

UNIVERSITY OF LIÈGE
Faculty of Applied Sciences
Department of Electrical Engineering & Computer Science

PhD dissertation

CONTRIBUTIONS TO GENERATIVE MODELING
FROM DEEP MODELS TO HYBRID APPROACHES

by ANTOINE WEHENKEL



Advisor: Prof. GILLES LOUPPE
Sometime in 2022

JURY MEMBERS

???, Professor at the Université de Liège (President);

GILLES LOUPPE, Professor at the Université de Liège (Advisor);

???, Professor at the Université de Liège;

???, ???;

???, ???;

???, ???;

ABSTRACT

To be completed.

ACKNOWLEDGMENTS

To be completed.

CONTENTS

1	INTRODUCTION	3
1.1	Research question	5
1.2	Outline and structure	6
1.3	Publications	7
1.4	Additional publications	7
i	BACKGROUND	11
2	PROBABILISTIC MODELLING	15
2.1	Introduction	15
2.1.1	Probabilistic model	16
2.1.2	Learning	17
2.2	Probabilistic graphical modelling	21
2.2.1	The curses of dimensionality	21
2.2.2	Directed graphical models - Bayesian networks	21
2.2.3	Undirected graphical models	21
2.2.4	Inference	21
2.2.5	Learning	21
2.2.6	Discussion	21
2.3	Deep generative modelling	21
2.3.1	The inductive bias of neural networks	21
2.3.2	Energy based models	21
2.3.3	Autoregressive models	21
2.3.4	Normalizing flows	21
2.3.5	Diffusion models	21
2.3.6	Variational auto-encoders	21
2.3.7	Discussion	21
2.4	The multiple definitions of hybrid modelling	21
2.5	Challenges and opportunities	21
2.6	Summary	21
ii	DEEP GENERATIVE MODELLING	23
3	DEEP LATENT VARIABLES GENERATIVE MODELS	25
3.1	Introduction	25
3.2	Variational Auto Encoders	26
3.3	Diffusion Models	27
3.4	Related work	29
3.5	Experiments	29
3.6	Conclusion and future work	32
4	LIMITATIONS OF AFFINE NORMALIZING FLOWS	33
4.1	Introduction	33
4.2	Background	34
4.2.1	Normalizing flows	34

4.2.2	Bayesian networks	34
4.3	Normalizing flows as Bayesian networks	36
4.3.1	Autoregressive conditioners	36
4.3.2	Coupling conditioners	36
4.3.3	Stacking transformation steps	37
4.4	Affine normalizing flows unlock their capacity with 3 transformation steps	38
4.5	Affine normalizing flows are not universal density ap- proximators	39
4.6	Summary	40
5	MONOTONIC NORMALIZING FLOWS	43
5.1	Introduction	43
5.2	Unconstrained monotonic neural networks	44
5.3	UMNN autoregressive models	46
5.3.1	Normalizing flows	46
5.3.2	Autoregressive transformations	46
5.3.3	UMNN autoregressive transformations (UMNN- MAF)	47
5.4	Related work	49
5.5	Experiments	50
5.5.1	2D toy problems	50
5.5.2	Density estimation	51
5.5.3	Variational auto-encoders	53
5.6	Discussion and summary	54
iii	HYBRID GENERATIVE MODELLING	57
6	STRUCTURED DENSITY ESTIMATION WITH NORMAL- IZING FLOWS	61
6.1	Introduction	61
6.2	Background	62
6.3	Normalizing flows as Bayesian networks	64
6.4	Graphical normalizing flow	65
6.4.1	Graphical conditioners	65
6.4.2	Learning the topology	66
6.4.3	Stochastic adjacency matrix	67
6.4.4	Optimization	67
6.5	Experiments	68
6.5.1	On the importance of graph topology	68
6.5.2	Density estimation benchmark	71
6.6	Discussion	73
6.7	Epilogue	76
6.7.1	On the relationship between smoothness and in- dependence	76
7	HYBRID GENERATIVE MODELS	77
7.1	Introduction	77
7.2	Hybrid learning	79

7.2.1	Hybrid generative modelling	80
7.3	Robust hybrid learning	82
7.3.1	Expert augmentation	84
7.4	Experiments	85
7.4.1	Problem description	85
7.4.2	Results	86
7.5	Related work	88
7.5.1	Hybrid modelling	88
7.5.2	Combining hybrid modelling and data augmentation	89
7.5.3	Robust ML and Invariant Learning	89
7.6	Discussion	90
7.7	Conclusion	91
iv	CONCLUSION AND OPPORTUNITIES	93
8	CONCLUSION	95
v	APPENDIX	97
A	NOTATIONS	99
B	REFERENCES	103

Now I'm going to discuss how we would look for a new law. In general, we look for a new law by the following process. First, we guess it (audience laughter), no, don't laugh, that's the truth. Then we compute the consequences of the guess, to see what, if this is right, if this law we guess is right, to see what it would imply and then we compare the computation results to nature or we say compare to experiment or experience, compare it directly with observations to see if it works. If it disagrees with experiment, it's wrong. In that simple statement is the key to science. It doesn't make any difference how beautiful your guess is, it doesn't matter how smart you are who made the guess, or what his name is ... If it disagrees with experiment, it's wrong. That's all there is to it.

RICHARD FEYNMAN

The physical laws, in their observable consequences, have a finite limit of precision.

KURT GÖDEL

All generalizations, with the possible exception of this one, are false.

KURT GÖDEL

INTRODUCTION

Ever since the beginning of their existence, humans have always been curious to understand the world's complexity. This is not only true at the individual level, as we keep building and improving our knowledge over our lifespan, but also at the level of humankind, where knowledge has never really stopped rising since the beginning of civilisation. As an objectification of our curiosity, science aims to ground the construction of this common knowledge on rationality. In particular, the modern scientific method arguably relies on five pillars, as depicted in Figure 1.1, which ensure discoveries are made out of rigour and on the basis of current scientific knowledge. Although the scientific method can answer questions in the context of a specified model of the world as stated by the hypotheses, the more fundamental goal of science is to refine these models by criticising their ability to predict the real world. In this context, this thesis aims to explore and contribute to modern techniques for building or refining models of the world.

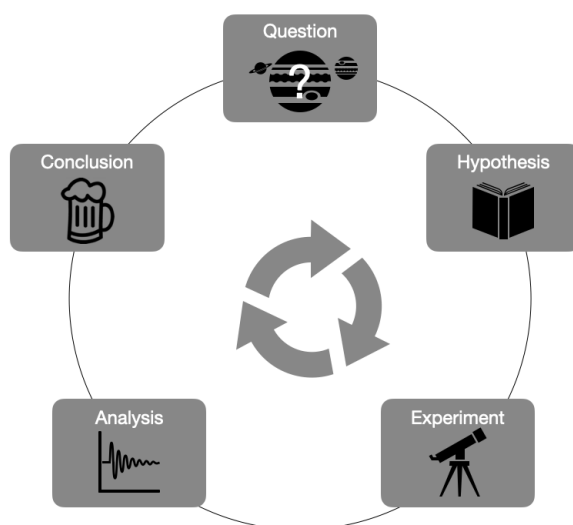


Figure 1.1: Illustration of the scientific inquiry as a simplified 5-steps process. The pictograms sketch a cartoon of the TRAPPIST-1 planetary system discovery made by astronomers from LiGège University in 2016 [Gillon et al., 2017]. Scientists first formulate a *research question* and a set of reasonable *hypotheses*, together they define a class of hypothetical models of the world. In order to answer their question, scientists gather data from *experiments* and analyze them in the context of their modelling assumptions. Checking the consistency of different hypothetical models allows them to *provide an answer* to their initial question up to a certain degree of certainty.

Classically, scientists or engineers use their domain expertise to build incrementally complex models and improve their faithfulness to reality. The scientific method is then applied to validate or reject the new class of models. This strategy has not only led to today’s state of science but also to the uttermost engineering accomplishments of modern times. In engineering, these successes impact our daily life, such as by enabling nuclear electricity - as predicted by special relativity - or allowing one to read this manuscript on a smart tablet - thanks to Maxwell equations’ implications on the design of modern computers. More abstract but arguably as impactful on our vision of the universe, the classical model refinement strategy led to all modern scientific breakthroughs. As an example, these models allow astronomers to observe the furthest human-known objects, thanks to general relativity and gravitational lensing and enable the indirect observations of black holes thanks to gravitational waves theory.

Despite these numerous achievements, machine learning has recently challenged the classical modelling approach. Where humans fall short of finding patterns in a large amount of data and build arbitrarily complex models by themselves, machines can automatically perform these tasks day and night. The shift from hand-crafted to automated modelling happens in a world where the most ambitious scientific experiments generate petabytes (10^{15}) of data per second [CERN]. On the one hand, while the computing capacity continuously increases, the human brain is not better wired to apprehend vast amounts of data than it was when Galileo Galilei set the basis of modern science 400 years ago. On the other hand, deep learning has recently proven its ability to build accurate generative models that outperform the ones designed out of human expertise. This is, for example, true in the context of voice synthesis [Van Den Oord et al., 2016], text-translation [Brown et al., 2020; Devlin et al., 2018], chatbots, or even text-to-image synthesis [??], to cite a few. Under these circumstances, the paradigm shift seems natural, even in a scientific context. Add fancy images of Dalle and flamingo

A paragraph on the progress in deep generative models over the last years and the history behind it, from deep belief networks to diffusion models. Most of the progress made for images and text. A bit for sound as well. These are signals for which we can easily gather tons of data. And this showcase that deep models can learn a useful representation from this large dataset. But this is far from being able to say that automated modelling can challenge classical modelling in all settings. In particular, there are two classes of complication that come jointly with all current approaches: 1) All current generative models require are very difficult to train and require expertise of these models, the tricks required can vary with the data modalities and so there are still huge area for improvement to simplify the broad utilization of these algorithms. 2) These algorithms only rely on architectural inductive bias, which might be sufficient for images, text and sounds that are

well suited for convolutions and attentions but they are not made to use other kind of prior knowledge we may have of the process of interest, this is a piece that miss to overpass the classical approach on many important modelling tasks.

Explicit vs. implicit models

In general, accurate modelling requires accounting for uncertainty. This is important as, by definition, models are simplified representations of reality and so cannot explicitly represent all possible sources of perturbations deterministically. As an example, the precision of any sensor is determined by fitting a model that accounts stochastically for internal and external perturbation that may arise in practical settings such as temperature, humidity or pressure. This is true for small models that are usually simplistic and larger models that build on top of smaller stochastic models. These models define stochastic generative processes that synthesise observations \mathbf{x} conditionally on the models' parameters $\boldsymbol{\theta}$. When it comes to the classical approach, the complexity of the models depends on the granularity of the different effects taken into account. And while small models correspond to simple statistical assumptions and provide access to the likelihood function $p(\mathbf{x}|\boldsymbol{\theta})$, larger models are computer programs for which the likelihood is usually intractable because of the multiple sources of randomness. We use the terms *explicit* and *implicit* to distinguish between models that provide direct access to the likelihood function and those that do not. The same distinction goes for deep generative models that provide access to the likelihood or solely to the generative process.

1.1 RESEARCH QUESTION

Motivated by this paradigm shift and the remaining challenges in deep generative modelling, this thesis contributes to answering the following research question:

How can we improve the construction of faithful generative models with deep learning algorithms?

In pursuing this objective, we explore several directions that study and improve different aspects of deep generative models. In particular, we distinguish between data-driven models that only rely on the inductive bias of the learning algorithm and the availability of representative data and hybrid models that embed solid prior knowledge (e.g. independence assumptions or partial understanding of the underlying physics) into deep generative models. We argue that contributions to both aspects are complementary and will improve the range of application of deep generative models. On the one hand, data-driven algorithms are well suited to modelling tasks for which we do not have strong domain knowledge but access to a lot of representative data. On the other hand, hybrid modelling allows the combination of domain knowledge with

data and is thus better equipped against small or poorly representative datasets.

1.2 OUTLINE AND STRUCTURE

Before diving into the core contributions of this thesis, a review of probabilistic modelling in Chapter 2 naturally follows this introduction. It provides all notions necessary for the appreciation of this thesis for a reader equipped with a technical background. Then, Part ii focuses on improving the expressivity and learning algorithms of implicit and explicit deep generative models. To this end, we explore two independent directions respectively on implicit and explicit generative models. First, Chapter 3 studies the complementarity of two implicit generative models, namely the variational auto-encoders and the diffusion models, and show that diffusion models are a suitable replacement for the simple Gaussian distributions classically used to model a prior over the latent variables of variational auto-encoders. We then start our exploration of explicit generative models in Chapter 4 by showing that affine normalizing flows, a class of explicit models, cannot represent arbitrary generative models. The first part ends in Chapter 5 that addresses the expressivity issues of affine normalizing flows by introducing a new neural architecture to model monotonic functions. Part iv explores hybrid modelling, a class of models that embrace the opportunity to combine pre-existing domain knowledge within deep generative models. In particular, Chapter 6 studies the modelling and implications of independence assumptions in normalizing flows, while Chapter 7 studies the generalisation capabilities of deep generative models equipped with a partial model of the studied process. As a conclusion and summary, Chapter 8 reflects upon the contributions of this thesis and current challenges faced by deep generative models.

1.3 PUBLICATIONS

Setting aside this introduction, an original primer on probabilistic modelling in Chapter 2 and the conclusion, the scientific content of this manuscript is exclusively borrowed to original contributions made to generative modelling over the last 4 years. Each selected contribution sets its own chapter complemented with a prologue and an epilogue. The manuscript builds upon the following list of publications:

List of publications

[Wehenkel and Louppe, 2019] *Unconstrained monotonic neural networks*, **Wehenkel Antoine** and Louppe Gilles.

Advances in neural information processing systems, 2019.

→ Chapter 5.

[Wehenkel and Louppe, 2020] *You say Normalizing Flows I see Bayesian Networks*, **Wehenkel Antoine** and Louppe Gilles.

International Conference on Machine Learning, Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models, 2020.

→ Chapter 4.

[Wehenkel and Louppe, 2021b] *Graphical Normalizing Flows*, **Wehenkel Antoine** and Louppe Gilles.

International Conference on Artificial Intelligence and Statistics, 2021.

→ Chapter 6.

[Wehenkel and Louppe, 2021a] *Diffusion Priors In Variational Autoencoders*, **Wehenkel Antoine** and Louppe Gilles.

International Conference on Machine Learning, Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models, 2021.

→ Chapter 3.

[Wehenkel et al., 2022] *Robust Hybrid Learning With Expert Augmentation*, **Wehenkel Antoine**, Behrmann Jens, Hsu Hsiang, Sapiro Guillermo, Louppe Gilles, and Jacobsen Jörn-Henrik.

In preparation, arXiv preprint arXiv:2202.03881.

→ Chapter 7.

1.4 ADDITIONAL PUBLICATIONS

Along the pursuit of my PhD degree, I had the chance to take part in fruitful collaborations not directly related to the scope of this the-

sis. The following list of co-authored publications stemmed from these collaborations:

Extra contributions

[Pesah et al., 2018] *Recurrent Machines For Likelihood-free Inference*, Arthur Pesah, **Wehenkel Antoine** and Louppe Gilles.

Advances in neural information processing systems, MetaLearn Workshop, 2018.

[Wehenkel et al., 2020] *Parameter Estimation Of Three-phase Untransposed Short Transmission Lines From Synchrophasor Measurements*, **Wehenkel Antoine**, Mukhopadhyay Arpan, Le Boudec Jean-Yves, Paolone Mario.

IEEE Transactions on Instrumentation and Measurement, 2020.

[Vecoven et al., 2020] *Introducing Neuromodulation In Deep Neural Networks To Learn Adaptive Behaviours*, Vecoven Nicolas, Ernst Damien, **Wehenkel Antoine**, Drion Guillaume.

PloS one 15 (1), e0227922.

[Vandegar et al., 2021] *Neural Empirical Bayes: Source Distribution Estimation and its Applications to Simulation-Based Inference*, Vandegar Maxime, Kagan Michael, **Wehenkel Antoine** and Louppe Gilles.

International Conference on Artificial Intelligence and Statistics, 2021.

[Delaunoy et al., 2020] *Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization*, Delaunoy Arnaud, **Wehenkel Antoine**, Hinderer Tanja, Nisanke Samaya, Weniger Christoph, Williamson Andrew R and Louppe Gilles.

Advances in neural information processing systems, ML4Science Workshop, 2020.

[Hermans et al., 2021] *Averting A Crisis In Simulation-based Inference*, Hermans Joeri, Delaunoy Arnaud, Rozet François, **Wehenkel Antoine** and Louppe Gilles.

In preparation, arXiv preprint arXiv:2110.06581.

[Dumas et al., 2021] *A Probabilistic Forecast-driven Strategy For A Risk-aware Participation In The Capacity Firming Market*, Dumas Jonathan, Cointe Colin, **Wehenkel Antoine**, Sutura Antonio, Fettweis Xavier and Cornélusse

Bertrand.

IEEE Transactions on Sustainable Energy, 2021.

[Dumas et al., 2022] *A Deep Generative Model For Probabilistic Energy Forecasting In Power Systems: Normalizing Flows*, Dumas Jonathan, **Wehenkel Antoine**, Lanaspeze Damien, Cornélusse Bertrand and Sutera Antonio. Applied Energy, 2022.

