
COUNTERFACTUAL TRAINING: TEACHING MODELS PLAUSIBLE AND ACTIONABLE EXPLANATIONS

A PREPRINT

Patrick Altmeyer 

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

p.altmeyer@tudelft.nl

Aleksander Buszydlik

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Arie van Deursen

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Cynthia C. S. Liem

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

March 9, 2025

ABSTRACT

We propose a novel training regime called Counterfactual Training that leverages counterfactual explanations to increase the explanatory capacity of models. Counterfactual explanations have emerged as a popular post-hoc explanation method for opaque machine learning models: they inform how factual inputs would need to change in order for a model to produce some desired output. To be useful in real-word decision-making systems, counterfactuals should be plausible with respect to the underlying data and actionable with respect to stakeholder requirements. Much existing research has therefore focused on developing post-hoc methods to generate counterfactuals that meet these desiderata. In this work, we instead hold models directly accountable for this desired end goal: Counterfactual Training employs counterfactuals ad-hoc during the training phase to minimize the divergence between learned representations and plausible, actionable explanations. We demonstrate empirically and theoretically that our proposed method facilitates training models that deliver inherently desirable explanations while maintaining high predictive performance.

Keywords Counterfactual Training • Counterfactual Explanations • Algorithmic Recourse • Explainable AI • Representation Learning

1 Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by advances in **representation learning**: instead of relying on features and rules that are carefully hand-crafted by humans, modern machine learning (ML)

18 models are tasked with learning the representations directly from data, guided by narrow objectives such as predictive
 19 accuracy (I. Goodfellow, Bengio, and Courville 2016). Modern advances in computing have made it possible to
 20 provide such models with ever growing degrees of freedom to achieve that task, which has often led them to outperform
 21 traditionally more parsimonious models. Unfortunately, in doing so, the models learn increasingly complex and highly
 22 sensitive representations that we can no longer easily interpret.

23 The trend towards complexity for the sake of performance has come under serious scrutiny in recent years. At the
 24 very cusp of the deep learning revolution, Szegedy et al. (2013) showed that artificial neural networks (ANN) are
 25 sensitive to adversarial examples: counterfactuals of model inputs that yield vastly different model predictions despite
 26 being “imperceptible” in that they are semantically indifferent from their factual counterparts. Although some partially
 27 effective mitigation strategies have been proposed, for example **adversarial training** (I. J. Goodfellow, Shlens, and
 28 Szegedy 2014), truly robust deep learning (DL) remains unattainable even for models that are considered shallow by
 29 today’s standards (Kolter 2023).

30 Part of the problem is that the high degrees of freedom provide room for many solutions that are locally optimal with
 31 respect to narrow objectives (Wilson 2020).¹ Indeed, recent work on the so called “lottery tickets” suggests that modern
 32 neural networks can be pruned by up to 90% while preserving their predictive performance (Frankle and Carbin 2019)
 33 and generalizability (Morcos et al. 2019). Similarly, Zhang et al. (2021) showed that state-of-the-art neural networks
 34 are so expressive that they can fit randomly labeled data. Thus, looking at the predictive performance, the solutions may
 35 seem to provide compelling explanations for the data, when in fact they are based on purely associative, semantically
 36 meaningless patterns. This poses two related challenges. Firstly, there is no dependable way to verify if such complex
 37 representations correspond to meaningful and plausible explanations. Secondly, even if we could resolve the first
 38 challenge, it remains undecided how to ensure that models can *only* learn valuable explanations.

39 The first challenge has attracted an abundance of research on **explainable AI** (XAI) which aims to develop tools to
 40 derive explanations from complex model representations. This can mitigate a scenario in which we deploy opaque
 41 models and blindly rely on their predictions. On countless occasions, this scenario has occurred in practice and caused
 42 real harm to people who were affected adversely and often unfairly by automated decision-making (ADM) systems
 43 involving opaque models (O’Neil 2016; McGregor 2021). Effective XAI tools can aid us in monitoring models and
 44 providing recourse to individuals to turn adverse outcomes (e.g., “loan application rejected”) into positive ones (e.g.,
 45 “application accepted”). Wachter, Mittelstadt, and Russell (2017) propose **counterfactual explanations** (CE) as an
 46 effective approach to achieve this goal: CEs explain how factual inputs need to change in order for some fitted model
 47 to produce some desired output, typically involving minimal perturbations.

