Robustness of Nonlinear Representation Learning

Simon Buchholz [1 2] **Bernhard Schölkopf** [1 2 3]

Abstract

We study the problem of unsupervised representation learning in slightly misspecified settings, and thus formalize the study of robustness of nonlinear representation learning. We focus on the case where the mixing is close to a local isometry in a suitable distance and show based on existing rigidity results that the mixing can be identified up to linear transformations and small errors. In a second step, we investigate Independent Component Analysis (ICA) with observations generated according to $x = f(s) = As + h(s)$ where A is an invertible mixing matrix and h a small perturbation. We show that we can approximately recover the matrix A and the independent components. Together, these two results show approximate identifiability of nonlinear ICA with almost isometric mixing functions. Those results are a step towards identifiability results for unsupervised representation learning for real-world data that do not follow restrictive model classes.

1. Introduction

One of the fundamental problems of data analysis is the unsupervised learning of representations. Modern machine learning algorithms excel at learning accurate representations of very complex data distributions. However, those representations are, in general, not related to the true underlying latent factors of variations that generated the data. Nevertheless, it is desirable to not only match the training distribution, but also to identify the underlying causal structure because such representations are expected to improve the downstream performance and improve explainability and robustness (Schölkopf et al., 2021).

Generally, we can only hope to learn the ground truth model in identifiable settings, i.e., when, up to certain symmetries, there is a unique model in the considered class generating the observations. The simplest example of such a setting is linear ICA where we observe linear mixtures of independent variables. Identifiability of linear ICA was shown under mild assumptions in Comon (1994). On the other hand, it is well known that for general nonlinear functions ICA is not identifiable (Hyvärinen & Pajunen, 1999). However, recently, a flurry of results was proved that showed various identifiability results under additional assumptions. Those results rely among others on restrictions of the function class (Taleb & Jutten, 1999; Horan et al., 2021; Gresele et al., 2021; Buchholz et al., 2022; Kivva et al., 2022), multi-environment data (Khemakhem et al., 2020; Hyvarinen et al., 2019), or interventional data (Ahuja et al., 2022; Squires et al., 2023; von Kügelgen et al., 2023; Buchholz et al., 2023), for a broader overview we refer to the recent survey Hyvärinen et al. (2023).

While those results brought significant progress to the field of Causal Representation Learning (CRL) they also generally require strong assumptions which will typically only hold approximately for real data. Thus, it is of great importance to understand the various robustness properties of such identifiability results. A first step in this direction is the investigation of whether the latent variables can still be approximately identified in settings where a small amount of misspecification is allowed, i.e., when the model assumptions are mildly violated. Our analysis here focuses on theoretical questions regarding identifiability in misspecified settings, but this has nevertheless profound implications for the empirical side because it clarifies what assumptions generate or do not generate useful learning signals that can be exploited by suitable algorithms. This is particularly important since there is still a lack of algorithms that uncover the true latent structure for complex data beyond toy settings.

In this work, we study the problem of approximate identifiability in a setting where the mixing function is close to a local isometry and the latent variables are independent (ICA). In particular, our results can be seen as a generalization of the identifiability result of Horan et al. (2021). The main contributions of this work can be summarized as follows:

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Tübingen AI Center, Tübingen, Germany [3]ELLIS Institute, Tübingen, Germany. Correspondence to: Simon Buchholz <sbuchholz@tue.mpg.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. We initiate the study of robustness for representation learning and clarify that some existing results are not robust to arbitrarily small amount of misspecification.

2. We show based on existing rigidity results from material science that we can identify the latent variables up to a linear transformation and a small perturbation when the mixing function f is close to a local isometry.

3. Then we carefully investigate slightly non-linear ICA, a setting of independent interest. We show that then the linear part of the mixing can be identified up to a small error, which depends on the strength of the perturbation.

4. Finally, we combine the previous two results to show robustness for representation learning problems with independent latent variables and a mixing function close to a local isometry.

The rest of this paper is structured as follows. In Section 2 we introduce the general setting of representation learning, ICA, and local isometries, and we then consider the robustness of identifiability in representation learning in Section 3. Our results on approximate linear identifiability for approximate local isometries can be found in Section 4 followed by our analysis of perturbed linear ICA and a combination of the two results in Section 5. We conclude in Section 6. In Appendix A we collect some notation. In this work, C denotes a generic constant that is allowed to change from line to line. We denote by $\overset{\mathcal{D}}{=}$ equality of the distributions of two random variables.

2. Setting

In this section, we introduce and motivate our main setting. We assume that we have d-dimensional latent variables S such that their distribution satisfies $\mathbb{P} \in \mathcal{P}$ for some class of probability distributions $\mathcal{P}$. For our analysis of local isometries we make the following mild assumption.

Assumption 2.1. We assume that the class of probability distributions $\mathcal{P}$ has bounded support $\Omega \subset \mathbb{R}^d$ with density lower and upper bounded.

In the second part of this work we focus on independent component analysis, i.e., we make the following assumption.

Assumption 2.2. All $\mathbb{P} \in \mathcal{P}$ have independent components, i.e., $\mathbb{P} = \bigotimes_{i=1}^d \mathbb{P}_i$ for some measures $\mathbb{P}_i$ on $\mathbb{R}$ and we furthermore assume that the $\mathbb{P}_i$ are non-Gaussian and have connected support.

The latent variables S are hidden, and we assume that we observe $X = f(S)$ for some mixing function $f \in \mathcal{F}(\Omega)$. Note that the distribution of X is then given by the push-forward $f_*\mathbb{P}$. Here $\mathcal{F}$ is some function class consisting of functions $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^D$ and we will always assume without further notice that f is a diffeomorphism on its image (i.e., injective, differentiable and with differentiable inverse on its domain which is a submanifold of $\mathbb{R}^D$). Our main interest concerns the function class of local isometries, i.e.,

$$\mathcal{F}_{\text{iso}} = \{f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^D : \; Df^\top(s)Df(s) = \text{Id} \quad \text{for all } s \in \Omega\} \tag{1}$$

for some connected domain $\Omega \subset \mathbb{R}^d$. The class of local isometries has attracted substantial attention in representation learning because it locally preserves the structure of the data, which is a desirable feature in many settings (Tenenbaum et al., 2000; Donoho & Grimes, 2003; Belkin & Niyogi, 2003). Closely connected notions like the restricted isometry property play a crucial role in signal processing (Candes & Tao, 2005), and several works show that making (parts) of neural networks isometric improves performance (Qi et al., 2020; Liu et al., 2021; Miyato et al., 2018). In Gresele et al. (2021) it is argued based on the independent mechanism principle that the closely related function class of orthogonal coordinate transformations is a natural function class for representation learning. Essentially, the argument relies on the well-known fact that two isotropic random vectors are very close to orthogonal up to an exponentially small probability. In Appendix J we show a construction of a family of random functions that becomes increasingly isometric as $D \to \infty$. Let us finally remark that the investigation of the inductive bias of VAEs in Rolínek et al. (2019); Zietlow et al. (2021); Reizinger et al. (2022) revealed that their loss function and architecture promote that the encoder implements an orthogonal coordinate transformation. The empirical success of VAEs on disentanglement tasks is therefore a further motivation to study theoretical properties of orthogonal coordinate transformations and local isometries.

For ICA with locally isometric mixing function the following identifiability result was shown.

Theorem 2.3 (Theorem 1, Horan et al. (2021))**.** *We assume that $\mathcal{P}$ satisfies Assumption 2.2. Suppose $X = f(S)$ where $S \sim \mathbb{P} \in \mathcal{P}$ and $f \in \mathcal{F}_{\text{iso}}$. If $X \overset{\mathcal{D}}{=} \tilde{f}(\tilde{S})$ for some $\tilde{f} \in \mathcal{F}_{\text{iso}}$ and $\tilde{S} \sim \tilde{\mathbb{P}} \in \mathcal{P}$, then $f = \tilde{f} \circ P$ for some linear map P which is a product of a permutation and reflections.*

Note that Horan et al. (2021) only claimed disentanglement, but the stronger version stated presently follows as sketched below and is also a special case of Theorem 2 of Buchholz et al. (2022). The proof of this result proceeds in two steps. First, we can identify f up to an orthogonal linear transformation A by general results. Indeed, $\tilde{f}^{-1} \circ f : \Omega \to \mathbb{R}^d$ is a local isometry and local isometries from $\Omega \subset \mathbb{R}^d \to \mathbb{R}^d$ are affine. This is a result first shown by Liouville in the 19th

century, a simple proof can be found in the recent review of Hyvärinen et al. (2023). A slightly different viewpoint showing the same result was given by Tenenbaum et al. (2000). After identifying S up to linear transformations, we can as a second step apply the standard identifiability result for linear ICA to identify A up to permutations and reflection (there is no scale ambiguity here because we consider local isometries).

We emphasize that this separation into two steps, where first identifiability up to linear transformations and then full identifiability are shown, is very common, e.g., a similar proof strategy has been applied for polynomial mixing functions (Ahuja et al., 2022), piecewise linear functions (Kivva et al., 2022), or for general mixing functions and interventional data (Buchholz et al., 2023).

3. Robustness and approximate Identifiabilility

Theorem 2.3 proves identifiability, i.e., all data representations agree up to permutation and reflections. However, this uses crucially that the assumptions are exactly satisfied, i.e., f is a true local isometry, the coordinates of S are perfectly independent, and we know the distribution of X exactly.

In real-world settings, typically none of these assumptions will hold exactly but at best approximately ('all models are wrong' (Box, 1976)). Thus, identifiability results can only be relevant for applications when they are robust, i.e., they continue to hold in some approximate sense when the assumptions are mildly violated. For, e.g., supervised learning or distribution learning, there is a large body of research (essentially the field of learning theory) addressing in particular the dependence on the sample size but also misspecification attracted considerable attention, e.g., White (1980; 1982). For representation learning there are much fewer results. The sample complexity of linear ICA algorithms has been studied, e.g., by Yau & Bresler (1992); Tichavsky et al. (2006); Wei (2015). Moreover, Zhang & Chan (2008) investigated ICA for a perturbed linear model, although mostly empirically and without stringent quantitative results. To the best of our knowledge, no robustness results are known for non-linear mixing functions f. The main focus of this work is to show a first robustness result for a misspecification of the function class, while we do not consider misspecification of independence of the ground truth samples and finite sample effects here.

Let us emphasize that it is an important question because not all identifiability results have a meaningful extension to slight misspecification. As a simple motivating example, consider analytic functions (i.e., functions that can be expressed as a power series) and smooth functions (i.e., functions that have arbitrary many derivatives). While those function classes are superficially similar, they exhibit profound differences: Analytic functions $f : \mathbb{R} \to \mathbb{R}$ have the property that there is a unique continuation of $f|_U$ from any arbitrarily small open set U to a maximal domain of definition (this can be seen as arbitrary far o.o.d. generalization) while for smooth functions virtually no information about $f|_{U^\complement}$ can be deduced from $f|_U$. This indicates that minor misspecification (analytic vs. smooth) can render identifiability results void, while the representation capacity remains almost the same and the difference can hardly be detected from data. To connect this more closely to identifiability in representation learning, consider, e.g., an identifiability result such that the following properties hold (a similar discussion can be found in the Appendix of Buchholz et al. (2023))

- When assuming that the mixing f is in some function class $\mathcal{F}$ then f is identifiable (up to certain symmetries, e.g., permutations or linear maps)
- When the mixing is just assumed to be an injective and continuous function f is not identifiable, i.e., for $f \in \mathcal{F}$ there is f' giving rise to the same observational distribution (for suitable latent distributions) but f and f' are entirely different data representations that are not related by the allowed symmetries.
- The function space $\mathcal{F}$ is dense in the space of injective continuous functions, i.e., every injective and continuous function f' can be approximated to arbitrary precision by $f \in \mathcal{F}$.

Note that the last property implies that for $f' \notin \mathcal{F}$ no meaningful distance to $\mathcal{F}$ can be defined, as it can be approximated arbitrarily well by functions in $\mathcal{F}$. This also makes the identifiability result brittle: While $f \in \mathcal{F}$ can be identified, there is a sequence of functions $\mathcal{F} \ni f_i \to f'$ converging to the spurious solution f'. Thus, it can be arbitrarily difficult to decide in practice whether f generated the data or an approximation $f_i \in \mathcal{F}$ of the spurious solution f'. Concrete examples of this setting include polynomial mixing functions and piecewise linear mixing functions which can be used to identify, e.g., Gaussian or rectangular base distributions (Ahuja et al., 2022; Kivva et al., 2022). We emphasize that these results nevertheless are important theoretical contributions, however, these identifiability results might be difficult to exploit in practice. Our observations here are in spirit similar to a no free lunch theorem (Wolpert & Macready, 1997) because it clarifies the fundamental tradeoff between representation capacity of the function class and identifiability. Note that the lack of identifiability can also harm downstream performance, as shown by Saengkyongam et al. (2023).

Let us try to make the point above concrete in a toy setting. We consider two-dimensional Gaussian latent variables Z.

Figure 1. (Left) Color map of Gaussian latent variable Z, (Center) Color map of the transformed data $X = f(Z)$ where f is a piecewise linear approximation of a radius dependent rotation (i.e., $f(Z) \overset{\mathcal{D}}{\approx} Z$), (Right) Representation $X' = f'(Z)$ learned by a VAE with ReLU activation functions initialized with f and f^{-1} for decoder and encoder and small variance.

Let f be a piecewise linear mixing function f that approximates a radius dependent rotation m (Hyvärinen & Pajunen, 1999) which is a map such that $m(Z) \overset{\mathcal{D}}{=} Z$. Consider observations $X = f(Z) \overset{\mathcal{D}}{\approx} Z$. Note that f and the identity map generate similar observations so it is hard to distinguish between the mixing function f and the identity id even though f is identifiable for Gaussian latent variables up to linear transformations (see Kivva et al. (2022)). However, if we train a VAE with ReLU activations that is initialized with f and f^{-1} as decoder and encoder, i.e., we initialize it with the true mixing, it nevertheless fails to preserve this representation. Instead, the VAE learns a simple orthogonal transformation of the data, i.e., even though the map f is in theory identifiable the learning signal is not sufficient to overcome the inductive bias of the learning method (here the VAE) towards simple functions that, e.g., preserve simple geometric properties. An illustration of the learned mixing can be found in Figure 1. Additional details for the setting can be found in Appendix H.

The main purpose of this work is to show that the function class $\mathcal{F}_{\text{iso}}$ behaves differently, i.e., robust representation learning results can be proved. To formalize this, we first need to introduce a way to measure the distance $\Theta(f)$ of a function f to the space of local isometries $\mathcal{F}_{\text{iso}}$. This quantity should vanish ($\Theta(f) = 0$) when $f \in \mathcal{F}_{\text{iso}}$ and our goal is to show that when $\Theta(f)$ is small we can still approximately identify f. For functions $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^d$ we use the following quantity

$$\Theta_p^p(f, \Omega) = \int_\Omega \mathrm{dist}(Df(s), \mathrm{SO}(d))^p + \mathrm{dist}\big((Df)^{-1}(s), \mathrm{SO}(d)\big)^p \, \mathrm{d}s. \tag{2}$$

Here $\mathrm{dist}^2(A, \mathrm{SO}(d)) = \min_{Q \in \mathrm{SO}(d)} |Q - A|^2$ refers to the euclidean distance to the space $\mathrm{SO}(d) = \{A \in \mathbb{R}^{d \times d} : AA^\top = \mathrm{Id}, \det A = 1\}$ of all rotations. We discuss the definition and properties of the distance in more detail in Appendix C. The distance Θ (which is not a distance in the mathematical sense) measures how close a function f is to a local isometry pointwise and integrates this. If $f \in \mathcal{F}_{\text{iso}}(\Omega)$ is a local isometry and if $\det Df > 0$ everywhere (which is no loss of generality by reflecting one coordinate), then indeed $\Theta_p(f, \Omega) = 0$ because $Df \in \mathrm{SO}(d)$ pointwise. For our results we need the second term in (2), i.e., we measure the local deviation from being an isometry of the map and its inverse. One could also define the metric by integrating with respect to the latent distribution $\mathbb{P}$ instead of the Lebesgue measure, but we think that the latter is slightly more natural because a common assumption based on the independent mechanism assumption (Janzing & Schölkopf, 2010; Schölkopf et al., 2012; Janzing et al., 2016) is that $\mathbb{P}$ and f are sampled independently. This also simplifies our analysis slightly.

Our main interest concerns high dimensional embeddings, i.e., mixing function $f : \mathbb{R}^d \to M \subset \mathbb{R}^D$ for $d \ll D$ where M is an embedded submanifold. In this case, the definition (2) needs to be slightly adapted, by essentially restricting the target of Df to the tangent space of M. For $D \geq d$ we define

$$\Theta_p^p(f, \Omega) = \int_\Omega \mathrm{dist}^p(Df(z), \mathrm{SO}(d, T_{f(z)}M)) + \mathrm{dist}^p\big((Df)^{-1}(z), \mathrm{SO}(T_{f(z)}M, d)\big) \, \mathrm{d}z \tag{3}$$

where $T_{f(z)}M$ denotes the tangent space of M at $f(z)$ and $\mathrm{SO}(d, T_{f(z)}M)$ denotes the set of orthogonal matrices $Q \in \mathbb{R}^{D \times d}$ (i.e., $Q^\top Q = \mathrm{Id}_d$) with range in $T_{f(z)}M$ that are orientation preserving (here we fix any orientation

on M). To interpret the second summand, we need to associate certain matrix representations to the maps $(Df)^{-1}$ and $\mathrm{SO}(T_{f(z)}M, d)$ and we refer to Appendix D for a careful definition.

Let us now discuss and motivate the choice of distance in a bit more detail. First, we emphasize that (2) does not (directly) quantify how well f is globally approximated by a local isometry, i.e., whether there exists a $g \in \mathcal{F}_{\text{iso}}$ such that, e.g., $\|f - g\|_1$ is small. We note that from a modelling standpoint, it seems more reasonable to assume that (2) is small for the ground truth mixing function than assuming that f can be globally well approximated by a local isometry because the former corresponds to the common assumption that f preserves the structure of the data locally (in particular distances) while it is not clear why f should be globally constrained. In Appendix J we show that for $d \ll D$ and certain classes of random functions Θ will be indeed small. This quantifies the observation that those random functions become isometric as $D \to \infty$ (similar results can be found in Reizinger et al. (2023)).

One difficulty in the setting is that as soon as we allow misspecification of f the mixing function f will typically be no longer identifiable (otherwise the existing identifiability result could be strengthened to a larger function class). We focus mostly on the setting of nonlinear ICA and in this case, it is well known that f is not identifiable. Thus, we can only hope to prove a version of approximate identifiability that gracefully recovers the classical identifiability result as the misspecification disappears. We here define approximate identifiability of the latent variables based on the mean correlation coefficient (MCC) which has been used before to evaluate the empirical performance of representation learning algorithms. For a pair of d dimensional random variables $(S, \tilde{S})$ the MCC is defined by

$$\mathrm{MCC}(S, \tilde{S}) = \max_{\pi \in S_d} d^{-1} \sum_{i=1}^{d} |\rho(S_i, \tilde{S}_{\pi(i)})| \tag{4}$$

where $\rho(X, Y) = \mathrm{Cov}(X, Y)/(\mathrm{Var}(X)\,\mathrm{Var}(Y))^{1/2}$ denotes the correlation coefficient. Note that $\mathrm{MCC}(S, \tilde{S}) = 1$ implies that $S_i = \lambda_i \tilde{S}_{\pi(i)}$, for a vector $\lambda \in \mathbb{R}^d$, i.e., we recover Z up to permutation and scale. For MCC values close to 1 this is approximately true, in particular $\tilde{S} = PS + \varepsilon$ holds for a scaled permutation P and some error ε that vanishes as the MCC goes to 1. More generally, one could also allow coordinatewise reparametrizations of $\tilde{S}$ when those are not identifiable (Gresele et al., 2021), but this is not necessary when considering (perturbed) local isometries.

Let us elaborate on the difference between proper identifiability (Xi & Bloem-Reddy, 2023) and approximate identifiability. Recall that a proper identifiability results is of the form $f_*\mathbb{P} = \tilde{f}_*\tilde{\mathbb{P}}$ for $f, \tilde{f} \in \mathcal{F}$ and $\mathbb{P}, \tilde{\mathbb{P}} \in \mathcal{P}$ implies that $\tilde{f}^{-1} \circ f$ is a tolerable ambiguity, e.g., a combination of permutation and reflection in Theorem 2.3. In this case, an equivalent statement is that $f_*\mathbb{P} = \tilde{f}_*\tilde{\mathbb{P}}$ implies $\mathrm{MCC}(\tilde{f}^{-1} \circ f(S), S) = 1$ for $S \sim \mathbb{P}$. In contrast, for approximate identifiability results we cannot simply take any $\tilde{f}$ such that $\tilde{f}_*\tilde{\mathbb{P}} = f_*\mathbb{P}$ because there will typically exist arbitrarily chaotic spurious solutions with no guarantee on $\mathrm{MCC}(\tilde{f}^{-1} \circ f(S), S)$ for $S \sim \mathbb{P}$. Instead, we need to make a specific choice for the function $\tilde{f}$ such that $\tilde{f}_*\tilde{\mathbb{P}} = f_*\mathbb{P}$ which ensures that $\mathrm{MCC}(\tilde{f}^{-1} \circ f(S), S) \approx 1$ for a suitable class of f. In our case those will be functions such that $\Theta_p(f, \Omega)$ is small. Of course, our choice of $\tilde{f}$ is only allowed to depend on the observations $X = f(S)$ but not on f itself. Here the basic idea is to choose $\tilde{f}$ as close as possible to being a local isometry, i.e., we roughly chose $\tilde{f}$ such that $\Theta(\tilde{f}, \tilde{\Omega})$ is minimal under the constraint $\tilde{f}_*\tilde{\mathbb{P}} = f_*\mathbb{P}$. Let us now provide a very informal version of our final approximate identifiability result.