Part I

BACKGROUND

You can't do inference without making assumption.

THE UNKNOWN BAYESIAN

Outline

he previous chapter established that modelling has played a major role in the progress of science and engineering. In this chapter, we argue that accounting for uncertainty is crucial for many, if not all, modelling tasks.

This chapter is on *probabilistic modelling* and shall answer the following questions:

- (a) What is a probabilistic model?
- (b) How do we build a probabilistic model?
- (c) How can we use a probabilistic model?
- (d) What are the technical challenges considered in this thesis?

2.1 INTRODUCTION

The invention of computers enabled the automatization of many tasks historically accomplished by humans. One key ingredient of this revolution is the ability to let computer reason - to make an informed judgment based on logical arguments and observations. The set of hypotheses on which this logic builds is refereed as a *model*. Both humans and machines relies on reasoning to perform tasks. As an example, let us consider that we want to chose a bottle of wine at the restaurant. We first need some hypotheses, e.g., - What kind of wine do people like at the table? Red or white? Strong or delicate? - What is the budget? - What are we going to eat tonight? - What is the winelist? etc. Provided this model, we can make an informed choice: remove the wines that are too expensive or would unpleased someone at the table, and, *wine not* pick one of the remaining that goes well with tonight's dinner?! More seriously, a scientist needs a model of the earth and its atmosphere to provide an explanation to the climate change. Youtube considers your videos historic and a model that relates watched videos to recommendations to suggest videos. In these contexts, and in many others, the model plays a central part. We want good models, ones that lead to the right judgment. One that leads to the right wine, one that fits the climate over the last centuries, one that makes you click on one of the suggested videos.

Of course, the definition of a good model depends on its end application. However certain classes of models are usually strictly better than

others. In particular, a model that embeds notions of uncertainty is more powerful than the fully deterministic version. It is because deterministic models do not faithfully represent phenomena that exhibit some forms of randomness which is inherent to our life. Although it is unknown whether the laws that rule our reality are deterministic or stochastic, modelling requires to make simplifying assumptions inducing randomness. These assumptions often handle hidden factors that we cannot directly observe, noisy measurements or simplifying approximations. They are inherent to the modeling of complex phenomenon, hence is the induced stochasticity.

We thus need a language to express this stochasticity. The language of *probability* is the one to make statements about uncertain events. It allows to contrast between what is possible versus what is plausible. The distinction is important as it allows us to eventually make deterministic decisions by neglecting the most unlikely events and to focus on events that are plausible. An old bottle has higher chance to be corked than a recent one, on the opposite it might also taste better. We can use probability to express this fact and select a bottle that is to our taste with high probability.

2.1.1 Probabilistic model

A probabilistic model is a model that describes a phenomenon of interest in probabilistic terms. Practically, it defines a probability distribution between the set of variables considered useful to describe the phenomenon (e.g., x, y, z). The probability can be the joint between all variables (e.g., $p(x, y, z)$) when it models the joint observations. It can also be a conditional distribution (e.g., $p(x|y, z)$) if we can consider some of the variables known when using this model. For simplicity but with no loss of generality, we mostly limit our discussion to the former case except when mentioned otherwise. For discrete variables the mathematical objects $p(x, y, z)$ is a probability, and it is a density for continuous variables. In the following of this chapter, we will often use the symbol p with no additional mention of the variables type (discrete vs continuous) to generalize the discussion to both types.

One goal of building a probabilistic model, arguably the main one, is to perform *inference*. That is to answer questions in the context of the models. These questions come in different flavours. What is the most likely value of y if we are to observe x ? What is the conditional distribution of y in this case? Do we want to evaluate the value $p(y|x)$ or just sample from it? To these purposes, the probabilistic models may have to handle different type of queries: *marginalization*, *conditioning*, *sampling*, and *probability evaluation*.

Certain representations are appropriate for a subset of queries and not for others. As an example, we can represent the discrete distribution between x, y, z with a 3D table where each entry store the corresponding

joint probability. The evaluation of the joint probability is very efficient with this representation. However, evaluating a conditional distribution requires to go over each entry that corresponds to the conditioning value and to re-normalize the probabilities by their sum. Sampling becomes very inefficient as the number of entries in the table grow. And this number grows exponentially with the number of dimensions. Fortunately, there exist other representations that have different advantages and drawbacks. We will review those important to this thesis later in the chapter.

The next two sections provide an overview of two main classes of probabilistic models, namely **Graphical models** and **Deep generative models**. As the name says, the former class aims for a graphical representation of the distribution. It allows to easily handle and understand modelling assumptions such as independence hypotheses. The latter focuses on models whose internal representations use deep neural networks. These models are usually well suited for sampling. We will discuss the particularities of each class of models and will provide a thorough description of the main models. We will also details some algorithms to perform the different queries aforementioned on these models.

In this manuscript, we will argue at multiple occasions that we shall not make a rigid distinction between graphical models and deep generative models as they are just different representations of the same mathematical object. Some deep generative models have a direct correspondance within the graphical family which enables abstract reasoning independent from the neural networks architectures. However, for the sake of clarity we will first introduce probabilistic graphical models. And then borrow the newly introduced notations and concepts to describe several deep generative modelling algorithms.

2.1.2 *Learning*

Before jumping into the description of different classes of models we can keep our discussion general longer. Until now, we have implicitly assumed the model was given to us. However, this is not realistic in many interesting settings. For example, how can I create an accurate model of my winetaste? Certainly I cannot simply come by myself with a list ranking all wines in the world. It might be too expensive to make and would not help me finishing this dissertation. However, I could answer a long list of questions and then figure out a summary of my tastes with respect to the principal characteristics of wine. The task of summarizing my answers is what is called learning - from observations (my answers) we build a compressed representation. Afterwards we can use this representation to perform informed guess, this representation is a model, a simplified representation, of my winetaste. Learning is thus the task of instantiating a model from data.

In practice, we do not perform *learning* without additional assumptions. Instead, we define a set of models among which we believe at least one would be a good representation of the phenomenon of interest. If we are Bayesians we even add a probability that summarizes whether or not we believe the model is good. We will come back to this later. For now let us assume we do not have such a priori.

The class of possible models can be finite, e.g. the class contains two models - one for people who like red wine and white wine; the other for people who only like red wine. Compressing my taste into one of these models would go with a big loss of information but might already be useful in certain settings. The class of models can be infinite, e.g. if parameterized by real values. For example, we can summarize winetaste by attributing an affinity score to each of the main features of wine.

Maximum likelihood estimation. The learning strategy can be seen as a set of rules that leads to the instantiation of a model from data. When there is only a finite number of models there exists a simple strategy. We test the predictive performance of each model and select the one that is the most consistent with our data. If the models describe the phenomenon with discrete events then we maximize the probability. If it considers a continuous set of events we maximize the density. In the case where one of the model is *correct* - it is the one that generates the data - this selection algorithm will eventually select the right model as the number of independently and identically distributed (iid) data points tends to ∞ . This selection technique thus is said *consistent*.

We can use a similar approach when the models are parameterized by a real vector θ . In this case, our goal is to estimate a good value for θ . One approach, denoted maximum likelihood estimation (MLE), is to select the model's parameter θ that maximizes the joint distribution (density or probability) of the data provided its value. This quantity is called the likelihood function of the parameter θ , denoted $\mathcal{L}(\theta) \triangleq p(data|\theta)$. Hence the MLE estimator is formally defined as

$$\theta_{\text{MLE}} = \arg \max_{\theta} p(data|\theta). \quad (2.1)$$

In the presence of iid data this estimator is consistent - provided a class of models that contains the 'true' generative process, it eventually recovers the 'true' model as the number of points tends to ∞ . Formally, the consistency property is a convergence in probability of the estimator to the exact value and requires additional assumptions that ensures the model is identifiable and the likelihood function is well behaved (compactness and continuity with respect to θ , and dominance).

The consistency of the MLE learning method makes it appealing. However, we must consider the main assumption very carefully! While ensuring that the model class contains the true generative process is ok in artificial settings. For real data, this assumption is almost a metaphysical question. The law of large numbers saves us if we only look

at things in average and are provided with a lot of data but is not applicable to all modeling tasks. Sometimes, we know this assumption does not hold but we would still like to learn a good model, the MLE principle does not say much in this context. Even when we can be sure that the model class contains the right model (e.g. we consider a parametric universal density approximator), we know nothing about the convergence speed. And this is without mentioning the optimization procedure when there is no closed form solution to Equation (2.1).

Learning as inference. The strict delimitation between possible and impossible models is another limitation of the MLE approach. The model is either part of the class of models and it is as likely as others to be the right one, or it is certainly irrelevant and it is ignored. We can avoid the strict separation between possible and ignored models by reformulating learning. Instead of interpreting learning as model selection, we see learning as a special inference task. We are thus back in a setting where the model is provided, however it is fine as we can consider the model as a very generic description of the phenomenon. As an example, the model can be a parametric function, exactly as when we define a class of models in the MLE approach. The distinction between learning as inference and MLE lies in the interpretation we make of the parameters. In the former we consider the parameter θ as part of the model rather than defining the class of models. While learning via MLE is usually associated to a frequentist interpretation of probability, learning as inference is seen as Bayesian.

Let us consider the case where we want to learn a model that represents the conditional distribution $\arg \max_y p(y|\mathbf{x})$ that I like a wine with features $\mathbf{x} \in \mathbb{R}^d$, where $y \in \mathbb{R}$ says if I like (∞) or hate ($-\infty$) a wine. We could consider a neural network that is parameterized by θ and represents a parametric density function conditioned on a input x : $f(\cdot, \mathbf{x}; \theta) : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}^+$. Learning aims at summarizing the information in the data in order to perform a task of interest, here to predict $p(y|\mathbf{x}, data)$. If we state that f represents the conditional distribution $p(y|\mathbf{x}, \theta)$, this gives

$$p(y|\mathbf{x}, data) = \int p(y|\mathbf{x}, \theta) p(\theta|data) d\theta \quad (2.2)$$

$$= \int f(y, \mathbf{x}; \theta) p(\theta|data) d\theta. \quad (2.3)$$

We notice the data are only used through the posterior distribution of the parameters θ provided the data. The posterior summarizes our belief about which models describe the best the phenomenon of interest in the lights of the data observed and our initial belief. The goal of learning is thus to compute $p(\theta|data)$, which is an inference task. **how shall I mention functional form, I am not sure about the right vocabulary. We replace $p(\theta)$ by $p(f)$ - mention kernel methods that naturally work**

in a function space. very general hypothesis space might be the class of smooth functions \mathbb{C}_∞ from \mathbb{R}^d to \mathbb{R} .

In practice it is often cumbersome to keep track of the complete posterior distribution. We can avoid this by selecting the maximum a posteriori (MAP) sub-model. In the case of a parametric model this means freezing the parameter θ to their most plausible value $\theta_{MAP} = \arg \max_{\theta} p(\theta|data)$. Seeing learning as inference strictly generalizes the MLE principle to settings where the prior knowledge is more subtle than possible or impossible. When we consider a non-informative prior the method is strictly equivalent to the MLE. And when we do have a good prior, this strategy will naturally reduce the importance of bad model instantiation and mostly rely on the instantiations that were plausible and describes well the data. Moreover, this approach is also consistent and eventually select the ‘right’ instantiation.

Learning as inference is sometimes criticised by practitioners who do not like the concept of attributing plausibility to different instantiation of the model. However, this criticism is empty from my point of view. The superiority of this approach relies on the obligation to explicit the modeling assumptions and the plausibility associated to each learnable component of the model. It forces us to acknowledge that learning a model is a subjective task. As an example, the Occam’s razzor says that among potential explanations we should always favor the simplest. The bayesian approach naturally handles the occam’s razzor by attributing higher plausibility to simpler model instantiations. This is not true for the MLE approach that requires to add an ad-hoc rule or regularization objective to the optimization problem.

Machine learning = probabilistic modeling

Bayesian vs frequentist interpretation

Two views oppose each others on the interpretation of probabilities. In the above discussion we brought probabilities as a formal expression of our uncertainty about the truthness of facts. This is the *Bayesian* interpretation of probabilities. In this context, we start with a *prior* belief about the truthness of a set of possible facts and use the Bayes’ rule to update the belief when evidence comes, hence the reference to Sir Bayes. In this context, the prior is part of the model and impacts its quality. The main drawback of the Bayesian interpretation is thus that it may be hard to define the prior well. The second interpretation views a probability as a frequency of events and is refereed as *frequentist*. With this interpretation, a probability does not quantify uncertainty, instead it expresses intrinsic randomness. Frequentistm has obviously no concept of prior belief which has pros and cons. In general, there is no inteprepretation better than the other. However, the Bayesian inteprepretation is arguably the most common in machine learning and is the one we will mostly implicitly use in our discussions. At the same time, we will

also heavily use the term likelihood which is usually associated to the frequentist view.

2.2 PROBABILISTIC GRAPHICAL MODELLING

2.2.1 *The curses of dimensionality*

2.2.2 *Directed graphical models - Bayesian networks*

2.2.3 *Undirected graphical models*

2.2.4 *Inference*

2.2.5 *Learning*

2.2.6 *Discussion*

2.3 DEEP GENERATIVE MODELLING

2.3.1 *The inductive bias of neural networks*

2.3.2 *Energy based models*

2.3.3 *Autoregressive models*

2.3.4 *Normalizing flows*

2.3.5 *Diffusion models*

2.3.6 *Variational auto-encoders*

2.3.7 *Discussion*

2.4 THE MULTIPLE DEFINITIONS OF HYBRID MODELLING

2.5 CHALLENGES AND OPPORTUNITIES

2.6 SUMMARY

Add a schematic view of how different models are related to each others and where we made the connections/contributions

Part II

DEEP GENERATIVE MODELLING

DEEP LATENT VARIABLES GENERATIVE MODELS

OUTLINE

Among likelihood-based approaches for deep generative modelling, variational autoencoders (VAEs) offer scalable amortized posterior inference and fast sampling. However, VAEs are also more and more outperformed by competing models such as normalizing flows (NFs), deep-energy models, or the new denoising diffusion probabilistic models (DDPMs). In this preliminary work, we improve VAEs by demonstrating how DDPMs can be used for modelling the prior distribution of the latent variables. The diffusion prior model improves upon Gaussian priors of classical VAEs and is competitive with NF-based priors. Finally, we hypothesize that hierarchical VAEs could similarly benefit from the enhanced capacity of diffusion priors.

3.1 INTRODUCTION

Over the last few years, the interest of the deep learning community for generative modelling has increased steadily. Among the likelihood-based approaches for deep generative modelling, variational autoencoders [Kingma and Welling, 2013, VAEs] stand as one of the most popular, although competing approaches now demonstrate better performance. In particular, Ho et al. [2020]; Nichol and Dhariwal [2021]; Dhariwal and Nichol [2021] recently showed that denoising diffusion probabilistic models (DDPMs) are competitive deep generative models, obtaining samples quality similar to those of the best implicit deep generative models such as ProgressiveGAN [Karras et al., 2017] and StyleGAN [Karras et al., 2019]. Similarly to VAEs, DDPMs train on a variational bound and may be interpreted under the encoding-decoding framework.

In the original formulation of VAEs, the prior and the posterior distributions over the latent variables are assumed to be both Gaussian. However, these assumptions are often incompatible and are thus limiting the performance of VAEs for complex modelling tasks [Tomczak and Welling, 2018; Chen et al., 2018a]. A natural solution to this problem is to parameterize the prior, sometimes also the posterior, with more expressive distributions. In this preliminary work, we improve VAEs by demonstrating how DDPMs can be used for modelling the prior distribution of the latent variables. In addition to boosting DDPM with the compression properties of VAEs, combining the two models should

eventually lead to greater generative performance by enabling complex generative modelling even with simple decoder architecture. Finally, working in the latent space should eventually reduce the computational burden associated with diffusion generative models.

3.2 VARIATIONAL AUTO ENCODERS

We want to learn a generative model of an unknown distribution $p(\mathbf{x})$ given a dataset $X \in \mathbb{R}^{N \times d}$ of N i.i.d observations \mathbf{x} sampled from this unknown distribution. The original VAE postulates a two-step generative process in which some unobserved variables $\mathbf{z} \in \mathbb{R}^h$ are first sampled from a prior distribution $p(\mathbf{z})$ and then observations \mathbf{x} are generated from a conditional distribution $p_{\theta}(\mathbf{x}|\mathbf{z})$. The generative process can be expressed mathematically as

$$\mathbf{z} \sim p(\mathbf{z}) \quad \text{and} \quad \mathbf{x} \sim p_{\theta}(\mathbf{x}|\mathbf{z}). \quad (3.1)$$

The prior $p(\mathbf{z})$ is chosen Gaussian while the likelihood $p_{\theta}(\mathbf{x}|\mathbf{z})$ is modeled with a neural network. The likelihood model decodes latent variables into observations and is thus usually refereed as the decoder in the literature. In its original formulation, the likelihood is parameterized with a multivariate Gaussian $\mathcal{N}(\mu_{\theta}(\mathbf{z}), \text{diag}(\sigma_{\theta}^2(\mathbf{z})))$ when the observations are continuous, and a categorical distribution when they are discrete.

