Independent Mechanism Analysis and the Manifold Hypothesis

Shubhangi Ghosh [1,*] **Luigi Gresele** [2] **Julius von Kügelgen** [2,3]

Michel Besserve [2] **Bernhard Schölkopf** [2]

[1] Columbia University, USA
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3] University of Cambridge, United Kingdom

shubhangi.ghosh@columbia.edu
{luigi.gresele,jvk,besserve,bs}@tue.mpg.de

Abstract

Independent Mechanism Analysis (IMA) seeks to address non-identifiability in nonlinear Independent Component Analysis (ICA) by assuming that the Jacobian of the mixing function has orthogonal columns. As typical in ICA, previous work focused on the case with an equal number of latent components and observed mixtures. Here, we extend IMA to settings with a larger number of mixtures that reside on a manifold embedded in a higher-dimensional space—in line with the *manifold hypothesis* in representation learning. For this setting, we show that IMA still circumvents several non-identifiability issues, suggesting that it can also be a beneficial principle for higher-dimensional observations when the manifold hypothesis holds. Further, we prove that the IMA principle is approximately satisfied with high probability (increasing with the number of observed mixtures) when the directions along which the latent components influence the observations are chosen independently at random. This provides a new and rigorous statistical interpretation of IMA.

1 Introduction

Nonlinear Independent Component Analysis (ICA) provides a principled approach to representation learning (Gresele et al., 2020; Hyvärinen and Morioka, 2016; Khemakhem et al., 2020). It postulates that the observed variables are nonlinear mixtures of independent latent components, and focuses on whether it is possible to reconstruct the latent components from the mixtures—formalized through the notion of *identifiability*. When the mixing is nonlinear, the model is provably non-identifiable without additional assumptions (Hyvärinen and Pajunen, 1999), i.e., the latent variables cannot be recovered. Independent Mechanism Analysis (IMA; Gresele et al., 2021) seeks to address this problem by restricting the class of considered mixing functions. Specifically, IMA postulates that the columns of the Jacobian of the mixing function, which describe how each latent component *influences* the observed mixtures, are orthogonal. This can be viewed as encoding a non-statistical notion of independence among these influences which is inspired by the principle of Independent Causal Mechanisms (ICM) (Janzing and Schölkopf, 2010; Peters et al., 2017).

While identifiability of IMA remains an open question, Gresele et al. (2021) showed that IMA can circumvent certain non-identifiability issues arising in nonlinear ICA by ruling out well-known counterexamples or *spurious solutions* (Darmois, 1951; Hyvärinen and Pajunen, 1999; Locatello et al.,

[*] S.G. mostly worked on this project while at ETH, Zürich & the MPI for Intelligent Systems, Tübingen.

Accepted for the NeurIPS 2023 Workshop on Causal Representation Learning.

2019). Buchholz et al. (2022) then proved that IMA is, in fact, *locally identifiable*. Further, Reizinger et al. (2022) showed that IMA may also provide a way to explain the empirical success of Variational Autoencoders (VAEs; Kingma and Welling, 2014) in disentangled representation learning. However, all the aforementioned works on IMA consider a setting with an equal number of latent components and observed mixtures, the one most typically studied in ICA (Hyvärinen et al., 2001, 2023).[2] As a result, they do not directly apply to cases in which the observed data is high-dimensional and the latents low-dimensional, as is often the case in representation learning—e.g., for images or biomedical data.

In this work, we address this shortcoming of previous theory and generalize IMA to higher-dimensional observations. In particular, we adopt the *manifold hypothesis* (Becker and Hinton, 1992; Bengio et al., 2013; Vincent and Bengio, 2002) which posits that many high-dimensional data sets that occur in the real world actually lie along low-dimensional manifolds inside that high-dimensional space (Cayton, 2005; Fefferman et al., 2016).[3] In this spirit, we extend the analysis of IMA to the setting in which observations lie on a low-dimensional Riemannian manifold, with dimension equal to that of the latent space, embedded in a higher-dimensional observation space.

We show that IMA still helps circumvent non-identifiability issues in this scenario, in the sense that it rules out several kinds of spurious solutions when the generative model satisfies the manifold hypothesis. This suggests that IMA may also be useful for more realistic representation learning settings involving dimensionality reduction. This insight is consistent with work by Cunningham et al. (2022) which also provides empirical evidence illustrating the benefits of an orthogonality constraint akin to IMA for unsupervised representation learning with dimensionality reduction—albeit from a different perspective than the one based on nonlinear ICA and identifiability which we adopt here.

According to Gresele et al. (2021), IMA can intuitively be interpreted as "Nature choosing the direction of the influence of each source component in the observation space independently and from an isotropic prior". Based on the manifold hypothesis, we provide a quantitative argument that formalizes this statement: when the observations lie on a low-dimensional manifold in the higher-dimensional ambient space, we show that the IMA principle is approximately satisfied with high probability if the influence directions are sampled independently and isotropically in the high-dimensional space, with increasing probability as its dimensionality grows. The argument is based on a concentration inequality—*Levy's Lemma* (see, e.g., Janzing et al., 2010, Lemma 1)—and relies on a construction which generates smoothened piecewise-affine functions. These functions also play an important role in the theoretical analysis of deep neural networks (e.g., Montúfar et al., 2014). Our work thus shows that, under the manifold hypothesis, the IMA principle can be considered the consequence of a *genericity* assumption on the data generating process (Besserve et al., 2018; Freeman, 1994; Janzing et al., 2010).

Structure and Main Contributions:

- § 2 briefly reviews independent component analysis and Independent Mechanism Analysis (IMA).
- In § 3, we introduce an extension of the IMA principle under the *manifold hypothesis*.
- In § 4 we then show that certain common counterexamples to identifiability are ruled out by our extension of IMA to manifolds.
- In § 5, we show that, when the manifold hypothesis holds, the IMA principle follows from a *genericity* assumption on the data-generating process.

2 Background

Independent Comonent Analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001) assumes a data-generating process where *observed mixtures* $\mathbf{x} \in \mathbb{R}^d$ are generated by a smooth and invertible *mixing function* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ belonging to a function class $\mathcal{F}$, which takes as input a vector $\mathbf{s} \in \mathbb{R}^d$ sampled from a distribution with independent components, i.e.,

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \qquad \mathbf{s} \sim p(\mathbf{s}) = \prod_{i=1}^{d} p_i(s_i)\,. \tag{1}$$

[2] For linear ICA, identifiability has also been established for the case of more mixtures than sources (Eriksson and Koivunen, 2004). For nonlinear ICA, Khemakhem et al. (2020) extended existing identifiability results relying on additional supervision in the form of an auxiliary variable (Gresele et al., 2020; Hälvä and Hyvarinen, 2020; Hyvarinen and Morioka, 2016, 2017; Hyvarinen et al., 2019) to the high-dimensional observation setting.

[3] In this work, we mostly assume that observations lie exactly *on* a low-dimensional manifold, not *close to it*.

Can we recover the independent components based only on the observed mixtures? Unfortunately, when $\mathbf{f}$ is nonlinear, the model is non-identifiable without additional constraints (Darmois, 1951; Hyvärinen and Pajunen, 1999)—i.e., the independent components cannot be unambiguously recovered from the observed mixtures. This can be shown through suitable *spurious solutions*, which transform the mixtures $\mathbf{x}$ into independent components which *may themselves be mixtures of the true ones*.

Definition 2.1 (Darmois construction (Darmois, 1951; Hyvärinen and Pajunen, 1999))**.** *The Darmois construction $\mathbf{g}^{\mathbf{D}}:\mathbb{R}^d\to(0,1)^d$ transforms a given distribution to the uniform distribution by recursively applying the conditional Cumulative Distribution Function (CDF) transform,*

$$g_i^D(\mathbf{x}_{1:i})=\int_{-\infty}^{x_i}p(x_i'|\mathbf{x}_{1:i-1})dx_i'\,. \tag{2}$$

Definition 2.2 (Rotated-Gaussian measure preserving Automorphism (MPA) (Khemakhem et al., 2020; Locatello et al., 2019))**.** *The rotated-Gaussian MPA transforms a given density into a Gaussian density by the CDF-transform, applies as orthonormal rotation, and inverts the preceding transformation. Let $\mathbf{F_s}(\mathbf{s})$ and $\mathbf{\Phi}(\mathbf{z})$ denote the CDFs of the latent source distribution and the multivariate Gaussian distribution respectively. For an orthogonal matrix, $\mathbf{R}\in O(d)$, the "rotated-Gaussian" MPA $\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ is defined as,*

$$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})=\mathbf{F}_{\mathbf{s}}^{-1}\circ\mathbf{\Phi}\circ\mathbf{R}\circ\mathbf{\Phi}^{-1}\circ\mathbf{F}_{\mathbf{s}}\,. \tag{3}$$

Towards the goal of achieving identifiability, Independent Mechanism Analysis (IMA) (Gresele et al., 2021) constraints the mixing function class $\mathcal{F}$. IMA postulates that the influence directions of individual latent sources in the mixing process, given by $\frac{\partial\mathbf{f}}{\partial s_i}$, are orthogonal—i.e.,

$$\log|\mathbf{J_f}(\mathbf{s})|=\sum_{i=1}^d\log\left\|\frac{\partial\mathbf{f}}{\partial s_i}(\mathbf{s})\right\|\,. \tag{4}$$

While identifiability of IMA is still an open question, it provably rules out both the Darmois construction and the rotated-Gaussian MPA (Gresele et al., 2021); empirically, IMA allows recovery of the ground truth sources when the data-generating process satisfies the IMA principle (Gresele et al., 2021), as well as under mild model misspecification (Sliwa et al., 2022). Moreover, IMA was proved to be *locally identifiable* (Buchholz et al., 2022).

3 IMA under the Manifold Hypothesis

In the following, we revisit the generative process in (1) as follows: the observed mixtures lie on a d-dimensional Riemannian manifold (a smooth manifold with a d-dimensional tangent space), $\mathbb{X}\subseteq\mathbb{R}^m, m\geq d$, which is embedded in the m-dimensional Euclidean space. The mixing function $\mathbf{f}:\mathbb{R}^d\to\mathbb{X}$ is a *diffeomorphism*[4] from the latent space to the observation manifold: the observations therefore lie on a low-dimensional manifold within the high-dimensional ambient space, in line with the *manifold hypothesis* (Becker and Hinton, 1992; Bengio et al., 2013; Vincent and Bengio, 2002). To study IMA in this setting, the main definitions in (Gresele et al., 2021) need to be adapted, since they were originally tailored to the setting where latent and observation spaces have the same dimension.

We start by extending the IMA principle (Gresele et al., 2021, Principle 4.1). We say that when the manifold hypothesis holds, the IMA principle implies the following equality:

$$\sum_{i=1}^d\log\left\|\frac{\partial\mathbf{f}}{\partial s_i}(\mathbf{s})\right\|=\frac{1}{2}\log\left|\mathbf{J}_{\mathbf{f}}^\top(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})\right|\quad\forall\,\mathbf{s}\in\mathbb{R}^d, \tag{5}$$

where $\mathbf{J_f}(\mathbf{s})$ is the Jacobian of $\mathbf{f}$ at $\mathbf{s}$. Equation (5) therefore states that the area element on the Riemannian manifold $\mathbb{X}$ at $\mathbf{x}=\mathbf{f}(\mathbf{s})$, given by $\sqrt{\left|\mathbf{J}_{\mathbf{f}}^\top(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})\right|}$, equals the product of the norms of the *influences* $\|\frac{\partial\mathbf{f}}{\partial s_i}\|$ that span that element. It is therefore an orthogonality condition, similar to the one expressed in eq. (4) for $m=d$: note however that eq. (5) is meaningful for any $m\geq d$.

Based on (5), we redefine the *local IMA contrast* (Gresele et al., 2021, Def. 4.2), which quantifies the violation of the IMA principle at a given point $\mathbf{x}=\mathbf{f}(\mathbf{s})$, and state two of its useful properties.

[4] A diffeomorphism is an invertible function between two manifolds such that both the function and its inverse are continuously differentiable.

Definition 3.1 (Local IMA contrast)**.** *The local IMA contrast,* $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$*, of* $\mathbf{f}$ *at point* $\mathbf{s}$*, is defined as*

$$c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) = \sum_{i=1}^{d} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| - \frac{1}{2} \log \left| \mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s}) \mathbf{J}_{\mathbf{f}}(\mathbf{s}) \right| . \tag{6}$$

Proposition 3.2. *The local IMA contrast satisfies the following properties:*

(i) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$ *with equality iff. all columns* $\frac{\partial \mathbf{f}}{\partial u_i}(\mathbf{s})$ *of* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *are orthogonal.*

(ii) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ *is invariant to left multiplication of* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *by an orthogonal matrix and to right multiplication by permutation and diagonal matrices.*

Property *(i)* is a geometric condition: given the vectors that span a parallelepiped, the largest volume is obtained when the spanning vectors are orthogonal. Property *(ii)* states that permutation and rescaling of the latent factors, or any orthonormal basis transformation applied to the columns of the Jacobian $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$, do not affect the value of c_{IMA}. Next, we redefine the *global IMA contrast* (Gresele et al., 2021, Def. 4.5).

Definition 3.3 (Global IMA contrast)**.** *The global IMA contrast,* C_{IMA}*, is defined as the expected value of the local IMA contrast with respect to the source distribution,* $p_{\mathbf{s}}$*.*

$$C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = \mathbb{E}_{\mathbf{s} \sim p_{\mathbf{s}}}[c_{\text{IMA}}(\mathbf{f}, \mathbf{s})] = \int c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) p_{\mathbf{s}}(\mathbf{s}) d\mathbf{s}. \tag{7}$$

Proposition 3.4. *The global IMA contrast (*$C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ *satisfies:*

1. $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) \geq 0$ *with equality iff.* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *has orthogonal columns almost surely wrt* $\mathbf{p_s}$*.*
2. $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = C_{\text{IMA}}(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ *for any* $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ *and* $\tilde{\mathbf{s}} = \mathbf{Ph}(\mathbf{s})$ *where* $\mathbf{P} \in \mathbb{R}^{d \times d}$ *is a permutation matrix and* $\mathbf{h}(\mathbf{s}) = (h_1(s_1), h_2(s_2), ..., h_d(s_d))$ *is an invertible element-wise function.*

Property *(i)* states that C_{IMA} can be used to verify whether the IMA condition holds almost surely with respect to the latent distribution, $p_{\mathbf{s}}$. Property *(ii)* states that the IMA constrast for high-dimensional observations is blind to permutation and element-wise transformation of the sources.

4 Ruling out "spurious solutions" under the Manifold Hypothesis

Measure preserving automorphisms (MPAs) applied in the latent space can be used to construct spurious solutions, by composition with the true mixing function (Xi and Bloem-Reddy, 2023). Below, we show that the global IMA contrast defined above rules out spurious solutions based on different MPAs. All proofs of the results in this section can be found in App. A.

Gaussian-rotated MPA, Defn. 2.2. We prove that IMA rules out the MPA in Defn. 2.2 for the case in which the mixing function is a *conformal*[5] (angle-preserving) map—a special case of the IMA function class which, for the $m = d$ case, was proved to be identifiable in (Buchholz et al., 2022).

Theorem 4.1. *Consider* $(\mathbf{f}, p_{\mathbf{s}})$ *such that* $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = 0$*, and moreover* $\mathbf{f} : \mathbb{R}^d \to \mathbb{X}$ *is a conformal map. Given* $\mathbf{R} \in \mathbf{O}(n)$*, assume additionally that* $\exists$ *at least one non-Gaussian* s_i *whose associated canonical basis vector* $\mathbf{e}_i$ *is not transformed by* $\mathbf{R}^{-1} = \mathbf{R}^{\top}$ *into another canonical basis vector* $\mathbf{e}_j$*. Then,* $C_{\text{IMA}}(\mathbf{f} \circ a^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) > 0$*.*

Measure preserving automorphism based on the Darmois construction The Darmois construction (Defn. 2.1) does not directly yield a spurious solution when the dimension of the observed space does not match one of the latent space.[6] Instead, we construct a spurious solution by applying the Darmois construction to an orthonormal rotation of the latent distribution. We show that the IMA contrast defined in our work can distinguish between such a counterexample and the ground truth (up to tolerable ambiguities like permutation and element-wise transformations).

[5] For a formal definition of a *conformal* map, refer to App. A

[6] This is because the CDF transform cannot define a map between a distribution on a higher-dimensional ambient space (observations) to the uniform distribution on a lower-dimensional space (latent sources).

Theorem 4.2. *Let $(\mathbf{f}, p_\mathbf{s}) \in \mathcal{M}_{\text{IMA}}$ where $\mathbf{f} : \mathbb{R}^d \to \mathbb{X}$ is a bijective map and the sources $\mathbf{s}$ are such that at most one factor s_i is Gaussian. Further, we assume that $\mathbf{f}$ is a conformal map. Consider an orthonormal transformation $\mathbf{O} \in \mathbb{R}^{d\times d}$ applied on $\mathbf{s}$, $\tilde{\mathbf{x}} = \mathbf{O}\mathbf{s} \in \mathbb{R}^d$. We further consider that the orthonormal transfomation given by $\mathbf{O}$ is not trivial, i.e. it does not correspond to a permutation or an element-wise scaling. The observations $\mathbf{x} \in \mathbb{X}$ and the transformed variables $\tilde{\mathbf{x}}$ have a bijective relationship. Then any Darmois solution $(\tilde{\mathbf{f}}^D, p_\mathbf{u})$ based on applying $\mathbf{g}^D$ to $\tilde{\mathbf{x}}$ satisfies $C_{\text{IMA}}(\tilde{\mathbf{f}}^D, p_\mathbf{u}) > 0$. Here, $\tilde{\mathbf{f}}^D = \mathbf{f} \circ \mathbf{O}^\top \circ {\mathbf{g}^D}^{-1}$. Thus a solution satisfying $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$ can be distinguished from $(\tilde{\mathbf{f}}^D, p_\mathbf{u})$ based on the contrast C_{IMA}.*

Thm. 4.1 and Thm. 4.2 therefore suggest that IMA may be beneficial for identifiable representation learning even when the manifold hypothesis holds, extending previous findings for $m = d$.

5 Genericity of IMA under the Manifold Hypothesis

In this section, we provide a formal interpretation and justification of IMA as the consequence of a *genericity* assumption—i.e., that the IMA principle is typically satisfied when "Nature [chooses] the direction of the influence of each source component in the observation space independently" (Gresele et al., 2021). We do so by defining a process to sample mixing functions from a lower-dimensional source space to a higher-dimensional observation space, and show that the IMA principle in equation (5) is typically approximately satisfied if the influences are sampled independently from an isotropic prior. While this may not be the only way to sample mixing functions which are typically close to the IMA function class, our proposed construction provides the first rigorous statistical argument that justifies the non-statistical notion of independence expressed in (5).

We construct mixing functions $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ with $m \gg d$ such that, locally, the Jacobian of $\mathbf{f}$ has columns $\mathbf{J}_{\mathbf{f},i}(\mathbf{s}) := \mathbf{J}_\mathbf{f}(\mathbf{s})[:, i]$ that are sampled independently and isotropically, i.e., from a spherically invariant distribution $p_\mathbf{r}$ over $\mathbb{R}^m$:

$$\mathbf{J}_{\mathbf{f},1}(\mathbf{s}), \mathbf{J}_{\mathbf{f},2}(\mathbf{s}), \ldots, \mathbf{J}_{\mathbf{f},d}(\mathbf{s}) \overset{\text{i.i.d}}{\sim} p_\mathbf{r} \,.$$

The i-th column of the Jacobian, $\mathbf{J}_{\mathbf{f},i}(\mathbf{s})$, represents the influence of the i-th source on the observations: this sampling procedure formalizes the intuition that every source influences the mixtures independently. We then proceed to show that typical samples from this process satisfy the IMA principle with high probability. Note that orthogonality of the Jacobian columns is not enforced from the outset: rather, it emerges as a property typically (approximately) satisfied by samples from this process.

In § 5.1, we prove an upper bound on the global IMA contrast $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$, satisfied with high probability by linear maps. Next, in § 5.2, we show how to generate nonlinear maps with locally independent influences, and prove a bound for $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$ of maps sampled from this procedure. For detailed proofs and additional technical details on the results in this section, see App. B.

5.1 Bound on the local IMA contrast

Theorem 5.1. *Consider linear maps, $\mathbf{f}(\mathbf{s}) = \mathbf{J}\mathbf{s}$, where the columns of $\mathbf{J} \in \mathbb{R}^{m\times d}$ are sampled from a spherically symmetric distribution $p_\mathbf{r}$ over $\mathbb{R}^m$; $\mathbf{J}_1, \mathbf{J}_2, ..., \mathbf{J}_d \overset{i.i.d}{\sim} p_\mathbf{r}$. For such maps, the IMA contrast satisfies for $m \gg d$ and $\delta > 0$:*

$$\Pr\left[C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s}) \leq \delta\right] \geq 1 - \min\left\{1, \exp\left(2\log d - \kappa(m-1)\frac{\delta^2}{d^2}\right)\right\} \,.$$

This theorem is based on a concentration result for isotropic priors given by *Levy's lemma* (Janzing et al., 2010, Lemma 1). Levy's lemma shows that a smooth function of an isotropically sampled direction concentrated around its mean with probability growing exponentially in the dimension of the sample space. We observe that in our sampling process, each sampled *influence direction* is orthogonal to the other sampled directions in expectation, i.e. the pairwise inner products of the sampled directions is equal to zero in expectation. We employ Levy's lemma to derive a concentration result on the inner products to obtain the result in Thm. 5.1.

5.2 Bound on the global IMA contrast

We now describe a sampling procedure, and derive bounds for $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$, for non-linear maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$. We retain the principle that locally the columns of the Jacobian of $\mathbf{f}$ are sampled from a spherically invariant distribution. Our procedure therefore samples piece-wise linear functions. We restrict the latent domain to be bounded, in particular the d-dimensional unit cube, $[0,1]^d$, and consider a grid-like partition on the same. On each grid unit, we sample a linear function by the previously described sampling process in § 5.1.

5.2.1 Defining non-linear functions as composition of affine functions

Definition 5.2. *On the source domain* $\mathbf{s} \in [0,1]^d$*, define an axis-aligned square grid partition, with width* $\delta \in \mathbb{R}$*; the number of grid parts along a dimension,* $k \in [d]$ *is therefore equal to* $p = \lceil \frac{1}{\delta} \rceil + 1$.[7] *Consider matrices* $\mathbf{J}^{(1)}, \mathbf{J}^{(2)}, ..., \mathbf{J}^{(p)} \in \mathbb{R}^{m \times d}$ *with columns sampled from a spherically symmetric distribution,* $p_{\mathbf{r}}$*;* $\mathbf{J}_1^{(i)}, \mathbf{J}_2^{(i)}, ..., \mathbf{J}_d^{(i)} \overset{i.i.d}{\sim} p_{\mathbf{r}} \forall i \in [p]$*. The sampled function,* $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$*, is specified as as a sum of* coordinate-wise functions $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d \mathbf{f}_k(s_k)$*, where*

$$\mathbf{f}_k(s_k) := \sum_{t=1}^{p} (\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta) 1_{s_k \in ((t-1)\delta, t\delta]} \, . \tag{8}$$

Observe that the Jacobian of the sampled function, $\mathbf{J_f}(\mathbf{s})$, has independent and isotropic Jacobian columns almost everywhere by construction; therefore we expect that the local IMA contrast, $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ is small almost everywhere. To derive a bound on the global IMA contrast, $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$, we require the Jacobian, $\mathbf{J_f}(\mathbf{s})$, to be defined everywhere—i.e., $\mathbf{f}$ should be continuous, injective and smooth. We briefly explain that the resulting $\mathbf{f}$ is indeed continuous and injective, see App. B for details. We then consider a smooth approximation of the sampled function.

Continuity of sampled functions. Note that coordinate-wise functions, $f_k : [0,1] \to \mathbb{R}^m, k \in [d]; \mathbf{f}(\mathbf{s}) = \sum_{k=1}^d \mathbf{f}_k(s_k)$ are piece-wise linear and continuous by definition. The sampled function $\mathbf{f}$ is therefore continuous as it is a sum of continuous functions.

Injectivity of sampled functions. We show that the subspaces spanned by the images of coordinate-wise functions $f_i : [0,1] \to \mathbb{R}^m$ of $\mathbf{f}$ are linearly independent, said to be in *direct sum*, and that the coordinate-wise functions are injective. We then show that the injectivity of coordinate-wise functions that are in direct sum entails the injectivity of $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d \mathbf{f}_k(s_k)$.

Smooth approximation of sampled functions. We discuss a smooth approximation to the sampled functions from Defn. 5.2 by means of a sinusoidal approximation to the step function.

Definition 5.3 (Smooth step function)**.** *We define the smooth step function as* $\tilde{1}_\epsilon : \mathbb{R} \to \mathbb{R}$ *as*

$$\tilde{1}_\epsilon(s) = \begin{cases} 0\,, & s \leq -\epsilon\,, \\ \frac{1}{2}\sin\left(\frac{\pi s}{2\epsilon}\right) + \frac{1}{2}\,, & -\epsilon < s \leq \epsilon\,, \\ 1\,, & s > \epsilon\,. \end{cases}$$

Definition 5.4 (Smooth approximation to grid-wise linear functions)**.** *We define the smooth approximation of* $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ *as* $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^d \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$ *for* $0 < \epsilon \ll \delta$ *where*

$$\tilde{\mathbf{f}}_{\epsilon,k}(s_k) := \sum_{t=1}^{p} \left(\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta \right) .(\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta))$$

We show that $\tilde{\mathbf{f}}_\epsilon(\mathbf{s})$ obtained in Defn. 5.4 is continuous and injective. For a detailed exposition on this section, refer to App. B.

Theorem 5.5 (Properties of smoothened functions)**.** *Functions* $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$ *defined in Defn. 5.4 are continuously differentiable in* $\mathbb{R}^d$*, in addition to being continuous and injective, are continuously differentiable for* $0 < \epsilon \ll \delta$ *arbitrarily small.*

[7] $\lceil . \rceil$ is the ceiling function.

We can now prove the following bound on the global IMA contrast $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ for the class of nonlinear functions specified in Defn. 5.4.

Theorem 5.6. *Consider the map* $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$ *sampled randomly from the procedure 5.4. Then the map* $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ *for* $0 < \epsilon \ll \delta$ *and any finite probability density* $p_{\mathbf{s}}$ *defined over* $[0,1]^d$ *satisfies the following bound on the global IMA contrast* $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}})$ *for* $m \gg d$ *and* $\delta > 0$*:*

$$\Pr\left[C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}}) \leq \delta\right] \geq 1 - \min\left\{1, \exp\left(2\log d - \kappa(m-1)\frac{\delta^2}{d^2}\right)\right\}$$

Thm. 5.6 shows that for smooth piecewise-linear functions $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$ sampled according to Defn. 5.4, the IMA principle (eq. (5)) is *typically approximately satisfied*: i.e., the probability of the columns of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}(\mathbf{s})$ being close to orthogonal increases as the dimension of the observation space grows. We achieve this result by using the previously derived bound on the IMA contrast for linear functions (Thm. 5.1), and applying it locally on the interior of the grid. We then show that the local IMA contrast, $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$, is finite on the grid boundary and can be neglected since the volume of the boundary is small.

