On Data Manifolds Entailed by Structural Causal Models

Ricardo Dominguez-Olmedo [1] **Amir-Hossein Karimi** [1 2] **Georgios Arvanitidis** [3] **Bernhard Schölkopf** [1]

Abstract

The geometric structure of data is an important inductive bias in machine learning. In this work, we characterize the data manifolds entailed by structural causal models. The strengths of the proposed framework are twofold: firstly, the geometric structure of the data manifolds is causally informed, and secondly, it enables causal reasoning about the data manifolds in an interventional and a counterfactual sense. We showcase the versatility of the proposed framework by applying it to the generation of causally-grounded counterfactual explanations for machine learning classifiers, measuring distances along the data manifold in a differential geometric-principled manner.

1. Introduction

The manifold hypothesis states that most naturally occurring datasets lie near a non-linear manifold embedded in the feature space (Hastie & Stuetzle, 1989; Smola et al., 2001; Belkin & Niyogi, 2003). The geometric structure of the data manifold is a powerful inductive bias for a variety of machine learning tasks, such as classification, clustering, density estimation, representation learning, and transfer learning; across a diverse set of application domains, including computer vision (Tosi et al., 2014; Arvanitidis et al., 2018; 2021), robotics (Scannell et al., 2021; Beik-Mohammadi et al., 2021; 2022), human motion capture (Tosi et al., 2014) and protein sequencing (Detlefsen et al., 2022).

Generative models offer an appealing framework to approximately learn the data manifold from data (Arvanitidis et al., 2018). Under certain smoothness conditions, a generative model's entailed data manifold is amenable to the study of differential geometry. This allows to formally define a set of fundamental operations on the data manifold (e.g., interpolations, distances, measures) that are meaningfully informed by its geometric structure (Hauberg, 2018).

In this work, we characterize the data manifolds entailed by structural causal models (SCMs) (Pearl, 2009). SCMs are generative models that, besides modeling the data distribution, incorporate additionally knowledge about the causal relationships between the variables of the modeled system. Importantly, SCMs provide a formal framework for reasoning about the data distribution under interventions to the variables of the system, including counterfactual statements pertaining to what would have or could have been given that something else was actually observed.

Analogous to the observational, interventional, and counterfactual distributions entailed by an SCM, we first derive sufficient conditions for an SCM to induce observational, interventional, and counterfactual smooth manifolds (§3); and we show that SCM classes with broad causal identifiability results (namely additive noise models, post-nonlinear models, and location-scale noise models) satisfy such sufficient conditions. Having established the foundations for a differential geometric study of SCMs, we secondly discuss Riemannian metrics that are informed by the causal knowledge embedded in an SCM (§4.1). Thirdly, we endow the smooth manifolds induced by an SCMs with a Riemannian metric, thus characterizing as Riemannian manifolds the observational, interventional, and counterfactual data manifolds entailed by an SCM (§4.2.1). This characterization allows us to define operations on the data manifold that are informed by the causal structure of the data.

Lastly, we leverage the proposed framework to generate counterfactual explanations for machine learning classifiers. We propose methods to generate counterfactual explanations that are both causally grounded and close to the data manifold (§5). In contrast to previous approaches, we measure distances along the data manifold in a differential geometric-principled manner. We demonstrate the effectiveness of the proposed methods on two real-world datasets (§6).

2. Background

2.1. Structural causal models

A structural causal model (SCM) $\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$ (Pearl, 2009) over a set $\mathbf{X} = \{X_1, \dots, X_d\}$ of d random variables

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zürich, Zürich, Switzerland [3]Technical University of Denmark, Lyngby, Denmark. Correspondence to: Ricardo Dominguez-Olmedo <ricardo.olmedo@tuebingen.mpg.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

(the *endogenous* variables) consists of:

(i) A set $\mathbf{S} = \left\{X_i := f_i\left(\mathbf{X}_{\text{pa}(i)}, U_i\right)\right\}_{i=1}^{d}$ of *structural assignments*, each describing as a function f_i the causal relationship between a variable X_i, its direct causal parents $\mathbf{X}_{\text{pa}(i)}$, and a random variable U_i representing unobserved background factors of variation.

(ii) A joint distribution $P_{\mathbf{U}}(U_1, \ldots, U_d)$ over the set of *exogenous* noise variables $\mathbf{U} = \{U_1, \ldots, U_d\}$.

We henceforth make two common assumptions:

Assumption 2.1 (Acyclicity)**.** *The causal graph $\mathcal{G}$ implied by the structural assignments $\mathbf{S}$, with nodes $\mathbf{X} \cup \mathbf{U}$ and edges $\{(v, \mathbf{X}_i) : v \in \mathbf{X}_{pa(i)} \cup \mathbf{U}_i\}_{i=1}^{d}$, is acyclic.*

Under acyclicity, each realization u of the exogenous variables $\mathbf{U}$ entails a unique realization $x = f(u)$ of the endogenous variables $\mathbf{X}$, where f is the *reduced-form mapping* obtained by recursive substitution of the structural assignments $\mathbf{S}$ in topological order of the causal graph $\mathcal{G}$. The SCM $\mathcal{M}$ then entails a unique joint distribution $P_{\mathbf{X}}$ over the endogenous variables $\mathbf{X}$, i.e., the *observational distribution*.

Definition 2.2 (Entailed distribution)**.** *The entailed distribution $P_{\mathbf{X}}$ of an SCM $\mathcal{M} := (\mathbf{S}, P_{\mathbf{U}})$ is the pushforward measure of $P_{\mathbf{U}}$ through the reduced-form mapping f of $\mathbf{S}$*

$$P_{\mathbf{X}}(\mathbf{X} = x) := P_{\mathbf{U}}(\mathbf{U} = f^{-1}(x)) \tag{1}$$

where f^{-1} is the preimage of the reduced-form mapping f.

We additionally assume that there are no hidden confounders causally affecting more than one of the observables $\mathbf{X}$:

Assumption 2.3 (Causal sufficiency)**.** *The exogenous variables $U_1, \ldots, U_D$ are independent, that is, their distribution factorizes as $P_{\mathbf{U}} = P_{U_1} \times \ldots \times P_{U_D}$.*

Interventions SCMs allow for modeling and evaluating the effect of external manipulation of the system modeled by the SCM (Pearl, 2009). Interventions are modeled by modifying some of the structural assignments, resulting in a modified SCM $\mathcal{M}^{\mathcal{I}} = (\mathbf{S}^{\mathcal{I}}, P_{\mathbf{U}}^{\mathcal{I}})$. Its entailed distribution $P_{\mathbf{X}}^{\mathcal{I}}$ is then an *interventional distribution*. *Hard interventions* $\mathcal{I} := do(\mathbf{X}_{\mathcal{I}} = \theta)$ (Pearl, 2009) fix the values of a subset $\mathcal{I} \subseteq \{1, \ldots, d\}$ of the endogenous variables to some given value $\theta \in \mathbb{R}^{|\mathcal{I}|}$, such that $\mathbf{S}_{\mathcal{I}_i}^{\mathcal{I}} := \mathbf{X}_{\mathcal{I}_i} = \theta_i$ for $i \in \{1, \ldots, |\mathcal{I}|\}$ and $\mathbf{S}_i^{\mathcal{I}} := \mathbf{S}_i \; \forall i \notin \mathcal{I}$. Notably, hard interventions sever the causal relationship between intervened-upon variables and all their causal ancestors.

Counterfactuals SCMs offer a principled framework to reason about counterfactuals, that is, what would have happened under certain hypothetical interventions all else being equal (Pearl, 2009). Formally, the *counterfactual distribution* $P_{\mathbf{X}|x}^{\mathcal{I}}$ pertaining to some observation x under some hypothetical intervention $\mathcal{I}$ is defined to be the entailed distribution of the modified SCM $\mathcal{M}^{\mathcal{I}, X=x} := (\mathbf{S}^{\mathcal{I}}, P_{\mathbf{U}|x})$, where $P_{\mathbf{U}|x}$ is posterior over $\mathbf{U}$ given the observed x. Essentially, the posterior $\mathbf{U}|x$ amounts to the background conditions likely to have resulted in the observation x, which are propagated through the intervened-upon structural assignments $S^{\mathcal{I}}$. Under some hard intervention $\mathcal{I}$, if f is invertible then each observable x is mapped to exactly one counterfactual x^{CF} since the posterior $\mathbf{U}|x$ collapses to a single realization of the exogenous variables. For a hypothetical intervention $\mathcal{I}$, we denote the map from factuals to counterfactuals as $x^{\text{CF}} = \mathbb{CF}(x, \mathcal{I})$.

Causal identifiability Identifying the true causal relationships between variables (i.e., the causal graph $\mathcal{G}$) from observational data alone is in general not possible without further assumptions, such as restrictions on the function class of the structural assignments $\mathbf{S}$. Classes of SCMs identifiable from observational data alone include additive noise models of the form $\mathbf{S}_i := f_i(\mathbf{X}_{\text{pa}(i)}) + U_i$ (Peters et al., 2017)[1], post-nonlinear models (Zhang & Hyvärinen, 2009), and location-scale noise models (Immer et al., 2022).