Theorem 3.1 (Informal sketch)**.** *Suppose $S \sim \mathbb{P}$ where $\mathbb{P}$ has independent components, and we observe $X = f(S)$ for some mixing function f. Then we can find $\tilde{f}$ such that $\hat{S} = \tilde{f}^{-1}(X)$ satisfies for some $C > 0$, $p > 1$*

$$\mathrm{MCC}(\hat{S}, S) \geq 1 - C\Theta_p^2(f) \tag{5}$$

where C depends on everything except f.

Note that if f is far away from a local isometry, i.e., $\Theta_p(f)$ is large then the right-hand side is negative, and the obtained recovery guarantee is void which is not surprising as nonlinear ICA is not identifiable. The actual statement can be found in Theorem 5.11. We emphasize again that robustness is an aspect of representation learning that has previously attracted little attention, and we study one specific example, namely robustness of learning approximate local isometries. However, there are many alternative robustness questions and possible misspecifications, e.g, we could consider latent sources that are only approximately independent or settings where MCC might not be the right measure of identifiability.

4. Approximate linear Identifiability for approximate local Isometries

We now extend linear identifiability for locally isometric embeddings to approximate isometric embeddings. Restricting the ambiguity to a linear transformation is already an important first step which can be in principle combined with any result on causal representation learning for linear mixing functions, examples of recent results for linear mixing functions include Ahuja et al. (2023); Squires et al. (2023); Varici et al. (2023). The key ingredient for our results is the rigidity statement Theorem E.1 in Appendix E which played an important role in the mathematical analysis of elastic materials (see Ciarlet (1997) for an overview).

The main consequence of this theorem that is relevant for

our work here is the bound

$$\min_L \|u - L\|_{L^q(\Omega)} \leq C(\Omega, p) \left(\int_\Omega \operatorname{dist}(Du(s), \mathrm{SO}(d))^p \, \mathrm{d}s \right)^{\frac{1}{p}} \tag{6}$$

for all u and $q = pd/(d-p)$ where the minimum is over functions of the form $L(s) = As + b$ with $A \in \mathrm{SO}(d)$. This result shows that when a map u has gradient pointwise close to any, potentially varying, rotation, it is globally close to an affine map. We emphasize that this is highly non-trivial, naively one would expect that the rotation could change with s. Note that when the right hand-side is zero, i.e., f is a local isometry then we conclude that u is affine recovering the fact that local isometries from $\mathbb{R}^d$ to $\mathbb{R}^d$ are already affine which is the key observation underlying Theorem 2.3. This rigidity property of almost isometries renders them almost identifiable. This then generalizes Theorem 2.3.

For simplicity, we mostly present our results for the case where $D = d$ in the main paper. The additional technicalities for $d < D$ (which is the main interest of our results) will be discussed in Appendix D. For a measure $\mathbb{P} \in \mathcal{P}$ with support $\Omega \subset \mathbb{R}^d$ we introduce the set of models

$$\mathcal{M}(f_*\mathbb{P}) = \{(g, \mathbb{Q}, \Omega') : g \in \mathcal{F}(\Omega'), \text{ where } g_*\mathbb{Q} = f_*\mathbb{P}, \operatorname{supp}(\mathbb{Q}) = \Omega'\} \tag{7}$$

that generate the observed distribution of $X = f(S)$. We now consider for a fixed p the triple

$$(g, \mathbb{Q}, \Omega') \in \operatorname*{argmin}_{(\bar{g}, \bar{Q}, \bar{\Omega}) \in \mathcal{M}(f_*\mathbb{P})} \int_{\bar{\Omega}} \operatorname{dist}((D\bar{g})^{-1}(g(s)), \mathrm{SO}(d))^p \, \bar{\mathbb{Q}}(\mathrm{d}s), \tag{8}$$

i.e., we pick $(g, \mathbb{Q}, \Omega')$ such that $X \sim g_*\mathbb{Q}$ so that it generates the observational distribution and in addition make its inverse as isometric as possible. For $D > d$ we replace the distance to $\mathrm{SO}(d)$ by the distance to $\mathrm{SO}(T_{g(s)}M, d)$ (see (73) in Appendix E for details). We emphasize that $\mathcal{M}(f_*\mathbb{P})$ and therefore g only depends on the observational distribution $f_*\mathbb{P}$ but not on f directly. Note that we could also minimize a variant of $\Theta(g, \Omega')$ where we integrate with respect to $\mathbb{Q}$. We remark in passing that we do not prove the existence of a minimizer of (8) (although this should be possible using the direct method of the calculus of variations). Since the functional is lower bounded, a minimizing sequence g_n exists and we can instead approximate the infimum up to arbitrarily small ε adding an arbitrarily small additional error term to Theorem 4.1 below (see Appendix E for details).

Theorem 4.1. *Suppose we have a latent distribution $\mathbb{P} \in \mathcal{P}$ satisfying Assumption 2.1 with support $\Omega \subset \mathbb{R}^d$ where Ω is a bounded connected Lipschitz domain. The observational distribution is given by $X = f(S) \in M \subset \mathbb{R}^D$ where $S \sim \mathbb{P}$ and $f \in \mathcal{F}(\Omega)$. Fix a $1 < p < \infty$. Define $(g, \mathbb{Q}, \Omega')$ as in (8) (as in (73) in Appendix D in the undercomplete case). Then there is $A \in \mathrm{SO}(d)$ such that $g^{-1} \circ f(s) = As + h(s) + b$ and h satisfies for $p < d$ and $q = pd/(d-p)$ the bound*

$$\|h\|_{\mathbb{P},q} \leq C_1 \Theta_p(f, \Omega). \tag{9}$$

Here C_1 is a constant depending on d, p, Ω, and the lower and upper bound on the density of $\mathbb{P}$.

The proof of this result, including the extension to the undercomplete case can be found in Appendix E. Let us continue our discussion on the meaning of approximate identifiability in the context of this theorem. First, we note that if $\Theta_p(f, \Omega)$ is not small we obtain no useful statement, except that our transformed data is some function of the original data. This is not surprising: We cannot hope to recover any mixing function because the problem of learning f from $f_*\mathbb{P}$ is not identifiable (even when assuming $\mathbb{P}$ known). Moreover, our statement only applies to one specific unmixing g which is again unavoidable for the same reason. What we show is that if f and g are both close to being locally isometric, then the concatenation $g^{-1}f$ will be close to a linear (even orthogonal) map. Note that g does not appear on the right-hand side of (9) because we choose g to be the maximally isometric representation of our observations, but we know that this representation is more isometric than any alternative representation, in particular more isometric than f. This allows us to bound the non-linearity h in terms of f only. While this result does not have the simplicity of a standard identifiability result it provides a more general viewpoint. Indeed, if $\Theta_p(f, \Omega) = 0$, i.e., when f is a local isometry we have $h = 0$ and we recover f up to a linear transformation which is the standard identifiability result for local isometries (see Theorem 2 of Horan et al. (2021)). Our result extends this gracefully to functions that are approximate local isometries in the sense that $\Theta_p(f, \Omega)$ is small. Let us add some remarks about this result.

Remark 4.2.
- The optimization problem for g in (8) is non-convex, and difficult to optimize in practice.
- For $D = d$ the introduction of g is not necessary, instead we can directly apply Theorem E.1 to f and just work with the original data X and directly apply Theorem E.1 to $X = f(S)$.
- The assumptions that Ω is connected and that the density of $\mathbb{P}$ is lower bounded are necessary. In particular the result does not apply to distributions $\mathbb{P}$ with disconnected support.
- There are alternative assumptions that allow us to remove (or bound) the second term in (2), e.g., assuming

$Df^\top Df > c_1 > 0$, i.e., the smallest singular value of Df is bounded below is sufficient (see Lemma C.2).

5. Perturbed linear ICA

In this section, we consider the problem of independent component analysis where the mixing is a slight perturbation of a linear function. This is a problem of general interest beyond the main setting considered in this paper because in typical real-world applications the mixing will only be approximately linear, so understanding the effect of the non-linear part is important. Concretely, we assume that data is generated by a perturbed linear model

$$x = f(s) = As + \eta h(s) \tag{10}$$

where $h : \mathbb{R}^d \to \mathbb{R}^d$ is a non-linear function and $\eta \in \mathbb{R}$ is a small constant. We can assume that h is centered, i.e., $\mathbb{E}(h(S)) = 0$ and that A is the regression matrix when regressing X on S. This is equivalent to the relation $\mathbb{E}_S(Sh(S)^\top) = 0_{d\times d}$.

Remark 5.1. We decided to express the nonlinear part of the model as ηh which allows us to consider $\eta \to 0$ to shrink the size of the perturbation. As we check carefully in the proofs, all bounds only depend on h through its norm $\|h\|_{\mathbb{P},q}$ for some q specified below.

It is clear that f cannot be identified from the distribution of X as the mixing is nonlinear. Instead, we investigate how well the linear part given by the matrix A can be recovered and to what degree we can recover the ground truth sources S. Our results show that as $\eta \to 0$ we recover the identifiability results for linear ICA. One issue that creates substantial theoretical problems in this work and in general is that there is a gap between the statistical and computational aspects of ICA, in the sense that linear ICA is identifiable for non-Gaussian latent variables, however ICA algorithms typically require a stronger non-degeneracy condition on the latent distribution than non-Gaussianity of the latent variables. In the misspecified setting we need to define a (not necessarily computationally feasible) algorithm to pick a mixing function for the observed data, which then results in the same limitations as the conventional algorithms face.

Let us now quickly review how the independent components can be identified. Most ICA algorithms consider a function $H : S^{d-1} \to \mathbb{R}$ defined by

$$H(w) = \mathbb{E}G\left(w^\top \Sigma_X^{-\frac{1}{2}} X\right) \tag{11}$$

where Σ_X denotes the covariance matrix of X so that $\Sigma_X^{-\frac{1}{2}} X$ is whitened and G is the so-called contrast function. Then, under a suitable degeneracy assumption, the independent components correspond to an orthogonal set of local extrema of H.

Indeed, as a motivation for the nonlinear case we discuss in Appendix F the well-known calculation that for linear mixing functions (i.e., $\eta = 0$) H has a local extremum if $w^\top \Sigma_X^{-\frac{1}{2}} X = S_i$ for some i, i.e., when $A^\top \Sigma_X^{-\frac{1}{2}} w = e_i$. Note that for linear mixing A the relation $\Sigma_X = AA^\top$ holds, and the vectors

$$\bar{w}_i = (AA^\top)^{\frac{1}{2}} A^{-\top} e_i \tag{12}$$

thus satisfy $\bar{w}_i^\top \Sigma_X^{-\frac{1}{2}} X = S_i$ and $|\bar{w}_i| = 1$. We also consider the matrix $\bar{W} = A^{-1}(AA^\top)^{\frac{1}{2}}$ which has rows $\bar{w}_i = \bar{W}^\top e_i$ and satisfies $\bar{W}\Sigma_X^{-\frac{1}{2}} X = S$ for $\eta = 0$.

The main goal of this section is to show that this general picture remains approximately true in the perturbed setting, i.e., under minor regularity assumptions on G there is a matrix W close to $\bar{W}$ such that its rows are local extrema of H. Let us collect the necessary assumptions for our results.

Assumption 5.2. The function G is even, three times differentiable, and there are constants C_g and d_g such that for $k \le 3$

$$|G^{(k)}(x)| \le C_g(1 + |x|)^{\max(d_g - k, 0)} \tag{13}$$

where $G^{(k)}$ denotes the k-th derivative of G.

We will write $g = G'$ from now on following the notation in the field. Note that commonly used contrast functions like $G = \ln \circ \cosh$ or $G = |\cdot|^4$ satisfy this assumption. We need some regularity assumption on the source variables.

Assumption 5.3. Write $q = \max(d_g, 3)$. Assume that the latent sources S satisfy for some constant $M < \infty$

$$\mathbb{E}(|S|^q) = M \tag{14}$$

Finally, we need a condition that ensures that the contrast function can single out the latent variable S_i which is also necessary in the linear case.

Assumption 5.4. The latent variables S satisfy for an $\alpha_i \neq 0$

$$\mathbb{E}(S_i g(S_i) - g'(S_i)) = \alpha_i. \tag{15}$$

Then we have the following result.

Theorem 5.5. *Suppose that X and S are random variables satisfying the relation (10) and the distribution $\mathbb{P}$ of S has independent components (see Assumption 2.2) and $\mathbb{E}(S) = 0$ and $\mathbb{E}(S_i^2) = 1$. Assume G is an even contrast function that satisfies the pointwise bounds in Assumption 5.2 and that S and G satisfy Assumptions 5.3 and 5.4 for some $1 \le i \le d$. Assume that $\|h\|_{\mathbb{P},q} \le 1$ where $q = \max(d_g, 3)$ is defined in Assumption 5.3. Let $\bar{w}_i$ be defined as in (12). Then, there is an $\eta_0 > 0$ (depending on d, α_d, A, M, C_g, and q but not*

on h) such that for $\eta < \eta_0$ there is a local extremum w_i of H close to $\bar{w}_i$ satisfying $|\bar{w}_i - w_i| = O(\eta)$. In addition, H is strictly convex or concave (depending on the sign of α_i) in a neighborhood $B_\kappa(\bar{w}_i)$ with κ independent of $\eta < \eta_0$.

Remark 5.6. In fact, we can characterize w_d a bit more precisely by

$$w_d = \Sigma_X^{\frac{1}{2}} A^\top \begin{pmatrix} \eta \alpha_d^{-1} v \\ 1 \end{pmatrix} + O(\eta^2) = \tilde{w}_d + O(\eta^2) \quad (16)$$

where $v_i = \mathbb{E}((A^{-1}h)_d(S)g'(S_d)S_i) + \mathbb{E}((A^{-1}h)_i g(S_d))$ for $i < d$ and $\alpha_d = \mathbb{E}(S_d g(S_d) - g'(S_d))$ as in (15) and similar expressions hold for $i < d$.

The proof can be found in Appendix F. We can apply the result to all coordinates simultaneously and obtain the following slight generalization.

Theorem 5.7. *Under the same assumptions as in Theorem 5.5 where we assume that the Assumption 5.4 holds for all S_i with $1 \leq i \leq d$ there is a matrix W whose rows $w_i = W^\top e_i$ satisfy $|w_i| = 1$, w_i is a local extremum of H, $A^\top \Sigma_X^{-\frac{1}{2}} w_i = e_i + O(\eta)$, and $|W - \bar{W}| = O(\eta)$.*

The main message of these results is that all standard ICA algorithms based on a contrast function G can extract w_i approximately in this setting, and the resulting matrix W is close to $\bar{W}$ the ground truth unmixing of the linear part. Let us clarify this through the example of gradient-based algorithms.

Corollary 5.8. *Under the same assumptions as in Theorem 5.5 and for $\eta < \eta_0$ gradient ascent (if $\alpha_i > 0$) or descent (if $\alpha_i < 0$) with sufficiently small step size will converge to w_i locally.*

This result is a direct consequence of Theorem 5.5 and convergence of gradient descent for strictly convex functions. We provide some more details in Appendix F. The results above established that we can approximately find A as $\eta \to 0$. Next we show that we do not only approximately recover the linear mixing, but we can also approximately recover the sources S.

Theorem 5.9. *Consider the same assumptions as in Theorem 5.7 and let W be (for $\eta < \eta_0$) the matrix as in the statement of Theorem 5.7. Define $\hat{S} = W\Sigma_X^{-\frac{1}{2}} X$. Then there is a constant C_2 depending on d, α_i, A, M, C_g, and q such that*

$$\mathrm{MCC}(\hat{S}, S) \geq 1 - C_2 \eta^2. \quad (17)$$

The proof of this result can be found in Appendix F. It is based on a general bound for the MCC stated in Lemma B.2 in Appendix B. We provide some experimental evidence that the derived bounds have the optimal scaling in η in Appendix I.

Remark 5.10. The results might appear a bit unsatisfactory because we only show existence of local extrema of H that allow us to recover the sources approximately and local convergence of gradient descent or ascent to these extrema. However, we do not claim that these are the only extrema or there is an efficient algorithm to find all extrema in polynomial time. This problem is, however, not a feature of the perturbed setting, but the same issue arises already for linear ICA (Hyvarinen, 1999). Note that for the special case of the kurtosis-based contrast function $G(s) = s^4 - 3$ with linear mixing it can be shown (under additional assumptions) that the vectors $\bar{w}_i$ are the only local maxima (or minima) of H and they can be found in polynomial time as shown by Arora et al. (2012). We expect that this result generalizes to our perturbed setting.

Finally, we combine the two parts of the analysis to show approximate identifiability of ICA with approximately locally isometric mixing. To simplify the analysis we assume that the support of $\mathbb{P}$ is compact. Then we can combine our previous results to obtain the following theorem.

Theorem 5.11. *Assume that $S \sim \mathbb{P}$ with bounded and connected support Ω satisfies Assumptions 2.1 and 2.2. Suppose observations are given by $X = f(S) \in M \subset \mathbb{R}^D$ for some $f \in \mathcal{F}(\Omega)$. Assume that $\mathbb{P}$ satisfies Assumption 5.2, 5.3, and 5.4 for some contrast function G. Let $q = \max(d_g, 3)$ and $p = dq/(d+q)$. Define $(g, \mathbb{Q}, \Omega')$ as in (8) (for $d = D$ and as in (73) in Appendix D for $d < D$) and shift g such that $\tilde{X} = g(X)$ is centered. Define $\hat{S} = W\Sigma_{\tilde{X}}^{-\frac{1}{2}} \tilde{X}$ as in Theorem 5.7 for observations $\tilde{X}$. Then the following bound holds for $\Theta_p(f, \Omega)$ sufficiently small and some $C_3 > 0$*

$$\mathrm{MCC}(\hat{S}, S) \geq 1 - C_3 \Theta_p^2(f, \Omega). \quad (18)$$

Here C_3 depends on C_1 and C_2 from Theorem 4.1 and 5.9 and the bounds on the density of $\mathbb{P}$.

The informal version of this theorem is that by learning a maximally locally isometric transformation of our data followed by running a linear ICA-algorithm we can approximately recover the true sources if the ground truth mixing is close to a local isometry. A proof of this theorem can be found in Appendix G. Let us add a few remarks about this result.

- For local isometries ($\Theta_p(f, \Omega) = 0$) we essentially recover Theorem 2 in (Horan et al., 2021).
- The assumption that $\mathbb{P}$ has bounded support combined with the independence assumption implies that $\mathbb{P}$ is supported on a cuboid. It was shown in (Ahuja et al., 2023; Roth et al., 2023) that then linear transformations of $\mathbb{P}$ are identifiable (up to scaling and permutation) and we expect that this can be generalized to perturbed

linear maps, but it is non-trivial to prove this and to obtain a quantitative statement as we obtain above.

- In Appendix G we also show how to relax the condition that $\mathbb{P}$ has compact support, which comes at a price of additional technical difficulties.

6. Conclusion

In this paper, we initialized the theoretical investigation of robustness results for representation learning with misspecified models. While we showed robustness results for linear ICA and for almost isometric embeddings there are several closely related questions, of which we name a few. First, it would be of interest to show for the perturbed linear ICA model finite sample guarantees, local convergence guarantees for algorithms not relying on gradient descent, e.g., for the fast-ICA algorithm, and analyze the effects of additive noise. Secondly, one limitation of this work is that we assume that the observational distribution $f_*\mathbb{P}$ and the manifold it is supported on are exactly known in the undercomplete case. Note that by the Nash embedding theorem (Nash, 1954) we can approximate uniformly any submanifold by a sequence of local isometries. Thus, here additional regularization by, e.g., penalizing higher order derivatives is necessary to extend robustness results to, e.g., finite sample size or noisy settings. The broader questions that this work motivates is what class of identifiability results are brittle and which are robust to misspecification (or finite sample bias). A deeper understanding of this landscape (and its intersection with real-world datasets) will help to understand which identifiability results give rise to useful learning signals, and thus provide some guidance for the development of novel algorithms for (causal) representation learning.