Thm. 5.6 thus enables us to view the IMA principle as the consequence of a genericity assumption for the sampling process in Defn. 5.4. This is because the functions sampled in accordance with Defn. 5.4 do not satisfy the IMA principle by construction: instead, it is the typical draws from the sampling procedure that approximately satisfy the principle when the observation space is high-dimensional.

6 Conclusion

We extended IMA theory under the *manifold hypothesis*, revisiting the definitions from previous works, and show that IMA provably circumvents non-identifiability issues even in this setting. Our results pave the way for an application of IMA to realistic representation learning involving dimensionality reduction. We also showed that the IMA principle can be seen as the consequence of a *genericity* assumption when the manifold hypothesis holds: this clarifies the interpretation of IMA. In particular, when the latent data-generating factors influence the observed mixture *independently*, the orthogonality condition given by IMA *typically* holds.

Acknowledgments

We thank Georgios Arvanitis, Dominik Janzing, and Joanna Sliwa for interesting discussions.

This work was supported by the Tübingen AI Center and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC number 2064/1, Project number 390727645. L.G. was supported by the VideoPredict project, FKZ: 01IS21088.

References

S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992. [Cited on pages 2 and 3.]

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. [Cited on pages 2 and 3.]

M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 557–565. PMLR, 2018. [Cited on page 2.]

R. P. Brent, J.-A. H. Osborn, and W. D. Smith. Bounds on determinants of perturbed diagonal matrices. *arXiv preprint arXiv:1401.7084*, 2014. [Cited on page 17.]

S. Buchholz, M. Besserve, and B. Schölkopf. Function classes for identifiable nonlinear independent component analysis. *Advances in Neural Information Processing Systems*, 35:16946–16961, 2022. [Cited on pages 2, 3, and 4.]

L. Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005. [Cited on page 2.]

E. Çinlar. *Probability and stochastics*, volume 261. Springer, 2011. [Cited on pages 28, 36, and 43.]

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. [Cited on page 2.]

E. Cunningham, A. Cobb, and S. Jha. Principal manifold flows. *arXiv preprint arXiv:2202.07037*, 2022. [Cited on page 2.]

G. Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951. [Cited on pages 1 and 3.]

J. Eriksson and V. Koivunen. Identifiability, separability, and uniqueness of linear ica models. *IEEE signal processing letters*, 11(7):601–604, 2004. [Cited on page 2.]

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. [Cited on page 2.]

W. T. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368 (6471):542–545, 1994. [Cited on page 2.]

L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR, 2020. [Cited on pages 1 and 2.]

L. Gresele, J. von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, Dec. 2021. [Cited on pages 1, 2, 3, 4, 5, 10, and 11.]

H. Hälvä and A. Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020. [Cited on page 2.]

A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016. [Cited on page 2.]

A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017. [Cited on page 2.]

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. [Cited on pages 1 and 3.]

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Ltd, 2001. [Cited on page 2.]

A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019. [Cited on page 2.]

A. Hyvärinen, I. Khemakhem, and R. Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *arXiv preprint arXiv:2302.02672*, 2023. [Cited on page 2.]

A. Hyvärinen and H. Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA, 2016. [Cited on page 1.]

Y. Ishii. On conharmonic transformations. *Tensor, NS*, 11:73–80, 1957. [Cited on page 12.]

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. [Cited on pages 1 and 16.]

D. Janzing, P. O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *International Conference on Machine Learning*, 2010. [Cited on pages 2, 5, and 16.]

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. [Cited on pages 1, 2, and 3.]

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014. [Cited on page 2.]

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. [Cited on pages 1 and 3.]

W. lodzimierz Bryc. Normal distribution characterizations with applications. *Lecture Notes in Statistics*, 100, 1995. [Cited on page 17.]

J. Milnor and D. W. Weaver. *Topology from the differentiable viewpoint*, volume 21. Princeton university press, 1997. [Cited on page 10.]

G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2924–2932, 2014. [Cited on page 2.]

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. [Cited on page 1.]

P. Reizinger, L. Gresele, J. Brady, J. Von Kügelgen, D. Zietlow, B. Schölkopf, G. Martius, W. Brendel, and M. Besserve. Embrace the gap: Vaes perform independent mechanism analysis. *Advances in Neural Information Processing Systems*, 35:12040–12057, 2022. [Cited on page 2.]

S. Roman, S. Axler, and F. Gehring. *Advanced linear algebra*, volume 3. Springer, 2005. [Cited on page 22.]

W. Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964. [Cited on pages 20, 26, 27, 30, 33, 34, 35, 38, 41, and 43.]

J. Sliwa, S. Ghosh, V. Stimper, L. Gresele, and B. Schölkopf. Probing the robustness of independent mechanism analysis for representation learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. [Cited on page 3.]

S. Stepanov and I. Tsyganok. Theorems on conformal mappings of complete riemannian manifolds and their applications. *Balkan Journal of Geometry and Its Applications*, 22(1):81–86, 2017. [Cited on page 12.]

P. Vincent and Y. Bengio. Manifold parzen windows. *Advances in neural information processing systems*, 15, 2002. [Cited on pages 2 and 3.]

Wikipedia contributors. Rank–nullity theorem — Wikipedia, the free encyclopedia, 2022. [Online; accessed 2-June-2022]. [Cited on page 20.]

Q. Xi and B. Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, pages 6912–6939. PMLR, 2023. [Cited on page 4.]

APPENDIX

Overview

- Appendix A provides proofs of technical results in Section 3.
- Appendix B provides a detailed exposition on Section 5.

A Proof of technical results in Section 3

A.1 Preliminaries

We provide some preliminary definitions that we refer to in the main text and the remainder of the Appendix.

Definition A.1 (Diffeomorphism, Chapter 1 (Milnor and Weaver, 1997))**.** *A diffeomorphism is an invertible function,* $\mathbf{f}$*, which maps one differentiable manifold onto another such that both the function and its inverse are smooth.*

Definition A.2 (Spherically symmetric distribution)**.** *A distribution* $p_{\mathbf{r}}$ *on the* m*-dimensional Lebesgue measure is said to be spherically symmetric if* $\forall \mathbf{x} \in \mathbb{R}^m, \mathbf{x} \overset{p_{\mathbf{r}}}{\sim} \mathbf{O}\mathbf{x}$*, where* $\mathbf{O} \in \mathbb{R}^{m \times m}$ *is an orthonormal matrix.*

A.2 Properties of the local IMA contrast under the Manifold Hypothesis

Proposition 3.2. *The local IMA contrast satisfies the following properties:*

(i) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$ *with equality iff. all columns* $\frac{\partial \mathbf{f}}{\partial u_i}(\mathbf{s})$ *of* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *are orthogonal.*

(ii) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ *is invariant to left multiplication of* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *by an orthogonal matrix and to right multiplication by permutation and diagonal matrices.*

Proof.

(i) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) = \frac{1}{2} D^{l}_{KL}(\mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})) \geq 0$ with equality iff. $\mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ is a diagonal matrix. $\mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ is diagonal iff. the columns of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$, $\frac{\partial \mathbf{f}}{\partial u_i}(\mathbf{s})$ are orthogonal.

Thus, we have shown that $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$ with equality iff all columns $\frac{\partial \mathbf{f}}{\partial u_i}(\mathbf{s})$ of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ are orthogonal.
Remark: To show that $D^{l}_{KL}(\mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})) \geq 0$, Hadamard's determinant inequality is used in a different form than in (Gresele et al., 2021). In particular, for positive definite matrices, in our case $\mathbf{W} = \mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})$, the determinant is upper-bounded by the product of its diagonal entries.

$$|\det(\mathbf{W})| \leq \prod_{i=1}^{d} w_{ii} \tag{9}$$

with equality iff. $\mathbf{W}$ is diagonal. This is obtained by considering the Cholesky decomposition of $\mathbf{W} = \mathbf{N}\mathbf{N}^{\top}$ which uniquely exists for any real positive definite matrix, where $\mathbf{N} \in \mathbb{R}^{d \times d}$

$$|\det(\mathbf{W})| = |\det(\mathbf{N})|^2 \leq \prod_{i=1}^{d} \|\mathbf{n}_i\|^2 = \prod_{i=1}^{d} w_{ii}$$

where $\mathbf{n}_i$ are the columns of $\mathbf{N}$. Equality holds iff. the columns of $\mathbf{N}$ are orthogonal i.e. $\mathbf{W}$ is diagonal.

(ii) *Invariance to left multiplication by orthogonal matrix*
$\mathbf{W} = \mathbf{J}_{\mathbf{f}}(\mathbf{s}) \in \mathbb{R}^{m\times d}$, $\mathbf{O} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. $\tilde{\mathbf{W}} = \mathbf{O}\mathbf{W}$.

$$\begin{aligned}\frac{1}{2}D^{l}_{KL}(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}) &= \frac{1}{2}D^{l}_{KL}(\mathbf{W}^{\top}\mathbf{O}^{\top}\mathbf{O}\mathbf{W})\\ &= \frac{1}{2}D^{l}_{KL}(\mathbf{W}^{\top}\mathbf{I}_{m}\mathbf{W})\\ &= \frac{1}{2}D^{l}_{KL}(\mathbf{W}^{\top}\mathbf{W})\end{aligned}$$

This corresponds to a change of basis in the observation space.

Invariance to right multiplication by a permutation matrix
Let $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{P}$ where $\mathbf{P} \in \mathbb{R}^{d\times d}$ is a permutation matrix. Then $\tilde{\mathbf{W}}$ is just $\mathbf{W}$ with permuted columns. Clearly, the sum of log-column-norms does not change with the order of the summands. Further, $\log|\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}| = \log|\mathbf{P}^{\top}\mathbf{W}^{\top}\mathbf{W}\mathbf{P}| = \log|\mathbf{P}^{\top}| + \log|\mathbf{W}^{\top}\mathbf{W}| + \log|\mathbf{P}| = \log|\mathbf{W}^{\top}\mathbf{W}|$ because the absolute value of the determinant of the permutation matrix is one.

Invariance to right multiplication by a diagonal matrix
Let $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{D}$ where $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix.
For the first term, we know that the columns of $\tilde{\mathbf{W}}$ are scaled versions of the columns of $\mathbf{W}$, i.e. $\tilde{\mathbf{w}}_i = \mathbf{w}_i$, $\|\tilde{\mathbf{w}}_i\| = |d_i|\|\mathbf{w}_i\|$. For the second term, we use the decomposition of the determinant:

$$\begin{aligned}\log|\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}| &= \log|\mathbf{D}^{\top}\mathbf{W}^{\top}\mathbf{W}\mathbf{D}|\\ &= 2\log|\mathbf{D}| + \log|\mathbf{W}^{\top}\mathbf{W}|\\ &= \log|\mathbf{W}^{\top}\mathbf{W}| + 2\sum_{i=1}^{d}\log|d_i|\end{aligned}$$

Taken together, we obtain:

$$\begin{aligned}\sum_{i=1}^{d}\log\|\tilde{\mathbf{w}}_i\| - \frac{1}{2}\log|\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}| &= \sum_{i=1}^{d}\log|d_i|\|\mathbf{w}_i\| - \frac{1}{2}\left(\log|\mathbf{W}^{\top}\mathbf{W}| + 2\sum_{i=1}^{d}\log|d_i|\right)\\ &= \sum_{i=1}^{d}\log\|\mathbf{w}_i\| + \sum_{i=1}^{d}\log|d_i| - \frac{1}{2}\log|\mathbf{W}^{\top}\mathbf{W}|\\ &\quad - \sum_{i=1}^{d}\log|d_i|\\ &= \sum_{i=1}^{d}\log\|\mathbf{w}_i\| - \frac{1}{2}\log|\mathbf{W}^{\top}\mathbf{W}|\end{aligned}$$

□

A.3 Properties of the global IMA contrast under the Manifold Hypothesis

Proposition 3.4. *The global IMA contrast ($C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ satisfies:*

1. *$C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) \geq 0$ with equality iff. $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ has orthogonal columns almost surely wrt $\mathbf{p}_{\mathbf{s}}$.*
2. *$C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = C_{\text{IMA}}(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ for any $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ and $\tilde{\mathbf{s}} = \mathbf{P}\mathbf{h}(\mathbf{s})$ where $\mathbf{P} \in \mathbb{R}^{d\times d}$ is a permutation matrix and $\mathbf{h}(\mathbf{s}) = (h_1(s_1), h_2(s_2), ..., h_d(s_d))$ is an invertible element-wise function.*

In (Gresele et al., 2021), for property (i) $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ can be expressed as $\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ where $\mathbf{O}(\mathbf{s}), \mathbf{D}(\mathbf{s})$ are orthogonal and diagonal matrices respectively in the condition for equality. This is no longer possible in the case for high dimensional observations because the Jacobian $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ is no longer a square matrix.

Proof.

(i) From property (i) of Proposition 4.4, we know that $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$. Hence, $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) \geq 0$ follows as a direct consequence of integrating the non-negative quantity $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$.

Equality is attained iff. $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) = 0$ almost surely wrt $p_{\mathbf{s}}$ which holds when $\mathbf{J_f}(\mathbf{s})$ has orthogonal columns almost surely wrt $p_{\mathbf{s}}$.

(ii) $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ and $\tilde{\mathbf{s}} = \mathbf{Ph}(\mathbf{s})$ where $\mathbf{P} \in \mathbb{R}^{d \times d}$ is a permutation matrix and $\mathbf{h}(\mathbf{s}) = (h_1(s_1), h_2(s_2), ..., h_d(s_d))$ is an invertible element-wise function. Consider

$$C_{\text{IMA}}(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) = \int c_{\text{IMA}}(\tilde{\mathbf{f}}, \tilde{\mathbf{s}}) p_{\tilde{\mathbf{s}}}(\tilde{\mathbf{s}}) d\tilde{\mathbf{s}} = \int c_{\text{IMA}}(\tilde{\mathbf{f}}, \tilde{\mathbf{s}}) p_{\mathbf{s}}(\mathbf{s}) d\mathbf{s} \tag{10}$$

where for the second equality we have used $p_{\tilde{\mathbf{s}}}(\tilde{\mathbf{s}}) d\tilde{\mathbf{s}} = p_{\mathbf{s}}(\mathbf{s}) d\mathbf{s}$ since $P \circ h$ is an invertible transformation. It thus suffices to show that

$$c_{\text{IMA}}(\tilde{\mathbf{f}}, \tilde{\mathbf{s}}) = c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \tag{11}$$

at any point $\tilde{\mathbf{s}} = \mathbf{P} \circ \mathbf{h}(\mathbf{s})$. To show this we write

$$\begin{aligned}
\mathbf{J}_{\tilde{\mathbf{f}}}(\tilde{\mathbf{s}}) &= \mathbf{J}_{\mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}}(\mathbf{Ph}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f} \circ \mathbf{h}^{-1}}(\mathbf{P}^{-1} \circ \mathbf{Ph}(\mathbf{s})) \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{Ph}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f} \circ \mathbf{h}^{-1}}(\mathbf{h}(\mathbf{s})) \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{Ph}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{h}^{-1} \circ \mathbf{h}(\mathbf{s})) \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{s})) \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{Ph}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{s}) \mathbf{D}(\mathbf{s}) \mathbf{P}^{-1}
\end{aligned}$$

where we have used the chain rule for differentiability, $\mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{s}))$ is a diagonal matrix $\mathbf{D}(\mathbf{s})$ and $\mathbf{J}_{\mathbf{P}^{-1}} = \mathbf{P}^{-1}$ for any $\mathbf{s}$. Note that $\mathbf{P}^{-1}$ is also a permutation matrix.

The equality in (11) follows from applying (ii) from Proposition 4.4. Substituting (11) into (10), we finally obtain

$$C_{\text{IMA}}(\tilde{\mathbf{f}}, \mathbf{p}_{\tilde{\mathbf{s}}}) = C_{\text{IMA}}(\mathbf{f}, \mathbf{p}_{\mathbf{s}})$$

□

A.4 Ruling out “spurious solutions" under the manifold hypothesis

“Spurious solutions" to nonlinear ICA on an observation manifold are constructed by composing *conformal* maps – a subclass of the IMA function class – with measure preserving automorphisms on the latent space. We define conformal maps between Riemannian manifolds below, and comment on the Jacobian of conformal maps.

Definition A.3 (Conformal map between Riemannian manifolds (Ishii, 1957; Stepanov and Tsyganok, 2017))**.** *A diffeomorphism* $\mathbf{f} : (\mathbb{M}, \mathbf{g}) \to (\bar{\mathbb{M}}, \bar{\mathbf{g}})$ *between two Riemannian manifolds,* $\mathbb{M}$ *and* $\bar{\mathbb{M}}$*, equipped with the Riemannian metric tensors,* $\mathbf{g}$ *and* $\bar{\mathbf{g}}$*, is called conformal if it preserves the angles between any pair curves. In this case, the metric tensors* $\mathbf{g}$ *and* $\bar{\mathbf{g}}$ *are related as* $\bar{\mathbf{g}} = e^{2\sigma}\mathbf{g}$ *for some scalar function* $\sigma \in C^2\mathbb{M}$*.*

We make an observation on conformal maps from the d-dimensional Euclidean space to a d-dimensional Riemannian manifold living in a higher m-dimensional Euclidean space ($m \geq d$). In overview, this observation derives from the definition of conformal maps on manifolds A.3 that the columns of the Jacobian $\mathbf{J_f}(\mathbf{s})$ are equal in norm for all values in the domain of $\mathbf{f}$, here $\mathbf{s} \in \mathbb{R}^d$, which is equivalent to the condition that $\mathbf{J}_{\mathbf{f}}^{\top}(\mathbf{s})\mathbf{J_f}(\mathbf{s})$ is a scalar multiple of the identity matrix.

In our scenario, we consider a map between the Riemannian manifolds, $\mathbb{M} \equiv \mathbb{R}^d$ to $\bar{\mathbb{M}} \equiv \mathcal{X} \subset \mathbb{R}^m$, where $m \gg d$. Note that the d-dimensional Euclidean space is also a Riemannian manifold.

The tangent space of $\mathbb{M} \equiv \mathbb{R}^d$ is set of the canonical basis vectors, $e_1, e_2, ..., e_d$ at all points in $\mathbb{R}^d$. Hence, the metric tensor associated with $\mathbb{M} \equiv \mathbb{R}^d$ is the identity matrix $\mathbf{g} \equiv \mathbb{I}_d$ at all points in $\mathbb{R}^d$.

The metric tensor associated with the Riemannian manifold, $\bar{\mathbb{M}} \equiv \mathcal{X}$ is written as $\bar{\mathbf{g}} \equiv \mathbf{J}_{\mathbf{f}}^\top(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ at the point $\mathbf{f}(\mathbf{s}) \in \mathcal{X}$, where $\mathbf{s} \in \mathbb{R}^d$ upon which the bijective map $\mathbf{f} : \mathbb{R}^d \to \mathcal{X}$ acts.

For the map, $\mathbf{f}$ to be conformal A.3, we require that $\bar{\mathbf{g}} = e^{2\sigma}\mathbf{g}$ for some scalar function $\sigma \in C^2\mathbb{M}$. This is equivalent to the condition that $\mathbf{J}_{\mathbf{f}}^\top(\mathbf{s})\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = t(\mathbf{s})\mathbb{I}_d$, where $t : \mathbb{R}^d \to \mathbb{R}^+$ is a positive scalar function. This is further equivalent to $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ having orthogonal columns with the same norm.

We define $\lambda(\mathbf{s}) = \sqrt{t(\mathbf{s})}$ as the conformal factor (equal to the norm of the column vectors of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$).

Rotated-Gaussian measure preserving automorphism We now present the theorem which shows that the IMA contrast rules out rotated-Gaussian measure preserving automorphism solutions.

Theorem 4.1. *Consider* $(\mathbf{f}, p_{\mathbf{s}})$ *such that* $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = 0$*, and moreover* $\mathbf{f} : \mathbb{R}^d \to \mathbb{X}$ *is a conformal map. Given* $\mathbf{R} \in \mathbf{O}(n)$*, assume additionally that* $\exists$ *at least one non-Gaussian* s_i *whose associated canonical basis vector* $\mathbf{e}_i$ *is not transformed by* $\mathbf{R}^{-1} = \mathbf{R}^\top$ *into another canonical basis vector* $\mathbf{e}_j$*. Then,* $C_{\text{IMA}}(\mathbf{f} \circ a^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) > 0$.

Proof. Recall the definition

$$a^{\mathbf{R}}(p_{\mathbf{s}}) = \mathbf{f}_{\mathbf{s}}^{-1} \circ \Phi \circ \mathbf{R} \circ \Phi^{-1} \circ \mathbf{f}_{\mathbf{s}}$$

For notational convenience, we denote $\sigma = \Phi^{-1} \circ \mathbf{f}_{\mathbf{s}}$ and write

$$a^{\mathbf{R}}(p_{\mathbf{s}}) = \sigma^{-1} \circ \mathbf{R} \circ \sigma$$

Note that since both $\mathbf{f}_{\mathbf{s}}$ and Φ are element-wise transformations so is σ.

First by using property *(ii)* of Prop. 3.4 (invariance of C_{IMA} to element-wise transformations), we obtain

$$C_{\text{IMA}}(\mathbf{f} \circ a^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) = C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ \mathbf{R} \circ \sigma, p_{\mathbf{s}}) = C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ \mathbf{R}, p_{\mathbf{z}})$$

with $\mathbf{z} = \sigma(\mathbf{s})$ such that $p_{\mathbf{z}}$ is an isotropic Gaussian distribution.

Suppose *for a contradiction* that $C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ \mathbf{R}, p_{\mathbf{z}}) = 0$.

According to property *(i)* of Prop. 3.4, this entails that the matrix

$$\mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z}) = \mathbf{R}^\top J_{\sigma^{-1}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))\mathbf{J}_{\sigma^{-1}}(\mathbf{z})\mathbf{R} \tag{12}$$

is diagonal almost surely w.r.t $p_{\mathbf{z}}$. Moreover, smoothness of $p_{\mathbf{s}}$ and $\mathbf{f}$ implies the matrix expression of 12 is a continuous function of $\mathbf{z}$. Thus $\mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z})$ needs to be diagonal for all $\mathbf{z} \in \mathbb{R}^d$.

Since $\mathbf{f}$ is a conformal map,

$$\mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))$$

is diagonal. Moreover, since σ is an element-wise transformation, $\mathbf{J}_{\sigma^{-1}}(\mathbf{z})^\top$ and $\mathbf{J}_{\sigma^{-1}}(\mathbf{z})$ are also diagonal. Taken together, this implies that

$$\mathbf{J}_{\sigma^{-1}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z})) J_{\sigma^{-1}}(\mathbf{z}) \tag{13}$$

is diagonal (i.e. 12 is of the form $\mathbf{R}^\top \mathbf{D}(\mathbf{z})\mathbf{R}$ for some diagonal matrix $\mathbf{D}(\mathbf{z})$).

Without loss of generality, we assume the first dimension s_1 of $\mathbf{s}$ is non-Gaussian and satisfies the assumptions relative to $\mathbf{R}$ (axis not invariant nor sent to another canonical axis).

Now, since both the Gaussian CDF Φ and the CDF $\mathbf{f}_{\mathbf{s}}$ are smooth (the latter by the assumption that $p_{\mathbf{s}}$ is a smooth density), σ is a smooth function and thus has continuous partial derivatives.

By continuity of the partial derivative and the non-Gaussianity of s_1, the first diagonal element $\frac{\partial \sigma_1^{-1}}{\partial z_1}$ of $\mathbf{J}_{\sigma^{-1}}$ must be strictly monotonic in a neighborhood of some z_1^0.

On the other hand, our assumptions related to $\mathbf{R}$ entail that there are at least two non-vanishing coefficients in the first row of $\mathbf{R}$. Let us call $i \neq j$ such pairs of coordinates, i.e. $r_{1i} \neq 0$ and $r_{1j} \neq 0$.

Now consider the off-diagonal term (i, j) of $\mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z})$ which we assumed must be 0. Since the term in 13 is diagonal, this off-diagonal term is given by

$$\sum_{k=1}^{n} \left(\frac{\partial \sigma_k^{-1}}{dz_k}(z_k)\right)^2 \left\| \frac{\partial \mathbf{f}}{ds_k} \circ \sigma^{-1}(\mathbf{z}) \right\|^2 r_{ki} r_{kj} = \sum_{k=1}^{n} \left(\frac{\partial \sigma_k^{-1}}{dz_k}(z_k)\right)^2 \lambda(\sigma^{-1}(\mathbf{z}))^2 r_{ki} r_{kj} = 0$$

By definition of conformal map between Riemannian manifolds A.3, the square of the conformal factor is a strictly positive function.

$$\lambda(\sigma^{-1}(\mathbf{z}))^2 > 0 \forall \mathbf{z}$$

Thus, for all $\mathbf{z}$ we must have

$$\sum_{k=1}^{n} \left(\frac{\partial \sigma_k^{-1}}{d z_k}(z_k) \right)^2 r_{ki} r_{kj} = 0 \tag{14}$$

Now consider the first term $\left(\frac{\partial \sigma_1^{-1}}{d z_1}(z_1) \right)^2 r_{1i} r_{1j}$ in the sum.

Recall that $r_{1i} r_{1j} \neq 0$, and that $\frac{\partial \sigma_1^{-1}}{d z_1}(z_1)$ is strictly monotonic on a neighborhood of z_1^0.

As a consequence, $\left(\frac{\partial \sigma_1^{-1}}{d z_1}(z_1) \right)^2 r_{1i} r_{1j}$ is also strictly monotonic with respect to z_1 on a neighborhood of z_1^0 (where the other variables $(z_2, z_3, ..., z_n)$ are left constant), while the other terms in 14 are left constant because σ is an element-wise transformation.

This leads to a contradiction as 14 (which should be satisfied for all $\mathbf{z}$) cannot constantly stay zero as z_1 varies within the neighborhood of z_1^0.

Hence, our assumption that $C_{\text{IMA}}(f \circ \sigma^{-1} \circ \mathbf{R}, p_{\mathbf{z}}) = 0$ cannot hold.

We conclude that $C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ R, p_{\mathbf{z}}) > 0$.

□

Measure preserving automorphism based on the Darmois construction

Following are helper lemmata to prove Theorem 4.2, which rules out a counterexample based on the Darmois construction.