2.2. Riemannian manifolds

A d-dimensional smooth manifold $\mathcal{M}$ is a topological space which locally resembles $\mathbb{R}^d$ and has a smooth structure. A Riemannian manifold (do Carmo, 1992) is a smooth manifold $\mathcal{M}$ equipped with a *Riemannian metric*:

Definition 2.4 (Riemannian metric)**.** *A Riemannian metric $\mathbf{M} : \mathcal{M} \to \mathcal{S}_{++}^{d}$ is a smooth function that assigns a symmetric positive definite matrix to any point in $\mathcal{M}$.*

Intuitively, the Riemannian metric defines an infinitesimal notion of distance on the manifold $\mathcal{M}$, thus endowing the manifold $\mathcal{M}$ with a particular metric structure. The length of a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ on $\mathcal{M}$ is then defined as

$$\mathcal{L}(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^T \, \mathbf{M}(\gamma(t)) \, \dot{\gamma}(t)} \, \mathrm{d}t \tag{2}$$

where $\dot{\gamma}(t) := \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)$ denotes the velocity of the curve. A natural notion of interpolation between two points $p, q \in \mathcal{M}$ on the manifold $\mathcal{M}$ is the shortest curve on $\mathcal{M}$ that connects p and q. Distances between points on the manifold $\mathcal{M}$ are then defined as length of the shortest curve connecting them:

Definition 2.5 (Riemannian distance)**.** *The distance $d_{\mathbf{M}}(p, q)$ between two points $p, q \in \mathcal{M}$ on a Riemannian manifold $(\mathcal{M}, \mathbf{M})$ is defined as the infimum of the length of all smooth curves $\gamma : [0, 1] \to \mathcal{M}$ connecting p and q*

$$d_{\mathbf{M}}(p, q) = \inf\{\mathcal{L}(\gamma) \mid \gamma(0) = p, \; \gamma(1) = q\} \tag{3}$$

[1] Both for non-Gaussian $P_{\mathbf{U}}$ and for non-linear f_i.

The Riemannian volume measure The Riemannian volume measure $\mathrm{Vol}_{\mathbf{M}}(p) := \sqrt{\det \mathbf{M}(p)}$ provides a measure of the magnitude of the local distortion of space at $p \in \mathcal{M}$. Curves which traverse regions of the manifold with large volume measure tend to have large lengths, and shortest paths between points of the manifold tend to avoid, if possible, such regions with large volume measure.

The pullback metric Let $\phi : \mathcal{W} \to \mathcal{M}$ be a smooth mapping between two smooth manifolds $\mathcal{W}$, $\mathcal{M}$, and let $\mathcal{M}$ be equipped with a Riemannian metric $\mathbf{M}$. Then, the metric $\mathbf{M}$ can be "pulled back" to $\mathcal{W}$ via the *pullback metric*:

$$\mathbf{W}(w) := (D\phi(w))^T \mathbf{M}(\phi(w)) D\phi(w) \tag{4}$$

where $D\phi(w)$ is the Jacobian of ϕ at $w \in \mathcal{W}$. Via the pullback $\mathbf{W}$, the manifold $\mathcal{W}$ inherits the infinitesimal notion of distance of the manifold $(\mathcal{M}, \mathbf{M})$. If ϕ is an immersion[2], then it holds that $\mathbf{W}$ is a Riemannian metric. If and only if ϕ is additionally injective (i.e., a diffeomorphism[3]), then ϕ is an isometry and the manifolds $(\mathcal{W}, \mathbf{W})$ and $(\mathcal{M}, \mathbf{M})$ induce identical Riemannian distances:

$$d_{\mathbf{W}}(p, q) = d_{\mathbf{M}}(\phi(p), \phi(q)) \tag{5}$$

Conformal equivalence Two metrics $\mathbf{M}$, $\mathbf{M}'$ in $\mathcal{M}$ are conformally equivalent if there exists some smooth function $\lambda : \mathcal{M} \to (0, \infty)$ such that $\mathbf{M}'(p) = \lambda(p)\mathbf{M}(p)\ \forall p \in \mathcal{M}$. The function λ is commondly denoted as the *conformal factor*. Intuitively, $\mathbf{M}$ and $\mathbf{M}'$ are identical "up to scale".

3. SCMs Entail Smooth Manifolds

In this section, we present sufficient conditions for an SCM to induce observational, interventional, and counterfactual smooth manifolds. We will then equip these entailed smooth manifolds with a suitable Riemannian metric in order to characterize the entailed data manifolds of an SCM (§4).

Our starting point of study is the exogenous space $\mathcal{U}$, that is, the space of realizations of the exogenous variables $\mathbf{U}$. We argue that the exogenous space $\mathcal{U}$ is a smooth manifold for typical modeling choices of the exogenous distribution $P_{\mathbf{U}}$. In particular, under causal sufficiency it suffices that the support of every marginal P_{U_i} is a smooth manifold.

Observation 3.1. *Under causal sufficiency, if the support of every marginal P_{U_i} is a d_i-dimensional smooth manifold, then the exogenous space $\mathcal{U}$ is a d-dimensional smooth manifold, where $d = \sum_i d_i$.*

For SCMs with real-valued variables, typical modeling choices of the marginal distributions P_{U_i} (e.g., Gaussian or Gamma distributions) satisfy the conditions of Observation 3.1, since their support is a non-empty open interval of $\mathbb{R}$, which is trivially a 1-dimensional smooth manifold.

Consequently, we argue that in practice the exogenous space $\mathcal{U}$ generally admits a differential geometric treatment. We will now consider the endogenous space $\mathcal{X}$.

3.1. The endogenous space $\mathcal{X}$

For acyclic SCMs, the endogenous space $\mathcal{X}$ (i.e., the space of realizations of the endogenous variables $\mathbf{X}$) is precisely the image of the exogenous space $\mathcal{U}$ through the reduced-form mapping f of the SCM, that is, $\mathcal{X} := f(\mathcal{U})$. We now present sufficient conditions under which the endogenous space $\mathcal{X}$ is a smooth manifold.

Lemma 3.2. *Under acyclicity, if for all structural assignments it holds that f_i is differentiable and its partial derivative $\partial_{U_i} f_i(X_{pa(i)}, U_i)$ is nonvanishing in $\mathcal{U}$, then the reduced-form map $f : \mathcal{U} \to \mathcal{X}$ is an immersion.*

Lemma 3.3. *Under acyclicity, if for all $i \in \{1, \dots, d\}$ and all $u^{(1)}, u^{(2)} \in \mathcal{U}$ such that $u_i^{(1)} \neq u_i^{(2)}$ it holds that*

$$f_i\left(x_{pa(i)}, u_i^{(1)}\right) \neq f_i\left(x_{pa(i)}, u_i^{(2)}\right) \tag{6}$$

then the reduced-form mapping $f : \mathcal{U} \to \mathcal{X}$ is injective.

Proposition 3.4. *Under the conditions of Lemma 3.2 and Lemma 3.3, if the exogenous space $\mathcal{U}$ is a smooth manifold then the endogenous space $\mathcal{X} := f(\mathcal{U})$ is a smooth manifold and the map $f : \mathcal{U} \to \mathcal{X}$ is a diffeomorphism.*

Corollary 3.5. *For additive noise models, post-nonlinear models, and location-scale noise models, if the exogenous space $\mathcal{U}$ is a smooth manifold and the structural assignments f_i are differentiable, then the endogenous space $\mathcal{X} := f(\mathcal{U})$ is a smooth manifold and $f : \mathcal{U} \to \mathcal{X}$ is a diffeomorphism.*

Note that the condition of Lemma 3.3 is weaker than injectivity of every f_i, since the functions f_i are only required to be injective with respect to U_i. Corollary 3.5 is a significant result, since additive noise models and post-nonlinear models are precisely the classes of SCMs with strong causal identifiability guarantees (§2.1). Note that the differentiability condition in Corollary 3.5 is generally required to establish causal identifiability (Peters et al., 2017; Zhang & Hyvärinen, 2009), and in that sense amounts to a relatively mild condition on the structural assignments of the SCM.

We have so far presented sufficient conditions for the endogenous space $\mathcal{X}$ induced by an SCM to be a smooth manifold, and showed that particularly notable classes of SCMs satisfy such conditions. Analogous to the interventional and counterfactual distributions entailed by an SCM, we now present sufficient conditions for an SCM to entail interventional and counterfactual smooth manifolds.

[2] The rank of the Jacobian $D\phi(g)$ of ϕ at every point $p \in \mathcal{W}$ is equal to the dimensionality of the manifold $\mathcal{W}$. This ensures that the pullback metric induces a positive-definite inner product.

[3] Alternatively, both ϕ and its inverse ϕ^{-1} are differentiable.

3.2. Interventional smooth manifolds

An intervention $\mathcal{I}$ to some SCM $\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$ results in a modified SCM $\mathcal{M}^{\mathcal{I}} = (\mathbf{S}^{\mathcal{I}}, P^{\mathcal{I}}_{\mathbf{U}})$. Similarly to §3.1, we define the endogenous space $\mathcal{X}^{\mathcal{I}}$ under an intervention $\mathcal{I}$ as the image of the modified exogenous space $\mathcal{U}^{\mathcal{I}} := \operatorname{supp}(P^{\mathcal{I}}_{\mathbf{U}})$ through the reduced-form mapping $f^{\mathcal{I}}$ of the intervened-upon structural assignments $\mathbf{S}^{\mathcal{I}}$, that is, $\mathcal{X}^{\mathcal{I}} := f^{\mathcal{I}}(\mathcal{U}^{\mathcal{I}})$. If the modified SCM $\mathcal{M}^{\mathcal{I}}$ satisfies the conditions of Proposition 3.4, the interventional space $\mathcal{X}^{\mathcal{I}}$ is a smooth manifold, which denote as an *interventional manifold*.

