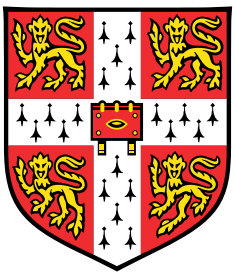
Abstractions, Representations and Latent Spaces



Paul Kishan Rubenstein

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Jesus College

November 2019

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Paul Kishan Rubenstein
November 2019

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Contents

List of Figures

List of Tables

Nomenclature

Roman Symbols

F	complex function

Greek Symbols

γ	a simply closed curve on a complex plane
ι	unit imaginary number $\sqrt{-1}$
π	$\simeq 3.14\ldots$

Superscripts

j	superscript index

Subscripts

0	subscript index
crit	Critical state

Other Symbols

$\oint_\gamma$	integration around a curve γ

Acronyms / Abbreviations

ALU	Arithmetic Logic Unit
BEM	Boundary Element Method
CD	Contact Dynamics
CFD	Computational Fluid Dynamics
CIF	Cauchy's Integral Formula
CK	Carman - Kozeny

DEM Discrete Element Method

DKT Draft Kiss Tumble

DNS Direct Numerical Simulation

EFG Element-Free Galerkin

FEM Finite Element Method

FLOP Floating Point Operations

FPU Floating Point Unit

FVM Finite Volume Method

GPU Graphics Processing Unit

LBM Lattice Boltzmann Method

LES Large Eddy Simulation

MPM Material Point Method

MRT Multi-Relaxation Time

PCI Peripheral Component Interconnect

PFEM Particle Finite Element Method

PIC Particle-in-cell

PPC Particles per cell

RVE Representative Elemental Volume

SH Savage Hutter

SM Streaming Multiprocessors

USF Update Stress First

USL Update Stress Last

Chapter 1

Introduction

1.1 Non-technical introduction

Non-technical introduction that should be accessible to people who don't know about machine learning, e.g. my parents.

- Describe machine learning as way to program computers.
- When humans look at a picture, we don't see pixels. We immediately see higher level concepts.
- An important topic in ML, and the subject of this thesis, is, roughly speaking, how machines can learn high level concepts.
- The example of images is easy to grasp because we are familiar with the idea of objects like cats and dogs. In fact, the difficult thing to understand is that images (on a screen) are fundamentally an array of numbers.
- Another example: audio. If I showed you a picture of a wave form, you wouldn't understand what it is. But if I play it to you, you'd be able to decompose the continuous stream into different parts (voice, drums, ...)
- But this ability to understand higher level concepts is not something that machines have, where inputs are just arrays of numbers (we also wouldn't have this with a printed list of numbers).
- One of the key features of human perception is the ability to understand the world at different scales of detail. For instance, (image of car) is at its simplest just a car. But a car consists of doors, windows, a wind shield, wheels. Each of these components can be more closely inspected: each wheel has the metal central part and rubber tyres, and we know that out of view there is a complicated steering mechanism connects the

wheels to the rest of the car. If we inspected the tyres closely we might have interesting things to say about the tread, and so on. Similarly, an album of music consists of songs that are related, each song consists of chorus and verse, within each of these there is a progression of chords, and so on. Chapter 3 presents the first major topic of my PhD, which considers how to mathematically describe the fact that there is no one objective level at which we understand any system; rather, we are aware that any understanding exists at some particular scale, and that depending on what we are doing or trying to achieve, thinking in more or less detailed ways may be appropriate.

- Another feature of human perception is our ability to synthesise together different streams of perceptual information into one conscious experience. For example, each of our eyes sees a 2D image. Yet we perceive the world in 3D because our brains automatically merge these two distinct streams together. (This is chapter 4, ICA stuff)

- When we look at a red car, we are able to understand that the 'redness' and the 'carness' are two independent features: same same car in green is fundamentally still the same kind of car even though the colour is different, while a red jumper shares little in common with the car, despite having the same colour. Moreover, if I were to present you with a picture of a red car and a green jumper, you could probably imagine what the car would look like in the jumper's shade of green, and also what the jumper would look like in the car's shade of red. Roughly speaking, this is the topic of Chapter 5, which considers a family of methods known as *Wasserstein Autoencoders*.

- The topic of Chapter 6 is more difficult to explain by analogy to common human experience, since it is a more focussed and technical contribution to the field. The use of Wasserstein Autoencoders as in Chapter 5 requires solving a particular mathematical problem. Usually when this problem is encountered, it is very difficult to solve. In Chapter 6 we study this problem in the specific case of Wasserstein Autoencoders in great detail, and show that in this case it is actually not so hard to solve.

1.2 Technical introduction

This is the 'proper' introduction.

- A human looking at an image understands its content not in terms of the pixels that are directly observed, but at higher conceptual levels such as that objects exist and relate to one another. This understanding can fluidly shift between multiple scales, so that most objects can be decomposed into smaller objects in a hierarchical fashion. Different sensory streams can be merged into a single richer conscious experience, so that we perceive the world in three dimensions despite each of our eyes seeing only in

two dimensions. The fact that these happen is a consequence both of evolution as well as a life-time of experience.

- Machine learning models, in contrast, generally have neither of these from which to benefit. This is a thesis about how high-level concepts and representations can nonetheless be modelled and learned. It examines three different ways in which representations occur, though there are others that are not treated here.

Causality:

- The first of these comes from the field of *causal inference*, the goal of which is to learn causal relations between random variables from either observational or experimental data. The asking of causal questions is ubiquitous in the social and natural sciences. Does smoking cause cancer? Does cutting corporation taxes cause economic growth? Will a sugar tax reduce the prevalence of obesity?

- It is troubling, however, that scientists often seek to discover causal relations that are ill-defined in the real world. For instance, do blood cholesterol levels causally influence the risk of heart disease?

- For a long time, researchers investigated this question to find contradictory conclusions. Some found that raising cholesterol levels caused increased risk of heart disease, while others found the opposite. The resolution of these conflicting results came with the realisation that there are two types of blood cholesterol with opposite effects on heart disease risk. Thus, raising one type protects against heart disease, while raising the other raises its risk, yet both of these interventions would be registered as an increase to blood cholesterol levels.

- Even when the variables under investigation are well-defined, in many cases a non-trivial set of assumptions is implicitly made. Consider the link between smoking and prevalence of cancer. In reality, the human body is a complex time-evolving system consisting of numerous individual cells. The effect of regular smoking is the accumulation of small perturbations to this complex system which collectively lead to increased risk that, at some point in time, one or more of the many cells malfunction and a cancerous growth appears.

- Thus, although it is true to say that "smoking causes cancer", hiding behind this simple statement is a very complicated causal relationship at the level of molecules and cells.

- The first contribution of this thesis is the formalisation of a theory of *causal abstraction* which builds on the mathematical language for modelling causal structure known as Structural Equation Models (SEMs), providing an understanding of when it is legitimate

to model causal relationships in the world at a coarser level of detail than the 'true' level at which causal relations hold. The main outcome of this research was the paper *Causal Consistency of Structural Equation Models* which is adapted and presented in Chapter 3. A follow-up paper, *From Deterministic ODEs to Dynamic Structural Causal Models*, is not presented in this thesis.

Non-linear ICA:

- The second type of representation comes from *Independent Component Analysis* (ICA).
- The *cocktail party problem* is often used to describe this problem: at a party where many people are speaking, listening to the voice of a single person requires the separation of many mixed audio signals.
- In the classical ICA problem setting, independent sources $\mathbf{s}$ are mixed through an unknown function f giving rise to a vector of observations $\mathbf{x} = f(\mathbf{s})$. Given the observations $\mathbf{x}$, the goal is to learn a function g that inverts f up to possibly tolerable ambiguities, thus approximately recovering the original independent sources $\mathbf{s}$.
- The resulting $g(\mathbf{x})$ is a representation of the observed data that should have more appealing properties compared to the raw observations themselves.
- In the usual single view setting, very strong assumptions need to be made on f and $\mathbf{s}$ in order to prove that recovery is possible.
- The contribution of this thesis to the ICA literature is to consider a problem setting in which multiple different *views* of the same sources is given. For example, different functions f_1 and f_2 give rise to different observations $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ which can both be used to recover $\mathbf{s}$. In this multi-view setting, it is proven that recovery of $\mathbf{s}$ can be made under much weaker assumptions compared to the single-view setting. This significantly increases the applicability of ICA methods, since the multi-view setting arises naturally in many real scenarios where, for example, different data modalities (audio, vision) of the same system being observed may exist.

WAE:

- The third and final type of representation comes from *Wasserstein Autoencoders* (WAEs), a type of generative model.
- Autoencoders are a broad class of method that enable learning low-dimensional representations of high-dimensional raw data. An encoder e mapping from the data space $\mathcal{X}$ to the low dimensional representation space $\mathcal{Z}$ and generator or decoder g mapping the reverse direction are simultaneously learned by minimising a reconstruction loss

$L(x, g(e(x)))$. The result is that $e(x)$ is a compressed representation of the original data x.

- WAEs additionally specify a *prior* distribution over the latent space and impose a regularisation term that penalises deviation from this by the *aggregate posterior*, the distribution of encoded data, with respect to some divergence or distance on distributions.
- WAEs can thus be used as a generative model, since samples can be generated by sampling from the prior and passing them through the generator, as well as a way to get representations.
- This thesis contains two contributions to the study of WAEs. The first is the proposal to use *probabilistic* encoders that map a single datum to a distribution over latent codes. It is shown that this leads to improved generative modelling performance as well as better properties for the learned representations.
- The second is to analyse in detail the problem of estimating the deviation between the prior and aggregate posterior distributions for a family of divergences known as *f-divergences*.

Scrap:
Unify notation: x or $\mathbf{x}$ for data, z or $\mathbf{z}$ for latent space. Encoder e and generator g.

1.3 Outline and Contributions

Summarise structure of thesis and outline what are the key contributions of each chapter.

- Chapter 2 is literature review. Fill this in after making proper literature review plan.
- Chapter 3 is causality section. Contribution is introducing theory of causal abstraction.
- Chapter 4 is ICA. Contribution is providing first identifiability results in multi-view setting.
- Chapter 5 is WAE. Contribution is introducing use of probabilistic encoders.
- Chapter 6 is WAE/learning theory. Contribution is showing that divergence estimation is *much* easier in AE setting than in agnostic setting.
- Chapter 7 is conclusion and discussion of where the field is going.

Summary of PhD work not included in this thesis

Chapter 2

Literature review

In this chapter we review the literature relevant to the thesis.

- This is a thesis about representations in a general sense, covering three different subfields of modern machine learning research. As such, in this literature review it is necessary to cover a lot of ground. The literature review begins with a broad overview of representation learning and the different ways that representations arise across machine learning.
- In the major part of the literature review, we cover the background of each of the three areas separately in detail.
- We give an introduction to causality and causal inference to set the stage of Chapter 3.
- ICA is covered next in order to introduce Chapter 4.
- Finally, we discuss generative modelling in order to set up Chapters 5 and 6.
- Each of the above named chapters additionally begins with a smaller literature review that is more focussed to the particular problem setting considered.

2.1 Representations in machine learning

- What do we mean by representations?
  - The terms *representation* and *representation learning* in machine learning are somewhat ill-defined, with their precise meaning depending on the particular context or niche of research under consideration. Common to these subtly different meanings is the existence of a data space $\mathcal{X}$ of raw data and a function $f : \mathcal{X} \to \mathcal{Z}$ mapping to another space $\mathcal{Z}$. In many cases, $\mathcal{X}$ may be high dimensional natural data such as images or audio, while $\mathcal{Z}$ is a lower dimensional vector space.

  - The function f is often referred to as a *feature extractor* and *representations* as *features*.

- Where do representations occur?

  - Across machine learning, many subfields involve learning representations. The reasons for doing so can be broadly divided into two main categories:
  - the unsupervised setting, where the representation learning is in some sense the explicit end goal, such as in unsupervised structure learning algorithms such as PCA, ICA, (interpretability? disentanglement? data analysis?) ... to a lesser extent, also autoencoder based generative models such as VAEs and WAEs where representation arise as a serendipitous by-product of another goal.
  - the supervised setting, where the representations learned are a means to solve some other goal, such as in supervised deep learning models, transfer learning / domain adaptation...

- Why do machine learners care about representations?

  - Possibly the main pragmatic reason is that many of the bread-and-butter problems in both academia and industry for which machine learning methods are applied are supervised in nature, in which case the performance of popular algorithms may be very sensitive to the featurisation of the data. For instance, the performance of simple methods such as linear and logistic regression can be significantly improved by first appropriately featurising the raw data. Indeed, the famous class of techniques known as *kernel methods* can be viewed as performing simple (e.g. linear) algorithms on top of rich features induced by the kernel of choice (Rasmussen and Williams, 2006; Schölkopf and Smola, 2001), while binary classification using neural networks can be viewed as performing logistic regression on top of learned features (Goodfellow et al., 2016).
  - An even more basic reason is that many successful machine learning algorithms typically require vectors as inputs. For sequential discrete data such as text or genome data, the raw data must first be processed into a form that can be digested by these methods. [cite BERT, word2vec and other NLP preprocessing techniques]
  - A different class of reasons are cases in which the goal is to *understand* the raw data, either by directly learning features, or by then using the features in some downstream statistical analysis. For instance, high dimensional data often lies on a low dimensional manifold of the data space. A basic algorithm for identifying the low dimensional structure can be found in PCA, which for any k finds the k-dimensional subspace in which the data varies most. Identification of this low-dimensional structure is useful in a variety of ways: for instance, as a first step

of *data exploration* to assist humans in understanding how to model the data, or indeed as a preprocessing / data cleaning technique to avoid overfitting to high-dimensional noise when fitting models. More advanced versions of this technique include manifold learning,... possibly connect to autoencoders.

  - Another example comes from applications of ICA, ranging from neuroimaging to astronomy and finance, where recorded data may be a superposition of many independent signals [cite papers from incomplete rosetta stone paper]. The problem of *blind source separation* is to separate out these signals, which may be of interest in and of themselves (e.g. detecting the changing luminescence of celestial bodies) or as an input to some other statistical analysis (e.g. correlating activity in some brain region with a controlled stimulus).

- Although this thesis is primarily about representations in the context of causality, ICA and generative modelling, this section would be incomplete without discussing recent advances in the application of representation learning to boosting the performance of deep supervised learning. We thus discuss this in the next section before moving on to background of the main topics.

2.2 Transfer learning and deep unsupervised representation learning / Recent advances in improving supervised learning with deep unsupervised representation learning

- In the past decade, rapid innovation and progress has occurred in deep learning. This is epitomised by progress in the Imagenet challenge, a supervised learning benchmark of natural images, for which the classification error has plummeted to better-than-human performance. Improvements have come from a variety of areas, including architectures [resnets, relu], optimizers [adam], hardware [gpus, tpus] and software [torch, pytorch, tensorflow].

- Performance on benchmark supervised learning tasks continue to improve, despite concerns of the community overfitting to these tasks [recht, ludwig schmidt cifar10 and imagenet paprs]. Nonetheless, the diminishing returns have caused many to switch their focus to reducing the sample complexity of solving supervised problems, i.e. the amount of labelled data required to achieve a given level of performance.

- Doing so is an academic problem with significant industrial implications. There are many supervised tasks which could in principle be solved, but for which the cost of collecting a sufficiently large dataset are not offset by the economic benefits of doing so. For such problems, reducing the amount of data required can entail economic feasibility.

- We have lots of unlabelled data, as well as labelled data from different datasets. These can be used 'for free'.
- Recent methods have been proposed for exploiting this, falling under different categories.
- A common and straightforward strategy for classification is to use a pre-trained classifier on some previous dataset (e.g. Imagenet) to provide embeddings which are then fine-tuned on the new dataset. The penultimate layer of the classifier is a common choice to provide the embeddings, on top of which a new linear layer or small MLP is added mapping to the new class outputs. The entire network can then be trained using the new dataset (this is fine-tuning).
- When it is not possible or not desirable to use existing datasets, other methods exist to exploit datasets for which only a small number of labels are given. Semi-supervised learning combines unsupervised learning on a large corpus of unlabelled data, with supervised learning on a small subset of the same dataset with labels. The additional unlabelled samples are used to model the marginal distribution of the input.... Examples include pre-training an autoencoder to get features (also, back in the day this was used as a way to do normal supervised learning [greedy layerwise unsupervised training, see bengio 2013 representation review paper])
- Recently there has been interest in *self-supervised* methods, which are fully unsupervised methods that generate a synthetic 'pretext' task. Simple examples of this include rotation, in which a random 0, 90, 180 or 270 degree rotation is applied and then predicted [cite rotation paper], or jigsaw, in which the relative location of patches of an image are applied [cite jigsaw paper]. More sophisticated methods have been proposed [CPC, deep infomax] based on the paradigm of mutual information maximisation, though recent work has demonstrated that theoretical understanding of these methods is not yet complete [our ICML paper]. [s4l] combines both elements of self-supervision and semi-supervision.
- In academic scenarios, no single agreed-upon metrics exist for evaluating the performance of representation learning algorithms. A common approach has been to evaluate performance on 1% and 10% of a dataset such as Imagenet to understand how the algorithms perform in the low data regime, though recently a more comprehensive benchmark has been released permitting the evaluation on a large number of synthetic and natural image datasets [VTAB].

Further discussion on these topics will be deferred to Chapter [conclusion and discussion]. In the remainder of this chapter we will discuss the backgrounds required for Chapters 3-6.

2.3 Causality

The goal of this section is to set the scene for Chapter 2. We introduce Structural Equation Models, the basic mathematical framework of the Pearl school of causality, and survey the literature on causal inference, the learning of causal structure from data.

2.3.1 Structural Equation Models

- SEMs
- Interventions
- Distributions implied by SEMs (observational and interventional), this is discussed in more detail in Chapter 3.
- Cyclicity. (Discussed in more detail in Chapter 3).

2.3.2 Causal Inference

- Conditional independence based methods
- SEM / residual methods

2.3.3 Causal Variables

- Causal inference algorithms suppose that we are given a tuple of meaningful variables over which causal structure is to be learned. But in reality, meaningful variables have to be inferred from low level sensory data.
- Discuss work that talks about where these variables come from. See intro of causal consistency paper as well as papers that cite this from Google scholar.

2.4 Independent Component Analysis

Tutorial: http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/

- Problem statement and impossibility result.

2.4.1 Results for classical single view setting

- Assumptions on distribution of s
- Assumptions on mixing function f
- Linear vs nonlinear.
- Intuition of proof techniques.

2.4.2 Recent advances

- Aapo's string of papers, including the one that our work built on.
- VAE paper with Dirk and Aapo.
- Discussion of new types of assumptions being explored.

2.5 Generative Modelling with Latent Variable Models

- Fundamental problem of generative modelling: given iid samples from some distribution Q_X, learn an distribution $P_X \approx Q_X$ from which new samples can be drawn.

2.5.1 Latent Variable Models

- A convenient way to specify distributions over high dimensional spaces is with a latent variable model which can express complex distributions by combining simpler parts. Idea to specify simple *prior* distribution P_Z over low dimensional *latent space* and a family of conditional distributions $P_{X|Z}$. For any fixed value $Z = z$, the distribution $P_{X|Z=x}$ may be simple, e.g. isotropic Gaussian, but my marginalising over the prior we can get arbitrarily complex distributions $P_X = \int P_{X|Z=z} P_Z(z) dz$. Give precise mathematical definition

2.5.2 Divergences

- The goal that P_X should be approximately equal to Q_X is made precise by demanding that some divergence $D(P_X \| Q_X)$ should be small. A divergence is a measure of dissimilarity on distributions. Give precise mathematical definition. Note that this is weaker than a *distance*, which has a triangle inequality.
- Examples of divergences include Integral Probability Metrics and f-divergences.
- In practice, minimising such divergences between high dimensional distributions is often intractable. Thus, two broad families of methods exist to get tractable surrogate losses. General strategy is to get variational lower bound on loss (GANs), tractable upper bound (VAE), or other approximate loss (WAE).

2.5.3 Evaluating Generative Models

- FID scores

2.5.4 GANs

- Describe GANs. We don't use them in this thesis, but they have been so important throughout generative modelling that any review of this area is required to cover them. In their vanilla form, GANs don't have encoders, though some extensions have included encoders [cite encoder GAN project and references therein, e.g. BiGAN].

2.5.5 VAEs

- VAEs are another popular method. Describe them. They do have encoders, though the name autoencoder is perhaps somewhat of a misnomer [cite blogpost].

2.5.6 WAEs

- WAEs are a recently introduced method which will be examined in more detail in Chapters 5 and 6. Describe that they start from the optimal transport perspective, which has some advantages over KL in VAE. Idea is that OT cost factors into latent space, and so problem reduces to reconstruction + distribution matching.

Chapter 3

Causal Abstractions

This chapter is based on the paper *Causal Consistency of Structural Equation Models* published at UAI 2017.

3.1 Introduction

Physical systems or processes in the real world are complex and can be understood at various levels of detail. For instance, a gas in a volume consists of a large number of molecules. But instead of modelling the motions of each particle individually (micro-level), we may choose to consider macroscopic properties of their motions such as temperature and pressure. Our decision to use such macroscopic properties is first necessitated by practical considerations. Indeed, for all but extremely simple cases, making a measurement of all the individual molecules is practically impossible and our resources insufficient for modelling the $\sim 10^{22}$ particles present per litre of ideal gas. Furthermore, the decision for a macroscopic description level is also a pragmatic one: if we only wish to reason about temperature and pressure, a model of 10^{22} particles is ill-suited.

Statistical physics explains how higher-level concepts such as temperature and pressure arise as statistical properties of a system of a large number of particles, justifying the use of a macro-level model as a useful transformation of the micro-level model (Balian, 1992). However, in many cases aggregate or indirect measurements of a complex system form the basis of a macroscopic description of the system, with little theory to explain whether this is justified or how the micro- and macro-descriptions stand in relation to each other.

Due to deliberate modelling choice or the limited ability to observe a system, differing levels of model descriptions are ubiquitous and occur, amongst possibly others, in the following three settings:

(a) Models with large numbers of variables versus models in which the 'irrelevant' or unobservable variables have been marginalised out (Bongers et al., 2016); e. g. modelling

blood cholesterol levels and risk of heart disease while ignoring other blood chemicals or external factors such as stress.

(b) Micro-level models versus macro-level models in which the macro-variables are aggregate features of the micro-variables (Chalupka et al., 2015, 2016; Hoel et al., 2013; Iwasaki and Simon, 1994; Simon and Ando, 1961); e. g. instead of modelling the brain as consisting of 100 billion neurons it can be modelled as averaged neuronal activity in distinct functional brain regions.

(c) Dynamical time series models versus models of their stationary behaviour (Dash and Druzdzel, 2001; Fisher, 1970; Iwasaki and Simon, 1994; Lacerda et al., 2008; Mooij and Heskes, 2013; Mooij et al., 2013); e. g. modelling only the final ratios of reactants and products of a time evolving chemical reaction.

In the context of causal modelling, such differing model levels should be consistent with one another in the sense that they agree in their predictions of the effects of interventions. The particular causal models we focus on in this paper are Structural Equation Models (SEMs, Section 3.2, Section 3.3) (Pearl, 2009; Spirtes et al., 2000).

In Section 3.4, we introduce the notion of an exact transformation between two SEMs, providing us with a general framework to evaluate when two models can be thought of as causal descriptions of the same system. An important novel idea of this paper is to explicitly make use of a natural ordering on the set of interventions. On a high level, if an SEM can be viewed as an exact transformation of another SEM, we are provided with an explicit correspondence between the two models in such a way that causal reasoning on both levels is consistent. We discuss this notion of consistency in detail in Sections 3.4.4 and 3.4.5.

In Section 3.5 we apply this mathematical framework and prove the exactness of transformations belonging to each of the three categories listed above, with practical implications for the following questions in causal modelling: When can we model only a subsystem of a more complex system? When does a micro-level system admit a causal description in terms of macro-level features? How do cyclic SEMs arise? The fact that these distinct problems can all be considered using the language of transformations between SEMs demonstrates the generality of our approach. We close in Section 3.6 with a discussion.

3.1.1 A historical motivation: Cholesterol and Heart Disease

In the following we give an example of the problems that can arise when there exists no consistent correspondence between two causal models, i. e. neither model can be viewed as an exact transformation of the other. This example falls into category (b) of the differing model levels listed above and was used by Spirtes and Scheines (2004) to illustrate problems in the causal modelling process.

Historically, the level of total cholesterol in the blood (TC) was thought to be an important variable in determining risk of heart disease (HD). To investigate this, different experiments

(a) (b)

Figure 3.1 As illustrated by (a), the current consensus is that LDL (resp. HDL) has a negative (resp. positive) effect on heart disease (HD). Considering TC = LDL + HDL to be a causal variable as in (b) leads to problems: two diets promoting raised LDL levels and raised HDL levels have the same effect on TC but opposite effects on heart disease. Hence different studies come to contradictory conclusions about the effect of TC on heart disease.

were carried out in which patients were assigned to different diets in order to raise or lower TC. Conflicting evidence was found by different experiments: some found that higher TC had the effect of lowering HD, while others found the opposite (cf. Figure 3.1b) (Steinberg, 2011; Truswell, 2010).

From our point of view, this problem (seemingly conflicting studies) arose from trying to perform an 'invalid' transformation of the 'true' underlying model (cf. Figure 3.1a). According to the American Heart Association, the current scientific consensus is that the two types of blood cholesterol, low-density lipoprotein (LDL) and high-density lipoprotein (HDL), have a negative and positive effect on HD respectively. Assigning diets that raise LDL or HDL both raise TC but have different effects on HD. It is therefore not possible to transform the model in Figure 3.1a into the model in Figure 3.1b without leading to conflict: in order to reason about the causes of HD we need to consider the variables LDL and HDL separately.

3.2 Structural Equation Models

SEMs are a widely used framework in causal modelling, with applications in neuroscience, economics and the social sciences (Bollen, 2014; Pearl, 2009). In this section we introduce them as an abstract mathematical object; in Section 3.3 we describe their use as a causal modelling tool. Readers already familiar with SEMs should note that our definition is more general and deviates from the standard definition of SEMs in the following ways: we do not require that all possible perfect interventions be modelled; we do not assume independence of exogenous variables;[1] and we do not require acyclicity.

Definition 1 (Structural Equation Model (SEM))**.** *Let $\mathbb{I}_X$ be an index set. An SEM $\mathcal{M}_X$ over variables $X = (X_i : i \in \mathbb{I}_X)$ taking value in $\mathcal{X}$ is a triple $(\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_E)$ where*

- *$\mathcal{S}_X$ is a set of structural equations, i. e. it is a set of equations $X_i = f_i(X, E_i)$ for $i \in \mathbb{I}_X$;*

[1] Exogenous variables are also referred to as *noise variables* in the literature. Our relaxation of the assumption of independent exogenous variables means our models may be considered a type of semi-Markovian causal model.

- *$(\mathcal{I}_X, \leq_X)$ is a subset of all perfect interventions equipped with a natural partial ordering (see below), i. e. it is an index set where each index corresponds to a particular perfect intervention on some of the X variables;*
- *$\mathbb{P}_E$ is a distribution over the exogenous variables $E = (E_i : i \in \mathbb{I}_X)$;*
- *with $\mathbb{P}_E$-probability one, under any intervention $i \in \mathcal{I}_X$ there is a unique solution $x \in \mathcal{X}$ to the intervened structural equations. This ensures that for any intervention $i \in \mathcal{I}_X$, $\mathcal{M}_X$ induces a well-defined distribution over $\mathcal{X}$.*[2]

In an SEM, each X_i is a function of the X-variables and the exogenous variable E_i. In this mathematical model, a perfect intervention on a single variable $\text{do}(X_i = x_i)$ is realised by replacing the structural equation for variable X_i in $\mathcal{S}_X$ with $X_i = x_i$. Perfect interventions on multiple variables, e.g. $\text{do}(X_i = x_i, X_j = x_j)$, are similarly realised by replacing the structural equations for each variable individually. Elements of $\mathcal{I}_X$ correspond to perfectly intervening on a subset of the X variables, setting them to some particular combination of values.

$\mathcal{I}_X$ has a natural partial ordering in which, for interventions $i, j \in \mathcal{I}_X$, $i \leq_X j$ if and only if i intervenes on a subset of the variables that j intervenes on and sets them equal to the same values as j. For example, $\text{do}(X_i = x_i) \leq_X \text{do}(X_i = x_i, X_j = x_j)$.[3] The observation that this structure is important is a contribution of this paper. We make crucial use of it in the next section.

The purpose of the following example is to illustrate how SEMs are written in our notation and to provide and example of a restricted set of interventions $\mathcal{I}_X$.

Example 2. *Consider the following SEM defined over the variables $\{B_1, B_2, L\}$*

$$\begin{aligned}
\mathcal{S}_X &= \{B_1 = E_1,\ B_2 = E_2,\ L = \text{OR}(B_1, B_2, E_3)\} \\
\mathcal{I}_X &= \{\varnothing,\ \text{do}(B_1 = 0),\ \text{do}(B_2 = 0), \\
&\qquad \text{do}(B_1 = 0, B_2 = 0)\}, \\
\{E_1, E_2, E_3\} &\overset{iid}{\sim} \text{Bernoulli}(0.5)
\end{aligned}$$

where by the element $\varnothing \in \mathcal{I}$ we denote the null-intervention corresponding to the unintervened SEM.

[2] That is, with probability one over the exogenous variables E, for each draw $E = e$ there exists a unique value $x \in \mathcal{X}$ such that e and x satisfy the intervened structural equations. The distribution of E in conjunction with $\mathcal{S}_X$ then implies a distribution over $\mathcal{X}$ for each intervention $i \in \mathcal{I}_X$ via these unique solutions. If the SEM is acyclic, this is always satisfied; we impose this condition because we also consider *cyclic* SEMs Bongers et al. (2016).

[3] Informally, this means that j can be performed after i without having to change or undo any of the changes to the structural equations made by i. Not all pairs of elements must be comparable: for instance, if $i = \text{do}(X_1 = x_1)$ and $j = \text{do}(X_2 = x_2)$, then neither $i \leq_X j$ nor $j \leq_X i$.

3.3 SEMs for Causal Modelling

In addition to being abstract mathematical objects, SEMs are used in causal modelling to describe distributions of variables and how they change under interventions (Pearl, 2009). The do-interventions as abstract manipulations of SEMs are understood as corresponding to actual (or potentially only hypothetical) physical implementations in the real world, i. e. the model is 'rooted in reality'. For instance, if a binary variable B_1 in an SEM reflects whether a light bulb is emitting light, then $\mathrm{do}(B_1 = 0)$ could be achieved by flipping the light switch or by removing the light bulb.

The SEM in Example 2 could be thought of as a simple causal model of two light bulbs B_1 and B_2 and the presence of light L in a room with a window. Suppose that we have no access to the light switch and there are no curtains in the room but that we can intervene by removing the light bulbs. We can model this restricted set of interventions by $\mathcal{I}_X$, i. e. the do-intervention on the SEM side $\mathrm{do}(B_1 = 0)$ corresponds to removing the light bulb B_1.

The partial ordering of $\mathcal{I}_X$ corresponds to the ability to compose physical implementations of interventions. The fact that we can first remove light bulb B_1 ($\mathrm{do}(B_1 = 0)$) and then afterwards remove light bulb B_2 (resulting in the combined intervention $\mathrm{do}(B_1 = 0, B_2 = 0)$) is reflected in the partial ordering via the relation $\mathrm{do}(B_1 = 0) \leq_X \mathrm{do}(B_1 = 0, B_2 = 0)$.

3.4 Transformations between SEMs

We now work towards our definition of an exact transformation between SEMs. Our core idea is to analyse the correspondence between different levels of modelling by considering one model to be a transformation of the other. We discuss in Section 3.4.4 how causal reasoning in two SEMs relate when one SEM can be viewed as an exact transformation of the other and in Section 3.4.5 we illustrate what can go wrong when this is not the case.

3.4.1 Distributions implied by an SEM

Usually, a statistical model implies a single joint distribution over all variables once its parameters are fixed. SEMs are different in that, once the parameters are fixed, an SEM implies a family of joint distributions over the random variables, one for each intervention. That is, for each intervention $i \in \mathcal{I}_X$, the SEM $\mathcal{M}_X$ defines a distribution over $\mathcal{X}$ which we denote by $\mathbb{P}_X^{\mathrm{do}(i)}$. Throughout, we will denote the null-intervention corresponding to the unintervened setting by $\varnothing \in \mathcal{I}_X$. We can write the poset of all distributions implied by the SEM $\mathcal{M}_X$ as

$$\mathcal{P}_X := \left(\left\{\mathbb{P}_X^{\mathrm{do}(i)} \;:\; i \in \mathcal{I}_X\right\}, \leq_X\right)$$

where $\leq_X$ is the partial ordering inherited from $\mathcal{I}_X$, i. e. $\mathbb{P}_X^{\text{do}(i)} \leq_X \mathbb{P}_X^{\text{do}(j)} \iff i \leq_X j$.[4]

Note that $\mathcal{P}_X$ contains all of the information in $\mathcal{M}_X$ about the different distributions implied by the SEM and, importantly, how they are related via the interventions.[5]

3.4.2 Transformations of random variables

Suppose we have a function $\tau : \mathcal{X} \to \mathcal{Y}$ which maps the variables of the SEM $\mathcal{M}_X$ to another space $\mathcal{Y}$. Observe that since X is a random variable, $\tau(X)$ is also a random variable. For any distribution $\mathbb{P}_X$ on $\mathcal{X}$ we thus obtain the distribution of the variable $\tau(X)$ on $\mathcal{Y}$ as $\mathbb{P}_{\tau(X)} = \tau\left(\mathbb{P}_X\right)$ via the push-forward measure.

In particular, for each intervention $i \in \mathcal{I}_X$ we can define the induced distribution $\mathbb{P}^i_{\tau(X)} = \tau\left(\mathbb{P}_X^{\text{do}(i)}\right)$. We can write the poset of distributions on $\mathcal{Y}$ that are induced by the original SEM $\mathcal{M}_X$ and the transformation τ as

$$\mathcal{P}_{\tau(X)} := \left(\left\{\mathbb{P}^i_{\tau(X)} \;:\; i \in \mathcal{I}_X\right\}, \leq_X\right)$$

where $\leq_X$ is the partial ordering inherited from $\mathcal{P}_X$ (and in turn from $\mathcal{I}_X$).

$\mathcal{P}_{\tau(X)}$ is just a structured collection of distributions over $\mathcal{Y}$, indexed by interventions $\mathcal{I}_X$ on the $\mathcal{X}$-level; importantly, the indices are *not* interventions on the $\mathcal{Y}$-level.

3.4.3 Exact Transformations between SEMs

Although $\mathcal{P}_{\tau(X)}$ is a poset of distributions over $\mathcal{Y}$, there does not necessarily exist an SEM $\mathcal{M}_Y$ over $\mathcal{Y}$ that implies it. For instance, if there is some intervention $i \in \mathcal{I}_X \setminus \{\varnothing\}$ such that none of the variables Y_i is constant under the distribution $\mathbb{P}^i_{\tau(X)}$, then $\mathbb{P}^i_{\tau(X)}$ could not possibly be expressed as arising from a do-intervention $j \in \mathcal{I}_Y \setminus \{\varnothing\}$ in any SEM over $\mathcal{Y}$.[6]

The case in which there *does* exist an SEM $\mathcal{M}_Y$ that implies $\mathcal{P}_{\tau(X)}$ is special, motivating our main definition.

Definition 3 (Exact Transformations between SEMs)**.** *Let $\mathcal{M}_X$ and $\mathcal{M}_Y$ be SEMs and $\tau : \mathcal{X} \to \mathcal{Y}$ be a function. We say $\mathcal{M}_Y$ is an* exact *τ-transformation of $\mathcal{M}_X$ if there exists a* surjective order-preserving *map $\omega : \mathcal{I}_X \to \mathcal{I}_Y$ such that*

$$\mathbb{P}^i_{\tau(X)} = \mathbb{P}_Y^{\text{do}(\omega(i))} \quad \forall i \in \mathcal{I}_X$$

[4] More formally, one would need to define $\mathcal{P}_X$ to be the poset of *tuples* $\left(i, \mathbb{P}_X^{\text{do}(i)}\right)$ to avoid problems in the case that $\mathbb{P}_X^{\text{do}(i)} = \mathbb{P}_X^{\text{do}(j)}$ for some $i \neq_X j$. Doing so would not require a change to Definition 3 or affect the further results of this paper. To avoid notational burden in our exposition, we omit this treatment.

[5] For example, the distribution over the variables X in the observational setting, $\mathbb{P}_X^{\varnothing}$, changes to $\mathbb{P}_X^{\text{do}(i)}$ if we implement the intervention $\text{do}(i)$, and the partial ordering contains all information about which interventions can be composed.

[6] This problem is elaborated upon in Eberhardt (2016).

where $\mathbb{P}^{i}_{\tau(X)}$ *is the distribution of the* $\mathcal{Y}$*-valued random variable* $\tau(X)$ *with* $X \sim \mathbb{P}_X^{\mathrm{do}(i)}$.

Order-preserving means that $i \leq_X j \implies \omega(i) \leq_Y \omega(j)$. It is important that the converse need not in general hold as this would imply that ω is injective,[7] and hence also bijective. This would constrain the ways in which $\mathcal{M}_Y$ can be 'simpler' than $\mathcal{M}_X$.[8] That ω is surjective ensures that for any do-intervention $j \in \mathcal{I}_Y$ on $\mathcal{M}_Y$ there is at least one corresponding intervention on the $\mathcal{M}_X$ level, namely an element of $\omega^{-1}(\{j\}) \subseteq \mathcal{I}_X$. The following two results follow immediately from the definition (cf. proofs in Appendix 3.7).

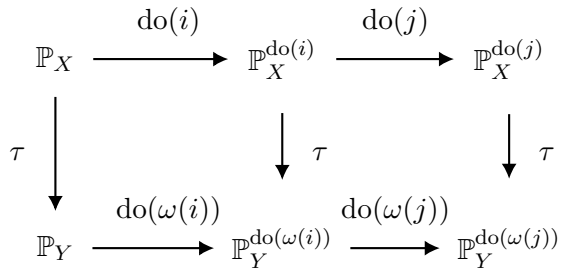
Lemma 4. *The identity mapping and permuting the labels of variables are both exact transformations.*

This is a good sanity check; it would be problematic if this were not the case and the labelling of our variables mattered. Similarly, compositions of exact transformations are also exact.

Lemma 5 (Transitivity of exact transformations)**.** *If* $\mathcal{M}_Z$ *is an exact* τ_{ZY}*-transformation of* $\mathcal{M}_Y$ *and* $\mathcal{M}_Y$ *is an exact* τ_{YX}*-transformation of* $\mathcal{M}_X$*, then* $\mathcal{M}_Z$ *is an exact* $(\tau_{ZY} \circ \tau_{YX})$*-transformation of* $\mathcal{M}_X$*.*

The following theorem is a consequence of the fact that ω is order-preserving. This is a mathematical formalisation of the sense in which an exact transformation preserves causal reasoning, which will be elaborated upon in the next subsection.

Theorem 6 (Causal consistency under exact transformations)**.** *Suppose that* $\mathcal{M}_Y$ *is an exact* τ*-transformation of* $\mathcal{M}_X$ *and* ω *is a corresponding surjective order-preserving mapping between interventions. Let* $i, j \in \mathcal{I}_X$ *be interventions such that* $i \leq_X j$*. Then the following diagram commutes:*



Proof. Let $i, j \in \mathcal{I}_X$ be interventions with $i \leq_X j$. The commutativity of the left square of the diagram follows immediately from the definition of an exact transformation. It remains to be shown that the right square of the diagram commutes. By definition we

[7] Since $\omega(i) = \omega(j) \iff (\omega(i) \leq_Y \omega(j)) \wedge (\omega(j) \leq_Y \omega(i))$, which, if the converse held, would imply that $(i \leq_X j) \wedge (j \leq_X i)$, which is equivalent to $i = j$.

[8] For instance, if it were necessary that ω be bijective, Theorems 9 and 11 would not hold.

have that $\tau\left(\mathbb{P}_X^{\mathrm{do}(i)}\right) = \mathbb{P}_Y^{\mathrm{do}(\omega(i))}$ and $\tau\left(\mathbb{P}_X^{\mathrm{do}(j)}\right) = \mathbb{P}_Y^{\mathrm{do}(\omega(j))}$. Thus, we only have to show that $\mathbb{P}_Y^{\mathrm{do}(\omega(i))} \leq_Y \mathbb{P}_Y^{\mathrm{do}(\omega(j))}$ as elements of $\mathcal{P}_Y$, i. e. that the arrow $\mathbb{P}_Y^{\mathrm{do}(\omega(i))} \xrightarrow{\mathrm{do}(\omega(j))} \mathbb{P}_Y^{\mathrm{do}(\omega(j))}$ exists. This follows from the order-preservingness of ω. □

3.4.4 Causal Interpretation of Exact Transformations

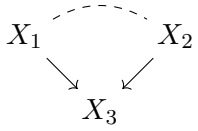
The notion of an exact transformation between SEMs was motivated by the desire to analyse the correspondence between two causal models describing the same system at different levels of detail. The purpose of this section is to show that if one SEM can be viewed as an exact transformation of the other, then both can sensibly be thought of as causal models of the same system. In the following, we assume that $\mathcal{M}_Y$ is an exact τ-transformation of $\mathcal{M}_X$ with ω the corresponding map between interventions.

Surjectivity of ω ensures that any intervention in $\mathcal{I}_Y$ can be viewed as an $\mathcal{M}_Y$-level representative of some intervention on the $\mathcal{M}_X$-level. Consequently, if do-interventions on the $\mathcal{M}_X$-level are in correspondence with physical implementations, then surjectivity of ω ensures that do-interventions on the $\mathcal{M}_Y$-level have at least one corresponding physical implementation, i. e. if $\mathcal{M}_X$ is 'rooted in reality', then so is $\mathcal{M}_Y$.

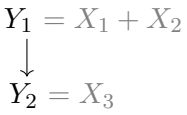
Commutativity of the left hand part of the diagram ensures that the effects of interventions are consistently modelled by $\mathcal{M}_X$ and $\mathcal{M}_Y$. Suppose we want to reason about the effects on the $\mathcal{M}_Y$-level caused by the intervention $j \in \mathcal{I}_Y$. For example, we may wish to reason about how the temperature and pressure of a volume of gaseous particles is affected by being heated. We could perform this reasoning by considering any corresponding $\mathcal{M}_X$-level intervention $i \in \omega^{-1}(\{j\})$ and considering the distribution this implies over $\mathcal{Y}$ via τ. In our example, this would correspond to considering how heating the volume of gas could be modelled by changing the motions of all the gaseous particles and then computing the temperature and pressure of the volume of particles. Commutativity of the left hand part of the diagram implies that $\mathcal{M}_X$ and $\mathcal{M}_Y$ are consistent in the sense that $\mathcal{M}_Y$ allows us to immediately reason about the effect of the intervention $j \in \mathcal{I}_Y$ while being equivalent to performing the steps above. That is, we can reason directly about temperature and pressure when heating a volume of gas without having to perform the intermediate steps that involve the microscopic description of the system.

Commutativity of the right hand side of the diagram ensures that once an intervention that fixes a subset of the variables has been performed, we can still consistently reason about the effects of further interventions on the remaining variables in $\mathcal{M}_X$ and $\mathcal{M}_Y$. Furthermore, it ensures that compositionality of do-interventions on the $\mathcal{M}_X$-level carries over to the $\mathcal{M}_Y$-level, i. e. if the intervention j on the $\mathcal{M}_X$-level can be performed additionally to the intervention i in $\mathcal{M}_X$—that is, $i \leq_X j$—, then the same is true of their representations in $\mathcal{M}_Y$.

If $\mathcal{M}_X$ and $\mathcal{M}_Y$ are models of the same system and it has been established that $\mathcal{M}_Y$ is an exact τ-transformation of $\mathcal{M}_X$ for some mapping τ, then the commutativity of the whole diagram in Theorem 6 ensures that they are causally consistent with one another in the sense



(a) SEM $\mathcal{M}_X$



(b) SEM $\mathcal{M}_Y$

Figure 3.2 Graphical illustration of parent-child relationships for the examples in Section 3.4.5. The micro-level model $\mathcal{M}_X$ depicted in (a) is to be transformed into the macro-level model $\mathcal{M}_Y$ depicted in (b) which is a coarser descriptions as in it only considers the sum of X_1 and X_2. In Section 3.4.5 we give examples of what can go wrong if the transformation is not exact.

described in the preceding paragraphs. If we wish to reason about the effects of interventions on the $\mathcal{Y}$-variables then it suffices to use the model $\mathcal{M}_Y$, rather than the (possibly more complex) model $\mathcal{M}_X$. In particular, this means that we can view the $\mathcal{Y}$-variables as causal entities, rather than only functions of underlying 'truly' causal entities. Only if this is the case, causal statements such as 'raising temperature increases pressure' or 'LDL causes heart disease' are meaningful.

3.4.5 What can go wrong when a transformation is not exact?

In the previous section we argued that our definition of exact transformations between SEMs is a sensible formalisation of causal consistency. In this section we will try to give the reader an intuition for why weakening the conditions of our definition would be problematic. In particular we focus on the requirement that ω be order-preserving, which we view as one of the core ideas of our paper.

The requirement that ω be surjective is, as discussed above, required so that all interventions on the $\mathcal{M}_Y$-level have a corresponding intervention on the $\mathcal{M}_X$-level. If we were to only require that ω be surjective (but not order-preserving), the observational distribution of $\mathcal{M}_X$ may be mapped to an interventional distribution of $\mathcal{M}_Y$, as illustrated by the following example (cf. Figure 3.2 for an illustration).

