# Toward Universal Laws of Outlier Propagation

**Aram Ebtekar**[1,*] **Yuhao Wang**[2,3,*] **Dominik Janzing**[3]

[1]Independent Researcher [2]National University of Singapore [3]Amazon Research Tübingen, Germany

## Abstract

When a variety of anomalous features motivate flagging different samples as *outliers*, Algorithmic Information Theory (AIT) offers a principled way to unify them in terms of a sample's *randomness deficiency*. Subject to the Independence of Mechanisms Principle, we show that for a joint sample on the nodes of a causal Bayesian network, the randomness deficiency decomposes into a sum of randomness deficiencies at each causal mechanism. Consequently, anomalous observations can be attributed to their root causes, i.e., the mechanisms that behaved anomalously. As an extension of Levin's law of randomness conservation, we show that weak outliers cannot cause strong ones. We show how these information theoretic laws clarify our understanding of outlier detection and attribution, in the context of more specialized outlier scores from prior literature.

## 1 INTRODUCTION

Generically speaking, an *outlier* is an anomalous or atypical event, suggesting possible interference and/or downstream effects. Anomaly detection plays a crucial role in business, technology, and medicine. Typical use cases range from fraud detection in finance and online trading [Donoho, 2004], performance drops in manufacturing lines [Susto et al., 2017] and cloud computing applications [Gan et al., 2021, Ma et al., 2020, Hardt et al., 2023], health monitoring in intensive care units [Maslove et al., 2016], to explaining extreme weather and climate events [Zscheischler et al., 2022]. It has motivated a vast effort towards developing methodologies relevant to outlier analysis. For examples, we refer the reader to some of the early works in statistics and computer science [Freeman, 1995, Rocke and Woodruff, 1996, Rousseeuw and Leroy, 2003, Aggarwal, 2017].

In complex systems, an anomaly may cause a large cascade of related anomalies [Panjei et al., 2022]. To mitigate them, it is not enough to merely *detect* anomalies; we must also identify which of the anomalies was the root cause [Budhathoki et al., 2022, Ikram et al., 2022, Li et al., 2022, Hardt et al., 2023, Wang et al., 2023b,a]. Thus, we implicitly face the *counterfactual* question of what conditions could have been different to prevent the (usually undesired) anomalous event.

To render a complex system accessible to human understanding, we begin with a causal model of its relevant mechanisms, specifying not only their default behavior, but also their behavior under modifications called *interventions*. Such a model should be modular in two respects. First, we may want to understand the *causal pathway*, along which a perturbation of any part of the system propagates through its components until it generates the event. Second, we want to "blame" some component(s) of the system, while acknowledging that others worked as expected.

Causal Bayesian networks offer such a modular description, specifying causal relations via a directed acyclic graph (DAG) $G$ with random variables $X_1, \ldots, X_n$ as nodes [Pearl, 2009, Spirtes et al., 1993]. Under the causal Markov condition [Pearl, 2009], the joint distribution factorizes according to

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_j \mid \mathrm{PA}_j), \tag{1}$$

where $\mathrm{PA}_j$ denotes the parents of $X_j$ in $G$, i.e., its direct causes. We will think of each conditional distribution $P(X_j \mid \mathrm{PA}_j)$ as an *independent mechanism* of the system, which can in principle be changed or replaced without changing the others (see 2.1 and 2.2 in Peters et al. [2017] for a historical overview).

[*]These authors contributed equally to this work.

Provided that *most* of the mechanisms operate normally, we can blame an anomalous joint observation $x_1, \ldots, x_n$ on a small number of mechanisms as root causes, in alignment with Schölkopf et al. [2021]'s "sparse mechanism shift hypothesis". Budhathoki et al. [2022] provide a most elaborate formalization of the idea of *attributing anomalies to mechanisms*. We develop our concepts starting from this baseline.

### 1.1 OUTLIER SCORES FROM P-VALUES

Budhathoki et al. [2022] introduce what they call an Information Theoretic (IT) outlier score[1] via

$$\lambda_\tau(x) := -\log P(\tau(X) \geq \tau(x)), \tag{2}$$

where $x$ denotes an observation of the random variable $X$, and $\tau : \mathcal{X} \to \mathbb{R}$ is a feature statistic whose highest values we consider anomalous.

$\lambda_\tau$ can be viewed as a statistical test of the null hypothesis that $x$ was sampled from $P$: setting the base of the logarithm to 2 yields the p-value $2^{-\lambda_\tau(x)}$. A small p-value (or large $\lambda_\tau$) corresponds to an unusual sample under $P$, which can be labeled an outlier. Since $x$ is a single observation, anomaly scoring thus reduces to classical hypothesis testing on a sample size of 1 [Shen, 2020, Vovk, 2020][2].

### 1.2 QUANTITATIVE ROOT CAUSE ANALYSIS IN CAUSAL NETWORKS

In order to quantify the contribution of different upstream mechanisms to an anomalous observation, the arXiv version of Budhathoki et al. [2022] defines the *conditional outlier scores*:

$$\lambda_\tau(x_j \mid \mathrm{pa}_j) := -\log P(\tau(X_j) \geq \tau(x_j) \mid \mathrm{PA}_j = \mathrm{pa}_j).$$

They demonstrate that this score can be interpreted as measuring the anomalousness of the mechanism's noise term,[3] whenever $\lambda_\tau(X_j \mid PA_j)$ is independent of $PA_j$. The feature functions $\tau_j$ can be node-specific, which is essential when the variables $X_j$ have different characteristics (e.g., different dimensionality or data types). When the specific choice of $\tau$ is not crucial for the discussion, we will simplify notation by dropping the subscript. The arXiv version of Budhathoki et al. [2022] extends this framework by introducing a joint outlier score through "convolution":

$$\begin{aligned} \lambda(x_1, \ldots, x_n) \quad &:= \quad \sum_{j=1}^{n} \lambda(x_j \mid \mathrm{pa}_j) \\ &- \quad \log \sum_{i=1}^{n-1} \frac{(\sum_{j=1}^{n} \lambda(x_j \mid \mathrm{pa}_j))^i}{i!}. \end{aligned} \tag{3}$$

The second term serves as a correction to ensure the score maintains the properties of an IT score, provided that the conditional distributions have densities with respect to the Lebesgue measure.

### 1.3 MONOTONICITY OF SCORES

Okati et al. [2024] formalize an intuitive principle of anomaly propagation: unless the connecting mechanisms also behave anomalously, a moderate outlier (measured by $\lambda(x_1)$) should not *cause* an extreme outlier (measured by $\lambda(x_2)$). This can be stated for the bivariate causal relation $X_1 \to X_2$ as follows:

**Lemma 1.1** (Weak IT outliers rarely cause strong ones)**.** *If $\lambda$ are IT scores and $X_1$ is a continuous variable, the following inequality holds:*

$$P(\lambda(X_2) \geq \lambda(x_2) \mid \lambda(X_1) \geq c) \leq 2^{c-\lambda(x_2)}.$$

In other words, whenever we generate an anomaly $x_1$ randomly from $P(X_1 \mid \lambda(X_1) \geq c)$ and then generate $X_2$ via its mechanism $P(X_2 \mid X_1)$, the resulting outlier score is unlikely to be much larger than $c$. The reason is that the probability of an event of the form $\lambda(X_j) \geq c$ decreases exponentially in $c$.

Note that Lemma 1.1 says nothing about any individual $x_1$ – some values of $x_1$ may generate stronger outliers with high probability, although "most" do not. For instance, it's easy to construct a deterministic relation $X_2 = f(X_1)$, whose non-linearity maps a tiny region of non-anomalous values $x_1$ to anomalous values $x_2$. On the other hand, whenever this occurs, it must either be the case that the nonlinearity of $f$ is specially "tuned" to pick out those specific $x_1$; or else, $x_1$ is anomalous in some other sense not captured by the feature $\tau_1$. Our Independence of Mechanisms Principle will rule out the former case, prompting the need for a general outlier score to capture the latter case.

### 1.4 LIMITATIONS OF CURRENT APPROACHES

**Rigid definition of outliers** $\lambda_\tau$ defines outliers in terms of a feature statistic $\tau$ that is chosen in advance.

[1] While Budhathoki et al. [2022] use the natural base $e$, we use base 2 logarithms to align with binary program lengths in algorithmic information theory. In effect, we express the outlier score in units of bits [Frank, 2005].

[2] Note that Tartakovsky et al. [2012] discuss anomalies as change points over multiple observations, whereas we focus on anomalies confined to an individual observation.