Acknowledgements

This work was supported by the Tübingen AI Center.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adams, R. and Fournier, J. *Sobolev Spaces*. ISSN. Elsevier Science, 2003. ISBN 9780080541297. URL https://books.google.de/books?id=R5A65Koh-EoC.

Ahuja, K., Mahajan, D., Syrgkanis, V., and Mitliagkas, I. Towards efficient representation identification in supervised learning. *arXiv preprint arXiv:2204.04606*, 2022.

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. Interventional causal representation learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Arora, S., Ge, R., Moitra, A., and Sachdeva, S. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

Boumal, N. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164.

Box, G. E. P. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. ISSN 01621459. URL http://www.jstor.org/stable/2286841.

Buchholz, S., Besserve, M., and Schölkopf, B. Function classes for identifiable nonlinear independent component analysis. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=DpKaP-PY8bK.

Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., and Ravikumar, P. Learning linear causal representations from interventions under general nonlinear mixing, 2023.

Candes, E. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979.

Ciarlet, P. G. *Mathematical Elasticity: Volume II: Theory of Plates*. ISSN. Elsevier Science, 1997. ISBN 9780080535913.

Comon, P. Independent component analysis, a new concept? *Signal Processing*, 36 (3):287–314, 1994. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0165168494900299. Higher Order Statistics.

Conti, S. and Schweizer, B. Rigidity and gamma convergence for solid-solid phase transitions with so(2) invariance. *Communications on Pure and Applied Mathematics*, 59(6):830–868, 2006. doi: https://doi.org/10.1002/cpa.20115. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20115.

Donoho, D. L. and Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003. doi: 10.1073/pnas.1031596100. URL https://www.pnas.org/doi/abs/10.1073/pnas.1031596100.

Friesecke, G., James, R. D., and Müller, S. A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity. *Communications on Pure and Applied Mathematics*, 55 (11):1461–1506, 2002. doi: https://doi.org/10.1002/cpa.10048. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.10048.

Gresele, L., Von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 34, 2021.

Horan, D., Richardson, E., and Weiss, Y. When is unsupervised disentanglement possible? In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5150–5161. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/29586cb449c90e249f1f09a0a4ee245a-Paper.pdf.

Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. doi: 10.1109/72.761722.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5): 411–430, 2000.

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

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

Hyvärinen, A., Khemakhem, I., and Monti, R. Identifiability of latent-variable and structural-equation models: from linear to nonlinear, 2023.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. doi: 10.1109/TIT.2010.2060095.

Janzing, D., Chaves, R., and Schölkopf, B. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9):093052, sep 2016. doi: 10.1088/1367-2630/18/9/093052. URL https://dx.doi.org/10.1088/1367-2630/18/9/093052.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

Kohn, R. V. New integral estimates for deformations in terms of their nonlinear strains. *Archive for Rational Mechanics and Analysis*, 78(2):131–172, 1982.

Kollo, T. and von Rosen, D. *Advanced Multivariate Statistics with Matrices*. Mathematics and Its Applications. Springer Netherlands, 2006. ISBN 9781402034190. URL https://books.google.de/books?id=27mA99VQTi4C.

Lee, J. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York, 2013. ISBN 9780387217529. URL https://books.google.de/books?id=w4bhBwAAQBAJ.

Liu, W., Lin, R., Liu, Z., Rehg, J. M., Paull, L., Xiong, L., Song, L., and Weller, A. Orthogonal over-parameterized training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7251–7260, June 2021.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Nash, J. C1 isometric imbeddings. *Annals of Mathematics*, 60(3):383–396, 1954. ISSN 0003486X. URL http://www.jstor.org/stable/1969840.

Qi, H., You, C., Wang, X., Ma, Y., and Malik, J. Deep isometric learning for visual recognition. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7824–7835. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/qi20a.html.

Reizinger, P., Gresele, L., Brady, J., von Kügelgen, J., Zietlow, D., Schölkopf, B., Martius, G., Brendel, W., and Besserve, M. Embrace the gap: Vaes perform independent mechanism analysis, 2022. URL https://arxiv.org/abs/2206.02416.

Reizinger, P., Li, H.-B., Ravuri, A., Huszár, F., and Lawrence, N. D. Independent mechanism analysis in GPLVMs. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023. URL https://openreview.net/forum?id=WgK0RYP-6H.

Rolínek, M., Zietlow, D., and Martius, G. Variational autoencoders pursue PCA directions (by accident). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 12406–12415. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01269. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Rolinek_Variational_Autoencoders_Pursue_PCA_Directions_by_Accident_CVPR_2019_paper.html.

Roth, K., Ibrahim, M., Akata, Z., Vincent, P., and Bouchacourt, D. Disentanglement of correlated factors via hausdorff factorized support. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, May 2023. URL https://openreview.net/pdf?id=OKcJhpQiGiX.

Saengkyongam, S., Rosenfeld, E., Ravikumar, P., Pfister, N., and Peters, J. Identifying representations for intervention extrapolation, 2023.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1255–1262, New York, NY, USA, 2012. Omnipress.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. arXiv:2102.11107.

Squires, C., Seigal, A., Bhate, S. S., and Uhler, C. Linear causal disentanglement via interventions. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32540–32560. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/squires23a.html.

Taleb, A. and Jutten, C. Source separation in post-nonlinear mixtures. *IEEE Trans. Signal Process.*, 47(10):2807–2820, 1999. doi: 10.1109/78.790661. URL https://doi.org/10.1109/78.790661.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

Tichavsky, P., Koldovsky, Z., and Oja, E. Performance analysis of the fastica algorithm and crame´r-rao bounds for linear independent component analysis. *Trans. Sig. Proc.*, 54(4):1189–1203, oct 2006. ISSN 1053-587X. doi: 10.1109/TSP.2006.870561. URL https://doi.org/10.1109/TSP.2006.870561.

Varici, B., Acarturk, E., Shanmugam, K., Kumar, A., and Tajer, A. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

von Kügelgen, J., Besserve, M., Liang, W., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions, 2023.

Wei, T. A convergence and asymptotic analysis of the generalized symmetric fastica algorithm. *IEEE Transactions on Signal Processing*, 63(24):6445–6458, 2015. doi: 10.1109/TSP.2015.2468686.

White, H. Using least squares to approximate unknown regression functions. *International Economic Review*, 21 (1):149–170, 1980. ISSN 00206598, 14682354. URL http://www.jstor.org/stable/2526245.

White, H. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

Wolpert, D. and Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

Xi, Q. and Bloem-Reddy, B. Indeterminacy in generative models: Characterization and strong identifiability. In

Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6912–6939. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/xi23a.html.

Yau, S. and Bresler, Y. A compact cramer-rao bound expression for parametric estimation of superimposed signals. *IEEE Transactions on Signal Processing*, 40(5):1226–1230, 1992. doi: 10.1109/78.134484.

Zhang, K. and Chan, L. Minimal nonlinear distortion principle for nonlinear independent component analysis. *Journal of Machine Learning Research*, 9(81):2455–2487, 2008. URL http://jmlr.org/papers/v9/zhang08b.html.

Zietlow, D., Rolínek, M., and Martius, G. Demystifying inductive biases for (beta-)vae based architectures. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12945–12954. PMLR, 2021. URL http://proceedings.mlr.press/v139/zietlow21a.html.

Supplementary Material

Here we collect additional results, details, and proofs for the main paper. This appendix is structured as follows. We first introduce some notation and definitions used throughout the paper in Appendix A then we state and prove a lemma that can be used to bound the MCC in our setting in Appendix B. We gather some elementary results about the distance of matrices to the orthogonal group in Appendix C, and we extend these results to mappings to submanifolds $M \subset \mathbb{R}^D$ and define our measure of non-isometry also in this setting in Appendix D. In Appendix E we prove our results on approximate linear identifiability for approximate local isometries, followed by our careful analysis of perturbed linear ICA in Appendix F. The proof of our combination of the two results can be found in Appendix G. In Appendix H we provide the missing details for the experimental illustration of non-robustness and in Appendix I we confirm empirically the theoretical convergence rates which we derived for perturbed linear ICA. Finally, in Appendix J we show that a certain class of random functions becomes increasingly isometric as the ambient dimension grows.

A. Notation and Mathematical Definitions

In the paper we use notation from different fields. Let us collect a few definitions and references to the literature.

Linear Algebra We denote by $|A| = |A|_F$ the Frobenius norm on matrices, i.e., $|A|^2 = \operatorname{Tr} AA^\top = \sum_{i,j} |A_{ij}|^2$. We will frequently use unitary invariance $|X|_F = |QX|_F$ for $Q \in \mathrm{SO}(d)$ of the Frobenius norm. The symbol $A^{-\top} = (A^{-1})^\top$ denotes the inverse of the transpose.

Probability Theory We denote by $f_*\mathbb{P}$ the pushforward of a measure $\mathbb{P}$ along f. It is defined by $f_*\mathbb{P}(A) = \mathbb{P}(f^{-1}A)$ and the distribution of $f(Z)$ is $f_*\mathbb{P}$ if $Z \sim \mathbb{P}$. We denote by $\|f\|_{\mathbb{P},p} = (\mathbb{E}_{\mathbb{P}}|f||^p)^{1/p}$ the L^p norm.

Sobolev spaces For differentiable functions $u : \Omega \to \mathbb{R}$ we define the Sobolev norm

$$\|u\|^p_{W^{1,p}(\Omega)} = \int_\Omega |u|^p(s) + |Du|^p(s)\,\mathrm{d}s. \tag{19}$$

The Sobolev embedding $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ is continuous for $q = pd/(d-p)$. For details and a proper definition of Sobolev spaces we refer to Adams & Fournier (2003). We denote by $\|f\|_p$ the usual L^p norm with respect to the Lebesgue measure.

Differential Geometry We consider mixing functions $f : \mathbb{R}^d \to M \subset \mathbb{R}^D$ where M is an embedded submanifold. In this case, we need some notions from differential geometry such as the tangent space T_pM whose definitions can be found in, e.g., Lee (2013).

B. Bounding MCC for almost Permutations

Here, we provide a Lemma that allows us to control the MCC between two data representations when the transformation is perturbed linear and the linear part is close to a permutation matrix.

Lemma B.1. *Let $\mathbb{P}$ be a centered probability measure. Assume that for $Z \sim \mathbb{P}$ the bound*

$$\max_{i,j} \frac{\|Z_i\|_{\mathbb{P},2}}{\|Z_j\|_{\mathbb{P},2}} \leq c_1 \tag{20}$$

holds for some constant c_1. Let $T(z) = Az + h(z)$ where A is a matrix such that there is a permutation ρ and a constant c_2 with

$$\max_i \frac{1}{|A_{i,\rho(i)}|} \sum_{j \neq \rho(i)} |A_{i,j}| \leq c_2 \tag{21}$$

and h has the property that for some constant c_3

$$\frac{\|h_i\|_{\mathbb{P},2}}{|A_{i,\rho(i)}|\|Z_{\rho(i)}\|_{\mathbb{P},2}} \leq c_3. \tag{22}$$

Then the bound

$$\mathrm{MCC}(T(Z), Z) \geq 1 - 2c_3 - 2c_1c_2 \tag{23}$$

holds.

Proof. We lower bound the expression (4) for the permutation $\pi = \rho^{-1}$ in the statement of the lemma. Let $Z \sim \mathbb{P}$ and $\tilde{Z} = T(Z)$. Consider the correlation coefficients

$$\rho(Z_{\rho(i)}, \hat{Z}_i) = \frac{\mathrm{Cov}(Z_{\rho(i)}, \hat{Z}_i)}{\sigma_{Z_{\rho(i)}}\sigma_{\hat{Z}_i}}. \tag{24}$$

We find, using that $Z_{\rho(i)}$ is centered

$$\begin{aligned}
\left|\mathrm{Cov}(Z_{\rho(i)}, \hat{Z}_i) - A_{i,\rho(i)}\mathbb{E}(Z^2_{\rho(i)})\right| &= \left|\mathbb{E}(Z_{\rho(i)}\hat{Z}_i) - A_{i,\rho(i)}\mathbb{E}(Z^2_{\rho(i)})\right| \\
&= \left|\sum_j A_{i,j}\mathbb{E}(Z_{\rho(i)}Z_j) + \mathbb{E}(Z_{\rho(i)}h_i(Z)) - A_{i,\rho(i)}\mathbb{E}(Z^2_{\rho(i)})\right| \\
&= \left|\sum_{j\neq\rho(i)} A_{i,j}\mathbb{E}(Z_{\rho(i)}Z_j) + \mathbb{E}(Z_{\rho(i)}h_i(Z))\right| \\
&\leq \|Z_{\rho(i)}\|_{\mathbb{P},2}\left(\|h_i(Z)\|_{\mathbb{P},2} + \sum_{j\neq\rho(i)}|A_{i,j}|\|Z_i\|_{\mathbb{P},2}\right) \\
&\leq |A_{i,\rho(i)}|\|Z_{\rho(i)}\|^2_{\mathbb{P},2}(c_3 + c_2c_1).
\end{aligned} \tag{25}$$

Note that the first terms vanish for independent latent variables. Next we control the variance of $\hat{Z}_i$. We find using the triangle inequality

$$\begin{aligned}
\sigma^2_{\hat{Z}_i} \leq \mathbb{E}(\hat{Z}_i^2) &\leq \left(\|h_i(Z)\|_{\mathbb{P},2} + \sum_j |A_{ij}|\|Z_j\|_{\mathbb{P},2}\right)^2 \\
&\leq \|Z_{\rho(i)}\|^2_{\mathbb{P},2}\left(c_3|A_{i,\rho(i)}| + |A_{i,\rho(i)}| + c_1\sum_j |A_{i,j}|\right)^2 \\
&\leq \|Z_{\rho(i)}\|^2_{\mathbb{P},2}|A_{i,\rho(i)}|^2\left(1 + c_3 + c_2c_1\right)^2.
\end{aligned} \tag{26}$$

We now combine (25) and (26) with $(1+x)^{-1} \geq 1 - x$ and find

$$|\rho(Z_{\rho(i)}, \hat{Z}_i)| \geq \frac{|A_{i,\rho(i)}|\|Z_{\rho(i)}\|^2_{\mathbb{P},2}(1 - c_3 - c_1c_2)}{\|Z_{\rho(i)}\|_{\mathbb{P},2}\cdot\|Z_{\rho(i)}\|_{\mathbb{P},2}|A_{i,\rho(i)}|\left(1 + c_3 + c_2c_1\right)} \geq (1 - c_3 - c_1c_2)^2 \geq 1 - 2c_3 - 2c_1c_2. \tag{27}$$

The claim follows by bounding

$$\mathrm{MCC}(T(Z), Z) \geq d^{-1}\sum_i |\rho(Z_{\rho(i)}, \hat{Z}_i)| \geq 1 - 2c_3 - 2c_1c_2. \tag{28}$$

□

The previous lemma has minimal assumptions on $\mathbb{P}$ and h. We now show that the lemma can be improved when assuming independence of the coordinates of $\mathbb{P}$. Otherwise, the statement and proof are similar.

Lemma B.2. *Let $\mathbb{P}$ be a centered probability measure on $\mathbb{R}^d$ satisfying Assumption 2.2, i.e., with independent components. Assume that for $Z \sim \mathbb{P}$ the bound*

$$\max_{i,j} \frac{\|Z_i\|_{\mathbb{P},2}}{\|Z_j\|_{\mathbb{P},2}} \leq c_1 \tag{29}$$

holds for some constant c_1. Suppose that $T(z) = Az + h(z)$ and $\mathbb{E}_{\mathbb{P}}(Z_i h_j(Z)) = 0$ for all $1 \leq i, j \leq d$. Assume that A is a matrix such that there is a permutation ρ and a constant c_2 with

$$\max_i \frac{1}{|A_{i,\rho(i)}|^2} \sum_{j \neq \rho(i)} |A_{i,j}|^2 \leq c_2^2 \tag{30}$$

and h has the property that for some constant c_3

$$\frac{\|h_i\|_{\mathbb{P},2}}{|A_{i,\rho(i)}| \|Z_{\rho(i)}\|_{\mathbb{P},2}} \leq c_3. \tag{31}$$

Then the bound

$$\mathrm{MCC}(T(Z), Z) \geq 1 - \frac{c_3^2}{2} - \frac{(c_1 c_2)^2}{2} \tag{32}$$

holds.

Proof. Again, we lower bound the expression (4) for the permutation $\pi = \rho^{-1}$ in the statement of the lemma. Let $Z \sim \mathbb{P}$ and $\tilde{Z} = T(Z)$. We first control

$$\rho(Z_{\rho(i)}, \hat{Z}_i) = \frac{\mathrm{Cov}(Z_{\rho(i)}, \hat{Z}_i)}{\sigma_{Z_{\rho(i)}} \sigma_{\hat{Z}_i}}. \tag{33}$$

Using that $Z_{\rho(i)}$ is centered followed by $\mathbb{E}(Z_i Z_j) = \mathbb{E}(Z_i h_j(Z)) = 0$ (by independence and the assumption for h) we get

$$\begin{aligned} \mathrm{Cov}(Z_{\rho(i)}, \hat{Z}_i) &= \mathbb{E}(Z_{\rho(i)} \hat{Z}_i) \\ &= \sum_j A_{i,j} \mathbb{E}(Z_{\rho(i)} Z_j) + \mathbb{E}(Z_{\rho(i)} h_i(Z)) \\ &= A_{i,\rho(i)} \mathbb{E}(Z_{\rho(i)}^2). \end{aligned} \tag{34}$$

Next we control the variance of $\hat{Z}_i$. We expand the product and find (exploiting again independence and that h and Z are uncorrelated)

$$\begin{aligned} \sigma_{\hat{Z}_i}^2 &= \mathbb{E}(\hat{Z}_i^2) = \mathbb{E}\left(h_i(Z) + \sum_j A_{i,j} Z_j\right)^2 \\ &= \|h_i\|_{\mathbb{P},2}^2 + \sum_{j \neq \rho(i)} |A_{i,j}|^2 \|Z_j\|_{\mathbb{P},2}^2 + |A_{i,\rho(i)}|^2 \|Z_{\rho(i)}\|_{\mathbb{P},2}^2 \\ &\leq |A_{i,\rho(i)}|^2 \cdot \|Z_{\rho(i)}\|_{\mathbb{P},2}^2 \left(1 + c_3^2 + c_1^2 \frac{1}{|A_{i,\rho(i)}|^2} \sum_j |A_{i,j}|^2\right) \\ &\leq \|Z_{\rho(i)}\|_{\mathbb{P},2}^2 |A_{i,\rho(i)}|^2 \left(1 + c_3^2 + (c_2 c_1)^2\right). \end{aligned} \tag{35}$$

Combining again (34) and (35) with $\sqrt{1+x} \leq 1 + x/2$ (for $x > 0$) and $(1+x)^{-1} \geq 1 - x$, we find

$$|\rho(Z_{\rho(i)}, \hat{Z}_i)| \geq \frac{|A_{i,\rho(i)}| \cdot \|Z_{\rho(i)}\|_{\mathbb{P},2}^2}{\|Z_{\rho(i)}\|_{\mathbb{P},2} \cdot |A_{i,\rho(i)}| \cdot \|Z_{\rho(i)}\|_{\mathbb{P},2} \sqrt{1 + c_3^2 + (c_2 c_1)^2}} \geq \left(1 + \frac{c_3^2}{2} + \frac{(c_2 c_1)^2}{2}\right)^{-1} \geq 1 - \frac{c_3^2}{2} - \frac{(c_2 c_1)^2}{2}. \tag{36}$$

We finish by noting

$$\mathrm{MCC}(T(Z), Z) \geq d^{-1} \sum_i |\rho(Z_{\rho(i)}, \hat{Z}_i)| \geq 1 - \frac{c_3^2}{2} - \frac{(c_2 c_1)^2}{2}. \tag{37}$$

□

C. Auxiliary result on distances to the orthogonal group

We need some elementary and well-known facts on the distance of matrices to the orthogonal group. Let $A \in \mathbb{R}^{d\times d}$ be a matrix with $\det A > 0$. Then we consider a Singular Value Decomposition (SVD) $A = U\Sigma V^\top$ where $U, V \in \mathrm{SO}(d)$ and $\Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_d) \in \mathrm{Diag}(d)$ contains the singular values $\sigma_i > 0$. The following Lemma holds.

Lemma C.1. *For $A \in \mathbb{R}^{d\times d}$ with $\det A > 0$ and any SVD $A = U\Sigma V^\top$ as above, the relation*

$$\mathrm{dist}(A, \mathrm{SO}(d))^2 = |(A^\top A)^{\frac{1}{2}} - \mathrm{Id}|_F^2 = |\Sigma - \mathrm{Id}|_F^2 = \sum_{i=1}^d (\sigma_i - 1)^2 \tag{38}$$

holds, moreover the minimizer of $|A - Q|_F^2$ is given by $Q = UV^\top$.