Example 7. *Consider the SEM* $\mathcal{M}_X = \{\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_E\}$ *over* $\mathcal{X} = \mathbb{R}^3$ *where*

$$\begin{aligned}\mathcal{S}_X &= \{X_1 = E_1,\ X_2 = E_2,\ X_3 = X_1 + X_2 + E_3\}\\ \mathcal{I}_X &= \{\varnothing,\ \mathrm{do}(X_2 = 0),\ \mathrm{do}(X_1 = 0,\, X_2 = 0)\},\\ &\quad E_1 \sim \mathbb{P}_{E_1},\ \ E_2 = -E_1,\ \ E_3 \sim \mathbb{P}_{E_3}\end{aligned}$$

where $\mathbb{P}_{E_1}$ *and* $\mathbb{P}_{E_3}$ *are arbitrary distributions. Let* $\tau : \mathcal{X} \to \mathcal{Y} = \mathbb{R}^2$ *be the mapping such that*

$$\tau\left(x_1, x_2, x_3\right) = \left(y_1, y_2\right) = \left(x_1 + x_2, x_3\right)$$

Let $\mathcal{M}_Y = \{\mathcal{S}_Y, \mathcal{I}_Y, \mathbb{P}_F\}$ *be an SEM over* $\mathcal{Y}$ *with*

$$\begin{aligned}\mathcal{S}_Y &= \{Y_1 = F_1,\ Y_2 = Y_1 + F_2\}\\ \mathcal{I}_Y &= \{\varnothing,\ \mathrm{do}(Y_1 = 0)\},\\ F_1 \sim \mathbb{P}_{E_1},\ F_2 &\sim \mathbb{P}_{E_3}\end{aligned}$$

Let $\omega : \mathcal{I}_X \to \mathcal{I}_Y$ *be defined by*

$$\omega : \begin{cases} \varnothing & \mapsto \mathrm{do}(Y_1 = 0)\\ \mathrm{do}(X_2 = 0) & \mapsto \varnothing\\ \mathrm{do}(X_1 = 0,\, X_2 = 0) & \mapsto \mathrm{do}(Y_1 = 0)\end{cases}$$

Then it is true that $\mathbb{P}^i_{\tau(X)} = \mathbb{P}_Y^{\mathrm{do}(\omega(i))}$ *for all* $i \in \mathcal{I}_X$*, while* ω *is not order-preserving and* $\omega(\varnothing) \neq \varnothing$*.*

If the SEMs in the above example were used to model the same system, it would be problematic that the observational setting of $\mathcal{M}_X$—a description of the system when not having physically performed any intervention—would correspond to an interventional setting in $\mathcal{M}_Y$, conversely suggesting that the system *had* been intervened upon.

To avoid the above conflict, we could demand in addition to surjectivity that ω map the null intervention of $\mathcal{M}_X$ to the null intervention of $\mathcal{M}_Y$. This additional assumption would ensure commutativity of the left-hand part of the diagram in Theorem 6. However, as the following example shows, this would not ensure that the right-hand part of the diagram commutes for all pairs of interventions $i \leq_X j$, since in this case the arrow from $\mathbb{P}_Y^{\mathrm{do}(\omega(i))}$ to $\mathbb{P}_Y^{\mathrm{do}(\omega(j))}$ may not exist.[9]

Example 8. *Let* $\mathcal{X}, \mathcal{Y}$ *and* τ *be as in Example 7. Consider the SEM* $\mathcal{M}_X = \{\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_E\}$ *where*

$$\begin{aligned}\mathcal{S}_X &= \{X_1 = E_1,\ X_2 = E_2,\ X_3 = X_1 + X_2 + E_3\}\\ \mathcal{I}_X &= \{\varnothing,\ \mathrm{do}(X_2 = 0),\ \mathrm{do}(X_1 = 0,\, X_2 = 0)\},\\ E_1 = 1,\ E_2 &\sim \mathbb{P}_{E_2},\ E_3 \sim \mathbb{P}_{E_3}\end{aligned}$$

[9] By definition of the poset $\mathcal{P}_Y$, this arrow exists if and only if $\omega(i) \leq_Y \omega(j)$.

where $\mathbb{P}_{E_2}$ *and* $\mathbb{P}_{E_3}$ *are arbitrary distributions. Let* $\mathcal{M}_Y = \{\mathcal{S}_Y, \mathcal{I}_Y, \mathbb{P}_F\}$ *be the SEM over* $\mathcal{Y}$ *with*

$$\begin{aligned}\mathcal{S}_Y &= \{Y_1 = 1 + F_1,\ Y_2 = Y_1 + F_2\}\\ \mathcal{I}_Y &= \{\varnothing,\ \mathrm{do}(Y_1 = 0),\ \mathrm{do}(Y_1 = 1)\},\\ F_1 &\sim \mathbb{P}_{E_2},\ \ F_2 \sim \mathbb{P}_{E_3}\end{aligned}$$

Let $\omega : \mathcal{I}_X \to \mathcal{I}_Y$ *be defined by*

$$\omega : \begin{cases} \varnothing & \mapsto \varnothing \\ \mathrm{do}(X_2 = 0) & \mapsto \mathrm{do}(Y_1 = 1) \\ \mathrm{do}(X_1 = 0,\ X_2 = 0) & \mapsto \mathrm{do}(Y_1 = 0) \end{cases}$$

Then it is true that $\mathbb{P}^i_{\tau(X)} = \mathbb{P}_Y^{\mathrm{do}(\omega(i))}$ *for all* $i \in \mathcal{I}_X$ *and* $\omega(\varnothing) = \varnothing$*, although* ω *is not order-preserving.*

If the above SEMs were used as models of the same system, they would not suffer from the problem illustrated in Example 7. Suppose now, however, that we have performed the intervention $\mathrm{do}(X_2 = 0)$ in $\mathcal{M}_X$, corresponding to the intervention $\mathrm{do}(Y_1 = 1)$ in $\mathcal{M}_Y$. If we wish to reason about the effect of the intervention $\mathrm{do}(X_1 = 0,\ X_2 = 0)$ in $\mathcal{M}_X$, we run into a problem. $\mathcal{M}_X$ suggests that $\mathrm{do}(X_1 = 0,\ X_2 = 0)$ could be implemented by performing an additional action on top of $\mathrm{do}(X_2 = 0)$. In contrast, $\mathcal{M}_Y$ suggests that implementing the corresponding intervention $\mathrm{do}(Y_1 = 0)$ would conflict with the already performed intervention $\mathrm{do}(Y_1 = 1)$.

3.5 Examples of exact transformations

In the introduction we motivated the problem considered in this paper by listing three settings in which differing model levels naturally occur. Having now introduced the notion of an exact transformation between SEMs, we provide in this section examples of exact transformations falling into each of these categories. The fact that a single framework can be used to draw an explicit correspondence between differing model levels in each of these settings demonstrates the generality of our framework.

Observe that in each of the following examples, the particular set of interventions considered is important. If we were to allow larger sets of interventions $\mathcal{I}_X$ in the SEM $\mathcal{M}_X$, the transformations given would not be exact. This highlights the importance to the causal modelling process of carefully considering the set of interventions. All proofs are found in the Appendix.

3.5.1 Marginalisation of variables

In the following two Theorems we consider two operations that can be performed on SEMs, namely marginalisation of childless or non-intervened variables, and prove that these are exact transformations. That is, an SEM can be simplified into an SEM with fewer variables by either of these operations without losing any causal content concerning the remaining variables.

Thus if the SEM $\mathcal{M}_Y$ can be obtained from another SEM $\mathcal{M}_X$ by successively performing the operations in the following theorems, then $\mathcal{M}_Y$ is an exact transformation of $\mathcal{M}_X$ and hence the two models are causally consistent. This formally explains why we can sensibly consider causal models that focus on a subsystem $\mathcal{M}_Y$ of a more complex system $\mathcal{M}_X$ (cf. Figure 3.3). For a measure-theoretic treatment of marginalisation in SEMs, see Bongers et al. (2016).

Theorem 9 (Marginalisation of childless variables)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_E)$ be an SEM and suppose that $\mathbb{I}_Z \subset \mathbb{I}_X$ is a set of indices of variables with no children, i. e. if $i \in \mathbb{I}_Z$ then X_i does not appear in the right-hand side of any structural equation in $\mathcal{S}_X$. Let $\mathcal{Y}$ be the set in which $Y = (X_i : i \in \mathbb{I}_X \setminus \mathbb{I}_Z)$ takes value. Then the transformation $\tau : \mathcal{X} \to \mathcal{Y}$ mapping*

$$\tau : (x_i : i \in \mathbb{I}_X) = x \mapsto y = (x_i : i \in \mathbb{I}_X \setminus \mathbb{I}_Z)$$

naturally gives rise to an SEM $\mathcal{M}_Y$ that is an exact τ-transformation of $\mathcal{M}_X$, corresponding to marginalising out the childless variables X_i for $i \in \mathbb{I}_Z$.

Theorem 10 (Marginalisation of non-intervened variables)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_E)$ be an* acyclic *SEM and suppose that $\mathbb{I}_Z \subset \mathbb{I}_X$ is a set of indices of variables that are not intervened upon by any intervention $i \in \mathcal{I}_X$. Let $\mathcal{Y}$ be the set in which $Y = (X_i : i \in \mathbb{I}_X \setminus \mathbb{I}_Z)$ takes value. Then the transformation $\tau : \mathcal{X} \to \mathcal{Y}$ mapping*

$$\tau : (x_i : i \in \mathbb{I}_X) = x \mapsto y = (x_i : i \in \mathbb{I}_X \setminus \mathbb{I}_Z)$$

naturally gives rise to an SEM $\mathcal{M}_Y$ that is an exact τ-transformation of $\mathcal{M}_X$, corresponding to marginalising out the never-intervened-upon variables X_i for $i \in \mathbb{I}_Z$.

The assumption of acyclicity made in Theorem 10 can be relaxed to allow marginalisation of non-intervened variables in cyclic SEMs, at the expense of extra technical conditions (see Section 3 of Bongers et al. (2016)).

We remind the reader that our definition of an SEM does not require that the exogenous E-variables be independent. Theorem 10 would not hold if this restriction were made (which is usually the case in the literature); marginalising out a common parent node will in general result in its children having dependent exogenous variables.

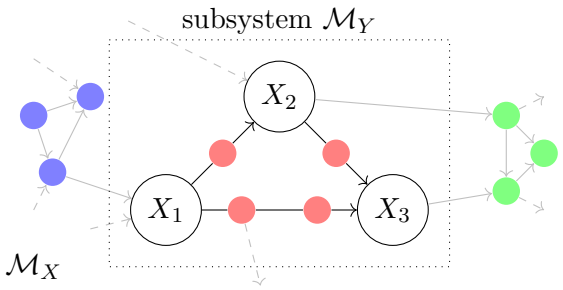


Figure 3.3 Suppose that there is a complex model $\mathcal{M}_X$ but that we only wish to model the distribution over X_1, X_2, X_3 and how it changes under some interventions on X_1, X_2, X_3. By Theorem 9, we can ignore downstream effects (●) after grouping them together as one multivariate variable and by Theorem 10 we can ignore intermediate steps of complex mechanisms (●) and treat upstream causes as noise fluctuations (●). That is, we can exactly transform the complex SEM $\mathcal{M}_X$ into a simpler model $\mathcal{M}_Y$ by marginalisation.

3.5.2 Micro- to macro-level

Transformations from micro- to macro-levels may arise in situations in which the micro-level variables can be observed via a 'coarse' measurement device, represented by the function τ, e. g. we can use a thermometer to measure the temperature of a gas, but not the motions of the individual particles. They may also arise due to deliberate modelling choice when we wish to describe a system using higher level features, e. g. viewing the motor cortex as a single entity responsible for movements, rather than as a collection of individual neurons.

In such situations, our framework of exact transformations allows one to investigate whether such a macro-level model admits a causal interpretation. The following theorem provides an exact transformation between a micro-level model $\mathcal{M}_X$ and a macro-level model $\mathcal{M}_Y$ in which the variables are aggregate features of variables in $\mathcal{M}_X$ obtained by averaging (cf. Figure 3.4).

Theorem 11 (Micro- to macro-level)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_{E,F})$ be a linear SEM over the variables $W = (W_i : 1 \leq i \leq n)$ and $Z = (Z_i : 1 \leq i \leq m)$ with*

$$\begin{aligned}\mathcal{S}_X = \{&W_i = E_i : 1 \leq i \leq n\} \\ &\cup \left\{ Z_i = \sum_{j=1}^{n} A_{ij} W_j + F_i : 1 \leq i \leq m \right\} \\ \mathcal{I}_X = \Big\{&\varnothing,\ \mathrm{do}(W = w),\ \mathrm{do}(Z = z), \\ &\mathrm{do}(W = w, Z = z) : w \in \mathbb{R}^n,\ z \in \mathbb{R}^m \Big\}\end{aligned}$$

and $(E, F) \sim \mathbb{P}$ where $\mathbb{P}$ is any distribution over $\mathbb{R}^{n+m}$ and A is a matrix.

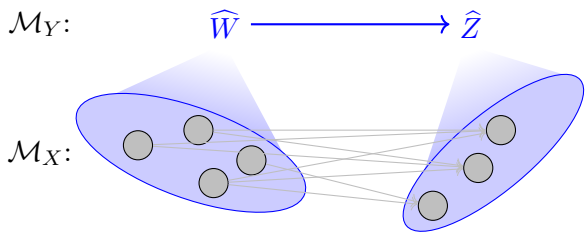


Figure 3.4 An illustration of the setting considered in Theorem 11. The micro-variables $W_1, \ldots, W_n$ and $Z_1, \ldots, Z_m$ in the SEM $\mathcal{M}_X$ can be averaged to derive macro-variables $\widehat{W}$ and $\widehat{Z}$ in such a way that the resulting macro-level SEM $\mathcal{M}_Y$ is an exact transformation of the micro-level SEM $\mathcal{M}_X$.

Assume that there exists an $a \in \mathbb{R}$ such that each column of A sums to a. Consider the following transformation that averages the W and Z variables:

$$\begin{aligned}
&\tau : \mathcal{X} \to \mathcal{Y} = \mathbb{R}^2 \\
&\begin{pmatrix} W \\ Z \end{pmatrix} \mapsto \begin{pmatrix} \widehat{W} \\ \widehat{Z} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} W_i \\ \frac{1}{m}\sum_{j=1}^{m} Z_j \end{pmatrix}
\end{aligned}$$

Further, let $\mathcal{M}_Y = \left(\mathcal{S}_Y, \mathcal{I}_Y, \mathbb{P}_{\widehat{E},\widehat{F}}\right)$ over the variables $\left\{\widehat{W}, \widehat{Z}\right\}$ be an SEM with

$$\begin{aligned}
\mathcal{S}_Y &= \left\{\widehat{W} = \widehat{E},\ \widehat{Z} = \frac{a}{m}\widehat{W} + \widehat{F}\right\} \\
\mathcal{I}_Y &= \Big\{\varnothing,\ \mathrm{do}(\widehat{W} = \widehat{w}),\ \mathrm{do}(\widehat{Z} = \widehat{z}), \\
&\qquad \mathrm{do}(\widehat{W} = \widehat{w}, \widehat{Z} = \widehat{z}) : \widehat{w} \in \mathbb{R},\ \widehat{z} \in \mathbb{R}\Big\} \\
\widehat{E} &\sim \frac{1}{n}\sum_{i=1}^{n} E_i, \quad \widehat{F} \sim \frac{1}{m}\sum_{i=1}^{m} F_i
\end{aligned}$$

Then $\mathcal{M}_Y$ is an exact τ-transformation of $\mathcal{M}_X$.

3.5.3 Stationary behaviour of dynamical processes

In this section we provide an example of an exact transformation between an SEM $\mathcal{M}_X$ describing a time-evolving system and another SEM $\mathcal{M}_Y$ describing the system after it has equilibrated. In this setting, τ could be thought of as representing our ability to only measure the time-evolving system at a single point in time, after the transient dynamics have taken place.

In particular, we consider a discrete-time linear dynamical system with identical noise and provide the explicit form of an SEM that models the distribution of the equilibria under each intervention (cf. Figure 3.5).[10]

Theorem 12 (Discrete-time linear dynamical process with identical noise)**.** *Let* $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, \mathbb{P}_E)$ *over the variables* $\{X_t^i \,:\, t \in \mathbb{Z},\, i \in \{1,\dots,n\}\}$ *be a linear SEM with*

$$\mathcal{S}_X = \left\{ X_{t+1}^i = \sum_{j=1}^{n} A_{ij} X_t^j + E_t^i \;:\; i \in \{1,\dots,n\}, t \in \mathbb{Z} \right\}$$

$$\textit{i. e. } X_{t+1} = AX_t + E_t$$

$$\mathcal{I}_X = \left\{ \mathrm{do}(X_t^j = x_j \;\forall t \in \mathbb{Z}, \forall j \in J) : x \in \mathbb{R}^{|J|},\, J \subseteq \{1,\dots,n\} \right\}$$

$$E_t = E \;\forall t \in \mathbb{Z} \textit{ where } E \sim \mathbb{P}$$

where $\mathbb{P}$ *is any distribution over* $\mathbb{R}^n$ *and* A *is a matrix.*

Assume that the linear mapping $v \mapsto Av$ *is a contraction. Then the following transformation is well-defined under any intervention* $i \in \mathcal{I}_X$*:*[11]

$$\tau : \mathcal{X} \to \mathcal{Y}$$
$$(x_t)_{t\in\mathbb{Z}} \mapsto y = \lim_{t\to\infty} x_t$$

Let $\mathcal{M}_Y = (\mathcal{S}_Y, \mathcal{I}_Y, \mathbb{P}_F)$ *be the (potentially cyclic) SEM over the variables* $\{Y^i \,:\, i \in \{1,\dots,n\}\}$ *with*

$$\mathcal{S}_Y = \left\{ Y^i = \frac{\sum_{j\neq i} A_{ij} Y^j}{1 - A_{ii}} + \frac{F^i}{1 - A_{ii}} \;:\; i \in \{1,\dots,n\} \right\}$$

$$\mathcal{I}_Y = \left\{ \mathrm{do}(Y^j = y_j \;\forall j \in J) : y \in \mathbb{R}^{|J|},\, J \subseteq \{1,\dots,n\} \right\}$$

$$F \sim \mathbb{P}$$

Then $\mathcal{M}_Y$ *is an exact* τ*-transformation of* $\mathcal{M}_X$*.*

The above theorem demonstrates how a linear additive SEM can arise as a result of making observations of a dynamical process. This supports one interpretation of SEMs as a description

[10] Note that the assumption that the transition dynamics be linear can be relaxed to more general non-linear mappings. In this case, however, the structural equations of $\mathcal{M}_Y$ can only be written in terms of implicit solutions to the structural equations of $\mathcal{M}_X$. For purposes of exposition, we stick here to the simpler case of linear dynamics.

[11] In Appendix 3.10.1 we show that A being a contraction mapping ensures that the sequence $(X_t)_{t\in\mathbb{Z}}$ defined by $\mathcal{M}_X$ converges everywhere under any intervention $i \in \mathcal{I}_X$. That is, for any realisation $(x_t)_{t\in\mathbb{Z}}$ of this sequence, its limit $\lim_{t\to\infty} x_t$ as a sequence of elements of $\mathbb{R}^n$ exists.

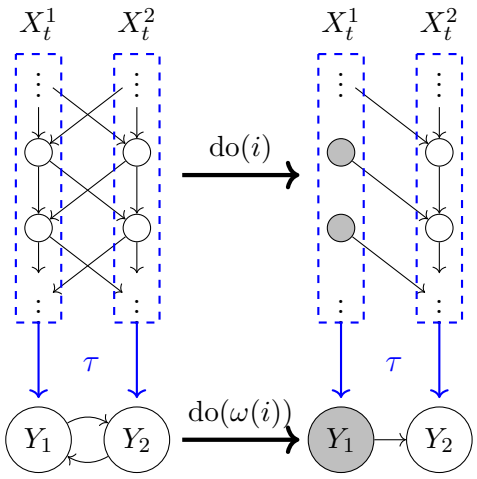


Figure 3.5 An illustration of the setting considered in Theorem 12. The discrete-time dynamical process is exactly transformed into a model describing its equilibria.

of a dynamical process that equilibrates quickly compared to its external environment.[12] The framework of exact transformations allows us to explain in a precise way the sense in which such equilibrium models can be used as causal descriptions of an underlying dynamical process.

This result also sheds light on the interpretation of cyclic causal models. One interpretation of the structural equations of an acyclic SEM is that they represent a temporally ordered series of mechanisms by which data are generated. This is not possible in the case that the SEM exhibits cycles: there does not exist a partial ordering on the variables and hence one cannot think of each variable being generated temporally downstream of its parents. By showing that cyclic SEMs can arise as exact transformations of *acyclic* SEMs, we provide an interpretation of cyclic SEMs that does not suffer from the above problem.

3.6 Discussion and Future work

It's turtles all the way down! There is no such thing as a 'correct' model, but in this paper we introduced the notions of exact transformations between SEMs to evaluate when two SEMs can be viewed as causally consistent models of the same system. Illustrating how these notions can be used in order to relate differing model levels, we proved in Section 3.5 the exactness of transformations occurring in three different settings. These have implications for the following questions in causal modelling: When can we model only a subsystem of a more complex system? When does a micro-level system admit a causal description in terms of macro-level features? How do cyclic causal models arise?

[12] This interpretation corresponds to the assumption that the noise in the dynamical model is constant through time, and is used by e. g. Hyttinen et al. (2012); Lacerda et al. (2008); Mooij et al. (2011, 2013) and Mooij and Heskes (2013) to meaningfully interpret cyclic SEMs.

Our work has implications for other problems in causal modelling. It suggests that ambiguous manipulations (Spirtes and Scheines, 2004) may be thought of as arising due to the application of an inexact transformation to an SEM $\mathcal{M}_X$. This was illustrated in Section 3.1.1 in which LDL and HDL cholesterol were only measured via their sum TC, resulting in a model that suffered from the problem of ambiguous manipulations (cf. Figure 3.1b) since it was not an exact transformation of the underlying model (cf. Figure 3.1a). This is related to the problem of causal variable definition as studied by Eberhardt (2016).

A future line of enquiry would be to generalise the notion of an exact transformation in order to analyse the trade-off between model accuracy and model complexity for causal modelling using SEMs. For a transformation to be exact, we require that the posets $\mathcal{P}_{\tau(X)}$ and $\mathcal{P}_Y$ be equal. One could imagine a 'softening' of this requirement such that the distributions in the posets are required to be only approximately equal. A slightly inaccurate model with a small number of variables may be preferable to an accurate but complex model.

We discussed the importance of an order-preserving ω to ensure a notion of causal consistency between two SEMs. It would be interesting to better understand the conditions under which different properties of consistency between causal models hold – for instance, counterfactual reasoning, which we have not discussed in this paper.

While we have introduced the notion of an exact transformation, we have not provided any criterion to choose from amongst the set of all possible exact transformations of an SEM. Foundational work in a similar direction to ours has been done by Chalupka et al. (2015, 2016), who consider a particular discrete setting. They provide algorithms to learn a transformation of a micro-level model to a macro-level model with desirable information-theoretic properties. We conjecture that our framework may lead to extensions of their work, e. g. to the continuous setting.

Finally, suppose that we have made observations of an underlying system $\mathcal{M}_X$ via a measurement device τ, and that we want to fit an SEM $\mathcal{M}_Y$ from a restricted model class to our data. By using our framework, asking whether or not $\mathcal{M}_Y$ admits a causal interpretation consistent with $\mathcal{M}_X$ reduces to asking whether the transformation is exact. More generally, by fixing any two of $\mathcal{M}_X$, τ and $\mathcal{M}_Y$, we can ask what properties must be fulfilled by the third in order for the two models to be causally consistent. We hope that this may lead to the practical use of SEMs being theoretically grounded.

Acknowledgements

We thank Tobias Mistele for valuable early feedback. Stephan Bongers was supported by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^o 639466).

3.7 Proofs for Section 3.4.3: elementary exact transformations

Lemma 4. *The identity mapping and permuting the labels of variables are both exact transformations. That is, if $\mathcal{M}_X$ is an SEM and $\pi : \mathbb{I}_X \to \mathbb{I}_X$ is a bijection then the transformation*

$$\tau : \mathcal{X} \to \mathcal{Y}$$
$$(x_i : i \in \mathbb{I}_X) \mapsto (x_{\pi(i)} : i \in \mathbb{I}_X)$$

naturally gives rise to an SEM $\mathcal{M}_Y$ that is an exact τ-transformation of $\mathcal{M}_X$, corresponding to relabelling the variables.

Proof of Lemma 4. Consider the SEM $\mathcal{M}_Y$ obtained from $\mathcal{M}_X$ by replacing, for all $i \in \mathbb{I}_X$, any occurrence of X_i in the structural equations $\mathcal{S}_X$ and interventions $\mathcal{I}_X$ by $Y_{\pi(i)}$ and leaving the distribution over the exogenous variables unchanged. □

Proof of Lemma 5 (Transitivity of exact transformations). Let $\omega_{ZY} : \mathcal{I}_Y \to \mathcal{I}_Z$ and $\omega_{YX} : \mathcal{I}_X \to \mathcal{I}_Y$ be the mappings between interventions corresponding to the exact transformations τ_{ZY} and τ_{YX} respectively and define $\omega_{ZX} = \omega_{ZY} \circ \omega_{YX} : \mathcal{I}_X \to \mathcal{I}_Z$. Then ω_{ZX} is surjective and order-preserving since both ω_{ZY} and ω_{YX} are surjective and order-preserving. Since τ_{ZY} and τ_{YX} are exact it follows that for all $i \in \mathcal{I}_X$

$$\mathbb{P}^i_{\tau_{ZX}(X)} = \mathbb{P}^{\omega_{ZY}(\omega_{YX}(i))}_{\tau_{ZY}(\tau_{YX}(X))} = \mathbb{P}^{\mathrm{do}(\omega_{ZX}(i))}_Z$$

i. e. $\mathcal{M}_Z$ is an τ_{ZX}-exact transformation of $\mathcal{M}_X$. □

3.8 Proofs for Section 3.5.1: Marginalisation of variables

Proof of Theorem 9 (Marginalisation of childless variables). By Lemma 5 it suffices to proof this for marginalisation of one childless variable. Without loss of generality, let X_1 be the childless variable to be marginalised out.

Let $\mathcal{M}_Y = (\mathcal{S}_Y, \mathcal{I}_Y, \mathbb{P}_F)$ be the SEM where

- the structural equations $\mathcal{S}_Y$ are obtained from $\mathcal{S}_X$ by removing the structural equation corresponding to the childless variable X_1;
- $\mathcal{I}_Y$ is the image of the map $\omega : \mathcal{I}_X \to \mathcal{I}_Y$ that drops any reference to the variable X_1 (e. g. $\mathrm{do}(X_1 = x_1, X_2 = x_2) \in \mathcal{I}_X$ would be mapped to $\mathrm{do}(X_2 = x_2) \in \mathcal{I}_Y$);
- $F = (E_i : \ i \in \mathbb{I}_X \setminus \{1\})$ are the remaining noise variables distributed according to their marginal distribution under $\mathbb{P}_E$.

By construction, ω is surjective and order-preserving. Let $i \in \mathcal{I}_X$ be any intervention. The variable X_1 being childless ensures that the law on the remaining variables $X_k, k \in \mathbb{I}_X \setminus \{1\}$ that we obtain by *marginalisation* of the childless variable, i. e. $\mathbb{P}^i_{\tau(X)}$, is equivalent to the law one obtains by simply *dropping* the childless variable, which is exactly what the law under $\mathcal{M}_Y$ amounts to, i. e. $\mathbb{P}_Y^{\omega(\mathrm{do}(i))}$. □

Proof of Theorem 10 (Marginalisation of non-intervened variables). By Lemma 5 it suffices to proof this for marginalisation of one never-intervened-upon variable. Without loss of generality, let X_1 be the never-intervened-upon variable to be marginalised out. By acyclicity of the SEM $\mathcal{M}_X$, the structural equation corresponding to variable X_1 is of the form $X_1 = f_1\left(\mathbf{X}_{\mathrm{pa}(1)}, E_1\right)$ and X_1 does not appear in the structural equation for any of its ancestors.

Now let $\mathcal{M}_Y = (\mathcal{S}_Y, \mathcal{I}_Y, \mathcal{P}_F)$ be the SEM where

- $\mathcal{I}_Y = \mathcal{I}_X$;
- $F_i = ((E_i, E_1) : \ i \in \mathbb{I}_X \setminus \{1\})$ are the noise variables distributed as implied by $\mathbb{P}_E$;
- the structural equations $\mathcal{S}_Y$ are obtained from $\mathcal{S}_X$ by removing the structural equation of X_1 and replacing any occurrence of X_1 in the right-hand side of the structural equations of children of X_1 by $f_1\left(\mathbf{X}_{\mathrm{pa}(1)}, E_1\right)$, yielding $X_i = f_i\left(f_1\left(\mathbf{X}_{\mathrm{pa}(1)}, E_1\right),\ \mathbf{X}_{\mathrm{pa}(i)},\ E_i\right)$.

Note that the structural equations of the resulting SEM are still acyclic and are all of the form $X_i = h_i\left(\mathbf{X}_{\setminus i},\ F_i\right)$.

Then $\mathcal{M}_Y$ is, by construction, an τ-exact transformation of $\mathcal{M}_X$ for $\omega = \mathrm{id}$. □

3.9 Proof for Section 3.5.2: Micro- to macro-level

Proof of Theorem 11. We begin by defining a mapping between interventions

$$\begin{aligned}
\omega : \mathcal{I}_X &\to \mathcal{I}_Y \\
\varnothing &\mapsto \varnothing \\
\mathrm{do}(W = w) &\mapsto \mathrm{do}\left(\widehat{W} = \frac{1}{n}\sum_{i=1}^{n} w_i\right) \\
\mathrm{do}(Z = z) &\mapsto \mathrm{do}\left(\widehat{Z} = \frac{1}{m}\sum_{i=1}^{m} z_i\right) \\
\mathrm{do}(W = w, Z = z) &\mapsto \mathrm{do}\left(\widehat{W} = \frac{1}{n}\sum_{i=1}^{n} w_i,\ \widehat{Z} = \frac{1}{m}\sum_{i=1}^{m} z_i\right)
\end{aligned}$$

Note that ω is surjective and order-preserving (in fact, it is an order embedding). Therefore, it only remains to show that the distributions implied by $\tau(X)$ under any intervention $i \in \mathcal{I}_X$

agree with the corresponding distributions implied by $\mathcal{M}_Y$. That is, we have to show that

$$\mathbb{P}^i_{\tau(X)} = \mathbb{P}^{\mathrm{do}(\omega(i))}_Y \quad \forall i \in \mathcal{I}_X$$

In the observational setting, the distribution over $\mathcal{Y}$ is implied by the following equations:

$$\widehat{W} = \frac{1}{n}\sum_{i=1}^{n} W_i = \frac{1}{n}\sum_{i=1}^{n} E_i$$
$$\widehat{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}W_j + F_i\right) = \frac{a}{m}\widehat{W} + \frac{1}{m}\sum_{i=1}^{m} F_i$$

Since the distributions of the exogenous variables in $\mathcal{M}_Y$ are given by $\widehat{E} \sim \frac{1}{n}\sum_{i=1}^{n} E_i$, $\widehat{F} \sim \frac{1}{m}\sum_{i=1}^{m} F_i$, it follows that $\mathbb{P}^{\mathrm{do}(\varnothing)}_{\tau(X)}$ and $\mathbb{P}^{\mathrm{do}(\varnothing)}_Y$ agree. Similarly, the push-forward measure on $\mathcal{Y}$ induced by the intervention $\mathrm{do}(W = w) \in \mathcal{I}_X$ is given by

$$\widehat{W} = \frac{1}{n}\sum_{i=1}^{n} W_i = \frac{1}{n}\sum_{i=1}^{n} w_i$$
$$\widehat{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}W_j + F_i\right) = \frac{a}{m}\widehat{W} + \frac{1}{m}\sum_{i=1}^{m} F_i$$

which is the same as the distribution induced by the ω-corresponding intervention $\mathrm{do}\left(\widehat{W} = \frac{1}{n}\sum_{i=1}^{n} w_i\right)$ in $\mathcal{M}_Y$.

Similar reasoning shows that this also holds for the interventions $\mathrm{do}(Z = z)$ and $\mathrm{do}(W = w, Z = z)$.

□

3.10 Proof for Section 3.5.3: stationary behaviour of dynamical processes

Proof of Theorem 12. We begin by defining a mapping between interventions

$$\omega : \mathcal{I}_X \to \mathcal{I}_Y$$
$$\mathrm{do}(X^j_t = x_j \ \forall t \in \mathbb{Z},\ \forall j \in J) \mapsto \mathrm{do}(Y^j = x_j \ \forall j \in J)$$

Note that ω is surjective and order-preserving (in fact, it is an order embedding). Therefore, it only remains to show that the distributions implied by $\tau(X)$ under any intervention $i \in \mathcal{I}_X$ agree with the corresponding distributions implied by $\mathcal{M}_Y$. That is, we have to show that

$$\mathbb{P}^i_{\tau(X)} = \mathbb{P}^{\mathrm{do}(\omega(i))}_Y \quad \forall i \in \mathcal{I}_X$$

For this we consider, without loss of generality, the distribution arising from performing the $\mathcal{M}_X$-level intervention

$$i = \mathrm{do}(X_t^j = x_j \ \forall t \in \mathbb{Z}, \forall j \leq m \leq n) \in \mathcal{I}_X$$

for $m \in [n]$ (for $m = 0$ this amounts to the null-intervention).

Since A is a contraction mapping, it follows from Lemma 15 that for any intervention in $\mathcal{I}_X$, the sequence of random variables X_t defined by $\mathcal{M}_X$ converges everywhere. That is, there exists a random variable X_* such that $X_t \xrightarrow[t\to\infty]{\text{everywhere}} X_*$. In the case of the intervention i above, the random variable X_* satisfies:

$$\begin{cases} X_*^k = x_k & \text{if } k \leq m \\ X_*^k = \sum_j A_{kj} X_*^j + E^k & \text{if } m < k \leq n \end{cases} \tag{3.1}$$

Since $\tau(X) = \lim_{t\to\infty} X_t$, it follows from the definition of X_* that $\tau(X) = X_*$, and hence $\tau(X)$ also satisfies the equations above. It follows (rewriting the second line in Equation 3.1 above) that under the push-forward measure $\mathbb{P}^i_{\tau(X)} = \tau\left(\mathbb{P}_X^{\mathrm{do}(i)}\right)$ the distribution of the random variable $\tau(X) = X_*$ is given by:

$$\begin{cases} X_*^k = x_k & \text{if } k \leq m \\ X_*^k = \frac{\sum_{j\neq k} A_{kj} X_*^j}{1-A_{kk}} + \frac{E^k}{1-A_{kk}} & \text{if } m < k \leq n \end{cases}$$

We need to compare this to the law of Y as implied by $\mathcal{M}_Y$ under the intervention $\omega(i)$, i. e. $\mathbb{P}_Y^{\mathrm{do}(\omega(i))}$. The $\mathcal{M}_Y$-level intervention $\omega(i)$ corresponding to i is

$$\omega(i) = \mathrm{do}(Y^j = x_j \ \forall j \leq m \leq n) \in \mathcal{I}_Y$$

and so the structural equations of $\mathcal{M}_Y$ under the intervention $\omega(\mathrm{do}(i))$ are

$$\begin{cases} Y^k = x_k & \text{if } k \leq m \\ Y^k = \frac{\sum_{j\neq k} A_{kj} Y^j}{1-A_{kk}} + \frac{F^k}{1-A_{kk}} & \text{if } m < k \leq n \end{cases}$$

Since $F \sim E$ it indeed follows that $\tau(X) \sim Y$, i. e. $\mathbb{P}^i_{\tau(X)} = \mathbb{P}_Y^{\mathrm{do}(\omega(i))}$.

Thus $\mathcal{M}_Y$ is an exact τ-transformation of $\mathcal{M}_X$. □

3.10.1 Contraction mapping and convergence

The following Lemmata show that A being a contraction mapping ensures that the sequence $(X_t)_{t\in\mathbb{Z}}$ defined by $\mathcal{M}_X$ in Theorem 12 converges everywhere under any intervention $i \in \mathcal{I}_X$.

That is, for any realisation $(x_t)_{t\in\mathbb{Z}}$ of this sequence, its limit $\lim_{t\to\infty} x_t$ as a sequence of elements of $\mathbb{R}^n$ exists.

Lemma 13. *Suppose that the function*

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R}^m \\ x &\mapsto f(x) \end{aligned}$$

is a contraction mapping. Then, for any $e \in \mathbb{R}^m$, so is the function

$$\begin{aligned} f^* : \mathbb{R}^n &\to \mathbb{R}^m \\ x &\mapsto f(x) + e \end{aligned}$$

Proof. By definition, there exists $c < 1$ such that for any $x, y \in \mathbb{R}^n$,

$$\|f^*(x) - f^*(y)\| = \|(f(x) + e) - (f(y) + e)\| = \|f(x) - f(y)\| \leq c\|x - y\|$$

and hence f^* is a contraction mapping. □

Lemma 14. *Suppose that the function*

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R}^n \\ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} &\mapsto \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} \end{aligned}$$

is a contraction mapping. Then for any $m \leq n$, and $x_i^* \in \mathbb{R}$, $i \in [m]$, so is the function

$$\begin{aligned} f^* : \mathbb{R}^n &\to \mathbb{R}^n \\ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} &\mapsto \begin{pmatrix} x_1^* \\ \vdots \\ x_m^* \\ f_{m+1}(x) \\ \vdots \\ f_n(x) \end{pmatrix} \end{aligned}$$

Proof. By definition, there exists $c < 1$ such that for any $x, y \in \mathbb{R}^n$,

$$\|f^*(x) - f^*(y)\| = \left\| \begin{pmatrix} x_1^* \\ \vdots \\ x_m^* \\ f_{m+1}(x) \\ \vdots \\ f_n(x) \end{pmatrix} - \begin{pmatrix} x_1^* \\ \vdots \\ x_m^* \\ f_{m+1}(y) \\ \vdots \\ f_n(y) \end{pmatrix} \right\| = \left\| \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f_{m+1}(x) - f_{m+1}(y) \\ \vdots \\ f_n(x) - f_n(y) \end{pmatrix} \right\| \leq \left\| \begin{pmatrix} f_1(x) - f_1(y) \\ \vdots \\ f_n(x) - f_n(y) \end{pmatrix} \right\|$$

$$= \|f(x) - f(y)\|$$

$$\leq c\|x - y\|$$

and hence f^* is a contraction mapping. □

Lemma 15. *Consider the SEM $\mathcal{M}_X$ in Theorem 12, and suppose that the linear map $A : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction mapping. Then, for any intervention $i \in \mathcal{I}_X$, the sequence of X_t converges everywhere.*

Proof. Consider, without loss of generality, the intervention

$$\mathrm{do}(X_t^j = x_j \ \forall t \in \mathbb{Z}, \forall j \leq m \leq n) \in \mathcal{I}_X$$

for $m \in [n]$ (for $m = 0$ this amounts to the null-intervention). The structural equations under this intervention are

$$\begin{cases} X_{t+1}^k = x_k & \text{if } k \leq m \\ X_{t+1}^k = \sum_j A_{kj} X_t^j + E^k & \text{if } m < k \leq n \end{cases}$$

and thus the sequence X_t can be seen to transition according to the function $f = g \circ h$, where

$$h : \mathbb{R}^n \to \mathbb{R}^n$$
$$v \mapsto w = Av + E$$

$$g : \mathbb{R}^n \to \mathbb{R}^n$$
$$w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ \vdots \\ x_m \\ w_{m+1} \\ \vdots \\ w_n \end{pmatrix}$$

By Lemma 13 and Lemma 14, f is a contraction mapping for any fixed E. Thus, by the contraction mapping theorem, the sequence of X_t converges everywhere to a unique fixed point. □

Chapter 4

Nonlinear Independent Component Analysis

This chapter is based on the paper *The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-View Nonlinear ICA* published at UAI 2019.

4.1 Introduction

We consider the setting described by the following generative model

$$\boldsymbol{x}_1 = \boldsymbol{f}_1(\boldsymbol{s}) \tag{4.1}$$

$$\boldsymbol{x}_2 = \boldsymbol{f}_2(\boldsymbol{s}) \tag{4.2}$$

$$p(\boldsymbol{s}) = \prod_i p_i(s_i)\,, \tag{4.3}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{s} \in \mathbb{R}^D$ and $\boldsymbol{f}_1, \boldsymbol{f}_2$ are arbitrary smooth and invertible transformations of the latent variable $\boldsymbol{s} = (s_1, \ldots, s_D)$ with mutually independent components. The goal is to recover $\boldsymbol{s}$, undoing the mixing induced by the $\boldsymbol{f}_i$, in the case where only observations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are available.

The two decoupled problems defined by considering pairs of Equations 4.1, 4.3 and 4.2, 4.3 separately are instances of Independent Component Analysis (ICA). This unsupervised learning method aims at providing a principled approach to disentanglement of independent latent components, blind source separation, and feature extraction Hyvärinen and Oja (2000). Its applications are ubiquitous, including neuroimaging McKeown and Sejnowski (1998), signal processing Sawada et al. (2003), text mining Honkela et al. (2010), astronomy Nuzillard and Bijaoui (2000) and financial time series analysis Oja et al. (2000). An ICA problem is identifiable when it is provably possible to simultaneously undo the mixing and recover the sources $\boldsymbol{s}$ up to tolerable ambiguities. Proofs of identifiability are crucial for the characteri-

zation of reliable ICA methods; in absence of these, we cannot be confident that a method successfully retrieves the true sources, even within controlled settings.

The case in which $\boldsymbol{f}_i$ is a linear function, called linear ICA, has been shown to be identifiable if at most one of the latent components is Gaussian Comon (1994); Darmois (1953); Skitovich (1954). This triggered the development of algorithms and encouraged their application. In contrast, the nonlinear ICA problem was shown to be provably unidentifiable without further assumptions on the data generating process Hyvärinen and Pajunen (1999). Much research in this field has thus attempted to characterize the assumptions under which identifiability holds. Such assumptions may be grouped into two main categories: (i) those regarding properties of the sources (e.g. non-stationarity or time correlation in time series settings Cardoso (2001); Singer and Coifman (2008)); and (ii) those restricting the functional form of the mixing functions (e.g., post-nonlinear mixing Taleb and Jutten (1999)).

A recent breakthrough was to leverage a technique known as contrastive learning, a method recasting the problem of unsupervised learning as a supervised one Gutmann and Hyvärinen (2010); Hyvarinen and Morioka (2016); Hyvärinen and Morioka (2017); Hyvarinen et al. (2019). This is a powerful proof technique, which additionally provides algorithms which can be practically implemented using modern deep learning frameworks. The setup in Hyvarinen and Morioka (2016); Hyvärinen and Morioka (2017); Hyvarinen et al. (2019) makes strong assumptions on the data generating mechanism, but allows for arbitrary nonlinear mixing of the sources. However, the unconditional independence assumption of the sources (Equation 4.3) is replaced by a *conditional* independence statement, and requires observations of the additional variable conditioned on.

In this paper, we employ contrastive learning to address the setting specified by Equations 4.1–4.3, where in contrast to Hyvarinen et al. (2019), no observations of parent variables of the sources are available. This corresponds to cases in which multiple recordings of the same process, acquired with different instruments and possibly different modalities, are available, and the goal is to find an unambiguous representation of the latent state common to all. Multiview settings of this sort are common in large biomedical and neuroimaging datasets Allen et al. (2012); Miller et al. (2016); Shafto et al. (2014); Van Essen et al. (2013), motivating the need for reliable statistical tools enabling simultaneous handling of multiple sets of variables.

As a metaphor for such a setting, consider the story of the Rosetta Stone, a stele discovered during Napoleon's campaign in Egypt in 1799, inscribed with three versions of a decree issued at Memphis in 196 BC. The realization that the stone reported the same text translated into three different languages led the French philologist Champollion to succeed in translating two unknown languages (Ancient Egyptian, in hieroglyphic script and Demotic script) by exploiting a known one (Ancient Greek). Rather, we consider the radically unsupervised task in which, given a Rosetta Stone with only two texts, both in unknown languages, we want to learn an unambiguous common representation for both of them.

The main contribution of this paper is to show that jointly addressing multiple demixing problems allows for identifiability with assumptions which do not directly refer to the sources, nor to restriction of the class of mixing functions, but rather to the conditional probability distribution of one observation given the other. This provides identifiability results in a novel setting, with assumptions entailing a different interpretation - namely, that the views have to be sufficiently diverse.

The remainder of this paper is organized as follows. In Section 4.2 we provide background information about the technique of contrastive learning for ICA and briefly review recent work that employs it. In Section 4.3 we present our main results, providing identifiability for different multi-view settings. In Section 4.4 we discuss other relevant works in the literature. Finally, we summarize and discuss our results in Section 4.5.

4.2 Nonlinear ICA with contrastive learning

Consider the nonlinear ICA setting, where observations of a variable $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$ are available, where $\boldsymbol{f}$ is an arbitrary nonlinear invertible mixing. The proof of non-identifiability for the general case with unconditionally independent sources was an important negative result Hyvärinen and Pajunen (1999). We review it briefly in Appendix 4.6.

A proposed modification of this setting Hyvarinen et al. (2019) involves an auxiliary observed variable $\boldsymbol{u}$ and a change in the independence properties. If the *unconditional independence* is substituted with a *conditional independence* given the auxiliary variable $\boldsymbol{u}$, i.e.

