Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice

David Watson [*1] Limor Gultchin [*2 3] Ankur Taly [4] Luciano Floridi [5 3]

Abstract

Necessity and sufficiency are the building blocks of all successful explanations. Yet despite their importance, these notions have been conceptually underdeveloped and inconsistently applied in explainable artificial intelligence (XAI), a fast-growing research area that is so far lacking in firm theoretical foundations. Building on work in logic, probability, and causality, we establish the central role of necessity and sufficiency in XAI, unifying seemingly disparate methods in a single formal framework. We provide a sound and complete algorithm for computing explanatory factors with respect to a given context, and demonstrate its flexibility and competitive performance against state of the art alternatives on various tasks.

1. Introduction

Machine learning algorithms are increasingly used in a variety of high-stakes domains, from credit scoring to medical diagnosis. However, many such methods are *opaque*, in that humans cannot understand the reasoning behind particular predictions. Post-hoc, model-agnostic local explanation tools (e.g., feature attributions, rule lists, and counterfactuals) are at the forefront of a fast-growing area of research variously referred to as *interpretable machine learning* or *explainable artificial intelligence* (XAI).

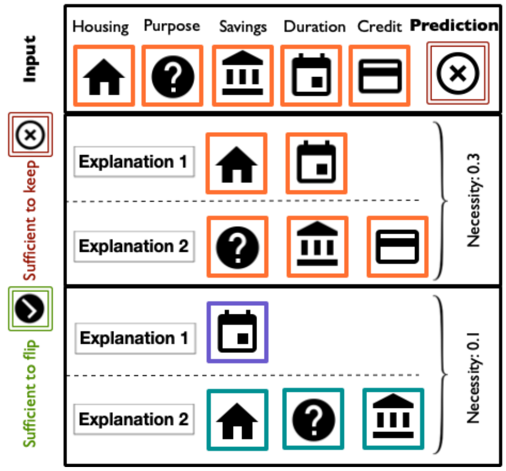

Figure 1. We describe minimal sufficient factors (here, sets of features) for a given input (top row), with the aim of preserving or flipping the original prediction. We report a sufficiency score for each set and a cumulative necessity score for all sets, indicating the proportion of paths towards the outcome that are covered by the explanation. Feature colors indicate source of feature values (input or reference).

Many authors have pointed out the inconsistencies between popular XAI tools, raising questions as to which method is more reliable in particular cases (Mothilal et al., 2020a; Ramon et al., 2020; Fernández-Loría et al., 2020). Theoretical foundations have proven elusive in this area, perhaps due to the perceived subjectivity inherent to notions such as "intelligible" and "relevant" (Watson and Floridi, 2020). Practitioners often seek refuge in the axiomatic guarantees of Shapley values, which have become the de facto standard in many XAI applications, due in no small part to their attractive theoretical properties (Bhatt et al., 2020). However, ambiguities regarding the underlying assumptions of the method (Kumar et al., 2020) and the recent proliferation of mutually incompatible implementations (Sundararajan and Najmi, 2019; Merrick and Taly, 2020) have complicated this picture. Despite the abundance of alternative XAI tools (Molnar, 2021), a dearth of theory persists. This has led some to conclude that the goals of XAI are underspecified (Lipton, 2018), and even that post-hoc methods do more harm than good (Rudin, 2019)

We argue that this lacuna at the heart of XAI should be filled by a return to fundamentals – specifically, to *necessity* and *sufficiency*. As the building blocks of all successful explanations, these dual concepts deserve a privileged position in the theory and practice of XAI. We operationalize this insight with a unified framework (Sect. 3) that reveals

[*] Equal contribution [1] Department of Statistical Science, University College London, London, United Kingdom [2] Department of Computer Science, University of Oxford, Oxford, United Kingdom [3] The Alan Turing Institute, London, United Kingdom [4] Google Inc., Mountain View, USA [5] Oxford Internet Institute, University of Oxford, Oxford, United Kingdom. Correspondence to: David Watson <david.watson@ucl.ac.uk>, Limor Gultchin <limor.gultchin@gmail.com>.

Preprint.

unexpected affinities between various XAI tools and probabilities of causation (Sect. 4). We proceed to implement a novel procedure for computing model explanations that improves upon the state of the art in various quantitative and qualitative comparisons (Sect. 5).

We make three main contributions. (1) We present a formal framework for XAI that unifies several popular approaches, including feature attributions, rule lists, and counterfactuals. (2) We introduce novel measures of necessity and sufficiency that can be computed for any feature subset. The method enables users to incorporate domain knowledge, search various subspaces, and select a utility-maximizing explanation. (3) We present a sound and complete algorithm for identifying explanatory factors, and illustrate its performance on a range of tasks. The method compares favorably to state of the art XAI tools.

2. Necessity and Sufficiency

Necessity and sufficiency have a long philosophical tradition (Halpern and Pearl, 2005b; Brennan, 2017; Miller, 2019), including logical, probabilistic, and causal variants. In predicate logic, we say that x is a sufficient condition for y iff $(\forall x \forall y)\ x \rightarrow y$, and x is a necessary condition for y iff $(\forall x \forall y)\ y \rightarrow x$. By the law of contraposition, both definitions have alternative formulations, whereby sufficiency may be written as $(\forall x \forall y)\ \neg y \rightarrow \neg x$, and necessity as $(\forall x \forall y)\ \neg x \rightarrow \neg y$.

These logical formulae suggest probabilistic versions that relax the universal quantifiers, measuring x's sufficiency for y by $P(y|x)$ and x's necessity for y by $P(x|y)$. Because there is no probabilistic law of contraposition, these quantities are generally uninformative w.r.t. $P(\neg x|\neg y)$ and $P(\neg y|\neg x)$, which may be of independent interest. We revisit the distinction between inverse and converse formulations in Sect. 4.

These definitions struggle to track our intuitions when we consider causal explanations (Pearl, 2000; Tian and Pearl, 2000). It may make sense to say in logic that if x is a necessary condition for y, then y is a sufficient condition for x. It does not follow that if x is a necessary *cause* of y, then y is a sufficient *cause* of x. We may amend both concepts using *counterfactual probabilities* – e.g., the probability that Alice would still have a headache if she had not taken an aspirin, given that she does not have a headache and did take an aspirin. Let $P(y_x|x', y')$ denote such a quantity, to be read as "the probability that Y would equal y under an intervention that sets X to x, given that we observe $X = x'$ and $Y = y'$." Then, according to Pearl (2000), the probability that x is a sufficient cause of y is given by $\texttt{suf}(x, y) := P(y_x|x', y')$, and the probability that x is a necessary cause of y is given by $\texttt{nec}(x, y) := P(y'_{x'}|x, y)$.

Analysis becomes more difficult in higher dimensions, where contingent events or potential confounders may block or unblock causal pathways. Halpern (2016) provides various criteria to distinguish different notions of "actual causality", as well as "but-for" (similar to necessary) and sufficient causes. The definitions are too long and varied to recapitulate here, although we incorporate their most essential elements into our proposals below. Neither Pearl (2000) nor Halpern (2016) implement their ideas on any real-world datasets, which raise very different challenges than Boolean thought experiments. Operationalizing these theories in a practical method is one of our primary contributions.

Necessity and sufficiency have begun to receive explicit attention in the XAI literature. Ribeiro et al. (2018a) propose a bandit procedure for identifying a minimal set of Boolean conditions that entails a predictive outcome (more on this in Sect. 4). Dhurandhar et al. (2018) propose an autoencoder for learning pertinent negatives and positives, i.e. features whose presence or absence is decisive for a given label, while Zhang et al. (2018) develop a technique for generating symbolic corrections to alter model outputs. Both methods are optimized for neural networks, unlike the model-agnostic approaches we prioritize here.

In a recent paper, Mothilal et al. (2020a) build on probabilistic concepts of necessity and sufficiency to critique popular XAI tools, proposing a new feature attribution method with some purported advantages. We improve on this work in several respects. First, we evaluate feature *subsets*, rather than individual predictors, allowing us to detect potential interaction effects. Second, we consider not only cases in which features are mutually independent, but also more realistic settings in which dependencies between predictors are present. Our formal results clarify the relationship between existing XAI methods and probabilities of causation, while our empirical results demonstrate their applicability to a wide array of tasks and datasets.