[3] This aligns in spirit with Von Kügelgen et al. [2023] where events are also backtracked by attributing them to the noise variables.

Sometimes, we only know what makes a sample abnormal after seeing it, so we would like to be able to choose $\tau$ with hindsight. For example, suppose $P$ is univariate Gaussian. It seems natural to define $\tau(x) := |x - \mu|$, so that the anomaly detector flags extremely high or low values. However, we might also like to flag observations such as $x = 0$ or $x = \mu$, as it would be a surprising coincidence to see precisely these values. While the probability *density* at these points may be high, intuitively it seems much more likely to guess these specific numbers as a result of some alternative process different from $P$, than it is to sample them from $P$ (which indeed occurs with probability zero).

As another example, suppose each component of a continuous multivariate observation $x$ represents the reading of a different sensor. The signal transmission from all $d$ sensors may be broken in such a way that all components equal to the same constant "idle" state $c$. Such coincidences also indicate an unusual event.

Aggarwal [2016] describe a broad variety of different outliers beyond the above toy examples: These can be unusual frequencies of words in text documents [Mohotti and Nayak, 2020], unexpected patterns in images [Yakovlev et al., 2021], unusually large cliques in graphs [Hooi et al., 2016], or points lying in low density regions [Breunig et al., 2000].

**No general decomposition rule** While (3) nicely decomposes the joint outlier score into mechanism-specific conditional scores, a key limitation is that this decomposition relies on *defining* the joint score based on a sum of conditional scores. This does not suggest that any reasonable outlier score (e.g., using a generic feature function $\tau$) for the joint observation can be decomposed in this manner.

### 1.5 OUR CONTRIBUTIONS

This paper's contributions are purely conceptual. We do not propose a practical algorithm for outlier detection or root cause analysis, as our core definitions can only be computed in the infinite-runtime limit. Instead, we provide a theoretical framework for calibrating and interpreting outlier scores. The framework ensures that outlier scores meet three critical criteria: (i) Comparability across diverse probability spaces and data modalities. (ii) Non-increasingness along causal chains of downstream effects, regardless of variable modalities. (iii) Well-defined attribution of joint system outlier scores to anomalies of mechanisms.

Our ideas were guided by the following general working hypothesis, which we believe applies far beyond the subject of this paper:

**Principle 1.2** (Information Theory as a Guide)**.** *Good information theoretic concepts enable many nice theorems, but they are often hard to work with in practice. However, together with distributional assumptions (e.g. Gaussianity) they can boil down to simple concepts (e.g. linear algebraic expressions). The resulting formulae may be valid and useful beyond the distributional assumptions (e.g. by virtue of linear algebra).*

While Shannon information is sometimes hard to estimate from small sample sizes, *algorithmic* information is even worse: Kolmogorov complexity is not even computable. Furthermore, its identities hold only up to machine-dependent additive constants. Therefore, we owe it to the reader to show that our algorithmic information theoretic concepts trigger insights that can be applied in practice, as we will try in Section 4.

The paper is structured as follows. Section 2 reviews concepts from statistics, information theory, and causality that we build upon. Section 3 derives the decomposition, and uses it to show that weak outliers do not cause stronger ones. Section 4 discusses simple examples, and Section 5 a toy experiment. For a cleaner exposition, we defer some formal proofs and definitions to the supplementary material.

## 2 KEY INGREDIENTS

### 2.1 STATISTICAL TESTING WITH E-VALUES INSTEAD OF P-VALUES

While p-values are the most famous measure of evidence in statistical testing, e-values are recently gaining popularity for their superior ability to aggregate evidence across multiple tests [Ramdas and Wang, 2024]. These values are inconsistently scaled in the literature, with e-values being comparable to reciprocals of p-values and exponentials of algorithmic randomness scores. To remove any obfuscation arising from scaling conventions, we introduce the following definitions.

**Definition 2.1.** *A probability-bounded test (**p-test**) in ratio form is a statistic $\Lambda : \mathcal{X} \to [0, \infty]$ satisfying $\forall \varepsilon > 0, P(\Lambda(X) \geq 1/\varepsilon) \leq \varepsilon$. An expectation-bounded test (**e-test**) in ratio form is a statistic $\Lambda : \mathcal{X} \to [0, \infty]$ satisfying $\mathbf{E}_{X \sim P}(\Lambda(X)) \leq 1$.*

We say a statistic $\Lambda : \mathcal{X} \to [0, \infty]$ is a p-test (or e-test) in *probability form*, if $1/\Lambda$ is a p-test (or e-test) in ratio form. Note that what is commonly called a "p-value" is a p-test in probability form, satisfying $\Lambda(X) \leq \epsilon$ with probability at most $\epsilon$. In contrast, what the literature calls an "e-value" is an e-test in *ratio* form.

Similarly, a statistic $\lambda : \mathcal{X} \to [-\infty, \infty]$ is a p-test (or e-test) in *log form*, if $2^\lambda$ is a p-test (or e-test) in ratio

form. Our convention is to use lowercase symbols like $\lambda$ to indicate log form. For more details on tests, see Section 7 in the supplementary material.

p-tests provide a straightforward way to control the Type I error rate, i.e., false positives. While we want to be flexible about what kinds of anomalies to consider, we should insist that *most* samples from $P$ are not outliers. To ensure that samples from $P$ are labeled as outliers at a rate no larger than a desired threshold $\epsilon$, we choose a p-test and flag only those samples whose scores are above $1/\epsilon$ (i.e., below $\epsilon$ when expressed in probability form).

By Markov's inequality, every e-test is also a p-test. Hence, e-tests achieve the same false positive rate; however, they are more conservative. In return, e-tests offer many advantages related to composability and optional stopping [Grünwald et al., 2020, Ramdas et al., 2023, Ramdas and Wang, 2024]. Our main result, Theorem 3.2, concerns a general-purpose e-test that conveniently decomposes into a sum of contributions, attributable to a causal network's mechanisms.

### 2.2 BASIC NOTIONS FROM ALGORITHMIC INFORMATION THEORY

In a very general sense, an observation $x$ may be deemed an outlier whenever it belongs to a low-probability set that is flagged by some computable test. This condition can be rephrased in terms of the length of a compressed description of $x$.

To formalize description lengths, we fix a universal prefix-free Turing machine, and discretize our variables so that their values can be written as strings. The conditional *Kolmogorov complexity* $K(x \mid y)$ is the bit length of the shortest program $p$ that outputs $x$ when given access to another string $y$ [Li and Vitányi, 2019]. It satisfies the Kraft inequality $\sum_x 2^{-K(x|y)} < 1$. When $y$ is the empty string, we simply write $K(x)$. Just as Shannon's entropy measures information content for a probability distribution, $K(x)$ measures it for an individual sample $x$.

Using a standard prefix-free encoding of $n$-tuples, Li and Vitányi [2019] also define the joint Kolmogorov complexity $K(x_1, \ldots, x_n)$. By analogy with Shannon's mutual information, they define the algorithmic mutual information between $x$ and $y$, conditional on $z^*$, by

$$\begin{aligned} I(x : y \mid z^*) &:= K(x \mid z^*) + K(y \mid z^*) - K(x, y \mid z^*) \\ &\stackrel{+}{=} K(x \mid z^*) - K(x \mid (y, z)^*). \end{aligned}$$

Here, $z^*$ denotes a shortest program that outputs $z$, and $\stackrel{+}{=}$ denotes equality up to a constant dependent on the universal machine, but not on $x$ or $y$[4]. We say $x$ and $y$ are *conditionally independent* given $z^*$, abbreviated as $x \perp\!\!\!\perp y \mid z^*$, if $I(x : y \mid z^*) \stackrel{+}{=} 0$.

### 2.3 UNIVERSAL TESTS

Semicomputable p- and e-tests are called Martin-Löf and Levin tests, respectively. Each class of tests can be aggregated, resulting in universal tests that effectively combine every conceivable anomaly scoring algorithm, in order to detect the broadest possible variety of outliers. Section 8 in the supplementary material contains more details; here we give a brief overview.

**Definition 2.2** (Domination property)**.** *For two statistical tests $\lambda_1$ and $\lambda_2$ expressed in log form, we say* $\lambda_1$ dominates $\lambda_2$ *if there exists a constant $c \in \mathbb{R}$ such that for all observations $x$ in the sample space,*

$$\lambda_1(x) \geq \lambda_2(x) - c.$$

*Intuitively, this means $\lambda_1$ can detect at least all the anomalies that $\lambda_2$ can detect, up to a constant term.*

The domination property provides a natural way to compare the power of different statistical tests. Since Definition 2.1 restricts us from increasing a test everywhere, it is perhaps surprising that the class of semicomputable tests contains a *universal* test that dominates all the others.

