Counterfactual Explanations for Model Ensembles Using Entropic Risk Measures

Erfaun Noorani
University of Maryland
College Park, United States
enoorani@umd.edu

Pasan Dissanayake
University of Maryland
College Park, United States
pasand@umd.edu

Faisal Hamman
University of Maryland
College Park, United States
fhamman@umd.edu

Sanghamitra Dutta
University of Maryland
College Park, United States
sanghamd@umd.edu

ABSTRACT

Counterfactual explanations indicate the smallest change in input that can translate to a different outcome for a machine learning model. Counterfactuals have generated immense interest in high-stakes applications such as finance, education, hiring, etc. In several use-cases the decision-making process often relies on an ensemble of models rather than just one. Despite significant research on counterfactuals for one model, the problem of generating a single counterfactual explanation for an ensemble of models has received limited interest. Each individual model might lead to a different counterfactual, whereas trying to find a counterfactual accepted by all models might significantly increase cost (effort). We propose a novel strategy to find the counterfactual for an ensemble of models using the perspective of entropic risk measure. Entropic risk is a convex risk measure that satisfies several desirable properties. We incorporate our proposed risk measure into a novel constrained optimization to generate counterfactuals for ensembles that stay valid for several models. The main significance of our measure is that it provides a knob that allows for the generation of counterfactuals that stay valid under an adjustable fraction of the models. We also show that a limiting case of our entropic-risk-based strategy yields a counterfactual valid for all models in the ensemble (worst-case min-max approach). We study the trade-off between the cost (effort) for the counterfactual and its validity for an ensemble by varying degrees of risk aversion, as determined by our risk parameter knob. We validate our performance on real-world datasets.

KEYWORDS

Counterfactual Explanations; Ensemble Models; Entropic Risk

ACM Reference Format:
Erfaun Noorani, Pasan Dissanayake, Faisal Hamman, and Sanghamitra Dutta. 2025. Counterfactual Explanations for Model Ensembles Using Entropic Risk Measures. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

The widespread adoption of machine learning models in critical decision-making, from education to finance [8, 12, 14, 25], has raised concerns about the explainability of these models [45, 52]. To address this issue, a recently-emerging category of explanations that has gained tremendous interest is: *counterfactual explanation* [63]. Given a specific data point and a model, a counterfactual explanation (also referred to as "counterfactual") is a feature vector leading to a different model outcome. Typically, counterfactuals are based on the *closest* point on the other side of the decision boundary of the model, also referred to as the closest counterfactual (also see surveys [4, 37, 50, 65]). For example, in automated lending, a counterfactual can inform a denied loan applicant about specific changes such as increasing collateral by 10K can lead to loan approval.

In several applications, multiple models with distinct architectures and training processes can be trained for a specific prediction task, potentially yielding different predictions for the same input. Ensemble models in machine learning combine the predictions of multiple such models to improve overall prediction (goes way back to random forests [9]; also used for neural networks [24, 40]). Ensembles aggregate insights from a set of models, often leading to more reliable outcomes. Evidence suggests averaging ensembles work because each model will make some errors independent of one another due to the high variance inherent in neural networks with large number of parameters [24, 40]. Additionally, ensembling might also be beneficial when different models capture distinct facets of the input data, akin to how multiple interviewers might offer varied perspectives on a candidate. Sometimes, ensembling several smaller models has also been found to be more useful than training one large model [39]. By leveraging ensembles, machine learning systems can mitigate biases and errors inherent in individual models, enhancing performance. Ensembling techniques are also commonly employed to address the issue of predictive multiplicity where multiple equally-well-performing models lead to different predictions on certain data points [6, 28, 64].

Providing recourse for model ensembles can be challenging since each individual model would lead to a different closest counterfactual. Finding a single closest counterfactual with a reasonable cost that remains valid across all models in the ensemble is nontrivial. This *worst-case* approach is overly conservative leading to counterfactuals that are potentially quite far from the original point; sometimes, it may not even identify any counterfactual if there is

no region where all the acceptance regions overlap. To make sure counterfactual explanations are useful and actionable to the users, we not only need them to be close but also require them to stay valid under a reasonable portion of the models within the ensemble. In general, it might even be impossible to guarantee the existence of a counterfactual that stays valid for all possible models in the ensemble. However, one might be able to ensure acceptance for a subset of models. This generates a need for an adjustable knob to obtain counterfactuals that accommodate varying fraction of models within the ensemble.

Balancing the cost and validity of counterfactuals across an ensemble of models is crucial due to potential disparities in the cost of counterfactuals across the ensemble. Understanding this trade-off will allow practitioners to tailor explanations for specific constraints and needs effectively. By providing a flexible mechanism to adjust this trade-off, machine learning systems can better manage the complexity of ensemble scenarios, ensuring that counterfactuals are both feasible and aligned with practical considerations.

Our Contributions: In this work, we propose a novel entropic risk measure to quantify the reliability of the counterfactual for an ensemble of models. Entropic risk is a convex risk measure and satisfies several desirable properties. Furthermore, we incorporate our proposed risk measure in the generation of reliable counterfactuals. A significance of our measure is its ability to establish a unifying connection between a worst-case (min-max optimization) approach and risk-constrained counterfactuals. Our proposed measure is rooted in large deviation theory and mathematical finance [20]. Our contributions can be concisely listed as follows:

An Entropic Risk Measure for Counterfactuals in Model Ensembles: We propose a novel entropic risk measure to quantify the reliability of counterfactuals in an ensemble. Our measure is convex and satisfies several desirable properties. It has a "knob"– the risk parameter– that can be adjusted to trade off between risk-constrained and worst-case approaches. While risk-constrained accounts for general ensemble in an expected sense, the worst-case scenario prioritizes the worst model within the ensemble, thus having a higher cost.

Formulation of Constrained-Optimization to Find Counterfactuals for Model Ensembles: Our proposed entropic risk measure enables us to obtain risk-constrained reliable counterfactuals (see constrained optimization P3). The significance of our strategy is that it enables one to tune "how much" a user wants to prioritize the worst model by trading off cost (effort). Since calculating expectations over ensembles, especially with infinitely many models, can be impractical, we use empirical averages instead. This approach forms the basis of our main formulation in P4, which is crucial for developing our algorithm.

Connection to Min-Max Optimization: We show that the worst-case approaches are, in fact, a limiting case of our entropic-risk-based approach (see Theorem 1). The extreme value of the knob (risk parameter) maps our measure back to a min-max (adversarial) approach. By establishing this connection, we show that our proposed measure is not postulated and stems from the mathematical connection with worst-case analysis.

Experimental Results: We include an algorithm that leverages our relaxed risk measure and finds counterfactuals for model ensembles. We provide a trade-off analysis between the cost (distance) and the validity of the counterfactual on real-world datasets, namely, HELOC [19], German Credit [26], and Adult Income [5].

