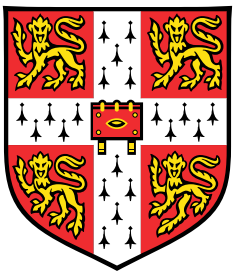
Advances in Latent Variable and Causal Models



Paul Kishan Rubenstein

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Jesus College

July 2020

To Matthew and Nanaji

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit of 65,000 words for the Engineering Degree Committee, including appendices, footnotes, tables and equations.

Paul Kishan Rubenstein
July 2020

Acknowledgements

This thesis is the culmination of many years of study, during which time I have been blessed to have interacted with numerous outstanding people. I thank my supervisors Carl Edward Rasmussen and Bernhard Schölkopf, without whose trust I would not have taken the long journey leading to this work. Ilya Tolstikhin has been a friend and mentor and has guided me through both the ups and downs of the PhD and I thank him for the support, encouragement, feedback and advice he has given me throughout. He has been a role model both within research and without, and I am eternally grateful. Joris Mooij gave me valuable supervision early in my PhD, and I learned a great deal through my collaboration with him and Stephan Bongers. Sebastian Weichwald was an excellent collaborator who showed me how productive and satisfying a close collaboration can be, and that digging deep into a topic is best done with company. I thank him and Anna Weininger für deren Geduld, als ich nach Tübingen umgezogen bin und kaum Deustch konnte. Luigi Gresele confirmed again that close collaborations are the most enjoyable and productive, and I learned a great deal from him, as well as from all of my other collaborators: Josip Djolonga, Carlos Riquelme, Olivier Bousquet, Arash Mehrjou, Francesco Locatello, Dominik Janzing, Moritz Grosse-Wentrup, Philipp Hennig, Dominik Roblek, Yunpeng Li, Sylvain Gelly, Michael Tschannen, Mario Lucic and Julius von Kügelgen.

There are also many people with whom I have not collaborated, but have nonetheless enriched my life during my PhD. I thank Nilesh Tripuraneni, James Heald, Hannah Sheahan, Mohsen Sadhegi, Matej Balog, Mark van der Wilk and the rest of CBL, who made my time in Cambridge memorable. From the MPI in Tübingen, I thank Mateo Rojas-Carulla, Eduardo Pérez-Pellitero, Diego Agudelo-España, Sebastian Gomez-Gonzalez, Dieter Büchler, Niki Kilbertus, Giambattista Parascandolo, John Bradshaw, Alessandro Ialongo, Alex Neitz, Matthias Bauer, Adam Scibior, Jonas Kübler and the rest of the team.

Magda has made me a better person, and I thank her for supporting me at all times, particularly during my time in Zurich. Finally, I thank Mum, Dad, Josh and the rest of my family who give me a solid foundation based on unconditional love.

Abstract

This thesis considers three different areas of machine learning concerned with the modelling of data, extending theoretical understanding in each of them. First, the estimation of f-divergences is considered in a setting that is naturally satisfied in the context of autoencoders. By exploiting structural assumptions on the distributions of concern, the proposed estimator is shown to exhibit fast rates of concentration and bias-decay. In contrast, in much of the existing f-divergence estimation literature, fast rates are only obtainable under strong conditions that are difficult to verify in practice. Next, novel identifiability results are presented for nonlinear Independent Component Analysis (ICA) in a multi-view setting, extending the scarce literature of known identifiability results for nonlinear ICA. A result of particular note is that if one noiseless view of the sources is supplemented by a second view that is appropriately corrupted by source-level noise, the sources can be fully reconstructed from the observations up to tolerable ambiguities. This setting is applicable to areas such as neuroimaging, where multiple data modalities may be available. Finally, a framework is introduced to evaluate when two causal models are consistent with one another, meaning that a correspondence can be established between them such that reasoning about the effects of interventions in both models agree. This can be used to understand when two models of the same system at different levels of detail are consistent, and has application to the problem of causal variable definition. This work has broad implications to the causal modelling process in general, as there is often a mismatch between the level at which measurements are made and the level at which the underlying 'true' causal structure exists, yet causal inference algorithms generally seek to discover causal structure at the level of measurements.

Contents

List of Figures

List of Tables

Nomenclature

General Abbreviations

i.i.d. Independent and identically distributed

CCA Canonical Correlation Analysis

ELBO Evidence Lower Bound

GAN Generative Adversarial Network

HD Heart Disease

HDL High-Density Lipoprotein

ICA Independent Component Analysis

IPM Integral Probability Metric

LDL Low-Density Lipoprotein

LVM Latent Variable Model

MC Monte-Carlo

MI Mutual Information

MMD Maximum Mean Discrepancy

MWS Minibatch Weighted Sample

RAM Random Mixture estimator

RAM-MC Random Mixture estimator with Monte-Carlo sampling

RKHS Reproducing Kernel Hilbert Space

SDV Sufficiently Distinct Views

SEM Structural Equation Model

TC Total Correlation

VAE Variational Autoencoder

WAE Wasserstein Autoencoder

f-divergence Abbreviations

H^2 Squared-Hellinger

JS Jensen-Shannon

KL Kullback-Leibler

TV Total Variation

Mathematical Notation

$\#$ Push-forward operator

$\text{do}(X_i = x_i)$ Do-intervention setting variable X_i to value x_i

$\hat{Q}_Z^N$, $\hat{q}_N(z)$ Empirical approximation to aggregate posterior based on N samples, density

$\mathbb{E}$ Expectation operator

$\mathbb{P}$ Probability operator

$\mathbf{X}^N$ Set of N observations of X

$\mathbf{Z}^M$ Set of M observations of Z

$\mathcal{G}$ Causal graph

$\mathcal{H}$ Function class

$\mathcal{I}_X$ Interventions in $\mathcal{M}_X$

$\mathcal{M}_X$ Structural Equation Model over X

$\mathcal{N}(\mu, \Sigma)$ Gaussian distribution with mean μ and covariance matrix Σ

$\mathcal{P}(\mathcal{X})$ The set of probability distributions over $\mathcal{X}$

$\mathcal{S}_X$ Structural equations in $\mathcal{M}_X$

$\mathcal{X}$ Data space

$\mathcal{Z}$ Latent space

ω Function mapping between sets of interventions

τ Function mapping between causal variables

A Matrix

D, D_f Divergence

E Exogenous noise variable

$H(Q_Z)$ Differential entropy of Q_Z

$I(X, Z)$ Mutual information of variables X and Z

$L(\theta, \phi)$ Loss function with parameters θ and ϕ

N Noise variable

OT_c Optimal transport distance with respect to metric c

P_X^θ, $p^\theta(x)$ Model distribution with parameter θ, density

$P_X^{\mathrm{do}(\cdot)}$ Distribution over X after intervention $\mathrm{do}(\cdot)$

P_Z, $p(z)$ Prior distribution over latent space, density

Q_X, $q(x)$ Data distribution, density

Q_Z, $q(z)$ Aggregate posterior distribution over latent space, density

$TC(Q_Z)$ Total correlation of Q_Z

X Random variable over data space

x Particular value taken by X

Z Random variable over latent space

Chapter 1

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. For instance, we understand 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 this happens is a consequence both of evolution as well as a lifetime of experience. Machine learning models, in contrast, generally have neither of these from which to benefit, and in spite of the significant advances that have been made in recent years, the modelling of structured data remains a broad and active area of research.

This thesis considers three different areas of machine learning concerned with the modelling of data, extending theoretical understanding in each of them. These areas are:

1. Autoencoders, a family of generative models, the goal of which is to model the unknown distribution of data given i.i.d. samples;
2. Independent Component Analysis, the goal of which is to unmix or separate signals from independent sources that have been mixed together; and
3. Causality, a broad area concerned with the modelling and inference of causal relations between random variables.

Autoencoders are a family of models that involve the introduction of a *latent space* and associated *encoder* and *generator*. The encoder maps from the data space to the latent space and the generator maps in the reverse direction. In recent years, substantial advances have been made in this area as part of the general progress in machine learning and deep learning in particular. Despite this, fundamental questions remain.

The estimation and minimisation of divergences between distributions in the latent spaces of autoencoders are important problems in modern research. Of particular interest are divergences between distributions known as the *prior* and *aggregate posterior*, the former being a user-specified distribution and the latter being induced by the unknown data distribution in conjunction with the learned encoder. Chapter 3 presents and studies an estimator for a class of divergences known as *f-divergences*, deriving bounds on the rate of decay of the bias and finite sample concentration bounds. Although this estimator may be applied in other settings, the structural assumptions required for its use are naturally satisfied in the estimation of divergences between priors and aggregate posteriors in the autoencoder setting.

Much of the literature on f-divergence estimation considers settings in which weak knowledge is assumed about the distributions for which the divergence is being estimated. In such settings, the number of samples required to estimate the divergence typically grows exponentially in the dimension of the space over which the distributions are defined, unless the associated densities satisfy strong smoothness assumptions that are difficult to verify in practice. By exploiting the natural structure present in the autoencoder setting, superior rates are obtained in Chapter 3 with only mild and easily verifiable additional assumptions. These results additionally have implications to existing work elsewhere in the literature by giving a rigorous foundation to heuristic proposals in related settings.

Independent Component Analysis (ICA) assumes that data are generated by independent sources that are mixed together. This is formalised by introducing a latent space over which a factorised *source distribution* is assumed. Observations are obtained by passing the sources through a *mixing function*, each coordinate of which is a function of several sources. Given a dataset of observations, the goal of ICA is to invert the unknown mixing function and recover the independent sources.

In addition to the derivation of practical algorithms, one of the main lines of enquiry in the ICA community is the search for *identifiability results*. Specifying an ICA problem requires making assumptions, for instance on the source distribution or the mixing functions, thus defining a restricted family of models. An ICA problem is *unidentifiable* if there are multiple fundamentally different models in this restricted family that result in the same data distribution, and is *identifiable* otherwise. Identifiability results allow us to understand conditions under which it is in principle possible—or impossible—to recover the latent sources up to tolerable ambiguities.

At the one extreme of very strong assumptions, identifiability holds if the mixing functions are linear and at most one component of the latent sources is Gaussian. At the other extreme, the ICA problem is unidentifiable if no assumptions on the mixing functions or sources are made. This line of research thus seeks to identify settings between these extremes where identifiability still holds. Chapter 4 extends the scarce literature of identifiability results for

nonlinear ICA by considering a novel setting in which *multiple views* of the sources through different mixing functions are available. In particular, if one noiseless view of the sources is supplemented by a second view that is appropriately corrupted by source-level noise, the sources can be fully reconstructed from the observations up to tolerable ambiguities. These theoretical results have important practical implications, as in many applications such as neuroimaging, multiple data modalities may be available. These results show that jointly using these different modalities can result in recovery of the sources under weaker conditions than when using them separately.

The field of causality is largely concerned with the inference of causal relationships between variables from data. These are distinct from statistical relationships, since if two variables are statistically dependent, either one may causally influence the other or they could both be influenced by a third. Causal inference algorithms generally operate on vectorial data in which each component is assumed to be a variable that exhibits causal relations with some subset of the other components, the goal then being to identify these relations. But in many realistic scenarios, there is no guarantee that the components of collected data are well-defined causal variables in this sense.

For example, early investigations into the influence of blood cholesterol on the risk of developing heart disease found conflicting results. Some studies found that raising total blood cholesterol levels increased the risk, while others found the opposite. Later, it was discovered that there are in fact two types of cholesterol, one reducing the risk, the other raising it. Attempting to identify the causal relation between their *sum* and the risk was therefore doomed to failure: total blood cholesterol is not a well-defined causal variable in relation to the risk of developing heart disease.

More generally, though any physical system in the real world exhibits causal structure at some level of detail, measurements are typically made at a more coarse-grained or abstract level. Chapter 5 seeks to identify when these 'higher-level' variables admit an interpretation as well-defined causal variables. This is done by introducing a framework for understanding when two causal models are consistent with one another: if a high-level model is consistent with a low-level model that is physically grounded in real causal relations, it admits a causal interpretation through its connection to the low-level model. This work has implications to causal modelling generally by giving a handle on the problem of defining causal variables.

1.1 Outline and Contributions

This thesis is organised as follows.

- Chapter 2 is an overview of the literature on generative modelling with Latent Variable Models (LVMs), providing important background for Chapters 3 and 4. This defines and discusses LVMs, divergences including the family of f-divergences, density ratio estimation and examples of generative models including autoencoders.
- Chapter 3 presents and studies an estimator for f-divergences between distributions with particular application to the setting of autoencoders. The natural structural assumptions that hold in this setting make it possible to estimate the divergences with fast rates. In contrast, in much of the existing f-divergence estimation literature, fast rates are only attainable under strong assumptions that would be hard to verify in practice. This chapter is based on the *NeurIPS* conference paper Rubenstein et al., 2019.
- Chapter 4 presents novel identifiability results for nonlinear ICA, extending the scarce literature of such results. A multi-view setting is considered in which multiple observations of the sources are simultaneously available through different mixing functions. In particular, supplementing a noiseless view of the sources with a second appropriately corrupted view leads to the model being identifiable. This has application to practical scenarios in which multiple data modalities are available. This chapter is based on the *UAI* conference paper Gresele, Rubenstein et al., 2019.
- Chapter 5 presents a framework for understanding when two casual models at different levels of detail are consistent with one another, showing how high-level causal variables can arise as functions of lower-level variables. This has implications to the understanding of causal modelling by shedding light on the definition of causal variables, and in particular highlights the importance of considering interventions as part of the causal modelling process. This chapter is based on the *UAI* conference paper Rubenstein, Weichwald et al., 2017.
- Chapter 6 concludes and additional materials for the main chapters are included in Appendices A, B and C.

All of the work presented in this thesis was conducted with my numerous collaborators. The conference papers cited above on which the chapters of this thesis are based are repeated with a full list of collaborators:

> Paul K Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme and Ilya Tolstikhin. "Practical and Consistent Estimation of f-Divergences". *Advances in Neural Information Processing Systems (NeurIPS)*. 2019

> Luigi Gresele*, Paul K Rubenstein*, Arash Mehrjou, Francesco Locatello and Bernhard Schölkopf. "The Incomplete Rosetta Stone Problem: Identifiability

Results for Multi-view Nonlinear ICA". *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*. *Joint first authorship. 2019

Paul K Rubenstein*, Sebastian Weichwald*, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup and Bernhard Schölkopf. "Causal consistency of structural equation models". *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. *Joint first authorship. 2017

In addition, the following papers were also written during my PhD but are not discussed in this thesis.

Paul K Rubenstein, Stephan Bongers, Bernhard Schölkopf and Joris M Mooij. "From deterministic ODEs to dynamic structural causal models". *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*. 2018

Paul K Rubenstein, Ilya Tolstikhin, Philipp Hennig and Bernhard Schölkopf. "Probabilistic Active Learning of Functions in Structural Causal Models". *Causality Workshop of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. 2017

Paul K Rubenstein, Bernhard Schölkopf and Ilya Tolstikhin. "Learning Disentangled Representations with Wasserstein Auto-Encoders". *International Conference on Learning Representations (ICLR), Workshop Track*. 2018

Paul K Rubenstein, Bernhard Schölkopf and Ilya Tolstikhin. "Wasserstein auto-encoders: Latent dimensionality and random encoders". *International Conference on Learning Representations (ICLR), Workshop Track*. 2018

Paul K Rubenstein, Bernhard Schoelkopf and Ilya Tolstikhin. "On the Latent Space of Wasserstein Auto-Encoders". *arXiv preprint arXiv:1802.03761* (2018)

Paul K Rubenstein, Yunpeng Li and Dominik Roblek. "An Empirical Study of Generative Models with Encoders". *arXiv preprint arXiv:1812.07909* (2018)

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly and Mario Lucic. "On mutual information maximization for representation learning". *International Conference on Learning Representations (ICLR)*. 2020

Dominik Janzing, Paul Rubenstein and Bernhard Schölkopf. "Structural causal models for macro-variables in time-series". *arXiv preprint arXiv:1804.03911* (2018)

Julius von Kügelgen, Paul K Rubenstein, Bernhard Schölkopf and Adrian Weller. "Optimal experimental design via Bayesian optimization: active causal structure

learning for Gaussian process networks". *NeurIPS 2019 Workshop "Do the right thing": Machine Learning and Causal Inference for Improved Decision Making.* 2019

Chapter 2

Generative Modelling with Latent Variable Models

This chapter introduces key ideas in the literature of generative modelling with latent variable models relevant to this thesis: Chapter 3 presents learning theoretic results for divergence estimation in the latent spaces of autoencoders, a type of latent variable generative model; Chapter 4, which concerns Independent Component Analysis (ICA), presents identifiability results for a particular class of latent variable models.

2.1 Introduction

Suppose that a dataset of samples, drawn independently and identically distributed (i.i.d.) from some unknown data distribution Q_X, is given. The high level goal of generative modelling is to learn a model distribution P_X that approximates the unknown data distribution Q_X based on these samples. Latent variable models are a flexible way to specify such distributions, and work by composing simple distributions over (unobserved) latent variables with mappings to the observed data space. The applications of latent variable generative modelling are diverse, since having such a model of the data may be desirable for a variety of reasons, for instance: to artificially generate new samples of data (Goodfellow et al., 2014; Oord et al., 2016); to perform compression (Townsend et al., 2019a; Townsend et al., 2019b); to unsupervisedly learn features for transfer to other tasks (Tschannen et al., 2018; Donahue and Simonyan, 2019); or to extract latent structure present in the data (Hyvärinen and Oja, 2000).

The goal of this chapter is to introduce the necessary background and context for Chapters 3 and 4 of this thesis. Although both concern latent variable generative models, these chapters are set in different niches of the machine learning literature. Chapter 3 presents learning

theoretic results that are relevant to the *deep generative modelling community*, a field centred around the generation of artificial data, for which Sections 2.2, 2.4 and 2.6 of this chapter are relevant. Chapter 4 concerns the *Independent Component Analysis (ICA) community*, in which the goal is the inference of latent sources, for which Sections 2.2 and 2.5 of this chapter are relevant.

In the deep generative modelling community, the problem of generative modelling is made precise with specification of a choice of divergence D and family of distributions P_X^θ with parameter $\theta \in \Theta$, the goal then being

$$\min_{\theta \in \Theta} : \; D\left(P_X^\theta, Q_X\right). \tag{2.1}$$

There are two main challenges in practically implementing and solving this problem. First, when the data are drawn from complex high dimensional distributions, how can this be modelled with a parameterised distribution that is computationally tractable? Second, what are appropriate choices of divergences, and how can they be estimated or minimised with respect to the parameters θ?

In general, Q_X is unknown and can only be approximated as an empirical distribution based on available samples, often denoted $\widehat{Q}_X$. Thus, although the goal is to minimise a divergence between the model distribution and the true data distribution Q_X, the problem often reduces to minimising some divergence or loss function involving instead $\widehat{Q}_X$. It may be tempting simply to replace Q_X with $\widehat{Q}_X$ in Equation 2.1, however the distinction between the underlying data distribution and empirical distribution can sometimes require subtle reasoning for two main reasons. First, one could easily overfit if care is not taken, learning only to reproduce the observed data. Second, in some cases divergences between model and empirical distributions may not be well-defined. For instance, maximum likelihood learning is equivalent to minimising the KL-divergence in Equation 2.1 when P_X^θ is absolutely continuous with respect to Q_X. But when P_X^θ is a continuous distribution and $\widehat{Q}_X$ is an empirical distribution, the KL-divergence is not well-defined, although the maximum likelihood interpretation still is. For other choices of divergences, for instance Integral Probability Metrics (see Section 2.4.1), this latter issue may not arise.

In the ICA community, the goal is to use the learned model P_X^θ to infer the values of latent variables. Since latent variables are by definition unknown, it is important that if multiple solutions to the generative modelling problem exist, they should correspond to similar models with regard to the use of the latent variables. That is, if distinct parameters $\theta_1 \neq \theta_2$ are such that $P_X^{\theta_1} = P_X^{\theta_2} = Q_X$, the corresponding models should be strongly related; for instance, corresponding to the same model but with permuted coordinates over the latent variables. Such results are known as *identifiability results* and are important in the theoretical study

of ICA algorithms. ICA will be discussed in more detail in Chapter 4, in which novel identifiability results are presented.

2.2 Latent Variable Models

A Latent Variable Model (LVM) is a way to specify complex distributions over potentially high-dimensional spaces using simple components. These are flexible models that are used widely in the machine learning literature and as such, different parts of the literature often use different terminology to describe fundamentally similar ideas. In the following, two sets of nomenclature will be introduced for the deep generative modelling and ICA communities.

Definition 2.1 (Latent Variable Model)**.** *A Latent Variable Model (LVM) over a data space $\mathcal{X}$ consists of a distribution P_Z over a low dimensional latent space $\mathcal{Z}$ together with conditional distributions $P_{X|Z}$. Together, these induce a distribution P_X over the data space.*

In the deep generative modelling community, P_Z is referred to as a *prior* or *noise distribution* and is usually fixed to be some simple distribution such as a unit Gaussian or uniform distribution. In the ICA community, P_Z is referred to as a *source distribution* and may be specified only implicitly through some assumed properties, such as being a factorised distribution. Usually the letter S denotes the corresponding sources in the ICA literature, but for consistency this thesis will use Z.

The conditional distributions can be thought of as a mapping $g : \mathcal{Z} \to \mathcal{P}(\mathcal{X})$ from elements of $\mathcal{Z}$ to distributions over $\mathcal{X}$. If all of these distributions are Dirac delta distributions (where all probability mass is placed at a single point), the mapping g can be seen as a function $g : \mathcal{Z} \to \mathcal{X}$.

In the deep generative modelling community, the conditional distributions are referred to as *generators* or *decoders*, and when given parameter θ may be written as $P^\theta_{X|Z}$ or g^θ. If the conditional distributions they correspond to are Dirac delta distributions, the generators are called *deterministic*, otherwise they are *stochastic* (sometimes *probabilistic*). In the ICA community, the generators are generally deterministic and are known as *mixing functions*, and are usually denoted by f.

When practically implemented in modern applications, generators are often realised as neural networks. This is straightforward in the deterministic case: g^θ is simply a function and can thus be represented as a deep network with parameters θ. If the generator is stochastic, it can still be explicitly represented as a neural network provided that the conditional distributions are sufficiently structured. For instance, if the conditional distributions are Gaussian with varying mean and covariance, g^θ can be represented as a neural network with two outputs, one for the mean and one for the covariance.

For a fixed choice of parameter θ, P_Z and $P^\theta_{X|Z}$ specify a joint distribution P^θ_{XZ} over $\mathcal{X} \times \mathcal{Z}$ and thus a distribution P^θ_X over the data space $\mathcal{X}$. For simplicity, it will be assumed that densities of all relevant distributions exist, so that this is equivalent to specifying P^θ_X via the integral

$$p^\theta(x) = \int p^\theta(x|z)p(z)dz.$$

In some special cases (e.g. normalising flows in Section 2.3.1), the density $p^\theta(x)$ may be tractable. In most cases in the deep generative modelling community, the integral is intractable, meaning that P^θ_X has unknown density in practice. Despite this, LVMs are useful here because samples from P^θ_X can be drawn easily: one samples first a value $z \sim P_Z$ and then $x \sim P^\theta_{X|Z=z}$. All relevant distributions can be chosen so that these sampling procedures are simple, e.g. if P_Z and all $P^\theta_{X|Z}$ are Gaussian.

2.3 Other generative models

This section briefly covers two other generative models that are encountered across the machine learning literature. The first, *normalising flows*, are a specialisation of the general formulation of LVMs as introduced in the previous section. The second, *autoregressive models*, do not use latent variables and are therefore fundamentally different.

Both of these families of models have in common that the likelihood of observed data can be calculated directly, and thus can be fit straightforwardly by maximum likelihood learning. In practice, this typically means using the log-likelihood as an objective function which is maximised with respect to the model parameters using stochastic gradient methods.

2.3.1 Normalising Flows

Normalising flows, first introduced by Tabak and Vanden-Eijnden, 2010 and Tabak and Turner, 2013 and subsequently popularised within the machine learning community by Dinh et al., 2014 and Rezende and Mohamed, 2015, are type of LVM for which the generator f^θ is deterministic and invertible with known Jacobian and inverse. By the change of variable formula, the density $p^\theta(x)$ can be explicitly calculated in terms of the prior $p(z)$ and generator f^θ

$$p^\theta(x) = p(z) \left| \det \frac{\partial f^\theta}{\partial z} \right|^{-1}, \tag{2.2}$$

where $x = f^\theta(z)$. The setting of normalising flows applies naturally to the problem of ICA due to the assumption that f^θ be invertible.

By the chain rule, if two functions have known Jacobian and inverse, their composition also has known Jacobian and inverse. Thus, f^θ can be specified by composing many layers of simple components. One of the lines of research within the normalising flow community is the development of such layers, for instance Planar Flows (Rezende and Mohamed, 2015), Nonlinear Independent Components Estimation (NICE) (Dinh et al., 2014), Real Non-Volume Preserving (RealNVP) (Dinh et al., 2016), Masked Autoregressive Flow (Papamakarios et al., 2017) and Inverse Autoregressive Flow layers (Kingma et al., 2016).

The advantage of normalising flows over the more general LVMs introduced in the previous chapter is that the likelihood of input data can be calculated exactly via Equation 2.2, while it is still possible to featurise inputs and generate samples via f^θ and its inverse. The challenge of these models is that 'standard' architectures (e.g. convolutions) cannot be straightforwardly applied due to the invertibility constraint. Moreover, the assumption that the data and latent space have the same dimension may be restrictive in some settings, for example when modelling high-dimensional data such as images.

2.3.2 Autoregressive Models

In contrast to LVMs, autoregressive models do not involve the introduction of a latent variable and prior distribution. Instead, the distribution over the observable variables is modelled directly by making use of the factorisation of the joint distribution

$$p^\theta(x) = \prod_{i=1}^{n} p^\theta(x_i|x_{j<i}), \tag{2.3}$$

where x is an n-dimensional vector and $x_{j<i}$ is the vector of components with index smaller than i. Any joint distribution can be factorised this way so no assumptions need to be made in order to use this decomposition, though part of the modelling process may be to drop some dependencies: for example, rather than depending on all previous components, x_i may depend only on some subset of previous components. Each component $p^\theta(x_i|x_{j<i})$ can be modelled with a neural network, with inputs $x_{j<i}$ and output a distribution over x_i. Such models can be used for both discrete and continuous data by varying the choice of output distributions.

The main considerations in the modelling process are deciding the order in which to enumerate the components x_i in Equation 2.3, whether to drop any dependencies, and precisely how to model the factors $p^\theta(x_i|x_{j<i})$.

Prominent methods in the literature include PixelRNN (Van Oord et al., 2016) which models images by enumerating each pixel and colour-channel by row and column and using a recurrent neural network to model the distribution over each pixel conditioned on all previous pixels. PixelCNN (Van den Oord et al., 2016) instead models the distribution of a pixel using a

masked convolution centred at that pixel, such that it is a function only of previous pixels which are also spatially close. WaveNet (Oord et al., 2016) models audio by modelling the wave amplitude at each time step as a function of those at previous time steps by using dilated convolutions, achieving long-range dependencies with a convolutional architecture.

Similar to normalising flow models, autoregressive models have the advantage over LVMs that it is possible to exactly calculate the likelihood of any observation, since this reduces to just evaluating each of the factors in Equation 2.3. They can also be used to generate samples, though this can be significantly slower than for LVMs, since each component needs to be sampled sequentially given previous components. In particular, this means that high-dimensional data such as images or audio can be slow to sample. In contrast to LVMs, autoregressive models cannot be used to featurise data.

2.4 Divergences

A divergence is a notion of dissimilarity between pairs of distributions that is weaker than a metric.

Definition 2.2. *A divergence D is a mapping $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R} \cup \{\infty\}$ such that*

- $D(P, Q) \geq 0$ *for any distributions* $P, Q \in \mathcal{P}(\mathcal{X})$,
- $D(P, Q) = 0$ *if and only if* $P = Q$,

where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability distributions on $\mathcal{X}$.

As a technical side-note, this definition is very general, and does not assume that the distributions admit densities; the only requirement is that P and Q be probability measures over $\mathcal{X}$ with the same σ-algebra Σ over $\mathcal{X}$. Σ is the set of all subsets or 'events' assigned probability mass under the distributions. For most typical settings, where $\mathcal{X}$ would be a subset of some Euclidean space $\mathbb{R}^d$, Σ would canonically be the set of Lebesgue-measurable subsets. $P = Q$ if and only $P(A) = Q(A)$ for all $A \in \Sigma$, meaning that they assign the same probability mass to each possible event. If P and Q admit densities p and q, then $P = Q$ if and only if $p = q$ almost everywhere as functions $\mathcal{X} \to \mathbb{R}$.

A metric is additionally symmetric and obeys the triangle inequality. For a divergence $D(P_X^\theta, Q_X)$ to be useful in the context of generative modelling, it must be possible to minimise it with respect to the parameters θ. There are two main families of divergences that are used in the machine learning literature generally. At a high level, Integral Probability Metrics (IPMs) can be thought of as comparing distributions by considering the pointwise difference between their densities, while f-divergences can be thought of as considering the ratio of their densities. These two families are discussed in the rest of this section.

2.4.1 Integral Probability Metrics

Definition 2.3. *An Integral Probability Metric (IPM) is a divergence that can be written as*

$$D_{\mathcal{H}}(P,Q) = \sup_{h\in\mathcal{H}} \left| \int h(x)dP(x) - \int h(x)dQ(x) \right|,$$

for some restricted function class $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R}\}$.

Elements of $\mathcal{H}$ are referred to as *witness functions*. If $\mathcal{H}$ is too small, $D_{\mathcal{H}}$ may not be a divergence. For instance, taking $\mathcal{H}$ to contain only the constant 0 function results in $D_{\mathcal{H}}(P,Q) = 0$ for any P and Q. On the other hand, if $\mathcal{H}$ is too rich then $D_{\mathcal{H}}$ may not be useful: taking $\mathcal{H}$ to be the set of all real-valued measurable functions leads to the trivial case (Theorem 1, Sriperumbudur et al., 2009)

$$D_{\mathcal{H}}(P,Q) = \begin{cases} 0 & \text{if } P = Q, \\ \infty & \text{if } P \neq Q. \end{cases}$$

Provided $\mathcal{H}$ is sufficiently rich that $D_{\mathcal{H}}$ is a divergence, it is in fact a metric, obeying the triangle inequality and symmetry. These properties are inherited from the function $d(x,y) = |x-y|$.

Commonly encountered IPMs include (Müller, 1997; Sriperumbudur et al., 2009):

- The 1-Wasserstein distance, where $\mathcal{H}$ is the set of all functions with Lipschitz constant 1 with respect to some base metric;
- The Total Variation distance, where $\mathcal{H}$ is the set of all functions with infinity norm 1; and
- The Maximum Mean Discrepancy, where $\mathcal{H}$ is the set of functions in a reproducing kernel Hilbert space with norm at most 1 induced by some kernel k (Gretton et al., 2012).

2.4.2 f-divergences

f-divergences, sometimes called ϕ-divergences, are a family of divergences that compare pairs of distributions via their density ratio, and are widespread and important in the statistics literature (Csiszár and Shields, 2004; Liese and Vajda, 2006; Tsybakov, 2009). Their estimation is studied in Chapter 3.

Definition 2.4. *Let f be a convex real-valued function defined on $(0,\infty)$ such that $f(1) = 0$, and let P and Q be distributions with densities $p(x)$ and $q(x)$. The f-divergence between P*

and Q is defined as

$$D_f(P, Q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx.$$

This is defined for distributions P and Q for which P is absolutely continuous with respect to Q, meaning informally that the density ratio $p(x)/q(x)$ is finite for all x with mass under Q, and is usually taken to be ∞ otherwise.

A useful property of f-divergences is that for any constant c, replacing $f(u)$ by $\tilde{f}(u) = f(u) + c(u-1)$ does not change the divergence D_f:

$$\begin{aligned} D_{\tilde{f}}(P, Q) &= \int \left[f\left(\frac{p(x)}{q(x)}\right) + c\left(\frac{p(x)}{q(x)} - 1\right)\right] q(x)dx \\ &= \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx + c\int p(x) - q(x)dx \\ &= D_f(P, Q), \end{aligned}$$

where the last equality holds since $p(x)$ and $q(x)$ integrate to 1. It is often convenient to work with $f_0(u) := f(u) - f'(1)(u-1)$ which is decreasing on $(0, 1)$, increasing on $(1, \infty)$ and satisfies $f_0'(1) = 0$.

To see that f-divergences are indeed divergences, consider first non-negativity, which follows from convexity of f and Jensen's inequality:

$$\begin{aligned} D_f(P, Q) &= \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx \\ &\geq f\left(\int \frac{p(x)}{q(x)} q(x)dx\right) \\ &= f(1) = 0. \end{aligned} \tag{2.4}$$

To show that $D_f(P, Q) = 0$ if and only if $P = Q$, note first that if $P = Q$ then $D_f(P, Q) = 0$, since $f(1) = 0$. To see that $D_f(P, Q) > 0$ for any $P \neq Q$, observe that Inequality 2.4 above is strict for any f that is strictly convex. This also holds for any f that does not have constant gradient in any neighbourhood of 1, since the f_0 associated to any such f is strictly positive on $\mathbb{R}_+ \setminus \{1\}$. For any $P \neq Q$, the distribution Q must put positive mass in areas for which $p(x)/q(x) \neq 1$ and so $D_{f_0}(P, Q)$ must be positive. It follows that $D_f = D_{f_0}$ is also positive. Hence 'most' choices of f lead to valid f-divergences.

Different choices of f yield several commonly encountered divergences, including the Kullback-Leibler (KL), Jensen-Shannon (JS), Total Variation (TV), χ^2 and α-divergences as well as the lesser known β-divergences (Osterreicher and Vajda, 2003). This family of symmetric

Table 2.1 f and f_0 of divergences referenced in this thesis.

f-divergence	$f_0(x)$	Other typical $f(x)$
Kullback-Leibler (KL)	$x \log x - x + 1$	$x \log x$
Total Variation (TV)	$\frac{1}{2}\lvert 1 - x \rvert$	-
χ^2	$x^2 - 2x$	$(x-1)^2,\ x^2 - 1$
Squared-Hellinger (H^2)	$2(1 - \sqrt{x})$	$(\sqrt{x} - 1)^2$
Jensen-Shannon (JS)	$(1+x)\log(\frac{2}{1+x}) + x \log x$	-
α-divergence, $-1 < \alpha < 1$	$\frac{4}{1-\alpha^2}\left(1 - x^{\frac{1+\alpha}{2}}\right) - \frac{2(x-1)}{\alpha - 1}$	$\frac{4}{1-\alpha^2}\left(1 - x^{\frac{1+\alpha}{2}}\right)$
β-divergence, $\beta > 0,\ \beta \neq \frac{1}{2}$	$\frac{1}{1-\frac{1}{\beta}}\left[(1+x^\beta)^{\frac{1}{\beta}} - 2^{\frac{1}{\beta}-1}(1+x)\right]$	-

divergences is parameterized by $\beta \in (0, \infty]$ and includes the squared-Hellinger (H^2, $\beta = \frac{1}{2}$), Jensen-Shannon ($\beta = 1$) and Total Variation ($\beta = \infty$). Table 2.1 lists the f_0 and other commonly encountered choices of f for the divergences considered in Chapter 3.

2.5 Density Ratio Estimation

A natural way to distinguish between two distributions P and Q with overlapping support is to consider the problem of classifying between draws from each distribution. Suppose that samples from P are labelled as class 1, and samples from Q as class 0. This classification problem can be implemented as logistic regression, in which case a function $c : \mathcal{X} \to [0, 1]$ is introduced and trained to minimise the objective

$$\begin{aligned} L(c) &= \mathbb{E}_{x \sim P}\left[-\log c(x)\right] + \mathbb{E}_{x \sim Q}\left[-\log(1 - c(x))\right] \\ &= \int -\log c(x) p(x) - \log(1 - c(x)) q(x) dx. \end{aligned}$$

For any particular x, the integrand is minimised by $c^*(x) = \frac{p(x)}{q(x)+p(x)}$, and so the optimal classifier assigns x to class 1 with the posterior probability that it was drawn from P (Proposition 1, Goodfellow et al., 2014). If c is parametrised as $\frac{1}{1+\exp(-r(x))}$ where $r : \mathcal{X} \to \mathbb{R}$, then the optimal c^* corresponds to $r^*(x) = \log\left(p(x)/q(x)\right)$.