3. A Unifying Framework

We propose a unifying framework that highlights the role of necessity and sufficiency in XAI. Its constituent elements are described below.

Target function. Post-hoc explainability methods assume access to a target function $f : \mathcal{X} \mapsto \mathcal{Y}$, i.e. the model whose prediction(s) we seek to explain. For simplicity, we restrict attention to the binary setting, with $Y \in \{0, 1\}$. Multi-class extensions are straightforward, while continuous outcomes may be accommodated via discretization.

Context. The context $\mathcal{D}$ is a probability distribution over which we quantify sufficiency and necessity. Contexts may be constructed in various ways but always consist of at least some input (point or space) and reference (point or space). For instance, we may want to compare $\boldsymbol{x}_i$ with all other samples, or else just those perturbed along one or two axes, perhaps based on some conditioning event(s).

In addition to predictors and outcomes, we optionally include information exogenous to f. For instance, if any events were conditioned upon to generate a given reference sample, this information may be recorded among a set of auxiliary variables $\boldsymbol{W}$. Other examples of potential auxiliaries include metadata or engineered features such as those learned via neural embeddings. This augmentation allows us to evaluate the necessity and sufficiency of factors beyond those found in $\boldsymbol{X}$. Contextual data take the form $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{W}) \sim \mathcal{D}$. The distribution may or may not encode dependencies between (elements of) $\boldsymbol{X}$ and (elements of) $\boldsymbol{W}$. We extend the target function to augmented inputs by defining $f(\boldsymbol{z}) := f(\boldsymbol{x})$.

Factors. Factors pick out the properties whose necessity and sufficiency we wish to quantify. Formally, a factor $c : \mathcal{Z} \mapsto \{0, 1\}$ indicates whether its argument satisfies some criteria with respect to predictors or auxiliaries. For instance, if $\boldsymbol{x}$ is an input to a credit lending model, and $\boldsymbol{w}$ contains information about the subspace from which data were sampled, then a factor could be $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\mathsf{gender} =$ "female"$] \wedge \boldsymbol{w}[do(\mathsf{income} > \$50\text{k})]]$, i.e. checking if $\boldsymbol{z}$ is female and drawn from a context in which an intervention fixes income at greater than $50k. We use the term "factor" as opposed to "condition" or "cause" to suggest an inclusive set of criteria that may apply to predictors $\boldsymbol{x}$ and/or auxiliaries $\boldsymbol{w}$. Such criteria are always observational w.r.t. $\boldsymbol{z}$ but may be interventional or counterfactual w.r.t. $\boldsymbol{x}$. We assume a finite space of factors $\mathcal{C}$.

Partial order. When multiple factors pass a given necessity or sufficiency threshold, users will tend to prefer some over others. For instance, factors with fewer conditions are generally preferable to those with more, all else being equal; factors that change a variable by one unit as opposed to two are preferable, and so on. Rather than formalize this preference in terms of a distance metric, which unnecessarily constrains the solution space, we treat the partial ordering as primitive and require only that it be complete and transitive. This covers not just distance-based measures but also more idiosyncratic orderings that are unique to individual agents. Ordinal preferences may be represented by cardinal utility functions under modest assumptions (see, e.g., (von Neumann and Morgenstern, 1944)).

We are now ready to formally specify our framework.

Definition 1 (Basis)**.** A *basis* for computing necessary and sufficient factors for model predictions is a tuple $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle$, where f is a target function, $\mathcal{D}$ is a context, $\mathcal{C}$ is a set of factors, and $\preceq$ is a partial ordering on $\mathcal{C}$.

3.1. Explanatory Measures

For some fixed basis $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle$, we define the following measures of sufficiency and necessity, with probability taken over $\mathcal{D}$.

Definition 2 (Probability of Sufficiency)**.** The probability that c is a sufficient factor for outcome y is given by:

$$PS(c, y) := P(f(\boldsymbol{z}) = y \mid c(\boldsymbol{z}) = 1).$$

The probability that factor set $C = \{c_1, \ldots, c_k\}$ is sufficient for y is given by:

$$PS(C, y) := P(f(\boldsymbol{z}) = y \mid \sum_{i=1}^{k} c_i(\boldsymbol{z}) \geq 1).$$

Definition 3 (Probability of Necessity)**.** The probability that c is a necessary factor for outcome y is given by:

$$PN(c, y) := P(c(\boldsymbol{z}) = 1 \mid f(\boldsymbol{z}) = y).$$

The probability that factor set $C = \{c_1, \ldots, c_k\}$ is necessary for y is given by:

$$PN(C, y) := P(\sum_{i=1}^{k} c_i(\boldsymbol{z}) \geq 1 \mid f(\boldsymbol{z}) = y).$$

Remark 1. These probabilities can be likened to the "precision" (positive predictive value) and "recall" (true positive rate) of a (hypothetical) classifier that predicts whether $f(\boldsymbol{z}) = y$ based on whether $c(\boldsymbol{z}) = 1$. By examining the confusion matrix of this classifier, one can define other related quantities, e.g. the true negative rate $P(c(\boldsymbol{z}) = 0 | f(\boldsymbol{z}) \neq y)$ and the negative predictive value $P(f(\boldsymbol{z}) \neq y | c(\boldsymbol{z}) = 0)$, which are contrapositive transformations of our proposed measures. We can recover these values exactly via $PS(1 - c, 1 - y)$ and $PN(1 - c, 1 - y)$, respectively.

3.2. Minimal Sufficient Factors

We introduce Local Explanations via Necessity and Sufficiency (LENS), a procedure for computing explanatory factors with respect to a given basis $\mathcal{B}$ and threshold parameter τ (see Alg. 1). First, we calculate a factor's probability of sufficiency (see probSufficiency) by drawing n samples from $\mathcal{D}$ and taking the maximum likelihood estimate $\hat{PS}(c, y)$. Next, we sort the space of factors w.r.t. $\preceq$ in search of those that are τ-minimal.

Definition 4 (τ-minimality)**.** We say that c is τ-minimal iff (i) $PS(c, y) \geq \tau$ and (ii) there exists no factor c' such that $PS(c', y) \geq \tau$ and $c' \prec c$.

This determination may be assisted with statistical inference procedures (see Thm. 2 below), although we assume access to population proportions in our experiments.

Since a factor is necessary to the extent that it covers all possible pathways towards a given outcome, our next step is to span the τ-minimal factors and compute their cumulative PN (see probNecessity). As a minimal factor c stands for all c' such that $c \preceq c'$, in reporting probability of necessity, we expand C to its upward closure. As Thms. 1 and 2 state, this procedure has certain desirable properties (see Appendix A for all proofs).

Theorem 1. With oracle estimates of $PS(c, y)$ for all $c \in \mathcal{C}$, Alg. 1 is sound and complete. That is, for any C returned by Alg. 1 and all $c \in \mathcal{C}$, c is τ-minimal iff $c \in C$.

Theorem 2. With sample estimates of $\hat{PS}(c, y)$ for all $c \in \mathcal{C}$, Alg. 1 is uniformly most powerful. That is, Alg. 1 identifies the most τ-minimal factors of any method with fixed type I error α.

Remark 2. We take it that the main quantity of interest in most applications is sufficiency, be it for the original or alternative outcome, and therefore define τ-minimality w.r.t. sufficient (rather than necessary) factors. However, necessity serves an important role in tuning τ, as there is an inherent trade-off between the parameters. More factors are excluded at higher values of τ, thereby inducing lower cumulative PN; more factors are included at lower values of τ, thereby inducing higher cumulative PN. See Appendix B for a discussion.