**Theorem 2.3** (Universality of randomness deficiency)**.** *Let $\mathcal{X}$ be a discrete space that can be interpreted as a subset of $\{0,1\}^*$ in a canonical way, and $P$ be a computable probability distribution on $\mathcal{X}$ (i.e., with finite description length). Then, the randomness deficiency of $x \in \mathcal{X}$, defined by*

$$\delta(x) := -\log P(x) - K(x \mid P^*), \tag{4}$$

*is a universal e-test, dominating all other semicomputable e-tests.*

The intuition behind Eq. (4) is that typical samples from a distribution $P$ are optimally compressed by encodings of length $-\log P(x)$. When a sample $x$ can be compressed beyond this theoretical limit, we consider it "anomalous". This includes the case where we observe $x = 0$ from a discretized centered Gaussian distribution. Despite being the mode of the distribution, 0 is considered anomalous because its negative log-likelihood $-\log P(x)$ is substantially larger than its Kolmogorov complexity $K(0) \stackrel{+}{=} 0$.

[4] Note that $x^*$ contains more information than $x$, because $x$ is easily generated from $x^*$ but not vice versa. Conditioning on $x$ instead of $x^*$ would result in error terms that are often constant, and at worst logarithmic in the length of $x$ [Gács, 2021].

**Example 1** (Uniform distribution)**.** *When $P$ is uniform over all $2^d$ binary strings of length $d$, (4) becomes $\delta(x) = d - K(x \mid P^*) \stackrel{+}{=} d - K(x \mid d)$, measuring the degree to which $x$ is compressible.*

When a distributional estimate $\hat{P}$ is inferred from data, [Bishop, 1993] uses the "reconstruction loss" $-\log \hat{P}(x)$ as an outlier score. Without further considerations, this score is not calibrated, because there may be a huge region with low density, with high total probability. However, if the term $K(x \mid P^*)$ is also small, then we can say that $x$ is special within this huge set.[5]

### 2.4 ALGORITHMIC MARKOV CONDITION AND INDEPENDENCE OF MECHANISMS

In order to identify root causes of an anomaly, we would like to decompose the randomness deficiency into a sum of contributions from each causal mechanism in a Bayesian network. Our framework for doing so requires an algorithmic analogue of the causal Markov condition. Specifically, Janzing and Schölkopf [2010] propose an adaptation of (1) that characterizes algorithmic dependencies between *individual objects*, rather than statistical dependencies between *random variables*:

**Postulate 2.4** (Algorithmic Markov condition)**.** *Let $x_1, \ldots, x_n$ be binary words describing objects whose causal relations are given by the DAG $G$. Then the joint complexity of $(x_1, \ldots, x_n)$ decomposes as*

$$K(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j \mid \mathrm{pa}_j^*).$$

*Furthermore, for any three sets $R, S, T$ of nodes such that $R$ and $S$ are d-separated by $T$, we have*

$$R \perp\!\!\!\perp S \mid T^*.$$

Janzing and Schölkopf [2010] argue that, within a causal Bayesian network whose nodes represent random variables, the mechanisms $P_{X_j|\mathrm{PA}_j}$ constitute independent information-bearing objects. Given $m$ observations of $n$-tuples $(x_1^1, \ldots, x_n^1), \ldots, (x_1^m, \ldots, x_n^m)$, they define a DAG $G^m$ that connects the $n$ mechanisms $P_{X_i|\mathrm{PA}_i}$ with the $n \times m$ observations in the following way: Each $x_i^l$ has the observations $\mathrm{pa}_i^l$ and the *node* $P_{X_i|\mathrm{PA}_i}$ as parents, formalizing the idea that the mechanisms $P_{X_i|\mathrm{PA}_i}$ determine how the parents in $G$ influence the respective observation. Figure 1 shows $G^m$ for the DAG $X \to Y$. For our purpose, it is sufficient to consider $G^1$, where we have only one observation $x_j$ from each node $X_j$ in $G$. Accordingly, we construct $G^1$ from $G$ as follows: replace each variable $X_j$ with its observation $x_j$, and make $P_{X_j|\mathrm{PA}_j}$ a parent of $X_j$. Each $P_{X_j|\mathrm{PA}_j}$ becomes a root node, and these are the only root nodes in $G^1$. This placement of $P_{X_j|\mathrm{PA}_j}$ as root nodes, which implies their algorithmic independence, formalizes the Independence of Mechanisms[6]. Applying the algorithmic Markov condition to $G^1$ yields:

**Lemma 2.5** (Conditional irrelevance of other mechanisms and predecessors when parents are given)**.** *Let $X_1, \ldots, X_n$ be causally ordered. Given its parents $\mathrm{pa}_j$ and the mechanism $P_{X_j|\mathrm{PA}_j}$, none of the other mechanisms $(P_{X_i|\mathrm{PA}_i})_{i \neq j}$ and none of the causal predecessors $(x_i)_{i<j}$ enable further compression of $x_j$. That is,*

$$x_j \perp\!\!\!\perp (x_i)_{i<j}, (P_{X_i|\mathrm{PA}_i})_{i \neq j} \mid (\mathrm{pa}_j, P_{X_j|\mathrm{PA}_j})^*.$$

## 3 DECOMPOSITION OF RANDOMNESS DEFICIENCY

We start by presenting the decomposition of randomness deficiency in the bivariate case. Specifically, consider the DAG $X \to Y$ between a cause $X$ and its effect $Y$. Extending Eq. (4) to joint and conditional distributions, define the joint randomness deficiency of outcomes $(x, y)$ with respect to $P_{X,Y}$ by

$$\delta(x, y) := -\log P(x, y) - K(x, y \mid (P_{X,Y})^*), \quad (5)$$

and the conditional randomness deficiency by

$$\delta(y \mid x) := -\log P(y \mid x) - K(y \mid (x, P_{Y|X})^*). \quad (6)$$

**Lemma 3.1.** *(Decomposition of randomness deficiency for a cause-effect pair) For any two random variables $X \to Y$ (i.e., $X$ being the cause of $Y$), and for individual observations $x$ and $y$, the following equality holds under Postulate 2.4 for the DAG in Figure 1:*

$$\delta(x, y) \stackrel{+}{=} \delta(x) + \delta(y \mid x)$$

The proof relies on Lemma 2.5, and is provided in Section 9 of the supplementary. The following example shows why Independence of Mechanisms is essential in order for the randomness deficiency to be additive.

[5] Note that conformal p-values [Bates et al., 2021] do not require any parametric assumptions or density estimation, since their coverage guarantees rely on exchangeability alone. However, statements on *conditional* coverage [Barber et al., 2019] are too weak for our purpose of calibrating *conditional* outlier scores.

[6] Janzing and Schölkopf [2010] call the algorithmic independence of different $P_{X_j|\mathrm{PA}_j}$ *Independence of Mechanisms*. Here, we prefer using this term for the underlying idea of applying the Algorithmic Markov Condition with the mechanisms as root nodes.

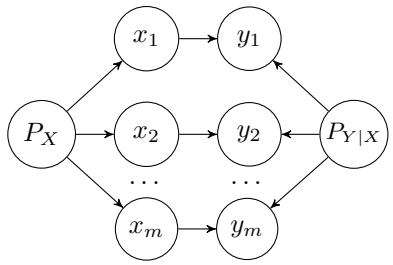


Figure 1: $x_1, \ldots, x_m$ sampled from $P_X$, $y_1, \ldots, y_m$ sampled from $P_{Y|X}$.