Training the generative model is achieved by finding the parameters θ of the decoder that maximize the sum of the marginal likelihoods of individual points

$$p_{\theta}(X) = \sum_{\mathbf{x} \in X} \log \int p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}.$$

These integrals are intractable but the introduction of an encoder network that approximates the posterior distribution $q_{\phi}(\mathbf{z}|\mathbf{x})$ allows maximizing the associated evidence lower bound

$$\text{ELBO} := \mathbb{E}_q \left[\log \frac{p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_{\psi}(\mathbf{z}|\mathbf{x})} \right] \quad (3.2)$$

$$= \log p_{\theta}(\mathbf{x}) - \mathbb{KL}[q_{\psi}(\mathbf{z}|\mathbf{x})||p_{\theta}(\mathbf{z}|\mathbf{x})] \quad (3.3)$$

$$\leq \log p_{\theta}(\mathbf{x}). \quad (3.4)$$

The ELBO becomes tighter as the approximate posterior $q_{\psi}(\mathbf{z}|\mathbf{x})$ gets closer to the true posterior. Learning the generative model is finally performed by jointly optimizing the parameters θ of the decoder and ϕ of the approximate posterior via stochastic gradient ascent. In the original VAE, the encoder models the approximate posterior as a conditional multivariate Gaussian distribution $\mathcal{N}(\mu_{\phi}(\mathbf{x}), \text{diag}(\sigma_{\phi}^2(\mathbf{x})))$.

The ELBO loss presents two antagonistic goals to the encoder. It should be able to both encodes the data accurately while being as close

as possible to the prior distribution. Consequently, the Gaussian assumptions made on both the prior and the posterior distributions are often incompatible and limit the generative performance. A possible solution consists in learning a prior distribution that is compatible with the learned posteriors. For example, [Habibian et al. \[2019\]](#) and [Chen et al. \[2017\]](#) respectively showed that autoregressive models and normalizing flows [[Rezende and Mohamed, 2015](#), NFs] greatly improve the performance of VAEs when used as prior distributions. In the following we present how denoising diffusion probabilistic models can be used to improve the performance of classical VAEs.

3.3 DIFFUSION MODELS

Inspired by non-equilibrium statistical physics, [Sohl-Dickstein et al. \[2015\]](#) originally introduced DDPMs while [Ho et al. \[2020\]](#) demonstrated only more recently how to train these models for image synthesis, achieving results close to the state-of-the-art on this task. DDPMs formulate generative modelling as the reverse operation of diffusion, a physical process which progressively destroys information. Formally, the reverse process is a latent variable model of the form

$$p_\phi(\mathbf{x}_0) := \int p_\phi(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T},$$

where $\mathbf{x}_0 := \mathbf{x}$ denotes the observations and $\mathbf{x}_1, \dots, \mathbf{x}_T$ denote latent variables of the same dimensionality as \mathbf{x}_0 . The joint distribution $p_\phi(\mathbf{x}_{0:T})$ is modelled as a first order Markov chain with Gaussian transitions, that is

$$p_\phi(\mathbf{x}_{0:T}) := p_\phi(\mathbf{x}_T) \prod_{t=1}^T p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad (3.5)$$

$$p_\phi(\mathbf{x}_T) := \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3.6)$$

$$p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}). \quad (3.7)$$

Similar to VAEs, the reverse Markov chain is trained on an ELBO. However, the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is fixed to a diffusion process that is also a first order Markov chain with Gaussian transitions,

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (3.8)$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (3.9)$$

where β_1, \dots, β_T are the variance schedule that is either fixed as training hyper-parameters or learned. The ELBO is then given by

$$\text{ELBO} := \mathbb{E}_q \left[\log \frac{p_\phi(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \leq \log p_\phi(\mathbf{x}_0). \quad (3.10)$$

Provided that the variance schedule β_t is small and that the number of timesteps T is large enough, the Gaussian assumptions on the generative process p_ϕ are reasonable. [Ho et al. \[2020\]](#) take advantage of the Gaussian transitions to express the ELBO as

$$\mathbb{E}_q \left[\mathbb{KL} [q(\mathbf{x}_T | \mathbf{x}_0) || p(\mathbf{x}_T)] - \log p_\phi(\mathbf{x}_0 | \mathbf{x}_1) + \sum_{t=2}^T \mathbb{KL} [q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p_\phi(\mathbf{x}_{t-1} | \mathbf{x}_t)] \right]. \quad (3.11)$$

The inner sum in Equation (3.11) is made of comparisons between the Gaussian generative transitions $p_\phi(\mathbf{x}_{t-1} | \mathbf{x}_t)$ and the conditional forward posterior $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ which can also be expressed in closed form as Gaussians $\mathcal{N}(\tilde{\mu}_t(\mathbf{x}_0, \mathbf{x}_t), \tilde{\beta}_t \mathbf{I})$, where $\tilde{\beta}_t$ are functions of the variance schedule. The KL can thus be calculated with closed form expressions which reduces the variance of the final expression. In addition, [Ho et al. \[2020\]](#) empirically demonstrate that it is sufficient to take optimization steps on uniformly sampled terms of the sum instead of computing it completely. The final objective closely resembles denoising score matching over multiple noise levels [[Song and Ermon, 2019](#)]. These observations combined with additional simplifications leads to a simplified loss

$$L_{\text{DDPM}}(\mathbf{x}_0; \phi) := \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_t} \left[\frac{1}{2\sigma_t^2} \|\boldsymbol{\mu}_\phi(\mathbf{x}_t, t) - \tilde{\mu}_t(\mathbf{x}_0, \mathbf{x}_t)\|^2 \right], \quad (3.12)$$

where $\tilde{\mu}_t(\mathbf{x}_0, \mathbf{x}_t)$ is the mean of $q(\mathbf{x}_{t-1} | \mathbf{x}_0, \mathbf{x}_t)$, the forward diffusion posterior conditioned on the observation \mathbf{x}_0 .

We now introduce our contribution which consists in using a DDPM for modelling the prior distribution in VAEs. We formulate the generative model as

$$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (3.13)$$

$$\mathbf{z}_{t-1|t} \sim p_\phi(\mathbf{z}_{t-1} | \mathbf{z}_t) \quad \forall t \in [T, \dots, 1] \quad (3.14)$$

$$\mathbf{x} \sim p_\theta(\mathbf{x} | \mathbf{z}_0), \quad (3.15)$$

where ϕ denotes the parameters of the reverse diffusion model encoding the prior distribution. Equations (3.13) and (3.14) implicitly define a prior distribution over the usual latent variables \mathbf{z}_0 which is modelled with a reverse diffusion process.

Unfortunately, we cannot train a VAE with a diffusion prior directly on the ELBO as expressed in Equation (3.2) as $p_\phi(\mathbf{z}_0)$ cannot be evaluated. However, Equation (3.2) can be further developed as

$$\mathbb{E}_{q_\psi} [\log p_\theta(\mathbf{x} | \mathbf{z}_0)] - \mathbb{E}_{q_\psi} [\log q(\mathbf{z}_0 | \mathbf{x})] + \mathbb{E}_{q_\psi} [\log p_\phi(\mathbf{z}_0)] \quad (3.16)$$

in which a lower bound on the last term can be expressed by Equation (3.10). This finally leads to the following expression

$$\mathbb{E}_{q_\psi} \left[\log p_\theta(\mathbf{x} | \mathbf{z}_0) - \log q(\mathbf{z}_0 | \mathbf{x}) + \mathbb{E}_q \left[\log \frac{p_\phi(\mathbf{z}_{0:T})}{q(\mathbf{z}_{1:T} | \mathbf{z}_0)} \right] \right] \quad (3.17)$$

$$\leq \mathbb{E}_{q_\psi} [\log p_\theta(\mathbf{x}|\mathbf{z}_0) - \log q(\mathbf{z}_0|\mathbf{x}) + \log p_\phi(\mathbf{z}_0)] \quad (3.18)$$

$$\leq \log p_\theta(\mathbf{x}), \quad (3.19)$$

which is a valid ELBO. Finally, the diffusion prior p_ϕ is trained jointly with the approximate posterior q_ψ and the likelihood models p_θ which are optimized as in a classical VAE. This leads to the following loss function:

$$\mathcal{L}(\mathbf{x}; \phi, \theta, \psi) := \mathbb{E}_{q_\psi} \left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})}{q_\psi(\mathbf{z}|\mathbf{x})} \right] + \mathbb{E}_{q_\psi} [L_{\text{DDPM}}(\mathbf{z}_0; \phi)]. \quad (3.20)$$

3.4 RELATED WORK

Various approaches have been proposed to improve the modelling capacity and the training of VAEs. As a first example, some state-of-the-art deep generative models based on VAEs model the posterior with normalizing flows or autoregressive models [Kingma et al., 2016; Vahdat and Kautz, 2020]. Autoregressive models are also often used as a replacement of the original likelihood parameterization, which assumes conditional independencies that are often unrealistic [Oord et al., 2016]. Another popular improvement made to the original VAE is the embedding of structure in the latent variables. In particular, hierarchical VAEs [Sønderby et al., 2016; Kingma et al., 2016] combined with careful training demonstrate impressive results on generative modelling for images [Vahdat and Kautz, 2020].

Vahdat et al. [2021] concurrently proposed to use diffusion for modelling the prior distributions of VAEs. They obtain state-of-the-art results on image synthesis by combining continuous diffusion models and VAEs. Not as close to our work but related, Chen et al. [2017] proposed to learn the prior as a solution to the mismatch between the approximate and the true posteriors. They model the prior with an autoregressive flow, which also closely relates to modelling the posterior distribution with an inverse autoregressive flow [Kingma et al., 2016]. Tomczak and Welling [2018] takes inspiration from the aggregated posterior $\frac{1}{N} \sum_{i=1}^N q_\psi(z|x)$ [Hoffman and Johnson, 2016; Makhzani et al., 2015] to introduce the VampPrior defined as a mixture of learned pseudo-inputs. An orthogonal line of work suggests that the mismatch between the approximate posterior and the exact posterior can be reduced by over-weighting the terms related to the prior and to the approximate posterior in the ELBO [Higgins et al., 2016; Chen et al., 2018a].

3.5 EXPERIMENTS

We now compare VAEs for different choices of priors, including the original Gaussian prior, an NF prior, and the proposed diffusion prior. All models share a same backbone encoder-decoder architecture inspired from DCGAN [Radford et al., 2015]. Optimization is performed with

Table 3.1: FID scores for different models for prior modelling in VAEs and for different latent size. *Diffusion priors outperform classical VAE on CelebA but are slightly worse than NFs. FID scores do not reveal the superiority of any method on CIFAR10.*

<i>Dataset</i>	CelebA			CIFAR10		
<i>Latent Size</i>	40	100	200	40	100	200
<i>Gaussian</i>	154.3	149.4	139.1	176.0	126.2	123.9
<i>NF</i>	72.9	59.49	54.7	167.6	129.1	129.6
<i>Diffusion</i>	114.8	67.95	88.3	177.9	160.5	153.1

Adam for 250 epochs with a learning rate set to 0.0005. After each epoch, the models are evaluated on a validation set used to select the best one for each training setting. We compare the models on the CIFAR10 and CelebA datasets for 3 different latent variables dimensionality (40, 100, 200). The NF used in our experiments is a 3-step autoregressive affine flow with simple MLP backbones similar to the one used to model the transition function of DDPM.

Table 3.1 presents the FID scores for the different models. We first notice the large scores reached by all models on the CIFAR10 dataset. This can be explained by the simplicity of the models trained in our experiments. We believe these scores could be greatly improved by using a more sophisticated likelihood model such as a PixelCNN [Oord et al., 2016]. Although FID scores suggest that the Gaussian prior outperforms the diffusion prior in terms of generative performance, the visual inspection of Figure 3.1 shows that the diffusion prior results in samples slightly more realistic than those of the classical VAE. The best FID score is achieved by the NF prior, although its samples do not seem to reflect this superiority. In this case, we believe the FID scores are not entirely informative about the quality of the images synthesized by the models and should be interpreted with a grain of salt. Although learned priors seem to improve generative performance on CIFAR10, additional work is needed to reach results that would justify using a diffusion prior for this dataset.

On CelebA however, we observe in Table 3.1 that diffusion priors outperform the Gaussian prior. This is in line with the visual inspection of Figure 3.2a and Figure 3.2c. As for CIFAR10, the NF prior outperforms the Gaussian and diffusion priors in terms of FID scores, although the visual inspection of the corresponding samples in Figure 3.2b does not reveal a much better quality of images when compared to those resulting from the diffusion prior. We conclude from these observations that diffusion priors offer an interesting alternative to NFs for modelling the prior in a VAE.

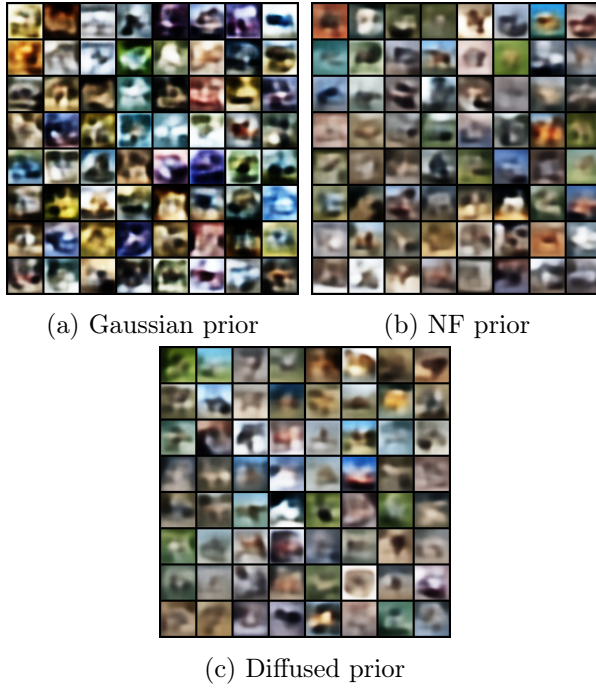


Figure 3.1: Samples generated by a VAE trained on CIFAR10 for three different prior models. *The diffusion prior leads to slightly better sampling quality than the Gaussian distribution and similar to the NF prior.*

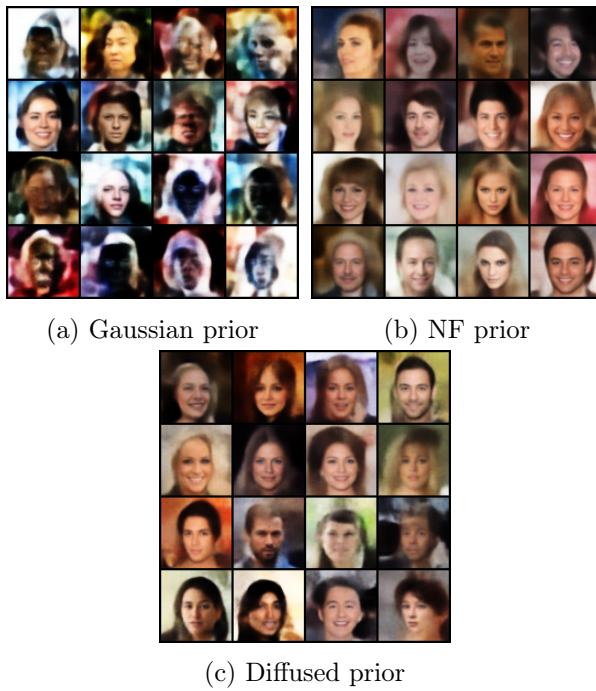


Figure 3.2: Samples generated by a VAE trained on CelebA for three different prior models. *The diffusion prior leads to better sampling quality than the Gaussian distribution and similar to the NF prior.*

3.6 CONCLUSION AND FUTURE WORK

This preliminary work presents how denoising diffusion probabilistic models can be used as a new class of learnable priors for VAEs. As a notable contribution, we empirically demonstrate that optimizing implicitly a prior on an ELBO can be performed jointly to training the encoder and the decoder of the VAE. In addition, our results suggest DDPM performs on par with NFs for modelling prior distribution.

A large spectrum of future research directions could benefit from the basic idea expressed in this preliminary work. As an example, recent advances in diffusion models such as the continuous formulation [Song et al., 2020] or improvement to the training procedure of DDPM [Nichol and Dhariwal, 2021] could be implemented in the prior model. Similarly, many improvements could be made to the architectures used for the VAE and to the training procedure. In particular, image synthesis with hierarchical VAEs which organizes the latent variables into multiple scales images could reveal the full potential of diffusion priors. This would indeed combine the structural knowledge embed by such type of VAEs with the impressive performance of DDPM for modelling distributions over images. Finally, diffusion does not constrain the neural networks architectures and so enables the embedding of a larger choice of inductive biases in the prior distribution compared to autoregressive models and NFs.

