ECCCos from the Black Box: Faithful Explanations through Energy-Constrained Conformal Counterfactuals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Counterfactual Explanations offer an intuitive and straightforward way to explain black-box models and offer Algorithmic Recourse to individuals. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn how the input data is distributed. This effectively reallocates the task of learning realistic representations of the data from the model itself to the surrogate. Consequently, the generated explanations may seem plausible to humans but need not necessarily faithfully describe the behaviour of the black-box model. We formalise this notion of faithfulness through the introduction of a tailored evaluation metric and propose a novel algorithmic framework for generating **E**nergy-**C**onstrained **C**onformal **Co**unterfactuals that are only as plausible as the model permits. Through extensive empirical studies involving multiple synthetic and real-world datasets, we demonstrate that **ECCCo** reconciles the need for plausibility and faithfulness. In particular, we show that it is possible to achieve state-of-the-art plausibility for models with gradient access without the need for surrogate models. To do so, ECCCo relies solely on properties defining the black-box model itself by leveraging recent advances in energy-based modelling and conformal inference. Through this work, we also shine new light on the explanatory properties of Joint Energy Models. Our framework is intuitive, flexible and fully open-sourced. By highlighting the need for faithfulness in the context of Counterfactual Explanations, we believe that in the short term, our work will enable researchers and practitioners to better distinguish trustworthy from unreliable models. We further anticipate that ECCCo can serve as a baseline for future research directed at providing plausible but faithful Counterfactual Explanations.

1 Introduction

Counterfactual Explanations provide a powerful, flexible and intuitive way to not only explain black-box models but also enable affected individuals to challenge them through the means of Algorithmic Recourse. Instead of opening the black box, Counterfactual Explanations work under the premise of strategically perturbing model inputs to understand model behaviour [29]. Intuitively speaking, we generate explanations in this context by asking simple what-if questions of the following nature: 'Our credit risk model currently predicts that this individual's credit profile is too risky to offer them a loan. What if they reduced their monthly expenditures by 10%? Will our model then predict that the individual is credit-worthy'?

This is typically implemented by defining a target outcome $\mathbf{y}^* \in \mathcal{Y}$ for some individual $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ described by D attributes, for which the model $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ initially predicts a different outcome:

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

$M_\theta(\mathbf{x}) \neq \mathbf{y}^*$. Counterfactuals are then searched by minimizing a loss function that compares the predicted model output to the target outcome: $\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}^*)$. Since Counterfactual Explanations (CE) work directly with the black-box model, valid counterfactuals always have full local fidelity by construction [17]. Fidelity is defined as the degree to which explanations approximate the predictions of the black-box model. This is arguably one of the most important evaluation metrics for model explanations, since any explanation that explains a prediction not actually made by the model is useless [16].

In situations where full fidelity is a requirement, CE therefore offers a more appropriate solution to Explainable Artificial Intelligence (XAI) than other popular approaches like LIME [22] and SHAP [12], which involve local surrogate models. But even full fidelity is not a sufficient condition for ensuring that an explanation faithfully describes the behaviour of a model. That is because multiple very distinct explanations can all lead to the same model prediction, especially when dealing with heavily parameterized models like deep neural networks which are typically underspecified by the available data [30].

In the context of CE, the idea that no two explanations are the same arises almost naturally. A key focus in the literature has therefore been to identify those explanations and algorithmic recourses that are deemed most appropriate based on a myriad of desiderata such as sparsity, actionability and plausibility. In this work, we draw closer attention to the insufficiency of model fidelity as an evaluation metric for the faithfulness of counterfactual explanations. Our key contributions are as follows: firstly, we introduce a new notion of faithfulness that is suitable for counterfactuals and propose a novel evaluation measure that draws inspiration from recent advances in Energy-Based Modelling (EBM); secondly, we a novel algorithmic approach for generating Energy-Constrained Conformal Counterfactuals (ECCCo) that explicitly address the need for faithfulness; finally, we provide illustrative examples and extensive empirical evidence demonstrating that ECCCos faithfully explain model behaviour without sacrificing existing desidarata like plausibility and sparsity.

2 Background and Related Work

In this section, we provide some background on Counterfactual Explanations and our motivation for this work. To start, we briefly introduce the methodology underlying most state-of-the-art (SOTA) counterfactual generators.