Lemma A.4. *Jacobian of the Darmois construction, $\mathbf{g}^D(\mathbf{x})$, in Definition 2.1 is lower triangular.*

Proof. On applying the recursive Darmois construction, we obtain latent variables $\mathbf{z} = \mathbf{g}^D(\mathbf{x}) \sim$ Unif$([0,1]^d)$. The Darmois construction is invertible since the (conditional) cumulative distribution function is injective. Consider the inverse of the Darmois construction, $\mathbf{f}^D$ such that $\mathbf{X} = \mathbf{f}^D(\mathbf{z})$. We observe from 2.1 that x_1 is related to all the coordinates of $\mathbf{z} = (z_1, z_2, ..., z_d)$, z_2 is related to $\mathbf{Z}_{\geq 2} = (z_2, z_3, ..., z_d)$ and so on. Hence, we take note of the observation that the Jacobian of $\mathbf{f}^D$ is lower-triangular. □

Lemma A.5. *Consider a matrix $\tilde{\mathbf{A}} = \mathbf{AO}$, where $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times d}, \mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{O} \in \mathbb{R}^{d \times d}$. $\mathbf{A}$ is a tall matrix which has orthogonal columns with unit norm i.e. $\mathbf{A}^\top \mathbf{A} = \mathbb{I}_d$, and $\mathbf{O}$ is an orthonormal matrix. Then, $\tilde{\mathbf{A}}$ has orthogonal columns with unit norm i.e. $\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} = \mathbb{I}_d$.*

Proof.

$$\begin{aligned} \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} &= (\mathbf{AO})^\top (\mathbf{AO}) \\ &= \mathbf{O}^\top \mathbf{A}^\top \mathbf{AO} \\ &= \mathbf{O}^\top \mathbf{O} = \mathbb{I}_d \end{aligned}$$

□

Lemma A.6. *Consider a matrix, $\tilde{\mathbf{A}} = \mathbf{AT}$, where $\mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{T} \in \mathbb{R}^{d \times d}$. $\mathbf{A}$ has orthogonal columns with unit norm i.e. $\mathbf{A}^\top \mathbf{A} = \mathbb{I}_d$, and $\mathbf{T}$ is a lower-triangular matrix. $\tilde{\mathbf{A}}$ has orthogonal columns iff. $\mathbf{T}$ is diagonal.*

Proof. *$\mathbf{T}$ is diagonal. $\implies$ $\tilde{\mathbf{A}}$ has orthogonal columns.*

$$\begin{aligned} \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} &= (\mathbf{AT})^\top (\mathbf{AT}) \\ &= \mathbf{T}^\top \mathbf{A}^\top \mathbf{AT} \\ &= \mathbf{T}^\top \mathbf{T} = \mathbf{T}^2 \end{aligned}$$

$\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is a diagonal matrix, hence $\tilde{\mathbf{A}}$ has orthogonal columns. $\tilde{\mathbf{A}}$ *has orthogonal columns.* $\implies$ $\mathbf{T}$ *is diagonal.*

We know that $\mathbf{D} = \tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is diagonal, by definition of orthogonality of the columns of $\tilde{\mathbf{A}}$.

$$\begin{aligned} D = \tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} &= \mathbf{T}^{\mathbf{T}}\mathbf{A}^\top\mathbf{A}\mathbf{T} \\ &= \mathbf{T}^\top\mathbf{T} \end{aligned}$$

Consider the determinant of $\mathbf{D}$, $|\mathbf{D}|$, and the determinant of $\mathbf{T}^\top\mathbf{T}$, $|\mathbf{T}^\top\mathbf{T}|$. $|\mathbf{D}| = \prod_{i=1}^d D_{ii} = \prod_{i=1}^d \|\mathbf{T}_i\|^2$. Also, $|\mathbf{T}^\top\mathbf{T}| = |\mathbf{T}|^2 = \prod_{i=1}^d T_{ii}^2$, since the determinant of a triangular matrix is the product of its diagonal elements.

$$\begin{gathered} \mathbf{D} = \mathbf{T}^\top\mathbf{T},\ |\mathbf{D}| = |\mathbf{T}^\top\mathbf{T}| \\ \prod_{i=1}^d \|\mathbf{T}_i\|^2 = \prod_{i=1}^d T_{ii}^2 \\ \implies \mathbf{T} \text{ is diagonal.} \end{gathered}$$

□

Lemma A.7. *A smooth function* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ *whose Jacobian is diagonal everywhere is an element-wise function,* $f(\mathbf{s}) = (f_1(s_1), f_2(s_2), ..., f_d(s_d))$.

Proof. Let $\mathbf{f}$ be a smooth function with a diagonal Jacobian everywhere.

Consider the function $f_i(\mathbf{s})$ for any $i \in 1, 2, ..., d$. Suppose *for a contradiction* that f_i depends on s_j for some $j \neq i$. Then there must be at least one point $\mathbf{s}*$ such that $\frac{\partial f_i}{s_j}(\mathbf{s}*) \neq 0$. However, this contradicts the assumption that $\mathbf{J}_\mathbf{f}$ is diagonal everywhere (since $\frac{\partial f_i}{s_j}$ is an off-diagonal element for $i \neq j$). Hence, f_i can only depend on s_i for all i. i.e. f is an element-wise function. □

Theorem 4.2. *Let* $(\mathbf{f}, p_\mathbf{s}) \in \mathcal{M}_{\text{IMA}}$ *where* $\mathbf{f} : \mathbb{R}^d \to \mathbb{X}$ *is a bijective map and the sources* $\mathbf{s}$ *are such that at most one factor* s_i *is Gaussian. Further, we assume that* $\mathbf{f}$ *is a conformal map. Consider an orthonormal transformation* $\mathbf{O} \in \mathbb{R}^{d\times d}$ *applied on* $\mathbf{s}$*,* $\tilde{\mathbf{x}} = \mathbf{O}\mathbf{s} \in \mathbb{R}^d$*. We further consider that the orthonormal transfomation given by* $\mathbf{O}$ *is not trivial, i.e. it does not correspond to a permutation or an element-wise scaling. The observations* $\mathbf{x} \in \mathbb{X}$ *and the transformed variables* $\tilde{\mathbf{x}}$ *have a bijective relationship. Then any Darmois solution* $(\tilde{\mathbf{f}}^D, p_\mathbf{u})$ *based on applying* $\mathbf{g}^D$ *to* $\tilde{\mathbf{x}}$ *satisfies* $C_{\text{IMA}}(\tilde{\mathbf{f}}^D, p_\mathbf{u}) > 0$*. Here,* $\tilde{\mathbf{f}}^D = \mathbf{f} \circ \mathbf{O}^\top \circ \mathbf{g}^{D^{-1}}$*. Thus a solution satisfying* $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$ *can be distinguished from* $(\tilde{\mathbf{f}}^D, p_\mathbf{u})$ *based on the contrast* C_{IMA}.

Proof. The theorem follows the following bijective maps:

$$\mathbf{x} \in \mathcal{X} \underset{(i)}{\longleftrightarrow} \mathbf{s} \in \mathbb{R}^d \underset{(ii)}{\longleftrightarrow} \tilde{\mathbf{x}} \in \mathbb{R}^d \underset{(iii)}{\longleftrightarrow} \tilde{\mathbf{s}} \in \mathbb{R}^d$$

The bijective maps are described as follows:

(i) $\mathbf{x} = \mathbf{f}(\mathbf{s}), \mathbf{s} = \mathbf{f}^{-1}(\mathbf{x})$

(ii) $\tilde{\mathbf{x}} = \mathbf{O}\mathbf{s}, \mathbf{O} \in \mathbb{R}^{d\times d}$

(iii) $\tilde{\mathbf{s}} = \mathbf{g}^D(\tilde{\mathbf{x}})$ by the Darmois construction 2.1

$\tilde{\mathbf{s}}$ is mixed with respect to $\mathbf{s}$ since $\mathbf{g}^D \neq \mathbf{O}^\top$ as the Jacobians cannot match, $\mathbf{J}_{\mathbf{g}^D} \neq \mathbf{J}_{\mathbf{O}^\top}$ unless $\mathbf{g}^D$ is an element-wise transformation. $\mathbf{J}_{\mathbf{g}^D}$ is a triangular matrix by A.4, and $\mathbf{J}_{\mathbf{O}^\top} = \mathbf{O}^\top$ is an orthonormal matrix.

We want to check if the solution, $(\tilde{\mathbf{f}}^D, p_\mathbf{u})$ satisfies IMA, i.e. $C_{\text{IMA}}(\tilde{\mathbf{f}}^D, p_\mathbf{u}) = 0$. This is satisfied if $J_{\tilde{\mathbf{f}}^D}(\tilde{\mathbf{s}})$ has orthogonal columns almost surely.

$$\begin{aligned}\mathbf{J}_{\tilde{\mathbf{f}}^D}(\tilde{\mathbf{s}}) &= \mathbf{J}_{\mathbf{f}\circ\mathbf{O}\circ\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}}) \\ &= \mathbf{J}_{\mathbf{f}}(\mathbf{O}\circ\mathbf{g}^{D^{-1}}(\tilde{\mathbf{s}}))\mathbf{J}_{\mathbf{O}}(\mathbf{g}^{D^{-1}}(\tilde{\mathbf{s}}))\mathbf{J}_{\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}}) \\ &= \mathbf{J}_{\mathbf{f}}(\mathbf{O}\circ\mathbf{g}^{D^{-1}}(\tilde{\mathbf{s}}))\mathbf{O}\mathbf{J}_{\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}})\end{aligned}$$

Consider $\mathbf{A} = \mathbf{J}_{\mathbf{f}}(\mathbf{O}\circ\mathbf{g}^{D^{-1}}(\tilde{\mathbf{s}})), \mathbf{T} = \mathbf{J}_{\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}})$. Since $\mathbf{f}$ is a conformal map, $\mathbf{A}$ has orthogonal columns with the same norm. Without loss of generality, we consider that the norm of the columns of $\mathbf{A}$ is one. $\mathbf{T}$ is a lower triangular matrix by A.4.

$$\begin{aligned}\mathbf{J}_{\tilde{\mathbf{f}}^D}(\tilde{\mathbf{s}}) &= \mathbf{AOT} \\ &= \tilde{\mathbf{A}}\mathbf{T} && \tilde{\mathbf{A}} \text{ has orthogonal columns with unit norm by A.5}\end{aligned}$$

For $\mathbf{J}_{\tilde{\mathbf{f}}^D}(\tilde{\mathbf{s}})$ to have orthogonal columns, $\mathbf{T}$ is diagonal (by A.6).

Thus, for $C_{\text{IMA}}(\tilde{\mathbf{f}}^D, p_{\mathbf{u}}) = 0$, $\mathbf{T}$ has to be almost surely diagonal w.r.t $p_{\tilde{\mathbf{s}}}$.

Consider an off-diagonal element of $\mathbf{T} = \mathbf{J}_{\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}}) = \frac{\partial g^{D^{-1}}_i}{\tilde{\mathbf{s}}_j}$ for $i \neq j$, and because continuous functions which are zero almost everywhere must be zero everywhere, we conclude that $\frac{\partial g^{D^{-1}}_i}{\tilde{s}_j} = 0$ everywhere for $i \neq j$, i.e. the Jacobian $\mathbf{J}_{\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}})$ is *diagonal everywhere*.

Hence, we conclude from Lemma A.7 that $\mathbf{g}^{D^{-1}}$ must be an element-wise function, $\mathbf{g}^{D^{-1}}(\tilde{\mathbf{s}}) = (g^{D^{-1}}_1(\tilde{s}_1), g^{D^{-1}}_2(\tilde{s}_2), ..., g^{D^{-1}}_d(\tilde{s}_d))$.

Since $\tilde{\mathbf{s}}$ has independent components by construction, it follows that $\tilde{x}_i = (g^{D^{-1}}_i(\tilde{s}_i)$ and $\tilde{x}_j = (g^{D^{-1}}_j(\tilde{s}_j)$ are independent for any $i \neq j$. This implies that $\mathbf{O}$ is a trivial matrix, i.e. a permutation or element-wise scaling. This is a contradiction to our theorem assumption.

We conclude that $\mathbf{J}_{\mathbf{g}^{D^{-1}}}(\tilde{\mathbf{s}})$ cannot be diagonal almost everwhere, and hence, $C_{\text{IMA}}(\tilde{\mathbf{f}}^D, p_{\mathbf{u}}) > 0$.

Thus a solution satisfying $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ can be distinguished from $(\tilde{\mathbf{f}}^D, p_{\mathbf{u}})$ based on the contrast C_{IMA}.

□

B Genericity of IMA under the manifold hypothesis

In this section, we provide a detailed exposition of the genericity arugument for IMA under the manifold hypothesis, presented in Section 5 of the main text.

B.1 Levy's Lemma

Genericity claims typically rely on high-dimensional concentration results (Janzing and Schölkopf, 2010; Janzing et al., 2010). In our work, we heavily use Levy's lemma, which is concentration result on smooth functions of vectors sampled from spherically symmetric priors around their mean.

Lemma B.1 (Lévy's Lemma (Janzing et al., 2010))**.** *Let g: $\mathbb{U}_m \to \mathbb{R}$ be a L-Lipschitz continuous function on the m-dimensional sphere. If a point $\mathbf{u}$ on $\mathbb{D}_m$ is randomly chosen according to an $\mathbf{O}(m)$-invariant prior, it satisfies*

$$|g(\mathbf{u}) - \bar{g}| \leq \epsilon$$

with probability at least $1 - \exp(-\kappa(m-1)\epsilon^2/L^2)$ for some constant κ where $\bar{g}$ can be interpreted as the median or average of $g(\mathbf{u})$.

B.2 Bound on the local IMA contrast

Following are helper lemmata for proving Theorem 5.1, which presents a high probability upper bound on the local IMA contrast, $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$, on functions, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, sampled according to a statistical process which tries to emulate the IMA princicple, see section 5.1 in the main text.

Lemma B.2. *Consider a random matrix $\mathbf{V} \in \mathbb{R}^{m \times d}$ with columns $\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_d \overset{i.i.d}{\sim} p_\mathbf{r}$ where $p_\mathbf{r}$ is a finite spherically symmetric distribution (A.2) on the Lebesgue measure over $\mathbb{R}^m$. Then $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_d$ are non-zero linearly independent with probability 1.*

Proof. We prove the statement by induction. $\mathbf{v} \sim p_\mathbf{r} \neq 0$ with probability 1 since the probability mass of $p_\mathbf{r}$ at $\mathbf{v} = 0$ is infinitesimally small. By the induction hypothesis, $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k \overset{i.i.d}{\sim} p_\mathbf{r}$ are linearly independent, i.e. they span a k-dimensional subspace in $\mathbb{R}^m$, $\mathbb{D}_k$. Consider $\mathbf{v}_{k+1} \sim p_\mathbf{r}$. The probability that $\mathbf{v}_{k+1} \in \mathbb{D}_k$ is infinitesimally small. In fact, $p_\mathbf{r}$ is finite at all points and the volume of a k-dimensional linear subspace with respect to the Lebesgue measure defined on $\mathbb{R}^m$ is infinitesimally small. Thus, the vectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k, \mathbf{v}_{k+1}$ are linearly independent and span a $(k+1)$-dimensional linear subspace in $\mathbb{R}^m$. Hence, we conclude that if $\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_d \overset{i.i.d}{\sim} p_\mathbf{r}$, $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_d$ are linearly independent with probability 1. □

Lemma B.3 (Spherically invariant distributions (lodzimierz Bryc, 1995))**.** *Suppose $\mathbf{X}$ is an m-dimensional random vector with spherically symmetric distribution. Then, $\mathbf{X} = R\mathbf{U}$, where the random variable $\mathbf{U} \sim$ Unif$(\mathbb{U}_m)$ is uniformly distributed on the unit sphere $\mathbb{U}_m \subset \mathbb{R}^m$, $R \triangleq \|\mathbf{X}\| \geq 0$ is real valued, and random variables $(R, \mathbf{U})$ are statistically independent.*

Lemma B.4. *Consider $\mathbf{W}_i, \mathbf{W}_j \sim$ Unif$(\mathbb{U}_m)$ where $\mathbb{U}_m$ is the unit sphere in $\mathbb{R}^m$. Then, $\langle \mathbf{W}_i, \mathbf{W}_j \rangle \triangleq \langle \mathbf{w}_i, \mathbf{W}_j \rangle$ for any $\mathbf{w}_i \in \mathbb{U}_m$ where $\triangleq$ represents being congruent in distribution.*

Proof. First we show that $\langle \mathbf{w}_i, \mathbf{W}_j \rangle$ has the same distribution for all $\mathbf{w}_i \in \mathbb{U}_m$ i.e. $\langle \mathbf{w}_i, \mathbf{W}_j \rangle \sim p_d \; \forall \mathbf{w}_i \in \mathbb{U}_m$.

For any orthonormal matrix $\mathbf{O} \in \mathbb{R}^{m \times m}$,

$$\begin{aligned} \langle \mathbf{w}_i, \mathbf{W}_j \rangle &\triangleq \langle \mathbf{w}_i, \mathbf{O}\mathbf{W}_j \rangle \qquad && \mathbf{W}_j \sim \text{Unif}(\mathbb{U}_m) \text{ which is a spherical invariant distribution} \\ &\cong \left\langle \mathbf{O}^\top \mathbf{w}_i, \mathbf{W}_j \right\rangle \end{aligned}$$

Since any $\mathbf{w}_i, \mathbf{w}_k \in \mathbb{U}_m$ are related through an orthonormal transformation, $\langle \mathbf{w}_i, \mathbf{W}_j \rangle \sim p_d \; \forall \mathbf{w}_i \in \mathbb{U}_m$.

To show $\langle \mathbf{W}_i, \mathbf{W}_j \rangle \cong \langle \mathbf{w}_i, \mathbf{W}_j \rangle$ for any $\mathbf{w}_i \in \mathbb{U}_m$:

$$\begin{aligned} \mathbb{P}(\langle \mathbf{W}_i, \mathbf{W}_j \rangle = c) &= \int_{\mathbf{w}_i} \mathbb{P}(\langle \mathbf{W}_i = \mathbf{w}_i, \mathbf{W}_j \rangle = c)\mathbb{P}(\mathbf{W}_i = \mathbf{w}_i) d\mathbf{w}_i \\ &= \mathbb{P}(\langle \mathbf{w}_i, \mathbf{W}_j \rangle = c) \int_{\mathbf{w}_i} \mathbb{P}(\mathbf{W}_i = \mathbf{w}_i) d\mathbf{w}_i \quad \text{for any } \mathbf{w}_i \in \mathbb{U}_m \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{By } \langle \mathbf{w}_i, \mathbf{W}_j \rangle \sim p_d \; \forall \mathbf{w}_i \in \mathbb{U}_m \\ &= \mathbb{P}(\langle \mathbf{w}_i, \mathbf{W}_j \rangle = c) \end{aligned}$$

Hence, $\langle \mathbf{W}_i, \mathbf{W}_j \rangle \triangleq \langle \mathbf{w}_i, \mathbf{W}_j \rangle$ for any $\mathbf{w}_i \in \mathbb{U}_m$.

□

Lemma B.5 (Lower bounds on determinants of matrices, Corollary 3 in (Brent et al., 2014))**.** *If $\mathbf{A} = \mathbb{I} - \mathbf{E} \in \mathbb{R}^{n \times n}$, $|E_{ij}| \leq \epsilon$ for $1 \leq i, j \leq n, E_{ii} = 0$ for $1 \leq i \leq n$, and $(n-1)\epsilon \leq 1$, then*

$$|\mathbf{A}| \geq (1 - (n-1)\epsilon)(1 + \epsilon)^{n-1}$$

and the inequality is sharp. A non-sharp lower bound is as follows:

$$|\mathbf{A}| \geq 1 - n\epsilon \tag{15}$$

Note that the non-sharp bound in (15) holds when the diagonal elements of $\mathbf{E}$, E_{ii} are non-zero.

Theorem 5.1. *Consider linear maps, $\mathbf{f}(\mathbf{s}) = \mathbf{J}\mathbf{s}$, where the columns of $\mathbf{J} \in \mathbb{R}^{m \times d}$ are sampled from a spherically symmetric distribution $p_\mathbf{r}$ over $\mathbb{R}^m$; $\mathbf{J}_1, \mathbf{J}_2, ..., \mathbf{J}_d \overset{i.i.d}{\sim} p_\mathbf{r}$. For such maps, the IMA contrast satisfies for $m \gg d$ and $\delta > 0$:*

$$\Pr\left[C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s}) \leq \delta\right] \geq 1 - \min\left\{1, \exp\left(2\log d - \kappa(m-1)\frac{\delta^2}{d^2}\right)\right\}.$$

Remark: A sharper lower bound for the local IMA contrast is as follows,

$$c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \leq \frac{1}{2}(-\log(1-(d-1)\epsilon) - (d-1)\log(1+\epsilon))$$

with (high) probability $\geq 1 - \min\left\{1, \exp(2\log d - \kappa(m-1)\epsilon^2)\right\}$ for $m \gg d$ and $\epsilon > 0$.

Proof. Computing $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$(6) relies on $\mathbf{J_f}(\mathbf{s})$ being full column-rank. This holds by Lemma B.2. Further, $|\mathbf{J}_\mathbf{f}^\top(\mathbf{s})\mathbf{J_f}(\mathbf{s})|$ is finite as all the columns of $\mathbf{J_f}(\mathbf{s}), \mathbf{J}_{\mathbf{f},1}(\mathbf{s}), \mathbf{J}_{\mathbf{f},2}(\mathbf{s}), ..., \mathbf{J}_{\mathbf{f},d}(\mathbf{s})$ are non-zero.

By Lemma B.3, $\mathbf{J_f}(\mathbf{s}) = \mathbf{WD}$, where $\mathbf{D} = diag(\|\mathbf{J}_{\mathbf{f},1}(\mathbf{s})\|, \|\mathbf{J}_{\mathbf{f},2}(\mathbf{s})\|, ..., \|\mathbf{J}_{\mathbf{f},d}(\mathbf{s})\|)$ and the columns of $\mathbf{W}$, $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d \sim$ Unif$(\mathbb{U}_m)$ where $\mathbb{U}_m$ is the unit sphere in $\mathbb{R}^m$.

Hence, $\mathbf{W}^\top\mathbf{W} = \mathbb{I}_d + \mathbf{E}$ where $E_{ii} = 0$. Consider the off-diagonal elements $E_{ij} = \langle \mathbf{W}_i, \mathbf{W}_j \rangle$ for $i \neq j$. $\langle \mathbf{W}_i, \mathbf{W}_j \rangle$ is congruent in distribution to $\langle \mathbf{w}_i, \mathbf{W}_j \rangle$ for any $\mathbf{w}_i \in \mathbb{U}_m$ by Lemma B.4.

Consider $g(\mathbf{W}_j) = \langle \mathbf{w}_i, \mathbf{W}_j \rangle$ for a given $\mathbf{w}_i \in \mathbb{U}_m$. $\mathbb{E}_{\mathbf{W}_j}(g(\mathbf{W}_j)) = 0$ since $\mathbf{W}_j$ comes from a spherically invariant distribution centered at 0. Further, $g(.)$ is Lipschitz with $L = 1$ since $\|\mathbf{w}_i\| = 1$.

By Lévy's Lemma B.1,

$$\mathbb{P}(|g(\mathbf{W}_j)| \leq \epsilon) \geq 1 - \exp(-\kappa(m-1)\epsilon^2) \text{ for arbitrarily small } \epsilon\,.$$

Since $E_{ij} \triangleq g(\mathbf{W}_j)$,

$$\mathbb{P}(|E_{ij}| \leq \epsilon) \geq 1 - \exp(-\kappa(m-1)\epsilon^2) \text{ for arbitrarily small } \epsilon\,. \tag{16}$$

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i,j\in[d], i\neq j} E_{ij} \leq \epsilon\right) &= 1 - \mathbb{P}\left(\bigcup_{i,j\in[d], i\neq j} E_{ij} \geq \epsilon\right) \\
&\geq 1 - \min\left\{1, \Sigma_{i,j\in[d], i\neq j}\mathbb{P}(E_{ij} \geq \epsilon)\right\} \quad \text{By union bound of probability} \\
&\geq 1 - \min\left\{1, d^2 e^{-\kappa(m-1)\epsilon^2}\right\} \qquad \text{By 16} \qquad (17) \\
&= 1 - \min\left\{1, e^{2\log d - \kappa(m-1)\epsilon^2}\right\} \qquad (18)
\end{aligned}$$

Hence, for $m \gg d$, $\bigcap_{i,j\in[d], i\neq j} E_{ij} \leq \epsilon$ with *high* probability.

We write the local IMA contrast $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ as a function of $\mathbf{W}^\top\mathbf{W}$ and the column norms of the Jacobian, $\mathbf{J_f}(\mathbf{s})$ so that we can bound it.

□

B.3 Bound on the global IMA contrast

Defining non-linear functions as composition of *two* affine functions We consider the initial scenario of partitioning the domain of the map, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, into two half-spaces $\mathbb{P}^{(0)}$ and $\mathbb{P}^{(1)}$ defined as follows. Given a non-zero vector $\mathbf{w} \in \mathbb{R}^d$ and $c \in \mathbb{R}$,

$$\mathbb{P}^{(0)} = \{\mathbf{s} \in \mathbb{R}^d, \mathbf{w}^\top\mathbf{s} \leq c\} \quad \text{and} \quad \mathbb{P}^{(1)} = \mathbb{R}^d \setminus \mathbb{P}^{(0)} = \{\mathbf{s} \in \mathbb{R}^d, \mathbf{w}^\top\mathbf{s} > c\}\,.$$

To define $\mathbf{f}$, we glue together two affine maps across the partition boundary. The affine maps are defined by the following local Jacobians,

$$\mathbf{J_f}(\mathbf{s}) = \begin{cases} \mathbf{J}^{(0)}, & \mathbf{s} \in \mathbb{P}^{(0)} \\ \mathbf{J}^{(1)}, & \mathbf{s} \in \mathbb{P}^{(1)} \end{cases}.$$

As previously mentioned, we locally retain the sampling procedure defined in the previous section—i.e., locally, the columns of the Jacobian of $\mathbf{f}, \mathbf{J_f}(\mathbf{s})$ are sampled from a spherically invariant distribution. Let us denote $\mathbf{J}_{:,k}$ the k-th column of $\mathbf{J}$. We sample i.i.d. the columns of matrices $\mathbf{J}^{(0)}$ and

$\mathbf{J}^{(0)} \in \mathbb{R}^{m \times d}$ as follows

$$\mathbf{J}^{(0)}_1(\mathbf{s}), \mathbf{J}^{(0)}_2(\mathbf{s}), ..., \mathbf{J}^{(0)}_d(\mathbf{s}) \overset{i.i.d}{\sim} p_{\mathbf{r}}\,,$$
$$\mathbf{J}^{(1)}_1(\mathbf{s}), \mathbf{J}^{(1)}_2(\mathbf{s}), ..., \mathbf{J}^{(1)}_d(\mathbf{s}) \overset{i.i.d}{\sim} p_{\mathbf{r}}\,.$$

where $p_{\mathbf{r}}$ is a spherically symmetric distribution in $\mathbb{R}^m$.

We thereby consider the resulting maps of the form,

$$\mathbf{f}(\mathbf{s}) = \begin{cases} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}\,, & \mathbf{w}^\top \mathbf{s} \leq c\,, \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1\,, & \mathbf{w}^\top \mathbf{s} > c\,. \end{cases}$$

Before we derive an upper bound on the global IMA contrast, $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$, we introduce lemmas to ensure that the function is well-defined and well-behaved at every point in the domain:

1. First, we provide the conditions for defining a continuous map $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$ by composing two affine maps.
2. Second, we discuss how to ensures injectivity of the composition of two affine maps.
3. Finally, we define a smooth approximation to the composition of two affine maps, and show that such approximation is continuously differentiable (in addition to being continuous and injective).

In the following lemma, we derive the condition required for local bases across the partition boundary, $\mathbf{w}^\top \mathbf{s} = c$, given by the Jacobians, $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}$, to be able to define a continuous map.