Hard interventions $\mathcal{I} := do(\mathbf{X}_{\mathcal{I}} = \theta)$ are arguably the most common modeling choice for interventions. We show that for hard interventions $\mathcal{I}$, the interventional space $\mathcal{X}^{\mathcal{I}}$ is a smooth manifold under strictly weaker conditions than those presented in §3.1 for the observational setting.

Proposition 3.6. *Let $\mathcal{I} := do(\mathbf{X}_{\mathcal{I}} = \theta)$ be a hard intervention on the variables $\mathbf{X}_{\mathcal{I}}$. If the structural assignments $f_j \ \forall j \notin \mathcal{I}$ satisfy the conditions of Lemma 3.2 and Lemma 3.3, and $\mathcal{U}^{\mathcal{I}}$ is a smooth manifold, then interventional space $\mathcal{X}^{\mathcal{I}} := f^{\mathcal{I}}(\mathcal{U}^{\mathcal{I}})$ is a $(d-m)$-dimensional smooth manifold embedded in $\mathbb{R}^d$, where d is the number of endogenous variables and $m := |\mathcal{I}|$ is the number of intervened-upon variables. Additionally, the reduced-form mapping $f^{\mathcal{I}} : \mathcal{U}^{\mathcal{I}} \to \mathcal{X}^{\mathcal{I}}$ is a diffeomorphism.*

Corollary 3.7. *If an SCM $\mathcal{M}$ satisfies the conditions of Proposition 3.4, then $\mathcal{M}$ entails interventional smooth manifolds $\mathcal{X}^{\mathcal{I}}$ under hard interventions $\mathcal{I} := do(\mathbf{X}_{\mathcal{I}} = \theta)$.*

Consequently, when considering hard interventions, the interventional space $\mathcal{X}^{\mathcal{I}}$ admits a differential geometric study without requiring additional assumptions compared to the observational setting discussed in §3.1.

3.3. Counterfactual smooth manifolds

The sufficient conditions presented in §3.1 and §3.2 imply that the reduced-form mapping f is invertible. Then, given some observation x, abduction results in the posterior over exogenous variables $\mathbf{U}|x$ collapsing to a single realization $u = f^{-1}(x)$. Consequently, under non-stochastic interventions $\mathcal{I}$ (i.e., hard interventions) the entailed counterfactual distributions collapse to single realizations of the endogenous variables x^{CF}. In order to meaningfully define a notion of counterfactual manifolds given that counterfactual distributions collapse to single counterfactuals, we instead consider counterfactuals under a space of interventions $\mathcal{H}$.

We argue that such counterfactual manifolds commonly arise in counterfactual reasoning when considering the effects of competing hypothetical interventions; such as with the query "What dietary intervention would most favorably improve the health outcomes of some particular individual?". For simplicity, we restrict our analysis to spaces of hard interventions of the form $\mathcal{H} := \{do(\mathbf{X}_{\mathcal{I}} = \theta) \mid \theta \in \Delta\}$.

As introduced in §2.1, for any given SCM $\mathcal{M}$ let us denote by $\mathbb{CF}$ the mapping between factual and counterfactuals for some hard intervention $\mathcal{I} := do(\mathbf{X}_{\mathcal{I}} = \theta)$, such that $x^{\text{CF}} = \mathbb{CF}(x, \mathcal{I})$. We define the space of counterfactuals $\mathcal{X}^{\mathcal{H}|x}$ for some observable x under some space of interventions $\mathcal{H}$ as the image of $\mathcal{H}$ through the counterfactual mapping $\mathbb{CF}$, that is, $\mathcal{X}^{\mathcal{H}|x} := \mathbb{CF}(x, \mathcal{H})$. We now present sufficient conditions on $\mathcal{M}$ for $\mathcal{X}^{\mathcal{H}|x}$ to be a smooth manifold.

Proposition 3.8. *Let $\mathcal{H} := \{do(\mathbf{X}_{\mathcal{I}} = \theta) \mid \theta \in \Delta\}$ be a space of hard interventions on the variables $\mathbf{X}_{\mathcal{I}}$, and let $\mathbf{X}_{\mathcal{D}}$ be the causal descendants of $\mathbf{X}_{\mathcal{I}}$, excluding $\mathbf{X}_{\mathcal{I}}$. Under the conditions of Lemma 3.3 (injectivity of f), if the structural assignments f_i corresponding to the causal descendants $i \in \mathcal{D}$ are differentiable and Δ is a m-dimensional smooth manifold, then the counterfactual space $\mathcal{X}^{\mathcal{H}|x} := \mathbb{CF}(x, \mathcal{H})$ is an m-dimensional smooth manifold, and the mapping $\mathbb{CF}(x, \cdot) : \mathcal{H} \to \mathcal{X}^{\mathcal{H}|x}$ is a diffeomorphism.*

Corollary 3.9. *If an SCM $\mathcal{M}$ satisfies the conditions of Proposition 3.4, then $\mathcal{M}$ entails counterfactual smooth manifolds $\mathcal{X}^{\mathcal{H}|x}$ under spaces of hard interventions $\mathcal{H} := \{do(\mathbf{X}_{\mathcal{I}} = \theta) \mid \theta \in \Delta\}$, where Δ is a smooth manifold.*

Note that only the causal descendants $\mathcal{D}$ are required to have differentiable structural assignments, and that there are no requirements on the exogenous variables $\mathbf{U}$. In particular, the causal ancestors of the intervened-upon variables need not be real-valued for the space of counterfactuals to admit a differential geometric treatment. Such scenarios are common in socioeconomic settings, where root variables often include categorical variables (e.g., gender or nationality).

Analogously to the interventional setting, when considering hard interventions, the counterfactual space $\mathcal{X}^{\mathcal{H}|x}$ admits a differential geometric study without requiring additional assumptions compared to the observational setting (§3.1).

4. SCMs Entail Data Manifolds

In the previous section, we derived sufficient conditions on an SCM $\mathcal{M}$ such that it induces observational, interventional, and counterfactual smooth manifolds. In this section, we discuss the inductive biases of different Riemannian metrics informed by the SCM. Such metrics allow us to endow the aforementioned smooth manifolds with metric structures that are meaningfully informed by the causal structure of the data. We then characterize as Riemannian manifolds the observational, interventional, and counterfactual data manifolds entailed by an SCM.

4.1. Riemannian metrics and their inductive biases

Insofar as geometric judgments are hypotheses about the world (Riemann, 1868), the choice of Riemannian metric is a fundamental modeling tool towards encoding appropriate inductive biases for the system being modeled. Prior works

in manifold learning typically considers locally Euclidean metrics, which are then regularized to have large volume measure in regions of the feature space far away from the observed data (Hauberg et al., 2012; Tosi et al., 2014; Hauberg, 2018; Arvanitidis et al., 2016; 2018; 2022). Such modeling choice encodes the inductive bias that shortest paths on the data manifold should remain close to the observed data, since curves crossing regions of the feature space with low data density will then necessarily have large length.

4.1.1. Locally Euclidean metrics

The sufficient conditions presented in §3 establish an isometry between the exogenous space $\mathcal{U}$ and the endogenous space $\mathcal{X}$. Consequently, for any Riemannian metric $\mathbf{M}_{\mathcal{U}}$ (resp. $\mathbf{M}_{\mathcal{X}}$) defined in $\mathcal{U}$ (resp. $\mathcal{X}$), there exists an equivalent metric in $\mathcal{X}$ (resp. $\mathcal{U}$) defined by the pushforward (resp. pullback) via the reduced-form map f of the structural assignments of the SCM, such that $\mathcal{X}$ (resp. $\mathcal{U}$) inherits the infinitesimal notion of distance defined by $\mathbf{M}_{\mathcal{U}}$ (resp. $\mathbf{M}_{\mathcal{X}}$).

Locally Euclidean $\mathbf{M}_{\mathcal{X}} := I$ Encodes the inductive bias that background conditions (i.e., the exogenous variables $\mathbf{U}$) should be considered similar if they lead to similar observations (i.e., the endogenous variables $\mathbf{X}$) in a locally Euclidean sense. Loosely, the resulting metric structure places more weight on differences in outcomes rather than differences in causes. Such choice of metric is additionally well-justified for SCMs whose exogenous variables merely encode the stochasticity in the relation between causal variables, but their numerical values are arbitrary (e.g., the model would be equally useful under elementwise reparametrization of the exogenous variables). Importantly, pulling back $\mathbf{M}_{\mathcal{X}}$ allows us to define a metric in the exogenous space $\mathcal{U}$ that is grounded on the observed space $\mathcal{X}$, and that is invariant to diffeomorphic reparametrizations of $\mathcal{U}$.