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_i q_i(s_i, \boldsymbol{u}) \,, \tag{4.4}$$

for some functions q_i, the model becomes identifiable. The conditional independence statement in Equation 4.4 can be interpreted as positing that $\boldsymbol{u}$ is a parent of the sources $\boldsymbol{s}$. A further assumption on the effect of variations in $\boldsymbol{u}$ on $\boldsymbol{x}$, called *variability* in the paper, is required. Intuitively, it demands that $\boldsymbol{u}$ has a sufficiently diverse influence on $\boldsymbol{x}$.

In the setting described above, a constructive proof of identifiability is attained by exploiting contrastive learning Gutmann and Hyvärinen (2010).

This technique transforms a density ratio estimation problem into one of supervised function approximation. This idea has a long history Friedman et al. (2001), and has attracted attention in machine learning in recent years Goodfellow et al. (2014); Gutmann and Hyvärinen (2010). We recapitulate the method in Appendix 4.7.

In the setting of nonlinear ICA with auxiliary variables, contrastive learning can be exploited by training a classifier to distinguish between a tuple sampled from the joint distribution, which we denote as $(\boldsymbol{x}, \boldsymbol{u})$, and one where $\boldsymbol{u}^*$ is a sample generated from the marginal $p(\boldsymbol{u})$ independently of $\boldsymbol{x}$, $(\boldsymbol{x}, \boldsymbol{u}^*)$. Intuitively, tuples drawn from the former distribution correspond to the same sources $\boldsymbol{s}$, and thus share information, while tuples

from the latter correspond to different sources and thus do not share information. Since the marginals of both distributions are equal, the classifier must learn to distinguish between them based on the common information shared by $\boldsymbol{x}$ and $\boldsymbol{u}$; that is, ultimately, $\boldsymbol{s}$.

With this method, the reconstruction of $\boldsymbol{s}$ is only possible up to an invertible scalar "gauge" transformation. This is due to a fundamental ambiguity in the setup of nonlinear ICA and does not represent a limitation of their results; it can therefore be considered a trivial one. We further comment on this in Appendix 4.6.3.

4.3 Nonlinear ICA with multiple views

We described how naively splitting Equations 4.1, 4.2 and 4.3 into two separate nonlinear ICA problems renders both problems non-identifiable, unless strong assumptions are made on the $\boldsymbol{f}_i$ or the distribution of $\boldsymbol{s}$.

In the Rosetta stone story, awareness that different texts reported on the stele were linked by a common topic helped solving the translation problem; similarly, in our setting, matched observations of the two views are linked through the shared latent variable $\boldsymbol{s}$. Thus the central question we investigate is whether these assumptions can be relaxed by exploiting the structure of the generative model; that is, whether jointly observing $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ provides sufficient constraints to the inverse problem, thus removing the ambiguities present in the vanilla nonlinear ICA setting. We consider a contrastive learning task in which a classifier is trained to distinguish between pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ corresponding to the same $\boldsymbol{s}$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2^*)$ corresponding to different realizations of $\boldsymbol{s}$. As discussed in Section 4.2, the classifier will be forced to employ the information shared by the simultaneous views in order to distinguish the two classes. As we show, this ultimately results in recovering $\boldsymbol{s}$ (up to unavoidable ambiguities).

For technical reasons discussed in Appendix 4.7, our method requires some stochasticity in the relationship between $\boldsymbol{s}$ and at least one of the $\boldsymbol{x}_i$. However this is not a significant constraint in practice; in most real settings observations are corrupted by noise, and a truly deterministic relationship between $\boldsymbol{s}$ and the $\boldsymbol{x}_i$ would be unrealistic. We will consider a component-wise independent corruption of our sources, i.e. $\boldsymbol{x}_1 = \boldsymbol{f}_1 \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ with $g_{1i}(\boldsymbol{s}, \boldsymbol{n}_1) = g_{1i}(s_i, n_{1i})$, where the components of $\boldsymbol{n}_1$ are mutually independent, and similar for $\boldsymbol{x}_2$. The noise variables $\boldsymbol{n}_1$, $\boldsymbol{n}_2$ and the sources $\boldsymbol{s}$ are assumed to be mutually independent. Note that this only puts constraints on the way the signal is corrupted by the noise, namely $\boldsymbol{g}$, and not on the mixing $\boldsymbol{f}$. We will refer to such $\boldsymbol{g}$ as *component-wise corrupter* throughout, and to its output as *corruption*. In the the vanilla ICA setting, inverting the mixing and recovering the sources $\boldsymbol{s}$ are equivalent; in the setting that we consider, the inversion of the mixing $\boldsymbol{f}$ only implies recovering the sources up to the effect of the corrupter $\boldsymbol{g}$.

We will consider three instances of the general setting, providing identifiability results for each.

1. First we consider the case that only one of the observations, $\boldsymbol{x}_2$, is corrupted with noise. This corresponds, for instance, to a setting in which one accurate measurement device is supplemented with a second noisy device. We show that in this setting it is possible to fully reconstruct $\boldsymbol{s}$ using the noiseless variable (Section 4.3.1).

2. Next, we consider the case that both variables are corrupted with noise. In this setting, it is possible to recover $\boldsymbol{s}$ up to the corruptions. Furthermore, we show that $\boldsymbol{s}$ can be recovered with arbitrary precision in the limit that the corruptions go to zero (Section 4.3.2).

3. Finally, we consider the case of having N simultaneous views of the source $\boldsymbol{s}$ rather than just two. When considering the limit $N \to \infty$, we prove sufficient conditions under which it is possible to reconstruct $\boldsymbol{s}$ even if each observation is corrupted by noise (Section 4.3.3).

To the best of our knowledge, no result of identifiability of latent sources in the case in which only corrupted, mixed versions are observed has been given before.

4.3.1 One noiseless view

Consider the generative model

$$\boldsymbol{x}_1 = \boldsymbol{f}_1(\boldsymbol{s}) \tag{4.5}$$

$$\boldsymbol{x}_2 = \boldsymbol{f}_2(\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})) \tag{4.6}$$

$$p(\boldsymbol{s}) = \prod_i p_i(s_i)$$

$$p(\boldsymbol{n}) = \prod_i p_i(n_i) \tag{4.7}$$

where $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are invertible, $\boldsymbol{g}$ is a component-wise corrupter, $\boldsymbol{n} \perp\!\!\!\perp \boldsymbol{s}$ and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are observed. This is represented in Figure 4.1.

Subject to some assumptions, it is possible to recover $\boldsymbol{s}$ up to the component-wise invertible ambiguity.

Theorem 16. *The difference of the log joint probability and log product of marginals of the observed variables in the generative model specified by Equations 4.5-4.7 admits the following*

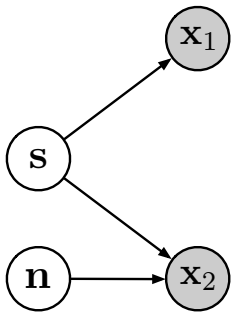


Figure 4.1 The setting considered in Section 4.3.1. Two views of the sources are available, one of which, $\boldsymbol{x}_1$, is not corrupted by noise. In this and all other figures, each node is a deterministic function of all its parents in the graph.

factorisation:

$$\begin{aligned}
&\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2) \\
&= \log p(\boldsymbol{x}_2|\boldsymbol{x}_1) - \log p(\boldsymbol{x}_2) \\
&= \left(\sum_i \alpha_i(s_i, g_i(s_i, n_i)) + \log\det J \right) \\
&\qquad - \left(\sum_i \delta_i(g_i(s_i, n_i)) + \log\det J \right) \\
&= \sum_i \alpha_i(s_i, g_i(s_i, n_i)) - \sum_i \delta_i(g_i(s_i, n_i))
\end{aligned} \tag{4.8}$$

where $s_i = f_{1i}^{-1}(\boldsymbol{x}_1)$, $g_i = f_{2i}^{-1}(\boldsymbol{x}_2)$, *and* J *is the Jacobian of the transformation* f_2^{-1} *(note that the introduced Jacobians cancel). Suppose that*

1. $\boldsymbol{\alpha}$ *satisfies the* Sufficiently Distinct Views *assumption (see after this theorem).*

2. *We train a classifier to discriminate between*

$$(\boldsymbol{x}_1, \boldsymbol{x}_2) \text{ vs. } (\boldsymbol{x}_1, \boldsymbol{x}_2^*),$$

   where $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ *correspond to the same realization of* $\boldsymbol{s}$ *and* $(\boldsymbol{x}_1, \boldsymbol{x}_2^*)$ *correspond to different realizations of* $\boldsymbol{s}$.

3. *The classifier is constrained to use a regression function of the form*

$$r(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \psi_i(h_i(\boldsymbol{x}_1), \boldsymbol{x}_2)$$

   where $\boldsymbol{h} = (h_1, \ldots, h_n)$ *are invertible, smooth and have smooth inverse.*

Then, in the limit of infinite data and with universal approximation capacity, $\boldsymbol{h}$ inverts $\boldsymbol{f}_1$ in the sense that the $h_i(\boldsymbol{x}_1)$ recover the independent components of $\boldsymbol{s}$ up to component-wise invertible transformations.

The proof can be found in Appendix 4.9.1. The assumption of invertibility for $\boldsymbol{h}$ could be satisfied by, e.g., the use of normalizing flows Chen et al. (2018c); Rezende and Mohamed (2015) or deep invertible networks Jacobsen et al. (2018).

We remark that at several points in this paper we consider the difference between two log-probabilities. In all of these cases, the Jacobians introduced by a change of variables cancel out as in Equation 4.8. For brevity we omit explanation of this fact in the rest of the results.

The *Sufficiently Distinct Views (SDV)* assumption specifies in a technical sense that the two views available are sufficiently different from one another, resulting in more information being available in totality than from each view individually. In the context of Theorem 16, it is an assumption about the log-probability of the *corruption* conditioned on the source. Informally, it demands that the probability distribution of the corruption should vary significantly as a result of conditioning on different values of the source.

Definition 17 (Sufficiently Distinct Views)**.** *Let $\alpha_i(y_i, t_i)$, $i = 1, \ldots, N$ be functions of two arguments. Denote by $\boldsymbol{\alpha}$ the vector of functions and define*

$$\alpha_i'(y_i, t_i) = \partial \alpha_i(y_i, t_i)/\partial t, \tag{4.9}$$

$$\alpha_i''(y_i, t_i) = \partial^2 \alpha_i(y_i, t_i)/\partial t^2 \tag{4.10}$$

$$\boldsymbol{w}_{\boldsymbol{\alpha}}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D'). \tag{4.11}$$

We say that $\boldsymbol{\alpha}$ satisfies the assumption of Sufficiently Distinct Views (SDV) *if for any value of $\boldsymbol{y}$, there exist $2D$ distinct values $\boldsymbol{t}_j$, $j = 1, \ldots, 2D$ such that the vectors $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$ are linearly independent.*

This is closely related to the Assumption of Variability in Hyvarinen et al. (2019). We provide simple cases of conditional log-probability density functions satisfying and violating the SDV assumption in Appendix 4.8.

Theorem 16 shows that by jointly considering the two views, it is possible to recover $\boldsymbol{s}$, in contrast to the single-view setting. This result can be extended to learn the inverse of $\boldsymbol{f}_2$ up to component-wise invertible functions.

Corollary 18. *Consider the setting of Theorem 16, and the alternative factorization of the log joint probability given by*

$$\begin{aligned}&\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)\\ &= \log p(\boldsymbol{x}_1|\boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)\\ &= \sum_i \gamma_i(s_i, g_i(s_i, n_i)) - \sum_i \beta_i(s_i))\,. \end{aligned} \tag{4.12}$$

Suppose that $\boldsymbol{\gamma}$ satisfies the SDV assumption. Replacing the regression function with

$$r(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \psi_i(\boldsymbol{x}_1, h_i(\boldsymbol{x}_2))$$

results in $\boldsymbol{h}$ inverting $\boldsymbol{f}_2$ in the sense that the $h_i(\boldsymbol{x}_2)$ recover the independent components of $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ up to component-wise invertible transformations.

The proof can be found in Appendix 4.9.2. These two results together mean that it is possible to learn inverses $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ of $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, and therefore to recover $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$, up to component-wise intertible functions. Note, however, that doing so requires running two separate algorithms. Furthermore, there is no guarantee that the learned inverses $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are 'aligned' in the sense that for each i the components $\boldsymbol{h}_{1i}(\boldsymbol{x}_1)$ and $\boldsymbol{h}_{2i}(\boldsymbol{x}_2)$ correspond to the same components of $\boldsymbol{s}$.

This problem of misalignment can be resolved by changing the form of the regression function.

Theorem 19. *Consider the settings of Theorem 16 and Corollary 18. Suppose that both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ satisfy the SDV assumption. Replacing the regression function with*

$$r(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \psi_i(h_{1,i}(\boldsymbol{x}_1), h_{2,i}(\boldsymbol{x}_2)) \tag{4.13}$$

results in $\boldsymbol{h}_1$, $\boldsymbol{h}_2$ inverting $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ in the sense that the $h_{1,i}(\boldsymbol{x}_1)$ and $h_{2,i}(\boldsymbol{x}_2)$ recover the independent components of $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,

$$h_{1,i}(\mathbf{x}_1) \perp\!\!\!\perp h_{2,j}(\mathbf{x}_2)$$

The proof can be found in Appendix 4.10.

Note that Theorem 19 is *not* a generalisation of Theorem 16 or Corollary 18, since it makes stricter assumptions by imposing the SDV assumption on both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. In contrast, Theorem 16 and Corollary 18 require that only one is valid for each.

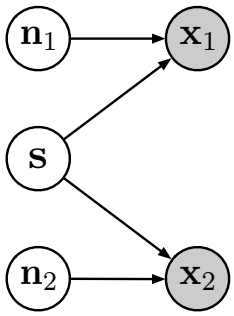


Figure 4.2 Setting with two views of the sources $\boldsymbol{s}$, both corrupted by noise.

For cases in which finding aligned representations for $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ are desired, Theorem 19 should be applied. If the only goal is recovery of $\boldsymbol{s}$, the assumptions of Theorem 16 are simpler to verify.

In practical applications, the multi-view scenario is useful in multimodal datasets where one of the two acquisition modalities has much higher signal to noise ratio than the other one (e.g., in neuroimaging, when simultaneous fMRI and Optical Imaging recordings are compared). In such cases, jointly exploiting the multiple modalities would help to discern a meaningful and identifiable latent representation which could not be attained through analysis of the more reliable modality alone.

Equivalence with Permutation Contrastive Learning for Time Dependent Sources

Note that the analysis of Theorem 16 covers the case of temporally dependent stationary sources analyzed in Hyvärinen and Morioka (2017). Indeed, if it is further assumed that $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ are uniformly dependent Hyvärinen and Morioka (2017), they can be seen as a pair of subsequent time points of an ergodic stationary stochastic process for which the analysis of Theorem 1 of Hyvärinen and Morioka (2017) would hold. In other words, we can define a stochastic process as $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t) := p(\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})|\boldsymbol{s})$. Note that while the two formulations are theoretically equivalent, our view offers a wider applicability as it covers the asynchronous sensing of $\boldsymbol{s}$, provided that multiple measurements (i.e. $\boldsymbol{x}_1, \boldsymbol{x}_2$) are available; additionally, our *Sufficiently Distinct Views* assumption does not necessarily imply uniform dependency. Furthermore, while Hyvärinen and Morioka (2017) considers a generative model of the form $\boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{s}(t))$, thus constraining the mixing function to be the same for any two data points $\boldsymbol{x}(t_1)$, $\boldsymbol{x}(t_2)$, in our setting we consider two different mixing functions, $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, for the two different views. Finally, we study this setting as an intermediate step for the following two sections, in which no deterministic function of the sources is observed, learning to invert any of the $\boldsymbol{f}_i$ can only recover $\boldsymbol{s}$ up to the corruption operated by $\boldsymbol{g}$.

4.3.2 Two noisy views

We next consider the setting in which both variables are corrupted by noise. Consider the following generative model (represented in Figure 4.2):

$$\begin{aligned}\boldsymbol{x}_1 &= \boldsymbol{f}_1(\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)) \\ \boldsymbol{x}_2 &= \boldsymbol{f}_2(\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2))\,,\end{aligned}$$

where all variables take value in $\mathbb{R}^D$, and $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are nonlinear, invertible, deterministic functions, $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are component-wise corrupters, and $\boldsymbol{s}$ and the $\boldsymbol{n}_i$ are independent with independent components. This class of models generalizes the setting of Section 4.3.1 since by taking $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1) = \boldsymbol{s}$ we reduce to the case of one noiseless observation.

The difference $\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$ admits similar factorizations to those given in Equations 4.8 and 4.12:

$$\begin{aligned}&\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2) \\ &= \log p(\boldsymbol{x}_1|\boldsymbol{x}_2) - \log p(\boldsymbol{x}_1) \\ &= \sum_i \eta_i(g_{1i}(s_i, n_{i1}), g_{2i}(s_i, n_{2i})) - \sum_i \theta_i(g_{1i}(s_i, n_{1i}) \qquad (4.14) \\ &= \log p(\boldsymbol{x}_2|\boldsymbol{x}_1) - \log p(\boldsymbol{x}_2) \\ &= \sum_i \lambda_i(g_{2i}(s_i, n_{2i}), g_{1i}(s_i, n_{1i})) - \sum_i \mu_i(g_{2i}(s_i, n_{2i})) \qquad (4.15)\end{aligned}$$

Since we only have access to corrupted observations, exact recovery of $\boldsymbol{s}$ is not possible. Nonetheless, a generalization of Theorem 19 holds showing that the $\boldsymbol{f}_i$ can be inverted and $\boldsymbol{s}$ recovered up to the corruptions induced by the $\boldsymbol{n}_i$ via $\boldsymbol{g}_i$.

Theorem 20. *Suppose that $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ satisfy the SDV assumption. The algorithm described in Theorem 16 with regression function specified in Equation 4.13 results in $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ inverting $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ in the sense that the $h_{1,i}(\boldsymbol{x}_1)$ and $h_{2,i}(\boldsymbol{x}_2)$ recover the independent components of $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,*

$$h_{1,i}(\mathbf{x}_1) \perp\!\!\!\perp h_{2,j}(\mathbf{x}_2)$$

The proof can be found in Appendix 4.10.

We can thus recover the common source $\boldsymbol{s}$ up to the corruptions $\boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)$. In the limit of the magnitude of one of the noise variables going to zero, the reconstruction of the sources $\boldsymbol{s}$ attained through the corresponding view is exact up to the component-wise invertible functions, as stated in the following corollary.

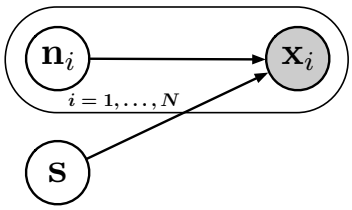


Figure 4.3 Setting with N corrupted views of the sources.

Corollary 21. *Let $\boldsymbol{n}_1^{(k)} = \frac{1}{k} \cdot \tilde{\boldsymbol{n}}$ for $k \in \mathbb{N}$, where $\tilde{\boldsymbol{n}} \in \mathbb{R}^D$ is a fixed random variable, and $\boldsymbol{n}_2$ be a random variable that does not depend on k. Let $\boldsymbol{h}_1^{(k)}, \boldsymbol{h}_2^{(k)}$ be the output of the algorithm specified by Theorem 20 with noise variables $\boldsymbol{n}_1^{(k)}$ and $\boldsymbol{n}_2$.*

Suppose that the corrupters $\boldsymbol{g}_i$ satisfy the following two criteria:

1. $\exists \boldsymbol{a} \in \mathbb{R}^D_{>0}$ *s.t.* $\left|\frac{\partial \boldsymbol{g}_1(\boldsymbol{s},\boldsymbol{n})}{\partial \boldsymbol{n}}\right|_{\boldsymbol{n}=0} \leq \boldsymbol{a}$ *for all* $\boldsymbol{s}$
2. $\exists \boldsymbol{b} \in \mathbb{R}^D_{>0}$ *s.t.* $0 < \frac{\partial \boldsymbol{g}_1(\boldsymbol{s},0)}{\partial \boldsymbol{s}} \leq \boldsymbol{b}$

Then, denoting by $\boldsymbol{E}$ the set of all scalar, invertible functions, we have that

$$\lim_{k\to\infty} \inf_{\boldsymbol{e}\in\boldsymbol{E}} \left\| \boldsymbol{s} - \boldsymbol{e}(\boldsymbol{h}_1^{(k)}(\boldsymbol{x}_1)) \right\| = 0$$

The proof can be found in Appendix 4.11.

Corollary 21 implies that in the limit of small noise, the sources $\boldsymbol{s}$ can be recovered exactly. Condition *i)* upper bounds the influence of $\boldsymbol{n}$ on the corruption: we can not hope to retrieve $\boldsymbol{s}$ if $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ contains too little signal. Condition *ii)* ensures that the function $\boldsymbol{g}$ is invertible with respect to $\boldsymbol{s}$ when $\boldsymbol{n}$ is equal to zero. If this were not satisfied, some information about $\boldsymbol{s}$ would be washed out by $\boldsymbol{g}$ even in absence of noise. This would make recovery of $\boldsymbol{s}$ trivially impossible.

4.3.3 Multiple noisy views

The results of Section 4.3.2 state that in the two noisy view setting, $\boldsymbol{s}$ can be recovered up to the corruptions. In the limit that the magnitude of the noises goes to zero, the uncorrupted $\boldsymbol{s}$ can be recovered. The intuition is that the less noise there is, the more information each observation provides about $\boldsymbol{s}$.

In this section we consider the multi-view setting, where N distinct noisy views of $\boldsymbol{s}$ are available,

$$\boldsymbol{x}_i = \boldsymbol{f}_i(\boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)) \ , i = 1, \ldots, N \,,$$

and the noise variables $\boldsymbol{n}_i$ are mutually independent, as represented in Figure 4.3. Since each view provides additional information about $\boldsymbol{s}$, we ask: in the limit as $N \to \infty$, is it possible to reconstruct $\boldsymbol{s}$ exactly?

By applying Theorem 20 to the pair $(\boldsymbol{x}_1, \boldsymbol{x}_i)$ it is possible to recover $(\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1), \boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i))$ such that the components are aligned, but up to different component-wise invertible functions

$\boldsymbol{k}_1$ and $\boldsymbol{k}_i$. Running the algorithm on a different pair $(\boldsymbol{x}_1, \boldsymbol{x}_j)$ will result in recovery up to different component-wise invertible functions $\boldsymbol{k}'_1$ and $\boldsymbol{k}'_j$.

Note that these will *not* necessarily result in $\boldsymbol{k}_i \circ \boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)$ and $\boldsymbol{k}'_j \circ \boldsymbol{g}_j(\boldsymbol{s}, \boldsymbol{n}_j)$ being aligned with each other. However, the components of $\boldsymbol{k}_1 \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{k}'_1 \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ are the same, up to permutation and component-wise invertible functions. This permutation can therefore be undone by performing independence testing between each pair of components. Components that are 'different' will be independent; those that are the same will be deterministically related. Therefore, they can be used as a reference to permute the components of $\boldsymbol{k}'_j$ and make it aligned with $\boldsymbol{k}_i$.

The problem is then how to combine the information from each aligned $\boldsymbol{k}_i \circ \boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)$ to more precisely identify $\boldsymbol{s}$. The fact that the components are recovered up to *different* scalar invertible functions makes combining information from different views non-trivial.

As a first step in this direction, we consider the special case that each $\boldsymbol{g}_i$ acts additively and each $\boldsymbol{n}_i$ is zero mean and each of $\boldsymbol{s}$ and the $\boldsymbol{n}_i$ are independent with independent components.

$$\left.\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{f}_i(\boldsymbol{s} + \boldsymbol{n}_i) \\ \mathbb{E}\boldsymbol{n}_i &= 0 \end{aligned}\right\} \quad i \in \mathbb{N} \tag{4.16}$$

Suppose to begin with that we are able to recover each $\boldsymbol{s}+\boldsymbol{n}_i$ *without* the usual component-wise invertible functions. Then, writing $\boldsymbol{n}$ to denote all of the $\boldsymbol{n}_i$, it is possible to estimate $\boldsymbol{s}$ as

$$\boldsymbol{s} \approx \Omega^N(\boldsymbol{s}, \boldsymbol{n}) = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{s} + \boldsymbol{n}_i) \,.$$

Subject to mild conditions on the rate of growth of the variances $\mathrm{Var}(\boldsymbol{n}_i)$ as $i \to \infty$, Kolmogorov's strong law implies that $\Omega^N(\boldsymbol{s}, \boldsymbol{n})$ is a good approximation to $\boldsymbol{s}$ as $N \to \infty$ in the sense that $\Omega^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} \boldsymbol{s}$. This implies moreover that it is possible to reconstruct the $\boldsymbol{n}_i$ by considering the residue $R_i^N(\boldsymbol{s}, \boldsymbol{n}) = (\boldsymbol{s} + \boldsymbol{n}_i) - \Omega^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} \boldsymbol{n}_i$.

In the presence of the unknown functions $\boldsymbol{k}_i$, we would be able to reconstruct $\boldsymbol{s}$ and the $\boldsymbol{n}_i$ if we were able to identify the inverses $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$ for each i. For any component-wise invertible functions $\boldsymbol{e}_i$, define

$$\begin{aligned} \Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n}) &= \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) \\ R_{\boldsymbol{e},i}^N(\boldsymbol{s}, \boldsymbol{n}) &= \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}). \end{aligned}$$

$\boldsymbol{e}_i$ is something we can choose and $\boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) = \boldsymbol{h}_i(\boldsymbol{x}_i)$ is the output of the algorithm, and hence $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ and $R_{\boldsymbol{e},i}^N(\boldsymbol{s}, \boldsymbol{n})$ are random variables with known distributions. Subject to mild

conditions, the dependence of these quantities on most or all of the $\boldsymbol{n}_i$ becomes increasingly small as N grows and disappears in the limit $N \to \infty$.

Lemma 22. *Suppose that the sequence* $\mathbb{E}_{\boldsymbol{n}}[\Omega^N_{\mathbf{e}}(\boldsymbol{s}, \boldsymbol{n})] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i)]$ *converges as* $N \to \infty$ *for almost all* $\boldsymbol{s}$*, and write*

$$\Omega_{\mathbf{e}}(\boldsymbol{s}) = \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{n}}[\Omega^N_{\mathbf{e}}(\boldsymbol{s}, \boldsymbol{n})].$$

Suppose further that there exists K *such that* $V_{\mathbf{e}_i} = \mathrm{Var}\left(\boldsymbol{e}_i \circ \boldsymbol{g}_i(\boldsymbol{s} + \boldsymbol{n}_i)\right) \leq K$ *for all* i*. Then*

$$\begin{aligned}
\Omega^N_{\mathbf{e}}(\boldsymbol{s}, \boldsymbol{n}) &\xrightarrow{a.s.} \Omega_{\mathbf{e}}(\boldsymbol{s}) \\
R^N_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}) &\xrightarrow{a.s.} R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) = \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) - \Omega_{\mathbf{e}}(\boldsymbol{s})
\end{aligned}$$

The proof can be found in Appendix 4.12. Given some choice of $\boldsymbol{e}$, we can think of $\Omega_{\mathbf{e}}(\boldsymbol{s})$ and $R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i)$ as our putative candidates for $\boldsymbol{s}$ and $\boldsymbol{n}_i$ respectively. As discussed earlier, if we could identify $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$, then we would have $\Omega_{\mathbf{e}}(\boldsymbol{s}) = \boldsymbol{s}$ and $R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) = \boldsymbol{n}_i$, and thus $\Omega_{\mathbf{e}}$ and $R_{\mathbf{e},i}$ would satisfy the same independences and other statistical properties as $\boldsymbol{s}$ and $\boldsymbol{n}_i$ respectively. Can we use these properties as criteria to identify good choices of $\boldsymbol{e}_i$?

The following theorem gives a set of sufficient conditions under which each $\boldsymbol{e}_i$ inverts $\boldsymbol{k}_i$ up to some affine ambiguity which is the same for every i.

Theorem 23. *Suppose there exists* $C > 0$ *such that* $Var(\boldsymbol{n}_i) \leq C$ *for all* i *and let* $\mathcal{G}_K = \{\{\boldsymbol{e}_i\}$ *s.t.*

$$V_{\boldsymbol{e}_i} \leq K \ \forall i \tag{4.17}$$

$$\Omega_{\boldsymbol{e}}(\boldsymbol{s}) < \infty \ \ \textit{for almost all } \boldsymbol{s} \tag{4.18}$$

$$R_{\boldsymbol{e},i} \perp\!\!\!\perp R_{\boldsymbol{e},j} \ \forall i \neq j, \tag{4.19}$$

$$\tag{4.20}$$

$$\mathbb{E}R_{\boldsymbol{e},i} = 0 \ \forall i \tag{4.21}$$

$$R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) = R_{\boldsymbol{e},i}(\boldsymbol{n}_i) \ \forall i \ \} \tag{4.22}$$

Then,

$$\mathcal{G}_K \subseteq \left\{\{\boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}\} \ : \ \boldsymbol{\alpha} \in \mathbb{R}^D_{\neq 0}, \ \boldsymbol{\beta} \in \mathbb{R}^D\right\}$$

where $\boldsymbol{\alpha}\boldsymbol{k}_i^{-1}$ *denotes the element-wise product with the scalar elements of* $\boldsymbol{\alpha}$*. If* $K \geq Var(\boldsymbol{s}) + C$*, then* $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$*, and so* $\mathcal{G}_K$ *is non-empty for* K *sufficiently large.*

The proof can be found in Appendix 4.13. It follows that it is possible recover $\boldsymbol{s}$ and $\boldsymbol{n}_i$ up to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via $\Omega_{\mathbf{e}}(\boldsymbol{s}) = \boldsymbol{\alpha}\boldsymbol{s} + \boldsymbol{\beta}$ and $R_{\boldsymbol{e},i}(\boldsymbol{n}_i) = \boldsymbol{\alpha}\boldsymbol{n}_i$.

We remark that each of the conditions 4.17–4.21 can be verified from known information. We conjecture that condition 4.22 can be relaxed to assuming the verifiable condition of independence between $\Omega_{\mathbf{e}}(\boldsymbol{s})$ and $R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i)$ for all i along with additional regularity assumptions on the functional form of $R_{\mathbf{e},i}$ (e.g. smoothness).

To conclude, Theorem 8 provides sufficient conditions under which it is possible to fully reconstruct $\boldsymbol{s}$ with corrupted views. In contrast to previous results in Sections 4.3.1 and 4.3.2, this result leverages infinitely many corrupted views rather than vanishingly small corruption of finitely many views.

4.4 Related work

A central concept in our work is that of multiple simultaneous views and joint extraction of features from them. We briefly review some related work considering similar settings.

4.4.1 Canonical Correlation Analysis

Given two (or more) random variables, the goal of Canonical Correlation Analysis (CCA) Hotelling (1992) is to find a corresponding pair of linear subspaces that have high cross-correlation, so that each component within one of the subspaces is correlated with a single component from the other subspace Bishop (2006). In dealing with correlation instead of independence, CCA is more closely related to Principal Component Analysis (PCA) than to ICA.

CCA can be interpretated probabilistically Bach and Jordan (2005) and is equivalent to maximum likelihood estimation in a graphical model which is a special case of that depicted in Figure 4.2. The differences compared to our setting are (i) the latent components retrieved in CCA are forced to be uncorrelated, whereas our method is retrieves independent components; (ii) in CCA, mappings between the sources $\boldsymbol{s}$ and $\boldsymbol{x}$ are linear, whereas our method allows for nonlinear mappings.

At a high level, the model we consider in Section 4.3.2 is to CCA as nonlinear ICA is to PCA. Nonlinear extensions of the basic CCA framework have been proposed Andrew et al. (2013); Fukumizu et al. (2007); Lai and Fyfe (2000); Michaeli et al. (2016), but identifiability results in the sense we consider in this paper are lacking.

4.4.2 Multi-view latent variable models

Bearing a strong resemblance to our considered setting, Lederman and Talmon (2018) proposes a sequence of diffusion maps to find the common source of variability captured by multiple sensors, discarding irrelevant sensor-specific effects. It computes the distance among the samples measured by different sensors to form a similarity matrix for the measurements of each sensor; each similarity matrix is then associated to a diffusion operator, which is a Markov matrix by construction. A Markov chain is then run by alternately applying these

Markov matrices on the initial state. During these Markovian dynamics, sensor specific information will eventually vanish, and the final state will only contain information on the common source. While the method focuses on recovering the common information in the form of a parametrization of the common variable, our method both inverts the mixing mechanisms of each view and recovers the common latent variables.

Song et al. (2014) proves identifiability for multi-view, latent variable models, unifying previously proposed spectral techniques Anandkumar et al. (2014). However, while the setting is similar to the one considered in this work, both the objectives and the employed methods are different. The paper considers the setting in which L variables X_l, $l = 1, \ldots, L$ are observed; additionally, there exists an unobserved latent variable H, such that conditional distributions $P(X_l|H)$ are independent. While the setting bears obvious similarities with our multi-view ICA, the method proposed in Song et al. (2014) is aimed at learning the mixture parameters, rather than the exact realization of latent variables. Their method is based on the mean embedding of distributions in a Reproducing Kernel Hilbert Space and a result of identifiability for the parameters of the mean embeddings of $P(H)$ and $P(X|H)$ is proved. Another related field of study is multi-view clustering, which considers a multiview setting and aims at performing clustering on a given dataset, see e.g. De Sa (2005) and Kumar et al. (2011). While related to our setting, this line of work is different from it in two key ways. Firstly, clustering can be thought of as assigning a discrete latent label per datapoint. In contrast, our setting seeks to recover a continuous latent vector per datapoint. Second, since no underlying generative model with discrete latent variable is assumed, identifiability results are not given.

4.4.3 Half-sibling regression

Half-sibling regression Schölkopf et al. (2016) is a method to reconstruct a source from noisy observations by exploiting other sources that are affected by the same noise process but otherwise independent from it.

Suppose that a latent variable of interest Q is not directly available, and that we can only observe corrupted versions of it, denoted as Y, where the corruption is due to a noise N. Without knowledge of N, it is impossible to reconstruct Q. However, if one or more additional variables X, also influenced by N, are observed, we can exploit them to model the effect of N on Y by regressing Y on X.

Subtracting this from the observed Y recovers the latent variable Q up to a constant offset, provided that (1) the additivity assumption

$$Y = Q + f(N)$$

holds, and (2) that Y contains sufficient information about $f(N)$. Analogous to our aim of recovering $\boldsymbol{s}$, the goal of half-sibling regression is not to infer only the distribution of Q, but rather the random variable itself (almost surely).

4.5 Discussion and conclusion

We presented identifiability results in a novel setting by extending the formalism of nonlinear ICA. We have investigated different scenarios of multi-view latent variable models and provided theoretical proofs on the possibility of inverting the mixing function and recovering the sources in each case. Our results thus extend the scarce literature on identifiability for nonlinear ICA models.

In the classical noiseless ICA setting, the deterministic relationship between the sources and observations means that inverting the mixing function and recovering the sources are equivalent. In contrast, we consider views of corrupted versions of the common sources, resulting in the decoupling of the demixing and retrieval of the sources. Remarkably, Theorem 23 points towards the possibility of simultaneously solving the two problems in the limit of infinitely many views.

Classical nonlinear ICA is provably non-identifiable because a single view is not sufficiently informative to resolve non-trivial ambiguities when recovering the sources. While many papers in the ICA literature have explored placing restrictions either on the source distribution or on the form of the mixing to resolve these ambiguities, in this paper we consider exploiting additional views to constrain the inverse problem. Clearly, if a second view is identical to the first, then nothing is gained by its observation. Hence, in order for the second view to assist in resolving ambiguity, it must be sufficiently different from the first. This is the intuition behind the technical assumption of *sufficiently distinct views*.

Typically, noise is a nuisance variable that would be preferably non-existent. In our setting, however, the noise variables acting on the sources are a crucial component, without which the contrastive learning approach could not be applied. Furthermore, the assumption of sufficiently distinct views is ultimately an assumption about the complexity of the joint distribution of the (corrupted) sources corresponding to each view. Without the noise variables the sufficiently distinct views assumption could not hold.

Our setting is relevant in a number of practical real-world applications, namely in all datasets that include multiple distinct measurements of related phenomena. In practice, it may be better to think of the noise variables rather as intrinsic sources of variability specific to each view. In most practical applications this would probably not be a significant limitation due to the prevalence of stochasticity in real-world systems.

An exemplary application of our method can be found in the field of neuroimaging. Consider a study involving a cohort of subjects (perceivers), measuring their response to the presentation of the same stimulus. One of the key problems in the field is how to extract a shared response from all subjects despite high inter-subject variability and complex nonlinear mappings between latent source and observation Chen et al. (2015); Haxby et al. (2011). Our results provide principled ways to extract and decompose the components of the shared response. In particular, the setting described in our model is suited to account for the high

variability of the responses throughout the cohort, since the measurement corresponding to each subject is given by a combination of individual variability and shared response.

Looking to the future, we note that Theorem 23 builds on the setting of Theorem 20 which only makes use of pairwise information from the observations. A natural extension of this work should investigate algorithms that explicitly make use of $N > 2$ views, which we conjecture would allow relaxation of the additivity assumption on the corruptions. Furthermore, Theorem 23 provides results that only hold for the asymptotic limit as the number of views becomes large. Other extensions to this result could include analysis of the case of finitely many views.

Acknowledgements

Thanks to Krikamol Muandet for providing his office for fruitful discussions, to Matthias Bauer and Manuel Wüthrich for proofreading and to Lucia Busso for interesting input about linguistics.

4.6 On the unidentifiability of nonlinear ICA

The purpose of this section is to briefly review the proof of unidentifiablity of nonlinear ICA as Hyvärinen and Pajunen (1999): In this section we assume the most general conventional form of nonlinear ICA where the generative model follows:

$$\mathbf{x} = \mathbf{f}(\boldsymbol{s}) \tag{4.23}$$

where $\boldsymbol{s}$ are the independent sources and $\mathbf{x}$ are mixed signals. In the following, we show how to construct a function $\mathbf{g} : R^n \to R^n$ so that the components $\mathbf{y} = \mathbf{g}(\mathbf{x})$ are independent. More importantly, we show that this construction is by no means unique.

4.6.1 Existence

The proposed method in Hyvärinen and Pajunen (1999) is a generalization of the famous Gram-Schmidt orthogonalization. Given m independent variables, $y_1, \ldots, y_m$ and a variable x, one constructs a new variable $y_{m+1} = g(y_1, \ldots, y_m, x)$ so that the set $y_1 s, \ldots, y_{m+1}$ is mutually independent. The construction process is defined recursively as follows. Assume we have m independent random variables $y_1, \ldots, y_m$ with uniform distribution in $[0, 1]^m$. x is any random variable and $a_1, \ldots, a_m, b$ are some nonrandom scalars. Next, we define

$$\begin{aligned} g\left(a_1, \ldots, a_m, b; p_{y,x}\right) &= p(x \leq b | y_1 = a_1, \ldots, y_m = a_m) \\ &= \frac{\int_{-\infty}^{b} p_{y,x}\left(a_1, \ldots, a_m, \xi\right) \mathrm{d}\xi}{p_y\left(a_1, \ldots, a_m\right)} \end{aligned} \tag{4.24}$$

Theorem 1 of Hyvärinen and Pajunen (1999) says that the random variable defined as $y_{m+1} = g(y_1, \ldots, y_m, x)$ is independent from the $y_1 \ldots, y_m$ and $y_1, \ldots, y_{m+1}$ are uniformly distributed in the unit cube $[0, 1]^{m+1}$.

4.6.2 Non-uniqueness

In the previous section, it was shown that there exists a mapping $\mathbf{g}$ that transforms any random vector $\mathbf{x}$ into a uniformly distributed random vector $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Here, we show that the construction of $\mathbf{g}$ is not unique and this non-Uniqueness can be caused by several factors.

- A linear transformation $\mathbf{x}'$ can precede the nonlinear map $\mathbf{f}$ and then compute the independent components $\mathbf{y}' = \mathbf{g}'(\mathbf{x}')$ where $\mathbf{g}'$ is computed as describe in the previous section. The new map $\mathbf{g}'$ gives a new decomposition of $\mathbf{x}$ into independent components $\mathbf{y}'$ which can not be trivially reduced to $\mathbf{y}$.

- An element-wise function $\mathbf{h}$ can apply on the independent sources $\boldsymbol{s}$ first to give new sources $\boldsymbol{s}'$ such that $s'_i = h_i(s_i)$. Constructing the solution g for these new scaled version of sources gives a new decomposition into independent components.

- Assume a class of measure-preserving automorphisms $\mathbf{h} : [0, 1]^n \to [0, 1]^n$. The mapping $\mathbf{h}$ does not change the probability distribution of a uniformly distributed random variable in n-dimensional hypercube. The composition $\mathbf{h} \circ \mathbf{g}$ gives another solution to nonlinear ICA. Therefore, the class of measure-preserving automorphisms gives a parameterization of the solutions to nonlinear ICA introducing a class of non-trivial indeterminacies.

If only independence among the components matters, it is possible to construct a mapping $\boldsymbol{y} = G(\boldsymbol{x})$ such that y_i is independent of y_j for $i \neq j$ and uniformely distributed in $[0, 1]^n$. This shows that at least one solution exists. The non-uniqueness of the solution can be shown by parameterising a class of infinitely many solutions. Once $\boldsymbol{y}$ is found with above conditions, any measure-preserving automorphism $f : [0, 1]^n \to [0, 1]^n$ can be used to parameterize G as $f \circ G$, suggesting that there are infinitely many solutions to nonlinear ICA whose relations are nontrivial.

4.6.3 The scalar invertible function gauge

Another indeterminacy is element-wise functions f_i applying on y_i which suggets another dimension of ambiguity. Non-Gaussianity cannot help here since we can construct any marginal distribution by combining the CDF of the observed variable with the inverse CDF of the target marginal distribution. This indeterminacy is in some sense unavoidable and is related to the fact that in linear ICA recovery of the sources is possible up to a scalar multiplicative ambiguity.

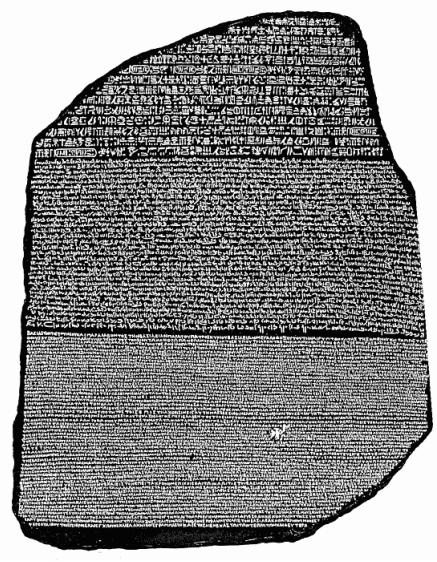


Figure 4.4 The Rosetta Stone, a stele found in 1799, inscribed with three versions of a decree issued at Memphis, Egypt in 196 BC. The top and middle texts are in Ancient Egyptian using hieroglyphic script and Demotic script, respectively, while the bottom is in Ancient Greek. (Source: Wikipedia)

4.7 Why does classification result in the log ratio?

Let us suppose that a variable X is drawn with equal probability from two distributions P_0 and P_1 with densities $p_0(x)$ and $p_1(x)$ respectively. We train a classifier $D : x \mapsto [0, 1]$ to estimate the posterior probability that a particular realization of X was drawn from P_0 with the cross entropy loss, i.e. the parameters of D are chosen to minimize

$$L(D) = \mathbb{E}_{X\sim P_0}\left[-\log D(X)\right] + \mathbb{E}_{X\sim P_1}\left[-\log(1 - D(X))\right].$$

As shown in, for instance, Goodfellow et al. (2014), the global optimum of this loss occurs when $D(x) = \frac{p_0(x)}{p_0(x)+p_1(x)}$, which can be rewritten as

$$D(x) = \frac{1}{1 + p_1(x)/p_0(x)} \tag{4.25}$$

$$= \frac{1}{1 + \exp(-\log(p_0(x)/p_1(x)))} \tag{4.26}$$

$$\tag{4.27}$$

Recall that in our setting, the function $r(x_1, x_2)$ is trained to classify between the two cases that (x_1, x_2) is drawn from the joint distribution $\mathbb{P}_{x_1,x_2}$ (*class* 0) or the product of

marginals $\mathbb{P}_{x_1}\mathbb{P}_{x_2}$ (*class* 1). $r(x_1, x_2)$ is trained so that $\frac{1}{1+\exp(-r(x_1,x_2))}$ estimates the posterior probability of (x_1, x_2) belonging to class 0. By comparing to Equation 4.26, it can be seen that

$$\begin{aligned} r(x_1, x_2) &= \log\left(p(x_1, x_2)/p(x_1)p(x_2)\right) \\ &= \log p(x_1|x_2) - \log p(x_1) \\ &= \log p(x_2|x_1) - \log p(x_2) \end{aligned}$$

Note that in order for the classification trick of contrastive learning to be useful, the variables x_1 and x_2 cannot be deterministically related. If this is the case, the log-ratio is everywhere either 0 or ∞ and hence the learned features are not useful.

To see why this is the case, suppose that x_1, and x_2 are each N-dimensional vectors. If they are deterministically related, $p(x_1, x_2)$ puts mass on an N-dimensional submanifold of a $2N$-dimensional space. On the other hand, $p(x_1)p(x_2)$ will put mass on a $2N$-dim manifold since it is the product of two distributions each of which are N-dimensional.

In this case, the distributions $p(x_1, x_2)$ and $p(x_1)p(x_2)$ are therefore not absolutely continuous with respect to one another and thus the log-ratio is ill-defined: $p(x_1, x_2)/p(x_1)p(x_2) = \infty$ at any point (x_1, x_2) at which $p(x_1, x_2)$ puts mass and zero at points where $p(x_1)p(x_2)$ puts mass and $p(x_1, x_2)$ does not.