This provides a useful tool for cases in which only samples from two distributions are available and estimation of their density ratio is desired, as is the case in Chapter 4. Moreover, solving this classification problem is closely related to divergence estimation, since for any choice of classifier c,

$$L(c) \geq \log 4 - 2 \cdot D_{\mathrm{JS}}(P, Q).$$

This follows straightforwardly from the definition of the Jensen-Shannon divergence and the fact that equality is attained by the optimal c^*, see Goodfellow et al., 2014 for details. Rearranging, it follows that the Jensen-Shannon divergence between two distributions can be estimated by maximising the lower bound

$$D_{\mathrm{JS}}(P, Q) \geq \log 2 - \frac{L(c)}{2}.$$

Recall that the Jensen-Shannon divergence is an f-divergence. It can be similarly shown that any f-divergence between two distributions can be estimated by maximisation of a lower bound corresponding to a classification problem, and that doing so results in a function of the density ratio being estimated. All f-divergences admit a variational form as a result of convex conjugacy (Nguyen et al., 2010). Any convex function $f(u)$ has a conjugate $f^*(t)$ defined as

$$f^*(t) = \sup_{u \in \mathrm{dom}\,(f)} \{ut - f(u)\}.$$

The resulting function f^* is itself convex, and provided that f is continuous, f and f^* are dual in the sense that $f^{**} = f$. This means that f can be written as

$$f(u) = \sup_{t \in \mathrm{dom}\,(f^*)} \{ut - f^*(t)\}.$$

Plugging this into the definition of f-divergences yields

$$\begin{aligned} D_f(P, Q) &= \int q(x) \sup_{t \in \mathrm{dom}\,(f^*)} \left\{ t \frac{p(x)}{q(x)} - f^*(t) \right\} dx \\ &\geq \sup_{T \in \mathcal{T}} \left\{ \int p(x) T(x) dx - \int q(x) f^*(T(x)) \right\} dx \\ &= \sup_{T \in \mathcal{T}} \mathbb{E}_{x \sim P} \left[T(x) \right] - \mathbb{E}_{x \sim Q} \left[f^*(T(x)) \right], \end{aligned}$$

where $\mathcal{T}$ is an arbitrary class of functions $\mathcal{X} \to \text{dom}\,(f^*) \subseteq \mathbb{R}$. It can be shown that the optimal T^* attaining the supremum satisfies

$$T^*(x) = f'\left(\frac{p(x)}{q(x)}\right),$$

subject to mild conditions on f (Lemma 1, Nguyen et al., 2010). In this sense, T^* estimates (a function of) the density ratio $p(x)/q(x)$.

2.6 Methods for fitting Latent Variable Models

Solving the generative modelling problem as posed in the introduction requires the minimisation of $D(P_X^\theta, Q_X)$ with respect to the parameters θ of the LVM in a computationally tractable way. In general it is infeasible to estimate $D(P_X^\theta, Q_X)$ directly, or even to compute gradients of it with respect to θ.

There are two main families of methods that introduce auxiliary functions as computational tricks to bound or estimate $D(P_X^\theta, Q_X)$ in a computationally tractable way. These are *Generative Adversarial Networks (GANs)*, which introduce a discriminator $d : \mathcal{X} \to [0, 1]$, and *autoencoders*, which introduce an encoder $e : \mathcal{X} \to \mathcal{Z}$.

The remainder of this section discusses GANs and two types of autoencoders, Variational Autoencoders (VAEs) and Wasserstein Autoencoders (WAEs), showing how specific choices of divergences can be approximated.

2.6.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a family of methods that approximately minimise any f-divergence and some choices of IPMs. In the case of f-divergences, the key ideas have been already introduced in Section 2.5.

Goodfellow et al., 2014 introduced the idea of training a discriminator $d^\phi : \mathcal{X} \to [0, 1]$ to classify between 'real' samples from Q_X and 'fake' samples from P_X^θ, which then provides a surrogate loss for a generator g^θ, trained simultaneously to maximise the loss of the discriminator. In Goodfellow et al., 2014, the loss is implemented as logistic regression:

$$L(\theta, \phi) = \mathbb{E}_{x \sim Q_X}\left[-\log d^\phi(x)\right] + \mathbb{E}_{x \sim P_Z}\left[-\log(1 - d^\phi(g^\theta(z))\right].$$

As previously discussed in Section 2.5, this is a lower bound on the Jensen-Shannon divergence $D_{\text{JS}}(P_X^\theta, Q_X)$, up to constants and scalar factors.

Stochastic gradients of this loss can be taken with respect to both ϕ and θ by using minibatch samples of data to approximate the outer expectations when d^ϕ and g^θ are both differentiably parametrised neural networks. Thus, although actually computing $D_{\mathrm{JS}}(P_X^\theta, Q_X)$ is intractable, it can be approximately minimised with respect to θ by maximising $L(\theta, \phi)$ with respect to ϕ and minimising this surrogate loss with respect to θ.

Nowozin et al., 2016, building on the work of Nguyen et al., 2010, generalised this to arbitrary f-divergences using the variational formulation of f-divergences to yield the surrogate loss

$$L_f(\phi, \theta) = \mathbb{E}_{x \sim Q_X}\left[T^\phi(x)\right] - \mathbb{E}_{z \sim P_Z}\left[f^*(T^\phi(g^\theta(x)))\right] \leq D_f(P_X^\theta, Q_X),$$

where the function T^ϕ is implemented as a neural network. As with the original GAN objective, stochastic gradients of this surrogate loss can be computed. The function T^ϕ plays the role of the discriminator introduced by Goodfellow et al., 2014. Nowozin et al., 2016 thus demonstrated how the GAN 'trick' can be applied to other f-divergences to approximately minimise $D_f(P_X^\theta, Q_X)$ with respect to the LVM parameters θ.

A similar idea can also be used to approximately minimise IPMs $D_{\mathcal{H}}(P_X^\theta, Q_X)$, provided that the function class $\mathcal{H}$ can be practically parameterised. If $\mathcal{H}$ is such that $h \in \mathcal{H}$ implies that $-h \in \mathcal{H}$, $D_{\mathcal{H}}$ can be written without the inner absolute function, leading to

$$\begin{aligned} D_{\mathcal{H}}(P, Q) &= \sup_{h \in \mathcal{H}} \int h(x) dP(x) - \int h(x) dQ(x) \\ &= \sup_{h \in \mathcal{H}} \left\{\mathbb{E}_{x \sim P}\left[h(x)\right] - \mathbb{E}_{x \sim Q}\left[h(x)\right]\right\} \\ &\geq \mathbb{E}_{x \sim P}\left[h(x)\right] - \mathbb{E}_{x \sim Q}\left[h(x)\right], \end{aligned}$$

where the inequality holds for any $h \in \mathcal{H}$, where h now plays the role of the discriminator. If the set $\mathcal{H}$ can be differentiably parametrised, stochastic gradients of the loss can be obtained with respect to its parameters. In contrast to the f-divergence case, the discriminator h must belong to a particular class of functions, which can complicate specifying the parametrisation. One example of an IPM-based GAN is the Wasserstein GAN of Arjovsky et al., 2017. Here, the 1-Wasserstein distance is used, corresponding to h having Lipschitz constant at most 1. For certain neural network architectures, including those composed of fully-connected and convolutional layers, this can be enforced by weight clipping.

It should be noted that while the ultimate goal is to minimise the divergence $D(P_X^\theta, Q_X)$ with respect to the parameters θ, the surrogate losses provided by the above techniques result in *lower bounds* to $D(P_X^\theta, Q_X)$. This leads to several potential issues. For example, it is difficult to make rigorous claims about the value of $D(P_X^\theta, Q_X)$ and whether it indeed becomes smaller throughout training. Furthermore, GANs are famously unstable to train, requiring a delicate balance between updates to the generator and discriminator. Minimisation of upper bounds

on the divergence $D(P_X^\theta, Q_X)$ are preferable, and one such method that does so in the case of the KL-divergence is discussed in the next section. Other works such as Zhang et al., 2019 investigate other upper bounds for more general choices of divergences to avoid the problems associated with GANs.

2.6.2 Variational Autoencoders

Variational Autoencdoers (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014) are a method to minimise the KL-divergence between model and data distributions, defined as

$$\begin{aligned} D_{\mathrm{KL}}(Q_X, P_X^\theta) &= \int q(x) \log\left(\frac{q(x)}{p^\theta(x)}\right) dx \\ &= \int q(x) \log q(x) - \int q(x) \log p^\theta(x) dx. \end{aligned}$$

The first term above, the negative of the *differential entropy* of Q_X, often written $H(Q_X)$, cannot be estimated without knowledge of the density $q(x)$. However, since it is constant as a function of θ, it can be ignored. The term $\log p^\theta(x)$ inside the second integral is known as the *log-likelihood* or *evidence*, and maximisation of this quantity is known as *maximum likelihood estimation*. Although the density $p^\theta(x)$ is intractable, the evidence can be tractably lower bounded leading to the so-called *evidence lower bound* (ELBO), which in turn leads to a tractable *upper* bound on $D_{\mathrm{KL}}(Q_X, P_X^\theta)$. This will be derived next.

First, observe that the log-likelihood can be written

$$\log p^\theta(x) = \log\left(\int p^\theta(x|z)p(z)dz\right).$$

Given any distribution $q^\phi(z|x)$ depending on parameter ϕ and the value of x, we can multiply and divide inside the integral, leaving its value unchanged. This leads to

$$\begin{aligned} \log p^\theta(x) &= \log\left(\int p^\theta(x|z)\frac{p(z)}{q^\phi(z|x)}q^\phi(z|x)dz\right) \\ &= \log\left(\mathbb{E}_{q^\phi(z|x)}\left[p^\theta(x|z)\frac{p(z)}{q^\phi(z|x)}\right]\right) \\ &\geq \mathbb{E}_{q^\phi(z|x)}\left[\log p^\theta(x|z) + \log p(z) - \log q^\phi(z|x)\right] \\ &= \mathbb{E}_{q^\phi(z|x)} \log p^\theta(x|z) - D_{\mathrm{KL}}\left(q^\phi(z|x), p(z)\right), \end{aligned}$$

where the inequality follows from Jensen's inequality due to the concavity of log. The distribution $q^\phi(z|x)$ is a variational approximation to the true posterior $p^\theta(z|x)$; it can be shown that the gap introduced by Jensen's inequality is equal to $D_{\mathrm{KL}}\left(q^\phi(z|x), p^\theta(z|x)\right)$. $q^\phi(z|x)$ is often referred to as an encoder as it maps elements of the data space $\mathcal{X}$ to

distributions over the latent space $\mathcal{Z}$. Putting things together yields

$$D_{\mathrm{KL}}(Q_X, P_X^\theta) \leq -H(Q_X) - \mathbb{E}_{q(x)}\mathbb{E}_{q^\phi(z|x)} \log p^\theta(x|z) + \mathbb{E}_{q(x)} D_{\mathrm{KL}}\left(q^\phi(z|x), p(z)\right),$$

where $H(Q_X)$ is the constant differential entropy, leading to the VAE loss

$$L_{\mathrm{VAE}}(\theta, \phi) = \underbrace{\mathbb{E}_{q(x)}\mathbb{E}_{q^\phi(z|x)}\left[-\log p^\theta(x|z)\right]}_{\text{(i)}} + \underbrace{\mathbb{E}_{q(x)} D_{\mathrm{KL}}\left(q^\phi(z|x), p(z)\right)}_{\text{(ii)}}. \tag{2.5}$$

Up to constants, this is an upper bound on $D_{\mathrm{KL}}(Q_X, P_X^\theta)$. It is common for the prior to be a standard Gaussian, the generator to output Gaussians with mean $\mu^\theta(z)$ and fixed isotropic covariance, and the encoder to map to Gaussians with mean $\mu^\phi(x)$ and diagonal covariance $\Sigma^\phi(x)$. In this case, term (i) above can be interpreted as an average reconstruction loss and (ii) as a regulariser, hence making the model a type of regularised autoencoder[1].

The encoder $Q^\phi_{Z|X}$ together with the data distribution Q_X induce the push-forward distribution Q^ϕ_Z known as the *aggregate posterior*. The term (ii) was shown by Hoffman and Johnson, 2016 to be equivalent to $D_{\mathrm{KL}}(Q^\phi_Z, P_Z) + I(X, Z)$ where $I(X, Z) = D_{\mathrm{KL}}(Q^\phi_{XZ}, Q_X Q^\phi_Z)$ is the mutual information of a data sample and its encoding. Chapter 3 concerns estimation of f-divergences (and hence the KL-divergence) between priors and aggregate posteriors, and so is directly relevant to mutual information estimation via this equivalence.

2.6.3 Wasserstein Autoencoders

Wasserstein Autoencoders (WAEs) (Tolstikhin et al., 2018) approximately minimise optimal transport distances between model and data distributions. These were discussed briefly in Section 2.4.1 as they can be expressed as IPMs, but here an alternative formulation of these distances will be used.

Let c be any non-negative function on $\mathcal{X} \times \mathcal{X}$ satisfying $c(x, x) = 0$. For intuition, $c(x, x')$ may be thought of as a function specifying the cost of transporting a point from x to x'. The optimal transport distance between two distributions P and Q over $\mathcal{X}$ is then the minimal cost incurred by transporting the probability mass of P to that of Q.

Formally, let Γ be the set of joint distributions over $\mathcal{X} \times \mathcal{X}$ with marginals P and Q. That is, any element $\gamma(x, x') \in \Gamma$ is a joint distribution satisfying $\gamma(x) = p(x)$ and $\gamma(x') = q(x')$.

[1] Note however that the interpretation of VAEs as a type of autoencoder breaks down when more powerful classes of generators are used, as discussed in http://paulrubenstein.co.uk/variational-autoencoders-are-not-autoencoders/.

Then, the optimal transport distance is defined as

$$OT_c(P, Q) = \min_{\gamma \in \Gamma} \mathbb{E}_{x,x' \sim \gamma} \left[c(x, x') \right].$$

This can equivalently be written as

$$OT_c(P, Q) = \min_{\gamma \in \Gamma} \mathbb{E}_{x \sim Q} \mathbb{E}_{x' \sim \gamma(x'|x)} \left[c(x, x') \right],$$

where γ can be thought of as a conditional distribution specifying how an element of probability mass at x should be moved and spread over the points x'. Specified this way, each element of Γ should satisfy $\int \gamma(x'|x) q(x) dx = p(x')$. In the generative modelling setting, we thus have

$$OT_c(P_X^\theta, Q_X) = \min_{\gamma \in \Gamma} \mathbb{E}_{x \sim Q_X} \mathbb{E}_{x' \sim \gamma(x'|x)} \left[c(x, x') \right].$$

In practice, this minimisation problem cannot be solved directly: given a candidate conditional distribution $\gamma(x'|x)$ it is not practically possible to verify whether or not $\int \gamma(x'|x) q_X(x) dx = p_X^\theta(x')$ since only samples from Q_X are available and the density $p_X^\theta(x)$ is intractable.

Tolstikhin et al., 2018 proved the following result, giving a handle on this problem: If the generator $p^\theta(x|z)$ is deterministic, any valid γ can be written as a composition $\gamma(x'|x) = \int p^\theta(x'|z) q^\phi(z|x) dz$ where $q^\phi(z|x)$ is a conditional distribution satisfying $\int q^\phi(z|x) q(x) dx = p(z)$. That is, any γ can be 'factored through' the latent space by introducing an encoder $q^\phi(z|x)$, replacing the constraint of distribution matching in the data-space ($\int \gamma(x'|x) q_X(x) dx = p_X^\theta(x')$) with distribution matching in the latent space ($\int q^\phi(z|x) q(x) dx = p(z)$).

Intuitively, P_Z is a different parametrisation of P_X^θ and so if Q_X pushed through the encoder results in P_Z, pushing Q_X through the composition of the encoder and generator will result in P_X^θ. While it is clear that the composition of any such encoder and generator induces a valid γ, it is non-trivial that any γ can be decomposed as such. This is proved in Theorem 1 of Tolstikhin et al., 2018.

Writing $Q_Z = \int Q_{Z|X=x} q(x) dx$ to be the latent space distribution obtained by pushing the data through the encoder and g for the deterministic generator yields the following alternative statement of the optimal transport distance in the LVM setting:

$$OT_c(P_X^\theta, Q_X) = \min_{Q_{Z|X} : Q_Z = P_Z} \mathbb{E}_{x \sim Q_X} \mathbb{E}_{z \sim Q_{Z|X=x}} \left[c(x, g(z)) \right]. \tag{2.6}$$

Again this optimisation problem is not feasible to solve due to the hard constraint, but it can be made computationally feasible by relaxing the constraint to obtain the WAE objective

$$L_{\mathrm{WAE}}^{\lambda,D}(\theta,\phi) = \underbrace{\mathbb{E}_{x\sim Q_X}\mathbb{E}_{z\sim Q_{Z|X=x}^{\phi}}\left[c(x,g^{\theta}(z))\right]}_{\text{(i)}} + \underbrace{\lambda D\left(Q_Z^{\phi}, P_Z\right)}_{\text{(ii)}}, \tag{2.7}$$

where D is some divergence, λ is a positive scalar and the encoder is given parameter ϕ. While the generator is required to be deterministic, the encoder may be deterministic or stochastic (Rubenstein et al., 2018c). For general choices of λ and D, $\inf_{\phi} L_{\mathrm{WAE}}^{\lambda,D}(\theta,\phi)$ is neither an upper nor lower bound on the original objective $OT_c(P_X^{\theta}, Q_X)$, but a heuristic approximation due to the relaxation of the constraint.

Term (i) of the WAE loss is simply a reconstruction loss, corresponding to the average distance between data sampled from Q_X and the reconstruction obtained by pushing through the encoder and generator. This is therefore simple to estimate and optimise with respect to the parameters ϕ and θ. Term (ii) is potentially challenging to estimate depending on the choice of D. Tolstikhin et al., 2018 propose two choices for D which can be estimated based on samples from P_Z and Q_Z^{ϕ}: the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) which can be estimated directly, leading to WAE-MMD, and a GAN style estimation of the Jensen-Shannon divergence by introducing an additional discriminator to obtain a lower bound on the divergence, leading to WAE-GAN.

Estimating a divergence between P_Z and Q_Z^{ϕ} using sample-based methods does not make use of a significant degree of structure that is present in the problem. P_Z is typically chosen to be a simple distribution with known density. While Q_Z^{ϕ} may be complex, its density can be decomposed as $q^{\phi}(z) = \mathbb{E}_{x\sim Q_X} q^{\phi}(z|x)$ where $q^{\phi}(z|x)$ is also typically chosen to be simple (e.g. Gaussian) and Q_X can be sampled. The main contribution of Chapter 3 is to propose and analyse an estimator making use of this structure for the case that D in (ii) is chosen to be an f-divergence.

2.7 Conclusion

This chapter introduced key ideas in the literature of generative modelling with LVMs that are relevant to this thesis, in particular the following two chapters. Next, Chapter 3 presents learning theoretic results for divergence estimation in the latent spaces of autoencoders. Following that, Chapter 4 presents identifiability results for ICA models, a type of LVM. Chapter 5, the third and final main chapter of research content, concerns causality and is less relevant to the ideas presented here, though there are strong connections between more general probabilistic modelling and causality.

Chapter 3

Latent Space Learning Theory

This chapter presents and analyses RAM-MC, an f-divergence estimator that is applicable to estimating divergences between particular distributions in the latent spaces of autoencoder models such as Variational Autoencoders and Wasserstein Autoencoders. Learning theoretic analyses of sampled-based f-divergence estimators usually yield poor rates of convergence unless strong assumptions are made. By exploiting the natural structure present in the autoencoder setting, RAM-MC exhibits fast rates under mild assumptions.

This work is an important contribution to the literature for two main reasons. First, it demonstrates that the f-divergences considered can be estimated with a practical number of samples. Second, it provides a rigorous foundation to heuristically proposed methods for estimating and minimising Total Correlation and Mutual Information in the context of Variational Autoencoders.

The main technical content of this chapter has been published in the paper:

> Paul K Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme and Ilya Tolstikhin. "Practical and Consistent Estimation of f-Divergences". *Advances in Neural Information Processing Systems (NeurIPS)*. 2019

3.1 Introduction

The estimation and minimisation of divergences between probability distributions based on samples are fundamental problems of machine learning. The previous chapter discussed generative modelling, but there are numerous other applications across the literature. For example, in variational inference, an intractable posterior $p(z|x)$ is approximated with a tractable distribution $q(z)$ chosen to minimise the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q(z), p(z|x))$. The mutual information between two variables $I(X, Y)$, core to information theory and

Bayesian machine learning, is equivalent to $D_{\mathrm{KL}}(P_{X,Y}, P_X P_Y)$. 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 minimised, where $\phi_{\#}$ represents the push-forward operation (Ben-David et al., 2007; Ganin et al., 2016).

This chapter concerns the estimation of f-divergences, introduced in Section 2.4.2. A significant body of work exists studying estimation of $D_f(P, Q)$ for general probability distributions P and Q. While the majority of this focuses on α-divergences and closely related Rényi-α divergences (Poczos and Schneider, 2011; Singh and Poczos, 2014; Krishnamurthy et al., 2014), many works address specifically the KL-divergence (Perez-Cruz, 2008; Wang et al., 2009) with fewer considering f-divergences in full generality (Nguyen et al., 2010; Kanamori et al., 2012; Moon and Hero, 2014a; Moon and Hero, 2014b). Although the KL-divergence is the most frequently encountered f-divergence in the machine learning literature, in recent years there has been growing interest in other f-divergences (Nowozin et al., 2016; Zhang et al., 2019), in particular in the variational inference community where they have been employed to derive alternative evidence lower bounds (Li and Turner, 2016; Dieng et al., 2017; Chen et al., 2018a).

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

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 $p(x)$ and $q(x)$ (Singh and Poczos, 2014; Krishnamurthy et al., 2014) or the density ratio $p(x)/q(x)$ (Nguyen et al., 2010; Kanamori et al., 2012) from samples. Both of these estimation problems are provably hard (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 $N^{-1/d}$ is

known 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 (Nguyen et al., 2010; Krishnamurthy et al., 2014) 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 (Wang et al., 2009; Poczos and Schneider, 2011; Kanamori et al., 2012).

Motivated by recent advances in machine learning, this chapter considers a setting in which structural assumptions are made about the distributions. Although the assumptions are strong, they are naturally satisfied in the setting of autoencoders with probabilistic encoders, including Wasserstein Autoencoders and variants of Variational Autoencoders.

3.1.1 Summary of setting and results

Let $\mathcal{X}$ and $\mathcal{Z}$ be two finite dimensional Euclidean spaces. This chapter studies estimation of the divergence $D_f(Q_Z, P_Z)$ between two probability distributions P_Z and Q_Z, both defined over $\mathcal{Z}$. It is assumed that P_Z has known density $p(z)$, while Q_Z with density $q(z)$ admits the factorization $q(z) = \int q(z|x)q(x)dx$. 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 the language of autoencoders, $\mathcal{X}$ 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*, though the theory presented in this work does not apply exclusively to this setting.

Given independent observations $X_1, \dots, X_N$ from Q_X, the goal is to estimate $D_f(Q_Z, P_Z)$. The main contribution of this chapter is to use the finite mixture $\hat{Q}_Z^N := \frac{1}{N}\sum_{i=1}^N Q_{Z|X_i}$ as a surrogate for the continuous mixture Q_Z, to use this to approximate $D_f(Q_Z, P_Z)$ with $D_f(\hat{Q}_Z^N, P_Z)$, and to theoretically study conditions under which this approximation is reasonable.

$D_f(\hat{Q}_Z^N, P_Z)$ is denoted the *Random Mixture (RAM)* estimator and rates at which it converges to $D_f(Q_Z, P_Z)$ as N grows are derived. Similar guarantees are also provided for *RAM-MC*, a practical Monte-Carlo based version of RAM. By side-stepping the need to perform density estimation, one obtains *parametric* rates of order $N^{-\gamma}$, where γ is independent of the dimension (see Tables 3.1 and 3.2), 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.

These results have immediate implications to existing literature. For the particular case of the KL divergence, a similar approach has been heuristically applied for estimation of mutual information (Poole et al., 2018) and total correlation (Chen et al., 2018b) in the

context of Variational Autoencoders. Both works have application to representation learning, the latter in particular to disentangled representation learning. The results presented in this chapter thus provide strong theoretical grounding for these existing methods through rigorous analysis lacking in the original proposals. In these works, the quantities being estimated and minimised are $D_{\mathrm{KL}}(Q_{Z,X}, Q_Z Q_X)$ and $D_{\mathrm{KL}}(Q_Z, \prod_i Q_{Z_i})$ respectively. Each of these quantities can be rewritten in terms of $D_{\mathrm{KL}}(Q_Z, P_Z)$ and so RAM-MC can be used to estimate them. In doing so, one recovers the estimators proposed by the authors, thus results concerning the convergence properties of RAM-MC transfer to results about their estimators. See Section 3.5 for details.

Moreover, while this work considers estimation of $D_f(Q_Z, P_Z)$, minimisation of this quantity is the key challenge in the training of Wasserstein Autoencoders (see Section 2.6.3). Preliminary experiments not included in this thesis did not find the proposed RAM-MC estimator to lead to improved training over existing methods, but future research could investigate this more thoroughly.

Section 3.2 presents known results from the literature on f-divergences and basic results in learning theory that are used in the proofs of novel results presented in this chapter. Following this, Section 3.3 introduces the RAM and RAM-MC estimators and presents the main theoretical results, including rates of convergence for the bias (Theorems 3.11 and 3.12) and tail bounds (Theorems 3.13 and 3.14). Section 3.4 validates the results in both synthetic and real-data experiments. Section 3.5 discusses applications of these results to the literature, and Section 3.6 concludes. The results presented in Section 3.3 have long proofs; short sketches are presented in the main text, while the full proofs can be found in Appendix A.

3.2 Background results

This section presents basic results from the literature that are used in the proofs of novel results in this chapter. These include bounds relating different f-divergences, closed-form expressions for some f-divergences in the case of Gaussian probability distributions, and basic concentration inequalities. Proofs are omitted for referenced results.

3.2.1 f-divergence bounds

The results presented here are inequalities relating the values taken by f-divergences. These are mostly used in the proof of Theorem 3.11. A comprehensive treatment of the relationships between different f-divergences can be found in Tsybakov, 2009, from which many of the results below are taken.

Lemma 3.1 (Lemma 2.4, Tsybakov, 2009)**.** *Let A and B be probability distributions. Then,*

$$\mathrm{H}^2(A, B) \leq \mathrm{KL}(A, B).$$

Lemma 3.2 (Pinsker's inequality, Lemma 2.5, Tsybakov, 2009)**.** *Let A and B be probability distributions. Then,*

$$\mathrm{TV}(A, B) \leq \sqrt{\frac{1}{2}\mathrm{KL}(A, B)}.$$

Lemma 3.3 (Lemma 2.7, Tsybakov, 2009)**.** *Let A and B be probability distributions. Then,*

$$\mathrm{KL}(A, B) \leq e^{\mathrm{KL}(A,B)} - 1 \leq \chi^2(A, B).$$

Lemma 3.4 (Theorem 2, Osterreicher and Vajda, 2003)**.** *Let A and B be probability distributions. For any value $\beta \geq 0$, there exists a scalar $\psi(\beta)$ such that*

$$D_{f_\beta}(A, B) \leq \psi(\beta)\mathrm{TV}(A, B).$$

Lemma 3.5. *Suppose that $D_f^{\frac{1}{2}}$ satisfies the triangle inequality. Let A_N be a sequence of probability distributions, and let B and C be fixed probability distributions. Then for any $\lambda > 0$,*

$$D_f\left(A_N, C\right) - D_f\left(B, C\right) \leq (1 + \lambda) D_f\left(A_N, B\right) + \frac{1}{\lambda} D_f\left(B, C\right).$$

If, furthermore, $D_f\left(A_N, B\right) = O\left(\frac{1}{N^k}\right)$ for some $k > 0$, then

$$D_f\left(A_N, C\right) - D_f\left(B, C\right) = O\left(\frac{1}{N^{k/2}}\right).$$

Proof. The first inequality follows from the triangle inequality applied to $D_f^{\frac{1}{2}}(A_N, C)$, 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 $\lambda = N^{-\frac{k}{2}}$. □

3.2.2 Closed-form expressions for f-divergences between Gaussians

The closed-form expressions for f-divergences presented here are used for the experiments in Section 3.4, as well as to understand cases in which the assumptions of all results hold in practical scenarios, discussed at the end of Section 3.3.

Lemma 3.6 (Exercise 1.6.11, Pardo, 2005)**.** *The KL-divergence between two d-variate Gaussians is*

$$\mathrm{KL}(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{2}\left(tr\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^\intercal\Sigma_2^{-1}(\mu_2-\mu_1) - d + \log\frac{|\Sigma_2|}{|\Sigma_1|}\right).$$

Lemma 3.7 (Exercise 1.6.14, Pardo, 2005)**.** *The squared Hellinger (H^2) divergence between two multivariate Gaussians is*

$$\begin{aligned}&\mathrm{H}^2(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2))\\ &\quad= 1 - \frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{1/2}}\exp\left\{-\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)\right\}.\end{aligned}$$

Lemma 3.8 (Lemma 1, Nielsen and Nock, 2014)**.** *Suppose that P_1 and P_2 are members of the same exponential family of distributions with log-partition function F and natural parameters θ_1 and θ_2 respectively. Then*

$$\chi^2(P_1,P_2) = e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))} - 1,$$

and is finite provided that $2\theta_2-\theta_1$ belongs to the natural parameter space. In the particular case of Gaussians,

$$\begin{aligned}\chi^2(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)) = {}& \frac{\det(\Sigma_2^{-1})}{\sqrt{\det(2\Sigma_2^{-1}-\Sigma_1^{-1})\det(\Sigma_1^{-1})}}\exp\left(\frac{1}{2}\mu_2'\Sigma_1^{-1}\mu_2 - \mu_1'\Sigma_2^{-1}\mu_1\right)\\ &\times\exp\left(-\frac{1}{4}(2\mu_1'\Sigma_2^{-1}-\mu_2'\Sigma_1^{-1})(\frac{1}{2}\Sigma_1^{-1}-\Sigma_2^{-1})^{-1}(2\Sigma_2^{-1}\mu_1-\Sigma_1^{-1}\mu_2)\right) - 1.\end{aligned}$$

3.2.3 Concentration inequalities

Concentration inequalities provide bounds on the probability with which a random variable deviates from its expectation. Here two such results are outlined; a comprehensive study can be found in Boucheron et al., 2013. One basic result is *Chebyshev's inequality.*

Lemma 3.9 (Chebyshev's inequality, Section 2.1, Boucheron et al., 2013)**.** *Let X be a random variable with finite expectation and variance. Then, for any $t>0$,*

$$\mathbb{P}\left(|X-\mathbb{E}X|\geq t\right) \leq \frac{\mathrm{Var}(X)}{t}.$$

A much stronger concentration result, *McDiarmid's inequality* (sometimes called the *bounded difference inequality*), provides an exponential bound if the *bounded difference property* is satisfied. This result forms the basis of the proof of Theorem 3.13.

Theorem 3.10 (McDiarmid's inequality, Theorem 6.2, Boucheron et al., 2013)**.** *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

$$\mathbb{P}\left(|\phi(X_1, \ldots, X_N) - \mathbb{E}\phi| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^N c_i^2}\right).$$

3.3 Random mixture estimator and convergence results

This section introduces the proposed f-divergence estimator and presents theoretical guarantees for it. The existence is assumed 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, but a set of i.i.d. samples $\mathbf{X}^N = \{X_1, \ldots, X_N\}$ from Q_X are given. The ultimate goal is to estimate the f-divergence

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

This cannot be directly computed since Q_Z is unknown. 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 the proposed *Random Mixture estimator (RAM)*:

$$D_f(\hat{Q}_Z^N, P_Z) := D_f\Big(\tfrac{1}{N}\textstyle\sum_{i=1}^N Q_{Z|X_i}, P_Z\Big). \tag{3.1}$$

Although $\hat{Q}_Z^N = \hat{Q}_Z^N(\mathbf{X}^N)$ is a function of the i.i.d. samples $\mathbf{X}^N$, this explicit dependence is omitted for notational brevity. The true f-divergence $D_f(Q_Z, P_Z)$ is a real-valued scalar, but $D_f(\hat{Q}_Z^N, P_Z)$ is a random variable whose randomness is inherited from the i.i.d. samples.

The rest of this section is devoted to the exploration of conditions under which $D_f(\hat{Q}_Z^N, P_Z)$ is a 'good' estimator of $D_f(Q_Z, P_Z)$. More formally, conditions are established under which the estimator is asymptotically unbiased, concentrates to its expected value and can be practically estimated using Monte-Carlo sampling.

3.3.1 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. It follows that

the *bias* of RAM, the difference between its expectation and the true value of the quantity being estimated, is positive and decreasing in N.

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

$$D_f(Q_Z, P_Z) \;\leq \mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N, P_Z)] \;\leq\; \mathbb{E}_{\mathbf{X}^M}[D_f(\hat{Q}_Z^M, P_Z)]. \tag{3.2}$$

Proof sketch (full proof in Appendix A.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 $\mathbf{X}^M$ can be drawn by sub-sampling (without replacement) M entries of $\mathbf{X}^N$, and by applying Jensen's inequality again. □

As a function of N, the expectation of RAM is a decreasing sequence that is bounded below. By the monotone convergence theorem, the sequence converges. Theorems 3.11 and 3.12 below 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 fs and provide rates at which this happens, summarised in Table 3.1. The two theorems are proved using different techniques and assumptions. These assumptions, along with those of other methods (see Table 3.3) are discussed in Section 3.3.5.

Theorem 3.11 (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 3.1.*

Proof sketch (full proof in Appendix A.2). The proofs vary slightly for each choice of f but there are two key steps. The first is to bound the bias in terms of $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N, Q_Z)]$. The second step is to bound $\mathbb{E}_{\mathbf{X}^N}[D_f(\hat{Q}_Z^N, Q_Z)]$ in terms of $\mathbb{E}_{\mathbf{X}^N}[\chi^2(\hat{Q}_Z^N, Q_Z)]$. From the definition of the χ^2 divergence, the latter quantity is the variance of the average of N i.i.d. random variables and therefore decomposes as $\mathbb{E}_{X\sim Q_X}[\chi^2(Q_{Z|X}, Q_Z)]/N = O(N^{-1})$.

For KL, the first bound is an equality provided that $\mathrm{KL}(Q_Z, P_Z)$ is finite, and the second follows from the fact that $\mathrm{KL} \leq \chi^2$ (Lemma 3.3). For TV, the first bound holds because it is a metric, after which Pinsker's inequality (Lemma 3.2) can be used to upper bound in terms of the rate for the KL.

For D_{f_β}, which includes H^2 and JS as special cases, the first bound is derived by applying Lemma 3.5, using the property that $D_{f_\beta}^{1/2}$ satisfies the triangle inequality for $\beta \geq \frac{1}{2}$ (Hein and Bousquet, 2005). The rate for H^2 can then be related to that of the KL by the relation $\mathrm{H}^2 \leq \mathrm{KL}$ (Lemma 3.1), while for the other D_{f_β} (including JS), it can be related to that of TV via the relation $D_{f_\beta} \leq \psi(\beta)\mathrm{TV}$ for some scalar $\psi(\beta)$ (Lemma 3.4). □

Theorem 3.12 (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 3.1.*

Table 3.1 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 3.11	N^{-1}	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{4}}$	$N^{-\frac{1}{4}}$	$N^{-\frac{1}{4}}$	-
Theorem 3.12	$N^{-\frac{1}{3}} \log N$	$N^{-\frac{1}{2}}$	N^{-1}	$N^{-\frac{1}{5}}$	$N^{-\frac{1}{3}} \log N$	$N^{-\frac{1}{3}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{\alpha+1}{\alpha+5}}$

Proof sketch (full proof in Appendix A.4). The exact proof is different for each choice of f. For the χ^2-divergence, this bias can be bounded directly in terms of the assumed finite expectation $\mathbb{E}_{X\sim Q_X, Z\sim P_Z}\left[q^4(Z|X)/p^4(Z)\right]$.