Locally Euclidean $\mathbf{M}_{\mathcal{U}} := I$ Encodes the inductive bias that the similarity of observables should be measured in terms of the similarity of the background conditions which gave rise to said observables, in a locally Euclidean sense. Intuitively, if the differences in "income" and "savings" of two individuals can be explained solely due to their difference in "income" (which causally affects "savings"), then their dissimilarity may be smaller than if we were to consider two individuals for which their differences cannot be explained in terms of a single common cause. However, we emphasize that for $\mathbf{M}_{\mathcal{U}} := I$ to be a meaningful metric, the exogenous variables themselves must be intrinsically meaningful (i.e., not merely a device to encode the stochasticity in the relation between causal variables). One prominient class of models for which exogenous variables may be intrinsically meaningful are additive noise models, where they indicate the deviation of an observed variable from its "expected" state given its observed causal parents.

4.1.2. Regularizing the Riemannian metric

As stated previously, prior works in manifold learning regularize the metric to have large volume measure in regions of the feature space with low data density. SCMs model the probability distribution $P_{\mathbf{X}}$ of the endogenous variables, formally defined as the pushforward measure of $P_{\mathbf{U}}$ through the reduced-form mapping f of the SCM. We draw inspiration from Arvanitidis et al. (2022), and propose to scale the Riemannian metric $\mathbf{M}_{\mathcal{X}}$ by the conformal factor $\lambda_{\mathbf{X}}(x) := (\alpha \cdot p_{\mathbf{X}}(x) + \beta)^{-2/d}$, where $p_{\mathbf{X}}$ is the density of $P_{\mathbf{X}}$ and $\alpha, \beta > 0$ upper and lower bound $\lambda_{\mathbf{X}}$. The conformal factor $\lambda_{\mathbf{X}}$ scales the volume measure of the manifold inversely proportionally to the data density $p_{\mathbf{X}}$, such that

$$\mathrm{Vol}_{\lambda_{\mathbf{X}}\mathbf{M}_{\mathcal{X}}}(x) = \frac{\mathrm{Vol}_{\mathbf{M}_{\mathcal{X}}}(x)}{\alpha \cdot p_{\mathbf{X}}(x) + \beta} \tag{7}$$

The parameters α and β determine the local curvature of the manifold as a function of the data density $p_{\mathbf{X}}$; determining how strongly shortest paths are pulled towards regions of the space with large data density. The values of α and β are consequently a further modeling choice with which practitioners may encode appropriate inductive biases.

4.2. Data manifolds entailed by SCMs

We now characterize the data manifolds entailed by an SCM, by equipping the smooth manifolds described in §3 with a Riemannian metric $\mathbf{M}$ which is regularized by the conformal factor motivated in §4.1.2. Note that we make no assumptions on $\mathbf{M}$ and treat it as a modeling choice.

Definition 4.1 (Entailed data manifold)**.** *An entailed data manifold of an SCM $\mathcal{M} := (\mathbf{S}, P_{\mathbf{U}})$ is a Riemmanian manifold $(\mathcal{X}, \lambda_{\mathbf{X}}\mathbf{M})$ comprised of a smooth manifold $\mathcal{X} := f(\mathcal{U})$ (the endogenous space) equipped with a Riemannian metric $\mathbf{M}$ scaled by some conformal factor $\lambda_{\mathbf{X}}$*

$$\lambda_{\mathbf{X}} := (\alpha \cdot p_{\mathbf{X}}(x) + \beta)^{-2/d} \quad \alpha, \beta > 0 \tag{8}$$

where $p_{\mathbf{X}}$ is the density of the probability distribution $P_{\mathbf{X}}$ entailed by the SCM $\mathcal{M}$.

Analogous to an SCM's entailed distribution (Definition 2.2), it is possible to reason about an SCM's interventional and counterfactual data manifolds.

4.2.1. Interventional data manifolds

We define an SCM's interventional data manifold under some intervention $\mathcal{I}$ as an entailed manifold of the modified SCM $\mathcal{M}^{\mathcal{I}} = (\mathbf{S}^{\mathcal{I}}, P^{\mathcal{I}}_{\mathbf{U}})$. In particular, such interventional data manifold is comprised of the interventional smooth

manifold $\mathcal{X}^{\mathcal{J}} := f^{\mathcal{J}}(\mathcal{U}^{\mathcal{J}})$ equipped with a Riemannian metric $\mathbf{M}^{\mathcal{J}}$ scaled by some conformal factor $\lambda_{\mathbf{X}}^{\mathcal{J}}$ inversely proportional to the density of the interventional distribution $P_{\mathbf{X}}^{\mathcal{J}}$.

Note that an intervention $\mathcal{J}$ alters the observational data manifold in multiple ways: firstly, by modifying the space of possible realizations $\mathcal{X}^{\mathcal{J}}$ of the endogenous variables; secondly, by modifying the distribution over observations $P_{\mathbf{X}}^{\mathcal{J}}$ (i.e., the contents of the interventional space) which curves the data manifold; and thirdly, by modifying the Riemannian metric $\mathbf{M}^{\mathcal{J}}$ which endows the manifold with an infinitesimal notion of distance (e.g., when considering the pushforward metric through the modified reduced map $f^{\mathcal{J}}$).

4.2.2. Counterfactual data manifolds

We define an SCM's entailed counterfactual data manifold by endowing the counterfactual smooth manifolds $\mathcal{X}^{\mathcal{H}|x}$ introduced in §3.3 with a Riemannian metric $\mathbf{M}^{\mathcal{H}|x}$. Note that as discussed in §3.3, since we assume the reduced-form mapping f to be invertible, the entailed counterfactual distributions of the SCM collapse to single counterfactuals. We argue that a suitable alternative may be to simply consider the conformal factor $\lambda_{\mathbf{X}}$ corresponding to the observational data density $p_{\mathbf{X}}$, if the counterfactual space $\mathcal{X}^{\mathcal{H}|x}$ is a subset of the support of $P_{\mathbf{X}}$; that is, the if the hypothetical counterfactuals are somewhat consistent with the data distribution.

We argue that such choice of conformal factor is particularly well justified for systems where the observed distribution $P_{\mathbf{X}}$ is a "snapshot" of some process in which individual "units" $x \in \mathcal{X}$ naturally experience interventions, such as socioeconomic systems. For instance, people regularly experience or decide on "interventions" such as changes in their employment status. In such settings, we argue that the data distribution implicitly contains some counterfactual information. For instance, if the counterfactual x^{CF} of some observed x under an intervention $\mathcal{J}$ has arbitrarily small density $p_{\mathbf{X}}(x)$, then it might be likely that the intervention $\mathcal{J}$ on the individual x is not realistically feasible.

5. Implications for counterfactual explanations

Machine learning classifiers are increasingly being deployed in consequential decision-making settings. Prior work has argued that for such systems to be trustworthy, algorithmic decisions should be accompanied by an explanation, such that individuals are able to understand and possibly contest such decisions (Wachter et al., 2017; Venkatasubramanian & Alfano, 2020). *Counterfactual explanations*[4] (CFEs) have gained much popularity in recent years, as CFEs are generally thought to be both easily comprehensible and compliant with some regulatory frameworks (Wachter et al., 2017).

[4]Despite their name, counterfactual explanations in ML are generally not counterfactuals in the causal sense of Pearl (2009).

Consider the setting where a classifier $h : \mathcal{X} \to \{0, 1\}$ is used to assign either favourable or unfavourable outcomes to individuals $x \in \mathcal{X}$. For a negatively classified individual x, counterfactual explanations seek to explain the classifier's decision by searching for the "closest" individual x' that would have been favourably classified, that is,

$$\begin{aligned} \underset{x' \in \mathcal{X}}{\arg\min} \quad & d(x, x') \\ \text{s.t.} \quad & h(x') = 1 \end{aligned} \tag{9}$$

The distance function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ amounts to a modeling choice encoding the set of desiderata as to what amounts to an "effective" counterfactual explanation x'. We focus on two desiderata that have received much attention in the literature: counterfactual explanations should be realistic (i.e., well supported by the observed data) and consistent with the underlying causal structure of the world (Verma et al., 2020; Karimi et al., 2022).

Realistic CFEs Prior works argue that for a CFE x' to be realistic, there must exist some plausible path of change between the negatively classified individual x and the offered CFE x' (Joshi et al., 2019; Poyiadzi et al., 2020). This desideratum is particularly notable if the intent of the CFE is not only to aid the understanding of the classifier's decision, but also to inform individuals of what features they can realistically change in order to obtain a favourable classification in the future; a setting known as *algorithmic recourse* (AR). For instance, it may not be reasonable to recommend to an unsuccessful life insurance applicant to play more sports if said applicant is physically disabled (Poyiadzi et al., 2020). While prior works aim to model the data manifold in order to search for realistic CFEs (Joshi et al., 2019; Mahajan et al., 2019; Pawelczyk et al., 2020; Downs et al., 2020; Poyiadzi et al., 2020; Antoran et al., 2021), none adopt a principled differential geometric approach. In contrast, we propose to model the data manifold as a Riemannian manifold and to consider Riemannian distances along the manifold as the distance function d. Such principled approach ensures that there exists some "minimally costly" path of change (i.e., shortest curve) connecting an individual x and the counterfactual example x' offered to them.

Causally grounded CFEs Prior works have proposed methods to generate counterfactual examples within the rigorous causality framework of Pearl (2009). In particular, Karimi et al. (2020; 2021); Dominguez-Olmedo et al. (2022) search for a set of hypothetical hard interventions on the features of the individual x such that the resulting counterfactual would be favourably classified. In contrast, von Kügelgen et al. (2023) argue for backtracking counterfactuals, a non-interventional notion of causal counterfactuals where one instead searches over different "background contions" (i.e., exogenous variables) that would have given rise

to a favourably classified counterfactual x'. However, neither of the two approaches consider the underlying data manifold when generating counterfactual explanations.