Algorithm 1 LENS

```
Input: $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle, \tau$

sample $\hat{D} := \{\boldsymbol{z}_i\}_{i=1}^n \sim \mathcal{D}$
n(c&y) = $\sum_{i=1}^n \mathbb{1}[c(\boldsymbol{z}_i) = 1 \wedge f(\boldsymbol{z}_i) = y]$

function probSufficiency($c$, $y$)
    n(c) = $\sum_{i=1}^n \mathbb{1}[c(\boldsymbol{z}_i) = 1]$
    return n(c&y) / n(c)

function probNecessity($C$, $y$, upward_closure_flag)
    if upward_closure_flag then
        $C = \{c \mid c \in \mathcal{C} \wedge \exists\, c' \in C : c' \preceq c\}$
    end if
    n(y) = $\sum_{i=1}^n \mathbb{1}[f(\boldsymbol{z}_i) = y]$
    return n(c&y) / n(y)

function minimalSufficientFactors($y$, $\tau$)
    sorted_factors = topological_sort($\mathcal{C}, \preceq$)
    cands = []
    for $c$ in sorted_factors do
        if $\exists(c', \_) \in$ cands : $c' \preceq c$ then
            continue
        end if
        ps = probSufficiency($c$, $y$)
        if ps $\geq \tau$ then
            cands.append($c$, ps)
        end if
    end for
    cum_pn=probNecessity($\{c|(c, \_) \in$ cands$\}, y$, True)
    return cands, cum_pn
```

4. Encoding Existing Measures

Explanatory measures can be shown to play a central role in many seemingly unrelated XAI tools, albeit under different assumptions about the basis tuple $\mathcal{B}$. In this section, we relate our framework to a number of existing methods.

Feature attributions. Several popular feature attribution algorithms are based on Shapley values (Shapley, 1953), which decompose the predictions of any target function as a sum of weights over d input features:

$$f(\boldsymbol{x}_i) = \phi_0 + \sum_{j=1}^{d} \phi_j, \tag{1}$$

where ϕ_0 represents a baseline expectation and ϕ_j the weight assigned to X_j at point $\boldsymbol{x}_i$. Let $v : 2^d \mapsto \mathbb{R}$ be a value function such that $v(S)$ is the payoff associated with feature subset $S \subseteq [d]$ and $v(\{\emptyset\}) = 0$. Define the complement $R = [d] \backslash S$ such that we may rewrite any $\boldsymbol{x}_i$ as a pair of subvectors, $(\boldsymbol{x}_i^S, \boldsymbol{x}_i^R)$. Payoffs are given by:

$$v(S) = \mathbb{E}[f(\boldsymbol{x}_i^S, \boldsymbol{X}^R)], \tag{2}$$

although this introduces some ambiguity regarding the reference distribution for $\boldsymbol{X}^R$ (more on this below). The Shapley value ϕ_j is then j's average marginal contribution to all subsets that exclude it:

$$\phi_j = \sum_{S \subseteq [d] \backslash \{j\}} \frac{|S|!(d - |S| - 1)!}{d!} v(S \cup \{j\}) - v(S). \tag{3}$$

It can be shown that this is the unique solution to the attribution problem that satisfies certain desirable properties, including efficiency, linearity, sensitivity, and symmetry.

Reformulating this in our framework, we find that the value function v is a sufficiency measure. To see this, let each $\boldsymbol{z} \sim \mathcal{D}$ be a sample in which a random subset of variables S are held at their original values, while remaining features R are drawn from a fixed distribution $\mathcal{D}(\cdot|S)$.[1]

Proposition 1. Let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x} \subseteq \boldsymbol{z}$ was constructed by holding $\boldsymbol{x}^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\cdot|S)$.

[1] The diversity of Shapley value algorithms is largely due to variation in how this distribution is defined. Popular choices include the marginal $P(\boldsymbol{X}^R)$ (Lundberg and Lee, 2017); conditional $P(\boldsymbol{X}^R|\boldsymbol{x}^S)$ (Aas et al., 2019); and interventional $P(\boldsymbol{X}^R|do(\boldsymbol{x}^S))$ (Heskes et al., 2020) distributions.

Then $v(S) = PS(c_S, y)$.

Thus, the Shapley value ϕ_j measures X_j's average marginal increase to the sufficiency of a random feature subset. The advantage of our method is that, by focusing on particular subsets instead of weighting them all equally, we disregard irrelevant permutations and home in on just those that meet a τ-minimality criterion. Kumar et al. (2020) observe that, "since there is no standard procedure for converting Shapley values into a statement about a model's behavior, developers rely on their own mental model of what the values represent" (p. 8). By contrast, necessary and sufficient factors are more transparent and informative, offering a direct path to what Shapley values crudely summarize.

Rule lists. Rule lists are sequences of if-then statements that describe a hyperrectangle in feature space, creating partitions that can be visualized as decision or regression trees. Rule lists have long been popular in XAI. While early work in this area tended to focus on global methods (Friedman and Popescu, 2008; Letham et al., 2015), more recent efforts have prioritized local explanation tasks (Lakkaraju et al., 2019; Sokol and Flach, 2020).

We focus in particular on the Anchors algorithm (Ribeiro et al., 2018a), which learns a set of Boolean conditions A (the eponymous "anchors") such that $A(\boldsymbol{x}_i) = 1$ and

$$P_{\mathcal{D}_{(\boldsymbol{x}|A)}}(f(\boldsymbol{x}_i) = f(\boldsymbol{x})) \geq \tau. \tag{4}$$

The lhs of Eq. 4 is termed the *precision*, prec(A), and probability is taken over a synthetic distribution in which the conditions in A hold while other features are perturbed. Once τ is fixed, the goal is to maximize *coverage*, formally defined as $\mathbb{E}[A(\boldsymbol{x}) = 1]$, i.e. the proportion of datapoints to which the anchor applies.

The formal similarities between Eq. 4 and Def. 2 are immediately apparent, and the authors themselves acknowledge that Anchors are intended to provide "sufficient conditions" for model predictions.

Proposition 2. Let $c_A(\boldsymbol{z}) = 1$ iff $A(\boldsymbol{x}) = 1$. Then $\text{prec}(A) = PS(c_A, y)$.

While Anchors outputs just a single explanation, our method generates a ranked list of candidates, thereby offering a more comprehensive view of model behavior. Moreover, our necessity measure adds another dimension of explanatory information entirely lacking in Anchors.

Counterfactuals. Counterfactual explanations identify one or several nearest neighbors with different outcomes, e.g. all datapoints $\boldsymbol{x}$ within an ϵ-ball of $\boldsymbol{x}_i$ such that labels $f(\boldsymbol{x})$ and $f(\boldsymbol{x}_i)$ differ (for classification) or $f(\boldsymbol{x}) > f(\boldsymbol{x}_i) + \delta$ (for regression).[2] The optimization problem is:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \text{CF}(\boldsymbol{x}_i)}{\text{argmin}} \; cost(\boldsymbol{x}_i, \boldsymbol{x}), \tag{5}$$

where $\text{CF}(\boldsymbol{x}_i)$ denotes a counterfactual space such that $f(\boldsymbol{x}_i) \neq f(\boldsymbol{x})$ and *cost* is a user-supplied cost function, typically equated with some distance measure. Wachter et al. (2018) recommend using generative adversarial networks to solve Eq. 5, while others have proposed alternatives designed to ensure that counterfactuals are coherent and actionable (Ustun et al., 2019; Karimi et al., 2020a; Wexler et al., 2020). As with Shapley values, the variation in these proposals is reducible to the choice of context $\mathcal{D}$.

For counterfactuals, we rewrite the objective as a search for minimal perturbations sufficient to flip an outcome.

Proposition 3. Let *cost* be a function representing $\preceq$, and let c be some factor spanning reference values. Then the counterfactual recourse objective is:

$$c^* = \underset{c \in \mathcal{C}}{\text{argmin}} \; cost(c) \;\text{ s.t. } PS(c, 1 - y) \geq \tau, \tag{6}$$

where τ denotes a decision threshold. Counterfactual outputs will then be any $\boldsymbol{z} \sim \mathcal{D}$ such that $c^*(\boldsymbol{z}) = 1$.