48 To our surprise, the second challenge has not yet attracted any major consolidated research effort. Specifically, there
 49 has been no concerted effort towards improving models’ explanatory capacity, which we will henceforth
 50 simply call “explainability”, defined as the degree to which learned representations correspond to explanations that are
 51 interpretable and deemed **plausible** by humans (see Definition 3.1). Instead, the choice has typically been to improve
 52 the ability of XAI tools to identify the subset explanations that are both plausible and valid for any given model,
 53 independent of whether the learned representations are also compatible with implausible explanations (Altmeyer et
 54 al. 2024). Fortunately, recent findings indicate that explainability can arise as byproduct of regularization techniques
 55 aimed at other objectives such as robustness, generalization, and generative capacity (Schut et al. 2021; Augustin,
 56 Meinke, and Hein 2020; Altmeyer et al. 2024).

57 Building on these findings, we introduce **counterfactual training**: a novel training regime explicitly geared towards
 58 aligning model representations with plausible explanations. Our contributions are as follows:

- 59 • We discuss existing related work on improving models and consolidate it through the lens of counterfactual
 explanations (Section 2).
- 60 • We present our proposed methodological framework that leverages faithful counterfactual explanations during
 the training phase of models to achieve the explainability objective (Section 3).
- 61 • Through extensive experiments we demonstrate the counterfactual training improve model explainability
 while maintaining high predictive performance. We run ablation studies and grid searches to understand
 how the underlying model components and hyperparameters affect outcomes. (Section 4).

66 Despite some limitations of our approach discussed in Section 5, we conclude in Section 6 that counterfactual training
 67 provides a practical framework for researchers and practitioners interested in making opaque models more trustworthy.
 68 We also believe that this work serves as an opportunity for XAI researchers to re-evaluate the trend of improving XAI
 69 tools without improving the underlying models.

¹We follow the standard ML convention, where “degrees of freedom” refer to the number of parameters estimated from data.

70 2 Related Literature

71 To the best of our knowledge, our proposed framework of counterfactual training represents the first attempt to use
 72 counterfactual explanations during training to improve model explainability. In high-level terms, we define model
 73 explainability as the extent to which valid explanations derived for an opaque model are also deemed plausible with
 74 respect to the underlying data and stakeholder requirements. To make this more concrete, we follow Augustin, Meinke,
 75 and Hein (2020) in tying the concept of explainability to the quality of counterfactual explanations that we can generate
 76 for a given model. The authors show that counterfactual explanations—understood here as minimal input perturbations
 77 that yield some desired model prediction—are generally more meaningful if the underlying model is more robust to
 78 adversarial examples. We can make intuitive sense of this finding when looking at adversarial training (AT) through
 79 the lens of representation learning with high degrees of freedom: by inducing models to “unlearn” representations that
 80 are susceptible to worst-case counterfactuals (i.e., adversarial examples), AT effectively removes some implausible
 81 explanations from the solution space.

82 2.1 Adversarial Examples are Counterfactual Explanations

83 This interpretation of the link between explainability through counterfactuals on one side, and robustness to adversarial
 84 examples on the other, is backed by empirical evidence. Sauer and Geiger (2021) demonstrate that using counterfactual
 85 images during classifier training improves model robustness. Similarly, Abbasnejad et al. (2020) argue that counter-
 86 factuals represent potentially useful training data in machine learning, especially in supervised settings where inputs
 87 may be reasonably mapped to multiple outputs. They, too, demonstrate the augmenting the training data of image
 88 classifiers can improve generalization. Teney, Abbasnejad, and Hengel (2020) propose an approach using counterfac-
 89 tuals in training that does not rely on data augmentation: they argue that counterfactual pairs typically already exist
 90 in training datasets. Specifically, their approach relies on identifying similar input samples with different annotations
 91 and ensuring that the gradient of the classifier aligns with the vector between such pairs of counterfactual inputs using
 92 the cosine distance as the loss function.

93 In the natural language processing (NLP) domain, counterfactuals have similarly been used to improve models through
 94 data augmentation: Wu et al. (2021), propose *Polyjuice*, a general-purpose counterfactual generator for language mod-
 95 els. They demonstrate empirically that augmenting training data through *Polyjuice* counterfactuals improves robust-
 96 ness in a number of NLP tasks. Balashankar et al. (2023) also use *Polyjuice* to augment NLP datasets through diverse
 97 counterfactuals and show that classifier robustness improves up to 20%. Finally, Luu and Inoue (2023) introduce
 98 Counterfactual Adversarial Training (CAT), which also aims at improving generalization and robustness of language
 99 models. Specifically, they propose to proceed as follows: firstly, they identify training samples that are subject to
 100 high predictive uncertainty; secondly, they generate counterfactual explanations for those samples; and, finally, they
 101 fine-tune the given language model on the augmented dataset that includes the generated counterfactuals.