We propose to leverage SCMs entailed data manifolds (§4) to generate counterfactual examples that both respect the underlying data manifold and are causally grounded. We present methods to generate observational counterfactual examples (i.e., "backtracking") and interventional counterfactual examples (i.e., causal algorithmic recourse).

5.1. Backtracking CFEs on the data manifold

Backtracking involves reasoning about counterfactuals by varying the exogenous variables $\mathbf{U}|x$ (i.e., the background conditions that plausibly gave rise to some observation x) without altering the structural assignments $\mathbf{S}$. Since $\mathbf{S}$ is not intervened-upon, backtracking counterfactuals are also know as "observational" counterfactuals, in contrast to the "interventional" counterfactuals of Pearl (2009).

We assume access to a know SCM $\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$ over the features $\mathbf{X}$ used for the classification of individuals $x \in \mathcal{X}$. We additionally assume that the SCM $\mathcal{M}$ satisfies the sufficient conditions of Proposition 3.4, such that the endogenous space $\mathcal{X}$ is a smooth manifold. The reduced-form mapping $f : \mathcal{U} \to \mathcal{X}$ is then invertible, and searching for backtracking CFEs as formalized by von Kügelgen et al. (2023) reduces to the following optimization problem:

$$\begin{aligned} \min_{u \in \mathcal{U}} \quad & d\left(f^{-1}(x), u\right) \\ \text{s.t.} \quad & h(f(u)) = 1 \end{aligned} \tag{10}$$

where $x \in \mathcal{X}$ is a negatively classified individual for which we seek an explanation, h is the decision-making classifier, $d : \mathcal{U} \times \mathcal{U} \to [0, \infty)$ is a distance function on the exogenous space $\mathcal{U}$, and $x' := f(u)$ is the backtracking CFE. We propose to instead search for backtracking counterfactual examples along the data manifold induced by the SCM $\mathcal{M}$. Since by assumption the reduced-form map f is a diffeomorphism, the exogenous space $\mathcal{U}$ and the endogenous space $\mathcal{X}$ are isomorphic. For any Riemannian metric in $\mathcal{X}$, there exists an equivalent Riemannian metric in $\mathcal{U}$, and vice versa. A differential geometric viewpoint thus provides a new perspective on the extent to which backtracking counterfactuals are "observational": searching for counterfactuals across the observed features $\mathcal{X}$ and across the exogenous space $\mathcal{U}$ is equivalent for equivalent choices of metric.

Without loss of generally, we search for backtracking counterfactuals along the exogenous space $\mathcal{U}$, for some choice of metric $\mathbf{M} : \mathcal{U} \to \mathcal{S}^d_{++}$. We note that, as discussed in §4.1.1, the exogenous space $\mathcal{U}$ may lack an intrinsically meaningful metric structure, and it might be appropriate to consider the pullback of a metric defined on $\mathcal{X}$. We scale the metric $\mathbf{M}$ with a conformal factor $\lambda_{\mathbf{U}}$ inversely proportional to the density of $P_{\mathbf{U}}$, as motivated in §4.1.2. Consequently, searching for counterfactual examples along the data manifold entailed by the SCM $\mathcal{M}$ is equivalent to solving the following optimization problem:

$$\begin{aligned} \min_{u \in \mathcal{U}} \quad & d_{\lambda_{\mathbf{U}}\mathbf{M}}\left(f^{-1}(x), u\right) \\ \text{s.t.} \quad & h(f(u)) = 1 \end{aligned} \tag{11}$$

where $d_{\lambda_{\mathbf{U}}\mathbf{M}}$ is the Riemannian distance function induced by the Riemannian manifold $(\mathcal{U}, \lambda_{\mathbf{U}}\mathbf{M})$.

5.2. Causal algorithmic recourse on the data manifold

Causal algorithmic recourse models recourse recommendations as hard interventions on some subset $\mathbf{X}_{\mathcal{I}}$ of the features $\mathbf{X}$ of individuals $x \in \mathcal{X}$, thus reasoning in a causally-principled manner about the downstream causal effects of the recourse recommendations offered to individuals. Let $\mathcal{H} := \{do(\mathbf{X}_{\mathcal{I}} = \theta) \mid \theta \in \Delta\}$ be the set of hard interventions actionable for some negatively classified individual x. The causal algorithm recourse problem is formalized as the following optimization problem (Karimi et al., 2021):

$$\begin{aligned} \min_{\mathcal{I} \in \mathcal{H}} \quad & d\left(x, \mathbb{CF}(x, \mathcal{I})\right) \\ \text{s.t.} \quad & h(\mathbb{CF}(x, \mathcal{I})) = 1 \end{aligned} \tag{12}$$

where $\mathbb{CF}(x, \cdot) : \mathcal{H} \to \mathcal{X}^{\mathcal{H}|x}$ denotes the mapping between factuals and counterfactuals under hard interventions on $\mathbf{X}_{\mathcal{I}}$, and typically $d\left(x, \mathbb{CF}(x, do(\mathbf{X}_{\mathcal{I}} = \theta))\right) = \|x_{\mathcal{I}} - \theta\|$.

We propose to instead search for recourse interventions $\mathcal{I}$ along a counterfactual data manifold entailed by the SCM $\mathcal{M}$. We assume that $\mathcal{M}$ satisfies the conditions stated in Proposition 3.8, such that the space of counterfactuals $\mathcal{X}^{\mathcal{H}|x}$ is a smooth manifold. For some appropriate choice of metric $\mathbf{M} : \mathcal{X}^{\mathcal{H}|x} \to \mathcal{S}^d_{++}$, we propose to scale such metric by a conformal factor $\lambda_{\mathbf{X}}$ inversely proportional to the density of the observational distribution $P_{\mathbf{X}}$, as motivated in §4.2.2. We then consider the pullback metric $\mathbf{M}'$ of $\lambda_{\mathbf{X}}\mathbf{M}$ via the counterfactual mapping $\mathbb{CF}(x, \cdot)$. Consequently, we formalize the search for algorithmic recourse along the counterfactual data manifold entailed by the SCM $\mathcal{M}$ as:

$$\begin{aligned} \min_{\mathcal{I} \in \mathcal{H}} \quad & d_{\mathbf{M}'}\left(x, \mathbb{CF}(x, \mathcal{I})\right) \\ \text{s.t.} \quad & h(\mathbb{CF}(x, \mathcal{I})) = 1 \end{aligned} \tag{13}$$

where $d_{\mathbf{M}'}$ is the Riemannian distance function induced by the Riemannian manifold $\left(\mathcal{X}^{\mathcal{H}|x}, \mathbf{M}'\right)$, and $\mathbf{M}'$ is the pullback metric of $\lambda_{\mathbf{X}}\mathbf{M}$ via the mapping $\mathbb{CF}(x, \cdot)$.

6. Experiments

We evaluate the methods proposed in §5 against a variety of previously proposed counterfactual explanation methods.

We open source our implementation and experiments[5]. We consider the following prior art:

- Wachter et al. (2017): considers the objective function $\min_\delta \lambda \|\delta\|_2 + \ell(h(x+\delta), 1)$, where ℓ is the cross-entropy loss and λ is gradually annealed.
- REVISE (Joshi et al., 2019): similar Wachter et al. (2017) but the optimization problem is solved in the latent space $\mathcal{Z}$ of a VAE trained to reconstruct the data.
- FACE (Poyiadzi et al., 2020): searches for the closest counterfactual along a weighted nearest neighbour graph constructed from the observed data.
- Causal recourse (Karimi et al., 2021): solves the optimization problem presented in Equation 11.
- Backtracking counterfactuals (von Kügelgen et al., 2023): solves the optimization problem presented in Equation 10. We consider Euclidean distances in $\mathcal{U}$.

For CFEs, we include REVISE as a representative method of the several VAE-based approaches in the literature for realistic CFE generation. We include FACE because it considers some approximate notion of distance along the data manifold via shortest paths in a nearest neighbour graph. For causal AR, we consider the standard approach of Karimi et al. (2021). For our proposed approach, we consider both the setting where $\mathcal{U}$ is assumed locally Euclidean (*Ours-$\mathcal{U}$*) and the setting where $\mathcal{X}$ is locally Euclidean (*Ours-$\mathcal{X}$*).

Optimizing along the data manifold We solve the optimization problems in Equation 11 and Equation 13 using gradient descent. We compute Riemannian distances by solving for the geodesic $\gamma^* : \gamma : [0,1] \to \mathcal{M}$ which connects the two points of interest $\gamma^*(0) = u_0$, $\gamma^*(1) = u_1$, such that $d_{\mathbf{M}}(u_0, u_1) := \mathcal{L}(\gamma^*)$. We compute the geodesic γ^* by solving the boundary value problem (BVP)

$$\ddot{\gamma}_t = g(\gamma_t, \dot{\gamma}_t) \quad \gamma(0) = u_0,\ \gamma(1) = u_1 \tag{14}$$

where g is a system of ordinary differential equations determined by the Riemannian metric of the manifold (do Carmo, 1992). We use the versatile automatic differentiation system of JAX (Bradbury et al., 2018) to differentiate through the boundary conditions of the BVP, which we solve using a fourth order collocation algorithm with residual control similar to Kierzenka & Shampine (2001).