2.1 Gradient-Based Counterfactual Search

While Counterfactual Explanations can be generated for arbitrary regression models [24], existing work has primarily focused on classification problems. Let $\mathcal{Y} = (0, 1)^K$ denote the one-hot-encoded output domain with K classes. Then most SOTA counterfactual generators rely on gradient descent to optimize different flavours of the following counterfactual search objective:

$$\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} \{\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^*) + \lambda \text{cost}(f(\mathbf{Z}'))\} \tag{1}$$

Here yloss denotes the primary loss function already introduced above and cost is either a single penalty or a collection of penalties that are used to impose constraints through regularization. Equation 1 restates the baseline approach to gradient-based counterfactual search proposed by Wachter et al. [29] in general form where $\mathbf{Z}' = \{\mathbf{z}_l\}_L$ denotes an L-dimensional array of counterfactual states [2]. This is to explicitly account for the multiplicity of explanations and the fact that we may choose to generate multiple counterfactuals and traverse a latent encoding $\mathcal{Z}$ of the feature space $\mathcal{X}$ where we denote $f^{-1} : \mathcal{X} \mapsto \mathcal{Z}$. Encodings may involve simple feature transformations or more advanced techniques involving generative models, as we will discuss further below. The baseline approach, which we will simply refer to as **Wachter** [29], searches a single counterfactual directly in the feature space and penalises its distance between the original factual.

Solutions to Equation 1 are considered valid as soon as the predicted label matches the target label. A stripped-down counterfactual explanation is therefore little different from an adversarial example. In Figure 1, for example, we have applied Wachter to MNIST data (centre panel) where the underlying classifier M_θ is a simple Multi-Layer Perceptron (MLP) with above 90 percent test accuracy. For the generated counterfactual $\mathbf{x}'$ the model predicts the target label with high confidence (centre panel in Figure 1). The explanation is valid by definition, even though it looks a lot like an Adversarial Example [6]. Schut et al. [23] make the connection between Adversarial Examples and Counterfactual Explanations explicit and propose using a Jacobian-Based Saliency Map Attack (JSMA) to solve Equation 1. They demonstrate that this approach yields realistic and sparse counterfactuals for Bayesian, adversarially robust classifiers. Applying their approach to our simple MNIST classifier does not yield a realistic counterfactual but this one, too, is valid (right panel in Figure 1).

2.2 From Adversial Examples to Plausible Explanations

The crucial difference between Adversarial Examples (AE) and Counterfactual Explanations is one of intent. While an AE is intended to go unnoticed, a CE should have certain desirable properties. The literature has made this explicit by introducing various so-called *desiderata* that counterfactuals should meet in order to properly serve both AI practitioners and individuals affected by AI decision-making systems. The list of desiderate includes but is not limited to the following: sparsity, proximity [29], actionability [27], diversity [17], plausibility [9, 21, 23], robustness [26, 20, 2] and causality [11].

Researchers have come up with various ways to meet these desiderata, which have been extensively surveyed and evaluated in various studies [28, 10, 19, 4, 8]. Perhaps unsurprisingly, the different desiderata are often positively correlated. For example, Artelt et al. [4] find that plausibility typically also leads to improved robustness. Similarly, plausibility has also been connected to causality in the sense that plausible counterfactuals respect causal relationships [13].

2.2.1 Plausibility through Surrogates

Arguably, the plausibility of counterfactuals has been among the primary concerns and some have focused explicitly on this goal. Joshi et al. [9], for example, were among the first to suggest that instead of searching counterfactuals in the feature space $\mathcal{X}$, we can instead traverse a latent embedding $\mathcal{Z}$ (Equation 1) that implicitly codifies the data generating process (DGP) of $\mathbf{x} \sim \mathcal{X}$. To learn the latent embedding, they introduce a surrogate model. In particular, they propose to use the latent embedding of a Variational Autoencoder (VAE) trained to generate samples $\mathbf{x}^* \leftarrow \mathcal{G}(\mathbf{z})$ where $\mathcal{G}$ denotes the decoder part of the VAE. Provided the surrogate model is well-trained, their proposed approach —REVISE— can yield compelling counterfactual explanations like the one in the centre panel of Figure 2.

Others have proposed similar approaches. Dombrowski et al. [5] traverse the base space of a normalizing flow to solve Equation 1, essentially relying on a different surrogate model for the generative task. Poyiadzi et al. [21] use density estimators ($\hat{p} : \mathcal{X} \mapsto [0,1]$) to constrain the counterfactuals to dense regions in the feature space. Karimi et al. [11] argue that counterfactuals should comply with the causal model that generates the data. All of these different approaches share a common goal: ensuring that the generated counterfactuals comply with the true and unobserved DGP. To summarize this broad objective, we propose the following definition:

Definition 2.1 (Plausible Counterfactuals)**.** *Let $\mathcal{X}|\mathbf{y}^*$ denote the true conditional distribution of samples in the target class $\mathbf{y}^*$. Then for $\mathbf{x}'$ to be considered a plausible counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^*$.*

Surrogate models offer an obvious solution to achieve this objective. Unfortunately, surrogates also introduce a dependency: the generated explanations no longer depend exclusively on the black-box model itself, but also on the surrogate model. This is not necessarily problematic if the primary objective is not to explain the behaviour of the model but to offer recourse to individuals affected by it. It may become problematic even in this context if the dependency turns into a vulnerability. To illustrate this point, we have used REVISE [9] with an underfitted VAE to generate the counterfactual in the right panel of Figure 2: in this case, the decoder step of the VAE fails to yield plausible values ($\{\mathbf{x}' \leftarrow \mathcal{G}(\mathbf{z})\} \not\sim \mathcal{X}|\mathbf{y}^*$) and hence the counterfactual search in the learned latent space is doomed.