4.8 The sufficiently distinct views assumption

We give the following two examples to provide intuition about the Sufficiently Distinct Views (SDV) assumption - one regarding a case in which it does not hold, and another one in which it does.

A simple case in which the assumption does not hold is when the conditional probability of $\boldsymbol{z}$ given $\boldsymbol{s}$ is Gaussian, as in

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z}\exp\left[-\sum_i (z_i - s_i)^2/(2\sigma_i^2)\right], \tag{4.28}$$

where Z is the normalization factor, $Z = (2\pi)^{n/2}\prod_i \sigma_i$. Since taking second derivatives of the log-probability with respect to s_i results in constants, it can be easily shown that there is no way to find $2D$ vectors $\boldsymbol{z}_j$, $j = 1, \ldots, 2D$, such that the corresponding $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z}_j)$ (see Definition 1) are linearly independent.

The fact that the assumption breaks down in this case is reminiscent of the breakdown in the case of Gaussianity for linear ICA. Interestingly, in our work, the true latent sources **are** allowed to be Gaussian. In fact, the distribution of $\boldsymbol{s}$ does not enter the expression above.

An example in which the SDV assumption does hold is a conditional pdf given by

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})} \exp\left[-\sum_i (z_i^2 s_i^2 + z_i^4 s_i^4)\right], \tag{4.29}$$

where $Z(\boldsymbol{s})$ is again a normalization function. Proving that this distribution satisfies the SDV assumption requires a few lines of computation. The idea is that $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z})$ can be written as the product of a matrix and vector which are functions only of $\boldsymbol{s}$ and $\boldsymbol{z}$ respectively. Once written in this form, it is straightforward to show that the columns of the matrix are linearly independent for almost all values of $\boldsymbol{s}$ and that $2D$ linearly independent vectors can be realized by different choices of $\boldsymbol{z}$.

4.9 Proof of Theorem 16 and Corollary 18

4.9.1 Proof of Theorem 16

This proof is mainly inspired by the techniques employed by Hyvarinen et al. (2019).

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ are s.t.

$$h_i(\boldsymbol{x}_1) \perp\!\!\!\perp h_j(\boldsymbol{x}_1), \forall i \neq j$$

We start by writing the difference in log-densities of the two classes:

$$\begin{aligned}\sum_i \psi_i(h_i(\boldsymbol{x}_1), \boldsymbol{x}_2) = \sum_i \alpha_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) + \\ - \sum_i \delta_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2))\end{aligned}$$

We now make the change of variables

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{x}_1) \\ \boldsymbol{v}(\boldsymbol{y}) &= \boldsymbol{f}_1^{-1}(\boldsymbol{h}^{-1}(\boldsymbol{y})) \\ \boldsymbol{t} &= \boldsymbol{f}_2^{-1}(\boldsymbol{x}_2))\end{aligned}$$

and rewrite the first equation in the following form:

$$\sum_i \psi_i(y_i, \boldsymbol{x}_2) = \sum_i \alpha_i(v_i(\boldsymbol{y}), t_i) \tag{4.30}$$

$$- \sum_i \delta_i(t_i) \tag{4.31}$$

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation 4.39. Adopting the conventions in 4.9 and 4.10 and

$$v_i^j(\boldsymbol{y}) = \partial v_i(\boldsymbol{y})/\partial y_j \tag{4.32}$$

$$v_i^{jj'}(\boldsymbol{y}) = \partial^2 v_i(\boldsymbol{y})/\partial y_j \partial y_{j'}\,, \tag{4.33}$$

we have

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y}) + \alpha_i'(v_i(\boldsymbol{y}), t_i) v^{jj'}(\boldsymbol{y}) = 0\,,$$

where taking derivative w.r.t. y_j and y_j' for $j \neq j'$ makes LHS equal to zero, since the LHS has functions which depend only one y_i each. If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i)\boldsymbol{a}_i(\boldsymbol{y}) + \alpha_i'(v_i(\boldsymbol{y}), t_i))\boldsymbol{b}_i(\boldsymbol{y}) = 0\,.$$

The above expression can be recast in matrix form,

$$\boldsymbol{M}(\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0\,,$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_n'', \alpha_1', \ldots, \alpha_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0\,,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one y_j.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

□

4.9.2 Proof of Corollary 18

Proof. This follows exactly by repeating the proof of Theorem 16 where the roles of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are exchanged and the regression function in the statement of the corollary is used. □

4.10 Proof of Theorems 19 AND 20

Theorem 19 is a special case of Theorem 20 by considering the case $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1) = \boldsymbol{s}$. We therefore prove only the more general Theorem 20.

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ and $k_i(\boldsymbol{x}_2)$ are such that

$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{1,j}(\boldsymbol{x}_1), \forall i \neq j \tag{4.34}$$

$$h_{2,i}(\boldsymbol{x}_2) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j \tag{4.35}$$

$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j. \tag{4.36}$$

We start by exploiting Equations 4.14 and 4.15 to write the difference in log-densities of the two classes

$$\begin{aligned} &\sum_i \psi_i(h_{1,i}(\boldsymbol{x}_1), h_{2,i}(\boldsymbol{x}_2)) \\ =&\sum_i \eta_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) - \sum_i \theta_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1)) \qquad (4.37) \\ =&\sum_i \lambda_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2), \boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1)) - \sum_i \mu_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) \qquad (4.38) \end{aligned}$$

We now make the change of variables

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{h}_1(\boldsymbol{x}_1) \\ \boldsymbol{t} &= \boldsymbol{h}_2(\boldsymbol{x}_2) \\ \boldsymbol{v}(\boldsymbol{y}) &= \boldsymbol{f}_1^{-1}(\boldsymbol{h}_1^{-1}(\boldsymbol{y})) \\ \boldsymbol{u}(\boldsymbol{t}) &= \boldsymbol{f}_2^{-1}(\boldsymbol{h}_2^{-1}(\boldsymbol{t})) \end{aligned}$$

and rewrite equation 4.37 in the following form:

$$\begin{aligned} &\sum_i \psi_i(y_i, t_i) \\ &= \sum_i \eta_i(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) - \sum_i \theta_i(v_i(\boldsymbol{y})) \qquad (4.39) \end{aligned}$$

We first want to prove the condition in Equation 4.34. We will show this is true by proving that

$$v_i(\boldsymbol{y}) \equiv v_i(y_{\pi(i)}) \tag{4.40}$$

for some permutation of the indices π with respect to the indexing of the sources $\boldsymbol{s} = (s_1, \ldots, s_D)$.

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation 4.39, yielding

$$\begin{aligned}&\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y}) \\ &+ \sum_i \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))v^{jj'}(\boldsymbol{y}) = 0\end{aligned}$$

If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\begin{aligned}&\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))\boldsymbol{a}_i(\boldsymbol{y}) \\ &+ \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))\boldsymbol{b}_i(\boldsymbol{y}) = 0\,.\end{aligned}$$

Again following Hyvarinen et al. (2019), we recast the above formula in matrix form,

$$\boldsymbol{M}(\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0\,, \tag{4.41}$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\eta_1'', \ldots, \eta_n'', \eta_1', \ldots, \eta_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption on $\boldsymbol{\eta}$. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0\,,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one $y_j = y_{\pi(i)}$.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

For the condition in Equation 4.35, we need

$$u_i(\boldsymbol{t}) \equiv u_i(t_{\tilde{\pi}(i)})\,, \tag{4.42}$$

where the permutation $\tilde{\pi}$ doesn't need to be equal to π. By symmetry, exactly the same argument as used to prove the condition in Equation 4.40 holds, by replacing $(\boldsymbol{v}, \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\theta})$ with $(\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, noting that the SDV assumption is also assumed for $\boldsymbol{\lambda}$.
We have shown that $\boldsymbol{y} = \boldsymbol{h}_1(\boldsymbol{x}_1)$ and $\boldsymbol{t} = \boldsymbol{h}_2(\boldsymbol{x}_2)$ estimate $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$ up to two different gauges of all possible scalar invertible functions.

A remaining ambiguity could be that the two representations might be misaligned; that is, defining $\boldsymbol{z}_1 = \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{z}_2 = \boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$, while

$$z_{1,i} \perp\!\!\!\perp z_{2,j} \forall i \neq j \tag{4.43}$$

we might have

$$y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \forall i \neq j \,,$$

where $\pi(i)$, $\tilde{\pi}(i)$ are two different permutations of the indices $i = 1, \ldots, n$. We want to show that this ambiguity is also resolved; that means, our goal is to show that

$$y_i \perp\!\!\!\perp t_j, \;\; \forall i \neq j \tag{4.44}$$

We recall that, by definition, we have $v_i(y_{\pi(i)}) = z_{1,i}$ and $u_j(t_{\tilde{\pi}(j)}) = z_{2,j}$. Then, due to equation 4.43,

$$v_i(y_{\pi(i)}) \perp\!\!\!\perp u_j(t_{\tilde{\pi}(j)}) \;\; \forall i \neq j \tag{4.45}$$

$$\implies y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \;\; \forall i \neq j \tag{4.46}$$

$$\implies y_i \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(j)} \;\; \forall i \neq j \,, \tag{4.47}$$

where the implication 4.45-4.46 follows from invertibility of v_i and u_j, and the implication 4.46-4.47 follows from considering that, given that we know 4.46, we can define $l = \pi(j)$ and $k = \pi(i)$ and have

$$y_k \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(l)} \;\; \forall k \neq l.$$

Define

$$\tau = \tilde{\pi} \circ \pi^{-1}$$

and note that it is a permutation. Then

$$y_i \perp\!\!\!\perp t_{\tau(j)} \forall i \neq j \tag{4.48}$$

Fix any particular i. Our goal is to show that for any $j \neq i$ the independence relation in Equation 4.44 holds. There are two possibilities:

1. $\tau(i) = i$
2. $\tau(i) \neq i$

In the first case, τ restricted to the set $\{1,\ldots,D\}\setminus\{i\}$ is still a permutation, and thus considering the independences of Equation 4.48 for all $j \neq i$ implies each of the independences of Equation 4.44 and we are done.

Let us consider the second case. Then,

$$\exists l \in \{1,\ldots,D\}\setminus\{i\} \text{ s.t. } l = \tau(i)\,.$$

We then need to prove

$$y_i \perp\!\!\!\perp t_l\,, \tag{4.49}$$

which is the only independence implied by Equation 4.44 which is not implied by Equation 4.48.

In order to do so, we rewrite equation 4.39, yielding

$$\begin{aligned}&\sum_m \psi_m(y_m, t_m)\\ =&\sum_m \eta_m(v_m(y_{\pi(m)}), u_m(t_{\tilde{\pi}(m)})) - \sum_m \theta_i(v_m(y_{\pi(m)}))\end{aligned} \tag{4.50}$$

We now take derivative with respect to y_i and t_l in 4.49; noting that $\tilde{\pi}^{-1}(l) = \pi^{-1}(i)$, we get

$$\begin{aligned}0 =&\frac{\partial^2}{\partial v_{\pi^{-1}(i)}\partial u_{\pi^{-1}(i)}}\eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l))\\ &\times\frac{\partial}{\partial y_i}v_{\pi^{-1}(i)}(y_i)\frac{\partial}{\partial t_l}u_{\pi^{-1}(i)}(t_l)\end{aligned} \tag{4.51}$$

Since $v_{\pi^{-1}(i)}(y_i)$ is a smooth and invertible function of its argument, the set of y_i such that $\frac{\partial}{\partial y_i}v_{\pi^{-1}(i)}(y_i) = 0$ has measure zero. Similarly, $\frac{\partial}{\partial t_l}u_{\pi^{-1}(i)}(t_l) = 0$ on a set of measure zero.

It therefore follows that

$$\frac{\partial}{\partial y_i}v_{\pi^{-1}(i)}(y_i)\frac{\partial}{\partial t_l}u_{\pi^{-1}(i)}(t_l) \neq 0$$

almost everywhere and hence that

$$\frac{\partial^2}{\partial v_{\pi^{-1}(i)}\partial u_{\pi^{-1}(i)}}\eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) = 0\,. \tag{4.52}$$

almost everywhere. We can thus conclude that

$$\begin{aligned}&\eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) =\\ &\eta^y_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i)) + \eta^t_{\pi^{-1}(i)}(u_{\pi^{-1}(i)}(t_l))\end{aligned}$$

This in turn implies that, for some functions A and B, we can write

$$\begin{aligned}&\log p(z_{1,\pi^{-1}(i)}|z_{2,\pi^{-1}(i)}) - \log p(z_{1,\pi^{-1}(i)})\\ &= A(v_{\pi^{-1}(i)}(y_i)) + B(u_{\pi^{-1}(i)}(t_l))\end{aligned}$$

and therefore

$$\log p(z_{1,\pi^{-1}(i)}, z_{2,\pi^{-1}(i)}) = C(v_{\pi^{-1}(i)}(y_i)) + D(u_{\pi^{-1}(i)}(t_l))$$

for some functions C and D. This decomposition of the log-pdf implies

$$\begin{aligned}&z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\pi^{-1}(i)}\\ \implies\ &z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\tilde{\pi}^{-1}(l)}\\ \implies\ &v_{\pi^{-1}(i)}(y_i) \perp\!\!\!\perp u_{\tilde{\pi}^{-1}(l)}(t_l)\\ \implies\ &y_i \perp\!\!\!\perp t_l\,,\end{aligned}$$

where the last implication holds due to invertibility of $v_{\pi^{-1}(i)}$ and $u_{\tilde{\pi}^{-1}(l)}$.

We have thus concluded the proof.

□

4.11 Proof of Corollary 21

Proof. Denoting by $\boldsymbol{d}_1^{(k)}$ the component-wise invertible ambiguity up to which $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})$ is recovered, we have that

$$\inf_{\boldsymbol{e}\in\boldsymbol{E}} \mathbb{E}_{\boldsymbol{x}_1}\left[\left\|\boldsymbol{s} - \boldsymbol{e}(\boldsymbol{h}_1^{(k)}(\boldsymbol{x}_1))\right\|_2^2\right] \tag{4.53}$$

$$= \inf_{\boldsymbol{e}\in\boldsymbol{E}} \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}\circ\boldsymbol{d}_1^{(k)}\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{4.54}$$

$$= \inf_{\tilde{\boldsymbol{e}}\in\boldsymbol{E}} \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \tilde{\boldsymbol{e}}\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{4.55}$$

$$\leq \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{4.56}$$

The lower bound holds for any $\boldsymbol{e}^* \in \boldsymbol{E}$ by definition of infimum and in particular for $\boldsymbol{e}^* = \boldsymbol{g}_1|_{\boldsymbol{n}=0}^{-1}$, the existence of which is guaranteed by the assumptions on $\boldsymbol{g}_1$. Taking a Taylor

expansion of $\boldsymbol{e}^* \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})$ around $\boldsymbol{n}_1^{(k)} = 0$ yields

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{n}_1^{(k)}, \boldsymbol{s})} \Bigg[\Bigg\| \boldsymbol{s} - \boldsymbol{e}^* \circ \boldsymbol{g}_1(\boldsymbol{s}, 0) \\
&\quad + \frac{\partial \boldsymbol{e}^*}{\partial \boldsymbol{g}_1} \frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{n}_1^{(k)}} \cdot \boldsymbol{n}_1^{(k)} + \mathcal{O}(\|\boldsymbol{n}_1^{(k)}\|^2) \Bigg\|_2^2 \Bigg] \\
&= \mathbb{E}_{(\boldsymbol{n}_1^{(k)}, \boldsymbol{s})} \left[\left\| \frac{\partial \boldsymbol{e}^*}{\partial \boldsymbol{g}_1} \frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{n}_1^{(k)}} \cdot \boldsymbol{n}_1^{(k)} + \mathcal{O}(\|\boldsymbol{n}_1^{(k)}\|^2) \right\|_2^2 \right] \\
&\longrightarrow 0 \text{ as } k \longrightarrow \infty
\end{aligned}$$

where the last equality follows from fact that $\boldsymbol{e}^* = \boldsymbol{g}|_{\boldsymbol{n}=0}^{-1}$ and the convergence follows from the fact that $\boldsymbol{n}_1^{(k)} \longrightarrow 0$ as $k \to \infty$. □

4.12 Proof of Lemma 22

We will make crucial use of *Kolmogorov's strong law*:

Theorem 24. *Suppose that X_n is a sequence of independent (but not necessarily identically distributed) random variables with*

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \mathrm{Var}[X_n] < \infty$$

Then,

$$\frac{1}{N} \sum_{n=1}^{N} X_n - \mathbb{E}[X_n] \xrightarrow{a.s.} 0$$

Fix $\boldsymbol{s}$ and consider $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ as a random variable with randomness induced by $\boldsymbol{n}$. We will show that for almost all $\boldsymbol{s}$ this converges $\boldsymbol{n}$-almost surely to a constant, and hence $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ converges almost surely to a function of $\boldsymbol{s}$.

The law of total expectation says that

$$\begin{aligned}
&\mathrm{Var}_{\boldsymbol{s}, \boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)] \\
&= \mathbb{E}_{\boldsymbol{s}}\left[V_i(\boldsymbol{s})\right] + \mathrm{Var}_{\boldsymbol{s}}\left[\mathbb{E}_{\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)]\right] \\
&\geq \mathbb{E}_{\boldsymbol{s}}\left[V_i(\boldsymbol{s})\right].
\end{aligned}$$

Since by assumption $\mathrm{Var}_{\boldsymbol{s},\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i)] \leq K$, we have that

$$\mathbb{E}_{\boldsymbol{s}}\left[\sum_{i=1}^{\infty}\frac{V_i(\boldsymbol{s})}{i^2}\right] \leq \frac{K\pi^2}{6}$$

and therefore $\sum_{i=1}^{\infty}\frac{V_i(\boldsymbol{s})}{i^2} < \infty$ with probability 1 over $\boldsymbol{s}$, else the expectation above would be unbounded since $V_i(\boldsymbol{s}) \geq 0$.

We have further that for almost all $\boldsymbol{s}$,

$$\Omega_{\boldsymbol{e}}(\boldsymbol{s}) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}E_{\boldsymbol{e}_i}(\boldsymbol{s})$$

exists. Therefore, for almost all s the conditions of Kolmogorov's strong law are met by $\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n})$ and so

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n}) - \mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n})] \overset{\boldsymbol{n}-a.s.}{\longrightarrow} 0$$

Since $\mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n})] \overset{\boldsymbol{n}-a.s.}{\longrightarrow} \Omega_{\boldsymbol{e}}(\boldsymbol{s})$, it follows that

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n}) \overset{\boldsymbol{n}-a.s.}{\longrightarrow} \Omega_{\boldsymbol{e}}(\boldsymbol{s}).$$

Since this holds with probability 1 over $\boldsymbol{s}$, we have that

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n}) \overset{\boldsymbol{n}-a.s.}{\longrightarrow} \Omega_{\boldsymbol{e}}(\boldsymbol{s}).$$

It follows that we can write

$$\begin{aligned} R_{\boldsymbol{e},i}^N(\boldsymbol{s},\boldsymbol{n}) &= \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n}) \\ &\overset{a.s.}{\longrightarrow} R_{\boldsymbol{e},i}(\boldsymbol{s},\boldsymbol{n}_i) := \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}(\boldsymbol{s}) \end{aligned}$$

4.13 Proof of Theorem 23

We will begin by showing that if $K \geq \mathrm{Var}(\boldsymbol{s}) + C$ then $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$.

For $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$, we have that

$$\begin{aligned} \Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{s}+\boldsymbol{n}_i &\overset{a.s.}{\longrightarrow} \boldsymbol{s} = \Omega_{\boldsymbol{e}}^N(\boldsymbol{s}) \\ R_i^N = \boldsymbol{s}+\boldsymbol{n}_i - \Omega_{\boldsymbol{e}}(\boldsymbol{s},\boldsymbol{n}) &\overset{a.s.}{\longrightarrow} \boldsymbol{n}_i = R_{\boldsymbol{e},i}(\boldsymbol{n}_i) \end{aligned}$$

where the convergences follow from application of Kolmogorov's strong law, using the fact that $\text{Var}(\boldsymbol{n}_i) \leq C$ for all i. Satisfaction of condition 4.17 follows from the fact that $\text{Var}_{\boldsymbol{s},\boldsymbol{n}_i}(\boldsymbol{s}+\boldsymbol{n}_i) \leq C + \text{Var}(\boldsymbol{s}) \leq K$. Since $\boldsymbol{s}$ is a well-defined random variable, $\Omega_{\boldsymbol{e}}(\boldsymbol{s}) < \infty$ with probability 1, satisfying condition 4.18. It follows from the mutual independence of $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$ that $R_{\boldsymbol{e},i}$ and $R_{\boldsymbol{e},j}$ satisfy condition 4.19. Condition 4.21 follows from the fact that $\mathbb{E}[\boldsymbol{n}_i] = 0$ Condition 4.22 follows from $R_{\boldsymbol{e},i}$ being constant as a function of $\boldsymbol{s}$.

It therefore follows that $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$ for K sufficiently large.

We will next show that if $\{\boldsymbol{e}_i\} \in \mathcal{G}_K$ then there exist a matrix $\boldsymbol{\alpha}$ and vector $\boldsymbol{\beta}$ such that $\boldsymbol{e}_i = \boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}$ for all i. Since $\boldsymbol{e}_i$ acts coordinate-wise, it moreover follows that $\boldsymbol{\alpha}$ is diagonal.

First, we will show that each $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, i.e. there exist potentially different $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ such that $\boldsymbol{e}_i = \boldsymbol{\alpha}_i\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}_i$ for each i.

Then we will show that we must have $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i, j.

To see that $\boldsymbol{e}_i$ is affine, we make use of that fact that $R_{\boldsymbol{e},i}$ is constant as a function of $\boldsymbol{s}$. It follows that for any x and y

$$\begin{aligned}
\boldsymbol{e}_i \circ \boldsymbol{k}_i(x+y) &= R_{\boldsymbol{e},i}(x) + \Omega_{\boldsymbol{e}}(y) \\
&= R_{\boldsymbol{e},i}(x) + \Omega_{\boldsymbol{e}}(0) + R_{\boldsymbol{e},i}(0) + \Omega_{\boldsymbol{e}}(y) \\
&\quad - (R_{\boldsymbol{e},i}(0) + \Omega_{\boldsymbol{e}}(0)) \\
&= \boldsymbol{e}_i \circ \boldsymbol{k}_i(x) + \boldsymbol{e}_i \circ \boldsymbol{k}_i(y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)
\end{aligned}$$

It therefore follows that $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, since if we define

$$\begin{aligned}
L(x+y) &= \boldsymbol{e}_i \circ \boldsymbol{k}_i(x+y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0) \\
&= (\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)) \\
&\quad + (\boldsymbol{e}_i \circ \boldsymbol{k}_i(y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)) \\
&= L(x) + L(y)
\end{aligned}$$

then L is linear and we can write $\boldsymbol{e}_i \circ \boldsymbol{k}_i(x)$ as the sum of a linear function and a constant:

$$\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) = L(x) + \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)$$

Thus $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, and we have some (diagonal) matrix $\boldsymbol{\alpha}_i$ and vector $\boldsymbol{\beta}_i$ such that for any x

$$\begin{aligned}
&\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) = \boldsymbol{\alpha}_i x + \boldsymbol{\beta}_i \\
\implies &\boldsymbol{e}_i(x) = \boldsymbol{\alpha}_i \boldsymbol{k}_i^{-1} x + \boldsymbol{\beta}_i.
\end{aligned}$$

Next we show that for the set of $\{\boldsymbol{e}_i = \boldsymbol{\alpha}_i\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}_i\}$, it must be the case that each $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$.

Observe that

$$\begin{aligned}\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n}) &= \frac{1}{N}\sum_{i=1}^N \boldsymbol{\alpha}_i\boldsymbol{s} + \boldsymbol{\alpha}_i\boldsymbol{n}_i + \boldsymbol{\beta}_i \\ &= \left(\frac{1}{N}\sum_{i=1}^N \boldsymbol{\alpha}_i\right)\boldsymbol{s} + \frac{1}{N}\sum_{i=1}^N \boldsymbol{\beta}_i + \frac{1}{N}\sum_{i=1}^N \boldsymbol{\alpha}_i\boldsymbol{n}_i \\ \mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n})] &= \left(\frac{1}{N}\sum_{i=1}^N \boldsymbol{\alpha}_i\right)\boldsymbol{s} + \frac{1}{N}\sum_{i=1}^N \boldsymbol{\beta}_i\end{aligned}$$

Define

$$\begin{aligned}\boldsymbol{\alpha} &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N \boldsymbol{\alpha}_i \\ \boldsymbol{\beta} &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N \boldsymbol{\beta}_i\end{aligned}$$

which exist by the assumption that $\Omega_{\boldsymbol{e}}^N(\boldsymbol{s},\boldsymbol{n})$ converges as $N\to\infty$. Thus

$$\begin{aligned}\Omega_{\boldsymbol{e}}(\boldsymbol{s}) &= \boldsymbol{\alpha}\boldsymbol{s} + \boldsymbol{\beta} \\ R_{\boldsymbol{e},i}(\boldsymbol{s},\boldsymbol{n}_i) &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha})\boldsymbol{s} + \boldsymbol{\alpha}_i\boldsymbol{n}_i + \boldsymbol{\beta}_i - \boldsymbol{\beta}\end{aligned}$$

Now, suppose that there exist i and j such that such that $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}_j$. It follows that

$$\begin{aligned}R_{\boldsymbol{e},i}(\boldsymbol{s},\boldsymbol{n}_i) &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha})\boldsymbol{s} + \boldsymbol{\alpha}_i\boldsymbol{n}_i + \boldsymbol{\beta}_i - \boldsymbol{\beta} \\ R_{\boldsymbol{e},j}(\boldsymbol{s},\boldsymbol{n}_j) &= (\boldsymbol{\alpha}_j - \boldsymbol{\alpha})\boldsymbol{s} + \boldsymbol{\alpha}_j\boldsymbol{n}_j + \boldsymbol{\beta}_j - \boldsymbol{\beta}\end{aligned}$$

There are two cases. If $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}$, then $R_{\boldsymbol{e},i}(\boldsymbol{s},\boldsymbol{n}_i)$ is not a constant function of $\boldsymbol{s}$. But if $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$, then $\boldsymbol{\alpha}_j \neq \boldsymbol{\alpha}$ and so $R_{\boldsymbol{e},j}(\boldsymbol{s},\boldsymbol{n}_j)$ is not a constant function of $\boldsymbol{s}$. This is a contradiction, and so $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ for all i,j.

Suppose similarly that there exist $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$. If $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}$, then $\mathbb{E}[R_{\boldsymbol{e},i}(\boldsymbol{n}_i)] = \boldsymbol{\beta}_i - \boldsymbol{\beta}$ which is non-zero. If $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, then $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}$ and so $\mathbb{E}[R_{\boldsymbol{e},j}(\boldsymbol{n}_j)] = \boldsymbol{\beta}_j - \boldsymbol{\beta}$ is non-zero. This is a contradiction, and so $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i,j.

We have thus proven that set $\{\boldsymbol{e}_i\} \in \mathcal{G}_K$ is of the form $\boldsymbol{e}_i = \boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}$ for all i.

Chapter 5

Generative modelling / autoencoders

This chapter is based on the paper *On the Latent Space of Wasserstein Auto-Encoders*. This work was published as two separate workshop papers at ICLR 2018.

5.1 Introduction

Unsupervised generative modeling is increasingly attracting the attention of the machine learning community. Given a collection of unlabelled data points S_X, the ultimate goal of the task is to learn a model capable of generating sets of synthetic points S_G which *look similar to* S_X. The closely related field of unsupervised representation learning in addition aims to produce semantically meaningful representations (or features) of the data points S_X. There are various ways of defining the notion of *similarity* between two sets of data points. The most common approach assumes that both S_X and S_G are sampled independently from two probability distributions P_X and P_G respectively, and employ some of the known divergence measures for distributions.

Two major approaches currently dominate this field. Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014) minimize the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(P_X, P_G)$, which is equivalent to maximizing the *marginal log-likelihood* of the model P_G. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) employ a framework, commonly referred to as *adversarial training*, which is suitable for many different divergence measures, including (but not limited to) f-divergences (Nowozin et al., 2016), 1-Wasserstein distance (Arjovsky et al., 2017), and Maximum Mean Discrepancy (MMD) (Binkowski et al., 2018; Dziugaite et al., 2015; Li et al., 2017).

Both approaches have their pros and cons. VAEs can both generate and *encode* (featurize) data points, are stable to train, and typically manage to cover all modes of the data distribution. Unfortunately, they often produce examples that are far from the true data manifold. This

is especially true for structured high-dimensional datasets such as natural images, where VAEs generate *blurry* images. GANs, on the other hand, are good at producing realistic looking examples (landing on or very close to the manifold), however they cannot featurize the points, often cover only few modes of the data distribution, and are highly unstable to train (Salimans et al., 2016).

A number of recent papers modify the VAE framework with the goal of improving sample quality. Approaches include using more flexible approximate posteriors (Kingma et al., 2016), priors (Tomczak and Welling, 2018) and different ways of blending with adversarial training in the hope of combining the strengths of GANs and VAEs (Larsen et al., 2015; Makhzani et al., 2016; Mescheder et al., 2017).

Taking a different approach by switching focus from the KL objective to the optimal transport distance, Tolstikhin et al. (2018a) introduce the Wasserstein Auto-Encoder. This architecture shares many of the desirable properties of VAEs while providing samples of better quality. Importantly, WAEs still allow for adversary-free versions, resulting in a min-min training objective leading to stable training. In this work we aim at further improving the quality of generative modeling and representation learning techniques, focussing on the adversary-free WAE-MMD architecture as we find the instability of the adversarial training to be an unfortunate obstacle when it comes to controlled reproducible experiments. Our main contributions are threefold.

First, we demonstrate that a mismatch between the latent space dimensionality $d_{\mathcal{Z}}$ and the intrinsic data dimensionality $d_{\mathcal{I}}$ may harm the performance of WAEs when using deterministic encoders as considered in all experiments of Tolstikhin et al. (2018a). For auto-encoder architectures, the quality of generated samples can be no better than that of reconstructed images. Therefore we are interested in training WAEs with larger latent bottlenecks as reconstruction error generally improves with increased latent dimension. However, we show that generated sample quality actually *degrades* beyond some ideal latent dimension due to the difficulty of distribution matching in higher dimensions.

Second, we argue that WAEs can be made adaptive to the unknown $d_{\mathcal{I}}$ by using *probabilistic* encoders. Somewhat unexpectedly, the use of probabilistic encoders with WAEs is not trivial, as we empirically found that in high latent dimensions the variances of the encoding distributions tend to zero, effectively collapsing to deterministic encoders. We overcome this problem by introducing an additional regulariser; however, how to best train probabilistic-encoder WAEs remains an open question that we would encourage interested researchers to pursue.

Third, we evaluate the usefulness of probabilistic-encoder WAEs for representation learning using the *dSprites* dataset (Matthey et al., 2017), a benchmark task in learning *disentangled representations*. We evaluate our method using metrics introduced by three recent papers (Higgins et al., 2017; Kim and Mnih, 2018; Kumar et al., 2018), finding that our

method achieves competitive performance on this task compared to state-of-the-art methods introduced by the aforementioned papers.

Short introduction to Wasserstein auto-encoders

Similarly to VAEs, WAEs describe a particular way to train probabilistic *latent variable models* (LVMs) P_G. LVMs act by first sampling a code (feature) vector Z from a *prior distribution* P_Z defined over the latent space $\mathcal{Z}$ and then mapping it to a random input point $X \in \mathcal{X}$ using a conditional distribution $P_G(X|Z)$ also known as *the decoder* or *generator*. We will be mostly working with image datasets, so for simplicity we set $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$, $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, and refer to points $x \in \mathcal{X}$ as pictures, images, or inputs interchangeably.

Instead of minimizing the KL divergence between the LVM P_G and the unknown data distribution P_X as done by VAEs, WAEs aim at minimizing the optimal transport distance between them. Given any non-negative cost function $c(x, x')$ between two images, WAEs minimize the following objective with respect to parameters of the *deterministic* decoder $P_G(X|Z = z) = \delta_{G(z)}$ mapping[1] codes $z \in \mathcal{Z}$ to pictures $G(z) \in \mathcal{X}$:

$$\min_{Q(Z|X)} \mathop{\mathbb{E}}_{P_X} \mathop{\mathbb{E}}_{Q(Z|X)} \big[c(X, G(Z))\big] + \lambda \cdot \mathcal{D}_Z(Q_Z, P_Z), \tag{5.1}$$

where the conditional distributions $Q(Z|X)$ are commonly known as *encoders*, $Q_Z(Z) := \int Q(Z|X)P_X(X)dX$ is *the aggregated posterior* distribution, $\mathcal{D}_Z$ is any divergence measure between two distributions over $\mathcal{Z}$, and $\lambda > 0$ is a regularization coefficient. In practice $Q(Z|X = x)$ and $G(z)$ are often parametrized with deep nets, in which case back propagation can be used with stochastic gradient descent techniques to optimize the objective.

An important design choice that must be made when using a WAE is whether the encoder should map an image $x \in \mathcal{X}$ to a *distribution* $Q(Z|X = x)$ over the latent space or to a single code $z = \varphi(x) \in \mathcal{Z}$, i.e. $Q(Z|X = x) = \delta_{\varphi(x)}$. We refer to the former type as *probabilistic* encoders and the latter as *deterministic* encoders. In practice, the reparametrisation trick, commonly employed in the training of VAEs, may be used in the case of probabilistic encoders.

The objective (5.1) is similar to that of the VAE and has two terms. The first *reconstruction term* aligns the encoder-decoder pair so that the encoded images can be accurately reconstructed by the decoder as measured by the cost function c (we will only use the *cross-entropy loss* throughout).

The second regularization term is different from VAEs: it forces the aggregated posterior Q_Z to match the prior distribution P_Z rather than asking point-wise posteriors $Q(Z|X = x)$ to match P_Z simultaneously for all data points x. To better understand the difference, note that Q_Z is the distribution obtained by averaging conditional distributions $Q(Z|X = x)$ for all different points x drawn from the data distribution P_X. This means that WAEs explicitly

[1] Here δ_t is a point distribution supported on t.

control the shape of the *entire* encoded dataset while VAEs constrain every input point separately.

Both existing versions of the algorithm—WAE-GAN based on adversarial training and the adversary-free WAE-MMD based on the maximum mean discrepancy, only the latter of which we use in this paper—allow for *any* prior distributions P_Z and encoders $Q(Z|X)$ as long as P_Z and Q_Z can be efficiently sampled. As a result, the WAE model may be easily endowed with prior knowledge about the possible structure of the dataset through the choice of P_Z.

Notation. We denote $\varphi(x) = \mathbb{E}\left[Q(Z|X = x)\right]$ to be the mean of the encoding distribution for a given input x. By $\varphi_i(x)$ we mean the ith coordinate of $\varphi(x)$, $1 \leq i \leq d_{\mathcal{Z}}$, and by $P_Z(Z_i)$ and $Q_Z(Z_i)$ the marginal distributions of the prior and aggregated posteriors over the ith dimension of $\mathcal{Z}$ respectively.

A note on implementation and code. Throughout, we train all WAE models using Algorithm 2 (WAE-MMD) of Tolstikhin et al. (2018a) optionally with an additional regulariser that we introduce in Section 5.2.1. Experimental details not present in the main text may be found in the Supplementary Materials. All code for all experiments will be made public upon successful publication.

5.2 Dimension mismatch and probabilistic encoders

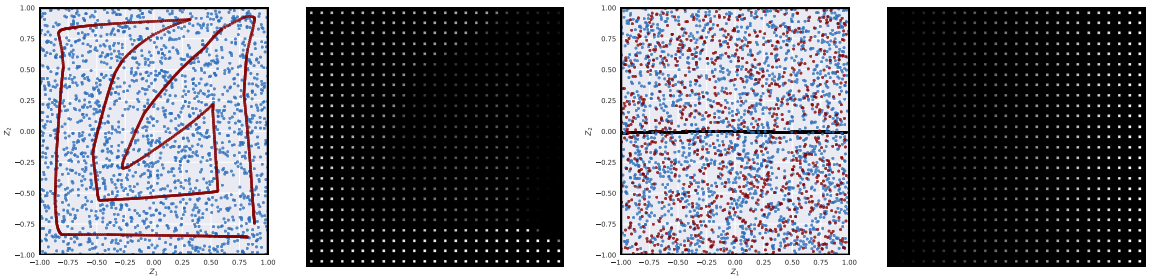
What happens if a deterministic-encoder WAE is trained with a latent space of dimension $d_{\mathcal{Z}}$ that is larger than the intrinsic dimensionality $d_{\mathcal{I}}$[2] of the data distribution? If the encoder is continuous then the data distribution P_X will be mapped to Q_Z supported on a latent manifold of dimension at most $d_{\mathcal{I}} < d_{\mathcal{Z}}$ while the regularizer in (5.1) will encourage the encoder to fill the latent space similarly to the prior P_Z as much as possible. This is a hard task for the encoder for the same reason that it is hard to fill the plane with a one dimensional curve.

To empirically investigate this setting, we introduce the simple synthetic *fading squares* dataset consisting of 32×32 pixel images of centred, 6×6 pixel grey squares on a black background. The value of this colour varies uniformly from 0 (black) to 1 (white) in steps of 10^{-3}. The intrinsic dimensionality of this dataset is therefore 1, as each image in the dataset can be uniquely identified by the value of the colour of its grey square.

We trained a deterministic-encoder WAE with a uniform prior over $[-1, 1]^2$. Since $d_{\mathcal{Z}} = 2$, we can easily visualise the learned embedding of the data into the latent space and the output of the decoder across the whole latent space. This is displayed in Figure 5.1 (left two plots) for one such WAE.

The WAE is forced to reconstruct the images well, while at the same time trying to fill the latent space uniformly with the 1-dimensional data manifold. The only way to do this is by curling up the data-manifold in the latent space. In practice, the WAE must only fill

[2] $d_{\mathcal{I}}$ is informally the minimum number of parameters required to continuously parametrise the data manifold.



Deterministic WAE (encoder, decoder) Probabilistic WAE (encoder, decoder)

Figure 5.1 Visualisations of the 2-dimensional latent space of the WAE trained on the fading squares dataset with deterministic and probabilistic encoders and a uniform prior P_Z over the box. Within each pair of plots, the left shows 1000 points sampled from the aggregated posterior Q_Z (**dark red**) and prior P_Z (**blue**); for the probabilistic encoder **black** points show data points x mapped to the mean values of the encoder $\mathbb{E}[Q(Z|X=x)]$. Right plots show decoder outputs at the corresponding points of the latent space.

the space to the extent that it fools the divergence measure which sees only a mini-batch of samples at each training step. We found that larger mini-batches resulted in tighter curling of the manifold supporting Q_Z, suggesting that mini-batch size may strongly affect the performance of WAEs via estimation of $D(Q_Z, P_Z)$.

We repeated the same experiment with a probabilistic-encoder WAE, for which the encoder maps an input image to a uniform distribution over the axis aligned box with centre $(\varphi_1(x), \varphi_2(x))$ and side lengths $(\sigma_1(x), \sigma_2(x))$. Figure 5.1 (right two plots) shows the resulting behaviour of the learned encoder and decoder. In contrast to the deterministic-encoder WAE, the probabilistic-encoder WAE is robust to the fact that $d_{\mathcal{Z}} > d_{\mathcal{I}}$ and can use one dimension to encode useful information to the decoder while filling the other with noise. That is, a single image gets mapped to a thin and tall box in the latent space. In this way, the probabilistic-encoder WAE is able to 'inflate' the latent manifold and properly match the aggregated posterior Q_Z to the prior distribution P_Z.

To what extent is it actually a problem that the deterministic WAE represents the data as a curved manifold in the latent space? There are two issues.

Poor sample quality: Only a small fraction of the total volume of the latent space is covered by the deterministic encoder. Hence the decoder is only trained on this small fraction, because under the objective (5.1) the decoder learns to act only on the encoded training images. While it appears in this 2-dimensional toy example that the quality of decoder samples is nonetheless good everywhere, in high dimensions, such 'holes' may be significantly more problematic. This possibly explains the results presented in Table 5.1, in which we find that large latent dimensions decrease the quality of the samples produced by deterministic WAEs, and has implications for understanding the better sample quality for WAE-GAN than WAE-MMD reported in Tolstikhin et al. (2018a), discussed in Supplementary Materials 5.5.2.

Incorrect proportions of generated images Although in this toy example all of the samples generated by the deterministic-encoder WAE are of good quality, we verify empirically in Supplementary Materials 5.5.3 that they are not produced in the correct proportions. By analogy, this would correspond to a model trained on MNIST producing too few 3s and too many 7s.

5.2.1 Probabilistic encoders with large $d_{\mathcal{Z}}$

To test our new intuitions about the behaviour of deterministic- and probabilistic-encoder WAEs with different latent dimensions, we next consider the *CelebA* dataset. All experiments reported in this section used Gaussian priors and, for the probabilistic-encoder WAEs, Gaussian encoders. A fixed convolutional architecture with cross-entropy reconstruction loss was used for all experiments. To keep computation time feasible, we used small networks.

Table 5.1 shows the results of training 5 probabilistic- and 5 deterministic-encoder WAEs for each value of $d_{\mathcal{Z}} \in \{32, 64, 128, 256\}$. Deterministic- and probabilistic-encoder WAEs exhibit similar behaviour: test reconstruction error decreases as $d_{\mathcal{Z}}$ increases, while the FID scores (Heusel et al., 2017) of generated samples first decrease to some minimum and then subsequently increase (lower FID scores indicate better sample quality).

For deterministic encoders, this agrees with the intuition we gained from the *fading squares* experiment. Unable to fill the whole latent space when $d_{\mathcal{I}} < d_{\mathcal{Z}}$, the encoder leaves large holes in the latent space on which the decoder is never trained. When $d_{\mathcal{I}} \ll d_{\mathcal{Z}}$, these holes occupy most of the total volume, and thus most of the samples produced by the decoder from draws of the prior are poor.

For probabilistic encoders we did not expect this behaviour. We observed that the probabilistic encoders would 'collapse' to deterministic encoders when $d_{\mathcal{I}} \ll d_{\mathcal{Z}}$, rather than using noise to expand the latent manifold, making Q_Z accurately match P_Z. More precisely, the *log-variances* for each dimension of $Q(Z|X = x)$ averaged over a mini-batch of data would typically continually decrease throughout training to less than -20. Identifying this problem is one of the crucial points of this paper, and we next propose a solution to this.

Resolving variance collapse through regularization The precise cause of this variance collapse is uncertain to us. It is clear that the reconstruction loss will encourage small variances (since this makes the task of the decoder easier), and we suspect that the MMD fails to prevent this from happening due to the weakness of the topology on distributions it induces as a distance.

Nonetheless, we found we could effectively eliminate this issue by adding an L_p penalty on the log-variances, providing encouragement for the variances to remain closer to 1 and thus for the encoder to remain stochastic. More precisely, we added the following term to

Table 5.1 FID scores and test reconstructions for deterministic- and probabilistic-encoder WAEs trained on *CelebA* for various latent dimensions $d_{\mathcal{Z}}$. Test reconstructions get better with increased dimension, while FID scores suffer for $d_{\mathcal{Z}} \gg d_{\mathcal{I}}$.

$d_{\mathcal{Z}}$	FID score		Test reconstruction	
	Det.	Prob.	Det.	Prob.
32	75.0± 0.7	74.8± 0.5	6457.0± 10.4	6445.5± 7.5
64	71.6 ± 0.8	71.1 ± 1.0	6364.4± 7.4	6365.0± 5.4
128	76.8± 1.3	76.8± 1.2	6300.5± 6.6	6309.3± 9.7
256	147.6± 2.3	139.8± 4.2	6265.3± 9.5	6262.6± 6.7

the objective function to be minimised:

$$\frac{\lambda_p}{N} \sum_{n=1}^{N} \sum_{i=1}^{d_{\mathcal{Z}}} \left| \log \left(\sigma_i^2(x_n) \right) \right|^p \tag{5.2}$$

where i indexes the dimensions of the latent space $\mathcal{Z}$, n indexes the inputs x_n in a mini-batch, and $\lambda_p \geq 0$ is the new regularization coefficient.

We experimented with both L_1 and L_2 regularisation and found both to give similar qualitative behaviour, but L_1 regularisation gave better performance and thus we report only these results. There are two interpretations we have for why L_1 regularisation might be sensible.

First, note that an L_1 penalty on the log-variances should encourage the encoder/decoder pair to *sparsely* use latent dimensions to code useful information. Indeed, the L_1 penalty will encourage sparsely many dimensions to have non-zero log-variances, and if the variance of $Q(Z_i|X)$ in some dimension i is always 1 then in order for the marginal $Q_Z(Z_i)$ to match $P_Z(Z_i)$, we must have that $\varphi_i(x) = 0$ for all inputs x, meaning that $\varphi_i(X)$ contains no information about X.

The second interpretation of this term comes from the observation that $|\log(\sigma^2)|$ and $\sigma^2 - \log(\sigma^2)$ both attain their minimum at $\sigma^2 = 1$. The first term is our regulariser, while the second term is the log-variance part of the KL regulariser of the VAE objective. In this sense, our new objective is related to that of DIP-VAE (Kumar et al., 2018), which takes the VAE objective and adds a term to penalise a divergence $D(P_Z, Q_Z)$.

Using latent dimensions 32 and 256 to consider both the case of under- and over-shooting the intrinsic dimensionality of the dataset,[3] we trained 5 L_1-regularized probabilistic-encoder WAEs for a variety of values for λ_1. Figure 5.2 shows the test reconstruction errors and FID scores obtained at the end of training.

