Backtracking Counterfactuals for Deep Structural Causal Models

Klaus-Rudolf Kladny[1] **Julius von Kügelgen**[2] **Bernhard Schölkopf**[1] **Michael Muehlebach**[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]ETH Zurich, Switzerland

Abstract

Counterfactuals answer questions of what would have been observed under altered circumstances and can therefore offer valuable insights. Whereas the classical interventional interpretation of counterfactuals has been studied extensively, *backtracking* constitutes a less studied alternative where all causal laws are kept intact. In the present work, we introduce a practical method called *deep backtracking counterfactuals* (DeepBC) for computing backtracking counterfactuals in structural causal models that consist of deep generative components. We employ constrained optimization to generate counterfactuals for high-dimensional data and conduct experiments on a modified version of MNIST.

1 INTRODUCTION

The classical literature in causality constructs counterfactuals by actively manipulating causal relationships (*interventional counterfactuals*), which has been contested by some psychologists and philosophers (Rips, 2010; Gerstenberg et al., 2013; Lucas & Kemp, 2015). Instead, they have proposed an account of counterfactuals where alternate worlds are derived by tracing changes back to background conditions while leaving all causal mechanisms intact. This type of counterfactual is termed *backtracking counterfactual* (Lewis, 1979; Khoo, 2017).

Recently, von Kügelgen et al. (2023) have formalized backtracking counterfactuals within the structural causal model (SCM; Pearl (2009)) framework. However, implementing this formalization for deep SCMs (Pawlowski et al., 2020) poses computational challenges due to steps such as marginalizations and the evaluation of distributions that are intractable. The present work addresses these challenges and offers a computationally efficient implementation by framing the generation of counterfactuals as a constrained optimization problem by only computing a single, "most likely" solution. A more comprehensive account that also includes a strategy for sampling backtacking counterfactuals is presented by Kladny et al. (2023).

"What would have been, had the intensity (i) been higher?"

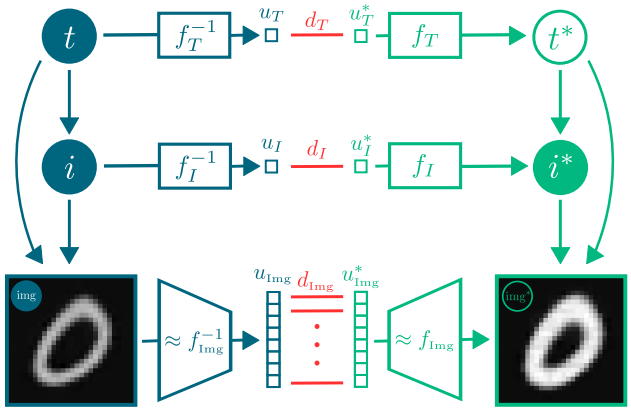

Figure 1: **Visualization of DeepBC for Morpho-MNIST.** We generate a counterfactual (green) image img* and thickness t^* with antecedent intensity i^* for the factual, observable realizations (filled blue) img, t, i. Our approach finds new latent variables $\mathbf{u}^*$ that are close with respect to distances d_i to the factual latents $\mathbf{u}$, subject to rendering the antecedent i^* true. The causal mechanisms in the factual world remain unaltered in the counterfactual world. In this specific distribution, thickness and intensity are positively related, thus rendering the image both more intense and thicker in the counterfactual. Dependence of f_i on graphical parents is omitted for simplifying visual appearance.

1.1 INVERTIBLE SCMS

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a collection of potentially high-dimensional observable "endogenous" random variables. For instance, these variables could be high-dimensional objects such as images (e.g., the MNIST im-

age in Fig. 1) or scalar feature variables (such as t and i in Fig. 1). The causal relationships among the X_i are specified by a directed acyclic graph G that is known. An SCM (Pearl, 2009) is characterized by a collection of structural equations $X_i \leftarrow f_i(\mathbf{X}_{\text{pa}(i)}, U_i)$, for $i = 1, 2, ..., n$, where $\mathbf{X}_{\text{pa}(i)}$ are the causal parents of X_i as specified by G and $\mathbf{U} = (U_1, U_2, ..., U_n)$ are exogenous latent variables. The acyclicity of G ensures that for all i, we can recursively solve for X_i to obtain a deterministic expression in terms of $\mathbf{U}$. Thus, there exists a unique function that maps $\mathbf{U}$ to $\mathbf{X}$, which we denote by $\mathbf{F}$,

$$\mathbf{X} = \mathbf{F}(\mathbf{U}), \tag{1}$$

and which is known as the reduced-form expression. We see that $\mathbf{F}$ induces a distribution over observables $\mathbf{X}$, for any given distribution over the latents $\mathbf{U}$. For DeepBC, we consider all structural equations f_i as deep invertible generative models like normalizing flows and variational autoencoders that are learned from data.