For all other divergences, the proofs have the following general outline. A convex function lies above any supporting hyperplane, and so $f(u+t) \geq f(u) + f'(u)t$ for any u, t in the scalar case. Taking $a = u$, $b = u+t$ and rearranging yields the inequality $f(a) - f(b) \leq (a-b)f'(a)$. Denoting by $\hat{q}_N(z)$ the density of $\hat{Q}_Z^N$,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f(\hat{Q}_Z^N, P_Z)\right] - D_f(Q_Z, P_Z) \\
&\quad = \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[f\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right] - \mathbb{E}_{P_Z}\left[f\left(\frac{q(z)}{p(z)}\right)\right] \\
&\quad = \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[f\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f\left(\frac{q(z)}{p(z)}\right)\right] \\
&\quad \leq \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\frac{\hat{q}_N(z) - q(z)}{p(z)} f'\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right] \\
&\quad \leq \sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z) - q(z)}{p(z)}\right)^2\right]} \sqrt{\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[f'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right]},
\end{aligned}
$$

where the second upper bound follows by Cauchy-Schwartz. The left term can be bounded in terms of $\mathbb{E}_{X\sim Q_X, Z\sim P_Z}\left[q^4(Z|X)/p^4(Z)\right]$, while the right hand term is bounded by controlling f'.

Subtle treatment is required for the case that f' diverges as the density ratio $\hat{q}_N(z)/p(z)$ approaches zero. In this case, the bias is written as a sum of integrals over separate ranges of values for $\hat{q}_N(z)/p(z)$. For small values of $\hat{q}_N(z)/p(z)$, the integral is controlled directly, while for sufficiently large values the above inequalities are used. □

3.3.2 Tail bounds for RAM

Theorems 3.11 and 3.12 describe the convergence of the expectation of the random variable $D_f(\hat{Q}_Z^N, P_Z)$. In practical scenarios, the spread of its distribution will also be of interest, because evaluation based on a single set of i.i.d. samples $\mathbf{X}^N$ corresponds to a single observation

Table 3.2 Rate $\psi(N)$ of high probability bounds for $D_f(\hat{Q}_Z^N, P_Z)$ (Theorem 3.13).

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} $\frac{1}{3}<\alpha<1$
$\psi(N)$	$N^{-\frac{1}{6}}\log N$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{6}}\log N$	$N^{-\frac{1}{6}}$	$N^{-\frac{1}{2}}$	$N^{\frac{1-3\alpha}{\alpha+5}}$

of the random variable $D_f(\hat{Q}_Z^N, P_Z)$. If the spread were large, then even if the bias were small, one could not confidently conclude that a single draw of $D_f(\hat{Q}_Z^N, P_Z)$ would be close to the true divergence $D_f(Q_Z, P_Z)$. Fortunately, the following result shows that RAM rapidly concentrates to its expectation.

Theorem 3.13 (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 3.2.*

Proof sketch (full proof in Appendix A.5). These results follow by McDiarmid's inequality applied to $D_f(\hat{Q}_Z^N, P_Z)$ (Theorem 3.10). To apply it, it needs to be shown that RAM viewed as a function of $\mathbf{X}^N$ exhibits the bounded differences property. That is, when changing a single coordinate X_i of $\mathbf{X}^N = (X_1, X_2, \ldots, X_N)$, the value of $D_f(\hat{Q}_Z^N(\mathbf{X}^N), P_Z)$ changes by at most a constant $c_{i,N}$ that may depend on N. This constant is shown to be $O(N^{-1/2}\psi(N))$ for all values of i, from which the result follows directly from McDiarmid.

Proof of the bounded difference property proceeds similarly to the proof of Theorem 3.12. Let $\mathbf{X}^N$ and $\mathbf{X}^{N\prime}$ be two vectors that differ only in their first coordinate, so that $X_1 \neq X_1'$, but $X_i = X_i'$ for all $j > 1$. Denote by $\hat{q}_N$ and $\hat{q}_N'$ the densities of $\hat{Q}_Z^N(\mathbf{X}^N)$ and $\hat{Q}_Z^N(\mathbf{X}^{N\prime})$

respectively. Then,

$$
\begin{aligned}
&D_f(\hat{Q}_Z^N(\mathbf{X}^N), P_Z) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime}), P_Z\right) \\
&= \mathbb{E}_{P_Z}\left[f\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right] - \mathbb{E}_{P_Z}\left[f\left(\frac{\hat{q}'_N(z)}{p(z)}\right)\right] \\
&= \mathbb{E}_{P_Z}\left[f\left(\frac{\hat{q}_N(z)}{p(z)}\right) - f\left(\frac{\hat{q}'_N(z)}{p(z)}\right)\right] \\
&\leq \mathbb{E}_{P_Z}\left[\frac{\hat{q}_N(z) - \hat{q}'_N(z)}{p(z)} f'\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right] \\
&\leq \sqrt{\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z) - \hat{q}'_N(z)}{p(z)}\right)^2\right]} \sqrt{\mathbb{E}_{P_Z}\left[f'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right]} \\
&= \sqrt{\mathbb{E}_{P_Z}\left[\left(\frac{1}{N}\frac{q(z|X_1) - q(z|X'_1)}{p(z)}\right)^2\right]} \sqrt{\mathbb{E}_{P_Z}\left[f'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right]} \\
&= \frac{1}{N}\sqrt{\mathbb{E}_{P_Z}\left[\left(\frac{q(z|X_1) - q(z|X'_1)}{p(z)}\right)^2\right]} \sqrt{\mathbb{E}_{P_Z}\left[f'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\right]}.
\end{aligned}
$$

A similar bound can be derived by swapping the role of $\mathbf{X}^N$ and $\mathbf{X}^{N\prime}$, thus the absolute value of the difference is upper bounded by the maximum of these bounds. By symmetry, it suffices to control one of them. The left hand term is controlled by the assumption that $\chi^2\left(Q_{Z|x}, P_Z\right) \leq C < \infty$. The right hand term requires separate treatment for each choice of f. Similar to the proof of Theorem 3.12, special care is required for the case that f diverges as the density ratio $\hat{q}_N(z)/p(z)$ goes to zero. □

3.3.3 Practical estimation with RAM-MC

In practice it may not be possible to evaluate $D_f(\hat{Q}_Z^N, P_Z)$ analytically as this would require solving a potentially complicated integral. However, since both densities $\hat{q}_N(z)$ and $p(z)$ are known, Monte-Carlo (MC) sampling can be used to estimate the integral. In particular, consider importance sampling with proposal distribution $\pi(z|\mathbf{X}^N)$, where 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 $\pi(z|\mathbf{X}^N)$:

$$
\hat{D}_f^M(\hat{Q}_Z^N, P_Z) := \frac{1}{M}\sum_{m=1}^{M} f\left(\frac{\hat{q}_N(Z_m)}{p(Z_m)}\right)\frac{p(Z_m)}{\pi\left(Z_m|\mathbf{X}^N\right)}. \tag{3.3}
$$

Theorem 3.14 (RAM-MC is unbiased and consistent)**.** *For any proposal distribution π, RAM-MC is unbiased:*

$$\mathbb{E}[\hat{D}_f^M(\hat{Q}_Z^N, P_Z)] = \mathbb{E}[D_f(\hat{Q}_Z^N, P_Z)].$$

If the hypothesis of Theorem 3.13 holds and moreover either of the following conditions are satisfied:

$$(i)\begin{cases}\pi(z|\mathbf{X}^N) = p(z),\\ \mathbb{E}_{Q_X}\int f\left(\dfrac{q(z|X)}{p(z)}\right)^2 p(z)dz < \infty,\\ \mathbb{E}_{Q_X}\int \left(\dfrac{q(z|X)}{p(z)}\right)^2 p(z)dz < \infty,\end{cases}$$

$$(ii)\begin{cases}\pi(z|\mathbf{X}^N) = \hat{q}_N(z),\\ \mathbb{E}_{Q_X}\int f\left(\dfrac{q(z|X)}{p(z)}\right)^2 \left(\dfrac{p(z)}{q(z|X)}\right)^2 q(z|X)dz < \infty,\\ \mathbb{E}_{Q_X}\int \left(\dfrac{p(z)}{q(z|X)}\right)^2 q(z|X)dz < \infty,\end{cases}$$

then denoting by $\psi(N)$ the rate given in Table 3.2, the variance of RAM-MC decays as

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

Proof sketch (proof in Appendix A.6). For unbiasedness, observe that

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}^M,\mathbf{X}^N}\hat{D}_f^M(\hat{Q}_Z^N, P_Z) &= \mathbb{E}_{\mathbf{X}^N}\left[\mathbb{E}_{\mathbf{Z}^M \overset{i.i.d.}{\sim} \pi(z|\mathbf{X}^N)}\hat{D}_f^M(\hat{Q}_Z^N, P_Z)\right]\\ &= \mathbb{E}_{\mathbf{X}^N}\left[\mathbb{E}_{z\sim\pi(z|\mathbf{X}^N)} f\left(\frac{\hat{q}_N(z)}{p(z)}\right)\frac{p(z)}{\pi(z|\mathbf{X}^N)}\right]\\ &= \mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N, P_Z\right)\right].\end{aligned}$$

By the law of total variance, the variance can be decomposed as

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

The first of these terms is $O(M^{-1})$ by standard results on MC integration, subject to the finiteness assumptions. The concentration results of Theorem 3.13 imply bounds on the

Table 3.3 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., 2014	-	-	-	-	-	-	-	$N^{-\frac{1}{2}}+N^{\frac{-3s}{2s+d}}$
Nguyen et al., 2010	$N^{-\frac{1}{2}}$	-	-	-	-	-	-	-
Moon and Hero, 2014a	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$

second term, since for a random variable X,

$$\begin{aligned}\mathrm{Var}X &= \mathbb{E}(X-EX)^2 \\ &= \int_0^\infty \mathbb{P}\left((X-\mathbb{E}X)^2 > t\right) dt \\ &= \int_0^\infty \mathbb{P}\left(|X-\mathbb{E}X| > \sqrt{t}\right) dt.\end{aligned}$$

For the second term, observe first that the tail bound of Theorem 3.13 can be rewritten as

$$\mathbb{P}\left(\left|D_f\left(\hat{Q}_Z^N, P_Z\right) - \mathbb{E}D_f\left(\hat{Q}_Z^N, P_Z\right)\right| > K\psi(N)\sqrt{\log\frac{2}{\delta}}\right) \leq \delta.$$

Taking $\sqrt{t} = K\psi(N)\sqrt{\log\frac{2}{\delta}}$ implies $\delta = 2e^{\frac{-t}{K^2\psi(N)^2}}$ and so plugging into the above formula for the variance yields

$$\begin{aligned}\mathrm{Var}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N, P_Z\right)\right] &\leq \int_0^\infty 2\exp\left(-\frac{1}{K^2\psi(N)^2}t\right) dt \\ &= 2K^2\psi(N)^2 \\ &= O\left(\psi(N)^2\right).\end{aligned}$$

□

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. Through Chebyshev's inequality (Lemma 3.9) it is also possible to derive confidence bounds for RAM-MC of the form similar to Theorem 3.13, but with an additional dependence on M and the worse dependence on δ of $1/\delta$ instead of $\sqrt{\log(2/\delta)}$.

3.3.4 Discussion about assumptions

Although the data distribution Q_X will generally be unknown, in some practical scenarios such as 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 3.12, 3.13 and 3.14 can be satisfied by making suitable restrictions (we conjecture also for Theorem 3.11).

For example, a common architectural choice would be to take $P_Z = \mathcal{N}(0, I_d)$ and $Q_{Z|X} = \mathcal{N}(\mu(X), \Sigma(X))$ with Σ diagonal. If furthermore there exist constants $K, \epsilon > 0$ such that $\|\mu(X)\| \leq K$ and $\Sigma_{ii}(X) \in [\epsilon, 1]$ for all i, then the assumptions of Theorems 3.12, 3.13 and 3.14 hold.

Indeed, $\chi^2(Q_{Z|x}, P_Z)$ can be written in terms of $\mu(X)$ and $\Sigma(X)$ and is finite for all $x \in \mathcal{X}$ by Lemma 3.8. Since both $\mu(X)$ and $\Sigma(X)$ take value in compact sets, it follows that there exists $C < \infty$ such that $\chi^2(Q_{Z|x}, P_Z) \leq C$ and thus the setting of Theorem 3.13 holds.

A similar argument based on compactness shows that the density ratio is uniformly bounded in z and x, so that $q(z|x)/p(z) \leq C'$ for some $C' < \infty$ for all values of z and x. It follows that the conditions of Theorems 3.12 and 3.14 hold. For the former this is because $\int q^4(z|x)/p^4(z) dP(z) < C'^4 < \infty$, and for the latter this is because the terms inside the integrals of condition (i) are bounded and thus the norms and expectations are finite.

The existence of such an ϵ and K are not particularly strong assumptions in practice, since numerical stability often requires the diagonal entries of Σ to be lower bounded by a small number (e.g. 10^{-6}), and if $\mathcal{X}$ is compact (as is the case for images) then such a K is guaranteed to exist; if not, choosing K very large yields an insignificant constraint.

We conjecture that the strong boundedness assumptions on $\mu(X)$ and $\Sigma(X)$ also imply the setting of Theorem 3.11 for which it is required that $\mathbb{E}_X[\chi^2(Q_{Z|X}, Q_Z)] < \infty$. Since the divergence Q_Z explicitly depends on the data distribution, this is more difficult to verify than the conditions of Theorems 3.12 and 3.13. The crude upper bound provided by convexity

$$\mathbb{E}_X[\chi^2(Q_{Z|X}, Q_Z)] \leq \mathbb{E}_X \mathbb{E}_{X'}[\chi^2(Q_{Z|X}, Q_{Z|X'})]$$

means that finiteness of the right hand side would imply that the assumptions of Theorem 3.11 hold. This would be the case, for instance, if $\|\mu(X)\| \leq K$ and $\Sigma_{ii}(X) \in [\frac{1}{2} + \epsilon, 1]$ for all i, however this is a rather strong and unrealistic assumption on $\Sigma(X)$.

3.3.5 Summary

All of the rates presented for RAM and RAM-MC 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.

Table 3.3 summarises the rates of bias for some existing methods. In contrast to the proposals of this work, 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 *periodic Hölder class of smoothness s* (see Definition 1 of Krishnamurthy et al., 2014), 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, see Section III.A of Nguyen et al., 2010) is 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 the assumptions of Krishnamurthy et al., 2014), and f must have derivatives of order at least d. Observe that the bounded support assumption does not hold in the case of autoencoders with Gaussian priors or encoders.

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.

3.4 Empirical evaluation

The previous section discussed theoretical properties of the proposed RAM-MC estimator. This section demonstrates its empirical performance. First, its behaviour is probed in a synthetic, controlled setting where all distributions and divergences are known. After this, a more realistic setting is considered in which the goal is to estimate a divergence between the aggregate posterior Q_Z and prior P_Z in pretrained autoencoder models.

3.4.1 Synthetic experiments

The aim is to investigate the behaviour of the RAM-MC estimator for various $d = \dim(\mathcal{Z})$ and f-divergences in a synthetic controlled setting. This is done by considering a setting in which Q_Z^λ, parametrised by a scalar λ, and P_Z are both d-variate Gaussians for $d \in \{1, 4, 16\}$. In this case $D_f(Q_Z^\lambda, P_Z)$ can be computed analytically for the KL, χ^2, and H^2 divergences (see Section 3.2.2). These analytic values can be used as baselines to compare with the

estimates obtained through use of RAM-MC. Specifically, take

$$\begin{aligned} P_Z &= \mathcal{N}(0, I_d), \\ Q^{\lambda}_{Z|X=x} &= \mathcal{N}\left(A_\lambda x + b_\lambda, \epsilon^2 I_d\right), \\ P_X &= \mathcal{N}\left(0, I_{20}\right), \end{aligned}$$

where the constant $\epsilon = 0.5$.[1] This results in $Q^{\lambda}_Z = \mathcal{N}\left(b_\lambda, A_\lambda A_\lambda^\intercal + \epsilon^2 I_d\right)$. b_λ was chosen by randomly sampling a vector v from the unit sphere and setting $b_\lambda = \lambda v$. A_λ was chosen by randomly sampling a $(d, 20)$-dimensional matrix A_0 with i.i.d. Gaussian entries and normalising it to have unit Frobenius norm, taking A_1 to be the similarly sized matrix with 1 on the main diagonal and 0 elsewhere, and setting $A_\lambda = \frac{1}{2}A_1 + \lambda A_0$.

Figure 3.1 shows the behaviour of RAM-MC with $N \in \{1, 500\}$ and M=128 compared to the ground truth as $\lambda \in [-2, 2]$ is varied. The columns of Figure 3.1 correspond to different dimensions $d \in \{1, 4, 16\}$, and rows to the KL, χ^2 and H^2 divergences, respectively. For each column, the values of A_0 and v were randomly sampled so that the distributions being compared are the same within columns and different between columns. Two other baseline methods are included for comparison. 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 (see Section 2.5).

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 $\hat{Q}^N_Z$ and RAM-MC (Equation 3.3) was evaluated. Using $\hat{Q}^N_Z$ as the proposal distribution resulted in significantly better results compared to using P_Z as the proposal distribution for all divergences. 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. Note that this density estimation approach ignores all of the structure present, and thus demonstrates the poor performance of an agnostic method. It would also be possible to exploit part of the problem structure and use one of the known density $p(z)$ or $\widehat{Q}^N_Z$, rather than estimating both by density estimation; these cases led to performance better than the naive density estimation approach, but still significantly worse than RAM-MC and were omitted from Figure 3.1 to avoid clutter. The M1 estimator involves solving a convex linear program in N variables to maximise a lower bound on the true divergence, see Nguyen et al., 2010 for more details. Although the M1 estimator can

[1] Multiple values of ϵ were experimented with, and although changing its value does change the scale of the plots in Figure 3.1, the general shapes and conclusions drawn from the results are the same.

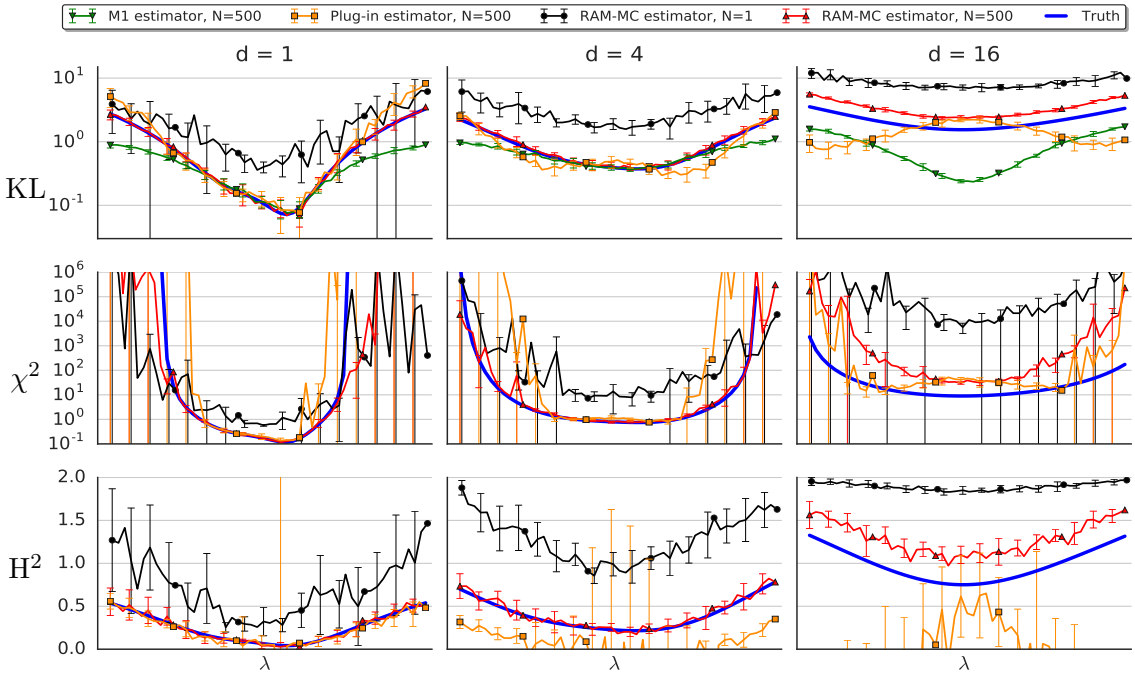


Figure 3.1 Results of synthetic experiments, Section 3.4.1. Estimating $D_f(\mathcal{N}(\mu_\lambda, \Sigma_\lambda), \mathcal{N}(0, I_d))$ for various f, d, and parameters μ_λ and Σ_λ indexed by $\lambda \in \mathbb{R}$. Horizontal axis correspond to $\lambda \in [-2, 2]$, columns to $d \in \{1, 4, 16\}$ and rows to KL, χ^2, and H^2 divergences respectively. **Blue** are true divergences, **black** and **red** are RAM-MC estimators (Equation 3.3) 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.

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.

The results of this experiment empirically support Proposition 1 and Theorems 3.11, 3.12, and 3.14: (i) in expectation, RAM-MC upper bounds the true divergence; (ii) increasing N from 1 to 500 clearly decreases 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. By side-stepping the issue of density estimation—that is, samples are drawn at the level of $\mathcal{X}$ so that Q_Z is approximated as a mixture of known components, rather than performing kernel density estimation based on samples at the level of $\mathcal{Z}$—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 due to the large variance.

3.4.2 Real-data experiments

To investigate the behaviour of RAM-MC in a more realistic setting, this experiment considers the estimation of divergences in the context of Variational Autoencoders (VAEs) and Wasserstein Autoencoders (WAEs), introduced in Section 2.6. Similar to the synthetic experiments, we are purely concerned with estimating $D_f(Q_Z, P_Z)$ with pre-trained models here, not actually training models from scratch. Recall that both VAEs and WAEs have a prior P_Z over the latent space and involve learning an encoder $Q^{\phi}_{Z|X}$ with parameter ϕ mapping from the data to latent space. Although the divergence $D_f(Q_Z, P_Z)$ does not appear in the VAE objective function, pretrained VAEs can nonetheless be used alongside WAEs as more realistic, higher dimensional settings to investigate estimation of this quantity compared to the simple synthetic setting considered in the previous section.

Pretrained models that were trained on the *CelebA* dataset (Liu et al., 2015) were used to evaluate the RAM-MC estimator as follows. The test dataset is taken as the ground-truth Q_X, and this is embedded into the latent space via the trained encoder. Since all models considered have Gaussian encoders, the resulting *empirical aggregate posterior* is a ∼20k-component Gaussian mixture for Q_Z, one component for each item in the test dataset. 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 (10^4 samples were used). This is computationally costly as it involves evaluating $2 \cdot 10^4$ Gaussian densities for each of the 10^4 MC points. This evaluation was repeated 10 times, and the means and standard deviations are reported in Figures 3.2 and 3.3 for the KL and H^2 divergences respectively. 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. This procedure was performed for the KL and H^2 divergences on six models that were chosen to have latent dimension $d \in \{32, 64, 128\}$ and were selected from the classes VAE, WAE-MMD and WAE-GAN.

Figure 3.2 shows the result of performing this for the KL divergence on six different models. 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 large. There is evidence supporting Theorem 3.14, which informally states 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. It was found that there are two general modes of behaviour of RAM-MC across the six trained models considered. In the bottom row of Figure 3.2, the decrease in bias with N is very obvious, supporting Proposition 1 and Theorems 3.11 and 3.12. 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 of Moon and Hero, 2014a and M1 of Nguyen et al., 2010) perform so poorly that their inclusion 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.

Figure 3.3 displays similar results for the H^2-divergence. Since $\mathrm{H}^2(A, B) \in [0, 2]$ for any probability distributions A and B and all computed estimates were close to 2, considerations of scale mean that the estimated values $\log\left(2 - \hat{D}^M_{\mathrm{H}^2}(\hat{Q}^N_Z, P_Z)\right)$ were plotted instead. Decreasing bias in N of RAM-MC therefore manifests itself as the lines *increasing* in Figure 3.3. 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 the results for the KL, these also support the theoretical findings presented in the previous section.

Additionally, the same experiment was attempted using the χ^2-divergence but numerical issues were encountered. 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 (Lemma 3.3). From Figure 3.2 it can be seen that the KL-divergence reaches values higher than 1000, making the corresponding value of the χ^2-divergence larger than can be represented using double-precision floats.

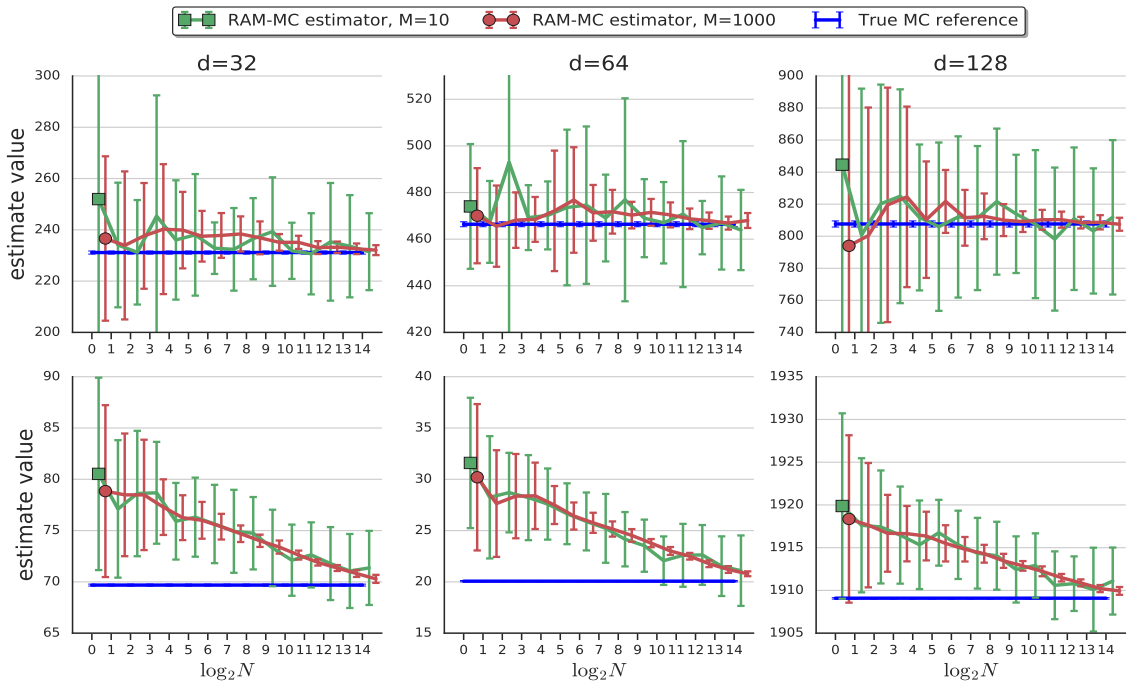


Figure 3.2 Results of real-data experiments, Section 3.4.2. Estimates of $\mathrm{KL}(Q_Z^\theta, P_Z)$ for pretrained autoencoder models with RAM-MC as a function of N for M=10 (**green**) and M=1000 (**red**) compared to an accurate MC estimate of the ground truth (**blue**). Lines and error bars represent means and standard deviations over 50 trials.

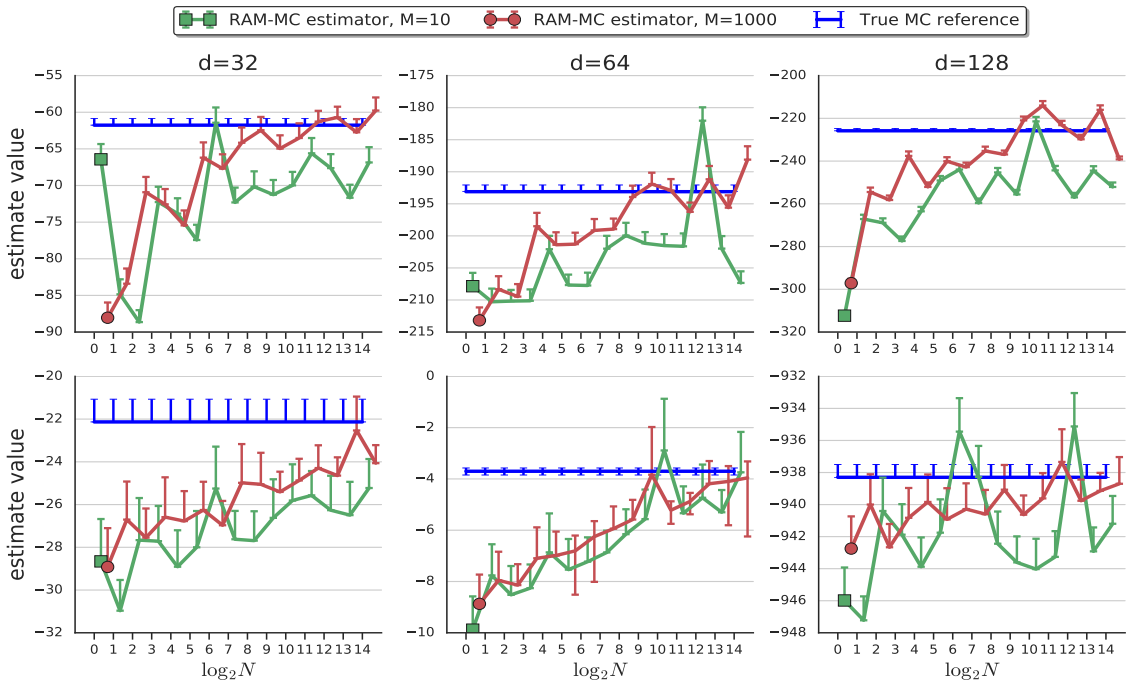


Figure 3.3 Results of real-data experiments, Section 3.4.2. Estimating $\mathrm{H}^2(Q_Z^\theta, P_Z)$ in pretrained autoencoder models with RAM-MC as a function of N for $M = 10$ (**green**) and M=1000 (**red**) compared to ground truth (**blue**). Lines and error bars represent means and standard deviations over 50 trials. Plots depict $\log\left(2 - \hat{D}^M_{\mathrm{H}^2}(\hat{Q}_Z^N, P_Z)\right)$ 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).

3.5 Applications

This section details some direct consequences of the proposed estimator and its theoretical guarantees to existing literature, illustrating the applicability of this work. We remind the reader that these results apply specifically to the setting of LVMs with probabilistic encoders.

3.5.1 Entropy estimation

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

$$\begin{aligned}
&H(Q_Z) - \mathbb{E}_{\mathbf{X}^N} H(\hat{Q}_Z^N) \\
&\quad = -\int q(z) \log q(z) dz + \mathbb{E}_{\mathbf{X}^N} \int \hat{q}_N(z) \log \hat{q}_N(z) dz \\
&\quad = -\int q(z) \log q(z) dz + \int q(z) \log p(z) dz \\
&\qquad - \int \mathbb{E}_{\mathbf{X}^N} \hat{q}_N(z) \log p(z) dz + \mathbb{E}_{\mathbf{X}^N} \int \hat{q}_N(z) \log \hat{q}_N(z) dz \\
&\quad = -\int q(z) \log \frac{q(z)}{p(z)} dz + \mathbb{E}_{\mathbf{X}^N} \int \hat{q}_N(z) \log \frac{\hat{q}_N(z)}{p(z)} dz \\
&\quad = \mathbb{E}_{\mathbf{X}^N} D_{\mathrm{KL}} \left(\hat{Q}_Z^N, P_Z \right) - D_{\mathrm{KL}} \left(Q_Z, P_Z \right).
\end{aligned}$$

Therefore, Theorems 3.11 and 3.12 provide sufficient conditions under which $H(\hat{Q}_Z^N)$ is an asymptotically unbiased estimator of $H(Q_Z)$. Note that it suffices for the assumptions of these results to hold for *any* choice of P_Z, suggesting that a more direct analysis of the behaviour of $H(\hat{Q}_Z^N)$ could yield milder sufficient conditions.

3.5.2 Total correlation estimation

The results on entropy estimation have consequences for some existing VAE literature. The Total Correlation (TC) of a distribution Q_Z, defined in terms of the KL-divergence, can be

written in terms of differential entropy:

$$\begin{aligned} TC(Q_Z) &:= D_{\mathrm{KL}}\left(Q_Z, \prod_{i=1}^{d_Z} Q_{Z_i}\right) \\ &= -\int q(z) \log\left(\frac{q(z)}{\prod_i q(z_i)}\right) dz \\ &= -\int q(z) \log q(z) dz + \int q(z) \sum_i \log q(z_i) dz \\ &= -\int q(z) \log q(z) dz + \sum_i \int q(z_i) \log q(z_i) dz \\ &= \sum_{i=1}^{d_Z} H(Q_{Z_i}) - H(Q_Z), \end{aligned}$$

where Q_{Z_i} is the ith marginal of Q_Z. This is considered by Chen et al., 2018b, who subtract it from the VAE loss function (see Section 2.6). Since a non-negative quantity is subtracted from a lower bound on the evidence, the resulting loss function is still an evidence lower bound, with additional encouragement for Q_Z to be factorised. This is named the β-TC-VAE algorithm. By the identities above, estimation of TC can be reduced to estimation of $H(Q_Z)$ with only slight modifications needed to treat $H(Q_{Z_i})$.

Two methods are proposed by 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. The results presented in this chapter therefore imply *inconsistency* of the MWS method due to the constant $\log D$ offset. This inconsistency is fact not problematic in the context of Chen et al., 2018b since they are concerned with minimising (not estimating) the TC, and a constant offset does not affect gradient-based optimization techniques. Interestingly, although their derivations suppose a data distribution of finite support, the results presented here show that minor modifications result in an estimator suitable for both finite and infinite support data distributions.

3.5.3 Mutual information estimation

The mutual information (MI) between variables with joint distribution $Q_{Z,X}$ is defined as

$$I(Z, X) := D_{\mathrm{KL}}\left(Q_{Z,X}, Q_Z Q_X\right) = \mathbb{E}_X D_{\mathrm{KL}}\left(Q_{Z|X}, Q_Z\right).$$

Several recent papers have estimated or optimised this quantity in the context of autoencoder architectures, coinciding with the setting considered here (Hoffman and Johnson, 2016; Alemi et al., 2018; Dieng 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:

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

The gap in this inequality can be written as

$$I(Z,X) - I_{TCPC}^N(Z,X) = \mathbb{E}_{\mathbf{X}^N} D_{\mathrm{KL}}\left(\hat{Q}_Z^N, P_Z\right) - D_{\mathrm{KL}}\left(Q_Z, P_Z\right)$$

where P_Z is *any* distribution. Therefore, the results in this chapter also provide sufficient conditions under which I_{TCPC}^N is an asymptotically unbiased estimator of the true mutual information.

3.5.4 Related, but fundamentally different work

Burda et al., 2015 propose to reduce the gap introduced by Jensen's inequality in the derivation of the classical ELBO by using multiple Monte-Carlo samples from the approximate posterior $Q_{Z|X}$. This is similar in flavour to the approach considered in this chapter, but is fundamentally different since the approach taken here uses multiple samples from the *data distribution* to reduce a different Jensen gap.

To avoid confusion, note that replacing the 'regulariser' term $\mathbb{E}_X[D_{\mathrm{KL}}(Q_{Z|X}, P_Z)]$ of the classical ELBO with expectation of the proposed estimator $\mathbb{E}_{\mathbf{X}^N}[D_{\mathrm{KL}}(\hat{Q}_Z^N, P_Z)]$ results in an upper bound of the classical ELBO (by Proposition 1) but is itself not in general an evidence lower bound:

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

3.6 Conclusion

This chapter 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 $\hat{Q}_Z^N = \frac{1}{N}\sum_n Q_{Z|X_n}$, and RAM-MC is the version of this estimator where $D_f(\hat{Q}_Z^N, P_Z)$ is estimated with MC sampling.

Rates of convergence and concentration were proved for both RAM and RAM-MC, in terms of sample size N and MC samples M under a variety of choices of f. Due to the strong structural assumptions made on the forms of the distributions in question, the fast rates presented here hold under relatively mild and verifiable further assumptions, thus making them applicable to the estimation of divergences in the latent spaces of autoencoders. In contrast, in the existing

literature on f-divergence estimation, which generally makes few structural assumptions, fast rates are only obtained under strong assumptions on the smoothness of densities or density ratios of the distributions.

Synthetic and real-data experiments strongly support the validity of the proposal in practice, and the theoretical results provide guarantees for methods previously proposed heuristically in existing literature for mutual information and total correlation estimation, thus extending understanding of the conditions under which these quantities can be estimated with practical numbers of samples.