102 There have also been several attempts at formalizing the relationship between counterfactual explanations and adver-
 103 sarial examples (AE). Pointing to clear similarities in how CE and AE are generated, Freiesleben (2022) makes the
 104 case for jointly studying the opaqueness and robustness problem in representation learning. Formally, AE can be seen
 105 as the subset of CE for which misclassification is achieved (Freiesleben 2022). Similarly, Pawelczyk et al. (2022)
 106 show that CE and AE are equivalent under certain conditions and derive theoretical upper bounds on the distances
 107 between them.

108 Two recent works are closely related to ours in that they use counterfactuals during training with the explicit goal
 109 of affecting certain properties of post-hoc counterfactual explanations. Firstly, Ross, Lakkaraju, and Bastani (2024)
 110 propose a way to train models that are guaranteed to provide recourse for individuals to move from an adverse outcome
 111 to some positive target class with high probability. Their approach builds on adversarial training, where in this context
 112 susceptibility to targeted adversarial examples for the positive class is explicitly induced. The proposed method allows
 113 for imposing a set of actionability constraints ex-ante: for example, users can specify that certain features (e.g., *age*,
 114 *gender*, ...) are immutable. Secondly, Guo, Nguyen, and Yadav (2023) are the first to propose an end-to-end training
 115 pipeline that includes counterfactual explanations as part of the training procedure. In particular, they propose a
 116 specific network architecture that includes a predictor and CE generator network, where the parameters of the CE
 117 generator network are learnable. Counterfactuals are generated during each training iteration and fed back to the
 118 predictor network. In contrast to Guo, Nguyen, and Yadav (2023), we impose no restrictions on the neural network
 119 architecture at all.

120 2.2 Beyond Robustness

121 Improving the adversarial robustness of models is not the only path towards aligning representations with plausible
 122 explanations. In a work closely related to this one, Altmeyer et al. (2024) show that explainability can be improved
 123 through model averaging and refined model objectives. The authors propose a way to generate counterfactuals that
 124 are maximally **faithful** to the model in that they are consistent with what the model has learned about the underlying

125 data. Formally, they rely on tools from energy-based modelling to minimize the divergence between the distribution
 126 of counterfactuals and the conditional posterior over inputs learned by the model. Their proposed counterfactual
 127 explainer, *ECCo*, yields plausible explanations if and only if the underlying model has learned representations that
 128 align with them. They find that both deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and joint energy-
 129 based models (JEMs) (Grathwohl et al. 2020) tend to do well in this regard.

130 Once again it helps to look at these findings through the lens of representation learning with high degrees of freedom.
 131 Deep ensembles are approximate Bayesian model averages, which are most called for when models are underspecified
 132 by the available data (Wilson 2020). Averaging across solutions mitigates the aforementioned risk of relying on a
 133 single locally optimal representations that corresponds to semantically meaningless explanations for the data. Previous
 134 work by Schut et al. (2021) similarly found that generating plausible (“interpretable”) counterfactual explanations is
 135 almost trivial for deep ensembles that have also undergone adversarial training. The case for JEMs is even clearer:
 136 they involve a hybrid objective that induces both high predictive performance and generative capacity (Grathwohl et al.
 137 2020). This is closely related to the idea of aligning models with plausible explanations and has inspired our proposed
 138 counterfactual training objective, as we explain in Section 3.

139 3 Counterfactual Training

140 Counterfactual training combines ideas from adversarial training, energy-based modelling and counterfactuals expla-
 141 nations with the explicit objective of aligning representations with plausible explanations that comply with user re-
 142 quirements. In the context of CEs, plausibility has broadly been defined as the degree to which counterfactuals comply
 143 with the underlying data generating process (Poyiadzi et al. 2020; Guidotti 2022; Altmeyer et al. 2024). Plausibility
 144 is a necessary but insufficient condition for using CEs to provide algorithmic recourse (AR) to individuals affected
 145 by opaque models in practice. This is because for recourse recommendations to be **actionable**, they need to not only
 146 result in plausible counterfactuals but also be attainable. A plausible CE for a rejected 20-year-old loan applicant, for
 147 example, might reveal that their application would have been accepted, if only they were 20 years older. Ignoring
 148 all other features, this would comply with the definition of plausibility if 40-year-old individuals were in fact more
 149 credit-worthy on average than young adults. But of course this CE does not qualify for providing actionable recourse
 150 to the applicant since *age* is not a (directly) mutable feature. For our intents and purposes, counterfactual training aims
 151 to improve model explainability by aligning models with counterfactuals that meet both desiderata, plausibility and
 152 actionability. Formally, we define explainability as follows:

153 **Definition 3.1** (Model Explainability). Let $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denote a supervised classification model that maps from the
 154 D -dimensional input space \mathcal{X} to representations $\phi(\mathbf{x}; \theta)$ and finally to the K -dimensional output space \mathcal{Y} . Assume
 155 that for any given input-output pair $\{\mathbf{x}, \mathbf{y}\}_i$ there exists a counterfactual $\mathbf{x}' = \mathbf{x} + \Delta : M_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = M_\theta(\mathbf{x})$
 156 where $\arg \max_y \mathbf{y}^+ = y^+$ and y^+ denotes the index of the target class.

157 We say that M_θ is **explainable** to the extent that faithfully generated counterfactuals are plausible (i.e. consistent with
 158 the data) and actionable. Formally, we define these properties as follows:

- 159 1. (Plausibility) $\int^A p(\mathbf{x}' | \mathbf{y}^+) d\mathbf{x} \rightarrow 1$ where A is some small region around \mathbf{x}' .
- 160 2. (Actionability) Permutations Δ are subject to some actionability constraints.

161 We consider counterfactuals as faithful to the extent that they are consistent with what the model has learned about the
 162 input data. Let $p_\theta(\mathbf{x} | \mathbf{y}^+)$ denote the conditional posterior over inputs, then formally:

- 163 3. (Faithfulness) $\int^A p_\theta(\mathbf{x}' | \mathbf{y}^+) d\mathbf{x} \rightarrow 1$ where A is defined as above.

164 The definitions of faithfulness and plausibility in Definition 3.1 are the same as in Altmeyer et al. (2024), with adapted
 165 notation. Actionability constraints in Definition 3.1 vary and depend on the context in which M_θ is deployed. In this
 166 work, we focus on domain and mutability constraints for individual features x_d for $d = 1, \dots, D$. We limit ourselves
 167 to classification tasks for reasons discussed in Section 5.

168 3.1 Our Proposed Objective

169 Let \mathbf{x}'_t for $t = 0, \dots, T$ denote a counterfactual explanation generated through gradient descent over T iterations
 170 as initially proposed by Wachter, Mittelstadt, and Russell (2017). For our purposes, we let T vary and consider the
 171 counterfactual search as converged as soon as the predicted probability for the target class has reached a pre-determined
 172 threshold, $\tau : \mathcal{S}(M_\theta(\mathbf{x}'))[y^+] \geq \tau$, where \mathcal{S} is the softmax function.²

²For detailed background information on gradient-based counterfactual search and convergence see ?@sec-app-ce.

173 To train models with high explainability as defined in Definition 3.1, we propose to leverage counterfactuals in the
 174 following objective:

$$\min_{\theta} \text{yloss}(\mathbf{M}_{\theta}(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}} \text{div}(\mathbf{x}, \mathbf{x}'_T, y; \theta) + \lambda_{\text{adv}} \text{advloss}(\mathbf{M}_{\theta}(\mathbf{x}'_{t \leq T}), \mathbf{y}) \quad (1)$$

175 where $\text{yloss}(\cdot)$ is any conventional classification loss that induces discriminative performance (e.g., cross-entropy).
 176 The two additional components in Equation 1 are explained in more detail below. For now, they can be sufficiently
 177 described as inducing explainability directly and indirectly by penalizing: (1) the contrastive divergence, $\text{div}(\cdot)$, be-
 178 tween counterfactuals \mathbf{x}'_T and observed samples x and, (2) the adversarial loss, $\text{advloss}(\cdot)$, with respect to nascent
 179 counterfactuals $\mathbf{x}'_{t \leq T}$. The tradeoff between the different components can be governed by adjusting the strengths of
 180 the penalties λ_{div} and λ_{adv} .

181 3.1.1 Directly Inducing Explainability through Contrastive Divergence

182 Grathwohl et al. (2020) observe that any classifier can be re-interpreted as a joint energy-based model (JEM)
 183 that learns to discriminate output classes conditional on inputs and generate inputs. They show that JEMs can be
 184 trained to perform well at both tasks by directly maximizing the joint log-likelihood factorized as $\log p_{\theta}(\mathbf{x}, \mathbf{y}) =$
 185 $\log p_{\theta}(\mathbf{y}|\mathbf{x}) + \log p_{\theta}(\mathbf{x})$. The first factor can be optimized using conventional cross-entropy as in Equation 1. To
 186 optimize $\log p_{\theta}(\mathbf{x})$ Grathwohl et al. (2020) minimize the contrastive divergence between samples drawn from $p_{\theta}(\mathbf{x})$
 187 and training observations, i.e., samples from $p(\mathbf{x})$.