**Example 2** (No additivity without Independence of Mechanisms)**.** *Let $\mathcal{X} = \mathcal{Y} = \{0,1\}^d$ and $P_X$ be the uniform distribution. Furthermore, let $P_{Y|X}$ be the deterministic mechanism*

$$P(y \mid x) = 1 \quad \textit{iff} \quad y = x \oplus x_0,$$

*where $\oplus$ denotes bitwise XOR and $x_0$ is an algorithmically random string with $K(x_0) = d$. Since $P_{Y|X}$ is deterministic, we always have $\delta(y \mid x) \stackrel{+}{=} 0$.*

*Consider the input $x = x_0$, which violates the conditional independence relation $x \perp\!\!\!\perp P_{Y|X} \mid P_X$. Then, $\delta(x) \stackrel{+}{=} 0$ because $x$ is random with respect to $P_X$, but $\delta(x,y) \stackrel{+}{=} d \gg 0$ because $(x,y) = (x_0, 0^d)$ is a simple function of $x_0$, which is easily deciphered from $P_{X,Y}$. Since $x$ is non-generic relative to $P_{Y|X}$, we find that the randomness deficiency of $(x,y)$ cannot be attributed to $x$, nor to the mechanism generating $y$ from $x$.*

The multivariate generalization of Lemma 3.1 is this paper's main result:

**Theorem 3.2** (Decomposition of multivariate joint randomness deficiency)**.** *Let the set of strings $x_1, x_2, \ldots, x_n$ be causally connected by a directed acyclic graph $G$, so that the causal Markov condition holds for $G^m$. Then, the joint randomness deficiency of $x_1, x_2, \ldots, x_n$ decomposes into a sum of conditional randomness deficiencies of the mechanisms:*

$$\delta(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j),$$

*where $\delta(x_j \mid \mathrm{pa}_j)$ denotes the randomness deficiency of $x_j$ given its parents.*

The proof is by induction over the number $n$ of nodes; in order to chain the $\stackrel{+}{=}$ equalities, we treat $n$ as a constant. See Section 9 in the supplementary for details, as well as a proof of the following corollary:

**Corollary 3.3** (Weak outliers do not cause stronger ones)**.** *If there is a unique root cause $j \in \{1, \ldots, n\}$, in the sense that*

$$\delta(x_i \mid \mathrm{pa}_i) \stackrel{+}{=} 0 \quad \textit{for } i \neq j,$$

*and the conditions of Theorem 3.2 are met, then*

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_j \mid \mathrm{pa}_j) \quad \forall i \in \{1, \ldots, n\}. \tag{7}$$

In essence, (7) states that none of the nodes has a randomness deficiency exceeding that of the root cause. In a nutshell, "weak outliers cannot cause strong ones". – We have thus found an AIT version of Lemma 1.1.[7] It is worth noting that Levin [1984] (see also Gács [2021]) demonstrated a similar principle, calling it *randomness conservation*. It says that the output of a mechanism cannot exhibit a substantially larger randomness deficiency than the input - under the condition that the mechanism itself has a constant description length. In physics, randomness deficiency corresponds to a lack of entropy [Zurek, 1989, Gács, 1994]. Therefore, the second law of thermodynamics amounts to an instance of randomness conservation [Ebtekar and Hutter, 2025]. In our context, *simplicity* of mechanisms is generalized by the *independence* of mechanisms.

## 4 RELATION TO COMPUTABLE ANOMALY SCORES

In practice, we cannot directly use the randomness deficiency to detect outliers, because it is not computable. However, it is *semicomputable*: by searching over descriptions of $x$, one can gradually improve the upper bound on its Kolmogorov complexity, thus obtaining approximations that monotonically approach the randomness deficiency in the limit. Since a generic search would be extremely slow, the search should focus on features or models suitable to the domain in question. We leave the development of good domain-specific anomaly search algorithms to future work.

On the other hand, the search need not be performed algorithmically. During statistical analysis, one may identify any computable feature, perhaps even *after* observing $x$, and from it derive a lower bound on the randomness deficiency.

This section describes simple scenarios in which such bounds boil down to simple and well-known outlier scores. Thus, the randomness deficiency provides a practical recipe for obtaining and calibrating computable scores. Moreover, its theoretical properties motivate us to seek computable scores with the same desirable properties: *decomposition of joint scores into scores of the mechanisms* and *monotonicity under marginalization*.

[7] Hence, in the language of modern resource theories in physics [Coecke et al., 2016], Definition 5.1, an anomaly is a "resource" and $\delta$ a "monotone" measuring its value.

**Example 3** (z-score)**.** *For a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma)$, the squared z-score reads $z^2(x) := (x-\mu)^2/\sigma^2$. At a fixed level of precision, $K(x \mid P^*) \overset{+}{\leq} 2\log|x-\mu|$. Substituting the log likelihood*

$$-\log P(x) = \log\sqrt{2\pi}\sigma + \frac{\log e}{2} z^2(x),$$

*and treating $\sigma$ as a constant, yields*

$$\delta(x) \overset{+}{\geq} \frac{\log e}{2} z^2(x) - 2\log|x-\mu|.$$

Note that while the well-known identity $\mathbf{E}[z^2] = 1$ tells us that $z^2$ is an e-value, the universal e-value $2^\delta \approx e^{z^2/2}$ provides far stronger evidence at the tails. Now we generalize to $n > 1$ dimensions:

**Example 4** (Decomposing squared Mahalanobis distance)**.** *The randomness deficiency of a random vector $\boldsymbol{x} \in \mathbb{R}^n$, drawn from a centered Gaussian, satisfies*

$$\delta(\boldsymbol{x}) \overset{+}{\geq} \frac{\log e}{2} \boldsymbol{x}^T \Sigma_{\boldsymbol{X}}^{-1} \boldsymbol{x} - O(\log \|\boldsymbol{x}\|_\infty),$$

*where $\Sigma_{\boldsymbol{X}}$ denotes the covariance matrix and $n$ is treated as a constant. The leading term is $(\log e)/2$ times the squared Mahalanobis distance (M-distance, for short).*

*Let $\boldsymbol{X}$ be generated by a causal Bayesian network with linear structural equations $\boldsymbol{X} = A\boldsymbol{X} + \mathbf{N}$, where $A$ is strictly lower triangular and $N_i$ are independent noise variables with variance $\sigma_i^2$. Using the transformation $\mathbf{N} = (I - A)\boldsymbol{X}$, we obtain a diagonal form*

$$\boldsymbol{x}^T \Sigma_{\boldsymbol{X}}^{-1} \boldsymbol{x} = \sum_{i=1}^{n} \frac{n_i^2}{\sigma_i^2}.$$

*For large $n_i$, the expression $n_i^2/\sigma_2$ is roughly proportional to the randomness deficiency of the mechanism $P(X_i \mid PA_i)$. Hence, decomposition of randomness deficiencies translates asymptotically to decompositions of the squared M-distance (as used for multivariate anomaly detection Aggarwal [2016]) into $z^2$-scores. Furthermore, M-distance is non-increasing with respect to marginalization to a subset of variables (see Section 10 in the supplementary), resembling the monotonicity of randomness deficiency.*

Example 4 supports Principle 1.2: a conservative bound on the randomness deficiency yields a computable e-value, as a function of the $z^2$-scores. Moreover, we verified using linear algebra that the $z^2$-score of any variable cannot exceed the sum of all the "noise scores".

Note that the upper bound in Corollary 3.3 is a *conditional* randomness deficiency. The following example shows that a root cause may have a smaller *marginal* randomness deficiency than its downstream effects.

**Example 5** (Root cause with small marginal score)**.** *With $N_i \sim \mathcal{N}(0,1)$, consider the three-node model*

$$\begin{aligned} X_1 &:= N_1, \\ X_2 &:= 2X_1 + N_2, \\ X_3 &:= X_1 - X_2 + N_3. \end{aligned}$$

*Then, $X_2 \sim \mathcal{N}(0, \sqrt{5})$. Starting from a typical observation, suppose we perturb $x_2$ to some very large value. Then, its marginal randomness deficiency $\delta(x_2)$ increases with $x_2^2/5$, while its conditional randomness deficiency $\delta(x_2 \mid x_1)$ increases with $x_2^2/1$ (where we've ignored the factor of $\frac{\log e}{2}$ from Example 3). Since $X_3 \sim \mathcal{N}(0, \sqrt{3})$, $\delta(x_3)$ increases with $x_2^2/3$. Thus, for large perturbations of $x_2$, we have $\delta(x_2) \ll \delta(x_3) \ll \delta(x_2 \mid x_1) \approx \delta(x_1, x_2)$. If we use marginal scores, it would seem as though $x_2$ caused a stronger outlier $x_3$.*

This paradox resembles Example 2.1 in Li et al. [2024], where it is phrased in terms of $z^2$-scores. Following Principle 1.2, we can describe it entirely in terms of M-distances (which are just $z^2$-scores in the one-dimensional case): the M-distance of $x_3$ can be larger than the M-distance of its root cause $x_2$, but not larger than the M-distance of the pair $(x_1, x_2)$, i.e., the full cause of $x_3$. In other words, to fully quantify the anomaly introduced by perturbing $x_2$, we must also account for the destroyed coupling between $x_1$ and $x_2$, not only for the size of the value $x_2$ itself.