[3]The fact that the deterministic WAE produced samples with better FIDs with $d_{\mathcal{Z}} = 64$ than $d_{\mathcal{Z}} = 32$ suggested to us that the intrinsic dimensionality of the CelebA datset is greater than 32.

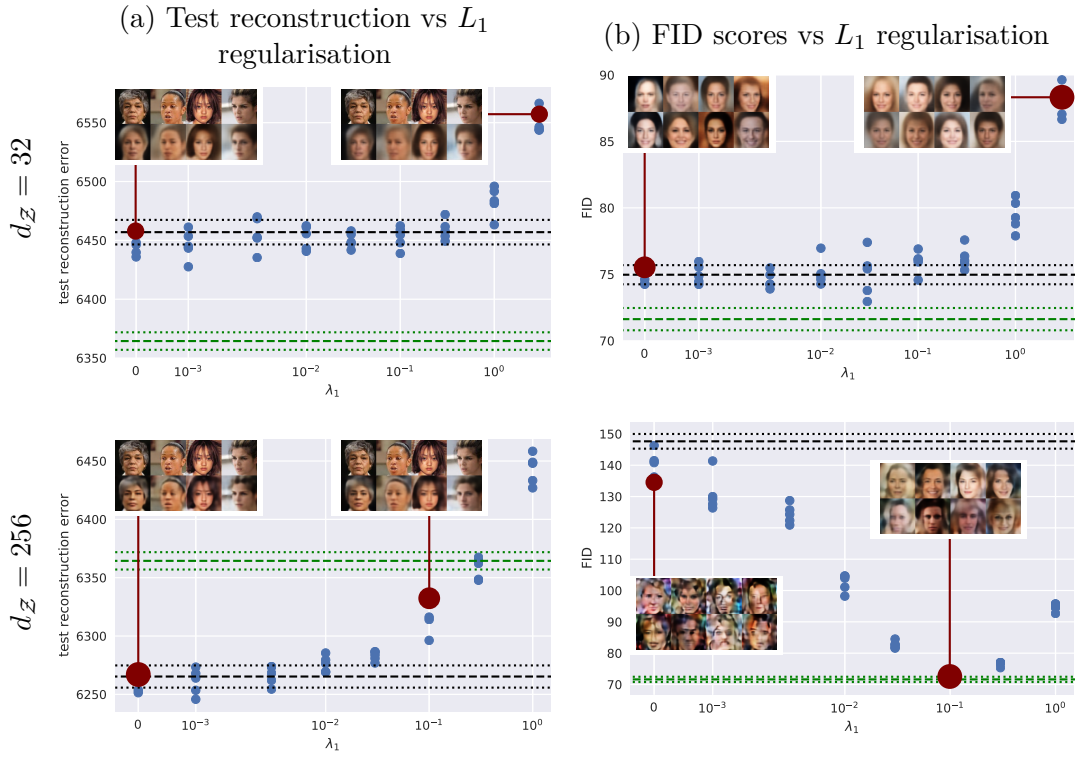


Figure 5.2 FID scores and test reconstruction errors for probabilistic-encoder WAEs with latent space dimension $d_{\mathcal{Z}} = 32$ (**first row**) and $d_{\mathcal{Z}} = 256$ (**second row**) for different L_1 regularisation coefficients λ_1. In each plot, the dashed/dotted black lines represent the mean $\pm$ s.d. for deterministic-encoder WAEs with the same $d_{\mathcal{Z}}$ (i.e. 32 or 256). The dashed/dotted green lines represent the mean $\pm$ s.d. for deterministic WAEs $d_{\mathcal{Z}} = 64$, for which the FID scores were best amongst all latent dimensions we tested. Overlaid images are (a) test reconstructions and (b) random samples coming from experiments indicated by the red circle. These plots show that when $d_{\mathcal{Z}} < d_{\mathcal{I}}$, (i) probabilistic-encoder WAEs perform comparably to deterministic WAEs and (ii) when appropriately regularised ($\lambda = 10^{-1}$), probabilistic encoders with high dimensional latent spaces can produce samples of similar quality to deterministic encoders with tuned latent dimension. At the same time, the test reconstruction errors are lower.

When $d_{\mathcal{Z}}$ is large, L_1 regularisation can significantly improve the performance of probabilistic-encoder WAEs as measured by FID scores compared to their deterministic counterparts. In particular, tuning for the best λ_1 parameter results in samples of quality comparable to deterministic encoders with the best latent dimension size, while simultaneously achieving lower test reconstruction errors. Through appropriate regularisation, probabilistic-encoder WAEs are able to adapt to large $d_{\mathcal{Z}}$ and still perform well.

When $d_{\mathcal{Z}} < d_{\mathcal{I}}$, L_1 regularisation does not improve test reconstruction error and FID scores and the probabilistic-encoder WAEs perform at best the same as deterministic-encoder WAEs. This makes sense: if in the deterministic case the WAE is already having to perform

'lossy compression' by reducing the effective dimensionality of the dataset, then the optimal probabilistic encoder cannot do better than becoming deterministic. Thus, forcing the encoder to be more random can only harm performance.

We have demonstrated a way to train probabilistic-encoder WAEs without collapsing to deterministic encoders. In doing so we have provided a way of making WAEs capable of adapting to the intrinsic data dimensionality. However, it could be argued that our approach merely substitutes the problem of searching for the 'right' latent dimensionality $d_{\mathcal{Z}}$ with the problem of searching for the 'right' regularisation λ_1. Future directions of research include exploring divergence measures other than MMD and whether the L_p regularisation coefficients λ_p can be adaptively adjusted by the learning machine itself.

5.3 Learned representation and disentanglement

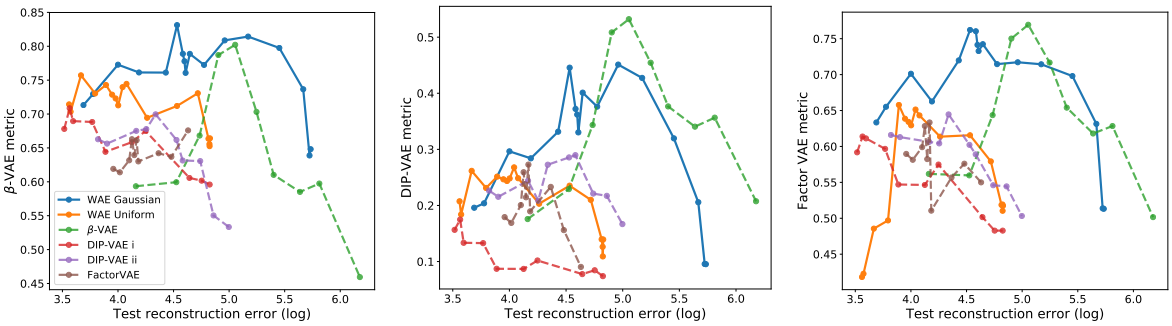
Disentangled representation learning is closely related to the more general problem of *manifold learning* for which auto-encoding architectures are often employed. The goal, though not precisely defined, is to learn representations of datasets such that individual coordinates in the feature space correspond to human-interpretable generative factors (also referred to as *factors of variation* in the literature). It is argued by Bengio et al. (2013) and Lake et al. (2017) that learning such representations is essential for significant progress in machine learning research.

Recently, Higgins et al. (2017) proposed the synthetic *dSprites* dataset and a metric to evaluate algorithms on their ability to learn disentangled representations. The dataset consists of 2-dimensional white shapes on a black background with 5 factors of variation: shape, size, rotation, x-position and y-position.

The proposed metric assumes ground truth labels for the generative factors are given. We provide here an intuition of what the metric does; see Higgins et al. (2017) for full details. Given a trained feature map $\varphi\colon \mathcal{X} \to \mathcal{Z}$ from the image space to the latent space, we ask the following question. Suppose we are given two images x_1 and x_2 which have exactly one latent factor whose value is the same—say they are both the same shape, but different in size, position and rotation. By looking at the *absolute values of the difference in feature vectors* $|\varphi(x_1) - \varphi(x_2)| \in \mathbb{R}^{d_{\mathcal{Z}}}$, is it possible to identify that it is the *shape* that they share in common, and not any other factor?

Kim and Mnih (2018) point out that the metric of Higgins et al. (2017) has a failure mode: it is possible to attain 100% accuracy with this metric without having representations of all the generative factors. They propose a modified version of this metric that does not exhibit this failure mode.

Kumar et al. (2018) propose yet another disentanglement metric, based on the linear correlation between generative factors and feature coordinates. Under their metric, a feature map scores highly if each generative factor is highly correlated with one feature coordinate, and uncorrelated with the rest.



(a) (b) (c)

Figure 5.3 Test reconstruction error against disentanglement metrics of **(a)** Higgins et al. (2017), **(b)** Kumar et al. (2018) and **(c)** Kim and Mnih (2018). Solid lines are WAE models, dashed lines are baselines. Each line shows how test reconstruction and metric scores vary as the single hyper-parameter for each model was varied (averages over 10 random seeds for each hyper-parameter setting were taken). **In all plots, up-and-left is better**. WAEs are competitive against all other methods. For Gaussian WAE, $\lambda_1 = 2.5$ attained the highest disentanglement under all metrics.

Each of these three papers additionally proposes a model based on a modification to the VAE objective. The β-VAE of Higgins et al. (2017) multiplies the KL regulariser by a scalar β. FactorVAE of Kim and Mnih (2018) augments the ELBO by adding a *Total Correlation* penalty encouraging Q_Z to be factorised, which is estimated adversarially. DIP-VAE of Kumar et al. (2018) adds a term similar to the WAE regulariser, penalising a divergence $D(Q_Z, P_Z)$. Two different choices of divergence give rise to DIP-VAE-i and DIP-VAE-ii. In all of these cases the objective function is still a lower bound on the marginal log-likelihood, provided that the additional term is non-negative.

Given these metrics, we have a well-defined task: learn representations that score highly on each metric while simultaneously achieving low test reconstruction. The additional regularisers for each model encourage disentanglement in the learned representation, but in practice increase test reconstruction error since they further constrain the auto-encoder.

Quantitative Experiments We evaluated the performance of probabilistic-encoder WAEs on the disentanglement task as measured by each of the three metrics. WAEs are flexible models for which it is possible to specify any encoding distribution and prior, as long as it is possible to sample from them. We considered two configurations, one with a Gaussian prior and encoder, the other with a uniform prior and encoder. As baselines we also evaluated the performances of the β-VAE, DIP-VAE, and FactorVAE. The results are summarised in Figure 5.3. We describe our experimental procedure here in brief; see Supplementary Materials 5.7 for further details and results.

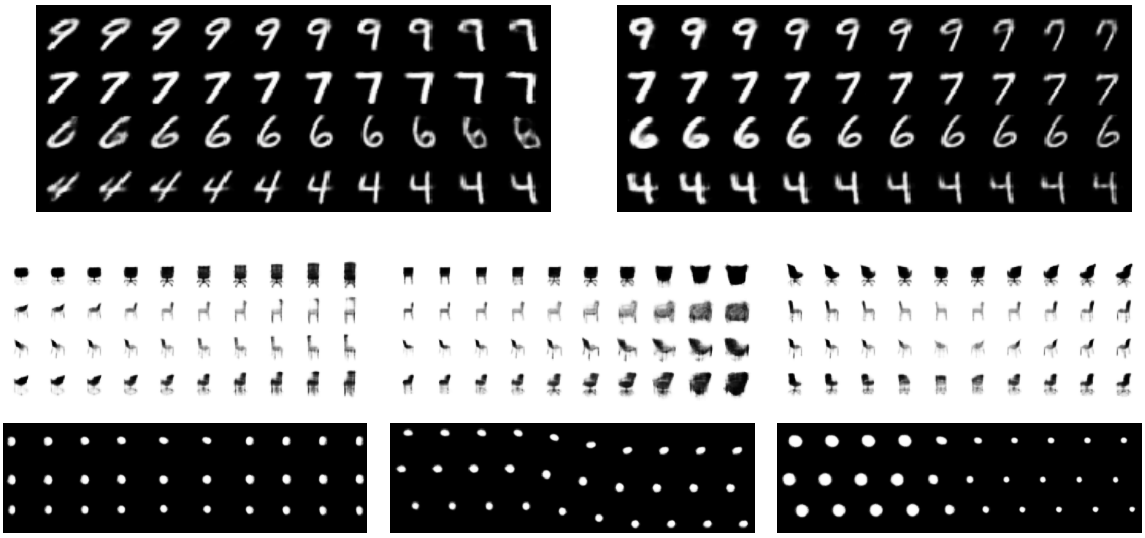


Figure 5.4 Various factors in different datasets learned by Gaussian-encoder WAEs. **Top (MNIST):** slant and thickness ($\lambda_1 = 1$); **Middle (3D Chairs):** Back style, size, azimuth ($\lambda_1 = 0.5$); **Bottom (dSprites):** X-pos., Y-pos., size ($\lambda_1 = 2.5$). Images generated by fixing one code per row and varying one coordinate from left to right within one block of images.

For all models we used the same fixed fully-connected architecture used by Higgins et al. (2017) with a Bernoulli (cross-entropy) reconstruction loss and 10 latent dimensions. All models considered have one tunable hyper-parameter[4] corresponding to the weighting of the additional regulariser. For several settings of each hyper-parameter, we trained 10 models with different random seeds. After training, we calculated the average test reconstruction error and score for each metric over these 10 runs. These averages were then plotted in Figure 5.3 (so one point in each plot corresponds to one hyper-parameter setting for one model, averaged over 10 random seeds).

The two WAE models were competitive with all other methods, able to achieve a good trade-off between test reconstruction error and disentanglement under all three metrics. In particular, the WAE with Gaussian encoder and prior performed well.

Qualitative experiments We ran probabilistic encoder WAEs on other datasets for which no ground truth generative factors are available. In such cases, no metric exists to evaluate disentanglement so this is instead commonly done by inspecting *latent traversals*. Such figures are generated by sampling a point from the latent space and then showing how the output of the decoder changes as we vary the individual coordinates of the latent code. Figure 5.3 shows latent traversals demonstrating learned factors when training on MNIST, 3D Chairs and dSprites.

[4]For WAE models we kept the weighting of the MMD term $\lambda = 400$ fixed and varied only the weighting of the L_1 regularisation λ_1

5.4 Conclusion and future directions

We investigated the problems that can arise when there is a mismatch between the dimension $d_{\mathcal{Z}}$ of the latent space of a WAE and the intrinsic dimension $d_{\mathcal{I}}$ of the dataset on which it is trained. We propose to use probabilistic encoders rather than deterministic encoders to mitigate these problems. In practice, we found that when probabilistic encoders are used, the original WAE-MMD formulation fails to train properly but that this can be resolved with additional regularisation on the variances of the encoding distributions. With this regularisation, probabilistic-encoder WAEs are able to adapt to larger latent dimensions. We applied regularised probabilistic-encoder WAEs to a benchmark disentangled representation learning task on which WAEs performed competitively against state-of-the-art baselines.

One direction for future research is to investigate whether it is possible for probabilistic-encoder WAEs to automatically adapt to $d_{\mathcal{Z}}$ without any hyper-parameter tuning. Approaches to this include deriving theoretically justified regularisation to prevent variance collapse and considering other divergence measures that take into account the encoding distribution variances. The results of our experiments on the disentanglement benchmark combined with the flexibility of the WAE framework indicate that WAEs have the potential to learn useful semantically meaningful representations of data.

Supplementary Materials

5.5 Details for Fading Squares experiment

5.5.1 Experimental details

We trained WAEs with deterministic and random encoders. The architectures we used were:

Deterministic encoder: Input $\mathbb{R}^{32\times32} \to$ FC 1200 units, ReLU $\to$ FC 1200 units, ReLU $\to$ FC $\mathbb{R}^2$

Probabilistic encoder: Input $\mathbb{R}^{32\times32} \to$ FC 1200 units, ReLU $\to$ FC 1200 units, ReLU $\to$ FC $\mathbb{R}^{2\times2}$

Decoder: $\mathbb{R}^2 \to$ FC 1200 units, tanh $\to$ FC 1200 units, tanh $\to$ FC 1200 units, tanh $\to$ FC $\mathbb{R}^{32\times32}$

The probabilistic encoder outputs the mean $\mu(X)$ and *log-side-lengths* $\log(\ell(X))$ of an axis aligned box. This parametrises a uniform distribution over an axis aligned box

$$Q(Z|X) = \text{Unif}\left(\prod_i [\mu_i(X) - \ell_i(X), \mu_i(X) + \ell_i(X)]\right)$$

We use the reparametrisation trick to allow back-propagation through sampling $z_n \sim Q(Z|x_n)$

$$\epsilon \sim \text{Unif}\left([-1,1]^{d_\mathcal{Z}}\right)$$
$$z_n = \mu(x_n) + \epsilon \odot \ell(X)$$

where $\odot$ is the element-wise product of two vectors.

Models were trained with the vanilla WAE-MMD Algorithm 2 of Tolstikhin et al. (2018a), with $\lambda = 50$, Bernoulli loss and batch size 100. We used the Adam optimizer with learning rate 10^{-3} for 40,000 iterations.

5.5.2 Preliminary results: implications for WAE-GAN

Preliminary experimentation suggests that the better quality of samples from the WAE-GAN compared to the WAE-MMD reported in Tolstikhin et al. (2018a) could be a result of the instability of the associated adversarial training. We found that when training a deterministic-encoder WAE-GAN on the *fading squares* dataset, the 1-D embedding of the data-manifold (the support of Q_Z) would move constantly through the support of P_Z throughout training without converging. This means that the decoder is trained on a much larger fraction of the total volume of P_Z compared to the WAE-MMD, for which the stability of training means that convergence to the final manifold (constituting a small fraction of P_Z) is quick.

5.5.3 Incorrect proportions of generated images

To see that the decoded images were generated in incorrect proportions, consider the mean pixel value of an image in the toy *fading squares* dataset. It is a 1-dimensional random variable, uniformly distributed on the interval $[0, 36/1024]$, where $36/1024$ is the mean (over the whole image) in the case of a white square. We trained 5 deterministic- and probabilistic-encoder WAEs, and for each one estimated the cumulative distribution function (CDF) of the mean pixel values with 100,000 sampled images. As a baseline, we also ran this procedure using 5 VAEs with the same architecture.

This is summarised in Figure 5.5, which displays the deviation from the theoretical CDF for each of the models trained. This shows that the deviation from the theoretical distribution for deterministic-encoder WAEs is consistently worse than for the probabilistic-encoder WAEs and VAEs, which fair comparably with one another. Note that while observing a uniform distribution here does not prove that images are generated with the correct frequencies, deviation from it does indeed indicate failure.

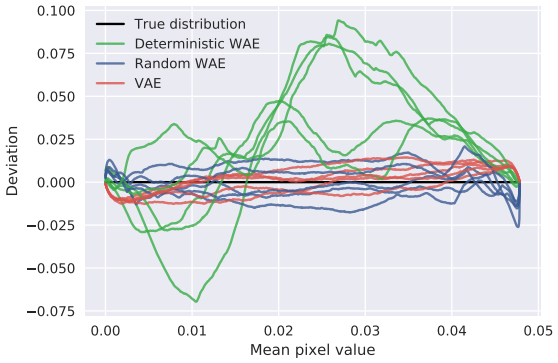


Figure 5.5 Deviation from the correct cumulative distribution of the mean pixel values for models trained on the *fading squares* dataset. If images were generated using the correct frequencies, the deviations should be close to 0. The deterministic WAE does not meet this goal.

5.6 Details for CelebA experiments

We preprocessed the CelebA dataset by centre-cropping and down-sampling so that all images are 64×64 pixels. In all experiments ran using the CelebA dataset (i.e. those summarised in Table 5.1 and Figure 5.2) we used DC-GAN architectures with 3 layers, 4×4 kernels and up to 128 filters. Specifically:

Deterministic Encoders: Input $\mathbb{R}^{64\times64\times3} \to 4 \times 4$ convolution 32 filters, stride 2, batch norm, ReLU $\to 4 \times 4$ convolution 64 filters, stride 2, batch norm, ReLU $\to 4 \times 4$ convolution 128 filters, stride 2 $\to$ FC $\mathbb{R}^{d_\mathcal{Z}}$

Probabilistic Encoders: Input $\mathbb{R}^{64\times64\times3} \to 4 \times 4$ convolution 32 filters, stride 2, batch norm, ReLU $\to 4 \times 4$ convolution 64 filters, stride 2, batch norm, ReLU $\to 4 \times 4$ convolution 128 filters, stride 2 $\to$ FC $\mathbb{R}^{d_\mathcal{Z}\times2}$

Decoders: Input $\mathbb{R}^{d_\mathcal{Z}} \to$ FC $8 \times 8 \times 128$ ReLU $\to$ transposed convolution, 64 filters, stride 2, batch norm, ReLU $\to$ transposed convolution, 32 filters, stride 2, batch norm, ReLU $\to$ transposed convolution, 3 filters, stride 2.

We used Gaussian encoders for all probabilistic encoders. The same as in a standard VAE, we used the encoder to parametrise the mean $\mu(X)$ and log-variances $\log(\Sigma(X))$ of a diagonal-constrained Gaussian. We use the reparametrisation trick to back-propagate through sampling $z \sim \mathcal{N}(\mu(X), \Sigma(X))$

$$\epsilon \sim \mathcal{N}(0, I)$$
$$z = \mu(X) + \epsilon \odot \exp(\log \Sigma(X)/2)$$

Models were trained with the vanilla WAE-MMD Algorithm 2 of Tolstikhin et al. (2018a), with the objective function modified with the addition of our new L_1 regulariser in the case of the probabilistic encoders. Additionally, we found that with little or no L_1 regularisation, the log-variances of the probabilistic encoders would become very negative and would cause numerical issues when back-propagating. To avoid these problems we clipped the log-variances to be no less than -20.

In all cases we used $\lambda = 400$, Bernoulli reconstruction loss and batch size 100. We used the Adam optimizer with learning rate 10^{-4} for 16 epochs, then 10^{-5} for a further 16 epochs (hence 32 epochs in total).

For the experiments to make Figure 5.2 we used $\lambda_1 \in \{0, 1\times10^{-3}, 3\times10^{-3}, 1\times10^{-2}, 3\times10^{-2}, 1\times10^{-1}, 3\times10^{-1}, 1, 3\}$

5.7 Details for disentanglement experiments

For all experiments in the disentanglement section (Figures 5.3 and 5.3), we used the same architecture as proposed by Higgins et al. (2017) with latent dimension $d_{\mathcal{Z}} = 10$ and Bernoulli loss.

Encoder: Input $\mathbb{R}^{64\times64}$ $\rightarrow$ FC 1200 units, ReLU $\rightarrow$ FC 1200 units, ReLU $\rightarrow$ FC $\mathbb{R}^{10\times2}$

Decoder: $\mathbb{R}^{10}$ $\rightarrow$ FC 1200 units, tanh $\rightarrow$ FC 1200 units, tanh $\rightarrow$ FC 1200 units, tanh $\rightarrow$ FC $\mathbb{R}^{64\times64}$

We downsampled the 3D Chairs images to 64×64 pixels. We padded the MNIST images to make them 32×32 pixels. For MNIST, we used the same architectures but with 32×32 inputs and outputs for the encoder and decoders respectively.

5.7.1 WAE specific details

On the dSprites task, we tested four different types of WAE.

WAE Uniform: We used a uniform prior over the box $[-1, 1]^{d_{\mathcal{Z}}}$ and parametrised the uniform distribution over an axis-aligned box with the encoder, using the reparametrisation trick as described in Appendix 5.5. Additionally we constrained the mean $\mu(X)$ to be within the set $[-1, 1]^{d_{\mathcal{Z}}}$ by adding a tanh activation.

WAE Gaussian: We used a Gaussian prior and parametrised a diagonally-constrained Gaussian with the encoder, using the reparametrisation trick as described in Appendix 5.6. Additionally we constrained the mean $\mu(X)$ to be within the set $[-1, 1]^{d_{\mathcal{Z}}}$ by adding a tanh activation.

Results from the above two models were presented in Figure 5.3. We additionally tested the following two models, which are the same except we do not constrain the means to be within the set $[-1, 1]^{d_{\mathcal{Z}}}$.

WAE Uniform (no tanh): We used a uniform prior over the box $[-1, 1]^{d_Z}$ and parametrised the uniform distribution over an axis-aligned box with the encoder, using the reparametrisation trick as described in Appendix 5.5.

WAE Gaussian (no tanh): We used a Gaussian prior and parametrised a diagonally-constrained Gaussian with the encoder, using the reparametrisation trick as described in Appendix 5.6.

For each WAE model, we fixed $\lambda = 400$ and varied λ_1. For each value of λ_1, we trained 10 models with different random seeds. All models were trained with batch size 100 with the Adam Optimiser for 30,000 iterations with learning rate 10^{-3} and a further 30,000 iterations with learning rate 10^{-4}.

For each of these 10 models per hyper-parameter setting we calculated each of the disentanglement metrics as well as the test reconstruction, and took their average. For the β-VAE metric which involves stochastically training a classifier, we evaluated the metric three times and took the maximum to be the value of the metric.

For all WAE models we searched over $\lambda_1 \in \{0, 0.1, 0.5, 1, 2, 2.5, 2.8, 3, 3.2, 3.5, 5, 8, 12, 18, 25, 40, 60\}$

5.7.2 Baseline specific details

Open source implementations were not available for FactorVAE, DIP-VAE-I or DIP-VAE-II, so we implemented these models ourselves based on the detailed descriptions given in Kim and Mnih (2018) and Kumar et al. (2018). **Our implementations will be open sourced upon publication of this work.**

FactorVAE: It should be noted that the evaluations performed by Kim and Mnih (2018) use a different architecture to that used in this paper (we use the same architectures used in Higgins et al. (2017) and Kumar et al. (2018)) and hence the disentanglement scores we obtain in this work are different to those reported by the authors themselves.

FactorVAE has a single hyper-parameter γ which is the weighting of the regulariser they introduce additional to the original VAE objective. We searched over $\gamma \in \{1, 5, 10, 20, 35, 50, 100, 200, 500, 1000\}$.

The additional regulariser of FactorVAE is adversarially estimated, and thus training involves alternating steps to optimise the auto-encoder and the adversary. We trained with batch size 100 and the Adam Optimiser for both the auto-encoder and the adversary. We trained for 30,000 iterations with learning rates 10^{-3} and 10^{-4} for the auto-encoder and adversary respectively, then a further 30,000 iterations with learning rates 10^{-4} and 10^{-5}.

DIP-VAE: The two versions of DIP-VAE involve adding an additional regulariser to the VAE objective. For both versions there are two hyper-parameters λ_d and λ_{od} (weightings of 'diagonal' and 'off-diagonal' terms in the regulariser). Kumar et al. (2018) propose that for DIP-VAE-i one should set $\lambda_d = 10\lambda_{od}$ and for DIP-VAE-ii $\lambda_d = \lambda_{od}$ for the dSprites dataset. For both DIP-VAE-i and DIP-VAE-ii we searched over $\lambda_{od} \in \{1, 2, 5, 10, 20, 50, 100, 500, 1000, 2000\}$.

We trained with batch size 100 and Adam Optimiser with learning rate 10^{-3} for 30,000 iterations and 10^{-4} for a further 30,000 iterations.

β-VAE: We searched over $\beta \in \{1, 3, 10, 20, 30, 40, 50, 75, 100, 150, 200, 300\}$.

5.7.3 A note on the disentanglement metrics

Open source implementations were not available for any of the disentanglement metrics, so we implemented them ourselves based on the detailed descriptions given in the papers. **Our implementations will be open sourced upon publication of this work.**

5.7.4 Additional results

In Figure 5.6 we summarise the disentanglement scores for all models considered.

In particular, observe that the WAE models *without* the tanh-constrained means perform rather poorly. For the Uniform WAE, this may be because the tanh makes training easier by constraining the mean to land where the prior has support. It is curious, however, that tanh-constraint improves the performance of Gaussian WAE. We suspect this is because constraining the means to lie in the interval $[-1, 1]$ forces the model to have non-zero variances in order to put probability mass outside of this interval where P_Z has support.

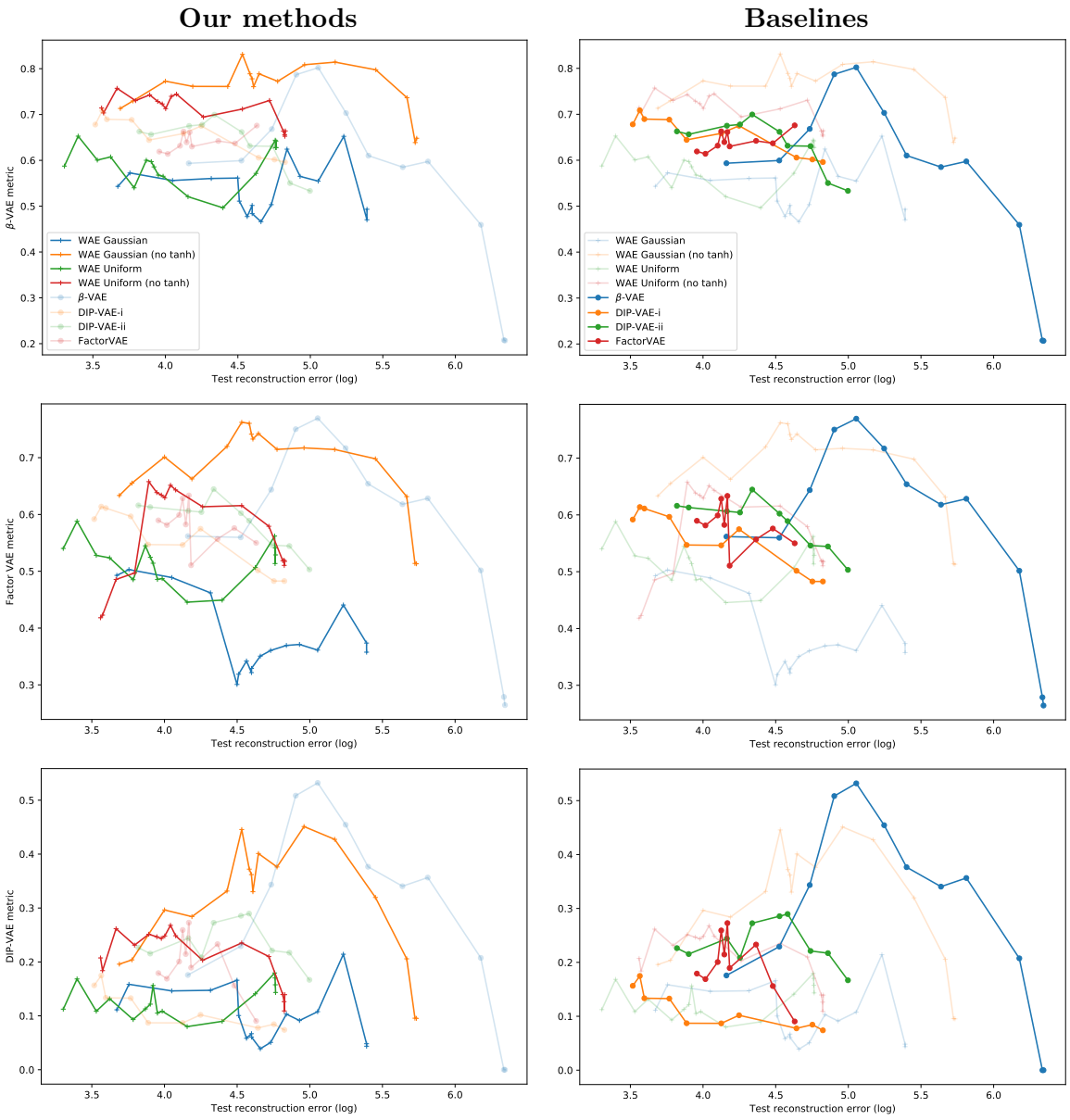


Figure 5.6 Results of disentanglement metric experiments on all evaluated models.

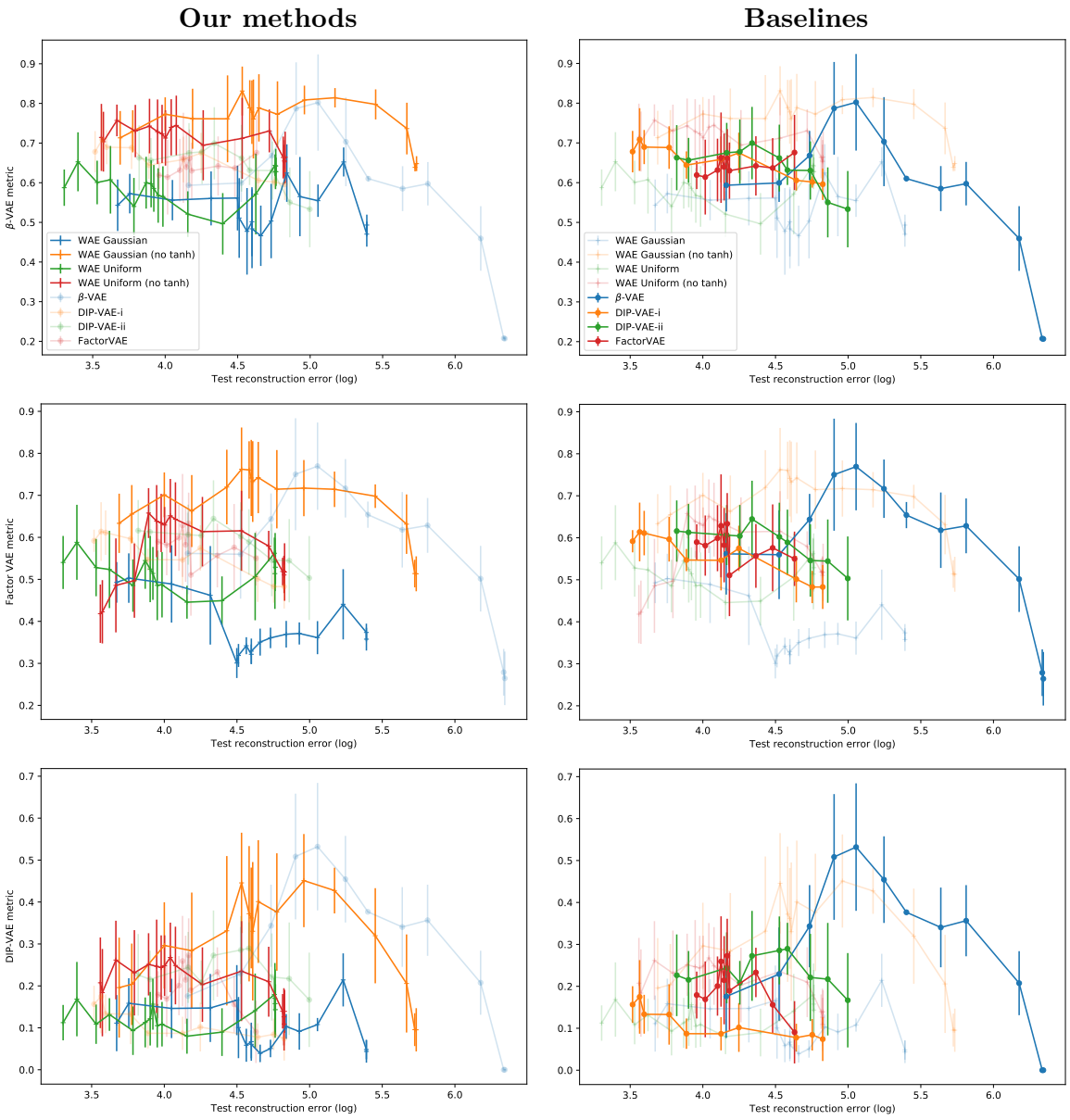


Figure 5.7 The same as Figure 5.6 but with error bars showing ± one standard deviation.

Chapter 6

Latent space learning theory

This chapter is based on the paper *Practical and Consistent Estimation of f-Divergences* published at NeurIPS 2019.

6.1 Introduction and related literature

The estimation and minimization of divergences between probability distributions based on samples are fundamental problems of machine learning. For example, maximum likelihood learning can be viewed as minimizing the Kullback-Leibler divergence $\mathrm{KL}(P_{\text{data}} \| P_{\text{model}})$ with respect to the model parameters. More generally, generative modelling—most famously Variational Autoencoders and Generative Adversarial Networks Goodfellow et al. (2014); Kingma and Welling (2013)—can be viewed as minimizing a divergence $D(P_{\text{data}} \| P_{\text{model}})$ where P_{model} may be intractable. In variational inference, an intractable posterior $p(z|x)$ is approximated with a tractable distribution $q(z)$ chosen to minimize $\mathrm{KL}(q(z) \| p(z|x))$. The mutual information between two variables $I(X, Y)$, core to information theory and Bayesian machine learning, is equivalent to $\mathrm{KL}(P_{X,Y} \| P_X P_Y)$. Independence testing often involves estimating a divergence $D(P_{X,Y} \| P_X P_Y)$, while two-sample testing (does $P = Q$?) involves estimating a divergence $D(P \| Q)$. Additionally, one approach to domain adaptation, in which a classifier is learned on a distribution P but tested on a distinct distribution Q, involves learning a feature map ϕ such that a divergence $D\left(\phi_{\#} P \| \phi_{\#} Q\right)$ is minimized, where $\phi_{\#}$ represents the push-forward operation Ben-David et al. (2007); Ganin et al. (2016).

In this work we consider the well-known family of f-divergences Csiszár et al. (2004); Liese and Vajda (2006) that includes amongst others the KL, Jensen-Shannon (JS), χ^2, and α-divergences as well as the Total Variation (TV) and squared Hellinger (H^2) distances, the latter two of which play an important role in the statistics literature B. Tsybakov (2009). A significant body of work exists studying the estimation of the f-divergence $D_f(Q \| P)$ between general probability distributions Q and P. While the majority of this focuses on α-divergences and closely related Rényi-α divergences (Krishnamurthy et al., 2014; Poczos and Schneider,

2011; Singh and Poczos, 2014), many works address specifically the KL-divergence (Perez-Cruz, 2008; Wang et al., 2009) with fewer considering f-divergences in full generality Kanamori et al. (2012); Moon and Hero (2014a,b); Nguyen et al. (2010). Although the KL-divergence is the most frequently encountered f-divergence in the machine learning literature, in recent years there has been a growing interest in other f-divergences Nowozin et al. (2016), in particular in the variational inference community where they have been employed to derive alternative evidence lower bounds Chen et al. (2018a); Dieng et al. (2017); Li and Turner (2016).

The main challenge in computing $D_f(Q\|P)$ is that it requires knowledge of either the densities of both Q and P, or the density ratio dQ/dP. In studying this problem, assumptions of differing strength can be made about P and Q. In the weakest *agnostic* setting, we may be given only a finite number of i.i.d samples from the distributions without any further knowledge about their densities. As an example of stronger assumptions, both distributions may be mixtures of Gaussians Durrieu et al. (2012); Hershey and Olsen (2007), or we may have access to samples from Q and have full knowledge of P (Hero et al., 2001, 2002) as in e.g. model fitting.

Most of the literature on f-divergence estimation considers the weaker agnostic setting. The lack of assumptions makes such work widely applicable, but comes at the cost of needing to work around estimation of either the densities of P and Q Krishnamurthy et al. (2014); Singh and Poczos (2014) or the density ratio dQ/dP (Kanamori et al., 2012; Nguyen et al., 2010) from samples. Both of these estimation problems are provably hard (B. Tsybakov, 2009; Nguyen et al., 2010) and suffer rates—the speed at which the error of an estimator decays as a function of the number of samples N—of order $N^{-1/d}$ when P and Q are defined over $\mathbb{R}^d$ unless their densities are sufficiently smooth. This is a manifestation of the *curse of dimensionality* and rates of this type are often called *nonparametric*. One could hope to estimate $D_f(P\|Q)$ without explicitly estimating the densities or their ratio and thus avoid suffering nonparametric rates, however a lower bound of the same order $N^{-1/d}$ was recently proved for α-divergences (Krishnamurthy et al., 2014), a sub-family of f-divergences. While some works considering the agnostic setting provide rates for the bias and variance of the proposed estimator Krishnamurthy et al. (2014); Nguyen et al. (2010) or even exponential tail bounds (Singh and Poczos, 2014), it is more common to only show that the estimators are asymptotically unbiased or consistent without proving specific rates of convergence Kanamori et al. (2012); Poczos and Schneider (2011); Wang et al. (2009).

Motivated by recent advances in machine learning, we study a setting in which much stronger structural assumptions are made about the distributions. Let $\mathcal{X}$ and $\mathcal{Z}$ be two finite dimensional Euclidean spaces. We estimate the divergence $D_f(Q_Z\|P_Z)$ between two probability distributions P_Z and Q_Z, both defined over $\mathcal{Z}$. P_Z has known density $p(z)$, while Q_Z with density $q(z)$ admits the factorization $q(z) := \int_{\mathcal{X}} q(z|x)q(x)dx$ where access to independent samples from the distribution Q_X with unknown density $q(x)$ and full knowledge of the conditional distribution $Q_{Z|X}$ with density $q(z|x)$ are assumed. In most cases Q_Z is intractable due to the integral and so is $D_f(Q_Z\|P_Z)$. As a concrete example, these

assumptions are often satisfied in applications of modern unsupervised generative modeling with deep autoencoder architectures, where $\mathcal{X}$ and $\mathcal{Z}$ would be *data* and *latent* spaces, P_Z the *prior*, Q_X the *data distribution*, $Q_{Z|X}$ the *encoder*, and Q_Z the *aggregate posterior*.

Given independent observations $X_1, \ldots, X_N$ from Q_X, the finite mixture $\hat{Q}_Z^N := \frac{1}{N}\sum_{i=1}^N Q_{Z|X_i}$ can be used to approximate the continuous mixture Q_Z. **Our main contribution** is to approximate the intractable $D_f(Q_Z\|P_Z)$ with $D_f(\hat{Q}_Z^N\|P_Z)$, a quantity that can be estimated to arbitrary precision using Monte-Carlo sampling since both distributions have known densities, and to theoretically study conditions under which this approximation is reasonable. We call $D_f(\hat{Q}_Z^N\|P_Z)$ the Random Mixture (RAM) estimator and derive rates at which it converges to $D_f(Q_Z\|P_Z)$ as N grows. We also provide similar guarantees for RAM-MC—a practical Monte-Carlo based version of RAM. By side-stepping the need to perform density estimation, we obtain *parametric* rates of order $N^{-\gamma}$, where γ is independent of the dimension (see Tables 6.1 and 6.2), although the constants may still in general show exponential dependence on dimension. This is in contrast to the agnostic setting where *both* nonparametric rates and constants are exponential in dimension.

Our results have immediate implications to existing literature. For the particular case of the KL divergence, a similar approach has been *heuristically* applied independently by several authors for estimating the mutual information Poole et al. (2018) and total correlation Chen et al. (2018b). Our results provide strong theoretical grounding for these existing methods by showing sufficient conditions for their consistency.

A final piece of related work is Burda et al. (2015), which proposes to reduce the gap introduced by Jensen's inequality in the derivation of the classical evidence lower bound (ELBO) by using multiple Monte-Carlo samples from the approximate posterior $Q_{Z|X}$. This is similar in flavour to our approach, but fundamentally different since we use multiple samples from the *data distribution* to reduce a different Jensen gap. To avoid confusion, we note that replacing the "regularizer" term $\mathbb{E}_X[\mathrm{KL}(Q_{Z|X}\|P_Z)]$ of the classical ELBO with expectation of our estimator $\mathbb{E}_{\mathbf{X}^N}[\mathrm{KL}(\hat{Q}_Z^N\|P_Z)]$ results in an upper bound of the classical ELBO (see Proposition 1) but is itself not in general an evidence lower bound:

$$\mathbb{E}_X\Big[\mathbb{E}_{Q_{Z|X}}\log p(X|Z) - \mathrm{KL}(Q_{Z|X}\|P_Z)\Big] \leq \mathbb{E}_X\Big[\mathbb{E}_{Q_{Z|X}}\log p(X|Z)\Big] - \mathbb{E}_{\mathbf{X}^N}\Big[\mathrm{KL}(\hat{Q}_Z^N\|P_Z)\Big].$$

The remainder of the paper is structured as follows. In Section 6.2 we introduce the RAM and RAM-MC estimators and present our main theoretical results, including rates of convergence for the bias (Theorems 26 and 27) and tail bounds (Theorems 28 and 29). In Section 6.3 we validate our results in both synthetic and real-data experiments. In Section 6.4 we discuss further applications of our results. We conclude in Section 6.5.

6.2 Random mixture estimator and convergence results

In this section we introduce our f-divergence estimator, and present theoretical guarantees for it. We assume the existence of probability distributions P_Z and Q_Z defined over $\mathcal{Z}$ with known density $p(z)$ and intractable density $q(z) = \int q(z|x)q(x)dx$ respectively, where $Q_{Z|X}$ is known. Q_X defined over $\mathcal{X}$ is unknown, however we have an i.i.d. sample $\mathbf{X}^N = \{X_1, \ldots, X_N\}$ from it. Our ultimate goal is to estimate the intractable f-divergence $D_f(Q_Z\|P_Z)$ defined by:

Definition 25 (f-divergence)**.** *Let f be a convex function on $(0, \infty)$ with $f(1) = 0$. The f-divergence D_f between distributions Q_Z and P_Z admitting densities $q(z)$ and $p(z)$ respectively is*

$$D_f(Q_Z\|P_Z) := \int f\left(\frac{q(z)}{p(z)}\right) p(z)dz.$$

Many commonly used divergences such as Kullback–Leibler and χ^2 are f-divergences. All the divergences considered in this paper together with their corresponding f can be found in Appendix 6.6. Of them, possibly the least well-known in the machine learning literature are f_β-divergences Osterreicher and Vajda (2003). These symmetric divergences are continuously parameterized by $\beta \in (0, \infty]$. Special cases include squared-Hellinger (H^2) for $\beta = \frac{1}{2}$, Jensen-Shannon (JS) for $\beta = 1$, Total Variation (TV) for $\beta = \infty$.

