
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

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

February 10, 2025

ABSTRACT

Counterfactual Explanations have emerged as a popular tool to explain predictions made by opaque machine learning models: they explain how factual inputs need to change in order for some fitted model to produce some desired output. Much existing research has focused on identifying explanations that are not only valid but also deemed plausible and desirable with respect to the underlying data and stakeholder requirements. Recent work has shown that under this premise, the task of learning plausible explanations is effectively reassigned from the model itself to the (post-hoc) counterfactual explainer. Building on that work, we propose a novel model objective that leverages counterfactuals during the training phase (ad-hoc) in order to minimize the divergence between learned representations and plausible explanations. Through extensive experiments, we demonstrate that our proposed methodology facilitates training models that inherently deliver plausible explanations while maintaining high predictive performance.

Keywords Counterfactual Explanations • 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) models are tasked with learning these representations from scratch, guided by narrow objectives such as predictive accuracy (I. Goodfellow, Bengio, and Courville 2016). Modern advances in computing have made it possible to provide such models with ever greater degrees of freedom to achieve that task, which has often led them to outperform traditionally more parsimonious models. Unfortunately, in doing so they also learn increasingly complex and highly sensitive representations that we can no longer easily interpret.

This trend towards complexity for the sake of performance has come under serious scrutiny in recent years. At the very cusp of the deep learning revolution, Szegedy et al. (2013) showed that artificial neural networks (ANN) are sensitive

23 to adversarial examples: counterfactuals of model inputs that yield vastly different model predictions despite being
 24 “imperceptible” in that they are semantically indifferent from their factual counterparts. Despite partially effective
 25 mitigation strategies such as **adversarial training** (I. J. Goodfellow, Shlens, and Szegedy 2014), truly robust deep
 26 learning (DL) remains unattainable even for models that are considered shallow by today’s standards (Kolter 2023).

27 Part of the problem is that high degrees of freedom provide room for many solutions that are locally optimal with
 28 respect to narrow objectives (Wilson 2020)¹. Based purely on predictive performance, these solutions may seem to
 29 provide compelling explanations for the data, when in fact they are based on purely associative, semantically mean-
 30 ingless patterns. This poses two related challenges: firstly, it makes these models inherently opaque, since humans
 31 cannot simply interpret what type of explanation the complex learned representations correspond to; secondly, even
 32 if we could resolve the first challenge, it is not obvious how to mitigate models from learning representations that
 33 correspond to meaningless and implausible explanations.

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

43 To our surprise, the second challenge has not yet attracted any consolidated research effort. Specifically, there has
 44 been no concerted effort towards improving model **explainability**, which we define here as the degree to which learned
 45 representations correspond to explanations that are interpretable and deemed **plausible** by humans (see Definition 3.1).
 46 Instead, the choice has typically been to improve the capacity of XAI tools to identify the subset explanations that are
 47 both plausible and valid for any given model, independent of whether the learned representations are also compatible
 48 with implausible explanations (Altmeyer et al. 2024). Fortunately, recent findings indicate that explainability can arise
 49 as byproduct of regularization techniques aimed at other objectives such as robustness, generalization and generative
 50 capacity Altmeyer et al. (2024).

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

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

60 Despite limitations of our approach discussed in Section 5, we conclude that counterfactual training provides a practi-
 61 cal framework for researchers and practitioners interested in making opaque models more trustworthy Section 6. We
 62 also believe that this work serves as an opportunity for XAI researchers to reevaluate the premise of improving XAI
 63 tools without improving models.

64 2 Related Literature

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

¹For clarity: we follow standard ML convention in using “degrees of freedom” to refer to the number of parameters estimated from data.

74 representations that are susceptible to worst-case counterfactuals (i.e. adversarial examples), AT effectively removes
 75 some implausible explanations from the solution space.