188 A key empirical finding in Altmeyer et al. (2024) was that JEMs tend to do well with respect to the plausibility objec-
 189 tive in Definition 3.1. If we consider samples drawn from $p_{\theta}(\mathbf{x})$ as counterfactuals, this is an expected finding, because
 190 the JEM objective effectively minimizes the divergence between the conditional posterior and $p(\mathbf{x}|y^+)$. To generate
 191 samples, Grathwohl et al. (2020) rely on Stochastic Gradient Langevin Dynamics (SGLD) using an uninformative
 192 prior for initialization. This is where we depart from their methodology: instead of generating samples through SGLD,
 193 we propose using counterfactual explainers to generate counterfactuals for observed training samples. Specifically, we
 194 have

$$\text{div}(\mathbf{x}, \mathbf{x}'_T, y; \theta) = \mathcal{E}_{\theta}(\mathbf{x}, y) - \mathcal{E}_{\theta}(\mathbf{x}'_T, y) \quad (2)$$

195 where $\mathcal{E}_{\theta}(\cdot)$ denotes the energy function. In particular, we set $\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{y}) = -\mathbf{M}_{\theta}(\mathbf{x})[y^+]$ where y^+ denotes the index of
 196 the target class. We generate samples \mathbf{x}'_T by first randomly sampling the target class $y^+ \sim p(y)$ and then generating
 197 a counterfactual explanation for that target over T iterations using a gradient-based counterfactual generator. This is
 198 similar to how conditional sampling is used to draw from $p_{\theta}(\mathbf{x})$ in Grathwohl et al. (2020).

199 Intuitively, the gradient of Equation 2 decreases the energy of observed training samples (positive samples) while at
 200 same time increasing the energy of counterfactuals (negative samples) (Du and Mordatch 2020). As the generated
 201 counterfactuals get more plausible (Definition 3.1) over the course of training, these two opposing effects gradually
 202 balance each out (Lippe 2024).

203 The departure from SGLD allows us to tap into the vast repertoire of explainers that have been proposed in the literature
 204 to meet different desiderata. Typically, these methods facilitate the imposition of domain and mutability constraints,
 205 for example. In principle, any existing approach for generating counterfactual explanations is viable, so long as it does
 206 not violate the faithfulness condition. Like JEMs (Murphy 2022), counterfactual training can be considered as a form
 207 of contrastive representation learning.

208 3.1.2 Indirectly Inducing Explainability through Adversarial Robustness

209 Based on our analysis in Section 2, counterfactuals \mathbf{x}' can be repurposed as additional training samples (Luu and Inoue
 210 2023; Balashankar et al. 2023) or adversarial examples (Freiesleben 2022; Pawelczyk et al. 2022). This leaves some
 211 flexibility with respect to the exact choice for $\text{advloss}(\cdot)$ in Equation 1. An intuitive functional form to use, though
 212 likely not the only reasonable choice, is inspired by adversarial training:

$$\begin{aligned} \text{advloss}(\mathbf{M}_{\theta}(\mathbf{x}'_{t \leq T}), \mathbf{y}; \varepsilon) &= \text{yloss}(\mathbf{M}_{\theta}(\mathbf{x}'_{t_\varepsilon}), \mathbf{y}) \\ t_\varepsilon &= \max_t \{t : \|\Delta_t\|_\infty < \varepsilon\} \end{aligned} \quad (3)$$

213 Under this choice, we consider nascent counterfactuals $\mathbf{x}'_{t \leq T}$ as adversarial examples as long as the magnitude of the
 214 perturbation to any individual feature is at most ε . This is closely aligned with Szegedy et al. (2013), who define an
 215 adversarial attack as an “imperceptible non-random perturbation”. Thus, we choose to work with a different distinction
 216 between CE and AE than Freiesleben (2022), who considers misclassification as the key distinguishing feature of AE.

217 One of the key observations in this work is that we can leverage counterfactual explanations during training and get
 218 adversarial examples, essentially for free.