Notably, in agent-based systems, agents often operate in environments with incomplete or uncertain information and must make decisions that are robust to varying strategies of other agents. For example, agents make decisions based on models predicting the behavior of other agents, but these models can differ due to varying assumptions or strategies. In a similar vein, our approach uses counterfactual reasoning with an ensemble of models to explore alternative scenarios, helping agents understand potential outcomes under different conditions. By ensuring counterfactuals are robust across a range of models, agents can make more reliable decisions despite uncertainty and diverse behaviors in the system.

Related Works: Counterfactuals have been extensively studied in the literature, with numerous papers exploring methodologies and applications within the context of single models (see surveys [4, 37, 50, 65]). However, ensemble models are widely recognized and extensively used in machine learning for their effectiveness in improving predictive performance. By combining multiple base models—often diverse in architecture or training data—ensemble methods harness the collective wisdom of individual models to produce more accurate and reliable predictions. Popular techniques like bagging, boosting, and stacking leverage this diversity to mitigate biases, reduce variance, and enhance overall model generalization. For a survey on ensembles, we refer the readers to [48].

Despite significant research on counterfactuals for one model, the problem of finding counterfactuals for an ensemble has received limited interest. A well-studied research direction involves robust counterfactuals that account for changes in models [1, 7, 18, 21–23, 33, 34, 58, 61, 62]. Other works examine counterfactual robustness to small feature variations (noisy implementation) [10, 15, 17, 42, 46, 53, 57] and and distribution shifts [13, 47, 58]. We refer to recent surveys on robust counterfactual explanations [36, 51]. Applying such techniques to ensembles, where the constituent models are known a priori, may be considered excessive. Unlike scenarios where the model identities are unknown or variable, our problem involves a fixed ensemble, allowing for more targeted approaches.

Closely related is robustness under model multiplicity [21, 22, 38, 43, 56]. [56] suggests that counterfactuals within the data manifold are more resilient against model multiplicity compared to closest counterfactuals. [21, 22] introduce a stability measure to quantify the robustness of counterfactuals under model multiplicity and provide probabilistic guarantees. [38] proposes using Pareto improvement, a multi-objective optimization to generate robust counterfactuals under model multiplicity. [43] and [35] propose approaches to compute robust counterfactuals that hold across all models within an ensemble of neural networks. Our contribution lies in first developing a rigorous quantification of reliability that is specifically tailored to generate counterfactual explanations for ensembles. We use entropic risk measures that arrive with a knob to tradeoff cost and validity of counterfactuals across the models in the ensemble. Our contribution lies in developing a methodology specifically tailored to generate reliable counterfactual explanations for ensembles using entropic risk measures with a knob that trades off the cost and overall validity of counterfactuals, aiming to provide counterfactual with reasonable costs that stay valid under as many models as possible in the ensemble.

Entropic risk measure has been the cornerstone of risk-sensitive control (see [2, 3, 29–32, 41, 54, 55, 60]) and risk-sensitive Markov decision processes (see [27]). The connection between risk-sensitive control and robust control has been shown in its full generality in [32], establishing that the entropic risk measure emerges from the mathematical analysis of H-infinity output robust control for general non-linear systems and has been used to trade off robustness and performance in feedback control design. Further analytical development of such mathematical analysis for financial applications has been studied extensively; see [20] and references therein.

2 PRELIMINARIES

Here, we provide some contextual details, definitions, and background materials, and set our notation. We consider machine learning models $m \in \mathcal{M}$, where $\mathcal{M}$ is a non-empty set of ensemble models for binary classification that takes an input value $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and outputs a probability between 0 and 1. Let $\mathcal{S} = \{x_i \in \mathcal{X}\}_{i=1}^n$ be a dataset of n independent and identically distributed data points generated from an unknown density over $\mathcal{X}$.

Definition 1 (Closest Counterfactual $C_p(x, m)$)**.** *A closest counterfactual with respect to the model $m(\cdot)$ of a given point $x \in \mathbb{R}^d$ such that $m(x) < 0.5$ is a point $x' \in \mathbb{R}^d$ such that $m(x') \geq 0.5$ and the cost in terms of p-norm $\|x - x'\|_p$ is minimized.*

$$C_p(x, m) = \arg\min_{x' \in \mathbb{R}^d} c(x, x') \quad s.t. \quad m(x') \geq 0.5.$$

For example, norm $p = 1$ results in counterfactuals with as few feature changes as possible, enforcing a sparsity constraint (also referred to as "sparse" counterfactuals [56]).

In this work, our **goal** is to generate a single counterfactual that is accepted by as many models as possible in the ensemble while minimizing the cost. Towards this goal, we propose an entropic risk measure as a systematic measure of the reliability of counterfactuals. Our objective involves: (i) arriving at a measure for the reliability of a counterfactual x under a given ensemble $\mathcal{M}$, that satisfies desirable properties; (ii) establishing the connection between our entropic-risk-based approach and the worst-case approaches, and (iii) showing the algorithmic impacts of our measure by developing a constrained-optimization-based algorithm for generating counterfactuals for model ensembles based on our reliability measure which allows for a tunable knob that allows one to trade off between the cost (effort) and potential validity on multiple models.

3 MAIN RESULTS: RELIABILITY VIA ENTROPIC RISK MEASURE

For a single model, the counterfactual x' would simply be the closest point to the original instance x that lies on the accepted side. We would minimize the ℓ_2-norm (i.e., the "cost" $c(x, x') = \|x - x'\|_2$ remains low). However, when we introduce the ensemble, ensuring that the counterfactual remains valid across a specified fraction of models in the ensemble can often make it move further from x to satisfy this additional requirement. This results in a higher ℓ_2-norm, increasing the cost $c(x, x')$. When generating a counterfactual for a reference model m_r and an ensemble of models, we also seek to ensure its validity across multiple models within the ensemble, defining its *reliability*. The higher the required fraction of models that must validate the counterfactual, the more robust (reliable) it must be. However, this robustness comes at a cost: as more models must agree, the counterfactual tends to shift further from the original instance. A trade-off arises between the counterfactual's distance from x and the robustness constraint. To balance this tradeoff, we formulate a general multi-objective optimization that hedges against the worst-case models while managing both cost and robustness, i.e.,

$$\min_{x' \in \mathbb{R}^d} \left(c(x, x'), \max_{m \in \mathcal{M}} \ell(m(x'))\right) \quad \text{s.t.} \quad m_r(x') \geq 0.5. \tag{P}$$

Here $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the cost of changing an instance x to x', e.g., $c(x, x') = \|x - x'\|_p$, where $1 \leq p \leq \infty$, and $\ell : \mathcal{M} \times \mathcal{X} \to R_+$ is a differentiable loss function that ensures that $m(x')$ is close to the desired value of 1, e.g., $\ell(m(x)) = 1 - m(x)$. We denote the ensemble as $\mathcal{M}$, where any model within the ensemble is represented by m. The reference model of interest, denoted as m_r, is a specific, fixed model within $\mathcal{M}$. In P, neither m nor m_r are treated as random variables. The second objective function $\max_{m \in \mathcal{M}} \ell(m(x'))$ is the worst-case loss over the ensemble set $\mathcal{M}$.