76 2.1 Adversarial Examples are Counterfactual Explanations

77 This interpretation of the link between explainability through counterfactuals on one side, and robustness to adversarial
 78 examples on the other, is backed by empirical evidence. Sauer and Geiger (2021) demonstrate that using counterfactual
 79 images during classifier training improves model robustness. Similarly, Abbasnejad et al. (2020) argue that counterfactuals
 80 represent potentially useful training data in machine learning, especially in supervised settings where inputs may
 81 be reasonably mapped to multiple outputs. They, too, demonstrate the augmenting the training data of image classifi-
 82 cers can improve generalization. Teney, Abbasnejad, and Hengel (2020) propose an approach using counterfactuals
 83 in training that does not rely on data augmentation: they argue that counterfactual pairs typically already exist in train-
 84 ing datasets. Specifically, their approach relies on, firstly, identifying similar input samples with different annotations
 85 and, secondly, ensuring that the gradient of the classifier aligns with the vector between pairs of counterfactual inputs
 86 using the cosine distance as a loss function. In the natural language processing (NLP) domain, counterfactuals have
 87 similarly been used to improve models through data augmentation: Wu et al. (2021), propose *POLYJUICE*, a general-
 88 purpose counterfactual generator for language models. They demonstrate empirically that augmenting training data
 89 through *POLYJUICE* counterfactuals improves robustness in a number of NLP tasks. Luu and Inoue (2023) introduce
 90 Counterfactual Adversarial Training (CAT), which also aims at improving generalization and robustness of language
 91 models. Specifically, they propose to proceed as follows: firstly, they identify training samples that are subject to
 92 high predictive uncertainty; secondly, they generate counterfactual explanations for those samples; and, finally, they
 93 fine-tune the given language model on the augmented dataset that includes the generated counterfactuals.

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

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

112 2.2 Beyond Robustness

113 Improving the adversarial robustness of models is not the only path towards aligning representations with plausible
 114 explanations. In a work closely related to this one, Altmeyer et al. (2024) show that explainability can be improved
 115 through model averaging and refined model objectives. The authors propose a way to generate counterfactuals that
 116 are maximally **faithful** to the model in that they are consistent with what the model has learned about the underlying
 117 data. Formally, they rely on tools from energy-based modelling to minimize the divergence between the distribution
 118 of counterfactuals and the conditional posterior over inputs learned by the model. Their proposed counterfactual
 119 explainer, *ECCo*, yields plausible explanations if and only if the underlying model has learned representations that
 120 align with them. They find that both deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and joint energy-
 121 based models (JEMs) (Grathwohl et al. 2020) tend to do well in this regard.

122 Once again it helps to look at these findings through the lens of representation learning with high degrees of freedom.
 123 Deep ensembles are approximate Bayesian model averages, which are most called for when models are underspecified
 124 by the available data (Wilson 2020). Averaging across solutions mitigates the aforementioned risk of relying on a
 125 single locally optimal representations that corresponds to semantically meaningless explanations for the data. Previous
 126 work by Schut et al. (2021) similarly found that generating plausible (“interpretable”) counterfactual explanations is
 127 almost trivial for deep ensembles that have also undergone adversarial training. The case for JEMs is even clearer:
 128 they involve a hybrid objective that induces both high predictive performance and generative capacity (Grathwohl et al.

129 This is closely related to the idea of aligning models with plausible explanations and has inspired our proposed
 130 counterfactual training objective, as we explain in Section 3.

131 3 Counterfactual Training

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

145 **Definition 3.1** (Model Explainability). Let $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denote a supervised classification model that maps from the
 146 D -dimensional input space \mathcal{X} to representations $\phi(\mathbf{x}; \theta)$ and finally to the K -dimensional output space \mathcal{Y} . Assume that
 147 for any given input-output pair $\{\mathbf{x}, \mathbf{y}\}_i$ there exists a counterfactual $\mathbf{x}' = \mathbf{x} + \Delta : \mathbf{M}_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = \mathbf{M}_\theta(\mathbf{x})$ where
 148 \mathbf{y}^+ denotes some target output. We say that \mathbf{M}_θ is **explainable** to the extent that faithfully generated counterfactuals
 149 are plausible (i.e. consistent with the data) and actionable. Formally, we define these properties as follows:

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

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

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

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

159 3.1 Our Proposed Objective

160 To train models with high explainability as defined in Definition 3.1, we propose the following objective,

$$\text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}} \text{div}(\mathbf{x}, \mathbf{x}', \mathbf{y}; \theta) + \lambda_{\text{adv}} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}) \quad (1)$$

161 where $\text{yloss}(\cdot)$ denotes any conventional classification loss function (e.g. cross-entropy) that induces discriminative
 162 (predictive) performance. The two additional components in Equation 1 are explained in more detail below. For now,
 163 they can be sufficiently described as inducing explainability directly and indirectly by penalizing: 1) the contrastive
 164 divergence, $\text{div}(\cdot)$, between counterfactuals x' and observed samples x and, 2) the adversarial loss, $\text{advloss}(\cdot)$, with
 165 respect to counterfactuals. The tradeoff between the different components can be governed by adjusting the strengths
 166 of the penalties λ_{div} and λ_{adv} .