219 3.2 Encoding Actionability Constraints

220 Many existing counterfactual explainers support domain and mutability constraints out-of-the-box. In fact, both types
 221 of constraints can be implemented for any counterfactual explainer that relies on gradient descent in the feature space
 222 for optimization (Altmeyer, van Deursen, and Liem 2023). In this context, domain constraints can be imposed by
 223 simply projecting counterfactuals back to the specified domain, if the previous gradient step resulted in updated feature
 224 values that were out-of-domain. Mutability constraints can similarly be enforced by setting partial derivatives to zero
 225 to ensure that features are only mutated in the allowed direction, if at all.

226 Since actionability constraints are binding at test time, we should also impose them when generating \mathbf{x}' during each
 227 training iteration to align model representations with user requirements. Through their effect on \mathbf{x}' , both types of
 228 constraints influence model outcomes through Equation 2. Here it is crucial that we avoid penalizing implausibility
 229 that arises due to mutability constraints. For any mutability-constrained feature d this can be achieved by enforcing
 230 $\mathbf{x}[d] - \mathbf{x}'[d] := 0$ whenever perturbing $\mathbf{x}'[d]$ in the direction of $\mathbf{x}[d]$ would violate mutability constraints. Specifically,
 231 we set $\mathbf{x}[d] := \mathbf{x}'[d]$ if:

- 232 1. Feature d is strictly immutable in practice.
- 233 2. We have $\mathbf{x}[d] > \mathbf{x}'[d]$ but feature d can only be decreased in practice.
- 234 3. We have $\mathbf{x}[d] < \mathbf{x}'[d]$ but feature d can only be increased in practice.

235 From a Bayesian perspective, setting $\mathbf{x}[d] := \mathbf{x}'[d]$ can be understood as assuming a point mass prior for $p(\mathbf{x})$ with
 236 respect to feature d . Intuitively, we think of this simply in terms ignoring implausibility costs with respect to immutable
 237 features, which effectively forces the model to instead seek plausibility with respect to the remaining features. This
 238 in turn results in lower overall sensitivity to immutable features, which we demonstrate empirically for different
 239 classifiers in Section 4. Under certain conditions, this results holds theoretically.³

240 **Proposition 3.1** (Protecting Immutable Features). *Let $f_\theta(\mathbf{x}) = \mathcal{S}(\mathbf{M}_\theta(\mathbf{x})) = \mathcal{S}(\Theta\mathbf{x})$ denote a linear classifier with
 241 softmax activation \mathcal{S} (i.e., multinomial logistic regression) where $y \in \{1, \dots, K\} = \mathcal{K}$ and $\mathbf{x} \in \mathbb{R}^D$. If we assume
 242 multivariate Gaussian class densities with common diagonal covariance matrix $\Sigma_k = \Sigma$ for all $k \in \mathcal{K}$, then protecting
 243 an immutable feature from the contrastive divergence penalty (Equation 2) will result in lower classifier sensitivity to
 244 that feature relative to the remaining features, provided that at least one of those is mutable and discriminative.*

245 It is worth highlighting that Proposition 3.1 assumes independence of features. This raises a valid concern about the
 246 effect of protecting immutable features in the presence of proxy features that remain unprotected. We discuss this
 247 limitation in Section 5.

248 3.3 Illustration

249 To better convey the intuition underlying our proposed method, we illustrate different model outcomes in Example 3.1.

250 **Example 3.1** (Prediction of Consumer Credit Default). Suppose we are interested in predicting the likelihood that
 251 loan applicants default on their credit. We have access to historical data on previous loan takers comprised of a binary
 252 outcome variable ($y \in \{1 = \text{default}, 2 = \text{no default}\}$) two input features: (1) the subjects' *age*, which we define as
 253 immutable, and (2) the subjects' existing level of *debt*, which we define as mutable.

254 We have simulated this scenario using synthetic data with independent features and Gaussian class-conditional densities
 255 in Figure 1. The four panels in Figure 1 show the outcomes for different training procedures using the same model
 256 architecture each time (a linear classifier). In each case, we show the linear decision boundary (green) and the training
 257 data colored according to their ground-truth label: orange points belong to the target class, $y^+ = 2$, blue points belong
 258 to the non-target class, $y^- = 1$. Stars indicate counterfactuals in the target class generated at test time using generic
 259 gradient descent until convergence.

260 In panel (a), we have trained our model conventionally, and we do not impose mutability constraints at test time. The
 261 generated counterfactuals are all valid, but not plausible: they are clearly distinguishable from the ground-truth data. In
 262 panel (b), we have trained our model with counterfactual training, once again not imposing mutability constraints at test
 263 time. We observe that the counterfactuals are clearly plausible, therefore meeting the first objective of Definition 3.1.

