coalition features $\mathbf{X}_{\bar{\mathcal{S}}}$ where needed. This is a clear advantage in contrast to conditional Shapley value methods that need access to $P(\mathbf{X}_{\bar{\mathcal{S}}}|\mathbf{X}_s)$ for all possible coalitions $2^{|\mathcal{D}|}$. Note that the pairwise exchangeability fulfilled by knockoffs is needed to guarantee on-manifold data in the imputation step, which is why other conditional sampling methods cannot be calculated upfront. This suggests a lower computational complexity for the knockoff imputed Shapley values in comparison to conditional Shapley values, however, the exact complexity will depend on the knockoff generating algorithm used. Further, even though we may want to sample N_{ko} knockoffs in advance to reduce bias, a reasonable number for N_{ko} is typically $N_{ko} \ll 2^{|D|}$.

However, the benefit of being able to sample knockoffs upfront comes at the cost of enforcing a restrictive set of conditioning events. At a first glance, knockoff imputation and calculating conditional Shapley values, i.e. using $\mathcal{R} = P(\mathbf{X}_{\bar{\mathcal{S}}} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$, may appear interchangeable. However, knockoffs implicitly condition on all the feature values of the observation, which is inevitable since the exchangeability property must hold for *any* set of variables. This subtle difference yields the following expression for the game that uses knockoffs $\tilde{\mathbf{X}}_{\bar{\mathcal{S}}}$ as imputation for the out-of-coalition features in set $\bar{\mathcal{S}}$:

$$v_{\text{ko}}(\mathcal{S}) = \mathbb{E}_{p(\tilde{\mathbf{X}}_{\bar{\mathcal{S}}} | \mathbf{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}})} \left[f(\mathbf{x}_{\mathcal{S}}, \tilde{\mathbf{X}}_{\bar{\mathcal{S}}}) \right]. \tag{5}$$

To elaborate on the consequences of the expectation taken w.r.t. $P(\tilde{\mathbf{X}}_{\bar{\mathcal{S}}} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}, \mathbf{X}_{\bar{\mathcal{S}}} = \mathbf{x}_{\bar{\mathcal{S}}})$, imagine a dataset with three variables, i.e. X_1, X_2, X_3, where X_1 is in-coalition and the task is to impute values for the out-of-coalition features X_2 and X_3. When using knockoff $\tilde{X}_2$ for imputation, this knockoff has been generated from a knockoff sampler that was fitted on the observed values of all three variables in the dataset. For the Shapley value calculation however, the data for imputation is required to condition on the observed value of X_1 only. Hence, the procedure leaks information from the out-of-coalition feature X_3 during the imputation of X_2. As a consequence, the range of values sampled for imputing out-of-coalition values will be too narrow, i.e. conditioned on more features than necessary, which reduces the entropy of the predicted values in $f(\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}})$. To be clear, the generated data throughout Shapley value calculation will still be on the same data manifold as the original samples, so this procedure effectively defends against adversarial attacks, which is what we focus on in this paper. We can interpret the restrictive conditioning as a form of regularization imposed through the data sampling mechanism. We therefore expect estimated Shapley values of lower magnitude when using knockoff imputation. As a result, on the one hand, conditioning on variables in the out-of-coalition set may introduce bias due to information leakage from other covariates; on the other hand, this will also lead to a reduction in variance of Shapley values that are estimated by approximation instead of exact calculation, which may be advantageous. We encourage future research to investigate potential trade-offs.

4 Experiments

4.1 Unfooling SHAP

We start off the section on experiments by illustrating that knockoff imputed Shapley values indeed are able to prevent adversarial attacks that make use of extrapolation. We replicate and extend the German Credit [9] experiments conducted by [23], where the task is to determine whether clients will be good customers ($\texttt{GoodCustomer} = 1$) or not ($\texttt{GoodCustomer} = -1$). We demonstrate that with knockoff imputation, the adversarial attack is no longer successful.

The fooling strategy of [23] can be illustrated in an intuitive manner: Imagine an unethical bank that aims to exploit sensitive information, such as the `Gender` of the credit applicant, to assess creditworthiness with a prediction model f. However, in the country the bank operates in, credit assessment tools are legally required to pass a so-called fairness audit prior to model deployment. During the audit, black-box models will be evaluated via SHAP [17]. If high importance scores are attributed to sensitive features, the model fails the audit and its deployment is legally prohibited. In spite of this danger, the bank is determined to use the discriminatory model f in its daily business, and comes up with a strategy to fool the auditors and make them believe that the bank uses an innocent prediction algorithm ψ, which utilises the non-discriminatory feature `LoanRateAsPercentOfIncome` only.