Future work should investigate the use of the proposals for optimization loops, in contrast to pure estimation. When $Q^{\phi}_{Z|X}$ depends on parameter ϕ and the goal is to minimise $D_f(Q^{\phi}_Z, P_Z)$ with respect to ϕ, RAM-MC provides a practical surrogate loss that can be minimised using stochastic gradient methods. The obvious setting to apply this is in training WAEs, where the main problem is minimisation of a divergence term $D(Q^{\phi}_Z, P_Z)$, discussed in Section 2.6.3. Indeed, the research presented in this chapter was originally motivated by the study of WAEs. Preliminary experiements not included in this thesis found little improvement when using RAM-MC as the latent-space matching penalty compared to existing methods (WAE-MMD and WAE-GAN). It is possible that this lack of improvement is attributable to deficiencies in the proposed method, or to the use of f-divergences in this setting more generally. It may instead be due to the insufficiently thorough nature of those preliminary experiments.

Another more theoretical direction of research relating the proposed RAM-MC estimator to its application in training WAEs would be to investigate how bounds on the divergence $D_f(Q^{\phi}_Z, P_Z)$ in combination with the reconstruction error of a WAE relate to the optimal transport distance $OT(P^{\theta}_X, Q_X)$ that a WAE is ultimately supposed to minimise. Work in this direction has been done by Patrini et al., 2018, who relate the reconstruction error and latent-space optimal transport distance $OT(Q^{\phi}_Z, P_Z)$ with the data-space optimal transport distance $OT(P^{\theta}_X, Q_X)$. However, it is not clear whether such a connection must hold for other choices of latent-space divergences, since the 'relaxed' WAE objective (Equation 2.7) is itself a heuristic approximation to the original 'constrained' optimal transport formulation (Equation 2.6).

Chapter 4

Multi-view Nonlinear Independent Component Analysis

This chapter presents identifiability results in a novel multi-view nonlinear ICA setting. In the usual single-view setting, identifiability holds only under strong assumptions on the source distribution and mixing functions. The results presented here show that when multiple distinct views of the sources are available, identifiability holds under much weaker assumptions. This work is an important contribution to the literature as it extends the few known identifiability results for nonlinear ICA models.

The main technical content of this chapter has been published in the paper:

> Luigi Gresele*, Paul K Rubenstein*, Arash Mehrjou, Francesco Locatello and Bernhard Schölkopf. "The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-view Nonlinear ICA". *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*. *Joint first authorship. 2019.

4.1 Introduction

Independent Component Analysis (ICA) is often motivated by the so-called *cocktail-party problem*. When two conversations at a party are happening simultaneously, a listener will hear different mixtures of the two audio streams produced by the speakers in each of their ears. Despite both ears receiving mixtures of the conversations, the listener is able to focus on either of the conversations separately, hearing and understanding one while ignoring the other. This is due to the brain's ability to separate out the mixed audio streams into the separate underlying sources, one for each conversation.

More generally, given data that are mixtures of independent underlying sources, the goal of ICA is to 'unmix' the data, thus recovering the sources. This provides a principled approach to the disentanglement of independent latent components, blind source separation, and feature extraction (Hyvärinen and Oja, 2000). The applications of ICA 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).

The ICA problem can be written formally by defining the latent variable model

$$Z \sim p(z) = \prod_i p(z_i), \tag{4.1}$$

$$X = f(Z), \tag{4.2}$$

where Z is a vector of independent *sources*, X is a vector of *observations* or *mixtures* and f is the vector of *mixing functions* expressing how each coordinate of X depends on all of the coordinates of Z. Crucially, f is assumed to be invertible and thus X and Z are of the same dimension. Given a dataset of observations of X, the goal of ICA is to recover the corresponding unknown values of Z by learning to invert the unknown f. In the general case, no assumptions are made about $p(z)$ other than that it factorises. The latent variable model in this setting should thus be thought of as a true descriptive model for the world, in contrast to the neural sampler setting, where latent variable models with factorised priors merely provide a convenient way to specify distributions over the observable variables.

An ICA problem is known as *identifiable* when it is possible to recover the sources Z up to tolerable ambiguities. For instance, it is generally acceptable to recover Z up to linear rescaling or permuted coordinates. Although identifiability results generally assume access to unlimited samples of data, they are crucial for ensuring the reliability of ICA methods in practical scenarios; in the absence of these, there is no guarantee that a proposed method will successfully reconstruct the true sources, even in controlled settings.

It was proved by Hyvärinen and Pajunen, 1999 that the 'vanilla' ICA setting, in which only independence of the sources and invertibility of f are assumed, is *non-identifiable*. Thus, much of the research in this field has attempted to characterise the assumptions under which identifiability holds. Such assumptions may be made either on the mixing functions or on the distributions of the sources.

(a) (b)

Figure 4.1 Graphical models depicting (a) the usual single-view ICA setting; (2) the multi-view setting considered in this work. Z is a vector of unobserved latent sources and the X_i are observed mixtures of the components of Z. In both settings, all variables are of the same dimension.

4.1.1 Summary of results

The central contribution of the work presented in this chapter is the derivation of identifiability results in a novel *multi-view* setting, in which multiple observations of the same underlying sources through different mixing functions are given (see Figure 4.1b):

$$\begin{aligned} Z &\sim \prod_i p(z_i), \\ X_1 &= f_1(Z), \\ X_2 &= f_2(Z), \end{aligned}$$

where again each f_i is invertible and X_1, X_2 and Z are of the same dimension. In this setting, identifiability holds subject to much weaker assumptions than are required for the single-view setting. These results are of practical importance for applications in which multiple data modalities may be simultaneously available.

In particular, variants on the model above are considered in which noise processes occur at the source level in one or both views. Subject to certain assumptions holding, formalised in the *Sufficiently Distinct Views* assumption (see Definition 4.2), model identifiability is proved, meaning that it is in principle possible to recover the sources up to the noisy corruptions (Theorems 4.1–4.5 and Corollary 4.3). Although some infinitesimally small amount of noise is required for the results to hold, the low-noise limiting case is also analysed (Corollary 4.6). Finally, one might hope that even in the presence of large amounts of noise, having access to a larger number of views should be beneficial. First results in this direction are given in in Theorem 4.8.

For a high-level intuition of the results, consider the perception of 3D structure through eyesight. Each of our eyes only sees a 2D projection of the true state of the 3D world. With only a single eye available, it may be possible to infer the 3D structure present in a scene

by exploiting known cues, such as the objects in the scene being familiar to the viewer or the presence of shadows. But with two eyes together, a human can perceive 3D structure immediately, even without such cues. Analogously, the results presented in this work show that inference of latent structure can be performed under much weaker assumptions when more than a single view is available.

The remainder of this chapter is structured as follows. Section 4.2 provides an introduction to ICA and the literature surrounding it, as well as literature from other areas that is relevant to the setting considered in this work. Section 4.3 presents the main results. Section 4.4 discusses the assumptions of these results, and Section 4.5 concludes.

4.2 Overview of ICA and related literature

This section provides an overview of different ICA settings for which identifiability results are known, and discusses other literature relevant to the multi-view setting considered in this work.

4.2.1 Linear ICA

Linear ICA refers to the setting in which the vector of mixing functions f is a linear map, in which case the ICA model can be written

$$\begin{aligned} Z &\sim \prod_i p(z_i), \\ X &= AZ, \end{aligned}$$

where A is a square matrix. This problem has been extensively studied and has been shown to be identifiable if at most one of the latent components is Gaussian (Darmois, 1953; Skitovich, 1954; Comon, 1994). Nonidentifiability in the case of more than one Gaussian component is a consequence of the fact that an isotropic Gaussian is invariant under orthogonal linear mappings. Thus if the diagonal matrix Λ rescales the Gaussian components of Z to be unit variance, and U is an orthogonal matrix mixing these components, X can be rewritten

$$X = \left(A\Lambda^{-1}U^{-1}\right)(U\Lambda Z).$$

Since both Z and $U\Lambda Z$ have independent components, it is impossible to tell which of Z and $U\Lambda Z$ corresponds to the true source distribution. For most such U, $U\Lambda Z$ nontrivially mixes the components of Z, and thus the problem is nonidentifiable due to the existence of multiple plausible, yet fundamentally different, solutions.

Note that even in the non-Gaussian case, taking $U = I$ or a permutation matrix also results in $U\Lambda Z$ being a valid solution. This, however, corresponds to a 'trivial' indeterminacy of the linear ICA problem, since in this case $U\Lambda Z$ still recovers the separate components of Z, only linearly rescaled and in a different order.

The non-Gaussianity assumption is exploited by linear ICA algorithms by seeking linear maps W such that the transformed data WX have maximally non-Gaussian components. Intuition for why such an approach works can be seen in the Central Limit Theorem which, informally, states that an average of i.i.d. random variables becomes more Gaussian-like as the number of variables in the average increases. Similarly, in a sense that can be made formal (Hyvärinen and Oja, 2000), linearly mixing random variables makes them more Gaussian-like, meaning that appropriate measures of Gaussianity can be used as objective functions for de-mixing.

Linear ICA methods can be used for causal discovery in linear causal models, as will be discussed in Section 5.3.2.

4.2.2 Nonlinear ICA

It was proved by Hyvärinen and Pajunen, 1999 that if only independence of the sources and invertibility of f are assumed, the nonlinear ICA problem is unidentifiable. Specifically, given any distribution over the observable variables X that admits density with respect to Lebesgue measure, there exist many vector-valued invertible mappings g with the property that the components of $g(X)$ are independent, and these many solutions are non-trivially different.

This is proved by first demonstrating the existence of a function g with the property that $Y = g(X)$ is uniformly distributed on the unit cube $[0, 1]^n$, a generalisation of the result that, for a one-dimensional random variable U with cumulative distribution function F_U, the random variable $F_U(U)$ is uniformly distributed on $[0, 1]$. Next, non-uniqueness is proved by demonstrating the existence of an infinite class of functions h which are *measure-preserving* maps $[0, 1]^n \to [0, 1]^n$. That is, if Y is uniformly distributed on $[0, 1]^n$ then so is $Y' = h(Y)$. It follows that $h \circ g$ thus provides a valid solution to the nonlinear ICA problem and thus there are infinitely many solutions. Such a class of measure-preserving functions is given explicitly in the case of $n = 2$ dimensions; by extending such functions to the identity mapping on extra dimensions and composing, such a class can be generated for any n.

Note that any function $k : \mathbb{R}^n \to \mathbb{R}^n$ that acts coordinate-wise and is invertible—that is, for each i, $k(X)_i = k_i(X_j)$ for some j with k_i invertible—can be composed with g to result in the random vector $Y' = k \circ g(X)$ having any desired factorised distribution. This is in some sense a 'trivial' indeterminacy of nonlinear ICA, analogous to the scalar and permutation indeterminacy of linear ICA. All of the novel identifiability results presented in Section 4.3

hold only up to such functions, which are referred to throughout as *component-wise invertible transformations*.

Other works have shown that identifiability is possible when additional assumptions are made. Mostly these assume that the observations correspond not to i.i.d. samples of the sources, but rather time series with temporal structure (Singer and Coifman, 2008; Sprekeler et al., 2014). In contrast, Taleb and Jutten, 1999 prove identifiability under the rather strong *post-nonlinear mixing* assumption on the mixing functions, corresponding to linear mixing followed by a nonlinear component-wise invertible function.

4.2.3 Nonlinear ICA with auxiliary variables

Hyvärinen et al., 2019, generalising the results of Hyvärinen and Morioka, 2016 and Hyvärinen and Morioka, 2017, study a modification of the typical ICA setting where an additional observed auxiliary variable U is introduced. U is assumed to be always observed with the sources Z being *conditionally* independent given U, resulting in the model

$$Z|U \sim p(z|u) = \prod_i p_i(z_i|u), \tag{4.3}$$

$$X = f(Z). \tag{4.4}$$

This general model includes temporally dependent sources as a special case, taking $(U, X) = (X_t, X_{t+1})$. Hyvärinen et al., 2019 prove identifiability results under conditions on both the conditional distributions $p_i(z_i|u)$, the relationships between sources and auxiliary variables, and subject to U having a sufficiently diverse influence on X in a sense that is formalised as the *assumption of variability*.

Identifiability in this model is proved constructively by considering classification between tuples (x, u), sampled from the joint distribution $p(x, u)$, and tuples (x, u^*) sampled from the product of marginals $p(u)p(x)$. Tuples from the former distribution correspond to the same value of the sources z, and thus share information, while tuples from the latter correspond to different sources and thus do not share information. By appropriately constraining the form of the regression function used in this classification, it is shown that the optimal classifier extracts z up to component-wise invertible functions.

The results presented in Section 4.3 build on this approach, extending the results to a novel multi-view setting.

4.2.4 Other related work

A central concept of the work presented in this chapter is the extraction of features from multiple simultaneous views. This section briefly reviews some related work considering similar settings outside of the ICA literature.

Canonical Correlation Analysis

Given two vector-valued random variables, the goal of Canonical Correlation Analysis (CCA) is to find a 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 (Hotelling, 1992; Bishop, 2006). CCA admits a probabilistic interpretation (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 4.1b.

The main differences compared to the setting of this chapter are that the latent components retrieved in CCA are forced to be uncorrelated, whereas ICA is concerned with independent components; and in CCA, mappings between the sources and observations are linear, whereas this work considers nonlinear mappings. In dealing with correlation instead of independence, CCA is more closely related to Principal Component Analysis (PCA) than to ICA. Nonlinear extensions of the basic CCA framework have been proposed (Lai and Fyfe, 2000; Fukumizu et al., 2007; Andrew et al., 2013; Michaeli et al., 2016), but identifiability results in the sense considered in this work are lacking.

Multi-view latent variable models

Song et al., 2014 prove identifiability for multi-view, discrete latent variable models. While the setting they consider is similar to that of this work, their proposed method is aimed at estimating model parameters with the goal of performing density estimation, rather than estimating the values of (continuous) latent variables. The paper considers a setting in which L variables X_l, $l = 1, \ldots, L$ are observed; additionally, there exists an unobserved discrete latent variable H, such that conditional distributions $P(X_l|H)$ are independent. 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. This line of work differs from the setting considered here in two key ways. First, clustering can be thought of as assigning a discrete latent label per observation. In contrast,

the setting considered here is concerned with recovery of a continuous latent vector for each observation. Second, since no underlying generative model with discrete latent variables is assumed, identifiability results are not given.

Half-sibling regression

Half-sibling regression (Schölkopf et al., 2016) is a method to reconstruct a source from noisy observations by exploiting observations of other sources that are affected by the same noise process. In contrast to the multi-view ICA setting, in which the sources to be reconstructed are common to the multiple views, in half-sibling regression it is the *noise* that is common to both views, with the desired sources being separate for each observation.

Schölkopf et al., 2016 study this problem under an additive noise assumption. By regressing one observation against the other, this common noise can be identified and hence subtracted, recovering the desired sources.

4.3 Nonlinear ICA with multiple views

This section presents the main contribution of this chapter, in which identifiability results for variations on the following setting are given:

$$Z \sim p(z) = \prod_i p_i(z_i), \tag{4.5}$$

$$X_1 = f_1(Z), \tag{4.6}$$

$$X_2 = f_2(Z), \tag{4.7}$$

where $X_1, X_2, Z \in \mathbb{R}^D$ and f_1, f_2 are arbitrary smooth and invertible transformations of the latent variable $Z = (Z_1, \ldots, Z_D)$ with smooth inverse. X_1 and X_2 are referred to as different *views* of the sources Z. Since each f_i is invertible, X_1, X_2 and Z are of the same dimension. Given observations of X_1 and X_2, the goal is to recover Z, undoing the mixing induced by the f_i.

The two problems defined by separately considering the pairs of Equations 4.5, 4.6 and 4.5, 4.7 are instances of the usual single-view nonlinear ICA setting. As previously discussed, unless strong assumptions are made on the f_i or the distribution of Z, these problems are separately unidentifiable.

The key contribution of this chapter is the derivation of identifiability results with relaxed assumptions by exploiting the fact that the two views are connected through the shared

latent variable Z. That is, observing X_1 and X_2 together provides sufficient information to remove the ambiguities present in the vanilla nonlinear ICA setting.

This section considers three instances of the general setting described above, providing identifiability results for each. Specifically:

- Section 4.3.1 considers the case that only one of the observations, X_2, is corrupted with noise, showing that it is possible to fully reconstruct Z using the noiseless variable. This corresponds to a setting in which one accurate measurement device is supplemented with a second noisy device.
- Section 4.3.2 considers the case that both variables are corrupted with noise, showing that it is possible to recover Z up to the corruptions. Furthermore, it is shown that Z can be recovered with arbitrary precision in the limit that the corruptions go to zero.
- Section 4.3.3 considers the case of M simultaneous views of the source Z rather than just two. When considering the limit $M \to \infty$, sufficient conditions are provided under which it is possible to reconstruct Z even if each observation is corrupted by noise.

The setting considered in this work is related to that of Hyvärinen et al., 2019, discussed in Section 4.2.3, and the approach to proving the identifiability results presented here builds on the technique presented in that work. This approach is to classify between pairs (X_1, X_2) corresponding to the same Z and (X_1, X_2^*) corresponding to different realisations of Z. This classification problem can only be solved by employing the information shared by the simultaneous views in order to distinguish the two classes. By placing constraints on the regression function used in such a classifier, it can be shown that the representation obtained by taking an intermediate layer of this classifier recovers Z up to tolerable ambiguities.

In contrast to the complete setting considered by Hyvärinen et al., 2019, the model considered here is undercomplete by a factor of two. That is, the number of observed dimensions is twice that of the number of latent dimensions. As such, one may expect identifiability under more general conditions due to the increased number of constraints.

For technical reasons discussed in Section 4.4.2, the results require some stochasticity in the relationship between Z and at least one of the X_i. This is not a significant constraint in practice; in most real settings observations are corrupted by noise, and a truly deterministic relationship between Z and the X_i would be unrealistic. Component-wise independent corruptions of the sources are considered, i.e. $\mathbb{R}^D$-valued noise vectors N_1 and N_2 are introduced, and $X_1 = f_1 \circ g_1(Z, N_1)$ with $g_{1i}(Z, N_1) = g_{1i}(Z_i, N_{1i})$, where the components of N_1 are mutually independent, and similar for N_2 and X_2. The noise variables N_1, N_2 and the sources Z are assumed to be mutually independent. This constrains the way the source is corrupted by noise, namely the g_i, and not the mixing functions f_i. In the the vanilla ICA setting, inversion of the mixing function and recovery of the sources Z are equivalent; in the

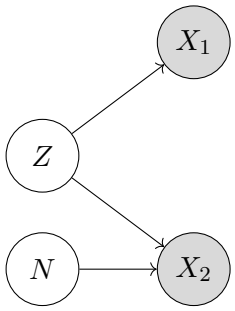


Figure 4.2 The setting considered in Section 4.3.1. Two views of the sources are available, one of which, X_1, is not corrupted by noise. In this and all subsequent figures in this chapter, each node is a deterministic function of its parents in the graph.

setting considered here, inversion of the mixing f_i only implies recovering the sources up to the effect of the corrupter g_i.

Such g_i as described in the previous paragraph are referred to as *component-wise corrupters* throughout, and the corresponding output as *corruptions*. All identifiability results hold only up to *component-wise invertible transformations*, meaning that the components of Z are recovered, but possibly reparametrised and in a permuted order.

4.3.1 One noiseless view

Consider the following model in which one noiseless and one noisy view of the sources are given, represented in Figure 4.2,

$$Z \sim p(z) = \prod_i p(z_i), \tag{4.8}$$
$$N \sim p(n) = \prod_i p(n_i),$$
$$X_1 = f_1(Z), \tag{4.9}$$
$$X_2 = f_2(g(Z, N)), \tag{4.10}$$

where f_1 and f_2 are invertible, g is a component-wise corrupter, $N \perp\!\!\!\perp Z$ and X_1 and X_2 are observed. The following theorem demonstrates assumptions under which identifiability in this model holds. This result is quite involved; it will first be stated, then discussed.

Theorem 4.1. *The difference between the log joint probability and log product of marginals of the observed variables in the model given in Equations 4.8-4.10 admits the following*

factorisation:

$$\begin{aligned}
&\log p(x_1, x_2) - \log p(x_1)p(x_2) \\
&= \log p(x_2|x_1) - \log p(x_2) \\
&= \left(\sum_i \alpha_i(z_i, g_i(z_i, n_i)) + \log \det J \right) \\
&\quad - \left(\sum_i \delta_i(g_i(z_i, n_i)) + \log \det J \right) \\
&= \sum_i \alpha_i(z_i, g_i(z_i, n_i)) - \sum_i \delta_i(g_i(z_i, n_i)),
\end{aligned} \tag{4.11}$$

where $z_i = f_{1i}^{-1}(x_1)$, $g_i(z_i, n_i) = f_{2i}^{-1}(x_2)$, *and* J *is the Jacobian of the transformation* f_2^{-1} *(note that the introduced Jacobians cancel*[1]*). Suppose that*

1. α *satisfies the* Sufficiently Distinct Views *assumption (see after this theorem).*
2. *A classifier is trained to discriminate between*

$$(X_1, X_2) \text{ vs. } (X_1, X_2^*),$$

   where (X_1, X_2) *correspond to the same realisation of* Z *and* (X_1, X_2^*) *correspond to different realisations of* Z*.*

3. *The classifier minimises the logistic regression loss, and is constrained to use a regression function of the form*

$$r(x_1, x_2) = \sum_i \psi_i(h_i(x_1), x_2),$$

   where $h = (h_1, \ldots, h_n)$ *is invertible, smooth and has smooth inverse.*

Then, in the limit of infinite data and with universal approximation capacity, h *inverts* f_1 *in the sense that the* $h_i(X_1)$ *recover the independent components of* Z *up to component-wise invertible transformations.*

An outline of the proof for this result is provided below after discussing some of the assumptions; full proof can be found in Appendix B.1.1.

The assumption of invertibility for h could be satisfied by, e.g., the use of normalizing flows (Rezende and Mohamed, 2015; Chen et al., 2018c) or deep invertible networks (Jacobsen et al., 2018).

[1]Several subsequent results in this section consider the difference between two log-probabilities. In all of these cases, the Jacobians introduced by the change of variables cancel out. For brevity these Jacobians are omitted henceforth.

If X_1 and X_2 were always equal, the multiple view setting would reduce to the normal nonlinear ICA setting. The *Sufficiently Distinct Views (SDV)* assumption formalises a sense in which the two views must be sufficiently different from one another, resulting in more information being available in totality than from each view individually. In the context of Theorem 4.1, 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 4.2 (Sufficiently Distinct Views)**.** *Let $\alpha_i(y_i, t_i)$, $i = 1, \ldots, D$ be functions of two arguments. Denote by α the vector of functions and define*

$$\alpha_i'(y_i, t_i) = \partial \alpha_i(y_i, t_i)/\partial y_i, \tag{4.12}$$

$$\alpha_i''(y_i, t_i) = \partial^2 \alpha_i(y_i, t_i)/\partial y_i^2, \tag{4.13}$$

$$w_\alpha(y, t) = (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D'). \tag{4.14}$$

α satisfies the assumption of Sufficiently Distinct Views (SDV) *if for any value of y, there exist $2D$ distinct values t^j, $j = 1, \ldots, 2D$ such that the vectors $w_\alpha(y, t^j)$ are linearly independent.*

This is closely related to the Assumption of Variability in Hyvärinen et al., 2019. The SDV assumption is discussed in further detail in Section 4.4.1, where simple cases of conditional log-probability density functions satisfying and violating the assumption are presented.

Sketch proof of Theorem 4.1. The first observation to be made is that for logistic regression, the optimal regression function for the logit $r(x_1, x_2)$ is equal to the log density-ratio between the two distributions being distinguished, namely $\log\left(p(x_1, x_2)/p(x_1)p(x_2)\right) = \log p(x_1, x_2) - \log p(x_1)p(x_2)$ (as discussed in Section 2.5). Thus, in the limit of infinite data and with universal approximation capacity, the following equality holds:

$$\sum_i \psi_i(h_i(x_1), x_2) = \sum_i \alpha_i(z_i, g_i(z_i, n_i)) - \sum_i \delta_i(g_i(z_i, n_i)).$$

By performing the change of variables $y = h(x_1)$, $t = f_2^{-1}(x_2)$, and defining $v(y) := f_1^{-1}(h^{-1}(y)) = z$, this equation can be rewritten

$$\sum_i \psi_i(y_i, f_2(t)) = \sum_i \alpha_i(v_i(y), t_i) - \sum_i \delta_i(t_i). \tag{4.15}$$

The goal is to show that $v_i(y)$ depends on exactly one coordinate of y, so that $v_i(y) = v_i(y_j)$ for some j. Since $z = v(y)$, this implies that $z_i = v_i(y_j)$ is a function only of y_j, which in turn implies that z_i is a function only of $h_j(x_1)$. Since $v = f_1^{-1} \circ h^{-1}$ is the composition of two invertible functions, it is itself invertible, and since each component of v depends only

one component of its input, each of the components are also invertible. It follows that z_i is an invertible function of $h_j(x_1)$, and so $h(x_1)$ recovers z up to permutations and coordinate-wise invertible transformations.

Showing that $v_i(y) = v_i(y_j)$ for some j is somewhat technically involved, and it is here that the SDV assumption is required. It is proved by taking partial derivatives of Equation 4.15 with respect to y_j and $y_{j'}$ for $j \neq j'$. This results in an expression involving first- and second-order derivatives of α_i and v_i in which the expressions in the SDV assumption appear. If the SDV assumption holds, it follows that the derivative of v_i with respect to y_j is non-zero for at most one one value of j, meaning that v_i does not depend on all other $y_{j'}$. □

See the full proof in Appendix B.1.1 for further details; proofs of subsequent results Corollary 4.3 and Theorems 4.4 and 4.5 proceed similarly to this, and thus sketches of these results will be omitted.

Theorem 4.1 shows that by jointly considering the two views, it is possible to recover Z, in contrast to the single-view setting. This result can be extended to learn the inverse of f_2 up to component-wise invertible functions.

Corollary 4.3. *Consider the setting of Theorem 4.1 with the alternative factorisation of the log joint probability*

$$\begin{aligned}
&\log p(x_1, x_2) - \log p(x_1)p(x_2) \\
&= \log p(x_1|x_2) - \log p(x_1) \\
&= \sum_i \gamma_i(z_i, g_i(z_i, n_i)) - \sum_i \beta_i(z_i)) \,.
\end{aligned} \tag{4.16}$$

Suppose that γ satisfies the SDV assumption. Replacing the regression function with

$$r(x_1, x_2) = \sum_i \psi_i(x_1, h_i(x_2))$$

results in h inverting f_2 in the sense that the $h_i(X_2)$ recover the independent components of the $g(Z, N)$ up to component-wise invertible transformations.

The proof can be found in Appendix B.1.2. Theorem 4.1 and Corollary 4.3 together mean that it is possible to learn inverses h_1 and h_2 of f_1 and f_2, and therefore to recover Z and $g(Z, N)$, 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 h_1 and h_2 are 'aligned' in the sense that for each i the components $h_{1i}(X_1)$ and $h_{2i}(X_2)$ correspond to the same components of Z.

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

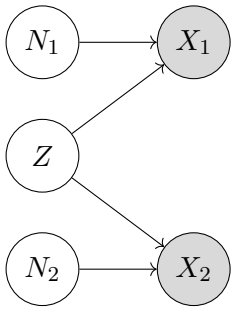


Figure 4.3 Setting with two views of the sources Z, both corrupted by noise.

Theorem 4.4. *Consider the settings of Theorem 4.1 and Corollary 4.3. Suppose that both α and γ satisfy the SDV assumption. Replacing the regression function with*

$$r(x_1, x_2) = \sum_i \psi_i(h_{1,i}(x_1), h_{2,i}(x_2)) \tag{4.17}$$

results in h_1, h_2 inverting f_1, f_2 in the sense that the $h_{1,i}(X_1)$ and $h_{2,i}(X_2)$ recover the independent components of Z and $g(Z, N)$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,

$$h_{1,i}(X_1) \perp\!\!\!\perp h_{2,j}(X_2).$$

The proof can be found in Appendix B.2.1. Note that Theorem 4.4 is *not* a generalisation of Theorem 4.1 or Corollary 4.3, since it makes stricter assumptions by imposing the SDV assumption on both α and γ. In contrast, Theorem 4.1 and Corollary 4.3 require that only one is valid for each. For cases in which finding aligned representations for Z and $g(Z, N)$ are desired, Theorem 4.4 should be applied. If the only goal is recovery of Z, the assumptions of Theorem 4.1 are easier to satisfy.

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, these results show that jointly exploiting the multiple modalities can lead to identification of the true underlying sources in a manner not attainable through use of the more reliable modality alone.

4.3.2 Two noisy views

Consider next the setting in which both variables are corrupted by noise, depicted in Figure 4.3 and described by the following model:

$$X_1 = f_1(g_1(Z, N_1)),$$
$$X_2 = f_2(g_2(Z, N_2)),$$

where all variables take value in $\mathbb{R}^D$, f_1 and f_2 are nonlinear, invertible, deterministic functions, g_1 and g_2 are component-wise corrupters, and Z and the N_i are independent with independent components. This class of models generalises the setting of Section 4.3.1, since by taking $g_1(Z, N_1) = Z$ it reduces to the case of one noiseless observation.

The log density-ratio $\log p(x_1, x_2) - \log p(x_1)p(x_2)$ admits similar factorisations to those given in Equations 4.11 and 4.16:

$$\begin{aligned}
&\log p(x_1, x_2) - \log p(x_1)p(x_2) \\
&= \log p(x_1|x_2) - \log p(x_1) \\
&= \sum_i \eta_i(g_{1i}(z_i, n_{i1}), g_{2i}(z_i, n_{2i})) - \sum_i \theta_i(g_{1i}(z_i, n_{1i}) \qquad (4.18) \\
&= \log p(x_2|x_1) - \log p(x_2) \\
&= \sum_i \lambda_i(g_{2i}(z_i, n_{2i}), g_{1i}(z_i, n_{1i})) - \sum_i \mu_i(g_{2i}(z_i, n_{2i})). \qquad (4.19)
\end{aligned}$$

Since access is only given to corrupted observations, exact recovery of Z is not possible. Nonetheless, a generalisation of Theorem 4.4 holds showing that the f_i can be inverted and Z recovered up to the corruptions induced by the N_i via the g_i.

Theorem 4.5. *Suppose that η and λ satisfy the SDV assumption. The algorithm described in Theorem 4.1 with regression function specified in Equation 4.17 results in h_1 and h_2 inverting f_1 and f_2 in the sense that the $h_{1,i}(X_1)$ and $h_{2,i}(X_2)$ recover the independent components of $g_1(Z, N_1)$ and $g_2(Z, N_2)$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,*

$$h_{1,i}(X_1) \perp\!\!\!\perp h_{2,j}(X_2).$$

The proof can be found in Appendix B.2.1.

The common source Z can thus be recovered up to the corruptions $g_i(Z, N_i)$. In the limit of the magnitude of one of the noise variables going to zero, the reconstruction of the sources Z attained through the corresponding view is exact up to the component-wise invertible functions, as stated in the following corollary.

Corollary 4.6. *Let $N_1^{(k)} = \frac{1}{k} \cdot \tilde{N}$ for $k \in \mathbb{N}$, where $\tilde{N} \in \mathbb{R}^D$ is a fixed random variable with finite variance, and let N_2 be a random variable that does not depend on k. Let $h_1^{(k)}, h_2^{(k)}$ be the output of the algorithm specified by Theorem 4.5 with noise variables $N_1^{(k)}$ and N_2.*

Suppose that the corrupters g_i satisfy the following two criteria:

1. $\exists a \in \mathbb{R}^D_{>0}$ *s.t.* $\left|\frac{\partial g_1(z,n)}{\partial n}\right|_{n=0} \leq a$ *for all* z,
2. $\exists b \in \mathbb{R}^D_{>0}$ *s.t.* $0 < \frac{\partial g_1(z,0)}{\partial z} \leq b$.

Then, denoting by $\mathcal{E}$ the set of all component-wise, invertible functions, it holds that

$$\inf_{e \in \mathcal{E}} \left\| Z - e(h_1^{(k)}(X_1)) \right\| \xrightarrow[k \to \infty]{p} 0,$$

where p denotes convergence in probability.

Proof sketch; see Appendix B.2.2 for full proof. The key idea of the proof is to rewrite $e(h_1^{(k)}(X_1))$ as $\tilde{e} \circ g_1(Z, N_1^{(k)})$ for some $\tilde{e} \in \mathcal{E}$, and to Taylor expand $g_1(Z, N_1^{(k)})$ in its second argument. Together with the assumptions on g_1, it is proved that the random variable converges to 0 in mean, which implies that it converges to 0 in probability. □

Corollary 4.6 implies that in the limit of small noise, the sources Z can be recovered exactly. Condition 1 upper bounds the influence of N_1 on the corruption: one cannot not hope to recover Z if $g_1(Z, N_1)$ contains too little signal. Condition 2 ensures that the function g_1 is invertible with respect to z when n_1 is equal to zero. If this were not satisfied, some information about Z would be washed out by g_1 even in the absence of noise, which would make recovery of Z trivially impossible. These conditions are satisfied, for example, by additive noise.

4.3.3 Multiple noisy views

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

This section considers the multi-view setting, where M distinct noisy views of Z are available,

$$X_i = f_i(g_i(Z, N_i)), \quad i = 1, \ldots, M\,,$$

and the noise variables N_i are mutually independent, as represented in Figure 4.4. Since each view provides additional information about Z, the question naturally arises: in the limit as $M \to \infty$, is it possible to reconstruct Z exactly?

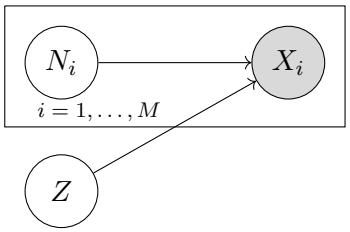


Figure 4.4 The setting of Section 4.3.3 with M corrupted views of the sources.

By applying Theorem 4.5 to the pair (X_1, X_i) it is possible to recover $(g_1(Z, N_1), g_i(Z, N_i))$ such that the components are aligned, but up to different component-wise invertible functions k_1 and k_i. Running the algorithm on a different pair (X_1, X_j) will result in recovery up to different component-wise invertible functions k_1' and k_j'.

Note that these will *not* necessarily result in $k_i \circ g_i(Z, N_i)$ and $k_j' \circ g_j(Z, N_j)$ being aligned with each other. However, the components of $k_1 \circ g_1(Z, N_1)$ and $k_1' \circ g_1(Z, N_1)$ 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 k_j' and make it aligned with k_i.

The problem is then how to combine the information from each aligned $k_i \circ g_i(Z, N_i)$ to more precisely identify Z. 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, consider the special case that each g_i acts additively, each N_i is zero mean and each of Z and the N_i are independent with independent components:

$$\left.\begin{aligned} X_i &= f_i(Z + N_i), \\ \mathbb{E}[N_i] &= 0, \end{aligned}\right\} \quad i \in \mathbb{N}. \tag{4.20}$$

Suppose to begin with that it is possible to recover each $Z + N_i$ *without* the usual component-wise invertible functions. Then, writing N to denote all of the N_i, it is possible to estimate Z as

$$Z \approx \Omega^M(Z, N) = \frac{1}{M} \sum_{i=1}^{M} (Z + N_i)\,.$$

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

In the presence of the unknown functions k_i, we would be able to reconstruct Z and the N_i if we were able to identify the inverses $e_i = k_i^{-1}$ for each i. For any component-wise invertible functions e_i, define

$$\Omega_e^M(Z, N) = \frac{1}{M} \sum_{i=1}^{M} e_i \circ k_i(Z + N_i),$$
$$R_{e,i}^M(Z, N) = e_i \circ k_i(Z + N_i) - \Omega_e^M(Z, N).$$

e_i is something we can choose and $k_i(Z + N_i) = h_i(X_i)$ is the output of the algorithm, and hence $\Omega_e^M(Z, N)$ and $R_{e,i}^M(Z, N)$ are random variables with known distributions. Subject to mild conditions, the dependence of these quantities on most or all of the N_i becomes increasingly small as M grows and disappears in the limit $M \to \infty$.