OUTLINE

Normalizing flows have emerged as an important family of deep neural networks for modelling complex probability distributions. In this note, we revisit their coupling and autoregressive transformation layers as probabilistic graphical models and show that they reduce to Bayesian networks with a pre-defined topology and a learnable density at each node. From this new perspective, we provide three results. First, we show that stacking multiple transformations in a normalizing flow relaxes independence assumptions and entangles the model distribution. Second, we show that a fundamental leap of capacity emerges when the depth of affine flows exceeds 3 transformation layers. Third, we prove the non-universality of the affine normalizing flow, regardless of its depth.

4.1 INTRODUCTION

Normalizing flows [NF, [Rezende and Mohamed, 2015](#)] have gained popularity in the recent years because of their unique ability to model complex data distributions while allowing both for sampling and exact density computation. This family of deep neural networks combines a base distribution with a series of invertible transformations while keeping track of the change of density that is caused by each transformation.

Probabilistic graphical models (PGMs) are well-established mathematical tools that combine graph and probability theory to ease the manipulation of joint distributions. They are commonly used to visualize and reason about the set of independencies in probabilistic models. Among PGMs, Bayesian networks [BN, [Pearl, 2011](#)] offer a nice balance between readability and modelling capacity. Reading independencies stated by a BN is simple and can be performed graphically with the d-separation algorithm [[Geiger et al., 1990](#)].

In this note, we revisit NFs as Bayesian networks. We first briefly review the mathematical grounds of these two worlds. Then, for the first time in the literature, we show that the modelling assumptions behind coupling and autoregressive transformations can be perfectly expressed by distinct classes of BNs. From this insight, we show that stacking multiple transformation layers relaxes independencies and entangles the model distribution. Then, we show that a fundamental change of regime emerges when the NF architecture includes 3 transformation steps or more. Finally, we prove the non-universality of affine normalizing flows.

4.2 BACKGROUND

4.2.1 Normalizing flows

A normalizing flow is defined as a sequence of invertible transformation steps $\mathbf{g}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ($k = 1, \dots, K$) that are composed together to create an expressive invertible mapping $\mathbf{g} = \mathbf{g}_1 \circ \dots \circ \mathbf{g}_K : \mathbb{R}^d \rightarrow \mathbb{R}^d$. This mapping can be used to perform density estimation, using $\mathbf{g}(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ to map a sample $\mathbf{x} \in \mathbb{R}^d$ to a latent vector $\mathbf{z} \in \mathbb{R}^d$ equipped with a density $p_{\mathbf{z}}(\mathbf{z})$. The transformation \mathbf{g} implicitly defines a density $p(\mathbf{x}; \boldsymbol{\theta})$ as given by the change of variables formula,

$$p(\mathbf{x}; \boldsymbol{\theta}) = p_{\mathbf{z}}(\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})) |\det J_{\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})}|,$$

where $J_{\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})}$ is the Jacobian of $\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})$ with respect to \mathbf{x} . The resulting model is trained by maximizing the likelihood of the data $\{\mathbf{x}^1, \dots, \mathbf{x}^N\}$. NFs can also be used for data generation tasks while keeping track of the density of the generated samples such as to improve the latent distribution in variational auto-encoders [Rezende and Mohamed, 2015]. In the rest of this paper, we will not distinguish between \mathbf{g} and \mathbf{g}_k when the discussion will be focused on only one of these steps \mathbf{g}_k .

In general, steps \mathbf{g} can take any form as long as they define a bijective map. Here, we focus on a sub-class of normalizing flows for which these steps can be mathematically described as

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g^1(x_1; \mathbf{c}^1(\mathbf{x})) & \dots & g^d(x_d; \mathbf{c}^d(\mathbf{x})) \end{bmatrix},$$

where the \mathbf{c}^i are denoted as the **conditioners** and constrain the structure of the Jacobian of \mathbf{g} . The functions g^i , partially parameterized by their conditioner, must be invertible with respect to their input variable x_i . These are usually defined as affine or strictly monotonic functions, with the latter being the most general class of invertible scalar continuous functions. In this note, we mainly discuss affine normalizers that can be expressed as $g(x; m, s) = x \exp(s) + m$ where $m \in \mathbb{R}$ and $s \in \mathbb{R}$ are computed by the conditioner.

4.2.2 Bayesian networks

Bayesian networks allow for a compact and natural representation of probability distributions by exploiting conditional independence. More precisely, a BN is a directed acyclic graph (DAG) which structure encodes for the conditional independencies through the concept of d-separation [Geiger et al., 1990]. Equivalently, its skeleton supports an efficient factorization of the joint distribution.

A BN is able to model a distribution p if and only if it is an I-map with respect to p . That is, iff the set of independencies stated by the BN structure is a subset of the independencies that holds for p . Equivalently,

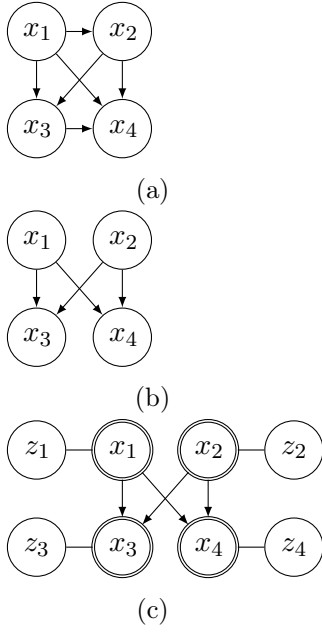


Figure 4.1: Bayesian networks for single-step normalizing flows on a vector $\mathbf{x} = [x_1, x_2, x_3, x_4]^T$. **(a)** BN for an autoregressive conditioner. **(b)** BN for a coupling conditioner. **(c)** Pseudo BN for a coupling conditioner, with the latent variables shown explicitly. Double circles stand for deterministic functions of the parents and non-directed edges stand for bijective relationships.

a BN is a valid representation of a random vector \mathbf{x} iff its density $p_{\mathbf{x}}(\mathbf{x})$ can be factorized by the BN structure as

$$p_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^d p(x_i | \mathcal{P}_i), \quad (4.1)$$

where $\mathcal{P}_i = \{j : A_{i,j} = 1\}$ denotes the set of parents of the vertex i and $A \in \{0, 1\}^{d \times d}$ is the adjacency matrix of the BN. As an example, Fig. 6.1a is a valid BN for any distribution over \mathbf{x} because it does not state any independence, leading to a factorization that results in the chain rule.

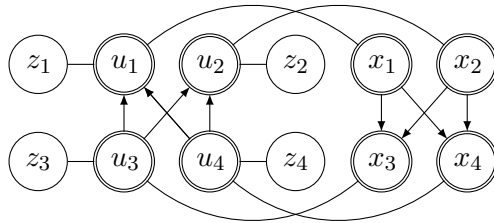


Figure 4.2: A Bayesian network equivalent to a 2-step normalizing flow based on coupling layers. Independence statements are relaxed by the second step.

4.3 NORMALIZING FLOWS AS BAYESIAN NETWORKS

4.3.1 Autoregressive conditioners

Autoregressive conditioners can be expressed as

$$\mathbf{c}^i(\mathbf{x}) = \mathbf{h}^i \left(\begin{bmatrix} x_1 & \dots & x_{i-1} \end{bmatrix}^T \right),$$

where $\mathbf{h}^i : \mathbb{R}^{i-1} \rightarrow \mathbb{R}^l$ are functions of the first $i-1$ components of \mathbf{x} and whose output size depends on architectural choices. These conditioners constrain the Jacobian of \mathbf{g} to be lower triangular, making the computation of its determinant $\mathcal{O}(d)$. The multivariate density $p(\mathbf{x}; \boldsymbol{\theta})$ induced by $\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})$ and $p_{\mathbf{z}}(\mathbf{z})$ can be expressed as a product of d univariate conditional densities,

$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \boldsymbol{\theta}) \prod_{i=2}^d p(x_i | \mathbf{x}_{1:i-1}; \boldsymbol{\theta}). \quad (4.2)$$

When $p_{\mathbf{z}}(\mathbf{z})$ is a factored distribution $p_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^d p(z_i)$, we identify that each component z_i coupled with the corresponding function g^i encodes for the conditional $p(x_i | \mathbf{x}_{1:i-1}; \boldsymbol{\theta})$. An explicit connection between BNs and autoregressive conditioners can be made if we define $\mathcal{P}_i = \{x_1, \dots, x_{i-1}\}$ and compare (6.4) with (6.1). Therefore, and as illustrated in Fig. 6.1a, autoregressive conditioners can be seen as a way to model the conditional factors of a BN that does not state any independence.

4.3.2 Coupling conditioners

Coupling conditioners Dinh et al. [2017] are another popular type of conditioners used in normalizing flows. The conditioners \mathbf{c}^i made from coupling layers are defined as

$$\mathbf{c}^i(\mathbf{x}) = \begin{cases} \underline{\mathbf{h}}^i & \text{if } i < k \\ \mathbf{h}^i(\mathbf{x}_{<k}) & \text{if } i \geq k \end{cases}$$

where the $\underline{\mathbf{h}}^i$ symbol define constant values. As for autoregressive conditioners, the Jacobian of \mathbf{g} made of coupling layers is lower triangular. Assuming a factored latent distribution, the density associated with these conditioners can be written as follows:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{k-1} p(x_i) \prod_{i=k}^d p(x_i | \mathbf{x}_{<k}),$$

$$\text{where } p(x_i) = p(g^i(x_i; \underline{\mathbf{h}}^i)) \frac{\partial g^i(x_i; \underline{\mathbf{h}}^i)}{\partial x_i}$$

$$\text{and } p(x_i | \mathbf{x}_{<k}) = p(g^i(x_i; \mathbf{h}^i(\mathbf{x}_{<k}))) \frac{\partial g^i(x_i; \mathbf{h}^i(\mathbf{x}_{<k}))}{\partial x_i}.$$

The factors define valid 1D conditional probability distributions because they can be seen as 1D changes of variables between z_i and x_i . This factorization can be graphically expressed by a BN as shown in Fig. 6.1b. In addition, we can see Fig. 6.1b as the marginal BN of Fig. 6.1c which fully describes the stochastic process modeled by a NF that is made of a single transformation step and a coupling conditioner. In contrast to autoregressive conditioners, coupling layers are not by themselves universal density approximators, even when associated with very expressive normalizers g^i . Indeed, d-separation reveals independencies stated by this class of BN, such as the conditional independence between each pair in $\mathbf{x}_{\geq k}$ knowing $\mathbf{x}_{<k}$. These independence statements do not hold in general.

4.3.3 Stacking transformation steps

In practice, the invertible transformations discussed above are often stacked together in order to increase the representation capacity of the flow, with the popular good practice of permuting the vector components between two transformation steps. The structural benefits of this stacking strategy can be explained from the perspective of the underlying BN.

First, a BN that explicitly includes latent variables is faithful as long as the sub-graph made only of those latent nodes is an I-map with respect to their distribution. Normalizing flows composed of multiple transformation layers can therefore be viewed as single transformation flows whose latent distribution is itself recursively modeled by a normalizing flow. As an example, Fig. 4.2 illustrates a NF made of two transformation steps with coupling conditioners. It can be observed that the latent vector \mathbf{u} is itself a normalizing flow whose distribution can be factored out by a class of BN.

Second, from the BN associated to a NF, we observe that additional layers relax the independence assumptions defined by its conditioners. The distribution modeled by the flow gets more entangled at each additional layer. For example, Fig. 4.2 shows that for coupling layers, the additional steps relax the strong conditional independencies between x_1 and x_2 of the single transformation NF of Fig. 6.1c. Indeed, we can observe from the figure that x_1 and x_2 have common ancestors (z_3 and z_4) whereas they are clearly assumed independent in Fig. 6.1b.

In general, we note that edges between two nodes in a BN do not model dependence, only the absence of edges does model independence. However, because some of the relationship between nodes are bijective, this implies that these nodes are strictly dependent on each other. We represent these relationships with undirected edges in the BN, as it can be seen in Fig. 4.2.

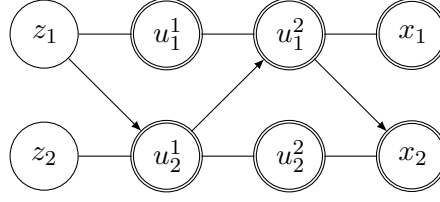


Figure 4.3: The Bayesian network of a three-steps normalizing flow on vector $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^4$. It can be observed that the distribution of the intermediate latent variables, and at the end of the vector \mathbf{x} , becomes more entangled at each additional transformation step. Considering the undirected edges as affine relationships, we see that while u_1^1 and u_2^1 are affine transformations of the latent \mathbf{z} , the vector \mathbf{x} cannot be expressed as a linear function of the latent \mathbf{z} .

4.4 AFFINE NORMALIZING FLOWS UNLOCK THEIR CAPACITY WITH 3 TRANSFORMATION STEPS

We now show how some of the limitations of affine normalizers can be relaxed by stacking multiple transformation steps. We also discuss why some limitations cannot be relaxed even with a large number of transformation steps. We intentionally put aside monotonic normalizers because they have already been proven to lead to universal density approximators when the conditioner is autoregressive [Huang et al. \[2018\]](#). We focus our discussion on a multivariate normal with an identity covariance matrix as base distribution $p_{\mathbf{z}}(\mathbf{z})$.

We first observe from Fig. 6.1 that in a NF with a single transformation step at least one component of \mathbf{x} is a function of only one latent variable. If the normalizer is affine and the base distribution is normal, then this necessarily implies that the marginal distribution of this component is normal as well, which will very likely not lead to a good fit. We easily see that adding steps relaxes this constraint. A more interesting question to ask is what exactly the modelling capacity gain for each additional step of affine normalizer is. Shall we add steps to increase capacity or shall we increase the capacity of each step instead? We first discuss a simple 2-dimensional case, which has the advantage of unifying the discussion for autoregressive and coupling conditioners, and then extend it to a more general setting.

Affine NFs made of a single transformation step induce strong constraints on the form of the density. In particular, these models implicitly assume that the data distribution can be factorized as a product of conditional normal distributions. These assumptions are relaxed when accumulating steps in the NF. As an example, Fig. 4.3 shows the equivalent BN of a 2D NF composed of 3 steps. This flow is mathematically described with the following set of equations:

$$\begin{aligned} u_1^1 &:= z_1 & u_2^1 &:= \exp(s_2^1(z_1))z_2 + m_2^1(z_1) \\ u_2^2 &:= u_2^1 & u_1^2 &:= \exp(s_1^2(u_2^1))z_1 + m_1^2(u_2^1) \end{aligned}$$

$$x_1 := u_1^2 \quad x_2 := \exp(s_2^3(u_1^2))u_2^2 + m_2^3(u_1^2)$$

From these equations, we see that after one step the latent variables u_1^1 and u_2^1 are respectively normal and conditionally normal. This is relaxed with the second step, where the latent variable u_1^2 is a non-linear function of two random variables distributed normally (by assumption on the distribution of z_1 and z_2). However, u_2^2 is a stochastic affine transformation of a normal random variable. In addition, we observe that the expression of u_1^2 is strictly more expressive than the expression of u_2^2 . Finally, x_1 and x_2 are non-linear functions of both latent variables z_1 and z_2 . Assuming that the functions s_j^i and m_j^i are universal approximators, we argue that the stochastic process that generates x_1 and the one that generates x_2 are as expressive as each other. Indeed, by making the functions arbitrarily complex the transformation for x_1 could be made arbitrarily close to the transformations for x_2 and vice versa. This is true because both transformations can be seen as an affine transformation of a normal random variables whose scaling and offset factors are non-linear arbitrarily expressive transformations of all the latent variables. Because of this equilibrium between the two expressions, additional steps do not improve the modelling capacity of the flow. The same observations can be made empirically as illustrated in Fig. 4.3 for 2-dimensional toy problems. A clear leap of capacity occurs from 2-step to 3-step NFs, while having 4 steps or more does not result in any noticeable improvement when s_j^i and m_j^i already have enough capacity.

For $d > 2$, autoregressive and coupling conditioners do not correspond to the same set of equations or BN. However, if the order of the vector is reversed between two transformation steps, the discussion generalizes to any value of d for both conditioners. Indeed, in both cases each component of the intermediate latent vectors can be seen as having a set of conditioning variables and a set of independent variables. At each successive step the indices of the non-conditioning variables are exchanged with the conditioning ones and thus any output vector's component can be expressed either as a component of the vector form of x_1 or of x_2 .

4.5 AFFINE NORMALIZING FLOWS ARE NOT UNIVERSAL DENSITY APPROXIMATORS

We argue that affine normalizers do not lead to universal density approximators in general, even for an infinite number of steps. In the following, we assume again that the latent variables are distributed according to a normal distribution with a unit covariance matrix.