Lemma B.6 (Conditions on local bases for a continuous map)**.** *Consider a map $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$, with full column rank $m \times d$ matrices $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ such that*

$$\mathbf{f}(\mathbf{s}) = \begin{cases} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}\,, & \mathbf{w}^\top \mathbf{s} \leq c\,, \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1\,, & \mathbf{w}^\top \mathbf{s} > c\,, \end{cases}$$

where $\mathbf{w} \in \mathbb{R}^d, c \in \mathbb{R}, \mathbf{c}_1 \in \mathbb{R}^m$ are given. The local bases of $\mathbf{f}^{(0)}$ and $\mathbf{f}^{(1)}$, i.e. the columns of $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$, are sampled from a spherically invariant distribution, the columns of $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ are mutually independent. $\mathbf{f}$ is continuous non-linear only if colrank $\left[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\right] = 1$.

Further, if colrank $\left[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\right] = 1$, *then $\exists \mathbf{w} \in \mathbb{R}^d, c \in \mathbb{R}, \mathbf{c}_1 \in \mathbb{R}^m$ such that $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$,*

$$\mathbf{f}(\mathbf{s}) = \begin{cases} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}\,, & \mathbf{w}^\top \mathbf{s} \leq c\,, \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s}\,, + \mathbf{c}_1 & \mathbf{w}^\top \mathbf{s} > c\,, \end{cases}$$

is a continuous non-linear function.

Proof. Consider $\mathbf{s} \in \mathbb{R}^d$ such that $\mathbf{w}^\top \mathbf{s} > c$. We define the intersection point $i(\mathbf{s}) = \lambda(\mathbf{s})\mathbf{s}, \lambda(\mathbf{s}) \in (0, 1)$, of the segment between $\mathbf{0}$ and $\mathbf{s}$ with the partition boundary, $\mathbf{w}^\top \mathbf{s} = c, c > 0$ such that $\mathbf{w}^\top(\lambda(\mathbf{s})\mathbf{s}) = c$. Observe that the Jacobian of the map $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is as follows:

$$\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \begin{cases} \mathbf{J}^{(0)}\,, & \mathbf{w}^\top \mathbf{s} \leq c\,, \\ \mathbf{J}^{(1)}\,, & \mathbf{w}^\top \mathbf{s} > c\,. \end{cases}$$

If we assume $\mathbf{f}$ is continuous, in addition to be by definition *continuously differentiable* within each half-space, the following holds:

$$\begin{aligned} \mathbf{f}(\mathbf{s}) &= \int_{\mathbf{0}}^{i(\mathbf{s})} \mathbf{J}^{(0)} d\mathbf{s} + \int_{i(\mathbf{s})}^{\mathbf{s}} \mathbf{J}^{(1)} d\mathbf{s} \\ &= \mathbf{J}^{(0)} i(\mathbf{s}) + \mathbf{J}^{(1)}(\mathbf{s} - i(\mathbf{s})) \end{aligned}$$

such that

$$\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{J}^{(1)} + (\mathbf{J}^{(0)} - \mathbf{J}^{(1)}) \frac{\partial i(\mathbf{s})}{\partial \mathbf{s}}, \text{ whenever } \mathbf{w}^\top \mathbf{s} > c\,. \tag{19}$$

The partition boundary in the domain of $\mathbf{f}$, is defined by $\mathbf{w}^\top \mathbf{s} = c, c > 0$, and is thus a $(d-1)$-dimensinoal affine space whose associated vector space is the span ($(d-1)$-dimensional) of all $\frac{\partial i(\mathbf{s})}{\partial \mathbf{s}}$. To obtain the correct Jacobian, $\mathbf{J}_\mathbf{f}(\mathbf{s}) = \mathbf{J}^{(1)}$, in the half-space $\mathbf{w}^\top \mathbf{s} > c$., we need

$$(\mathbf{J}^{(0)} - \mathbf{J}^{(1)}) \frac{\partial i(\mathbf{s})}{\partial \mathbf{s}} = 0 \tag{20}$$

Thus, to obtain a *continuous* map $\mathbf{f}$, $\dim \big(\text{Null} \left[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\right]\big) \geq (d-1)$. Moreover, to have a non-linear function, we need the two Jacobian values to be different, such that we require $\dim \big(\text{Null} \left[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\right]\big) = (d-1)$.

By the Rank-Nullity Theorem (Wikipedia contributors, 2022),

$$\text{colrank}\left([\mathbf{J}^{(0)} - \mathbf{J}^{(1)}]\right) + \dim\left(\text{Null}[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}]\right) = \#\text{cols. of } [\mathbf{J}^{(0)} - \mathbf{J}^{(1)}]$$

Leading to $\text{colrank}\big([\mathbf{J}^{(0)} - \mathbf{J}^{(1)}]\big) = d - (d-1) = 1$ Hence, $\mathbf{f}$ is continuous non-linear *only if* $\text{colrank}\big[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\big] = 1$.

To show the reverse direction, consider the existence of $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}$ s.t. $\text{colrank}\big[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\big] = 1$. Consider the singular value decomposition of $\big[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\big] = \sigma \mathbf{u}\mathbf{w}^\top, \mathbf{u} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^d$. To construct a function, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, we start by partitioning the domain into two half space by a hyperplane normal to $\mathbf{w}$, say $\mathbf{w}^\top \mathbf{s} = c$ for $c \in \mathbb{R}$, leading to the half-spaces definition

$$\mathbb{P}^{(0)} = \{\mathbf{s} \in \mathbb{R}^d, \mathbf{w}^\top \mathbf{s} \leq c\} \quad \text{and} \quad \mathbb{P}^{(1)} = \mathbb{R}^d \setminus \mathbb{P}^{(0)} = \{\mathbf{s} \in \mathbb{R}^d, \mathbf{w}^\top \mathbf{s} > c\}\,.$$

For the partition boundary, $\mathbb{K} := \big\{\mathbf{s} : \mathbf{w}^\top \mathbf{s} = c\big\}$, $\big(\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\big)\, \mathbf{s} = \sigma \mathbf{u}\mathbf{w}^\top \mathbf{s} = c\sigma \mathbf{u}$. $\mathbf{c}_1 := c\sigma \mathbf{u}$ is a constant vector in $\mathbb{R}^m$ such that $\forall \mathbf{s} \in \mathbb{K}, \mathbf{J}^{(0)}\mathbf{s} = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1$. $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$,

$$\mathbf{f}(\mathbf{s}) = \begin{cases} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}\,, & \mathbf{w}^\top \mathbf{s} \leq c\,, \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s}\,, + \mathbf{c}_1 & \mathbf{w}^\top \mathbf{s} > c\,, \end{cases}$$

is a continuous function since, for the exhaustive cases for $\mathbf{s} \in \mathbb{R}^d$:

1. $\mathbf{w}^\top \mathbf{s} = c$

   $\mathbf{f}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}$, and in limit also equal to $\mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1$. Since we have shown that the limit of the function is equal to the value assumed by the function, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is continuous $\forall \mathbf{s} : \mathbf{w}^\top \mathbf{s} = c$ (Theorem 4.6, (Rudin et al., 1964)).

2. $\mathbf{w}^\top \mathbf{s} < c$

   $\mathbf{f}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}$ is affine in this region, and hence continuous.

3. $\mathbf{w}^\top \mathbf{s} > c$

   $\mathbf{f}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1$ is affine in this region, and hence continuous.

Hence, if $\exists \mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}$ s.t. $\text{colrank}\big[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\big] = 1$, then $\exists \mathbf{w} \in \mathbb{R}^d, c \in \mathbb{R}, \mathbf{c}_1 \in \mathbb{R}^m$ s. t. $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$,

$$\mathbf{f}(\mathbf{s}) = \begin{Bmatrix} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s} & \mathbf{w}^\top \mathbf{s} \leq c \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1 & \mathbf{w}^\top \mathbf{s} > c \end{Bmatrix}$$

is a continuous function.

□

This result leads to one possible way to sample nonlinear mixings approximately satisfying the global IMA principle.

Sampling procedure for piecewise affine continuous maps

Observation B.7 (Partition boundaries of continuous maps). For functions $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$,

$$\mathbf{f}(\mathbf{s}) = \begin{cases} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}\,, & \mathbf{w}^\top \mathbf{s} \leq c\,, \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1\,, & \mathbf{w}^\top \mathbf{s} > c\,, \end{cases},$$

where $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m\times d}$, one way to achieve colrank $\left[\mathbf{J}^{(0)} - \mathbf{J}^{(1)}\right] = 1$ is the following:

1. Sample the columns of $\mathbf{J}^{(0)}$ independently from the mentioned spherically symmetric distribution $p_\mathbf{r}$, $\mathbf{J}_1^{(0)}, \mathbf{J}_2^{(0)}, ..., \mathbf{J}_d^{(0)} \overset{i.i.d}{\sim} p_\mathbf{r}$.
2. To construct $\mathbf{J}^{(1)}$, retain any $(d-1)$ columns of $\mathbf{J}^{(0)}$ and sample the remaining column, $\mathbf{J}_k^{(1)} \sim p_\mathbf{r}$.

Notice that the sampling procedure described above is locally equivalent to the one described in Theorem 5.1. We deliberately retain the same sampling procedure so that we can derive a similar upper bound to the IMA contrast in the case of non-affine functions defined by joining two contiguous affine maps (Definition B.8).

We will now show that a consequence of this sampling procedure is that the boundary of the partition of the domain is constrained to be axis-aligned. The alignment of the partition boundary $\mathbf{w}^\top \mathbf{s} = c$, given by $\mathbf{w}$, is determined by the choice of $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$.

Without loss of generality, consider that the last column is newly sampled (B.7) in $\mathbf{J}^{(1)}$. Consider $\mathbf{s} \in \mathbb{R}^d$ such that $\mathbf{w}^\top \mathbf{s} > c$. By (20), for continuous $\mathbf{f}$, $(\mathbf{J}^{(0)} - \mathbf{J}^{(1)})\frac{\partial i(\mathbf{s})}{\partial \mathbf{s}} = 0$. We thus have the following constraints,

$$\mathbf{w}^\top \left(\frac{\partial i(\mathbf{s})}{\partial \mathbf{s}}\right) = \mathbf{0}; \left[\frac{\partial i(\mathbf{s})}{\partial \mathbf{s}}\right]_{d.} = \mathbf{0} \tag{21}$$

(21) is achieved *iff.* $\mathbf{w}$ defines an axis-aligned $(d-1)$-dimensional subspace normal to the canonical basis vector associated with the index of the column that changes from $\mathbf{J}^{(0)}$ to $\mathbf{J}^{(1)}$ (here, the last column), i.e.

$$\mathbf{w} = \alpha[0, 0, ..., 1]^\top$$

Thus, the aforementioned sampling procedure of local bases (B.7) (i.e. the columns of $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$) to ensure continuity of the resultant manifold, leads to a constraint on the partition of the domain of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, i.e. the partition can only be axis-aligned.

Later in this chapter (Section B.3), when will extend this construction to more expressive maps where the partition of the input domain is defined as a grid. We will show that axis-alignment of the partition still allows some degree of expressivity for the resulting class of maps. Those can approximate a large family of Riemannian manifolds embedded in $\mathbb{R}^m$ isomorphic to the d-dimensional Euclidean space.

Hence, consider the following definition of maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$ by composing two affine maps, incorporating the axis alignement contraint,

Definition B.8 (Maps defined by composing two affine maps)**.** *Consider the columns of matrices $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ are sampled by the following procedure:*

1. *Sample the d columns of $\mathbf{J}^{(0)}$ independently from the mentioned spherically symmetric distributio, $p_\mathbf{r}$, $\mathbf{J}_{:,1}^{(0)}, \mathbf{J}_{:,2}^{(0)}, ..., \mathbf{J}_{:,d}^{(0)} \overset{i.i.d}{\sim} p_\mathbf{r}$.*
2. *To construct $\mathbf{J}^{(1)}$, retain any $(d-1)$ columns of $\mathbf{J}^{(0)}$ and sample the remaining k-th column, $\mathbf{J}_{:,k}^{(1)} \sim p_\mathbf{r}$.*

Consider the map $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$,

$$\mathbf{f}(\mathbf{s}) = \begin{cases} \mathbf{f}^{(0)}(\mathbf{s}) = \mathbf{J}^{(0)}\mathbf{s}\,, & s_k \leq c\,, \\ \mathbf{f}^{(1)}(\mathbf{s}) = \mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1\,, & s_k > c\,, \end{cases}$$

where $c \in \mathbb{R}$ is given and $\mathbf{c}_1 \in \mathbb{R}^m$ is set by continuity at the boundary to $\mathbf{c}_1 = c\left(\mathbf{J}_{:,k}^{(0)} - \mathbf{J}_{:,k}^{(1)}\right)$.

Note that since the partition boundary for the change of Jacobian is axis-aligned, the map $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ can be alternatively written as a sum of coordinate-wise functions, $\mathbf{f}(\mathbf{s}) = \sum_{i=1}^{d} \mathbf{f}_i(s_i)$ for $\mathbf{s} = \{s_1, s_2, ..., s_d\}$. Without loss of generality, we describe the case $k = d$. The coordinate-wise functions $\mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m$ are then defined as follows:

1. $\mathbf{f}_i(s_i) = \mathbf{J}^{(0)}_{:,i} s_i = \mathbf{J}^{(1)}_{:,i} s_i \forall i \in \{1, 2, ..., (d-1)\}$
2. $\mathbf{f}_d(s_d) = \begin{cases} \mathbf{J}^{(0)}_{:,d} s_d & s_d \le t_d \\ \mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d & s_d > t_d \end{cases}$

for $t_d \in \mathbb{R}$ determined by $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}$.

Next, we show that the functions above defined (Definition B.8) are injective. The idea is that injectivity of coordinate-wise functions of $\mathbf{f}$, $\mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m$ that are in direct sum entails of $\mathbf{f}(\mathbf{s}) = \sum_{i=1}^{d} \mathbf{f}_i(s_i)$. We first show that the subspaces spanned by the images of $\mathbf{f}_i : \mathbb{R} \to \mathbb{R}^d$ are linearly independent, and thereby are in (internal) direct sum (Definition B.9) with respect to the image of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$. Then, we show that the coordinate-wise functions, $\mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m$ are injective, and conclude that $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is injective.

Definition B.9 (Internal Direct Sum of Subspaces, Chapter 1 (Roman et al., 2005))**.** *Let $\mathbb{V}$ be a vector space. We say that $\mathbb{V}$ is an (internal) direct sum of the family $\mathcal{F} = \{\mathbf{S}_i | i \in K\}$ of subspaces of $\mathbb{V}$ if every vector $\mathbf{v} \in \mathbb{V}$ can be written in a unique way (except for order) as a finite sum of vectors from the subspaces in $\mathcal{F}$, that is, if for all $\mathbf{v} \in \mathbb{V}$,*

$$\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_n$$

for $\mathbf{u}_i \in \mathbf{S}_i$ and furthermore, if

$$\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_n$$

where $\mathbf{w}_i \in \mathbf{S}_i$, then $m = n$ and (after reindexing if necessary) $\mathbf{w}_i = \mathbf{u}_i$ for all $i = 1, 2, ..., n$. If $\mathcal{F} = \{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_n\}$ is a finite family, we write $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_n$.

Lemma B.10. *Consider $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}$ as sampled in Definition (B.8). With probability 1, the vector space $\mathbb{V} = span\{\mathbf{J}^{(0)}_{:,1} = \mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(0)}_{:,2} = \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,(d-1)} = \mathbf{J}^{(1)}_{:,(d-1)}, \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}$ is the direct sum of the family $\mathcal{F} = \{\mathbf{S}_1 = span\{\mathbf{J}^{(0)}_{:,1} = \mathbf{J}^{(1)}_{:,1}\}, \mathbf{S}_2 = span\{\mathbf{J}^{(0)}_{:,2} = \mathbf{J}^{(1)}_{:,2}\}, ..., \mathbf{S}_{d-1} = span\{\mathbf{J}^{(0)}_{:,d-1} = \mathbf{J}^{(1)}_{:,d-1}\}, \mathbf{S}_d = span\{\mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}\}$.*

Proof. From Lemma B.2, for the scenario $m \gg d$, the vectors $\{\mathbf{J}^{(0)}_{:,1} = \mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(0)}_{:,2} = \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,(d-1)} = \mathbf{J}^{(1)}_{:,(d-1)}, \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}$ are non-zero and linearly independent with probability 1.

Consider $\mathbf{v} \in \mathbb{V}, \mathbf{u}_1, \mathbf{w}_1 \in \mathbf{S}_1, \mathbf{u}_2, \mathbf{w}_2 \in \mathbf{S}_2, ..., \mathbf{u}_d, \mathbf{w}_d \in \mathbf{S}_d$ such that

$$\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_d, \mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_d$$

By definition B.9, to show $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$, we need to show $\mathbf{u}_1 = \mathbf{w}_1, \mathbf{u}_2 = \mathbf{w}_2, ..., \mathbf{u}_d = \mathbf{w}_d$.

Let

- $u_1 = c_1 \mathbf{J}^{(0)}_{:,1}, w_1 = c_1' \mathbf{J}^{(0)}_{:,1}$
- $u_2 = c_2 \mathbf{J}^{(0)}_{:,2}, w_2 = c_2' \mathbf{J}^{(0)}_{:,2}$

  $\vdots$
- $u_{d-1} = c_{d-1} \mathbf{J}^{(0)}_{:,d-1}, w_{d-1} = c_{d-1}' \mathbf{J}^{(0)}_{:,d-1}$
- $u_d = c_d^{(0)} \mathbf{J}^{(0)}_{:,d} + c_d^{(1)} \mathbf{J}^{(1)}_{:,d}, v_d = c_d^{(0)'} \mathbf{J}^{(0)}_{:,d} + c_d^{(1)'} \mathbf{J}^{(1)}_{:,d}$

$$\begin{aligned}
&\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_d = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_d \\
&c_1 \mathbf{J}^{(0)}_{:,1} + c_2 \mathbf{J}^{(0)}_{:,2} + ... + c_{d-1} \mathbf{J}^{(0)}_{:,d-1} + c_d^{(0)} \mathbf{J}^{(0)}_{:,d} + c_d^{(1)} \mathbf{J}^{(1)}_{:,d} = \\
&c'_1 \mathbf{J}^{(0)}_{:,1} + c'_2 \mathbf{J}^{(0)}_{:,2} + ... + c_{d-1} \mathbf{J}^{(0)}_{:,d-1} + c_d^{(0)'} \mathbf{J}^{(0)}_{:,d} + c_d^{(1)'} \mathbf{J}^{(1)}_{:,d} \\
&(c_1 - c'_1) \mathbf{J}^{(0)}_{:,1} + (c_2 - c'_2) \mathbf{J}^{(0)}_{:,2} + ... + (c_{d-1} - c'_{d-1}) \mathbf{J}^{(0)}_{:,d-1} \\
&+ (c_d^{(0)} - c_d^{(0)'}) \mathbf{J}^{(0)}_{:,d} + (c_d^{(1)} - c_d^{(1)'}) \mathbf{J}^{(1)}_{:,d} = \mathbf{0}
\end{aligned}$$

Since $\{\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(0)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,d-1}, \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}$ are nonzero and linearly independent with probability 1 (Lemma B.2),

$$(c_1 - c'_1) = (c_2 - c'_2) = ... = (c_{d-1} - c'_{d-1}) = (c_d^{(0)} - c_d^{(0)'}) = (c_d^{(1)} - c_d^{(1)'}) = 0$$

Hence, it follows that $\mathbf{u}_1 = \mathbf{w}_1, \mathbf{u}_2 = \mathbf{w}_2, ..., \mathbf{u}_d = \mathbf{w}_d$ and $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ... \mathbf{S}_d$.

□

Lemma B.11. *Consider maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$, where $\mathbf{f}$ can be written as sum of coordinate-wise functions, $\mathbf{f}(\mathbf{s}) = \sum_{i=1}^d \mathbf{f}_i(s_i)$ for $\mathbf{s} = \{s_1, s_2, ..., s_d\}$. We define $\mathbb{V} = span(Im(\mathbf{f})), \mathbf{S}_1 = span(Im(\mathbf{f}_1)), \mathbf{S}_2 = span(Im(\mathbf{f}_2)), ..., \mathbf{S}_d = span(Im(\mathbf{f}_d))$, where $Im(\mathbf{f}), Im(\mathbf{f}_i)$ denote the images of the functions $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, \mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m \quad \forall i \in [d]$. If $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ... \bigoplus \mathbf{S}_d$, the injectivity of the coordinate-wise functions $\mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m$ implies the injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$.*

Proof. To show injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, we need to show the following:

$$\forall \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^d : \mathbf{f}(\mathbf{s}_1) = \mathbf{f}(\mathbf{s}_2) \implies \mathbf{s}_1 = \mathbf{s}_2 \tag{22}$$

We show (22) by contradiction. Let

$$\exists \mathbf{s}^{(1)} \neq \mathbf{s}^{(2)} \in \mathbb{R}^d \text{ s.t.} \mathbf{f}(\mathbf{s}^{(1)}) = \mathbf{f}(\mathbf{s}^{(2)}) \tag{23}$$

Observe that $\text{Im}(\mathbf{f}) \subseteq \mathbb{V}, \text{Im}(\mathbf{f}_i) \subseteq \mathbf{S}_i \forall i \in [d]$. By Lemma B.10, the vector space $\mathbb{V}$ is the direct sum of the subspaces $\mathbf{S}_i, \forall i \in [d]$, i.e. $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ... \bigoplus \mathbf{S}_d$.

Hence by definition of direct sum (Definition B.9), it follows that

$$\mathbf{f}(\mathbf{s}^{(1)}) = \mathbf{f}(\mathbf{s}^{(1)}) \implies \mathbf{f}_1(s_1^{(1)}) = \mathbf{f}_1(s_1^{(2)}), \mathbf{f}_1(s_2^{(1)}) = \mathbf{f}_1(s_2^{(2)}), ..., \mathbf{f}_1(s_d^{(1)}) = \mathbf{f}_1(s_d^{(2)})$$

By injectivity of the coordinate-wise functions,

$$\begin{aligned}
\mathbf{f}_1(s_1^{(1)}) = \mathbf{f}_1(s_1^{(2)}) &\implies s_1^{(1)} = s_1^{(2)} \\
\mathbf{f}_1(s_2^{(1)}) = \mathbf{f}_1(s_2^{(2)}) &\implies s_2^{(1)} = s_2^{(2)} \\
\vdots \mathbf{f}_1(s_d^{(1)}) = \mathbf{f}_1(s_d^{(2)}) &\implies s_d^{(1)} = s_d^{(2)} \\
\implies \mathbf{s}^{(1)} = \mathbf{s}^{(2)}
\end{aligned}$$

We arrive at a contradiction to (23), hence (22) holds. Injectivity of the coordinate-wise functions, $\mathbf{f}_1 : \mathbb{R} \to \mathbb{R}^m, \mathbf{f}_2 : \mathbb{R} \to \mathbb{R}^m, ..., \mathbf{f}_d : \mathbb{R} \to \mathbb{R}^m$ implies injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ as defined in B.8.

□

Lemma B.12 (Injectivity of maps defined as a composition of two affine spaces)**.** *Maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.8 are continuous and injective with probability one.*

Proof. $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is continuous by lemma B.6.

By definition (B.8), $\mathbf{f}(\mathbf{s}) = \sum_{i=1}^{d} f_i(s_i)$ for $\mathbf{s} = \{s_1, s_2, ..., s_d\}$. Consider the coordinate-wise functions $\mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m$:

1. $\mathbf{f}_i(s_i) = \mathbf{J}^{(0)}_{:,i} s_i = \mathbf{J}^{(1)}_{:,i} s_i \forall i \in \{1, 2, ..., (d-1)\}$
2. $\mathbf{f}_d(s_d) = \begin{cases} \mathbf{J}^{(0)}_{:,d} s_d & s_d \leq t_d \\ \mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d & s_d > t_d \end{cases}$

for $t_d \in \mathbb{R}$.

1. $\mathbf{f}_i : \mathbb{R} \to \mathbb{R} \forall i \in \{1, 2, ..., (d-1)\}$ are injective since they are affine.
2. To show that $\mathbf{f}_d : \mathbb{R} \to \mathbb{R}^m$ is injective, we need to show the following:

$$\forall s_d^{(1)}, s_d^{(2)} \in \mathbb{R} : \; \mathbf{f}_d(s_d^{(1)}) = \mathbf{f}_d(s_d^{(2)}) \implies s_d^{(1)} = s_d^{(2)} \tag{24}$$

As usual, we show (24) by contradiction. Let

$$\exists s_d^{(1)} \neq s_d^{(2)} \in \mathbb{R} \text{ s.t. } \mathbf{f}_d(s_d^{(1)}) = \mathbf{f}_d(s_d^{(2)}) \tag{25}$$

Consider the following cases,

(a) $s_d^{(1)}, s_d^{(2)} \leq t_d$ Then,

$$\begin{aligned} \mathbf{f}_d(s_d^{(1)}) &= \mathbf{f}_d(s_d^{(2)}) \\ \mathbf{J}^{(0)}_{:,d} s_d^{(1)} &= \mathbf{J}^{(0)}_{:,d} s_d^{(2)} \\ s_d^{(1)} &= s_d^{(2)} && \because \mathbf{J}^{(0)}_{:,d} \in \mathbb{R}^m \text{ is non-zero w.p. 1. (Lemma B.2)} \end{aligned}$$

Hence, we arrive at a contradiction to (25) in this case.

(b) $s_d^{(1)} \leq t_d, s_d^{(2)} > t_d$ Then,

$$\begin{aligned} \mathbf{f}_d(s_d^{(1)}) &= \mathbf{f}_d(s_d^{(2)}) \\ \mathbf{J}^{(0)}_{:,d} s_d^{(1)} &= \mathbf{J}^{(1)}_{:,d}(s_d^{(2)} - t_d) + \mathbf{J}^{(0)}_{:,d} t_d \\ \mathbf{J}^{(0)}_{:,d}(s_d^{(1)} - t_d) &= \mathbf{J}^{(1)}_{:,d}(s_d^{(2)} - t_d) \\ [\mathbf{J}^{(0)}_{:,d} - \mathbf{J}^{(1)}_{:,d}] \begin{bmatrix} s_d^{(1)} - t_d \\ s_d^{(2)} - t_d \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \implies \begin{bmatrix} s_d^{(1)} - t_d \\ s_d^{(2)} - t_d \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ & \because \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d} \text{ are non-zero} \\ & \text{and linearly independent w. p. 1 (Lemma B.2)} \\ \implies s_d^{(1)} = s_d^{(2)} &= t_d \end{aligned}$$

Thus, we arrive at a contradiction since $s_d^{(2)} > t_d$.

(c) $s_d^{(1)}, s_d^{(2)} > t_d$ Then,

$$\begin{aligned}
\mathbf{f}_d(s_d^{(1)}) &= \mathbf{f}_d(s_d^{(2)}) \\
\mathbf{J}^{(1)}_{:,d}(s_d^{(1)} - t_d) + \mathbf{J}^{(0)}_{:,d} t_d &= \mathbf{J}^{(1)}_{:,d}(s_d^{(2)} - t_d) + \mathbf{J}^{(0)}_{:,d} t_d \\
\mathbf{J}^{(1)}_{:,d} s_d^{(1)} &= \mathbf{J}^{(1)}_{:,d} s_d^{(2)} \\
s_d^{(1)} &= s_d^{(2)} \\
&\because \mathbf{J}^{(1)}_{:,d} \in \mathbb{R}^m \text{ is non-zero w.p. 1.} \\
&\text{(Lemma B.2)}
\end{aligned}$$

Hence, we also arrive at a contradiction to (25) in this case.

