Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution

Tsuyoshi Idé
tide@us.ibm.com
IBM Research, Thomas J. Watson Research Center
Yorktown Heights, New York, USA

Naoki Abe
nabe@us.ibm.com
IBM Research, Thomas J. Watson Research Center
Yorktown Heights, New York, USA

ABSTRACT

We address the task of probabilistic anomaly attribution in the black-box regression setting, where the goal is to compute the probability distribution of the attribution score of each input variable, given an observed anomaly. The training dataset is assumed to be unavailable. This task differs from the standard XAI (explainable AI) scenario, since we wish to explain the anomalous deviation from a black-box prediction rather than the black-box model itself.

We begin by showing that mainstream model-agnostic explanation methods, such as the Shapley values, are not suitable for this task because of their "deviation-agnostic property." We then propose a novel framework for probabilistic anomaly attribution that allows us to not only compute attribution scores as the predictive mean but also quantify the uncertainty of those scores. This is done by considering a generative process for perturbations that counter-factually bring the observed anomalous observation back to normalcy. We introduce a variational Bayes algorithm for deriving the distributions of per variable attribution scores. To the best of our knowledge, this is the first probabilistic anomaly attribution framework that is free from being deviation-agnostic.

CCS CONCEPTS

• **Mathematics of computing** → **Computing most probable explanation**; **Variational methods**.

KEYWORDS

explainable AI (XAI), anomaly attribution, generative model, variational inference, Shapley value, integrated gradient

ACM Reference Format:
Tsuyoshi Idé and Naoki Abe. 2023. Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599365

1 INTRODUCTION

Over the last decade, we have witnessed a dramatic resurgence of deep neural networks (DNNs) and numerous attempts to use DNNs in real-world applications. Despite their remarkable achievements, growing concerns are also expressed regarding the lack of transparency in advanced machine learning (ML) algorithms, making explainable artificial intelligence (XAI) an active research area in the data mining community. While early XAI studies tended to focus on the psychological aspects of how AI should be made explainable, the bulk of research interest is now shifting towards *actionability* in business and industrial applications, as the adoption of AI is becoming more widespread [15, 22].

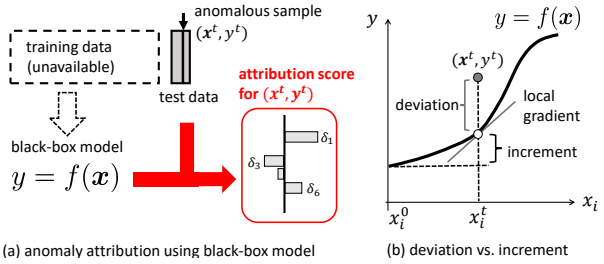

Figure 1: Problem setting and motivation. (a) Given a black-box deterministic regression model and anomalous sample(s), our goal is to find the *probability distribution* of input variables' responsibility scores without access to the training data. (b) Existing attribution methods attempt to explain either the local gradient or the increment from a reference point x^0, rather than the deviation of the sample in question.

One important problem in this context is how to explain an unusual event, observed as a significant discrepancy from the prediction of an ML model. Although this problem encompasses various different scenarios, we are particularly interested in the task of **anomaly attribution in the doubly black-box regression setting** (see Fig. 1 (a)): We are given a black-box regression model $y = f(\boldsymbol{x})$, where y is the real-valued noisy output (such as miles per gallon) and $\boldsymbol{x}$ is a vector of noisy real-valued input variables (such as driver's weight and average speed). We have access to the API (application programming interface) of $f(\cdot)$ but do not have access to either its parametric form or training data (hence, "doubly"). Given a limited amount of test samples, we ask: how can we quantify the contribution of each input variable in the face of an unexpected deviation between observation and prediction?

This question has typically been addressed with one of the following three model-agnostic post-hoc XAI methods in the literature: 1) Local linear surrogate modeling, which is best known under the name LIME (Local Interpretable Model-agnostic Explanations) [33]; 2) Shapley value (SV), which was first introduced to the ML community by [41]; and 3) integrated gradient (IG) [44].

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599365

Despite their popularity, however, there are two major limitations with those methods. One is that all of them are, in fact, "**deviation-agnostic**," meaning that they explain the black-box function $f(\cdot)$ itself in the form of the local gradient or an increment, not the observed deviation, as illustrated in Fig. 1 (b). Here, note that, unlike the standard XAI scenarios, we seek explanations *relative to* the deviation from a black-box prediction, as we will discuss in detail later. The other limitation is that they have limited capabilities of quantifying the uncertainty of the attribution scores. Motivated by the requirements from industrial applications, e.g., [29], uncertainty quantification (UQ) of attribution scores is becoming a major topic in XAI research. In the black-box setting without access to the training data, however, this problem is considered extremely challenging and limited work has been done to date. Existing works include empirical comparative studies, e.g., [48, 50], and semi-theoretical analysis based on known results of probabilistic linear regression [12, 16, 39, 49].

In this paper, we propose a novel probabilistic framework called the **generative perturbation analysis** (GPA) for anomaly attribution in the black-box regression setting, which we believe is the first fully probabilistic black-box attribution algorithm. The key idea is to consider a counterfactual data generative process including perturbation $\boldsymbol{\delta}$ as a model parameter, and reduce the task of attribution to that of statistical parameter estimation. In this way, the uncertainty in attribution is naturally evaluated by finding its posterior distribution. Here we additionally introduce a novel idea of using variational Bayes inference to decompose the contribution of each of the input variables.

To summarize, our contributions are: 1) to mathematically show that the existing attribution methods have the deviation-agnostic property; 2) to uncover their interrelationship that has been hitherto unnoticed; and 3) to propose the first generative framework for anomaly attribution.

2 RELATED WORK

Anomaly attribution has been studied as a sub-task of anomaly detection in the ML community, typically in the white-box unsupervised setting. In the supervised setting, the majority of prior works are about either model- or classification-specific algorithms. For example, saliency maps [35, 36] and layer-wise relevance propagation [26] are well-known *model-specific* attribution methods. Sainyam et al. [11] leveraged a counterfactual framework [14] for probabilistic black-box explanations in the classification setting with binary variables. Similar approaches have been discussed under the terms like perturbation-based or mask-based (e.g. [8, 9, 32]), but most of them are for classification without the capability of computing the distribution of attribution score and are not directly applicable to the present setting.

In the *model-agnostic* regression setting, 1) local linear modeling, 2) SV, and 3) IG have been widely used for black-box attribution, as summarized in Table 1, along with three additional methods: The expected integrated gradient (EIG) [6], which is a generalized version of IG, the Z-score, which is a standard outlier detection metric in the *unsupervised* setting, and likelihood compensation (LC) [20], which conducts a semi-probabilistic analysis for attribution. In the context of *anomaly* attribution, LIME and its variant have been applied to anomaly explanation [13, 47]. SV is used in sensor fault diagnosis [18] and for explaining unexpected observations in crop yield analysis [25] and unusual warranty claims [1]. Also, the use of IG for anomaly explanation is discussed by Sipple [37, 38].

Interestingly, it has been suggested that these attribution methods may have some mutual connection. Prior work along this line includes Deng et al. [6], which attempted to characterize IG using Taylor expansion and gave the first definition of EIG. Also, Sundararajan and Najmi [43] proposed a unified attribution framework, where they pointed out that there can be a few different definitions for SV and discussed the relationship with IG in a qualitative manner. Lundberg and Lee [24] reintroduced the SV-based attribution method [41] to propose a hybrid method between SV and LIME. Inspired by these works, we go one step further in this paper: We explicitly show a mathematical relationship between those existing attribution methods, and show that the deviation-agnostic property (see the 'y-sensitive' column in Table 1) is an inherent consequence of the common mathematical structure.

Another important contribution of this paper is the proposal of a principled framework for probabilistic prediction of attribution scores. Most of the existing works tackling this problem [12, 39, 49] under settings similar to ours use the standard result of probabilistic linear regression (see, e.g., Chap. 3 of [3]) to evaluate uncertainty in the regression coefficients as the LIME attribution score (the 'built-in UQ' column in Table 1). However, the black-box model $f(\boldsymbol{x})$ is generally highly nonlinear; It is not clear to what extent the theoretical results of the linear model apply. Also, it is not clear how the distribution of the attribution score is computed for each input variable (hence 'yes/no' in the table). In fact, BayLIME [49]'s posterior covariance is a constant that depends only on the hyperparameters independently of $f(\cdot)$ (See Sec. 6.3). LC [20] shares a similar starting point with ours but differs fundamentally in that it is not able to compute the probability distribution of the attribution score. Guo et al. [16] used a Dirichlet-enhanced probabilistic linear regression mixture but it is intended for global model explanations rather than local anomaly attribution.

3 PROBLEM SETTING

As mentioned earlier, we focus on the task of anomaly attribution in the *regression* setting rather than classification or unsupervised settings. Figure 1 (a) summarizes the overall problem setting. Suppose we have a (deterministic) regression model $y = f(\boldsymbol{x})$ in the *doubly black-box setting*: Neither the training data set $\mathcal{D}_{\text{train}}$ nor the (true) distribution of $\boldsymbol{x}$ is available (see the 'training-data-free' column in Table 1). Throughout the paper, the input variable $\boldsymbol{x} \in \mathbb{R}^M$ and the output variable $y \in \mathbb{R}$ are assumed to be *noisy real-valued*, where M is the dimensionality of the input vector. We also assume that queries to get the response $f(\boldsymbol{x})$ can be performed cheaply at any $\boldsymbol{x}$.

In practice, anomaly attribution is typically coupled with anomaly detection: When we observe a test sample $(\boldsymbol{x}, y) = (\boldsymbol{x}^t, y^t)$, we first compute an anomaly score $a^t = a(\boldsymbol{x}^t, y^t)$ to quantify how anomalous it is. Then, if $a^t \in \mathbb{R}$ is high enough, we go to the next step of anomaly attribution. In this scenario, the task of anomaly attribution is defined as follows.

Definition 1 (**probabilistic anomaly attribution**). *Given a black-box regression model $y = f(\boldsymbol{x})$ and observed test sample(s),*

Table 1: Comparison of model-agnostic attribution methods in the regression setting.

	model-agnostic	training-data-free	baseline-input-free	y-sensitive	built-in UQ	reference point
LIME [33]	yes	yes	yes	no	yes/no	infinitesimal vicinity
SV [41, 42]	yes	no	yes	no	no	globally distributional
IG [37, 44]	yes	yes	no	no	no	arbitrary
EIG [6]	yes	no	yes	no	no	globally distributional
Z-score [5]	yes	no	yes	no	no	global mean of predictors
LC [20]	yes	yes	yes	yes	no	maximum likelihood point
GPA	**yes**	**yes**	**yes**	**yes**	**yes**	maximum a posteriori point

compute the distribution of the score for each input variable indicative of the extent to which that variable is responsible for the sample being anomalous.