To prove the non-universality of affine normalizing flows, one only needs to provide a counter-example. Let us consider the simple setup in which one component x_I of the random vector \mathbf{x} is independent from all the other components. Let us also assume that x_I is distributed under a

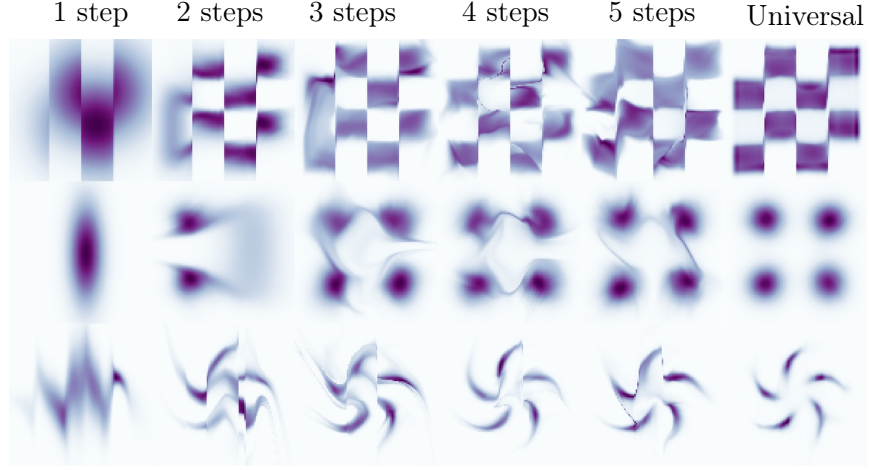


Figure 4.4: Evolution of an affine normalizing flow’s capacity as the number of steps increases. For comparison, the density learned by a universal density approximator is shown on the last column.



Figure 4.5: The equivalent BN of a component with a unique latent variable as ancestor.

non-normal distribution. We can then consider two cases. First, x_I has only one component of the latent vector \mathbf{z} as an ancestor. This implies that the equivalent BN would be as in Fig. 4.5, hence that x_I is a linear function of this ancestor and is therefore normally distributed. Else, x_I has n components of the latent vector as ancestors. However, this second case would imply that at least one undirected edge is removed from the original BN considered in Section 4.3.3. This cannot happen since it would deadly hurt the bijectivity of the flow.

Besides proving the non-universality of affine NFs, this discussion provides the important insight that when affine normalizers must transform non-linearly some latent variables they introduce dependence in the model of the distribution. In some sense, this means that the additional disorder required to model this non-normal component is performed at the cost of some loss in entropy caused by mutual information between the random vector components.

4.6 SUMMARY

In this preliminary work, we have revisited normalizing flows from the perspective of Bayesian networks. We have shown that stacking multiple transformations in a normalizing flow relaxes independence assumptions and entangles the model distribution. Then, we have shown that affine normalizing flows benefit from having at least 3 transformation

layers. Finally, we demonstrated that they remain non-universal density approximators regardless of their depths.

We hope these results will give practitioners more intuition in the design of normalizing flows. We also believe that this work may lead to further research. First, unifying Bayesian networks and normalizing flows could be pushed one step further with conditioners that are specifically designed to model Bayesian networks. Second, the study could be extended for other type of normalizing flows such as non-autoregressive monotonic flows. Finally, we believe this study may spark research at the intersection of structural equation modelling, causal networks and normalizing flows.

OUTLINE

Monotonic neural networks have recently been proposed as a way to define invertible transformations. These transformations can be combined into powerful autoregressive flows that have been shown to be universal approximators of continuous probability distributions. Architectures that ensure monotonicity typically enforce constraints on weights and activation functions, which enables invertibility but leads to a cap on the expressiveness of the resulting transformations. In this work, we propose the Unconstrained Monotonic Neural Network (UMNN) architecture based on the insight that a function is monotonic as long as its derivative is strictly positive. In particular, this latter condition can be enforced with a free-form neural network whose only constraint is the positiveness of its output. We evaluate our new invertible building block within a new autoregressive flow (UMNN-MAF) and demonstrate its effectiveness on density estimation experiments. We also illustrate the ability of UMNNs to improve variational inference.

5.1 INTRODUCTION

Monotonic neural networks have been known as powerful tools to build monotone models of a response variable with respect to individual explanatory variables [Archer and Wang, 1993-01; Sill, 1998; ?; Gupta et al., 2016; You et al., 2017]. Recently, strictly monotonic neural networks have also been proposed as a way to define invertible transformations. These transformations can be combined into effective autoregressive flows that can be shown to be universal approximators of continuous probability distributions. Examples include Neural Autoregressive Flows [NAF, Huang et al., 2018] and Block Neural Autoregressive Flows [B-NAF, De Cao et al., 2020]. Architectures that ensure monotonicity typically enforce constraints on weight and activation functions, which enables invertibility but leads to a cap on the expressiveness of the resulting transformations. For neural autoregressive flows, this does not impede universal approximation but typically requires either complex conditioners or a composition of multiple flows.

Nevertheless, autoregressive flows defined as stacks of reversible transformations have proven to be quite efficient for density estimation of empirical distributions [Papamakarios et al., 2019, 2017; Huang et al., 2018], as well as to improve posterior modelling in Variational Auto-Encoders (VAE) [Germain et al., 2015; Kingma et al., 2016; Huang et al.,

2018]. Practical successes of these models include speech synthesis [Van Den Oord et al., 2016, 2018], likelihood-free inference [Papamakarios et al., 2019], probabilistic programming [van de Meent et al., 2018] and image generation [Kingma and Dhariwal, 2018]. While stacking multiple reversible transformations improves the capacity of the full transformation to represent complex probability distributions, it remains unclear which class of reversible transformations should be used.

In this work, we propose a class of reversible transformations based on a new Unconstrained Monotonic Neural Network (UMNN) architecture. We base our contribution on the insight that a function is monotonic as long as its derivative is strictly positive. This latter condition can be enforced with a free-form neural network whose only constraint is for its output to remain strictly positive.

We summarize our contributions as follows:

- We introduce the Unconstrained Monotonic Neural Network (UMNN) architecture, a new reversible scalar transformation defined via a free-form neural network.
- We combine UMNN transformations into an autoregressive flow (UMNN-MAF) and we demonstrate competitive or state-of-the-art results on benchmarks for normalizing flows.
- We empirically illustrate the scalability of our approach by applying UMNN on high dimensional density estimation problems.

5.2 UNCONSTRAINED MONOTONIC NEURAL NETWORKS

Our primary contribution consists in a neural network architecture that enables learning arbitrary monotonic functions. More specifically, we want to learn a strictly monotonic scalar function $F(x; \psi) : \mathbb{R} \rightarrow \mathbb{R}$ without imposing strong constraints on the expressiveness of the hypothesis class. In UMNNs, we achieve this by only imposing the derivative $f(x; \psi) = \frac{\partial F(x; \psi)}{\partial x}$ to remain of constant sign or, without loss of generality, to be strictly positive. As a result, we can parameterize the bijective mapping $F(x; \psi)$ via its strictly positive derivative $f(x; \psi)$ as

$$F(x; \psi) = \int_0^x f(t; \psi) dt + \underbrace{F(0; \psi)}_{\beta}, \quad (5.1)$$

where $f(t; \psi) : \mathbb{R} \rightarrow \mathbb{R}_+$ is a strictly positive parametric function and $\beta \in \mathbb{R}$ is a scalar. We make f arbitrarily complex using an unconstrained neural network whose output is forced to be strictly positive through an ELU activation unit increased by 1. ψ denotes the parameters of this neural network.

Forward integration The forward evaluation of $F(x; \psi)$ requires solving the integral in Equation (5.1). While this might appear daunting,

such integrals can often be efficiently approximated numerically using Clenshaw-Curtis quadrature. The better known trapezoidal rule, which corresponds to the two-point Newton-Cotes quadrature rule, has an exponential convergence when the integrand is periodic and the range of integration corresponds to its period. Clenshaw-Curtis quadrature takes advantage of this property by using a change of variables followed by a cosine transform. This extends the exponential convergence of the trapezoidal rule for periodic functions to any Lipschitz continuous function. As a result, the number of evaluation points required to reach convergence grows with the Lipschitz constant of the function.

Backward integration Training the integrand neural network f requires evaluating the gradient of F with respect to its parameters. While this gradient could be obtained by backpropagating directly through the integral solver, this would also result in a memory footprint that grows linearly with the number of integration steps. Instead, the derivative of an integral with respect to a parameter ω can be expressed with the Leibniz integral rule:

$$\frac{d}{d\omega} \left(\int_{a(\omega)}^{b(\omega)} f(t; \omega) dt \right) = f(b(\omega); \omega) \frac{d}{d\omega} b(\omega) - f(a(\omega); \omega) \frac{d}{d\omega} a(\omega) + \int_{a(\omega)}^{b(\omega)} \frac{\partial}{\partial \omega} f(t; \omega) dt. \quad (5.2)$$

Applying Equation (5.2) to evaluate the derivative of Equation (5.1) with respect to the parameters ψ , we find

$$\begin{aligned} \nabla_{\psi} F(x; \psi) &= f(x; \psi) \nabla_{\psi} (x) - f(0; \psi) \nabla_{\psi} (0) + \int_0^x \nabla_{\psi} f(t; \psi) dt + \nabla_{\psi} \beta \\ &= \int_0^x \nabla_{\psi} f(t; \psi) dt + \nabla_{\psi} \beta. \end{aligned} \quad (5.3)$$

When using a UMN block in a neural architecture, it is also important to be able to compute its derivative with respect to its input x . In this case, applying Equation (5.2) leads to

$$\frac{d}{dx} F(x; \psi) = f(x; \psi). \quad (5.4)$$

Equations (5.3) and (5.4) make the memory footprint for the backward pass independent from the number of integration steps, and therefore also from the desired accuracy. Indeed, instead of computing the gradient of the integral (which requires keeping track of all the integration steps), we integrate the gradient (which is memory efficient, as this corresponds to summing gradients at different evaluation points). We provide the pseudo-code of the forward and backward passes using Clenshaw-Curtis quadrature in Appendix ??.

Numerical inversion In UMMNs, the modeled monotonic function F is arbitrary. As a result, computing its inverse cannot be done analytically. However, since F is strictly monotonic, it admits a unique inverse

x for any point $y = F(x; \psi)$ in its image, therefore inversion can be computed efficiently with common root-finding algorithms. In our experiments, search algorithms such as the bisection method proved to be fast enough.

5.3 UMN AUTOREGRESSIVE MODELS

5.3.1 Normalizing flows

A Normalizing Flow [NF, ?] is defined as a sequence of invertible transformations $\mathbf{u}_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ($i = 1, \dots, k$) composed together to create an expressive invertible mapping $\mathbf{u} = \mathbf{u}_1 \circ \dots \circ \mathbf{u}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$. It is common for normalizing flows to stack the same parametric function \mathbf{u}_i (with different parameters values) and to reverse variables ordering after each transformation. For this reason, we will focus on how to build one of these repeated transformations, which we further refer to as $\mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^d$.

Density estimation NFs are most commonly used for density estimation, that map empirical samples to unstructured noise. Using normalizing flows, we define a bijective mapping $\mathbf{u}(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ from a sample $\mathbf{x} \in \mathbb{R}^d$ to a latent vector $\mathbf{z} \in \mathbb{R}^d$ equipped with a density $p_Z(\mathbf{z})$. The transformation \mathbf{u} implicitly defines a density $p(\mathbf{x}; \boldsymbol{\theta})$ as given by the change of variables formula,

$$p(\mathbf{x}; \boldsymbol{\theta}) = p_Z(\mathbf{u}(\mathbf{x}; \boldsymbol{\theta})) |\det J_{\mathbf{u}(\mathbf{x}; \boldsymbol{\theta})}|, \quad (5.5)$$

where $J_{\mathbf{u}(\mathbf{x}; \boldsymbol{\theta})}$ is the Jacobian of $\mathbf{u}(\mathbf{x}; \boldsymbol{\theta})$ with respect to \mathbf{x} . The resulting model is trained by maximizing the likelihood of the data $\{\mathbf{x}^1, \dots, \mathbf{x}^N\}$.

Variational auto-encoders NFs are also used in VAE to improve posterior modelling. In this case, a normalizing flow transforms a distribution p_Z into a complex distribution q which can better model the variational posterior. The change of variables formula yields

$$q(\mathbf{u}(\mathbf{z}; \boldsymbol{\theta})) = p_Z(\mathbf{z}) |\det J_{\mathbf{u}(\mathbf{z}; \boldsymbol{\theta})}|^{-1}. \quad (5.6)$$

5.3.2 Autoregressive transformations

To be of practical use, NFs must be composed of transformations for which the determinant of the Jacobian can be computed efficiently, otherwise its evaluation would be running in $\mathcal{O}(d^3)$. A common solution consists in making the transformation \mathbf{g} autoregressive, i.e., such that $\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})$ can be rewritten as a vector of d scalar functions,

$$\mathbf{g}(\mathbf{x}; \boldsymbol{\theta}) = \begin{bmatrix} g^1(x_1; \boldsymbol{\theta}) & \dots & g^i(\mathbf{x}_{1:i}; \boldsymbol{\theta}) & \dots & g^d(\mathbf{x}_{1:d}; \boldsymbol{\theta}) \end{bmatrix},$$

where $\mathbf{x}_{1:i} = [x_1 \ \dots \ x_i]^T$ is the vector including the i first elements of the full vector \mathbf{x} . The Jacobian of this function is lower triangular, which makes the computation of its determinant $\mathcal{O}(d)$. Enforcing the bijectivity of each component g^i is then sufficient to make \mathbf{g} bijective as well.

For the multivariate density $p(\mathbf{x}; \boldsymbol{\theta})$ induced by $\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})$ and $p_Z(\mathbf{z})$, we can use the chain rule to express the joint probability of \mathbf{x} as a product of d univariate conditional densities,

$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \boldsymbol{\theta}) \prod_{i=1}^{d-1} p(x_{i+1} | \mathbf{x}_{1:i}; \boldsymbol{\theta}). \quad (5.7)$$

When $p_Z(\mathbf{z})$ is a factored distribution $p_Z(\mathbf{z}) = \prod_{i=1}^d p(z_i)$, we identify that each component z_i coupled with the corresponding function g^i encodes for the conditional $p(x_i | \mathbf{x}_{1:i-1}; \boldsymbol{\theta})$. Autoregressive transformations strongly rely on the expressiveness of the scalar functions g^i . In this work, we propose to use UMNns to create powerful bijective scalar transformations.

5.3.3 UMN autoregressive transformations (UMN-MAF)

We now combine UMNns with an embedding of the conditioning variables to build invertible autoregressive functions g^i . Specifically, we define

$$\begin{aligned} g^i(\mathbf{x}_{1:i}; \boldsymbol{\theta}) &= F^i(x_i, \mathbf{h}^i(\mathbf{x}_{1:i-1}; \boldsymbol{\phi}^i); \boldsymbol{\psi}^i) \\ &= \int_0^{x_i} f^i(t, \mathbf{h}^i(\mathbf{x}_{1:i-1}; \boldsymbol{\phi}^i); \boldsymbol{\psi}^i) dt + \beta^i(\mathbf{h}^i(\mathbf{x}_{1:i-1}; \boldsymbol{\phi}^i)), \end{aligned} \quad (5.8)$$

where $\mathbf{h}^i(\cdot; \boldsymbol{\phi}^i) : \mathbb{R}^{i-1} \rightarrow \mathbb{R}^q$ is a q -dimensional neural embedding of the conditioning variables $\mathbf{x}_{1:i-1}$ and $\beta(\cdot)^i : \mathbb{R}^{i-1} \rightarrow \mathbb{R}$. Both degenerate into constants for $g^1(\mathbf{x}_1)$. The parameters $\boldsymbol{\theta}$ of the whole transformation $\mathbf{g}(\cdot; \boldsymbol{\theta})$ is the union of all parameters $\boldsymbol{\phi}^i$ and $\boldsymbol{\psi}^i$. For simplicity we remove the parameters of the networks by rewriting $f^i(\cdot; \boldsymbol{\psi}^i)$ as $f^i(\cdot)$ and $\mathbf{h}^i(\cdot; \boldsymbol{\phi}^i)$ as $\mathbf{h}^i(\cdot)$.

In our implementation, we use a Masked Autoregressive Network [Germain et al., 2015; Kingma et al., 2016; Papamakarios et al., 2017] to simultaneously parameterize the d embeddings. In what follows we refer to the resulting UMN autoregressive transformation as UMN-MAF. Figure 5.1 summarizes the complete architecture.

Log-density The change of variables formula applied to the UMN autoregressive transformation results in the log-density

$$\begin{aligned} \log p(\mathbf{x}; \boldsymbol{\theta}) &= \log p_Z(\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})) |\det J_{\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})}| \\ &= \log p_Z(\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})) + \log \left| \prod_{i=1}^d \frac{\partial F^i(x_i, \mathbf{h}^i(\mathbf{x}_{1:i-1}))}{\partial x_i} \right| \end{aligned}$$

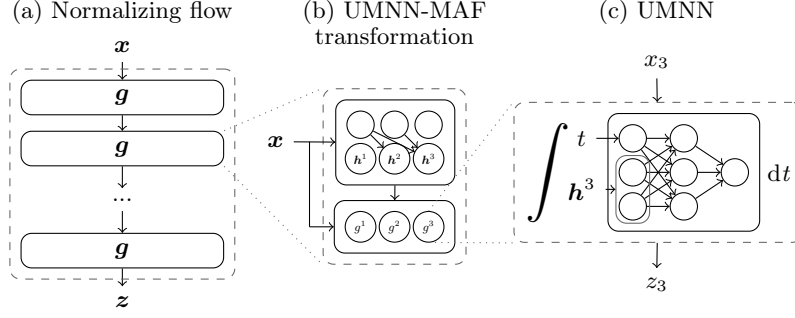


Figure 5.1: **(a)** A normalizing flow made of repeated UMNN-MAF transformations g with identical architectures. **(b)** A UMNN-MAF which transforms a vector $x \in \mathbb{R}^3$. **(c)** The UMNN network used to map x_3 to z_3 conditioned on the embedding $h^3(x_{1:2})$.

$$= \log p_Z(g(x; \theta)) + \sum_{i=1}^d \log f^i(x_i, h^i(x_{1:i-1})). \quad (5.9)$$

Therefore, the transformation leads to a simple expression of (the determinant of) its Jacobian, which can be computed efficiently with a single forward pass. This is different from FJORD [Grathwohl et al., 2018] which relies on numerical methods to compute both the Jacobian and the transformation between the data and the latent space. Therefore our proposed method makes the computation of the Jacobian exact and efficient at the same time.