Since we arrive at a contradiction to (25) in the aforementioned exhaustive and mutually exclusive cases for $s_d^{(1)}, s_d^{(2)} \in \mathbb{R}$, we conclude that $\mathbf{f}_d : \mathbb{R} \to \mathbb{R}^m$ is injective.

Hence, we have shown that the coordinate wise functions, $f_i : \mathbb{R} \to \mathbb{R}^m$ are injective. By Lemma B.11, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is injective.

□

We now develop a smooth approximation to the map defined in B.8.

Definition B.13 (Smooth step function)**.** *We define the smooth step function as* $\tilde{1}_\epsilon : \mathbb{R} \to \mathbb{R}$ *as*

$$\tilde{1}_\epsilon(s) = \begin{cases} 0\,, & s \leq -\epsilon\,, \\ \frac{1}{2}\sin\left(\frac{\pi s}{2\epsilon}\right) + \frac{1}{2}\,, & -\epsilon < s \leq \epsilon\,, \\ 1\,, & s > \epsilon\,. \end{cases}$$

Definition B.14 (Smoothing composition of affine maps)**.** *Consider maps* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *defined in B.8. Such maps can be written as,*

$$\mathbf{f}(\mathbf{s}) = (\mathbf{J}^{(0)}\mathbf{s})1_{\mathbf{w}^\top \mathbf{s} \leq c} + (\mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1)1_{\mathbf{w}^\top \mathbf{s} > c}$$

where $\mathbf{w} \in \mathbb{R}^d, c \in \mathbb{R}, \mathbf{c}_1 \in \mathbb{R}^m$ *are given.*

We define the smoothened version of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *as* $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ *as,*

$$\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = (\mathbf{J}^{(0)}\mathbf{s})\tilde{1}_\epsilon(c - \mathbf{w}^\top \mathbf{s}) + (\mathbf{J}^{(1)}\mathbf{s} + \mathbf{c}_1)\tilde{1}_\epsilon(\mathbf{w}^\top \mathbf{s} - c)$$

Note that since $\mathbf{w} \in \mathbb{R}^d$ *is an axis-aligned vector, without loss of generality for* $\mathbf{w} = \mathbf{e}_d = (0, 0, ..., 1) \in \mathbb{R}^d$*, the function* $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ *can be defined as a sum of coordinate-wise functions,* $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{i=1}^d \tilde{\mathbf{f}}_{\epsilon,i}(s_i)$ *where* $\mathbf{s} \in \mathbb{R}^d = (s_1, s_2, ..., s_d)$*:*

1. $\tilde{\mathbf{f}}_{\epsilon,i}(s_i) = \mathbf{J}^{(0)}_{:,i} s_i = \mathbf{J}^{(1)}_{:,i} s_i \forall i \in \{1, 2, ..., (d-1)\}$
2. $\tilde{\mathbf{f}}_{\epsilon,d}(s_d) = \mathbf{J}^{(0)}_{:,d} s_d \tilde{1}_\epsilon(t_d - s_d) + (\mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d)\tilde{1}_\epsilon(s_d - t_d)$

for $t_d \in \mathbb{R}$ *determined by* $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}, \mathbf{w} \in \mathbb{R}^d, \mathbf{c}_1 \in \mathbb{R}^m$.

Finally, before we present the theorem with the high probability bound on the global IMA contrast, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$ for any finite probability density, $p_\mathbf{s}$ on $\mathbb{R}^d$, we introduce a lemma to show that maps $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.14 are continuous, injective and continuously differentiable. The objective of the following lemma is to ensure that the Jacobian of $\tilde{\mathbf{f}}_\epsilon$, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon} \in \mathbb{R}^{m \times d}$, is well-defined for at all points in the domain of $\tilde{\mathbf{f}}_\epsilon$ such that the IMA contrast, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$ can be computed for maps, $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$, can be computed with respect to all finite distribuitions, $p_\mathbf{s}$ on $\mathbb{R}^d$.

Lemma B.15. *Functions* $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ *defined in B.14 are continuously differentiable in* $\mathbb{R}^d$*, in addition to being continuous and injective, are continuously differentiable with* $\epsilon > 0$ *arbitrarily small.*

Proof. We show successively that $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ is continuous, injective and continuously differentiable.

Continuity of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$

Consider the coordinate-wise decomposition of $\tilde{\mathbf{f}}_\epsilon; \tilde{f}_{\epsilon,i}(\mathbf{s}) = \mathbf{J}^{(0)}_{i,:}\mathbf{s}\tilde{\mathbb{1}}_\epsilon(c - \mathbf{w}^\top \mathbf{s}) + (\mathbf{J}^{(1)}_{i,:}\mathbf{s} + c_{1,i})\tilde{\mathbb{1}}_\epsilon(\mathbf{w}^\top \mathbf{s} - c) \forall i \in [m]$.

$\mathbf{J}^{(0)}_{i,:}\mathbf{s}, \mathbf{J}^{(1)}_{i,:}\mathbf{s} + c_{1,i}$ are continuous in $\mathbf{s} \in \mathbb{R}^d$ since they are affine.

Note that $\tilde{\mathbb{1}}_\epsilon : \mathbb{R} \to \mathbb{R}$ is continuous by definition. $\tilde{\mathbb{1}}_\epsilon(c - \mathbf{w}^\top \mathbf{s}), \tilde{\mathbb{1}}_\epsilon(\mathbf{w}^\top \mathbf{s} - c)$ are compositions of a continuous function with affine functions (thereby continuous), and hence are continuous (Theorem 4.9, (Rudin et al., 1964)).

$\tilde{f}_{\epsilon,i} : \mathbb{R}^d \to \mathbb{R}$, being a sum of continuous functions, is continuous for all $i \in [m]$ (Theorem 4.9, (Rudin et al., 1964)).

Since the coordinate functions of $\tilde{\mathbf{f}} : \mathbb{R}^d \to \mathbb{R}^m$, $\tilde{f}_i : \mathbb{R}^d \to \mathbb{R}$ are continuous, $\tilde{\mathbf{f}}$ is continuous (Theorem 4.10, (Rudin et al., 1964)).

Injectivity of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$

Consider the coordinate-wise functions as defined in B.14.

1. By Lemma B.12, the functions $\tilde{\mathbf{f}}_{\epsilon,i} = \mathbf{f}_i : \mathbb{R} \to \mathbb{R}^m \forall i \in \{1, 2, ..., d-1\}$ are injective.
2. We now show that $\tilde{\mathbf{f}}_{\epsilon,d} : \mathbb{R} \to \mathbb{R}^m$ is injective.

$$\begin{aligned}
\tilde{\mathbf{f}}_{\epsilon,d}(s_d) &= \mathbf{J}^{(0)}_{:,d} s_d \tilde{\mathbb{1}}_\epsilon(t_d - s_d) + (\mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \\
&= \mathbf{J}^{(0)}_{:,d} s_d (1 - \tilde{\mathbb{1}}_\epsilon(s_d - t_d)) + (\mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \\
&\quad \because \tilde{\mathbb{1}}_\epsilon(s_d) + \tilde{\mathbb{1}}_\epsilon(-s_d) = 1 \\
&= \mathbf{J}^{(0)}_{:,d}(s_d - (s_d - t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d)) + \mathbf{J}^{(0)}_{:,d}(s_d - t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \\
&= [\mathbf{J}^{(0)}_{:,d}\ \mathbf{J}^{(0)}_{:,d}] \begin{bmatrix} s_d - (s_d - t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \\ (s_d - t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \end{bmatrix} \qquad (26)
\end{aligned}$$

Define $\mathbf{t}_d : \mathbb{R} \to \mathbb{R}^d$ such that $\mathbf{t}_d(s_d) = \begin{bmatrix} s_d - (s_d - t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \\ (s_d - t_d)\tilde{\mathbb{1}}_\epsilon(s_d - t_d) \end{bmatrix}$. To show that $\tilde{\mathbf{f}}_{\epsilon,d} : \mathbb{R} \to \mathbb{R}^m$ is injective, we need to show the following:

$$\forall s_d^{(1)}, s_d^{(2)} \in \mathbb{R} : \ \tilde{\mathbf{f}}_{\epsilon,d}(s_d^{(1)}) = \tilde{\mathbf{f}}_{\epsilon,d}(s_d^{(2)}) \implies s_d^{(1)} = s_d^{(2)} \qquad (27)$$

As usual, we show (27) by contradiction. Let

$$\exists s_d^{(1)} \neq s_d^{(2)} \in \mathbb{R} \text{ s.t. } \tilde{\mathbf{f}}_{\epsilon,d}(s_d^{(1)}) = \tilde{\mathbf{f}}_{\epsilon,d}(s_d^{(2)}) \qquad (28)$$

then we deduce

$$\begin{aligned}
\tilde{\mathbf{f}}_{\epsilon,d}(s_d^{(1)}) &= \tilde{\mathbf{f}}_{\epsilon,d}(s_d^{(2)}) \\
[\mathbf{J}^{(0)}_{:,d}\ \mathbf{J}^{(0)}_{:,d}]\mathbf{t}(s_d^{(1)}) &= [\mathbf{J}^{(0)}_{:,d}\ \mathbf{J}^{(0)}_{:,d}]\mathbf{t}(s_d^{(2)}) \\
\implies \mathbf{t}(s_d^{(1)}) &= \mathbf{t}(s_d^{(2)}) \\
&\quad \because [\mathbf{J}^{(0)}_{:,d}\ \mathbf{J}^{(0)}_{:,d}] \text{ is full column rank, Lemma B.2} \\
\begin{bmatrix} s_d^{(1)} - (s_d^{(1)} - t_d)\tilde{\mathbb{1}}_\epsilon(s_d^{(1)} - t_d) \\ (s_d^{(1)} - t_d)\tilde{\mathbb{1}}_\epsilon(s_d^{(1)} - t_d) \end{bmatrix} &= \begin{bmatrix} s_d^{(2)} - (s_d^{(2)} - t_d)\tilde{\mathbb{1}}_\epsilon(s_d^{(2)} - t_d) \\ (s_d^{(2)} - t_d)\tilde{\mathbb{1}}_\epsilon(s_d^{(2)} - t_d) \end{bmatrix} \\
\implies s_d^{(1)} &= s_d^{(2)}
\end{aligned}$$

Hence, we arrive at a contradiction to (28). Thereby, $\tilde{\mathbf{f}}_{\epsilon,d} : \mathbb{R} \to \mathbb{R}^m$ is injective.

Hence, we have shown that the coordinate-wise functions of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$, $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{i=1}^d \tilde{\mathbf{f}}_{\epsilon,i}(s_i)$ where $\mathbf{s} \in \mathbb{R}^d = (s_1, s_2, ..., s_d)$, given by $\tilde{\mathbf{f}}_{\epsilon,i} : \mathbb{R} \to \mathbb{R}^m$ are injective. We now show the above statement implies the injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d f_k(s_k)$. Observe that by definition B.8

1. $\mathbf{S}_1$ = span(Im($\mathbf{f}_1$)) = span($\mathbf{J}^{(0)}_{:,1}$) = span($\mathbf{J}^{(1)}_{:,1}$)
2. $\mathbf{S}_2$ = span(Im($\mathbf{f}_2$)) = span($\mathbf{J}^{(0)}_{:,2}$) = span($\mathbf{J}^{(1)}_{:,2}$)

   $\vdots$
3. $\mathbf{S}_{d-1}$ = span(Im($\mathbf{f}_{d-1}$)) = span($\mathbf{J}^{(0)}_{:,d-1}$) = span($\mathbf{J}^{(1)}_{:,d-1}$)
4. $\mathbf{S}_d$ = span(Im($\mathbf{f}_d$)) = span($\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1}$)

Consider $\mathbb{V} = \text{span}(\text{Im}(\mathbf{f}))$. By Lemma B.10, $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$. Further, by Lemma B.11, injectivity of $\mathbf{f}_i : \mathbb{R}^d \to \mathbb{R}^m \forall i \in [d]$ implies injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$.

Continuity of derivatives of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$

Consider the derivatives of $\tilde{\mathbf{f}}_\epsilon(\mathbf{s})$ with respect to the coordinates of $\mathbf{s} = (s_1, s_2, ..., s_d)$.

1. By definition B.14, $\tilde{\mathbf{f}}_{\epsilon,i}(s_i) = \mathbf{J}^{(0)}_{:,i} s_i = \mathbf{J}^{(1)}_{:,i} s_i \forall i \in \{1, 2, ..., (d-1)\}$. Therefore, $\frac{\partial \tilde{\mathbf{f}}_\epsilon}{\partial s_i} = \mathbf{J}^{(0)}_{:,i} = \mathbf{J}^{(1)}_{:,i} \forall i \in \{1, 2, ..., (d-1)$. Thus, the derivatives $\frac{\partial \tilde{f}_{\epsilon,j}}{\partial s_i} \forall j \in [m], i \in \{1, 2, ..., (d-1)\}$ are continuous.
2. Consider the derivative of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ with respect to s_d. By (26),

$$\tilde{\mathbf{f}}_{\epsilon,d}(s_d) = [\mathbf{J}^{(0)}_{:,d} \ \mathbf{J}^{(0)}_{:,d}] \begin{bmatrix} s_d - (s_d - t_d)\tilde{1}_\epsilon(s_d - t_d) \\ (s_d - t_d)\tilde{1}_\epsilon(s_d - t_d) \end{bmatrix}$$

$$\tilde{\mathbf{f}}'_{\epsilon,d}(s_d) = [\mathbf{J}^{(0)}_{:,d} \ \mathbf{J}^{(0)}_{:,d}] \begin{bmatrix} s_d - \tilde{1}_\epsilon(s_d - t_d) - (s_d - t_d)\tilde{1}'_\epsilon(s_d - t_d) \\ \tilde{1}_\epsilon(s_d - t_d) + (s_d - t_d)\tilde{1}'_\epsilon(s_d - t_d) \end{bmatrix}$$

   where by definition B.13 $\tilde{1}'_\epsilon(s) = \begin{cases} 0 & s \leq -\epsilon \\ \frac{1}{2}\cos\left(\frac{\pi s}{2\epsilon}\right)\frac{\pi}{2\epsilon} & -\epsilon < s \leq \epsilon \\ 0 & s > \epsilon \end{cases}$.

   Notice that $\tilde{1}'_\epsilon : \mathbb{R} \to \mathbb{R}$ is continuous in $\mathbb{R}$. $\tilde{\mathbf{f}}'_{\epsilon,d}(s_d)$ is continuous since it is composed by a sum and product of continuous functions (Theorem 4.9, (Rudin et al., 1964)). Notice also that the term $(s_d - t_d)\tilde{1}'_\epsilon(s_d - t_d) = \frac{1}{2}\cos\left(\frac{\pi(s_d - t_d)}{2\epsilon}\right)\frac{\pi}{2\epsilon}.(s_d - t_d)$ is non-zero only when $-\epsilon < (s_d - t_d) \leq \epsilon$, hence this term is finite even for $\epsilon > 0$ arbitrarily small. The other terms in $\tilde{\mathbf{f}}'_{\epsilon,d}(s_d)$ are also finite be definition. Thus, the derivatives $\frac{\partial \tilde{f}_{\epsilon,j}}{\partial s_d} \forall j \in [m]$ are continuous for $\epsilon > 0$ arbitrarily small.

Since all the partial derivatives of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ are continuous, $\tilde{\mathbf{f}}$ is continuously differentiable (Theorem 9.21, (Rudin et al., 1964)).

□

We now present the theorem that introduces a bound on the global IMA contrast for non-affine maps, $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$, composed by smoothly joining two affine maps with local bases sampled isotropically as defined here B.14.

Theorem B.16. *Consider the map $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ sampled randomly from the procedure B.14 and any finite probability density, $p_\mathbf{s}$, defined over $\mathbb{R}^d$.*

Then, for $\epsilon > 0$ arbitrarily small, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s}) \leq \delta$ with (high) probability $\geq 1 - \min\left\{1, \exp(2\log d - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ for $m \gg d$ where $\delta < \frac{1}{2}$ is arbitrarily small.

Proof. We show that the condition of Theorem 5.1, the columns of the Jacobian of $\tilde{\mathbf{f}}_\epsilon$ are locally sampled isotropically i.e. , is still satisfied for the domain of $\tilde{\mathbf{f}}_\epsilon$, i.e. $\forall \mathbf{s} \in \mathbb{R}^d$ almost surely w.r.t finite probability measure, $p_\mathbf{s}$ over $\mathbb{R}^d$.

Following from B.14, consider the following partition of the domain,

$$\begin{cases} \mathbf{s} \in \mathbb{P}^{(0)} & \iff \mathbf{w}^\top \mathbf{s} \leq c - \epsilon \\ \mathbf{s} \in \mathbb{P}^{(1)} & \iff \mathbf{w}^\top \mathbf{s} > c + \epsilon \\ \mathbf{s} \in \mathbb{B} & \iff c - \epsilon < \mathbf{w}^\top \mathbf{s} \leq c + \epsilon \end{cases}.$$

1. $\forall \mathbf{s} \in \mathbb{P}^{(0)}, \mathbf{J_f}(\mathbf{s}) = \mathbf{J}^{(0)},$ with $\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(0)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,d} \overset{i.i.d}{\sim} p_\mathbf{r}$.

2. $\forall \mathbf{s} \in \mathbb{P}^{(1)}, \mathbf{J_f}(\mathbf{s}) = \mathbf{J}^{(1)},$ with $\mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(1)}_{:,d} \overset{i.i.d}{\sim} p_\mathbf{r}$.

3. The region $\mathbb{B}$ sandwiching the boundary of the partitions has arbitrarily small probability measure since:

   (a) $\mathbb{B}$ is an ϵ-sandwich of a $(d-1)$-dimensional subspace of a d-dimensional domain. The Lebesgue measure on $\mathbb{B}$ is equal to the volumne element associated with $\mathbb{B}$ (3.3, (Çinlar, 2011)), thus, $\lambda(\mathbb{B}) = \Theta(\epsilon)$[8] where $\lambda(.)$ denotes the Lebesgue measure.

   (b) $p_\mathbf{s}$ is finite at all points.

   Hence, $p(\mathbb{B}) = \int_\mathbb{B} p_\mathbf{s} \lambda(\mathbf{s}) = \Theta(\epsilon)$, is arbitrarily small for suitably chosen ϵ.

   To derive a bound on the global IMA contrast of $\tilde{\mathbf{f}}_\epsilon$, $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$, we need that in region, $\forall \mathbf{s} \in \mathbb{B}$, the value of the local IMA contrast $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})$ is finite. This is equivalent to showing that the Jacobian, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ is full column-rank for all $\mathbf{s} \in \mathbb{B}$. To show this, consider the alternate definition of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ in terms of coordinate-wise functions (Definition B.14), $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{i=1}^d \tilde{\mathbf{f}}_{\epsilon,i}(s_i)$ where $\mathbf{s} \in \mathbb{R}^d = (s_1, s_2, ..., s_d)$:

   (a) $\tilde{\mathbf{f}}_{\epsilon,i}(s_i) = \mathbf{J}^{(0)}_{:,i} s_i = \mathbf{J}^{(1)}_{:,i} s_i \forall i \in \{1, 2, ..., (d-1)\}$

   (b) $\tilde{\mathbf{f}}_{\epsilon,d}(s_d) = \mathbf{J}^{(0)}_{:,d} s_d \tilde{1}_\epsilon(t_d - s_d) + (\mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d) \tilde{1}_\epsilon(s_d - t_d)$

   for $t_d \in \mathbb{R}$

   From the above definition, we see that the first $(d-1)$ columns of the Jacobian, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ are defined as $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,i} = \frac{\partial \tilde{\mathbf{f}}_\epsilon(\mathbf{s})}{\partial s_i} = \mathbf{J}^{(0)}_{:,i} = \mathbf{J}^{(0)}_{:,i}$ for $i \in \{1, 2, ..., d-1\}$. By Lemma B.2, the first $(d-1)$ columns of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ are nonzero and linearly independent with probability 1. Consider the d-th column of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$.

   $$\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,d} = \mathbf{J}^{(0)}_{:,d} \tilde{1}_\epsilon(t_d - s_d) + \mathbf{J}^{(1)}_{:,d} \tilde{1}_\epsilon(s_d - t_d) - \mathbf{J}^{(0)}_{:,d} s_d \tilde{1}'_\epsilon(t_d - s_d) \\ + (\mathbf{J}^{(1)}_{:,d}(s_d - t_d) + \mathbf{J}^{(0)}_{:,d} t_d) \tilde{1}'_\epsilon(s_d - t_d)$$

   thus

   $$\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,d} = \mathbf{J}^{(0)}_{:,d}((t_d - s_d)\tilde{1}'_\epsilon(t_d - s_d) \\ + \tilde{1}_\epsilon(t_d - s_d)) + \mathbf{J}^{(1)}_{:,d}((s_d - t_d)\tilde{1}'_\epsilon(s_d - t_d) \\ + \tilde{1}_\epsilon(s_d - t_d))$$

   Observe that $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,d}$ is a linear combination of $\mathbf{J}^{(0)}_{:,d}$ and $\mathbf{J}^{(1)}_{:,d}$. Since by Lemma B.2, $\mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}$ are nonzero and linearly independent with respect to each other and $\mathbf{J}^{(0)}_{:,i}, \mathbf{J}^{(1)}_{:,i}$ $\forall i \in \{1, 2, ..., (d-1)\}$, the only possibility for $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ to not be full column-rank is

[8] Refer to big theta notation $\Theta(.)$ here.

$$\begin{aligned}
&\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,d} = \mathbf{0} \\
&\quad\implies (t_d - s_d)\tilde{1}'_\epsilon(t_d - s_d) + \tilde{1}_\epsilon(t_d - s_d) = (s_d - t_d)\tilde{1}'_\epsilon(s_d - t_d) + \tilde{1}_\epsilon(s_d - t_d) = 0 \\
&\qquad\quad \because \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d} \text{ are linearly independent.}
\end{aligned}$$

Consider the function, $q : \mathbb{R} \to \mathbb{R}$ such that $q(s) = s\tilde{1}'_\epsilon(s) + \tilde{1}_\epsilon(s)$. Observe that $q(s) \geq 0$ for $s \geq 0$. Thus, for $(t_d - s_d)\tilde{1}'_\epsilon(t_d - s_d) + \tilde{1}_\epsilon(t_d - s_d) = (s_d - t_d)\tilde{1}'_\epsilon(s_d - t_d) + \tilde{1}_\epsilon(s_d - t_d) = 0$, we need that $s_d = t_d$. At $s_d = t_d$, $q(s_d - t_d) = q(t_d - s_d) = \frac{1}{2} \neq 0$. Hence, we have shown that $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,d} \neq \mathbf{0}$, thereby $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ is full column-rank and $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})$ is finite for all $\mathbf{s} \in \mathbb{B}$.

Hence,

$$\begin{aligned}
C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}}) &= \int_{\mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} \\
&= \int_{\mathbb{P}^{(0)}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \int_{\mathbb{P}^{(1)}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \int_{\mathbb{B}} c_{\text{IMA}}\; p_{\mathbf{s}} d\mathbf{s} \\
&= \int_{\mathbb{P}^{(0)}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \int_{\mathbb{P}^{(1)}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \Theta(\epsilon) \\
&\approx \int_{\mathbb{P}^{(0)}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \int_{\mathbb{P}^{(1)}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} \quad \text{for } \epsilon \text{ arbitrarily small.} \\
&\leq \max_{\mathbf{s} \in \mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s}) \int_{\mathbb{R}^d} p_{\mathbf{s}} d\mathbf{s} \\
&\leq \max_{\mathbf{s} \in \mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s}) \leq \delta \quad \text{w. p. } \geq 1 - \min\left\{1, \exp(2logd - \kappa(m-1)\frac{\delta^2}{d^2})\right\}
\end{aligned}$$

by Theorem 5.1.

Thus, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}}) \leq \delta$ for $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.14 with (high) probability $\geq 1 - \min\left\{1, \exp(2logd - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ for $m \gg d$ where $\delta < \frac{1}{2}$ is arbitrarily small.

□

Defining non-linear functions by gluing 2^d affine functions We generalize the above construct for the map, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, where $\mathbf{f}$ was constructed by stitching together *two* affine functions. In the following construct, the domain $\mathbb{R}^d$ is split into 2^d axis-aligned parts using one split point per coordinate. The map $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is now constructed by stitching together 2^d affine functions defined on each part of the domain, mapping the domain to more complex manifolds embedded in the observation space $\mathbb{R}^m$.

Definition B.17 (Maps defined as a composition of 2^d affine maps on an axis-aligned domain partition)**.** *The maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ we consider are defined as follows:*

1. *Consider the map, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ applied to $\mathbf{s} \in \mathbb{R}^d$.*

2. *For any $\mathbf{t} = (t_1, t_2, ..., t_d) \in \mathbb{R}^d$,a partition of the domain of $\mathbf{f}$ is defined by the binary vector, $\mathbf{b} : \mathbb{R}^d \to \{0,1\}^d$ where $\mathbf{b}_k(\mathbf{s}) := 1_{s_k > t_k}$, $\mathbb{R}^d = \mathbb{P}_{\mathbb{R}^d} = \bigcup_{\mathbf{b} \in \{0,1\}^d} \mathbb{P}^{(\mathbf{b})}$, where $\mathbb{P}^{(\mathbf{b})} := \{\mathbf{s} \quad | \quad \mathbf{b}(\mathbf{s}) = \mathbf{b}\}$. Note that the partition defined is axis-aligned to the canonical basis in $\mathbb{R}^d$. This follows to extend the continuity argument from the two-partition case in Lemma B.6, observation B.7.*

3. *Consider the two matrices, $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m \times d}$, used to define the Jacobian in each part, $\mathbb{P}^{(\mathbf{b})}$ $\forall \mathbf{b} \in \{0,1\}^d$. Sample the columns of $\mathbf{J}^{(0)}, \mathbf{J}^{(1)}$ independently from the mentioned spherically symmetric distribution (5.1) $p_{\mathbf{r}}$, $\mathbf{J}^{(0)}_1, \mathbf{J}^{(0)}_2, ..., \mathbf{J}^{(0)}_d \overset{i.i.d}{\sim} p_{\mathbf{r}}$, $\mathbf{J}^{(1)}_1, \mathbf{J}^{(1)}_2, ..., \mathbf{J}^{(1)}_d \overset{i.i.d}{\sim} p_{\mathbf{r}}$.*

4. *For* $\mathbf{s} \in \mathbb{R}^d$ *with* $\mathbf{b}(\mathbf{s}) = \mathbf{b} \in \{0,1\}^d$, $\mathbf{J_f}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}$ *such that* $\begin{Bmatrix} \mathbf{J}^{(\mathbf{b})}_{:,k} = \mathbf{J}^{(0)}_{:,k} & b_k = 0 \\ \mathbf{J}^{(\mathbf{b})}_{:,k} = \mathbf{J}^{(1)}_{:,k} & b_k = 1 \end{Bmatrix}$.