Probabilities of causation. Our framework can describe Pearl (2000)'s aforementioned probabilities of causation, however in this case $\mathcal{D}$ must be constructed with care.

Proposition 4. Consider the bivariate Boolean setting, as in Sect. 2. We have two counterfactual distributions: an input space $\mathcal{I}$, in which we observe x, y but intervene to set $X = x'$; and a reference space $\mathcal{R}$, in which we observe x', y' but intervene to set $X = x$. Let $\mathcal{D}$ denote a uniform mixture over both spaces, and let auxiliary variable W tag each sample with a label indicating whether it comes from the original ($W = 1$) or contrastive ($W = 0$) counterfactual space. Define $c(\boldsymbol{z}) = w$. Then we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, y')$.

In other words, we regard Pearl's notion of necessity as *sufficiency of the negated factor for the alternative outcome*. By contrast, Pearl (2000) has no analogue for our probability of necessity. This is true of any measure that defines sufficiency and necessity via inverse, rather than converse probabilities. Whereas we can recover all four explanatory measures, corresponding to the classical definitions and their contrapositive forms, definitions that prohibit conditioning on outcomes are constrained.

Remark 3. We have assumed that factors and outcomes are Boolean throughout. Results can be extended to probabilistic versions of either or both variables, so long as

[2] The term "counterfactual" is used in XAI to refer to any point with an alternative outcome, whereas in the causal literature it denotes a space characterized by incompatible conditioning events. We generally intend the former unless explicitly stated otherwise.

Experiment	Datasets	f	$\mathcal{D}$	$\mathcal{C}$	$\preceq$
Attribution comparison	`German`, `SpamAssassins`	`Extra-Trees`	R2I, I2R	Intervention targets	-
Anchors comparison: Brittle predictions	`IMDB`	`LSTM`	R2I, I2R	Intervention targets	$\preceq_{subset}$
Anchors comparison: PS and Prec	`German`	`Extra-Trees`	R2I	Intervention targets	$\preceq_{subset}$
Counterfactuals: Adverserial	`SpamAssassins`	`MLP`	R2I	Intervention targets	$\preceq_{subset}$
Counterfactuals: Recourse, DiCE comparison	`Adult`	`MLP`	I2R	Full interventions	$\preceq_{cost}$
Counterfactuals: Recourse, causal vs. non-causal	`German`	`Extra-Trees`	I2R$_{causal}$	Full interventions	$\preceq_{cost}$

Table 1. Overview of experimental settings by basis configuration.

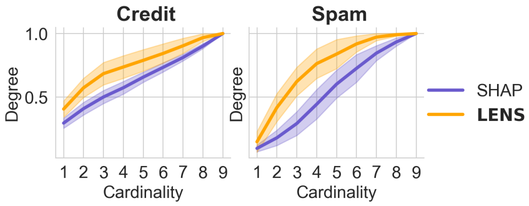

Figure 2. Comparison of SHAP scores, for top k ranked by SHAP, against the best performing LENS subset of size k in terms of $PS(c, y)$. `German` results are over 50 inputs; `SpamAssassins` results are over 25 inputs.

$c(\boldsymbol{Z}) \perp\!\!\!\perp Y \mid \boldsymbol{Z}$. This conditional independence holds whenever $\boldsymbol{W} \perp\!\!\!\perp Y \mid \boldsymbol{X}$, which is true for all cases we consider. However, we defend the Boolean assumption on the grounds that it is (a) well motivated by contrastivist epistemologies (Kahneman and Miller, 1986; Lipton, 1990; Blaauw, 2013) and (b) not especially restrictive, given that partitions of arbitrary complexity may be defined over $\boldsymbol{Z}$ and Y.

5. Experiments

In this section, we demonstrate the use of LENS on a variety of tasks and compare results with popular XAI tools, using the basis configurations detailed in Table 1. A comprehensive discussion of experimental design, including datasets and pre-processing pipelines, is left to Appendix C. Code for reproducing all results will be made publicly available.

Contexts. We consider a range of contexts $\mathcal{D}$ in our experiments. For the input-to-reference (I2R) setting, we replace input values with reference values for feature subsets S; for the reference-to-input (R2I) setting, we replace reference values with input values. We use R2I for examining sufficiency/necessity of the original model prediction, and I2R for examining sufficiency/necessity of a contrastive model prediction. We sample from the empirical data in all experiments, except in Sect. 5.3, where we assume access to a structural causal model (SCM).

Partial Orderings. We consider two types of partial orderings in our experiments. The first, $\preceq_{subset}$, evaluates subset relationships. For instance, if $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\text{gender} = \text{“female”}]]$ and $c'(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\text{gender} = \text{“female”} \wedge \text{age} \geq 40]]$, then we say that $c \preceq_{subset} c'$. The second, $c \preceq_{cost} c' := c \preceq_{subset} c' \ \wedge \ cost(c) \leq cost(c')$, adds the additional constraint that c has a lower cost than c'. The cost function could be arbitrary. Here, we consider distance measures over either the entire state space or just the intervention targets corresponding to c.

5.1. Feature Attributions

Feature attributions are often used to identify the top-k most important features for a given model outcome (Barocas et al., 2020). However, we argue that these feature sets may not be explanatory with respect to a given prediction. To show this, we compute R2I and I2R sufficiency – i.e., $PS(c, y)$ and $PS(1-c, 1-y)$, respectively – for the top-k most influential features ($k \in [1, 9]$) as identified by SHAP (Lundberg and Lee, 2017) and LENS. Fig. 2 shows results from the R2I setting for `German` credit (Dua and Graff, 2017) and `SpamAssassin` datasets (SpamAssassin, 2006). Our method attains higher PS for all cardinalities. We repeat the experiment over 50 inputs, plotting means and 95% confidence intervals for all k. Results indicate that our ranking procedure delivers more informative explanations than SHAP at any fixed degree of sparsity. Results from the I2R setting are in Appendix C.

5.2. Rule Lists

Sentiment sensitivity analysis. Next, we use LENS to study model weaknesses by considering minimal factors with high R2I and I2R sufficiency in text models. Our goal is to answer questions of the form, “What are words with/without which our model would output the original/opposite prediction for an input sentence?” For this experiment, we train an LSTM network on the `IMDB` dataset for sentiment analysis (Maas et al., 2011). If the model mislabels a sample, we investigate further; if it does not, we inspect the most explanatory factors to learn more about model behavior. For the purpose of this example, we only inspect sentences of length 10 or shorter. We provide two examples below and compare with Anchors (see Table 2).

Consider our first example: READ BOOK FORGET MOVIE is a sentence we would expect to receive a negative prediction, but our model classifies it as positive. Since we are investigating a positive prediction, our reference space is conditioned on a negative label. For this model, the clas-

Inputs		Anchors		**LENS**	
Text	Original model prediction	Suggested anchors	Precision	Sufficient R2I factors	Sufficient I2R factors
’read book forget movie’	wrongly predicted positive	[read, movie]	0.94	[read, forget, movie]	read, forget, movie
’you better choose paul verhoeven even watched’	correctly predicted negative	[choose, better, even, you, paul, verhoeven]	0.95	choose, even	better, choose, paul, even

Table 2. Example prediction given by an LSTM model trained on the IMDB dataset. We compare τ-minimal factors identified by LENS (as individual words), based on $PS(c, y)$ and $PS(1 - c, 1 - y)$, and compare to output by Anchors.