Sampling Generating samples require evaluating the inverse transformation $g^{-1}(z; \theta)$. The components of the inverse vector $x^{\text{inv}} = g^{-1}(z; \theta)$ can be computed recursively by inverting each component of $g(x; \theta)$:

$$x_1^{\text{inv}} = (g^1)^{-1}(z_1; h^1) \quad \text{if } i = 1 \quad (5.10)$$

$$x_i^{\text{inv}} = (g^i)^{-1}(z_i; h^i(x_{1:i-1}^{\text{inv}})) \quad \text{if } i > 1 \quad (5.11)$$

where $(g^i)^{-1}$ is the inverse of g^i . Another approach to invert an autoregressive model would be to approximate its inverse with another autoregressive network [Van Den Oord et al., 2018]. In this case, the evaluation of the approximated inverse model is as fast as the forward model.

Universality Since the proof is straightforward, we only sketch that UMNN-MAF is a universal density approximator of continuous random variables. We rely on the inverse sampling theorem to prove that UMNNs are universal approximators of continuously derivable (\mathbb{C}^1) monotonic functions. Indeed, if UMNNs can represent any \mathbb{C}^1 monotonic function, then they can also represent the (inverse) cumulative distribution function of any continuous random variable. Any continuously derivable function $f : \mathcal{D} \rightarrow \mathcal{I}$ can be expressed as the following

integral: $f(x) = \int_a^x \frac{df}{dx} dx + f(a)$, $\forall x, a \in \mathcal{D}$. The derivative $\frac{df}{dx}$ is a continuous positive function and the universal approximation theorem of NNs ensures it can be successfully approximated with a NN of sufficient capacity (such as those used in UMNNS).

5.4 RELATED WORK

The most similar work to UMN-MAF are certainly Neural Autoregressive Flow [NAF, [Huang et al., 2018](#)] and Block Neural Autoregressive Flow [B-NAF, [De Cao et al., 2020](#)], which both rely on strictly monotonic transformations for building bijective mappings. In NAF, transformations are defined as neural networks which activation functions are all constrained to be strictly monotonic and which weights are the output of a strictly positive and autoregressive HyperNetwork [?]. [Huang et al. \[2018\]](#) shows that NAFs are universal density approximators. In B-NAF, the authors improve on the scalability of the NAF architecture by making use of masking operations instead of HyperNetworks. They also present a proof of the universality of B-NAF, which extends to UMN-MAF. Our work differs from both NAF and B-NAF in the sense that the UMN monotonic transformation is based on free-form neural networks for which no constraint, beyond positiveness of the output, is enforced on the hypothesis class. This leads to multiple advantages: it enables the use of any state-of-the-art neural architecture, simplifies weight initialization, and leads to a more lightweight evaluation of the Jacobian.

More generally, UMN-MAF relates to works on normalizing flows built upon autoregressive networks and affine transformations. [Germain et al. \[2015\]](#) first introduced masking as an efficient way to build autoregressive networks, and proposed autoregressive networks for density estimation of high dimensional binary data. Masked Autoregressive Flows [[Papamakarios et al., 2017](#)] and Inverse Autoregressive Flows [[Kingma et al., 2016](#)] have generalized this approach to real data, respectively for density estimation and for latent posterior representation in variational auto-encoders. More recently, [Oliva et al. \[2018\]](#) proposed to stack various autoregressive architectures to create powerful reversible transformations. Meanwhile, [Jaini et al. \[2019\]](#) proposed a new Sum-of-Squares flow that is defined as the integral of a second order polynomial parametrized by an autoregressive NN.

With NICE, ? introduced coupling layers, which correspond to bijective transformations splitting the input vector into two parts. They are defined as

$$\mathbf{z}_{1:k} = \mathbf{x}_{1:k} \quad \text{and} \quad \mathbf{z}_{k+1:d} = e^{\boldsymbol{\sigma}(\mathbf{x}_{1:k})} \odot \mathbf{x}_{k+1:d} + \boldsymbol{\mu}(\mathbf{x}_{1:k}), \quad (5.12)$$

where $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ are two unconstrained functions $\mathbb{R}^{d-k} \rightarrow \mathbb{R}^{d-k}$. The same authors introduced RealNVP [[Dinh et al., 2017](#)], which combines coupling layers with normalizing flows and multi-scale architectures for

image generation. Glow [Kingma and Dhariwal, 2018] extends RealNVP by introducing invertible 1x1 convolutions between each step of the flow. In this work we have used UMNNs in the context of autoregressive architectures, however UMNNs could also be applied to replace the linear transformation in coupling layers.

Finally, our architecture also shares a connection with Neural Ordinary Differential Equations [NODE, Chen et al., 2018b]. The core idea of this architecture is to learn an ordinary differential equation which dynamic is parameterized by a neural network. Training can be carried out by backpropagating efficiently through the ODE solver, with constant memory requirements. Among other applications, NODE can be used to model a continuous normalizing flow with a free-form Jacobian as in FFJORD [Grathwohl et al., 2018]. Similarly, a UMNN transformation can be seen as a structured neural ordinary differential equation in which the dynamic of the vector field is separable and can be solved efficiently by direct integration.

5.5 EXPERIMENTS

In this section, we evaluate the expressiveness of UMNN-MAF on a variety of density estimation benchmarks, as well as for approximate inference in variational auto-encoders. The source code is accessible at <https://github.com/AWehenkel/UMNN>.

Experiments were carried out using the same integrand neural network in the UMNN component – i.e., in Equation 5.8, $f^i = f$ with shared weights $\psi^i = \psi$ for $i \in \{1, \dots, d\}$. The functions β^i are taken to be equal to one of the outputs of the embedding network. We observed in our experiments that sharing the same integrand function does not impact performance. Therefore, the neural embedding function \mathbf{h}^i must produce a fixed size output for $i \in \{1, \dots, d\}$.

5.5.1 2D toy problems

We first train a UMNN-MAF on 2-dimensional toy distributions, as defined by Grathwohl et al. [2018]. To train the model, we minimize the negative log-likelihood of observed data

$$L(\theta) = - \sum_{n=1}^N \left[\log p_Z(\mathbf{g}(\mathbf{x}^n; \theta)) + \sum_{i=1}^d \log f(x_i^n, \mathbf{h}^i(\mathbf{x}_{1:i-1}^n)) \right]. \quad (5.13)$$

The flow used to solve these tasks is the same for all distributions and is composed of a single transformation. More details can be found in Appendix ??.

Figure 5.2 demonstrates that our model is able to learn a change of variables that warps a simple isotropic Gaussian into multimodal and/or discontinuous distributions. We observe from the figure that

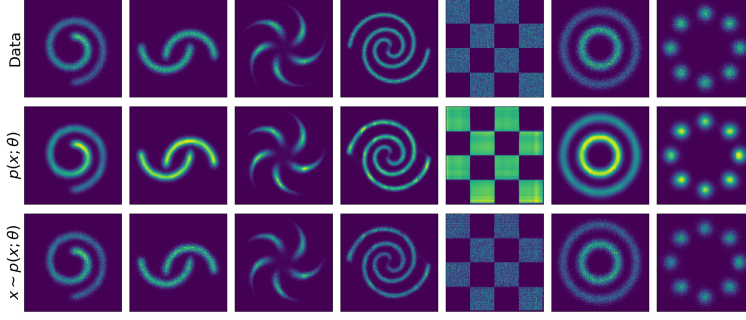


Figure 5.2: Density estimation and sampling with a UMNN-MAF network on 2D toy problems. **Top:** Samples from the empirical distribution $p(\mathbf{x})$. **Middle:** Learned density $p(\mathbf{x}; \theta)$. **Bottom:** Samples drawn by numerical inversion. UMNN-MAF manages to precisely capture multi-modal and/or discontinuous distributions. Sampling is possible even if the model is not invertible analytically.

our model precisely captures the density of the data. We also observe that numerical inversion for generating samples yields good results.

5.5.2 Density estimation

We further validate UMNN-MAF by comparing it to state-of-the-art normalizing flows. We carry out experiments on tabular datasets (POWER, GAS, HEPMASS, MINIBOONE, BSDS300) as well as on MNIST. We follow the experimental protocol of [Papamakarios et al. \[2017\]](#). All training hyper-parameters and architectural details are given in Appendix ?? . For each dataset, we report results on test data for our best performing model (selected on the validation data). At testing time we use a large number of integration steps (100) to compute the integral, this ensures its correctness and avoids misestimating the performance of UMNN-MAF.

Table 6.3 summarizes our results, where we can see that on tabular datasets, our method is competitive with other normalizing flows. For POWER, our architecture slightly outperforms all others. It is also better than other monotonic networks (category (c)) on 3 tabular datasets over 5. From these results, we could conclude that Transformation Autoregressive Networks [TAN, [Oliva et al., 2018](#)] is overall the best method for density estimation. It is however important to note that TAN is a flow composed of many heterogeneous transformations (both autoregressive and non-autoregressive). For this reason, it should not be directly compared to the other models which respective results are specific to a single architecture. However, TAN provides the interesting insight that combining heterogeneous components into a flow leads to better results than an homogeneous flow.

Notably, we do not make use of a multi-scale architecture to train our model on MNIST. On this task, UMNN-MAF slightly outperforms all other models by a reasonable margin. Samples generated by a condi-

Table 5.1: Average negative log-likelihood on test data over 3 runs, error bars are equal to the standard deviation. Results are reported in nats for tabular data and bits/dim for MNIST; lower is better. The best performing architecture for each dataset is written in bold and the best performing architecture per category is underlined. (a) Non-autoregressive models, (b) Autoregressive models, (c) Monotonic and autoregressive models. UMNN outperforms other monotonic transformations on 4 tasks over 6 and is the overall best performing model on 2 tasks over 6.

Dataset	POWER	GAS	HEPMASS	MINIBOONE	BSDS300	MNIST
RealNVP - Dinh et al. [2017]	$-0.17_{\pm.01}$	$-8.33_{\pm.14}$	$18.71_{\pm.02}$	$13.55_{\pm.49}$	$-153.28_{\pm1.78}$	-
(a) Glow - Kingma and Dhariwal [2018]	$-0.17_{\pm.01}$	$-8.15_{\pm.40}$	$19.92_{\pm.08}$	$11.35_{\pm.07}$	$-155.07_{\pm.03}$	-
FFJORD - Grathwohl et al. [2018]	<u>$-0.46_{\pm.01}$</u>	<u>$-8.59_{\pm.12}$</u>	<u>$14.92_{\pm.08}$</u>	<u>$10.43_{\pm.04}$</u>	<u>$-157.40_{\pm.19}$</u>	-
MADE - Germain et al. [2015]	$3.08_{\pm.03}$	$-3.56_{\pm.04}$	$20.98_{\pm.02}$	$15.59_{\pm.50}$	$-148.85_{\pm.28}$	$2.04_{\pm.01}$
(b) MAF - Papamakarios et al. [2017]	$-0.24_{\pm.01}$	$-10.08_{\pm.02}$	$17.70_{\pm.02}$	$11.75_{\pm.44}$	$-155.69_{\pm.28}$	$1.89_{\pm.01}$
TAN - Oliva et al. [2018]	<u>$-0.60_{\pm.01}$</u>	<u>$-12.06_{\pm.02}$</u>	$13.78_{\pm.02}$	<u>$11.01_{\pm.48}$</u>	<u>$-159.80_{\pm.07}$</u>	<u>$1.19_{\pm.01}$</u>
NAF - Huang et al. [2018]	$-0.62_{\pm.01}$	$-11.96_{\pm.33}$	$15.09_{\pm.40}$	$8.86_{\pm.15}$	$-157.73_{\pm.30}$	-
(c) B-NAF - De Cao et al. [2020]	$-0.61_{\pm.01}$	<u>$-12.06_{\pm.09}$</u>	$14.71_{\pm.38}$	$8.95_{\pm.07}$	$-157.36_{\pm.03}$	-
SOS - Jaini et al. [2019]	$-0.60_{\pm.01}$	$-11.99_{\pm.41}$	$15.15_{\pm.1}$	$8.90_{\pm.11}$	$-157.48_{\pm.41}$	$1.81_{\pm.01}$
UMNN-MAF (ours)	<u>$-0.63_{\pm.01}$</u>	$-10.89_{\pm.7}$	<u>$13.99_{\pm.21}$</u>	$9.67_{\pm.13}$	<u>$-157.98_{\pm.01}$</u>	<u>$1.13_{\pm.01}$</u>

tional model are shown on Figure 5.3, for which it is worth noting that UMNN-MAF is the first monotonic architecture that has been inverted to generate samples. Indeed, MNIST can be considered as a high dimensional dataset ($d = 784$) for standard feed forward neural networks which autoregressive networks are part of. NAF and B-NAF do not report any result for this benchmark, presumably because of memory explosion. In comparison, BSDS300, which data dimension is one order of magnitude smaller than MNIST ($63 \ll 784$), are the largest data they have tested on. Table 5.2 shows the number of parameters used by UMNN-MAF in comparison to B-NAF and NAF. For bigger datasets, UMNN-MAF requires less parameters than NAF to reach similar or better performance. This could explain why NAF has never been used for density estimation on MNIST.

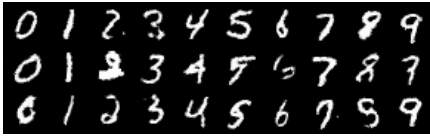


Figure 5.3: Samples generated by numerical inversion of a conditional UMNN-MAF trained on MNIST. Samples z are drawn from an isotropic Gaussian with $\sigma = .75$. See Appendix ?? for more details.

Table 5.2: Comparison of the number of parameters between NAF, B-NAF and UMNN-MAF. In high dimensional datasets, UMNN-MAF requires fewer parameters than NAF and a similar number to B-NAF.

Dataset	NAF	B-NAF	UMNN-MAF
POWER ($d = 6$)	4.14e5	3.07e5	5.09e5
GAS ($d = 8$)	4.02e5	5.44e5	8.15e5
HEPMASS ($d = 21$)	9.27e6	3.72e6	3.62e6
MINIBOONE ($d = 43$)	7.49e6	4.09e6	3.46e6
BSDS300 ($d = 63$)	3.68e7	8.76e6	1.56e7

5.5.3 Variational auto-encoders

To assess the performance of our model, we follow the experimental setting of ? for VAE. The encoder and the decoder architectures can be found in the appendix of their paper. In VAE it is usual to let the encoder output the parameters of the flow. For UMNN-MAF, this would cause the encoder output’s dimension to be too large. Instead, the encoder output is passed as additional entries of the UMNN-MAF. Like other architectures, the UMNN-MAF also takes as input a vector of noise drawn from an isotropic Gaussian of dimension 64.

Table 5.3 presents our results. It shows that on MNIST and Omniglot, UMNN-MAF slightly outperforms the classical VAE as well as planar flows. Moreover, on these datasets and Freyfaces, IAF, B-NAF and UMNN-MAF achieve similar results. FFJORD is the best among all, however it is worth noting that the roles of encoder outputs in FFJORD, B-NAF, IAF and Sylvester are all different. We believe that

Table 5.3: Average negative evidence lower bound of VAEs over 3 runs, error bars are equal to the standard deviation. Results are reported in bits per dim for Freyfaces and in nats for the other datasets; lower is better. UMNN-NAF is performing slightly better than IAF but is outperformed by B-NAF. We believe that the gap in performance between B-NAF and UMNN is due to the way the NF is conditioned by the encoder’s output.

Dataset	MNIST	Freyfaces	Omniglot	Caltech 101
VAE - ?	86.65 \pm .06	4.53 \pm .02	104.28 \pm .39	110.80 \pm .46
Planar - ?	86.06 \pm .32	4.40 \pm .06	102.65 \pm .42	109.66 \pm .42
(a) IAF - Kingma et al. [2016]	84.20 \pm .17	4.47 \pm .05	102.41 \pm .04	111.58 \pm .38
Sylvester - ?	83.32 \pm .06	4.45 \pm .04	99.00 \pm .04	104.62 \pm .29
FFJORD - Grathwohl et al. [2018]	82.82 \pm .01	4.39 \pm .01	98.33 \pm .09	104.03 \pm .43
(b) B-NAF - De Cao et al. [2020]	83.59 \pm .15	4.42 \pm .05	100.08 \pm .07	105.42 \pm .49
UMNN-MAF (ours)	84.11 \pm .05	4.51 \pm .01	100.98 \pm .13	110.45 \pm .69

the heterogeneity of the results could be, at least in part, due to the different amortizations.