SCMs and datasets We consider two real-world data sets: the COMPAS recidivism dataset (Larson et al., 2016) and the Adult demographic dataset (Kohavi & Becker, 1996), for which we assume the causal graphs presented in Nabi & Shpitser (2018). We assume additive noise model SCMs, and we regress the structural equations using MLPs with one hidden layer. We model the probability density of the residuals (i.e., $\mathbf{U}$) using kernel density estimation.

Evaluation metrics We evaluate the counterfactuals generated by each of the methods with the following metrics:

- L_2: ℓ_2 distance between the factual and counterfactual, or norm of the recommended feature change $\|\delta\|_2$.
- $L_{\mathcal{U}}$ (resp. $L_{\mathcal{X}}$): Riemannian distance induced by the data manifold entailed by the SCM $\mathcal{M}$, where the Riemannian metric is locally Euclidean in $\mathcal{U}$ (resp. $\mathcal{X}$) and scaled by a conformal factor $\lambda_{\mathbf{U}}$ (resp. $\lambda_{\mathbf{X}}$) inversely proportional to the density of $P_{\mathbf{U}}$ (resp. $P_{\mathbf{X}}$).
- $L_{\mathcal{M}}$: Riemannian distance induced by a data manifold constructed using kernel density estimation, with a locally Euclidean metric in feature space. We include this metric to test whether, despite the restrictive functional assumptions made on the SCM (i.e., additive noise), the CFEs generated generalize well to manifolds learned without such functional assumptions.

Other experimental details For prediction, we train both logistic regression (LR) classifiers as well as neural network (NN) classifiers with two hidden layers. We search for counterfactuals for the negatively classified individuals in the test set. When searching for counterfactuals, we only allow changes to real-valued features. The experimental results are averaged over five random seeds.

6.1. Results for counterfactual explanations

We present the results for CFE generation in Table 1. As expected, methods that do not explicitly consider the underlying geometry of the data manifold (Wachter and backtracking counterfactuals) achieve lowest L_2 distances for most classifiers and datasets. We additionally observe that FACE and REVISE generally result in counterfactuals that are closer along the data manifold compared to the method of Wachter et al. (2017). In contrast, backtracking produces competitive results compared to FACE and REVISE, indicating that searching for counterfactuals along the exogenous space of the SCM may be an effective approach to search for counterfactuals along the data manifold.

As expected, our proposed methods result in counterfactuals which are closer along the entailed data manifolds of the SCM ($L_{\mathcal{U}}$ and $L_{\mathcal{X}}$), since they precisely optimize for such Riemannian distances. We observe that our proposed methods also fare favourably in terms of the the data manifold learned without the SCM ($L_{\mathcal{M}}$), indicating that, despite the functional assumptions on the SCM (i.e., additive noise models), the generated CFEs generalize well.

[5] https://github.com/RicardoDominguez/data-manifolds-scms

Table 1. Experimental results: Counterfactual examples.

	LINEAR CLASSIFIER								NN CLASSIFIER							
	ADULT				COMPAS				ADULT				COMPAS			
METHOD	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$
WACHTER	7.38	**1.65**	5.76	5.86	2.47	**0.80**	3.00	2.66	3.83	**1.88**	6.59	6.89	2.90	**0.81**	2.75	2.68
BACKTR	3.12	1.69	5.47	6.07	4.11	0.83	2.85	2.80	3.51	1.92	6.40	7.00	2.53	0.85	2.83	2.81
FACE	3.29	1.85	5.50	5.69	2.31	0.85	2.88	2.71	5.01	2.10	7.02	6.78	2.25	0.85	3.73	2.54
REVISE	5.64	2.18	9.02	8.71	2.22	0.92	2.57	2.53	3.87	2.21	6.35	6.46	2.55	0.96	2.83	2.90
OURS $L_{\mathcal{U}}$	2.79	1.71	**3.21**	**3.48**	2.77	0.84	**2.33**	2.33	**3.25**	1.95	**4.02**	4.58	2.74	0.86	2.51	2.52
OURS $L_{\mathcal{X}}$	**2.75**	1.70	3.43	**3.48**	**2.18**	0.81	2.35	**2.27**	3.64	1.94	4.29	**4.36**	**2.19**	0.83	**2.41**	**2.51**

Table 2. Experimental results: Algorithmic recourse.

	LINEAR CLASSIFIER								NN CLASSIFIER							
	ADULT				COMPAS				ADULT				COMPAS			
METHOD	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$	$L_{\mathcal{M}}$	L_2	$L_{\mathcal{U}}$	$L_{\mathcal{X}}$
KARIMI ET AL.	2.68	**1.49**	4.04	4.05	1.33	**0.75**	2.62	2.63	3.47	**1.84**	5.63	5.66	1.37	**0.79**	2.68	2.69
OURS $L_{\mathcal{U}}$	1.29	1.58	1.48	1.48	1.19	0.79	**2.25**	2.29	**0.86**	1.92	**1.15**	**1.15**	**1.20**	0.85	**2.20**	2.23
OURS $L_{\mathcal{X}}$	**1.09**	1.58	**1.31**	**1.32**	**1.17**	0.79	2.27	**2.27**	1.13	1.91	1.52	1.52	1.22	0.85	2.23	**2.19**

6.2. Results for algorithmic recourse

We present the results for causal algorithmic recourse in Table 2. As expected, the approach of Karimi et al. (2021), which precisely seeks to minimize the magnitude of the recourse intervention, generates recourse recommendations with smaller magnitude (L_2) compared to our proposed methods. However, our proposed methods achieve better results for all distance measures that are informed by the data manifold ($L_{\mathcal{U}}, L_{\mathcal{X}}, L_{\mathcal{M}}$). Insofar minimal distances along the data manifold is a desideratum for algorithmic recourse, our differential geometric-principled approach bridges the gap between causal algorithmic recourse and previously proposed manifold-based non-causal methods.

7. Conclusion and Outlook

In this work, we have analyzed SCMs from a novel differential geometric perspective. We first derived sufficient conditions for SCMs to admit a differential geometric study (i.e., induce smooth manifolds) in the observational, interventional and counterfactual settings; and showed that these conditions are satisfied by well-studied classes of SCMs with broad idenfibiality results, namely additive noise models, post-nonlinear models, and location-scale models. Drawing inspiration from the prior works in manifold learning, we then proposed a Riemannian characterization of the data manifolds entailed by SCMs. This characterization enables us to define operations on the data manifold (e.g., distance computations) that are informed by the causal structure of the data; and it enables to causally reason about the data manifold in an interventional and counterfactual sense.

We then leveraged the proposed framework to generate counterfactual explanations for machine learning classifiers. In contrast to previous manifold-based methods for generating counterfactual explanation, we measure distances along the data manifold in a differential-geometric principled manner, leveraging the pertinent entailed observational data manifolds. Lastly, we novelly consider the problem of manifold-based causal algorithmic recourse, for which we instead leverage the entailed counterfactual data manifolds.

The study of causal models from a differential geometric perspective is a promising avenue of future research, since both causality and manifold learning allow the introduction of strong inductive biases for machine learning. In this work, we characterize the data manifolds entailed by SCMs as deterministic manifolds; however, future work may consider different causal models (i.e., causal graphical models) and/or manifold characterizations (i.e., statistical manifolds). Lastly, in this work we leveraged the observational and counterfactual data manifolds entailed by SCMs to generate counterfactual explanations. Future works may instead consider tasks that require reasoning about the interventional data manifolds entailed by SCMs.

Acknowledgments The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Ricardo Dominguez-Olmedo. Amir-Hossein Karimi acknowledges generous founding support from NSERC, CLS, and Google.

References

Antoran, J., Bhatt, U., Adel, T., Weller, A., and Hernández-Lobato, J. M. Getting a clue: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*, 2021.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. A locally adaptive normal distribution. *Advances in Neural Information Processing Systems*, 29, 2016.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: On the curvature of deep generative models. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Arvanitidis, G., Hauberg, S., and Schölkopf, B. Geometrically enriched latent spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 631–639. PMLR, 2021.

Arvanitidis, G., Georgiev, B. M., and Schölkopf, B. A prior-based approximate latent riemannian metric. In *International Conference on Artificial Intelligence and Statistics*, pp. 4634–4658. PMLR, 2022.

Beik-Mohammadi, H., Hauberg, S., Arvanitidis, G., Neumann, G., and Rozo, L. Learning riemannian manifolds for geodesic motion skills. In *Robotics: Science and Systems*, 2021.

Beik-Mohammadi, H., Hauberg, S., Arvanitidis, G., Neumann, G., and Rozo, L. Reactive motion generation on learned riemannian manifolds. *arXiv preprint arXiv:2203.07761*, 2022.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373 – 1396, 2003.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Crampin, M. and Pirani, F. A. E. *Applicable differential geometry*. Cambridge University Press, 1994.

Detlefsen, N. S., Hauberg, S., and Boomsma, W. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1–12, 2022.

do Carmo, M. P. *Riemannian Geometry*. Birkhäuser, 1992.

Dominguez-Olmedo, R., Karimi, A. H., and Schölkopf, B. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pp. 5324–5342. PMLR, 2022.

Downs, M., Chu, J. L., Yacoby, Y., Doshi-Velez, F., and Pan, W. Cruds: Counterfactual recourse using disentangled subspaces. In *ICML Workshop on Human Interpretability in Machine Learning*, 2020.