From	To	Subject	First Sentence	Last Sentence
resumevalet info resumevalet com	yyyy cv spamassassin taint org	adv put resume back work	dear candidate	professionals online network inc
jacqui devito goodroughy ananzi co za	picone linux midrange com	enlargement breakthrough zibdrzpay	recent survey conducted	increase size enter detailsto come open
rose xu email com	yyyyac idt net	adv harvest lots target email address quickly	want	advertisement persons 18yrs old

Gaming options	Feature subsets for value changes	
1	From	To
	crispin cown crispin wirex com	example com mailing... list secprog securityfocus... moderator
2	From	First Sentence
	crispin cowan crispin wirex com	scott mackenzie wrote
3	From	First Sentence
	tim one comcast net tim peters	tim

Table 3. (Top) A selection of emails from SpamAssassins, correctly identified as spam by an MLP. The goal is to find minimal perturbations that result in non-spam predictions for them. (Bottom) Minimal subsets of features-value assignments that achieve non-spam predictions with respect to the emails above.

sic UNK token receives a positive prediction. Thus we opt for an alternative, PLATE. Performing interventions on all possible combinations of words with our token, we find the conjunction of READ, FORGET, and MOVIE is a sufficient factor for a positive prediction (R2I). We also find that changing any of READ, FORGET, or MOVIE to PLATE would result in a negative prediction (I2R). Anchors, on the other hand, perturbs the data stochastically (see Appendix C), suggesting the conjunction READ AND BOOK.

Next, we investigate the sentence: YOU BETTER CHOOSE PAUL VERHOEVEN EVEN WATCHED. Since the label here is negative, we use the UNK token. We find that this prediction is brittle – a change of most single words would be sufficient to flip the outcome. Anchors, on the other hand, reports a conjunction including most words in the sentence. Taking the R2I view, we still find a more concise explanation: CHOOSE or EVEN would be enough to attain a negative prediction. These brief examples illustrate how LENS may be used to find brittle predictions across samples, search for similarities between errors, or test for model reliance on sensitive attributes (e.g., gender pronouns).

Anchors comparison. Anchors also includes a tabular variant, against which we compare LENS’s performance in terms of R2I sufficiency. We present the results of this comparison in Fig. 3, and include additional comparisons in Appendix C. We sample 100 inputs from the German dataset, and query both methods with $\tau = 0.9$ using the classifier from Sect. 5.1. Anchors satisfies a PAC bound controlled by parameter δ. At the default value $\delta = 0.1$, Anchors fails to meet the τ threshold on 14% of samples; LENS meets it on 100% of samples. This result accords with Thm. 1, and demonstrates the benefits of algorithmic soundness and completeness. Note that we also go beyond Anchors in providing multiple explanations instead of just a single output, as well as a cumulative probability measure with no analogue in their algorithm.

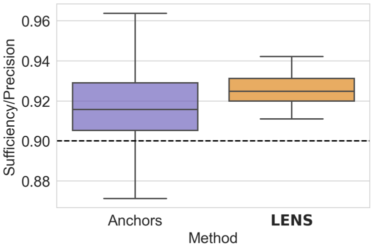

Figure 3. We compare $PS(c, y)$ against precision scores attained by the output of LENS and Anchors for examples from German. We repeat the experiment for 100 inputs, and each time consider the single example generated by Anchors against the mean $PS(c, y)$ among LENS’s candidates. Dotted line indicates $\tau = 0.9$.

5.3. Counterfactuals

Adversarial examples: spam emails. R2I sufficiency answers questions of the form, “What would be sufficient for the model to predict y?”. This is particularly valuable in cases with unfavorable outcomes y'. Inspired by adversarial interpretability approaches (Ribeiro et al., 2018b; Lakkaraju and Bastani, 2020), we train an MLP classifier on the SpamAssassin dataset and search for minimal factors sufficient to relabel a sample of spam emails as non-spam.

Our examples follow some patterns common to spam emails: received from unusual email addresses, includes suspicious keywords such as ENLARGEMENT or ADVERTISEMENT in the subject line, etc. We identify minimal changes that will flip labels to non-spam with high probability. Options include altering the incoming email address to more common domains, and changing the subject or first sentences (see Tables 3). These results can improve understanding of both a model’s behavior and a dataset’s properties.

Diverse counterfactuals. Our explanatory measures can also be used to secure algorithmic recourse. For this experiment, we benchmark against DiCE (Mothilal et al., 2020b), which aims to provide diverse recourse options for any underlying prediction model. We illustrate the differences

input									I2R		I2R$_{causal}$	
Age	Sex	Job	Housing	Savings	Checking	Credit	Duration	Purpose	$\tau-$minimal factors ($\tau = 0$)	Cost	$\tau-$minimal factors ($\tau = 0$)	Cost
23	Male	Skilled	Free	Little	Little	1845	45	Radio/TV	Job: Highly skilled	1	Age: 24	0.07
									Checking: NA	1	Sex: Female	1
									Duration: 30	1.25	Job: Highly skilled	1
									Age: 65, Housing: Own	4.23	Housing: Rent	1
									Age: 34, Savings: N/A	1.84	Savings: N/A	1

Table 4. Recourse example comparing causal and non-causal (i.e., feature independent) $\mathcal{D}$. We sample a single input example with a negative prediction, and 100 references with the opposite outcome. For I2R$_{causal}$ we propagate the effects of interventions through a user-provided SCM.

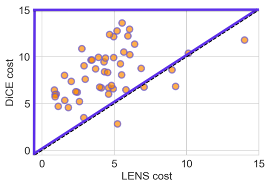

Figure 4. A comparison of mean cost of outputs by LENS and DiCE for 50 inputs sampled from the `Adult` dataset.

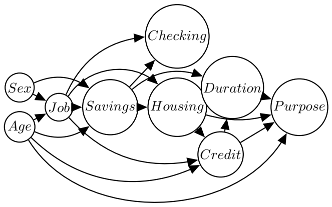

Figure 5. Example DAG for `German` dataset.

between our respective approaches on the `Adult` dataset (Kochavi and Becker, 1996), using an MLP and following the procedure from the original DiCE paper.

According to DiCE, a diverse set of counterfactuals is one that differs in *values* assigned to features, and can thus produce a counterfactual set that includes different interventions on the same variables (e.g., CF1: $\texttt{age} = 91, \texttt{occupation}$ = "retired"; CF2: $\texttt{age} = 44, \texttt{occupation}$ = "teacher"). Instead, we look at diversity of counterfactuals in terms of intervention *targets*, i.e. features changed (in this case, from input to reference values) and their effects. We present minimal cost interventions that would lead to recourse for each feature set, but we summarize the set of paths to recourse via subsets of features changed. Thus, DiCE provides answers of the form "Because you are not 91 and retired" or "Because you are not 44 and a teacher"; we answer "Because of your age and occupation", and present the lowest cost intervention on these features sufficient to flip the prediction.

With this intuition in mind, we compare outputs given by DiCE and LENS for various inputs. For simplicity, we let all features vary independently. We consider two metrics for comparison: (a) the mean cost of proposed factors, and (b) the number of minimally valid candidates proposed, where a factor c from a method M is *minimally valid* iff for all c' proposed by M', $\neg(c' \prec_{cost} c)$ (i.e., M' does not report a factor better than c). We report results based on 50 randomly sampled inputs from the `Adult` dataset, where references are fixed by conditioning on the opposite prediction. The cost comparison results are shown in Fig. 4, where we find that LENS identifies lower cost factors for the vast majority of inputs. Furthermore, DiCE finds no minimally valid candidates that LENS did not already account for. Thus LENS emphasizes *minimality* and *diversity* of intervention targets, while still identifying low cost intervention values.

Causal vs. non-causal recourse. When a user relies on XAI methods to plan interventions on real-world systems, causal relationships between predictors cannot be ignored. In the following example, we consider the DAG in Fig. 5, intended to represent dependencies in the `German` credit dataset. For illustrative purposes, we assume access to the structural equations of this data generating process. (There are various ways to extend our approach using only partial causal knowledge as input (Karimi et al., 2020b; Heskes et al., 2020).) We construct D by sampling from the SCM under a series of different possible interventions. Table 4 describes an example of how using our framework with augmented causal knowledge can lead to different recourse options. Computing explanations under the assumption of feature independence results in factors that span a large part of the DAG depicted in Fig. 5. However, encoding structural relationships in D, we find that LENS assigns high explanatory value to nodes that appear early in the topological ordering. This is because intervening on a single root factor may result in various downstream changes once effects are fully propagated.