264 In panel (c), we have used conventional training again, this time imposing the mutability constraint on *age* at test time.
 265 Counterfactuals are valid but involve some substantial reductions in *debt* for some individuals (very young applicants).

³For the proof, see the supplementary appendix.

266 By comparison, counterfactual paths are shorter on average in panel (d), where we have used counterfactual training
 267 and protected immutable features as described in Section 3.2. In particular, we observe that due to the classifier's
 268 lower sensitivity to *age*, recourse recommendations with respect to *debt* are much more homogenous, in that they do
 269 not disproportionately punish younger individuals. The counterfactuals are also plausible with respect to the mutable
 270 feature. Thus, we consider the model in panel (d) as the most explainable according to Definition 3.1.

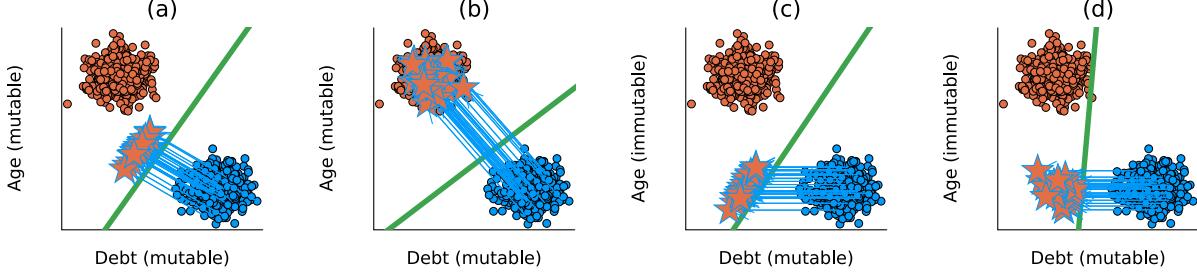


Figure 1: Visual illustration of how counterfactual training improves explainability. See Example 3.1 for details.

271 4 Experiments

272 In this section, we present experiments that we have conducted in order to answer the following research questions:

273 **Research Question 4.1** (Plausibility). *Does our proposed counterfactual training objective (Equation 1) induce mod-
 274 els to learn plausible explanations?*

275 **Research Question 4.2** (Actionability). *Does our proposed counterfactual training objective (Equation 1) yield more
 276 favorable algorithmic recourse outcomes in the presence of actionability constraints?*

277 Beyond this, we are also interested in understanding how robust our answers to RQ 4.1 and RQ 4.2 are:

278 **Research Question 4.3** (Hyperparameters). *What are the effects of different hyperparameter choices with respect to
 279 Equation 1?*

280 4.1 Experimental Setup

281 4.2 Experimental Results

282 5 Discussion

283 5.1 Approach/Future Directions

- 284 1. Limited to classification models.
- 285 2. Training instabilities.
- 286 3. Hyperparameter sensitivity -> can we do better than grid search? (Bayes opt, ...)

287 5.2 Limitations

- 288 3. Proxy attributes of immutable features.
- 289 4. Increased training time.
- 290 5. Fairness and caveats (aware it's not a classical approach in this context, but there is a clear link).

291 6 Conclusion

292 References

- 293 Abbasnejad, Ehsan, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. “Counterfactual
 294 Vision and Language Learning.” In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition
 295 (CVPR)*, 10041–51. <https://doi.org/10.1109/CVPR42600.2020.01006>.
- 296 Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia CS Liem. 2024. “Faithful Model Explanations
 297 Through Energy-Constrained Conformal Counterfactuals.” In *Proceedings of the AAAI Conference on Artificial
 298 Intelligence*, 38:10829–37. 10.
- 299 Altmeyer, Patrick, Arie van Deursen, and Cynthia C. S. Liem. 2023. “Explaining Black-Box Models Through Coun-
 300 terfactuals.” In *Proceedings of the JuliaCon Conferences*, 1:130. 1.