In our setting Q_Z is intractable and so is $D_f(Q_Z\|P_Z)$. Substituting Q_Z with a sample-based finite mixture $\hat{Q}_Z^N := \frac{1}{N}\sum_{i=1}^N Q_{Z|X_i}$ leads to our proposed **Random Mixture estimator (RAM)**:

$$D_f(\hat{Q}_Z^N\|P_Z) := D_f\Big(\tfrac{1}{N}\textstyle\sum_{i=1}^N Q_{Z|X_i}\|P_Z\Big). \tag{6.1}$$

Although $\hat{Q}_Z^N$ is a function of $\mathbf{X}^N$ we omit this dependence in notation for brevity. In this section we identify sufficient conditions under which $D_f(\hat{Q}_Z^N\|P_Z)$ is a "good" estimator of $D_f(Q_Z\|P_Z)$. More formally, we establish conditions under which the estimator is asymptotically unbiased, concentrates to its expected value and can be practically estimated using Monte-Carlo sampling.

6.2.1 Convergence rates for the bias of RAM

The following proposition shows that $D_f(\hat{Q}_Z^N\|P_Z)$ upper bounds $D_f(Q_Z\|P_Z)$ in expectation for any finite N, and that the upper bound becomes tighter with increasing N:

Proposition 1. *Let $M \leq N$ be integers. Then*

$$D_f(Q_Z\|P_Z) \;\leq \mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N\|P_Z)] \;\leq\; \mathbb{E}_{\mathbf{X}^M}[D_f(\hat{Q}_Z^M\|P_Z)]. \tag{6.2}$$

Proof sketch (full proof in Appendix 6.7.1). The first inequality follows from Jensen's inequality, using the facts that f is convex and $Q_Z = \mathbb{E}_{\mathbf{X}^N}[\hat{Q}_Z^N]$. The second holds since a sample

Table 6.1 Rate of bias $\mathbb{E}_{\mathbf{X}^N} D_f(\hat{Q}_Z^N \| P_Z) - D_f(Q_Z \| P_Z)$.

f-divergence	KL	TV	χ^2	H^2	JS	D_{f_β} $\frac{1}{2}<\beta<1$	D_{f_β} $1<\beta<\infty$	D_{f_α} $-1<\alpha<1$
Theorem 1	N^{-1}	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{4}}$	$N^{-\frac{1}{4}}$	$N^{-\frac{1}{4}}$	-
Theorem 2	$N^{-\frac{1}{3}} \log N$	$N^{-\frac{1}{2}}$	N^{-1}	$N^{-\frac{1}{5}}$	$N^{-\frac{1}{3}} \log N$	$N^{-\frac{1}{3}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{\alpha+1}{\alpha+5}}$

$\mathbf{X}^M$ can be drawn by sub-sampling (without replacement) M entries of $\mathbf{X}^N$, and by applying Jensen again. □

As a function of N, the expectation is a decreasing sequence that is bounded below. By the monotone convergence theorem, the sequence converges. Theorems 26 and 27 in this section give sufficient conditions under which the expectation of RAM converges to $D_f(Q_Z \| P_Z)$ as $N \to \infty$ for a variety of f and provide rates at which this happens, summarized in Table 6.1. The two theorems are proved using different techniques and assumptions. These assumptions, along with those of existing methods (see Table 6.3) are discussed at the end of this section.

Theorem 26 (Rates of the bias)**.** *If* $\mathbb{E}_{X\sim Q_X}[\chi^2(Q_{Z|X}, Q_Z)]$ *and* $\mathrm{KL}(Q_Z \| P_Z)$ *are finite then the bias* $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N \| P_Z)] - D_f(Q_Z \| P_Z)$ *decays with rate as given in the first row of Table 6.1.*

Proof sketch (full proof in Appendix 6.7.2). There are two key steps to the proof. The first is to bound the bias by $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N, Q_Z)]$. For the KL this is an equality. For D_{f_β} this holds because for $\beta \geq 1/2$ it is a *Hilbertian metric* and its square root satisfies the triangle inequality (Hein and Bousquet, 2005). The second step is to bound $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N, Q_Z)]$ in terms of $\mathbb{E}_{\mathbf{X}^N}[\chi^2(\hat{Q}_Z^N, Q_Z)]$, which is the variance of the average of N i.i.d. random variables and therefore decomposes as $\mathbb{E}_{X\sim Q_X}[\chi^2(Q_{Z|X}, Q_Z)]/N$. □

Theorem 27 (Rates of the bias)**.** *If* $\mathbb{E}_{X\sim Q_X, Z\sim P_Z}[q^4(Z|X)/p^4(Z)]$ *is finite then the bias* $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N \| P_Z)] - D_f(Q_Z \| P_Z)$ *decays with rate as given in the second row of Table 6.1.*

Proof sketch (full proof in Appendix 6.7.4). Denoting by $\hat{q}_N(z)$ the density of $\hat{Q}_Z^N$, the proof is based on the inequality $f(\hat{q}_N(z)/p(z)) - f(q(z)/p(z)) \leq \frac{\hat{q}_N(z)-q(z)}{p(z)} f'(\hat{q}_N(z)/p(z))$ due to convexity of f, applied to the bias. The integral of this inequality is bounded by controlling f', requiring subtle treatment when f' diverges when the density ratio $\hat{q}_N(z)/p(z)$ approaches zero. □

6.2.2 Tail bounds for RAM and practical estimation with RAM-MC

Theorems 26 and 27 describe the convergence of the *expectation* of RAM over $\mathbf{X}^N$, which in practice may be intractable. Fortunately, the following shows that RAM rapidly concentrates to its expectation.

Table 6.2 Rate $\psi(N)$ of high probability bounds for $D_f(\hat{Q}_Z^N \| P_Z)$ (Theorem 3).

f-divergence	KL	TV	χ^2	H^2	JS	D_{f_β} $\frac{1}{2}<\beta<1$	D_{f_β} $1<\beta<\infty$	D_{f_α} $\frac{1}{3}<\alpha<1$
$\psi(N)$	$N^{-\frac{1}{6}} \log N$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{6}} \log N$	$N^{-\frac{1}{6}}$	$N^{-\frac{1}{2}}$	$N^{\frac{1-3\alpha}{\alpha+5}}$

Theorem 28 (Tail bounds for RAM)**.** *Suppose that* $\chi^2\left(Q_{Z|x} \| P_Z\right) \leq C < \infty$ *for all* x *and for some constant* C*. Then, the RAM estimator* $D_f(\hat{Q}_Z^N \| P_Z)$ *concentrates to its mean in the following sense. For* $N > 8$ *and for any* $\delta > 0$*, with probability at least* $1 - \delta$ *it holds that*

$$\left| D_f(\hat{Q}_Z^N \| P_Z) - \mathbb{E}_{\mathbf{X}^N}\left[D_f(\hat{Q}_Z^N \| P_Z)\right] \right| \leq K \cdot \psi(N) \sqrt{\log(2/\delta)},$$

where K *is a constant and* $\psi(N)$ *is given in Table 6.2.*

Proof sketch (full proof in Appendix 6.7.5). These results follow by applying McDiarmid's inequality. To apply it we need to show that RAM viewed as a function of $\mathbf{X}^N$ has bounded differences. We show that when replacing $X_i \in \mathbf{X}^N$ with X_i' the value of $D_f(\hat{Q}_Z^N \| P_Z)$ changes by at most $O(N^{-1/2}\psi(N))$. Proof of this proceeds similarly to the one of Theorem 27. □

In practice it may not be possible to evaluate $D_f(\hat{Q}_Z^N \| P_Z)$ analytically. We propose to use Monte-Carlo (MC) estimation since both densities $\hat{q}_N(z)$ and $p(z)$ are assumed to be known. We consider importance sampling with proposal distribution $\pi(z|\mathbf{X}^N)$, highlighting the fact that π can depend on the sample $\mathbf{X}^N$. If $\pi(z|\mathbf{X}^N) = p(z)$ this reduces to normal MC sampling. We arrive at the **RAM-MC estimator** based on M i.i.d. samples $\mathbf{Z}^M := \{Z_1, \ldots, Z_M\}$ from $\pi(z|\mathbf{X}^N)$:

$$\hat{D}_f^M(\hat{Q}_Z^N \| P_Z) := \frac{1}{M} \sum_{m=1}^{M} f\left(\frac{\hat{q}_N(Z_m)}{p(Z_m)}\right) \frac{p(Z_m)}{\pi\left(Z_m|\mathbf{X}^N\right)}. \tag{6.3}$$

Theorem 29 (RAM-MC is unbiased and consistent)**.** $\mathbb{E}[\hat{D}_f^M(\hat{Q}_Z^N \| P_Z)] = \mathbb{E}[D_f(\hat{Q}_Z^N \| P_Z)]$ *for any proposal distribution* π*. If* $\pi(z|\mathbf{X}^N) = p(z)$ *or* $\pi(z|\mathbf{X}^N) = \hat{q}_N(z)$ *then under mild assumptions** *on the moments of* $q(Z|X)/p(Z)$ *and denoting by* $\psi(N)$ *the rate given in Table 6.2, we have*

$$Var_{\mathbf{X}^N, \mathbf{Z}^M}\left[\hat{D}_f^M(\hat{Q}_Z^N \| P_Z)\right] = O\left(M^{-1}\right) + O\left(\psi(N)^2\right).$$

Proof sketch (**full statement and proof in Appendix 6.7.6).* By the law of total variance,

$$\mathrm{Var}_{\mathbf{X}^N, \mathbf{Z}^M}\left[\hat{D}_f^M\right] = \mathbb{E}_{\mathbf{X}^N}\left[\mathrm{Var}\left[\hat{D}_f^M \,|\mathbf{X}^N\right]\right] + \mathrm{Var}_{\mathbf{X}^N}\left[D_f(\hat{Q}_Z^N \| P_Z)\right].$$

Table 6.3 Rate of bias for other estimators of $D_f(P,Q)$.

f-divergence	KL	TV	χ^2	H^2	JS	D_{f_β} $\frac{1}{2}<\beta<1$	D_{f_β} $1<\beta<\infty$	D_{f_α} $-1<\alpha<1$
Krishnamurthy et al. [22]	-	-	-	-	-	-	-	$N^{-\frac{1}{2}}+N^{\frac{-3s}{2s+d}}$
Nguyen et al. [28]	$N^{-\frac{1}{2}}$	-	-	-	-	-	-	-
Moon and Hero [26]	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$

The first of these terms is $O(M^{-1})$ by standard results on MC integration, subject to the assumptions on the moments. Using the fact that $\mathrm{Var}[Y] = \int_0^\infty \mathbb{P}(|Y - \mathbb{E}Y| > \sqrt{t})dt$ for any random variable Y we bound the second term by integrating the exponential tail bound of Theorem 28. □

Through use of the Efron-Stein inequality—rather than integrating the tail bound provided by McDiarmid's inequality—it is possible for some choices of f to weaken the assumptions under which the $O(\psi(N)^2)$ variance is achieved: from uniform boundedness of $\chi^2(Q_{Z|X}\|P_Z)$ to boundedness in expectation. In general, a variance better than $O(M^{-1})$ is not possible using importance sampling. However, the constant and hence practical performance may vary significantly depending on the choice of π. We note in passing that through Chebyshev's inequality, it is possible to derive confidence bounds for RAM-MC of the form similar to Theorem 28, but with an additional dependence on M and worse dependence on δ. For brevity we omit this.

6.2.3 Discussion: assumptions and summary

All the rates in this section are independent of the dimension of the space $\mathcal{Z}$ over which the distributions are defined. However the constants may exhibit some dependence on the dimension. Accordingly, for fixed N, the bias and variance may generally grow with the dimension.

Although the data distribution Q_X will generally be unknown, in some practical scenarios such as deep autoencoder models, P_Z may be chosen by design and $Q_{Z|X}$ learned subject to architectural constraints. In such cases, the assumptions of Theorems 27 and 28 can be satisfied by making suitable restrictions (we conjecture also for Theorem 26). For example, suppose that P_Z is $\mathcal{N}(0, I_d)$ and $Q_{Z|X}$ is $\mathcal{N}(\mu(X), \Sigma(X))$ with Σ diagonal. Then the assumptions hold if there exist constants $K, \epsilon > 0$ such that $\|\mu(X)\| < K$ and $\Sigma_{ii}(X) \in [\epsilon, 1]$ for all i (see Appendix 6.7.7). In practice, numerical stability often requires the diagonal entries of Σ to be lower bounded by a small number (e.g. 10^{-6}). If $\mathcal{X}$ is compact (as for images) then such a K is guaranteed to exist; if not, choosing K very large yields an insignificant constraint.

Table 6.3 summarizes the rates of bias for some existing methods. In contrast to our proposal, the assumptions of these estimators may in practice be difficult to verify. For the

estimator of Krishnamurthy et al. (2014), both densities p and q must belong to the Hölder class of smoothness s, be supported on $[0,1]^d$ and satisfy $0 < \eta_1 < p, q < \eta_2 < \infty$ on the support for known constants η_1, η_2. For that of Nguyen et al. (2010), the density ratio p/q must satisfy $0 < \eta_1 < p/q < \eta_2 < \infty$ and belong to a function class G whose *bracketing entropy* (a measure of the complexity of a function class) is properly bounded. The condition on the bracketing entropy is quite strong and ensures that the density ratio is well behaved. For the estimator of Moon and Hero (2014a), both p and q must have the same bounded support and satisfy $0 < \eta_1 < p, q < \eta_2 < \infty$ on the support. p and q must have *continuous bounded* derivatives of order d (which is stronger than assumptions of [22]), and f must have derivatives of order at least d.

In summary, the RAM estimator $D_f(\hat{Q}_Z^N \| P_Z)$ for $D_f(Q_Z \| P_Z)$ is **consistent** since it concentrates to its expectation $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N \| P_Z)]$, which in turn converges to $D_f(Q_Z \| P_Z)$. It is also **practical** because it can be efficiently estimated with Monte-Carlo sampling via RAM-MC.

6.3 Empirical evaluation

In the previous section we showed that our proposed estimator has a number of desirable theoretical properties. Next we demonstrate its practical performance. First, we present a synthetic experiment investigating the behaviour of RAM-MC in controlled settings where all distributions and divergences are known. Second, we investigate the use of RAM-MC in a more realistic setting to estimate a divergence between the aggregate posterior Q_Z and prior P_Z in pretrained autoencoder models. For experimental details not included in the main text, see Appendix 6.8[1].

6.3.1 Synthetic experiments

The data model. Our goal in this subsection is to test the behaviour of the RAM-MC estimator for various $d = \dim(\mathcal{Z})$ and f-divergences. We choose a setting in which Q_Z^λ parametrized by a scalar λ and P_Z are both d-variate normal distributions for $d \in \{1, 4, 16\}$. We use RAM-MC to estimate $D_f(Q_Z^\lambda, P_Z)$, which can be computed analytically for the KL, χ^2, and squared Hellinger divergences in this setting (see Appendix 6.8.1). Namely, we take P_Z and Q_X to be standard normal distributions over $\mathcal{Z} = \mathbb{R}^d$ and $\mathcal{X} = \mathbb{R}^{20}$ respectively, and $Z \sim Q_{Z|X}^\lambda$ be a linear transform of X plus a fixed isotropic Gaussian noise, with the linear function parameterized by λ. By varying λ we can interpolate between different values for $D_f(Q_Z^\lambda \| P_Z)$.

The estimators. In Figure 6.1 we show the behaviour of RAM-MC with $N \in \{1, 500\}$ and M=128 compared to the ground truth as λ is varied. The columns of Figure 6.1 correspond to different dimensions $d \in \{1, 4, 16\}$, and rows to the KL, χ^2 and H^2 divergences,

[1] A python notebook to reproduce all experiments is available at https://github.com/google-research/google-research/tree/master/f_divergence_estimation_ram_mc.

respectively. We also include two baseline methods. First, a plug-in method based on kernel density estimation Moon and Hero (2014a). Second, and only for the KL case, the M1 method of Nguyen et al. (2010) based on density ratio estimation.

The experiment. To produce each plot, the following was performed 10 times, with the mean result giving the bold lines and standard deviation giving the error bars. First, N points $\mathbf{X}^N$ were drawn from Q_X. Then M=128 points $\mathbf{Z}^M$ were drawn from $\hat{Q}_Z^N$ and RAM-MC (6.3) was evaluated. For the plug-in estimator, the densities $\hat{q}(z)$ and $\hat{p}(z)$ were estimated by kernel density estimation with 500 samples from Q_Z and P_Z respectively using the default settings of the Python library `scipy.stats.gaussian_kde`. The divergence was then estimated via MC-sampling using 128 samples from Q_Z and the surrogate densities. The M1 estimator involves solving a convex linear program in N variables to maximize a lower bound on the true divergence, see Nguyen et al. (2010) for more details. Although the M1 estimator can in principle be used for arbitrary f-divergences, its implementation requires hand-crafted derivations that are supplied only for the KL in Nguyen et al. (2010), which are the ones we use.

Discussion. The results of this experiment empirically support Proposition 1 and Theorems 26, 27, and 29: (i) in expectation, RAM-MC upper bounds the true divergence; (ii) by increasing N from 1 to 500 we clearly decrease both the bias and the variance of RAM-MC. When the dimension d increases, the bias for fixed N also increases. This is consistent with the theory in that, although the rates are independent of d, the constants are not. We note that by side-stepping the issue of density estimation, RAM-MC performs favourably compared to the plug-in and M1 estimators, more so in higher dimensions ($d = 16$). In particular, the shape of the RAM-MC curve follows that of the truth for each divergence, while that of the plug-in estimator does not for larger dimensions. In some cases the plug-in estimator can even take negative values because of the large variance.

6.3.2 Real-data experiments

The data model. To investigate the behaviour of RAM-MC in a more realistic setting, we consider Variational Autoencoders (VAEs) and Wasserstein Autoencoders (WAEs) Kingma and Welling (2013); Tolstikhin et al. (2018b). Both models involve learning an *encoder* $Q_{Z|X}^\theta$ with parameter θ mapping from high dimensional data to a lower dimensional latent space and decoder mapping in the reverse direction. A prior distribution P_Z is specified, and the optimization objectives of both models are of the form "reconstruction + distribution matching penalty". The penalty of the VAE was shown by Hoffman and Johnson (2016) to be equivalent to $\mathrm{KL}(Q_Z^\theta \| P_Z) + I(X, Z)$ where $I(X, Z)$ is the mutual information of a sample and its encoding. The WAE penalty is $D(Q_Z^\theta \| P_Z)$ for any divergence D that can practically be estimated. Following Tolstikhin et al. (2018b), we trained models using the Maximum Mean Discrepency (MMD), a kernel-based distance on distributions, and a divergence estimated using a GAN-style classifier leading to WAE-MMD and WAE-GAN respectively Goodfellow

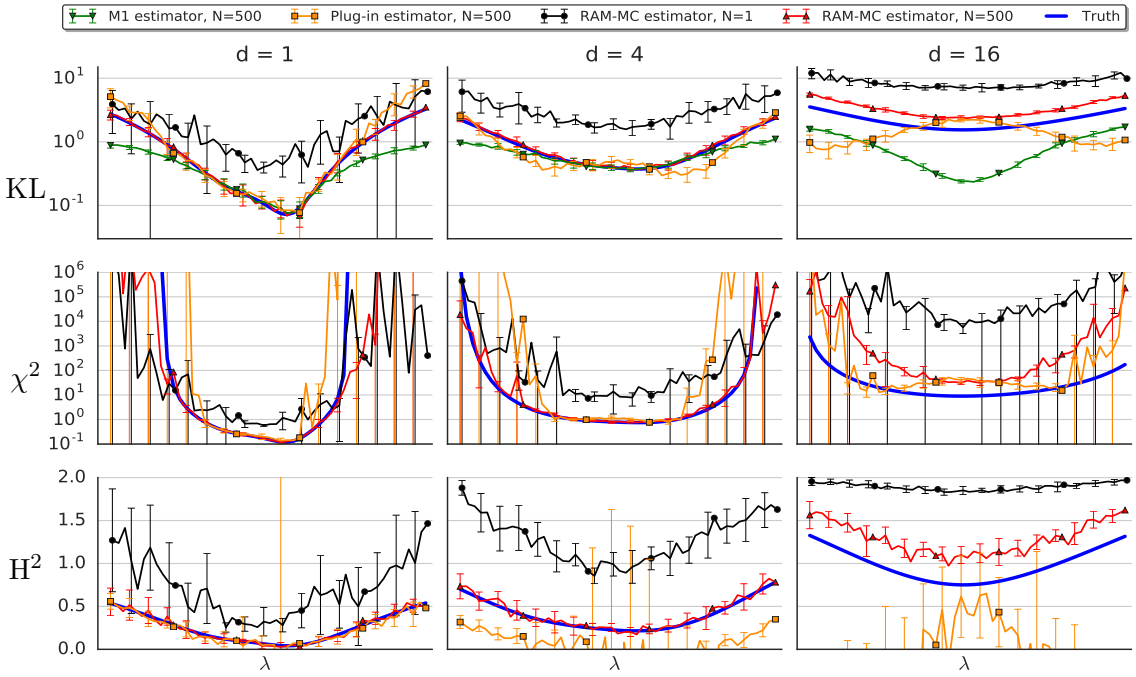


Figure 6.1 (Section 6.3.1) Estimating $D_f(\mathcal{N}(\mu_\lambda, \Sigma_\lambda), \mathcal{N}(0, I_d))$ for various f, d, and parameters μ_λ and Σ_λ indexed by $\lambda \in \mathbb{R}$. Horizontal axis correspond to $\lambda \in [-2, 2]$, columns to $d \in \{1, 4, 16\}$ and rows to KL, χ^2, and H^2 divergences respectively. **Blue** are true divergences, **black** and **red** are RAM-MC estimators (6.3) for $N \in \{1, 500\}$ respectively, **green** are M1 estimator of (Nguyen et al., 2010) and **orange** are plug-in estimates based on Gaussian kernel density estimation (Moon and Hero, 2014a). $N = 500$ and $M = 128$ in all the plots if not specified otherwise. Error bars depict one standard deviation over 10 experiments.

et al. (2014); Gretton et al. (2012). For more information about VAE and WAE, see Appendix 6.8.2.

The experiment. We consider models pre-trained on the *CelebA* dataset Liu et al. (2015), and use them to evaluate the RAM-MC estimator as follows. We take the test dataset as the ground-truth Q_X, and embed it into the latent space via the trained encoder. As a result, we obtain a $\sim$20k-component Gaussian mixture for Q_Z, the *empirical aggregate posterior*. Since Q_Z is a finite—not continuous— mixture, the true $D_f(Q_Z\|P_Z)$ can be estimated using a large number of MC samples (we used 10^4). Note that this is very costly and involves evaluating $2\cdot 10^4$ Gaussian densities for each of the 10^4 MC points. We repeated this evaluation 10 times and report means and standard deviations. RAM-MC is evaluated using $N \in \{2^0, 2^1, \ldots, 2^{14}\}$ and $M \in \{10, 10^3\}$. For each combination (N, M), RAM-MC was computed 50 times with the means plotted as bold lines and standard deviations as error bars. In Figure 6.2 we show the result of performing this for the KL divergence on six different models. For each dimension $d \in \{32, 64, 128\}$, we chose two models from the classes (VAE, WAE-MMD, WAE-GAN). See Appendix 6.8.2 for further details and similar plots for the H^2-divergence.

Discussion. The results are encouraging. In all cases RAM-MC achieves a reasonable accuracy with N relatively small, even for the bottom right model where the true KL divergence (≈ 1910) is very big. We see evidence supporting Theorem 29, which says that the variance of RAM-MC is mostly determined by the smaller of $\psi(N)$ and M: when N is small, the variance of RAM-MC does not change significantly with M, however when N is large, increasing M significantly reduces the variance. Also we found there to be two general modes of behaviour of RAM-MC across the six trained models we considered. In the bottom row of Figure 6.2 we see that the decrease in bias with N is very obvious, supporting Proposition 1 and Theorems 26 and 27. In contrast, in the top row it is less obvious, because the comparatively larger variance for M=10 dominates reductions in the bias. Even in this case, both the bias and variance of RAM-MC with M=1000 become negligible for large N. Importantly, the behaviour of RAM-MC does not degrade in higher dimensions.

The baseline estimators (plug-in Moon and Hero (2014a) and M1 Nguyen et al. (2010)) perform so poorly that we decided not to include them in the plots (doing so would distort the y-axis scale). In contrast, even with a relatively modest N=2^8 and M=1000 samples, RAM-MC behaves reasonably well in all cases.

6.4 Applications: total correlation, entropy, and mutual information estimates

In this section we describe in detail some direct consequences of our new estimator and its guarantees. Our theory may also apply to a number of machine learning domains where

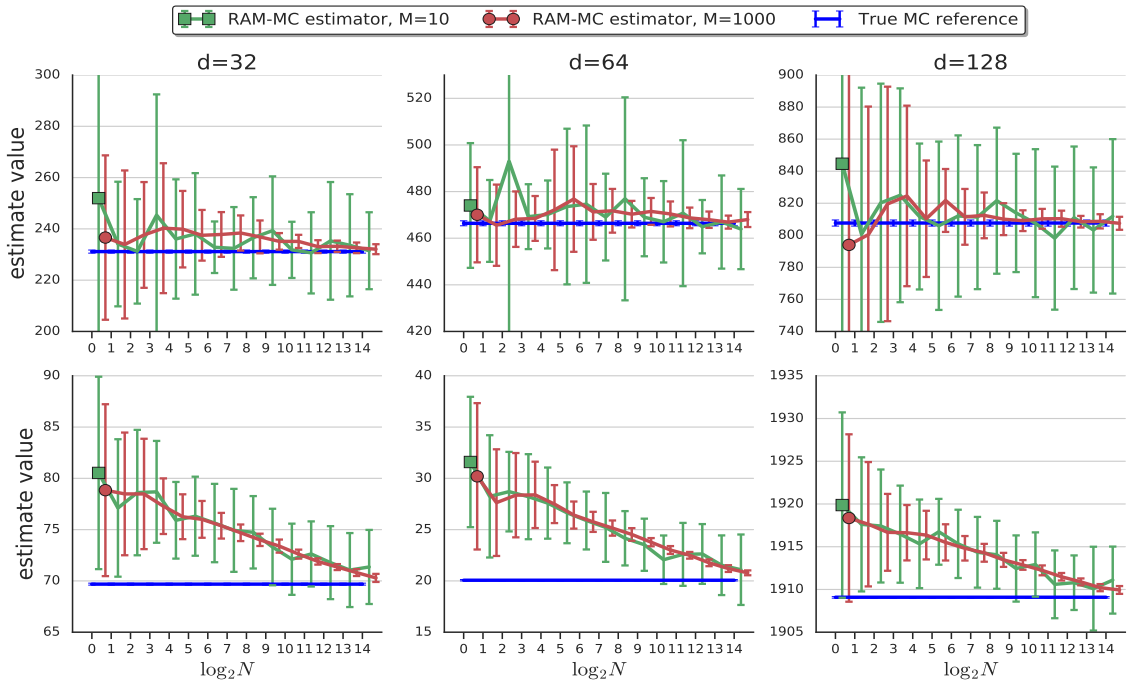


Figure 6.2 (Section 6.3.2) Estimates of $KL(Q_Z^\theta \| P_Z)$ for pretrained autoencoder models with RAM-MC as a function of N for M=10 (**green**) and M=1000 (**red**) compared to an accurate MC estimate of the ground truth (**blue**). Lines and error bars represent means and standard deviations over 50 trials.

estimating entropy, total correlation or mutual information is either the final goal or part of a broader optimization loop.

Total correlation and entropy estimation. The differential entropy, which is defined as $H(Q_Z) = -\int_{\mathcal{Z}} q(z) \log q(z) dz$, is often a quantity of interest in machine learning. While this is intractable in general, straightforward computation shows that for *any* P_Z

$$H(Q_Z) - \mathbb{E}_{\mathbf{X}^N} H(\hat{Q}_Z^N) = \mathbb{E}_{\mathbf{X}^N} \mathrm{KL}[\hat{Q}_Z^N \| P_Z] - \mathrm{KL}[Q_Z \| P_Z].$$

Therefore, our results provide sufficient conditions under which $H(\hat{Q}_Z^N)$ converges to $H(Q_Z)$ and concentrates to its mean. We now examine some consequences for Variational Autoencoders (VAEs).

Total Correlation is considered by Chen et al. (2018b), $TC(Q_Z) := \mathrm{KL}[Q_Z \| \prod_{i=1}^{d_Z} Q_{Z_i}] = \sum_{i=1}^{d_Z} H(Q_{Z_i}) - H(Q_Z)$ where Q_{Z_i} is the ith marginal of Q_Z. This is added to the VAE loss function to encourage Q_Z to be factorized, resulting in the β-TC-VAE algorithm. By the second equality above, estimation of TC can be reduced to estimation of $H(Q_Z)$ (only slight modifications are needed to treat $H(Q_{Z_i})$).

Two methods are proposed in Chen et al. (2018b) for estimating $H(Q_Z)$, both of which assume a finite dataset of size D. One of these, named *Minibatch Weighted Sample* (MWS), coincides with $H(\hat{Q}_Z^N) + \log D$ estimated with a particular form of MC sampling. Our results therefore imply *inconsistency* of the MWS method due to the constant $\log D$ offset. In the context of Chen et al. (2018b) this is not actually problematic since a constant offset does not affect gradient-based optimization techniques. Interestingly, although the derivations of Chen et al. (2018b) suppose a data distribution of finite support, our results show that minor modifications result in an estimator suitable for both finite and infinite support data distributions.

Mutual information estimation. The mutual information (MI) between variables with joint distribution $Q_{Z,X}$ is defined as $I(Z, X) := \mathrm{KL}\left[Q_{Z,X} \| Q_Z Q_X\right] = \mathbb{E}_X \mathrm{KL}\left[Q_{Z|X} \| Q_Z\right]$. Several recent papers have estimated or optimized this quantity in the context of autoencoder architectures, coinciding with our setting Alemi et al. (2018); Dieng et al. (2018); Hoffman and Johnson (2016); Oord et al. (2018). In particular, Poole et al. (2018) propose the following estimator based on replacing Q_Z with $\hat{Q}_Z^N$, proving it to be a lower bound on the true MI:

$$I_{TCPC}^N(Z, X) = \mathbb{E}_{\mathbf{X}^N}\left[\frac{1}{N} \sum_{i=1}^N \mathrm{KL}[Q_{Z|X_i} \| \hat{Q}_Z^N]\right] \leq I(Z, X).$$

The gap can be written as $I(Z, X) - I_{TCPC}^N(Z, X) = \mathbb{E}_{\mathbf{X}^N} \mathrm{KL}[\hat{Q}_Z^N \| P_Z] - \mathrm{KL}[Q_Z \| P_Z]$ where P_Z is *any* distribution. Therefore, our results also provide sufficient conditions under which I_{TCPC}^N converges and concentrates to the true mutual information.

6.5 Conclusion

We introduced a practical estimator for the f-divergence $D_f(Q_Z\|P_Z)$ where $Q_Z = \int Q_{Z|X} dQ_X$, samples from Q_X are available, and P_Z and $Q_{Z|X}$ have known density. The RAM estimator is based on approximating the true Q_Z with data samples as a random mixture via $\hat{Q}_Z^N = \frac{1}{N}\sum_n Q_{Z|X_n}$. We denote by RAM-MC the estimator version where $D_f(\hat{Q}_Z^N\|P_Z)$ is estimated with MC sampling. We proved rates of convergence and concentration for both RAM and RAM-MC, in terms of sample size N and MC samples M under a variety of choices of f. Synthetic and real-data experiments strongly support the validity of our proposal in practice, and our theoretical results provide guarantees for methods previously proposed heuristically in existing literature.

Future work will investigate the use of our proposals for optimization loops, in contrast to pure estimation. When $Q_{Z|X}^\theta$ depends on parameter θ and the goal is to minimize $D_f(Q_Z^\theta\|P_Z)$ with respect to θ, RAM-MC provides a practical surrogate loss that can be minimized using stochastic gradient methods.

Acknowledgements

Thanks to Alessandro Ialongo, Niki Kilbertus, Luigi Gresele, Giambattista Parascandolo, Mateo Rojas-Carulla and the rest of Empirical Inference group at the MPI, and Ben Poole, Sylvain Gelly, Alexander Kolesnikov and the rest of the Brain Team in Zurich for stimulating discussions, support and advice.

6.6 f for divergences considered in this paper

One of the useful properties of f-divergences that we make use of in the proofs of Theorems 27 and 28 is that for any constant c, replacing $f(x)$ by $f(x) + c(x-1)$ does not change the divergence D_f. It is often convenient to work with $f_0(x) := f(x) - f'(1)(x-1)$ which is decreasing on $(0,1)$ and increasing on $(1,\infty)$ and satisfies $f_0'(1) = 0$.

In Table 6.4 we list the forms of the function f_0 for each of the divergences considered in this paper.

6.7 Proofs

6.7.1 Proof of Proposition 1

Proposition 1. *Let $M \leq N$ be integers. Then*

$$D_f(Q_Z\|P_Z) \;\leq \mathbb{E}_{\mathbf{X}^N\sim Q_X^N} D_f(\hat{Q}_Z^N\|P_Z) \;\leq\; \mathbb{E}_{\mathbf{X}^M\sim Q_X^M} D_f(\hat{Q}_Z^M\|P_Z).$$

Table 6.4 f corresponding to divergences referenced in this paper.

f-divergence	$f_0(x)$
KL	$x \log x - x + 1$
TV	$\frac{1}{2}\lvert 1 - x \rvert$
χ^2	$x^2 - 2x$
H^2	$2(1 - \sqrt{x})$
JS	$(1+x)\log(\frac{2}{1+x}) + x \log x$
D_{f_β}, $\beta > 0$, $\beta \neq \frac{1}{2}$	$\frac{1}{1-\frac{1}{\beta}}\left[(1+x^\beta)^{\frac{1}{\beta}} - 2^{\frac{1}{\beta}-1}(1+x)\right]$
D_{f_α}, $-1 < \alpha < 1$	$\frac{4}{1-\alpha^2}\left(1 - x^{\frac{1+\alpha}{2}}\right) - \frac{2(x-1)}{\alpha-1}$

Proof. Observe that $\mathbb{E}_{\mathbf{X}^N}\hat{Q}_Z^N = Q_Z$. Thus,

$$\begin{aligned} D_f(Q_Z \| P_Z) &= \int f\left(\frac{\mathbb{E}_{\mathbf{X}^N}\hat{q}_N(z)}{p(z)}\right) dP_Z(z) \\ &\leq \mathbb{E}_{\mathbf{X}^N} \int f\left(\frac{\hat{q}_N(z)}{p(z)}\right) dP_Z(z) \\ &= \mathbb{E}_{\mathbf{X}^N \sim P_X^N} D_f(\hat{Q}_Z^N \| P_Z), \end{aligned}$$

where the inequality follows from convexity of f.

To see that $\mathbb{E}_{\mathbf{X}^N \sim P_X^N} D_f(\hat{Q}_Z^N \| P_Z) \leq \mathbb{E}_{\mathbf{X}^M \sim P_X^N} D_f(\hat{Q}_Z^M \| P_Z)$ for $N \geq M$, let $I \subseteq \{1, \ldots, N\}$, $|I| = M$ and write

$$\hat{Q}_Z^I = \frac{1}{M}\sum_{i \in I} Q_{Z|X_i}.$$

Letting I be a random subset chosen uniformly *without replacement*, observe that for any fixed I, $\mathbf{X}^I \sim \mathbb{P}_X^M$ (with the randomness coming from $\mathbf{X}^N \sim \mathbb{P}_X^N$). Thus

$$\begin{aligned}\hat{Q}_Z^N &= \frac{1}{N}\sum_{i=1}^{N} Q_{Z|X_i} \\ &= \mathbb{E}_I \frac{1}{M}\sum_{i\in I} Q_{Z|X_i} \\ &= \mathbb{E}_I \hat{Q}_Z^I\end{aligned}$$

and so again by convexity of f we have that

$$\mathbb{E}_{\mathbf{X}^N\sim P_X^N} D_f(\hat{Q}_Z^N\|P_Z) \leq \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_I D_f(\hat{Q}_Z^I\|P_Z) \tag{6.4}$$

$$= \mathbb{E}_{\mathbf{X}^M} D_f(\hat{Q}_Z^M\|P_Z) \tag{6.5}$$

with the last line following from the observation that $\mathbf{X}^I \sim \mathbb{P}_X^M$. □

6.7.2 Proof of Theorem 26

Lemma 30. *Suppose that $D_f^{\frac{1}{2}}$ satisfies the triangle inequality. Then for any $\lambda > 0$,*

$$D_f\left(\hat{Q}_Z^N\|P_Z\right) - D_f\left(Q_Z\|P_Z\right) \leq (1+\lambda) D_f\left(\hat{Q}_Z^N\|Q_Z\right) + \frac{1}{\lambda} D_f\left(Q_Z\|P_Z\right)$$

If, furthermore, $\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|Q_Z\right)\right] = O\left(\frac{1}{N^k}\right)$ for some $k > 0$, then

$$\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|P_Z\right)\right] - D_f\left(Q_Z\|P_Z\right) = O\left(\frac{1}{N^{k/2}}\right)$$

Proof. The first inequality follows from the triangle inequality for $D_f^{\frac{1}{2}}$ on $\hat{Q}_Z^N$ and P_Z, and the fact that $2\sqrt{ab} \leq \lambda a + \frac{b}{\lambda}$ for $a, b, \lambda > 0$. The second inequality follows from the first by taking $\lambda = N^{-\frac{k}{2}}$. □

* TODO: sort out theorem numbering*

Theorem 31 (Rates of the bias)**.** *If $\mathbb{E}_{X\sim Q_X}\left[\chi^2(Q_{Z|X}, Q_Z)\right]$ and $\mathrm{KL}\left(Q_Z\|P_Z\right)$ are finite then the bias $\mathbb{E}_{\mathbf{X}^N}\left[D_f(\hat{Q}_Z^N\|P_Z)\right] - D_f\left(Q_Z\|P_Z\right)$ decays with rate as given in the first row of Table 6.1.*

Proof. To begin, observe that

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}^N}\left[\chi^2(\hat{Q}_Z^N, Q_Z)\right] &= \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{Q_Z}\left[\left(\frac{\hat{q}_N(z)}{q(z)} - 1\right)^2\right] \\
&= \mathbb{E}_{Q_Z}\mathrm{Var}_{\mathbf{X}^N}\left[\frac{1}{N}\sum_{n=1}^{N}\frac{q(z|X_n)}{q(z)}\right] \\
&= \frac{1}{N}\mathbb{E}_{Q_Z}\mathrm{Var}_X\left[\frac{q(z|X)}{q(z)}\right] \\
&= \frac{1}{N}\mathbb{E}_X\left[\chi^2(Q_{Z|X}, Q_Z)\right]
\end{aligned}$$

where the introduction of the variance operator follows from the fact that $\mathbb{E}_{X_N}\left[\frac{\hat{q}_N(z)}{q(z)}\right] = 1$.

For the KL-divergence, using the fact that $\mathrm{KL} \leq \chi^2$ (Lemma 2.7 of B. Tsybakov (2009)) yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}^N}\left[\mathrm{KL}\left(\hat{Q}_Z^N \| P_Z\right)\right] - \mathrm{KL}\left(Q_Z \| P_Z\right) &= \mathbb{E}_{\mathbf{X}^N}\left[\mathrm{KL}\left(\hat{Q}_Z^N \| Q_Z\right)\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\chi^2(\hat{Q}_Z^N, Q_Z)\right] \\
&= \frac{1}{N}\mathbb{E}_X\left[\chi^2(Q_{Z|X}, Q_Z)\right] \\
&= O\left(\frac{1}{N}\right),
\end{aligned}$$

where the first equality can be verified by using the definition of KL and the fact that $Q_Z = \mathbb{E}_{\mathbf{X}^N}\hat{Q}_Z^N$.

For Total Variation, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}^N}\left[\mathrm{TV}\left(\hat{Q}_Z^N \| P_Z\right)\right] - \mathrm{TV}\left(Q_Z \| P_Z\right) &\leq \mathbb{E}_{\mathbf{X}^N}\left[\mathrm{TV}\left(\hat{Q}_Z^N \| Q_Z\right)\right] \\
&\leq \frac{1}{\sqrt{2}}\sqrt{\mathbb{E}_{\mathbf{X}^N}\left[\mathrm{KL}\left(\hat{Q}_Z^N \| Q_Z\right)\right]} \\
&= O\left(\frac{1}{\sqrt{N}}\right),
\end{aligned}$$

where the first inequality holds since TV is a metric and thus obeys the triangle inequality, and the second inequality follows by Pinsker's inequality combined with concavity of $\sqrt{x}$ (Lemma 2.5 of B. Tsybakov (2009)).

For D_{f_β} (including Jenson-Shannon) using the fact that $D_{f_\beta}^{1/2}$ satisfies the triangular inequality, we apply the second part of Lemma 30 in combination with the fact that $D_{f_\beta}\left(\hat{Q}_Z^N \| Q_Z\right) \leq \psi(\beta)\ \mathrm{TV}\left(\hat{Q}_Z^N \| Q_Z\right)$ for some scalar $\psi(\beta)$ (Theorem 2 of Osterreicher and Vajda (2003)) to obtain

$$\mathbb{E}_{\mathbf{X}^N}\left[D_{f_\beta}\left(\hat{Q}_Z^N \| P_Z\right)\right] - D_{f_\beta}\left(Q_Z \| P_Z\right) \leq O\left(\frac{1}{N^{1/4}}\right).$$

Although the squared Hellinger divergence is a member of the f_β-divergence family, we can use the tighter bound $\mathrm{H}^2\left(\hat{Q}_Z^N \| Q_Z\right) \leq KL\left(\hat{Q}_Z^N \| Q_Z\right)$ (Lemma 2.4 of B. Tsybakov (2009)) in combination with Lemma 30 to obtain

$$\mathbb{E}_{\mathbf{X}^N}\left[\mathrm{H}^2\left(\hat{Q}_Z^N \| P_Z\right)\right] - \mathrm{H}^2\left(Q_Z \| P_Z\right) \leq O\left(\frac{1}{\sqrt{N}}\right).$$

□

6.7.3 Upper bounds of f

We will make use of the following lemmas in the proof of Theorem 27 and 28.

Lemma 32. *Let $f_0(x) = x\log x - x + 1$, corresponding to $D_{f_0} = \mathrm{KL}$. Write $g(x) = f_0'^2(x) = \log^2(x)$.*

For any $0 < \delta < 1$, the function

$$h_\delta(x) := \begin{cases} g(\delta) + xg'(e) & x \in [0, e] \\ g(\delta) + eg'(e) + g(x) - g(e) & x \in [e, \infty) \end{cases}$$

is an upper bound of $g(x)$ on $[\delta, \infty)$, and is concave and non-negative on $[0, \infty)$.

Proof. First observe that h_δ is concave. It has continuous first and second derivatives:

$$h_\delta'(x) = \begin{cases} g'(e) & x \in [0, e] \\ g'(x) & x \in [e, \infty) \end{cases} \qquad h_\delta''(x) = \begin{cases} 0 & x \in [0, e] \\ g''(x) & x \in [e, \infty) \end{cases}$$

Note that $g''(x) = \frac{2}{x^2} - \frac{2\log(x)}{x^2} \leq 0$ for $x \geq e$ and $g''(e) = 0$. Therefore $h_\delta''(x)$ has non-positive second derivative on $[0, \infty)$ and is thus concave on this set.

To see that $h_\delta(x)$ is an upper bound of $g(x)$ for $x \in [\delta, \infty)$, use the fact that $g'(x) = \frac{2\log(x)}{x}$ and observe that

$$h_\delta(x) - g(x) = \begin{cases} \log^2(\delta) + \frac{2x}{e} - \log^2(x) & x \in [\delta, e] \\ \log^2(\delta) + 1 & x \in [e, \infty) \end{cases} > 0.$$

To see that $h_\delta(x)$ is non-negative on $[0, \infty)$, note that $h_\delta(x) > g(x) \geq 0$ on $[\delta, \infty)$. Moreover, $g'(e) = 2/e > 0$, and so for $x \in [0, \delta]$ we have that $h_\delta(x) = g(\delta) + 2x/e \geq g(\delta) \geq 0$. □

Lemma 33. *Let $f_0(x) = 2(1-\sqrt{x})$ corresponding to the square of the Hellinger distance. Write $g(x) = f_0'^2(x) = (1-\frac{1}{\sqrt{x}})^2$. For any $0 < \delta < 1$, the function*

$$h_\delta(x) = \frac{1}{\delta}(x-1)^2$$

is an upper bound of $g(x)$ on $[\delta, \infty)$.