2.2.2 Plausibility through Minimal Predictive Uncertainty

Schut et al. [23] show that to meet the plausibility objective we need not explicitly model the input distribution. Pointing to the undesirable engineering overhead induced by surrogate models, they propose that we rely on the implicit minimisation of predictive uncertainty instead. Their proposed methodology solves Equation 1 by greedily applying JSMA in the feature space with standard cross-entropy loss and no penalty at all. They demonstrate theoretically and empirically that their approach

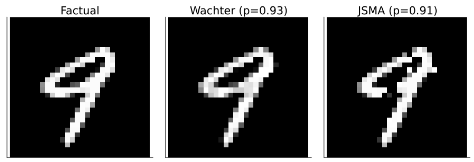

Figure 1: Explanations or Adversarial Examples? Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using Wachter et al. [29] (centre); and a counterfactual produced using the approach introduced by [23] that uses Jacobian-Based Saliency Map Attacks to solve Equation 1.

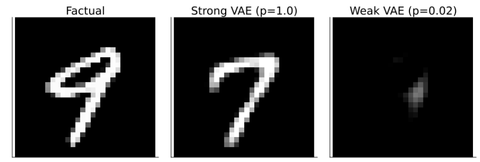

Figure 2: Using surrogates can improve plausibility, but also increases vulnerability. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using REVISE [9] with a well-specified surrogate (centre); and a counterfactual produced using REVISE [9] with a poorly specified surrogate (right).

yields counterfactuals for which the model M_θ predicts the target label $\mathbf{y}^*$ with high confidence. Provided the model is well-specified, these counterfactuals are plausible. Unfortunately, this idea hinges on the assumption that the black-box model provides well-calibrated predictive uncertainty estimates.

2.3 From Fidelity to Model Faithfulness

Above we explained that since Counterfactual Explanations work directly with the Black Box model, the fidelity of explanations as we defined it earlier is not a concern. This may explain why research has primarily focused on other desiderata, most notably plausibility (Definition 2.1). Enquiring about the plausibility of a counterfactual essentially boils down to the following question: 'Is this counterfactual consistent with the underlying data'? We posit a related, slightly more nuanced question: 'Is this counterfactual consistent with what the model has learned about the underlying data'? We will argue that fidelity is not a sufficient evaluation measure to answer this question and propose a novel way to assess if Counterfactual Explanations conform with model behaviour.

The word *fidelity* stems from the Latin word 'fidelis', which means 'faithful, loyal, trustworthy' [15]. As we explained in Section 2, model explanations are generally considered faithful if their corresponding predictions coincide with the predictions made by the model itself. Since this definition of faithfulness is not useful in the context of Counterfactual Explanations, we propose an adapted version:

Definition 2.2 (Faithful Counterfactuals)**.** *Let $\mathcal{X}_\theta|\mathbf{y}^* = p_\theta(\mathbf{X}_{\mathbf{y}^*})$ denote the conditional distribution of $\mathbf{x}$ in the target class $\mathbf{y}^*$, where θ denotes the parameters of model M_θ. Then for $\mathbf{x}'$ to be considered a conformal counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^*$.*

In words, conformal counterfactuals conform with what the predictive model has learned about the input data $\mathbf{x}$. Since this definition works with distributional properties, it explicitly accounts for the multiplicity of explanations we discussed earlier. To assess counterfactuals with respect to Definition 2.2, we need to be able to quantify the posterior conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}^*)$. This is very much at the core of our proposed methodological framework, which reconciles the notions of plausibility and model faithfulness and which we will introduce next.

3 Methodological Framework

The primary objective of this work has been to develop a methodology for generating maximally plausible counterfactuals under minimal intervention. Our proposed framework is based on the premise that explanations should be plausible but not plausible at all costs. Energy-Constrained Conformal Counterfactuals (ECCCo) achieve this goal in two ways: firstly, they rely on the Black Box itself for the generative task; and, secondly, they involve an approach to predictive uncertainty quantification that is model-agnostic.