6. Conclusion

We have presented a unified framework for XAI that foregrounds necessity and sufficiency, which we argue are the fundamental building blocks of successful explanations. We defined simple measures of both, and showed how they undergird various XAI methods. We illustrated illuminating connections between our explanatory measures and Pearl (2000)'s probabilities of causation, extending our scope to

SCMs. We introduced a sound and complete algorithm for identifying minimally sufficient factors, and demonstrated our method on a range of tasks and datasets. Our approach prioritizes completeness over efficiency, suitable for settings of limited dimensionality. Yet our experiments demonstrate that empirical sampling on datasets of moderate dimensionality can produce compelling results. Datasets of higher dimensionality can be represented as lower dimensional abstractions favorable to explanations (see discussion in Appendix B). Future research will explore approximations that perform well on larger datasets, model-specific variants optimized for, e.g., convolutional neural networks, and developing a graphical user interface.

Acknowledgments

David Watson was supported in part by ONR grant N62909-19-1-2096.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv* preprint, 1903.10464v2, 2019.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M F Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly, 2009.

Martijn Blaauw, editor. *Contrastivism in Philosophy*. Routledge, New York, 2013.

Andrew Brennan. Necessary and Sufficient Conditions. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2017.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, volume 31, pages 592–603, 2018.

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

C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by AI systems: The counterfactual approach. *arXiv* preprint, 2001.07417, 2020.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *Ann. Appl. Stat.*, 2(3):916–954, 2008.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, 2006.

Joseph Y Halpern. *Actual causality*. The MIT Press, Cambridge, MA, 2016.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *Br. J. Philos. Sci.*, 56(4):843–887, 2005a.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *Br. J. Philos. Sci.*, 56(4):889–911, 2005b.

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

Daniel Kahneman and Dale T. Miller. Norm theory: Comparing reality to its alternatives. *Psychol. Rev.*, 93(2): 136–153, 1986.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. *arXiv* preprint, 2010.04050, 2020a.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in Neural Information Processing Systems*, 2020b.

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

Ronny Kochavi and Barry Becker. Adult income dataset, 1996. URL `https://archive.ics.uci.edu/ml/datasets/adult`.

Indra Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures.

In *Proceedings of the 37th International Conference on Machine Learning*, pages 1–10, 2020.

Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, New York, NY, USA, 2020.

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

E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, third edition, 2005.

Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, 9(3):1350–1371, 2015.

Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

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

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 142–150, 2011.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction - 4th International Cross-Domain Conference (CD-MAKE)*, volume 12279, pages 17–38. Springer, 2020.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 101(2):343–352, 1955.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Münich, 2021. URL `https://christophm.github.io/interpretable-ml-book/`.

Ramaravind K. Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. *arXiv* preprint, 2011.04917, 2020a.

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

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Advances in Data Analysis and Classification*, 2020.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535, 2018a.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 856–865, 2018b.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

Lloyd Shapley. A value for *n*-person games. In *Contributions to the Theory of Games*, chapter 17, pages 307–317. Princeton University Press, Princeton, 1953.

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

Apache SpamAssassin, 2006. URL `https://spamassassin.apache.org/old/publiccorpus/`. Accessed 2021.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the ACM Conference*, New York, 2019.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, New York, NY, USA, 2019.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

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

David S Watson and Luciano Floridi. The explanation game: a formal framework for interpretable machine learning. *Synthese*, 2020.

J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020.

Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in Neural Information Processing Systems*, page 4879–4890, 2018.

Appendix

A. Proofs

A.1. Theorems

A.1.1. Proof of Theorem 1

Theorem. With oracle estimates of $PS(c, y)$ for all $c \in \mathcal{C}$, Alg. 1 is sound and complete.

Proof. Soundness and completeness follow directly from the specification of (P1) $\mathcal{C}$ and (P2) $\preceq$ in the algorithm's input $\mathcal{B}$, along with (P3) access to oracle estimates of $PS(c, y)$ for all $c \in \mathcal{C}$. Recall that the partial ordering must be complete and transitive, as noted in Sect. 3.

Assume that Alg. 1 generates a false positive, i.e. outputs some c that is not τ-minimal. Then by Def. 4, either the algorithm failed to properly evaluate $PS(c, y)$, thereby violating (P3), or failed to identify some c' such that (i) $PS(c', y) \geq \tau$ and (ii) $c' \prec c$. (i) is impossible by (P3), and (ii) is impossible by (P2). Thus there can be no false positives.

Assume that Alg. 1 generates a false negative, i.e. fails to output some c that is in fact τ-minimal. By (P1), this c cannot exist outside the finite set $\mathcal{C}$. Therefore there must be some $c \in \mathcal{C}$ for which either the algorithm failed to properly evaluate $PS(c, y)$, thereby violating (P3), or wrongly identified some c' such that (i) $PS(c', y) \geq \tau$ and (ii) $c' \prec c$. Once again, (i) is impossible by (P3), and (ii) is impossible by (P2). Thus there can be no false negatives.

A.1.2. Proof of Theorem 2

Theorem. With sample estimates of $\hat{PS}(c, y)$ for all $c \in \mathcal{C}$, Alg. 1 is optimal, in the frequentist sense.

Proof. A testing procedure is uniformly most powerful (UMP) if it attains the lowest type II error β of all tests with fixed type I error α. Let Θ_0, Θ_1 denote a partition of the parameter space into null and alternative regions. The goal in frequentist inference is to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$ for some parameter θ. Let $\psi(X)$ be a testing procedure of the form $\mathbb{1}[T(X) \geq c_\alpha]$, where X denotes a finite sample, $T(X)$ is a test statistic, and c_α is the critical value. This latter parameter defines a rejection region such that test statistics integrate to α under H_0. We say that $\psi(X)$ is UMP iff, for any other test $\psi'(X)$ such that

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi'(X)] \leq \alpha,$$

we have

$$(\forall \theta \in \Theta_1)\ \mathbb{E}_\theta[\psi'(X)] \leq \mathbb{E}_\theta[\psi(X)],$$

where $\mathbb{E}_{\theta \in \Theta_1}[\psi(X)]$ denotes the power of the test to detect the true θ, $1 - \beta_\psi(\theta)$. The UMP-optimality of Alg. 1 follows from the UMP-optimality of the binomial test (see (Lehmann and Romano, 2005), Ch. 3), which is used to decide between $H_0 : PS(c, y) < \tau$ and $H_1 : PS(c, y) \geq \tau$ on the basis of observed proportions $\hat{PS}(c, y)$, estimated from n samples for all $c \in \mathcal{C}$. The proof now takes the same structure as that of Thm. 1, with (P3) replaced by (P3′): access to UMP estimates of $PS(c, y)$. False positives are no longer impossible but bounded at level α; false negatives are no longer impossible but occur with frequency β. Because no procedure can find more τ-minimal factors for any fixed α, Alg. 1 is UMP.

A.2. Propositions

A.2.1. Proof of Proposition 1

Proposition. Let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x} \subseteq \boldsymbol{z}$ was constructed by holding $\boldsymbol{x}^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\cdot|S)$. Then $v(S) = PS(c_S, y)$.

As noted in the text, $\mathcal{D}(\boldsymbol{x}|S)$ may be defined in a variety of ways (e.g., via marginal, conditional, or interventional distributions). For any given choice, let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x}$ is constructed by holding $\boldsymbol{x}_i^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\boldsymbol{x}|S)$. Since we assume binary Y (or binarized, as discussed in 3), we can rewrite Eq. 2 as a probability:

$$v(S) = P_{\mathcal{D}(\boldsymbol{x}|S)}(f(\boldsymbol{x}_i) = f(\boldsymbol{x})),$$

where $\boldsymbol{x}_i$ denotes the input point. Since conditional sampling is equivalent to conditioning after sampling, this value function is equivalent to $PS(c_S, y)$ by Def. 2.

A.2.2. Proof of Proposition 2

Proposition. Let $c_A(\boldsymbol{z}) = 1$ iff $A(\boldsymbol{x}) = 1$. Then $\text{prec}(A) = PS(c_A, y)$.