We can readily generalize the problem to that of *collective* probabilistic anomaly detection and attribution. Specifically, given a test data set $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}^t, y^t) \mid t = 1, \ldots, N_{\text{test}}\}$, where t is the index for the t-th test sample and N_{test} is the number of test samples, we can consider anomaly score as well as attribution score distributions for the whole test set $\mathcal{D}_{\text{test}}$.

The standard approach to **anomaly detection** is to use the negative log-likelihood of the test sample(s) as the anomaly score (See, e.g., [23, 28, 40, 45, 46]). Assume that, from the deterministic regression model, we can somehow obtain $p(y \mid \boldsymbol{x})$, a probability density over y given the input signal $\boldsymbol{x}$. Under the i.i.d. assumption, the anomaly score can be written as

$$a(\boldsymbol{x}^t, y^t) = -\ln p(y^t \mid \boldsymbol{x}^t), \quad \text{or,} \tag{1}$$

$$a(\mathcal{D}_{\text{test}}) = -\tfrac{1}{N_{\text{test}}} \textstyle\sum_{t=1}^{N_{\text{test}}} \ln p(y^t \mid \boldsymbol{x}^t), \tag{2}$$

corresponding to the single sample and collective cases, respectively. **Anomaly attribution** is the task to attribute a high anomaly score to each of the input variables.

Notation. We use boldface to denote vectors. The i-th dimension of a vector $\boldsymbol{\delta}$ is denoted as δ_i. The ℓ_1 and ℓ_2 norms of a vector are denoted by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, and are defined as $\|\boldsymbol{\delta}\|_1 \triangleq \sum_i |\delta_i|$ and $\|\boldsymbol{\delta}\|_2 \triangleq \sqrt{\sum_i \delta_i^2}$. The sign function $\text{sign}(\delta_i)$ is defined as being 1 for $\delta_i > 0$, and -1 for $\delta_i < 0$. For $\delta_i = 0$, the function takes an indeterminate value in $[-1, 1]$. For a vector input, the definition applies element-wise, yielding a vector of the same size as the input vector. We distinguish between a random variable and its realizations via the absence or presence of a superscript. For notational simplicity, we use $p(\cdot)$ or $P(\cdot)$ as a proxy to represent different probability distributions, whenever there is no confusion. For instance, $p(\boldsymbol{x})$ is used to represent the probability density of a random variable $\boldsymbol{x}$ while $p(y \mid \boldsymbol{x})$ is a different distribution of another random variable y conditioned on $\boldsymbol{x}$.

4 EXISTING ATTRIBUTION METHODS ARE DEVIATION-AGNOSTIC

This section summarizes our remarkable new results on the existing attribution methods: 1) IG, SV, and LIME are inherently deviation-agnostic and are not appropriate for anomaly attribution, 2) SV is equivalent to EIG up to the second order in the power expansion, and 3) LIME can be derived as the derivative of IG or EIG in a certain limit. Throughout this subsection, we assume that the derivative of the black-box regression function $f(\cdot)$ is computable somehow to an arbitrary order.

Formally, the deviation-agnostic property is defined as follows:

Definition 2 (**deviation-agnostic**). *An anomaly attribution method A is said to be* deviation-agnostic *if for any black-box regression model $f(\cdot)$, observed test sample $(\boldsymbol{x}^t, y^t)$, deviation Δ and input variable i, $A_{f,i}(\boldsymbol{x}^t, y^t) = A_{f,i}(\boldsymbol{x}^t, y^t + \Delta)$, where $A_{f,i}(\boldsymbol{x}^t, y^t)$ denotes the attribution score computed by A for f, i and $(\boldsymbol{x}^t, y^t)$.*

We often drop the subscript f when it is clear from the context.

4.1 Deviation-agnostic properties

4.1.1 LIME. In general, the local linear surrogate modeling approach fits a linear regression model locally to explain a black-box function in the vicinity of a given test sample $(\boldsymbol{x}^t, y^t)$. For anomaly attribution, we need to consider the *deviation function* $F(\boldsymbol{x}, y) \triangleq f(\boldsymbol{x}) - y$ instead of $f(\boldsymbol{x})$. Algorithm 1 summarizes the local anomaly attribution procedure. Let β_i denote the i-th output by $\text{LIME}_i(\boldsymbol{x}^t, y^t)$. Rather unexpectedly, despite the modification to fit $F(\boldsymbol{x}, y)$ rather than $f(\boldsymbol{x})$, the following property holds:

Theorem 1. *LIME is deviation-agnostic:* $\text{LIME}_i(\boldsymbol{x}^t, y^t) = \text{LIME}_i(\boldsymbol{x}^t)$.

Proof. With ν being the ℓ_1 regularization strength, the loss function for LIME is written as

$$\begin{aligned}\Psi(\boldsymbol{\beta}, \beta_0) &= \frac{1}{N_s} \sum_{n=1}^{N_s} (z^{t[n]} - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{x}^{t[n]})^2 + \nu \|\boldsymbol{\beta}\|_1, \\ &= \frac{1}{N_s} \sum_{n=1}^{N_s} (f(\boldsymbol{x}^{t[n]}) - (y^t + \beta_0) - \boldsymbol{\beta}^\top \boldsymbol{x}^{t[n]})^2 + \nu \|\boldsymbol{\beta}\|_1,\end{aligned}$$

which is equivalent to the lasso objective for LIME with the intercept $y^t + \beta_0$. Since the lasso objective is convex, the solution $\boldsymbol{\beta}$ is unique. With an arbitrary adjusted intercept, the attribution score $\boldsymbol{\beta}$ remains unchanged. Hence, $\forall i$, $\text{LIME}_i(\boldsymbol{x}^t, y^t) = \text{LIME}_i(\boldsymbol{x}^t)$. □

In the local linear surrogate modeling approach, the final attribution score can vary depending on the nature of the regularization term. For the theoretical analysis below, we use a generic algorithm by setting $\nu \to 0_+$ in Algorithm 1, and call the resulting attribution score LIME_i^0 for $i = 1, \ldots, M$. As is well-known, LIME_i^0 is a local estimator of $\partial f / \partial x_i$ at $\boldsymbol{x} = \boldsymbol{x}^t$ if f is locally differentiable.

4.1.2 Integrated gradient. For anomaly attribution, which is an *input* attribution task, IG [37, 44] should be computed for the deviation function $F(\boldsymbol{x}, y) \triangleq f(\boldsymbol{x}) - y$ rather than f alone as

$$\text{IG}_i(\boldsymbol{x}^t, y^t \mid \boldsymbol{x}^0, y^0) \triangleq (x_i^t - x_i^0) \int_0^1 \mathrm{d}\alpha \left.\frac{\partial F}{\partial x_i}\right|_\alpha \tag{3}$$

Algorithm 1 Local linear surrogate model for $F(\boldsymbol{x}, y)$

Require: $f(\boldsymbol{x})$, test point $(\boldsymbol{x}^t, y^t)$, regularization parameter ν.

1: Randomly populate N_s points $\{\boldsymbol{x}^{t[1]}, \ldots, \boldsymbol{x}^{t[N_s]}\}$ in the vicinity of $\boldsymbol{x}^t$ ($N_s \sim 1000$).
2: Compute the deviation $z^{t[n]} \triangleq f(\boldsymbol{x}^{t[n]}) - y^t$ for all n.
3: Fit a linear model $z = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}$ using the ℓ_1 weight ν to the dataset $\{(\boldsymbol{x}^{t[n]}, z^{t[n]}) \mid n = 1, \ldots, N_s\}$.
4: **return** $\boldsymbol{\beta}$, which is the local attribution score at $(\boldsymbol{x}^t, y^t)$.

for $i = 1, \ldots, M$, where the gradient is estimated at $\boldsymbol{x} = \boldsymbol{x}^0 + (\boldsymbol{x}^t - \boldsymbol{x}^0)\alpha$ and $y = y^0 + (y^t - y^0)\alpha$. The baseline input $(\boldsymbol{x}^0, y^0)$ has to be determined from prior knowledge. We also define EIG by integrating out the baseline input:

$$\mathrm{EIG}_i(\boldsymbol{x}^t, y^t) \triangleq \int \mathrm{d}y^0 \int \mathrm{d}\boldsymbol{x}^0 P(\boldsymbol{x}^0, y^0) \mathrm{IG}_i(\boldsymbol{x}^t, y^t \mid \boldsymbol{x}^0, y^0), \quad (4)$$

where $P(\boldsymbol{x}, y)$ is the joint distribution of $\boldsymbol{x}$ and y, which is actually unavailable in our setting. The following property holds:

Theorem 2. *IG and EIG are deviation-agnostic:* $\mathrm{IG}_i(\boldsymbol{x}^t, y^t) = \mathrm{IG}_i(\boldsymbol{x}^t)$ *and* $\mathrm{EIG}_i(\boldsymbol{x}^t, y^t) = \mathrm{EIG}_i(\boldsymbol{x}^t)$.

Proof. We define

$$\mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0) \triangleq (x_i^t - x_i^0) \int_0^1 \mathrm{d}\alpha \left. \frac{\partial f}{\partial x_i} \right|_{\boldsymbol{x}^0 + (\boldsymbol{x}^t - \boldsymbol{x}^0)\alpha} \quad (5)$$

and $\mathrm{EIG}_i(\boldsymbol{x}^t) \triangleq \int \mathrm{d}\boldsymbol{x}^0 P(\boldsymbol{x}^0) \mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0)$. Since $\frac{\partial F}{\partial x_i} = \frac{\partial f}{\partial x_i}$, the statement about IG holds. Also, for EIG, the integration w.r.t. y^0 produces $\int \mathrm{d}y^0 P(\boldsymbol{x}^0, y^0) = P(\boldsymbol{x}^0)$, yielding $\mathrm{EIG}_i(\boldsymbol{x}^t, y^t) = \mathrm{EIG}_i(\boldsymbol{x}^t)$. □

4.1.3 Shapley value. There are a few different versions of SV in the literature [43]. Here we adopt the definition of the conditional expectation SV applied to the deviation function:

$$\mathrm{SV}_i(\boldsymbol{x}^t, y^t) = \frac{1}{M} \sum_{k=0}^{M-1} \binom{M-1}{k}^{-1} \sum_{\mathcal{S}_i : |\mathcal{S}_i| = k} \Delta f(\mathcal{S}_i), \quad (6)$$

where $\mathcal{S}_i$ denotes any subset of the variable indices $i \in \{1, \ldots, M\}$ excluding i. $|\mathcal{S}_i|$ is the size of $\mathcal{S}_i$. The second summation runs over all possible choices of $\mathcal{S}_i$ under the constraint $|\mathcal{S}_i| = k$ from the first summation. We also define the complement $\bar{\mathcal{S}}_i$, which is the subset of $\{1, \ldots, M\}$ excluding i and $\mathcal{S}_i$. For example, if $M = 12$, $i = 3$ and $\mathcal{S}_i = \{1, 2\}$, the complement $\bar{\mathcal{S}}_i$ will be $\{4, 5, \ldots, 12\}$. Corresponding to this division, we rearrange the M variables as $\boldsymbol{x} = (x_i, \boldsymbol{x}_{\mathcal{S}_i}, \boldsymbol{x}_{\bar{\mathcal{S}}_i})$. Finally, the $\Delta f(\mathcal{S}_i)$ term is defined as the difference between the expected values of F under two different conditions: One is $(x_i, \boldsymbol{x}_{\mathcal{S}_i}, y) = (x_i^t, \boldsymbol{x}_{\mathcal{S}_i}^t, y^t)$ with $\boldsymbol{x}_{\bar{\mathcal{S}}_i}$ to be integrated out. The other is $(\boldsymbol{x}_{\mathcal{S}_i}, y) = (\boldsymbol{x}_{\mathcal{S}_i}^t, y^t)$ with $(x_i, \boldsymbol{x}_{\bar{\mathcal{S}}_i})$ to be integrated out. We denote them by $\langle F \mid x_j^t, \boldsymbol{x}_{\mathcal{S}_j}^t, y^t \rangle$ and $\langle F \mid \boldsymbol{x}_{\mathcal{S}_j}^t, y^t \rangle$, respectively.