3.1 Quantifying the Model's Generative Property

Recent work by Grathwohl et al. [7] on Energy Based Models (EBM) has pointed out that there is a 'generative model hidden within every standard discriminative model'. The authors show that we can draw samples from the posterior conditional distribution $p_\theta(\mathbf{x}|\mathbf{y})$ using Stochastic Gradient Langevin Dynamics (SGLD). The authors use this insight to train classifiers jointly for the discriminative task using standard cross-entropy and the generative task using SGLD. They demonstrate empirically that among other things this improves predictive uncertainty quantification for discriminative models. Our findings in this work suggest that Joint Energy Models (JEM) also tend to yield more plausible Counterfactual Explanations. Based on the definition of plausible counterfactuals (Definition 2.1) this is not surprising.

Crucially for our purpose, one can apply their proposed sampling strategy during inference to essentially any standard discriminative model. Even models that are not explicitly trained for the joint objective learn about the distribution of inputs X by learning to make conditional predictions about the output y. We can leverage this observation to quantify the generative property of the Black Box model itself. In particular, note that if we fix $\mathbf{y}$ to our target value $\mathbf{y}^*$, we can sample from $p_\theta(\mathbf{x}|\mathbf{y}^*)$ using SGLD as follows,

$$\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j - \frac{\epsilon^2}{2}\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*) + \epsilon\mathbf{r}_j, \quad j = 1, ..., J \tag{2}$$

where $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ is typically polynomially decayed. The term $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*)$ denotes the energy function where we use $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*) = -M_\theta(\mathbf{x}_j)[\mathbf{y}^*]$, that is the negative logit corresponding to the target class label $\mathbf{y}^*$. Generating multiple samples in this manner yields an empirical distribution $\hat{\mathbf{X}}_{\theta,\mathbf{y}^*}$ that we use in our search for plausible counterfactuals, as discussed in more detail below. Appendix A provides additional implementation details for any tasks related to energy-based modelling.

3.2 Quantifying the Model's Predictive Uncertainty

To quantify the model's predictive uncertainty we use Conformal Prediction (CP), an approach that has recently gained popularity in the Machine Learning community [3, 14]. Crucially for our intended application, CP is model-agnostic and can be applied during inference without placing any restrictions on model training. Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty estimates by repeatedly sifting through the training data or a dedicated calibration dataset. Conformal classifiers produce prediction sets for individual inputs that include all output labels that can be reasonably attributed to the input. These sets tend to be larger for inputs that do not conform with the training data and are therefore characterized by high predictive uncertainty.

In order to generate counterfactuals that are associated with low predictive uncertainty, we use a smooth set size penalty introduced by Stutz et al. [25] in the context of conformal training:

$$\Omega(C_\theta(\mathbf{x};\alpha)) = \max\left(0, \sum_{\mathbf{y}\in\mathcal{Y}} C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha) - \kappa\right) \tag{3}$$

Here, $\kappa \in \{0, 1\}$ is a hyper-parameter and $C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha)$ can be interpreted as the probability of label $\mathbf{y}$ being included in the prediction set.

In order to compute this penalty for any black-box model we merely need to perform a single calibration pass through a holdout set $\mathcal{D}_{\text{cal}}$. Arguably, data is typically abundant and in most applications, practitioners tend to hold out a test data set anyway. Consequently, CP removes the restriction on the family of predictive models, at the small cost of reserving a subset of the available data for calibration. This particular case of conformal prediction is referred to as Split Conformal Prediction (SCP) as it involves splitting the training data into a proper training dataset and a calibration dataset. Details concerning our implementation of Conformal Prediction can be found in Appendix B.

3.3 Energy-Constrained Conformal Counterfactuals (ECCCo)

Our framework for generating ECCCos combines the ideas introduced in the previous two subsections. Formally, we extend Equation 1 as follows,

$$\begin{aligned}\mathbf{Z}' = \arg\min_{\mathbf{Z}' \in \mathcal{Z}^M} \{ &\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^*) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) \\ &+ \lambda_2 \text{dist}(f(\mathbf{Z}'), \hat{\mathbf{x}}_\theta) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \}\end{aligned} \tag{4}$$

where $\hat{\mathbf{x}}_\theta$ denotes samples generated using SGLD (Equation 2) and $\text{dist}(\cdot)$ is a generic term for a distance metric. Our default choice for $\text{dist}(\cdot)$ is the L1 Norm, or Manhattan distance, since it induces sparsity.

The first two terms in Equation 4 correspond to the counterfactual search objective defined in Wachter et al. [29] which merely penalises the distance of counterfactuals from their factual values. The additional two penalties in ECCCo ensure that counterfactuals conform with the model's generative property and lead to minimally uncertain predictions, respectively. The hyperparameters $\lambda_1, ..., \lambda_3$ can be used to balance the different objectives: for example, we may choose to incur larger deviations from the factual in favour of faithfulness with the model's generative property by choosing lower values of λ_1 and relatively higher values of λ_2. Figure 3 illustrates this balancing act for an example involving synthetic data: vector fields indicate the direction of gradients with respect to the different components our proposed objective function (Equation 4).