5.6 DISCUSSION AND SUMMARY

Static integral quadrature can be inaccurate. Computing the integral with static Clenshaw-Curtis quadrature only requires the evaluation of the integrand at predefined points. As such, batches of points can be processed all at once, which makes static Clenshaw-Curtis quadrature well suited for neural networks. However, static quadratures do not account for the error made during the integration. As a consequence, the quadrature is inaccurate when the integrand is not smooth enough and the number of integration steps is too small. In this work, we have reduced the integration error by applying the normalization described by ? in order to control the Lipschitz constant of the integrand and appropriately set the number of integration steps. We observed that as long as the Lipschitz constant of the network does not increase dramatically (< 1000), a reasonable number of integration steps (< 100) is sufficient to ensure the convergence of the quadrature. An alternative solution would be to use dynamic quadrature such as dynamic Clenshaw-Curtis.

Efficiency of numerical inversion. Architectures relying on linear transformations [Papamakarios et al., 2017; Kingma et al., 2016; Dinh et al., 2017; Kingma and Dhariwal, 2018] are trivially exactly and efficiently invertible. In contrast, the UMNN transformation has no analytic inverse. Nevertheless, it can be inverted numerically using root-finding algorithms. Since most such algorithms rely on multiple nested evalu-

ations of the function to be inverted, applying them naively to a numerical integral would quickly become very inefficient. However, the Clenshaw-Curtis quadrature is part of the nested quadrature family, meaning that the evaluation of the integral at multiple nested points can take advantage of previous evaluations and thus be implemented efficiently. As an alternative, [Van Den Oord et al. \[2018\]](#) have shown that an invertible model can always be distilled to learn its inverse, and thus make the inversion efficient whatever the cost of inversion of the original model.

Scalability and complexity analysis. UMNN-MAF is particularly well suited for density estimation because the computation of the Jacobian only requires a single forward evaluation of a NN. Together with the Leibniz integral rule, they make the evaluation of the log-likelihood derivative as memory efficient as usual supervised learning, which is equivalent to a single backward pass on the computation graph. By contrast, density estimation with previous monotonic transformations typically requires a backward evaluation of the computation graph of the transformer NN to obtain the Jacobian. Then, this pass must be evaluated backward again in order to obtain the log-likelihood derivative. Both NAF and B-NAF provide a method to make this computation numerically stable, however both fail at not increasing the size of the computation graph of the log-likelihood derivative, hence leading to a memory overhead. The memory saved by the Leibniz rule may serve to speed up the quadrature computation. In the case of static Clenshaw-Curtis, the function values at each evaluation point can be computed in parallel using batch of points. In consequence, when the GPU memory is large enough to store "meta-batches" of size $d \times N \times B$ (with d the dimension of the data, N the number of integration steps and B the batch size) the computation is approximately as fast as a forward evaluation of the integrand network.

Summary We have introduced Unconstrained Monotonic Neural Networks, a new invertible transformation built upon free-form neural networks allowing the use of any state-of-the-art architecture. Monotonicity is guaranteed without imposing constraints on the expressiveness of the hypothesis class, contrary to classical approaches. We have shown that the resulting integrated neural network can be evaluated efficiently using standard quadrature rule while its inverse can be computed using numerical algorithms. We have shown that our transformation can be composed into an autoregressive flow, with competitive or state-of-the-art results on density estimation and variational inference benchmarks. Moreover, UMNN is the first monotonic transformation that has been successfully applied for density estimation on high dimensional data distributions (MNIST), showing better results than the classical approaches.

We identify several avenues for improvement and further research. First, we believe that numerical integration could be fasten up during training, by leveraging the fact that controlled numerical errors can actually help generalization. Moreover, the UMNN transformation would certainly profit from using a dynamic integration scheme, both in terms of accuracy and efficiency. Second, it would be worth comparing the newly introduced monotonic transformation with common approaches for modelling monotonic functions in machine learning. On a similar track, these common approaches could be combined into an autoregressive flow as shown in Section 5.3.3. Finally, our monotonic transformation could be used within other neural architectures than generative autoregressive networks, such as multi-scale architectures [Dinh et al., 2017] and learnable 1D convolutions [Kingma and Dhariwal, 2018].

Part III

HYBRID GENERATIVE MODELLING

*If I have seen further, it is by standing
on shoulders of giants.*

ISAAC NEWTON (1675)

STRUCTURED DENSITY ESTIMATION WITH NORMALIZING FLOWS

OUTLINE

???

6.1 INTRODUCTION

Normalizing flows [NFs, [1, 2]] have proven to be an effective way to model complex data distributions with neural networks. These models map data points to latent variables through an invertible function while keeping track of the change of density caused by the transformation. In contrast to variational auto-encoders (VAEs) and generative adversarial networks (GANs), NFs provide access to the exact likelihood of the model's parameters, hence offering a sound and direct way to optimize the network parameters. Normalizing flows have proven to be of practical interest as demonstrated by [3] and [4] for speech synthesis, by [5] and [6] for variational inference or by [7] and [8] for simulation-based inference. Yet, their usage as a base component of the machine learning toolbox is still limited in comparison to GANs or VAEs. Recent efforts have been made by [9] and [10] to define the fundamental principles of flow design and by [11] to provide coding tools for modular implementations. We argue that normalizing flows would gain in popularity by offering stronger inductive bias as well as more interpretability.

Sometimes forgotten in favor of more data oriented methods, probabilistic graphical models (PGMs) have been popular for modelling complex data distributions while being relatively simple to build and read [12]. Among PGMs, Bayesian networks [BNs, [13]] offer an appealing balance between modelling capacity and simplicity. Most notably, these models have been at the basis of expert systems before the big data era (e.g. [14]) and were commonly used to merge qualitative expert knowledge and quantitative information together. On the one hand, experts stated independence assumptions that should be encoded by the structure of the network. On the other hand, data were used to estimate the parameters of the conditional probabilities/densities encoding the quantitative aspects of the data distribution. These models have progressively received less attention from the machine learning community in favor of other methods that scale better with the dimensionality of the data.

Driven by the objective of integrating intuition into normalizing flows and the proven relevance of BNs for combining qualitative and quantitative reasoning, we summarize our contributions as follows: **(i)** From the

insight that coupling and autoregressive transformations can be reduced to Bayesian networks with a fixed topology, we introduce the more general graphical conditioner for normalizing flows, featuring either a prescribed or a learnable BN topology; **(ii)** We show that using a correct prescribed topology leads to improvements in the modeled density compared to autoregressive methods. When the topology is not known we observe that, with the right amount of ℓ_1 -penalization, graphical conditioners discover relevant relationships; **(iii)** In addition, we show that graphical normalizing flows perform well in a large variety of density estimation tasks compared to classical black-box flow architectures.

6.2 BACKGROUND

Bayesian networks A Bayesian network is a directed acyclic graph (DAG) that represents independence assumptions between the components of a random vector. Formally, let $\mathbf{x} = [x_1, \dots, x_d]^T \in \mathbb{R}^d$ be a random vector distributed under $p_{\mathbf{x}}$. A BN associated to \mathbf{x} is a directed acyclic graph made of d vertices representing the components x_i of \mathbf{x} . In this kind of network, the absence of edges models conditional independence between groups of components through the concept of d-separation [?]. A BN is a valid representation of a random vector \mathbf{x} iff its density can be factorized as

$$p_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^d p(x_i | \mathcal{P}_i), \quad (6.1)$$

where $\mathcal{P}_i = \{j : A_{i,j} = 1\}$ denotes the set of parents of the vertex i and $A \in \{0, 1\}^{d \times d}$ is the adjacency matrix of the BN. As an example, Fig. 6.1a is a valid BN for any distribution over \mathbf{x} because it does not state any independence and leads to a factorization that corresponds to the chain rule. However, in practice we seek for a sparse and valid BN which models most of the independence between the components of \mathbf{x} , leading to an efficient factorization of the modeled probability distribution. It is worth noting that making hypotheses on the graph structure is equivalent to assuming certain conditional independence between some of the vector's components.

Normalizing flows A normalizing flow is defined as a sequence of invertible transformation steps $\mathbf{g}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ($k = 1, \dots, K$) composed together to create an expressive invertible mapping $\mathbf{g} := \mathbf{g}_1 \circ \dots \circ \mathbf{g}_K : \mathbb{R}^d \rightarrow \mathbb{R}^d$. This mapping can be used to perform density estimation, using $\mathbf{g}(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ to map a sample $\mathbf{x} \in \mathbb{R}^d$ onto a latent vector $\mathbf{z} \in \mathbb{R}^d$ equipped with a prescribed density $p_{\mathbf{z}}(\mathbf{z})$ such as an isotropic Normal. The transformation \mathbf{g} implicitly defines a density $p(\mathbf{x}; \boldsymbol{\theta})$ as given by the change of variables formula,

$$p(\mathbf{x}; \boldsymbol{\theta}) = p_{\mathbf{z}}(\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})) \left| \det J_{\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})} \right|, \quad (6.2)$$

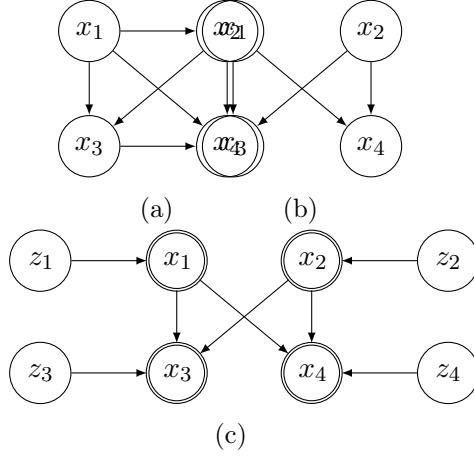


Figure 6.1: Bayesian networks equivalent to normalizing flows made of a single transformation step. (a) Autoregressive conditioner. (b) Coupling conditioner. (c) Coupling conditioner, with latent variables shown explicitly. Double circles stand for deterministic functions of the parents.

where $J_{\mathbf{g}(\mathbf{x};\boldsymbol{\theta})}$ is the Jacobian of $\mathbf{g}(\mathbf{x};\boldsymbol{\theta})$ with respect to \mathbf{x} . The resulting model is trained by maximizing the likelihood of the model’s parameters $\boldsymbol{\theta}$ given the training dataset $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$. Unless needed, we will not distinguish between \mathbf{g} and \mathbf{g}_k for the rest of our discussion.

In general the steps \mathbf{g} can take any form as long as they define a bijective map. Here, we focus on a sub-class of normalizing flows for which the steps can be mathematically described as

$$\mathbf{g}(\mathbf{x}) = \left[g^1(x_1; \mathbf{c}^1(\mathbf{x})) \quad \dots \quad g^d(x_d; \mathbf{c}^d(\mathbf{x})) \right]^T, \quad (6.3)$$

where the \mathbf{c}^i are the **conditioners** which role is to constrain the structure of the Jacobian of \mathbf{g} . The functions g^i , partially parameterized by their conditioner, must be invertible with respect to their input variable x_i . They are often referred to as transformers, however in this work we will use the term **normalizers** to avoid any confusion with attention-based transformer architectures.

The conditioners examined in this work can be combined with any normalizer. In particular, we consider affine and monotonic normalizers. An affine normalizer $g : \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ can be expressed as $g(x; m, s) = x \exp(s) + m$, where $m \in \mathbb{R}$ and $s \in \mathbb{R}$ are computed by the conditioner. There exist multiple methods to parameterize monotonic normalizers [????], but in this work we rely on Unconstrained Monotonic Neural Networks [UMNNs, ?] which can be expressed as $g(x; \mathbf{c}) = \int_0^x f(t, \mathbf{c}) dt + \beta(\mathbf{c})$, where $\mathbf{c} \in \mathbb{R}^{|\mathbf{c}|}$ is an embedding made by the conditioner and $f : \mathbb{R}^{|\mathbf{c}|+1} \rightarrow \mathbb{R}^+$ and $\beta : \mathbb{R}^d \rightarrow \mathbb{R}$ are two neural networks respectively with a strictly positive scalar output and a real scalar output. ? proved NFs built with autoregressive conditioners and monotonic normalizers are universal density approximators of continuous random variables.

6.3 NORMALIZING FLOWS AS BAYESIAN NETWORKS

Autoregressive conditioners Due to computing speed considerations, NFs are usually composed of transformations for which the determinant of the Jacobian can be computed efficiently, as otherwise its evaluation would scale cubically with the input dimension. A common solution is to use autoregressive conditioners, i.e., such that

$$\mathbf{c}^i(\mathbf{x}) = \mathbf{h}^i \left(\begin{bmatrix} x_1 & \dots & x_{i-1} \end{bmatrix}^T \right)$$

are functions \mathbf{h}^i of the first $i - 1$ components of \mathbf{x} . This particular form constrains the Jacobian of \mathbf{g} to be lower triangular, which makes the computation of its determinant $\mathcal{O}(d)$.

For the multivariate density $p(\mathbf{x}; \boldsymbol{\theta})$ induced by $\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})$ and $p_{\mathbf{z}}(\mathbf{z})$, we can use the chain rule to express the joint probability of \mathbf{x} as a product of d univariate conditional densities,

$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \boldsymbol{\theta}) \prod_{i=2}^d p(x_i | \mathbf{x}_{1:i-1}; \boldsymbol{\theta}). \quad (6.4)$$

When $p_{\mathbf{z}}(\mathbf{z})$ is a factored distribution $p_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^d p(z_i)$, we identify that each component z_i coupled with the corresponding functions g^i and embedding vectors \mathbf{c}^i encode for the conditional $p(x_i | \mathbf{x}_{1:i-1}; \boldsymbol{\theta})$. Therefore, and as illustrated in Fig. 6.1a, autoregressive transformations can be seen as a way to model the conditional factors of a BN that does not state any independence but relies on a predefined node ordering. This becomes clear if we define $\mathcal{P}_i = \{x_1, \dots, x_{i-1}\}$ and compare (6.4) with (6.1).

The complexity of the conditional factors strongly depends on the ordering of the vector components. While not hurting the universal representation capacity of normalizing flows, the arbitrary ordering used in autoregressive transformations leads to poor inductive bias and to factors that are most of the time difficult to learn. In practice, one often alleviates the arbitrariness of the ordering by stacking multiple autoregressive transformations combined with random permutations on top of each other.

Coupling conditioners Coupling layers [?] are another popular type of conditioners that lead to a bipartite structure. The conditioners \mathbf{c}^i made from coupling layers are defined as

$$\mathbf{c}^i(\mathbf{x}) = \begin{cases} \underline{\mathbf{h}}^i & \text{if } i \leq k \\ \mathbf{h}^i \left(\begin{bmatrix} x_1 & \dots & x_k \end{bmatrix}^T \right) & \text{if } i > k \end{cases}$$

where the underlined $\underline{\mathbf{h}}^i \in \mathbb{R}^{|\mathcal{C}|}$ denote constant values and $k \in \{1, \dots, d\}$ is a hyper-parameter usually set to $\lfloor \frac{d}{2} \rfloor$. As for autoregressive conditioners, the Jacobian of \mathbf{g} made of coupling layers is lower triangular. Again,

and as shown in Fig. 6.1b and 6.1c, these transformations can be seen as a specific class of BN where $\mathcal{P}_i = \{\}$ for $i \leq k$ and $\mathcal{P}_i = \{1, \dots, k\}$ for $i > k$. D-separation can be used to read off the independencies stated by this class of BNs such as the conditional independence between each pair in $\mathbf{x}_{k+1:d}$ knowing $\mathbf{x}_{1:k}$. For this reason, and in contrast to autoregressive transformations, coupling layers are not by themselves universal density approximators even when associated with very expressive normalizers g^i [?]. In practice, these bipartite structural independencies can be relaxed by stacking multiple layers, and may even recover an autoregressive structure. They also lead to useful inductive bias, such as in the multi-scale architecture with checkerboard masking [??].

Figure 6.2: Sampling

Algorithm 6.1 (H). 1: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$
 2: $\mathbf{x} \leftarrow \mathbf{z}$
 3: **repeat**
 4: $\mathbf{c}^i \leftarrow \mathbf{h}^i(\mathbf{x} \odot A_{i,:}) \quad \forall i \in \{1, \dots, d\}$
 5: $x_i \leftarrow (g^i)^{-1}(z_i; \mathbf{c}^i, \boldsymbol{\theta}) \quad \forall i \in \{1, \dots, d\}$
 6: **until** \mathbf{x} converged

6.4 GRAPHICAL NORMALIZING FLOW

6.4.1 Graphical conditioners

Following up on the previous discussion, we introduce the graphical conditioner architecture. We motivate our approach by observing that the topological ordering (a.k.a. ancestral ordering) of any BN leads to a lower triangular adjacency matrix whose determinant is equal to the product of its diagonal terms (proof in Appendix ??). Therefore, conditioning factors $\mathbf{c}^i(\mathbf{x})$ selected by following a BN adjacency matrix necessarily lead to a transformation \mathbf{g} whose Jacobian determinant remains efficient to compute.