The following property holds:

Theorem 3. *SV is deviation-agnostic:* $\mathrm{SV}_i(\boldsymbol{x}^t, y^t) = \mathrm{SV}_i(\boldsymbol{x}^t)$.

Proof. Since F is linear in y, we can easily see that $\langle F \mid x_j^t, \boldsymbol{x}_{\mathcal{S}_j}^t, y^t \rangle = \langle f \mid x_j^t, \boldsymbol{x}_{\mathcal{S}_j}^t \rangle - y^t$ and $\langle F \mid \boldsymbol{x}_{\mathcal{S}_j}^t, y^t \rangle = \langle f \mid \boldsymbol{x}_{\mathcal{S}_j}^t \rangle - y^t$ hold, which implies $\mathrm{SV}_i(\boldsymbol{x}^t, y^t) = \mathrm{SV}_i(\boldsymbol{x}^t)$. □

4.2 Relationship between IG, SV, and LIME

The fact that (E)IG, SV, and LIME share the same deviation-agnostic property suggests that they may share a common mathematical structure. In what follows, we show two results showing the interrelationship between them.

4.2.1 SV and EIG. First, let us consider the relationship between SV and EIG. The integral in IG and the combinatorial definition SV are major obstacles in getting deeper insights into what they really represent. This issue can be partially resolved by resorting to power expansion. The following remarkable property holds:

Theorem 4 (Equivalence of SV to EIG). *a)* SV_i *is equivalent to* EIG_i *for* $\forall i$ *up to the second order of the power expansion. b) SV and EIG satisfy exactly the same sum rule:*

$$\sum_{i=1}^{M} \mathrm{SV}_i(\boldsymbol{x}^t) = \sum_{i=1}^{M} \mathrm{EIG}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}) - \langle f \rangle, \quad (7)$$

where $\langle f \rangle \triangleq \int \mathrm{d}\boldsymbol{x}\, P(\boldsymbol{x}) f(\boldsymbol{x})$.

We leave the proof to our companion paper [19] due to space limitations. While the sum rule b) is known, Theorem 4 a) is the first result directly establishing the fact that $\forall i, \mathrm{SV}_i \approx \mathrm{EIG}_i$, to the best of our knowledge. In Sec. 6, we empirically show that indeed SV and EIG systematically give similar attribution scores.

4.2.2 LIME and EIG. Second, let us now consider the relationship between LIME and EIG. LIME, as a local linear surrogate modeling approach, differs from EIG and SV in two regards. First, LIME does not need the true distribution $P(\boldsymbol{x})$. Instead, it uses a local distribution to populate local samples. Second, LIME is defined as the gradient, not a differential increment. These observations lead us to an interesting question: Is the *derivative of EIG* in the local limit the same as the LIME attribution score? The following theorem answers this question affirmatively:

Theorem 5 (LIME and IG). *The derivative of IG and EIG is equivalent to LIME:*

$$\mathrm{LIME}_i^0(\boldsymbol{x}^t) = \lim_{\eta \to 0} \frac{\partial \mathrm{EIG}_i(\boldsymbol{x}^t)}{\partial x_i} = \lim_{\boldsymbol{x}^0 \to \boldsymbol{x}^t} \frac{\partial \mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0)}{\partial x_i}, \quad (8)$$

where the localized Gaussian $P(\boldsymbol{x}^0) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{x}^t, \eta \mathsf{I}_M)$ *is used in the definition of EIG.*

We leave the proof to our companion paper [19]. Since EIG, SV, and LIME can be derived from or associated with IG, it is legitimate to say that they are in the *integrated gradient (IG) family*. Since IG is deviation-agnostic, we conclude that the deviation-agnostic property is a common characteristic of the IG family.

4.2.3 Increment vs. deviation and local vs. global. Now let us consider the implications of these results in anomaly attribution. The definition of IG in Eq. (3) indicates that IG explains the *increment* of $f(\cdot)$ from the baseline point rather than the deviation, as illustrated in Fig. 1. The baseline is arbitrary. Hence, the increment is not directly relevant to the observed anomaly in general. EIG (and thus SV by Theorem 4) neutralizes this limitation by taking the expectation. However, it results in losing the locality of explanation because it attempts to explain the increment from *any* point in the domain, as suggested in [21] regarding SV. They are unsuitable for anomaly

attribution due to both their deviation-agnostic property and the lack of locality. LIME, on the other hand, maintains the locality by choosing the baseline input in the infinitesimal neighborhood, i.e., $\boldsymbol{x}^0 \to \boldsymbol{x}^t$, but it is still deviation-agnostic.

In general, we need a certain reference point to define anomalousness (cf. the 'reference point' column in Table 1). The above observations motivate us to explore a new idea in choosing a reference point. In this regard, the likelihood-based approach first proposed by the present authors [20] is quite suggestive. Inspired by [20], we propose a novel generative framework for anomaly attribution, where the notion of normalcy is equated to maximum a posteriori (MAP) estimation, as presented in the next section.

5 GENERATIVE PERTURBATION ANALYSIS

We have argued that the existing attribution methods are not suitable for anomaly attribution due to their deviation-agnostic property and/or limited built-in mechanism for evaluating the uncertainty of attribution. This section presents the method of generative perturbation analysis (GPA), a novel probabilistic framework for anomaly attribution that addresses these issues.

5.1 Generative model description

In a typical anomaly detection scenario, samples in the training dataset are assumed to have been collected under normal conditions, and hence, the learned function $y = f(\boldsymbol{x})$ represents normalcy as well. As discussed in Sec. 3, the canonical measure of anomalousness is the negative log likelihood $-\ln p(y \mid \boldsymbol{x})$. A low likelihood value signifies anomaly, and vice versa. From a geometric perspective, on the other hand, being an anomaly implies deviating from a certain normal value. We are interested in integrating these two perspectives.

5.1.1 Perturbation as explanation. Suppose we just observed a test sample $(\boldsymbol{x}^t, y^t)$ being anomalous because of a low likelihood value. Given the regression function $y = f(\boldsymbol{x})$, there are two possible geometric interpretations on the anomalousness (see Figs. 1 (b) and 2 (a)). One is to start with the input $\boldsymbol{x} = \boldsymbol{x}^t$, and observe the deviation $f(\boldsymbol{x}^t) - y^t$. In some sense, $(\boldsymbol{x}, y) = (\boldsymbol{x}^t, f(\boldsymbol{x}^t))$ is a reference point against which the observed sample $(\boldsymbol{x}^t, y^t)$ is judged. The other is to start with the output $y = y^t$, and move horizontally, looking for a perturbation $\boldsymbol{\delta}$ such that $\boldsymbol{x} = \boldsymbol{x}^t + \boldsymbol{\delta}$ gives the maximum possible fit to the normal model. In this case, the reference point is $(\boldsymbol{x}^t + \boldsymbol{\delta}, y^t)$ and $\boldsymbol{\delta}$ is the deviation measured horizontally. Since $\boldsymbol{\delta}$ is supposed to be zero if the sample is perfectly normal, each component $\delta_1, \ldots, \delta_M$ can be viewed as a value indicative of the responsibility of each input variable.

5.1.2 Generative model. Based on the intuition above, we define a novel probabilistic attribution approach through a data-generating process of observed data. The idea is that we write down a generative process for the observable variables $(\boldsymbol{x}, y)$ as a *parametric model of* $\boldsymbol{\delta}$. Then, *the whole task of anomaly attribution is reduced to a parameter estimation problem*, given an observed test point $(\boldsymbol{x}, y) = (\boldsymbol{x}^t, y^t)$. Specifically, the probabilistic regression model $p(y \mid \boldsymbol{x})$ is now viewed as a parametric model $p(y \mid \boldsymbol{x}, \boldsymbol{\delta})$ by setting $\boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{\delta}$. With an extra parameter λ representing the precision of the regression function and also prior distributions for λ and $\boldsymbol{\delta}$, we consider the following generative process:

$$p(y^t \mid \boldsymbol{x}^t, \boldsymbol{\delta}, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2}\right\} \tag{9}$$

$$p(\boldsymbol{\delta}) = \mathcal{N}(\boldsymbol{\delta} \mid \mathbf{0}, \eta^{-1}\mathrm{I}_M) \triangleq (2\pi)^{-\frac{M}{2}} \eta^{-\frac{1}{2}} \exp\left\{-\frac{\eta}{2}\|\boldsymbol{\delta}\|_2^2\right\}, \tag{10}$$

$$p(\lambda) = \mathrm{Gam}(\lambda \mid a_0, b_0) \triangleq \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp(-b_0 \lambda), \tag{11}$$

where $\mathcal{N}(\cdot \mid \cdot, \cdot)$ and $\mathrm{Gam}(\cdot \mid \cdot, \cdot)$ denote the Gaussian and gamma distributions, respectively, and η, a_0, b_0 are hyperparameters. As mentioned above, $\boldsymbol{\delta}$ plays the role of a model parameter here. Notice that Eq. (9) naturally represents the horizontal point-seeking mentioned above. If we point-estimated $\boldsymbol{\delta}$ with Eq. (9) alone, we would have the one that achieves $f(\boldsymbol{x}^t + \boldsymbol{\delta}) \approx y^t$. The challenge here is how to find the *distribution* of $\boldsymbol{\delta}$. The prior distribution $p(\boldsymbol{\delta})$ in Eq. (10) introduces potential variability of $\boldsymbol{\delta}$ to the model. Since $\boldsymbol{\delta} = \mathbf{0}$ represents the normal state, the use of zero-mean Gaussian makes sense. Other zero-mean distributions may work. In fact, we modify this prior a bit later, as discussed in Sec. 5.3.