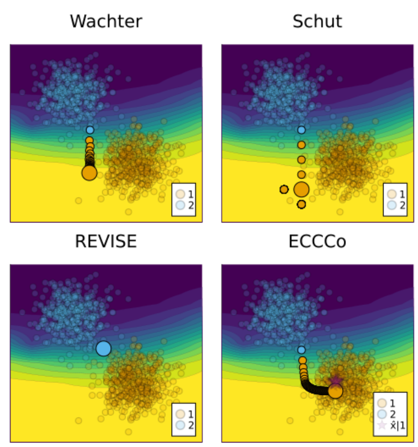

Figure 3: [PLACEHOLDER] Vector fields indicating the direction of gradients with respect to the different components of the ECCCo objective (Equation 4).

Algorithm 1: Generating ECCCos (For more details, see Appendix C)

Input: $\mathbf{x}, \mathbf{y}^*, M_\theta, f, \Lambda, \alpha, \mathcal{D}, T, \eta, n_\mathcal{B}, N_\mathcal{B}$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^*$
Output: $\mathbf{x}'$

```
Initialize z' ← f^{-1}(x)
Generate buffer B of N_B conditional samples x̂_θ|y* using SGLD (Equation 2)
Run SCP for M_θ using D
Initialize t ← 0
while not converged or t < T do
    x̂_{θ,t} ← rand(B, n_B)
    z' ← z' − η∇_{z'} L(z', y*, x̂_{θ,t}; Λ, α)
    t ← t + 1
end while
x' ← f(z')
```

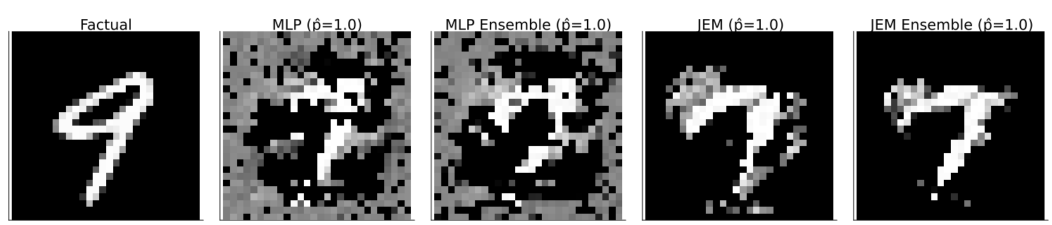

Figure 4: [SUBJECTO TO CHANGE] Original image (left) and ECCCos for turning an 8 (eight) into a 3 (three) for different Black Boxes from left to right: Multi-Layer Perceptron (MLP), Ensemble of MLPs, Joint Energy Model (JEM), Ensemble of JEMs.

The entire procedure for Generating ECCCos is described in Algorithm 1. For the sake of simplicity and without loss of generality, we limit our attention to generating a single counterfactual $\mathbf{x}' = f(\mathbf{z}')$ where in contrast to Equation 4 $\mathbf{z}'$ denotes a 1-dimensional array containing a single counterfactual state. That state is initialized by passing the factual $\mathbf{x}$ through the encoder f^{-1} which in our case corresponds to a simple feature transformer, rather than the encoder part of VAE as in REVISE [9]. Next, we generate a buffer of $N_{\mathcal{B}}$ conditional samples $\hat{\mathbf{x}}_\theta|\mathbf{y}^*$ using SGLD (Equation 2) and conformalise the model M_θ through Split Conformal Prediction on training data $\mathcal{D}$.

Finally, we search counterfactuals through gradient descent. Let $\mathcal{L}(\mathbf{z}', \mathbf{y}^*, \hat{\mathbf{x}}_{\theta,t}; \Lambda, \alpha)$ denote our loss function defined in Equation 4. Then in each iteration, we first randomly draw $n_{\mathcal{B}}$ samples from the buffer $\mathcal{B}$ before updating the counterfactual state $\mathbf{z}'$ by moving in the negative direction of that loss function. The search terminates once the convergence criterium is met or the maximum number of iterations T has been exhausted. Note that the choice of convergence criterium has important implications on the final counterfactual (for more detail on this see Appendix C).

Figure 4 presents ECCCos for the MNIST example from Section 2 for various black-box models of increasing complexity from left to right: a simple Multi-Layer Perceptron (MLP); an Ensemble of MLPs, each of the same architecture as the single MLP; a Joint Energy Model (JEM) based on the same MLP architecture; and finally, an Ensemble of these JEMs. Since Deep Ensembles have an improved capacity for predictive uncertainty quantification and JEMs are explicitly trained to learn plausible representations of the input data, it is intuitive to see that the plausibility of counterfactuals visibly improves from left to right. This provides some first anecdotal evidence that ECCCos achieve plausibility while maintaining faithfulness to the Black Box.