The proof for this proposition is essentially identical, except in this case our conditioning event is $A(\boldsymbol{x}) = 1$. Let $c_A = 1$ iff $A(\boldsymbol{x}) = 1$. Precision prec(A), given by the lhs of Eq. 3, is defined over a conditional distribution $\mathcal{D}(\boldsymbol{x}|A)$. Since conditional sampling is equivalent to conditioning after sampling, this probability reduces to $PS(c_A, y)$.

A.2.3. Proof of Proposition 3

Proposition. Let *cost* be a function representing $\preceq$, and let c be some factor spanning reference values. Then the counterfactual recourse objective is:

$$c^* = \operatorname*{argmin}_{c \in \mathcal{C}}\ cost(c)\ \text{ s.t. } PS(c, 1 - y) \geq \tau, \tag{7}$$

where τ denotes a decision threshold. Counterfactual outputs will then be any $\boldsymbol{z} \sim \mathcal{D}$ such that $c^*(\boldsymbol{z}) = 1$.

There are two closely related ways of expressing the counterfactual objective: as a search for optimal *points*, or optimal *actions*. We start with the latter interpretation, reframing actions as factors. We are only interested in solutions that flip the original outcome, and so we constrain the search to factors that meet an I2R sufficiency threshold, $PS(c, 1 - y_i) \geq \tau$. Then the optimal action is attained by whatever factor (i) meets the sufficiency criterion and (ii) minimizes cost. Call this factor c^*. The optimal point is then any $\boldsymbol{z}$ such that $c^*(\boldsymbol{z}) = 1$.

A.2.4. Proof of Proposition 4

Proposition. Consider the bivariate Boolean setting, as in Sect. 2. We have two counterfactual distributions: an input space $\mathcal{I}$, in which we observe x, y but intervene to set $X = x'$; and a reference space $\mathcal{R}$, in which we observe x', y' but intervene to set $X = x$. Let $\mathcal{D}$ denote a uniform mixture over both spaces, and let auxiliary variable W tag each sample with a label indicating whether it comes from the original ($W = 1$) or contrastive ($W = 0$) counterfactual space. Define $c(\boldsymbol{z}) = w$. Then we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, y')$.

Recall from Sect. 2 that (Pearl, 2000) defines $\texttt{suf}(x, y) := P(y_x | x', y')$ and $\texttt{nec}(x, y) := P(y'_{x'} | x, y)$. We may rewrite the former as $P_{\mathcal{R}}(y)$, where the reference space $\mathcal{R}$ denotes a counterfactual distribution conditioned on $x', y', do(x)$. Similarly, we may rewrite the latter as $P_{\mathcal{I}}(y')$, where the input space $\mathcal{I}$ denotes a counterfactual distribution conditioned on $x, y, do(x')$. Our context $\mathcal{D}$ is a uniform mixture over both spaces.

The key point here is that the auxiliary variable W indicates whether samples are drawn from $\mathcal{I}$ or $\mathcal{R}$. Thus conditioning on different values of W allows us to toggle between probabilities over the two spaces. Therefore, for $c(\boldsymbol{z}) = w$, we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, y')$.

B. Additional discussions of method

B.1. τ-minimality and necessity

As a follow up to Remark 2 in Sect. 3.2, we expand here upon the relationship between τ and cumulative probabilities of necessity, which is similar to a precision-recall curve quantifying and qualifying errors in classification tasks. In this case, as we lower τ, we allow more factors to be taken into account, thus covering more pathways towards a desired outcome in a cumulative sense. We provide an example of such a precision-recall curve in Fig. 6, using an R2I view of the German credit dataset. Different levels of cumulative necessity may be warranted for different tasks, depending on how important it is to survey multiple paths towards an outcome. Users can therefore adjust τ to accommodate desired levels of cumulative PN over successive calls to LENS.

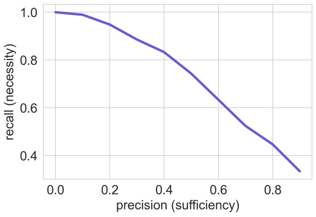

Figure 6. An example curve exemplifying the relationship between τ and cumulative probability necessity attained by selected τ-minimal factors.

B.2. Enumeration, low-dimensional explanations, and abstraction

While our proposed theoretical framework is broad enough to accommodate approximation of PS and sampling from any well-specified context $\mathcal{D}$, in our practical demonstration we focus on a full-enumeration based approach, and empirical values of all components. We manually apply all possible interventions under a basis $\mathcal{B}$. One criticism of this approach is that it is computationally infeasible in high dimensions, as the number of possible feature subsets is exponential in d. While we intend to work on statistical approximations of this approach that would be more favorable to higher dimensional settings in future work, we intentionally focused on lower-to-middle dimensional cases first. As noted in multiple sections of (Miller, 2019), “Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation.” Various authors cited in (Miller, 2019) emphasize that ideal explanations must be minimal, e.g., “ explanations that offer fewer causes and explanations that explain multiple observations are considered more believable and more valuable”, or “all things being equal, simpler explanations — those that cite fewer causes... are better explanations”.

Thus, we contend that higher-dimensional datasets are an obstacle not just computationally, but also semantically. Even if we could feasibly enumerate all τ-minimal factors for some very large value of d, it is not clear that such explanations would be helpful to humans, who famously struggle to hold more than seven objects in short-term memory at any given time (Miller, 1955). One strategy in such cases is to use a low-dimensional representation of a high-dimensional dataset. Explanations can now be offered at a higher level of abstraction. For instance, in our SpamAssassins experiments, we started with a pure text example that can be represented via high-dimensional vectors (e.g., word embeddings). However, we chose to pre-process the dataset and boil it down to its core elements: From and To email addresses, Subject line, First and Last sentences, etc. This way, we created a more abstract object and considered each segment as a potential intervention target, i.e. as candidate factors. We thus compressed a high-dimensional dataset into a 10-dimensional abstraction with only 10 targets of interventions. Similar strategies could be used in many cases, either through domain-knowledge or data-driven dimensionality reduction techniques.

C. Additional discussions of experimental results

C.1. Data pre-processing and model training

German Credit Risk. We first download the dataset from Kaggle[3], which is a slight modification of the UCI version (Dua and Graff, 2017). We follow the pre-processing steps from a Kaggle tutorial.[4] In particular, we map the categorical string variables in the dataset (Savings accounts, Checking account, Sex, Housing, Purpose and the outcome Risk) to numeric encodings, and mean-impute values missing values for Savings accounts and Checking account. We then train an Extra-Tree classifier (Geurts et al., 2006) using scikit-learn, with random state 0 and max depth 15. All other hyperparameters are at their default values. The model achieves a 71% accuracy.

German Credit Risk - Causal. We assume a partial ordering over the features in the dataset, as described in Fig. 5. We use this DAG to fit a structural causal model (SCM) based on the original data. In particular, we fit linear regressions for every continuous variable and a random forest classifier for every categorical variable. When sampling from $\mathcal{D}$, we let variables remain at their original values unless either (a) they are directly intervened on, or (b) one of their ancestors was intervened on. In the latter case, changes are propagated via the structural equations. We add stochasticity via Gaussian noise for continuous outcomes, with variance given by each model's residual mean squared error. For categorical variables, we perform multinomial sampling over predicted class probabilities. We use the same f model as for the non-causal `German` credit risk description above.

SpamAssassins. The original spam assassins dataset comes in the form of raw, multi-sentence emails captured on the Apache SpamAssassins project, 2003-2015.[5] To utilize our method for use with this dataset, we segmented the emails to the following "features": `From`, represents the from address line; `To`, represented the to address line; `Subject`, represents the subject line of the email; `Urls`, any URLs found in the email; `Emails`, any email address found in the email; `First Sentence`, represents the first sentence in the body of the email; `Second Sentence`, represents the second sentence in the body of the email; `Penult Sentence`, represents the penultimate sentence in the body of the email; `Last Sentence`, represents the last sentence in the body of the email. We use the original outcome label from the original dataset (indicated by which folder the different emails were saved). Once we obtain a dataset in the form above, we continue to pre-process by lower casing all characters, only keeping words or digits, clearing most punctuation (except for '-' and '_'), and removing stopwords based on nltk's provided list (Bird et al., 2009). Finally, we convert all clean strings to their mean 50-dim GloVe vector representation (Pennington et al., 2014). We train a standard MLP classifier using scikit-learn, with random state 1, max iteration 300, and all other hyperparameters set to their default values. [6] This model attains an accuracy of 98.3%.