Proof. First, we note that the last three expressions agree. For the last two this is obvious, and we find

$$\sqrt{A^\top A} = \sqrt{V\Sigma^2 V^\top} = V\sqrt{\Sigma^2}V^\top = V\Sigma V^\top. \tag{39}$$

Then we conclude by unitary invariance $|RX|_F = |X|_F$ for $R \in \mathrm{SO}(d)$ of the Frobenius norm that

$$|\sqrt{A^\top A} - \mathrm{Id}|_F = |V\Sigma V^\top - \mathrm{Id}|_F = |\Sigma - \mathrm{Id}|_F \tag{40}$$

It remains to show that also the first expression agrees with the other three expressions. We find for $Q \in \mathrm{SO}(d)$ using again unitary invariance that

$$\begin{aligned} |A - Q|_F^2 &= |U\Sigma V^\top - Q|_F^2 = |\Sigma - U^\top Q V|_F^2 \\ &= |\Sigma|_F^2 + |U^\top Q V|_F^2 - 2\,\mathrm{Tr}\,\Sigma(U^\top Q V) = |\Sigma|_F^2 + d - 2\sum_{i=1}^d \sigma_i (U^\top Q V)_{ii}. \end{aligned} \tag{41}$$

Here we used $|R|_F^2 = \mathrm{Tr}\, RR^\top = d$ in the last step. Since $\sigma_i > 0$ and since $R_{ii} \leq 1$ for $R \in \mathrm{SO}(d)$ we conclude

$$|A - Q|_F^2 \geq |\Sigma|_F^2 + d - \sum_{i=1}^d \sigma_i \tag{42}$$

with equality if $(U^\top Q V)_{ii} = 1$ for $1 \leq i \leq d$ and thus $U^\top Q V = \mathrm{Id}$ and therefore $Q = UV^\top$. From (41) we find

$$|A - UV^\top|_F^2 = |\Sigma|_F^2 + d - 2\sum_{i=1}^d \sigma_i = \sum_{i=1}^d (\sigma_i^2 - 2\sigma_i + 1). \tag{43}$$

□

This has the following consequence.

Lemma C.2. *Let $A \in \mathbb{R}^{d\times d}$ with $\det A > 0$ and denote the smallest singular value of A by $\sigma_{\min}$. Then*

$$\mathrm{dist}(A^{-1}, \mathrm{SO}(d))^2 \leq \frac{1}{\sigma_{\min}^2}\,\mathrm{dist}(A, \mathrm{SO}(d))^2 \tag{44}$$

Proof. Note that the singular values of A^{-1} are σ_i^{-1} where σ_i denote the singular values of A. Using Lemma C.1 we then find

$$\operatorname{dist}(A^{-1}, \mathrm{SO}(d))^2 = \sum_{i=1}^{d} (\sigma_i^{-1} - 1)^2 = \sum_{i=1}^{d} \sigma_i^{-2} (\sigma_i - 1)^2 \leq \sigma_{\min}^{-2} \operatorname{dist}(A, \mathrm{SO}(d))^2. \tag{45}$$

□

We need one further lemma that concerns the distance to $\mathrm{SO}(d)$ of a product of matrices.

Lemma C.3. *Let $A, B \in \mathbb{R}^{d\times d}$ with $\det A > 0$, $\det B > 0$. Then*

$$\operatorname{dist}(AB, \mathrm{SO}(d)) \leq \frac{3}{2} \operatorname{dist}(A, \mathrm{SO}(d)) + \frac{3}{2} \operatorname{dist}(B, \mathrm{SO}(d)). \tag{46}$$

This implies for $p > 1$

$$\operatorname{dist}^p(AB, \mathrm{SO}(d)) \leq 3^p \left(\operatorname{dist}^p(A, \mathrm{SO}(d)) + \operatorname{dist}^p(B, \mathrm{SO}(d)) \right). \tag{47}$$

Proof. Denote the projections of A and B on $\mathrm{SO}(d)$ by Q and R respectively. Then

$$\begin{aligned}
\operatorname{dist}(AB, \mathrm{SO}(d)) &\leq |AB - QR|_F \\
&= |(A-Q)R + Q(B-R) + (A-Q)(B-R)|_F \\
&\leq |(A-Q)R|_F + |Q(B-R)|_F + |(A-Q)(B-R)|_F \\
&\leq |A-Q|_F + |B-R|_F + \sqrt{|A-Q|_F |B-R|_F} \\
&\leq \frac{3}{2}|A-Q|_F + \frac{3}{2}|B-R|_F \\
&= \frac{3}{2} \operatorname{dist}(A, \mathrm{SO}(d)) + \frac{3}{2} \operatorname{dist}(B, \mathrm{SO}(d)).
\end{aligned} \tag{48}$$

Here we used again unitary equivalence of the Frobenius norm and the Cauchy-Schwarz inequality. This proves (46) and observe that by the generalized mean inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p) \leq 2^p(a^p + b^p)$ we obtain (47). □

D. Extension to the undercomplete case

The main interest of representation learning concerns the case where we actually learn a representation of a low dimensional submanifold embedded in a high dimensional space. This creates some additional technical difficulties that we address here. We always assume that the mixing $f : \Omega \to M \subset \mathbb{R}^D$ is a diffeomorphism on an embedded submanifold. We can identify the tangent space $T_{f(s)}M$ with an affine subspace of $\mathbb{R}^D$ and we center this subspace at 0 instead of $f(s)$ to obtain a linear subspace. We consider the standard Riemannian metric on $\mathbb{R}^d$ and $\mathbb{R}^D$ inducing also a metric on $T_{f(s)}M$. The differential $Df(s)$ defines a bijective linear map $\mathbb{R}^d \cong T_s\mathbb{R}^d \to T_{f(s)}M$.

To simplify the notation and restrict attention to the essential requirements we consider a linear subspace $H \subset \mathbb{R}^D$ and we assume that $A \in \mathbb{R}^{D\times d}$ is a full rank matrix with $\operatorname{range}(A) = H$, in particular $\dim(H) = d$. We now characterize the matrix representations of maps $\mathbb{R}^d \to H \subset \mathbb{R}^D$. For this we consider the unique polar decomposition $A = UP$ where $U : \mathbb{R}^d \to \mathbb{R}^D$ with $U \in \mathrm{SO}(d, D)$ and $P : \mathbb{R}^d \to \mathbb{R}^d$ is symmetric and positive definite. Note that P is given by the expression $P = \sqrt{A^\top A}$. In the following U will always denote the orthogonal matrix from such a polar decomposition.

Lemma D.1. *Let $A = UP$ be the polar decomposition of A. Then every linear map $T \in \mathbb{R}^{D\times d}$ with $\operatorname{Range}(T) \subset H = \operatorname{Range}(A)$ can be written uniquely as $T = UB$ for some $B \in \mathbb{R}^{d\times d}$.*

Proof. We first claim that $UU^\top T = T$. Let $x \in \mathbb{R}^d$ then $Tx = Uy$ for some $y \in \mathbb{R}^d$ because T maps to H and U is surjective on H. We thus find

$$UU^\top Tx = UU^\top Uy = Uy = Vx \tag{49}$$

where we used $U^\top U = \mathrm{Id}_d$ (recall $U \in \mathrm{SO}(d, D)$). Then we find $T = UB$ with $B = U^\top T \in \mathbb{R}^{d\times d}$. Uniqueness follows by injectivity of U. □

We now consider the space of isometries $\mathrm{SO}(d, H)$ consisting of all linear maps $Q : \mathbb{R}^d \to \mathbb{R}^D$ such that $\mathrm{Range}(Q) \subset H$ and $|Qv| = |v|$ for all $v \in \mathbb{R}^d$. Note that SO in addition requires preservation of orientation, but we can either just choose the orientation induced by A or ignore this issue and think of orthogonal matrices $\mathrm{O}(d, H)$. We nevertheless write SO to be consistent with the literature. Note that $\mathrm{SO}(d, H) \subset \mathrm{SO}(d, D)$. We can now define the distance

$$\mathrm{dist}(A, \mathrm{SO}(d, H)) = \min_{Q \in \mathrm{SO}(d,H)} |A - Q|_F. \tag{50}$$

This distance can be characterized as follows.

Lemma D.2. *Let $T \in \mathbb{R}^{D \times d}$ be a linear map with $\mathrm{Range}(T) \subset H$. Then*

$$\mathrm{dist}(T, \mathrm{SO}(d, H)) = \mathrm{dist}(B, \mathrm{SO}(d)). \tag{51}$$

where B is the unique matrix such that $T = UB$ (see Lemma D.1).

Proof. The crucial observation is that the relation

$$\mathrm{SO}(d, H) = U \cdot \mathrm{SO}(d) \tag{52}$$

holds where $U \cdot \mathrm{SO}(d) = \{UV : V \in \mathrm{SO}(d)\}$. We first consider the inclusion $\subset$. Consider $V \in \mathrm{SO}(d, H)$. We have seen in the proof of Lemma D.1 that $V = U(U^\top V)$ so we only need to show that $U^\top V \in \mathrm{SO}(d)$. We find

$$(U^\top V)^\top U^\top V = V^\top U U^\top V = V^\top V = \mathrm{Id}_d \tag{53}$$

where we used $UU^\top V = V$ and $V \in \mathrm{SO}(d, D)$. Therefore $U^\top V \in \mathrm{SO}(d)$ and thus $V \in U\,\mathrm{SO}(d)$. Now we consider the reverse inclusion $\supset$. Let $Q \in \mathrm{SO}(d)$. Then we find

$$\mathrm{Range}(UQ) = \mathrm{Range}(U) = \mathrm{Range}(UP) = \mathrm{Range}(A) = H. \tag{54}$$

Moreover $|UQv|^2 = v^\top Q^\top U^\top U Q v = v^\top Q^\top Q v = |v|^2$ and therefore $UQ \in \mathrm{SO}(d, H)$. Thus we find

$$\mathrm{dist}(T, \mathrm{SO}(d, H)) = \mathrm{dist}(UB, U\,\mathrm{SO}(d)) = \mathrm{dist}(B, \mathrm{SO}(d)) \tag{55}$$

where the last step used the unitary invariance of the Frobenius norm, i.e., $|UB| = |B|$ for $U \in \mathrm{SO}(d, D)$ and $B \in \mathbb{R}^{d \times d}$. □

Recall that in the polar decomposition $A = UP$ the matrix P is given by $P = (A^\top A)^{\frac{1}{2}}$. Then the lemma above implies that for A such that $H = \mathrm{range}(A)$

$$\mathrm{dist}^2(A, \mathrm{SO}(d, H)) = \mathrm{dist}^2((A^\top A)^{\frac{1}{2}}, \mathrm{SO}(d)) = \sum (\sigma_i - 1)^2 \tag{56}$$

where σ_i are the singular values of A and we used Lemma C.1 in the last step.

We also need to consider isometries mapping $H \to \mathbb{R}^d$ and the distance to these isometries. The space of $\mathrm{SO}(H, d)$ consists of all distance preserving linear maps $H \to \mathbb{R}^d$. Clearly, there is no unique matrix representation $\mathbb{R}^{d \times D}$ of such maps because the map is not defined on $H^\perp$, i.e., there are many extensions to $\mathbb{R}^D$. We essentially consider the extension by zero to the orthogonal complement, i.e, for a map $T : H \to \mathbb{R}^d$ we consider the extension $T : \mathbb{R}^D \to \mathbb{R}^d$ (not reflected in the notation) such that $Tv = 0$ for any $v \perp H$, i.e., $v^\top w = 0$ for all $w \in H$. Using this extension, we can identify such maps T with a unique matrix $\mathbb{R}^{d \times D}$. Then we define

$$\mathrm{dist}(T, \mathrm{SO}(H, d)) = \min_{Q \in \mathrm{SO}(H,d)} |T - Q|_F \tag{57}$$

where we identify T and Q with matrix representations as explained before. We have the following simple lemma.

Lemma D.3. *Let $T : H \to \mathbb{R}^d$ be a linear map, which we identify with a matrix representation $T \in \mathbb{R}^{D \times d}$ as explained above. Let $U : \mathbb{R}^d \to H \subset \mathbb{R}^D$ be an isometry. Then there is a matrix $B \in \mathbb{R}^{d \times d}$ such that $T = BU^\top$ and*

$$\mathrm{dist}(T, \mathrm{SO}(H, d)) = \mathrm{dist}(B, \mathrm{SO}(d)). \tag{58}$$

Proof. Since U is an isometry to H we find $U \in \mathrm{SO}(d, D)$ and $U^\top$ maps H isometrically to $\mathbb{R}^d$ (because $U^\top(Uv) = \mathrm{Id}_d v = v$). Therefore, there is a matrix $B \in \mathbb{R}^{d\times d}$ such that $Tv = BU^\top v$ for all $v \in H$. Let $w \in H^\perp$. Then for all $v \in \mathbb{R}^d$ we have $Uv \in H$ and thus $0 = \langle w, Uv\rangle = \langle U^\top w, v\rangle$ which implies $U^\top w = 0$. We conclude that $T = BU^\top$. Next we claim that similar to (52) the relation

$$\mathrm{SO}(H, d) = \mathrm{SO}(d) \cdot U^\top \tag{59}$$

holds. The inclusion $\supset$ follows, since the concatenation of isometries is an isometry (and $U^\top H^\perp = \{0\}$. The reverse inclusion can be shown by using that $Q \in \mathrm{SO}(H, d)$ can be expressed as $Q = BU^\top$ for some B and noting that since $U^\top \in \mathrm{SO}(H, d)$ we find that B must be an isometry on $\mathbb{R}^d$ and therefore $B \in \mathrm{SO}(d)$.

Plugging everything together, we find

$$\begin{aligned}
\mathrm{dist}^2(T, \mathrm{SO}(H, d)) &= \min_{Q\in\mathrm{SO}(H,d)} |Q - T|_F^2 = \min_{Q'\in\mathrm{SO}(d)} |BU^\top - Q'U^\top|_F^2 = \min_{Q'\in\mathrm{SO}(d)} \mathrm{Tr}\, U(B - Q')^\top (B - Q')U^\top \\
&= \min_{Q'\in\mathrm{SO}(d)} \mathrm{Tr}(B - Q')^\top (B - Q')U^\top U = \min_{Q'\in\mathrm{SO}(d)} \mathrm{Tr}(B - Q')^\top (B - Q')\mathrm{Id}_d \\
&= \min_{Q'\in\mathrm{SO}(d)} |B - Q'|_F^2 = \mathrm{dist}^2(B, \mathrm{SO}(d)).
\end{aligned} \tag{60}$$

This ends the proof. □

Note that an invertible matrix B induces a bijective map from $\mathbb{R}^d$ to H through the matrix representation UB. Its inverse has the matrix representation $B^{-1}U^\top$, indeed, then we find $B^{-1}U^\top UB = \mathrm{Id}_d$. Now we can define the extension of the distance Θ to the undercomplete case. We define for $f : \Omega \subset \mathbb{R}^d \to M \subset \mathbb{R}^D$

$$\Theta_p^p(f, \Omega) = \int_\Omega \mathrm{dist}^p(Df(z), \mathrm{SO}(d, T_{f(z)}M)) + \mathrm{dist}^p\big((Df)^{-1}(z), \mathrm{SO}(T_{f(z)}M, d)\big)\, \mathrm{d}z \tag{61}$$

where $(Df)^{-1}(z)$ denotes the inverse of the map $Df(z)$ viewed as a linear map with target $T_{f(z)}M$. Note that for $d = D$ this definition agrees with the definition in (2).

By our results so far, we can equivalently write

$$\begin{aligned}
\Theta_p^p(f, \Omega) &= \int_\Omega \mathrm{dist}^p(P(Df(s)), \mathrm{SO}(d)) + \mathrm{dist}^p\big(P^{-1}(Df(s)), \mathrm{SO}(d)\big)\, \mathrm{d}s \\
&= \int_\Omega \mathrm{dist}^p((Df(s)^\top Df(s))^{\frac{1}{2}}, \mathrm{SO}(d)) + \mathrm{dist}^p\big((Df(s)^\top Df(s))^{-\frac{1}{2}}, \mathrm{SO}(d)\big)\, \mathrm{d}s
\end{aligned} \tag{62}$$

where $P(Df(s)) = \sqrt{Df(s)^\top Df(s)}$ denotes the unique matrix P in the polar decomposition $Df(s) = UP$.

We need the following extension of Lemma C.3.

Lemma D.4. *Suppose $H = \mathrm{Range}(A)$ for a full rank matrix $A \in \mathbb{R}^{D\times d}$. Let $T \in \mathbb{R}^{D\times d}$ with $\mathrm{range}(T) \subset H$ and $S : H \to \mathbb{R}^d$ a linear map which we identify with its matrix representation $S \in \mathbb{R}^{d\times D}$ as explained above. Then ST is the matrix representation of the concatenation of the linear maps, and we assume that $\det(ST) > 0$. Then the bound*

$$\mathrm{dist}(ST, \mathrm{SO}(d)) \le \frac{3}{2}\,\mathrm{dist}(S, \mathrm{SO}(H, d)) + \frac{3}{2}\,\mathrm{dist}(T, \mathrm{SO}(d, H)) \tag{63}$$

holds. For $p \ge 1$ this implies

$$\mathrm{dist}^p(ST, \mathrm{SO}(d)) \le 3^p\,\mathrm{dist}^p(S, \mathrm{SO}(H, d)) + 3^p\,\mathrm{dist}^p(T, \mathrm{SO}(d, H)). \tag{64}$$

Proof. Denote as before the polar decomposition of A by $A = UP$. We have observed that there are matrices B, B' such that $S = B'U^\top$ and $T = UB$. Then we have $ST = B'U^\top UB = B'B$. Moreover, we have seen that

$$\mathrm{dist}(T, \mathrm{SO}(d, H)) = \mathrm{dist}(B, \mathrm{SO}(d)), \quad \mathrm{dist}(S, \mathrm{SO}(H, d)) = \mathrm{dist}(B', \mathrm{SO}(d)). \tag{65}$$

Thus we find using Lemma C.3 (for $A = B'$, $B = B$)

$$\begin{aligned}\operatorname{dist}(ST, \mathrm{SO}(d)) = \operatorname{dist}(B'B, \mathrm{SO}(d)) &\leq \frac{3}{2}\operatorname{dist}(B', \mathrm{SO}(d)) + \frac{3}{2}\operatorname{dist}(B, \mathrm{SO}(d)) \\ &= \frac{3}{2}\operatorname{dist}(T, \mathrm{SO}(d,H)) + \frac{3}{2}\operatorname{dist}(S, \mathrm{SO}(H,d)).\end{aligned} \tag{66}$$

□

E. Proof and extension of approximate linear identifiability for approximate isometries

In this section we provide the proofs for the results in Section 4. First, we state the key rigidity statement that we use in the proof of Theorem 4.1.

Theorem E.1 (Theorem 3.1 in Friesecke et al. (2002))**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain*[1]*. Then there is for $p > 1$ a constant $C(\Omega, p)$ such that for each $u \in W^{1,p}(\Omega, \mathbb{R}^d)$ there is a linear function $L(s) = As + b$ with $b \in \mathbb{R}^d$, $A \in \mathrm{SO}(d)$ such that*

$$\|u - L\|^p_{W^{1,p}(\Omega)} \leq C(\Omega, p) \int_\Omega \operatorname{dist}(Du(s), \mathrm{SO}(d))^p \,\mathrm{d}s. \tag{67}$$

Remark E.2.
- The reference above only stated the case $p = 2$ the simple extension to general p can be found in Section 2.4 in Conti & Schweizer (2006).
- The main interest for the study of elasticity is the bound on the gradient because this is related to the energy. Here, we are only interested in the deviation $L - u$ (where actually simpler proofs of similar results are possible (Kohn, 1982)). Using for $p < d$ the Sobolev embedding $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ for $q = pd/(d-p)$ the Theorem above implies

$$\begin{aligned}\|u - L\|_{L^q(\Omega)} &\leq C(\Omega, p)\|u - L\|_{W^{1,p}(\Omega)} \\ &\leq C(\Omega, p)\left(\int_\Omega \operatorname{dist}(Du(s), \mathrm{SO}(d))^p \,\mathrm{d}s\right)^{\frac{1}{p}}.\end{aligned} \tag{68}$$

Let us now discuss one lemma that we split off from the proof of Theorem 4.1 because we will use it in the proof of Theorem G.2 below again.

Lemma E.3. *Suppose that $d \leq D$ and $f : \Omega \subset \mathbb{R}^d \to M \subset \mathbb{R}^D$ and $g : \Omega' \to M$ are diffeomorphisms, $\mathbb{P}$ and $\mathbb{Q}$ are measures on Ω and Ω' such that $f_*\mathbb{P} = g_*\mathbb{Q}$. Suppose the density of $\mathbb{P}$ is lower bounded on Ω. Let $T = g^{-1} \circ f$, then*

$$\begin{aligned}&\int_\Omega \operatorname{dist}^p(DT(z), \mathrm{SO}(d)) \,\mathrm{d}z \\ &\qquad \leq 3^p \int_\Omega \operatorname{dist}^p(Df(s), \mathrm{SO}(d, T_{f(s)}M)) \,\mathrm{d}s + C \int_{\Omega'} \operatorname{dist}^p(Dg^{-1}(g(s)), \mathrm{SO}(T_{g(s)}M, d)) \,\mathbb{Q}(\mathrm{d}s)\end{aligned} \tag{69}$$

where $C > 0$ depends on the lower bound of the density and p.