Lemma 4.7. *Suppose that the sequence* $\mathbb{E}_N[\Omega_e^M(Z, N)] = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{N_i}[e_i \circ k_i(Z + N_i)]$ *converges as* $M \to \infty$ *for almost all* Z*, and write this limit as*

$$\Omega_e(Z) = \lim_{M \to \infty} \mathbb{E}_N[\Omega_e^M(Z, N)].$$

Suppose further that there exists K *such that* $V_{e_i} = \mathrm{Var}\left(e_i \circ k_i(Z + N_i)\right) \leq K$ *for all* i*. Then*

$$\Omega_e^M(Z, N) \xrightarrow{a.s.} \Omega_e(Z),$$
$$R_{e,i}^M(Z, N) \xrightarrow{a.s.} R_{e,i}(Z, N_i) = e_i \circ k_i(Z + N_i) - \Omega_e(Z).$$

Proof sketch; see Appendix B.3.1 for full proof. The result follows by applying Kolmogorov's strong law to $\Omega_e^M(Z, N)$ for each value of Z. Kolmogorov's strong law states that the average of a sequence of independent, but not necessarily identically distributed, random variables converges to the expectation of the average, provided that the variances of the random variables do not grow too quickly. This is ensured by the assumption on the V_{e_i}. □

Given some choice of e, the quantities $\Omega_e(Z)$ and $R_{e,i}(Z, N_i)$ can be thought of as putative candidates for Z and N_i respectively. As discussed earlier, if it were possible to identify $e_i = k_i^{-1}$, then it would be the case that $\Omega_e(Z) = Z$ and $R_{e,i}(Z, N_i) = N_i$, and thus Ω_e and $R_{e,i}$ would satisfy the same independences and other statistical properties as Z and N_i respectively. Can these properties be used as criteria to identify good choices of e_i?

The following theorem provides sufficient conditions which, if satisfied by a putative choice for the e_i, implies that they invert the k_i up to some affine ambiguity for all i.

Theorem 4.8. *Suppose there exists $C > 0$ such that $Var(N_i) \leq C$ for all i and let $\mathcal{G}_K$ be the set of $\{e_i\}_{i=1}^M$ s.t.*

$$V_{e_i} \leq K \ \forall i, \tag{4.21}$$

$$\Omega_e(Z) < \infty \ \text{ for almost all } Z, \tag{4.22}$$

$$R_{e,i} \perp\!\!\!\perp R_{e,j} \ \forall i \neq j, \tag{4.23}$$

$$\mathbb{E}R_{e,i} = 0 \ \forall i, \tag{4.24}$$

$$R_{e,i}(Z, N_i) = R_{e,i}(N_i) \ \forall i. \tag{4.25}$$

Then,

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

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

Proof sketch; see Appendix B.3.2 for full proof. The fact that $\{k_i^{-1}\} \in \mathcal{G}_K$ can be shown using Lemma 4.7. The fact that $\mathcal{G}_K \subseteq \left\{ \{\alpha k_i^{-1} + \beta\} \ : \ \alpha \in \mathbb{R}^D_{\neq 0}, \ \beta \in \mathbb{R}^D \right\}$ is proved by showing that for any e_i such that $\{e_i\} \in \mathcal{G}_K$, the composition $e_i \circ k_i$ is affine. It follows that $e_i = A_i k_i^{-1} + \beta_i$ for some matrix A_i and vector β_i. Finally, it is shown that A_i and β_i are equal for all choices of i, and A is a diagonal matrix, and thus Ak_i^{-1} can be written as an elementwise product with a vector α. □

It follows that it is possible recover Z and N_i up to α and β via $\Omega_e(Z) = \alpha Z + \beta$ and $R_{e,i}(N_i) = \alpha N_i$.

Each of the conditions 4.21–4.24 can be verified from known information. We conjecture that condition 4.25 can be relaxed to assuming the verifiable condition of independence between $\Omega_e(Z)$ and $R_{e,i}(Z, N_i)$ for all i along with additional regularity assumptions on the functional form of $R_{e,i}$ (e.g. smoothness).

To conclude, Theorem 4.8 provides sufficient conditions under which it is possible to fully reconstruct Z 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 Discussion about assumptions

This section discusses in further detail the Sufficiently Distinct Views (SDV) assumption and the necessity for source-level noise for at least one of the views.

Typically, noise is a nuisance variable that would preferably not exist. In the setting considered here however, the presence of some source-level noise is necessary, since without this the classification based approach cannot be applied. Furthermore, the SDV assumption is ultimately an assumption about how the corrupted sources corresponding to each view are related, and is by implication an assumption about the source corruptions themselves.

4.4.1 The Sufficiently Distinct Views assumption

Recall that the SDV assumption is a demand on how much the conditional probability distribution of the source of one view given another varies, e.g. how much $p(s_1|s_2)$ changes as a function of s_2 where $s_i = f_i^{-1}(x_i)$. To provide intuition, this section gives examples of cases in which the SDV assumption does and does not hold.

The SDV assumption is closely related to the *Assumption of Variability* of Hyvärinen et al., 2019, an analogous assumption that occurs in the context a different graphical model from the multi-view setting considered here; see that paper for further details.

An example violating SDV

Suppose that the conditional distribution of one corrupted source given the other is Gaussian, so that

$$\log p(s_1|s_2) = -\sum_i (s_{1i} - s_{2i})^2/(2\sigma_i^2) + C\,, \tag{4.26}$$

where C is a constant. Since taking second derivatives of the log-probability with respect to s_i results in constants, there is no way to find $2D$ vectors t_j, $j = 1, \ldots, 2D$, such that the corresponding $w(s_1, t_j)$ in Definition 4.2 are linearly independent.

This rules out the case in which one or both views correspond to the source being corrupted by additive Gaussian noise. Note that this result is distinct from the non-identifiability result in the case of Gaussian sources for ICA. The problem here is not that the conditional distribution is rotationally invariant, but that the connection it implies between the two variables is 'too simple'. In fact, the identifiability results presented here do not demand that the marginal distribution over the uncorrupted source be non-Gaussian.

An example satisfying SDV

By choosing a conditional distribution that is more complex, the SDV assumption can be satisfied. Consider

$$\log p(s_1|s_2) = -\sum_i (s_{1i}^2 s_{2i}^2 + s_{1i}^4 s_{2i}^4) + C(s_2)\,, \tag{4.27}$$

where $C(s_2)$ is a normalisation constant that depends only on s_2. Proof that this conditional distribution satisfies the SDV assumption requires a few lines of computation: since this polynomial expression is of order strictly greater than 2, the second derivatives are not constant. $w(s_1, s_2)$ can be written as the product of a matrix and vector which are functions only of s_1 and s_2 respectively. The columns of this matrix are linearly independent for almost all values of s_1 and $2D$ linearly independent vectors can be realised by different choices of s_2, and hence the assumption is satisfied.

4.4.2 Source noise

Noise on the sources is required for at least one of the views. This is a consequence of training a classifier as a way retrieve the the unmixed signals. The reasons for this are explained briefly here.

Recall from the discussion on density ratio estimation in Section 2.5 that if a classifier is trained with the logistic loss to classify between samples from two distributions P (class 1) and Q (class 0), the optimal classifier should output $c(x) = \frac{p(x)}{p(x)+q(x)}$ as the estimated probability that a sample is drawn from P. When the classifier is parametrised as $c(x) = \frac{1}{1+\exp(-r(x))}$, the corresponding optimal regression function r is $r(x) = \log(p(x)/q(x))$.

In the setting considered here, P and Q are the joint distribution $p(x_1, x_2)$ and product of marginals $p(x_1)p(x_2)$ of the views. Thus, at optimality

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

If the variables x_1 and x_2 are deterministically related, this log-ratio is everywhere either 0 or ∞. 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 $2N$-dimensional 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.

It follows that the method of classification used in the results considered in this chapter can only be applied when the different views X_1 and X_2 are not deterministically related. For this technical reason, the corruptions are necessary.

Contrast with observation-level noise

A more typical noise model across the machine learning literature would involve noise at the level of the observations, e.g. $X = f(Z) + \epsilon$ where ϵ could be a Gaussian random variable. The reader should be clear that the source noise model used in this work is fundamentally different to this more typical observation noise model.

The assumption in the observational noise model is that the observed variables are not measured with perfect accuracy. In contrast, in the source noise model, the assumption is that the two views share closely related, but non-identical sources. This could hold, for example, when the two views correspond to the same source at slightly different times, or if the two views correspond to measurements of different subjects that have similar state, in which case the source noise may account for inter-subject variability. Nonetheless, Corollary 4.6 demonstrates that the results hold for even infinitesimally small amounts of noise, and therefore it is conceivable that practical methods for ICA based on the setting of this work may still function even if the source noise assumption is violated.

The source noise model is in some sense orthogonal to the observation noise model, in that it is possible for both to be used simultaneously. The absence of observation noise in this work is unrealistic in the sense that it requires observations to be made in a truly noise-free way, something that is clearly not possible in practical settings, though it is possible that the results presented in this chapter could be extended to this setting.

4.5 Conclusion

The main contribution of this chapter was to present identifiability results in a novel multi-view nonlinear ICA setting. These results are an important contribution to the field since they extend the scarce literature on identifiability results for nonlinear ICA models.

Theorems 4.1 and 4.4 state that in contrast to the single-view setting, the two-view setting is identifiable when the mixing functions are arbitrary smooth, invertible nonlinear functions with smooth inverse, provided that one of the views is corrupted at the source level by sufficiently 'complex' noise such that the Sufficiently Distinct Views assumption is satisfied.

Identifiability results are also obtained in the case of corruptions on both views. Theorem 4.5 states that the sources can be recovered up to the corruptions, and Corollary 4.6 demonstrates that in the limit as one of the corruptions becomes small, the uncorrupted sources can be recovered.

Finally, initial results are presented in Theorem 4.8 providing conditions under which the uncorrupted sources are identifiable when a large number of views are available, even if these views are all corrupted by source noise.

The multi-view setting is relevant in a number of 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 as intrinsic sources of variability specific to each view, rather than as noise per se. In most practical applications this would probably not be a significant limitation due to the prevalence of stochasticity in real-world systems.

A specific example application of the work presented here can be found in the field of neuroimaging. Consider a study involving a cohort of subjects whose response to the presentation of the same stimulus is measured. One of the key problems in this 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 (Haxby et al., 2011; Chen et al., 2015). The results presented here provide principled approaches to the extraction and decomposition of the components of the shared response, by considering the measurements to be different views of an underlying shared response that is corrupted by inter-subject variability.

There are further directions to explore. Observe that Theorem 4.8 builds on the setting of Theorem 4.5, which only makes use of pairwise information from the observations. A natural extension of this work would be to investigate algorithms that explicitly make use of $N > 2$ views, which may allow relaxation of the additivity assumption on the corruptions. Furthermore, Theorem 4.8 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. In the direction of application, the results here prove that recovery of the latent sources is possible in a multi-view setting, but there may be many ways to actually perform this recovery in practice. The development of algorithms exploiting this setting are a natural direction to explore.

Chapter 5

Causal Modelling

This chapter introduces the notion of exact transformations *between Structural Equation Models (SEMs). This provides a framework to evaluate whether two SEMs are consistent with one another as causal models, meaning that a correspondence can be established between them such that reasoning about the effects of interventions in both models agree. In particular, this framework can be used to understand whether two models of the same system at different levels of detail are consistent, and whether measured variables derived from a lower-level causal model admit interpretation as causal variables themselves. This work has broad implications to the causal modelling process, as there is often a mismatch between the level at which measurements are made and the level at which the underlying 'true' causal structure exists, yet causal inference algorithms generally seek to discover causal structure at the level of measurements.*

The main technical content of this chapter has been published in paper:

> Paul K Rubenstein*, Sebastian Weichwald*, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup and Bernhard Schölkopf. "Causal consistency of structural equation models". *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI).* *Joint first authorship. 2017.

5.1 Introduction

Much of machine learning concerns the statistical relationships between random variables. In this context, the word *statistical* refers to the assumptions that a fixed but unknown probability distribution exists from which observed data are sampled i.i.d., and that new data at 'test time' will similarly be drawn i.i.d. from this distribution. Classification is the

canonical example of this, where given a set of i.i.d. samples from a joint distribution P_{XY} over input X and discrete target Y, the goal is to learn the conditional distribution $P_{Y|X}$ giving the probability distribution over targets for each possible input. Other problems such as density estimation can be phrased similarly.

Despite the great empirical successes of machine learning in practical and applied settings in recent years, there remain problems of interest that cannot be cast directly into the framework described above. The framework is limited in that it presupposes the existence of a single fixed joint distribution over all of the random variables of interest, with the operations of marginalisation and conditioning then providing the relationships connecting any subset of variables. But there are many examples of problems for which a single fixed joint distribution over all variables does not suffice. For many questions of scientific interest, this is because the problem either implicitly or explicitly concerns an *intervention* or *action* in the world that changes the joint distribution over the observable variables.

For example, we may be interested to understand the influence of diet on longevity, with the aim of improving public health by encouraging people to eat healthily. One might find the consumption of expensive imported fruits to be correlated with a longer life. This may well be due to the nutritiousness of such fruits; it could equally well be due to the fact that only wealthy people can afford such a diet, and that such wealth entails better access to medical treatment, sports facilities for exercise and so on. In the former case, intervening in the world by reducing tariffs on imported fruits to make them cheaper and thus encourage their consumption would have a positive effect on public health; in the latter, not. Similarly, we may observe in the population that taking over-the-counter painkillers is associated with an elevated risk of heart disease. This might be because such painkillers have a negative effect on the cardiovascular system, in which case acting to reduce access to such painkillers might have a positive impact on health outcomes. But if instead the association is because people who have poor health, and thus heightened risk of heart disease, tend to take more painkillers, then such a policy might have little effect other than to increase overall suffering.

These questions are concerned with understanding causal, not statistical, relationships in the world. The aim of *causality* is to study causal influence through the lens of a formal mathematical language, in much the same way that statistical machine learning uses the language of probability. *Causal inference* or *causal discovery*, a large part of the causality literature, concerns the identification of causal relationships using data. As the examples above demonstrate, this can be highly non-trivial, with the well-known phrase "correlation does not imply causation" standing testament to the simultaneous difficulty and ubiquity of this problem. Since correlation (or statistical dependence more generally) is a symmetric relation, an asymmetric causal relationship between two variables can never be inferred without other prior knowledge or assumptions. Moreover, simply identifying that two quantities tend to

co-occur does not itself imply a causal relation between the two, since both could be causally influenced by a third.

While causal inference is an important problem with a wide variety of applications ranging from astronomy to neuroscience and economics (Schölkopf et al., 2016; Ding et al., 2006; Hicks et al., 1980), the main contribution of this chapter is to extend and provide greater understanding of *Structural Equation Models (SEMs)*, one of the popular mathematical frameworks for formalising causal relationships between random variables and interventions, along with the variety of probability distributions these entail.

In particular, this work seeks to understand the implications of modelling causal structure at a different level of *abstraction* compared to the 'truth'. For instance, causal influence between variables of interest may be mediated by irrelevant variables that are ignored; interactions between low-level variables may instead be modelled at a macroscopic level, similar to the manner in which temperature and pressure arise as macroscopic properties of a large number of gaseous particles; and though time invariably plays a role in any causal influence in the real world, mathematical models of causal structure may often omit explicit reference to it.

This chapter is structured as follows. Sections 5.2 and 5.3 are an overview of the causality literature, providing context for this chapter. Section 5.4 discusses issues involved in defining causal variables in practice, as well as modelling the same system at different levels of details, followed by Section 5.5 which extends the definition of SEMs and provides a framework to analyse when two causal models at different levels of detail are consistent with one another. Section 5.6 provides examples of consistent models in a variety of settings. Section 5.7 discusses the implications of this work to the literature, as well as papers that have built on this since its original publication.

5.2 Structural Equation Models: A Language for Causality

This section introduces Structural Equations Models (SEMs), a mathematical formalism used to model causal influence. In Section 5.5, an extension to this definition will be presented. We will avoid going into measure-theoretic detail, and point the reader to Bongers et al., 2016 for a more rigorous measure theoretic introduction to SEMs.

An SEM over a tuple of random variables $X = (X_1, \dots, X_N)$ consists of equations so that each X_i is written as a function of a subset of the other X_j and an *exogenous noise variable* E_i. More formally:

Definition 5.1 (Structural Equation Model (SEM))**.** *Let $X = (X_1, \dots, X_N)$ and $E = (E_1, \dots E_N)$ with each X_i, E_i taking value in $\mathbb{R}$. An SEM $\mathcal{M}_X$ over X is a tuple $(\mathcal{S}_X, P_E)$ where*

- *$\mathcal{S}_X$ is a set of structural equations of the form $X_i = f_i(X_{\mathrm{pa}(i)}, E_i)$ for $i = 1, \ldots, N$, where* $\mathrm{pa}(i) \subset \{1, \ldots, N\}$ *and $X_{\mathrm{pa}(i)}$ is the corresponding subset of the variables X.*
- *The variables $E = (E_1, \ldots, E_N)$ have distribution P_E which factorises, i.e. the E_i are independent.*
- *The* causal graph *$\mathcal{G}$ is acyclic, where $\mathcal{G}$ is the directed graph with nodes X_i and edges $X_i \to X_j$ if and only if $i \in \mathrm{pa}(j)$.*

The requirement that $\mathcal{G}$ be acyclic ensures that the SEM implies a well-defined distribution P_X over the variables X. This is known as the *observational distribution.*

Lemma 5.2 (Well defined observational distribution)**.** *An SEM implies a well-defined observational distribution P_X over X.*

Proof. For any particular value e of the noise variables E, there is a unique vector x_e so that (x_e, e) solves the structural equations $\mathcal{S}_X$. To see this, observe that acyclicity of $\mathcal{G}$ means that the structural equations can be solved recursively, beginning with variables with no parents. It follows that each x_i can be written as a function of $e_{\mathrm{anc}(i)}$, where $\mathrm{anc}(i)$ are the indices of the *ancestors* of X_i in $\mathcal{G}$, that is the X_j for which there exists a path $X_j \to X_i$ using the edges in $\mathcal{G}$. Denote by $g(e) = x_e$ the function mapping from values of e to the unique solution x_e. Then g in combination with P_E induces the push-forward distribution $P_X := g_{\#} P_E$ over the X-variables, i.e. P_X is the distribution of the random variable $g(E)$, a function of the random variables E. □

As discussed in the introduction, an important part of causal relationships is a notion of behaviour under *interventions.* SEMs are equipped with a formal notion of such interventions which are termed *perfect interventions.* Intervening on a variable makes a change to the structural equation determining its value in the observational setting. Although there may be an effect on the distributions over variables downstream of the intervened variable, the structural equations determining those variables are unchanged. Such an intervention is realised by altering the structural equation of the intervened variable. This notion is easily generalised to interventions on multiple variables by replacing all of the corresponding equations.

Definition 5.3 (Perfect interventions and the do-operator)**.** *Let $\mathcal{M}_X$ be a SEM, $i \in \{1, \ldots, N\}$ and $x_i \in \mathbb{R}$. The perfect intervention setting X_i to take value x_i is denoted* $\mathrm{do}(X_i = x_i)$ *and is implemented by replacing the ith equation with $X_i = x_i$. The resulting set of structural equations is denoted $\mathcal{S}_X^{\mathrm{do}(X_i = x_i)}$ and the resulting SEM $\mathcal{M}_X^{\mathrm{do}(X_i = x_i)} = (\mathcal{S}_X^{\mathrm{do}(X_i = x_i)}, P_E)$. Perfect interventions over two or more variables are also valid, which are denoted for instance as* $\mathrm{do}(X_i = x_i, X_k = x_k)$.

We note in passing that the notion of a perfect intervention may also be extended to *imperfect* or *stochastic interventions* in which the intervened variable is set equal to some random variable rather than a constant. Formally, this is no different from the case of perfect interventions other than needing to additionally introduce distributions over the new random variables.

An intervened SEM is still just a SEM, since it has structural equations and a distribution over exogenous variables. The only difference is that the functions corresponding to an intervened variable $X_i = f_i(X_{\mathrm{pa}(i)}, E_i)$ will have $\mathrm{pa}(i) = \varnothing$ and f_i will be a constant function. Moreover, the causal graph $\mathcal{G}_{\mathrm{do}(\cdot)}$ has the same nodes but a *subset* of edges compared to $\mathcal{G}$, and thus inherits acyclicity. This implies the following lemma.

Lemma 5.4 (Well defined interventional distribution for any perfect intervention)**.** *Any perfect intervention* $\mathrm{do}(\cdot)$ *on a SEM (i.e. any subset of variables set to any particular values) implies a well-defined interventional distribution* $P_X^{\mathrm{do}(\cdot)}$ *over* X*.*

Proof. As discussed in the remark above, $\mathcal{M}_X^{\mathrm{do}(\cdot)}$ is a valid SEM. Thus, Lemma 5.2 applies to $\mathcal{M}_X^{\mathrm{do}(\cdot)}$ and so it has a well-defined observational distribution. The interventional distribution $P_X^{\mathrm{do}(\cdot)}$ of $\mathcal{M}_X$ is equal to the observational distribution of $\mathcal{M}_X^{\mathrm{do}(\cdot)}$. □

SEMs can thus be thought of as a way to model not just a single distribution over the variables of interest, but an entire family of related distributions, one for each possible perfect intervention. This is illustrated in the following simple example.

Example 5.5. *Consider the SEM* $\mathcal{M}_X = \{\mathcal{S}_X, P_E\}$ *where*

$$\mathcal{S}_X = \{X_1 = E_1,\ X_2 = X_1 + E_2\},$$
$$P_E = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right).$$

Then the observational distribution is given by

$$P_X = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}\right),$$

while the interventional distributions corresponding to $\mathrm{do}(X_1 = x_1)$ *and* $\mathrm{do}(X_2 = x_2)$ *are given by the degenerate Gaussians*

$$P_X^{\mathrm{do}(X_1=x_1)} = \mathcal{N}\left(\begin{pmatrix}x_1\\x_1\end{pmatrix}, \begin{pmatrix}0 & 0\\0 & 1\end{pmatrix}\right),$$
$$P_X^{\mathrm{do}(X_2=x_2)} = \mathcal{N}\left(\begin{pmatrix}0\\x_2\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}\right).$$

5.2.1 Connections to Bayesian networks

SEMs are closely related to Bayesian networks, a class of graphical models. There is a long debate between certain members of the causality community and the Bayesian and statistics communities as to whether there is in fact any difference between these two classes of models. Discussion of this is somewhat tangential to the main contribution of this chapter, but for completeness it is nonetheless important to give a brief outline of this debate. In the following, I attempt to give a neutral overview of the similarities and differences between them, followed by my personal opinion.

Definition 5.6. *A Bayesian network over variables $X = (X_1, \dots, X_N)$ with directed acyclic graph $\mathcal{G}$ specifies a joint distribution over X as a product of conditional distributions*

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i | X_{\mathrm{pa}(i)}),$$

where the nodes of $\mathcal{G}$ correspond to the variables X and there is an edge $X_j \to X_i$ in $\mathcal{G}$ if and only if $j \in \mathrm{pa}(i)$.

Since any product distribution can always be decomposed as $p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i | X_{j<i})$, it is the *absence* of edges in $\mathcal{G}$ that imposes structure on the probability distribution.

Bayesian networks can be endowed with a similar notion of perfect intervention as SEMs. To model the effect of the perfect intervention $\mathrm{do}(X_i = x_i)$, in the resulting joint distribution the factor $p(X_i | X_{\mathrm{pa}(i)})$ is replaced with a Dirac delta distribution $\delta_{X_i = x_i}$. Bayesian networks equipped with such notions are sometimes referred to as *causal* Bayesian networks. In the following, we will simply refer to them as Bayesian networks.

Note that it is also possible instead to introduce additional 'treatment' or 'indicator' variables representing the application of interventions. This may in general replace the factor $p(X_i | X_{\mathrm{pa}(i)})$ with a Dirac delta distribution or indeed any other distribution. A theory of such models, termed *decision theoretic statistical causality* is developed in Dawid, 2020 and the references therein.

Correspondence between SEMs and Bayesian networks

SEMs and Bayesian networks exist in correspondence with one another. For any SEM, there exists a Bayesian network over the same variables that induces the same observational and interventional distributions, and vice versa. We briefly explain this correspondence.

Any SEM induces a Bayesian network with the same graph $\mathcal{G}$. To see this, observe that for any fixed value of $X_{\mathrm{pa}(i)}$, the equation $X_i = f(X_{\mathrm{pa}(i)}, E_i)$ in combination with the distribution

over E_i induces a distribution over X_i which corresponds to $p(X_i|X_{\text{pa}(i)})$. The observational distributions of the SEM and Bayesian network are then equal. Further, since the equation $X_i = x_i$ associated with the intervention $\text{do}(X_i = x_i)$ corresponds to the distribution $\delta_{X_i=x_i}$, and similar for interventions or arbitrary subsets of variables, the interventional distributions also agree.

Showing that any Bayesian network induces a SEM is somewhat more technical, and is outlined here based on the proof of Proposition 7.1 from Peters et al., 2017. For each X_i with parents $X_{\text{pa}(i)}$, define the function

$$F_{X_i|X_{\text{pa}(i)}} : \mathbb{R} \to [0, 1]$$

to be the cumulative distribution function of $p(X_i|X_{\text{pa}(i)})$ as given by the Bayesian network. Define the exogenous variables of the SEM $E_1, \ldots, E_N$ to be uniformly distributed on the interval $[0, 1]$, and let the structural equations be defined by

$$X = f_i(X_{\text{pa}(i)}, E_i) := F^{-1}_{X_i|X_{\text{pa}(i)}}(E_i).$$

To see that this SEM induces the same observational distribution over the variables X as the Bayesian network, it suffices to note that the conditional distributions $X_i|X_{\text{pa}(i)}$ are the same for both the Bayesian network and the SEM, by definition of the cumulative distribution function. Similarly, replacing any equation with $X_i = x_i$ corresponds to the cumulative distribution function associated to $\delta_{X_i=x_i}$ and so any interventional distributions also agree.

Differences between SEMs and Bayesian networks

In an SEM the noise variables E are explicitly modelled, while in a Bayesian network they are generally not. Pearl, 2009 considers this to correspond to a 'quasi-deterministic' view of the world in which any observed randomness is a consequence of a lack of knowledge. In contrast, modelling with implicit noise in a Bayesian network corresponds to a view that the world is inherently stochastic.

As a consequence of the noise variables being explicitly modelled, SEMs are equipped with *counterfactual reasoning*. That is, once the variables X have been observed, the values for the exogenous variables E can generally be inferred. This means that one can answer questions such as "what would have happened had the intervention $\text{do}(X_i = x_i)$ been performed?" Whether or not this is useful in practice is debatable, since different SEMs may imply the same set of observational and interventional distributions, but nonetheless be counterfactually non-equivalent. That is, they may produce different answers to the same counterfactual

question, despite implying the same observational and interventional distributions. This means that such models cannot be distinguished based on data, be that from observational or interventional settings. Therefore one cannot hope to be able to answer counterfactual questions based on data without prior knowledge to distinguish between counterfactually non-equivalent models.

It can be easier to express assumptions on the mechanisms of causal influence within the SEM framework, for instance by assuming that the distribution P_E and the functions f_i belong to some restricted sets, though this is also possible in the Bayesian network setting. The main consequential difference between SEMs and Bayesian networks is that it is conceptually simple to extend the SEM framework to express cyclic dependencies as a result of the explicit modelling of the noise variables.

Personal opinion

Acyclic SEMs and Bayesian networks are fundamentally equivalent as mathematical models. By convention, Bayesian networks usually do not explicitly specify the noise variables, though there is no reason why they cannot do so. Indeed, the *reparameterisation trick*, also known as a *disturbance* representation, involves explicit use of noise variables rather than an implicit representation.

This equivalence breaks down in the case of cycles: while SEMs can be easily generalised to allow cyclic causal graphs, it is not clear how to do so for Bayesian networks unless the noise variables are explicitly specified, in which case they are essentially the same as SEMs. That said, most of the debate around SEMs and Bayesian networks has considered the acyclic case. The substance of this debate seems to revolve largely around differences of philosophy or culture, rather than formal mathematics. This can be summarised in two key differences.

First, the causality community thinks of the structural equations as physical mechanisms representing some fundamental, invariant truth about how the universe operates. The Bayesian and statistics communities are more accustomed to thinking of them as abstract descriptions that are not necessarily grounded in physical reality to the same degree.

Second, the causality community places a strong emphasis on the role of interventions or 'action'. As such, to do-operator is considered as a central feature of SEMs, while the statistics and Bayesian communities often see the do-operator as a feature that has been post-hoc tacked onto Bayesian networks.

In summary, my personal belief is that Bayesian networks and SEMs are essentialy equivalent, with some subtleties in the cyclic case. The fact that debate on this topic continues to exist despite the formal relation between the two model classes being fairly clear and well understood

is, in my opinion, caused largely by cultural differences leading to misunderstandings as well as academic politics.

5.2.2 Cyclic Structural Equation Models

Most causal systems in the real world involve some degree of feedback. Examples can be found in a wide variety of settings: molecular biology (e.g. gene-gene or gene-protein interactions in a cell), ecology (e.g. population dynamics), climate science (e.g. methane release from thawing permafrost) and public policy and economics (e.g. poverty traps). Studying the mathematical modelling of these cases is interesting in part because their treatment requires consideration of issues that are not present in the acyclic, feedback-free case.

In reality, any cyclic causal system will correspond to an acyclic system when 'unravelled' in time. Nonetheless, explicitly modelling systems as cyclic may be useful in scenarios in which it is not possible to make measurements or observations at the level of the acyclic structure. For instance, in many biological settings, it may not be possible to generate temporal data if making a measurement involves destroying the system being measured (e.g. sequencing the mRNA in a cell).

Recall that Lemmas 5.2 and 5.4 relied on acyclicity of the causal graph $\mathcal{G}$ to prove that an SEM induces well-defined observational and interventional distributions over the variables X. When generalising to *cyclic SEMs* by relaxing this acyclicity constraint, one must be careful to understand under which conditions the observational and interventional distributions are well-defined. The implied observational distribution is well-defined if and only if there is a unique solution $X(E)$ to the structural equations for P_E-almost all values of E. Similarly, an interventional distribution is well-defined if and only if the intervened structural equations have unique solution P_E-almost surely.

That is, one generates from a cyclic SEM in exactly the same way that one generates from an acyclic one. First, sample the values for the exogenous noise variables. Then, solve the structural equations to find the (unique) values of X satisfying the equations. This value of X is the generated value. Therefore, a cyclic SEM only implies well-defined observational and interventional distributions if, with probability 1 over the distribution of the exogenous noise variables E, a unique solution $X(E)$ exists to the (possibly intervened) structural equations.

The following example shows a simple case of an SEM with well-defined observational distribution for which some interventional distributions are well-defined but others are ill-defined.

Example 5.7. *Consider the cyclic SEM* $\mathcal{M}_X = \{\mathcal{S}_X, P_E\}$ *where*

$$\begin{aligned}\mathcal{S}_X = \{&X_1 = X_2 + X_3 + E_1,\\ &X_2 = X_1 + X_3 + E_2,\\ &X_3 = X_1 + X_2 + E_3\}\end{aligned}$$

and P_E *is any distribution. In the observational setting, the values*

$$(X_1, X_2, X_3) = \left(\frac{E_2 + E_3}{2}, \frac{E_1 + E_3}{2}, \frac{E_1 + E_2}{2}\right)$$

uniquely solve the structural equations, and hence the observational distribution P_X *is well-defined.*

Under the intervention $\mathrm{do}(X_1 = x_1)$*, the equations have either infinitely many solution, in the case that* $E_Y + E_Z = -2x_1$*, or zero solutions, if* $E_Y + E_Z \neq -2x_1$*. Hence the interventional distribution* $P_X^{\mathrm{do}(X_1=x_1)}$ *is ill-defined, and similarly for other interventions on a single variable.*

Under the intervention $\mathrm{do}(X_1 = x_1, X_2 = x_2)$*, the equations have unique solution*

$$(X_1, X_2, X_3) = (x_1, x_2, x_1 + x_2 + E_3)\,.$$

Thus the interventional distribution $P_X^{\mathrm{do}(X_1=x_1, X_2=x_2)}$ *is well defined, similarly for any intervention on two variables.*

The literature has not settled on a set of criteria for determining which cyclic SEMs should be considered 'valid'. All agree on the fact that observational distributions must be well-defined, but there is significant disagreement on interventional distributions. Notable works include Hyttinen et al., 2010, which requires well-defined distributions after any intervention, and Mooij et al., 2011, which state that for any cyclic SEM, any intervention leading to a well-defined interventional distribution may be given a causal interpretation. Other works in this area such as Lacerda et al., 2008 avoid discussion of this issue.

It will be argued in Sections 5.4 and 5.5 that it should be considered an integral part of the modelling process to choose a set of interventions being modelled. As such, it should be guaranteed that interventional distributions be well-defined for those that are part of the modelled intervention set; the behaviour outside of this set being considered outside of the modelled universe and thus irrelevant.

5.3 Methods of Causal Inference

The problem that is typically of greatest interest to practitioners in the context of causality is that of *causal inference*, sometimes referred to as *causal discovery*. The goal of a causal inference algorithm is to learn the causal graph $\mathcal{G}$, either with only observational data, or with a mixture of observational and interventional data. In the case of no latent confounders[1], once $\mathcal{G}$ has been identified, learning the functional relationships between parents and children reduces to solving independent regression problems. Although the main contribution of this chapter is to improve theoretical understanding and extend the mathematical framework of SEMs, some discussion of approaches to causal inference will provide important context. All methods focus on the acyclic case unless otherwise stated.

The problem is usually formalised thus: Given i.i.d. draws from a distribution P_X induced by an SEM with causal graph $\mathcal{G}$, estimate $\mathcal{G}$. Broadly speaking, there are two main categories of approaches: those which exploit a correspondence between statistical properties of P_X and properties of $\mathcal{G}$; and those that make additional assumptions on the noise variable distribution P_E and structural equation functions f_i. These are briefly outlined next.

5.3.1 Conditional independence based methods

The key idea of this family of approaches is to relate *conditional independences* of the joint distribution P_X to graphical properties of the (acyclic) causal graph $\mathcal{G}$ known as *d-separation*. This line of work is historically important, but only tangentially related to the topic of this chapter. As such, only a brief outline follows; see Pearl, 2009 or Peters et al., 2017 for an in-depth explanation.

In a directed acyclic graph, d-separation is a relation between triples of disjoint subsets of nodes. Informally, if a subset of nodes A d-separates subsets B and C, then A blocks any information flow between B and C. Graphs exhibiting the same set of d-separations form an equivalence relation, the classes of which are known as *Markov equivalence classes*. Subject to mild assumptions, there is a one-one correspondence between possible sets of conditional independences present in P_X, and Markov equivalence classes of $\mathcal{G}$. It is thus possible to identify $\mathcal{G}$ up to its Markov equivalence class by inferring the conditional independences present in P_X.

In practice, one may have access only to a finite number of samples from P_X, and so conditional independence tests must be performed to identify the list of conditional independences which are subsequently used to infer the Markov equivalence class of $\mathcal{G}$.

[1] A confounder is an unobserved variable that causally influences two or more observed variables, leading to those observed variables exhibiting statistical dependence without directly influencing one another.

Methods for causal inference using conditional independences have the advantage that they make few assumptions on the underlying distribution. This comes at the cost of needing to perform conditional independence tests, a hard problem in the general case (Shah and Peters, 2018), as well only being able to identify $\mathcal{G}$ up to an equivalence class of graphs. In particular, it is impossible to distinguish between the two models $X \to Y$ and $Y \to X$ using only conditional independences since both exhibit the same (trivial) set of conditional independences. The fact that even this simple case cannot be solved using conditional independences has led to other approaches in which further assumptions are made.

5.3.2 Structural Equation based methods

One way in which further assumptions can be specified is to use the language of SEMs. Placing suitable restrictions on the functional forms of the structural equations or the distribution over the exogenous variables can lead to identifiability results: unambiguous recovery of the SEM from the observational distribution, in contrast to recovery only up to an equivalence class. Of course, such identifiability results usually hold in the limit of infinite data, and so practical performance in realistic cases will differ even between methods for which identifiability holds under the same assumptions. In the following, some methods for inferring SEMs and the assumptions they require for identifiability are outlined. The main restricted model class that has been studied are *additive noise models*.

Definition 5.8. *An SEM is an* additive noise model *if the structural equations are deterministic functions of the parent variables with additive noise,*

$$X_i = f_i(X_{\mathrm{pa}(i)}, E_i) = g_i(X_{\mathrm{pa}(i)}) + E_i, \tag{5.1}$$

for some functions g_i.

Linear additive noise models

If the functions g_i in Equation 5.1 are linear, the model can be written as

$$X = AX + E, \tag{5.2}$$

where A is a strictly upper triangular matrix[2] for an acyclic SEM. The distribution over E induces a density on the observable variables X via

$$X = (I - A)^{-1}E,$$

[2]The diagonal and anything below or left of it is 0.

where I is the identity matrix. This is an instance of linear Independent Component Analysis (ICA), discussed in detail in Chapter 4. Recall that identification of A and the distribution of E is possible if at most one of the components of E has Gaussian distribution. The Linear Non-Gaussian Acyclic causal Model method (LiNGAM) of Shimizu et al., 2006 uses ICA to recover the both the matrix $I - A$ and the distribution over E from only observational data. Peters and Bühlmann, 2013 instead consider linear additive noise models under the assumption that all noise variables are independent Gaussians with equal (unknown) variance and prove identifiability from observational data in this case.