   Note that this corresponds to the observation B.7 where changing one column of $\mathbf{J_f}(\mathbf{s})$ *across a partition of the domain results in axis-aligned partitions.*

5. $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *is defined as:* $\left\{ \mathbf{f}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}(\mathbf{s}) + \mathbf{c}^{(\mathbf{b})} \middle| \quad \mathbf{b}(\mathbf{s}) = \mathbf{b} \quad \right\}$, *where* $\mathbf{c}^{(\mathbf{b})} \in \mathbb{R}^m \; \forall \; \mathbf{b} \in \{0,1\}^d$.

6. *We show that owing to axis-alignment of chosen partition,* $\mathbb{P}_{\mathbb{R}^d}$ *and resampling exactly one column of the Jacobian,* $\mathbf{J_f}(\mathbf{s})$ *at the boundary between two parts,* $\mathbf{f}$ *can be written as a product of submanifolds. Consider the functions* $\mathbf{f}_k(s_k) = \begin{Bmatrix} \mathbf{J}^{(0)}_{:,k} s_k & s_k \leq t_k \\ \mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k & s_k > t_k \end{Bmatrix}$ *which only act on one coordinate of the input,* $\mathbf{s} = (s_1, s_2, ..., s_d) \in \mathbb{R}^d$.

   Consider again $\left\{ \mathbf{f}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}(\mathbf{s}) + \mathbf{c}^{(\mathbf{b})} \middle| \quad \mathbf{b}(\mathbf{s}) = \mathbf{b} \quad \right\}$, *where* $\mathbf{c}^{(\mathbf{b})} \in \mathbb{R}^m \; \forall \; \mathbf{b} \in \{0,1\}^d$, $\mathbf{c}^{(\mathbf{0})} = \mathbf{0}, \mathbf{c}^{(\mathbf{b})}, \mathbf{b} \neq \mathbf{0}$ *are completely specified by* $\mathbf{t}$. $\mathbf{f}(\mathbf{s})$ *can be equivalently written as* $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} \mathbf{f}_k(s_k)$. *Here, we write* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *as a product of the submanifolds, or also referred to in latter parts of the note as a sum of the coordinate-wise functions,* $\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^m$.

In the following results, we show that the Jacobian of maps defined by stitching together 2^d affine functions is well-defined at all points in the domain. We start by showing that maps defined in B.17 are continuous and injective. We then define a smooth approximation to B.17, so that the map is differentiable also at the partition boundary given by the partition of domain defined in $\mathbb{P}_{\mathbb{R}^d}$. Futher, we show that the smooth approximation to B.17 is continuous, injective and continuously differentiable, which ensures that the Jacobian is well-defined at all points in the domain and the IMA contrast can be computed. Finally, we present a theorem which bounds the IMA contrast with high probability as the dimension of the observed space grows.

Lemma B.18 (Continuity of composition of 2^d affine maps)**.** *Consider maps* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *as defined in B.17. Such a map* $\mathbf{f}$ *is continuous.*

Proof. Consider $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} f_k(s_k)$, where

$$\mathbf{f}_k(s_k) = \begin{cases} \mathbf{J}^{(0)}_{:,k} s_k \,, & s_k \leq t_k \,, \\ \mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k \,, & s_k > t_k \,. \end{cases}$$

We show continuity of $\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^m \; \forall k \in [d]$. For a particular k, consider the cases:

1. $s_k < t_k$:

   $\mathbf{f}_k(s_k) = \mathbf{J}^{(0)}_{:,k} s_k$ is affine in the entire region and hence, is continuous.

2. $s_k = t_k$:

   $\mathbf{f}_k(s_k) = \mathbf{J}^{(0)}_{:,k} s_k$, and in limit equal to $\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k$. For $s_k = t_k$, we need to show that $\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k$ converges to $\mathbf{J}^{(0)}_{:,k} s_k$. This is easily seen by substituting $s_k = t_k$, $\mathbf{J}^{(0)}_{:,k} t_k = \mathbf{J}^{(1)}_{:,k}(t_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k$. $\mathbf{f}_k(s_k)$ is continuous at $s_k = t_k$ (Theorem 4.6, (Rudin et al., 1964)).

3. $s_k > t_k$:

   $\mathbf{f}_k(s_k) = \mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k$ is affine in the entire region and hence, is continuous.

$\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, \mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} f_k(s_k)$ is continuous since the sum of continuous functions is continuous (Theorem 4.9, (Rudin et al., 1964)).

□

To show injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.17, we follow a similar approach as in the case of maps defined by joining two affine functions (Definition B.8, Lemma B.12). We start by showing that the images of the coordinate-wise functions, $\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^m$ which compose $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, \mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} \mathbf{f}_k(s_k)$, are in direct sum with respect to the image of $\mathbf{f}$. Then, we show that the coordinate-wise functinos, $\mathbf{f}_k$, are injective. Finally, we use that in this scenario the injectivity of the coordinate-wise functions, $\mathbf{f}_k$, implies the injectiivity of $\mathbf{f}$ to conclude that $\mathbf{f}$ is injective (Lemma B.11).

Lemma B.19. *Consider $\mathbf{J}^{(0)}, \mathbf{J}^{(1)} \in \mathbb{R}^{m\times d}$ as sampled in Definition (B.17). The vector space $\mathbb{V} = span\{\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(0)}_{:,2}, \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,(d-1)}, \mathbf{J}^{(1)}_{:,(d-1)}, \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}$ is the direct sum of the family $\mathcal{F} = \{\mathbf{S}_1 = span\{\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1}\}, \mathbf{S}_2 = span\{\mathbf{J}^{(0)}_{:,2}, \mathbf{J}^{(1)}_{:,2}\}, ..., \mathbf{S}_{d-1} = span\{\mathbf{J}^{(0)}_{:,d-1}, \mathbf{J}^{(1)}_{:,d-1}\}, \mathbf{S}_d = span\{\mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}\}$.*

Proof. From Lemma B.2, for the scenario $m \gg d$ (here it is sufficient to have $(m > 2d)$), the vectors $\{\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(0)}_{:,2}, \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,(d-1)} = \mathbf{J}^{(1)}_{:,(d-1)}, \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}\}$ are non-zero and linearly independent with probability 1.

Consider $\mathbf{v} \in \mathbb{V}, \mathbf{u}_1, \mathbf{w}_1 \in \mathbf{S}_1, \mathbf{u}_2, \mathbf{w}_2 \in \mathbf{S}_2, ..., \mathbf{u}_d, \mathbf{w}_d \in \mathbf{S}_d$ such that

$$\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_d, \mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_d$$

By definition B.9, to show $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$, we need to show $\mathbf{u}_1 = \mathbf{w}_1, \mathbf{u}_2 = \mathbf{w}_2, ..., \mathbf{u}_d = \mathbf{w}_d$.

Let

- $\mathbf{u}_1 = c_1^{(0)}\mathbf{J}^{(0)}_{:,1} + c_1^{(1)}\mathbf{J}^{(1)}_{:,1}, \mathbf{w}_1 = c_1^{(0)'}\mathbf{J}^{(0)}_{:,1} + c_1^{(1)'}\mathbf{J}^{(1)}_{:,1}$
- $\mathbf{u}_2 = c_2^{(0)}\mathbf{J}^{(0)}_{:,2} + c_2^{(1)}\mathbf{J}^{(1)}_{:,2}, \mathbf{w}_2 = c_2^{(0)'}\mathbf{J}^{(0)}_{:,2} + c_2^{(1)'}\mathbf{J}^{(1)}_{:,2}$

  $\vdots$
- $\mathbf{u}_1 = c_{d-1}^{(0)}\mathbf{J}^{(0)}_{:,d-1} + c_{d-1}^{(1)}\mathbf{J}^{(1)}_{:,d-1}, \mathbf{w}_1 = c_{d-1}^{(0)'}\mathbf{J}^{(0)}_{:,d-1} + c_{d-1}^{(1)'}\mathbf{J}^{(1)}_{:,d-1}$
- $\mathbf{u}_d = c_d^{(0)}\mathbf{J}^{(0)}_{:,d} + c_d^{(1)}\mathbf{J}^{(1)}_{:,d}, \mathbf{w}_d = c_d^{(0)'}\mathbf{J}^{(0)}_{:,d} + c_d^{(1)'}\mathbf{J}^{(1)}_{:,d}$

$$\begin{aligned}
&\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_d = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_d\\
&(c_1^{(0)}\mathbf{J}^{(0)}_{:,1} + c_1^{(1)}\mathbf{J}^{(1)}_{:,1}) + (c_2^{(0)}\mathbf{J}^{(0)}_{:,2} + c_2^{(1)}\mathbf{J}^{(1)}_{:,2}) = (c_1^{(0)'}\mathbf{J}^{(0)}_{:,1} + c_1^{(1)'}\mathbf{J}^{(1)}_{:,1}) + (c_2^{(0)'}\mathbf{J}^{(0)}_{:,2} + c_2^{(1)'}\mathbf{J}^{(1)}_{:,2})\\
&+ ... + (c_d^{(0)}\mathbf{J}^{(0)}_{:,d} + c_d^{(1)}\mathbf{J}^{(1)}_{:,d}) \qquad\qquad\qquad\qquad + ... + (c_d^{(0)'}\mathbf{J}^{(0)}_{:,d} + c_d^{(1)'}\mathbf{J}^{(1)}_{:,d})\\
&(c_1^{(0)} - c_1^{(0)'})\mathbf{J}^{(0)}_{:,1} + (c_2^{(0)} - c_2^{(0)'})\mathbf{J}^{(0)}_{:,2} + ... + (c_d^{(0)} - c_d^{(0)'})\mathbf{J}^{(0)}_{:,d}\\
&+ (c_1^{(1)} - c_1^{(1)'})\mathbf{J}^{(1)}_{:,1} + (c_2^{(1)} - c_2^{(1)'})\mathbf{J}^{(1)}_{:,2} + ... + (c_d^{(1)} - c_d^{(1)'})\mathbf{J}^{(1)}_{:,d} = \mathbf{0}
\end{aligned}$$

Since $\{\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(0)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(1)}_{:,d}\}$ are nonzero and linearly independent with probability 1 (Lemma B.2),

$$\begin{aligned}
(c_1^{(0)} - c_1^{(0)'}) = (c_2^{(0)} - c_2^{(0)'}) = ... = (c_d^{(0)} - c_d^{(0)'})\\
= (c_1^{(1)} - c_1^{(1)'}) = (c_2^{(1)} - c_2^{(1)'}) = ... = (c_d^{(1)} - c_d^{(1)'}) = 0
\end{aligned}$$

Hence, it follows that $\mathbf{u}_1 = \mathbf{w}_1, \mathbf{u}_2 = \mathbf{w}_2, ..., \mathbf{u}_d = \mathbf{w}_d$ and $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$.

□

Lemma B.20 (Injectivity of composition of 2^d affine maps)**.** *Consider maps $\mathbf{f}:\mathbb{R}^d \to \mathbb{R}^m$ as defined in B.17. Such a map $\mathbf{f}$ is injective.*

Proof. Consider $\mathbf{f}:\mathbb{R}^d \to \mathbb{R}^m$ written as a sum of coordinate-wise functions (Definition B.17), $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} f_k(s_k)$, where

$$\mathbf{f}_k(s_k) = \begin{cases} \mathbf{J}^{(0)}_{:,k} s_k\,, & s_k \le t_k\,, \\ \mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k\,, & s_k > t_k\,. \end{cases}$$

In the following proof, we show injectivity of the coordinate-wise functions $\mathbf{f}_k:\mathbb{R}\to\mathbb{R}^m\ \forall k \in [d]$ and conclude that $\mathbf{f}:\mathbb{R}^d\to\mathbb{R}^m$ is injective by Lemma B.19 and Lemma B.11. For a particular k, to show that $\mathbf{f}_k:\mathbb{R}\to\mathbb{R}^m$ is injective, we need to show the following:

$$\forall s_k^{(1)}, s_k^{(2)} \in \mathbb{R}: \quad \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)} \tag{29}$$

As usual, we show (B.27) by contradiction. Let

$$\exists s_k^{(1)} \neq s_k^{(2)} \in \mathbb{R} s.t. \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \tag{30}$$

Consider the mutually exclusive, and exhaustive cases:

1. $s_k^{(1)} \neq s_k^{(2)} \in \mathbb{R}, s_k^{(1)}, s_k^{(2)} \le t_k$

$$\begin{aligned} \mathbf{f}_k(s_k^{(1)}) &= \mathbf{f}_k(s_k^{(2)}) \\ \mathbf{J}^{(0)}_{:,k} s_k^{(1)} &= \mathbf{J}^{(0)}_{:,k} s_k^{(2)} \\ \implies s_k^{(1)} &= s_k^{(2)} && \mathbf{J}^{(0)} \neq \mathbf{0} \text{ w. p. } 1 \end{aligned}$$

Thus, for $s_k^{(1)}, s_k^{(2)} \le t_k$, $\mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)}$, which contradicts (30).

2. $s_k^{(1)} \neq s_k^{(2)} \in \mathbb{R}, s_k^{(1)}, s_k^{(2)} > t_k$

$$\begin{aligned} \mathbf{f}_k(s_k^{(1)}) &= \mathbf{f}_k(s_k^{(2)}) \\ \mathbf{J}^{(1)}_{:,k}(s_k^{(1)} - t_k) + \mathbf{J}^{(0)}_{:,k} t_k &= \mathbf{J}^{(1)}_{:,k}(s_k^{(2)} - t_k) + \mathbf{J}^{(0)}_{:,k} t_k \\ \mathbf{J}^{(1)}_{:,k} s_k^{(1)} &= \mathbf{J}^{(1)}_{:,k} s_k^{(2)} \\ \implies s_k^{(1)} &= s_k^{(2)} && \mathbf{J}^{(1)} \neq \mathbf{0} \text{ w. p. } 1 \end{aligned}$$

Thus, for $s_k^{(1)}, s_k^{(2)} > t_k$, $\mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)}$, which contradicts (30).

3. $s_k^{(1)} \neq s_k^{(2)} \in \mathbb{R}, s_k^{(1)} \le t_k, s_k^{(2)} > t_k$

$$\begin{aligned} \mathbf{f}_k(s_k^{(1)}) &= \mathbf{f}_k(s_k^{(2)}) \\ \mathbf{J}^{(0)}_{:,k} s_k^{(1)} &= \mathbf{J}^{(1)}_{:,k}(s_k^{(2)} - t_k) + \mathbf{J}^{(0)}_{:,k} t_k \\ \begin{bmatrix} \mathbf{J}^{(0)}_{:,k} - \mathbf{J}^{(1)}_{:,k} \end{bmatrix} \begin{bmatrix} s_k^{(1)} - t_k \\ s_k^{(2)} - t_k \end{bmatrix} &= \mathbf{0} \\ \implies s_k^{(1)} = s_k^{(2)} &= t_k \quad \because \begin{bmatrix} \mathbf{J}^{(0)}_{:,k} - \mathbf{J}^{(1)}_{:,k} \end{bmatrix} \text{ is full column-rank w.p. } 1 \end{aligned}$$

We arrive at a contradiction since $s_k^{(2)} > t_k$.

$\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^m \; \forall k$ is injective for the exhaustive cases for $s_k^{(1)}, s_k^{(2)} \in \mathbb{R}$ in the above-mentioned points and hence, is injective.

We now show that the injectivity of the coordinate-wise functions $\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^d$ implies the injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} \mathbf{f}_k(s_k)$. Observe that by definition B.17

1. $\mathbf{S}_1 = \text{span}(\text{Im}(\mathbf{f}_1)) = \text{span}(\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1})$
2. $\mathbf{S}_2 = \text{span}(\text{Im}(\mathbf{f}_2)) = \text{span}(\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1})$

   $\vdots$
3. $\mathbf{S}_d = \text{span}(\text{Im}(\mathbf{f}_d)) = \text{span}(\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1})$

Consider $\mathbb{V} = \text{span}(\text{Im}(\mathbf{f}))$. By Lemma B.19, $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$. Further, by Lemma B.11, injectivity of $\mathbf{f}_k : \mathbb{R}^d \to \mathbb{R}^m \forall k \in [d]$ implies injectivity of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$.

□

We proceed to define a smooth approximation to the map, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.17.

Definition B.21. *Consider the decomposition of* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *as a sum of coordinate-wise functions,* $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^{d} f_k(s_k)$*,* $\mathbf{s} = (s_1, s_2, ..., s_d) \in \mathbb{R}^d$ *where* $\mathbf{f}_k(s_k) = \begin{Bmatrix} \mathbf{J}^{(0)}_{:,k} s_k & s_k \leq t_k \\ \mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k & s_k > t_k \end{Bmatrix}$. $\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^m$ *can be alternatively written as:*

$$\mathbf{f}_k(s_k) = (\mathbf{J}^{(0)}_{:,k} s_k) 1_{s_k \leq t_k} + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k) 1_{s_k > t_k}$$

We define the smoothened version of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *as* $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^{d} \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$ *for* $\epsilon > 0$ *arbitrarily small where*

$$\tilde{\mathbf{f}}_{\epsilon,k}(s_k) = (\mathbf{J}^{(0)}_{:,k} s_k) \tilde{1}_\epsilon(t_k - s_k) + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k) \tilde{1}_\epsilon(s_k - t_k)$$

As with the case of defining maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ by smoothly joining *two* affine maps (Definition B.14, Lemma B.15), before we present the theorem with the high probability bound on the global IMA contrast, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$ for any finite probability density, $p_\mathbf{s}$ on $\mathbb{R}^d$, we introduce a lemma to show that maps $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.14 are continuous, injective and continuously differentiable. The objective of the following lemma is to ensure that the Jacobian of $\tilde{\mathbf{f}}_\epsilon$, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon} \in \mathbb{R}^{m \times d}$, is well-defined for at all points in the domain of $\tilde{\mathbf{f}}_\epsilon$ such that the IMA contrast, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$ can be computed for maps, $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$, can be computed with respect to all finite distribuitions, $p_\mathbf{s}$ on $\mathbb{R}^d$.

Lemma B.22. *Functions* $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ *defined in B.21 are continuously differentiable in* $\mathbb{R}^d$*, in addition to being continuous and injective, with* $\epsilon > 0$ *arbitrarily small.*

Proof. We show that $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.21 is continuous, injective and continuously differentiable.

Continuity of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$

Consider the coordinate-wise decomposition of $\tilde{\mathbf{f}}_\epsilon$; $\tilde{f}_{\epsilon,i}(\mathbf{s}) = \sum_{k=1}^{d} \tilde{f}_{\epsilon,(i,k)}(s_k)$, where $\tilde{f}_{\epsilon,(i,k)} : \mathbb{R} \to \mathbb{R}$ is defined as $\tilde{f}_{\epsilon,(i,k)}(s_k) := (\mathbf{J}^{(0)}_{i,k} s_k) \tilde{1}_\epsilon(t_k - s_k) + (\mathbf{J}^{(1)}_{i,k}(s_k - t_k) + \mathbf{J}^{(0)}_{i,k} t_k) \tilde{1}_\epsilon(s_k - t_k)$

$(\mathbf{J}^{(0)}_{m,k} s_k)$, $(\mathbf{J}^{(1)}_{m,k}(s_k - t_k) + \mathbf{J}^{(0)}_{m,k} t_k)$ are continuous in $s_k \in \mathbb{R} \; \forall k \in [d]$ since they are affine.

Note that $\tilde{1} : \mathbb{R} \to \mathbb{R}$ is continuous by definition. $\tilde{1}_\epsilon(t_k - s_k), \tilde{1}_\epsilon(s_k - t_k)$ are compositions of a continuous function with affine functions (thereby continuous), and hence are continuous (Theorem 4.9, (Rudin et al., 1964)).

$\tilde{f}_{\epsilon,i} : \mathbb{R}^d \to \mathbb{R}$, being a sum of continuous functions, is continuous for all $i \in [m]$ (Theorem 4.9, (Rudin et al., 1964)).

Since the coordinate functions of $\tilde{\mathbf{f}} : \mathbb{R}^d \to \mathbb{R}^m$, $\tilde{f}_i : \mathbb{R}^d \to \mathbb{R}$ are continuous, $\tilde{\mathbf{f}}$ is continuous (Theorem 4.10, (Rudin et al., 1964)).

Injectivity of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$

Consider the coordinate-wise functions as defined in B.21,
$\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^{d} \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$, $\tilde{\mathbf{f}}_{\epsilon,k}(s_k) = (\mathbf{J}^{(0)}_{:,k} s_k)\tilde{1}_\epsilon(t_k - s_k) + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k)\tilde{1}_\epsilon(s_k - t_k)$. We now show that all the coordinate-wise functions, $\tilde{\mathbf{f}}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m\ \forall k \in [d]$ are injective. The following proof is the same as the proof of injectivity of $\tilde{\mathbf{f}}_{\epsilon,d} : \mathbb{R} \to \mathbb{R}^m$ in Lemma B.15.

$$\begin{aligned}
\tilde{\mathbf{f}}_{\epsilon,k}(s_k) &= \mathbf{J}^{(0)}_{:,k} s_k \tilde{1}_\epsilon(t_k - s_k) + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k)\tilde{1}_\epsilon(s_k - t_k) \\
&= \mathbf{J}^{(0)}_{:,k} s_k (1 - \tilde{1}_\epsilon(s_k - t_k)) + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k)\tilde{1}_\epsilon(s_k - t_k) \\
&\because \tilde{1}_\epsilon(s_k) + \tilde{1}_\epsilon(-s_k) = 1 \\
&= \mathbf{J}^{(0)}_{:,k}(s_k - (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k)) + \mathbf{J}^{(0)}_{:,k}(s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \\
&= [\mathbf{J}^{(0)}_{:,k}\ \mathbf{J}^{(0)}_{:,k}] \begin{bmatrix} s_k - (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \\ (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \end{bmatrix}
\end{aligned} \tag{31}$$

Define $\mathbf{t}_k : \mathbb{R} \to \mathbb{R}^d$ such that $\mathbf{t}_k(s_k) = \begin{bmatrix} s_k - (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \\ (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \end{bmatrix}$. To show that $\tilde{\mathbf{f}}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m$ is injective, we need to show the following:

$$\forall s_k^{(1)}, s_k^{(2)} \in \mathbb{R} :\ \tilde{\mathbf{f}}_{\epsilon,k}(s_k^{(1)}) = \tilde{\mathbf{f}}_{\epsilon,k}(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)} \tag{32}$$

As usual, we show (32) by contradiction. Let

$$\exists s_k^{(1)} \neq s_k^{(2)} \in \mathbb{R} \text{ s.t. } \tilde{\mathbf{f}}_{\epsilon,d}(s_k^{(1)}) = \tilde{\mathbf{f}}_{\epsilon,k}(s_k^{(2)}) \tag{33}$$

Then we obtain

$$\begin{aligned}
\tilde{\mathbf{f}}_{\epsilon,k}(s_k^{(1)}) &= \tilde{\mathbf{f}}_{\epsilon,k}(s_k^{(2)}) \\
[\mathbf{J}^{(0)}_{:,k}\ \mathbf{J}^{(0)}_{:,k}]\mathbf{t}(s_k^{(1)}) &= [\mathbf{J}^{(0)}_{:,k}\ \mathbf{J}^{(0)}_{:,k}]\mathbf{t}(s_k^{(2)}) \\
\implies \mathbf{t}(s_k^{(1)}) &= \mathbf{t}(s_k^{(2)}) \\
&\quad \text{because } [\mathbf{J}^{(0)}_{:,k}\ \mathbf{J}^{(0)}_{:,k}] \text{ is full column rank (Lemma B.2)} \\
\begin{bmatrix} s_k^{(1)} - (s_k^{(1)} - t_k)\tilde{1}_\epsilon(s_k^{(1)} - t_k) \\ (s_k^{(1)} - t_k)\tilde{1}_\epsilon(s_k^{(1)} - t_k) \end{bmatrix} &= \begin{bmatrix} s_k^{(2)} - (s_k^{(2)} - t_k)\tilde{1}_\epsilon(s_k^{(2)} - t_k) \\ (s_k^{(2)} - t_k)\tilde{1}_\epsilon(s_k^{(2)} - t_k) \end{bmatrix} \\
\implies s_k^{(1)} &= s_k^{(2)}
\end{aligned}$$

Hence, we arrive at a contradiction to (33). Thereby, $\tilde{\mathbf{f}}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m$ is injective $\forall k \in [d]$.

We now show the above statement implies that the injectivity of the coordinate-wise functions $\tilde{\mathbf{f}}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m$ implies the injectivity of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$, $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^{d} \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$. Observe that by definition B.21

1. $\mathbf{S}_1 = \text{span}(\text{Im}(\tilde{\mathbf{f}}_{\epsilon,1})) = \text{span}(\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1})$

2. $\mathbf{S}_2 = \text{span}(\text{Im}(\tilde{\mathbf{f}}_{\epsilon,2})) = \text{span}(\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1})$

   $\vdots$

3. $\mathbf{S}_d = \text{span}(\text{Im}(\tilde{\mathbf{f}}_{\epsilon,d})) = \text{span}(\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1})$

Consider $\mathbb{V} = \text{span}(\text{Im}(\tilde{\mathbf{f}}_\epsilon))$. By Lemma B.19, $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$. Further, by Lemma B.11, injectivity of $\tilde{\mathbf{f}}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m \forall k \in [d]$ implies injectivity of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$.

Continuity of derivatives of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$

Consider the derivatives of $\tilde{\mathbf{f}}_\epsilon(\mathbf{s})$ with respect to the coordinates of $\mathbf{s} = (s_1, s_2, ..., s_d)$. Since $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^d \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$, $\frac{\partial \tilde{\mathbf{f}}_\epsilon(\mathbf{s})}{\partial s_k} = \frac{d\tilde{\mathbf{f}}_{\epsilon,k}(s_k)}{ds_k} = \tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$. The proof of continuity of $\tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$ is the same as the proof of continuity of $\tilde{\mathbf{f}'}_{\epsilon,d}(s_d)$ as in Lemma B.15 and is rewritten here for easy readability.

By (31),

$$\tilde{\mathbf{f}}_{\epsilon,k}(s_k) = [\mathbf{J}^{(0)}_{:,k} \ \mathbf{J}^{(1)}_{:,k}] \begin{bmatrix} s_k - (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \\ (s_k - t_k)\tilde{1}_\epsilon(s_k - t_k) \end{bmatrix}$$
$$\tilde{\mathbf{f}'}_{\epsilon,k}(s_k) = [\mathbf{J}^{(0)}_{:,k} \ \mathbf{J}^{(1)}_{:,k}] \begin{bmatrix} 1 - \tilde{1}_\epsilon(s_k - t_k) - (s_k - t_k)\tilde{1'}_\epsilon(s_k - t_k) \\ \tilde{1}_\epsilon(s_k - t_k) + (s_k - t_k)\tilde{1'}_\epsilon(s_k - t_k) \end{bmatrix}$$

where by definition B.13 $\tilde{1'}_\epsilon(s) = \begin{cases} 0 & s \leq -\epsilon \\ \frac{1}{2}\cos\left(\frac{\pi s}{2\epsilon}\right)\frac{\pi}{2\epsilon} & -\epsilon < s \leq \epsilon \\ 0 & s > \epsilon \end{cases}$. Notice that $\tilde{1'}_\epsilon : \mathbb{R} \to \mathbb{R}$ is continuous in $\mathbb{R}$. $\tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$ is continuous since it is composed by a sum and product of continuous functions (Theorem 4.9, (Rudin et al., 1964)). Notice also that the term $(s_k - t_k)\tilde{1'}_\epsilon(s_k - t_k) = \frac{1}{2}\cos\left(\frac{\pi(s_k - t_k)}{2\epsilon}\right)\frac{\pi}{2\epsilon}.(s_k - t_k)$ is non-zero only when $-\epsilon < (s_k - t_k) \leq \epsilon$, hence this term is finite even for $\epsilon > 0$ arbitrarily small. The other terms in $\tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$ are also finite be definition. Thus, the derivatives $\frac{\partial \tilde{f}_{\epsilon,i}}{\partial s_k} \forall i \in [m], k \in [d]$ are continuous for $\epsilon > 0$ arbitrarily small.