Remark E.4. For $d = D$ the formula (69) simplifies to

$$\int_\Omega \operatorname{dist}(DT(z), \mathrm{SO}(d))^p \,\mathrm{d}z \leq 3^p \int_\Omega \operatorname{dist}^p(Df(s), \mathrm{SO}(d)) \,\mathrm{d}s + C \int_{\Omega'} \operatorname{dist}^p(Dg^{-1}(g(s)), \mathrm{SO}(d)) \,\mathbb{Q}(\mathrm{d}s). \tag{70}$$

Proof. First we note that $g^{-1} : M \to \Omega$ is by assumption differentiable and by the chain rule we have $Dg^{-1}(f(s))Df(s) = DT(s)$. We now find using Lemma D.4 that

$$\begin{aligned}\int_\Omega \operatorname{dist}^p(DT(s), \mathrm{SO}(d)) \,\mathrm{d}s &= \int_\Omega \operatorname{dist}^p(Dg^{-1}(f(s))Df(s), \mathrm{SO}(d)) \,\mathrm{d}s \\ &\leq 3^p \int_\Omega \operatorname{dist}^p(Dg^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d)) + \operatorname{dist}^p(Df(s), \mathrm{SO}(d, T_{f(s)}M)) \,\mathrm{d}s\end{aligned} \tag{71}$$

[1] A Lipschitz domain is slightly informally a set whose boundary $\partial\Omega$ can be expressed as the union of the graphs of Lipschitz functions (see (Adams & Fournier, 2003) for a complete definition).

Now we use the transformation formula for push-forward measures, and the assumption $g_*\mathbb{Q} = f_*\mathbb{P}$ and the lower bound on the density of $\mathbb{P}$

$$\begin{aligned}\int_\Omega \operatorname{dist}^p(Dg^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\,\mathrm{d}s &\le C\int_\Omega \operatorname{dist}^p(Dg^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\,\mathbb{P}(ds)\\ &= C\int_{f(\Omega)} \operatorname{dist}^p(Dg^{-1}(x), \mathrm{SO}(T_xM, d))\,(f_*\mathbb{P})(\mathrm{d}x)\\ &= C\int_{g(\Omega')} \operatorname{dist}^p(Dg^{-1}(x), \mathrm{SO}(T_xM, d))\,(g_*\mathbb{Q})(\mathrm{d}x)\\ &= C\int_{\Omega'} \operatorname{dist}^p(Dg^{-1}(g(s)), \mathrm{SO}(T_{g(s)}M, d))\,\mathbb{Q}(\mathrm{d}s).\end{aligned} \tag{72}$$

The last two displays together imply the claim. □

We now provide the extensions of Theorem 4.1 to $d < D$. In this case, we can still define the set $\mathcal{M}(f_*\mathbb{P})$ as in (7). We define g by

$$(g, \mathbb{Q}, \Omega') \in \operatorname*{argmin}_{(\bar g, \bar Q, \bar\Omega)\in\mathcal{M}(f_*\mathbb{P})} \int_{\bar\Omega} \operatorname{dist}(D\bar g^{-1}(x), \mathrm{SO}(T_xM, d))^p\,\bar g_*\bar{\mathbb{Q}}(\mathrm{d}x). \tag{73}$$

Here we need to integrate the deviation from an isometry over the observational distribution $\bar g_*\mathbb{Q} = f_*\mathbb{P}$. Again, this agrees with the definition of (8) for $d = D$. To avoid the assumption that the minimum exists, we also let $(g_\varepsilon, \mathbb{Q}, \Omega') \in \mathcal{M}(f_*\mathbb{P})$ be any function such that

$$\int_{\bar\Omega} \operatorname{dist}(Dg_\varepsilon^{-1}(x), \mathrm{SO}(T_xM, d))^p\,(g_\varepsilon)_*\bar{\mathbb{Q}}(\mathrm{d}x) \le \inf_{(\bar g, \bar Q, \bar\Omega)\in\mathcal{M}(f_*\mathbb{P})} \int_{\bar\Omega} \operatorname{dist}(D\bar g^{-1}(x), \mathrm{SO}(T_xM, d))^p\,\bar g_*\bar{\mathbb{Q}}(\mathrm{d}x) + \varepsilon. \tag{74}$$

Then we can state and prove the following complete version of Theorem 4.1.

Theorem E.5. *Suppose we have a latent distribution $\mathbb{P} \in \mathcal{P}$ satisfying Assumptions 2.1 with support $\Omega \subset \mathbb{R}^d$ where Ω is a bounded connected Lipschitz domain. The observational distribution is given by $X = f(Z) \in M \subset \mathbb{R}^D$ where $Z \sim \mathbb{P}$ and $f \in \mathcal{F}(\Omega)$. Fix a $1 < p < \infty$. Let $(g_\varepsilon, \mathbb{Q}, \Omega') \in \mathcal{M}(f_*\mathbb{P})$ be any function satisfying (74). Then there is $A \in \mathrm{SO}(d)$ and $b \in \mathbb{R}^d$ such that $g_\varepsilon^{-1} \circ f(z) = Az + b + h(z)$ and h satisfies the bound*

$$\|h\|_{\mathbb{P},q} \le C\Theta_p(f, \Omega) + C\varepsilon^{\frac{1}{p}} \tag{75}$$

for $q = pd/(d-p)$. Here C is a constant depending on d, p, Ω, and the lower and upper bound on the density of $\mathbb{P}$.

It is clear that this theorem is more general than Theorem 4.1.

Proof. As before, we call the transition function $T = g^{-1}f : \Omega \to \Omega'$. First, we observe that by definition of g_ε

$$\begin{aligned}\int_{\bar\Omega} \operatorname{dist}(Dg_\varepsilon^{-1}(x), \mathrm{SO}(T_xM, d))^p\,(g_\varepsilon)_*\bar{\mathbb{Q}}(\mathrm{d}x) &\le \inf_{(\bar g, \bar Q, \bar\Omega)\in\mathcal{M}(f_*\mathbb{P})} \int_{\bar\Omega} \operatorname{dist}(D\bar g^{-1}(x), \mathrm{SO}(T_xM, d))^p\,\bar g_*\bar{\mathbb{Q}}(\mathrm{d}x) + \varepsilon\\ &\le \int_\Omega \operatorname{dist}^p(Df^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\,\mathbb{P}(\mathrm{d}s) + \varepsilon.\end{aligned} \tag{76}$$

Here we used in the second step that $(f, \mathbb{P}, \Omega) \in \mathcal{M}(f_*\mathbb{P})$, i.e., f is a valid representation of our data so that this provides an upper bound on the infimum. We now find using Lemma E.3 and (76)

$$\begin{aligned}&\int_\Omega \operatorname{dist}^p(DT(s), \mathrm{SO}(d))^p\,\mathrm{d}s\\ &\le 3^p\int_\Omega \operatorname{dist}^p(Df(s), \mathrm{SO}(d, T_{f(s)}M))\,\mathrm{d}s + C\int_{\Omega'} \operatorname{dist}^p(Dg^{-1}(g(s)), \mathrm{SO}(T_{f(s)}M, d))\,\mathbb{Q}(\mathrm{d}s)\\ &\le 3^p\int_\Omega \operatorname{dist}^p(Df(s), \mathrm{SO}(d, T_{f(s)}M))\,\mathrm{d}s + C\int_\Omega \operatorname{dist}^p(Df^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\,\mathbb{P}(\mathrm{d}s) + C\varepsilon\\ &\le 3^p\int_\Omega \operatorname{dist}^p(Df(s), \mathrm{SO}(d, T_{f(s)}M))\,\mathrm{d}s + C\int_\Omega \operatorname{dist}^p(Df^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\,\mathrm{d}s + C\varepsilon\\ &= C\Theta_p^p(f, \Omega) + C\varepsilon.\end{aligned} \tag{77}$$

Note that in the second to last step we used the upper bound on the density of $\mathbb{P}$.

Now we apply Theorem E.1 (or rather its consequence (68) which states that there is a matrix $A \in \mathrm{SO}(d)$ and $b \in \mathbb{R}^d$ such that

$$\left(\int_\Omega |T(s) - As - b|^q \,\mathrm{d}s\right)^{\frac{1}{q}} \leq C(\Omega) \left(\int_\Omega \mathrm{dist}^p(DT(s), \mathrm{SO}(d)) \,\mathrm{d}s\right)^{\frac{1}{p}}. \tag{78}$$

Using the upper bound on the density of $\mathbb{P}$ and the last two displays we find

$$\begin{aligned}\left(\int_\Omega |T(s) - As - b|^q \,\mathbb{P}(\mathrm{d}s)\right)^{\frac{1}{q}} &\leq C \left(\int_\Omega |T(s) - As - b|^q \,\mathrm{d}s\right)^{\frac{1}{q}} \\ &\leq C \left(\int_\Omega \mathrm{dist}^p(DT(s), \mathrm{SO}(d)) \,\mathrm{d}s\right)^{\frac{1}{p}} \\ &\leq C\Theta_p(f, \Omega) + C\varepsilon^{\frac{1}{p}}.\end{aligned} \tag{79}$$

Here we used $(a+b)^{\frac{1}{p}} \leq (2a)^{\frac{1}{p}} + (2b)^{\frac{1}{p}}$ This completes the proof. □

F. Proofs for the results on perturbed linear ICA

In this section we prove Theorem 5.5 and the extension in Theorem 5.9. Two technical difficulties when proving this result is that we consider a function defined on the sphere and that we need to whiten the data. Therefore, we first prove a linearized result in Lemma F.1 from which the Theorem can be deduced after some technical algebraic manipulations. To motivate the calculations, let us first briefly sketch the well-known unperturbed case of linear ICA (Hyvärinen & Oja, 2000).

Proof sketch of the Linear Result We assume that X is whitened and $X = AS$. Let w_0 with $|w_0| = 1$ be such that $w_0 X = w_0 AS = e_d S = S_d$. Note that for w in a neighborhood of w_0 we can write for some ε_i (small)

$$wX = \sqrt{1 - (\varepsilon_1^2 + \ldots + \varepsilon_{d-1}^2)} S_d + \varepsilon_1 S_1 + \ldots + \varepsilon_{d-1} S_{d-1}. \tag{80}$$

Indeed, since S_i are independent and with unit variance, we find that the prefactor of S_d has to be $\sqrt{1 - (\varepsilon_1^2 + \ldots + \varepsilon_{d-1}^2)}$ to ensure that wX has unit variance. Note that

$$1 - \sqrt{1 - (\varepsilon_1^2 + \ldots + \varepsilon_{d-1}^2)} = \tfrac{1}{2} \sum_{i=1}^{d-1} \varepsilon_i^2 + O(\varepsilon^4) \tag{81}$$

where $\varepsilon^2 = \sum_{i=1}^{d-1} \varepsilon_i^2$ is the $l2$ norm. Then we can Taylor expand (denoting $G' = g$)

$$G(Xw) = G(S_d) + \left(\sum_{i=1}^{d-1} \varepsilon_i S_i - \frac{1}{2} S_d \sum_{i=1}^{d-1} \varepsilon_i^2\right) g(S_d) + \frac{1}{2} g'(S_1) \left(\sum_{i=1}^{d-1} \varepsilon_i S_i\right)^2 + O(\varepsilon^3). \tag{82}$$

Taking the expectation over this expression, we obtain

$$\begin{aligned}\mathbb{E}(G(Xw)) &= \mathbb{E}(G(S_d)) + \mathbb{E}\left(\left(\sum_{i=1}^{d-1} \varepsilon_i S_i - \frac{1}{2} S_d \sum_{i=1}^{d-1} \varepsilon_i^2\right) g(S_d)\right) + \frac{1}{2}\mathbb{E}\left(g'(S_d)\left(\sum_{i=1}^{d-1} \varepsilon_i S_i\right)^2\right) + O(\varepsilon^3) \\ &= \mathbb{E}(G(S_d)) - \frac{1}{2}\mathbb{E}\left(S_d g(S_d)\right) \sum_{i=1}^{d-1} \varepsilon_i^2 + \frac{1}{2}\mathbb{E}(g'(S_d)) \sum_{i=1}^{d-1} \varepsilon_i^2 + O(\varepsilon^3)\end{aligned} \tag{83}$$

where we used that $\mathbb{E}(S_i) = 0$, $\mathbb{E}(S_i^2) = 1$, and the fact that S_i and S_j are independent for $i \neq j$ so that $\mathbb{E}(S_i g(S_d)) = \mathbb{E}(S_i)\mathbb{E}(g(S_d)) = 0$. In particular, we obtain

$$\mathbb{E}(G(Xw)) = \mathbb{E}(G(S_d)) + \frac{1}{2}\left(\sum_i \varepsilon_i^2\right) (\mathbb{E}(g'(S_d) - \mathbb{E}\left(S_d g(S_d)\right) + O(\varepsilon^3). \tag{84}$$

We conclude that under the condition

$$\mathbb{E}(g'(S_d)) - \mathbb{E}\left(S_d g(S_d)\right) \neq 0 \tag{85}$$

the function $w \to \mathbb{E}(G(wX))$ has a local extremum at $w = w_0$ and it is strictly convex or concave around w_0.

The key lemma. We now generalize the reasoning above to the perturbed setting. To simplify this further, we first assume $A = \mathrm{Id}$ so that

$$X = S + \eta h(S). \tag{86}$$

We also remove the linear whitening operation involved in H and instead consider

$$\tilde{H}_{\eta,h}(w) = \mathbb{E}G\left(\frac{wX}{\sigma_w}\right) \tag{87}$$

where we used the shorthand $\sigma_w = \sqrt{\mathbb{E}((wX)^2)}$. Clearly the function $\tilde{H}$ is invariant under rescaling of the argument, i.e., homogeneous of degree 0. Instead of restricting it to the sphere, we define the function $\bar{H} : \mathbb{R}^{d-1} \to \mathbb{R}$ given by

$$\bar{H}(\varepsilon) = \tilde{H}_{\eta,h}((\varepsilon, 1)^\top) \tag{88}$$

around $\varepsilon = 0$. Then our goal is to show that $\bar{H}$ has an extremum close to $\varepsilon = 0$ which allows us to approximately recover the independent component S_d. We now prove the following Lemma.

Lemma F.1. *Assume the contrast function G and the distribution of the sources S satisfy the Assumptions 5.2, 5.3, and 5.4 and assume that $\|h\|_{\mathbb{P},q} \leq 1$ for $q = \max(3, d_g)$. Define $v \in \mathbb{R}^{d-1}$ by*

$$v_i = \mathbb{E}(h_d(S)g'(S_d)S_i + g(S_d)h_i(S)). \tag{89}$$

Then, there is $\eta_0 > 0$ (depending on all problem constants, e.g., d, α, q, M but not on h) such that for $\eta \leq \eta_0$ the function $\bar{H}$ has a local extremum ε_0 in the vicinity of $\varepsilon = 0$ such that $|\varepsilon_0| = O(\eta)$. This extremum satisfies

$$\varepsilon_0 = \frac{\eta v}{\alpha} + O(\eta^2). \tag{90}$$

Moreover there is $\eta_0' > 0$ (depending on the same quantities as η_0) such that for $\eta \leq \eta_0'$ there is a radius $\kappa > 0$ such that $\bar{H}$ is strictly convex or concave on $B_\kappa(0)$ and satisfies

$$\begin{aligned} D^2\bar{H}(\varepsilon) &\geq \frac{|\alpha_d|}{2} \cdot \mathrm{Id} \quad \textit{for } \alpha_d < 0 \\ D^2\bar{H}(\varepsilon) &\leq \frac{-\alpha_d}{2} \cdot \mathrm{Id} \quad \textit{for } \alpha_d > 0\,. \end{aligned} \tag{91}$$

Proof. Since we are interested in small ε we will always assume that $|\varepsilon| \leq 1$. We will denote $w = (\varepsilon, 1)^\top$. Then we have

$$wX = S_d + \sum_{i=1}^{d-1} \varepsilon_i S_i + \eta w \cdot h(S). \tag{92}$$

We define $\Omega = \mathbb{E}(h(S)^\top h(S))$, i.e., the covariance of the nonlinear part. We find

$$E((wX)^2) = \mathbb{E}\left(S_d + \sum_{i=1}^{d-1} \varepsilon_i S_i + \eta w \cdot h(S)\right)^2 = 1 + \sum_{i=1}^{d-1} \varepsilon_i^2 + \eta^2 w \cdot \Omega w \tag{93}$$

where we used that S_i are unit variance and uncorrelated and $\mathbb{E}(S^\top h(S)) = 0$. In particular, we conclude that

$$\sigma_w = \sqrt{1 + \sum_{i=1}^{d-1} \varepsilon_i^2 + \eta^2 w \cdot \Omega w}. \tag{94}$$

Note that $|\Omega| \leq \|h\|_{\mathbb{P},2}^2 \leq \|h\|_{\mathbb{P},q}^2 \leq 1$ by assumption. Let us introduce some notation. We consider the shorthand

$$S_\varepsilon = \sum_{i=1}^{d-1} \varepsilon_i S_i. \tag{95}$$

We also use

$$h_w(S) = w \cdot h(S), \quad h_\varepsilon(S) = \sum_{i=1}^{d-1} \varepsilon_i h_i(S). \tag{96}$$

Since $w = (\varepsilon_1, \ldots, \varepsilon_{d-1}, 1)^\top$ we have

$$h_w(S) = h_\varepsilon(S) + h_d(S). \tag{97}$$

Now we perform a second order Taylor expansion of G around S_d with remainder term. We obtain that for some $\xi \in [S_d, wX/\sigma_w]$

$$\begin{aligned} G\left(\frac{wX}{\sigma_w}\right) &= G\left(S_d\right) + g\left(S_d\right) \frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w} \\ &\quad + \frac{1}{2} g'\left(S_d\right) \left(\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w}\right)^2 + \frac{1}{6} g''(\xi) \left(\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w}\right)^3. \end{aligned} \tag{98}$$

Our goal is to extract the quadratic terms in ε and η of this expression. We now start to bound the error term and show that it is of order 3. Since $\xi \in [S_d, wX/\sigma_w]$ and $\sigma_w > 1$ we conclude that $|\xi| \leq \max(|wX|, |S_d|)$. Then we can control using Assumption 5.2

$$\begin{aligned} |g''(\xi)| \leq C_g(1 + |\xi|^{\max(d_g-3,0)}) &\leq C_g(1 + \max(|wX|, |S_d|)^{\max(d_g-3,0)}) \\ &\leq C(1 + |S|^{\max(d_g-3,0)} + |\eta h(S)|^{\max(d_g-3,0)}). \end{aligned} \tag{99}$$

We have the simple bound

$$|\sigma_w - 1| \leq |\varepsilon|^2 + |\eta|^2 |w|^2 \cdot |\Omega| \leq |\varepsilon|^2 + |\eta|^2 |w|^2. \tag{100}$$

Then we can bound (using $\eta^2|\varepsilon| + |\varepsilon|^2\eta \leq |\varepsilon|^3 + \eta^3$)

$$\left|\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w}\right|^3 \leq C(|\varepsilon|^3 + \eta^3)|S|^3 + \eta^3 |h(S)|^3 \tag{101}$$

This implies (recall that $q = \max(3, d_g)$)

$$\frac{1}{6}\left| g''(\xi) \left(\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w}\right)^3 \right| \leq C(d, C_g, q)(|\varepsilon|^3 + \eta^3)\left(1 + |S|^q + |h(S)|^q\right). \tag{102}$$

Next, we consider the second term

$$\frac{1}{2} g'\left(S_d\right) \left(\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w}\right)^2 \tag{103}$$

which we approximate to 2nd order in ε and η and put all terms of order 3 and higher in the error term. Note that $S_\varepsilon = O(\varepsilon)$. We expand $\eta h_w(S) = \eta h_d(S) + \eta h_\varepsilon(S)$ where $\eta h_d(S) = O(\eta)$ and $\eta h_\varepsilon(S) = O(\eta\varepsilon)$. Finally $(1 - \sigma_w) = O(|\varepsilon|^2 + \eta^2)$. We conclude that

$$\left| \left(\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w}\right)^2 - \left(\frac{S_\varepsilon + \eta h_d}{\sigma_w}\right)^2 \right| \leq C(|S|^2 + |h(S)|^2) \tag{104}$$

where we again bounded mixed terms by, e.g., $|\varepsilon|^2\eta \le |\varepsilon|^3 + \eta^3$. Using in addition that $1 - \sigma_w^{-1} \le C(|\varepsilon|^2 + \eta^2)$ to replace σ_w in the denominator by 1 up to third order error terms, we find

$$\left| g'(S_d) \left(\frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w} \right)^2 - g'(S_d) \left(S_\varepsilon + \eta h_d(S) \right)^2 \right| \le C(|\varepsilon|^3 + \eta^3)(1 + |S|^{\max(d_g-2,0)})(|S|^2 + |h(S)|^2). \tag{105}$$