To address a multi-objective optimization problem of this nature, we can seek the Pareto optimal front using established techniques, such as linear scalarization or the epsilon-constraint methods [49]. The linear scalarization approach, for instance, entails solving

$$\min_{x' \in \mathbb{R}^d} \max_{m \in \mathcal{M}} c(x, x') + \lambda \ell(m(x')) \quad \text{s.t.} \quad m_r(x') \geq 0.5 \tag{P1}$$

for different values of λ to generate Pareto optimal solutions (e.g., a relaxed variant of this approach is employed in [62]), meanwhile, the epsilon-constraint method addresses the problem by solving

$$\min_{x' \in \mathbb{R}^d} c(x, x') \quad \text{s.t.} \quad \max_{m \in \mathcal{M}} \ell(m(x')) < \tau, \quad m_r(x') \geq 0.5 \tag{P2}$$

for different values of τ (e.g., a relaxed variant of this approach is employed in [21]).

By varying λ in P1 or τ in P2, different points on the Pareto front can be obtained (also see the book [49]). To see the equivalence of the threshold τ and the multiplier λ, note that the sensitivities of the cost $c(x, x')$ with respect to changes in the threshold τ (evaluated at the optimal x'^*) is the negative of the optimal multiplier (dual variable) λ (for a background on multi-objective optimization, please refer to Appendix B [11]), i.e, $\partial c(x, x'^*)/\partial \tau = -\lambda^*$. Each λ and τ results in a point on the Pareto optimal front of the multi-objective optimization problem [11, 49]. Both P1 and P2 lead to the same Pareto front, and λ and τ can be chosen such that P1 and P2 have the same solutions. The Pareto front characterizes the trade-off between the cost and validity of the counterfactuals.

The worst-case loss $\max_{m \in \mathcal{M}} \ell(m(x'))$ hedges against the worst possible model within the ensemble, but can often lead to somewhat conservative counterfactuals, i.e., ones which are quite well within the boundary and have a high cost (distance). To mitigate this issue, we use a risk measure that allows us to *hedge against the models based on their probability of occurrence*. We assume M is a random model drawn from a probability distribution P over the set of models $\mathcal{M}$. We propose the entropic risk measure as a quantification of reliability for counterfactuals which is defined as follows:

Definition 2. *The entropic risk measure of model M with the risk aversion parameter $\theta > 0$ is denoted by $\rho_\theta^{ent}(\cdot)$ and is given by:*

$$\rho_\theta^{ent}(\ell(M(x'))) := \frac{1}{\theta}\log(\mathbb{E}_{M\sim P}[e^{\theta\ell(M(x'))}]), \quad \theta > 0. \quad (1)$$

The parameter θ is called the risk parameter. A positive risk parameter results in risk-averse behavior. Hence, we refer to a positive risk parameter as the risk-aversion parameter. We show in Theorem 1 that as we increase the risk-aversion parameter, our probabilistic method converges to a worst-case formulation. Definition 2 allows us to reformulate our problem as follows:

$$\min_{x'\in\mathbb{R}^d} c(x,x') \quad \text{s.t.} \quad \rho_\theta^{ent}(\ell(M(x'))) < \tau, \quad m_r(x') \geq 0.5. \quad (P3)$$

3.1 Properties of Entropic Risk Measure

Entropic risk measure is rooted in large deviation theory and is not postulated. This measure enables establishing a connection to worst-case approaches for finding counterfactuals. Taylor's expansion of the exponential shows that the entropic risk measure is the infinite sum of the moments of the distribution. Furthermore, it is well-known [20] that entropic risk measure is a *convex* risk measure and as such, for a positive risk parameter θ>0, satisfies the properties of (1) monotonicity, (2) translation-invariance, and (3) convexity.

(1) **Monotonicity.** For $\ell(M_1(\cdot)) \geq \ell(M_2(\cdot))$,

$$\rho_\theta^{ent}(\ell(M_1(\cdot))) \geq \rho_\theta^{ent}(\ell(M_2(\cdot))).$$

(2) **Translation invariance.** For constant $\alpha\in\mathbb{R}$,

$$\rho_\theta^{ent}(\ell(M(\cdot)) + \alpha) = \rho_\theta^{ent}(\ell(M(\cdot))) + \alpha.$$

(3) **Convexity.** For $\alpha \in [0,1]$,

$$\rho_\theta^{ent}(\alpha\ell(M_1(\cdot)) + (1-\alpha)\ell(M_2(\cdot))) \leq \alpha\rho_\theta^{ent}(\ell(M_1(\cdot))) + (1-\alpha)\rho_\theta^{ent}(\ell(M_2(\cdot))).$$

For the sake of simplicity, consider the choice of cost function $\ell(M(x)) = 1 - M(x)$. Then, the monotonicity implies that a model with greater output probabilities has less risk. The translation invariance implies that adding a constant to the output of the predictor effectively reduces the risk by the same amount. The convexity is quite desirable since it means that the risk for a combined model is lower than the risk for the two of them individually.

To gain a deeper understanding of the risk constraint described in P3, we examine distributions characterized by their analytical Moment Generating Functions (MGFs). Two notable examples are the Uniform and truncated Gaussian distributions. For simplicity, we use the cost function $\ell(M(x'))$=1−$M(x')$. In our formulation, this loss function is minimized, encouraging a counterfactual with a higher predicted value. When using this specific cost function, any value of the threshold τ outside the interval $[0,1]$ renders the problem infeasible. Given these choices for the cost and model distribution, we provide the explicit form of the constraint in P3.

Example 1. *Let the distribution of the output of the models in the ensemble at the counterfactual point, $M(x')$, follow a uniform distribution on a δ-ball around the output of a specific model $m(x')$, i.e., $M(x') \sim \mathcal{U}[m(x')-\delta, m(x')+\delta]$ for some $\delta > 0$. With these choices, the constraint in P3 becomes:*

$$m(x') > (1-\tau) + K_{\delta,\theta}, \quad K_{\delta,\theta} := \frac{1}{\theta}\log\left(\frac{e^{\theta\delta} - e^{-\theta\delta}}{2\theta\delta}\right).$$

For the Uniform distribution, due to the monotonicity of $K_{\delta,\theta}$ with respect to θ, as the value of θ increases, a higher value of $m(x')$ is required to satisfy the constraint. It can be verified that $K_{\delta,\theta}$ in limit of $\theta \to \infty$ is δ. Given this, for the case when $\theta \to \infty$, our constraint becomes $m(x') > 1-\tau+\delta$. As the value of θ approaches to 0, $K_{\delta,\theta}$ approaches 0 and the constraint becomes $m(x') > (1-\tau)$, i.e., finding counterfactual x' with just high $m(x')$.