Nonlinear additive noise models

If the g_i in Equation 5.1 are allowed to be nonlinear, the possible model complexity grows enormously compared to the linear case. As such, much of the literature on *nonlinear* additive noise models has considered the bivariate case, as this is the simplest non-trivial problem. Causal inference in this case is often referred to as the problem of *distinguishing between cause and effect*.

A common approach to this case is to assume that the g_i belong to some restricted class of functions and that the noise variables satisfy some statistical properties such as independence. For example, Hoyer et al., 2009; Peters et al., 2010; Mooij et al., 2010 and Peters et al., 2014 perform regression under both $X \longrightarrow Y$ and $Y \longrightarrow X$ and choose between these using one of a variety of independence scores to test that the 'regressor' or 'parent' variable is independent of the residuals. Zhang and Hyvärinen, 2008 and Zhang and Hyvärinen, 2009 extend the additive noise model by considering the *post-nonlinear (PNL)* model where $x = f_2(f_1(y + e))$ and f_2 is invertible.

Other approaches include using information theory (Janzing et al., 2012; Janzing and Schölkopf, 2010; Janzing et al., 2009a) and treating the problem as one of binary classification between $X \longrightarrow Y$ and $Y \longrightarrow X$, for which a classifier is trained using artificial data (Lopez-Paz et al., 2015). This problem is more challenging in the presence of confounders, corresponding to the noise variables being dependent, see Hoyer et al., 2008; Janzing et al., 2009b for more details.

Cyclic additive noise models

Linear cyclic additive noise models, corresponding to Equation 5.2 in which the matrix A need not be strictly upper diagonal, may be learned using ICA methods provided that $I - A$ is invertible. This is done by Lacerda et al., 2008, generalising the LiNGAM method, though identifiability holds here only up to SEMs that induce the same observational distribution.

(a) (b)

Figure 5.1 Effects of cholesterol on risk of heart disease. As illustrated by (a), the current consensus is that low-density lipoprotein (LDL) has a negative effect on heart disease (HD), while high-density lipoprotein (HDL) has a positive effect on heart disease. Considering total blood cholesterol (TC = LDL + HDL) to be a causal variable as in (b) leads to problems: two diets promoting raised LDL levels and raised HDL levels respectively have the same effect on TC but opposite effects on heart disease. Hence different studies may come to contradictory conclusions about the effect of TC on heart disease.

Scheines et al., 2010; Hyttinen et al., 2010; Hyttinen et al., 2012 and Hyttinen et al., 2013 extend this setting to consider dependent noise variables, additionally assuming access to interventional data.

Another line of research tries to relax the linearity assumption. As is the case for acyclic SEMs, removing this assumption leads to significant increase in possible model complexity, with the challenge of causal inference growing correspondingly. Mooij et al., 2011 and Mooij and Heskes, 2013 consider Gaussian distributed noise variables and use Gaussian processes to model the functions g_i.

5.4 What are causal variables?

The approaches to causal inference discussed in the previous section all make a crucial assumption that we have not yet discussed: we are presented with a vector of random variables which are individually 'causally meaningful' in the sense that causal relations between them exist and can thus be discovered. Clearly not all random variables are meaningful in this way, and thus cannot be endowed with a causal interpretation. A simple intuitive example of this can be found in images, where individual pixels are not meaningful, though higher level features such as the presence of an object may be.

It will be argued that for macro-variables that are functions of underlying causal micro-variables to be themselves considered causal entities, it is important to consider the set of interventions being modelled and the structure exhibited by these interventions.

The issue is best illustrated concretely by an example previously used by Spirtes and Scheines, 2004 to demonstrate problems in the causal modelling process.

Historically, the level of total blood cholesterol (TC) in a human subject was thought to be an important variable in determining their risk of developing heart disease (HD). To investigate

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

The reason for this apparent contradiction is understood with hindsight, but serves to illustrate the care that must be taken when seeking causal relations. The current scientific consensus is that there are two types of blood cholesterol, low-density lipoprotein (LDL) and high-density lipoprotein (HDL), which have a negative and positive effect on HD respectively (Figure 5.1a). A measurement of TC is in fact a measurement of the sum of LDL and HDL. Therefore two experiments, one raising LDL levels and the other raising HDL levels, would have the same effect on TC but opposite effects on HD (Figure 5.1b).

In this example, total blood cholesterol is too 'coarse' a variable to have a well-defined causal relation with risk of heart disease. However, if it had been possible to affect only one of LDL and HDL through diet, this issue may never have been discovered: if only HDL could be influenced by diet, the scientific consensus would be that total blood cholesterol is protective against heart disease and no contradictions would have been found.

Similarly, it is conceivable that there are in fact two different types of HDL: a more prevalent form which is protective against heart disease and one present in smaller quantities that has a detrimental impact. If the ratio of these two types is constant under any intervention through diet, the negative impact will always be outweighed by the positive impact and so we might never discover the detrimental subtype. If this were true, would the statement 'increased HDL levels cause reduced risk of heart disease' be rendered false? Arguably it would be an oversimplification of the more complicated truth, but it would not be false in that it would correctly predict the outcome of any diet-based intervention.

This example raises two main points. The first is that in the real world, measurements or observations are always made at some level of detail or coarseness that is somewhat arbitrary. The second is that whether or not a variable is 'causally meaningful' is intricately connected to the interventions being considered. If we consider a coarse variable such as TC, we can consistently model causal relations if the interventions considered are sufficiently restricted, but this breaks down if the interventions are too rich.

We elaborate on each of these points next, setting the stage to tackle our overarching goal of trying to answer the following questions: If the true causal mechanisms of the world operate at a very low level of detail (e.g. atoms), under what conditions can we speak of causal relations at higher levels (e.g. objects)? How can we formalise such a notion of consistent modelling at different levels of abstraction in the framework of SEMs?

5.4.1 Modelling at different levels of detail

All physical systems or processes in the real world are complex and can be understood at various levels of detail. We previously discussed the example of cholesterol levels and risk of heart disease. Although the true mechanisms by which cholesterol affects heart disease are surely very complicated – for instance, it may be important how cholesterol is distributed throughout the body – we sought to summarise the micro-level details into a small number of macro-level variables, the total levels of HDL and LDL.

Another example of micro-macro abstraction can be found in statistical physics. A gas in a volume consists of a large number of molecules, but instead of modelling the motions of each particle individually, we may choose to consider macroscopic properties of their motions such as temperature and pressure. As in the case of cholesterol, the decision to use such macroscopic properties may be necessitated primarily by practical considerations. Indeed, for all but extremely simple cases, making a measurement of all the individual molecules is practically impossible and computational 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 is a rigorous theory that 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 other cases where aggregate or indirect measurements of a complex system form the basis of a macroscopic description of the system (such as the cholesterol example) there is little theory to explain whether this is justified or how the micro- and macro-descriptions stand in relation to one another. As we saw, this lack of theory occasionally leads to apparent contradictions.

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 (Simon and Ando, 1961; Iwasaki and Simon, 1994; Hoel et al., 2013; Chalupka et al., 2015; Chalupka et al., 2016); 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 (Fisher, 1970; Iwasaki and Simon, 1994; Dash and Druzdzel, 2001; Lacerda et al., 2008; Mooij et al., 2013; Mooij and Heskes, 2013); e. g. modelling only the final ratios of reactants and products of a time evolving chemical reaction.

In each of these cases, the fine-grained model may be considered the 'truth' while the coarse-grained model is a convenient abstraction.[3] In the context of causal modelling, an intervention in the coarse-grained model must correspond in reality to some intervention in the fine-grained model. Such differing model levels should be consistent with one another in the sense that they agree in their predictions of the effects of corresponding interventions.

5.4.2 The importance of interventions

Recall the two models in Figure 5.1 in the cholesterol and heart disease example. Here, the micro-level model with variables HDL and LDL is inconsistent with the macro-level model with only the TC variable. The inconsistency arose because two different interventions in the mirco-level (raise HDL and raise LDL) have different effects on risk of heart disease yet map to the same intervention at the micro-level (raise TC). This would not have been noticed had it been possible to intervene only on one of HDL and LDL through dietary means, and in this case the TC model would have been considered valid. In other words, whether or not the two models are consistent with one another is in large part a question of which interventions in the micro-model are to be represented in the macro-model.

This is illustrative of a more general case in which the same system is modelled at two different levels of complexity: if the two models are to be considered consistent, the set of interventions modelled at the micro-level may need to be restricted, since a macro-model will generally have the capacity to express fewer interventions than a micro-model with a larger number of variables. This point cannot be formally expressed within the framework of classical SEMs as given by Definition 5.1, since these does not specify which interventions are valid.

Moreover, for a *collection* of macro-variables to be considered a causal model consistent with a micro-level model, we must also consider compositionality, an often-overlooked natural structure present in interventions. Given an SEM $\mathcal{M}_X$ over variables $X_1, \ldots, X_N$, the two interventions $\mathrm{do}(X_i = x_i)$ and $\mathrm{do}(X_j = x_j)$ for $i \neq j$ can be combined to form the intervention $\mathrm{do}(X_i = x_i, X_j = x_j)$. This corresponds to the intuitive notion that interventions on distinct causal variables can be combined, and will be formalised by imposing a *partial ordering* on the set of interventions modelled.

[3] Of coure, the fine-grained model may itself be a coarsening of an even more detailed model.

In summary, for macro-variables that are functions of underlying causal micro-variables to be themselves considered causal entities, it is important to consider the set of interventions being modelled and the structure exhibited by these interventions.

5.5 Transformations between Structural Equation Models

This section will first introduce an extended definition for SEMs that can capture restricted sets of interventions and the structure they exhibit, after which a notion of *exact transformations* between two SEMs will be formalised. The idea is that if one SEM can be viewed as an exact transformation of another, they can both be viewed as consistent causal models of the same underlying system at different levels of detail. Elementary transformations satisfying this definition will be examined, and the main result (Theorem 5.14), which states that causal reasoning is preserved under exact transformations, is presented and discussed in detail. Section 5.6 presents practical examples of exact transformations covering each of the three categories of modelling abstractions discussed in Section 5.4.1: marginalisation in systems of many variables, macro-variables derived from underlying micro-variables, and descriptions of the time-invariant behaviour of a dynamical system.

5.5.1 An extended definition for SEMs

The following definition extends that of the classical definition for SEMs. The main difference is the explicit introduction of the intervention set being modelled, as well as bringing to attention the structure of a partial ordering it exhibits. Additionally, for generality it is not assumed that the distribution P_E factorises. For convenience later on, the variables are labelled with an arbitrary index set $\mathbb{I}_X$ rather than $\{1, \ldots, N\}$, though this slight notational change has no formal impact on subsequent results. All mentions of SEMs henceforth refer to the following definition, rather than the classical definition.

Definition 5.9 (Updated definition for SEMs)**.** *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 $(\mathcal{S}_X, \mathcal{I}_X, P_E)$ where*

- *$\mathcal{S}_X$ is a set of structural equations $X_i = f_i\left(X_{\mathrm{pa}(i)}, E_i\right)$ for $i \in \mathbb{I}_X$;*
- *$(\mathcal{I}_X, \leq_X)$ 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;*
- *P_E is a (not necessarily factorised) distribution over $E = (E_i : i \in \mathbb{I}_X)$;*
- *with 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.*

As before, 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 $X_i = x_i$, while 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.

$\mathcal{I}_X$ is equipped with the natural partial ordering $\leq_X$ 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, $\text{do}(X_i = x_i) \leq_X \text{do}(X_i = x_i, X_j = x_j)$. 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 $j \leq_X i$. This structure is important and crucial use of it will be made use of in the next sections.

Aside from the treatment of interventions, Definition 5.9 additionally relaxes the usual assumption of acyclicity of the causal graph, Instead, the final condition in the definition ensures that for any intervention $i \in \mathcal{I}_X$, $\mathcal{M}_X$ induces a well-defined distribution over $\mathcal{X}$. This is always satisfied if the SEM is acyclic, but must be explicity included to also allow consideration of the cyclic case.

The following example illustrates how SEMs are written in this notation and provides an example of a restricted set of interventions $\mathcal{I}_X$.

Example 5.10. *Consider the following SEM defined over the variables* $\{B_1, B_2, L\}$

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

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

The SEM in Example 5.10 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 $\text{do}(B_1 = 0)$ 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 ($\text{do}(B_1 = 0)$) and then afterwards remove light bulb B_2 (resulting in the combined intervention $\text{do}(B_1 = 0, B_2 = 0)$) is reflected in the partial ordering via the relation $\text{do}(B_1 = 0) \leq_X \text{do}(B_1 = 0, B_2 = 0)$.

5.5.2 Partially ordered sets of distributions

For each intervention $i \in \mathcal{I}_X$, the SEM $\mathcal{M}_X$ induces a distribution over the observable variables X that we denote by $P_X^{\mathrm{do}(i)}$. Throughout, we will denote the empty- or null-intervention corresponding to the unintervened setting by $\varnothing \in \mathcal{I}_X$. For notational convenience, we will use P_X and $P_X^{\mathrm{do}(i)}$ interchangeably for the observational distribution.

$\mathcal{M}_X$ induces a set of joint distributions over $\mathcal{X}$, one for each intervention in $\mathcal{I}_X$, which moreover inherits the partial ordering from $\mathcal{I}_X$. This poset of distributions can be written as

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

where $\leq_X$ is the partial ordering inherited from $\mathcal{I}_X$, i. e. $P_X^{\mathrm{do}(i)} \leq_X P_X^{\mathrm{do}(j)} \iff i \leq_X j$.

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. For example, the distribution over the variables X in the observational setting, $P_X^{\varnothing}$, changes to $P_X^{\mathrm{do}(i)}$ under the intervention $\mathrm{do}(i)$, and the partial ordering contains all information about which distributions are subsequently attainable by composing with other interventions.

5.5.3 Exact transformations of SEMs

Suppose the function $\tau : \mathcal{X} \to \mathcal{Y}$ maps the variables of the SEM $\mathcal{M}_X$ to another space $\mathcal{Y}$. Since X is a random variable, $\tau(X)$ is also a random variable. For any distribution P_X on $\mathcal{X}$ we thus obtain the distribution of the variable $\tau(X)$ on $\mathcal{Y}$ as $P_{\tau(X)} = \tau_{\#} P_X$ via the push-forward measure.

In particular, any intervention $i \in \mathcal{I}_X$ induces the distribution $P^i_{\tau(X)} = \tau_{\#} P_X^{\mathrm{do}(i)}$. 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

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

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

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 $P^i_{\tau(X)}$, then $P^i_{\tau(X)}$ could not possibly be expressed as arising from a do-intervention $j \in \mathcal{I}_Y \setminus \{\varnothing\}$ in any SEM over $\mathcal{Y}$, an issue that is studied in detail by Eberhardt, 2016.

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 5.11 (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*

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

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

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,[4] and hence also bijective. This would constrain the ways in which $\mathcal{M}_Y$ can be 'simpler' than $\mathcal{M}_X$.[5] 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 give elementary properties of exact transformations following immediately from the definition.

Lemma 5.12. *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*

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

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. 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. Denote by ω the corresponding mapping on interventions obtained by replacing X_i with $Y_{\pi(i)}$. □

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

[4] 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 $i = j$.

[5] For instance, if it were necessary for ω to be bijective, Theorems 5.17 and 5.19 would not hold.

Lemma 5.13 (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 $\mathcal{M}_X$.*

Proof. 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 $i \in \mathcal{I}_X$.

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

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

5.5.4 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. The main technical result is the following theorem.

Theorem 5.14 (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:*

$$\begin{array}{ccccc}
P_X & \xrightarrow{\mathrm{do}(i)} & P_X^{\mathrm{do}(i)} & \xrightarrow{\mathrm{do}(j)} & P_X^{\mathrm{do}(j)} \\
\tau \downarrow & & \downarrow \tau & & \downarrow \tau \\
P_Y & \xrightarrow{\mathrm{do}(\omega(i))} & P_Y^{\mathrm{do}(\omega(i))} & \xrightarrow{\mathrm{do}(\omega(j))} & P_Y^{\mathrm{do}(\omega(j))}
\end{array}$$

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 have that $\tau_\# P_X^{\mathrm{do}(i)} = P_Y^{\mathrm{do}(\omega(i))}$ and $\tau_\# P_X^{\mathrm{do}(j)} = P_Y^{\mathrm{do}(\omega(j))}$. Thus, we only have to show that

$P_Y^{\mathrm{do}(\omega(i))} \leq_Y P_Y^{\mathrm{do}(\omega(j))}$ as elements of $\mathcal{P}_Y$, i. e. that the arrow $P_Y^{\mathrm{do}(\omega(i))} \xrightarrow{\mathrm{do}(\omega(j))} P_Y^{\mathrm{do}(\omega(j))}$ exists. This follows from the order-preservingness of ω. □

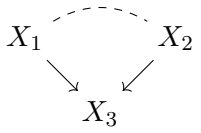
Suppose now that $\mathcal{M}_X$ is a causal model that is taken to be in some sense 'true'. We will examine the implications of another SEM $\mathcal{M}_Y$ being an exact τ-transformation of $\mathcal{M}_X$ with corresponding intervention map ω.

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. Thus, if $\mathcal{M}_X$ is physically grounded, so is $\mathcal{M}_Y$.

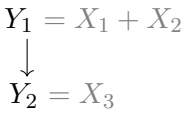
Commutativity of the left hand part of the diagram ensures that the effects of interventions are consistently modelled by $\mathcal{M}_X$ and $\mathcal{M}_Y$. Suppose we want to reason about the effects on the $\mathcal{M}_Y$-level caused by the intervention $j \in \mathcal{I}_Y$. For example, we may 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. That is, if the intervention j on the $\mathcal{M}_X$-level can be performed additionally to the intervention i in $\mathcal{M}_X$ (i. e. $i \leq_X j$), 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 5.14 ensures that they are causally consistent with one another in the sense 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



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



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

Figure 5.2 Graphical illustration of parent-child relationships for the examples in Section 5.5.5. 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 5.5.5 we give examples of what can go wrong if the transformation is not exact.

the case, causal statements such as 'raising temperature increases pressure' or 'LDL causes heart disease' are meaningful.

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

In the previous section it was argued that Definition 5.11 of exact transformations between SEMs is a sensible formalisation of causal consistency. This section provides intuition for why weakening the conditions of the definition would be problematic. Particular focus is paid to the requirement that ω be order-preserving, which is one of the core ideas of this work.

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 it were only required that ω be surjective (but not order-preserving), the observational distribution of $\mathcal{M}_X$ might be mapped to an interventional distribution of $\mathcal{M}_Y$, as illustrated by the following example, illustrated in Figure 5.2.

Example 5.15. *Consider the SEM* $\mathcal{M}_X = \{\mathcal{S}_X, \mathcal{I}_X, P_E\}$ *over* $\mathcal{X} = \mathbb{R}^3$ *where*

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

where P_{E_1} *and* 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 $\mathcal{M}_Y = \{\mathcal{S}_Y, \mathcal{I}_Y, P_F\}$ *be an SEM over* $\mathcal{Y}$ *with*

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

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 $P^i_{\tau(X)} = P_Y^{\mathrm{do}(\omega(i))}$ *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, it could be demanded 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 5.14. 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 $P_Y^{\mathrm{do}(\omega(i))}$ to $P_Y^{\mathrm{do}(\omega(j))}$ may not exist.[6]

Example 5.16. *Let* $\mathcal{X}, \mathcal{Y}$ *and* τ *be as in Example 5.15. Consider the SEM* $\mathcal{M}_X = \{\mathcal{S}_X, \mathcal{I}_X, P_E\}$ *where*

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

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

where P_{E_2} and P_{E_3} are arbitrary distributions. Let $\mathcal{M}_Y = \{\mathcal{S}_Y, \mathcal{I}_Y, P_F\}$ be the SEM over $\mathcal{Y}$ with

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

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 $P^i_{\tau(X)} = P_Y^{\mathrm{do}(\omega(i))}$ 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 5.15. 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 $\mathrm{do}(Y_1 = 1)$. The requirement that ω be order-preserving rules this pathology out.

5.5.6 Exact transformations as marginalisations in a larger model

This section provides an alternative intuition for how to think about exact transformations. Rather than thinking of them as transforming one SEM $\mathcal{M}_X$ into another $\mathcal{M}_Y$, it is possible to think of them as corresponding to marginalisation in a larger model containing both micro- and macro-variables X and Y, with some caveats.

Start with the base micro-level SEM $\mathcal{M}_X$. It is possible to add the macro-variables to this as deterministic functions of the micro-variables via the function τ. That is, one obtains an SEM $\mathcal{M}_{X,Y}$ over the variables X and Y, where the structural equations for the variables X_i are the same as in $\mathcal{M}_X$ and those for the variables Y_i are given by $Y_i = \tau_i(X)$ (note that these structural equations are deterministic and do not have exogenous variables). The model $\mathcal{M}_{X,Y}$ has the same interventions as $\mathcal{M}_X$, which induce distributions over all of the variables (X, Y).

Now, take the model $\mathcal{M}_{X,Y}$ and marginalise out the X variables, yielding the model $\widehat{\mathcal{M}}_Y$. Note that this may not be an SEM (hence that 'hat'), and the non-trivial part of this process is that there is in general no canonical way to write down the 'marginalised' structural equations—one could always simply write each Y_i as a function of the X-level exogenous noise variables, but then the structure between the Y_is would not be present.

Each intervention in $\mathcal{M}_{X,Y}$ will induce a distribution over the variables Y in the model $\widehat{\mathcal{M}}_Y$. This is analogous to the partially ordered set of distributions discussed in Section 5.5.2. If this set of distributions can be faithfully represented by some set of structural equations and interventions on these equations that respects the partial ordering on the original intervention set, then this model $\mathcal{M}_Y$ would be an exact transformation of $\mathcal{M}_X$.

In general, one could relax the notion of perfect interventions to allow probabilistic interventions, which would make it easier for the distributions of the marginalised model $\widehat{\mathcal{M}}_Y$ to be expressible in a single SEM. One could also use the framework of Dawid, 2020 to include the interventions as part of the set of variables being marginalised over, to make it clearer that interventions originally made on X propagate to interventions on Y after marginalisation.

However, the challenge would still remain that without further thought, marginalisation would result in a graph in which there are no direct arrows between the Y variables, and all dependence between them is due to confounders. E.g. in Theorem 5.19 and Figure 5.4 in the next section, marginalising naively would result in no arrow $\widehat{W} \to \widehat{Z}$, but rather a confounder between the two variables. Such a representation would miss the point of the whole exercise, which is to identify structure between variables at the macro-level. This is not a necessary outcome of the marginalisation view; this issue is raised only to bring attention to the fact that the representation of the relationships between the remaining variables would need to be carefully considered, in much the same way that it would in the view previously discussed in this chapter.

5.6 Examples of exact transformations

The problem of modelling at multiple levels of complexity was motivated in Section 5.4.1 by listing three settings in which differing model levels naturally occur: marginalisation of unobserved variables; macro-level models of underlying micro-variables; and time-invariant descriptions of dynamical systems. Having introduced the notion of an exact transformation between SEMs, this section provides examples of exact transformations falling into each of the categories.

Observe that in each example, the particular set of interventions considered is important. If one were to allow larger sets of interventions $\mathcal{I}_X$ in the SEM $\mathcal{M}_X$, the transformations given

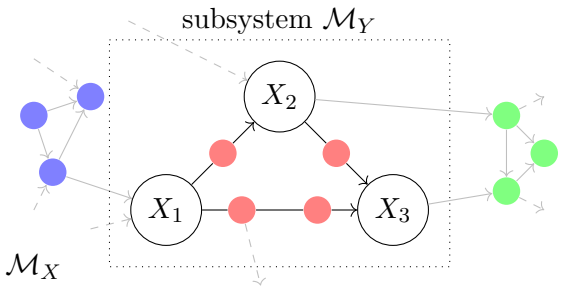


Figure 5.3 Suppose that a complex model $\mathcal{M}_X$ is given but that only the subsystem X_1, X_2, X_3 is of interest. By Theorem 5.17, downstream effects (●) can be ignored after grouping them together as one multivariate variable. By Theorem 5.18 intermediate steps of complex mechanisms (●) can be ignored and upstream causes (●) treated as exogenous noise. That is, the complex SEM $\mathcal{M}_X$ can be exactly transformed into a simpler model $\mathcal{M}_Y$ by marginalisation of the irrelevant variables.

would not be exact, highlighting the importance to the causal modelling process of carefully considering the set of interventions. All proofs can be found in Appendix C.

5.6.1 Marginalisation of variables

In the following two theorems it is shown that the marginalisation of childless or non-intervened variables is an exact transformation, illustrated in Figure 5.3. That is, an SEM can be simplified by marginalising out variables of either of these types 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 $\mathcal{M}_X$.

Theorem 5.17 (Marginalisation of childless variables)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_E)$ 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 5.18 (Marginalisation of non-intervened variables)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_E)$ 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*

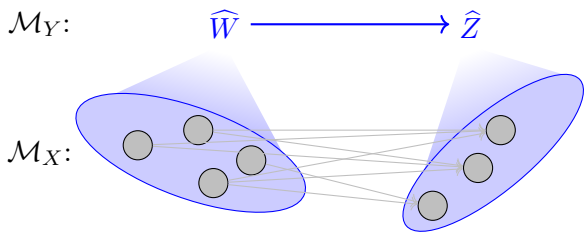


Figure 5.4 An illustration of the setting considered in Theorem 5.19. The micro-variables $W_1, \ldots, W_n$ and $Z_1, \ldots, Z_m$ in the SEM $\mathcal{M}_X$ can be averaged to derive macro-variables $\widehat{W}$ and $\widehat{Z}$ in such a way that the resulting macro-level SEM $\mathcal{M}_Y$ is an exact transformation of the micro-level SEM $\mathcal{M}_X$.

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 5.18 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.

Recall that Definition 5.9 does not require that the exogenous E-variables of a SEM be independent. Theorem 5.18 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.

5.6.2 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. For instance, 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, such as viewing the motor cortex as a single entity responsible for movements, rather than as a collection of individual neurons.

In such situations, the framework of exact transformations allows the investigation of whether such a macro-level model admits a causal interpretation. The following theorem, illustrated in Figure 5.4, 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.

Theorem 5.19 (Micro- to macro-level)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_{E,F})$ 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*

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

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

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

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

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

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

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

5.6.3 Stationary behaviour of dynamical processes

This section provides 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, illustrated in Figure 5.5. In this setting, τ could be thought of as representing our ability to measure the time-evolving system at only 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. The assumption that the transition dynamics are linear could be relaxed to

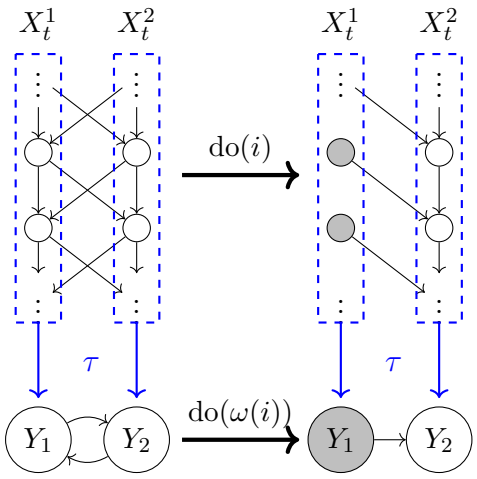


Figure 5.5 An illustration of the setting considered in Theorem 5.20. The discrete-time dynamical process is exactly transformed into a model describing its equilibria.

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, the simpler case of linear dynamics suffices.

Theorem 5.20 (Discrete-time linear dynamical process with identical noise)**.** *Let* $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_E)$ *over the variables* $\{X_t^i \,:\, t \in \mathbb{Z},\, i \in \{1,\dots,n\}\}$ *be a linear SEM with*

$$\begin{aligned}
\mathcal{S}_X &= \left\{ X_{t+1}^i = \sum_{j=1}^{n} A_{ij} X_t^j + E_t^i \;:\; i \in \{1,\dots,n\}, t \in \mathbb{Z} \right\}, \\
&\quad \textit{i. e. } X_{t+1} = AX_t + E_t \\
\mathcal{I}_X &= \Big\{ \mathrm{do}(X_t^j = x_j \;\; \forall t \in \mathbb{Z}, \forall j \in J) : x \in \mathbb{R}^{|J|},\; J \subseteq \{1,\dots,n\} \Big\}, \\
E_t &= E \;\forall t \in \mathbb{Z} \textit{ where } E \sim P,
\end{aligned}$$

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

Assume that the linear mapping $v \mapsto Av$ *is a contraction.*[7] *Then the following transformation is well-defined under any intervention* $i \in \mathcal{I}_X$*:*

$$\begin{aligned}
\tau : \mathcal{X} &\to \mathcal{Y}, \\
(x_t)_{t\in\mathbb{Z}} &\mapsto y = \lim_{t\to\infty} x_t.
\end{aligned}$$

[7]In Appendix C.3 it is shown 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.

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

$$\mathcal{S}_Y = \left\{ Y^i = \frac{\sum_{j \neq i} A_{ij} Y^j}{1 - A_{ii}} + \frac{F^i}{1 - A_{ii}} \,:\, i \in \{1, \ldots, n\} \right\},$$
$$\mathcal{I}_Y = \left\{ \mathrm{do}(Y^j = y_j \;\forall j \in J) : y \in \mathbb{R}^{|J|},\; J \subseteq \{1, \ldots, n\} \right\},$$
$$F \sim P.$$

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

5.7 Discussion

This section discusses the implications of this work to the causality literature, future directions for extending this work as well as outlining other papers have already built directly upon it since the publication of the paper on which this chapter is in part based (Rubenstein, Weichwald et al., 2017).

5.7.1 Implications to causality literature

Ambiguous manipulations and causal variable definition

The example of cholesterol and heart disease discussed in Section 5.4 was adapted from Spirtes and Scheines, 2004, who illustrate the problem of *ambiguous manipulations* of variables that are functions of underlying causal entities (e.g. TC = HDL + LDL). Although their work was primarily concerned with causal discovery in such a context, Eberhardt, 2016 demonstrated broader problematic implications to the idea of causal variable definition: in general, functions of causal variables cannot be considered to be themselves causal.

The framework of exact transformations between SEMs provides a partial solution to this issue, as it has been argued that if a higher level SEM $\mathcal{M}_Y$ is an exact transformation of a SEM $\mathcal{M}_X$ of causal variables, then the variables in $\mathcal{M}_Y$ can also be considered causal entities. The key idea missing in previous work was to consider not simply individual variables in isolation, but the entire SEM *along with the interventions being modelled* as the object being transformed.

Cyclic causal models

As previously discussed, there is no settled consensus within the causality literature on what constitutes a 'valid' cyclic SEM. Acyclic SEMs are usually interpreted as corresponding to a temporally ordered series of mechanisms by which data are generated; this is not possible for cyclic SEMs since there is no partial ordering on the variables. The most common interpretation is that they correspond to a dynamical system that converges quickly relative to its environment, represented by the exogenous noise variables (Mooij et al., 2013). This is supported by Theorem 5.20.

However, most real physical systems exhibit feedback, and many of these do not satisfy this assumption. This work provides a framework to also consider such cases. Since all physical systems evolve in time, they can be represented as an acyclic model $\mathcal{M}_X$. In general our ability to make measurements is imperfect, and can be represented by the function τ. Whether or not the observed variables $\tau(X)$ admit a causal interpretation is equivalent to asking whether there exists an $\mathcal{M}_Y$ that is an exact τ-transformation of $\mathcal{M}_X$. The framework of exact transformations between SEMs thus provides a way to think about cyclic SEMs, showing more generally how cyclicity can arise as a result of imperfect measurements of underlying acyclic models.

5.7.2 Extensions

A high-level model need not be a perfectly accurate and faithful representation of a low-level model in order to be useful. Indeed, all models are an approximation to some degree. In this regard, the theory of exact transformations is perhaps too strict because of the requirement that $\tau_{\#} P_X^{\mathrm{do}(i)} = P_Y^{\mathrm{do}(\omega(i))}$ should hold for all interventions $i \in \mathcal{I}_X$.

In the original publication from which this chapter is in part based (Rubenstein, Weichwald et al., 2017), one of the proposals for future directions of enquiry was to generalise the notion of an exact transformation to an *approximate* transformation in which the requirement of equality is relaxed to hold only approximately, something that could be rigorously formalised in a variety of ways.

Since publication of this work, Beckers et al., 2019 have investigated precisely this question, exploring several subtleties. For instance, divergences between pairs of distributions can be defined in multiple ways and a notion of nearness must hold in some sense over a set of interventions. They propose two ways of defining approximate abstractions and show how they are related, providing a way to quantify the trade-off between accuracy and abstraction. These ideas are illustrate in practice with application to climate models, showing how the El Niño event arises as a macroscopic property of wind and sea temperatures.

In a second paper, Beckers and Halpern, 2019 consider in great detail the notion of casual consistency proposed in this work. They argue that there exist examples of exact transformations that would nonetheless generally be considered to be inconsistent (or even unrelated) causal models. In essence, the issue is that it is possible to use arbitrary distributions P_E over the exogenous noise variables in order to construct pairs of SEMs for which an exact transformation can be established, even though the models have little in common. In the paper they propose a sequence of increasingly restrictive definitions for transformations between SEMs, the weakest of which corresponds to the exact transformations of this chapter, and the strongest of which rules out the aforementioned pathologies.

5.7.3 Future directions

In this work and that of Beckers et al., 2019 and Beckers and Halpern, 2019, no criteria to choose from amongst the set of all possible exact transformations of an SEM is provided. Foundational work by Chalupka et al., 2015; Chalupka et al., 2016 considered the construction of higher-level causal variables in a particular discrete setting, providing algorithms to learn a transformation of a micro-level model to a macro-level model with desirable information-theoretic properties. The framework proposed here may lead to extensions of their work, for example to the continuous setting.

Finally, suppose that observations of an underlying system $\mathcal{M}_X$ have been made via a measurement device τ, and that an SEM $\mathcal{M}_Y$ is fit from a restricted model class to the collected data. The framework proposed here, and extended by the aforementioned works, provides a method to ask whether $\mathcal{M}_Y$ admits a causal interpretation consistent with $\mathcal{M}_X$. Although verifying that a transformation is exact (or approximate) may be impossible in practical scenarios where $\mathcal{M}_X$ and τ may be very complex (e.g. measuring the brain with EEG), it may be possible to identify general cases in which this does or does not hold. This may lead to the practical use of SEMs being more theoretically grounded.

Chapter 6

Conclusion

This thesis presented theoretical advances in three niches of the machine learning literature related to the modelling of structured data. This chapter summarises the main contributions and discusses future directions of research.

6.1 Summary of contributions

Chapter 3 presented an estimator for f-divergences between pairs of distributions satisfying certain structural assumptions that are naturally satisfied in the setting of autoencoders. These assumptions enabled the derivation of fast rates for the decay of the bias and concentration of this estimator without additional strong assumptions on the distributions. This is in contrast to much of the existing f-divergence estimation literature, where fast rates are only attainable with strong assumptions that would be difficult to verify in practice.

Chapter 4 presented identifiability results for a novel multi-view nonlinear ICA setting, extending the few identifiability results known for nonlinear ICA. These results required at least one of the views to exhibit source-side noise, termed corruptions. In particular, if one noiseless view of the sources is supplemented by a second view that is appropriately corrupted by source-level noise, it was proved that the sources can be fully reconstructed from the observations up to tolerable ambiguities. This setting has application to practical scenarios in which multiple distinct data modalities are available, such as in neuroimaging.

Chapter 5 introduced the notion of *exact transformations* between Structural Equations Models (SEMs), providing a framework to understand when two SEMs can be viewed as consistent causal models of the same system at different levels of detail. This provides a way to formally understand when higher-level variables can be considered to be causal variables, and encompasses a wide range of settings in which such higher-level models arise.

Practically all measurements are made at a level of detail different to that at which 'true' causal structure exists, yet causal discovery algorithms typically seek causal relations at the level of measurements. Thus, this work has broad implications to the causality community in general and in particular to the problem of causal variable definition. It furthermore brings to attention the importance of the specification of interventions of interest as a part of the causal modelling process.