Finally we consider the term proportional to $g(S_d)$. Here we need the sharper bound

$$\left| \sigma_w - 1 - \frac{|\varepsilon|^2}{2} - \frac{\eta^2}{2} w \cdot \Omega w \right| \le 2|\varepsilon|^4 + 2\eta^4(|w|^2|\Omega|)^2 \le 2|\varepsilon|^4 + 2\eta^4|w|^4 \tag{106}$$

which implies using $w = e_d + O(\varepsilon)$ that $w \cdot \Omega w = \Omega_{dd} + O(\varepsilon)$ and moreover (recall $|\Omega| \le 1$)

$$\left| \sigma_w - 1 - \frac{|\varepsilon|^2}{2} - \frac{\eta^2}{2}\Omega_{dd} \right| \le 2|\varepsilon|^4 + 2\eta^4|w|^4 + 3|\varepsilon|\eta^2. \tag{107}$$

Then we get, again keeping terms up to order 2 in ε or η

$$\begin{aligned} &\left| g(S_d) \frac{S_\varepsilon + \eta h_w(S) + S_d(1-\sigma_w)}{\sigma_w} - g(S_d) \left(S_\varepsilon + \eta h_w(S) - \frac{1}{2}(|\varepsilon|^2 + \eta^2\Omega_{dd})S_d \right) \right| \\ &\qquad \le C(|\varepsilon|^3 + \eta^3)|S|^{\max(d_g-1,0)}(|S| + |h(S)|). \end{aligned} \tag{108}$$

Using the bounds (102), (105), and (108) in (98) we obtain

$$\begin{aligned} &\left| G(wX) - G(S_d) - g(S_d) \left(S_\varepsilon + \eta h_w(S) - \frac{1}{2}(|\varepsilon|^2 + \eta^2\Omega_{dd})S_d \right) - \frac{1}{2} g'(S_d) \left(S_\varepsilon + \eta h_d(S) \right)^2 \right| \\ &\qquad \le C(|\varepsilon|^3 + |\eta|^3)(1 + |S|^q + |h(S)|^q). \end{aligned} \tag{109}$$

Let us call the term between the absolute value on the left-hand side T. Then we can bound using Assumption 5.3

$$|\mathbb{E}T| \le \mathbb{E}|T| \le C(|\varepsilon|^3 + \eta^3)\mathbb{E}(|S|^q + |h(S)|^q) \le C(|\varepsilon|^3 + |\eta|^3)(M + \|h\|_{\mathbb{P},q}^q) = \Xi(|\varepsilon|^3 + |\eta|^3) \tag{110}$$

where $\Xi = C(M+2)$ was introduced for future reference and depends on C_g, d, q, and the moment bound M but is independent of ε and η and h. We observe next, using $\mathbb{E}(S_d S_\varepsilon) = 0$, $\mathbb{E}(g(S_d S_\varepsilon) = 0$ that

$$\begin{aligned} \mathbb{E}T &= \mathbb{E}G\left(\frac{wX}{\sigma_w}\right) - \mathbb{E}(G(S_d)) - \eta\mathbb{E}(h_w(S)g(S_d)) + \frac{1}{2}(|\varepsilon|^2 + \eta^2\Omega_{dd})\mathbb{E}(S_d g(S_d)) \\ &\quad - \frac{1}{2}|\varepsilon|^2\mathbb{E}(g'(S_d)) - \eta\mathbb{E}(h_d(S)S_\varepsilon g'(S_d)) - \frac{1}{2}\eta^2\mathbb{E}(g'(S_d)h_d(S)^2) \\ &= \mathbb{E}G\left(\frac{wX}{\sigma_w}\right) - \mathbb{E}(G(S_d)) - \eta\mathbb{E}(h_d(S)g(S_d)) + \frac{1}{2}(\eta^2\Omega_{dd})\mathbb{E}(S_d g(S_d)) - \frac{1}{2}\eta^2\mathbb{E}(g'(S_d)h_d(S)^2) \\ &\quad - \eta\mathbb{E}(h_d(S)S_\varepsilon g'(S_d)) - \eta\mathbb{E}(h_\varepsilon(S)g(S_d)) + \frac{1}{2}|\varepsilon|^2\left(\mathbb{E}g(S_d)S_1 - \mathbb{E}g'(S_d)\right). \end{aligned} \tag{111}$$

We thus obtained an expansion of $H(w) = \mathbb{E}G\left(wX/\sigma_w\right)$ up to second order in ε. Note that by plugging in $\varepsilon_i = 0$ for all i, i.e., $w = e_d$, in (110)

$$\left| \mathbb{E}G\left(\frac{e_d X}{\sigma_{e_d}}\right) - \mathbb{E}(G(S_d)) - \eta\mathbb{E}(h_d(S)g(S_d)) + \frac{1}{2}(\eta^2\Omega_{dd})\mathbb{E}(S_d g(S_d)) - \frac{1}{2}\eta^2\mathbb{E}(g'(S_d)h_d(S)^2) \right| \le \Xi\eta^3 \tag{112}$$

so we conclude using the triangle inequality from (110), (111), and (112) that

$$\left| \mathbb{E}G\left(\frac{wX}{\sigma_w}\right) - \mathbb{E}G\left(\frac{e_d X}{\sigma_{e_d}}\right) - \eta\mathbb{E}(h_d(S)S_\varepsilon g'(S_d)) - \eta\mathbb{E}(h_\varepsilon(S)g(S_d)) + \frac{1}{2}|\varepsilon|^2\left(\mathbb{E}g(S_d)S_d - \mathbb{E}g'(S_d)\right) \right| \le \Xi(|\varepsilon|^3 + 2\eta^3). \tag{113}$$

As a next step, we consider the derivatives of H. Since the reasoning is similar to the steps above, we provide slightly fewer details, i.e., we hide integrable terms in the O notation. First we observe that

$$\partial_i \sigma_w = \frac{\varepsilon_i + \eta^2 (\Omega w)_i}{\sigma_w}. \tag{114}$$

From here we obtain

$$\partial_i G\left(\frac{wX}{\sigma_w}\right) = \left(\frac{S_i + \eta h_i(S)}{\sigma_w} - \frac{(\varepsilon_i + \eta^2(\Omega w)_i)(wS + \eta h_w(S))}{\sigma_w^3}\right) g\left(\frac{wX}{\sigma_w}\right). \tag{115}$$

We apply a first order Taylor expansion to g and obtain the bound $g\left(\frac{wX}{\sigma_w}\right) = g(S_d) + O(|\varepsilon| + \eta)$. Thus, we find

$$\partial_i \bar{H}(\varepsilon) = \partial_i \mathbb{E} G\left(\frac{wX}{\sigma_w}\right) = \frac{1}{\sigma_w}\mathbb{E}(S_i g(S_d)) + O(|\varepsilon| + \eta) = O(|\varepsilon| + \eta) \tag{116}$$

where we used that S_j and S_d are independent.

Next we show using the bound (13) from Assumption 5.2 that the function $\varepsilon \to w = (\varepsilon, 1)^\top \to \mathbb{E}(G(wX/\sigma_w))$ is strictly convex or concave around $\varepsilon = 0$ for η sufficiently small. For this, we need to find an expression for the second derivatives of G. To keep the length of the formulas manageable, we hide the $\eta^2(\Omega w)_i$ as a $O(\eta^2)$ term. We find for $1 \le i, j \le d-1$ using (115) and (114)

$$\begin{aligned}
\partial_i \partial_j G\left(\frac{wX}{\sigma_w}\right) &= \partial_i \left(\left(\frac{S_j + \eta h_j(S)}{\sigma_W} - \frac{(\varepsilon_j + \eta^2(\Omega w)_j)(wS + \eta h_w(S))}{\sigma_w^3}\right) g\left(\frac{wX}{\sigma_w}\right)\right) \\
&= \left(\frac{S_j + \eta h_j(S)}{\sigma_w} + -\frac{\varepsilon_j(wS + \eta h_w(S))}{\sigma_w^3}\right)\left(\frac{S_i + \eta h_i(S)}{\sigma_w} - \frac{\varepsilon_i(wS + \eta h_w(S))}{\sigma_w^3}\right) g'\left(\frac{wX}{\sigma_w}\right) \\
&- \left(\frac{\varepsilon_j(S_i + \eta h_i(S)) + \delta_{ij}(wS + \eta h_w(S)) + \varepsilon_i(S_j + \eta h_j(S))}{\sigma_w^3} - \frac{3\varepsilon_i\varepsilon_j(wS + \eta h_w(S))}{\sigma_w^5}\right) g\left(\frac{wX}{\sigma_w}\right) + O(\eta^2) \\
&= S_j S_i g'\left(\frac{wX}{\sigma_w}\right) - \delta_{ij} S_1 g\left(\frac{wX}{\sigma_w}\right) + O(|\varepsilon| + \eta).
\end{aligned} \tag{117}$$

Here, we used again that $\sigma_w = 1 + O(|\varepsilon| + \eta)$. Now we apply a Taylor expansion to g and g' and obtain $g(wX/\sigma_w) = g(S_d) + O(|\varepsilon| + \eta)$ and similarly for g' (here similar power counting as in the first part implies that the highest moment that needs to be bounded is $q = \max(3, d_g)$, we do not show this in full detail here). Then we obtain

$$\begin{aligned}
\partial_i \partial_j \bar{H}(\varepsilon) = \partial_i \partial_j \mathbb{E} G\left(\frac{wX}{\sigma_w}\right) &= \mathbb{E}\left(S_i S_j g'\left(\frac{wX}{\sigma_w}\right)\right) - \delta_{ij}\mathbb{E}\left(S_d g\left(S_d\right)\right) + O(|\varepsilon| + \eta) \\
&= \delta_{ij}\mathbb{E}(g'(S_d) - S_d g(S_d)) + O(|\varepsilon| + \eta) = -\alpha_d \delta_{ij} + O(|\varepsilon| + \eta).
\end{aligned} \tag{118}$$

In particular, we find

$$|D^2 \bar{H}(\varepsilon) + \alpha_d \cdot \mathrm{Id}| = O(|\varepsilon| + \eta). \tag{119}$$

We conclude that for $\eta < \eta_0$ with η_0 sufficiently small, there is for some $\kappa > 0$ a neighborhood $B_\kappa(0)$ such that the function $B_\kappa(0) \ni \varepsilon \to \bar{H}(\varepsilon)$ is strictly convex or concave (depending on the sign of $\alpha_d = \mathbb{E}(S_d g(S_d) - g'(S_d))$) with $D^2\bar{H} \ge |\alpha_d|/2 \cdot \mathrm{Id}$ in the convex case. We emphasize that κ is independent of η as soon as $\eta \le \eta_0$ is sufficiently small.

It remains to prove the existence of a maximum or minimum and the expansion (90). To achieve this, we compare $\bar{H}$ with its expansion to second order, i.e., we define

$$f(\varepsilon) = \eta\mathbb{E}(h_d(S) S_\varepsilon g'(S_d)) + \eta\mathbb{E}(h_\varepsilon(S) g(S_d)) - \frac{1}{2}|\varepsilon|^2 \left(\mathbb{E} g(S_d) S_d - \mathbb{E} g'(S_d)\right) = \varepsilon\eta v - \frac{1}{2}\alpha|\varepsilon|^2 \tag{120}$$

where we recall that we defined $v \in \mathbb{R}^{d-1}$ by $v_i = \mathbb{E}(h_d(S) S_i g'(S_d)) + \mathbb{E}(h_i(S) g(S_d))$ for $i = 1, \dots, d-1$ and $\alpha_d = \mathbb{E} g(S_d) S_d - \mathbb{E} g'(S_d)$. Then we can rewrite (113) as

$$\left|\bar{H}(\varepsilon) - \bar{H}(0) - f(\varepsilon)\right| < \Xi(|\varepsilon|^3 + 2\eta^3). \tag{121}$$

In other words $\bar{H}$ agrees with f up to a constant term and error terms of order 3 in ε and η. If we just use this bound on $|\bar{H} - f|$ we could prove (90) but only with an error term of order $\eta^{\frac{3}{2}}$. To obtain the better rate η^2 we need to consider a second order expansion.

We assume now that $\alpha_d > 0$ such that $\bar{H}$ and f are concave in a neighborhood $B_\kappa(0)$ (the proof for $\alpha_d < 0$ is very similar) for $\eta < \eta_0$. For $\alpha_d < 0$ a similar reasoning applies and only some inequalities are reversed.

We first expand the function f. The relation $\alpha_d > 0$ implies that f is maximized at

$$\varepsilon_{\max} = \frac{\eta v}{\alpha_d}, \quad \text{and} \quad f(\varepsilon) \leq f(\varepsilon_{\max}) = \frac{\eta^2 |v|^2}{2\alpha_d}. \tag{122}$$

and we have for $\varepsilon = \varepsilon_{\max} + \Delta\varepsilon$ the expansion

$$f(\varepsilon) = f(\varepsilon_{\max} + \Delta\varepsilon) = f(\varepsilon_{\max}) - \frac{1}{2}\alpha_d |\Delta\varepsilon|^2. \tag{123}$$

Now we consider similar expansions for $\bar{H}$. For concreteness, we introduce the constant Ξ_2 such that

$$|D^2 \bar{H}(\varepsilon) + \alpha_d \cdot \mathrm{Id}| \leq \Xi_2(|\varepsilon| + \eta). \tag{124}$$

Such a constant exists by (119).

Recall that we assumed that $\alpha_d > 0$ such that $\bar{H}$ is concave in a neighborhood $B_\kappa(0)$ for $\eta < \eta_0$. Suppose ε_0 is the unique global maximum of $\bar{H}$ on $\overline{B_\rho(0)}$ for

$$\rho = \frac{2\eta(|v| + 1)}{|\alpha_d|} \tag{125}$$

where we assume that η is sufficiently small such that $\bar{H}$ is uniformly convex on B_ρ. Note that then either $D\bar{H}(\varepsilon_0) = 0$ or $\varepsilon_0 \in \partial B_\rho(0)$ and $D\bar{H}(\varepsilon_0)(\varepsilon - \varepsilon_0) < 0$ for all $\varepsilon \in B_\rho(0)$. Now, the general heuristic is that $\bar{H}$ roughly behaves like a parabola with vertex ε_0 and f is a parabola with vertex $\varepsilon_{\max}$ and both parabolas have approximately the same second derivative. Then, the distance between the two parabolas will increase for large arguments, leading to a contradiction to (121).

We now define the point ε_1 as the intersection of the ray from $\varepsilon_{\max}$ to ε_0 with the set $\partial B_{2\rho}(0)$. We define

$$\delta = |\varepsilon_{\max} - \varepsilon_0| \tag{126}$$

(the quantity we want to bound) and

$$\mu = |\varepsilon_0 - \varepsilon_1| \geq \rho \tag{127}$$

(since $\varepsilon_0 \in B_\rho(0)$ and $|\varepsilon_1| = 2\rho$). Note that ε_1 is on the ray from $\varepsilon_{\max}$ to ε_0 so we have $|\varepsilon_1 - \varepsilon_{\max}| = |\varepsilon_1 - \varepsilon_0| + |\varepsilon_0 - \varepsilon_{\max}| = \delta + \mu$. We then find using (123)

$$f(\varepsilon_0) - f(\varepsilon_1) == (f(\varepsilon_{\max}) - \frac{1}{2}\alpha_d \delta^2) - (f(\varepsilon_{\max}) - \frac{1}{2}\alpha_d(\delta + \mu)^2) = -\frac{1}{2}\alpha_d \delta^2 + \frac{1}{2}\alpha_d(\delta + \mu)^2 = \frac{1}{2}\alpha_d(\mu^2 + 2\mu\delta). \tag{128}$$

We now derive a similar bound for $\bar{H}$ for which we need the second order Taylor expansion with integral remainder which reads for $g : \mathbb{R}^d \to \mathbb{R}$ with $g \in C^2$ as follows

$$g(x) = g(x_0) + Dg(x_0)(x - x_0) + \int_0^1 (1 - t)(x - x_0)^\top D^2 g(x_0 + t(x - x_0))(x - x_0)\, \mathrm{d}t. \tag{129}$$

We apply this with ε_0 and ε_1. We observe that since $\varepsilon_1 - \varepsilon_0$ and $\varepsilon_{\max} - \varepsilon_0$ point in opposite directions, the relation $D\bar{H}(\varepsilon_0)(\varepsilon_{\max} - \varepsilon_0) \leq 0$ (since $\varepsilon_{\max} \in B_\rho(0)$) implies

$$D\bar{H}(\varepsilon_0)(\varepsilon_1 - \varepsilon_0) \geq 0. \tag{130}$$

Then we find from the second order expansion (129) using the last display and (119)

$$\begin{aligned}\bar{H}(\varepsilon_1) - \bar{H}(\varepsilon_0) &= D\bar{H}(\varepsilon_0)(\varepsilon_1 - \varepsilon_0) + \int_0^1 (1-t)(\varepsilon_1 - \varepsilon_0)^\top D^2\bar{H}(\varepsilon_0 + t(\varepsilon_1 - \varepsilon_0))(\varepsilon_1 - \varepsilon_0)\,\mathrm{d}t \\ &\geq 0 - (\alpha_d + \Xi_2(2\rho + \eta))\,(\varepsilon_1 - \varepsilon_0)^2 \int_0^1 (1-t)\,\mathrm{d}t \\ &= -\frac{1}{2}\,(\alpha_d + \Xi_2(2\rho+\eta))\,\mu^2.\end{aligned} \tag{131}$$

Using (121) followed by (128) and (131) we now find

$$\begin{aligned}4\Xi(4\rho^3 + \eta^3) &\geq |\bar{H}(\varepsilon_0) - \bar{H}(0) - f(\varepsilon_0)| + |\bar{H}(\varepsilon_1) - \bar{H}(0) - f(\varepsilon_1)| \\ &\geq -\big(\bar{H}(\varepsilon_0) - \bar{H}(0) - f(\varepsilon_0)\big) + \big(\bar{H}(\varepsilon_1) - \bar{H}(0) - f(\varepsilon_1)\big) \\ &= (\bar{H}(\varepsilon_1) - \bar{H}(\varepsilon_0)) + (f(\varepsilon_0) - f(\varepsilon_1)) \\ &\geq -\frac{1}{2}\,(\alpha_d + \Xi_2(2\rho+\eta))\,\mu^2 + \frac{1}{2}\alpha_d(\mu^2 + 2\mu\delta) \\ &= \alpha_d\mu\delta - \frac{1}{2}\Xi_2(2\rho+\eta)\mu^2.\end{aligned} \tag{132}$$

We conclude using $\rho \leq \mu \leq 2\rho$ and the definition (125) that

$$\begin{aligned}\delta &\leq \frac{4\Xi(4\rho^3+\eta^3)}{\alpha_d\mu} + \frac{1}{2\alpha_d}\Xi_2(2\rho+\eta)\mu \\ &\leq \frac{4\Xi}{\alpha_d}\left(4\rho^2 + \frac{\eta^3}{\rho}\right) + \frac{1}{\alpha_d}\Xi_2(2\rho^2 + \eta\rho) \\ &\leq \frac{4\Xi}{\alpha_d}\left(\frac{16\eta^2(|v|+1)^2}{\alpha_d^2} + \frac{\alpha_d\eta^2}{2(|v|+1)}\right) + \frac{\Xi_2}{\alpha_d}\left(\frac{8\eta^2(|v|+1)^2}{\alpha_d^2} + \frac{2\eta^2(|v|+1)}{\alpha_d}\right) \\ &\leq C\eta^2.\end{aligned} \tag{133}$$

For η sufficiently small this implies that $\delta < \rho/2$ and thus $\varepsilon_0 \in B_\rho(0)$ (i.e., in the interior and not on the boundary) so that ε_0 really is a local maximum of $\bar{H}$. The proof for $\alpha_d < 0$ follows similarly.

□

We now extend Lemma F.1 by considering $A \neq \mathrm{Id}$, including the whitening, and considering the function on the sphere, thus proving Theorem 5.5.