To fool the explanation, a supervised classification algorithm ω, in this case a random forest [4], is trained to distinguish in-distribution from OOD data. For in-distribution data, the adversarial model α uses the discriminatory prediction algorithm f, but for OOD data, α makes predictions according to ψ. For simplicity, the prediction model f here makes use of only the discriminatory variable `Gender`, whereas prediction model ψ determines the predictions according to the non-discriminatory variable `LoanRateAsPercentOfIncome`, see further Fig. 2, Panel **A**.

Extending the experiments of [23], we evaluate the adversarial model α not only with SHAP using marginal imputation with $kmeans(10)$ to summarize the background distribution, but also with knockoff imputed SHAP. We use sequential knockoffs [15] due to the presence of both categorical and continuous features in the dataset and generate $N_{ko} = 10$ knockoffs for each test data point. In brief, sequential knockoffs [15] are generated through an iterative procedure that samples values from a Gaussian (for continuous features) or multinomial (for categorical features) distribution. The distribution parameters are estimated from a penalized linear or multinomial logistic regression of the feature for which a knockoff copy should be generated on the other features and already sampled knockoffs from earlier iterations.

Figure 2 Panel **B** replicates results of [23], where we can see the aggregated explanations across test data instances suggesting `LoanRateAsPercentOfIncome` as most relevant for the model, hence the adversarial attack is successful. On the contrary, we can see from Panel **C** that with knockoff imputation, the explanation reveals the true importance of `Gender` for the model's prediction in the majority of test data points.

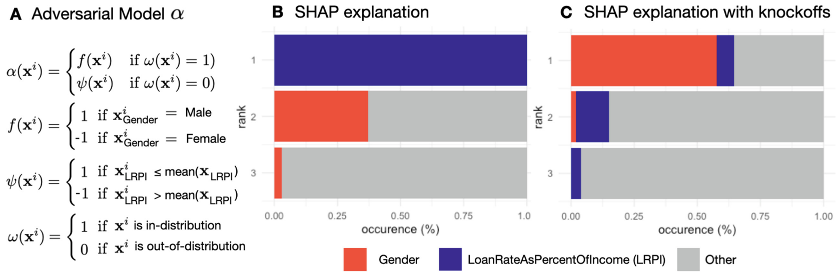

Fig. 2. Occurrences of features ranked amongst the top 3 features in SHAP explanations of test data, $N = 99$. Fidelity to detect real data as in-distribution with ω is 1.

Clearly, ω plays a crucial role for the success of the adversarial attack. In this experiment, we see that with knockoff imputation, ω is not able to detect OOD data as such, which is why the adversarial attack cannot unfold. At the same time, we have that ω always detects in-distribution data as such (fidelity = 1), so real world data will be reliably predicted with discriminatory model f. With this knowledge, advanced adversarial attackers might try to improve the model training of ω and optimize the decision threshold to reinstate a successful adversarial attack. However, with knockoff imputation – and conditional out-of-coalition-sampling in general – it is not possible to train an effective OOD classifier ω because the resulting data will lie on the same data manifold as the original data, hindering its detection as synthetic data. We discuss this issue further in Appendix B.

4.2 Unfooling SAGE

We now illustrate that global aggregates of Shapley values, SAGE values [8], suffer from the same vulnerability as local Shapley values and that knockoff imputation again can increase robustness. In this experiment, we simulate data, which further allows us to analyze key drivers in the data characteristics that affect the robustness against adversarial attacks.

We simulate data from $\mathbf{X}^{1000 \times 4} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = [0, \ldots, 0]$ and correlation matrix $\boldsymbol{\Sigma}$ with correlation coefficient $\rho \in [0, 0.9]$ in all off-diagonal elements. The target variable Y is defined by $Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{\beta} = [1, \ldots, 1]$ and $\epsilon \sim N(0, \sigma^2)$, such that σ^2 guarantees a signal to noise ratio $SNR = \frac{Var(Y)}{\sigma^2} = 2$. The adversarial model α is given in Fig. 3 Panel **A**.