4 Empirical Analysis

In this section, we present our empirical analysis and findings. Our goal is to shed line on the following questions:

Research Question 4.1 (Feasibility)**.** *Is it feasible to generate plausible Counterfactual Explanations through ECCCo without relying on surrogate models?*

Research Question 4.2 (Drivers)**.** *Subject to feasibility, what drives the performance of ECCCo? Is it sufficient to rely on energy-based modelling to quantify the model's generative property? Is it sufficient to rely on conformal prediction to quantify the model's uncertainty?*

In the following, we first briefly describe our evaluation framework and data, before presenting and discussing our results.

4.1 Key Evaluation Measures

Above we have defined plausibility (Definition 2.1) and faithfulness (Definition 2.2) for Counterfactual Explanations. These are the main criteria we use to evaluate counterfactuals in this study. In order to quantify the plausibility of counterfactuals we use a slightly adapted version of the implausibility metric proposed in Guidotti [8]. Formally, for a single counterfactual, we define implausibility as follows,

$$\text{impl} = \frac{1}{|\mathbf{x} \in \mathbf{X}_{\mathbf{y}^*}|} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbf{y}^*}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{5}$$

where $\mathbf{X}_{\mathbf{y}^*}$ is a subsample of the training data in the target class $\mathbf{y}^*$. This gives rise to a very similar evaluation measure for unfaithfulness. We merely swap out the subsample of individuals in the target class for a subset $\hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^*}$ of the generated conditional samples:

$$\text{unfaith} = \frac{1}{|\mathbf{x} \in \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^*}|} \sum_{\mathbf{x} \in \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^*}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{6}$$

Specifically, we form this subset based on the n_E generated samples associated with the lowest energy.

While we focus on these key evaluation metrics in the body of this paper, we also sporadically discuss outcomes with respect to other common measures used to evaluate the validity, proximity and sparsity of counterfactuals. Details can be found in Appendix D.

Table 1: Results for synthetic datasets. Standard deviations across samples are shown in parentheses. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (Wachter).

		Circles		Linearly Separable		Moons	
Model	Generator	Non-conformity ↓	Implausibility ↓	Non-conformity ↓	Implausibility ↓	Non-conformity ↓	Implausibility ↓
JEM	ECCCo	**0.63 (1.58)**	1.44 (1.37)	0.10 (0.06)**	0.19 (0.03)**	**0.57 (0.58)****	**1.29 (0.21)***
	ECCCo (no CP)	0.64 (1.61)	1.45 (1.38)	**0.10 (0.07)****	**0.19 (0.03)****	0.63 (0.64)*	1.30 (0.21)*
	ECCCo (no EBM)	1.41 (1.51)	1.50 (1.38)	0.37 (0.28)	0.38 (0.26)	1.73 (1.34)	1.73 (1.42)
	REVISE	0.96 (0.32)*	**0.95 (0.32)***	0.41 (0.02)**	0.41 (0.01)**	1.59 (0.55)	1.55 (0.20)
	Schut	0.99 (0.80)	1.28 (0.53)	0.66 (0.23)	0.66 (0.22)	1.55 (0.61)	1.42 (0.16)*
	Wachter	1.41 (1.50)	1.51 (1.35)	0.44 (0.16)	0.44 (0.15)	1.77 (0.48)	1.67 (0.15)
MLP	ECCCo	**0.37 (0.65)****	1.30 (0.68)	**0.03 (0.02)****	0.69 (0.10)	1.68 (1.74)	2.02 (0.86)
	ECCCo (no CP)	0.50 (0.85)*	1.28 (0.66)	**0.03 (0.02)****	0.68 (0.10)	**1.34 (1.66)**	2.11 (0.88)
	ECCCo (no EBM)	2.00 (1.46)	1.83 (1.00)	1.25 (0.87)	1.84 (1.10)	2.98 (1.89)	2.29 (1.75)
	REVISE	1.16 (1.05)	**0.95 (0.32)***	1.10 (0.10)	**0.40 (0.01)****	2.46 (1.05)	**1.54 (0.27)***
	Schut	1.60 (1.15)	1.24 (0.44)	0.81 (0.10)*	0.47 (0.24)	2.71 (1.15)	1.62 (0.42)
	Wachter	1.67 (1.05)	1.31 (0.43)	0.94 (0.11)	0.44 (0.15)	2.95 (1.42)	1.84 (1.33)

Table 2: Results for real-world datasets. Standard deviations across samples are shown in parentheses. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (Wachter).