The precision parameter λ in Eq. (9) describes potential noise that may have contaminated the data, as illustrated as the grey band in Fig. 2. As the preciseness of the measurements may vary from sample to sample, the use of a single λ value can be risky. The prior $p(\lambda)$ takes care of this aspect. As will be seen later, our model uses a mixture of Gaussians with different values of λ in some sense, which leads to the t-distribution instead of Gaussian for the observation model, adding extra capability of handling heavy noise.

Finally, we make two remarks about the proposed generative model. First, Bayesian (linear) regression models similar to the above have been considered in the literature, e.g., [39]. Our model is fundamentally different from them in that (1) $\boldsymbol{\delta}$ as an explanation is not linear regression coefficients, and (2) we do *not* approximate $f(\cdot)$ as a liner function. As for other Bayesian regression approaches, Moreira et al. [27] used the Gaussian process in the active learning setting, but not for attribution. Second, one might wonder whether the particular choice of a parametric form might lead to the loss of generality. Regarding this question, it is critical to understand that Eq. (9) is about the *deviation* or the *error* $f(\boldsymbol{x}^t) - y^t$. Although the variability of y over the entire domain obviously does not follow Gaussian in general, the error is often well-represented by Gaussian or t-distribution. This is exactly the same situation Carl Friedrich Gauss faced when he invented Gaussian-based fitting [4]: Planetary motions do not follow Gaussian, but the error does.

5.2 Inference approach

Given the generative model above, the task of probabilistic attribution is now turned to that of finding the posterior distribution of $\boldsymbol{\delta}$. However, there are two major differences from the standard Bayesian inference: 1) $f(\boldsymbol{x})$ is a black-box function. Exact inference is not possible. Approximating $f(\boldsymbol{x})$ with a specific functional form, such as the linear function, may not always be possible, either. 2) Posterior inference generally yields a joint distribution for $\boldsymbol{\delta}$, denoted by $Q(\boldsymbol{\delta})$. However, this is *not* what we want since it does *not directly explain* the contribution of the *individual* input variables. This section explains how we addressed these challenges.

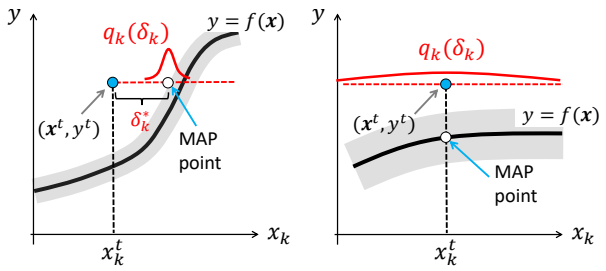

Figure 2: Illustration of the GPA solution. (a) The MAP estimate δ_k^* intuitively represents the deviation from the regression surface at the level of y^t. (b) If x_k is barely correlated with y, GPA tends to give a broad distribution around the MAP point, which is 0 almost surely.

5.2.1 Decomposing variable's contributions. One of the most important ideas of our probabilistic attribution framework is to assume a *factorized* form of posterior:

$$Q(\boldsymbol{\delta}) = Q(\delta_1, \ldots, \delta_M) \approx \prod_{k=1}^{M} q_k(\delta_k), \tag{12}$$

so that end-users can directly use $q_k(\delta_k)$ to get insights on the contribution of the k-th input variable (see Fig. 2). The factorized form (12) is reminiscent of what is assumed in the variational Bayes (VB) algorithm [3], and we can be guided by VB's general solution approach. Specifically, we find the unknown distributions $\{q_k\}$ by minimizing the KL (Kullback–Leibler) divergence between Q and $\prod_k q_k$. The key fact here is that Q is proportional to the complete likelihood by Bayes' rule. Since λ is an unobserved intermediate parameter, it can be marginalized. The integration can be performed analytically, yielding the following form of the likelihood:

$$Q(\boldsymbol{\delta}) \propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\text{test}}} \int_0^\infty \mathrm{d}\lambda \; p(y^t \mid \boldsymbol{x}^t, \boldsymbol{\delta}, \lambda) p(\lambda) \tag{13}$$

$$\propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\text{test}}} \frac{1}{\sqrt{b_0}} \left\{ 1 + \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2b_0} \right\}^{-(a_0 + \frac{1}{2})}, \tag{14}$$

where we assumed the collective attribution scenario for generality but note that N_{test} can be 1. The marginalization amounts to forming a weighted mixture of Gaussians. The resulting distribution (14) is the t-distribution with the degrees of freedom $2a_0$, the mean $f(\boldsymbol{x}^t + \boldsymbol{\delta})$, and the scale parameter $\sqrt{b_0/a_0}$, adding extra robustness to the model. The objective functional for $\{q_k\}$ is given by

$$\int \left(\prod_{k=1}^{M} \mathrm{d}\delta_k \; q_k(\delta_k) \right) \ln \frac{\prod_{l=1}^{M} q_l(\delta_l)}{Q(\boldsymbol{\delta})} + \sum_{k=1}^{M} \gamma_k \int \mathrm{d}\delta_k \; q_k(\delta_k), \tag{15}$$

where the first term is the KL divergence and the second term is to include the normalization condition with γ_k being Lagrange's multiplier. Note that the proportional coefficient in Eq. (14) has no effect here so we do not have to determine it.

By the calculus of variations w.r.t. q_k, it is straightforward to get the minimizer as

$$\ln q_k(\delta_k) = \text{c.} + \int \left(\prod_{j \neq k} \mathrm{d}\delta_j \; q_j(\delta_j) \right) \ln Q(\boldsymbol{\delta}), \tag{16}$$

Algorithm 2 Generative Perturbation Analysis

Require: $f(\boldsymbol{x})$, $\mathcal{D}_{\text{test}}$, parameters $\eta, \nu, \kappa, a_0, \{b(\boldsymbol{x}^t)\}$.

randomly initialize $\boldsymbol{\delta} \approx \boldsymbol{0}$.
repeat
  set $\boldsymbol{g} = \boldsymbol{0}$
  for all $(y^t, \boldsymbol{x}^t) \in \mathcal{D}_{\text{test}}$ **do**
    Compute the local gradient $\frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}}$
    Update $\boldsymbol{g} \leftarrow \boldsymbol{g} + \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{2b(\boldsymbol{x}^t) + [y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}$
  end for
  $\boldsymbol{g} \leftarrow (1 - \kappa\eta)\boldsymbol{\delta} + \kappa(2a_0 + 1)\boldsymbol{g}$
  $\boldsymbol{\delta} = \text{sign}(\boldsymbol{g}) \max\{0, |\boldsymbol{g}| - \eta\nu\}$
until convergence
set $\boldsymbol{\delta}^* = \boldsymbol{\delta}$
for all k **do**
  $q_k(\delta) = Q(\delta_1^*, \ldots, \delta_{k-1}^*, \delta, \delta_{k+1}^*, \delta, \delta_M^*)$
  $q_k(\cdot) \leftarrow q_k(\cdot) / \int \mathrm{d}\delta' q_k(\delta')$ with Eq. (18)
end for
return $\{q_k(\cdot) \mid k = 1, \ldots, M\}$ and $\boldsymbol{\delta}^*$

where c. is a symbol representing an unimportant constant in general. Since both q_k and q_j $(j \neq k)$ are unknown, this procedure is iterative in nature. Also, since q_k's are a functional of the black-box function $f(\cdot)$, analytically performing the integration is not possible. Although Monte Carlo techniques can be used in theory, they are not a preferable choice in realistic usage scenarios, where the end-users actively interact with the attribution tool with different test points.

5.3 Computing attribution score distribution

Here, we propose a practical solution to address these challenges. For attribution purposes, we do not necessarily need the posterior distribution over the entire domain. What we are interested in is how attribution score is distributed around the most probable value. Hence, we evaluate the expectation in Eq. (16) through the empirical distribution of δ_j $(j \neq k)$ with a sample at the maximum posteriori (MAP) point. In this approach, the variable-wise posterior is given simply by

$$q_k(\delta_k) \propto Q(\delta_1^*, \ldots, \delta_{k-1}^*, \delta_k, \delta_{k+1}^*, \ldots, \delta_M^*), \tag{17}$$

where $\boldsymbol{\delta}^*$ is the MAP solution $\boldsymbol{\delta}^* \triangleq \arg\max_{\boldsymbol{\delta}} \ln Q(\boldsymbol{\delta})$. Since this is a one-dimensional (1D) distribution and we know $\boldsymbol{\delta}$ distributes around zero, the normalization constant can be determined easily. Numerical integration is one approach. Otherwise, one may treat $q_k(\cdot)$ as a discrete distribution on a 1D grid. Specifically, we define 1D grid points $\delta^{[1]}, \ldots, \delta^{[N_{\text{g}}]}$ over $[-\delta_{\max}, \delta_{\max}]$, where N_{g} is an arbitrary number of grid points, such as 100, and $\delta_{\max}$ can be, for example, $\delta_{\max} \sim 1.1 \max_k |\delta_k^*|$. The distribution $q_k(\cdot)$ on the grid is obtained from its unnormalized version $\tilde{q}_k(\cdot)$ by

$$q_k(\delta^{[i]}) \approx \frac{1}{\sum_{i=1}^{N_{\text{g}}} \tilde{q}_k(\delta^{[i]})} \tilde{q}_k(\delta^{[i]}), \quad \text{for } i = 1, \ldots, N_{\text{g}}. \tag{18}$$

The inference procedure has now become a two-step process: MAP estimation and construction of $\{q_k\}$ with Eqs. (17)-(18). The

former problem is written as

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \{J(\boldsymbol{\delta}) + \eta\nu\|\boldsymbol{\delta}\|_1\}, \tag{19}$$

$$J(\boldsymbol{\delta}) \triangleq \frac{\eta}{2}\|\boldsymbol{\delta}\|_2^2 + \ln\left\{1 + \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2b(\boldsymbol{x}^t)}\right\}^{\frac{2a_0+1}{2}}, \tag{20}$$

where we have added an extra ℓ_1 term for better interpretability through sparsity. Here, ν is the strength of the ℓ_1 regularization relative to that of ℓ_2. With this modification, we need to use

$$p(\boldsymbol{\delta}) \propto \exp\left\{-\frac{\eta}{2}\|\boldsymbol{\delta}\|_2^2 - \eta\nu\|\boldsymbol{\delta}\|_1\right\} \tag{21}$$

in Eqs. (14) and (17). We have also included in Eq. (20) potential dependency of b_0 on $\boldsymbol{x}^t$ and denoted it as $b(\boldsymbol{x}^t)$. Corresponding to $a(\mathcal{D}_{\text{test}})$ in Eq. (2), if we wish to find the attribution distribution for a collection of test samples, $J(\boldsymbol{\delta})$ should be replaced with

$$J(\boldsymbol{\delta}) = \frac{\eta}{2}\|\boldsymbol{\delta}\|_2^2 + \sum_{t=1}^{N_{\text{test}}} \ln\left\{1 + \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2b(\boldsymbol{x}^t)}\right\}^{\frac{2a_0+1}{2}}. \tag{22}$$

We call the proposed probabilistic attribution framework the **generative perturbation analysis** (GPA) hereafter. As illustrated in Fig. 2, the GPA distribution $\{q_k\}$ is useful to have in order to evaluate the general informativeness of the input variables.