Proof of Theorem 5.5. Recall that we defined $H : S^{d-1} \to \mathbb{R}$ by

$$H_\eta(w) = \mathbb{E}G(w^\top \Sigma_X^{-\frac{1}{2}} X) \tag{134}$$

where $\Sigma_X = \mathbb{E}(XX^\top) = AA^\top + \eta^2\Omega$ with $\Omega = \mathbb{E}(h(S)h(S)^\top)$. Note that Σ_X depends implicitly on η so we will indicate this dependence in the following for some quantities. We now relate this to the setting in Lemma F.1. The function H defines a map on the manifold S^{d-1}. We analyze its properties by considering a suitable chart. We define the map $T : \mathbb{R}^{d-1} \to S^{d-1}$ by

$$T_\eta(\varepsilon) = \frac{\Sigma_X^{\frac{1}{2}} A^{-\top}\begin{pmatrix}\varepsilon \\ 1\end{pmatrix}}{\left|\Sigma_X^{\frac{1}{2}} A^{-\top}\begin{pmatrix}\varepsilon \\ 1\end{pmatrix}\right|}. \tag{135}$$

This map defines a chart locally around $\varepsilon = 0$. Recall that we defined in (88) (indicating the parameter dependence for clarity)

$$\bar{H}_{\eta,h}(\varepsilon) = \tilde{H}_{\eta,h}(w) = \mathbb{E}G\left(\frac{w^\top(S + \eta h(S))}{\sigma_w}\right) \tag{136}$$

where $w = (\varepsilon, 1)^\top$ and $\sigma_w = \mathbb{E}(wX)^2$ with $X = S + \eta h(S)$. Then the relation

$$H_\eta(T_\eta(\varepsilon)) = \bar{H}_{\eta, A^{-1}h}(\varepsilon) \tag{137}$$

holds. Indeed, writing $w = (\varepsilon, 1)^\top$ we obtain

$$\begin{aligned} H_\eta(T_\eta(\varepsilon)) &= \mathbb{E}G\left(\frac{\left(\Sigma_X^{\frac{1}{2}} A^{-\top} w\right)^\top \Sigma_X^{-\frac{1}{2}} X}{\left|\Sigma_X^{\frac{1}{2}} A^{-\top} w\right|}\right) \\ &= \mathbb{E}G\left(\frac{w^\top A^{-1}(AS + \eta h(S))}{\left|\Sigma_\eta^{\frac{1}{2}} A^{-\top} w\right|}\right) \\ &= \mathbb{E}G\left(\frac{w^\top (S + \eta A^{-1} h(S))}{\sigma_w}\right) = \bar{H}_{\eta, A^{-1}h}(\varepsilon) \end{aligned} \tag{138}$$

where we used that (recall $\mathbb{E}(Sh(S)^\top) = 0$)

$$\begin{aligned} \sigma_w^2 = \mathbb{E}((w^\top (S + \eta A^{-1} h(S)))^2) &= w \cdot \mathbb{E}\left((S + \eta A^{-1} h(S))(S^\top + \eta h(S)^\top A^{-\top})\right) w \\ &= w \cdot (\mathrm{Id} + \eta^2 A^{-1} \Omega A^{-\top} w = (A^{-\top} w) \cdot (AA^\top + \eta^2 \Omega) A^{-\top} w \\ &= (A^{-\top} w) \cdot \Sigma_X A^{-\top} w = \left|\Sigma_X^{\frac{1}{2}} A^{-\top} w\right|^2. \end{aligned} \tag{139}$$

Note that this is not surprising as both terms were chosen such that the argument of G has unit variance. Note that $|A^{-1}h(S)| \leq |A^{-1}| \cdot |h(S)|$ so we find $\|A^{-1}h\|_{\mathbb{P},q}^q \leq |A^{-1}| \cdot \|h\|_{\mathbb{P},q}$. Now we apply Lemma F.1 to the function $\bar{H}_{\eta, A^{-1}h}(\varepsilon) = \bar{H}_{|A^{-1}|\eta, |A^{-1}|^{-1}A^{-1}h}(\varepsilon)$ where $\||A^{-1}|^{-1}A^{-1}h\|_{\mathbb{P},q} \leq 1$. Lemma F.1 implies that $\bar{H}$ has a local extremum at some $\varepsilon_0 = \eta \alpha_d^{-1} v + O(\eta^{\frac{3}{2}})$ for $\eta < \eta_0$ and some η_0 where $\alpha_d = \mathbb{E}g(S_d)S_d - \mathbb{E}g'(S_d)$ and

$$v_i = \mathbb{E}((A^{-1}h)_d(S) g'(S_d) S_i + g(S_d)(A^{-1}h)_i(S)). \tag{140}$$

Thus H_η has a local extremum at

$$w_d = T_\eta(\varepsilon) \propto \Sigma_X^{\frac{1}{2}} A^{-\top} \begin{pmatrix} \eta \alpha_d^{-1} v \\ 1 \end{pmatrix} + O(\eta^2). \tag{141}$$

Since $\Sigma_X = AA^\top + O(\eta^2)$ we also find (recall $\bar{w_d} = (AA^\top)^{\frac{1}{2}} A^{-\top} e_d$ by (12)) that $|w_d - \bar{w_d}| = O(\eta)$. The estimated independent component is given by

$$w_d^\top \Sigma_X^{-\frac{1}{2}} X = \begin{pmatrix} \eta \alpha_d^{-1} v \\ 1 \end{pmatrix} \cdot S + \eta \begin{pmatrix} \eta \alpha_d^{-1} v \\ 1 \end{pmatrix} \cdot A^{-1} h(S) = S_d + O(\eta) \tag{142}$$

and

$$\Sigma_X^{\frac{1}{2}} w_0 = A^{-\top} \begin{pmatrix} \eta \alpha_d^{-1} v \\ 1 \end{pmatrix} + O(\eta^2) = (A^{-1})_{d,:} + O(\eta), \tag{143}$$

i.e., we recover the d-th row of the unmixing matrix up to errors of order η.

Finally, we prove the convexity of H around $\Sigma_X^{\frac{1}{2}} A^{-\top} e_d$ for $\alpha < 0$. Intuitively, this is not surprising as we proved the convexity of $\bar{H}$ around $\varepsilon = 0$ and the relation (137) should lift this to the map H. The formal proof requires tools from differential geometry, which we use freely. The proof relies on the following relation from Riemannian geometry

$$\mathrm{Hess}(H) = \sum_{i,j} \frac{\partial^2 \bar{H}}{\partial \varepsilon_i \, \partial \varepsilon_j} \mathrm{d}\varepsilon_i \otimes \mathrm{d}\varepsilon_j - \sum_{i,j,k} \Gamma_{ij}^k \frac{\partial \bar{H}}{\partial \varepsilon_k} \mathrm{d}\varepsilon_i \otimes \mathrm{d}\varepsilon_j. \tag{144}$$

Here Γ_{ij}^k denotes the Christoffel symbols expressed in the chart T_η where we use the induced metric on S^{d-1} as a submanifold of $\mathbb{R}^d$.

We have shown in Lemma F.1 that the matrix with entries $D^2\bar{H}$ satisfies for $\alpha < 0$

$$D^2\bar{H} \geq |\alpha| \cdot \mathrm{Id} + O(\eta + |\varepsilon|) \tag{145}$$

and that

$$D\bar{H} = O(|\varepsilon| + \eta). \tag{146}$$

It is straightforward to show using the calculations below that the Christoffel symbols are bounded (in fact small, but we do not need this). This and the last display imply that the last term in (144) is bounded $O(\eta + |\varepsilon|)$. We now consider the tangent vectors ∂_{ε_i} which can be identified with the vector $\partial_{\varepsilon_i} T_\eta(\varepsilon) \in \mathbb{R}^d$ that is tangential to S^{d-1}. We note that

$$\left|\Sigma_X^{\frac{1}{2}} A^{-\top} \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}\right|^2 = \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}^\top A^{-1}\Sigma_X A^{-\top} \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix} = \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}^\top (\mathrm{Id} + \eta^2 A^{-1}\Omega A^{-\top}) \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix} = 1 + |\varepsilon|^2 + \eta^2 P(\varepsilon) \tag{147}$$

where P denotes a quadratic polynomial in ε. This implies

$$\partial_{\varepsilon_i} T_\eta(\varepsilon) = \frac{\Sigma_X^{\frac{1}{2}} A^{-\top} e_i}{\left|\Sigma_X^{\frac{1}{2}} A^{-\top} \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}\right|} + O(|\varepsilon| + \eta). \tag{148}$$

This implies

$$g_{S^{d-1}}(\partial_{\varepsilon_i}, \partial_{\varepsilon_j}) = \frac{e_i^\top A^{-1}\Sigma_X A^{-\top} e_j}{\left|\Sigma_X^{\frac{1}{2}} A^{-\top} \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}\right|^2} + O(|\varepsilon| + \eta) = \frac{\delta_{ij} + O(\eta^2)}{1 + O(|\varepsilon|^2 + \eta^2} + O(|\varepsilon| + \eta) = \delta_{ij} + O(|\varepsilon| + \eta). \tag{149}$$

This means that, up to higher-order terms, the metric g agrees with the standard metric. We conclude that for any tangent vector $Y = \sum_{i=1}^{d-1} y_i \partial_{\varepsilon_i}$ we obtain

$$\frac{\mathrm{Hess}(H)(Y,Y)}{g_{S^{d-1}}(Y,Y)} \geq \frac{|\alpha| \cdot |y|^2}{|y|^2} + O(\eta + |\varepsilon|). \tag{150}$$

We conclude that for $\eta \leq \eta_0$ there is a neighborhood of $T_\eta(0) \propto \Sigma_X^{\frac{1}{2}} A^{-\top} e_d$ where $\mathrm{Hess}(H) \geq |\alpha|/2 \cdot \mathrm{Id}$, i.e. H is strictly convex. The proof for $\alpha > 0$ is similar.

□

The proof of Theorem 5.7 is now trivial.

Proof of Theorem 5.7. By the assumptions of the Theorem and Theorem 5.5 we find for each $1 \leq i \leq d$ a vector w_i with $|w_i| = 1$ such that H has a local extremum at w_i and $A^\top \Sigma_X^{-\frac{1}{2}} w_i = e_i + O(\eta)$ (when being pedantic, we apply the result to permuted data changing coordinates i and d). Then the matrix W with rows $w_i = W^\top e_i$ is the desired matrix. □

The proof of Corollary 5.8 follows from standard results in convex optimization.

Proof of Corollary 5.8. It is well known that gradient descent locally converges for a convex function and sufficiently small step size, a proof for the general case where we optimize over a Riemannian manifold (in our case the sphere) can be found, e.g., in (Boumal, 2023). Local convergence of projected gradient descent can also be shown. □

Finally, we prove that WX essentially recovers the true sources as stated in Theorem 5.9.

Proof of Theorem 5.9. We verify the assumption of Lemma B.2 for S and

$$\hat{S} = W\Sigma_X^{-\frac{1}{2}}X = W\Sigma_X^{-\frac{1}{2}}(AS + \eta h(S)), \tag{151}$$

i.e., we define $T(s) = (W\Sigma_X^{-\frac{1}{2}}A)s + \eta W\Sigma_X^{-\frac{1}{2}}h(s))$. Since we assume that S_i have unit variance, the bound (29) holds with $c_1 = 1$. Let us set $P = W\Sigma_X^{-\frac{1}{2}}A$ and verify condition (30) for P. By construction of W we have

$$e_i^\top P = e_i^\top W\Sigma_X^{-\frac{1}{2}}A = w_i^\top \Sigma_X^{-\frac{1}{2}}A = e_i + O(\eta). \tag{152}$$

This implies that when setting $\rho(i) = i$ then we find

$$\frac{1}{|P_{i,i}|^2}\sum_{j\neq i}|P_{i,j}|^2 \leq \frac{O(\eta^2)}{1 - O(\eta)} = O(\eta^2) \tag{153}$$

for sufficiently small η. Finally, we use the bound

$$\|\eta W\Sigma_X^{-\frac{1}{2}}h(S)\|_{\mathbb{P},2} \leq \eta|W| \cdot |\Sigma_X^{\frac{1}{2}}| \cdot \|h\|_{\mathbb{P},q} = O(\eta). \tag{154}$$

Here we used that $|W| = d$ since its rows are normalized. This implies

$$\frac{\|(\eta W\Sigma_X^{-\frac{1}{2}}h)_i(S)\|_{\mathbb{P},2}}{|P_{i,i}|\|Z\|_{\mathbb{P},2}} \leq O(\eta) \tag{155}$$

and therefore (31) holds with $c_3 = O(\eta)$. Now we can apply Lemma B.2 and conclude that

$$\mathrm{MCC}(S, \hat{S}) \geq 1 - C\eta^2 \tag{156}$$

for some constant $C > 0$. □

G. Proof for approximate identifiability for ICA with almost locally isometric mixing

In this section we prove Theorem 5.11 and then provide some extensions that allow us to consider $\mathbb{P}$ with unbounded support. However, we first prove one simple auxiliary lemma which allows us to transform a data representation $X = AS + h(S) - b$ to $X - b' = A'S + h'(S)$ such that h' is centered and $\mathbb{E}(Sh'(S)) = 0$.

Lemma G.1. *Suppose S has distribution $\mathbb{P}$ such that $\mathbb{E}(S) = 0$ and $\mathbb{E}SS^\top \geq \alpha^{-1}\mathrm{Id}$ in the sense of symmetric matrices for some $\alpha > 0$. We assume that $X = AS + h(S) + b$ for an orthogonal matrix $A \in \mathrm{SO}(d)$, some function $h : \mathbb{R}^d \to \mathbb{R}^d$ and $b \in \mathbb{R}^d$. Then there is a linear map A', $b' \in \mathbb{R}^d$ and $h' : \mathbb{R}^d \to \mathbb{R}^d$ such that $X - b' = A'S + h'(S)$ and*

$$\mathbb{E}_S(h'(S)) = 0, \quad \mathbb{E}_S(S^\top h'(S)) = 0. \tag{157}$$

Morevoew, for $q \geq 2$,

$$\|h'\|_{\mathbb{P},q} < 2\left(1 + \alpha\|S\|_{\mathbb{P},q}^2\right)\|h(S)\|_{\mathbb{P},q}. \tag{158}$$

and

$$\sigma_{\min}(A') \geq 1 - 2\alpha\|S\|_{\mathbb{P},2} \cdot \|h(S)\|_{\mathbb{P},q}. \tag{159}$$

Proof. First, we find

$$X - \mathbb{E}X = AS + (h(S) + b - \mathbb{E}X) = AS + h_c(S) \tag{160}$$

where we defined $h_c(S) = h(S) + b - \mathbb{E}X$. Note that $\mathbb{E}h_c(S) = \mathbb{E}X - \mathbb{E}X - A\mathbb{E}S = 0$ and thus $h_c(S) = h(S) - \mathbb{E}h(S)$. By the triangle inequality and Hölder's inequality (which implies $\|f\|_{\mathbb{P},p} \leq \|f\|_{\mathbb{P},q}$ for $p < q$) we get

$$\|h_c\|_{\mathbb{P},q} = \|h - \mathbb{E}_{\mathbb{P}}h\|_{\mathbb{P},q} \leq \|h\|_{\mathbb{P},q} + \|\mathbb{E}_{\mathbb{P}}h\|_{\mathbb{P},q} \leq \|h\|_{\mathbb{P},q} + |\mathbb{E}_{\mathbb{P}}h| \leq 2\|h\|_{\mathbb{P},q}. \tag{161}$$

Thus we find that h_c is centered, and its q norm is at most larger by a factor of 2 than the corresponding norm of h. Then we define

$$h'(S) = h_c(S) - \mathbb{E}(h_c(S)S^\top)\mathbb{E}(SS^\top)^{-1}S, \quad A' = A + \mathbb{E}(h_c(S)S^\top)\mathbb{E}(SS^\top)^{-1} \tag{162}$$

i.e., we regress $h_c(S)$ linearly on S so that $\mathbb{E}(h'(S)S^\top) = 0$. We bound using Cauchy-Schwarz and $q \geq 2$

$$\begin{aligned}\|\mathbb{E}(h_c(S)S^\top)\mathbb{E}(SS^\top)^{-1}S\|_{\mathbb{P},q} &\leq \alpha|\mathbb{E}(h_c(S)S^\top)|\|S\|_{\mathbb{P},q} \leq \alpha\mathbb{E}(|h_c(S)S^\top|)\|S\|_{\mathbb{P},q} \\ &\leq \alpha\mathbb{E}(|h_c(S)| \cdot |S|)\|S\|_{\mathbb{P},q} \leq \alpha\|h_c(S)\|_{\mathbb{P},2}\|S\|_{\mathbb{P},2}\|S\|_{\mathbb{P},q} \leq \alpha\|h_c(S)\|_{\mathbb{P},q}^2\|S\|_{\mathbb{P},q}.\end{aligned} \tag{163}$$

This implies

$$\|h'(S)\|_{\mathbb{P},q} \leq \left(1 + \alpha\|S\|_{\mathbb{P},q}^2\right)\|h_c(S)\|_{\mathbb{P},q} \leq 2\left(1 + \alpha\|S\|_{\mathbb{P},q}^2\right)\|h(S)\|_{\mathbb{P},q}. \tag{164}$$

Finally we conclude the lower bound on A'. We observe that

$$|A' - A| = |\mathbb{E}(h_c(S)S^\top)\mathbb{E}(SS^\top)^{-1}| \leq \alpha\|h_c(S)\|_{\mathbb{P},2} \cdot \|S\|_{\mathbb{P},2} \leq 2\alpha\|S\|_{\mathbb{P},2} \cdot \|h(S)_{\mathbb{P},q}. \tag{165}$$

Using $A \in \mathrm{SO}(d)$ we conclude that

$$\sigma_{\min}(A') \geq 1 - 2\alpha\|S\|_{\mathbb{P},2} \cdot \|h(S)_{\mathbb{P},q}. \tag{166}$$

□

We can now prove Theorem 5.11.

Proof of Theorem 5.11. The proof essentially follows by combining Theorem 4.1 and Theorem 5.9. We note that by assumption the support of $\mathbb{P}$ is connected and by independence of the S_i it is actually a cuboid and thus a Lipschitz domain. Using Theorem 4.1 we find that if we define g as in (8) (or (73) for $d < D$), then $\tilde{X} = g^{-1} \circ f(S) = g^{-1}(X)$ can be expressed as

$$\tilde{X} = g^{-1} \circ f(S) = AS + h(S) + b \tag{167}$$

for some $A \in \mathrm{SO}(d)$ and h satisfies the bound

$$\|h\|_{\mathbb{P},q} \leq C_1\Theta_p(f, \Omega) \tag{168}$$

where p and q are as in the statement of the Theorem, i.e, $q = \max(3, d_g)$ and $p = dq/(d+q)$ which is equivalent to $q = pd/(d-p)$, as required in Theorem 4.1. By shifting $\tilde{X}$ we can assume that $\tilde{X}$ is centered. Now we apply Lemma G.1 from above and find that

$$\tilde{X} = A'S + h'(S) \tag{169}$$

where

$$\sigma_{\min}(A') \geq 1 - 2\alpha\|S\|_{\mathbb{P},2} \cdot \|h(S)\|_{\mathbb{P},q} \geq 1 - 2\alpha\|S\|_{\mathbb{P},2}C_1\Theta_p(f, \Omega) > \frac{1}{2} \tag{170}$$

if $\Theta_p(f, \Omega)$ sufficiently small. Moreover, there is a constant $C_1' > 0$ depending on C_1 and the distribution $\mathbb{P}$ such that

$$\|h'\|_{\mathbb{P},q} < C_1'\Theta_p(f, \Omega). \tag{171}$$

We can write

$$h(S) = (C_1'\Theta_p(f, \Omega)) \cdot (C_1'\Theta_p(f, \Omega))^{-1}h(S) = \eta \cdot \frac{h(S)}{C_1'\Theta_p(f, \Omega)} \tag{172}$$

where $\eta = C_1'\Theta_p(f, \Omega)$. Now we apply Theorem 5.7. If $\eta = C_1'\Theta_p(f, \Omega) < \eta_0$ we find that there is a matrix $W \in \mathbb{R}^{d \times d}$ with rows w_i such that H (defined using the distribution of $\tilde{X}$) has local extrema at w_i. Moreover, by Theorem 5.9 the reconstructed sources $\hat{S} = W\tilde{X} = W\tilde{g}(X)$ satisfy

$$\mathrm{MCC}(\hat{S}, S) \geq 1 - C_2\eta^2 \geq 1 - C_3\Theta_p^2(f, \Omega) \tag{173}$$

for $C_3 = C_2C_1'^2$. □

We can also extend the results to ICA with approximately locally isometric mixing function and unbounded support of the sources. To simplify the analysis, we restrict our attention to functions that are perfect local isometries away from a compact set, i.e., we consider

$$\begin{aligned}\mathcal{F}_{c-\mathrm{iso}}(d,D,\Omega) = \{ f : \mathbb{R}^d \to M \subset \mathbb{R}^D : f|_{\Omega^\complement} \text{ is a} \\ \text{local isometry and } f(\omega) \text{ is not} \\ \text{isometric to } (\mathbb{R}^d, g_{\mathrm{stand}}) \text{ for any } \omega \subset \Omega \ \}.\end{aligned} \tag{174}$$

Here g_{stand} denotes the standard metric on $\mathbb{R}^d$. The definition essentially ensures that Ω is the maximal set such that outside of Ω the function f is a local isometry. Note that the second part of the condition is satisfied, e.g., when the curvature of $f(\Omega)$ is everywhere non-vanishing. Then we have the following result.