The insight from algorithmic information theory is that the anomaly scores of the vector $(x_1, x_2)$ and the scalar $x_3$ are indeed comparable, and also comparable to the conditional score of $x_2$ given $x_1$, because their calibration was guided by the randomness deficiency.

In the following example, which could be easily generalized to words over an arbitrary alphabet, randomness deficiency essentially boils down to relative entropy:

**Example 6** ($m$-bit binary word)**.** *Let us first consider an $m$-bit word whose bits are set to 1 independently with probability $p$. Counting the number of words $w$ with fixed Hamming weight $\ell$ yields the lower bound*

$$\delta(w) \overset{+}{\geq} -\ell\log p - (m-\ell)\log(1-p) - \log\binom{m}{\ell} - O(\log m).$$

*Using Stirling's approximation $\log m! = m\log m - m\log e + O(\log m)$, one can show that*

$$\delta(w) \overset{+}{\geq} m \cdot \mathrm{KL}\left(\frac{\ell}{m} \,||\, p\right) - O(\log m), \tag{8}$$

*with the binary Kullback-Leibler distance*

$$\mathrm{KL}(q \,||\, p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}.$$

*Here we see that words with unexpectedly low or unexpectedly Hamming weight are both assigned high outlier scores, enabling us to compare both types of outliers. This works despite the probability mass function being monotonic in the Hamming weight $\ell$.*[8]

Many works [Aggarwal, 2016, Akoglu et al., 2012, Noble and Cook, 2003] propose detecting anomalies by compression, but consider anomalies that have *higher* compression length than usual, seemingly in conflict with the randomness deficiency which flags *low* compression lengths. Example 6 resolves this paradox: for $p < 1/2$, any word with $K(w) > m \cdot H(p)$ must have Hamming weight larger than $mp$. Accordingly, its randomness deficiency is bounded from below by (8). Hence, paradoxically, an unusually *high* compression length also implies non-zero randomness deficiency, because it entails an increase of the log likelihood term that outweighs the observed increase in compression length.

More sophisticated notions of anomalies can be obtained, for instance, when $X$ is graph-valued and an anomaly is given by large cliques [Aggarwal, 2016] – e.g. when fraudsters work together frequently on illegal activities, their communication is densely connected. To estimate the random deficiency of a graph $G$ with a clique of size $k$, it suffices to define a probability distribution on the set of graphs with $m$ nodes, and bound a graph's description length by counting the number of graphs with cliques of size $k$.

## 5 EXPERIMENT WITH LEMPEL-ZIV COMPRESSION

So far we have accounted for the term $K(x \mid P^*)$ only by very rough upper bounds rather than trying to approximate it. Here we describe a toy scenario in which we can detect anomalies using the Lempel-Ziv compression algorithm; more difficult scenarios may demand more powerful compression algorithms that are closer to general AI.

Consider the causal DAG: $X_1 \to X_2 \to \cdots \to X_n$, with the structural equations

$$X_1 = N_1; \quad X_j = X_{j-1} + N_j \quad \text{for } j = 2, \ldots, n. \tag{9}$$

Moreover, suppose that every $N_j$ for $j = 1, \ldots, n$ is drawn from the uniform distribution on the set of numbers in $[0, 1]$ discretized to $d \gg 1$ digits of precision. By choosing the uniform distribution, the terms $-\log P(x_j \mid x_{j-1})$ become constant; moreover, we have $K(x_j \mid (x_{j-1}, P)^*) \stackrel{+}{=} K(x_j \mid x_{j-1})$. Since the conditional distribution of each $X_j$, given its parent, is uniform over the (discretized) interval $[x_{j-1}, x_{j-1} + 1]$, the conditional randomness deficiency of $x_j$ reads

$$\delta(x_j \mid x_{j-1}) \stackrel{+}{=} d \cdot \log 10 - K(x_j \mid x_{j-1}).$$

Suppose now we inject an anomaly at some node $j$ by setting $n_j$ to some numbers in $\{0, \ldots, 0.9\}$. As a result, $x_j$ and $x_{j-1}$ now coincide in at most 2 digits, so that $K(x_j \mid x_{j-1}) \stackrel{+}{=} 0$ and $\delta(x_j \mid x_{j-1}) \stackrel{+}{=} d \cdot \log 10$.

Following Steudel et al. [2010], we approximate $K(x_j \mid x_{j-1}^*) \stackrel{+}{=} K(x_j, x_{j-1}) - K(x_{j-1})$ by $R(x_j, x_{j-1}) - R(x_{j-1})$, where $R$ denotes the length of a compressed encoding using the Lempel-Ziv algorithm. Whenever $n_j$ corrupts to a 1-digit number, Lempel-Ziv recognizes that $x_j$ and $x_{j-1}$ coincide with respect to all but 1 or 2 digits. Thus, $R(x_j, x_{j-1}) \approx R(x_{j-1})$, which lets us infer the randomness deficiency of almost $d \log 10$.

We conducted experiments to verify our findings for $n = 4$, with the noise uniformly drawn from numbers in $[0, 1]$ with $d = 10$ digits. We randomly chose one of the 4 nodes as "root cause" and set all but one digit of its noise variables to zero. To detect the root cause, we selected the label $j$ that minimized Lempel-Ziv compression length and found the right one in 100 out of 100 runs.

From a theoretical point of view, the example shows that the joint observation can be anomalous (here in the sense of showing two variables whose digits coincide largely), although none of the variables show unexpected behaviour with respect to their *marginal* distribution. Hence, the root cause can only be found by inspecting which *mechanism* behaves unexpectedly.

## 6 CONCLUSIONS

The algorithmic randomness deficiency offers a principled unified definition of outliers, without prior specification of the feature that exposes the anomaly. On a causal Bayesian network, we saw that the randomness deficiency of a joint observation decomposes along individual causal mechanisms, subject to the Independent Mechanisms Principle. This allows us to trace anomalous observations back to their root causes, identifying specific mechanisms responsible for the anomaly. Moreover, we showed that weak outliers cannot cause strong outliers, extending Levin's law of randomness conservation. These foundational insights can help calibrate anomaly scores as well as inspire more versatile anomaly detection algorithms, supporting root cause analysis in complex systems.

[8] Cover and Thomas [2006], in Section 11.2, call a word "$\epsilon$-typical" when it satisfies $\mathrm{KL}\big(\frac{\ell}{m} \,||\, p\big) \leq \epsilon$. Their equation (11.67) finds that such a word occurs with probability at least $1 - (m+1)^2 \cdot e^{-\epsilon m}$.

**Acknowledgments:** DJ and YW would like to thank the authors of Li et al. [2024] for an interesting discussion about their Example 2.1.

## References

Stavros P Adam, Stamatios-Aggelos N Alexandropoulos, Panos M Pardalos, and Michael N Vrahatis. No free lunch theorem: A review. *Approximation and optimization: Algorithms, complexity and applications*, pages 57–82, 2019.

Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2nd edition, 2016. ISBN 3319475770.

Charu C Aggarwal. *An introduction to outlier analysis*. Springer, 2017.

Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. Fast and reliable anomaly detection in categorical data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 415–424, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311564.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 2019.

Stephen Bates, Emmanuel J. Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 2021.

D.S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas - Second Edition*. Princeton University Press, 2009.

C. M. Bishop. Novelty detection and neural network validation. In Stan Gielen and Bert Kappen, editors, *ICANN '93*, pages 789–794, London, 1993. Springer London.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International Conference on Machine Learning*, pages 2357–2369. PMLR, 2022.

Bob Coecke, Tobias Fritz, and Robert W. Spekkens. A mathematical theory of resources. *Information and Computation*, 250:59–86, 2016. ISSN 0890-5401. Quantum Physics and Logic.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2 edition, 2006.

Steve Donoho. Early detection of insider trading in option markets. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, 2004.

Aram Ebtekar and Marcus Hutter. Foundations of algorithmic thermodynamics. *Physical Review E*, 111:014118, 2025.

Michael P Frank. The indefinite logarithm, logarithmic units, and the nature of entropy. 2005.

Jim Freeman. Outliers in statistical data. *Journal of the Operational Research Society*, 46(8):1034–1035, 1995.

Péter Gács. The Boltzmann entropy and randomness tests. In *Proceedings Workshop on Physics and Computation. PhysComp'94*, pages 209–216. IEEE, November 1994.

Péter Gács. Lecture notes on descriptional complexity and randomness. 2021.

Yu Gan, Mingyu Liang, Sundar Dev, David Lo, and Christina Delimitrou. Sage: practical and scalable ml-driven performance debugging in microservices. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '21, page 135–151, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383172.

Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

Michaela Hardt, William Orchard, Patrick Blöbaum, Shiva Kasiviswanathan, and Elke Kirschbaum. The petshop dataset – finding causes of performance issues across microservices, 2023.

Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD international*

*conference on knowledge discovery and data mining*, pages 895–904, 2016.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. *Advances in Neural Information Processing Systems*, 35:31158–31170, 2022.

D Janzing and B Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Leonid A Levin. Uniform tests of randomness. In *Doklady Akademii Nauk*, volume 227, pages 33–35. Russian Academy of Sciences, 1976.

Leonid A Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61(1):15–37, 1984.

Jinzhou Li, Benjamin B. Chu, Ines F. Scheller, Julien Gagneur, and Marloes H. Maathuis. Root cause discovery via permutations and Cholesky decomposition. arxiv:2410.12151, 2024.

Ming Li and Paul M B Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 4 edition, 2019.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3230–3240, 2022.

Meng Ma, Jingmin Xu, Yuan Wang, Pengfei Chen, Zonghua Zhang, and Ping Wang. Automap: Diagnose your microservice-based web applications automatically. In *Proceedings of The Web Conference 2020*, pages 246–258, 2020.

Per Martin-Löf. The definition of random sequences. *Information and control*, 9(6):602–619, 1966.

David M Maslove, Joel A Dubin, Arvind Shrivats, and Joon Lee. Errors, omissions, and outliers in hourly vital signs measurements in intensive care. *Critical care medicine*, 44(11):e1021–e1030, 2016.

Wathsala Anupama Mohotti and Richi Nayak. Efficient outlier detection in text corpus using rare frequency and ranking. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(6):1–30, 2020.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, 1933.

Caleb C. Noble and Diane J. Cook. Graph-based anomaly detection. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 631–636, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137370.

Nastaran Okati, Sergio Hernan Garrido Mejia, William Roy Orchard, Patrick Blöbaum, and Dominik Janzing. Root cause analysis of outliers with missing structural knowledge. arxiv:2406.05014, 2024.

Egawati Panjei, Le Gruenwald, Eleazar Leal, Christopher Nguyen, and Shejuti Silvia. A survey on outlier explanations. *The VLDB Journal*, 31(5):977–1008, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Aaditya Ramdas and Ruodu Wang. *Hypothesis Testing with E-values*. 2024.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4): 576–601, 2023.

Samuel Rathmanner and Marcus Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6): 1076–1136, 2011.

David M Rocke and David L Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061, 1996.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2003.

B Schölkopf, F Locatello, S Bauer, N R Ke, N Kalchbrenner, A Goyal, and Y Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Alexander Shen. Randomness tests: theory and practice. In *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*, pages 258–290. Springer, 2020.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, NY, 1993.

B Steudel, D Janzing, and B Schölkopf. Causal Markov condition for submodular information measures. *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 464–476, 2010.

Gian Antonio Susto, Matteo Terzi, and Alessandro Beghi. Anomaly detection approaches for semiconductor manufacturing. *Procedia Manufacturing*, 11: 2018–2024, 2017.

Alexander G Tartakovsky, Aleksey S Polunchenko, and Grigory Sokolov. Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):4–11, 2012.

Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 11–14 Apr 2023.

Vladimir Vovk. Non-algorithmic theory of randomness. In *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*, pages 323–340. Springer, 2020.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Dongjie Wang, Zhengzhang Chen, Yanjie Fu, Yanchi Liu, and Haifeng Chen. Incremental causal graph learning for online root cause analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2269–2278, 2023a.

Dongjie Wang, Zhengzhang Chen, Jingchao Ni, Liang Tong, Zheng Wang, Yanjie Fu, and Haifeng Chen. Interdependent causal networks for root cause localization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5051–5060, 2023b.

Larry Wasserman and Kathryn Roeder. Weighted hypothesis testing. 2006.

David H Wolpert. The implications of the no-free-lunch theorems for meta-induction. *Journal for General Philosophy of Science*, 54(3):421–432, 2023.

Kirill Yakovlev, Imad Eddine Ibrahim Bekkouch, Adil Mehmood Khan, and Asad Masood Khattak. Abstraction-based outlier detection for image data. In *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 1*, pages 540–552. Springer, 2021.

Jakob Zscheischler, Jana Sillmann, and Lisa Alexander. Introduction to the special issue: Compound weather and climate events. *Weather and Climate Extremes*, 35:100381, 2022.

Wojciech H Zurek. Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731, 1989.

# Toward Universal Laws of Outlier Propagation (Supplementary Material)

**Aram Ebtekar**[1,*] **Yuhao Wang**[2,3,*] **Dominik Janzing**[3]

[1]Independent Researcher [2]National University of Singapore [3]Amazon Research Tübingen, Germany

## 7 P-TESTS AND E-TESTS

Table 1 summarizes key differences between p-tests and e-tests. While p-tests are linked to controlling false positive rates and p-values, e-tests are closely associated with likelihood ratios and betting game interpretations.

| | p-tests (Martin-Löf) | e-tests (Levin) |
|---|---|---|
| Intuition | False positives / p-values | Betting games / likelihood ratios |
| Defining property | $\forall\epsilon > 0,\ P\left(\Lambda(X) \geq 1/\epsilon\right) \leq \epsilon$ | $\mathbf{E}_{X\sim P}\left(\Lambda(X)\right) \leq 1$ |
| Prototypical example | $1/P(\tau(X) \geq \tau(x))$ | $Q(x)/P(x)$ |
| Combination | $\sup_i w_i\Lambda_i$ | $\sum_i w_i\Lambda_i$ |
| Combination (log form) | $\sup_i\{\lambda_i + \log w_i\}$ | $\log\left(\sum_i w_i 2^{\lambda_i}\right)$ |
| Universal test (log form) | $\sup_i\{\lambda_i(x) - K(i \mid P^*)\}$ | $-\log P(x) - K(x \mid P^*)$ |

Table 1: Summary of types of test statistics (outlier scores)

**Converting between p-Tests from e-Tests** By Markov's inequality, every e-test is also a p-test. Conversely, Vovk and Wang [2021] describe a number of ways to *calibrate* any given p-test into an e-test. A natural choice is the Ramdas calibration, which turns the p-test $\Lambda$ into the e-test

$$\Lambda'(x) := \frac{\Lambda(x) - \ln\Lambda(x) - 1}{\ln^2\Lambda(x)}. \tag{10}$$

**Example 7** (One-tailed p-value)**.** *For any feature statistic $\tau : \mathcal{X} \to \mathbb{R}$, the one-tailed p-value is given by*

$$\Lambda_\tau(x) := P(\tau(X) \geq \tau(x)). \tag{11}$$

*It is easily verified to be a p-test in probability form. The outlier score (2) is the same test as the p-value (11), but expressed in log form.*

**Example 8** (Likelihood ratio)**.** *For any alternative hypothesis described by a sub-probability distribution[1] $Q$, the likelihood ratio is given by*

$$\Lambda_Q(x) := \frac{Q(x)}{P(x)}. \tag{12}$$

[*]These authors contributed equally to this work.

[1]The algorithmic information theory literature refers to sub-probability distributions as *semimeasures*. They generalize probability measures by allowing their sum to be less than 1.

*It is easily verified to be an e-test in ratio form; in fact, all e-tests can be written this way. By the Neyman-Pearson lemma [Neyman and Pearson, 1933], when $Q$ is a probability distribution, $\Lambda_Q$ is the optimal test to distinguish between $P$ and $Q$.*

For the purposes of outlier detection, $\tau$ may be a feature for which a high value should be considered anomalous, while $Q$ might be a distribution that we expect would result from some kind of anomaly. In practice, we can design many tests, each specializing in a different kind of anomaly. The following lemma allows us to merge many such tests into one.