Example 2 (Truncated Gaussian)**.** *Let the distribution of the output of the models in the ensemble at the counterfactual point, $M(x')$, follow a truncated Gaussian distribution with a mean equal to the output of the original model $m(x')$ and a variance of σ^2 that lies between* 0 *and* 1*. With these choices, the constraint in P3 becomes:*

$$m(x') > (1-\tau) + \theta\frac{\sigma^2}{2} + \frac{1}{\theta}\log(K_\theta),$$
$$K_\theta := \frac{\Phi(\beta+\sigma\theta) - \Phi(\alpha+\sigma\theta)}{\Phi(\beta) - \Phi(\alpha)}$$

where $\alpha := \frac{-\mu}{\sigma}$ and $\beta := \frac{1-\mu}{\sigma}$ and $\Phi(x) = 1/2(1+\mathrm{erf}(x/\sqrt{2}))$. The error function, denoted by erf*, is defined as* $\mathrm{erf}\, z = 2/\sqrt{\pi}\int_0^z e^{-t^2}\,\mathrm{d}t$.

As the θ approaches 0, our constraint becomes $m(x') > 1-\tau$. As the value of θ increases, greater weight is placed on the variance term, emphasizing its importance. In both examples, when the distributions are unknown, determining the precise threshold for model output to satisfy the constraint becomes challenging. This is because higher values are more conservative (less risky), but incur higher costs. To address this challenge, we must devise techniques that do not rely on the explicit knowledge of the distribution, as explored further in the next subsections.

3.2 Connection of Entropic-Risk-Based Approach with Worst-Case Approach

We first establish the connection between our risk-based and the worst-case formulation (getting accepted by all models in the ensemble). The following theorem shows that the worst-case approach is the limiting case of our risk-based method as $\theta \to \infty$.

Theorem 1. *In the limit as the risk-aversion parameter θ approaches infinity, the optimization P3, which involves constraining the entropic risk measure associated with the reliability of models within an ensemble, asymptotically converges to the optimization problem P2, where the constraint pertains to the robustness of the worst model within the same ensemble.*

Theorem 1 shows how the entropic risk measure provides a single parameter (knob) that determines the risk-aversion of the counterfactual and can be used to study the effect of risk-aversion on the behavior of algorithms that generate reliable counterfactuals for an ensemble.

Proof Sketch of Theorem1 : We discuss the proof in Appendix A. The proof uses Vardhan's Lemma presented here. Such connections have been shown in the context of robust and risk-sensitive control and, more recently, risk-sensitive reinforcement learning.

Lemma 1. *[20] Let X be a random variable. The entropic risk measure is a convex risk measure and as such has a dual representation with*

the risk aversion parameter $\theta > 0$ is given by

$$\rho_\theta^{\text{ent}}(X) = \frac{1}{\theta}\log\left(\mathbb{E}_{X\sim P}[e^{\theta X}]\right) = \sup_{Q\ll P}\left\{E_Q[X] - \frac{1}{\theta}D(Q|P)\right\}$$

where $D(Q|P) := \mathbb{E}_Q[\log {}^{dQ}/_{dP}]$ is the Kullback-Libeler (KL) divergence between distributions P and Q, and $Q \ll P$ denotes the distribution Q is absolutely continuous with respect to P.

3.3 Formulation of the Constrained Optimization

We substitute the expectation in the risk measure with a computable empirical mean. This allows us to reformulate the problem as

$$\min_{x'\in\mathbb{R}^d} c(x,x') \tag{P4}$$
$$\text{s.t.} \quad \frac{1}{\theta}\log\left(\frac{1}{N}\sum_{i=1}^{N} e^{(1-\theta m_i(x'))}\right) < \tau, \quad m_r(x') \geq 0.5,$$

where m_i's are the sample models from the ensemble $\mathcal{M}$.

4 EXPERIMENTAL RESULTS

In this section, we experimentally demonstrate the effect of entropic risk minimization on generating counterfactual explanations. To this end, we observe that the risk aversion parameter θ plays a key role in the cost-validity trade-off.

Experimental Setup: We consider an ensemble of 20 models, each with three 128-neuron hidden layers and ReLU activations. The models are trained employing the Adam optimizer for 200 epochs with a batch size of 32. The same model architecture and hyperparameters were used for all the datasets, since it yielded satisfactory levels of accuracy on all of them. We evaluate the proposed method over three publicly available datasets namely HELOC [19], German Credit [26] and Adult Income [5] (see Appendix C for details). Each dataset is split into a training set and a test set. To generate our ensemble $\mathcal{M}$, we train each model m_i on a slightly different subset of the training split, generated by dropping k (a hyperparameter) randomly selected data points prior to training the model. The test split is used for evaluating model accuracies as well as for generating the counterfactuals. Table 1 summarizes dataset-specific details.

Table 1: Experimental setup: Standard deviations are given in parenthesis when applicable.

Property	HELOC	GERMAN	ADULT
Training set size	7844	670	15081
Test set size	2615	330	15081
# Rejected instances	889	186	10216
# Dropped Points (k)	1000	100	1000
Average model accuracy	0.66 (0.01)	0.72 (0.02)	0.82 (0.002)

Algorithm: In the experiments, we solve P4 through a two step process based on gradient descent. First, an ordinary counterfactual x' is generated for a randomly selected reference model m_r from the ensemble, using an existing counterfactual generating method (ℓ_1-norm closest counterfactual in our case). Then the counterfactual is updated until the entropic risk constraint $\rho_\theta^{\text{ent}}(x') < \tau$ is satisfied. This is done through a gradient descent process $x' \leftarrow x' - \eta\nabla_{x'}\rho_\theta^{\text{ent}}(x')$. Counterfactuals are generated only for the instances that were rejected under the said randomly selected model. Note that for some instances, the counterfactual generation method fails to render counterfactuals due to the finite number of gradient descent iterations. A workaround for this issue is to experiment with different hyperparameters such as the gradient descent step size η and the maximum number of iterations. Algorithm 1 presents the counterfactual generation steps concisely.

Algorithm 1 Entropic risk based counterfactual generation

Require: Input instance x, Model ensemble $\mathcal{M}$, $\theta > 0, \tau > 0$, Gradient descent step size η, max_iter $\in \mathbb{Z}^+$.
  Randomly select $\tilde{m} \in \mathcal{M}$.
  Generate ordinary counterfactual $x' \leftarrow C_p(x, \tilde{m})$.
  Initialize $i \leftarrow 0$.
  while $\rho_\theta^{\text{ent}}(x') \geq \tau$ **and** $i <$ max_iter **do**
    $x' \leftarrow x' - \eta\nabla_{x'}\rho_\theta^{\text{ent}}(x')$
    $i \leftarrow i + 1$
  end while
  if $\rho_\theta^{\text{ent}}(x') < \tau$ **then**
    return x' and exit
  else
    return Error (Invalid counterfactual) and exit
  end if

Metrics: We are interested in observing the trade-off between a counterfactual being easy to achieve (low cost) and being valid under an ensemble decision, e.g., majority vote (high validity). In this regard, for each set of parameters θ and τ, we compute the two metrics: (i) *Cost*: the ℓ_1-distance between the counterfactual and the corresponding original instance (ii) *Validity*: the ratio of models in the ensemble with respect to which the counterfactual is valid. These metrics are averaged over all the counterfactuals that satisfy the risk constraint. For comparison, we also include the results for the case $\tau = 1$ which is equivalent to not having any risk constraints (i.e., non-robust). In addition, we compute the average wall clock time taken to generate a counterfactual under each set of parameters θ and τ, which is an indicator of the difficulty in generating an explanation that satisfies the risk constraint.