Theorem G.2. *Suppose that the mixing $f \in \mathcal{F}_{c-\mathrm{iso}}(d, D, \Omega)$ for some bounded set Ω such that $\Omega^\complement$ is connected. Assume that $X = f(S)$ where $S \sim \mathbb{P}$. Assume that $\mathbb{P}$ of S satisfies Assumption 5.2, 5.3, and 5.4 for some contrast function G and the density of $\mathbb{P}$ is bounded above and everywhere positive on $\mathbb{R}^d$ and lower bounded on Ω. Let $q = \max(d_g, 3)$ and $p = dq/(d+q)$. Assume that $\Theta_p(f, \Omega)$ is sufficiently small. Then we can find a transformation $\tilde{g} : M \to \mathbb{R}^d$ such that the transformed data $\tilde{X} = \tilde{g}(X)$ is centered and has the property that there is a matrix $W \in \mathbb{R}^{d\times d}$ whose rows are d normalized vectors w_i which are local extrema of the function $w \to \mathbb{E}(G(w^\top \Sigma_{\tilde{X}}^{-\frac{1}{2}} \tilde{X}))$ such that $\hat{S} = W\tilde{X}$ satisfies*

$$\mathrm{MCC}(\hat{S}, S) \geq 1 - C\Theta_p^2(f, \Omega) \tag{175}$$

where C depends on Ω, d, and the parameters in the Assumptions 5.2, 5.3, and 5.4.

Remark G.3.
- Of course, we expect to recover the vectors w_i typically by applying an algorithm for linear ICA to the transformed data $\tilde{X}$, however, due to the lack of theoretical guarantees for ICA algorithms we cannot show a stronger statement (see Remark 5.10).
- We expect that the assumption that f is a local isometry outside Ω can be relaxed, e.g., by using bounded contrast functions like $G(s) = \exp(-s^2/2)$.

We now prove Theorem G.2. In its proof we need a slightly different rigidity result in this case (actually a simpler result because we in addition fix the boundary) which we now state. We denote by $\sigma_{\max}$ the largest singular value of a matrix.

Theorem G.4 (Theorem 2.1 in Kohn (1982))**.** *Let $1 \leq p \leq d$, suppose there is a rigid motion $L(z) = Az + b$ with $A \in \mathrm{SO}(d)$ and $u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)$. We assume that there is a bounded set Ω such that $u(z) = L(z)$ for $z \notin \Omega$. Then*

$$\|u - L\|_q \leq C(d,p)\|(\sigma_{\max}(Du) - 1)_+\|_p \tag{176}$$

where $q = np/(n-p)$.

Remark G.5.
1. Note that by Lemma C.1 we can bound

$$(\sigma_{\max}(Du) - 1)_+ \leq \sqrt{\sum_i (\sigma_i - 1)^2} = \mathrm{dist}(A, \mathrm{SO}(d)). \tag{177}$$

2. The exponent q agrees with the exponent for the Sobolev embedding result.
3. There is an extension to $p > d$ where $\|u - L\|_\infty$ can be bounded.
4. The result Theorem E.1 is more general than this result because it does not assume affine boundary values. We cannot directly apply Theorem E.1 in our context because we need that the linear map in (67) can be chosen as the affine boundary values L. While this is the case, it requires to inspect the proof carefully, and it is simpler to rely on this known result from the literature.

We can now continue with the Proof of Theorem G.2. To clarify the proof structure, we split the proof into two parts. First, we establish that the assumptions of the Theorem imply that we can reduce the problem to perturbed linear ICA. We summarize this result in the following proposition.

Proposition G.6. *Suppose that the mixing $f \in \mathcal{F}_{c-\mathrm{iso}}(d, D, \Omega)$ for some bounded set Ω and $\Omega^{\complement}$ connected. Suppose that $X = f(S)$ with $S \sim \mathbb{P}$. Assume that the density of S is globally positive and upper bounded and, moreover, lower bounded on Ω. Then we can find a map $g : M \to \mathbb{R}^d$ such that the recovered sources $X' = g(X)$ satisfies*

$$X' = AS + h(S) \tag{178}$$

for some rotation $A \in \mathrm{SO}(d)$ and function h such that

$$\|h\|_q \leq C\Theta_p(f, \Omega) \tag{179}$$

where $q = pd/(d-p)$ and C depends on d, p, and the bounds on the density of $\mathbb{P}$ on Ω.

Proof. For a diffeomorphism $g : \mathbb{R}^d \to M$ we let $I(g) \subset M$ be the maximal open set such that $g^{-1}|_{I(g)}$ is a local isometry. We consider the set $\mathcal{G}$ of diffeomorphisms from $\mathbb{R}^d$ to M such that $I(g)$ is maximal. By definition of the function class $\mathcal{F}_{c-\mathrm{iso}}(d, D, \Omega)$ there are functions g such that $I(g) = f(\mathring{\Omega}^{\complement})$ (e.g., $g = f$), moreover $I(g) \cap f(\Omega) = \emptyset$ for any g by definition of $\mathcal{F}_{c-\mathrm{iso}}(d, D, \Omega)$. Therefore, $\mathcal{G}$ is well-defined. Now we consider $g \in \mathcal{G}$ such that

$$g \in \operatorname*{argmin}_{\bar{g} \in \mathcal{G}} \int_{\mathbb{R}^d} \mathrm{dist}^p(D\bar{g}^{-1}(x), \mathrm{SO}(T_x M, d))\, f_* \mathbb{P}(\mathrm{d}x). \tag{180}$$

By shifting g we can assume that $g^{-1} \circ f(S)$ is centered. Then we consider the transition function $T = g^{-1} \circ f$. By assumption $f|_{\Omega^{\complement}}$ is a local isometry, and we have shown that $g^{-1}|_{f(\Omega^{\complement})}$ is a local isometry and thus $T|_{\Omega^{\complement}}$ is a local isometry and therefore $T(s) = As + b = L(s)$ for some $A \in \mathrm{SO}(d)$ by Theorem 2.3. We now consider the restriction $T|_{\Omega}$ and apply Lemma E.3 to find

$$\begin{aligned} &\int_{\Omega} \mathrm{dist}^p(DT(s), \mathrm{SO}(d))\, \mathrm{d}s \\ &\qquad \leq 3^p \int_{\Omega} \mathrm{dist}^p(Df(s), \mathrm{SO}(d, T_{f(s)}M))\, \mathrm{d}s + C \int_{\Omega'} \mathrm{dist}^p(Dg^{-1}(g(s)), \mathrm{SO}(T_{f(s)}M, d))\, \mathbb{Q}(\mathrm{d}s). \end{aligned} \tag{181}$$

Now since $f \in \mathcal{G}$ we can conclude (using the upper bound on the density of $\mathbb{P}$) as before that

$$\begin{aligned} \int_{\Omega'} \mathrm{dist}^p(Dg^{-1}(g(s)), \mathrm{SO}(T_{f(s)}M, d))\, \mathbb{Q}(\mathrm{d}s) &\leq \int_{\Omega} \mathrm{dist}^p(Df^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\, \mathbb{P}(\mathrm{d}s) \\ &\leq C \int_{\Omega} \mathrm{dist}^p(Df^{-1}(f(s)), \mathrm{SO}(T_{f(s)}M, d))\, \mathrm{d}s. \end{aligned} \tag{182}$$

Combining the last two displays we find

$$\int_{\Omega} \mathrm{dist}^p(DT(s), \mathrm{SO}(d))\, \mathrm{d}s \leq C\Theta_p^p(f, \Omega). \tag{183}$$

Now we apply Theorem G.4 (and the remark below this result) to conclude that

$$\|T - L\|_q \leq C\Theta_p(f, \Omega). \tag{184}$$

Defining $h = T - L$ and using the upper bound on the density of $\mathbb{P}$ we conclude that $X' = g^{-1}(X) = T(S)$ is given by

$$X = L(S) + h(S) \tag{185}$$

and

$$\|h\|_{\mathbb{P},q} \leq C\Theta_p(f, \Omega). \tag{186}$$

□

Now we can complete the proof of Theorem G.2.

Proof of Theorem G.2. The proof is now the same as the proof of Theorem 5.11 except that we first apply Proposition G.6 instead of Theorem 4.1. Indeed, by applying Proposition G.6 and the definition of g used in the proof we find that

$$\tilde{X} = g(X) = g(f(S)) = AS + h(S) + b \tag{187}$$

where $A \in \mathrm{SO}(d)$ and $\|h\|_{\mathbb{P},q} \leq C\Theta_p(f, \Omega) = C\Theta_p(f, \mathbb{R}^d)$. The rest of the proof is the same as the proof of Theorem 5.11, i.e., we apply Lemma G.1 to rewrite $\tilde{X} = A'S + h'(S)$ where $h'(S)$ is centered and $\mathbb{E}h'(S)S^\top) = 0$ and then apply Theorem 5.9. Note that those two results do not assume bounded support of S. □

H. Experimental Illustration of non-robustness of piecewise linear functions

In this appendix we collect some additional details on the experimental illustration of the non-robustness of learning piecewise linear mixing functions. Recall that we consider Z following a two-dimensional standard normal distribution. We let m be a radius dependent rotation (Hyvärinen & Pajunen, 1999) which is given by

$$m(z_1, z_2) = \begin{pmatrix} \cos(5\varphi(z^2))z_1 + \sin(5\varphi(z^2))z_2 \\ -\sin(5\varphi(z^2))z_1 + \cos(5\varphi(z^2))z_2 \end{pmatrix} \tag{188}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is a smooth bump function supported in $(0, 1)$ and given by

$$\varphi(t) = \begin{cases} \exp\left(-\frac{1}{1-(2t-1)^2}\right) & \text{for } 0 < t < 1 \\ 0 & \text{else.} \end{cases} \tag{189}$$

It can be checked explicitly that $m(Z) \overset{\mathcal{D}}{=} Z$. Now we consider a VAE model with encoder and decoder each having 3 hidden residual layers with 64 neurons. We train encoder and decoder in a supervised way on pairs $(m(Z), Z)$ and $(Z, m(Z))$ (where Z follows a standard normal distribution) using the mean squared error until the mean error is smaller than 0.001. We define the map implemented by the decoder as f. We generate data $f(Z)$ where Z is Gaussian and then train a VAE on samples $f(Z)$ and we initialize the VAE with the learned f and f^{-1} and give a large negative bias of -5 to the log-variance layer to make the encoder close to deterministic.

I. Experimental Illustration of Convergence Rate for perturbed linear ICA

To illustrate the results in Section 5, we provide a small scale experimental illustration. We consider data generated by $X = AS + \eta h(S)$ where S_i follow independent Laplace distributions and all entries of A follow a centered normal distribution with variance d^{-1} and $h_i(S) \propto S_i^3 - \beta_i S_i - \alpha_i$ where α_i, β_i are chosen such that $\mathbb{E}(h_i(S)S_j) = 0$, $E(h_i(S)) = 0$ and the proportionality is chosen such that $\mathbb{E}(h_i(S)^2) = 1$. Then we run the Fast-ICA algorithm (Hyvarinen, 1999) on $\hat{\Sigma}_X^{-\frac{1}{2}} X$ with $n = 10^6$ samples for $\eta \in [10^{-3}, 1]$ We initialize $w_{\text{init}} = \bar{w}_i$ (recall that $\bar{w}_i$ denote the rows of $A^{-1}(AA^\top)^{\frac{1}{2}}$, i.e., the true unmixing of the whitened linear part A) to avoid problems with spurious minima and matching of minima which are already present for linear ICA. While $\bar{w}_i$ cannot be inferred from X, this allows us to focus here on the essential differences of our setting to linear ICA. Recall that $\tilde{w}_i$ was defined in (16) and denotes the predicted next order expression of the recovered source. In Figure 2 we plot the distance of the output w_i of the fast-ICA algorithm to $\bar{w}_i$ and $\tilde{w}_i$ and evaluate the MCC of the recovered sources $\hat{S}$ and the true sources S as a function of η, the strength of the perturbation.

We find that the dependence on η roughly matches our theoretical results. Note that for $\eta \gtrsim .1$ the error is already of order 1 and thus independent of η while for $\eta \lesssim .01$ the distance $|w_i - \tilde{w}_i|$ no longer decreases due to final sample effects (note that roughly $|w_i - \tilde{w}_i| = O(n^{-\frac{1}{2}})$ for such η). We also refer to (Horan et al., 2021) where small scale real-world image data was investigated and the independent components recovered. Their mixing function is only approximately locally isometric, and therefore our results give a theoretical justification for their observations.

J. A Class of approximately isometric random Functions

In this section we illustrate how approximate local isometries naturally appear when considering random high dimensional functions. Note that a slightly more general class of random functions was considered in Reizinger et al. (2023). However, their work does not quantify approximate local isometries. We consider random functions $f : \mathbb{R}^d \to \mathbb{R}^D$ where we are mostly interested in the case $D \gg d$. Assume that each coordinate f_i is given by a random draw from a smooth Gaussian

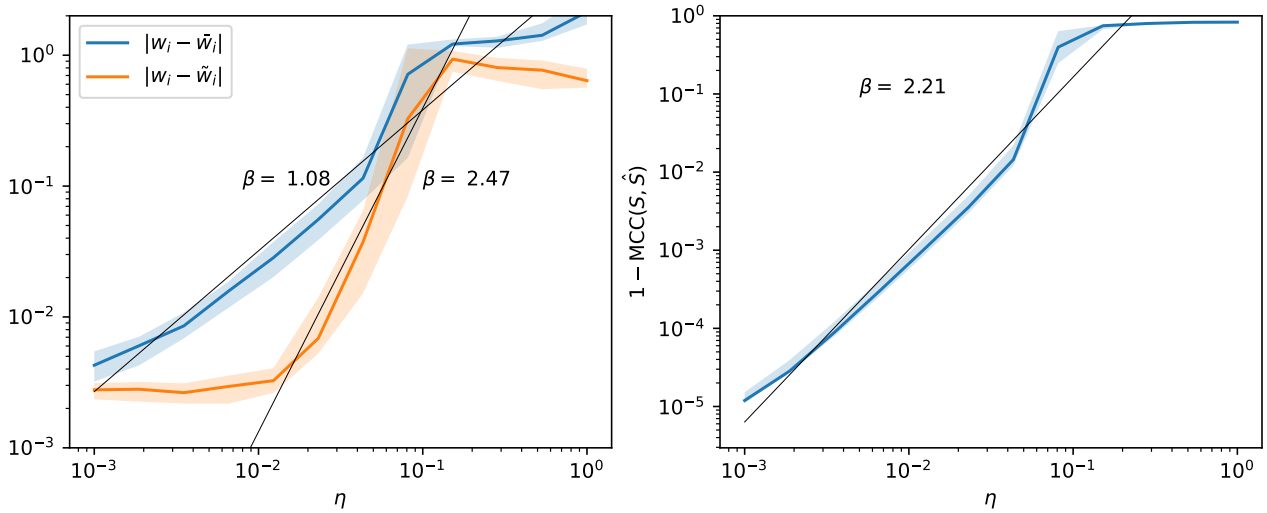

Figure 2. (Left) Distance of recovered linear unmixing w from $\bar{w}$ (see (12)) and $\tilde{w}$ (see (16)) as a function of η. Plotted is the median over 5 runs with $d = 5$ (i.e., 25 recovered components). The shaded area shows the range from the 32% to the 68% quantile. (Right) Difference of the Mean MCC over 5 runs of $\hat{S} = W\Sigma_X^{-\frac{1}{2}}X$ from perfect recovery (MCC $= 1$). Regression lines are obtained by linear regression in log-log space for $0.01 \leq \eta \leq 0.1$ and β values indicate their slope.

process, e.g., a mean zero process whose covariance is given by a Matérn kernel. Assume that the kernel is isotropic $K(x, y) = K(|x - y|)$ with $K'(0) = 0$ and $K''(0) = -1$ (can be obtained by scaling). Then the following holds for $k \neq l$

$$\begin{aligned}\mathbb{E}(\partial_k f_i(x)\partial_l f_i(x)) &= \lim_{h\to 0} h^{-2}\mathbb{E}((f_i(s + he_k) - f_i(s))(f_i(s + he_l) - f_i(s))) \\ &= \lim_{h\to 0} h^{-2}\mathbb{E}((f_i(s + he_k) - f_i(s))(f_i(s + he_l) - f_i(s))) \\ &= \lim_{h\to 0} h^{-2}(K(\sqrt{2}h) - 2K(h) + K(0)).\end{aligned} \tag{190}$$

Here, we used in the last step that $|(s + he_k) - (s + he_l)| = h|e_k - e_l| = \sqrt{2}h$. We then find using Taylor expansion of K (recall $K'(0) = 0$) that

$$\mathbb{E}(\partial_k f_i(s)\partial_l f_i(s)) = \lim_{h\to 0} h^{-2}\left((K(0) + h^2K''(0) + o(h^2)) - 2\left(K(0) - \frac{h^2}{2}K''(0)\right) + o(h^2)) + K(0)\right) = 0. \tag{191}$$

We also note that

$$\begin{aligned}\mathbb{E}(\partial_k f_i(s)\partial_k f_i(s)) &= \lim_{h\to 0} h^{-2}\mathbb{E}((f_i(s + he_k) - f_i(s))(f_i(s + he_k) - f_i(s))) \\ &= \lim_{h\to 0} h^{-2}\mathbb{E}(2K(0) - 2K(h)) = -K''(0) = 1.\end{aligned} \tag{192}$$

We conclude that the matrix $Df(s)$ has independent Gaussian entries with variance 1. To measure how isometric f is, we need to investigate the singular values of Df. This has been investigated extensively in the field of random matrix theory, e.g., the random matrices appearing here are a special case of the Wishart distribution. Indeed, there are explicit expressions for the joined eigenvalue density of $D^{-1}X^\top X$ where X_{ij} follow independent standard normal distributions and fine information on the spectrum is available even when $d, D \to \infty$ jointly with d/D fixed. Here we are only interested in the asymptotic scaling as $D \to \infty$, which is much simpler to obtain, and we now sketch this to provide some intuition. We denote $A = \frac{1}{D}Df(s)^\top Df(s)$, and we introduce $\tilde{f} = \sqrt{D}^{-1}f$. Using the results above, we find

$$\mathbb{E}A_{kl} = \mathbb{E}(D\tilde{f}^\top(s)D\tilde{f}(s))_{kl} = \frac{1}{D}\sum_m \mathbb{E}(Df(s))_{mk}(Df(s))_{ml} = \frac{1}{D}\sum_m \mathbb{E}\partial_k f_m(s)\partial_l f_m(s) = \delta_{kl}. \tag{193}$$

In other words $\mathbb{E}A = \mathrm{Id}_d$. By the law of large numbers we infer that

$$\lim_{D\to\infty} D\tilde{f}^\top(s)D\tilde{f}(s) = \mathrm{Id}_d. \tag{194}$$

In other words, we conclude that as $D \to \infty$ the sequence $\tilde{f}$ becomes increasingly isometric. This can also be quantified. Note that for $k \neq l$

$$\mathbb{E}(\partial_k f_i(s)\partial_l f_i(s))^2 = \mathbb{E}(\partial_k f_i(s))^2\mathbb{E}(\partial_l f_i(s))^2 = 1. \tag{195}$$

Here we used that uncorrelated Gaussian variables are independent. Similarly, we obtain for $k = l$

$$\mathbb{E}(\partial_k f_i(s)\partial_k f_i(s))^2 - \left(\mathbb{E}(\partial_k f_i(s)\partial_k f_i(s))\right)^2 = \mathbb{E}(\partial_k f_i(s))^4 - 1 = 3 - 1 = 2. \tag{196}$$

Linearity of the variance implies that

$$\mathrm{Var}A_{kl} \leq \frac{D}{D^2}(1+\delta_{kl}) = \frac{1}{D}(1+\delta_{kl}). \tag{197}$$

We conclude that

$$\mathbb{E}|A - \mathbb{E}(A)|_F^2 = \frac{d^2+d}{D} \tag{198}$$

Let $\sigma_i \geq 0$ be the singular values of $D\tilde{f}(s)$. Then we bound $|\sigma_i - 1| \leq |\sigma_i^2 - 1|$ for $\sigma_i > 0$ and find

$$\sum_i |\sigma_i - 1|^2 \leq \sum_i |\sigma_i^2 - 1|^2 = |A - \mathbb{E}(A)|_F^2 \tag{199}$$

Using (56) we conclude that

$$\mathbb{E}\,\mathrm{dist}^2(D\tilde{f}(s), \mathrm{SO}(d, T_{f(s)}M)) = \mathbb{E}\sum_i (\sigma_i - 1)^2 \leq \mathbb{E}|A - \mathbb{E}(A)|^2 = \frac{d^2+d}{D}. \tag{200}$$

The argument for the second term in (61) is similar but a bit more involved because we need to bound the inverse. Suitable bounds for A^{-1} can be found, e.g., in Theorem 2.4.14 in (Kollo & von Rosen, 2006). They then imply that

$$\mathbb{E}\,\mathrm{dist}^2(D\tilde{f}^{-1}(x), \mathrm{SO}(T_xM, d)) \leq \frac{C}{D}. \tag{201}$$

We then conclude that

$$\mathbb{E}\Theta_2(\tilde{f}, \Omega) \leq \sqrt{\mathbb{E}\Theta_2^2(\tilde{f}, \Omega)} \leq \frac{C}{\sqrt{D}} \tag{202}$$

and therefore the functions $\tilde{f}$ approach local isometries as $D \to \infty$ in a quantitative way with the expected rate. This result can be generalized to $p > 2$ using stronger concentration bounds for sub-Gaussian variables.