**Lemma 7.1** (Combination tests)**.** *Let $\Lambda_i$ (or $\lambda_i$) be a finite or countably infinite sequence of tests, with associated weights $w_i > 0$ summing to at most 1. Then:*

- *If $\Lambda_i$ are p-tests in prob. form, so is $\inf_i \frac{\Lambda_i}{w_i}$.*
- *If $\Lambda_i$ are p-tests in ratio form, so is $\sup_i w_i \Lambda_i$.*
- *If $\lambda_i$ are p-tests in log form, so is $\sup_i \{\lambda_i + \log w_i\}$.*
- *If $\Lambda_i$ are e-tests in prob. form, so is $\left(\sum_i \frac{w_i}{\Lambda_i}\right)^{-1}$.*
- *If $\Lambda_i$ are e-tests in ratio form, so is $\sum_i w_i \Lambda_i$.*
- *If $\lambda_i$ are e-tests in log form, so is $\log\left(\sum_i w_i 2^{\lambda_i}\right)$.*

*Proof.* Genovese et al. [2006] state a similar result for finitely many independent p-tests, calling the combination test a *weighted Bonferroni procedure*. Vovk and Wang [2021] state a similar result for equally weighted e-tests. It is fairly straightforward to extend these works to our setting; for completeness, the proof is as follows. First, given some p-tests $\Lambda_i$ in ratio form, we verify Definition 2.1 for their combination $\sup_i w_i \Lambda_i$:

$$P\left(\sup_i w_i \Lambda_i(X) \geq \frac{1}{\epsilon}\right) \leq \sum_i P\left(\Lambda_i(X) \geq \frac{1}{w_i \epsilon}\right) \leq \sum_i w_i \epsilon \leq \epsilon.$$

Next, given e-tests $\Lambda_i$ in ratio form, we verify Definition 2.1 for their combination $\sum_i w_i \Lambda_i$:

$$\mathbf{E}\left(\sum_i w_i \Lambda_i(X)\right) = \sum_i w_i \, \mathbf{E}(\Lambda_i(X)) \leq \sum_i w_i \leq 1.$$

Transforming the tests into probability and log form yields the remaining results. □

**Example 9** (Two-tailed p-value)**.** *By combining the one-tailed p-values (11) for the feature statistics $\tau$ and $-\tau$, each with weight 0.5, we obtain the two-tailed p-value*

$$\begin{aligned}\Lambda_{\pm\tau}(x) &:= \min\{\Lambda_\tau(x)/0.5,\ \Lambda_{-\tau}(x)/0.5)\} \\ &= 2\min\{\Pr(\tau(X) \geq \tau(x)),\ \Pr(\tau(X) \leq \tau(x)))\}.\end{aligned}$$

From now on, we express tests in log form except where stated otherwise. By Lemma 7.1, a combination test exceeds each of its component tests, up to the additive regularization term $\log w_i$ which does not depend on the sample $x$. Thus, by committing to a combination test prior to observing $x$, we effectively postpone the search over the component tests until after observing $x$.

Unfortunately, the regularization term does depend on $i$, and becomes arbitrarily large as the number of tests becomes large or infinite. If we want the combination test to be competitive with its most promising component tests, it becomes important to choose the weights well [Wasserman and Roeder, 2006]. The problem of choosing $w_i$ is analogous to that of choosing Bayesian priors, and is philosophically challenged by formal impossibility results in the theory of inductive inference [Adam et al., 2019, Wolpert, 2023]. Fortunately, computability poses a useful constraint on the set of permissible tests as well as priors [Rathmanner and Hutter, 2011]. It is suggestive that every computable sequence of weights $w_i$ can be turned into a computable binary code with lengths $\lceil -\log w_i \rceil$ [Cover and Thomas, 2006]. Thus, minimizing the regularization penalty amounts to finding the shortest computable encoding for $i$.

## 8 CONSTRUCTING THE UNIVERSAL E-TEST

To construct a universal e-test, we first need a few concepts from computability theory. We say a function $f : \{0,1\}^* \to \mathbb{R}$ is lower (or upper) semicomputable if it can be computably approximated from below (or above, respectively). For example, the Kolmogorov complexity $K$ is upper, but not lower, semicomputable: by running all programs in parallel, we gradually find shorter programs that output $x$. We say $f$ is computable if it is both lower and upper and semicomputable.

We say a p-test or e-test is semicomputable, if it is lower semicomputable when expressed as a function in either ratio or log form. A semicomputable p-test is called a Martin-Löf test [Martin-Löf, 1966], while a semicomputable e-test is called a Levin test [Levin, 1976]. Intuitively, semicomputable tests detect more anomalies as computation time increases. By restricting attention to semicomputable tests, it becomes possible to create universal combinations of them.

We return to the problem of optimizing the weights in Lemma 7.1: we would like to dominate not only each of the individual component tests, but also each of the combination tests obtainable by some computable sequence of weights $w_i$. Note that the constraints on $w_i$ amount to specifying a discrete sub-probability distribution. Among all lower semicomputable discrete sub-probability distributions,

$$m_i := 2^{-K(i|P^*)} \tag{13}$$

is *universal* in the sense that for all alternatives $w$ and all $i$,

$$\log m_i \overset{+}{\geq} \log w_i - K(w \mid P^*).$$

This also holds when $P$ is replaced by any other piece of prior knowledge.

Now, applying Lemma 7.1 with the universal weights (13), to any sequence of p-tests $\lambda_i$ in log form, yields their universal combination

$$\lambda(x) := \sup_i \{\lambda_i(x) - K(i \mid P^*)\}.$$

In the case where $\{\lambda_i\}_{i=1}^{\infty}$ is a computable enumeration of some Martin-Löf tests, $\lambda$ is itself a Martin-Löf test. It follows that if $\{\lambda_i\}_{i=1}^{\infty}$ enumerates *all* Martin-Löf tests, then $\lambda$ is universal among them. Every program $p$ that outputs $i$ with access to $P$, determines the feature $\lambda_p := \lambda_i$. Moreover, it satisfies $|p| \geq K(i \mid P^*)$, with equality for the shortest such program. Therefore, we can rewrite the universal Martin-Löf test as

$$\rho(x) := \sup_p \{\lambda_p(x) - |p|\}. \tag{14}$$

With e-tests, the situation is even nicer because Equation (12) provides a one-to-one correspondence between e-tests $\Lambda_Q$ and sub-probability distributions $Q$. Moreover, if we assume $P$ to be computable, an e-test for it is semicomputable (i.e., is a Levin test) iff its corresponding $Q$ is lower semicomputable. Applying Lemma 7.1 with the universal weights (13), to the likelihood ratios $\Lambda_{Q_i}$, yields their universal combination

$$\Lambda(x) := \sum_i m_i \Lambda_{Q_i}(x) = \frac{\sum_i m_i Q_i(x)}{P(x)}.$$

In the case where $\{Q_i\}_{i=1}^{\infty}$ enumerates all lower semicomputable sub-probability measures, Theorem 4.3.3 in Li and Vitányi [2019] implies $\sum_i m_i Q_i(x) \overset{\times}{=} m(x \mid P^*)$, where $\overset{\times}{=}$ indicate equality up to a multiplicative term. Switching to log form, we obtain the universal Levin test

$$\delta(x) := \log \frac{m(x \mid P^*)}{P(x)} = \log \frac{1}{P(x)} - K(x \mid P^*). \tag{15}$$

By the Kraft inequality, it is an e-test in log form. Note that (15) is the difference between a Shannon code length and a shortest program length for $x$. Intuitively, $\delta$ is high whenever the Shannon code derived from $P$ is inefficient at compressing $x$, indicating that $x$ possesses regularities that are atypical of $P$.

The algorithmic information theory literature uses the term *randomness deficiency* to refer to either the universal Martin-Löf test (14) or the universal Levin test (15). Keeping in mind conversions such as (10) between the two types of tests, we will develop our theory in terms of the latter.

**Example 10** (Feature Selection through Universal Tests)**.** *Consider an infinite sequence of basis feature functions* $(f_i)$*, where* $\log P$ *is expressed as a linear combination of finitely many features:* $\log P := \sum_i \alpha_i f_i$*. When an observation* $x$ *exhibits atypical feature values under* $P$*, it would typically have a substantially higher likelihood under some modified linear combination* $\log \tilde{P} := \sum_i \tilde{\alpha}_i f_i$*.*

*If we assume the description of* $P$ *is given in terms of the coefficient vector* $\alpha$*, we can bound the Kolmogorov complexity:*

$$K(x \mid P^*) \overset{+}{\leq} K(\tilde{\alpha} \mid \alpha) + \log \frac{1}{\tilde{P}(x)},$$

*This leads to a lower bound on the Levin test:*

$$\delta_P(x) = -\log P(x) - K(x \mid P^*) \overset{+}{\geq} \log \frac{\tilde{P}(x)}{P(x)} - K(\tilde{\alpha} \mid \alpha).$$

*The bound becomes large when the improvement in likelihood (first term) substantially exceeds the complexity cost* $K(\tilde{\alpha} \mid \alpha)$ *required to modify the coefficients. This demonstrates how the universal tests naturally detect feature-based anomalies: when certain feature statistics of* $x$ *are unusual under* $P$*, there exists an alternative distribution* $\tilde{P}$ *that better explains these feature values, leading to a high value of* $\delta_P(x)$*.*