SAGE with marginal imputation uses a background distribution summarized by $kmeans(10)$ and SAGE with knockoff imputation uses $N_{ko} = 10$ Gaussian knockoffs for the background distribution of each observation. Gaussian knockoffs are generated through sampling from a multivariate Gaussian distribution with parameters estimated from the observed features, see [5] for further details.

Figure 3 reveals that SAGE can be fooled just as SHAP and that knockoff imputation again prevents the adversarial attack, i.e. unfools the explanation.

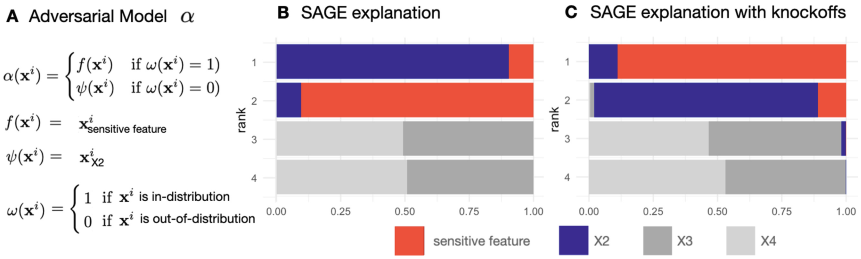

Fig. 3. Occurrences of features ranked amongst the top 4 in SAGE explanations with adversarial model α across 500 replicates. Correlation $\rho = 0.5$, fidelity of $\omega = 0.88$.

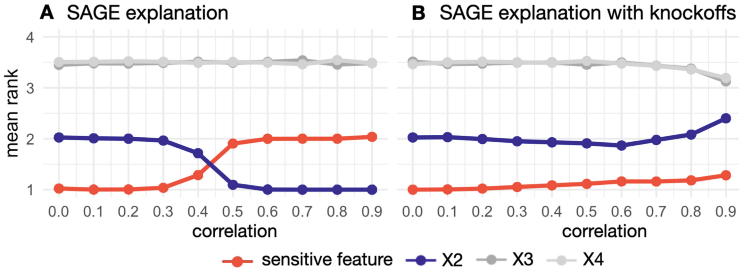

Fig. 4. Mean rank of features in SAGE explanations with adversarial classifier α across 500 replicates. We consider the adversarial attack successful if X_2 is ranked first, and unsuccessful if the sensitive feature is ranked first.

We can see from Fig. 4 that correlation strength plays a crucial role in the success of the attack. For marginal imputation (Fig. 4, Panel **A**), the adversarial attack is successful and hides the true rank of the sensitive feature (true rank is 1) for correlation levels greater than or equal to 0.5. Note that lower levels of correlation do not allow the adversarial attack to unfold regardless, because extrapolation is more pronounced and hence more likely to be detected by ω with medium or strongly correlated data (recall that $\rho = 0$ corresponds to feature independence). For SAGE with knockoffs (Fig. 4, Panel **B**), the model's usage of the sensitive feature is exposed for all levels of correlation. Hence, knockoff imputation effectively prevents the adversarial attack on SAGE explanations.

5 Discussion

We demonstrate that knockoff imputed Shapley values are robust against adversarial attacks that exploit extrapolated data. However, other adversarial attacks might be proposed. For example, because Shapley values are spread out across correlated features, the true importance of a sensitive feature could be toned down by adding correlated features to the model.

Further, the special characteristics of knockoffs may open up new trajectories in Shapley value research. One such example is SHAPLIT, which proposes conditional independence testing with FDR control for Shapley values [27]. Another promising approach could be to leverage the overly restrictive conditioning of knockoff imputed Shapley values for approximation tasks, where Shapley values are calculated with just a small fraction of all possible coalitions as opposed to exact Shapley value calculation. It is common in Shapley value software to optionally include some form of L_1 penalty on feature attributions to encourage sparse explanations, even when the underlying model f is not itself sparse [17]. Like many regularization methods, this effectively introduces bias in exchange for a decrease in variance. Knockoff imputed Shapley values may give a similar regularizing effect through the data sampling method rather than directly on the parameter estimation technique. This does not zero out feature attributions as the L_1 penalty does, but may serve to improve predictions for practitioners with limited computational budgets.