5.3.1 Solving MAP problem. One of the standard solution approaches to the optimization problem of the type Eq. (19) is proximal gradient descent [31], although the unavailability of closed-form expression of the gradient of $f(\cdot)$ makes the procedure a bit complicated. If a numerical estimation method for $\nabla f(\boldsymbol{x}^t + \boldsymbol{\delta})$ is available, Eq. (19) can be reduced to an iterative lasso regression problem:

$$\boldsymbol{\delta} \leftarrow \arg\min_{\boldsymbol{\delta}} \left\{\frac{1}{2\kappa}\|\boldsymbol{\delta} - \boldsymbol{\delta}' + \kappa\nabla J(\boldsymbol{\delta}')\|_2^2 + \eta\nu\|\boldsymbol{\delta}\|_1\right\}, \tag{23}$$

where κ is a constant corresponding to the learning rate and $\boldsymbol{\delta}'$ is the solution of the previous iteration round. By setting the subgradient zero, the solution of this problem is readily obtained as

$$\delta_i = \text{sign}(g_i)\max\{0, |g_i| - \eta\nu\}, \tag{24}$$

where we defined $\boldsymbol{g} \triangleq \boldsymbol{\delta}' - \kappa\nabla J(\boldsymbol{\delta}')$. Upon convergence, we set $\boldsymbol{\delta}^* = \boldsymbol{\delta}$. See lines 2-11 in Algorithm 2.

5.3.2 Algorithm summary. Algorithm 2 summarizes the entire algorithm of GPA. Whenever possible, it is recommended to standardize $\boldsymbol{x}$ somehow so that it distributes around zero with unit variance for each variable. For standardized data, the ℓ_2 strength η can be a value of O(1), such as 0.1, which can also be a reasonable starting point for κ. The ℓ_1 strength should be in the range $0 < \nu \leq 1$. We fixed $\nu = 0.5$ in our experiment. As $2a_0$ has the interpretation of degrees of freedom, one reasonable starting point is $2a_0 \sim N_{\text{test}} + 1$. As described in Appendix D, $b(\boldsymbol{x}^t)$ can be chosen as a constant $b_0 \sim a_0\sigma_{yf}^2/c_b$, where σ_{yf}^2 is an estimate of the variance of $y - f(\boldsymbol{x})$, or the maximizer of the marginalized likelihood, and c_b is the number of virtual samples, which can be O(10). As summarized in Table 2, we used $c_b = 1$ or 10, and also $2a_0 = 11$ in our experiments to simulate the variability of realistic cases.

It is easy to see that the complexity of the algorithm is $O(MN_{\text{test}})$ per iteration round. Note that $N_{\text{test}} = 1$ is the most common choice (i.e., local explanation) and the algorithm does not use any training data. Hence, typical scalability analysis about the data set size is irrelevant. The total computational time depends almost entirely on very low-level implementation details, such as how efficient the numerical gradient estimation routine is, how the black-box model $f(\cdot)$ is implemented, and to what extent the Python code is vectorized. Their detailed analysis is beyond the scope of the paper.

Table 2: Summary of the datasets and parameters used.

	N_{train}	N_{test}	M	$f(\boldsymbol{x})$	κ	c_b	η
`2Dsinusoidal`	∞	1	2	analytic	-	-	-
`Diabetes`	442	1	10	DNN	0.08	10	0.4
`Boston`	506	1	13	RF	0.08	10	0.1
`California`	20 640	3	8	GBT	$0.1/N_{\text{test}}$	1	$0.5N_{\text{test}}$

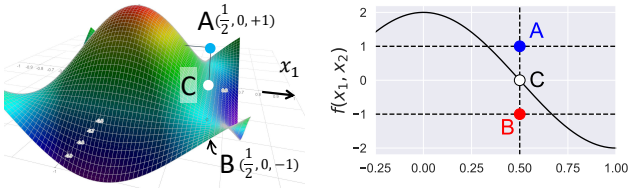

Figure 3: `2Dsinusoidal`: Surface plot and the $x_2 = 0$ slice. The points A, B, and C are at $y^t = 1, -1$, and 0, respectively, while they are at the same $\boldsymbol{x}^t = (1/2, 0)$.

6 EXPERIMENTS

This section presents empirical evaluation of the proposed anomaly attribution framework[1]. The goals of this evaluation are to 1) provide a clear picture of what deviation-sensitivity of an attribution method buys us; 2) demonstrate GPA's unique capability of providing the probability distribution of attribution scores; 3) quantitatively analyze the consistency and inconsistency among different attribution methods.

6.1 Datasets and baselines

Based on the datasets summarized in Table 2, we compared GPA with seven baselines: Six non-distributional attribution methods, LIME (Sec. 4.1.1), (E)IG (Sec. 4.1.2), SV (Sec. 4.1.3), LC [20], and the Z-score (e.g., [5]), as well as one distributional method, BayLIME [49]. For anomaly attribution, LIME, SV, IG, and EIG are applied to the deviation $f(\boldsymbol{x}) - y$ rather than $f(\boldsymbol{x})$. The Z-score is a standard univariate outlier detection metric in the unsupervised setting, and is defined as $Z_i \triangleq (x_i^t - m_i)/\sigma_i$ for the i-th variable, where m_i, σ_i are the mean and the standard deviation of x_i, respectively. In SV, we used the same sampling scheme as that proposed in [42] with the number of configurations limited to 100. In IG and EIG, we used the trapezoidal rule with 100 equally-spaced intervals to perform the integration w.r.t. α. For IG, EIG, LC, and GPA, we used the same gradient estimation algorithm described in Appendix C.

To compute SV, EIG, and the Z-score, we used the empirical distribution of the training data to approximate $P(\boldsymbol{x})$. Note that this is *actually not possible to do* in our doubly black-box setting. We are including SV, EIG, and the Z-score here for comparison purposes nonetheless.

[1]Python implementation is available at https://github.com/Idesan/gpa.

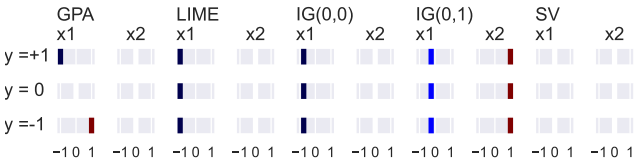

Figure 4: `2Dsinusoidal`: Comparison of normalized attribution scores at three test points (A, B and C in Fig. 3).

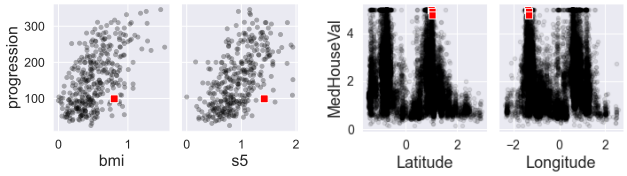

Figure 5: Scatter plot of selected input variables vs. y. Left: `Diabetes`. Right: `CaliforniaHousing`. The red squares highlight the detected top outliers.

6.2 Deviation-sensitivity

6.2.1 2Dsinusoidal. The first empirical evaluation uses a synthetic dataset named `2Dsinusoidal`, which is a newly proposed attribution benchmark model, defined by a 2-variate sinusoidal function

$$f(\boldsymbol{x}) = 2\cos(\pi x_1)\cos(\pi x_2). \tag{25}$$

One remarkable feature of this model is that it is possible to *calculate closed-form attribution scores*. See Appendix A for the details.

Suppose we have a test point at $\boldsymbol{x}^t = (1/2, 0)$ with three different y^t values, A ($y^t = 1$), B ($y^t = -1$), and C ($y^t = 0$), as illustrated in Fig. 3. In this case, SV and LIME attribution scores are $(0, 0)$ and $(-2\pi, 0)$, respectively. IG's scores are $(-2, 0)$ and $(-2/3, 8/3)$ if we choose $\boldsymbol{x}^0 = (0, 0)$ and $(0, 1)$, respectively. These do not depend on y^t due to the deviation-agnostic property, in contrast to GPA, which gives $\delta_1^* = (1/\pi)\arccos\left(y^t/2\right) - x_1^t$ and $\delta_2^* = 0$.

Figure 4 visualizes the attribution scores with what we call the '*litmus plot*,' where larger values get darker colors (0 gets white) and negative/positive values get blue/red colors. Due to space limitations, we omitted LC, which results in the same solution as that of GPA, and the Z-score. For GPA, the scores are normalized by dividing by $\max_k |\delta_k|$ for each test point. Similar normalization was done for the baselines with the convention $\frac{0}{0} = 0$.

In this example, A and B are outliers due to a shift in the x_1 direction, while C is normal in terms of deviation. Hence, an ideal attribution would be that x_1 *alone* gets a strong signal *only* for A and B. GPA precisely reproduces this, but all the baselines do not: They gave the same score for A, B, and C, as a consequence of the deviation-agnostic property. The figure also shows that IG's scores sensitively depend on the choice of $\boldsymbol{x}^0$, making IG trickier to use for the end-users. SV always satisfies $\mathrm{SV}_1 = \mathrm{SV}_2$ regardless of $\boldsymbol{x}^t$ in this case, and does not provide any clue for input attribution. This is a manifestation of the loss of locality in SV discussed in Sec. 4.2.3.

6.2.2 Diabetes. To test the deviation sensitivity of GPA on real-world data, we used `Diabetes` [7], which has a real-valued target variable ('progression') and $M = 10$ predictors including the body-mass index ('bmi'). For this dataset, we held out 20% of the samples and trained a deep neural network (DNN) on the rest as the black-box model $f(\cdot)$. We identified the top outlier using Eq. (1), which is highlighted in Fig 5.