- 301 Augustin, Maximilian, Alexander Meinke, and Matthias Hein. 2020. “Adversarial Robustness on in-and Out-
 302 Distribution Improves Explainability.” In *European Conference on Computer Vision*, 228–45. Springer.
- 303 Balashankar, Ananth, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi, Jilin Chen, and Alex Beutel. 2023.
 304 “Improving Classifier Robustness Through Active Generative Counterfactual Data Augmentation.” In *Findings of
 305 the Association for Computational Linguistics: EMNLP 2023*, 127–39.
- 306 Du, Yilun, and Igor Mordatch. 2020. “Implicit Generation and Generalization in Energy-Based Models.” <https://arxiv.org/abs/1903.08689>.
- 308 Frankle, Jonathan, and Michael Carbin. 2019. “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Net-
 309 works.” In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rJl-b3RcF7>.
- 310 Freiesleben, Timo. 2022. “The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples.”
 311 *Minds and Machines* 32 (1): 77–109.
- 312 Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2014. “Explaining and Harnessing Adversarial Examples.”
 313 <https://arxiv.org/abs/1412.6572>.
- 314 Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- 315 Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swer-
 316 sky. 2020. “Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One.” In *International
 317 Conference on Learning Representations*.
- 318 Guidotti, Riccardo. 2022. “Counterfactual Explanations and How to Find Them: Literature Review and Benchmark-
 319 ing.” *Data Mining and Knowledge Discovery*, 1–55.
- 320 Guo, Hangzhi, Thanh H. Nguyen, and Amulya Yadav. 2023. “CounterNet: End-to-End Training of Prediction Aware
 321 Counterfactual Explanations.” In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery
 322 and Data Mining*, 577–89. KDD ’23. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3580305.3599290>.
- 324 Kolter, Zico. 2023. “Keynote Addresses: SaTML 2023 .” In *2023 IEEE Conference on Secure and Trustworthy
 325 Machine Learning (SaTML)*, xvi–. Los Alamitos, CA, USA: IEEE Computer Society. <https://doi.org/10.1109/SaTML54575.2023.00009>.
- 327 Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. “Simple and Scalable Predictive Uncer-
 328 tainty Estimation Using Deep Ensembles.” *Advances in Neural Information Processing Systems* 30.
- 329 Lippe, Phillip. 2024. “UvA Deep Learning Tutorials.” <https://uvadlc-notebooks.readthedocs.io/en/latest/>.
- 330 Luu, Hoai Linh, and Naoya Inoue. 2023. “Counterfactual Adversarial Training for Improving Robustness of Pre-
 331 Trained Language Models.” In *Proceedings of the 37th Pacific Asia Conference on Language, Information and
 332 Computation*, 881–88.
- 333 McGregor, Sean. 2021. “Preventing repeated real world AI failures by cataloging incidents: The AI incident database.”
 334 In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:15458–63. 17.
- 335 Morcos, Ari S., Haonan Yu, Michela Paganini, and Yuandong Tian. 2019. “One Ticket to Win Them All: Gener-
 336 alizing Lottery Ticket Initializations Across Datasets and Optimizers.” In *Proceedings of the 33rd International
 337 Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.
- 338 Murphy, Kevin P. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.
- 339 O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.
 340 Crown.
- 341 Pawelczyk, Martin, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. “Exploring
 342 Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis.”
 343 In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, edited by Gustau
 344 Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, 151:4574–94. Proceedings of Machine Learning Research.
 345 PMLR. <https://proceedings.mlr.press/v151/pawelczyk22a.html>.
- 346 Poyiadzi, Rafael, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. “FACE: Feasible and
 347 Actionable Counterfactual Explanations.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*,
 348 344–50.
- 349 Ross, Alexis, Himabindu Lakkaraju, and Osbert Bastani. 2024. “Learning Models for Actionable Recourse.” In
 350 *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS ’21. Red
 351 Hook, NY, USA: Curran Associates Inc.
- 352 Sauer, Axel, and Andreas Geiger. 2021. “Counterfactual Generative Networks.” <https://arxiv.org/abs/2101.06046>.
- 353 Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. “Generating
 354 Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties.” In
 355 *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.
- 356 Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.
 357 2013. “Intriguing Properties of Neural Networks.” <https://arxiv.org/abs/1312.6199>.

- 358 Tenev, Damien, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. “Learning What Makes a Difference from
 359 Counterfactual Examples and Gradient Supervision.” In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part x 16*, 580–99. Springer.
 360
- 361 Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations Without Opening the Black
 362 Box: Automated Decisions and the GDPR.” *Harv. JL & Tech.* 31: 841. <https://doi.org/10.2139/ssrn.3063289>.
 363
- 364 Wilson, Andrew Gordon. 2020. “The Case for Bayesian Deep Learning.” <https://arxiv.org/abs/2001.10995>.
 365
- 366 Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. “Polyjuice: Generating Counterfactuals
 367 for Explaining, Evaluating, and Improving Models.” In *Proceedings of the 59th Annual Meeting of the Association
 368 for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing
 (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 6707–23. Online:
 369 Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.523>.
 370
- 371 Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. “Understanding Deep
 372 Learning (Still) Requires Rethinking Generalization.” *Commun. ACM* 64 (3): 107–15. <https://doi.org/10.1145/3446776>.