167 3.1.1 Directly Inducing Explainability through Contrastive Divergence

168 Grathwohl et al. (2020) observe that any classifier can be re-interpreted as a joint energy-based model (JEM)
 169 that learns to discriminate output classes conditional on inputs and generate inputs. They show that JEMs can be
 170 trained to perform well at both tasks by directly maximizing the joint log-likelihood factorized as $\log p_\theta(\mathbf{x}, \mathbf{y}) =$
 171 $\log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$. The first factor can be optimized using conventional cross-entropy as in Equation 1. To
 172 optimize $\log p_\theta(\mathbf{x})$ Grathwohl et al. (2020) minimize the contrastive divergence between samples drawn from $p_\theta(\mathbf{x})$
 173 and training observations, i.e. samples from $p(\mathbf{x})$.

174 A key empirical finding in Altmeyer et al. (2024) was that JEMs tend to do well with respect to the plausibility objec-
 175 tive in Definition 3.1. If we consider samples drawn from $p_\theta(\mathbf{x})$ as counterfactuals, this is an expected finding, because

176 the JEM objective effectively minimizes the divergence between the conditional posterior and $p(\mathbf{x}|\mathbf{y}^+)$. To generate
 177 samples, Grathwohl et al. (2020) rely on Stochastic Gradient Langevin Dynamics (SGLD) using an uninformative
 178 prior for initialization. This is where we depart from their methodology: instead of generating samples through SGLD,
 179 we propose using counterfactual explainers to generate counterfactuals for observed training samples. Specifically, we
 180 have

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

181 where $\mathcal{E}_\theta(\cdot)$ denotes the energy function. We generate samples \mathbf{x}' by first randomly sampling the target class $y^+ \sim p(y)$ and then generating a counterfactual explanation for that target, similar to how conditional sampling is used to
 182 draw from $p_\theta(\mathbf{x})$ in Grathwohl et al. (2020). In particular, we set $\mathcal{E}_\theta(\mathbf{x}, \mathbf{y}) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$ where y^+ denotes the
 183 index of the target class.

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

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

194 3.1.2 Indirectly Inducing Explainability through Adversarial Robustness

195 Based on our analysis in Section 2, counterfactuals \mathbf{x}' can be repurposed as additional training samples (Luu and Inoue
 196 2023) or adversarial examples (Friesleben 2022; Pawelczyk et al. 2022). This leaves some flexibility with respect to
 197 the exact choice for $\text{advloss}(\cdot)$ in Equation 1. An intuitive functional form to use, though likely not the only reasonable
 198 choice, is inspired by adversarial training:

$$\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}; \varepsilon) = \begin{cases} \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}) & \text{if } \|\Delta\|_\infty \leq \varepsilon \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

199 Under this choice we treat the counterfactual \mathbf{x}' as an adversarial example iff it is imperceptible, i.e. the magnitude of
 200 the perturbation of any individual feature is upper-bounded at ε .

201 3.2 Encoding Actionability Constraints

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

208 Since actionability constraints are binding at test time, we should also impose them when generating \mathbf{x}' during each
 209 training iteration to align model representations with user requirements. Through their effect on \mathbf{x}' , both types of
 210 constraints influence model outcomes through Equation 2. Here it is crucial that we avoid penalizing implausibility
 211 that arises due to mutability constraints. For any mutability-constrained feature d this can be achieved by enforcing
 212 $\mathbf{x}[d] - \mathbf{x}'[d] := 0$ whenever perturbing $\mathbf{x}'[d]$ in the direction of $\mathbf{x}[d]$ would violate mutability constraints. Specifically,
 213 we set $\mathbf{x}[d] := \mathbf{x}'[d]$ if

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

217 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
 218 respect to feature d . Intuitively, we think of this simply in terms ignoring implausibility costs with respect to im-
 219 mutable features, which effectively forces the model to instead seek plausibility with respect to the remaining features.
 220 This in turn results in lower overall sensitivity to immutable features, which we demonstrate empirically below and
 221 analytically for the linear case in Section B.1 of the supplementary appendix.