Proof. For $x = 1$, we have $g(1) = h_\delta(1)$. For $x \neq 1$,

$$\begin{aligned} & 0 \leq \frac{1}{\delta}(x-1)^2 - (1-\frac{1}{\sqrt{x}})^2 \\ \iff & \sqrt{\delta} \leq \frac{x-1}{1-\frac{1}{\sqrt{x}}} \end{aligned}$$

If $x \in [\delta, 1)$ then

$$\frac{x-1}{1-\frac{1}{\sqrt{x}}} = \sqrt{x} \cdot \frac{\frac{1}{\sqrt{x}} - \sqrt{x}}{\frac{1}{\sqrt{x}} - 1} \geq \sqrt{x} \geq \sqrt{\delta}.$$

If $x \in (1, \infty)$ then

$$\frac{x-1}{1-\frac{1}{\sqrt{x}}} = \sqrt{x} \cdot \frac{\sqrt{x} - \frac{1}{\sqrt{x}}}{1-\frac{1}{\sqrt{x}}} \geq \sqrt{x} \geq \sqrt{\delta}.$$

Thus $g(x) \leq h_\delta(x)$ for $x \in [\delta, \infty)$. □

Lemma 34. *Let $f_0(x) = \frac{4}{1-\alpha^2}\left(1 - x^{\frac{1+\alpha}{2}}\right) - \frac{2(x-1)}{\alpha-1}$ corresponding to the α-divergence with $\alpha \in (-1, 1)$. Write $g(x) = f_0'^2(x) = \frac{4}{(\alpha-1)^2}\left(x^{\frac{\alpha-1}{2}} - 1\right)^2$. For any $0 < \delta < 1$, the function*

$$h_\delta(x) = \frac{4\left(\delta^{\frac{\alpha-1}{2}} - 1\right)^2}{(\alpha-1)^2(\delta-1)^2} \cdot (x-1)^2$$

is an upper bound of $g(x)$ on $[\delta, \infty)$.

Proof. For $x = 1$, we have $g(1) = h_\delta(1)$. Consider now the case that $x \geq \delta$ and $x \neq 1$. Since $0 < \delta < 1$, we have that $1 - \delta > 0$. And because $(\alpha-1)/2 \in (-1, 0)$, we have that $\delta^{\frac{\alpha-1}{2}} - 1 > 0$. It follows by taking square roots that

$$\begin{aligned} & g(x) \leq h_\delta(x) \\ \iff & d(x) := \frac{x^{\frac{\alpha-1}{2}} - 1}{1-x} \leq \frac{\delta^{\frac{\alpha-1}{2}} - 1}{1-\delta} \end{aligned}$$

Now, $d(x)$ is non-increasing for $x > 0$. Indeed,

$$d'(x) = \frac{-1}{(1-x)^2}\left[1 - \frac{3-\alpha}{2}x^{\frac{\alpha-1}{2}} + \frac{1-\alpha}{2}x^{\frac{\alpha-3}{2}}\right]$$

and it can be shown by differentiating that the term inside the square brackets attains its minimum at $x = 1$ and is therefore non-negative. Since $(1-x)^2 \geq 0$ it follows that $d'(x) \leq 0$ and so $d(x)$ is non-increasing. From this fact it follows that $d(x)$ attains its maximum on $x \in [\delta, \infty)$ at $x = \delta$, and thus the desired inequality holds. □

Lemma 35. *Let* $f_0(x) = (1+x)\log 2 + x\log x - (1+x)\log(1+x)$ *corresponding to the Jensen-Shannon divergence. Write* $g(x) = f_0'^2(x) = \log^2 2 + \log^2\left(\frac{x}{1+x}\right) + 2\log 2 \log\left(\frac{x}{1+x}\right)$. *For* $0 < \delta < 1$, *the function*

$$h_\delta(x) = g(\delta) + 4\log^2 2$$

is an upper bound of $g(x)$ *on* $[\delta, \infty)$.

Proof. For $x \geq 1$, $\frac{x}{x+1} \in [0.5, 1)$ and so $\log\left(\frac{x}{1+x}\right) \in [-\log 2, 0)$. Therefore $g(x) \in \left(0, 4\log^2 2\right]$ for $x > 1$. It follows that for any value of δ, $h_\delta(x) \geq g(x)$ for $x \geq 1$. $f_0'(1) = 0$ and by differentiating again it can be shown that $f_0''(x) > 0$ for $x \in (0,1)$. Thus $f_0'(x) < 0$ and is increasing on $(0,1)$ and so $g(x) > 0$ and is decreasing on $(0,1)$. Thus $h_\delta(x) > g(\delta) \geq g(x)$ for $x \in [\delta, 1)$. □

Lemma 36. *Let* $f_0(x) = \frac{1}{1-\frac{1}{\beta}}\left[\left(1+x^\beta\right)^{\frac{1}{\beta}} - 2^{\frac{1}{\beta}-1}(1+x)\right]$ *corresponding to the* f_β*-divergence introduced in Osterreicher and Vajda (2003). We assume* $\beta \in \left(\frac{1}{2}, \infty\right) \setminus \{1\}$. *Write* $g(x) = f_0'^2(x) = \left(\frac{\beta}{1-\beta}\right)^2\left[\left(1+x^{-\beta}\right)^{\frac{1-\beta}{\beta}} - 2^{\frac{1}{\beta}-1}\right]^2$.

If $\beta \in \left(\frac{1}{2}, 1\right)$, *then* $\lim_{x\to\infty} g(x)$ *exists and is finite and for any* $0 < \delta < 1$, *we have that* $h_\delta(x) := g(\delta) + \lim_{x\to\infty} g(x) \geq g(x)$ *for all* $x \in [\delta, \infty)$.

If $\beta \in (1, \infty)$, *then* $\lim_{x\to 0} g(x)$ *and* $\lim_{x\to\infty} g(x)$ *both exist and are finite, and* $g(x) \leq \max\{\lim_{x\to 0} g(x), \lim_{x\to\infty} g(x)\}$ *for all* $x \in [0, \infty)$.

Proof. For any $\beta \in \left(\frac{1}{2}, \infty\right) \setminus \{1\}$, we have that $f_0''(x) = \frac{\beta}{(1-\beta)^2}\left[\frac{1}{x^{\beta+1}}\left(1+x^{-\beta}\right)^{\frac{1-2\beta}{\beta}}\right] > 0$ for $x > 0$. Since $f_0'(1) = 0$, it follows that $f_0'(x)$ is increasing everywhere, negative on $(0,1)$ and positive on $(1,\infty)$. It follows that $g(x)$ is decreasing on $(0,1)$ and increasing on $(1,\infty)$. $\beta > 0$ means that $1 + x^{-\beta} \to 1$ as $x \to \infty$. Hence $g(x)$ is bounded above and increasing in x, thus $\lim_{x\to\infty} g(x)$ exists and is finite.

For $\beta \in (\frac{1}{2}, 1)$, $\frac{1-\beta}{\beta} > 0$. It follows that $\left(1+x^{-\beta}\right)^{\frac{1-\beta}{\beta}}$ grows unboundedly as $x \to 0$, and hence so does $g(x)$. Since $g(x)$ is decreasing on $(0,1)$, for any $0 < \delta < 1$ we have that $h_\delta(x) \geq$

$g(x)$ on $(0,1)$. Since $g(x)$ is increasing on $(1,\infty)$ we have that $h_\delta(x) \geq \lim_{x\to\infty} g(x) \geq g(x)$ on $(1,\infty)$.

For $\beta \in (1,\infty)$, $\frac{1-\beta}{\beta} < 0$. It follows that $\left(1+x^{-\beta}\right)^{\frac{1-\beta}{\beta}} \to 0$ as $x \to 0$, and hence $\lim_{x\to 0} g(x)$ exists and is finite. Since $g(x)$ is decreasing on $(0,1)$ and increasing on $(1,\infty)$, it follows that $g(x) \leq \max\{\lim_{x\to 0} g(x), \lim_{x\to\infty} g(x)\}$ for all $x \in [0,\infty)$

□

6.7.4 Proof of Theorem 27

Theorem 37 (Rates of the bias)**.** *If* $\mathbb{E}_{X\sim Q_X, Z\sim P_Z}\left[q^4(Z|X)/p^4(Z)\right]$ *is finite then the bias* $\mathbb{E}_{\mathbf{X}^N}\left[D_f(\hat{Q}_Z^N\|P_Z)\right] - D_f\left(Q_Z\|P_Z\right)$ *decays with rate as given in the second row of Table 6.1.*

Proof. For each f-divergence we will work with the function f_0 which is decreasing on $(0,1)$ and increasing on $(1,\infty)$ with $D_f = D_{f_0}$ (see Appendix 6.6).

For shorthand we will sometimes use the notation $\|q(z|X)/p(z)\|^2_{L_2(P_Z)} = \int \frac{q(z|X)^2}{p(z)^2} p(z)dz$ and $\|q^2(z|X)/p^2(z)\|^2_{L_2(P_Z)} = \int \frac{q(z|X)^4}{p(z)^4} p(z)dz$.

We will denote $C := \mathbb{E}_{X\sim Q_X, Z\sim P_Z}\left[q^4(Z|X)/p^4(Z)\right]$ which is finite by assumption. This implies that the second moment $B := \mathbb{E}_{X\sim Q_X, Z\sim P_Z}\left[q^2(Z|X)/p^2(Z)\right]$ is also finite, thanks to Jensen's inequality:

$$\mathbb{E}[Y^2] = \mathbb{E}[\sqrt{Y^4}] \leq \sqrt{\mathbb{E}[Y^4]}.$$

The case that D_f is the χ^2-divergence: In this case, using $f(x) = x^2 - 1$, it can be seen that the bias is equal to

$$\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|P_Z\right)\right] - D_f\left(Q_Z\|P_Z\right) = \mathbb{E}_{\mathbf{X}^N}\left[\int_Z \left(\frac{\hat{q}_N(z) - q(z)}{p(z)}\right)^2 dP(z)\right]. \tag{6.6}$$

Indeed, expanding the right hand side and using the fact that $\mathbb{E}_{\mathbf{X}^N}\hat{q}_N(z) = q(z)$ yields

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[\int_Z \frac{\hat{q}_N^2(z) - 2\hat{q}_N(z)q(z) + q^2(z)}{p^2(z)} dP(z)\right] \\
&= \mathbb{E}_{\mathbf{X}^N}\left[\int_Z \frac{\hat{q}_N^2(z) - q^2(z)}{p^2(z)} dP(z)\right] \\
&= \mathbb{E}_{\mathbf{X}^N}\left[\int_Z \left(\frac{\hat{q}_N^2(z)}{p^2(z)} - 1\right) dP(z)\right] - \int_Z \left(\frac{q^2(z)}{p^2(z)} - 1\right) dP(z) \\
&= \mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|P_Z\right)\right] - D_f\left(Q_Z\|P_Z\right).
\end{aligned}$$

Again using the fact that $\mathbb{E}_{\mathbf{X}^N}\hat{q}_N(z) = q(z)$, observe that taking expectations over $\mathbf{X}^N$ in the right hand size of Equation 6.6 above (after changing the order of integration) can be viewed as taking the variance of $\hat{q}_N(z)/p(z)$, the average of N i.i.d. random variables, and so

$$\begin{aligned}\mathbb{E}_{\mathbf{X}^N}\left[\int_Z \left(\frac{\hat{q}_N(z)-q(z)}{p(z)}\right)^2 dP(z)\right] &= \int_Z \mathbb{E}_{\mathbf{X}^N}\left[\left(\frac{\hat{q}_N(z)-q(z)}{p(z)}\right)^2\right] dP(z)\\ &= \frac{1}{N}\int_Z \mathbb{E}_X\left[\left(\frac{q(z|X)-q(z)}{p(z)}\right)^2\right] dP(z)\\ &= \frac{1}{N}\mathbb{E}_X\chi^2\left(Q_{Z|X}\|P_Z\right) - \frac{1}{N}\chi^2\left(Q_Z\|P_Z\right)\\ &\leq \frac{B-1}{N}.\end{aligned}$$

The case that D_f is the Total Variation distance or D_{f_β} with $\beta > 1$: For these divergences, we only need the condition that the second moment $\mathbb{E}_X\|q(z|X)/p(z)\|^2_{L_2(P_Z)} < \infty$ is bounded.

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}^N}\left[D_{f_0}\left(\hat{Q}_Z^N\|P_Z\right)\right] - D_{f_0}\left(Q_Z\|P_Z\right)\\ &= \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f_0\left(\frac{q(z)}{p(z)}\right)\right]\\ &\leq \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)-q(z)}{p(z)}\right) f_0'\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right]\\ &\leq \underbrace{\sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)-q(z)}{p(z)}\right)^2\right]}}_{(i)} \times \underbrace{\sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right]\right)}}_{(ii)}\end{aligned}$$

where the first inequality holds due to convexity of f_0 and the second inequality follows by Cauchy-Schwartz. Then,

$$\begin{aligned}(i)^2 &= \mathbb{E}_{P_Z}\mathrm{Var}_{\mathbf{X}^N}\left[\frac{\hat{q}_N(z)}{p(z)}\right]\\ &= \frac{1}{N}\mathbb{E}_{P_Z}\mathrm{Var}_X\left[\frac{q(z|X)}{p(z)}\right]\\ &\leq \frac{1}{N}\mathbb{E}_X\mathbb{E}_{P_Z}\left[\frac{q^2(z|X)}{p^2(z)}\right] = \frac{1}{N}\mathbb{E}_X\left\|\frac{q(z|X)}{p(z)}\right\|^2_{L_2(P_Z)}\\ \implies (i) &= O\left(\frac{1}{\sqrt{N}}\right).\end{aligned}$$

For Total Variation, $f_0'^2(x) \leq 1$, so

$$(ii)^2 \leq 1.$$

For D_{f_β} with $\beta > 1$, Lemma 36 shows that $f_0'^2(x) \leq \max\{\lim_{x\to 0} f_0'^2(x), \lim_{x\to\infty} f_0'^2(x)\} < \infty$ and so

$$(ii)^2 = O(1).$$

Thus, for both cases considered,

$$\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right] - D_f\left(Q_Z \| P_Z\right) \leq O\left(\frac{1}{\sqrt{N}}\right).$$

All other divergences. We start by writing the difference as the sum of integrals over mutually exclusive events that partition $\mathcal{Z}$. Denoting by γ_N and δ_N scalars depending on N, write

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right] - D_f\left(Q_Z \| P_Z\right) \\
&= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f_0\left(\frac{q(z)}{p(z)}\right) dP_Z(z)\right] \\
&= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f_0\left(\frac{q(z)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} \leq \delta_N \text{ and } \frac{q(z)}{p(z)} \leq \gamma_N\right\}} dP_Z(z)\right] \qquad \text{(A)} \\
&\quad + \mathbb{E}_{\mathbf{X}^N}\left[\int f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f_0\left(\frac{q(z)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} \leq \delta_N \text{ and } \frac{q(z)}{p(z)} > \gamma_N\right\}} dP_Z(z)\right] \qquad \text{(B)} \\
&\quad + \mathbb{E}_{\mathbf{X}^N}\left[\int f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f_0\left(\frac{q(z)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} dP_Z(z)\right]. \qquad \text{(C)}
\end{aligned}$$

Consider each of the terms (A), (B) and (C) separately.

Later on, we will pick $\delta_N < \gamma_N$ to be decreasing in N. In the worst case, $N > 8$ will be sufficient to ensure that $\gamma_N < 1$, so in the remainder of this proof we will assume that $\delta_N, \gamma_N < 1$.

(A): Recall that $f_0(x)$ is decreasing on the interval $[0, 1]$. Since $\gamma_N, \delta_N \leq 1$, the integrand is at most $f_0(0) - f_0(\gamma_N)$, and so

$$\text{(A)} \leq f_0(0) - f_0(\gamma_N).$$

(B): The integrand is bounded above by $f_0(0)$ since $\delta_N < 1$, and so

$$\text{(B)} \leq f_0(0) \times \mathbb{P}_{Z,\mathbf{X}^N} \underbrace{\left\{\frac{\hat{q}_N(z)}{p(z)} \leq \delta_N \text{ and } \frac{q(z)}{p(z)} > \gamma_N\right\}}_{\text{(*)}}.$$

We will upper bound $\mathbb{P}_{Z,\mathbf{X}^N}(*)$: observe that if $\gamma_N > \delta_N$, then $(*) \implies \left|\frac{\hat{q}_N(z)-q(z)}{p(z)}\right| \geq \gamma_N - \delta_N$. It thus follows that

$$\begin{aligned}
\mathbb{P}_{Z,\mathbf{X}^N}(*) &\leq \mathbb{P}_{Z,\mathbf{X}^N}\left\{\left|\frac{\hat{q}_N(z)-q(z)}{p(z)}\right| \geq \gamma_N - \delta_N\right\} \\
&= \mathbb{E}_Z\left[\mathbb{P}_{\mathbf{X}^N}\left\{\left|\frac{\hat{q}_N(z)-q(z)}{p(z)}\right| \geq \gamma_N - \delta_N \mid Z\right\}\right] \\
&\leq \mathbb{E}_Z\left[\frac{\mathrm{Var}_{\mathbf{X}^N}\left[\frac{\hat{q}_N(z)}{p(z)}\right]}{(\gamma_N-\delta_N)^2}\right] \\
&= \frac{1}{N(\gamma_N-\delta_N)^2}\mathbb{E}_Z\left[\mathbb{E}_X\left[\frac{q^2(z|X)}{p^2(z)}\right] - \frac{q^2(z)}{p^2(z)}\right] \\
&\leq \frac{1}{N(\gamma_N-\delta_N)^2}\mathbb{E}_Z\mathbb{E}_X\left[\frac{q^2(z|X)}{p^2(z)}\right] \\
&\leq \frac{\sqrt{C}}{N(\gamma_N-\delta_N)^2}.
\end{aligned}$$

The second inequality follows by Chebyshev's inequality, noting that $\mathbb{E}_{\mathbf{X}^N}\frac{\hat{q}_N(z)}{p(z)} = \frac{q(z)}{p(z)}$. The penultimate inequality is due to dropping a negative term. The final inequality is due to the boundedness assumption $C = \mathbb{E}_X\left\|\frac{q^2(z|X)}{p^2(z)}\right\|^2_{L_2(P_Z)}$. We thus have that

$$(\mathrm{B}) \leq f_0(0)\frac{\sqrt{C}}{N(\gamma_N-\delta_N)^2}.$$

(C): Bounding this term will involve two computations, one of which (††) will be treated separately for each divergence we consider.

$$\begin{aligned}
(\mathrm{C}) &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f_0\left(\frac{q(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} dP_Z(z)\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int \left(\frac{\hat{q}_N(z)}{p(z)} - \frac{q(z)}{p(z)}\right) f_0'\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} dP_Z(z)\right] && \text{(Convexity of } f\text{)} \\
&\leq \underbrace{\sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_Z\left[\left(\frac{\hat{q}_N(z)}{p(z)} - \frac{q(z)}{p(z)}\right)^2\right]}}_{(\dagger)} \times \underbrace{\sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_Z\left[f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}}\right]}}_{(\dagger\dagger)} && \text{(Cauchy-Schwartz)}
\end{aligned}$$

Noting that $\mathbb{E}_X \frac{q(z|X)}{p(z)} = \frac{q(z)}{p(z)}$, we have that

$$\begin{aligned}(\dagger)^2 &= \mathbb{E}_Z \mathrm{Var}_{\mathbf{X}^N}\left[\frac{\hat{q}_N(z)}{p(z)}\right] \\ &= \frac{1}{N}\mathbb{E}_Z \mathrm{Var}_X\left[\frac{q(z|X)}{p(z)}\right] \\ &\leq \frac{1}{N}\mathbb{E}_X \left\|\frac{q(z|X)}{p(z)}\right\|^2_{L_2(P_Z)} \\ \implies (\dagger) &\leq \frac{\sqrt{B}}{\sqrt{N}}\end{aligned}$$

where $\sqrt{B} = \sqrt{\mathbb{E}_X \left\|\frac{q(z|X)}{p(z)}\right\|^2_{L_2(P_Z)}}$ is finite by assumption.

Term (††) will be bounded differently for each divergence, though using a similar pattern. The idea is to use the results of Lemmas 32-36 in order to upper bound $f_0'^2(x)$ with something that can be easily integrated.

KL. By Lemma 32, there exists a function $h_{\delta_N}(x)$ that is positive and concave on $[0, \infty)$ and is an upper bound of $f_0'^2(x)$ on $[\delta_N, \infty)$ with $h_{\delta_N}(1) = \log^2(\delta_N) + \frac{2}{e}$.

$$\begin{aligned}(\dagger\dagger)^2 &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] \\ &\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] && (h_{\delta_N} \text{ upper bounds } f'^2 \text{ on } (\delta_N, \infty)) \\ &\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right) p(z)dz\right] && (h_{\delta_N} \text{ non-negative on } [0, \infty)) \\ &\leq \mathbb{E}_{\mathbf{X}^N}\left[h_{\delta_N}\left(\int \frac{\hat{q}_N(z)}{p(z)} p(z)dz\right)\right] && (h_{\delta_N} \text{ concave}) \\ &= h_{\delta_N}(1) \\ &= \log^2(\delta_N) + \frac{2}{e} \\ \implies (\dagger\dagger) &= \sqrt{\log^2(\delta_N) + \frac{2}{e}}.\end{aligned}$$

Therefore,

$$\textcircled{C} \leq \sqrt{B}\sqrt{\frac{\log^2(\delta_N) + \frac{2}{e}}{N}}.$$

Putting everything together,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right] - D_f\left(Q_Z \| P_Z\right) \\
&\leq \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0(\gamma_N) + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \sqrt{B}\sqrt{\frac{\log^2(\delta_N) + \frac{2}{e}}{N}} \\
&= \gamma_N - \gamma_N \log \gamma_N + \frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \sqrt{B}\sqrt{\frac{\log^2(\delta_N) + \frac{2}{e}}{N}}.
\end{aligned}
$$

Taking $\delta_N = \frac{1}{N^{1/3}}$ and $\gamma_N = \frac{2}{N^{1/3}}$:

$$
\begin{aligned}
&= \frac{2}{N^{1/3}} - \frac{2}{N^{1/3}} \log\left(\frac{2}{N^{1/3}}\right) + \frac{\sqrt{C}}{N \cdot \frac{1}{N^{2/3}}} + \sqrt{B}\sqrt{\frac{\log^2\left(\frac{1}{N^{1/3}}\right) + \frac{2}{e}}{N}} \\
&= \frac{2 - 2\log 2}{N^{1/3}} + \frac{2}{3}\frac{\log N}{N^{1/3}} + \frac{\sqrt{C}}{N^{1/3}} + \sqrt{B}\sqrt{\frac{\frac{1}{4}\log^2(N) + \frac{2}{e}}{N}} \\
&= O\left(\frac{\log N}{N^{1/3}}\right)
\end{aligned}
$$

Squared-Hellinger. Lemma 33 provides a function h_δ that upper bounds $f'^2(x)$ for $x \in\in [\delta, \infty)$.

$$
\begin{aligned}
(\dagger\dagger)^2 &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] && (h_{\delta_N} \text{ upper bounds } f_0'^2 \text{ on } (\delta_N, \infty)) \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right) p(z)dz\right] && (h_{\delta_N} \text{ non-negative on } [0, \infty)) \\
&= \frac{1}{\delta_N}\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)}{p(z)} - 1\right)^2\right] \\
&\leq \frac{1}{\delta_N}\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)}{p(z)}\right)^2 + 1\right] \\
&= \frac{1}{\delta_N} + \frac{1}{\delta_N}\mathbb{E}_{\mathbf{X}^N}\left[\left\|\frac{\hat{q}_N(z)}{p(z)}\right\|_{L_2(P_Z)}^2\right] \\
&\leq \frac{B+1}{\delta_N} \\
\implies (\dagger\dagger) &= \frac{\sqrt{B+1}}{\sqrt{\delta_N}}.
\end{aligned}
$$

and thus

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right] - D_f\left(Q_Z \| P_Z\right) \\
&\leq \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0(\gamma_N) + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \frac{\sqrt{B}\sqrt{B+1}}{\sqrt{N\delta_N}} \\
&= 2\sqrt{\gamma_N} + \frac{2\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \frac{\sqrt{B}\sqrt{B+1}}{\sqrt{N\delta_N}}.
\end{aligned}$$

Setting $\gamma_N = \frac{2}{N^{2/5}}$ and $\delta_N = \frac{1}{N^{2/5}}$ yields

$$\begin{aligned}
&= \frac{2}{N^{1/5}} + \frac{2\sqrt{C}}{N^{1/5}} + \frac{\sqrt{B}\sqrt{B+1}}{N^{3/10}} \\
&= O\left(\frac{1}{N^{1/5}}\right)
\end{aligned}$$

α-divergence with $\alpha \in (-1, 1)$. Lemma 34 provides a function h_δ that upper bounds $f'^2(x)$ for $x \in\in [\delta, \infty)$.

$$
\begin{aligned}
(\dagger\dagger)^2 &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)}>\delta_N\right\}}p(z)dz\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)}>\delta_N\right\}}p(z)dz\right] && (h_{\delta_N} \text{ upper bounds } f_0'^2 \text{ on } (\delta_N,\infty)) \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)p(z)dz\right] && (h_{\delta_N} \text{ non-negative on } [0,\infty)) \\
&= \frac{4\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)^2}{(\alpha-1)^2(\delta_N-1)^2}\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)}{p(z)}-1\right)^2\right] \\
&\leq \frac{4\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)^2}{(\alpha-1)^2(\delta_N-1)^2}\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)}{p(z)}\right)^2+1\right] \\
&= \frac{4\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)^2}{(\alpha-1)^2(\delta_N-1)^2}\left(1+\mathbb{E}_{\mathbf{X}^N}\left[\left\|\frac{\hat{q}_N(z)}{p(z)}\right\|^2_{L_2(P_Z)}\right]\right) \\
&\leq \frac{4(1+B)\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)^2}{(\alpha-1)^2(\delta_N-1)^2} \\
\implies (\dagger\dagger) &= \frac{2\sqrt{1+B}\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)}{(\alpha-1)(\delta_N-1)}.
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|P_Z\right)\right] - D_f\left(Q_Z\|P_Z\right) \\
&\leq \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0(\gamma_N) + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N-\delta_N\right)^2} + \frac{2\sqrt{B}\sqrt{1+B}\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)}{(\alpha-1)(\delta_N-1)\sqrt{N}} \\
&\leq k_1\gamma_N^{\frac{\alpha+1}{2}} + k_2\gamma_N + \frac{k_3}{N(\gamma_N-\delta_N)^2} + \frac{k_4\delta_N^{\frac{\alpha-1}{2}}}{\sqrt{N}}.
\end{aligned}
$$

where each k_i is a positive constant independent of N.

Setting $\gamma_N = \frac{2}{N^{\frac{2}{\alpha+5}}}$ and $\delta_N = \frac{1}{N^{\frac{2}{\alpha+5}}}$ yields

$$
\begin{aligned}
&=\leq \frac{k_1}{N^{\frac{\alpha+1}{\alpha+5}}} + \frac{k_2}{N^{\frac{2}{\alpha+5}}} + \frac{k_3}{N^{\frac{\alpha+1}{\alpha+5}}} + \frac{k_4}{N^{\frac{7-\alpha}{2(\alpha+5)}}} \\
&= O\left(\frac{1}{N^{\frac{\alpha+1}{\alpha+5}}}\right)
\end{aligned}
$$

Jensen-Shannon. Lemma 35 provides a function h_δ that upper bounds $f'^2(x)$ for $x \in [\delta, \infty)$.

$$
\begin{aligned}
(\dagger\dagger)^2 &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] && (h_{\delta_N} \text{ upper bounds } f_0'^2 \text{ on } (\delta_N, \infty)) \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right) p(z)dz\right] && (h_{\delta_N} \text{ non-negative on } [0, \infty)) \\
&= 5\log^2 2 + \log^2\left(\frac{\delta_N}{1+\delta_N}\right) + 2\log 2 \log\left(\frac{\delta_N}{1+\delta_N}\right) \\
&= 5\log^2 2 + \log^2\left(1+\frac{1}{\delta_N}\right) - 2\log 2 \log\left(1+\frac{1}{\delta_N}\right) \\
&\leq 5\log^2 2 + 5\log^2\left(1+\frac{1}{\delta_N}\right) + 10\log 2 \log\left(1+\frac{1}{\delta_N}\right) \\
&= 5\left(\log\left(1+\frac{1}{\delta_N}\right) - \log 2\right)^2 \\
\implies (\dagger\dagger) &\leq \sqrt{5}\log\left(1+\frac{1}{\delta_N}\right) - \sqrt{5}\log 2 \\
&\leq \sqrt{5}\log\left(\frac{2}{\delta_N}\right) - \sqrt{5}\log 2 && (\text{since } \delta_N < 1) \\
&= -\sqrt{5}\log(\delta_N).
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right] - D_f\left(Q_Z \| P_Z\right) \\
&\leq \text{(A)} + \text{(B)} + \text{(C)} \\
&\leq f_0(0) - f_0(\gamma_N) + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} - \frac{\sqrt{5}\sqrt{B}\log\delta_N}{\sqrt{N}} \\
&\leq \gamma_N \log\left(\frac{1+\gamma_N}{2\gamma_N}\right) + \log(1+\gamma_N) + \frac{\log 2\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} - \frac{\sqrt{5}\sqrt{B}\log\delta_N}{\sqrt{N}}
\end{aligned}
$$

Using the fact that $\gamma_N \log(1+\gamma_N) \leq \gamma_N \log 2$ for $\gamma_N < 1$ and $\log(1+\gamma_N) \leq \gamma_N$, we can upper bound the last line with

$$
\leq \gamma_N\left(\log 2 + 1\right) - \gamma_N \log\gamma_N + \frac{\log 2\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} - \frac{\sqrt{5}\sqrt{B}\log\delta_N}{\sqrt{N}}
$$

Setting $\gamma_N = \frac{2}{N^{\frac{1}{3}}}$ and $\delta_N = \frac{1}{N^{\frac{1}{3}}}$ yields

$$
\begin{aligned}
&= \frac{k_1}{N^{\frac{1}{3}}} + \frac{k_2 \log N}{N^{\frac{1}{3}}} + \frac{k_3}{N^{\frac{1}{3}}} + \frac{k_4 \log N}{N^{\frac{1}{2}}} \\
&= O\left(\frac{\log N}{N^{\frac{1}{3}}}\right)
\end{aligned}
$$

where the k_i are positive constants independent of N.

f_β-divergence with $\beta \in (\frac{1}{2}, 1)$. Lemma 36 provides a function h_δ that upper bounds $f'^2(x)$ for $x \in [\delta, \infty)$.

$$
\begin{aligned}
(\dagger\dagger)^2 &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)}>\delta_N\right\}}p(z)dz\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)}>\delta_N\right\}}p(z)dz\right] \qquad (h_{\delta_N} \text{ upper bounds } f_0'^2 \text{ on } (\delta_N,\infty)) \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)p(z)dz\right] \qquad (h_{\delta_N} \text{ non-negative on } [0,\infty)) \\
&= \left(\frac{\beta}{1-\beta}\right)^2\left[\left(1+\delta_N^{-\beta}\right)^{\frac{1-\beta}{\beta}}-2^{\frac{1-\beta}{\beta}}\right]^2+\frac{\beta^2}{(1-\beta)^2}\left(2^{\frac{1-\beta}{\beta}}\right)^2 \\
&\leq 2\left(\frac{\beta}{1-\beta}\right)^2\left[\left(1+\delta_N^{-\beta}\right)^{\frac{1-\beta}{\beta}}+2^{\frac{1-\beta}{\beta}}\right]^2 \\
&\leq 2\left(\frac{\beta}{1-\beta}\right)^2\left[2\left(2\delta_N^{-\beta}\right)^{\frac{1-\beta}{\beta}}\right]^2 \qquad (\text{since } \delta_N<1 \text{ and } \beta>0 \text{ implies } \delta_N^{-\beta}>1) \\
&= 2^{\frac{2+\beta}{\beta}}\left(\frac{\beta}{1-\beta}\right)^2\delta_N^{2(\beta-1)} \\
\implies (\dagger\dagger) &\leq 2^{\frac{2+\beta}{2\beta}}\left(\frac{\beta}{1-\beta}\right)\delta_N^{\beta-1}
\end{aligned}
$$

(noting that $\frac{\beta^2}{(1-\beta)^2}\left(2^{\frac{1}{\beta}-1}\right)^2 = \lim_{x\to\infty} f_0'^2(x)$ as defined in Lemma 36). Thus

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|P_Z\right)\right]-D_f\left(Q_Z\|P_Z\right) \\
&\leq \text{(A)}+\text{(B)}+\text{(C)} \\
&\leq f_0(0)-f_0(\gamma_N)+f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N-\delta_N\right)^2}+\frac{\sqrt{B}}{\sqrt{N}}2^{\frac{2+\beta}{2\beta}}\left(\frac{\beta}{1-\beta}\right)\delta_N^{\beta-1} \\
&\leq \frac{\beta}{1-\beta}\left[1-\left(1+\delta_N^{\beta}\right)^{1/\beta}+2^{\frac{1-\beta}{\beta}}\delta_N\right]+f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N-\delta_N\right)^2}+\frac{\sqrt{B}}{\sqrt{N}}2^{\frac{2+\beta}{2\beta}}\left(\frac{\beta}{1-\beta}\right)\delta_N^{\beta-1} \\
&\leq \frac{\beta}{1-\beta}2^{\frac{1-\beta}{\beta}}\delta_N+f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N-\delta_N\right)^2}+\frac{\sqrt{B}}{\sqrt{N}}2^{\frac{2+\beta}{2\beta}}\left(\frac{\beta}{1-\beta}\right)\delta_N^{\beta-1} \\
&= k_1\delta_N+\frac{k_2}{N(\gamma_N-\delta_N)^2}+\frac{k_3\delta_N^{\beta-1}}{\sqrt{N}}
\end{aligned}
$$

where the k_i are positive constants independent of N.

Setting $\gamma_N = \frac{2}{N^{\frac{1}{3}}}$ and $\delta_N = \frac{1}{N^{\frac{1}{3}}}$ yields

$$= \frac{k_1}{N^{\frac{1}{3}}} + \frac{k_2}{N^{\frac{1}{3}}} + \frac{k_3}{N^{\frac{1}{2}+\frac{\beta-1}{3}}}$$
$$= O\left(\frac{1}{N^{\frac{1}{3}}}\right)$$

□

6.7.5 Proof of Theorem 28

We will make use of McDiarmid's theorem in our proof of Theorem 28:

Theorem (McDiarmid's inequality)**.** *Suppose that $X_1, \ldots, X_N \in \mathcal{X}$ are independent random variables and that $\phi : \mathcal{X}^N \to \mathbb{R}$ is a function. If it holds that for all $i \in \{1, \ldots, N\}$ and $x_1, \ldots, x_N, x_{i'}$,*

$$|\phi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_N) - \phi(x_1, \ldots, x_{i-1}, x_{i'}, x_{i+1}, \ldots, x_N)| \leq c_i,$$

then

$$\mathbb{P}\left(\phi(X_1, \ldots, X_N) - \mathbb{E}\phi \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^N c_i^2}\right)$$

and

$$\mathbb{P}\left(\phi(X_1, \ldots, X_N) - \mathbb{E}\phi \geq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^N c_i^2}\right)$$

In our setting we will consider $\phi(\mathbf{X}^N) = D_f\left(\hat{Q}_Z^N \| P_Z\right)$.

Theorem 38 (Tail bounds for RAM)**.** *Suppose that $\chi^2\left(Q_{Z|x} \| P_Z\right) \leq C < \infty$ for all x and for some constant C. Then, the RAM estimator $D_f(\hat{Q}_Z^N \| P_Z)$ concentrates to its mean in the following sense. For $N > 8$ and for any $\delta > 0$, with probability at least $1 - \delta$ it holds that*

$$\left| D_f(\hat{Q}_Z^N \| P_Z) - \mathbb{E}_{\mathbf{X}^N}\left[D_f(\hat{Q}_Z^N \| P_Z)\right] \right| \leq K \cdot \psi(N) \sqrt{\log(2/\delta)},$$

where K is a constant and $\psi(N)$ is given in Table 6.2.

Proof (Theorem 28). We will show that $D_f\left(\hat{Q}_Z^N \| P_Z\right)$ exhibits the bounded difference property as in the statement of McDiarmid's theorem. Since $\hat{q}_N(z)$ is symmetric in the indices of $\mathbf{X}^N$, we can without loss of generality consider only the case $i = 1$. Henceforth, suppose $\mathbf{X}^N, \mathbf{X}^{N'}$ are two batches of data with $\mathbf{X}_1^N \neq \mathbf{X}^{N'}{}_1$ and $\mathbf{X}_i^N = \mathbf{X}^{N'}{}_i$ for all $i > 1$. For the remainder of this proof we will write explicitly the dependence of $\hat{Q}_Z^N$ on $\mathbf{X}^N$. We will write $\hat{Q}_Z^N(\mathbf{X}^N)$ for the probability measure and $\hat{q}_N(z; \mathbf{X}^N)$ for its density.

We will show that $\left|D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right)\right| \leq c_N$ where c_N is a constant depending only on N. From this fact, McDiarmid's theorem and the union bound, it follows that:

$$\begin{aligned}
&\mathbb{P}\left(\left|D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - \mathbb{E}_{\mathbf{X}^N} D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right)\right| \geq t\right) \\
&= \mathbb{P}\Big(D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - \mathbb{E}_{\mathbf{X}^N} D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) \geq t \text{ or} \\
&\qquad\qquad D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - \mathbb{E}_{\mathbf{X}^N} D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) \leq -t\Big) \\
&\leq \mathbb{P}\left(D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - \mathbb{E}_{\mathbf{X}^N} D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) \geq t\right) + \\
&\qquad\qquad \mathbb{P}\left(D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - \mathbb{E}_{\mathbf{X}^N} D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) \leq -t\right) \\
&\leq 2\exp\left(\frac{-2t^2}{Nc_N^2}\right).
\end{aligned}$$

Observe that by setting $t = \sqrt{\frac{Nc_N^2}{2}\log\left(\frac{2}{\delta}\right)}$,

the above inequality is equivalent to the statement that for any $\delta > 0$, with probability at least $1-\delta$

$$\left|D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - \mathbb{E}_{\mathbf{X}^N} D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right)\right| < \sqrt{\frac{Nc_N^2}{2}}\sqrt{\log\left(\frac{2}{\delta}\right)}.$$

We will show that $c_N \leq kN^{-1/2}\psi(N)$ for k and $\psi(N)$ depending on f. The statement of Theorem 28 is of this form. Note that in order to show that

$$\left|D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right)\right| \leq c_N, \tag{6.7}$$

it is sufficient to prove that

$$D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \leq c_N \tag{6.8}$$

since the symmetry in $\mathbf{X}^N \leftrightarrow \mathbf{X}^{N\prime}$ implies that

$$-D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) + D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \leq c_N \tag{6.9}$$

and thus implies Inequality 6.7. The remainder of this proof is therefore devoted to showing that Inequality 6.8 holds for each divergence.

We will make use of the fact that $\chi^2\left(Q_{Z|x}\|P_Z\right) \leq C \implies \left\|\frac{q(z|x)}{p(z)}\right\|_{L_2(P_Z)} \leq C+1$

The case that D_f is the χ^2-divergence, Total Variation or D_{f_β} with $\beta > 1$:

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \\
&= \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) dP_Z(z) \\
&\leq \int \left(\frac{\hat{q}_N(z;\mathbf{X}^N) - \hat{q}_N(z;\mathbf{X}^{N\prime})}{p(z)}\right) f_0'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) dP_Z(z) \\
&\leq \left\|\frac{\hat{q}_N(z;\mathbf{X}^N) - \hat{q}_N(z;\mathbf{X}^{N\prime})}{p(z)}\right\|_{L_2(P_Z)} \times \left\|f_0'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right)\right\|_{L_2(P_Z)} \qquad \text{(Cauchy-Schwartz)} \\
&= \left\|\frac{1}{N}\frac{q(z|X_1) - q(z|X_1')}{p(z)}\right\|_{L_2(P_Z)} \times \left\|f_0'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right)\right\|_{L_2(P_Z)} \\
&\leq \frac{1}{N}\left(\left\|\frac{q(z|X_1)}{p(z)}\right\|_{L_2(P_Z)} + \left\|\frac{q(z|X_1')}{p(z)}\right\|_{L_2(P_Z)}\right) \times \left\|f_0'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right)\right\|_{L_2(P_Z)} \\
&\leq \frac{2(C+1)}{N}\left\|f_0'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right)\right\|_{L_2(P_Z)}.
\end{aligned}
$$

By similar arguments as made in the proof of Theorem 27 considering the term (ii), $\left\|f_0'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right)\right\|_{L_2(P_Z)} = \sqrt{\mathbb{E}_Z f_0'^2\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right)} = O(1)$ thus we have the difference is upper-bounded by $c_N = \frac{k}{N}$ for some constant k. The only modification needed to the proof in Theorem 27 is the omission of all occurrences of $\mathbb{E}_{\mathbf{X}^N}$.

This holds for any $N > 0$.

All other divergences. Similar to the proof of Theorem 27, we write the difference as the sum of integrals over different mutually exclusive events that partition $\mathcal{Z}$. Denoting by γ_N and δ_N scalars depending on N, we have that

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \\
&= \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) dP_Z(z) \\
&= \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) \mathbb{1}_{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\le\delta_N \text{ and } \frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\le\gamma_N\right\}} dP_Z(z) \qquad \text{(A)} \\
&\quad + \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) \mathbb{1}_{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\le\delta_N \text{ and } \frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)>\gamma_N\right\}} dP_Z(z) \qquad \text{(B)} \\
&\quad + \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) \mathbb{1}_{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)>\delta_N\right\}} dP_Z(z). \qquad \text{(C)}
\end{aligned}
$$

We will consider each of the terms (A), (B) and (C) separately.

Later on, we will pick γ_N and δ_N to be decreasing in N such that $\delta_N < \gamma_N$. We will require N sufficiently large so that $\gamma_N < 1$, so in the rest of this proof we will assume this to be the case and later on provide lower bounds on how large N must be to ensure this.

(A): Recall that $f_0(x)$ is decreasing on the interval $[0,1]$. Since $\gamma_N, \delta_N \le 1$, the integrand is at most $f_0(0) - f_0(\gamma_N)$, and so

$$
\text{(A)} \le f_0(0) - f_0(\gamma_N)
$$

(B): Since $\delta_N \le 1$, the integrand is at most $f_0(0)$ and so

$$
\text{(B)} \le f_0(0) \times \mathbb{P}_Z \underbrace{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z) \le \delta_N \text{ and } \frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z) > \gamma_N\right\}}_{(*)}
$$

We will bound $\mathbb{P}_Z(*) = 0$ using Chebyshev's inequality. Noting that

$$
\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} = \frac{\hat{q}_N(z;\mathbf{X}^{N\prime})}{p(z)} - \frac{1}{N}\frac{q(z|X_1')}{p(z)} + \frac{1}{N}\frac{q(z|X_1)}{p(z)},
$$

and using the fact that $\frac{q(z|X_1)}{p(z)} > 0$ it follows that

$$\begin{aligned}
\textcircled{*} &\implies \gamma_N - \frac{1}{N}\frac{q(z|X_1')}{p(z)} + \frac{1}{N}\frac{q(z|X_1)}{p(z)} < \delta_N \\
&\iff (\gamma_N - \delta_N)N + \frac{q(z|X_1)}{p(z)} < \frac{q(z|X_1')}{p(z)} \\
&\implies (\gamma_N - \delta_N)N < \frac{q(z|X_1')}{p(z)} \\
&\implies (\gamma_N - \delta_N)N - 1 < \frac{q(z|X_1')}{p(z)} - 1.
\end{aligned}$$

where the penultimate line follows from the fact that $q(z|X_1)/p(z) \geq 0$. It follows that

$$\begin{aligned}
\mathbb{P}_Z\textcircled{*} &\leq \mathbb{P}_Z\left\{\frac{q(z|X_1')}{p(z)} - 1 > (\gamma_N - \delta_N)N - 1\right\} \\
&\leq \mathbb{P}_Z\left\{\left|\frac{q(z|X_1')}{p(z)} - 1\right| > (\gamma_N - \delta_N)N - 1\right\}.
\end{aligned}$$

Denote by $\sigma^2(X) = \mathrm{Var}_Z\left[\frac{q(z|X)}{p(z)}\right] = \mathbb{E}_Z\frac{q^2(z|X)}{p^2(z)} - 1 \leq C$. We have by Chebyshev that for any $t > 0$,

$$\mathbb{P}_Z\left\{\left|\frac{q(z|X)}{p(z)} - 1\right| > t\right\} \leq \frac{\sigma^2(X)}{t^2}$$

and so setting $t = (\gamma_N - \delta_N)N - 1$ yields

$$\mathbb{P}_Z\textcircled{*} \leq \frac{\sigma^2(X)}{((\gamma_N - \delta_N)N - 1)^2} \leq \frac{C}{((\gamma_N - \delta_N)N - 1)^2}$$

It follow that

$$\textcircled{B} \leq f_0(0)\frac{C}{((\gamma_N - \delta_N)N - 1)^2}$$

(C): Similar to the proof of Theorem 27, we can upper bound this term by the product of two terms, one of which is independent of the choice of divergence. The other term will be treated separately for each divergence considered.

$$
\begin{aligned}
\textcircled{C} &= \int f_0\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) - f_0\left(\frac{\hat{q}_N(z;\mathbf{X}^{N\prime})}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} > \delta_N\right\}} dP_Z(z) \\
&\leq \int \left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} - \frac{\hat{q}_N(z;\mathbf{X}^{N\prime})}{p(z)}\right) {f_0}'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} > \delta_N\right\}} dP_Z(z) \quad \text{(Convexity of } f_0\text{)} \\
&= \int \frac{1}{N}\frac{q(z|X_1) - q(z|X_1')}{p(z)} {f_0}'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} > \delta_N\right\}} dP_Z(z) \\
&\leq \left\| \frac{1}{N}\frac{q(z|X_1) - q(z|X_1')}{p(z)} \right\|_{L_2(P_Z)} \left\| {f_0}'\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} > \delta_N\right\}} \right\|_{L_2(P_Z)} \quad \text{(Cauchy-Schwartz)} \\
&\leq \frac{2(C+1)}{N} \underbrace{\sqrt{\int {f_0}'^2\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} > \delta_N\right\}} p(z)dz}}_{\textcircled{*}} \quad \text{(Boundedness of } \left\|\tfrac{q(z|x)}{p(z)}\right\|_{L_2(P_Z)}\text{)}
\end{aligned}
$$

The term $\textcircled{*}$ will be treated separately for each divergence.