We want to emphasize that the use case for knockoff imputed Shapley values should be carefully chosen, since the method narrows down entropy of the target function, which may be disadvantageous in comparison to other methods when the computational capacity suffices to calculate exact conditional Shapley values.

Further, we want to highlight that a comparative benchmark study that analyzes variants for Shapley value calculation, including conditional Shapley value calculation, may be of great value for future research. For example, the knockoff-based approach proposed here could be compared with other conditional variants [1,2,20] both in terms of theory, e.g. analyzing the variance, and in empirical application, e.g. investigating the computational efficiency of the proposed algorithms and accuracy of estimates for different datasets. Such endeavors may further include novel methods that combine ideas from existing approaches. For example, one could use an overly strict conditioning set, as it is the case with knockoffs, for the conditional distribution based approaches to cut down the computational complexity of those approaches.

6 Conclusion

The paper presents an innovative approach to make Shapley explanations, such as SHAP [17] and SAGE [8], more robust against adversarial attacks by using model-X knockoffs. The discussion on theoretical guarantees and implications reveals that knockoffs can serve as a flexible and off-the-shelf methodology that effectively prevents extrapolation during Shapley value calculation. Through both real data and simulated data experiments, the paper demonstrates that vulnerability to adversarial attacks can be successfully reduced. It is worth emphasizing that the presented methodology can be used in conjunction with any valid knockoff sampling procedure and not only the deep [18], sequential [15] and Gaussian knockoffs [5] used in this paper, which further highlights the flexibility of the proposed approach. This, and the possibility to sample knockoffs upfront, which drastically reduces computational complexity, is a major advantage over conditional Shapley value calculation approaches that may otherwise be used for the prevention of adversarial attacks.

Acknowledgements. MNW and KB received funding for this project from the German Research Foundation (DFG), Emmy Noether Grant 437611051.

Data and Code availability. Reproducible code for the results presented in this paper is available at https://github.com/bips-hb/unfooling_shapley.

A Knockoff Imputed SAGE Values

Algorithm 2. Sampling-based approximation for SAGE values [8] with knockoff imputation

```
Input: data $(\mathbf{X}, Y)$, model $f$, loss function $l$, outer samples $n$, number of knockoffs
    $N_{ko}$, knockoff sampler $ko()$
Initialize $\hat{\phi}_1 = 0, \hat{\phi}_2 = 0, \ldots, \hat{\phi}_d = 0$
$\hat{y}_{\text{init}} = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i)$
train knockoff sampler $k(\mathbf{X})$
for i = 1 to n do
    Sample a data instance $(\mathbf{x}_i, y_i)$
    Sample instance $\pi$, a permutation of $\mathcal{D}$
    $\mathcal{S} = \emptyset$
    $\text{L}_{\text{prev}} = l(\hat{y}_{\text{init}}, \text{y})$
    for $j$ in $\mathcal{D}$ do
        $\mathcal{S} = \mathcal{S} \cup \{\pi[j]\}$
        $\hat{y} = 0$
        for $k = 1$ to $N_{ko}$ do
            Sample $\tilde{\mathbf{x}}^k = \tilde{x}_1^k, \ldots, \tilde{x}_d^k$ from $ko(\mathbf{X})$
            $\hat{y} = \hat{y} + f(\mathbf{x}_{\mathcal{S}}, \tilde{\mathbf{x}}^k_{\bar{\mathcal{S}}})$
        end for
        $\bar{\hat{y}} = \frac{\hat{y}}{N_{ko}}$
        $\text{L} = l(\bar{\hat{y}}, y)$
        $\Delta = \text{L}_{\text{prev}} - \text{L}$
        $\hat{\phi}_{\pi[j]} = \hat{\phi}_{\pi[j]} + \Delta$
        $\text{L}_{\text{prev}} = \text{L}$
    end for
end for
return  SAGE values $\frac{\hat{\phi}_1}{n}, \ldots, \frac{\hat{\phi}_d}{n}$
```

B The Role of $\boldsymbol{\omega}$

The detection of OOD data is the backbone of adversarial attacks which is why the performance of ω is crucial for the success of the attack. Let us take a closer look onto how ω is derived: In a first step, we take the training data and add the label $y = 1$ to it, because it is real world data. Then, we mimic the imputation procedure that occurs during Shapley value calculation. This means we treat some feature values as in-coalition, i.e. we take the actual value of these features for the respective observation, and impute the remaining out-of-coalition feature values with other quantities, for example, by sampling values from the marginal distribution. The resulting 'new' observations will then be labeled with $y = 0$, because this is generated data. In a second step, we fit a random forest classifier to distinguish in-distribution from OOD data, or technically speaking, to distinguish $y = 1$ from $y = 0$ data.