## 9 DECOMPOSITION OF RANDOMNESS DEFICIENCY

**Lemma 3.1.** *(Decomposition of randomness deficiency for a cause-effect pair) For any two random variables* $X \to Y$ *(i.e.,* $X$ *being the cause of* $Y$*), and for individual observations* $x$ *and* $y$*, the following equality holds under Postulate 2.4 for the DAG in Figure 1:*

$$\delta(x, y) \overset{+}{=} \delta(x) + \delta(y \mid x)$$

*Proof.* By Eqs. (5) and (6), we have:

$$\begin{aligned}
\delta(x, y) &\overset{+}{=} -\log P(x, y) - K(x, y \mid (P_{X,Y})^*) \\
&\overset{+}{=} -\log P(x, y) - K(x \mid (P_{X,Y})^*) - K(y \mid (x, P_{X,Y})^*), \\
\delta(x) + \delta(y \mid x) &\overset{+}{=} -\log P(x, y) - K(x \mid (P_X)^*) - K(y \mid (x, P_{Y|X})^*).
\end{aligned}$$

To complete the proof, it suffices to establish the following:

$$(1) \quad K(x \mid (P_{X,Y})^*) \overset{+}{=} K(x \mid (P_X)^*) \qquad \text{and} \qquad (2) \quad K(y \mid (x, P_{X,Y})^*) \overset{+}{=} K(y \mid (x, P_{Y|X})^*)$$

**Proof of (1):** Applying Lemma 2.5 to our bivariate case yields $x \perp\!\!\!\perp P_{Y|X} \mid (P_X)^*$, meaning that $P_{Y|X}$ becomes irrelevant when predicting $x$ from a shortest program for $P_X$. Hence,

$$K(x \mid (P_{X,Y})^*) \overset{+}{=} K(x \mid (P_{Y|X}, P_X)^*) \overset{+}{=} K(x \mid (P_X)^*).$$

**Proof of (2):** Again applying Lemma 2.5, $y \perp\!\!\!\perp P_X \mid (x, P_{Y|X})^*$, meaning that $P_X$ becomes irrelevant when predicting $y$ from a shortest program for $x$ and $P_{Y|X}$. Hence,

$$K(y \mid (x, P_{X,Y})^*) \overset{+}{=} K(y \mid (x, P_{Y|X}, P_X)^*) \overset{+}{=} K(y \mid (x, P_{Y|X})^*).$$

□

**Theorem 3.2** (Decomposition of multivariate joint randomness deficiency)**.** *Let the set of strings* $x_1, x_2, \ldots, x_n$ *be causally connected by a directed acyclic graph* $G$*, so that the causal Markov condition holds for* $G^m$*. Then, the joint randomness deficiency of* $x_1, x_2, \ldots, x_n$ *decomposes into a sum of conditional randomness deficiencies of the mechanisms:*

$$\delta(x_1, \ldots, x_n) \overset{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j),$$

*where* $\delta(x_j \mid \mathrm{pa}_j)$ *denotes the randomness deficiency of* $x_j$ *given its parents.*

*Proof.* We prove this theorem by induction on $n$.

**Base case:** When $n = 1$, we have $\mathrm{pa}_1 = \emptyset$, so the claim is trivial.

**Inductive Hypothesis:** Assume that for any sequence of $n-1$ strings $x_1, x_2, \ldots, x_{n-1}$,

$$\delta(x_1, \ldots, x_{n-1}) \stackrel{+}{=} \sum_{j=1}^{n-1} \delta(x_j \mid \mathrm{pa}_j).$$

**Inductive Step:** Now we must prove the statement holds for $n$ strings. The statistical Markov condition yields

$$P(x_n \mid x_1, \ldots, x_{n-1}) = P(x_n \mid \mathrm{pa}_n).$$

Meanwhile, the algorithmic Markov condition gives us Lemma 2.5, which implies

$$x_n \perp\!\!\!\perp x_1, \ldots, x_{n-1} \mid (\mathrm{pa}_n, P_{X_n|\mathrm{PA}_n})^*.$$

Together with $P_{X_n|X_1,\ldots,X_{n-1}} = P_{X_n|\mathrm{PA}_n}$, this yields

$$K(x_n \mid (x_1, \ldots, x_{n-1}, P_{X_n|X_1,\ldots,X_{n-1}})^*) \stackrel{+}{=} K(x_n \mid (\mathrm{pa}_n, P_{X_n|\mathrm{PA}_n})^*).$$

Putting these results together,

$$\begin{aligned}
\delta(x_n \mid x_1, \ldots, x_{n-1}) &:= -\log P(x_n \mid x_1, \ldots, x_{n-1}) - K(x_n \mid (x_1, \ldots, x_{n-1}, P_{X_n|X_1,\ldots,X_{n-1}})^*) \\
&\stackrel{+}{=} -\log P(x_n \mid \mathrm{pa}_n) - K(x_n \mid (\mathrm{pa}_n, P_{X_n|\mathrm{PA}_n})^*) \\
&= \delta(x_n \mid \mathrm{pa}_n).
\end{aligned}$$

Finally, we combine the inductive hypothesis with Lemma 3.1, substituting $x := (x_1, \ldots, x_{n-1})$ and $y := x_n$:

$$\begin{aligned}
\delta(x_1, \ldots, x_n) &\stackrel{+}{=} \delta(x_1, \ldots, x_{n-1}) + \delta(x_n \mid x_1, \ldots, x_{n-1}) \\
&\stackrel{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j).
\end{aligned}$$

This completes the inductive step and the proof. □

**Corollary 3.3** (Weak outliers do not cause stronger ones)**.** *If there is a unique root cause $j \in \{1, \ldots, n\}$, in the sense that*

$$\delta(x_i \mid \mathrm{pa}_i) \stackrel{+}{=} 0 \quad \textit{for } i \neq j,$$

*and the conditions of Theorem 3.2 are met, then*

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_j \mid \mathrm{pa}_j) \quad \forall i \in \{1, \ldots, n\}. \tag{7}$$

*Proof.* Corollary 4.1.11 of Gács [2021] states that $\delta$ is non-increasing under marginalization. That is, for all $i$,

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_1, \ldots, x_n).$$

Decomposing the joint randomness deficiency of $x_1, \ldots, x_n$ according to Theorem 3.2,

$$\begin{aligned}
\delta(x_1, \ldots, x_n) &\stackrel{+}{=} \sum_{i=1}^{n} \delta(x_i \mid \mathrm{pa}_i) \\
&= \delta(x_j \mid \mathrm{pa}_j) + \sum_{i \neq j,\, i=1}^{n} \delta(x_i \mid \mathrm{pa}_i) \\
&\stackrel{+}{=} \delta(x_j \mid \mathrm{pa}_j).
\end{aligned}$$

Putting these results together yields

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_j \mid \mathrm{pa}_j).$$

□

## 10 NON-INCREASINGNESS OF MAHALANOBIS DISTANCE

Let $\mathbf{x} \in \mathbb{R}^n$, and let $Q\mathbf{x}$ denote the projection of $\mathbf{x}$ onto the subspace spanned by the variables $X_{i_1}, \ldots, X_{i_k}$, where $k < n$. We express the covariance matrix $\Sigma_X$ in block matrix form as

$$\Sigma_X = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with index 1 referring to the variables $i_1, \ldots, i_k$ and 2 to the remaining variables. Using a known formula for inversion of $2 \times 2$ block matrices (see Proposition 3.9.7 in Bernstein [2009]), we obtain

$$\Sigma_X^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}A\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}A \\ -A\Sigma_{21}\Sigma_{11}^{-1} & A \end{pmatrix},$$

with $A := (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$. We now compute the difference between the squared Mahalanobis distances on $\mathbb{R}^n$ and $\mathbb{R}^k$:

$$\mathbf{x}^T\Sigma_X^{-1}\mathbf{x} - (Q\mathbf{x})^T\Sigma_{11}^{-1}Q\mathbf{x} = \mathbf{x}^T C\mathbf{x}, \tag{16}$$

with

$$C := \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12}A\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}A \\ -A\Sigma_{21}\Sigma_{11}^{-1} & A \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12} & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} A & A \\ A & A \end{pmatrix} \begin{pmatrix} \Sigma_{21}\Sigma_{11}^{-1} & 0 \\ 0 & -1 \end{pmatrix}.$$

Since $A$ is positive semi-definite, and the rightmost matrix in the product is the transpose of the leftmost matrix, it follows that $C$ is also positive semi-definite. Hence, (16) is non-negative.