Hyperparameter Selection: Values for the gradient descent step size η and the number of maximum iterations were selected empirically such that the algorithm would converge to a solution in a reasonable time for most of the input instances. Ranges for θ and τ were selected in a dataset-specific manner. For instance, for HELOC dataset, $\theta \in \{0.1, 1.0, 10.0\}$ results $\rho_\theta^{\text{ent}}(x') < 0.8$ for most of the input instances. Achieving $\rho_\theta^{\text{ent}}(x') < 0.1$ was seen to be difficult. These values were slightly different for the other datasets.

Results and Discussion: In addition to the real-world data, we conducted a synthetic experiment with a 2D dataset, to facilitate easy visualization. Figure 1 demonstrates the results corresponding to one of the input instances. Tables 2, 3, and 4 present the results of the experiment for HELOC, German Credit, and Adult Income datasets, respectively. Observe that the validity increases with increasing θ for a given value of τ, indicating how θ facilitates

Table 2: Experimental results for HELOC dataset. Standard deviations are given in parenthesis.

θ	$\tau = 0.1$		$\tau = 0.3$		$\tau = 0.5$		$\tau = 0.7$		$\tau = 1.0$	
	COST	VAL.	COST	VAL.	COST	VAL.	COST	VAL.	COST	VAL.
0.1	1.38 (1.29)	0.97 (0.06)	1.18 (1.24)	0.92 (0.11)	1.07 (1.22)	0.87 (0.15)	0.98 (1.22)	0.83 (0.19)	0.89 (1.22)	0.76 (0.24)
1.0	1.39 (1.27)	0.97 (0.06)	1.22 (1.24)	0.93 (0.10)	1.10 (1.24)	0.90 (0.13)	1.00 (1.22)	0.84 (0.18)	0.89 (1.22)	0.76 (0.25)
10.0	1.49 (1.22)	0.97 (0.09)	1.41 (1.23)	0.96 (0.09)	1.37 (1.24)	0.95 (0.11)	1.29 (1.25)	0.93 (0.12)	0.88 (1.23)	0.75 (0.25)

Table 3: Experimental results for German Credit dataset. Standard deviations are given in parentheses.

θ	$\tau = 0.3$		$\tau = 0.5$		$\tau = 0.7$		$\tau = 0.9$		$\tau = 1.0$	
	COST	VAL.	COST	VAL.	COST	VAL.	COST	VAL.	COST	VAL.
0.1	2.20 (1.49)	0.87 (0.07)	1.73 (1.48)	0.77 (0.13)	1.45 (1.41)	0.66 (0.20)	1.16 (1.38)	0.59 (0.28)	1.11 (1.38)	0.58 (0.29)
1.0	2.39 (1.51)	0.90 (0.05)	1.91 (1.47)	0.82 (0.10)	1.55 (1.44)	0.71 (0.16)	1.20 (1.38)	0.60 (0.27)	1.11 (1.38)	0.58 (0.29)
10.0	3.46 (1.49)	1.00 (0.00)	3.39 (1.55)	1.00 (0.00)	3.03 (1.51)	0.97 (0.02)	1.71 (1.47)	0.77 (0.12)	1.13 (1.38)	0.59 (0.29)

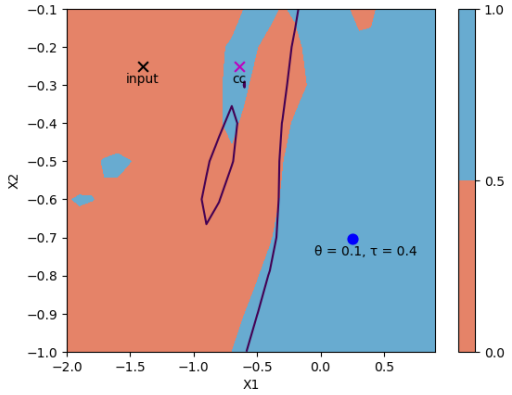

Figure 1: A 2D visualization of the proposed method. Black cross denotes a rejected input instance. Magenta cross is the closest counterfactual ("cc") in terms of ℓ_1−distance, generated w.r.t. reference model m_r. Blue dot represents the entropic-risk-based counterfactual generated with $\theta = 0.1$ and $\tau = 0.4$. Purple line is the decision boundary of another model $m \neq m_r$ in the ensemble. Observe that even though the closest counterfactual is valid under m_r, it is rejected under the other model m. In contrast, the entropic-risk-based counterfactual remains valid w.r.t. both models.

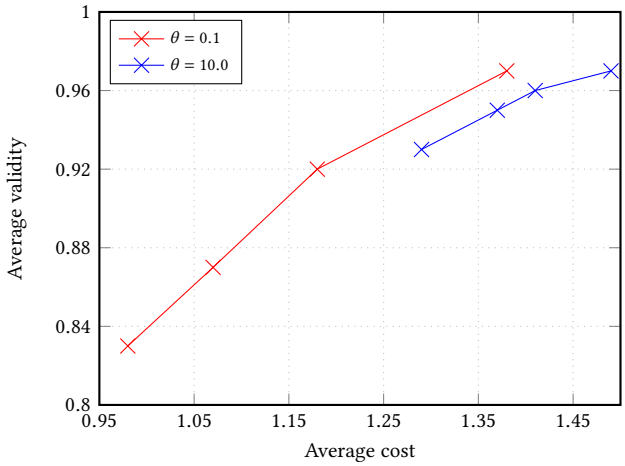

Figure 2: Cost-validity trade-off curves for different θ values on HELOC dataset. Each point on a given curve corresponds to a distinct $\tau \in \{0.1, 0.3, 0.5, 0.7\}$. Cost and validity increase monotonically with increasing τ.

a smooth trade-off between the two metrics. Figure 2 visualizes this trade-off. Tables 5, 6, and 7 show the average wall clock time taken to generate an explanation, which can be considered as a proxy for the difficulty of generating a valid counterfactual. Notice how the time decreases with increasing τ for each value of θ. Furthermore, notice that when the risk constraint is inactive (i.e., when $\tau = 1$, the averaged cost and validity values are almost the same for all values of θ.