In the experiments of Sect. 4.1, we train ω exactly like [23], where we impute the out-of-coalition features with values from a background distribution that is characterised by the center points of a $kmeans(c = 10)$ clustering of the data. Note that this is the exact same background distribution which is used for the marginal imputation in the SHAP explanations later on. With this ω, the adversarial attack is successful for marginal imputation, but with the very same ω, SHAP explanations using knockoff imputation prevents the attack.

To some readers, it might have been clear from the beginning that an ω which is trained to detect OOD samples that are generated with $kmeans(c = 10)$ will not perform well in detecting OOD samples that occur when calculated knockoff imputed SHAP, where the background distribution consists of knockoffs and not $kmeans(c = 10)$ cluster centers. An advanced adversarial attacker strategy could be to train ω for detecting OOD samples from many different background distributions, including a knockoff background distribution. However, the crucial point is that with conditional out-of-coalition-sampling, for example knockoffs, it is not possible to train an effective OOD classifier ω because the $y = 1$ and $y = 0$ data points will lie on the same data manifold. In other words, there is no difference in in-distribution and OOD data, which hinders its detection as synthetic data. When training ω on such data, the classifier cannot learn reasonable information from the data.

The implications of this are illustrated in Fig. 5. There, we train ω on data that is generated by knockoff imputation. We vary the hyperparameters for the random forest classifier to force model ω to overfit, i.e. be less (Fig. 5, Panel **A**) or more (Fig. 5, Panel **B**) conservative in predicting data as OOD. This can be achieved by varying the number of trees in the random forest classifier, and the number of $y = 0$ training samples we generate. We denote the hyperparameters with $\omega(\text{number of trees}, \text{number of samples generated})$.

For an adversarial attacker, the aim is high fidelity, i.e. a high percentage of true in-distribution classifications by ω and a high rank of the innocent feature `LoanRateAsPercentOfIncome` in the SHAP explanation. Different hyperparameter settings reveal that there is a trade-off between fidelity and the degree to which the innocent feature `LoanRateAsPercentOfIncome` is ranked up high. If the adversarial attacker is keen to predict real-world data with the discriminatory model, i.e. uses an ω that is conservative in classifying data as OOD, then knockoff imputed SHAP reveals the sensitive feature `Gender` as highly important (Fig. 5, Panel **B**). On the contrary, if the adversarial attacker prioritizes that the explanation should pretend that `LoanRateAsPercentOfIncome` is important, i.e.

uses an ω that is liberal in predicting data as OOD, then the fidelity of ω drops drastically (Fig. 5, Panel **A**). This is clearly in contrast to the overarching goal of adversarial attackers to use the discriminatory model for in-distribution (real world) applications, but fool the SHAP explanation such that the model appears innocent.

Consequently, when using knockoff imputed SHAP, the adversarial attacker is forced to use the fair model if the SHAP evaluation should suggest that the model is fair – in other words and recollecting the example stated in the main text before: The only way to pass a fairness audit that uses knockoff imputed SHAP explanations is using a fair model.

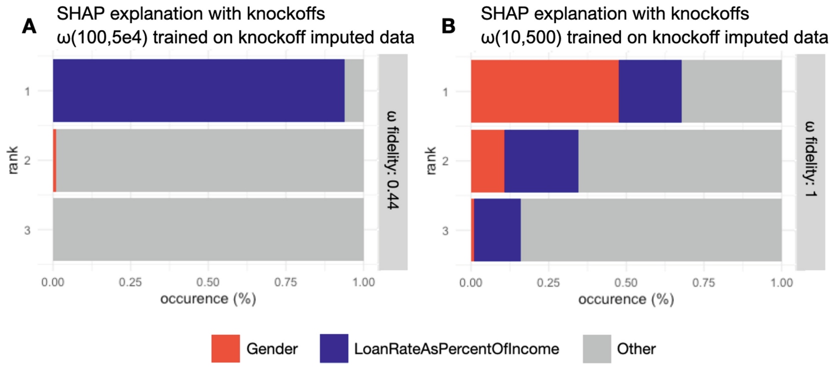

Fig. 5. Occurrences of features ranked amongst the top 3 features in SHAP explanations of $N = 99$ test data points.