Since all the partial derivatives of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ are continuous, $\tilde{\mathbf{f}}$ is continuously differentiable (Theorem 9.21, (Rudin et al., 1964)).

□

We now present the theorem that introduces a bound on the global IMA contrast for non-affine maps, $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$, composed by smoothly joining 2^d affine maps with local bases sampled isotropically as defined here B.21.

Theorem B.23. *Consider the map $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ sampled randomly from the procedure B.21.*

Then, the map $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$, for $\epsilon > 0$ arbitrarily small and any finite probability density, $p_\mathbf{s}$, defined over $\mathbb{R}^d$ satisfies the following bound on the global IMA contrast $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s}) \leq \delta$ with (high) probability $\geq 1 - \min\left\{1, \exp(2\log d - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ for $m \gg d$ where $\delta < \frac{1}{2}$ is arbitrarily small.

Proof. We show that the condition of Theorem 5.1, the columns of the Jacobian of $\tilde{\mathbf{f}}_\epsilon$ defined in B.21 are locally sampled isotropically i.e. , is still satisfied for the domain of $\tilde{\mathbf{f}}_\epsilon$, i.e. $\forall \mathbf{s} \in \mathbb{R}^d$ almost surely w.r.t finite probability measure, $p_\mathbf{s}$ over $\mathbb{R}^d$.

Following from Definition B.17 and Definition B.21, consider the partition of the domain of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$, $\mathbb{R}^d$ into the following regions,

1. $\text{int}(\mathbb{P}^{(\mathbf{b})}) := \{\mathbf{s} \mid \mathbf{b}(\mathbf{s}) = \mathbf{b}, s_i \notin (t_i - \epsilon, t_i + \epsilon] \ \forall i \in [d]\} \ \forall \mathbf{b} \in \{0,1\}^d$

   Notice that by Definition B.17 and Definition B.21, $\forall \mathbf{s} \in \text{int}(\mathbb{P}^{(\mathbf{b})})$, $\forall \mathbf{b} \in \{0,1\}^d$, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}$ (defined in B.17), where $\mathbf{J}^{(\mathbf{b})}_1, \mathbf{J}^{(\mathbf{b})}_2, ..., \mathbf{J}^{(\mathbf{b})}_d \overset{i.i.d}{\sim} p_\mathbf{r}$. Thus, the condition of

Theorem 5.1, the columns of the Jacobian of $\tilde{\mathbf{f}}_\epsilon$ are locally sampled isotropically, is still satisfied for these regions.

2. $\mathbb{B} := \mathbb{R}^d - \bigcup_{\mathbf{b} \in \{0,1\}^d} \text{int}(\mathbb{P}^{(\mathbf{b})})$

   As in the case where the map, $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ was defined as the smooth connection of *two* affine maps (Theorem B.16), the region $\mathbb{B}$ sandwiching the boundary of the partitions has arbitrarily small probability measure since:

   (a) $\mathbb{B}$ is an ϵ-sandwich of a $(d-1)$-dimensional region of a d-dimensional domain. The Lebesgue measure on $\mathbb{B}$ is equal to the volumne element associated with $\mathbb{B}$ (3.3, (Çinlar, 2011)), thus, $\lambda(\mathbb{B}) = \Theta(\epsilon)$[8] where $\lambda(.)$ denotes the Lebesgue measure.

   (b) $p_\mathbf{s}$ is finite at all points.

   Hence, $p(\mathbb{B}) = \int_\mathbb{B} p_\mathbf{s} \lambda(\mathbf{s}) = \Theta(\epsilon)$, is arbitrarily small for suitably chosen ϵ.

   Like in Theorem B.16, to derive a bound on the global IMA contrast of $\tilde{\mathbf{f}}_\epsilon$, $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s})$, we need that in region, $\forall \mathbf{s} \in \mathbb{B}$, the value of the local IMA contrast $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})$ is finite. This is equivalent to showing that the Jacobian, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ is full column-rank for all $\mathbf{s} \in \mathbb{B}$. Consider the definition of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ (Definition B.28) in terms of coordinate-wise functions, $\mathbf{f}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m, \forall k \in [d]$.

$$\tilde{\mathbf{f}}_{\epsilon,k}(s_k) = \mathbf{J}^{(0)}_{:,k} s_k \tilde{1}_\epsilon(t_k - s_k) + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k)\tilde{1}_\epsilon(s_k - t_k) \quad \forall k \in [d]$$

   Consider the k-th column of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ for any $k \in [d]$.

$$\begin{aligned}
\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} &= \mathbf{J}^{(0)}_{:,k}\tilde{1}_\epsilon(t_k - s_k) + \mathbf{J}^{(1)}_{:,k}\tilde{1}_\epsilon(s_k - t_k) - \mathbf{J}^{(0)}_{:,k} s_k \tilde{1}'_\epsilon(t_k - s_k) + (\mathbf{J}^{(1)}_{:,k}(s_k - t_k) + \mathbf{J}^{(0)}_{:,k} t_k)\tilde{1}'_\epsilon(s_k - t_k) \\
&= \mathbf{J}^{(0)}_{:,k}((t_k - s_k)\tilde{1}'_\epsilon(t_k - s_k) + \tilde{1}_\epsilon(t_k - s_k)) + \mathbf{J}^{(1)}_{:,k}((s_k - t_k)\tilde{1}'_\epsilon(s_k - t_k) + \tilde{1}_\epsilon(s_k - t_k))
\end{aligned}$$

   Observe that $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k}$ is a linear combination of $\mathbf{J}^{(0)}_{:,k}$ and $\mathbf{J}^{(1)}_{:,k} \quad \forall k \in [d]$. Since by Lemma B.2, $\mathbf{J}^{(0)}_{:,1}, \mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(0)}_{:,2}, \mathbf{J}^{(1)}_{:,2}, ..., \mathbf{J}^{(0)}_{:,d}, \mathbf{J}^{(1)}_{:,d}$ are all nonzero and linearly independent with respect to each other with probability 1, the only possibility for $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ to not be full column-rank is for $k \in [d]$,

$$\begin{aligned}
&\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} = \mathbf{0} \\
&\implies (t_k - s_k)\tilde{1}'_\epsilon(t_k - s_k) + \tilde{1}_\epsilon(t_k - s_k) = (s_k - t_k)\tilde{1}'_\epsilon(s_k - t_k) + \tilde{1}_\epsilon(s_k - t_k) = 0 \\
&\qquad \because \mathbf{J}^{(0)}_{:,k}, \mathbf{J}^{(1)}_{:,k} \text{ are linearly independent.}
\end{aligned}$$

   Consider the function, $q : \mathbb{R} \to \mathbb{R}$ such that $q(s) = s\tilde{1}'_\epsilon(s) + \tilde{1}_\epsilon(s)$. Observe that $q(s) \geq 0$ for $s \geq 0$. Thus, for $(t_k - s_k)\tilde{1}'_\epsilon(t_k - s_k) + \tilde{1}_\epsilon(t_k - s_k) = (s_k - t_k)\tilde{1}'_\epsilon(s_k - t_k) + \tilde{1}_\epsilon(s_k - t_k) = 0$, we need that $s_k = t_k$. At $s_k = t_k$, $q(s_k - t_k) = q(t_k - s_k) = \frac{1}{2} \neq 0$. Hence, we have shown that $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} \neq \mathbf{0} \forall k \in [d]$, thereby $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ is full column-rank and $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})$ is finite for all $\mathbf{s} \in \mathbb{B}$.

Hence,

$$\begin{aligned}
C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}}) &= \int_{\mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} \\
&= \bigcup_{\mathbf{b}\in\{0,1\}^d} \int_{\text{int}(\mathbb{P}^{(\mathbf{b})})} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \int_{\mathbb{B}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} \\
&= \bigcup_{\mathbf{b}\in\{0,1\}^d} \int_{\text{int}(\mathbb{P}^{(\mathbf{b})})} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} + \Theta(\epsilon) \\
&\approx \bigcup_{\mathbf{b}\in\{0,1\}^d} \int_{\text{int}(\mathbb{P}^{(\mathbf{b})})} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\; p_{\mathbf{s}} d\mathbf{s} \quad \text{for } \epsilon \text{ arbitrarily small.} \\
&\leq \max_{\mathbf{s}\in\mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s}) \int_{\mathbb{R}^d} p_{\mathbf{s}} d\mathbf{s} \\
&\leq \max_{\mathbf{s}\in\mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s}) \leq \delta \quad \text{w. p. } \geq 1 - \min\left\{1, \exp(2logd - \kappa(m-1)\frac{\delta^2}{d^2})\right\},
\end{aligned}$$

by Theorem 5.1.

Thus, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}}) \leq \delta$ for $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.14 with (high) probability at least $1 - \min\left\{1, \exp(2\log d - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ for $m \gg d$ where $\delta < \frac{1}{2}$ is arbitrarily small.

□

Defining non-linear functions as grid-wise affine functions In the previous subsection, we defined smooth nonlinear functions, by smoothly approximating affine functions defined on orthants across a given point (Definition B.17, B.21). In this subsection, we extend the previous sampling process for functions to consider functions which smoothly approximate functions which are piecewise affine across a grid-like partition of the domain. However, unlike the previous function sampling processes, we now restrict ourselves, in the current sampling process, to define functions on a bounded subset of the d-dimensional Euclidean space and without loss of generality, we choose our domain to be $[0,1]^d$. We make this restriction since upon defining a regular grid (with fixed grid width) on an unbounded domain, we would no longer be able to argue that the columns of the Jacobian of the to-be defined sampling process of functions (definition B.24), $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, defining the local bases, would no longer be linearly independent with probability 1, since there are infinitely many of them, and we can no longer apply Lemma B.2. Recall that we needed the linear independence of the local bases of the function with respect to one another to show injectivity of $\mathbf{f}$ (Lemmata B.12, B.15, B.27, B.22). In the context of signal processing, the assumption of bounded domain for the sources is often satisfied.

We proceed in the usual fashion, we first use grid-wise affine function and show their continuity and injectvity. Then, we define a smooth approximation to the grid-wise affine function and show that it is continuous, injective and continuously differentiable, and hence, the Jacobian of the function is defined at all points and the global IMA contrast can be computed. Then, we derive a high probability bound on the global IMA contrast in this scenario such the bound holds with growing probability as the dimensionality of the observed space increases.

Following is the definition of the grid-wise affine function.

Definition B.24. *The maps $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ we consider are defined as follows:*

1. *We define a function, $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ as a grid-wise affine function, applied to say $\mathbf{s} \in [0,1]^d$.*
2. *Consider a partition of the domain $[0,1]^d$ as a regular grid, with grid width $1 \geq \delta > \epsilon > 0, \delta, \epsilon \in \mathbb{R}$. The number of grid parts along a dimension, $k \in [d]$, of the domain $[0,1]^d$ is therefore equal to $p = \lceil \frac{1}{\delta} \rceil + 1$ where $\lceil . \rceil$ is the ceiling function.*
3. *To define partitions of the domain, $[0,1]^d$, consider the vector, $\mathbf{b}; [0,1]^d \to [p]^d, [p] = \{1, 2, ..., p\}$, where $\mathbf{b}(\mathbf{s}) = (b_1(\mathbf{s}), b_2(\mathbf{s}), ..., b_d(\mathbf{s}))$ such that $b_k(\mathbf{s}) := \lceil \frac{s_k}{\delta} \rceil$. The partition of the domain, $[0,1]^d$, we consider is defined as $[0,1]^d = \mathbb{P}_{[0,1]^d} = \bigcup_{\mathbf{b}\in[p]^d} \mathbb{P}^{(\mathbf{b})}$ where*

$\mathbb{P}^{(\mathbf{b})} := \{\mathbf{s} \quad | \quad \mathbf{b}(\mathbf{s}) = \mathbf{b}\}$. *Note that the partition defined is axis-aligned to the canonical basis in* $\mathbb{R}^d$*. This follows to extend the continuity argument from the two-partition case in Lemma B.6, observation B.7.*

4. *Consider the matrices,* $\mathbf{J}^{(1)}, \mathbf{J}^{(2)}, ..., \mathbf{J}^{(p)} \in \mathbb{R}^{m \times d}$*, used to define the Jacobian in each part,* $\mathbb{P}^{(\mathbf{b})}$ $\forall \mathbf{b} \in [p]^d$*. The columns of* $\mathbf{J}^{(1)}, \mathbf{J}^{(2)}, ..., \mathbf{J}^{(p)}$ *are sampled from a spherically symmetric distribution,* $p_{\mathbf{r}}$*,* $\mathbf{J}_1^{(i)}, \mathbf{J}_2^{(i)}, ..., \mathbf{J}_d^{(i)} \overset{i.i.d}{\sim} p_{\mathbf{r}} \forall i \in [p]$*, so that the pre-condition for Theorem 5.1 holds almost everywhere.*

5. *For* $\mathbf{s} \in \mathbb{R}^d$ *with* $\mathbf{b}(\mathbf{s}) = \mathbf{b} \in [p]^d$*,* $\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}$ *such that*

$$\left\{ \mathbf{J}^{(\mathbf{b})}_{:,k} = \mathbf{J}^{(i)}_{:,k}, \quad b_k = i\,,\ \forall\, k \in [d], i \in [p] \right\}$$

Note that this corresponds to Observation B.7 where changing one column of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *across a partition of the domain results in axis-aligned partitions, also akin to the definition in B.17, except we now have a regular grid where each dimension is split into* p *segments rather than orthants across a given point in the domain.*

6. $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *is defined as:* $\left\{ \mathbf{f}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}(\mathbf{s}) + \mathbf{c}^{(\mathbf{b})} | \quad \mathbf{b}(\mathbf{s}) = \mathbf{b} \quad \right\}$*, where* $\mathbf{c}^{(\mathbf{b})} \in \mathbb{R}^m \ \forall\ \mathbf{b} \in [p]^d$*.*

7. *Owing to the axis-alignment of the chosen partition,* $\mathbb{P}_[0,1]^d$*,* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *can be decomposed as a sum of functions acting upon individual coordinates (also called as coordinate-wise functions),* $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d \mathbf{f}_k(s_k)\ \forall\ k \in [d]$ *where* $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ *is defined as,*

$$\mathbf{f}_k(s_k) := \sum_{t=1}^{p} (\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta) 1_{s_k \in ((t-1)\delta, t\delta]} \tag{34}$$

We now show that maps, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.24 are continuous and injective.

Lemma B.25 (Continuity of grid-wise affine functions)**.** *Consider maps* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ *as defined in B.24. Such a map* $\mathbf{f}$ *is continuous.*

Proof. Consider $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d f_k(s_k)$, where $\mathbf{f}_k(s_k) = \mathbf{J}^{b_k(\mathbf{s})}_{:,k}(s_k - (b_k(\mathbf{s}) - 1)\delta) + \sum_{i=1}^{b_k(\mathbf{s})-1} \mathbf{J}^{(i)}_{:,k}(\delta)\ \forall\ k \in [d]$.

We show continuity of $\mathbf{f}_k : \mathbb{R} \to \mathbb{R}^m\ \forall k \in [d]$. For a particular k, consider the cases:

1. $\mathbb{B} := \{s_k = t\delta,\ t \in [p-1]\}$

   Let $s_k = t\delta,\ t \in [p-1], b_k(\mathbf{s}) = t$, by (34), $\mathbf{f}_k(s_k) = \sum_{i=1}^t \mathbf{J}^{(i)}_{:,k}\delta$, which is also equal to $\mathbf{f}_k(s_k)$ in the left limit. To show that $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ is continuous at $s_k = t\delta$, it remains to show that the right limit of $\mathbf{f}_k(s_k)$ is equal to the value of the function at $s_k = t\delta$. By (34), at the right limit of $s_k = t\delta$, $\mathbf{f}_k(s_k) = (\mathbf{J}^{(t+1)}_{:,k}(t\delta - t\delta) + \sum_{i=1}^t \mathbf{J}^{(i)}_{:,k}\delta) = \sum_{i=1}^t \mathbf{J}^{(i)}_{:,k}\delta = \mathbf{f}_k(s_k)$. Hence, $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ is continuous as $s_k = t\delta, t \in [p-1]$.

2. $s_k \in [0,1] - \mathbb{B}$

   For these values of s_k, $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ is affine. Let $\lceil \frac{s_k}{\delta} \rceil = t$, $\mathbf{f}_k(s_k) = \mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta$. Since, $\mathbf{f}_k(s_k)$ is affine for s_k in this region, $\mathbf{f}_k(s_k)$ is continuous in this region.

$\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, \mathbf{f}(\mathbf{s}) = \sum_{k=1}^d f_k(s_k)$ is continuous since the sum of continuous functions is continuous (Theorem 4.9, (Rudin et al., 1964)). □

To show that $\mathbf{f} : [0,1]^d \rightarrow \mathbb{R}^m$ as defined in B.24 is injective, we follow an analogous approach to the previous cases where non-linear functions were defined as a composition of two affine functions and 2^d affine functions respectively. We first show that the images of the coordinate-wise functions of $\mathbf{f}$, $\mathbf{f}_k : [0,1] \rightarrow \mathbb{R}^m$ are in direct sum with respect to the image of $\mathbf{f} : [0,1] \rightarrow \mathbb{R}^m$. Then, we show that the coordinate-wise functions, $\mathbf{f}_k$ are injective and use the direct sum property to conclude that injectivity of $\mathbf{f}$ is implied.

Lemma B.26. *Consider $\mathbf{J}^{(1)}, \mathbf{J}^{(2)}, ..., \mathbf{J}^{(p)} \in \mathbb{R}^{m\times d}$ as sampled in Definition B.24. The vector space $\mathbb{V} = span\{\bigcup_{i=1}^{p} cols(\mathbf{J}^{(i)})\}$ is the direct sum of the family*

$$\mathcal{F} = \{\mathbf{S}_1 = span\{\mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(2)}_{:,1}, ..., \mathbf{J}^{(p)}_{:,1}\}, \mathbf{S}_2 = span\{\mathbf{J}^{(1)}_{:,2}, \mathbf{J}^{(2)}_{:,2}, ..., \mathbf{J}^{(p)}_{:,2}\},$$
$$..., \mathbf{S}_{d-1} = span\{\mathbf{J}^{(1)}_{:,d-)}, \mathbf{J}^{(2)}_{:,d-1}, ..., \mathbf{J}^{(p)}_{:,d-1}\}, \mathbf{S}_d = span\{\mathbf{J}^{(1)}_{:,d}, \mathbf{J}^{(2)}_{:,d}, ..., \mathbf{J}^{(p)}_{:,d}\}\},$$

where cols(.) denotes the set of columns of a given matrix.

Proof. From Lemma B.2, for the scenario $m \gg d$ (here it is sufficient to have $(m > p.d)$), the set of vectors $\{\bigcup_{i=1}^{p} \text{cols}(\mathbf{J}^{(i)})\}$ are non-zero and linearly independent with probability 1.

Consider $\mathbf{v} \in \mathbb{V}, \mathbf{u}_1, \mathbf{w}_1 \in \mathbf{S}_1, \mathbf{u}_2, \mathbf{w}_2 \in \mathbf{S}_2, ..., \mathbf{u}_d, \mathbf{w}_d \in \mathbf{S}_d$ such that

$$\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_d, \mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_d$$

By definition B.9, to show $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$, we need to show $\mathbf{u}_1 = \mathbf{w}_1, \mathbf{u}_2 = \mathbf{w}_2, ..., \mathbf{u}_d = \mathbf{w}_d$.

Let

- $\mathbf{u}_1 = \sum_{i=1}^{p} c_1^{(i)} \mathbf{J}^{(i)}_{:,1},\ \mathbf{w}_1 = \sum_{i=1}^{p} c_1^{(i)'} \mathbf{J}^{(i)}_{:,1}$
- $\mathbf{u}_2 = \sum_{i=1}^{p} c_2^{(i)} \mathbf{J}^{(i)}_{:,2},\ \mathbf{w}_2 = \sum_{i=1}^{p} c_2^{(i)'} \mathbf{J}^{(i)}_{:,2}$

  $\vdots$
- $\mathbf{u}_{d-1} = \sum_{i=1}^{p} c_{d-1}^{(i)} \mathbf{J}^{(i)}_{:,d-1},\ \mathbf{w}_2 = \sum_{i=1}^{p} c_{d-1}^{(i)'} \mathbf{J}^{(i)}_{:,d-1}$
- $\mathbf{u}_d = \sum_{i=1}^{p} c_d^{(i)} \mathbf{J}^{(i)}_{:,d},\ \mathbf{w}_2 = \sum_{i=1}^{p} c_d^{(i)'} \mathbf{J}^{(i)}_{:,d}$

$$\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_d = \mathbf{w}_1 + \mathbf{w}_2 + ... + \mathbf{w}_d$$
$$\sum_{i=1}^{p} c_1^{(i)} \mathbf{J}^{(i)}_{:,1} + \sum_{i=1}^{p} c_2^{(i)} \mathbf{J}^{(i)}_{:,2} + ... + \sum_{i=1}^{p} c_d^{(i)} \mathbf{J}^{(i)}_{:,d} = \sum_{i=1}^{p} c_1^{(i)'} \mathbf{J}^{(i)}_{:,1} + \sum_{i=1}^{p} c_2^{(i)'} \mathbf{J}^{(i)}_{:,2} + ... + \sum_{i=1}^{p} c_d^{(i)'} \mathbf{J}^{(i)}_{:,d}$$
$$\sum_{i=1}^{p} (c_1^{(i)} - c_1^{(i)'}) \mathbf{J}^{(i)}_{:,1} + \sum_{i=1}^{p} (c_2^{(i)} - c_2^{(i)'}) \mathbf{J}^{(i)}_{:,2} + ... + \sum_{i=1}^{p} (c_d^{(i)} - c_d^{(i)'}) \mathbf{J}^{(i)}_{:,d} = \mathbf{0}$$

Since the set of vectors $\bigcup_{i=1}^{p} \text{cols}(\mathbf{J}^{(i)})$ are nonzero and linearly independent with probability 1 (Lemma B.2),

$$(c_1^{(i)} - c_1^{(i)'}) = (c_2^{(i)} - c_2^{(i)'}) = ... = (c_d^{(i)} - c_d^{(i)'}) = 0 \quad \forall i \in [p]$$

Hence, it follows that $\mathbf{u}_1 = \mathbf{w}_1, \mathbf{u}_2 = \mathbf{w}_2, ..., \mathbf{u}_d = \mathbf{w}_d$ and $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$.

□

Lemma B.27 (Injectivity of grid-wise affine maps)**.** *Consider maps $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^m$ as defined in B.24. Such a map $\mathbf{f}$ is injective.*

Proof. Consider $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ written as a sum of coordinate-wise functions (Definition B.24), $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d f_k(s_k)$, where $\mathbf{f}_k(s_k) = \sum_{t=1}^p (\mathbf{J}_{:,k}^{(t)}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}_{:,k}^{(i)}\delta)1_{s_k \in ((t-1)\delta, t\delta]}$.

In the following proof, we show injectivity of the coordinate-wise functions $\mathbf{f}_k : [0,1] \to \mathbb{R}^m \ \forall k \in [d]$ and conclude that $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ is injective by Lemma B.26 and Lemma B.11. For a particular k, to show that $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ is injective, we need to show the following:

$$\forall s_k^{(1)}, s_k^{(2)} \in [0,1] : \quad \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)} \tag{35}$$

As usual, we show (35) by contradiction. Let

$$\exists s_k^{(1)} \neq s_k^{(2)} \in [0,1] s.t. \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \tag{36}$$

Observe that $\mathbf{f}_k(s_k) \in \mathbf{S}_k = \text{span}(\mathbf{J}_{:,k}^{(1)}, \mathbf{J}_{:,k}^{(2)}, ..., \mathbf{J}_{:,k}^{(p)})$. We define the coefficient vector for a given $s_k \in \mathbb{R}$, $\lceil \frac{s_k}{\delta} \rceil = t+1$, $\mathbf{t} : [0,1] \to [0,1]^p, \mathbf{t}(s_k) := \sum_{i=1}^t \delta \mathbf{e}_i + (s_k - t\delta)\mathbf{e}_{t+1}$, $\mathbf{e}_i$ denotes the i-th canonical orthonormal basis vector in $\mathbb{R}^p$. Then we get

$$\begin{aligned} \mathbf{f}_k(s_k^{(1)}) &= \mathbf{f}_k(s_k^{(2)}) \\ [\mathbf{J}_{:,k}^{(1)}\ \mathbf{J}_{:,k}^{(2)}\ ...\ \mathbf{J}_{:,k}^{(p)}]\mathbf{t}(s_k^{(1)}) &= [\mathbf{J}_{:,k}^{(1)}\ \mathbf{J}_{:,k}^{(2)}\ ...\ \mathbf{J}_{:,k}^{(p)}]\mathbf{t}(s_k^{(2)}) \\ \implies \mathbf{t}(s_k^{(1)}) &= \mathbf{t}(s_k^{(2)}) \end{aligned}$$

because $\mathbf{J}_{:,k}^{(1)}, \mathbf{J}_{:,k}^{(2)}, ..., \mathbf{J}_{:,k}^{(p)}$ are linearly independent w. p. 1, Lemma B.2 This implies

$$\begin{aligned} \lceil \frac{s_k^{(1)}}{\delta} \rceil = \lceil \frac{s_k^{(2)}}{\delta} \rceil = t+1, s_k^{(1)} - t\delta = s_k^{(2)} - t\delta \qquad &\text{for some } t \in \mathbb{N} \\ \implies s_k^{(1)} = s_k^{(2)}. \end{aligned}$$

Thus, we arrive at a contradiction to (36), and therefore (35) holds, and $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ defined as in B.24 is injective $\forall k \in [d]$.