Hastie, T. J. and Stuetzle, W. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

Hauberg, S. Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*, 2018.

Hauberg, S., Freifeld, O., and Black, M. A geometric take on metric learning. *Advances in Neural Information Processing Systems*, 25, 2012.

Immer, A., Schultheiss, C., Vogt, J. E., Schölkopf, B., Bühlmann, P., and Marx, A. On the identifiability and estimation of causal location-scale noise models. *arXiv preprint arXiv:2210.09054*, 2022.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *Safe Machine Learning workshop at ICLR*, 2019.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, pp. 265–277, 2020.

Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *ACM Computing Surveys (CSUR)*, 2022.

Kierzenka, J. and Shampine, L. F. A bvp solver based on residual control and the maltab pse. *ACM Transactions on Mathematical Software (TOMS)*, 27(3):299–316, 2001.

Kohavi, R. and Becker, B. Uci adult data set. *UCI Meachine Learning Repository*, 6, 1996.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

Mahajan, D., Tan, C., and Sharma, A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *NeurIPS 2019 Workshop "Do the right thing": Machine Learning and Causal Inference for improved decision making*, 2019.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Pawelczyk, M., Broelemann, K., and Kasneci, G. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

Riemann, B. *Ueber die Hypothesen, welche der Geometrie zu Grunde liegen*. Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen 13, 1868.

Scannell, A., Ek, C. H., and Richards, A. Trajectory optimisation in learned multimodal dynamical systems via latent-ode collocation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12745–12751. IEEE, 2021.

Smola, A. J., Mika, S., Schölkopf, B., and Williamson, R. C. Regularized principal manifolds. *Journal of Machine Learning Research*, 1:179–209, 2001. http://www.jmlr.org.

Tosi, A., Hauberg, S., Vellido Alcacena, A., and Lawrence, N. D. Metrics for probabilistic geometries. In *Uncertainty in Artificial Intelligence: proceedings of the thirtieth conference (2014): July 23-27, 2014, Quebec City, Quebec, Canada*, pp. 800–808. AUAI Press (Association for Uncertainty in Artificial Intelligence), 2014.

Venkatasubramanian, S. and Alfano, M. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 284–293, 2020.

Verma, S., Dickerson, J., and Hines, K. Counterfactual explanations for machine learning: A review. *NeurIPS 2020 Workshop: ML-Retrospectives, Surveys Meta-Analyses*, 2020.

von Kügelgen, J., Mohamed, A., and Beckers, S. Backtracking counterfactuals. *2nd Conference on Causal Learning and Reasoning*, 2023.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 647–655. AUAI Press, 2009.

A. Proofs

A.1. Observation 3.1

Under causal sufficiency, the exogenous distribution $P_{\mathbf{U}}$ factorizes as $P_{\mathbf{U}} = P_{U_1} \times \ldots \times P_{U_d}$ and consequently its support is the Cartesian product of every marginal, that is, $\operatorname{supp} P_{\mathbf{U}} = \operatorname{supp} P_{U_1} \times \ldots \times \operatorname{supp} P_{U_1}$. The Cartesian product of n-many d_i-dimensional smooth manifolds is a d-dimensional smooth manifold, where $d = \sum_{i=0}^{n} d_i$.

A.2. Lemma 3.2

Under acyclicity of the causal graph it is possible to construct the reduced-form mapping $f : \mathcal{U} \to \mathcal{X}$ by recursive substitution of the structural assignments f_i in topological order of the causal graph. If all f_i are differentiable, then it follows that f is differentiable, since the composition of differentiable functions is differentiable. Therefore, its Jacobian $Df \in \mathbb{R}^{d \times d}$ is well-defined in its domain $\mathcal{U}$, where d is the number of exogenous variables.

Since the causal graph is acyclic, let us assume without loss of generality that the endogenous variables are ordered such that $i < j \implies X_j \notin \mathbf{X}_{\text{pa}(i)}$. Then, the partial derivatives $\partial_{U_i} f_j = 0$ vanish for all $i > j$, which implies that the Jacobian Df is lower-triangular, and consequently $\det(Df) = \prod_{i=1}^{d} \partial_{U_i} f_i$. Therefore, if all partial derivatives $\partial_{U_i} f_i$ are non-vanishing in $\mathcal{U}$, it holds that $\det Df(u) \neq 0 \; \forall u \in \mathcal{U}$; and consequently, $\operatorname{rank}(Df(u)) = d \;\; \forall u \in \mathcal{U}$. Since $\mathcal{X}$ is defined as the image of $\mathcal{U}$ through f, and f is differentiable and its Jacobian has everywhere rank d, under the assumption that $\mathcal{U}$ is a d-dimensional smooth manifold, then by definition $f : \mathcal{U} \to \mathcal{X}$ is an immersion (Crampin & Pirani, 1994, pp. 243).

A.3. Lemma 3.3

Under acyclicity of the causal graph it is possible to construct the reduced-form mapping $f : \mathcal{U} \to \mathcal{X}$ by recursive substitution of the structural assignments f_i in topological order of the causal graph. Without loss of generality, let us assume that the endogenous variables are ordered such that $i < j \implies X_j \notin \mathbf{X}_{\text{pa}(i)}$. Let us denote by $\tilde{f}_j$ the recursive substitution of the structural assignments up to X_j, where $\tilde{f}_j$ is a map from $\mathcal{U}_j = \operatorname{supp} P_{U_1} \times \ldots \times \operatorname{supp} P_{U_j}$ to $(X_i)_{i=1}^{j}$. Note that the reduced-form mapping f is by definition $\tilde{f}_d$.

By the condition in Equation 6 it trivially holds that $X_1 := f_1(U_1)$ is injective (note that necessarily $X_{\text{pa}(1)} = \emptyset$).

Let us assume that $\tilde{f}_{j-1}$ is injective. Let $u^*, u' \in \mathcal{U}_j$ such that $u^* \neq u'$. If $u'_j \neq u^*_j$, then $\tilde{f}_j(u') \neq \tilde{f}_j(u^*)$ since they must differ in X_j (by the condition of Equation 6). If $u' \neq u^*$ but $u'_j = u^*_j$, then $\tilde{f}_j(u') \neq \tilde{f}_j(u^*)$ since they must differ in some X_i for $i < j$ by assumption that $\tilde{f}_{j-1}$ is injective. Consequently, $\tilde{f}_j$ is injective.

By induction, $\tilde{f}_d$ is injective, where d is the number of exogenous variables. Therefore the reduced-form map f is injective.

A.4. Proposition 3.4

Under Lemma 3.2, if $\mathcal{U}$ is a d-dimensional smooth manifold then the reduced-form mapping $f : \mathcal{U} \to \mathcal{X}$ is an immersion. By Lemma 3.3, f is injective. Consequently, $f : \mathcal{U} \to \mathcal{X}$ is a smooth embedding, such that $\mathcal{U}$ is diffeomorphic to its image. By assumption $\mathcal{U}$ is a d-dimensional smooth manifold, it follows that $\mathcal{X}$ is a d-dimensional manifold and f a diffeomorphism.

A.5. Corollary 3.5

A.5.1. Additive Noise Models

For additive noise models of the form $X_i = f(X_{\text{pa}(i)}) + U_i$. It holds that $\partial_{U_i} f_i = 1$ for all $U_i \in \operatorname{supp} P_{U_i}$ and $i \in \{1, \ldots, d\}$, and therefore, the conditions of Lemma 3.2 (i.e., non-vanishing partial derivative in $\mathcal{U}$) are satisfied. Furthermore, for any $x_{\text{pa}(i)}$ it holds that $u^{(1)} \neq u^{(2)} \implies f_i(x_{\text{pa}(i)}, u^{(1)}) \neq f_i(x_{\text{pa}(i)}, u^{(2)})$ since $f_i(x_{\text{pa}(i)}) + u^{(1)} \neq f_i(x_{\text{pa}(i)}) + u^{(2)} \; \forall u^{(1)}, u^{(2)} \in \mathcal{U}$ s.t. $u^{(1)} \neq u^{(2)}$. Consequently, the conditions of Lemma 3.3 are satisfied, and therefore Proposition 3.4 holds.

A.5.2. POST-NONLINEAR MODELS

For post-nonlinear models, the structural equations take the form $f_i(X_{\text{pa}(i)}, U_i) = g_i^{(1)}(g_i^{(2)}(X_{\text{pa}(i)}) + U_i)$ for invertible $g_i^{(1)}$. For differentiable f_i, then $\partial_{U_i} f_i(X_{\text{pa}(i)}, U_i) = \partial_{U_i} g_i^{(1)}\left(g_i^{(2)}(X_{\text{pa}(i)})\right)$ which is non-vanishing per assumption that $g_i^{(1)}$ is invertible, thus satisfying the assumptions of Lemma 3.2. Furthermore, for any $x_{\text{pa}(i)}$ it holds that $u^{(1)} \neq u^{(2)} \implies f_i(x_{\text{pa}(i)}, u^{(1)}) \neq f_i(x_{\text{pa}(i)}, u^{(2)})$, since per $g_i^{(1)}$ invertible it holds that $f_i(x_{\text{pa}(i)}, u^{(1)}) \neq f_i(x_{\text{pa}(i)}, u^{(2)}) \iff g_i^{(2)}(x_{\text{pa}(i)}) + u^{(1)} \neq g_i^{(2)}(x_{\text{pa}(i)}) + u^{(1)}$, where the RHS is satisfied $\forall u^{(1)}, u^{(2)} \in \mathcal{U}$ s.t. $u^{(1)} \neq u^{(2)}$. Consequently, the conditions of Lemma 3.3 are satisfied, and therefore Proposition 3.4 holds.