222 **4 Experiments**

223 **4.1 Experimental Setup**

224 **4.2 Experimental Results**

225 **5 Discussion**

226 **6 Conclusion**

227 **References**

- 228 Abbasnejad, Ehsan, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. “Counterfactual
229 Vision and Language Learning.” In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition
230 (CVPR)*, 10041–51. <https://doi.org/10.1109/CVPR42600.2020.01006>.
- 231 Altmeyer, Patrick, Arie van Deursen, et al. 2023. “Explaining Black-Box Models Through Counterfactuals.” In
232 *Proceedings of the JuliaCon Conferences*, 1:130. 1.
- 233 Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia CS Liem. 2024. “Faithful Model Explanations
234 Through Energy-Constrained Conformal Counterfactuals.” In *Proceedings of the AAAI Conference on Artificial
235 Intelligence*, 38:10829–37. 10.
- 236 Augustin, Maximilian, Alexander Meinke, and Matthias Hein. 2020. “Adversarial Robustness on in-and Out-
237 Distribution Improves Explainability.” In *European Conference on Computer Vision*, 228–45. Springer.
- 238 Du, Yilun, and Igor Mordatch. 2020. “Implicit Generation and Generalization in Energy-Based Models.” <https://arxiv.org/abs/1903.08689>.
- 239 Freiesleben, Timo. 2022. “The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples.”
240 *Minds and Machines* 32 (1): 77–109.
- 241 Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2014. “Explaining and Harnessing Adversarial Examples.”
242 <https://arxiv.org/abs/1412.6572>.
- 243 Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- 244 Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swer-
245 sky. 2020. “Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One.” In *International
246 Conference on Learning Representations*.
- 247 Guidotti, Riccardo. 2022. “Counterfactual Explanations and How to Find Them: Literature Review and Benchmark-
248 ing.” *Data Mining and Knowledge Discovery*, 1–55.
- 249 Guo, Hangzhi, Thanh H. Nguyen, and Amulya Yadav. 2023. “CounterNet: End-to-End Training of Prediction Aware
250 Counterfactual Explanations.” In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery
251 and Data Mining*, 577–89. KDD ’23. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3580305.3599290>.
- 251 Kolter, Zico. 2023. “Keynote Addresses: SaTML 2023 .” In *2023 IEEE Conference on Secure and Trustworthy
252 Machine Learning (SaTML)*, xvi–. Los Alamitos, CA, USA: IEEE Computer Society. <https://doi.org/10.1109/SaTML54575.2023.00009>.
- 253 Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. “Simple and Scalable Predictive Uncer-
254 tainty Estimation Using Deep Ensembles.” *Advances in Neural Information Processing Systems* 30.
- 255 Lippe, Phillip. 2024. “UvA Deep Learning Tutorials.” <https://uvadlc-notebooks.readthedocs.io/en/latest/>.
- 256 Luu, Hoai Linh, and Naoya Inoue. 2023. “Counterfactual Adversarial Training for Improving Robustness of Pre-
257 Trained Language Models.” In *Proceedings of the 37th Pacific Asia Conference on Language, Information and
258 Computation*, 881–88.
- 259 Murphy, Kevin P. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.
- 260 O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.
261 Crown.
- 262 Pawelczyk, Martin, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. “Exploring
263 Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis.”
264 In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, edited by Gustau
265 Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, 151:4574–94. Proceedings of Machine Learning Research.
266 PMLR. <https://proceedings.mlr.press/v151/pawelczyk22a.html>.
- 267 Poyiadzi, Rafael, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. “FACE: Feasible and
268 Actionable Counterfactual Explanations.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*,
269 344–50.
- 270 Ross, Alexis, Himabindu Lakkaraju, and Osbert Bastani. 2024. “Learning Models for Actionable Recourse.” In
271 *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS ’21. Red
272 Hook, NY, USA: Curran Associates Inc.

- 277 Sauer, Axel, and Andreas Geiger. 2021. “Counterfactual Generative Networks.” <https://arxiv.org/abs/2101.06046>.
- 278 Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. “Generating
279 Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties.” In
280 *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.
- 281 Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.
282 2013. “Intriguing Properties of Neural Networks.” <https://arxiv.org/abs/1312.6199>.
- 283 Teney, Damien, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. “Learning What Makes a Difference from
284 Counterfactual Examples and Gradient Supervision.” In *Computer Vision–ECCV 2020: 16th European Confer-
285 ence, Glasgow, UK, August 23–28, 2020, Proceedings, Part x 16*, 580–99. Springer.
- 286 Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations Without Opening the Black
287 Box: Automated Decisions and the GDPR.” *Harv. JL & Tech.* 31: 841. <https://doi.org/10.2139/ssrn.3063289>.
- 288 Wilson, Andrew Gordon. 2020. “The Case for Bayesian Deep Learning.” <https://arxiv.org/abs/2001.10995>.
- 289 Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. “Polyjuice: Generating Counterfactuals
290 for Explaining, Evaluating, and Improving Models.” In *Proceedings of the 59th Annual Meeting of the Associa-
291 tion for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing
292 (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 6707–23. Online:
293 Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.523>.