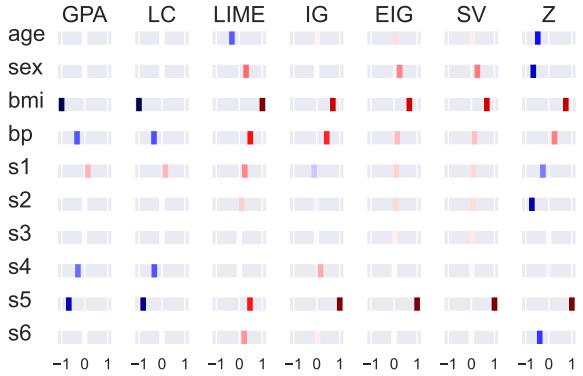

Figure 6: `Diabetes`: Comparison of normalized attribution scores in the litmus plot for the top outlier detected.

Figure 6 compares attribution scores for the top outlier. We set $\boldsymbol{x}^0 = \boldsymbol{0}$ for IG. All the attribution methods identify 'bmi' and 's5' as the top contributors. For both, GPA and LC get a large negative score since a smaller value is more typical for such a low y^t value, as shown in the scatter plot in Fig. 5. GPA gave $\delta_{\mathrm{bmi}} = -0.81$ and $\delta_{\mathrm{s5}} = -0.55$. Note that these values have actual meaning rather than just the magnitude of responsibility: A big negative in δ_{bmi} means that the BMI is too high for such a low y level. In other words, they would have looked normal if they were a little skinnier. Explainability of this kind is particularly useful in practice as the score provides *actionable insights* about how the status quo could be changed for the better. The alternative methods do not have such ability. LIME is positive for bmi because the slope is positive at the $\boldsymbol{x}^t$, regardless of the y^t value. Similarly, IG, EIG, and SV gave positive values for bmi because $f(\boldsymbol{x}^t)$ is higher than the mean of y, regardless of the specific value of y^t.

6.3 Distribution analysis

6.3.1 Comparison with BayLIME. Now let us discuss how GPA provides useful insights into the uncertainty of the attribution score. To the best of our knowledge, BayLIME [39, 49] is the only method in the literature applicable to our setting.

We used the `BostonHousing` data [2], where the task is to predict y, the median home price ('MEDV') of the districts in Boston, with $\boldsymbol{x}$, the input vector of size $M = 13$ including such features as the percentage of the lower status of the population ('LSTAT') and the average number of rooms ('RM')[2]. For this dataset, we held out 20% of the samples as $\mathcal{D}_{\mathrm{test}}$ and trained the random forest (RF) [17] on the rest as the black-box model $f(\cdot)$. We identified the top outlier using Eq. (1) and computed the score distribution for the top outlier.

The estimated distributions are shown in Fig. 7. As clearly seen from the figure, BayLIME gives the same curve for all the variables apart from the mean locations. In fact, the variance is given as a

[2]We excluded a variable named 'B' from attribution for ethical concerns [34].

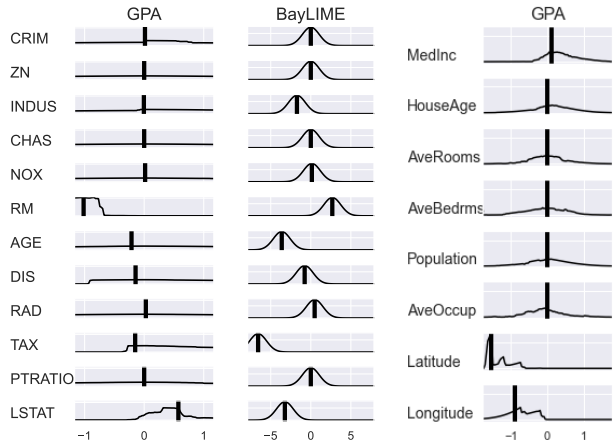

Figure 7: Estimated score distribution. Left: `BostonHousing`**. Right:** `CaliforniaHousing` **for collective attribution.**

constant $1/(\eta + \lambda N_s)$, where N_s is the number of virtual samples generated for estimating the regression coefficients and is 10 in our case (See Appendix B). In contrast, GPA provides variable-specific distributions. As illustrated in Fig. 2, less informative variables tend to produce a flatter distribution in GPA. In this case, we immediately see that RM and LSTAT are two dominating variables. Such an insight is not obtainable from BayLIME. It is interesting to see that the score distributions given by GPA tend to be piece-wise constant, reflecting the fact that RF is a collection of decision stumps.

6.3.2 Collective attribution. To show the unique capability of GPA for collective attribution, we used the `CaliforniaHousing` [30] dataset. The task is to predict the median house value of small geographical segments using predictor variables such the longitude and latitude. We held out 20% of the samples and trained gradient boosted trees (GBT) [10] on the rest. In this case, we identified *three* top outliers as shown in Fig. 5. The question is whether those outliers have common characteristics in their outlier-ness.

Figure 7 shows the computed distributions, where we omitted BayLIME due to its triviality. Very interestingly, 'Latitude' has a very sharp peak at a negative value. This indicates that the very high 'MedHouseVal' in Fig. 5 stands out in that latitude and they would look more common if they existed in a southern location.

6.4 Consistency analysis

We have compared GPA with seven alternative methods in a rather qualitative fashion so far. One important question in practice is how those methods are consistent or inconsistent among them overall. To answer this question, we identified five top outliers in the three real-world datasets (`BostonHousing`, `CaliforniaHousing`, `Diabetes`), and computed how their attribution scores are consistent with those of GPA in terms of four metrics: Kendall's τ, Spearman's ρ, the sign match ratio (SMR), and the hit ratio at 25% (h25). See Appendix E for the detail.

The result is summarized in Table 3. We omitted EIG because of Theorem 4. LC achieves very high consistency with GPA, although it lacks a built-in mechanism for UQ. This is understandable since it can be viewed as a point-estimation version of GPA in some sense.

Table 3: Result of consistency analysis. The mean and the standard deviation are shown in each cell, where 1 represents the highest consistency with GPA's MAP value.

		LC	LIME	IG	SV	Z-score
Bos.	τ	0.70 ± 0.14	0.25 ± 0.30	0.61 ± 0.11	0.43 ± 0.10	0.17 ± 0.30
	ρ	0.83 ± 0.09	0.32 ± 0.38	0.74 ± 0.08	0.57 ± 0.14	0.24 ± 0.35
	SMR	0.92 ± 0.11	0.71 ± 0.11	0.65 ± 0.12	0.69 ± 0.14	0.62 ± 0.17
	h25	0.80 ± 0.18	0.27 ± 0.28	0.73 ± 0.28	0.67 ± 0.00	0.20 ± 0.30
Cal.	τ	0.82 ± 0.15	0.67 ± 0.13	0.64 ± 0.07	0.73 ± 0.10	0.04 ± 0.20
	ρ	0.91 ± 0.11	0.76 ± 0.11	0.79 ± 0.06	0.83 ± 0.10	0.07 ± 0.27
	SMR	0.97 ± 0.06	0.95 ± 0.11	0.68 ± 0.07	0.68 ± 0.11	0.70 ± 0.14
	h25	0.80 ± 0.27	0.90 ± 0.22	1.00 ± 0.00	1.00 ± 0.00	0.30 ± 0.27
Dia.	τ	0.94 ± 0.06	0.31 ± 0.19	0.72 ± 0.08	0.58 ± 0.10	0.15 ± 0.15
	ρ	0.98 ± 0.03	0.41 ± 0.21	0.88 ± 0.04	0.75 ± 0.10	0.22 ± 0.20
	SMR	1.00 ± 0.00	0.62 ± 0.11	0.38 ± 0.08	0.62 ± 0.22	0.60 ± 0.10
	h25	1.00 ± 0.00	0.60 ± 0.22	0.90 ± 0.22	0.80 ± 0.27	0.30 ± 0.45

As expected, h25 generally has high scores, apart from the Z-score. This suggests that those attribution methods are a useful tool for selecting important features. Even in the other metrics, including the SMR, they produce reasonably consistent attributions in some cases. However, in some 20-30% of cases they are not necessarily consistent, which is a natural consequence of the fact that GPA is deviation-sensitive but the others are not.

6.5 Practical utility of the GPA framework

We remark further on the practical utility of the proposed framework, using the `BostonHousing` data as an example. Recall that the top detected outlier has two variables with dominating attribution scores. Depending on one's role, different insights may be obtained from this analysis: From the **end user**'s perspective, the outlier in Fig. 7 may point to a bargain since this house (district) has unusually more rooms and much fewer low-income neighbors than expected for the price range; For a **modeler** who is interested in debugging the model, the two dominating attribution scores may hint that the model may be failing to capture the relationship between the housing price and the variables RM and LSTAT, prompting the modeler to revise (e.g. contextualize) how these variables are defined. While the attribution scores may not decisively pinpoint the exact interpretation, the rich and accurate information given by GPA provides valuable clues in either usage scenario.

7 CONCLUSIONS

We have proposed GPA, a novel generative approach to probabilistic attribution of black-box regression models. The key idea is to reduce the attribution task to a statistical parameter estimation problem. This can be done by viewing the perturbation $\boldsymbol{\delta}$ as a model parameter of the generative process for the observed variables $(\boldsymbol{x}, y)$, where the posterior distribution gives the distribution of the attribution score. We proposed a variational inference algorithm to obtain variable-wise distributions.

We have also shown that the existing input attribution methods, namely integrated gradient IG), local linear surrogate modeling (LIME), and Shapley values (SV), are inherently deviation-agnostic and, thus, are not designed to be a viable solution for *anomaly attribution*. Unlike these methods, GPA is capable of providing directly interpretable insights in a deviation-sensitive and uncertainty-aware manner.

REFERENCES

[1] Liat Antwarg, Ronnie Mindlin Miller, Bracha Shapira, and Lior Rokach. 2021. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Systems with Applications* 186 (2021), 115736.

[2] David A. Belsley, Edwin Kuh, and Roy E. Welsch. 2005. *Regression diagnostics: Identifying influential data and sources of collinearity*. Vol. 571. John Wiley & Sons.

[3] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag.

[4] Richard G. Brereton. 2014. The normal distribution. *Journal of Chemometrics* 28, 11 (2014), 789–792.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Computing Survey* 41, 3 (2009), 1–58.

[6] Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, and Xia Hu. 2021. A Unified Taylor Framework for Revisiting Attribution Methods. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 21)*. 11462–11469.

[7] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least angle regression. *The Annals of statistics* 32, 2 (2004), 407–499.

[8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 19)*. 2950–2958.

[9] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 17)*. 3429–3437.

[10] Jerome H. Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38, 4 (2002), 367–378.

[11] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD 21)*. 577–590.

[12] Damien Garreau and Ulrike von Luxburg. 2020. Explaining the Explainer: A First Theoretical Analysis of LIME. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 20) (Proceedings of Machine Learning Research, Vol. 108)*. PMLR, 1287–1296.