IMDB. We follow the pre-processing and modeling steps taken in a standard tutorial on LSTM training for sentiment prediction with the IMDB dataset.[7] The CSV is included in the repository named above, and can be additionally downloaded from Kaggle or ai.standford.[8] In particular, these include removal of HTML-tags, non-alphabetical characters, and stopwords based on the the list provided in the ntlk package, as well as changing all alphabetical characters to lower-case. We then train a standard LSTM model, with 32 as the embedding dimension and 64 as the dimensionality of the output space of the LSTM layer, and an additional dense layer with output size 1. We use the sigmoid activation function, binary cross-entropy loss, and optimize with Adam (Kingma and Ba, 2015). All other hyperparameters are set to their default values as specified by Keras.[9] The model achieves an accuracy of 87.03%.

Adult Income. We obtain the adult income dataset via DiCE's implementation.[10] to maximize comparability of our methods DiCE in turn followed the pre-processing steps of Haojun Zhu 2016.[11] We further use a pretrained MLP model trained by DiCE for our recourse comparison, which is a single layer, non-linear model trained with TensorFlow, and stored in their repository as 'adult.h5'.

[3] See `https://www.kaggle.com/kabure/german-credit-data-with-risk?select=german_credit_data.csv`

[4] See `https://www.kaggle.com/vigneshj6/german-credit-data-analysis-python`.

[5] See `https://spamassassin.apache.org/old/credits.html`.

[6] See `https://scikit-learn.org/stable/modules/generated/sklearn.\neural_network.MLPClassifier.html`

[7] See `https://github.com/hansmichaels/sentiment-analysis-IMDB-Review-using-LSTM/blob/master/sentiment_analysis.py.ipynb`.

[8] See `https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews` or `http://ai.stanford.edu/~amaas/data/sentiment/`.

[9] See `https://keras.io`

[10] See `https://github.com/interpretml/DiCE`.

[11] See `https://rpubs.com/H_Zhu/235617`.

input								DiCE output		**LENS output**	
Age	Wrkcls	Edu.	Marital	Occp.	Race	Sex	Hrs/week	Targets of intervention	Cost	Targets of intervention	Cost
42	Govt.	HS-grad	Single	Service	White	Male	40	Age, Edu., Marital, Hrs/week	8.13	Edu.	1
								Age, Edu., Marital, Occp., Sex, Hrs/week	5.866	Martial	1
								Age, Wrkcls, Educ., Marital, Hrs/week	5.36	Occp., Hrs/week	19.3
								Age, Edu., Occp., Hrs/week	3.2	Wrkcls, Occp., Hrs/week	12.6
								Edu., Hrs/week	11.6	Age, Wrkcls, Occp., Hrs/week	12.2

Table 5. Recourse options for a single input given by DiCE and our method. We report here targets of interventions as suggested options, but they could correspond to different values of interventions. Our method tends to propose more minimal and diverse options in terms of targets of intervention. Note that all of DiCE's outputs are already subsets of LENS's two top suggestions, and due to τ-minimality LENS is forced to pick the next factors to be non-supersets of the two top rows. This explains the higher cost of LENS's bottom three rows.

C.1.1. TASKS

Comparison with attributions. For completeness, we also include here comparison of cumulative attribution scores per cardinality with probabilities of sufficiency as defined in LENS for I2R view, see Fig. 7.

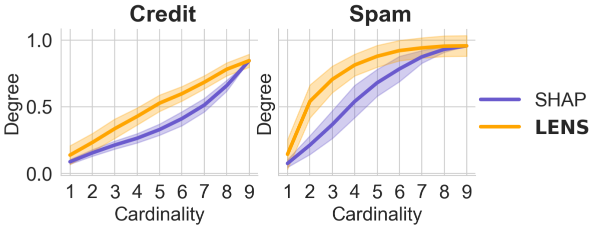

Figure 7. Comparison of degrees of sufficiency in I2R setting, for top k features based on SHAP scores, against the best performing subset for cardinality k identified by our method. Results for `German` are averaged over 50 inputs; results for `SpamAssassins` are averaged over 25 inputs.

Sentiment sensitivity analysis. We identify sentences in the original `IMDB` dataset that are up to 10 words long. Out of those, for the first example we only look at wrongly predicted sentences to identify a suitable example. For the other example, we simply consider a random example from the 10-word maximum length examples. We noted that Anchors uses stochastic word-level perturbations for this setting. This leads them to identify explanations of higher cardinality for some sentences, which include elements that are not strictly necessary. In other words, their outputs are not minimal, as required for descriptions of "actual causes" by (Halpern, 2016; Halpern and Pearl, 2005a).

Comparison with anchors. To complete the picture of our comparison with Anchors on the `German` Credit Risk dataset, we provide here additional results. In the main text, we included a comparison of Anchors's single output precision against the mean degree of sufficiency attained by our multiple suggestions per input. We sample 100 different inputs from the `German` Credit dataset and repeat this same comparison. Here we additionally consider the minimum and maximum probability of sufficiency $PS(c, y)$ attained by LENS against Anchors. Note that even when considering minimum PS suggestions by LENS, i.e. our worst output, the method shows more consistent performance. We qualify this discussion by noting that, by setting Anchors's δ hyperparameter to a lower value, one could possibly get comparable results to ours. We still see value in emphasizing this difference, however, as Ribeiro et al. (2018a) do not discuss this parameter in detail in either their original article or subsequent notebook guides. They use default settings in their own experiments, and we expect most practitioners will do the same.

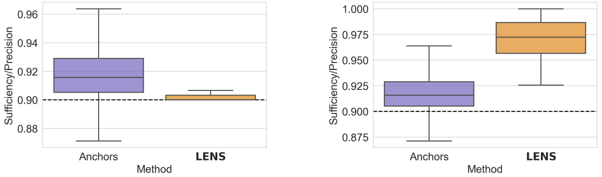

Figure 8. We compare degree of sufficiency against precision scores attained by the output of LENS and Anchors for examples from `German`. We repeat the experiment for 100 sampled inputs, and each time consider the single output by Anchors against the min (left) and max (right) $PS(c, y)$ among LENS's multiple candidates. Dotted line indicates $\tau = 0.9$, the threshold we chose for this experiment.

Recourse: DiCE comparison First, we provide a single illustrative example of the lack of diversity in intervention targets we identify in DiCE's output. Let us consider one example, shown in Table 5. While DiCE outputs are diverse in terms of values and target combinations, they tend to have great overlap in intervention targets. For instance, Age and Education appear in almost all of them. Our method would focus on minimal paths to recourse that would involve different combinations of features.

Next, we also provide additional results from our cost comparison with DiCE's output in Fig. 8. While in the main text we include a comparison of our mean cost output against DiCE's, here we additionally include a comparison of min and max cost of the methods' respective outputs. We see that even when considering minimum and maximum cost,

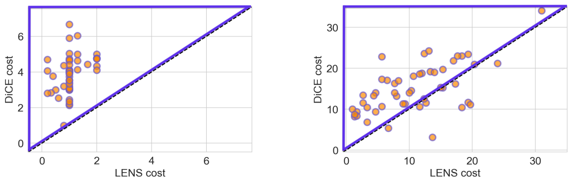

Figure 9. We show results over 50 input points sampled from the original dataset, and all possible references of the opposite class, across two metrics: the min cost (left) of counterfactuals suggested by our method vs. DiCE, and the max cost (right) of counterfactuals.

our method tends to suggest better cost recourse options. In particular, note that all of DiCE's outputs are already subsets of LENS's two top suggestions. The higher costs incurred by LENS for the next two lines are a reflection of this fact: due to τ-minimality, LENS is forced to find other intervention that are no longer supersets of options already listed above.