		California Housing		GMSC		MNIST	
Model	Generator	Non-conformity ↓	Implausibility ↓	Non-conformity ↓	Implausibility ↓	Non-conformity ↓	Implausibility ↓
JEM	ECCCo	**236.79 (51.16)**	39.78 (3.18)	**41.65 (17.24)****	40.57 (8.74)**	**116.09 (30.70)****	281.33 (41.51)**
	REVISE	284.51 (52.74)	**5.58 (0.81)****	74.89 (15.82)**	**6.01 (5.75)****	348.74 (65.65)**	**246.69 (36.69)****
	Schut	263.55 (60.56)	8.00 (2.03)	76.23 (15.54)**	6.02 (0.72)**	355.58 (64.84)**	270.06 (40.41)**
	Wachter	274.55 (51.17)	7.32 (1.80)	146.02 (64.48)	128.93 (74.00)	694.08 (50.86)	630.99 (33.01)
JEM Ensemble	ECCCo	**249.44 (58.53)**	35.09 (5.56)	**26.55 (12.94)****	33.65 (8.33)**	**89.89 (27.26)****	240.59 (37.41)**
	REVISE	268.45 (66.87)	**5.44 (0.74)****	52.47 (14.12)**	6.69 (3.37)**	292.52 (53.13)**	**240.50 (35.73)****
	Schut	279.38 (63.23)	7.64 (1.47)	56.34 (15.00)**	**6.27 (1.06)****	319.45 (59.02)**	266.80 (40.46)**
	Wachter	268.59 (68.66)	7.16 (1.46)	125.72 (70.80)	126.55 (93.75)	582.52 (58.46)	543.90 (44.24)
MLP	ECCCo	**230.92 (48.86)**	37.53 (5.40)	**46.90 (15.80)****	37.78 (8.40)**	**212.45 (36.70)****	649.63 (58.80)
	REVISE	281.10 (53.01)	**5.34 (0.67)****	81.08 (19.53)**	**4.60 (0.72)****	839.79 (77.14)*	**244.33 (38.69)****
	Schut	285.12 (56.00)	6.48 (1.18)**	90.67 (20.80)**	5.56 (0.81)**	842.80 (82.01)*	264.94 (42.18)**
	Wachter	262.50 (56.87)	9.21 (10.41)	191.68 (30.86)	200.23 (15.05)	982.32 (61.81)	561.23 (45.08)
MLP Ensemble	ECCCo	**212.47 (59.27)***	38.17 (6.18)	**74.65 (144.69)***	71.87 (145.19)	**162.21 (36.21)****	587.65 (95.01)
	REVISE	284.65 (49.52)	**5.64 (1.13)***	80.90 (14.59)**	**5.20 (1.52)****	741.30 (125.98)*	**242.76 (41.16)****
	Schut	269.19 (46.08)	7.30 (1.94)	85.63 (19.15)**	6.00 (0.99)**	754.35 (132.26)	266.94 (42.55)**
	Wachter	278.09 (73.65)	7.32 (1.75)	220.05 (17.41)	203.65 (14.77)	871.09 (92.36)	536.24 (48.73)

As noted by Guidotti [8], these distance-based measures are simplistic and more complex alternative measures may ultimately be more appropriate for the task. For example, we considered using statistical divergence measures instead. This would involve generating not one but many counterfactuals and comparing the generated empirical distribution to the target distributions in Definitions 2.1 and 2.2. While this approach is potentially more rigorous, generating enough counterfactuals is not always practical.

4.2 Data

4.3 Results

See Table 2

5 Discussion

5.1 Key Insights

Consistent with the findings in Schut et al. [23], we have demonstrated that predictive uncertainty estimates can be leveraged to generate plausible counterfactuals. Interestingly, Schut et al. [23] point out that this finding — as intuitive as it is — may be linked to a positive connection between the generative task and predictive uncertainty quantification. In particular, Grathwohl et al. [7] demonstrate that their proposed method for integrating the generative objective in training yields models that have improved predictive uncertainty quantification. Since neither Schut et al. [23] nor we have employed any surrogate generative models, our findings seem to indicate that the positive connection found in Grathwohl et al. [7] is bidirectional.

5.2 Limitations

- BatchNorm does not seem compatible with JEM
- Coverage and temperature impacts CCE in somewhat unpredictable ways
- It seems that models that are not explicitly trained for generative task, still learn it implictly
- Batch size seems to impact quality of generated samples (at inference, but not so much during JEM training)
- SGLD takes time
- REVISE has benefit of lower dimensional space
- For MNIST it seems that ECCCo is better at reducing pixel values than increasing them (better at erasing than writing)
- JEMs are more difficult to train
- There is a tradeoff: higher cost vs. higher faithfulness/plausibility
- Results are sensitive to choices of penalty strength and step size
- Counterfactuals may end up looking fairly homogenous
- For MNIST data we found CP to have little effect
- JEMs themselves are sensitive to scale
- ECCCo can backfire, in case generative property of model is poor

6 Conclusion

References

[1] Patrick Altmeyer. Conformal Prediction in Julia. URL `https://www.paltmeyer.com/blog/posts/conformal-prediction/`.