294 **A Notation**

- 295 • y^+ : The target class and also the index of the target class.
 296 • \mathbf{y}^+ : The one-hot encoded output vector for the target class.
 297 • θ : Model parameters (unspecified).
 298 • Θ : Matrix of parameters.

299 **B Training Details**300 **B.1 Details on Encoding Mutability Constraints**

301 In Section 3.2 we explained that to avoid penalizing implausibility that arises due to mutability constraints for any
 302 feature d , we impose a point mass prior on $p(\mathbf{x})$ for that features. This is best explained through a simple example
 303 involving a binary linear classifier.

304 Let $\mathbf{M}_\theta(\mathbf{x}) = \Theta\mathbf{x}$, with $\mathbf{x} \in \mathbb{R}^2$ and $y \in \{0, 1\}$. Then for any given counterfactual with $y^+ \sim p(y) = 1$ we have:

$$\nabla_\theta (\text{div}(\mathbf{x}, \mathbf{x}', \mathbf{y}; \theta)) = \nabla_\theta (-\mathbf{M}_\theta(\mathbf{x})[y^+]) - \nabla_\theta (-\mathbf{M}_\theta(\mathbf{x}')[y^+]) = \begin{pmatrix} 0 & 0 \\ \mathbf{x}'[1] - \mathbf{x}[1] & \mathbf{x}'[2] - \mathbf{x}[2] \end{pmatrix} \quad (4)$$

305 For the given counterfactual, contrastive divergence pushes down (up) on the coefficient $\Theta[y^+, d]$ by the amount that
 306 feature d that is lower (higher) for counterfactuals than samples. This has the intended effect of aligning counterfactual
 307 explanations with the data, as discussed above.

308 **Example B.1.** Let feature $d = 1$ in Equation 4 represents *age* and our model predicts the likelihood of credit default:
 309 $y \in \{0 := \text{'no default'}, 1 := \text{'default'}\}$. Then if $\mathbf{x}'[1] < \mathbf{x}[1]$, contrastive divergence will push down on the corre-
 310 sponding coefficient, $\Theta[\text{'default', 'age'}$]. Intuitively, this has the effect of decreasing the effect of *age* that the model
 311 attributes to the log odds of default, i.e. the model learns to predict higher default risk for younger individuals.

312 If $\Theta[y^+, d] < 0$ even in absence of this force, then contrastive divergence exacerbates the existing effect. But if *age* is
 313 immutable, then this corresponds to penalizing younger individuals for implausibility that arises due to a feature that
 314 they can change about themselves. This is precisely what we want to avoid, and therefore we impose $\mathbf{x}'[1] - \mathbf{x}[1] :=$
 315 0 in such cases. In doing so, we induce lower sensitivity with respect to *age* compared to other mutable features
 316 (Figure A1).

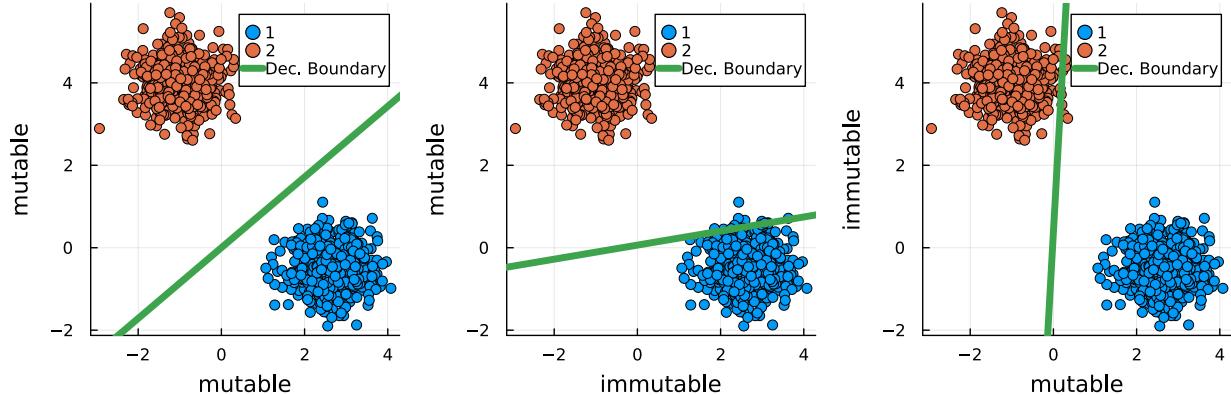


Figure A1: Visual illustration of the effect of imposing mutability constraints.