Moreover, we observe the effect of ensemble size N on the validity of a counterfactual, for a fixed set of parameters θ and τ. Intuitively, given that the model diversity is constant, the smaller the ensemble size the lower the chances of getting invalidated. Hence, when all the other parameters are constant, the validity should increase with reducing ensemble size. This is indeed the case observed empirically as reported in Table 8. Note how validity increases when the number of models in the ensemble is reduced from 20 to 10, corresponding to each value of τ.

5 CONCLUSION & LIMITATIONS

With our entropic risk measure, we showed that the risk-aversion parameter can be adjusted to balance the cost and validity of counterfactuals by considering the impact of the worst model. We showed that the worst-case approach is a limiting case of our approach based on entropic risk measures. This establishes the connection between our approach and a worst-case approach and explains the nature of the counterfactuals generated by our algorithm. Our research also

Table 4: Experimental results for Adult Income dataset. Standard deviations are given in parentheses.

θ	$\tau = 0.5$		$\tau = 0.7$		$\tau = 0.9$		$\tau = 1.0$	
	COST	VAL.	COST	VAL.	COST	VAL.	COST	VAL.
0.1	1.07 (3.13)	0.95 (0.14)	1.02 (3.13)	0.91 (0.20)	1.01 (3.13)	0.89 (0.22)	1.00 (3.14)	0.89 (0.23)
1.0	1.08 (3.13)	0.96 (0.12)	1.03 (3.13)	0.92 (0.19)	1.01 (3.13)	0.89 (0.22)	1.00 (3.14)	0.89 (0.23)
10.0	1.15 (3.12)	0.99 (0.07)	1.11 (3.13)	0.96 (0.14)	1.03 (3.14)	0.91 (0.20)	1.00 (3.14)	0.88 (0.24)

Table 5: Average wall clock time (in milliseconds) taken to generate a counterfactual – HELOC dataset.

θ	$\tau = 0.7$ (robust)	$\tau = 1.0$ (non-robust)
0.1	646	524
1.0	670	525
10.0	1180	520

Table 6: Average wall clock time (in milliseconds) taken to generate a counterfactual – German Credit dataset.

θ	$\tau = 0.7$ (robust)	$\tau = 1.0$ (non-robust)
0.1	1605	972
1.0	1843	960
10.0	4539	941

Table 7: Average wall clock time (in milliseconds) taken to generate a counterfactual – Adult Income dataset.

θ	$\tau = 0.7$ (robust)	$\tau = 1.0$ (non-robust)
0.1	5278	5189
1.0	5277	5181
10.0	5379	5191

Table 8: Effect of ensemble size N. Values shown for HELOC dataset with $\theta = 1.0$.

τ	N=10		N=20	
	COST	VAL.	COST	VAL.
0.1	1.38 (1.25)	0.96 (0.11)	1.39 (1.27)	0.97 (0.06)
0.3	1.22 (1.24)	0.93 (0.12)	1.22 (1.24)	0.93 (0.10)
0.5	1.11 (1.23)	0.89 (0.14)	1.10 (1.24)	0.90 (0.13)
0.7	1.01 (1.22)	0.83 (0.19)	1.00 (1.22)	0.84 (0.18)

makes a broader connection between the field of explainability and multi-objective optimization through the lens of risk measures.

Another related research direction is model reconstruction [16] where it has been found that counterfactuals lead to more efficient model reconstruction since they are quite close to the decision boundary. In this context, such strategies of generating counterfactuals for ensembles could also have potential applications in defending against such extraction attacks since they are less uniquely tied to a particular model, enhancing privacy.

The integration of machine learning systems into our daily lives has wide-ranging and complex implications. These implications range from economic to societal to ethical and legal considerations, necessitating a comprehensive approach to address the sociotechnical evolution driven by machine learning. While our current work represents a step towards trustworthy adoption, counterfactual explanations also suffer from a multitude of other limitations such as fairness, actionability, and personalization [37, 44, 59]Consider this scenario, when examining a loan approval, a counterfactual suggesting an increase in the value of the applicant's collateral might be perceived as more preferable for an applicant as opposed to a counterfactual suggesting an increase in education level even if they might have the same l_1 cost. Therefore, in our future work, we will explore approaches that incorporate additional metrics beyond explainability and reliability to generate counterfactuals, addressing other relevant considerations.

By ensuring the reliability and trustworthiness of counterfactuals from both user and institutional perspectives, we can foster greater trust in machine learning systems, leading to broader economic benefits and reliable adoption of machine learning in high-stakes applications. However, it is important to recognize that achieving the reliability of counterfactuals for ensembles requires solving computationally more expensive constrained optimization problems compared to the closest counterfactual for a single model. Therefore, future efforts should focus on devising more computationally efficient techniques to overcome this challenge and ensure the sustainability of counterfactual generation approaches.

A PROOF OF THEOREM 1

Theorem 1. *In the limit as the risk-aversion parameter θ approaches infinity, the optimization P3, which involves constraining the entropic risk measure associated with the reliability of models within an ensemble, asymptotically converges to the optimization problem P2, where the constraint pertains to the robustness of the worst model within the same ensemble.*

The proof of Theorem 1 uses the results in Lemma 1 and 2.

Lemma 1. *[20] Let X be a random variable. The entropic risk measure is a convex risk measure and as such has a dual representation with the risk aversion parameter $\theta > 0$ is given by*

$$\rho_\theta^{\text{ent}}(X) = \frac{1}{\theta}\log\left(\mathbb{E}_{X\sim P}[e^{\theta X}]\right) = \sup_{Q\ll P}\left\{E_Q[X] - \frac{1}{\theta}D(Q|P)\right\}$$

where $D(Q|P) := \mathbb{E}_Q\left[\log {}^{dQ}/_{dP}\right]$ is the Kullback-Libeler (KL) divergence between distributions P and Q, and $Q \ll P$ denotes the distribution Q is absolutely continuous with respect to P.

Note that Q is absolutely continuous with respect to P if $Q(x) = 0$ when $P(x) = 0$. This assumption ensures that the KL divergence is finite. Then, we have,

$$\lim_{\theta\to\infty} \rho^{\text{ent}}(X) = \sup_{Q\ll P} \left\{E_Q[X]\right\}. \tag{2}$$

For simplicity, we let both $Q(\tilde{m}) > 0$ and $P(\tilde{m}) > 0$ over the set of models $\mathcal{M}$ which is a compact and bounded set. Next, we show the following result.