[13] Ioana Giurgiu and Anika Schumann. 2019. Additive Explanations for Anomalies Detected from Multivariate Temporal Data. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 19)*. ACM, 2245–2248.

[14] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (2022), 1–55.

[15] David Gunning and David W. Aha. 2019. DARPA's explainable artificial intelligence program. *AI Magazine* 40, 2 (2019), 44.

[16] Wenbo Guo, Sui Huang, Yunzhe Tao, Xinyu Xing, and Lin Lin. 2018. Explaining Deep Learning Models–A Bayesian Non-parametric Approach. In *Advances in Neural Information Processing Systems (NIPS 18)*. 4514–4524.

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Springer.

[18] Chanwoong Hwang and Taejin Lee. 2021. E-SFD: Explainable sensor fault detection in the ics anomaly detection system. *IEEE Access* 9 (2021), 140470–140486.

[19] Tsuyoshi Idé and Naoki Abe. 2023. Black-Box Anomaly Attribution. *arXiv preprint arXiv:2305.18440* (2023).

[20] Tsuyoshi Idé, Amit Dhurandhar, JiříNavrátil, Moninder Singh, and Naoki Abe. 2021. Anomaly Attribution with Likelihood Compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 21)*, Vol. 35. 4131–4138.

[21] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning (ICML 20)*. PMLR, 5491–5500.

[22] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? — A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[23] Wenke Lee and Dong Xiang. 2000. Information-theoretic measures for anomaly detection. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy (SP 01)*. 130–143.

[24] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS 17)*. 4765–4774.

[25] Dennis A.-L. Mariadass, Ervin Gubin Moung, Maisarah Mohd Sufian, and Ali Farzamnia. 2022. Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE 22)*. IEEE, 219–224.

[26] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-Wise Relevance Propagation: An Overview*. Springer, 193–209.

[27] Catarina Moreira, Yu-Liang Chou, Mythreyi Velmurugan, Chun Ouyang, Renuka Sindhgatta, and Peter Bruza. 2021. LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems* 150 (2021), 113561.

[28] Keith Noto, Carly Brodley, and Donna Slonim. 2010. Anomaly detection using an ensemble of feature models. In *2010 IEEE International Conference on Data Mining (ICDM 10)*. IEEE, 953–958.

[29] Darian M. Onchis. 2020. Should I trust a deep learning condition monitoring prediction?. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 20)*. IEEE, 182–186.

[30] R. Kelley Pace and Ronald Barry. 1997. Sparse spatial autoregressions. *Statistics & Probability Letters* 33, 3 (1997), 291–297.

[31] Neal Parikh and Stephen Boyd. 2014. Proximal algorithms. *Foundations and Trends in Optimization* 1, 3 (2014), 127–239.

[32] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC 18)*.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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 (KDD 16)*. ACM, 1135–1144.

[34] scikit-learn 1.1.3. 2022. sklearn.datasets.load_boston. https://scikit-learn.org/1.1/modules/generated/sklearn.datasets.load_boston.html.

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 17)*. 618–626.

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[37] John Sipple. 2020. Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure. In *Proceedings of the 37th International Conference on Machine Learning (ICML 20)*. 9016–9025.

[38] John Sipple and Abdou Youssef. 2022. A general-purpose method for applying Explainable AI for Anomaly Detection. In *Proceeding of the International Symposium on Methodologies for Intelligent Systems (ISMIS 22)*. Springer, 162–174.

[39] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems (NeurIPS 21)* 34 (2021), 9391–9404.

[40] Stuart Staniford, James A. Hoagland, and Joseph M. McAlerney. 2002. Practical automated detection of stealthy portscans. *Journal of Computer Security* 10, 1-2 (2002), 105–136.

[41] Erik Štrumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research* 11, Jan (2010), 1–18.

[42] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.

[43] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. In *Proceedings of the 38th International Conference on Machine Learning (ICML 20)*. PMLR, 9269–9278.

[44] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 17)*. 3319–3328.

[45] Kenji Yamanishi, Jun ichi Takeuchi, Graham Williams, and Peter Milne. 2000. On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 00)*. 320–324.

[46] Kenji Yamanishi, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. 2004. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery* 8, 3 (2004), 275–300.

[47] Xiao Zhang, Manish Marwah, I-ta Lee, Martin Arlitt, Dan Goldwasser, et al. 2019. ACE—An Anomaly Contribution Explainer for Cyber-Security Applications. In *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data 19)*. IEEE, 1991–2000.

[48] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv:1904.12991* (2019).

[49] Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. 2021. BayLIME: Bayesian local interpretable model-agnostic explanations. In *Proceeding of the 37th Conference on Uncertainty in Artificial Intelligence (UAI 21)*. PMLR, 887–896.

[50] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2022. Do feature attribution methods correctly attribute features?. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 22)*. 9623–9633.

APPENDIX

A CLOSED-FORM SOLUTIONS FOR 2-VARIATE SINUSOIDAL MODEL

This section lists analytic expressions of a few attribution methods on the `2Dsinusoidal` model $f(x_1, x_2) = 2\cos(\pi x_1)\cos(\pi x_2)$.

A.1 LIME

Since LIME score is an estimator of the gradient w.r.t. the input variables x_1, x_2 in the limit of $\nu \to 0_+$, we have

$$\mathrm{LIME}^0(\boldsymbol{x}^t, y^t) = \begin{pmatrix} -2\pi \sin(\pi x_1)\cos(\pi x_2) \\ -2\pi \cos(\pi x_1)\sin(\pi x_2) \end{pmatrix} \tag{A.1}$$

for $\forall(\boldsymbol{x}^t, y^t)$, which obviously does not depend on y^t. If we choose $\boldsymbol{x}^t = (1/2, 0)^\top$, then $\mathrm{LIME}^0 = (-2\pi, 0)^\top$.

A.2 GPA

In GPA, the J function is given by

$$J(\boldsymbol{\delta}) \triangleq \frac{\eta}{2}\|\boldsymbol{\delta}\|_2^2 + \ln\left\{1 + \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2b(\boldsymbol{x}^t)}\right\}^{\frac{2a_0+1}{2}}$$

With $\Delta^t_{\boldsymbol{\delta}} \triangleq y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})$, the gradient is computed as

$$\frac{\partial J}{\partial \boldsymbol{\delta}} = \eta\boldsymbol{\delta} - \frac{(2a_0+1)\Delta^t_{\boldsymbol{\delta}}}{2b(\boldsymbol{x}^t) + (\Delta^t_{\boldsymbol{\delta}})^2}\frac{\partial f(\boldsymbol{x}+\boldsymbol{\delta})}{\partial \boldsymbol{\delta}},$$

where

$$\frac{\partial f(\boldsymbol{x}+\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \begin{pmatrix} -2\pi \sin(\pi(x_1^t + \delta_1))\cos(\pi(x_2^t + \delta_2)) \\ -2\pi \cos(\pi(x_1^t + \delta_1))\sin(\pi(x_2^t + \delta_2)) \end{pmatrix}.$$

Let us assume $\nu \to 0_+$ and $-2 < y^t < 2$. If we assume $x_2^t = 0$, then $\delta_2^* = 0$ should hold as long as $\boldsymbol{\delta}$ is initialized as $\boldsymbol{\delta} \approx \mathbf{0}$ and $\eta\nu > 0$ regardless of the sign of $\Delta^t_{\boldsymbol{\delta}}$. Given this partial solution, the condition of optimality for δ_1 is given by

$$\eta\delta_1 + (2a_0+1)\frac{2\pi\Delta^t_{\boldsymbol{\delta}}\sin(\pi(x_1^t+\delta_1))}{2b(\boldsymbol{x}^t) + (\Delta^t_{\boldsymbol{\delta}})^2} = 0.$$

If $x_1^t > 0$ and $\eta \to 0_+$, this equation yields a condition $y^t - 2\cos(\pi(x_1+\delta_1)) \approx 0$, leading to the solution

$$\delta_1^* = \frac{1}{\pi}\arccos\frac{y^t}{2} - x_1^t. \tag{A.2}$$

If we further choose $x_1^t = 1/2$ (i.e., $\boldsymbol{x}^t = (1/2, 0)$ again), we have $\boldsymbol{\delta}^* = (-\frac{1}{6}, 0)^\top, (0,0)^\top, (\frac{1}{6}, 0)^\top$ for $y^t = 1, 0, -1$, respectively.

A.3 LC

In LC, the J function in our notation is given by

$$J(\boldsymbol{\delta}) = \frac{1}{2}\eta\|\boldsymbol{\delta}\|_2^2 + \frac{1}{2}\lambda[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2. \tag{A.3}$$

With $\Delta^t_{\boldsymbol{\delta}} \triangleq y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})$, the gradient is computed as

$$\frac{\partial J}{\partial \boldsymbol{\delta}} = \eta\boldsymbol{\delta} - \lambda\Delta^t_{\boldsymbol{\delta}}\begin{pmatrix} -2\pi \sin(\pi(x_1^t + \delta_1))\cos(\pi(x_2^t + \delta_2)) \\ -2\pi \cos(\pi(x_1^t + \delta_1))\sin(\pi(x_2^t + \delta_2)) \end{pmatrix}.$$

Let us assume $\nu \to 0_+$ and $-2 < y^t < 2$. If we assume $x_2^t = 0$, again, $\delta_2^* = 0$ should hold as long as $\boldsymbol{\delta}$ is initialized as $\boldsymbol{\delta} \approx \mathbf{0}$ and $\eta\nu > 0$ regardless of the sign of $\Delta^t_{\boldsymbol{\delta}}$. Given this partial solution, the condition of optimality for δ_1 is written as

$$\eta\delta_1 + \lambda[y^t - 2\cos(\pi(x_1+\delta_1))]2\pi\sin(\pi(x_1^t+\delta_1)) = 0. \tag{A.4}$$

If $x_1^t > 0$ and $\eta \to 0_+$, we have a condition $y^t - 2\cos(\pi(x_1+\delta_1)) \approx 0$, which is the same as the MAP equation of GPA. Hence, LC gets the same attribution score as GPA's MAP value in this particular case.