⚠ Warning

@Cynthia, @Arie, the above still needs to be polished. I'm not sure how to demonstrate/prove this last point analytically.

317

318 **C Detailed Results**319 **C.1 Initial Grid Search**

320 For the initial round of experiments we

Table A1: Results for Linearly Separable data by energy penalty.

Table A2: Results for Linearly Separable data by cost penalty.

321 **C.1.1 Generator Parameters**

322 The hyperparameter grids for the first investigation of the effect of generator parameters are shown in Parameters C.1
 323 and Parameters C.2.

324 **Parameters C.1** (Training Phase).

325 **Parameters C.2** (Evaluation Phase).

326 **C.1.1.1 Linearly Separable**

- 327 • **Energy Penalty** (Table A1): *ECCo* generally does yield better results than *Vanilla* for higher choices of the
 328 energy penalty (10,15) during training. *Generic* performs poorly across the board. *Omni* seems to have an
 329 anchoring effect, in that it never performs terribly but also never as good as the best *ECCo* results. *REVISE*
 330 performs poorly across the board.
- 331 • **Cost** (Table A2): Results for all generators (except *Omni*) are quite bad, which can likely be attributed to
 332 extremely bad results for some choices of the **Energy Penalty** (results here are averaged). For *ECCo* and
 333 *Generic*, higher cost values generally lead to worse results.
- 334 • **Maximum Iterations**: No clear patterns recognizable, so it seems that smaller choices are ok.
- 335 • **Validity**: *ECCo* almost always valid except for very low values during training and high values at evaluation
 336 time. *Generic* often has poor validity.
- 337 • **Accuracy**: Seems largely unaffected.

338 **C.1.1.2 Moons**

- 339 • **Energy Penalty** (Table A3): *ECCo* consistently yields better results than *Vanilla*, except for very low choices
 340 of the energy penalty during training for which it performs abysmal. *Generic* performs quite badly across
 341 the board for high enough choices of the energy penalty at evaluation time. *Omni* has small positive effect.
 342 *REVISE* performs poorly across the board.
- 343 • **Cost (distance penalty)**: *Generic* generally does better for higher values, while *ECCo* does better for lower
 344 values.
- 345 • **Maximum Iterations**: No clear patterns recognizable, so it seems that smaller choices are ok.
- 346 • **Validity**: *ECCo* generally achieves full validity except for very low choices the energy penalty during training
 347 and high choices at evaluation time. *Generic* performs poorly for high choices of the energy penalty during
 348 evaluation.
- 349 • **Accuracy**: Largely unaffected although *ECCo* suffers a bit for very low choices the energy penalty during
 350 training. *REVISE* suffers a lot in general (around 10 percentage points).

351 **C.1.1.3 Circles**

- 352 • **Energy Penalty** (Table A4): *ECCo* consistently yields better results than *Vanilla*, though primarily for low to
 353 medium choices of the energy penalty (≤ 5) during training. The same goes for *Generic*, which sometimes
 354 outperforms *ECCo* (for small energy penalty at evaluation time). *Omni* does alright for lower energy penalty
 355 at evaluation time, but loses out for higher choices. *REVISE* performs poorly across the board (except very
 356 low choices at evaluation time).
- 357 • **Cost (distance penalty)**: *ECCo* and *Generic* generally achieve the best results when no cost penalty is used
 358 during training. Both *Omni* and *REVISE* are largely unaffected.
- 359 • **Maximum Iterations**: *ECCo* consistently yields better results for higher numbers of iterations. *Generic*
 360 generally does best for a medium number (50). *Omni* is sometimes invalid (???).
- 361 • **Validity**: *ECCo* tends to outperform its *Vanilla* counterpart, though primarily for low to medium choices of
 362 the energy penalty (≤ 5) during training and evaluation. *Vanilla* typically worse across the board.

Table A3: Results for Moons data by energy penalty.

Table A4: Results for Circles data by energy penalty.

- 363 • **Accuracy:** Mostly unaffected, but *REVISE* again consistently shows some deterioration and *ECCo* deteriorates for
364 high choices of energy penalty during training, reflecting other outcomes above.