KL**:** By Lemma 32, there exists a function $h_{\delta_N}(x)$ that is positive and concave on $[0,\infty)$ and is an upper bound of $f_0'^2(x)$ on $[\delta_N,\infty)$ with $h_{\delta_N}(1) = \log^2(\delta_N) + \frac{2}{e}$.

$$
\begin{aligned}
\textcircled{*}^2 &\leq \int h_{\delta_N}\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) \mathbb{1}_{\left\{\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} > \delta_N\right\}} p(z)dz \quad (h_{\delta_N} \text{ upper bounds } f'^2 \text{ on } (\delta_N,\infty)) \\
&\leq \int h_{\delta_N}\left(\frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)}\right) p(z)dz \quad (h_{\delta_N} \text{ non-negative on } [0,\infty)) \\
&\leq h_{\delta_N}\left(\int \frac{\hat{q}_N(z;\mathbf{X}^N)}{p(z)} p(z)dz\right) \quad (h_{\delta_N} \text{ concave}) \\
&= h_{\delta_N}(1) \\
&= \log^2(\delta_N) + \frac{2}{e} \\
\implies \textcircled{C} &\leq \frac{2(C+1)}{N}\sqrt{\log^2(\delta_N) + \frac{2}{e}}.
\end{aligned}
$$

Putting together the separate integrals and setting $\delta_N = \frac{1}{N^{2/3}}$ and $\gamma_N = \frac{2}{N^{2/3}}$, we have that

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \\
&= \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{2(C+1)}{N}\sqrt{\log^2(\delta_N) + \frac{2}{e}} \\
&= \gamma_N - \gamma_N \log \gamma_N + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{2(C+1)}{N}\sqrt{\log^2(\delta_N) + \frac{2}{e}} \\
&= \frac{2}{N^{2/3}} - \frac{2}{N^{2/3}} \log\left(\frac{2}{N^{2/3}}\right) + \frac{f_0(0)C}{\left(N^{1/3} - 1\right)^2} + \frac{2(C+1)}{N}\sqrt{\frac{4}{9}\log^2(N) + \frac{2}{e}} \\
&\leq \frac{2}{N^{2/3}} - \frac{2}{N^{2/3}} \log\left(\frac{2}{N^{2/3}}\right) + \frac{9f_0(0)C}{4N^{2/3}} + \frac{2(C+1)}{N}\sqrt{\frac{4}{9}\log^2(N) + \frac{2}{e}} \\
&= \frac{k_1}{N^{2/3}} + \frac{k_2 \log N}{N^{2/3}} + \frac{k_3\sqrt{\log^2 N + \frac{9}{2e}}}{N} \\
&\leq (k_1 + k_2 + 2k_3)\frac{\log N}{N^{2/3}}
\end{aligned}
$$

where k_1, k_2 and k_3 are constants depending on C. The second inequality holds if $N^{1/3} - 1 > \frac{N^{1/3}}{3} \iff N > \left(\frac{3}{2}\right)^3 < 4$ and the third inequality holds if $N \geq 4$

The assumption that $\delta_N, \gamma_N \leq 1$ holds if $N > 2^{3/2}$ and so holds if $N \geq 3$.

This leads to $Nc_N^2 = \frac{\log^2 N}{N^{1/3}}$ for $N > 3$.

Squared Hellinger. In this case similar reasoning to the other divergences leads to a bound that is worse than $O\left(\frac{1}{\sqrt{N}}\right)$ and thus Nc_N^2 is bigger than $O(1)$ leading to a trivial concentration result.

α-divergence with $\alpha \in (\frac{1}{3}, 1)$. Following similar reasoning to the proof of Theorem 27 for the α-divergence case, we use the function $h_{\delta_N}(x)$ provided by Lemma 34 to derive the following upper bound:

$$
\textcircled{C} \leq \frac{2(C+1)}{N} \cdot \frac{2\sqrt{1 + (C+1)^2}\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)}{(\alpha - 1)(\delta_N - 1)}.
$$

Setting $\delta_N = \frac{1}{N^{\frac{4}{\alpha+5}}}$ and $\gamma_N = \frac{2}{N^{\frac{4}{\alpha+5}}}$,

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \\
&= \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{f_0(0)C}{\left((\gamma_N-\delta_N)N-1\right)^2} + \frac{2(C+1)}{N}\frac{2\sqrt{1+(C+1)^2}\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)}{(1-\alpha)(1-\delta_N)} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{t^2 f_0(0)C}{(t-1)^2(\gamma_N-\delta_N)^2N^2} + \frac{2(C+1)}{N}\frac{2\sqrt{1+(C+1)^2}\left(\delta_N^{\frac{\alpha-1}{2}}-1\right)}{(1-\alpha)(1-\delta_N)} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{t^2 f_0(0)C}{(t-1)^2(\gamma_N-\delta_N)^2N^2} + \frac{2(C+1)}{N}\frac{4\sqrt{1+(C+1)^2}\delta_N^{\frac{\alpha-1}{2}}}{(1-\alpha)} \\
&\leq k_1\gamma_N^{\frac{\alpha+1}{2}} + k_2\gamma_N + \frac{k_3}{(\gamma_N-\delta_N)^2N^2} + \frac{k_4\delta_N^{\frac{\alpha-1}{2}}}{N} \\
&= \frac{k_1}{N^{\frac{2\alpha+2}{\alpha+5}}} + \frac{k_2}{N^{\frac{4}{\alpha+5}}} + \frac{k_3}{N^{\frac{2\alpha-2}{\alpha+5}}} + \frac{k_4}{N^{\frac{3\alpha+3}{\alpha+5}}} \\
&\leq \frac{k_1+k_2+k_3+k_4}{N^{\frac{2\alpha+2}{\alpha+5}}}
\end{aligned}
$$

where t is any positive number and where the second inequality holds if $N^{\frac{2\alpha+2}{\alpha+5}} - 1 > \frac{N^{\frac{2\alpha+2}{\alpha+5}}}{t} \iff N > (\frac{t}{t-1})^{\frac{\alpha+5}{2\alpha+21}}$. For $\alpha \in (\frac{1}{3}, 1)$ we have $\frac{\alpha+5}{2\alpha+2} \in (\frac{3}{2}, 2)$. If we take $t = 100$ then $N > 1$ suffices for any α.

The third inequality holds if $1 - \delta_N > \frac{1}{2} \iff N > 2^{\frac{\alpha+5}{4}}$ and so holds if $N > 3$.

The assumption that $\delta_N, \gamma_N \leq 1$ holds if $N > 4^{\frac{\alpha+5}{4}} \leq 8$ and so holds if $N > 8$.

Thus, this leads to $Nc_N^2 = \frac{k}{N^{\frac{3\alpha-1}{\alpha+5}}}$ for $N > 8$.

Jensen-Shannon. Following similar reasoning to the proof of Theorem 27 for the α-divergence case, we use the function $h_{\delta_N}(x)$ provided by Lemma 35 to derive the following upper bound:

$$
\textcircled{C} \leq \frac{2(C+1)}{N} \cdot \sqrt{5}\log\left(\frac{1}{\delta_N}\right).
$$

Setting $\delta_N = \frac{1}{N^{2/3}}$ and $\gamma_N = \frac{2}{N^{2/3}}$,

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \\
&= \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{2(C+1)}{N} \cdot \log\left(\frac{1}{\delta_N}\right) \\
&\leq \gamma_N \log\left(\frac{1+\gamma_N}{2\gamma_N}\right) + \log(1+\gamma_N) + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{2(C+1)}{N} \cdot \log\left(\frac{1}{\delta_N}\right).
\end{aligned}
$$

Using the fact that $\log(1+\gamma_N) \leq \gamma_N$, we obtain the following upper bound:

$$
\begin{aligned}
&\leq \gamma_N^2 + \gamma_N(1 - \log 2) - \gamma_N \log \gamma_N + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{2(C+1)}{N} \cdot \log\left(\frac{1}{\delta_N}\right) \\
&= \frac{k_1}{N^{4/3}} + \frac{k_2}{N^{2/3}} + \frac{k_3 \log N}{N^{2/3}} + \frac{k_4}{(N^{1/3}-1)^2} + \frac{k_5 \log N}{N^{2/3}} \\
&= \frac{k_1}{N^{4/3}} + \frac{k_2}{N^{2/3}} + \frac{k_3 \log N}{N^{2/3}} + \frac{k_4}{(N^{1/3}-1)^2} + \frac{k_5 \log N}{N^{2/3}} \\
&\leq \frac{k_1}{N^{4/3}} + \frac{k_2}{N^{2/3}} + \frac{k_3 \log N}{N^{2/3}} + \frac{100k_4}{81N^{2/3}} + \frac{k_5 \log N}{N^{2/3}} \\
&\leq (k_1 + k_2 + k_3 + k_4' + k_5)\frac{\log N}{N^{2/3}}
\end{aligned}
$$

where the penultimate inequality holds if $N^{1/3} - 1 > \frac{N^{1/3}}{10} \iff N > \left(\frac{10}{9}\right)^3$ which is satisfied if $N > 1$ and the last inequality is true if $N > 1$.

The assumption that $\delta_N, \gamma_N \leq 1$ holds if $N > 2^{3/2}$ and so holds if $N \geq 3$.

This leads to $Nc_N^2 = \frac{\log^2 N}{N^{1/3}}$ for $N > 2$.

f_β-divergence, $\beta \in (\frac{1}{2}, 1)$. Following similar reasoning to the proof of Theorem 27 for the α-divergence case, we use the function $h_{\delta_N}(x)$ provided by Lemma 36 to derive the following upper bound:

$$
\textcircled{C} \leq \frac{2(C+1)}{N} \cdot \frac{\beta}{1-\beta} \cdot 2^{\frac{2+\beta}{2\beta}} \delta_N^{\beta-1}.
$$

Setting $\delta_N = \frac{1}{N^{2/3}}$ and $\gamma_N = \frac{2}{N^{2/3}}$,

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N)\|P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime})\|P_Z\right) \\
&= \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{\beta}{1-\beta}\cdot 2^{\frac{2+\beta}{2\beta}}\delta_N^{\beta-1} \\
&\leq \frac{\beta}{\beta-1}2^{\frac{1-\beta}{\beta}}\gamma_N + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{\beta}{1-\beta}\cdot 2^{\frac{2+\beta}{2\beta}}\frac{\delta^{\beta-1}}{N} \\
&= \frac{k_1}{N^{2/3}} + \frac{k_2}{(N^{1/3}-1)^2} + \frac{k_3}{N^{\frac{2\beta+1}{3}}} \\
&\leq \frac{k_1}{N^{2/3}} + \frac{100k_2}{81N^{2/3}} + \frac{k_3}{N^{\frac{2\beta+1}{3}}} \\
&\leq \frac{k_1 + k_2' + k_3}{N^{2/3}}
\end{aligned}
$$

where the penultimate inequality holds if $N^{1/3} - 1 > \frac{N^{1/3}}{10} \iff N > \left(\frac{10}{9}\right)^3$ which is satisfied if $N > 1$.

The assumption that $\delta_N, \gamma_N \leq 1$ holds if $N > 2^{3/2}$ and so holds if $N \geq 3$.

This leads to $Nc_N^2 = \frac{1}{N^{1/3}}$ for $N > 2$.

□

6.7.6 Full statement and proof of Theorem 29

The statement of Theorem 29 in the main text was simplified for brevity. Below is the full statement, followed by its proof.

TODO sort out theorem numbering

Theorem 39. *For any π,*

$$\mathbb{E}_{\mathbf{Z}^M,\mathbf{X}^N}\left[\hat{D}_f^M(\hat{Q}_Z^N\|P_Z)\right] = \mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N\|P_Z\right)\right].$$

If either of the following conditions are satisfied:

$$
\begin{aligned}
&(i)\ \pi(z|\mathbf{X}^N) = p(z), \quad \mathbb{E}_X\left\|f\left(\tfrac{q(z|X)}{p(z)}\right)\right\|^2_{L_2(P_Z)} < \infty, \quad \mathbb{E}_X\left\|\tfrac{q(z|X)}{p(z)}\right\|^2_{L_2(P_Z)} < \infty \\
&(ii)\ \pi(z|\mathbf{X}^N) = \hat{q}_N(z), \quad \mathbb{E}_X\left\|f\left(\tfrac{q(z|X)}{p(z)}\right)\tfrac{p(z)}{q(z|X)}\right\|^2_{L_2(Q_{Z|X})} < \infty, \quad \mathbb{E}_X\left\|\tfrac{p(z)}{q(z|X)}\right\|^2_{L_2(Q_{Z|X})} < \infty
\end{aligned}
$$

then, denoting by $\psi(N)$ the rate given in Table 6.2, we have

$$Var_{\mathbf{Z}^M,\mathbf{X}^N}\left[\hat{D}_f^M(\hat{Q}_Z^N\|P_Z)\right] = O\left(M^{-1}\right) + O\left(\psi(N)^2\right)$$

In proving Theorem 29 we will make use of the following lemma.

Lemma 40. *For any $f_0(x)$, the functions $f_0(x)^2$ and $\frac{f_0(x)^2}{x}$ are convex on $(0, \infty)$.*

Proof. To see that $f_0(x)^2$ is convex, observe that

$$\frac{d^2}{dx^2} f_0(x)^2 = 2\left(f_0(x) f_0''(x) + f_0'(x)^2\right)$$

All of these terms are postive for $x > 0$. Indeed, since $f_0(x)$ is convex for $x > 0$, $f_0''(x) \geq 0$. By construction of f_0, $f_0(x) \geq 0$ for $x > 0$. Thus $f_0(x)^2$ has non-negative derivative and is thus convex on $(0, \infty)$.

To see that $\frac{f_0(x)^2}{x}$ is convex, observe that

$$\frac{d^2}{dx^2} \frac{f_0(x)^2}{x} = \frac{2}{x}\left(f_0(x) f_0''(x) + \left(f_0'(x) - \frac{f_0(x)}{x}\right)^2\right).$$

By the same arguments above, this is positive for $x > 0$ and thus $\frac{f_0(x)^2}{x}$ is convex for $x > 0$. □

Proof. (Theorem 29) For the expectation, observe that

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}^M, \mathbf{X}^N} \hat{D}_f^M(\hat{Q}_Z^N \| P_Z) &= \mathbb{E}_{\mathbf{X}^N}\left[\mathbb{E}_{\mathbf{Z}^M \overset{i.i.d.}{\sim} \pi(z|\mathbf{X}^N)} \hat{D}_f^M(\hat{Q}_Z^N \| P_Z)\right] \\
&= \mathbb{E}_{\mathbf{X}^N}\left[\mathbb{E}_{z \sim \pi(z|\mathbf{X}^N)} f\left(\frac{\hat{q}_N(z)}{p(z)}\right) \frac{p(z)}{\pi(z|\mathbf{X}^N)}\right] \\
&= \mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right].
\end{aligned}$$

For the variance, by the law of total variance we have that

$$\begin{aligned}
&\mathrm{Var}_{\mathbf{Z}^M, \mathbf{X}^N}\left[\hat{D}_f^M(\hat{Q}_Z^N \| P_Z)\right] \\
&= \mathbb{E}_{\mathbf{X}^N} \mathrm{Var}_{\mathbf{Z}^M \overset{i.i.d.}{\sim} \pi(z|\mathbf{X}^N)} \hat{D}_f^M(\hat{Q}_Z^N \| P_Z) + \mathrm{Var}_{\mathbf{X}^N} \mathbb{E}_{\mathbf{Z}^M \overset{i.i.d.}{\sim} \pi(z|\mathbf{X}^N)} \hat{D}_f^M(\hat{Q}_Z^N \| P_Z) \\
&= \frac{1}{M} \underbrace{\mathbb{E}_{\mathbf{X}^N} \mathrm{Var}_{\pi(z|\mathbf{X}^N)}\left[f\left(\frac{\hat{q}_N(z)}{p(z)}\right) \frac{p(z)}{\pi\left(z|\mathbf{X}^N\right)}\right]}_{(i)} + \underbrace{\mathrm{Var}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right]}_{(ii)}.
\end{aligned}$$

Consider term (ii). The concentration results of Theorem 28 imply bounds on (ii), since for a random variable X,

$$\begin{aligned}\mathrm{Var}X &= \mathbb{E}(X - EX)^2 \\ &= \int_0^\infty \mathbb{P}\left((X - \mathbb{E}X)^2 > t\right) dt \\ &= \int_0^\infty \mathbb{P}\left(|X - \mathbb{E}X| > \sqrt{t}\right) dt.\end{aligned}$$

It follows therefore that

$$\begin{aligned}\mathrm{Var}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N \| P_Z\right)\right] &\leq \int_0^\infty 2\exp\left(-\frac{k}{\psi(N)^2}t\right) dt \\ &= O\left(\psi(N)^2\right)\end{aligned}$$

where $\psi(N)$ is given by Table 6.2.

Next we consider (i) and show that it is bounded independent of N, and so the component of the variance due to this term is $O\left(\frac{1}{M}\right)$. In the case that $\pi(z|\mathbf{X}^N) = p(z)$,

$$\begin{aligned}(i) &\leq \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{p(z)}\left[f\left(\frac{\hat{q}_N(z)}{p(z)}\right)^2\right] \\ &= \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{p(z)}\left[\left(f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right) + f'(1)\left(\frac{\hat{q}_N(z)}{p(z)} - 1\right)\right)^2\right] \\ &\leq \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{p(z)}\left[f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right)^2\right] + f'(1)^2\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{p(z)}\left[\left(\frac{\hat{q}_N(z)}{p(z)} - 1\right)^2\right] \\ &\quad + 2f'(1)\sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{p(z)}\left[f_0\left(\frac{\hat{q}_N(z)}{p(z)}\right)^2\right]} \times \sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{p(z)}\left[\left(\frac{\hat{q}_N(z)}{p(z)} - 1\right)^2\right]} \\ &\leq \mathbb{E}_X\mathbb{E}_{p(z)}\left[f_0\left(\frac{q(z|X)}{p(z)}\right)^2\right] + f'(1)^2\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\frac{q(z|X)}{p(z)} - 1\right)^2\right] \\ &\quad + 2f'(1)\sqrt{\mathbb{E}_X\mathbb{E}_{p(z)}\left[f_0\left(\frac{q(z|X)}{p(z)}\right)^2\right]} \times \sqrt{\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\frac{q(z|X)}{p(z)} - 1\right)^2\right]}\end{aligned}$$

The penultimate inequality follows by application of Cauchy-Schwartz. The last inequality follows by Proposition 1 applied to $D_{f_0^2}$ and $D_{(x-1)^2}$, using the fact that the functions $f_0^2(x)$ and $(x-1)^2$ are convex and are zero at $x = 1$ (see Lemma 40). By assumption,

$\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\frac{q(z|X)}{p(z)}-1\right)^2\right]<\infty$. Consider the other term:

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_{p(z)}\left[f_0\left(\frac{q(z|X)}{p(z)}\right)^2\right] &= \mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(f\left(\frac{q(z|X)}{p(z)}\right)-f'(1)\left(\frac{q(z|X)}{p(z)}-1\right)\right)^2\right]\\ &\leq \mathbb{E}_X\mathbb{E}_{p(z)}\left[f\left(\frac{q(z|X)}{p(z)}\right)^2\right]+f'(1)^2\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\frac{q(z|X)}{p(z)}-1\right)^2\right]\\ &\quad+2f'(1)\sqrt{\mathbb{E}_X\mathbb{E}_{p(z)}\left[f\left(\frac{q(z|X)}{p(z)}\right)^2\right]}\times\sqrt{\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\frac{q(z|X)}{p(z)}-1\right)^2\right]}\\ &<\infty\end{aligned}$$

The inequality follows by Cauchy-Schwartz. All terms are finite by assumption. Thus $(i)\leq K<\infty$ for some K independent of N.

Now consider the case that $\pi(z|\mathbf{X}^N)=\hat{q}_N(z)$. Then, following similar (but algebraically more tedious) reasoning to the previous case, it can be shown that

$$\begin{aligned}(i)\leq\mathbb{E}_X\mathbb{E}_{p(z)}\left[f_0\left(\frac{q(z|X)}{p(z)}\right)^2\frac{p(z)}{q(z|X)}\right]+f'(1)^2\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\sqrt{\frac{q(z|X)}{p(z)}}-\sqrt{\frac{p(z)}{q(z|X)}}\right)^2\right]\\ +2f'(1)\sqrt{\mathbb{E}_X\mathbb{E}_{p(z)}\left[f_0\left(\frac{q(z|X)}{p(z)}\right)^2\frac{p(z)}{q(z|X)}\right]}\times\sqrt{\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\sqrt{\frac{q(z|X)}{p(z)}}-\sqrt{\frac{p(z)}{q(z|X)}}\right)^2\right]}\end{aligned}$$

where Proposition 1 is applied to $D_{\frac{f_0^2(x)}{x}}$ and $D_{(\sqrt{x}-\frac{1}{\sqrt{x}})^2}$, using the fact that the functions $f_0^2(x)/x$ and $(\sqrt{x}-\frac{1}{\sqrt{x}})^2$ are convex and are zero at $x=1$ (see Lemma 40). Noting that

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_{p(z)}\left[\left(\sqrt{\frac{q(z|X)}{p(z)}}-\sqrt{\frac{p(z)}{q(z|X)}}\right)^2\right] &= \mathbb{E}_X\mathbb{E}_{p(z)}\left[\frac{q(z|X)}{p(z)}+\frac{p(z)}{q(z|X)}-2\right]\\ &= \mathbb{E}_X\mathbb{E}_{p(z)}\left[\frac{p(z)}{q(z|X)}-1\right]<\infty\end{aligned}$$

where the inequality holds by assumption, it follows that

$$\begin{aligned}
&\mathbb{E}_X \mathbb{E}_{p(z)} \left[f_0 \left(\frac{q(z|X)}{p(z)} \right)^2 \frac{p(z)}{q(z|X)} \right] \\
&\leq \mathbb{E}_X \mathbb{E}_{p(z)} \left[f \left(\frac{q(z|X)}{p(z)} \right)^2 \frac{p(z)}{q(z|X)} \right] + f'(1)^2 \mathbb{E}_X \mathbb{E}_{p(z)} \left[\left(\sqrt{\frac{q(z|X)}{p(z)}} - \sqrt{\frac{p(z)}{q(z|X)}} \right)^2 \right] \\
&\quad + 2f'(1) \sqrt{\mathbb{E}_X \mathbb{E}_{p(z)} \left[f \left(\frac{q(z|X)}{p(z)} \right)^2 \frac{p(z)}{q(z|X)} \right]} \times \sqrt{\mathbb{E}_X \mathbb{E}_{p(z)} \left[\left(\sqrt{\frac{q(z|X)}{p(z)}} - \sqrt{\frac{p(z)}{q(z|X)}} \right)^2 \right]} \\
&< \infty.
\end{aligned}$$

where the first inequality holds by the definition of f_0 and Cauchy-Schwartz.

Thus $(i) \leq K < \infty$ for some K independent of N in both cases of π. □

6.7.7 Elaboration of Section 6.2.3: satisfaction of assumptions of theorems

Suppose that P_Z is $\mathcal{N}(0, I_d)$ and $Q_{Z|X}$ is $\mathcal{N}(\mu(X), \Sigma(X))$ with Σ diagonal. Suppose further that there exist constants $K, \epsilon > 0$ such that $\|\mu(X)\| \leq K$ and $\Sigma_{ii}(X) \in [\epsilon, 1]$ for all i.

By Lemma 41, it holds that $\chi^2(Q_{Z|x}, P_Z) < \infty$ for all $x \in \mathcal{X}$. By compactness of the sets in which $\mu(X)$ and $\Sigma(X)$ take value, it follows that there exists $C < \infty$ such that $\chi^2(Q_{Z|x}, P_Z) \leq C$ and thus the setting of Theorem 28 holds.

A similar argument based on compactness shows that the density ratio is uniformly bounded in z and x: $q(z|x)/p(z) \leq C'$ for some $C' < \infty$. It therefore follows that the condition of Theorem 27 holds: $\int q^4(z|x)/p^4(z) dP(z) < C'^4 < \infty$.

We conjecture that the strong boundedness assumptions on $\mu(X)$ and $\Sigma(X)$ also imply the setting of Theorem 26 $\mathbb{E}_X[\chi^2(Q_{Z|X}, Q_Z)] < \infty$. Since the divergence Q_Z explicitly depends on the data distribution, this is more difficult to verify than the conditions of Theorems 27 and 28.

The crude upper bound provided by convexity

$$\mathbb{E}_X[\chi^2(Q_{Z|X}, Q_Z)] \leq \mathbb{E}_X \mathbb{E}_{X'}[\chi^2(Q_{Z|X}, Q_{Z|X'})]$$

provides a sufficient (but very strong) set of assumptions under which it holds. Finiteness of the right hand side above would be implied, for instance, by demanding that $\|\mu(X)\| \leq K$ and $\Sigma_{ii}(X) \in [\frac{1}{2} + \epsilon, 1]$ for all i.

6.8 Empirical evaluation: further details

In this section with give further details about the synthetic and real-data experiments presented in Section 6.3.

6.8.1 Synthetic experiments

Analytical expressions for divergences between two Gaussians

The closed form expression for the χ^2-divergence between two d-variate normal distributions can be found in Lemma 1 of Nielsen and Nock (2014):

Lemma 41.

$$\chi^2(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)) = \frac{\det(\Sigma_1^{-1})}{\sqrt{\det(2\Sigma_1^{-1}-\Sigma_2^{-1})\det(\Sigma_2^{-1})}}\exp\left(\frac{1}{2}\mu_2'\Sigma_2^{-1}\mu_2 - \mu_1'\Sigma_1^{-1}\mu_1\right)\times$$
$$\times\exp\left(-\frac{1}{4}(2\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})(\frac{1}{2}\Sigma_2^{-1}-\Sigma_1^{-1})^{-1}(2\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2)\right)-1.$$

As a corollary, the following also holds:

Corollary 42. *Chi square divergence between two d-variate Gaussian distributions both having covariance matrices proportional to identity can be computed as:*

$$\chi^2(\mathcal{N}(\mu,\sigma^2 I_d),\mathcal{N}(0,\beta^2 I_d)) = \left(\frac{\beta^2}{\sigma^2\sqrt{2\beta^2/\sigma^2-1}}\right)^d e^{\frac{\|\mu\|^2}{2\beta^2-\sigma^2}}-1$$

assuming $2\beta^2>\sigma^2$. Otherwise the divergence is infinite.

The squared Hellinger divergence between two Gaussians is given in Pardo (2005):

Lemma 43.

$$H^2(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2))$$
$$= 1-\frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{1/2}}\exp\left\{-\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)\right\}.$$

The KL-divergence between two d-variate Gaussians is:

Lemma 44.

$$\mathrm{KL}(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{2}\left(tr\left(\Sigma_2^{-1}\Sigma_1\right)+(\mu_2-\mu_1)^\intercal\Sigma_2^{-1}(\mu_2-\mu_1)-d+\log\frac{|\Sigma_2|}{|\Sigma_1|}\right).$$

Further experimental details

We take $Q^\lambda_{Z|X=x}=\mathcal{N}\left(A_\lambda x+b_\lambda,\epsilon^2 I_d\right)$ and $P_X=\mathcal{N}\left(0,I_{20}\right)$. This results in $Q^\lambda_Z=\mathcal{N}\left(b_\lambda,A_\lambda A_\lambda^\intercal+\epsilon^2 I_d\right)$. We chose $\epsilon=0.5$ and used $\lambda\in[-2,2]$. $P_Z=\mathcal{N}(0,I_d)$.

A_λ and b_λ were determined as follows: Define A_1 to be the $(d,20)$-dimensional matrix with 1's on the main diagonal, and let A_0 be similarly sized matrix with entries randomly sampled i.i.d. unit Gaussians which is then normalised to have unit Frobenius norm. Let v be

a vector randomly sampled from the d-dimensional unit sphere. We then set $A_\lambda = \frac{1}{2}A_1 + \lambda A_0$ and $b_\lambda = \lambda v$.

A_0 and v are sampled once for each dimension $d \in \{1, 4, 16\}$, such that the within each column of Figure 6.1, the distributions used are the same.

6.8.2 Real-data experiments

Variational Autoencoders (VAEs) and Wasserstein Autoencoders (WAEs)

Autoencoders are a general class of models typically used to learn compressed representations of high-dimensional data. Given a *data-space* $\mathcal{X}$ and low-dimensional *latent space* $\mathcal{Z}$, the goal is to learn an *encoder* mapping $\mathcal{X} \to \mathcal{Z}$ and *generator* (or *decoder*[2]) mapping $\mathcal{Z} \to \mathcal{X}$. The objectives used to train these two components always involve some kind of reconstruction loss measuring how corrupted a datum becomes after mapping through both the encoder and generator, and often some kind of regularization.

Representing by θ and η the parameters of the encoder and generator respectively, the objective functions of VAEs and WAEs are:

$$L^{\text{VAE}}(\theta, \eta) = \mathbb{E}_X \left[\mathbb{E}_{q_\theta(Z|X)} \log p_\eta(X|Z) + \text{KL}\left(Q^\theta_{Z|X} \| P_Z\right) \right]$$
$$L^{\text{WAE}}(\theta, \eta) = \mathbb{E}_X \mathbb{E}_{q_\theta(Z|X)} c(X, G_\eta(Z)) + \lambda \cdot D(Q^\theta_Z \| P_Z)$$

For VAEs, both encoder $Q^\theta_{Z|X}$ and generator p_η are *stochastic* mappings taking an input and mapping it to a distribution over the output space. In WAEs, only the encoder $Q^\theta_{Z|X}$ is stochastic, while the generator G_η is deterministic. c is a cost function, λ is a hyperparameter and D is any divergence.

A common assumption made for VAEs is that the generator outputs a Gaussian distribution with fixed diagonal covariance and mean $\mu(z)$ that is a function of the input z. In this case, the $\log p_\eta(X|z)$ term can be written as the l_2^2 (i.e. square of the l_2 distance) between X and its reconstruction after encoding and re-generating $\mu(z)$. If the cost function of the WAE is chosen to be l_2^2, then the left hand terms of the VAE and WAE losses are the same. That is, in this particular case, L^{VAE} and L^{WAE} differ only in their regularizers.

The penalty of the VAE was shown by Hoffman and Johnson (2016) to be equivalent to $\text{KL}(Q^\theta_Z \| P_Z) + I(X, Z)$ where $I(X, Z)$ is the mutual information of a sample and its encoding. For the WAE penalty, there is a choice of which $D(Q^\theta_Z \| P_Z)$ to use; it must only be possible to practically estimate it. In the experiments used in this paper, we considered models trained with the Maximum Mean Discrepency (MMD) Gretton et al. (2012), a kernel-based distance

[2]In the VAE literature, the encoder and generator are sometimes referred to as the *inference network* and *likelihood model* respectively.

on distributions, and a divergence estimated using a GAN-style classifier Goodfellow et al. (2014) leading to WAE-MMD and WAE-GAN respectively, following Tolstikhin et al. (2018b).

Further experimental details

We took a corpus of VAE, WAE-GAN and WAE-MMD models that had been trained with a large variety of hyperparameters including learning rate, latent dimension (32, 64, 128), architecture (ResNet/DCGAN), scalar factor for regulariser, and additional algorithm-specific hyperparameters: kernel bandwidth for WAE-MMD and learning rate of discriminator for WAE-GAN. In total, 60 models were trained of each type (WAE-MMD, WAE-GAN and VAE) leading to 180 models in total.

The small subset of six models exposed in Figures 6.2 and 6.3 were selected by a heuristic that we next describe. However, we note that qualitatively similar behaviour was found in all other models tested, and so the choice of models to display was somewhat arbitrary; we describe it nonetheless for completeness.

Recall that the objective functions of WAEs and VAEs both include a divergence between Q_Z^θ and P_Z. We were interested in considering models from the two extremes of the distribution matching: some models in which Q_Z^θ and P_Z were close, some in which they were distant.

To determine whether Q_Z^θ and P_Z in a model are close, we made use of FID Heusel et al. (2017) scores as a proxy that is independent of the particular divergences for training. The FID score between two distributions over images is obtained by pushing both distributions through to an intermediate feature layer of the *Inception* network. The resulting push-through distributions are approximated with Gaussians and the *Fréchet* distance between them is calculated. Denote by $G_\#(Q_Z^\theta)$ the distribution over reconstructed images, $G_\#(P_Z)$ the distribution over model samples and Q_X the data distribution, where G is the generator and $\#$ denotes the push-through operator. The quantity $\text{FID}\left(Q_X, G_\#(Q_Z^\theta)\right)$ is a measure of quality (lower is better) of the reconstructed data, while $\text{FID}\left(Q_X, G_\#(P_Z)\right)$ is a measure of quality of model samples.

The two FID scores being very different is an indication that P_Z and Q_Z^θ are different. In contrast, if the two FID scores are similar, we cannot conclude that P_Z and Q_Z^θ are the same, though it provides some evidence towards that fact. Therefore, in order to select a model in which matching between P_Z and Q_Z^θ is poor, we pick one for which $\text{FID}\left(Q_X, G_\#(Q_Z^\theta)\right)$ is small but $\text{FID}\left(Q_X, G_\#(P_Z)\right)$ is large (good reconstructions; poor samples). In order to select a model in which matching between P_Z and Q_Z^θ is good, we pick one for both FIDs are small (good reconstructions; good samples). We will refer to these settings as *poor matching* and *good matching* respectively.

Our goal was to pick models according to the following criteria. The six chosen should include: two from each model class (VAE, WAE-GAN, WAE-MMD), of which one from each should exhibit poor matching and one good matching; two from each dimension

$d \in \{32, 64, 128\}$; three with the ResNet architecture and three with the DCGAN architecture. A set of models satisfying these criteria were selected by hand, but as noted previously we saw qualitatively similar results with the other models.

Additional results for squared Hellinger distance

Figure 6.3 we display similar results to those displayed in Figure 6.2 of the main paper but with the H^2-divergence instead of the KL. An important point is that $H^2(A, B) \in [0, 2]$ for any probability distributions A and B, and due to considerations of scale we plot the estimated values $\log\left(2 - \hat{D}^M_{H^2}(\hat{Q}^N_Z \| P_Z)\right)$. Decreasing bias in N of RAM-MC therefore manifests itself as the lines *increasing* in Figure 6.3. Concavity of log means that the reduction in variance when increasing M results in RAM-MC with M=1000 being above RAM-MC with M=10. Similar to those presented in the main part of the paper, these results therefore also support the theoretical findings of our work.

We additionally attempted the same experiment using the χ^2-divergence but encountered numerical issues. This can be understood as a consequence of the inequality $e^{\mathrm{KL}(A,B)} - 1 \leq \chi^2(A, B)$ for any distributions A and B. From Figure 6.2 we see that the KL-divergence reaches values higher than 1000 which makes the corresponding value of the χ^2-divergence larger than can be represented using double-precision floats.

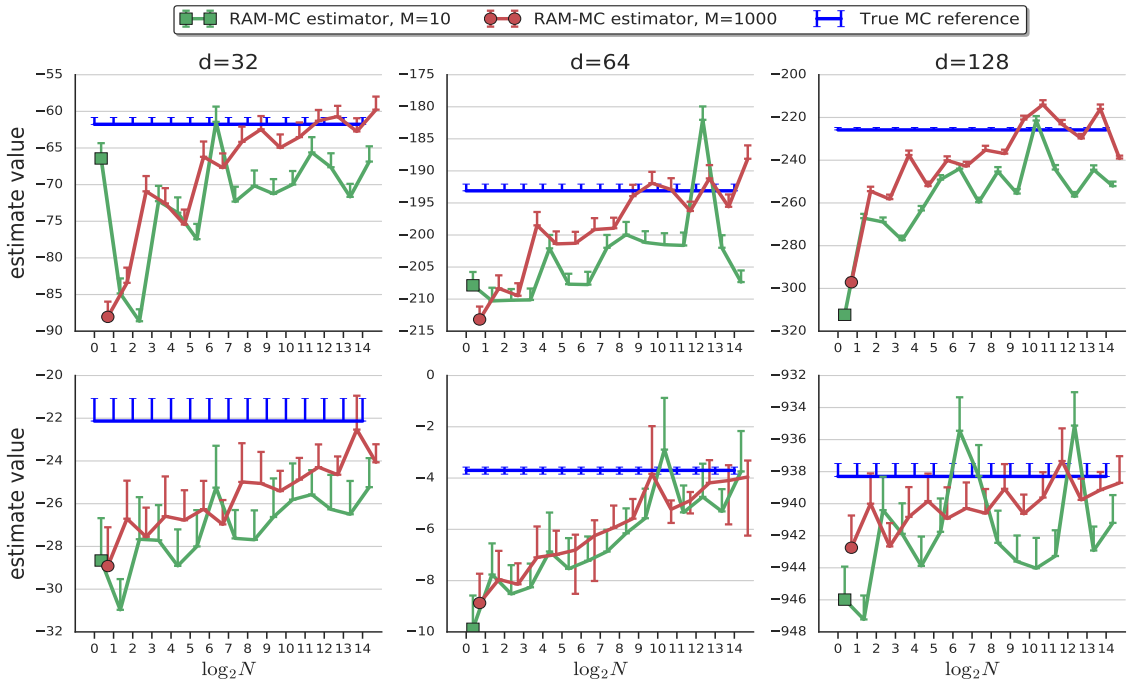


Figure 6.3 Estimating $H^2(Q_Z^\theta \| P_Z)$ in pretrained autoencoder models with RAM-MC as a function of N for $M = 10$ (**green**) and M=1000 (**red**) compared to ground truth (**blue**). Lines and error bars represent means and standard deviations over 50 trials. Plots depict $\log\left(2 - \hat{D}_{H^2}^M(\hat{Q}_Z^N \| P_Z)\right)$ since H^2 is close to 2 in all models. Omitted lower error bars correspond to error bars going to $-\infty$ introduced by log. Note that the approximately *increasing* behaviour evident here corresponds to the expectation of RAM-MC *decreasing* as a function of N. Due to concavity of log, the decrease in variance when increasing M manifests itself as the **red** line (M=1000) being consistently above the **green** line (M=10).

Chapter 7

Conclusion / Future directions

This chapter summarises the work presented in this thesis and discusses where the field is going.

Bibliography

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. (2018). Fixing a broken ELBO. In *ICML*, pages 159–168.

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., et al. (2012). Uk biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv:1701.07875*.

B. Tsybakov, A. (2009). Introduction to nonparametric estimation.

Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. *Technical report 688, Deparment of Statistics, UC Berkeley*.

Balian, R. (1992). *From microphysics to macrophysics*. Springer.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144.

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

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *ICLR*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bollen, K. A. (2014). *Structural equations with latent variables*. John Wiley & Sons.

Bongers, S., Peters, J., Schölkopf, B., and Mooij, J. M. (2016). Structural causal models: Cycles, marginalizations, exogenous reparametrizations and reductions. *arXiv preprint arXiv:1611.06221*.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Cardoso, J.-F. (2001). The three easy routes to independent component analysis; contrasts and geometry. In *Proc. ICA*, volume 2001.

Chalupka, K., Perona, P., and Eberhardt, F. (2015). Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 181–190. AUAI Press.

Chalupka, K., Perona, P., and Eberhardt, F. (2016). Multi-level cause-effect systems. In *The 19th International Conference on Artificial Intelligence and Statistics.*

Chen, L., Tao, C., Zhang, R., Henao, R., and Duke, L. C. (2018a). Variational inference and model selection with generalized evidence bounds. In *ICML*.

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., and Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, pages 460–468.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018b). Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018c). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6572–6583.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Csiszár, I., Shields, P. C., et al. (2004). Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528.

Darmois, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8.

Dash, D. and Druzdzel, M. J. (2001). Caveats for causal reasoning with equilibrium models. *Lecture notes in computer science*, pages 192–203.

De Sa, V. R. (2005). Spectral clustering with two views. In *ICML workshop on learning with multiple views*, pages 20–27.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. (2018). Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. (2017). Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741.

Durrieu, J.-L., Thiran, J.-P., and Kelly, F. (2012). Lower and upper bounds for approximation of the kullback-leibler divergence between gaussian mixture models. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4833–4836. Ieee.

Dziugaite, G., Roy, D., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference, UAI 2015*, pages 258–267.

Eberhardt, F. (2016). Green and grue causal variables. *Synthese*, 193(4):1029–1046.

Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica: Journal of the Econometric Society*, pages 73–92.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., and Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416.

Hein, M. and Bousquet, O. (2005). Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*.

Hero, A. O., Ma, B., Michel, O., and Gorman, J. (2001). Alpha divergence for classification, indexing and retrieval. *Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, Univ. Michigan, Ann Arbor, Tech. Rep. 328*.

Hero, A. O., Ma, B., Michel, O. J. J., and Gorman, J. (2002). Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*.

Hershey, J. R. and Olsen, P. A. (2007). Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a nash equilibrium. *arXiv:1706.08500*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). Beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Hoel, E. P., Albantakis, L., and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795.

Hoffman, M. D. and Johnson, M. J. (2016). ELBO surgery: yet another way to carve up the variational evidence lower bound.

Honkela, T., Hyvärinen, A., and Väyrynen, J. J. (2010). WordICA-emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16(3):277–308.

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(Nov):3387–3439.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773.

Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of Temporally Dependent Stationary Sources. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469, Fort Lauderdale, FL, USA. PMLR.

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

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

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR.

Iwasaki, Y. and Simon, H. A. (1994). Causality and model abstraction. *Artificial Intelligence*, 67(1):143–194.

Jacobsen, J.-H., Smeulders, A., and Oyallon, E. (2018). i-RevNet: Deep Invertible Networks. In *ICLR 2018 - International Conference on Learning Representations*, Vancouver, Canada.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2).

Kim, H. and Mnih, A. (2018). Disentangling by factorising. *arXiv preprint arXiv:1802.05983*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*.

Krishnamurthy, A., Kandasamy, A., Póczos, B., and Wasserman, L. (2014). Nonparametric estimation of Rényi divergence and friends. In *ICML*.

Kumar, A., Rai, P., and Daume, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, pages 1413–1421.

Kumar, A., Sattigeri, P., and Balakrishnan, A. (2018). Variational inference of disentangled latent concepts from unlabeled observations. In *ICLR*.

Lacerda, G., Spirtes, P. L., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence*.

Lai, P. L. and Fyfe, C. (2000). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.

Lederman, R. R. and Talmon, R. (2018). Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2200–2210.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081.

Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. (2016). Adversarial autoencoders. In *ICLR*.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

McKeown, M. J. and Sejnowski, T. J. (1998). Independent component analysis of fMRI data: examining the assumptions. *Human brain mapping*, 6(5-6):368–372.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv:1701.04722*.

Michaeli, T., Wang, W., and Livescu, K. (2016). Nonparametric canonical correlation analysis. In *International Conference on Machine Learning*, pages 1967–1976.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523.

Mooij, J. M. and Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence*.

Mooij, J. M., Janzing, D., Heskes, T., and Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24 (NIPS*2011)*, pages 639–647.

Mooij, J. M., Janzing, D., and Schölkopf, B. (2013). From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 440–448.

Moon, K. and Hero, A. (2014a). Ensemble estimation of multivariate f-divergence. In *2014 IEEE International Symposium on Information Theory*, pages 356–360.

Moon, K. and Hero, A. (2014b). Multivariate f-divergence estimation with confidence. In *NeurIPS*.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Information Theory*, 56(11):5847–5861.

Nielsen, F. and Nock, R. (2014). On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Process. Lett.*, 21(1):10–13.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*.

Nuzillard, D. and Bijaoui, A. (2000). Blind source separation and analysis of multispectral astronomical images. *Astronomy and Astrophysics Supplement Series*, 147(1):129–138.

Oja, E., Kiviluoto, K., and Malaroiu, S. (2000). Independent component analysis for financial time series. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pages 111–116. IEEE.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Osterreicher, F. and Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653.

Pardo, L. (2005). *Statistical inference based on divergence measures*. Chapman and Hall/CRC.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Perez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. In *IEEE International Symposium on Information Theory*.

Poczos, B. and Schneider, J. (2011). On the estimation of alpha-divergences. In *AISTATS*.

Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. (2018). On variational lower bounds of mutual information. In *ICML*.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

Sawada, H., Mukai, R., and Makino, S. (2003). Direction of arrival estimation for multiple source signals using independent component analysis. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 2, pages 411–414. IEEE.

Schölkopf, B., Hogg, D. W., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. (2016). Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398.

Schölkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J., Dalgleish, T., et al. (2014). The cambridge centre for ageing and neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1):204.

Simon, H. A. and Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica: journal of the Econometric Society*, pages 111–138.

Singer, A. and Coifman, R. R. (2008). Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239.

Singh, S. and Poczos, B. (2014). Generalized exponential concentration inequality for Rényi divergence estimation. In *ICML*.

Skitovich, V. P. (1954). Linear forms of independent random variables and the normal distribution law. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 18(2):185–200.

Song, L., Anandkumar, A., Dai, B., and Xie, B. (2014). Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*, pages 640–648.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P. and Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5):833–845.

Steinberg, D. (2011). *The Cholesterol Wars: The Skeptics vs the Preponderance of Evidence*. Academic Press.

Taleb, A. and Jutten, C. (1999). Source separation in post-nonlinear mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2018a). Wasserstein auto-encoders. In *ICLR*.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2018b). Wasserstein auto-encoders. In *ICLR*.

Tomczak, J. M. and Welling, M. (2018). VAE with a VampPrior. *AISTATS*.

Truswell, A. S. (2010). *Cholesterol and beyond: the research on diet and coronary heart disease 1900-2000*. Springer Science & Business Media.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5).

Appendix A

First Appendix

I'm not sure yet what will go here.

Appendix B

Second Appendix

I'm not sure yet what will go here.