We now show that the injectivity of the coordinate-wise functions $\mathbf{f}_k : [0,1] \to \mathbb{R}^d$ implies the injectivity of $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d \mathbf{f}_k(s_k)$. Observe that by definition B.24

1. $\mathbf{S}_1$ = span(Im($\mathbf{f}_1$)) = span($\mathbf{J}_{:,1}^{(1)}, \mathbf{J}_{:,1}^{(2)}, ..., \mathbf{J}_{:,1}^{(p)}$)
2. $\mathbf{S}_2$ = span(Im($\mathbf{f}_2$)) = span($\mathbf{J}_{:,2}^{(1)}, \mathbf{J}_{:,2}^{(2)}, ..., \mathbf{J}_{:,2}^{(p)}$)

   $\vdots$
3. $\mathbf{S}_d$ = span(Im($\mathbf{f}_d$)) = span($\mathbf{J}_{:,d}^{(1)}, \mathbf{J}_{:,d}^{(2)}, ..., \mathbf{J}_{:,d}^{(p)}$)

Consider $\mathbb{V} = \text{span}(\text{Im}(\mathbf{f}))$. By Lemma B.26, $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ... \mathbf{S}_d$. Further, by Lemma B.11, injectivity of $\mathbf{f}_k : [0,1]^d \to \mathbb{R}^m \forall k \in [d]$ implies injectivity of $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$.

□

We proceed to define a smooth approximation to $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ defined in B.24.

Definition B.28 (Smooth approximation to grid-wise affine maps)**.** *Consider the decomposition of* $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ *as a sum of coordinate-wise functions,* $\mathbf{f}(\mathbf{s}) = \sum_{k=1}^d f_k(s_k)$, $\mathbf{s} = (s_1, s_2, ..., s_d) \in [0,1]^d$ *where*

$$\mathbf{f}_k(s_k) = \sum_{t=1}^p (\mathbf{J}_{:,k}^{(t)}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}_{:,k}^{(i)}\delta)1_{s_k \in ((t-1)\delta, t\delta]}.$$

We define the smoothened version of $\mathbf{f} : [0,1]^d \to \mathbb{R}^m$ *as* $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^d \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$ *for* $\epsilon > 0$ *arbitrarily small where*

$$\tilde{\mathbf{f}}_{\epsilon,k}(s_k) := \sum_{t=1}^p \left(\mathbf{J}_{:,k}^{(t)}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}_{:,k}^{(i)}\delta \right) (\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta)).$$

Lemma B.29. *Functions $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$ defined in B.28 are continuously differentiable in $[0,1]^d$, in addition to being continuous and injective, with $\epsilon > 0$ arbitrarily small.*

Proof. We show that $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$ defined in B.28 is continuous, injective and continuously differentiable.

Continuity of $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$

Consider the coordinate-wise decomposition of $\tilde{\mathbf{f}}_\epsilon$; $\tilde{f}_{\epsilon,j}(\mathbf{s}) = \sum_{k=1}^d \tilde{f}_{\epsilon,(j,k)}(s_k)$, where $\tilde{f}_{\epsilon,(j,k)} : [0,1] \to \mathbb{R}$ is defined as

$$\tilde{f}_{\epsilon,(j,k)}(s_k) := \sum_{t=1}^{p} \left(\mathbf{J}_{j,k}^{(t)}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}_{j,k}^{(i)}\delta \right) (\tilde{\mathbb{1}}_\epsilon(s_k - (t-1)\delta) - \tilde{\mathbb{1}}_\epsilon(s_k - t\delta)) \,.$$

We note that $(\mathbf{J}_{j,k}^{(t)}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}_{j,k}^{(i)}\delta)$ is continuous in $s_k \in [0,1]$ $\forall t \in [p], k \in [d]$ since this is affine. Moreover $\tilde{\mathbb{1}} : \mathbb{R} \to \mathbb{R}$ is continuous by definition. $\tilde{\mathbb{1}}_\epsilon(s_k - (t-1)\delta), \tilde{\mathbb{1}}_\epsilon(s_k - t\delta)$ are compositions of a continuous function with affine functions (thereby continuous), and hence are contiinuous (Theorem 4.9, (Rudin et al., 1964)).

$\tilde{f}_{\epsilon,j} : [0,1]^d \to \mathbb{R}$, being a sum of continuous functions, is continuous for all $j \in [m]$ (Theorem 4.9, (Rudin et al., 1964)).

Since the coordinate functions of $\tilde{\mathbf{f}} : [0,1]^d \to \mathbb{R}^m$, $\tilde{f}_j : [0,1]^d \to \mathbb{R}$ $\forall j \in [m]$ are continuous, $\tilde{\mathbf{f}}_\epsilon$ is continuous (Theorem 4.10, (Rudin et al., 1964)).

Injectivity of $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$

We show injectivity of the coordinate-wise functions $\tilde{\mathbf{f}}_{\epsilon,k} : [0,1] \to \mathbb{R}^m$ $\forall k \in [d]$ and conclude that $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$ is injective by Lemma B.26 and Lemma B.11. For a particular k, to show that $\tilde{\mathbf{f}}_{\epsilon,k} : [0,1] \to \mathbb{R}^m$ is injective, we need to show the following:

$$\forall s_k^{(1)}, s_k^{(2)} \in [0,1] : \quad \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)} \tag{37}$$

As usual, we show (37) by contradiction. Let

$$\exists s_k^{(1)} \neq s_k^{(2)} \in [0,1] s.t. \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \tag{38}$$

Observe that $\mathbf{f}_k(s_k) \in \mathbf{S}_k = \text{span}(\mathbf{J}_{:,k}^{(1)}, \mathbf{J}_{:,k}^{(2)}, ..., \mathbf{J}_{:,k}^{(p)})$. We define the coefficient vector for a given $s_k \in \mathbb{R}$,

$$\mathbf{t} : [0,1] \to [0,1]^p, \mathbf{t}(s_k) := \sum_{t=1}^{p} \left(\sum_{i=1}^{t-1} \delta \mathbf{e}_i + (s_k - (t-1)\delta)\mathbf{e}_t \right) (\tilde{\mathbb{1}}_\epsilon(s_k - (t-1)\delta) - \tilde{\mathbb{1}}_\epsilon(s_k - t\delta)) \,,$$

where $\mathbf{e}_i$ denotes the i-th canonical orthonormal basis vector in $\mathbb{R}^p$. Then we get

$$\begin{gathered} \mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \\ [\mathbf{J}_{:,k}^{(1)} \; \mathbf{J}_{:,k}^{(2)} \; ... \; \mathbf{J}_{:,k}^{(p)}]\mathbf{t}(s_k^{(1)}) = [\mathbf{J}_{:,k}^{(1)} \; \mathbf{J}_{:,k}^{(2)} \; ... \; \mathbf{J}_{:,k}^{(p)}]\mathbf{t}(s_k^{(2)}) \\ \implies \mathbf{t}(s_k^{(1)}) = \mathbf{t}(s_k^{(2)}) \,, \end{gathered}$$

because $\mathbf{J}_{:,k}^{(1)}, \mathbf{J}_{:,k}^{(2)}, ..., \mathbf{J}_{:,k}^{(p)}$ are linearly independent with probability one (Lemma B.2).

Observe that the number of non-zero entries in $\mathbf{t}(s_k)$ is determined by the value of s_k. The number of nonzero entries in $\mathbf{t}(s_k)$ for $s_k \in [0,1]$ is as follows:

- $s_k \in \mathbb{P}_{[0,1]}^{(1)} := [0, \delta - \epsilon)$: No. of nonzero entries in $\mathbf{t}(s_k) = 1$
- $s_k \in \mathbb{P}_{[0,1]}^{(2)} := [\delta - \epsilon, 2\delta - \epsilon)$: No. of nonzero entries in $\mathbf{t}(s_k) = 2$

$\vdots$

- $s_k \in \mathbb{P}^{(p-1)}_{[0,1]} := [(p-2)\delta - \epsilon, (p-1)\delta - \epsilon)$: No. of nonzero entries in $\mathbf{t}(s_k) = p - 1$
- $s_k \mathbb{P}^{(p)}_{[0,1]} :=\in [(p-1)\delta - \epsilon, 1]$: No. of nonzero entries in $\mathbf{t}(s_k) = p$

Therefore, $\mathbf{t}(s_k^{(1)}) = \mathbf{t}(s_k^{(2)}) \implies s_k^{(1)}, s_k^{(2)} \in \mathbb{P}^{(r)}_{[0,1]}$ for $r \in [p]$.

1. Consider $s_k^{(1)}, s_k^{(2)} \in \mathbb{P}^{(r)}_{[0,1]}$ for $r \in \{2, 3, ..., p\}$.

   For $s_k \in \mathbb{P}^{(r)}_{[0,1]}, r \in \{2, 3, ..., p\}, \mathbf{t}(s_k) = \sum_{t=r-1}^{r} \left(\sum_{i=1}^{t-1} \delta \mathbf{e}_i + (s_k - (t-1)\delta)\mathbf{e}_t \right) (\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta))$. Observe that the coefficient of $\mathbf{e}_r$ is equal to $t_r(s_k) = (s_k - (r-1)\delta)\tilde{1}_\epsilon(s_k - (r-1)\delta)$. $t_r(s_k)$ is montonic in s_k.

   $\mathbf{t}(s_k^{(1)}) = \mathbf{t}(s_k^{(2)}) \implies t_r(s_k^{(1)}) = t_r(s_k^{(2)})$. Thus, $s_k^{(1)} = s_k^{(2)}$, since $t_r(.)$ is monotonic.

2. Consider $s_k^{(1)}, s_k^{(2)} \in \mathbb{P}^{(1)}_{[0,1]}$.

   For $s_k \in \mathbb{P}^{(1)}_{[0,1]}$, $\mathbf{t}(s_k) = s_k \mathbf{e}_1$. Thus, $\mathbf{t}(s_k^{(1)}) = \mathbf{t}(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)}$.

Thus, we see that, $\mathbf{t}(s_k^{(1)}) = \mathbf{t}(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)}$, $\mathbf{f}_k(s_k^{(1)}) = \mathbf{f}_k(s_k^{(2)}) \implies s_k^{(1)} = s_k^{(2)}$, thereby, $\mathbf{f}_k : [0,1] \to \mathbb{R}^m$ is injective $\forall k \in [d]$.

We now show that the injectivity of the coordinate-wise functions $\tilde{\mathbf{f}}_{\epsilon,k} : [0,1] \to \mathbb{R}^d$ implies the injectivity of $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$, $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^{d} \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$. Observe that by definition B.28

1. $\mathbf{S}_1$ = span(Im($\tilde{\mathbf{f}}_{\epsilon,1}$)) = span($\mathbf{J}^{(1)}_{:,1}, \mathbf{J}^{(2)}_{:,1}, ..., \mathbf{J}^{(p)}_{:,1}$)
2. $\mathbf{S}_2$ = span(Im($\tilde{\mathbf{f}}_{\epsilon,2}$)) = span($\mathbf{J}^{(1)}_{:,2}, \mathbf{J}^{(2)}_{:,2}, ..., \mathbf{J}^{(p)}_{:,2}$)

   ⋮
3. $\mathbf{S}_d$ = span(Im($\tilde{\mathbf{f}}_{\epsilon,d}$)) = span($\mathbf{J}^{(1)}_{:,d}, \mathbf{J}^{(2)}_{:,d}, ..., \mathbf{J}^{(p)}_{:,d}$)

Consider $\mathbb{V} = \text{span}(\text{Im}(\tilde{\mathbf{f}}_\epsilon))$. By Lemma B.26, $\mathbb{V} = \mathbf{S}_1 \bigoplus \mathbf{S}_2 \bigoplus ...\mathbf{S}_d$. Further, by Lemma B.11, injectivity of $\tilde{\mathbf{f}}_{\epsilon,k} : [0,1]^d \to \mathbb{R}^m \forall k \in [d]$ implies injectivity of $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$.

Continuity of derivatives of $\tilde{\mathbf{f}}_\epsilon : [0,1]^d \to \mathbb{R}^m$

Consider the derivatives of $\tilde{\mathbf{f}}_\epsilon(\mathbf{s})$ with respect to the coordinates of $\mathbf{s} = (s_1, s_2, ..., s_d)$. Since $\tilde{\mathbf{f}}_\epsilon(\mathbf{s}) = \sum_{k=1}^{d} \tilde{\mathbf{f}}_{\epsilon,k}(s_k)$, $\frac{\partial \tilde{\mathbf{f}}_\epsilon(\mathbf{s})}{\partial s_k} = \frac{d\tilde{\mathbf{f}}_{\epsilon,k}(s_k)}{ds_k} = \tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$. By Definition B.28,

$$\begin{aligned}
\tilde{\mathbf{f}}_{\epsilon,k}(s_k) &= \sum_{t=1}^{p} \left(\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta \right) (\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta)) \\
\tilde{\mathbf{f}'}_{\epsilon,k}(s_k) &= \sum_{t=1}^{p} \mathbf{J}^{(t)}_{:,k}(\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta)) \\
&\quad + \left(\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta \right) (\tilde{1'}_\epsilon(s_k - (t-1)\delta) - \tilde{1'}_\epsilon(s_k - t\delta)) \qquad (39)
\end{aligned}$$

where by definition B.13 $\tilde{1'}_{\epsilon}(s) = \begin{cases} 0 & s \leq -\epsilon \\ \frac{1}{2}\cos\left(\frac{\pi s}{2\epsilon}\right)\frac{\pi}{2\epsilon} & -\epsilon < s \leq \epsilon \\ 0 & s > \epsilon \end{cases}$. Notice that $\tilde{1'}_{\epsilon} : \mathbb{R} \to \mathbb{R}$ is continuous in $\mathbb{R}$. $\tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$ is continuous since it is composed by a sum and product of continuous functions (Theorem 4.9, (Rudin et al., 1964)).

To show that the derivative, $\tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$, is well-defined for $s_k \in [0, 1]$, we remark on terms containing $\tilde{1'}_{\epsilon}(s_k - t\delta)$ $\forall t \in [p - 1]$ since $\tilde{1'}_{\epsilon}(.)$ can be very large for small ϵ. Notice that the term $\tilde{1'}_{\epsilon}(s_k - p\delta)$ is always equal to zero for $s_k \in [0, 1]$ since $p\delta > 1, \epsilon > 0$ is arbitrarily small; and $\tilde{1'}_{\epsilon}(s_k - t\delta)$ is nonzero for $t\delta - \epsilon < s_k \leq t\delta + \epsilon$. The coefficient multiplied to $\tilde{1'}_{\epsilon}(s_k - t\delta)$ is equal to $\left(\mathbf{J}^{(t+1)}_{:,k}(s_k - t\delta) + \sum_{i=1}^{t} \mathbf{J}^{(i)}_{:,k}\delta - \mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) - \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta\right) = (\mathbf{J}^{(t+1)}_{:,k} - \mathbf{J}^{t}_{:,k})(s_k - t\delta)$. Thus for $t\delta - \epsilon < s_k \leq t\delta + \epsilon$, the term $(\mathbf{J}^{(t+1)}_{:,k} - \mathbf{J}^{t}_{:,k})(s_k - t\delta)\tilde{1'}_{\epsilon}(s_k - t\delta) = (\mathbf{J}^{(t+1)}_{:,k} - \mathbf{J}^{t}_{:,k})(s_k - t\delta)\frac{1}{2}\cos\left(\frac{\pi s}{2\epsilon}\right)$ is well-defined since $(s_k - t\delta) = \Theta(\epsilon)$[8].

We have shown that $\tilde{\mathbf{f}'}_{\epsilon,k}(s_k)$ is continuous and well-defined $\forall k \in [d]$. Thus, the derivatives $\frac{\partial \tilde{f}_{\epsilon,i}}{\partial s_k} \forall i \in [m], k \in [d]$ are continuous for $\epsilon > 0$ arbitrarily small.

Since all the partial derivatives of $\tilde{\mathbf{f}}_{\epsilon} : [0, 1]^d \to \mathbb{R}^m$ are continuous, $\tilde{\mathbf{f}}$ is continuously differentiable (Theorem 9.21, (Rudin et al., 1964)).

□

We now present the theorem that introduces a bound on the global IMA contrast for non-affine maps, $\tilde{\mathbf{f}}_{\epsilon} : \mathbb{R}^d \to \mathbb{R}^m, m \gg d$, defined as a smooth approximation to grid-wise affine maps B.28.

Theorem B.30. *Consider the map $\tilde{\mathbf{f}}_{\epsilon} : [0, 1]^d \to \mathbb{R}^m$ sampled randomly from the procedure B.28.*

Then, the map $\tilde{\mathbf{f}}_{\epsilon} : \mathbb{R}^d \to \mathbb{R}^m$, for $\epsilon > 0$ arbitrarily small and any finite probability density, $p_{\mathbf{s}}$, defined over $[0, 1]^d$ satisfies the following bound on the global IMA contrast $C_{\text{IMA}}(\tilde{\mathbf{f}}_{\epsilon}, p_{\mathbf{s}})$, $C_{\text{IMA}}(\tilde{\mathbf{f}}_{\epsilon}, p_{\mathbf{s}}) \leq \delta$ with (high) probability $\geq 1 - \min\left\{1, \exp(2logd - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ for $m \gg d$ where $\delta < \frac{1}{2}$ is arbitrarily small.

Proof. We show that the condition of Theorem 5.1, the columns of the Jacobian of $\tilde{\mathbf{f}}_{\epsilon}$ defined in B.21 are locally sampled isotropically i.e. , is still satisfied for the domain of $\tilde{\mathbf{f}}_{\epsilon}$, i.e. $\forall \mathbf{s} \in [0, 1]^d$ almost surely w.r.t finite probability measure, $p_{\mathbf{s}}$ over $[0, 1]^d$.

Following from Definition B.24 and Definition B.28, consider the partition of the domain of $\tilde{\mathbf{f}}_{\epsilon} : [0, 1]^d \to \mathbb{R}^m$, $[0, 1]^d$ into the following regions,

1. $\mathbb{I} := \{s_k - t\delta \neq (-\epsilon, \epsilon] \; \forall t \in [p - 1], k \in [d]\}$

   Notice that by Definition B.17 and Definition B.21, $\forall \mathbf{s} \in \mathbb{I}, \mathbf{b} \in \{0, 1\}^d$, such that $\mathbf{b}(\mathbf{s}) = \mathbf{b}$, $\mathbf{J}_{\tilde{\mathbf{f}}_{\epsilon}}(\mathbf{s}) = \mathbf{J}^{(\mathbf{b})}$. Recall that $\mathbf{b}(\mathbf{s})$ is defined such that, $b_k(\mathbf{s}) = \lceil \frac{s_k}{\delta} \rceil$ $\forall k \in [d]$. The columns of $\mathbf{J}^{(\mathbf{b})}$ are sampled independently from a spherically invariant distribution in $\mathbb{R}^m$, i.e. $\mathbf{J}^{(\mathbf{b})}_1, \mathbf{J}^{(\mathbf{b})}_2, ..., \mathbf{J}^{(\mathbf{b})}_d \overset{i.i.d}{\sim} p_{\mathbf{r}}$. Thus, the condition of Theorem 5.1, the columns of the Jacobian of $\tilde{\mathbf{f}}_{\epsilon}$ are locally sampled isotropically, is still satisfied for these regions.

2. $\mathbb{B} := [0, 1]^d - \mathbb{I}$

   As in the case where the map, $\tilde{\mathbf{f}}_{\epsilon} : \mathbb{R}^d \to \mathbb{R}^m$ was defined as the smooth connection of *two* affine maps (Theorem B.16), the region $\mathbb{B}$ sandwiching the boundary of the partitions has arbitrarily small probability measure since:

   (a) $\mathbb{B}$ is an ϵ-sandwich of a $(d-1)$-dimensional region of a d-dimensional domain. The Lebesgue measure on $\mathbb{B}$ is equal to the volumne element associated with $\mathbb{B}$ (3.3, (Çinlar, 2011)), thus, $\lambda(\mathbb{B}) = \Theta(\epsilon)$[8] where $\lambda(.)$ denotes the Lebesgue measure.

   (b) $p_{\mathbf{s}}$ is finite at all points.

Hence, $p(\mathbb{B}) = \int_{\mathbb{B}} p_{\mathbf{s}} \lambda(\mathbf{s}) = \Theta(\epsilon)$, is arbitrarily small for suitably chosen ϵ.

Like in Theorem B.16, to derive a bound on the global IMA contrast of $\tilde{\mathbf{f}}_\epsilon$, $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_{\mathbf{s}})$, we need that in region, $\forall \mathbf{s} \in \mathbb{B}$, the value of the local IMA contrast $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})$ is finite. This is equivalent to showing that the Jacobian, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ is full column-rank for all $\mathbf{s} \in \mathbb{B}$. Consider the definition of $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ (Definition B.28) in terms of coordinate-wise functions, $\mathbf{f}_{\epsilon,k} : \mathbb{R} \to \mathbb{R}^m, \forall k \in [d]$.

$$\tilde{\mathbf{f}}_{\epsilon,k}(s_k) := \sum_{t=1}^{p} \left(\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta \right) (\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta))$$

By (39), the k-th column of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ for any $k \in [d]$ is given by:

$$\begin{aligned}\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} = &\sum_{t=1}^{p} \mathbf{J}^{(t)}_{:,k}(\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta)) \\ &+ \left(\mathbf{J}^{(t)}_{:,k}(s_k - (t-1)\delta) + \sum_{i=1}^{t-1} \mathbf{J}^{(i)}_{:,k}\delta \right) (\tilde{1'}_\epsilon(s_k - (t-1)\delta) - \tilde{1'}_\epsilon(s_k - t\delta))\end{aligned}$$

For $s_k \in (t - \epsilon, t + \epsilon]$,

$$\begin{aligned}\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} &= \mathbf{J}^{(t)}_{:,k}(\tilde{1}_\epsilon(s_k - (t-1)\delta) - \tilde{1}_\epsilon(s_k - t\delta)) + \mathbf{J}^{(t+1)}_{:,k}(\tilde{1}_\epsilon(s_k - t\delta) - \tilde{1}_\epsilon(s_k - (t+1)\delta)) \\ &\quad + (\mathbf{J}^{(t+1)}_{:,k} - \mathbf{J}^{t}_{:,k})(s_k - t\delta)\tilde{1'}_\epsilon(s_k - t\delta) \\ &= \mathbf{J}^{(t)}_{:,k}(1 - \tilde{1}_\epsilon(s_k - t\delta) - (s_k - t\delta)\tilde{1'}_\epsilon(s_k - t\delta)) + \mathbf{J}^{(t+1)}_{:,k}(\tilde{1}_\epsilon(s_k - t\delta) + (s_k - t\delta)\tilde{1'}_\epsilon(s_k - t\delta)) \\ &\quad \because \tilde{1}_\epsilon(s_k - (t-1)\delta) = 1, \tilde{1}_\epsilon(s_k - (t+1)\delta) = 0 \\ &= \mathbf{J}^{(t)}_{:,k}(\tilde{1}_\epsilon(t\delta - s_k) + (t\delta - s_k)\tilde{1'}_\epsilon(t\delta - s_k)) + \mathbf{J}^{(t+1)}_{:,k}(\tilde{1}_\epsilon(s_k - t\delta) + (s_k - t\delta)\tilde{1'}_\epsilon(s_k - t\delta)) \\ &\quad \because \tilde{1}_\epsilon(s) + \tilde{1}_\epsilon(-s) = 1, \tilde{1'}_\epsilon(s) = \tilde{1'}_\epsilon(-s)\end{aligned}$$

Observe that $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k}$ is a linear combination of $\mathbf{J}^{(1)}_{:,k}, \mathbf{J}^{(2)}_{:,k}, ..., \mathbf{J}^{(p)}_{:,k} \quad \forall k \in [d]$. Since by Lemma B.2, $\bigcup_{i=1}^{p} \text{cols}(\mathbf{J}^{(i)})$ are all nonzero and linearly independent with respect to each other with probability 1, the columns of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ are linearly independent as long as they are all non-zero. Hence, the only possibility for $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ to not be full column-rank is for $k \in [d]$,

$$\begin{aligned}&\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} = \mathbf{0} \\ &\implies (\tilde{1}_\epsilon(t\delta - s_k) + (t\delta - s_k)\tilde{1'}_\epsilon(t\delta - s_k)) = 0, (\tilde{1}_\epsilon(s_k - t\delta) + (s_k - t\delta)\tilde{1'}_\epsilon(s_k - t\delta)) = 0 \\ &\quad \because \mathbf{J}^{(t)}_{:,k}, \mathbf{J}^{(t+1)}_{:,k} \text{ are linearly independent.}\end{aligned}$$

As in Theorem B.23, consider the function, $q : \mathbb{R} \to \mathbb{R}$ such that $q(s) = \tilde{1}_\epsilon(s) + s\tilde{1'}_\epsilon(s)$. Observe that $q(s) \geq 0$ for $s \geq 0$. Thus, for $(\tilde{1}_\epsilon(t\delta - s_k) + (t\delta - s_k)\tilde{1'}_\epsilon(t\delta - s_k)) = (\tilde{1}_\epsilon(s_k - t\delta) + (s_k - t\delta)\tilde{1'}_\epsilon(s_k - t\delta)) = 0$, we need that $s_k = t\delta$. At $s_k = t\delta$, $q(s_k - t\delta) = q(t\delta - s_k) = \frac{1}{2} \neq 0$. Hence, we have shown that $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,:,k} \neq \mathbf{0} \forall k \in [d]$, thereby $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}$ is full column-rank and $c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})$ is finite for all $\mathbf{s} \in \mathbb{B}$.

Hence,

$$\begin{aligned}
C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s}) &= \int_{\mathbb{R}^d} c_{\text{IMA}}\ p_\mathbf{s} d\mathbf{s} \\
&= \int_{\mathbb{I}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\ p_\mathbf{s} d\mathbf{s} + \int_{\mathbb{B}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\ p_\mathbf{s} d\mathbf{s} \\
&= \int_{\mathbb{I}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\ p_\mathbf{s} d\mathbf{s} + \Theta(\epsilon) \\
&\approx \int_{\mathbb{I}} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s})\ p_\mathbf{s} d\mathbf{s} \quad \text{for } \epsilon \text{ arbitrarily small.} \\
&\leq \max_{\mathbf{s} \in \mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s}) \int_{\mathbb{R}^d} p_\mathbf{s} d\mathbf{s} \\
&\leq \max_{\mathbf{s} \in \mathbb{R}^d} c_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, \mathbf{s}) \leq \delta\,,
\end{aligned}$$

with probability at least $1 - \min\left\{1, \exp(2\log d - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ by Theorem 5.1. Thus, $C_{\text{IMA}}(\tilde{\mathbf{f}}_\epsilon, p_\mathbf{s}) \leq \delta$ for $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.28 with (high) probability $\geq 1 - \min\left\{1, \exp(2logd - \kappa(m-1)\frac{\delta^2}{d^2})\right\}$ for $m \gg d$ where $\delta < \frac{1}{2}$ is arbitrarily small.

□

We have shown that for smoothened grid-wise affine maps, $\tilde{\mathbf{f}}_\epsilon : \mathbb{R}^d \to \mathbb{R}^m$ defined in B.28 where locally the columns of the Jacobian, $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}(\mathbf{s})$, are sampled independently from a spherically invariant distribution (statistical notion of independent influences), $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,1}(\mathbf{s}), \mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,2}(\mathbf{s}), ..., \mathbf{J}_{\tilde{\mathbf{f}}_\epsilon,d}(\mathbf{s}) \sim p_\mathbf{s}$, the IMA function class which formalizes the non-statistical notion of independent influences is "typical", i.e. the columns of $\mathbf{J}_{\tilde{\mathbf{f}}_\epsilon}(\mathbf{s})$ are close to orthogonal with high probability.