Lemma 2. *Let Q be any probability distribution over the set of models $\mathcal{M}$ such that $Q(\tilde{m}) > 0$ everywhere, and $\mathcal{M}$ be a compact and bounded set. Then we have,*

$$\sup_Q \mathbb{E}_Q[\ell(M)] = \max_{m_i\in\mathcal{M}} \ell(m_i)$$

We prove the equality by establishing two directions of the inequality. First, we note that the expected value of a set of values is always less than or equal to its maximum value. Thus,

$$\mathbb{E}_Q[\ell(M)] \le \max_{m\in\mathcal{M}} \ell(m), \quad \forall Q$$

Since it holds for all Q's we have

$$\sup_Q \mathbb{E}_Q[\ell(M)] \le \max_{m\in\mathcal{M}} \ell(m) \tag{3}$$

To prove the reverse direction, let Q_m be a probability distribution such that

$$Q_m(\tilde{m}) = \begin{cases} 1-\delta & \tilde{m} = m \\ \delta_{\tilde{m}} & \tilde{m} \neq m \end{cases}$$

where $\delta_{\tilde{m}} \neq 0$, for all $\tilde{m}\in\mathcal{M}$ and $\delta=\sum_{\tilde{m}\in\mathcal{M},\tilde{m}\neq m} \delta_{\tilde{m}}$. Then, we have

$$\mathbb{E}_{Q_m}[\ell(M)] = (1-\delta)\ell(m) + \sum_{\tilde{m}\in\mathcal{M},\tilde{m}\neq m} \delta_{\tilde{m}}\ell(\tilde{m}), \quad \forall m$$

Thus,

$$\begin{aligned}\sup_Q \mathbb{E}[\ell(M)] &\ge E_{Q_m}[\ell(M)] \\ &= (1-\delta)\ell(m) + \sum_{\tilde{m}\in\mathcal{M},\tilde{m}\neq m} \delta_{\tilde{m}}\ell(\tilde{m}), \quad \forall m\end{aligned}$$

Let $m^* = \arg\max_m \ell(m)$. Then we have,

$$\sup_Q \mathbb{E}[\ell(M)] \ge (1-\delta)\ell(m^*) + \sum_{\tilde{m}\in\mathcal{M},\tilde{m}\neq m^*} \delta_{\tilde{m}}\ell(\tilde{m})$$

By noting that δ can be made arbitrarily small, we have

$$\sup_Q \mathbb{E}[\ell(M)] \ge \max_{m\in\mathcal{M}} \ell(m) - \epsilon(\delta)$$

for an arbitrarily small $\epsilon(\delta) > 0$. Thus the result holds.

The set $\mathcal{M}$ needs to be such that the maximum exists, e.g., a bounded and compact set.

Now using Lemma 2, we have

$$\begin{aligned}\lim_{\theta\to\infty} \rho_\theta^{ent}(\ell(m(x'))) &:= \frac{1}{\theta}\log(\mathbb{E}_{M\sim P}[e^{\theta\ell(M(x)}]) \\ &\overset{(a)}{=} \sup_{Q\in\mathcal{M}_1} \left\{E_Q[\ell(M(x))]\right\} \\ &\overset{(b)}{=} \sup_{m\in\mathcal{M}} \ell(m(x')),\end{aligned}$$

where (a) holds since $\lim_{\theta\to\infty} \rho^{\text{ent}}(X) = \max_{Q\ll P}\left\{E_Q[X]\right\}$ as shown in equation 2 and (b) follows from Lemma 2.

B BACKGROUND ON MULTI-OBJECTIVE OPTIMIZATION

Consider a non-linear programming problem with inequality constraints such as:

$$\min_{x'} c(x, x') \qquad \text{subject to:} \quad R(x, x') \le \tau$$

where c and R are regular enough for the mathematical developments to be valid over the feasible region. It is also assumed that the problem has an optimum. Then the sensitivities of the objective function with respect to the threshold τ can be calculated using the following theorem:

Theorem 2. *[11] Assume that the solution of the above optimization problem is a regular point and that no degenerate inequality constraints exist. Then, the sensitivity of the objective function with respect to the parameter a is given by the gradient of the Lagrangian function*

$$L = c(x, x') + \lambda^T(R(x, x') - \tau)$$

with respect to τ evaluated at the optimal solution x^*, i.e.,

$$\frac{\partial c(x, x^*)}{\partial \tau} = \nabla_\tau L = -\lambda^*$$

where λ^* is the dual optimal solution. This shows how much the objective function value c changes when parameter τ changes.

C EXPERIMENTS

C.1 Datasets

HELOC. The FICO HELOC [19] dataset contains anonymized information about a home equity line of credit applications made by homeowners in the US, with a binary response indicating whether or not the applicant has ever been more than 90 days delinquent for a payment. It can be used to train a machine learning model to predict whether the homeowner qualifies for a line of credit or not. The dataset consists of 10459 rows and 40 features, which we have normalized to be between zero and one.

German Credit. The German Credit dataset [26] comprises 1000 entries, each representing an individual who has taken a credit from a bank. These entries are characterized by 20 categorical features, which are used to classify each person as a good or bad credit risk. To prepare the dataset, we one-hot encoded the data and normalized it such that all features fall between zero and one.

Adult Income. The Adult Income [5] dataset comprises entries for 48842 individuals with a collection of 14 features for each of them. The target is a binary variable that indicates whether the individual has an income exceeding $50,000 or not. All the features are normalized to lie between zero and one.

ACKNOWLEDGMENTS

This work was supported in part by Northrop Grumman Seed Grant and NSF CAREER Award 2340006.

REFERENCES

[1] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).

[2] J. S. Baras and M. R. James . 1997. Robust and Risk-sensitive Output Feedback Control for Finite State Machines and Hidden Markov Models. *Journal of Mathematical Systems, Estimation, and Control* 7, 3 (1997), 371–374.

[3] J. S. Baras and N. S. Patel. 1998. Robust Control of Set-valued Discrete-time Dynamical Systems. *IEEE Trans. Automat. Control* 43, 1 (1998), 61–75.

[4] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.

[5] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[6] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. 850–863.

[7] Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. 2021. Consistent counterfactuals for deep models. *arXiv preprint arXiv:2110.03109* (2021).

[8] Miranda Bogen. 2019. All the ways hiring algorithms can introduce bias. *Harvard Business Review* 6 (2019), 2019.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[10] Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. 2025. Coverage-validity-aware algorithmic recourse. *Operations Research* (2025).

[11] Enrique Castillo, Roberto Mínguez, and Carmen Castillo. 2008. Sensitivity analysis in optimization and reliability problems. *Reliability Engineering & System Safety* 93, 12 (2008), 1788–1800.

[12] Jiahao Chen. 2018. Fair lending needs explainable models for responsible recommendation. *arXiv preprint arXiv:1809.04684* (2018).

[13] Eoin Delaney, Derek Greene, and Mark T Keane. 2021. Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions. *arXiv preprint arXiv:2107.09734* (2021).

[14] Marguerite J Dennis. 2018. Artificial intelligence and recruitment, admission, progression, and retention. *Enrollment Management Report* 22, 9 (2018), 1–3.

[15] Amit Dhurandhar, Swagatam Haldar, Dennis Wei, and Karthikeyan Natesan Ramamurthy. 2024. Trust Regions for Explanations via Black-Box Probabilistic Certification. *arXiv preprint arXiv:2402.11168* (2024).