References

1. Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif. Intell. **298**, 103502 (2021)
2. Aas, K., Nagler, T., Jullum, M., Løland, A.: Explaining predictive models using Shapley values and non-parametric vine copulas. Depend. Model. **9**(1), 62–81 (2021)
3. Bates, S., Candès, E., Janson, L., Wang, W.: Metropolized knockoff sampling. J. Am. Stat. Assoc. **116**(535), 1413–1427 (2021)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Candès, E., Fan, Y., Janson, L., Lv, J.: Panning for gold: model-free knockoffs for high-dimensional controlled variable selection. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **80**(3), 551–577 (2018)
6. Chen, H., Covert, I.C., Lundberg, S.M., Lee, S.I.: Algorithms to estimate Shapley value feature attributions. arXiv:2207.07605 (2022)
7. Chen, H., Janizek, J.D., Lundberg, S., Lee, S.I.: True to the model or true to the data? arXiv:2006.16234 (2020)

8. Covert, I., Lundberg, S.M., Lee, S.I.: Understanding global feature contributions with additive importance measures. In: Advances in Neural Information Processing Systems, vol. 33, pp. 17212–17223 (2020)
9. Dua, D., Graff, C.: UCI machine learning repository (2017)
10. Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., Feige, I.: Shapley explainability on the data manifold. arXiv:2006.01272 (2020)
11. Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., Holmes, C.C.: On locality of local explanation models. In: Advances in Neural Information Processing Systems, vol. 34, pp. 18395–18407 (2021)
12. Gu, J., Yin, G.: Bayesian knockoff filter using Gibbs sampler. arXiv:2102.05223 (2021)
13. Hooker, G., Mentch, L., Zhou, S.: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat. Comput. **31**(6), 1–16 (2021)
14. Jordon, J., Yoon, J., van der Schaar, M.: KnockoffGAN: generating knockoffs for feature selection using generative adversarial networks. In: International Conference on Learning Representations (2019)
15. Kormaksson, M., Kelly, L.J., Zhu, X., Haemmerle, S., Pricop, L., Ohlssen, D.: Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool. Stat. Med. **40**(14), 3313–3328 (2021)
16. Liu, Y., Zheng, C.: Auto-encoding knockoff generator for FDR controlled variable selection. arXiv:1809.10765 (2018)
17. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Romano, Y., Sesia, M., Candès, E.: Deep knockoffs. J. Am. Stat. Assoc. **115**(532), 1861–1872 (2020)
19. Ribeiro, M. T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining ACM SIGKDD 22, pp. 1135–1144 (2016)
20. Redelmeier, A., Jullum, M., Aas, K.: Explaining predictive models with mixed features using Shapley values and conditional inference trees. In: Proceedings of the 4th International Cross-Domain Conference for Machine Learning and Knowledge Extraction CD-MAKE, pp. 117–137 (2020)
21. Sesia, M., Sabatti, C., Candès, E.J.: Gene hunting with hidden Markov model knockoffs. Biometrika **106**(1), 1–18 (2018)
22. Shapley, L.: A value for n-person games. In: Kuhn, H., Tucker, A. (eds.) Contributions to the Theory of Games II. Princeton University Press, Princeton (1953)
23. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180–186 (2020)
24. Štrumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. **11**, 1–18 (2010)
25. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2014)
26. Sudarshan, M., Tansey, W., Ranganath, R.: Deep direct likelihood knockoffs. In: Advances in Neural Information Processing Systems, vol. 33, pp. 5036–5046 (2020)

27. Teneggi, J., Bharti, B., Romano, Y., Sulam, J.: From Shapley back to Pearson: hypothesis testing via the Shapley value. arXiv:2207.07038 (2022)
28. Vreš, D., Robnik-Šikonja, M.: Preventing deception with explanation methods using focused sampling. Data Mining Knowl. Discov. (2022)
29. Williamson, B., Feng, J.: Efficient nonparametric statistical inference on population feature importance using Shapley values. In: International Conference on Machine Learning, pp. 10282–10291. PMLR (2020)