A.4 Integrated Gradient

The `2Dsinusoidal` model allows calculating IG analytically for any $(\boldsymbol{x}^t, \boldsymbol{x}^0)$ based on the definition (3) as

$$\mathrm{IG}_i(\boldsymbol{x}^t, y^t \mid \boldsymbol{x}^0, y^0) = d_i\left[G^t - G^0 - (-1)^i(H^t - H^0)\right] \tag{A.5}$$

with i being 1 or 2 and

$$G^k \triangleq \frac{\cos\pi(x_1^k + x_2^k)}{d_1 + d_2}, \quad H^k \triangleq \frac{\cos\pi(x_1^k - x_2^k)}{d_1 - d_2}, \tag{A.6}$$

where $d_1 \triangleq x_1^t - x_1^0$, $d_2 \triangleq x_2^t - x_2^0$ and k is either t or 0. Using elementary trigonometric formulas, one can verify the sum rule $\mathrm{IG}_1 + \mathrm{IG}_2 = f(\boldsymbol{x}^t) - f(\boldsymbol{x}^0)$. For $\boldsymbol{x}^t = (1/2, 0)^\top$, the IG values are

$$\mathrm{IG}(\boldsymbol{x}^t \mid (0,0)^\top) = (-2, 0)^\top, \quad \mathrm{IG}(\boldsymbol{x}^t \mid (0,1)^\top) = (-\tfrac{2}{3}, \tfrac{8}{3})^\top,$$

where we have omitted redundant y^t, y^0 from the arguments.

A.5 Shapley Value

The expected Shapley value depends on the true distribution $P(\boldsymbol{x})$. If $P(\boldsymbol{x})$ is the uniform distribution over $[-m, m]$ with m being an integer, the expectation of f is zero in `2Dsinusoidal`. The same applies to the conditional distributions. As a result, we have

$$\mathrm{SV}(\boldsymbol{x}^t) = \frac{1}{2}(f(\boldsymbol{x}^t), f(\boldsymbol{x}^t))^\top \tag{A.7}$$

for $\forall\boldsymbol{x}^t$ under the assumed uniform distribution.

B ATTRIBUTION SCORE DISTRIBUTION WITH BAYESIAN LIME

Equation (7) of BayLIME's paper [49] provides the posterior distribution of the regression coefficients. In our notation, the posterior covariance is given by

$$\Sigma^{-1} = \eta\mathsf{I}_M + \lambda\sum_{n=1}^{N_s}\boldsymbol{\xi}_n\boldsymbol{\xi}_n^\top, \tag{B.8}$$

where $\boldsymbol{\xi}_n$ is the n-th sample generated from $\mathcal{N}(\boldsymbol{\xi} \mid \mathbf{0}, \mathsf{I}_M)$, according to the authors. The paragraph after their Eq. (11) says that $\sum_{n=1}^{N_s}\boldsymbol{\xi}_n\boldsymbol{\xi}_n^\top \approx N_s\mathsf{I}_M$ holds. Hence, Σ can be computed as

$$\Sigma = \{\eta\mathsf{I}_M + \lambda N_s\mathsf{I}_M\}^{-1} = (\eta + \lambda N_s)^{-1}\mathsf{I}_M \tag{B.9}$$

and the posterior distribution of the attribution score $\boldsymbol{\beta}$ is given by $Q^{\mathrm{BayLIME}} \triangleq \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{\mathrm{BayLIME}}, \Sigma)$, where $\boldsymbol{\beta}^{\mathrm{BayLIME}}$ is the posterior mean. Since Σ is diagonal, β_is are statistically independent. The distribution of the attribution score of the k-th variable is given by

$$q_k^{\mathrm{BayLIME}}(\beta_k) = \mathcal{N}(\beta_k \mid \beta_k^{\mathrm{BayLIME}}, (\eta + \lambda N_s)^{-1}). \tag{B.10}$$

This is a one-dimensional distribution with the same variance for all the ks. Since the model evaluates the *variability of the generated samples based on an assumed distribution*, the variance does not have any explicit dependency on the black-box function $f(\cdot)$.

C ESTIMATING THE GRADIENT OF BLACK-BOX FUNCTION

To find the MAP solution for GPA, we need to numerically estimate the gradient of the black-box function $f(\cdot)$. To handle the potential non-differentiability of f, we define the gradient as the local mean of the slope function $[f(\boldsymbol{x}_\delta + h\boldsymbol{e}_i) - f(\boldsymbol{x}_\delta)]/h$, where $\boldsymbol{x}_\delta \triangleq \boldsymbol{x}^t + \boldsymbol{\delta}$, h is a small random perturbation, and $\boldsymbol{e}_i$ is a unit vector which takes 1 in the i-th entry and 0 otherwise. The local mean can be estimated by numerically evaluating

$$\frac{\partial f(\boldsymbol{x}_\delta)}{\partial \delta_i} = \int_{-\infty}^{\infty} \mathrm{d}h\, p(h) \frac{f(\boldsymbol{x}_\delta + h\boldsymbol{e}_i) - f(\boldsymbol{x}_\delta)}{h}, \tag{C.11}$$

where $p(h)$ is a local distribution for h around $\boldsymbol{x}_\delta$. One reasonable choice is $p(h) = \mathcal{N}(h \mid 0, \eta_1^2)$ with η_1 being the standard deviation of the perturbations. For numerical stability, we used $\eta_1 = 1$ in our experiments, where the input variables have been standardized. The number of Monte Carlo samples was set to 10, which was confirmed to provide sufficient convergence in our experiments.

D PARAMETER TUNING APPROACH

GPA's distribution can be used for verifying whether the computed MAP value has reached a satisfactory local maximum. In our experiments, we started with a default set of parameters: $\kappa = 0.1/N_{\text{test}}$, $\eta = 0.1N_{\text{test}}$, and $c_b = 10$. If any of the GPA distributions looked inconsistent with the MAP value, we gradually decreased c_b down to 1 and increased η up to 1. We kept ν fixed at 0.5, which turned out to achieve a sparsity level comparable to that of LIME.

We discuss how to initialize a_0, b_0 in the gamma prior below.

D.1 Gamma hyper-parameters: shape

Since $2a_0$ has the interpretation of the degree of freedom of the t-distribution, it makes sense to use

$$a_0 = (\tilde{N} + 1)/2. \tag{D.12}$$

Here, $\tilde{N}$ denotes the sample size and can be equated to N_{test}. We have added 1 so $a_0 = 1$ when $N_{\text{test}} = 1$. Otherwise, it can be interpreted as the virtual sample size, which can be a measure of the confidence level. Since UQ generally boils down to an ill-posed task that estimates uncertainty somehow when many samples are not available, $\tilde{N}$ can be viewed as a controllable parameter to simulate what would be seen when there were abundant samples. In such a case, $\tilde{N}$ could be a value like $1 \le \tilde{N} \lesssim 10$.

D.2 Gamma hyper-parameters: rate

Given a_0, the other parameter b_0 can be estimated by maximizing the log likelihood. For $\Delta_n \triangleq y^{(n)} - f(\boldsymbol{x}^{(n)})$, we solve

$$\begin{aligned}
&\max_b \sum_{n=1;n\neq t}^{N_{\text{test}}} w_n(\boldsymbol{x}^t) \left\{ \text{c.} - \frac{1}{2}\ln b - \left(a_0 + \frac{1}{2}\right) \ln \frac{1}{b}\left(b + \frac{\Delta_n^2}{2}\right)\right\} \\
&= \max_b \sum_{n=1;n\neq t}^{N_{\text{test}}} w_n(\boldsymbol{x}^t) \left\{ \text{c.} + a\ln b - \left(a_0 + \frac{1}{2}\right) \ln \left(b + \frac{\Delta_n^2}{2}\right)\right\}
\end{aligned}$$

to obtain an iterative formula

$$\frac{1}{b(\boldsymbol{x}^t)} \leftarrow \frac{2a_0 + 1}{a_0} \sum_{n=1;n\neq t}^{N_{\text{test}}} \frac{\tilde{w}_n(\boldsymbol{x}^t)}{2b(\boldsymbol{x}^t) + [y^{(n)} - f(\boldsymbol{x}^{(n)})]^2}, \tag{D.13}$$

where $\tilde{w}_n \triangleq \frac{w_n}{\sum_m w_m}$. For the kernel function, we can use, e.g.,

$$w_n(\boldsymbol{x}^t) = w_0 + \exp\left(-\frac{\|\boldsymbol{x}^{(n)} - \boldsymbol{x}^t\|^2}{2\eta_0^2}\right). \tag{D.14}$$

We need an initial estimate for b_0. One reasonable choice is obtained by replacing $[y^{(n)} - f(\boldsymbol{x}^{(n)})]^2$ with its average σ_{yf}^2, yielding

$$b_0 \approx a_0 \sigma_{yf}^2, \quad \text{where} \quad \sigma_{yf}^2 \triangleq \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} [y^t - f(\boldsymbol{x}^t)]^2. \tag{D.15}$$

Recall that the derived t-distribution has the scale parameter $\sqrt{b_0/a_0}$. As the scale parameter corresponds to the standard deviation, we see that the above relationship $b_0/a_0 \sim \sigma_{yf}^2$ is consistent with it.

Equation (D.15) can be also used as a constant approximation for $b(\boldsymbol{x}^t)$. However, for evaluating the probability density function of $\boldsymbol{\delta}$, it tends to give a bit too large value. This is understandable because if, e.g., $N_{\text{test}} = 1$, a majority of the probability mass is from the prior, giving a dull peak around zero. To reproduce a realistic distribution, we need to 'simulate' the situation where there are a reasonable number of test samples. This can be done by choosing a smaller b_0/a_0 because the precision (the reciprocal of the variance) linearly increases as a function of the sample size in Bayesian estimation. Hence, when estimating the distribution in GPA, we can include a correction factor c_b as

$$b_0 \sim a_0 \sigma_{yf}^2 / c_b. \tag{D.16}$$

Intuitively, c_b is interpreted as the number of virtual parameters. Typically, $c_b \sim 10$ gives a reasonable distribution but it should be viewed as a free parameter that can be tuned according to each use-case.

E COMPARING ATTRIBUTION SCORES

We computed the following four metrics to evaluate the consistency among different attribution methods. The first and second metrics are Kendall's τ and Spearman's ρ, calculated for two *absolute* attribution score vectors. They take a value of 1 if the orders are the same regardless of their values. The third metric is what we call the sign match ratio (SMR), which takes on 1 when all the signs are consistent between corresponding vector elements. When comparing an attribution score vector $\boldsymbol{u}$ against a reference score vector $\boldsymbol{r}$, SMR is defined as

$$(\text{SMR}) \triangleq 1 - \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}\left(\text{sign}(r_i)\,\text{sign}(u_i) = -1\right), \tag{E.17}$$

where $\mathbb{I}(\cdot)$ is the indicator function that takes on 1 when the argument is true, 0 otherwise. We define $\text{sign}(0) = 0$ in this case. Note that this favors sparse attribution scores: If $\boldsymbol{r} = \boldsymbol{0}$, then the score is always 1 regardless of $\boldsymbol{u}$. Finally, the fourth metric is what we call hit25, which gives 1 when the top 25% of the absolute entries perfectly match between $\boldsymbol{r}$ and $\boldsymbol{u}$, and 0 if none of the top 25% members of $\boldsymbol{r}$ is included in that of $\boldsymbol{u}$. As hit25 depends on neither the sign nor the rank, it quantifies simply the match of top contributors.