[16] Pasan Dissanayake and Sanghamitra Dutta. 2024. Model reconstruction using counterfactual explanations: A perspective from polytope theory. In *Neural Information Processing Systems (NeurIPS)*.

[17] Ricardo Dominguez-Olmedo, Amir H Karimi, and Bernhard Schölkopf. 2022. On the Adversarial Robustness of Causal Algorithmic Recourse. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 5324–5342. https://proceedings.mlr.press/v162/dominguez-olmedo22a.html

[18] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. 2022. Robust counterfactual explanations for tree-based ensembles. In *International Conference on Machine Learning*. PMLR, 5742–5756.

[19] FICO. 2018. FICO XML Challenge. https://community.fico.com/s/explainable-machine-learning-challenge.

[20] Hans Föllmer and Alexander Schied. 2002. Convex Measures of Risk and Trading Constraints. *Finance and stochastics* 6, 4 (2002), 429–447.

[21] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. 2023. Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees. In *International Conference on Machine Learning (ICML)*.

[22] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. 2024. Robust Algorithmic Recourse Under Model Multiplicity With Probabilistic Guarantees. *IEEE Journal on Selected Areas in Information Theory* (2024).

[23] Leif Hancox-Li. 2020. Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 640–647.

[24] Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence* 12, 10 (1990), 993–1001.

[25] Karen Hao and Jonathan Stray. 2019. Can you make AI fairer than a judge? Play our courtroom algorithm game. *MIT Technology Review* (2019).

[26] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[27] Ronald A Howard and James E Matheson. 1972. Risk-sensitive Markov decision processes. *Management science* 18, 7 (1972), 356–369.

[28] Hsiang Hsu and Flavio Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 28988–29000.

[29] David Jacobson. 1973. Optimal Stochastic Linear Systems With Exponential Performance Criteria and Their Relation to Deterministic Differential Games. *IEEE Trans. Automat. Control* 18, 2 (1973), 124–131.

[30] Matthew R James and JS Baras. 1996. Partially Observed Differential Games, Infinite-Dimensional Hamilton–Jacobi–Isaacs Equations, and Nonlinear H_∞ Control. *SIAM Journal on Control and Optimization* 34, 4 (1996), 1342–1364.

[31] M. R. James and J. S. Baras. 1995. Robust H_∞ output feedback control for nonlinear systems. *IEEE Trans. Automat. Control* 40, 6 (1995), 1007–1017.

[32] Matthew R James, John S Baras, and Robert J Elliott. 1994. Risk-sensitive Control and Dynamic Games for Partially Observed Discrete-time Nonlinear Systems. *IEEE transactions on automatic control* 39, 4 (1994), 780–792.

[33] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. In *Proceedings of the 15th Asian Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 222)*, Berrin Yanıkoğlu and Wray Buntine (Eds.). PMLR, 582–597. https://proceedings.mlr.press/v222/jiang24a.html

[34] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2022. Formalising the Robustness of Counterfactual Explanations for Neural Networks. *arXiv preprint arXiv:2208.14878* (2022).

[35] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Recourse under model multiplicity via argumentative ensembling. In *International Conference on Autonomous Agents and Multiagent Systems*. 954–963.

[36] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Robust counterfactual explanations in machine learning: A survey. *arXiv preprint arXiv:2402.01928* (2024).

[37] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *CoRR* abs/2010.04050 (2020). arXiv:2010.04050 https://arxiv.org/abs/2010.04050

[38] Keita Kinjo. 2025. Robust Counterfactual Explanations under Model Multiplicity Using Multi-Objective Optimization. *arXiv preprint arXiv:2501.05795* (2025).

[39] Dan Kondratyuk, Mingxing Tan, Matthew Brown, and Boqing Gong. 2020. When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570* (2020).

[40] Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems* 7 (1994).

[41] PR Kumar and JH Van Schuppen. 1981. On The Optimal Control of Stochastic Systems With an Exponential-of-integral Performance Index. *Journal of mathematical analysis and applications* 80, 2 (1981), 312–332.

[42] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2019. Issues with post-hoc counterfactual explanations: a discussion. *arXiv preprint arXiv:1906.04774* (2019).

[43] Francesco Leofante, Elena Botoeva, and Vineet Rajani. 2023. Counterfactual explanations and model multiplicity: a relational verification view. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, Vol. 19. 763–768.

[44] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 2022. Global Counterfactual Explanations: Investigations, Implementations and Improvements. *arXiv preprint arXiv:2204.06917* (2022).

[45] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[46] Donato Maragno, Jannis Kurtz, Tabea E Röber, Rob Goedhart, Ş Ilker Birbil, and Dick den Hertog. 2023. Finding regions of counterfactual explanations via robust optimization. *arXiv preprint arXiv:2301.11113* (2023).

[47] Anna P Meyer, Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. 2024. Verified Training for Counterfactual Explanation Robustness under Data Shift. *arXiv preprint arXiv:2403.03773* (2024).

[48] Ibomoiye Domor Mienye and Yanxia Sun. 2022. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* 10 (2022), 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

[49] Kaisa Miettinen. 1999. *Nonlinear multiobjective optimization*. Vol. 12. Springer Science & Business Media.

[50] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. 2021. A Survey on the Robustness of Feature Importance and Counterfactual Explanations. *arXiv e-prints* arXiv:2111.00358 (2021).

[51] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. 2021. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358* (2021).

[52] Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

[53] Tuan-Duy H Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. 2022. Robust bayesian recourse. In *Uncertainty in Artificial Intelligence*. PMLR, 1498–1508.

[54] Erfaun Noorani and John S Baras. 2021. Risk-sensitive reinforce: A monte carlo policy gradient algorithm for exponential performance criteria. In *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 1522–1527.

[55] Erfaun Noorani and John S Baras. 2021. Risk-sensitive reinforcement learning and robust learning for control. In *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2976–2981.
[56] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 809–818.
[57] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. 2022. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768* (2022).
[58] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. 2020. Can I Still Trust You?: Understanding the Impact of Distribution Shifts on Algorithmic Recourses. *arXiv preprint arXiv:2012.11788* (2020).
[59] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2019. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857* (2019).
[60] Jason Speyer, John Deyst, and D Jacobson. 1974. Optimization of Stochastic Linear Systems With Additive Measurement and Process Noise Using Exponential Performance Criteria. *IEEE Trans. Automat. Control* 19, 4 (1974), 358–366.
[61] Ignacy Stępka, Mateusz Lango, and Jerzy Stefanowski. 2024. Counterfactual Explanations with Probabilistic Guarantees on their Robustness to Model Change. *arXiv preprint arXiv:2408.04842* (2024).
[62] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. 2021. Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems* 34 (2021).
[63] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
[64] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10306–10314.
[65] Xuezhong Zhang, Libin Dai, Qingming Peng, Ruizhi Tang, and Xinwei Li. 2022. A Survey of Counterfactual Explanations: Definition, Evaluation, Algorithms, and Applications. In *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*. Springer, 905–912.