6.2 Future directions

Chapter 3 was fundamentally a learning theoretic study of f-divergence estimation under particular structural assumptions. One direction for future enquiry would be the use of the proposed RAM-MC estimator for optimisation, instead of pure estimation. A clear application of this would be to the training of Wasserstein Autoencoders, the regularisation term of which is any divergence between the prior and aggregate posterior, and naturally satisfies the structural assumptions considered in this chapter.

This work has broader implications as it demonstrates that there is interesting work to be done at the intersection of deep learning and learning theory. While the learning theory literature has tended to focus on settings in which as few assumptions are made as possible, this work shows that in some cases strong assumptions that naturally apply to modern deep learning settings can yield superior results. To give one specific example, it is known that in the general case, estimation of mutual information is a hard problem (McAllester and Stratos, 2018). Yet in many practical cases where mutual information is used, such as in representation learning (Hjelm et al., 2018; Oord et al., 2018; Tschannen et al., 2020), stronger assumptions may hold than in the general case. One such setting was encountered in Chapter 3, though others may exist.

The identifiability results presented in Chapter 4 show that ICA in the multi-view setting is in principle possible, and natural next steps would be the development of practical algorithms that actually work in application.

An emerging area within deep learning is *disentangled representation learning*. This empirically driven community shares similar goals to the ICA community but with a strong emphasis on image datasets. Despite this, few bridges have been built between the two communities, though recent work in disentanglement has begun to consider multi-view settings similar to that considered here (Shu et al., 2019).

One barrier to connecting the ICA and disentanglement communities is the pervasive assumption in ICA that the source and observation dimensions be the same. In high dimensional data such as images, this is clearly unrealistic as usually dozens of latent dimensions are

sufficient to explain the majority of variation in images with hundreds or thousands of pixels. Thus, attempting to relax this assumption would seem to be a possible way to give ICA wider applicability across the modern machine learning community.

Similarly, gaps exist between the causality literature and modern advances in deep learning. The fundamental assumption to almost all causal learning algorithms is that the individual components of the data are meaningful entities. In contrast, deep learning algorithms can be applied to data such as images and audio for which the components of raw data, i.e. individual pixels or amplitudes at a particular point in time, are themselves not meaningful, but where higher-level features such as objects, textures or syllables are. Causality is nonetheless gaining increasing attention outside of the traditional community, with authors such as Bengio et al., 2019 attempting to blend ideas from causality with deep learning.

The framework introduced in Chapter 5 allows one to reason about whether higher-level causal variables are consistent with the 'raw' variables from which they are derived, but it is not clear how such coarsenings can be learned automatically from data. My hope is that others may build on this work, leading to 'causal' feature learning. However, given the central importance of interventions in the causal setting, I have reservations about whether this is possible given the current paradigm of i.i.d. machine learning with large datasets. Reinforcement learning, however, could be a fruitful area in which to apply ideas from causality, given the centrality of action there.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous and Kevin Murphy. "Fixing a Broken ELBO". *ICML*. 2018, pp. 159–168.

Galen Andrew, Raman Arora, Jeff Bilmes and Karen Livescu. "Deep canonical correlation analysis". *International conference on machine learning*. 2013, pp. 1247–1255.

M. Arjovsky, S. Chintala and L. Bottou. "Wasserstein GAN". *arXiv:1701.07875* (2017).

Francis R Bach and Michael I Jordan. "A probabilistic interpretation of canonical correlation analysis". *Technical report 688, Deparment of Statistics, UC Berkeley* (2005).

R. Balian. *From microphysics to macrophysics*. Springer, 1992.

Sander Beckers, Frederick Eberhardt and Joseph Y Halpern. "Approximate Causal Abstraction". *Proceedings conference on Uncertainty in Artificial Intelligence (UAI)*. Vol. 2019. NIH Public Access. 2019.

Sander Beckers and Joseph Y Halpern. "Abstracting causal models". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 2678–2685.

Shai Ben-David, John Blitzer, Koby Crammer and Fernando Pereira. "Analysis of representations for domain adaptation". *Advances in neural information processing systems*. 2007, pp. 137–144.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal and Christopher Pal. "A meta-transfer objective for learning to disentangle causal mechanisms". *arXiv preprint arXiv:1901.10912* (2019).

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Stephan Bongers, Jonas Peters, Bernhard Schölkopf and Joris M Mooij. "Structural Causal Models: Cycles, Marginalizations, Exogenous Reparametrizations and Reductions". *arXiv preprint arXiv:1611.06221* (2016).

Stéphane Boucheron, Gábor Lugosi and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Yuri Burda, Roger Grosse and Ruslan Salakhutdinov. "Importance weighted autoencoders". *arXiv preprint arXiv:1509.00519* (2015).

Krzysztof Chalupka, Pietro Perona and Frederick Eberhardt. "Multi-level cause-effect systems". *The 19th International Conference on Artificial Intelligence and Statistics*. 2016.

Krzysztof Chalupka, Pietro Perona and Frederick Eberhardt. "Visual causal feature learning". *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2015, pp. 181–190.

Po-Hsuan Cameron Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby and Peter J Ramadge. "A reduced-dimension fMRI shared response model". *Advances in Neural Information Processing Systems*. 2015, pp. 460–468.

Liqun Chen, Chenyang Tao, Ruiyi Zhang, Ricardo Henao and Lawrence Carin Duke. "Variational Inference and Model Selection with Generalized Evidence Bounds". *ICML*. 2018. URL: http://proceedings.mlr.press/v80/chen18k.html.

Tian Qi Chen, Xuechen Li, Roger Grosse and David Duvenaud. "Isolating Sources of Disentanglement in Variational Autoencoders". *arXiv preprint arXiv:1802.04942* (2018).

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt and David K Duvenaud. "Neural ordinary differential equations". *Advances in Neural Information Processing Systems*. 2018, pp. 6572–6583.

Pierre Comon. "Independent component analysis, a new concept?" *Signal processing* 36.3 (1994), pp. 287–314.

Imre Csiszár and Paul C Shields. "Information theory and statistics: A tutorial". *Foundations and Trends® in Communications and Information Theory* 1.4 (2004), pp. 417–528.

George Darmois. "Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire". *Revue de l'Institut international de statistique* (1953), pp. 2–8.

Denver Dash and Marek J Druzdzel. "Caveats for causal reasoning with equilibrium models". *Lecture notes in computer science* (2001), pp. 192–203.

A Philip Dawid. "Decision-theoretic foundations for statistical causality". *arXiv preprint arXiv:2004.12493* (2020).

Virginia R De Sa. "Spectral clustering with two views". *ICML workshop on learning with multiple views*. 2005, pp. 20–27.

Adji B Dieng, Yoon Kim, Alexander M Rush and David M Blei. "Avoiding latent variable collapse with generative skip models". *arXiv preprint arXiv:1807.04863* (2018).

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley and David Blei. "Variational Inference via χ Upper Bound Minimization". *Advances in Neural Information Processing Systems*. 2017, pp. 2732–2741.

Mingzhou Ding, Yonghong Chen and Steven L Bressler. "Granger causality: basic theory and application to neuroscience". *Handbook of time series analysis: recent theoretical developments and applications* 437 (2006).

Laurent Dinh, David Krueger and Yoshua Bengio. "Nice: Non-linear independent components estimation". *arXiv preprint arXiv:1410.8516* (2014).

Laurent Dinh, Jascha Sohl-Dickstein and Samy Bengio. "Density estimation using real nvp". *arXiv preprint arXiv:1605.08803* (2016).

Jeff Donahue and Karen Simonyan. "Large scale adversarial representation learning". *Advances in Neural Information Processing Systems*. 2019, pp. 10541–10551.

J-L Durrieu, J-Ph Thiran and Finnian Kelly. "Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models". *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee. 2012, pp. 4833–4836.

Frederick Eberhardt. "Green and grue causal variables". *Synthese* 193.4 (2016), pp. 1029–1046.

Franklin M Fisher. "A correspondence principle for simultaneous equation models". *Econometrica: Journal of the Econometric Society* (1970), pp. 73–92.

Kenji Fukumizu, Francis R Bach and Arthur Gretton. "Statistical consistency of kernel canonical correlation analysis". *Journal of Machine Learning Research* 8.Feb (2007), pp. 361–383.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky. "Domain-adversarial training of neural networks". *The Journal of Machine Learning Research* 17.1 (2016), pp. 2096–2030.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. "Generative adversarial nets". *Advances in neural information processing systems*. 2014, pp. 2672–2680.

Luigi Gresele*, Paul K Rubenstein*, Arash Mehrjou, Francesco Locatello and Bernhard Schölkopf. "The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-view Nonlinear ICA". *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*. *Joint first authorship. 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf and Alexander Smola. "A kernel two-sample test". *Journal of Machine Learning Research* 13.Mar (2012), pp. 723–773.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke and Peter J Ramadge. "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron* 72.2 (2011), pp. 404–416.

M. Hein and O. Bousquet. "Hilbertian metrics and positive definite kernels on probability measures". *AISTATS*. 2005.

A. O. Hero, B. Ma, O. J. J. Michel and J. Gorman. "Applications of entropic spanning graphs". *IEEE Signal Processing Magazine* (2002).

A. O. Hero, B. Ma, O. Michel and J. Gorman. "Alpha divergence for classification, indexing and retrieval". *Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, Univ. Michigan, Ann Arbor, Tech. Rep. 328* (2001).

John R Hershey and Peder A Olsen. "Approximating the Kullback Leibler divergence between Gaussian mixture models". *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 4. IEEE. 2007, pp. IV–317.

John Hicks et al. *Causality in economics*. Canberra, ACT: Australian National University Press, 1980.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler and Yoshua Bengio. "Learning deep representations by mutual information estimation and maximization". *arXiv preprint arXiv:1808.06670* (2018).

Erik P Hoel, Larissa Albantakis and Giulio Tononi. "Quantifying causal emergence shows that macro can beat micro". *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19790–19795.

Matthew D Hoffman and Matthew J Johnson. "ELBO surgery: yet another way to carve up the variational evidence lower bound". 2016.

Timo Honkela, Aapo Hyvärinen and Jaakko J Väyrynen. "WordICA-emergence of linguistic representations for words by independent component analysis". *Natural Language Engineering* 16.3 (2010), pp. 277–308.

Harold Hotelling. "Relations between two sets of variates". *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters and Bernhard Schölkopf. "Nonlinear causal discovery with additive noise models". *Advances in neural information processing systems*. 2009, pp. 689–696.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen and Markus Palviainen. "Estimation of causal effects using linear non-Gaussian causal models with hidden variables". *International Journal of Approximate Reasoning* 49.2 (2008), pp. 362–378.

Antti Hyttinen, Frederick Eberhardt and Patrik O Hoyer. "Causal discovery for linear cyclic models with latent variables". *on Probabilistic Graphical Models* (2010), p. 153.

Antti Hyttinen, Frederick Eberhardt and Patrik O Hoyer. "Learning linear cyclic causal models with latent variables". *Journal of Machine Learning Research* 13.Nov (2012), pp. 3387–3439.

Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt and Matti Jarvisalo. "Discovering cyclic causal models with latent variables: A general SAT-based procedure". *arXiv preprint arXiv:1309.6836* (2013).

Aapo Hyvärinen and Hiroshi Morioka. "Unsupervised feature extraction by time-contrastive learning and nonlinear ICA". *Advances in Neural Information Processing Systems*. 2016, pp. 3765–3773.

Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". *Neural networks* 13.4-5 (2000), pp. 411–430.

Aapo Hyvärinen and Petteri Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results". *Neural Networks* 12.3 (1999), pp. 429–439.

Aapo Hyvärinen, Hiroaki Sasaki and Richard Turner. "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning". *Proceedings of Machine Learning Research*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 859–868. URL: http://proceedings.mlr.press/v89/hyvarinen19a.html.

AJ Hyvärinen and Hiroshi Morioka. "Nonlinear ICA of temporally dependent stationary sources". Proceedings of Machine Learning Research. 2017.

Yumi Iwasaki and Herbert A Simon. "Causality and model abstraction". *Artificial Intelligence* 67.1 (1994), pp. 143–194.

Jörn-Henrik Jacobsen, Arnold Smeulders and Edouard Oyallon. "i-RevNet: Deep Invertible Networks". *ICLR 2018 - International Conference on Learning Representations*. Vancouver, Canada, Apr. 2018. URL: https://hal.archives-ouvertes.fr/hal-01712808.

Dominik Janzing, Patrik O Hoyer and Bernhard Schölkopf. "Telling cause from effect based on high-dimensional observations". *arXiv preprint arXiv:0909.4386* (2009).

Dominik Janzing, Joris M Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel and Bernhard Schölkopf. "Information-geometric approach to inferring causal directions". *Artificial Intelligence* 182 (2012), pp. 1–31.

Dominik Janzing, Jonas Peters, Joris M Mooij and Bernhard Schölkopf. "Identifying confounders using additive noise models". *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2009, pp. 249–257.

Dominik Janzing, Paul Rubenstein and Bernhard Schölkopf. "Structural causal models for macro-variables in time-series". *arXiv preprint arXiv:1804.03911* (2018).

Dominik Janzing and Bernhard Schölkopf. "Causal inference using the algorithmic Markov condition". *IEEE Transactions on Information Theory* 56.10 (2010), pp. 5168–5194.

T. Kanamori, T. Suzuki and M. Sugiyama. "f-Divergence Estimation and Two-Sample Homogeneity Test under Semiparametric Density-Ratio Models". *IEEE Transactions on Information Theory* 58.2 (2012).

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever and Max Welling. "Improved variational inference with inverse autoregressive flow". *Advances in Neural Information Processing Systems*. 2016, pp. 4743–4751.

Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114* (2013).

A. Krishnamurthy, A. Kandasamy, B. Póczos and L. Wasserman. "Nonparametric estimation of Rényi divergence and friends". *ICML*. 2014.

Julius von Kügelgen, Paul K Rubenstein, Bernhard Schölkopf and Adrian Weller. "Optimal experimental design via Bayesian optimization: active causal structure learning for Gaussian process networks". *NeurIPS 2019 Workshop "Do the right thing": Machine Learning and Causal Inference for Improved Decision Making*. 2019.

Abhishek Kumar, Piyush Rai and Hal Daume. "Co-regularized multi-view spectral clustering". *Advances in neural information processing systems*. 2011, pp. 1413–1421.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey and Patrik O Hoyer. "Discovering cyclic causal models by independent components analysis". *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence*. 2008.

Pei Ling Lai and Colin Fyfe. "Kernel and nonlinear canonical correlation analysis". *International Journal of Neural Systems* 10.05 (2000), pp. 365–377.

Yingzhen Li and Richard E Turner. "Rényi divergence variational inference". *Advances in Neural Information Processing Systems*. 2016, pp. 1073–1081.

Friedrich Liese and Igor Vajda. "On divergences and informations in statistics and information theory". *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412.

Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang. "Deep Learning Face Attributes in the Wild". *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf and Ilya Tolstikhin. "Towards a learning theory of cause-effect inference". *Proceedings of the 32nd International Conference on Machine Learning, JMLR: W&CP, Lille, France*. 2015.

David McAllester and Karl Stratos. "Formal limitations on the measurement of mutual information". *arXiv preprint arXiv:1811.04251* (2018).

Martin J McKeown and Terrence J Sejnowski. "Independent component analysis of fMRI data: examining the assumptions". *Human brain mapping* 6.5-6 (1998), pp. 368–372.

Tomer Michaeli, Weiran Wang and Karen Livescu. "Nonparametric canonical correlation analysis". *International Conference on Machine Learning*. 2016, pp. 1967–1976.

Joris M Mooij and Tom Heskes. "Cyclic Causal Discovery from Continuous Equilibrium Data". *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence*. 2013.

Joris M Mooij, Dominik Janzing, Tom Heskes and Bernhard Schölkopf. "On causal discovery with cyclic additive noise models". *Advances in neural information processing systems*. 2011, pp. 639–647.

Joris M Mooij, Dominik Janzing and B Schölkopf. "Distinguishing between cause and effect." *NIPS Causality: Objectives and Assessment*. 2010, pp. 147–156.

Joris M Mooij, Dominik Janzing and Bernhard Schölkopf. "From Ordinary Differential Equations to Structural Causal Models: the deterministic case". *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence*. 2013, pp. 440–448.

K. Moon and A. Hero. "Ensemble estimation of multivariate f-divergence". *2014 IEEE International Symposium on Information Theory*. 2014, pp. 356–360.

K. Moon and A. Hero. "Multivariate f-divergence Estimation With Confidence". *NeurIPS*. 2014.

Alfred Müller. "Integral probability metrics and their generating classes of functions". *Advances in Applied Probability* 29.2 (1997), pp. 429–443.

XuanLong Nguyen, Martin J. Wainwright and Michael I. Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". *IEEE Trans. Information Theory* 56.11 (2010), pp. 5847–5861.

Frank Nielsen and Richard Nock. "On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences". *IEEE Signal Process. Lett.* 21.1 (2014), pp. 10–13.

Sebastian Nowozin, Botond Cseke and Ryota Tomioka. "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization". *NIPS*. 2016.

Danielle Nuzillard and Albert Bijaoui. "Blind source separation and analysis of multispectral astronomical images". *Astronomy and Astrophysics Supplement Series* 147.1 (2000), pp. 129–138.

Erkki Oja, Kimmo Kiviluoto and Simona Malaroiu. "Independent component analysis for financial time series". *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*. IEEE. 2000, pp. 111–116.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio". *arXiv preprint arXiv:1609.03499* (2016).

Aaron van den Oord, Yazhe Li and Oriol Vinyals. "Representation learning with contrastive predictive coding". *arXiv preprint arXiv:1807.03748* (2018).

Ferdinand Osterreicher and Igor Vajda. "A new class of metric divergences on probability spaces and its applicability in statistics". *Annals of the Institute of Statistical Mathematics* 55.3 (2003), pp. 639–653.

George Papamakarios, Theo Pavlakou and Iain Murray. "Masked autoregressive flow for density estimation". *Advances in Neural Information Processing Systems*. 2017, pp. 2338–2347.

Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2005.

Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein and Frank Nielsen. "Sinkhorn autoencoders". *arXiv preprint arXiv:1810.01118* (2018).

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

F. Perez-Cruz. "Kullback-Leibler divergence estimation of continuous distributions". *IEEE International Symposium on Information Theory*. 2008.

Jonas Peters and Peter Bühlmann. "Identifiability of Gaussian structural equation models with equal error variances". *Biometrika* (2013), ast043.

Jonas Peters, Dominik Janzing and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Jonas Peters, Dominik Janzing and Bernhard Schölkopf. "Identifying Cause and Effect on Discrete Data using Additive Noise Models." *AIStats*. 2010, pp. 597–604.

Jonas Peters, Joris M Mooij, Dominik Janzing, Bernhard Schölkopf, et al. "Causal discovery with continuous additive noise models." *Journal of Machine Learning Research* 15.1 (2014), pp. 2009–2053.

B. Poczos and J. Schneider. "On the estimation of alpha-divergences". *AISTATS*. 2011.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A Alemi and George Tucker. "On variational lower bounds of mutual information". *ICML*. 2018.

Danilo Jimenez Rezende, Shakir Mohamed and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". *International Conference on Machine Learning*. 2014, pp. 1278–1286.

Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1530–1538. URL: http://proceedings.mlr.press/v37/rezende15.html.

Paul K Rubenstein, Stephan Bongers, Bernhard Schölkopf and Joris M Mooij. "From deterministic ODEs to dynamic structural causal models". *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*. 2018.

Paul K Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme and Ilya Tolstikhin. "Practical and Consistent Estimation of f-Divergences". *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

Paul K Rubenstein, Yunpeng Li and Dominik Roblek. "An Empirical Study of Generative Models with Encoders". *arXiv preprint arXiv:1812.07909* (2018).

Paul K Rubenstein, Bernhard Schoelkopf and Ilya Tolstikhin. "On the Latent Space of Wasserstein Auto-Encoders". *arXiv preprint arXiv:1802.03761* (2018).

Paul K Rubenstein, Bernhard Schölkopf and Ilya Tolstikhin. "Learning Disentangled Representations with Wasserstein Auto-Encoders". *International Conference on Learning Representations (ICLR), Workshop Track*. 2018.

Paul K Rubenstein, Bernhard Schölkopf and Ilya Tolstikhin. "Wasserstein auto-encoders: Latent dimensionality and random encoders". *International Conference on Learning Representations (ICLR), Workshop Track*. 2018.

Paul K Rubenstein, Ilya Tolstikhin, Philipp Hennig and Bernhard Schölkopf. "Probabilistic Active Learning of Functions in Structural Causal Models". *Causality Workshop of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. 2017.

Paul K Rubenstein*, Sebastian Weichwald*, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup and Bernhard Schölkopf. "Causal consistency of structural equation models". *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. *Joint first authorship. 2017.

Hiroshi Sawada, Ryo Mukai and Shoji Makino. "Direction of arrival estimation for multiple source signals using independent component analysis". *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.* Vol. 2. IEEE. 2003, pp. 411–414.

Richard Scheines, Frederick Eberhardt and Patrik O Hoyer. "Combining experiments to discover linear cyclic models with latent variables" (2010).

Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel and Jonas Peters. "Modeling confounding by half-sibling regression". *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7391–7398.

Rajen D Shah and Jonas Peters. "The hardness of conditional independence testing and the generalised covariance measure". *arXiv preprint arXiv:1804.07203* (2018).

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen and Antti Kerminen. "A linear non-Gaussian acyclic model for causal discovery". *Journal of Machine Learning Research* 7.Oct (2006), pp. 2003–2030.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon and Ben Poole. "Weakly Supervised Disentanglement with Guarantees". *arXiv preprint arXiv:1910.09772* (2019).

Herbert A Simon and Albert Ando. "Aggregation of variables in dynamic systems". *Econometrica: journal of the Econometric Society* (1961), pp. 111–138.

Amit Singer and Ronald R Coifman. "Non-linear independent component analysis with diffusion maps". *Applied and Computational Harmonic Analysis* 25.2 (2008), pp. 226–239.

S. Singh and B. Poczos. "Generalized Exponential Concentration Inequality for Rényi Divergence Estimation". *ICML*. 2014.

Viktor Pavlovich Skitovich. "Linear forms of independent random variables and the normal distribution law". *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 18.2 (1954), pp. 185–200.

Le Song, Animashree Anandkumar, Bo Dai and Bo Xie. "Nonparametric estimation of multi-view latent variable models". *International Conference on Machine Learning*. 2014, pp. 640–648.

Peter Spirtes and Richard Scheines. "Causal inference of ambiguous manipulations". *Philosophy of Science* 71.5 (2004), pp. 833–845.

Henning Sprekeler, Tiziano Zito and Laurenz Wiskott. "An extension of slow feature analysis for nonlinear blind source separation". *The Journal of Machine Learning Research* 15.1 (2014), pp. 921–947.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf and Gert RG Lanckriet. "On integral probability metrics,\phi-divergences and binary classification". *arXiv preprint arXiv:0901.2698* (2009).

Daniel Steinberg. *The Cholesterol Wars: The Skeptics vs the Preponderance of Evidence*. Academic Press, 2011.

Esteban G Tabak and Cristina V Turner. "A family of nonparametric density estimation algorithms". *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164.

Esteban G Tabak and Eric Vanden-Eijnden. "Density estimation by dual ascent of the log-likelihood". *Communications in Mathematical Sciences* 8.1 (2010), pp. 217–233.

Anisse Taleb and Christian Jutten. "Source separation in post-nonlinear mixtures". *IEEE Transactions on signal Processing* 47.10 (1999), pp. 2807–2820.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly and Bernhard Schoelkopf. "Wasserstein auto-encoders". *ICLR*. 2018.

James Townsend, Thomas Bird, Julius Kunze and David Barber. "HiLLoC: Lossless Image Compression with Hierarchical Latent Variable Models". *arXiv preprint arXiv:1912.09953* (2019).

James Townsend, Tom Bird and David Barber. "Practical lossless compression with latent variables using bits back coding". *arXiv preprint arXiv:1901.04866* (2019).

A Stewart Truswell. *Cholesterol and beyond: the research on diet and coronary heart disease 1900-2000*. Springer Science & Business Media, 2010.

Michael Tschannen, Olivier Bachem and Mario Lucic. "Recent advances in autoencoder-based representation learning". *arXiv preprint arXiv:1812.05069* (2018).

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly and Mario Lucic. "On mutual information maximization for representation learning". *International Conference on Learning Representations (ICLR)*. 2020.

Alexandre B. Tsybakov. "Introduction to Nonparametric Estimation" (2009). DOI: 10.1007/b13794.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. "Conditional image generation with pixelcnn decoders". *Advances in neural information processing systems*. 2016, pp. 4790–4798.

Aaron Van Oord, Nal Kalchbrenner and Koray Kavukcuoglu. "Pixel Recurrent Neural Networks". *International Conference on Machine Learning*. 2016, pp. 1747–1756.

Q. Wang, S. R. Kulkarni and S. Verdú. "Divergence estimation for multidimensional densities via k-nearest-neighbor distances". *IEEE Transactions on Information Theory* 55.5 (2009).

Kun Zhang and Aapo Hyvärinen. "Distinguishing causes from effects using nonlinear acyclic causal models". *Journal of machine learning research, workshop and conference proceedings (NIPS 2008 causality workshop)*. Vol. 6. 2008, pp. 157–164.

Kun Zhang and Aapo Hyvärinen. "On the identifiability of the post-nonlinear causal model". *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press. 2009, pp. 647–655.

Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu and David Barber. "Variational f-divergence Minimization". *arXiv preprint arXiv:1907.11891* (2019).

Appendix A

Additional Materials for Chapter 3

A.1 Proof of Proposition 1

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

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

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

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

where the inequality follows from convexity of f.

To see that $\mathbb{E}_{\mathbf{X}^N \sim Q_X^N} D_f(\hat{Q}_Z^N, P_Z) \leq \mathbb{E}_{\mathbf{X}^M \sim Q_X^N} D_f(\hat{Q}_Z^M, P_Z)$ 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, $\mathbf{X}^I \sim Q_X^M$ (with the randomness coming from $\mathbf{X}^N \sim Q_X^N$). 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

$$\begin{aligned}\mathbb{E}_{\mathbf{X}^N\sim Q_X^N} D_f(\hat{Q}_Z^N, P_Z) &\le \mathbb{E}_{\mathbf{X}^N}\mathbb{E}_I D_f(\hat{Q}_Z^I, P_Z) \\ &= \mathbb{E}_{\mathbf{X}^M} D_f(\hat{Q}_Z^M, P_Z)\end{aligned}$$

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

A.2 Proof of Theorem 3.11

Theorem 3.11 (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 3.1 (see Page 31).*

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 3.3) 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 3.2).

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 3.5 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 $\psi(\beta)$ (Lemma 3.4) 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 3.1) in combination with Lemma 3.5 to obtain

$$\mathbb{E}_{\mathbf{X}^N}\left[\mathrm{H}^2\left(\hat{Q}_Z^N, P_Z\right)\right] - \mathrm{H}^2\left(Q_Z, P_Z\right) \leq O\left(\frac{1}{\sqrt{N}}\right).$$

□

A.3 Upper bounds of f

We will make use of the following lemmas in the proofs of Theorem 3.12 and 3.13.

Lemma A.1. *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 A.2. *Let $f_0(x) = 2(1 - \sqrt{x})$ corresponding to the squared-Hellinger divergence. 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 A.3. *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 A.4. *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 A.5. *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 $x \in [0, \infty)$. □

A.4 Proof of Theorem 3.12

Theorem 3.12 (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 3.1 (see Page 31).*

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 Section 2.4.2).

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 due 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

$$\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N, P_Z\right)\right] - D_f\left(Q_Z, P_Z\right) = \mathbb{E}_{\mathbf{X}^N}\left[\int_Z \left(\frac{\hat{q}_N(z) - q(z)}{p(z)}\right)^2 dP(z)\right]. \tag{A.1}$$

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 A.1 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 A.5 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

$$\text{(B)} \leq f_0(0) \times \mathbb{P}_{Z,\mathbf{X}^N} \underbrace{\left\{\frac{\hat{q}_N(z)}{p(z)} \leq \delta_N \text{ and } \frac{q(z)}{p(z)} > \gamma_N\right\}}_{(*)}.$$

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 A.1-A.5 in order to upper bound $f_0'^2(x)$ with something that can be easily integrated.

KL. By Lemma A.1, 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,

$$\text{(C)} \leq \sqrt{B}\sqrt{\frac{\log^2(\delta_N) + \frac{2}{e}}{N}}.$$

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 A.2 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 A.3 provides a function h_δ that upper bounds $f'^2(x)$ for $x \in\in [\delta, \infty)$.

$$\begin{aligned}
(\dagger\dagger)^2 &= \mathbb{E}_{\mathbf{X}^N}\left[\int f_0'^2\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right)\mathbb{1}_{\left\{\frac{\hat{q}_N(z)}{p(z)} > \delta_N\right\}} p(z)dz\right] && (h_{\delta_N} \text{ upper bounds } f_0'^2 \text{ on } (\delta_N, \infty)) \\
&\leq \mathbb{E}_{\mathbf{X}^N}\left[\int h_{\delta_N}\left(\frac{\hat{q}_N(z)}{p(z)}\right) p(z)dz\right] && (h_{\delta_N} \text{ non-negative on } [0, \infty)) \\
&= \frac{4\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)^2}{(\alpha-1)^2(\delta_N - 1)^2}\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)}{p(z)} - 1\right)^2\right] \\
&\leq \frac{4\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)^2}{(\alpha-1)^2(\delta_N - 1)^2}\mathbb{E}_{\mathbf{X}^N}\mathbb{E}_{P_Z}\left[\left(\frac{\hat{q}_N(z)}{p(z)}\right)^2 + 1\right] \\
&= \frac{4\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)^2}{(\alpha-1)^2(\delta_N - 1)^2}\left(1 + \mathbb{E}_{\mathbf{X}^N}\left[\left\|\frac{\hat{q}_N(z)}{p(z)}\right\|_{L_2(P_Z)}^2\right]\right) \\
&\leq \frac{4(1+B)\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)^2}{(\alpha-1)^2(\delta_N - 1)^2} \\
\implies (\dagger\dagger) &= \frac{2\sqrt{1+B}\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)}{(\alpha-1)(\delta_N - 1)}.
\end{aligned}$$

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 A.4 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

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N, P_Z\right)\right] - D_f\left(Q_Z, P_Z\right) \\
&\leq \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0(\gamma_N) + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} - \frac{\sqrt{5}\sqrt{B}\log\delta_N}{\sqrt{N}} \\
&\leq \gamma_N\log\left(\frac{1+\gamma_N}{2\gamma_N}\right) + \log(1+\gamma_N) + \frac{\log 2\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} - \frac{\sqrt{5}\sqrt{B}\log\delta_N}{\sqrt{N}}.
\end{aligned}
$$

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 A.5 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 A.5). Thus

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N, P_Z\right)\right] - D_f\left(Q_Z, P_Z\right) \\
&\leq \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0(\gamma_N) + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \frac{\sqrt{B}}{\sqrt{N}} 2^{\frac{2+\beta}{2\beta}} \left(\frac{\beta}{1-\beta}\right) \delta_N^{\beta-1} \\
&\leq \frac{\beta}{1-\beta}\left[1 - \left(1+\delta_N^\beta\right)^{1/\beta} + 2^{\frac{1-\beta}{\beta}}\delta_N\right] + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \frac{\sqrt{B}}{\sqrt{N}} 2^{\frac{2+\beta}{2\beta}} \left(\frac{\beta}{1-\beta}\right) \delta_N^{\beta-1} \\
&\leq \frac{\beta}{1-\beta} 2^{\frac{1-\beta}{\beta}}\delta_N + f_0(0)\frac{\sqrt{C}}{N\left(\gamma_N - \delta_N\right)^2} + \frac{\sqrt{B}}{\sqrt{N}} 2^{\frac{2+\beta}{2\beta}} \left(\frac{\beta}{1-\beta}\right) \delta_N^{\beta-1} \\
&= k_1\delta_N + \frac{k_2}{N(\gamma_N - \delta_N)^2} + \frac{k_3\delta_N^{\beta-1}}{\sqrt{N}}
\end{aligned}$$

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

$$\begin{aligned}
&= \frac{k_1}{N^{\frac{1}{3}}} + \frac{k_2}{N^{\frac{1}{3}}} + \frac{k_3}{N^{\frac{1}{2}+\frac{\beta-1}{3}}} \\
&= O\left(\frac{1}{N^{\frac{1}{3}}}\right).
\end{aligned}$$

□

A.5 Proof of Theorem 3.13

We will make use of McDiarmid's inequality (Theorem 3.10) in our proof of Theorem 3.13. In our setting we will consider $\phi(\mathbf{X}^N) = D_f\left(\hat{Q}_Z^N, P_Z\right)$.

Theorem 3.13 (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 3.2.

Proof. 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 $\mathbf{X}^N, \mathbf{X}^{N\prime}$ 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 $\left|D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}'^N), P_Z\right)\right| \leq c_N$ 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 3.13 is of this form. Note that in order to show that

$$\left|D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}'^N), P_Z\right)\right| \leq c_N, \tag{A.2}$$

it is sufficient to prove that

$$D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}'^N), P_Z\right) \leq c_N, \tag{A.3}$$

since the symmetry in $\mathbf{X}^N \leftrightarrow \mathbf{X}'^N$ implies that

$$-D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) + D_f\left(\hat{Q}_Z^N(\mathbf{X}'^N), P_Z\right) \leq c_N, \tag{A.4}$$

and thus implies Inequality A.2. The remainder of this proof is therefore devoted to showing that Inequality A.3 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 3.12 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 3.12 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 3.12, 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

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime}), P_Z\right) \\
&= \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) dP_Z(z) \\
&= \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) \mathbb{1}_{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\leq\delta_N \text{ and } \frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\leq\gamma_N\right\}} dP_Z(z) \qquad \text{(A)} \\
&\quad + \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) \mathbb{1}_{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\leq\delta_N \text{ and } \frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)>\gamma_N\right\}} dP_Z(z) \qquad \text{(B)} \\
&\quad + \int f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)\right) - f_0\left(\frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z)\right) \mathbb{1}_{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z)>\delta_N\right\}} dP_Z(z). \qquad \text{(C)}
\end{aligned}
$$

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 $\gamma_N, \delta_N \leq 1$, 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): Since $\delta_N \leq 1$, the integrand is at most $f_0(0)$ and so

$$\text{(B)} \leq f_0(0) \times \mathbb{P}_Z \underbrace{\left\{\frac{d\hat{Q}_Z^N(\mathbf{X}^N)}{dP_Z}(z) \leq \delta_N \text{ and } \frac{d\hat{Q}_Z^N(\mathbf{X}^{N\prime})}{dP_Z}(z) > \gamma_N\right\}}_{(*)}.$$

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 3.12, 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 A.1, 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 3.12 for the α-divergence case, we use the function $h_{\delta_N}(x)$ provided by Lemma A.3 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 $\gamma_N = \frac{2}{N^{\frac{4}{\alpha+5}}}$,

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime}), P_Z\right) \\
&= \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{2(C+1)}{N}\frac{2\sqrt{1+(C+1)^2}\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)}{(1-\alpha)(1-\delta_N)} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{t^2 f_0(0)C}{(t-1)^2(\gamma_N - \delta_N)^2 N^2} + \frac{2(C+1)}{N}\frac{2\sqrt{1+(C+1)^2}\left(\delta_N^{\frac{\alpha-1}{2}} - 1\right)}{(1-\alpha)(1-\delta_N)} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{t^2 f_0(0)C}{(t-1)^2(\gamma_N - \delta_N)^2 N^2} + \frac{2(C+1)}{N}\frac{4\sqrt{1+(C+1)^2}\delta_N^{\frac{\alpha-1}{2}}}{(1-\alpha)} \\
&\leq k_1\gamma_N^{\frac{\alpha+1}{2}} + k_2\gamma_N + \frac{k_3}{(\gamma_N - \delta_N)^2 N^2} + \frac{k_4\delta_N^{\frac{\alpha-1}{2}}}{N} \\
&= \frac{k_1}{N^{\frac{2\alpha+2}{\alpha+5}}} + \frac{k_2}{N^{\frac{4}{\alpha+5}}} + \frac{k_3}{N^{\frac{2\alpha-2}{\alpha+5}}} + \frac{k_4}{N^{\frac{3\alpha+3}{\alpha+5}}} \\
&\leq \frac{k_1 + k_2 + k_3 + k_4}{N^{\frac{2\alpha+2}{\alpha+5}}},
\end{aligned}
$$

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 3.12 for the α-divergence case, we use the function $h_{\delta_N}(x)$ provided by Lemma A.4 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 3.12 for the α-divergence case, we use the function $h_{\delta_N}(x)$ provided by Lemma A.5 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 $\gamma_N = \frac{2}{N^{2/3}}$,

$$
\begin{aligned}
&D_f\left(\hat{Q}_Z^N(\mathbf{X}^N), P_Z\right) - D_f\left(\hat{Q}_Z^N(\mathbf{X}^{N\prime}), P_Z\right) \\
&= \textcircled{A} + \textcircled{B} + \textcircled{C} \\
&\leq f_0(0) - f_0\left(\gamma_N\right) + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{\beta}{1-\beta} \cdot 2^{\frac{2+\beta}{2\beta}} \delta_N^{\beta-1} \\
&\leq \frac{\beta}{\beta - 1} 2^{\frac{1-\beta}{\beta}} \gamma_N + \frac{f_0(0)C}{\left((\gamma_N - \delta_N)N - 1\right)^2} + \frac{\beta}{1-\beta} \cdot 2^{\frac{2+\beta}{2\beta}} \frac{\delta^{\beta-1}}{N} \\
&= \frac{k_1}{N^{2/3}} + \frac{k_2}{(N^{1/3} - 1)^2} + \frac{k_3}{N^{\frac{2\beta+1}{3}}} \\
&\leq \frac{k_1}{N^{2/3}} + \frac{100 k_2}{81 N^{2/3}} + \frac{k_3}{N^{\frac{2\beta+1}{3}}} \\
&\leq \frac{k_1 + k_2' + k_3}{N^{2/3}},
\end{aligned}
$$

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 $N c_N^2 = \frac{1}{N^{1/3}}$ for $N > 2$. □

A.6 Proof of Theorem 3.14

Theorem 3.14 (RAM-MC is unbiased and consistent)**.** *For any proposal distribution π, RAM-MC is unbiased:*

$$\mathbb{E}[\hat{D}_f^M(\hat{Q}_Z^N, P_Z)] = \mathbb{E}[D_f(\hat{Q}_Z^N, P_Z)].$$

If the hypothesis of Theorem 3.13 holds and moreover either of the following conditions are satisfied:

$$(i)\begin{cases}\pi(z|\mathbf{X}^N)=p(z),\\ \mathbb{E}_{Q_X}\int f\left(\dfrac{q(z|X)}{p(z)}\right)^2 p(z)dz<\infty,\\ \mathbb{E}_{Q_X}\int \left(\dfrac{q(z|X)}{p(z)}\right)^2 p(z)dz<\infty,\end{cases}$$

$$(ii)\begin{cases}\pi(z|\mathbf{X}^N)=\hat{q}_N(z),\\ \mathbb{E}_{Q_X}\int f\left(\dfrac{q(z|X)}{p(z)}\right)^2\left(\dfrac{p(z)}{q(z|X)}\right)^2 q(z|X)dz<\infty,\\ \mathbb{E}_{Q_X}\int \left(\dfrac{p(z)}{q(z|X)}\right)^2 q(z|X)dz<\infty,\end{cases}$$

then denoting by $\psi(N)$ the rate given in Table 3.2, the variance of RAM-MC decays as

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

In proving Theorem 3.14 we will make use of the following lemma.