1.2 DEEPBC OPTIMIZATION OBJECTIVE

We compute the mode of the backtracking distribution $p(\mathbf{x}^* \mid \mathbf{x}^*_S, \mathbf{x})$, i.e., a single "most likely" counterfactual $\mathbf{x}^*$ for the factual realization $\mathbf{x}$ as a solution to the following constrained optimization problem:

$$\arg\min_{\mathbf{x}'} \sum_{i=1}^{n} d_i \left(\mathbf{F}_i^{-1}(\mathbf{x}'),\ \mathbf{F}_i^{-1}(\mathbf{x})\right) \tag{2}$$

$$\text{subject to} \quad \mathbf{x}'_S = \mathbf{x}^*_S, \tag{3}$$

where the d_i are differentiable distance functions. The variable $\mathbf{x}^*_S$ is the so-called antecedent, which is the explicitly altered counterfactual variable (filled green in Fig. 1).

2 EXPERIMENTS ON MORPHOMNIST

Setup. We use Morpho-MNIST, a modified version of MNIST proposed by Castro et al. (2019), to showcase how deep backtracking contrasts with its interventional counterpart (Pawlowski et al., 2020). The data set consists of three variables, two continuous scalars and an MNIST image of a handwritten digit, which all correspond to the observable variables (see § 1.1), depicted in Fig. 1. The first scalar variable T describes the thickness and the second variable I describes the intensity of the digit. They have a non-linear relationship and are positively correlated, as can be seen in Fig. 2 **(i)**, where the observational density of thickness and intensity is shown in blue. The known causal relationship between thickness and intensity is depicted in Fig. 2 (left); We train a normalizing flow for thickness and one for intensity conditionally on thickness, and model the image given T and I via a conditional β-VAE (Higgins et al., 2017). We here use $d_i(u'_i, u_i) = \|u'_i - u_i\|_2^2\ ,\ \forall i$ as the distance function and note that the u_i correspond to u_T, u_I and u_{img}.

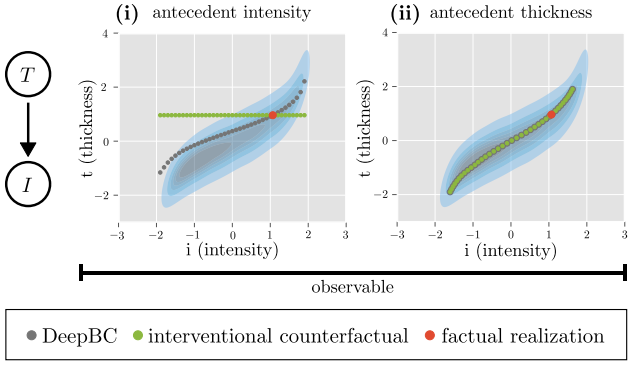

Figure 2: **Counterfactual Scalar Variables on Morpho-MNIST.** The blue shaded areas indicate the probability density of the data. **(i)** Interventional counterfactuals (green dots), in contrast to backtracking counterfactuals, leave t^* unchanged when the effect variable intensity is taken as antecendent. **(ii)** When treating thickness as the antecedent, counterfactual and backtracking counterfactuals yield identical solutions.

Results. Our results in Fig. 2 illustrate distinctive properties of the backtracking approach, in comparison to interventional counterfactuals. When choosing the effect variable intensity as the antecedent, backtracking preserves the causal laws and thus changes the upstream (cause) variable thickness accordingly to match the change in intensity as shown in Fig. 2 **(i)**. This leads to counterfactuals that resemble images from the original data set, where thickness and intensity change simultaneously.

In contrast, the interventional approach breaks the causal link from thickness to intensity when intensity is the antecedent and thus always leaves thickness unchanged, see the green dots in Fig. 2 **(i)**. This can be considered a weakness of the interventional approach, which does not yield faithful insights into the causal relationship underlying the data.

3 CONCLUSION

We presented DeepBC, a practical framework for computing backtracking counterfactuals for deep SCMs. We compared DeepBC to interventional counterfactuals. DeepBC is a general method for computing counterfactuals that measures distances between factual and counterfactual in the structured latent space of an underlying deep causal model, thus preserving the causal mechanisms in the generated counterfactuals. We hope that our approach will contribute to future developments of deep explanation methods that provide more faithful insights into the data generating process.

REFERENCES

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative Assessment and Diagnostics for Representation Learning. *Journal of Machine Learning Research*, 20 (178):1–29, 2019.

Tobias Gerstenberg, Christos Bechlivanidis, and David A. Lagnado. Back on track: Backtracking in counterfactual reasoning. In *Annual Meeting of the Cognitive Science Society*, volume 35, pp. 2386–2391, 2013.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

Justin Khoo. Backtracking Counterfactuals Revisited. *Mind*, 126(503):841–910, 2017.

Klaus-Rudolf Kladny, Julius von Kügelgen, Bernhard Schölkopf, and Michael Muehlebach. Deep Backtracking Counterfactuals for Causally Compliant Explanations. *arXiv preprint arXiv:2310.07665*, 2023.

David Lewis. Counterfactual Dependence and Time's Arrow. *Noûs*, pp. 455–476, 1979.

Christopher G. Lucas and Charles Kemp. An Improved Probabilistic Account of Counterfactual Reasoning. *Psychological Review*, 122(4):700, 2015.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Lance J. Rips. Two Causal Theories of Counterfactual Conditionals. *Cognitive Science*, 34(2):175–221, 2010.

Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking Counterfactuals. In *Conference on Causal Learning and Reasoning*, pp. 177–196, 2023.