A.5.3. LOCATION-SCALE NOISE MODELS

For location-scale noise models, the structural equations take the form $f_i(X_{\text{pa}(i)}, U_i) = g_i^{(1)}((X_{\text{pa}(i)})) + g_i^{(2)}(X_{\text{pa}(i)})U_i$ where $g_i^{(2)}$ is strictly positive (i.e., maps to $\mathbb{R}_+$). Consequently, $\partial_{U_i} f_i = g_i^{(2)}(X_{\text{pa}(i)})$ which is non-vanishing since $g_i^{(2)}$ is strictly positive, thus satisfying the assumptions of Lemma 3.2. Furthermore, for any $x_{\text{pa}(i)}$ it holds that $u^{(1)} \neq u^{(2)} \implies f_i(x_{\text{pa}(i)}, u^{(1)}) \neq f_i(x_{\text{pa}(i)}, u^{(2)})$, since per assumption that g_i is strictly positive $g_i^{(2)}(x_{\text{pa}(i)})u^{(1)} \neq g_i^{(2)}(x_{\text{pa}(i)})u^{(1)}$ $\forall u^{(1)}, u^{(2)} \in \mathcal{U}$ s.t. $u^{(1)} \neq u^{(2)}$. Consequently, the conditions of Lemma 3.3 are satisfied, and therefore Proposition 3.4 holds.

A.6. Corollary 3.7

The Corollary follows directly, since for any $\mathcal{I} \subseteq \{1, \dots, d\}$, if f_i $\forall i \in \{1, \dots, d\}$ satisfies the conditions of Lemma 3.2 and Lemma 3.3, then f_i $\forall i \notin \mathcal{I}$ necessarily satisfies the conditions of Lemma 3.2 and Lemma 3.3.

A.7. Proposition 3.6

Let us consider the a hard intervention $\mathcal{J} := do(\mathbf{X}_\mathcal{I} = \theta)$, where $\mathcal{I} \subset \{1, \dots, d\}$ with cardinality $m := |\mathcal{I}|$, and let us denote the remaining indices by $\mathcal{J} := \{1, \dots, d\} \setminus \mathcal{I}$ corresponding to the non-intervened-upon variables.

Since by assumption $P_\mathbf{U}$ satisfies the conditions of Proposition 3.1, then it also holds that $P_\mathcal{U}^\mathcal{J} := P_{U_{\mathcal{J}_1}} \times \ldots \times P_{U_{\mathcal{J}_{n-m}}}$ is a $(n-m)$-dimensional smooth manifold (i.e., the Cartesian product of $(n-m)$-many 1-dimensional smooth manifolds).

Consider the interventional structural assignments $\mathbf{S}^\mathcal{J}$, where $\mathbf{S}_\mathcal{I}^\mathcal{J} := \theta$ and $\mathbf{S}_\mathcal{J}^\mathcal{J} := \mathbf{S}_\mathcal{J}$. Note that, since by assumption the causal graph $\mathcal{G}$ of $\mathcal{M}$ is acyclic (Lemmas 3.2 and 3.3), then the interventional causal graph $\mathcal{G}^\mathcal{I}$ must also be acyclic, since edges are only removed from $\mathcal{G}$ to obtain $\mathcal{G}^\mathcal{J}$. Consequently, the interventional reduced-form map $f^\mathcal{J} : \mathcal{U}^\mathcal{J} \to \mathcal{X}_\mathcal{J}^\mathcal{J}$ can be readily obtained by recursive substitution of the structural assignments $\mathbf{S}^\mathcal{J}$ in topological order of the causal graph $\mathcal{G}^\mathcal{J}$. Since every f_i is differentiable (per assumption in Lemma A.2), and the composition of differentiable functions is a differentiable function, then it follows that $f^\mathcal{J}$ is differentiable. Following the same argument as Proposition 3.4, the Jacobian $Df^\mathcal{J} \in \mathbb{R}^{(n-m)\times(n-m)}$ has rank equal to $n-m$ if $\partial_{\mathbf{U}_\mathbf{i}} f_i$ is non-vanishing for $i \in \mathcal{J}$, which holds by assumption that the SCM $\mathcal{M}$ meets the conditions of Proposition 3.2.

If the injetivity conditions of Lemma 3.3 on f_i are satisfied, then the reduced-form map $f^\mathcal{J}$ is injective, following the same argument of Appendix A.3. Consequently, $f^\mathcal{J}$ is a smooth embedding, and the interventional space $\mathcal{X}^\mathcal{J}$ is a $(n-m)$-dimensional manifold embedded in $\mathbb{R}^d$.

A.8. Proposition 3.8

By assumption the reduced-form mapping f of the SCM is invertible, and consequently, the observed $x \in \mathcal{X}$ corresponds to a unique realization $u = f^{-1}(x)$ of the exogenous variables. Let $\mathcal{I} \subset \{1, \dots, d\}$ be the indices of the intervened-upon variables, with cardinality $m := |\mathcal{I}|$, and let us denote the remaining indices by $\mathcal{J} := \{1, \dots, d\} \setminus \mathcal{I}$. Additionally, let $\mathcal{D}$ be the set of indices corresponding to the causal descendants of $\mathbf{X}_\mathcal{I}$ (excluding $\mathbf{X}_\mathcal{I}$).

By assumption, the space of intervention values θ is a m-dimensional smooth manifold Δ. Consider the interventional structural equations $\mathbf{S}_\mathcal{I}^{do(\mathbf{X}_\mathcal{I}=\theta)} := \theta$ and $\mathbf{S}_\mathcal{J}^{do(\mathbf{X}_\mathcal{I}=\theta)} := \mathbf{S}_\mathcal{J}$. Recursive substitution of such structural equations in topological order of the acyclic interventional graph $\mathcal{G}^\mathcal{I}$ results in the counterfactual reduced-form map $f^{\text{CF}} : \Delta \to \mathbb{R}^{|I|+|\mathcal{D}|}$ from the intervention variables θ to the values of the intervened upon variables $\mathbf{X}_\mathcal{I}$ and its causal descendants $\mathbf{X}_\mathcal{D}$; where

the exogenous variables are kept fixed to u and the intervention values $\theta \in \Delta$ are allowed to vary.

By assumption, the structural assignments corresponding to the causal descendants $\mathbf{X}_{\mathcal{D}}$ are differentiable, and thus the Jacobian $Df^{\text{CF}} \in \mathbb{R}^{(|\mathcal{I}|+|\mathcal{D}|)\times|\mathcal{I}|}$ is well-defined. Without loss of generality, let us assume that the endogenous variables are ordered such that $i < j \implies X_j \notin \mathbf{X}_{\text{pa}(i)}$ in the interventional graph $\mathcal{G}^{\mathcal{I}}$ corresponding to hard intervening in $\mathbf{X}_{\mathcal{I}}$. Then, it follows that the partial derivatives $\partial_{\theta_i} f_j^{\text{CF}} = 0$ vanish for all $i > j$, and consequently since $\det(Df) = \prod_{i=1}^{m} \partial_{\theta_i} f_i^{\text{CF}} = 1$ everywhere, it holds that $\text{rank}\left(Df^{\text{CF}}\right) = m$. Consequently, f^{CF} is an immersion into its image in $\mathbb{R}^{|\mathcal{I}|+|\mathcal{D}|}$.

Let us denote by $\mathcal{D}^* = \{1, \ldots, d\} \setminus (\mathcal{I} \cup \mathcal{D})$ the indices of the endogenous variables that are neither intervened upon nor causal descendants of intervened-upon variables. Consider the modified reduced-form $f^{\text{CF}'} : \Delta \to \mathbb{R}^d$ such that $f^{\text{CF}'}(\theta) := \left(X_{\mathcal{D}^*}, f^{\text{CF}}(\theta)\right)$. Since f^{CF} is differentiable and its Jacobian has rank m everywhere, it trivially holds that $f^{\text{CF}'}$ also is differentiable and its Jacobian has rank m everywhere. Then, by definition $f^{\text{CF}'}$ is an immersion into its image in $\mathbb{R}^d$.

Such image is precisely defined as the space of counterfactual $\mathcal{X}^{\mathcal{H}|x} := f^{\text{RM}'}(\Delta)$. Since f is assumed invertible, then each f_i is injective and $f^{\text{CF}'}$ is also injective following the same argument as Lemma 3.3. It then holds that $f^{\text{CF}'}$ is a smooth embedding, and $\mathcal{X}^{\mathcal{H}|x}$ is a m-dimensional smooth manifold embedded in $\mathbb{R}^d$.

A.9. Corollary 3.9

The Corollary follows directly, since if an SCM satisfies the conditions of Proposition 3.4, then $f_i \; \forall i \in \{1, \ldots, d\}$ satisfy the conditions of Lemma 3.2 and Lemma 3.3, the latter implying that $f_i \; \forall i \in \mathcal{D}$ is differentiable for any given set of causal descendants $\mathcal{D}$ of the intervened-upon variables $\mathcal{I}$.