Lemma A.6. *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 second 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 3.14) 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 3.13 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

$$\begin{aligned}\mathrm{Var}_{\mathbf{X}^N}\left[D_f\left(\hat{Q}_Z^N,P_Z\right)\right] &\le \int_0^\infty 2\exp\left(-\frac{k}{\psi(N)^2}t\right)dt\\ &= O\left(\psi(N)^2\right),\end{aligned}$$

where $\psi(N)$ is given by Table 3.2.

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 A.6). 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

$$(i) \leq \mathbb{E}_X \mathbb{E}_{p(z)} \left[f_0 \left(\frac{q(z|X)}{p(z)} \right)^2 \frac{p(z)}{q(z|X)} \right] + f'(1)^2 \mathbb{E}_X \mathbb{E}_{p(z)} \left[\left(\sqrt{\frac{q(z|X)}{p(z)}} - \sqrt{\frac{p(z)}{q(z|X)}} \right)^2 \right]$$
$$+ 2f'(1) \sqrt{\mathbb{E}_X \mathbb{E}_{p(z)} \left[f_0 \left(\frac{q(z|X)}{p(z)} \right)^2 \frac{p(z)}{q(z|X)} \right]} \times \sqrt{\mathbb{E}_X \mathbb{E}_{p(z)} \left[\left(\sqrt{\frac{q(z|X)}{p(z)}} - \sqrt{\frac{p(z)}{q(z|X)}} \right)^2 \right]}$$

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 A.6). 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. □

Appendix B

Additional Materials for Chapter 4

B.1 Proofs for one noiseless view results (Section 4.3.1)

B.1.1 Proof of Theorem 4.1

Theorem 4.1. *The difference between the log joint probability and log product of marginals of the observed variables in the model given in Equations 4.8-4.9 admits the following factorisation:*

$$\begin{aligned}
&\log p(x_1, x_2) - \log p(x_1)p(x_2) \\
&= \log p(x_2|x_1) - \log p(x_2) \\
&= \left(\sum_i \alpha_i(z_i, g_i(z_i, n_i)) + \log\det J\right) \\
&\qquad - \left(\sum_i \delta_i(g_i(z_i, n_i)) + \log\det J\right) \\
&= \sum_i \alpha_i(z_i, g_i(z_i, n_i)) - \sum_i \delta_i(g_i(z_i, n_i)),
\end{aligned} \tag{B.1}$$

where $z_i = f_{1i}^{-1}(x_1)$, $g_i = f_{2i}^{-1}(x_2)$, *and* J *is the Jacobian of the transformation* f_2^{-1} *(note that the introduced Jacobians cancel). Suppose that*

1. α *satisfies the* Sufficiently Distinct Views *assumption.*

2. *A classifier is trained to discriminate between*

$$(X_1, X_2) \text{ vs. } (X_1, X_2^*),$$

where (X_1, X_2) *correspond to the same realisation of* Z *and* (X_1, X_2^*) *correspond to different realisations of* Z*.*

3. *The classifier minimises the logistic regression loss, and is constrained to use a regression function of the form*

$$r(x_1, x_2) = \sum_i \psi_i(h_i(x_1), x_2)$$

where $h = (h_1, \ldots, h_n)$ *is invertible, smooth and has smooth inverse.*

Then, in the limit of infinite data and with universal approximation capacity, h *inverts* f_1 *in the sense that the* $h_i(X_1)$ *recover the independent components of* Z *up to component-wise invertible transformations.*

This proof is inspired by the techniques employed by Hyvärinen et al., 2019.

Proof. The goal is to show that for the optimal classifier, $h_i(X_1)$ depends on exactly one component of Z.

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

$$\sum_i \psi_i(h_i(x_1), x_2) = \sum_i \alpha_i(f_{1,i}^{-1}(x_1), f_{2,i}^{-1}(x_2)) - \sum_i \delta_i(f_{2,i}^{-1}(x_2)).$$

Making the change of variables

$$\begin{aligned} y &= h(x_1), \\ v(y) &= f_1^{-1}(h^{-1}(y)), \\ t &= f_2^{-1}(x_2)), \end{aligned}$$

means that the first equation can be rewritten in the following form:

$$\sum_i \psi_i(y_i, x_2) = \sum_i \alpha_i(v_i(y), t_i) - \sum_i \delta_i(t_i). \tag{B.2}$$

Take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of each side of this equation. Adopting the notation from the SDV assumption definition

$$\begin{aligned}
\alpha_i'(y_i, t_i) &= \partial \alpha_i(y_i, t_i)/\partial y_i, \\
\alpha_i''(y_i, t_i) &= \partial^2 \alpha_i(y_i, t_i)/\partial y_i^2, \\
w_\alpha(y, t) &= (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D'),
\end{aligned}$$

and furthermore defining

$$\begin{aligned}
v_i^j(y) &= \partial v_i(y)/\partial y_j, \\
v_i^{jj'}(y) &= \partial^2 v_i(y)/\partial y_j \partial y_{j'},
\end{aligned}$$

we have

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

The left-hand side of this equation is 0 because each term ψ_i in Equation B.2 depends on exactly one y_i, and partial derivatives are taken with respect to two different y_i and y_j.

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

$$\sum_i \alpha_i''(v_i(y), t_i) a_i(y) + \alpha_i'(v_i(y), t_i)) b_i(y) = 0\,.$$

This can in turn be rewritten in matrix form,

$$M(y) w(y, t) = 0\,,$$

where $M(y) = (a_1(y), \ldots, a_D(y), b_1(y), \ldots, b_D(y))$ and $w(y, t) = (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D')$. $M(y)$ is therefore a $D(D-1)/2 \times 2D$ matrix, and $w(y, t)$ is a $2D$ dimensional vector.

To show that $M(y)$ is equal to zero, we invoke the SDV assumption. This implies the existence of $2D$ linearly independent $w(y, t_j)$. It follows that

$$M(y)[w(y, t_1), \ldots, w(y, t_{2D})] = 0\,,$$

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

Observe that $v(y) = z$. We have just proven that $v_i(y_{\pi(i)}) = z_i$ where π is some permutation. Since v_i is invertible, it follows that $h_{\pi(i)}(x_1) = y_{\pi(i)} = v_i^{-1}(z_i)$ and hence the components of $h(x_1)$ recover the components of z up to the invertible component-wise ambiguity given by v, and the permutation ambiguity.

□

B.1.2 Proof of Corollary 4.3

Corollary 4.3. *Consider the setting of Theorem 4.1 with the alternative factorisation of the log joint probability*

$$\begin{aligned}
&\log p(x_1, x_2) - \log p(x_1)p(x_2) \\
&= \log p(x_1|x_2) - \log p(x_1) \\
&= \sum_i \gamma_i(z_i, g_i(z_i, n_i)) - \sum_i \beta_i(z_i)) \,.
\end{aligned} \tag{B.3}$$

Suppose that γ satisfies the SDV assumption. Replacing the regression function with

$$r(x_1, x_2) = \sum_i \psi_i(x_1, h_i(x_2))$$

results in h inverting f_2 in the sense that the $h_i(X_2)$ recover the independent components of the $g(Z, N)$ up to component-wise invertible transformations.

Proof. This follows exactly by repeating the proof of Theorem 4.1 where the roles of x_1 and x_2 are exchanged and the regression function in the statement of the corollary is used. □

B.2 Proofs for two noisy view results (Section 4.3.2)

B.2.1 Proof of Theorems 4.4 and 4.5

Theorem 4.4 is a special case of Theorem 4.5 by considering the case $\boldsymbol{g}_1(\boldsymbol{z}, \boldsymbol{n}_1) = \boldsymbol{z}$. We therefore prove only the more general Theorem 4.5.

Theorem 4.5. *Suppose that η and λ satisfy the SDV assumption. The algorithm described in Theorem 4.1 with regression function specified in Equation 4.17 results in h_1 and h_2 inverting f_1 and f_2 in the sense that the $h_{1,i}(X_1)$ and $h_{2,i}(X_2)$ recover the independent components of $g_1(Z, N_1)$ and $g_2(Z, N_2)$ up to two different component-wise invertible transformations.*

Furthermore, the two representations are aligned, i.e. for $i \neq j$,

$$h_{1,i}(X_1) \perp\!\!\!\perp h_{2,j}(X_2).$$

Proof. This proof is similar to that of Theorem 4.1.

The goal is to show that for the optimal classifier, $h_{1i}(X_1)$ and $h_{2i}(X_2)$ each depend on exactly one component of Z. Moreover, in order to show that the representations are aligned, we will show that

$$h_{1,i}(x_1) \perp\!\!\!\perp h_{2,j}(x_2), \forall i \neq j. \tag{B.4}$$

We start by exploiting Equations 4.18 and 4.19 to write the difference in log-densities of the two classes

$$\sum_i \psi_i(h_{1,i}(x_1), h_{2,i}(x_2)) = \sum_i \eta_i(f_{1,i}^{-1}(x_1), f_{2,i}^{-1}(x_2)) - \sum_i \theta_i(f_{1,i}^{-1}(x_1)) \tag{B.5}$$

$$= \sum_i \lambda_i(f_{2,i}^{-1}(x_2), f_{1,i}^{-1}(x_1)) - \sum_i \mu_i(f_{2,i}^{-1}(x_2)) \tag{B.6}$$

and make the change of variables

$$\begin{aligned} y &= h_1(x_1), \\ t &= h_2(x_2), \\ v(y) &= f_1^{-1}(h_1^{-1}(y)), \\ u(t) &= f_2^{-1}(h_2^{-1}(t)). \end{aligned}$$

Equation B.5 can thus be rewritten as

$$\sum_i \psi_i(y_i, t_i) = \sum_i \eta_i(v_i(y), u_i(t)) - \sum_i \theta_i(v_i(y)). \tag{B.7}$$

To show that $h_{1i}(X_1)$ depends on exactly one component of Z, we will show that

$$v_i(y) = v_i(y_{\pi(i)}) \tag{B.8}$$

for some permutation of the indices π with respect to the indexing of the sources $z = (z_1, \ldots, z_D)$.

Taking derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of equation B.7 yields

$$0 = \sum_i \eta_i''(v_i(y), u_i(t)) v_i^j(y) v_i^{j'}(y) + \sum_i \eta_i'(v_i(y), u_i(t)) v^{jj'}(y).$$

Define vectors $a_i(y)$ collecting all entries $v_i^j(y)v_i^{j'}(y)$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $b_i(y)$ with the variables $v_i^j(y)v_i^{j'}(y)$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$0 = \sum_i \eta_i''(v_i(y), u_i(t))a_i(y) + \eta_i'(v_i(y), u_i(t))b_i(y).$$

As in the proof of Theorem 4.1, this can be rewritten in matrix form as

$$M(y)w(y,t) = 0\,, \tag{B.9}$$

where $M(y) = (a_1(y), \ldots, a_D(y), b_1(y), \ldots, b_D(y))$ and $w(y,t) = (\eta_1'', \ldots, \eta_D'', \eta_1', \ldots, \eta_D')$. $M(y)$ is therefore a $D(D-1)/2 \times 2D$ matrix, and $w(y,t)$ is a $2D$ dimensional vector.

To show that $M(y)$ is equal to zero, we invoke the SDV assumption on η. This implies the existence of $2D$ linearly independent $w(y,t_j)$. It follows that

$$M(y)[w(y,t_1), \ldots, w(y,t_{2D})] = 0\,,$$

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

Observe that $v(y) = z$. We have just proven that $v_i(y_{\pi(i)}) = z_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(x_1) = y_{\pi(i)} = v_i^{-1}(z_i)$ and hence the components of $h(x_1)$ recover the components of z up to the invertible component-wise ambiguity given by v, and the permutation ambiguity.

To prove that $h_{2i}(X_2)$ depends on exactly one component of Z, exactly the same argument can be applied, replacing (v, y, η, θ) with (u, t, λ, μ), noting that the SDV assumption is also assumed for λ. We thus see that

$$u_i(t) = u_i(t_{\tilde{\pi}(i)})\,, \tag{B.10}$$

where the permutation $\tilde{\pi}$ may be different from π.

We have shown that $y = h_1(x_1)$ and $t = h_2(x_2)$ estimate $g_1(z, n_1)$ and $g_2(z, n_2)$ 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 $s_1 = g_1(z, n_1)$ and $s_2 = g_2(z, n_2)$, while

$$s_{1,i} \perp\!\!\!\perp s_{2,j} \forall i \neq j \tag{B.11}$$

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 $i = 1, \ldots, n$. To show that this ambiguity is also resolved, we will show that

$$y_i \perp\!\!\!\perp t_j, \ \ \forall i \neq j. \tag{B.12}$$

We recall that, by definition, we have $v_i(y_{\pi(i)}) = s_{1,i}$ and $u_j(t_{\tilde{\pi}(j)}) = s_{2,j}$. Then, due to equation B.11,

$$v_i(y_{\pi(i)}) \perp\!\!\!\perp u_j(t_{\tilde{\pi}(j)}) \ \forall i \neq j \tag{B.13}$$

$$\implies y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \ \forall i \neq j \tag{B.14}$$

$$\implies y_i \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(j)} \ \forall i \neq j\,, \tag{B.15}$$

where the implication B.13-B.14 follows from invertibility of v_i and u_j, and the implication B.14-B.15 follows from considering that, given that we know B.14, 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

$$y_i \perp\!\!\!\perp t_{\tau(j)} \forall i \neq j. \tag{B.16}$$

Fix any particular i. Our goal is to show that for any $j \neq i$ the independence relation in Equation B.12 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 B.16 for all $j \neq i$ implies each of the independences of Equation B.12 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

$$y_i \perp\!\!\!\perp t_l\,, \tag{B.17}$$

which is the only independence implied by Equation B.12 which is not implied by Equation B.16.

In order to do so, we rewrite equation B.7, yielding

$$\sum_m \psi_m(y_m, t_m) = \sum_m \eta_m(v_m(y_{\pi(m)}), u_m(t_{\tilde{\pi}(m)})) - \sum_m \theta_i(v_m(y_{\pi(m)})).$$

We now take derivative with respect to y_i and t_l in B.17; noting that $\tilde{\pi}^{-1}(l) = \pi^{-1}(i)$, we get

$$0 = \frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) \times \frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l). \quad \text{(B.18)}$$

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

$$\frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) = 0 \quad \text{(B.19)}$$

almost everywhere. It thus follows that

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

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

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

and therefore

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

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

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

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

This concludes the proof.

□

B.2.2 Proof of Corollary 4.6

Corollary 4.6. *Let $N_1^{(k)} = \frac{1}{k} \cdot \tilde{N}$ for $k \in \mathbb{N}$, where $\tilde{N} \in \mathbb{R}^D$ is a fixed random variable with finite variance, and let N_2 be a random variable that does not depend on k. Let $h_1^{(k)}, h_2^{(k)}$ be the output of the algorithm specified by Theorem 4.5 with noise variables $N_1^{(k)}$ and N_2.*

Suppose that the corrupters g_i satisfy the following two criteria:

1. $\exists a \in \mathbb{R}^D_{>0}$ *s.t.* $\left|\frac{\partial g_1(z,n)}{\partial n}\right|_{n=0} \leq a$ *for all* z,
2. $\exists b \in \mathbb{R}^D_{>0}$ *s.t.* $0 < \frac{\partial g_1(z,0)}{\partial z} \leq b$.

Then, denoting by $\mathcal{E}$ the set of all component-wise, invertible functions, it holds that

$$\inf_{e \in \mathcal{E}} \left\| Z - e(h_1^{(k)}(X_1)) \right\| \xrightarrow[k\to\infty]{p} 0,$$

where p denotes convergence in probability.

Proof. To show that the random variable

$$\inf_{e \in \mathcal{E}} \left\| Z - e(h_1^{(k)}(X_1)) \right\|$$

converges to 0 in probability, we will prove the stronger statement that it converges in mean to 0. Denoting by $d_1^{(k)}$ the component-wise invertible ambiguity up to which $g(Z, N_1^{(k)})$ is recovered, we have that

$$\begin{aligned}
&\mathbb{E}_{Z,X_1} \left| \inf_{e \in \mathcal{E}} \left\| Z - e(h_1^{(k)}(X_1)) \right\| \right| && \text{(B.20)}\\
&= \mathbb{E}_{Z,X_1} \inf_{e \in \mathcal{E}} \left\| Z - e(h_1^{(k)}(X_1)) \right\| && \text{(B.21)}\\
&\leq \inf_{e \in \mathcal{E}} \mathbb{E}_{Z,X_1} \left\| Z - e(h_1^{(k)}(X_1)) \right\| && \text{(B.22)}\\
&= \inf_{e \in \mathcal{E}} \mathbb{E}_{Z,N_1^{(k)}} \left\| Z - e \circ d_1^{(k)} \circ g_1(Z, N_1^{(k)}) \right\| && \text{(B.23)}\\
&= \inf_{\tilde{e} \in \mathcal{E}} \mathbb{E}_{Z,N_1^{(k)}} \left\| Z - \tilde{e} \circ g_1(Z, N_1^{(k)}) \right\| && \text{(B.24)}\\
&\leq \mathbb{E}_{Z,N_1^{(k)}} \left\| Z - e^* \circ g_1(Z, N_1^{(k)}) \right\|, && \text{(B.25)}
\end{aligned}$$

where the first upper bound holds by concavity, and the second holds for any $e^* \in \mathcal{E}$ by definition of infimum and in particular for $e^* = g_1|_{n=0}^{-1}$, the existence of which is guaranteed

by the assumptions on g_1. Taking a Taylor expansion of $e^* \circ g_1(z, n_1^{(k)})$ around $n_1^{(k)} = 0$ yields

$$\begin{aligned}
&\mathbb{E}_{(Z,N_1^{(k)})}\left[\left\| Z - e^* \circ g_1(Z,0) + \frac{\partial e^*}{\partial g_1}\frac{\partial g_1(Z,0)}{\partial n_1^{(k)}} \cdot N_1^{(k)} + \mathcal{O}(\|N_1^{(k)}\|^2)\right\|\right] \\
&= \mathbb{E}_{(Z,N_1^{(k)})}\left[\left\| \frac{\partial e^*}{\partial g_1}\frac{\partial g_1(Z,0)}{\partial n_1^{(k)}} \cdot N_1^{(k)} + \mathcal{O}(\|N_1^{(k)}\|^2)\right\|\right] \\
&\longrightarrow 0 \text{ as } k \longrightarrow \infty,
\end{aligned}$$

where the last equality follows from fact that $e^* = g|_{\boldsymbol{n}=0}^{-1}$ and the convergence follows from the fact that $N_1^{(k)} = \frac{1}{k}N_1$ where N_1 has finite variance (and thus mean) and from the boundedness conditions on the partial derivatives of g_1. □

B.3 Proofs for multiple noisy views results (Section 4.3.3)

B.3.1 Proof of Lemma 4.7

Lemma 4.7. *Suppose that the sequence* $\mathbb{E}_N[\Omega_e^M(Z,N)] = \frac{1}{M}\sum_{i=1}^M \mathbb{E}_{N_i}[e_i \circ k_i(Z+N_i)]$ *converges as* $M \to \infty$ *for almost all* Z*, and write this limit as*

$$\Omega_e(Z) = \lim_{M\to\infty} \mathbb{E}_N[\Omega_e^M(Z,N)].$$

Suppose further that there exists K *such that* $V_{e_i} = \mathrm{Var}\left(e_i \circ k_i(Z+N_i)\right) \leq K$ *for all* i*. Then*

$$\begin{aligned}
&\Omega_e^M(Z,N) \xrightarrow{a.s.} \Omega_e(Z) \\
&R_{e,i}^M(Z,N) \xrightarrow{a.s.} R_{e,i}(Z,N_i) = e_i \circ k_i(Z+N_i) - \Omega_e(Z).
\end{aligned}$$

In proving this, we will make crucial use of *Kolmogorov's strong law*:

Theorem B.1. *Suppose that* X_m *is a sequence of independent (but not necessarily identically distributed) random variables with*

$$\sum_{m=1}^{\infty} \frac{1}{m^2}\mathrm{Var}[X_m] < \infty.$$

Then,

$$\frac{1}{M}\sum_{m=1}^{M} X_m - \mathbb{E}[X_m] \xrightarrow{a.s.} 0.$$

Proof of Lemma 4.7. Fix z and consider $\Omega_e^M(z,n)$ as a random variable with randomness induced by $n = (n_1, n_2, \ldots)$. We will show that for almost all z this converges n-almost surely to a constant, and hence $\Omega_e^M(z,n)$ converges almost surely to a function of z.

The law of total expectation says that

$$\begin{aligned}
&\mathrm{Var}_{z,n_i}[e_i \circ k_i(z+n_i)] \\
&= \mathbb{E}_z\left[V_i(z)\right] + \mathrm{Var}_z\left[\mathbb{E}_{n_i}[e_i \circ k_i(z+n_i)]\right] \\
&\geq \mathbb{E}_z\left[V_i(z)\right].
\end{aligned}$$

Since by assumption $\mathrm{Var}_{z,n_i}[e_i \circ k_i(z+n_i)] \leq K$, we have that

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

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

We have further that for almost all z,

$$\Omega_e(z) = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} E_{e_i}(z)$$

exists. Therefore, for almost all s the conditions of Kolmogorov's strong law are met by $\Omega_e^M(z,n)$ and so

$$\Omega_e^M(z,n) - \mathbb{E}_n[\Omega_e^M(z,n)] \overset{n-a.s.}{\longrightarrow} 0.$$

Since $\mathbb{E}_n[\Omega_e^M(z,n)] \overset{n-a.s.}{\longrightarrow} \Omega_e(z)$, it follows that

$$\Omega_e^M(z,n) \overset{n-a.s.}{\longrightarrow} \Omega_e(z).$$

Since this holds with probability 1 over z, we have that

$$\Omega_e^M(z,n) \overset{n-a.s.}{\longrightarrow} \Omega_e(z).$$

It follows that we can write

$$\begin{aligned}
R_{e,i}^M(z,n) &= e_i \circ k_i(z+n_i) - \Omega_e^M(z,n) \\
&\overset{a.s.}{\longrightarrow} R_{e,i}(z,n_i) := e_i \circ k_i(z+n_i) - \Omega_e(z).
\end{aligned}$$

□

B.3.2 Proof of Theorem 4.8

Theorem 4.8. *Suppose there exists $C > 0$ such that $Var(N_i) \leq C$ for all i and let $\mathcal{G}_K = \{\{e_i\}$ s.t.*

$$V_{e_i} \leq K \ \forall i \tag{4.21}$$
$$\Omega_e(Z) < \infty \ \text{ for almost all } Z \tag{4.22}$$
$$R_{e,i} \perp\!\!\!\perp R_{e,j} \ \forall i \neq j, \tag{4.23}$$

$$\mathbb{E}R_{e,i} = 0 \ \forall i \tag{4.24}$$
$$R_{e,i}(Z, N_i) = R_{e,i}(N_i) \ \forall i \ \} \tag{4.25}$$

Then,

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

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

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

For $e_i = k_i^{-1}$, we have that

$$\Omega_e^M(z, n) = \frac{1}{M}\sum_{i=1}^{M} z + n_i \xrightarrow{a.s.} z = \Omega_e^M(z),$$
$$R_i^M = z + n_i - \Omega_e(z, n) \xrightarrow{a.s.} n_i = R_{e,i}(n_i),$$

where the convergences follow from application of Kolmogorov's strong law, using the fact that $\mathrm{Var}(n_i) \leq C$ for all i. Satisfaction of condition 4.21 follows from the fact that $\mathrm{Var}_{z,n_i}(z + n_i) \leq C + \mathrm{Var}(z) \leq K$. Since z is a well-defined random variable, $\Omega_e(z) < \infty$ with probability 1, satisfying condition 4.22. It follows from the mutual independence of n_i and n_j that $R_{e,i}$ and $R_{e,j}$ satisfy condition 4.23. Condition 4.24 follows from the fact that $\mathbb{E}[n_i] = 0$ Condition 4.25 follows from $R_{e,i}$ being constant as a function of z.

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

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

First, we will show that each $e_i \circ k_i$ is affine, i.e. there exist potentially different α_i, β_i such that $e_i = \alpha_i k_i^{-1} + \beta_i$ for each i.

Then we will show that we must have $\alpha_i = \alpha_j$ and $\beta_i = \beta_j$ for all i, j.

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

$$\begin{aligned} e_i \circ k_i(x+y) &= R_{e,i}(x) + \Omega_e(y) \\ &= R_{e,i}(x) + \Omega_e(0) + R_{e,i}(0) + \Omega_e(y) \\ &\quad - (R_{e,i}(0) + \Omega_e(0)) \\ &= e_i \circ k_i(x) + e_i \circ k_i(y) - e_i \circ k_i(0). \end{aligned}$$

It therefore follows that $e_i \circ k_i$ is affine, since if we define

$$\begin{aligned} L(x+y) &= e_i \circ k_i(x+y) - e_i \circ k_i(0) \\ &= (e_i \circ k_i(x) - e_i \circ k_i(0)) \\ &\quad + (e_i \circ k_i(y) - e_i \circ k_i(0)) \\ &= L(x) + L(y), \end{aligned}$$

then L is linear and we can write $e_i \circ k_i(x)$ as the sum of a linear function and a constant:

$$e_i \circ k_i(x) = L(x) + e_i \circ k_i(0).$$

Thus $e_i \circ k_i$ is affine, and we have some (diagonal) matrix α_i and vector β_i such that for any x

$$\begin{aligned} e_i \circ k_i(x) &= \alpha_i x + \beta_i \\ \implies e_i(x) &= \alpha_i k_i^{-1} x + \beta_i. \end{aligned}$$

Next we show that for the set of $\{e_i = \alpha_i k_i^{-1} + \beta_i\}$, it must be the case that each $\alpha_i = \alpha_j$ and $\beta_i = \beta_j$.

Observe that

$$\begin{aligned} \Omega_e^M(z, n) &= \frac{1}{M} \sum_{i=1}^{M} \alpha_i z + \alpha_i n_i + \beta_i \\ &= \left(\frac{1}{M} \sum_{i=1}^{M} \alpha_i \right) z + \frac{1}{M} \sum_{i=1}^{M} \beta_i + \frac{1}{M} \sum_{i=1}^{M} \alpha_i n_i, \\ \mathbb{E}_n[\Omega_e^M(z, n)] &= \left(\frac{1}{M} \sum_{i=1}^{M} \alpha_i \right) z + \frac{1}{M} \sum_{i=1}^{M} \beta_i. \end{aligned}$$

Define

$$\alpha = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \alpha_i,$$
$$\beta = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \beta_i,$$

which exist by the assumption that $\Omega_e^M(z, n)$ converges as $M \to \infty$. Thus

$$\Omega_e(z) = \alpha z + \beta,$$
$$R_{e,i}(z, n_i) = (\alpha_i - \alpha)z + \alpha_i n_i + \beta_i - \beta.$$

Now, suppose that there exist i and j such that such that $\alpha_i \neq \alpha_j$. It follows that

$$R_{e,i}(z, n_i) = (\alpha_i - \alpha)z + \alpha_i n_i + \beta_i - \beta,$$
$$R_{e,j}(z, n_j) = (\alpha_j - \alpha)z + \alpha_j n_j + \beta_j - \beta.$$

There are two cases. If $\alpha_i \neq \alpha$, then $R_{e,i}(z, n_i)$ is not a constant function of z. But if $\alpha_i = \alpha$, then $\alpha_j \neq \alpha$ and so $R_{e,j}(z, n_j)$ is not a constant function of z. This is a contradiction, and so $\alpha_i = \alpha_j$ for all i, j.

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

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

Appendix C

Additional Materials for Chapter 5

This appendix provides proofs for the results in Section 5.6 in which examples of exact transformations are stated.

C.1 Marginalisation of variables (Section 5.6.1)

Theorem 5.17 (Marginalisation of childless variables)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_E)$ 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 $i \in \mathbb{I}_Z$.

Proof. By Lemma 5.13, compositions of exact transformations are exact and so it suffices to prove this result for the marginalisation of a single childless variable. Without loss of generality, let X_1 be the childless variable to be marginalised out.

Let $\mathcal{M}_Y = (\mathcal{S}_Y, \mathcal{I}_Y, P_F)$ 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. $\mathrm{do}(X_1 = x_1, X_2 = x_2) \in \mathcal{I}_X$ is 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 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 distribution of the remaining variables $X_k, k \in \mathbb{I}_X \setminus \{1\}$ that is obtained by *marginalisation* of the childless variable, i. e. $P^i_{\tau(X)}$, is equivalent to the distribution obtained by simply *dropping* the childless variable, which is exactly what the distribution under $\mathcal{M}_Y$ amounts to, i. e. $P_Y^{\omega(\mathrm{do}(i))}$. □

Theorem 5.18 (Marginalisation of non-intervened variables)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_E)$ 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 $i \in \mathbb{I}_Z$.

Proof. By Lemma 5.13, compositions of exact transformations are exact and so it suffices to prove this result for the marginalisation of a single 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 $X_1 = f_1\left(X_{\mathrm{pa}(1)}, E_1\right)$ and X_1 does not appear in the structural equation for any of its ancestors.

Now let $\mathcal{M}_Y = (\mathcal{S}_Y, \mathcal{I}_Y, P_F)$ 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 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 $f_1\left(X_{\mathrm{pa}(1)}, E_1\right)$, yielding $X_i = f_i\left(f_1\left(X_{\mathrm{pa}(1)}, E_1\right),\ X_{\mathrm{pa}(i)},\ E_i\right)$.

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

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

C.2 Micro- to macro-level (Section 5.6.2)

Theorem 5.19 (Micro- to macro-level)**.** *Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_{E,F})$ 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*

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

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

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

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

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

Proof. 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. 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

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

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

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

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 $P_{\tau(X)}^{\mathrm{do}(\varnothing)}$ and $P_Y^{\mathrm{do}(\varnothing)}$ agree. Similarly, the push-forward measure on $\mathcal{Y}$ induced by the intervention $\mathrm{do}(W = w) \in \mathcal{I}_X$ is given by

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

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 $\mathrm{do}(W = w, Z = z)$.

□

C.3 Stationary behaviour of dynamical processes (Section 5.6.3)

Theorem 5.20 (Discrete-time linear dynamical process with identical noise)**.** *Let* $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, P_E)$ *over the variables* $\{X_t^i \,:\, t \in \mathbb{Z},\, i \in \{1,\dots,n\}\}$ *be a linear SEM with*

$$\begin{aligned}
\mathcal{S}_X &= \left\{ X_{t+1}^i = \sum_{j=1}^{n} A_{ij} X_t^j + E_t^i \,:\, i \in \{1,\dots,n\}, t \in \mathbb{Z} \right\}, \\
&\quad \textit{i. e. } X_{t+1} = AX_t + E_t \\
\mathcal{I}_X &= \Big\{ \mathrm{do}(X_t^j = x_j \;\; \forall t \in \mathbb{Z}, \forall j \in J) : x \in \mathbb{R}^{|J|},\; J \subseteq \{1,\dots,n\} \Big\}, \\
E_t &= E \;\; \forall t \in \mathbb{Z} \textit{ where } E \sim P,
\end{aligned}$$

where 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* $i \in \mathcal{I}_X$*:*

$$\begin{aligned}
\tau : \mathcal{X} &\to \mathcal{Y}, \\
(x_t)_{t\in\mathbb{Z}} &\mapsto y = \lim_{t\to\infty} x_t.
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{S}_Y &= \left\{ Y^i = \frac{\sum_{j\neq i} A_{ij} Y^j}{1 - A_{ii}} + \frac{F^i}{1 - A_{ii}} \,:\, i \in \{1,\dots,n\} \right\}, \\
\mathcal{I}_Y &= \Big\{ \mathrm{do}(Y^j = y_j \;\; \forall j \in J) : y \in \mathbb{R}^{|J|},\; J \subseteq \{1,\dots,n\} \Big\}, \\
F &\sim P.
\end{aligned}$$

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

Before proving the above theorem, 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 5.20 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 C.1. *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 C.2. *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 $x, y \in \mathbb{R}^n$,

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

and hence f^* is a contraction mapping. □

Lemma C.3. *Consider the SEM $\mathcal{M}_X$ in Theorem 5.20, 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

$$\begin{aligned} h : \mathbb{R}^n &\to \mathbb{R}^n, \\ v &\mapsto w = Av + E, \end{aligned}$$

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

By Lemma C.1 and Lemma C.2, 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. □

Proof of Theorem 5.20. We begin by defining a mapping between interventions

$$\begin{aligned} \omega : \mathcal{I}_X &\to \mathcal{I}_Y, \\ \mathrm{do}(X_t^j = x_j \ \ \forall t \in \mathbb{Z}, \forall j \in J) &\mapsto \mathrm{do}(Y^j = x_j \ \forall j \in J). \end{aligned}$$

Note that ω is surjective and order-preserving. 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

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

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 C.3 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:

$$\begin{cases} X_*^k = x_k & \text{if } k \leq m, \\ X_*^k = \sum_j A_{kj} X_*^j + E^k & \text{if } m < k \leq n. \end{cases} \tag{C.1}$$

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 C.1 above) that under the push-forward measure $P^i_{\tau(X)} = \tau\left(P_X^{\mathrm{do}(i)}\right)$ 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 distribution of Y as implied by $\mathcal{M}_Y$ under the intervention $\omega(i)$, i. e. $P_Y^{\mathrm{do}(\omega(i))}$. 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. $P^i_{\tau(X)} = P_Y^{\mathrm{do}(\omega(i))}$.

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