[2] Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia Liem. Endogenous Macrodynamics in Algorithmic Recourse. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[3] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. 2021.

[4] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating Robustness of Counterfactual Explanations. Technical report, arXiv. URL `http://arxiv.org/abs/2103.02354`. arXiv:2103.02354 [cs] type: article.

[5] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.

[7] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. March 2020. URL `https://openreview.net/forum?id=Hkxzx0NtDB`.

[8] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. ISSN 1573-756X. doi: 10.1007/s10618-022-00831-6. URL `https://doi.org/10.1007/s10618-022-00831-6`.

[9] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. 2019.

[10] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. 2020.

[11] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[13] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. Technical report, arXiv. URL `http://arxiv.org/abs/1912.03277`. arXiv:1912.03277 [cs, stat] type: article.

[14] Valery Manokhin. Awesome conformal prediction.

[15] Merriam-Webster. "fidelity". URL `https://www.merriam-webster.com/dictionary/fidelity`.

[16] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

[17] 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*, pages 607–617, 2020.

[18] Kevin P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press.

[19] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. 2021.

[20] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768*, 2022.

[21] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[23] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR, 2021.

[24] Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. Counterfactual Explanations for Arbitrary Regression Models. 2021.

[25] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning Optimal Conformal Classifiers. May 2022. URL `https://openreview.net/forum?id=t8O-4LKFVx`.

[26] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse. 2021.

[27] 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, 2019.

[28] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. 2020.

[29] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[30] Andrew Gordon Wilson. The case for Bayesian deep learning. 2020.

Appendices

A JEM

While $\mathbf{x}_J$ is only guaranteed to distribute as $p_\theta(\mathbf{x}|\mathbf{y}^*)$ if $\epsilon \to 0$ and $J \to \infty$, the bias introduced for a small finite ϵ is negligible in practice [18, 7]. While Grathwohl et al. [7] use Equation 2 during training, we are interested in applying the conditional sampling procedure in a post-hoc fashion to any standard discriminative model.

B Conformal Prediction

The fact that conformal classifiers produce set-valued predictions introduces a challenge: it is not immediately obvious how to use such classifiers in the context of gradient-based counterfactual search. Put differently, it is not clear how to use prediction sets in Equation 1. Fortunately, Stutz et al. [25] have recently proposed a framework for Conformal Training that also hinges on differentiability. Specifically, they show how Stochastic Gradient Descent can be used to train classifiers not only for the discriminative task but also for additional objectives related to Conformal Prediction. One such objective is *efficiency*: for a given target error rate α, the efficiency of a conformal classifier improves as its average prediction set size decreases. To this end, the authors introduce a smooth set size penalty defined in Equation 3 in the body of this paper

Formally, it is defined as $C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha) := \sigma\left((s(\mathbf{x}_i,\mathbf{y}) - \alpha)T^{-1}\right)$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used for temperature scaling [25].

Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty estimates by repeatedly sifting through the data. It can be used to generate prediction intervals for regression models and prediction sets for classification models [1]. Since the literature on CE and AR is typically concerned with classification problems, we focus on the latter. A particular variant of CP called Split Conformal Prediction (SCP) is well-suited for our purposes, because it imposes only minimal restrictions on model training.

Specifically, SCP involves splitting the data $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n}$ into a proper training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. The former is used to train the classifier in any conventional fashion. The latter is then used to compute so-called nonconformity scores: $\mathcal{S} = \{s(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{D}_{\text{cal}}}$ where $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$ is referred to as *score function*. In the context of classification, a common choice for the score function is just $s_i = 1 - M_\theta(\mathbf{x}_i)[\mathbf{y}_i]$, that is one minus the softmax output corresponding to the observed label $\mathbf{y}_i$ [3].

Finally, classification sets are formed as follows,

$$C_\theta(\mathbf{x}_i;\alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \le \hat{q}\} \tag{7}$$

where $\hat{q}$ denotes the $(1-\alpha)$-quantile of $\mathcal{S}$ and α is a predetermined error rate. As the size of the calibration set increases, the probability that the classification set $C(\mathbf{x}_{\text{test}})$ for a newly arrived sample $\mathbf{x}_{\text{test}}$ does not cover the true test label $\mathbf{y}_{\text{test}}$ approaches α [3].

Observe from Equation 7 that Conformal Prediction works on an instance-level basis, much like Counterfactual Explanations are local. The prediction set for an individual instance $\mathbf{x}_i$ depends only on the characteristics of that sample and the specified error rate. Intuitively, the set is more likely to include multiple labels for samples that are difficult to classify, so the set size is indicative of predictive uncertainty. To see why this effect is exacerbated by small choices for α consider the case of $\alpha = 0$, which requires that the true label is covered by the prediction set with probability equal to 1.

C Conformal Prediction

D Results