Formally, given a BN with adjacency matrix $A \in \{0, 1\}^{d \times d}$, we define the **graphical conditioner** as being

$$\mathbf{c}^i(\mathbf{x}) = \mathbf{h}^i(\mathbf{x} \odot A_{i,:}), \quad (6.5)$$

where $\mathbf{x} \odot A_{i,:}$ is the element-wise product between the vector \mathbf{x} and the i^{th} row of A – i.e., the binary vector $A_{i,:}$ is used to mask on \mathbf{x} . NFs built with this new conditioner architecture can be inverted by sequentially inverting each component in the topological ordering. In our implementation the neural networks modelling the h^i functions are shared to save memory and they take an additional input that one-hot encodes the value i . An alternative approach would be to use a masking scheme similar to what is done by ? in MADE as suggested by ?.

The graphical conditioner architecture can be used to learn the conditional factors in a continuous BN while elegantly setting structural

independencies prescribed from domain knowledge. In addition, the inverse of NFs built with graphical conditioners is as simple as it is for autoregressive and coupling conditioners, Algorithm 6.2 describes an inversion procedure. We also now note how these two conditioners are just special cases in which the adjacency matrix reflects the classes of BNs discussed in Section 6.3.

6.4.2 Learning the topology

In many cases, defining the whole structure of a BN is not possible due to a lack of knowledge about the problem at hand. Fortunately, not only is the density at each node learnable, but also the DAG structure itself: defining an arbitrary topology and ordering, as it is implicitly the case for autoregressive and coupling conditioners, is not necessary.

Building upon *Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning* [NO TEARS, ?], we convert the combinatorial optimization of score-based learning of a DAG into a continuous optimization by relaxing the domain of A to real numbers instead of binary values. That is,

$$\begin{aligned} \max_{A \in \mathbb{R}^{d \times d}} F(A) \\ \text{s.t. } \mathcal{G}(A) \in \text{DAGs} \end{aligned} \iff \begin{aligned} \max_{A \in \mathbb{R}^{d \times d}} F(A) \\ \text{s.t. } w(A) = 0, \end{aligned} \quad (6.6)$$

where $\mathcal{G}(A)$ is the graph induced by the weighted adjacency matrix A and $F : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ is the log-likelihood of the graphical NF \mathbf{g} plus a regularization term, i.e.,

$$F(A) = \sum_{j=1}^N \log(p(\mathbf{x}^j; \boldsymbol{\theta})) + \lambda_{\ell_1} \|A\|_1, \quad (6.7)$$

where λ_{ℓ_1} is an ℓ_1 -regularization coefficient and N is the number of training samples \mathbf{x}^i . The likelihood is computed as

$$p(\mathbf{x}; \boldsymbol{\theta}) = p_{\mathbf{z}}(\mathbf{g}(\mathbf{x}; \boldsymbol{\theta})) \prod_{i=1}^d \left| \frac{\partial g^i(x_i; \mathbf{h}^i(\mathbf{x} \odot A_{i,:}), \boldsymbol{\theta})}{\partial x_i} \right|.$$

The function $w(A)$ that enforces the acyclicity is expressed as suggested by ? as

$$w(A) := \text{tr} \left((I + \alpha A)^d \right) - d \propto \text{tr} \left(\sum_{k=1}^d \alpha^k A^k \right),$$

where $\alpha \in \mathbb{R}_+$ is a hyper-parameter that avoids exploding values for $w(A)$. In the case of positively valued A , an element (i, j) of $A^k = \underbrace{AA \dots A}_{k \text{ terms}}$ is non-null if and only if there exists a path going from node j to node i that is made of exactly k edges. Intuitively $w(A)$ expresses to which extent the graph is cyclic. Indeed, the diagonal elements (i, i)

of A^k will be as large as there are many paths made of edges that correspond to large values in A from a node to itself in k steps.

In comparison to our work, ? use a quadratic loss on the corresponding linear structural equation model (SEM) as the score function $F(A)$. By attaching normalizing flows to topology learning, our method has a continuously adjustable level of complexity and does not make any strong assumptions on the form of the conditional factors.

6.4.3 Stochastic adjacency matrix

In order to learn the BN topology from the data, the adjacency matrix must be relaxed to contain reals instead of booleans. It also implies that the graph induced by A does not formally define a DAG during training. Work using NO TEARS to perform topology learning directly plug the real matrix A in (6.6) [????] however this is inadequate because the quantity of information going from node j to node i does not continuously relate to the value of $A_{i,j}$. Either the information is null if $A_{i,j} = 0$ or it passes completely if not. Instead, we propose to build the stochastic pseudo binary valued matrix A' from A , defined as

$$A'_{i,j} = \frac{e^{\frac{\log(\sigma(A_{i,j}^2)) + \gamma_1}{T}}}{e^{\frac{\log(\sigma(A_{i,j}^2)) + \gamma_1}{T}} + e^{\frac{\log(1 - \sigma(A_{i,j}^2)) + \gamma_2}{T}}},$$

where $\gamma_1, \gamma_2 \sim \text{Gumbel}(0, 1)$ and $\sigma(a) = 2(\text{sigmoid}(2a^2) - \frac{1}{2})$ normalizes the values of A between 0 and 1, being close to 1 for large values and close to zero for values close to 0. The hyper-parameter T controls the sampling temperature and is fixed to 0.5 in all our experiments. In contrast to directly using the matrix A , this stochastic transformation referred to as the Gumbel-Softmax trick in the literature [??] allows to create a direct and continuously differentiable relationship between the weights of the edges and the quantity of information that can transit between two nodes. Indeed, the probability mass of the random variables $A'_{i,j}$ is mainly located around 0 and 1, and its expected value converges to 1 when $A_{i,j}$ increases.

6.4.4 Optimization

We rely on the augmented Lagrangian approach to solve the constrained optimization problem (6.6) as initially proposed by ?. This optimization procedure requires solving iteratively the following sub-problems:

$$\max_A \mathbb{E}_{\gamma_1, \gamma_2} [F(A)] - \lambda_t w(A) - \frac{\mu_t}{2} w(A)^2, \quad (6.8)$$

where λ_t and μ_t respectively denote the Lagrangian multiplier and penalty coefficients of the sub-problem t .

We solve these optimization problems with mini-batch stochastic gradient ascent. We update the values of γ_t and μ_t as suggested by ? when

the validation loss does not improve for 10 consecutive epochs. Once $w(A)$ equals 0, the adjacency matrix is acyclic up to numerical errors. We recover an exact DAG by thresholding the elements of A while checking for acyclicity with a path finding algorithm. We provide additional details about the optimization procedure used in our experiments in Appendix ??.

6.5 EXPERIMENTS

In this section, we demonstrate some applications of graphical NFs in addition to unifying NFs and BN under a common framework. We first demonstrate how pre-loading a known or hypothesized DAG structure can help finding an accurate distribution of the data. Then, we show that learning the graph topology leads to relevant BNs that support generalization well when combined with ℓ_1 -penalization. Finally, we demonstrate that mono-step normalizing flows made of graphical conditioners are competitive density estimators.

6.5.1 On the importance of graph topology

The following experiments are performed on four distinct datasets, three of which are synthetic, such that we can define a ground-truth minimal Bayesian network, and the fourth is a causal protein-signaling network derived from single-cell data [?]. Additional information about the datasets are provided in Table 6.1 and in Appendix ?. For each experimental run we first randomly permute the features before training the model in order to compare autoregressive and graphical conditioners fairly.

Prescribed topology Rarely do real data come with their associated Bayesian network however oftentimes experts want to hypothesize a network topology and to rely on it for the downstream tasks. Sometimes the topology is known a priori, as an example the sequence of instructions in stochastic simulators can usually be translated into a graph topology (e.g. in probabilistic programming [?]). In both cases, graphical conditioners allow to explicitly take advantage of this to build density estimators while keeping the assumptions made about the network topology valid.

Table 6.2 presents the test likelihood of autoregressive and graphical normalizing flows on the four datasets. The flows are made of a single step and use monotonic normalizers. The neural network architectures are all identical. Further details on the experimental settings as well as additional results for affine normalizers are respectively provided in Appendix ? and ?. We observe how using correct BN structures lead to good test performance in Table 6.2. Surprisingly, the performance on the protein dataset are not improved when using the ground truth graph.

Table 6.1: Datasets description. d =Dimension of the data. V =Number of edges in the ground truth Bayesian Network.

Dataset	d	V	Train	Test
Arithmetic Circuit	8	8	10,000	5,000
8 Pairs	16	8	10,000	5,000
Tree	7	8	10,000	5,000
Protein	11	20	6,000	1,466
POWER	6	≤ 15	1,659,917	204,928
GAS	8	≤ 28	852,174	105,206
HEPMASS	21	≤ 210	315,123	174,987
MINIBOONE	43	≤ 903	29,556	3,648
BSDS300	63	$\leq 1,953$	1,000,000	250,000

Table 6.2: Graphical vs autoregressive conditioners combined with monotonic normalizers. Average log-likelihood on test data over 5 runs, under-scripted error bars are equal to the standard deviation. Results are reported in nats; higher is better. The best performing architecture for each dataset is written in bold. *Graphical conditioners clearly lead to improved density estimation when given a relevant prescribed topology in 3 out of the 4 datasets.*

Conditioner	Graphical	Autoreg.
Arithmetic Circuit	3.99\pm.16	3.06 \pm .38
8 Pairs	-9.40\pm.06	-11.50 \pm .27
Tree	-6.85\pm.02	-6.96 \pm .05
Protein	6.46 \pm .08	7.52\pm.10

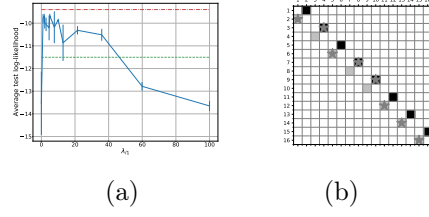


Figure 6.3: **(a)**: Test log-likelihood as a function of ℓ_1 -penalization on **8 pairs** dataset. The upper bound is the average result when given a pre-scribed topology, the lower bound is the result with an autoregressive conditioner. *Learning the right topology leads to better results than autoregressive conditioners.* **(b)**: The blacked cells corresponds to one correct topology of the 8 pairs dataset and the grey cells to the transposed adjacency matrix. The stars denote the edges discovered by the graphical conditioner when trained with $\lambda_{\ell_1} = 4$. *The optimization discovers a relevant BN (equivalent to the ground truth).*

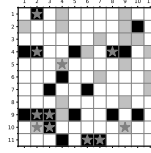


Figure 6.4: The adjacency matrix of the protein interaction network. The blacked cells are the directed connections proposed by domain experts and the grey is the transposed adjacency matrix. The stars denote the edges discovered by the graphical conditioner when trained with $\lambda_{\ell_1} = 12$. *In a realistic setting, the optimization leads to a graph that shares a lot with the one designed by experts.*

We observed during our experiments (see Appendix ??) that learning the topology from this dataset sometimes led to improved density estimation performance with respect to the ground truth graph. The limited dataset size does not allow us to answer if this comes from the limited capacity of the flow and/or from the erroneous assumptions in the ground truth graph. However, we stress out that the graphical flow respects the assumed causal structure in opposition to the autoregressive flow.

Learning the topology The λ_{ℓ_1} coefficient introduced in (6.7) controls the sparsity of the optimized BN, allowing to avoid some spurious connections. We now analyze the effect of sparsity on the density estimation performance. Fig. 6.3a shows the test log-likelihood as a function of the ℓ_1 -penalization on the 8 pairs dataset. We observe that the worst results are obtained when there is no penalization. Indeed, in this case the algorithm finds multiple spurious connections between independent vector’s components and then overfits on these incorrect relationships. Another extreme case shows up when the effect of penalization is too strong, in

this case, the learning procedure underfits the data because it ignores too many relevant connections. It can also be concluded from the plot that the optimal ℓ_1 -penalization performs on par with the ground truth topology, and certainly better than the autoregressive conditioner. Additional results on the other datasets provided in Appendix ?? lead to similar conclusions.

Protein network The adjacency matrix discovered by optimizing (6.8) (with $\lambda_{\ell_1} = 12$) on the protein network dataset is shown in Fig. 6.4. All the connections discovered by the method correspond to ground truth connections. Unsurprisingly, their orientation do not always match with what is expected by the experts. Overall we can see that the optimization procedure is able to find relevant connections between variables and avoids spurious ones when the ℓ_1 -penalization is optimized. Previous work on topology learning such as ??? compare their method to others by looking at the topology discovered on the same dataset. Here we do not claim that learning the topology with a graphical conditioner improves over previous methods. Indeed, we believe that the difference between the methods mainly relies on the hypothesis and/or inductive biased made on the conditional densities. Which method would perform better is dependent on the application. Moreover, the Structural Hamiltonian Distance (SHD) and the Structural Inference Distance (SID), often used to compare BNs, are not well motivated in general. On the one hand the SHD does not take into account the independence relationships modeled with Bayesian networks, as an example the SHD between the graph found on 8 pairs dataset of Fig. 6.3b and one possible ground truth is non zero whereas the two BNs are equivalent. On the other hand the SID aims to compare causal graphs whereas the methods discussed here do not learn causal relationships. In general BN topology identification is an ill posed problem and thus we believe using these metrics to compare different methods without additional downstream context is irrelevant. However, we conclude from Fig. 6.3b and Fig. 6.4 that the scores computed with graphical NFs can be used to learn relevant BN structures.

6.5.2 Density estimation benchmark

In these experiments, we compare autoregressive, coupling and graphical conditioners with no ℓ_1 -penalization on benchmark tabular datasets as introduced by ? for density estimation. See Table 6.1 for a description. We evaluate each conditioner in combination with monotonic and affine normalizers. We only compare NFs with a single transformation step because our focus is on the conditioner capacity. We observed during preliminary experiments that stacking multiple conditioners improves the performance slightly, however the gain is marginal compared to the loss of interpretability. To provide a fair comparison we have fixed in

Table 6.3: Average log-likelihood on test data over 3 runs, under-scripted error bars are equal to the standard deviation. Results are reported in nats; higher is better. The best performing architecture per category for each dataset is written in bold. (a) 1-step affine normalizers (b) 1-step monotonic normalizers. *Graphical normalizing flows outperform coupling and autoregressive architectures on most of the datasets.*

Dataset	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
Coup.	$-5.60 \pm .00$	$-3.05 \pm .01$	$-25.74 \pm .01$	$-38.34 \pm .02$	$57.33 \pm .00$
(a) Auto.	$-3.55 \pm .00$	$-0.34 \pm .01$	$-21.66 \pm .01$	$-16.70 \pm .05$	$63.74 \pm .00$
Graph.	$-2.80 \pm .01$	$1.99 \pm .02$	$-21.18 \pm .07$	$-19.67 \pm .06$	$62.85 \pm .07$
Coup.	$0.25 \pm .00$	$5.12 \pm .03$	$-20.55 \pm .04$	$-32.04 \pm .12$	$107.17 \pm .46$
(b) Auto.	$0.58 \pm .00$	$9.79 \pm .04$	$-14.52 \pm .16$	$-11.66 \pm .02$	$151.29 \pm .31$
Graph.	$0.62 \pm .04$	$10.15 \pm .15$	$-14.17 \pm .13$	$-16.23 \pm .52$	$155.22 \pm .11$

advance the neural architectures used to parameterize the normalizers and conditioners as well as the training parameters by taking inspiration from those used by ? and ?. The variable ordering of each dataset is randomly permuted at each run. All hyper-parameters are provided in Appendix ?? and a public implementation will be released on Github.

First, Table 6.3 presents the test log-likelihood obtained by each architecture. These results indicate that graphical conditioners offer the best performance in general. Unsurprisingly, coupling layers show the worst performance, due to the arbitrarily assumed independencies. Autoregressive and graphical conditioners show very similar performance for monotonic normalizers, the latter being slightly better on 4 out of the 5 datasets. Table 6.4 contextualizes the performance of graphical normalizing flows with respect to the most popular normalizing flow architectures. Comparing the results together, we see that while additional steps lead to noticeable improvements for affine normalizers (MAF), benefits are questionable for monotonic transformations. Overall, graphical normalizing flows made of a single transformation step are competitive with the best flow architectures with the added value that they can directly be translated into their equivalent BNs. *From these results we stress out that single step graphical NFs are able to model complex densities on par with SOTA while they offer new ways of introducing domain knowledge.*

Second, Table 6.5 presents the number of edges in the BN associated with each flow. For POWER and GAS, the number of edges found by the graphical conditioners is close or equal to the maximum number of edges. Interestingly, graphical conditioners outperform autoregressive conditioners on these two tasks, demonstrating the value of finding an appropriate ordering particularly when using affine normalizers. More-

Table 6.4: Average log-likelihood on test data over 3 runs, under-scripted error bars are equal to the standard deviation. Results are reported in nats, higher is better. The results followed by a star are copied from the literature and the number of steps in the flow is indicated in parenthesis for each architecture. *Graphical normalizing flows reach density estimation performance on par with the most popular flow architectures whereas it is only made of 1 transformation step.*

Dataset	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
Graph.-UMNN (1)	0.62 \pm .04	10.15 \pm .15	-14.17 \pm .13	-16.23 \pm .52	155.22 \pm .11
MAF (5)	0.14 \pm .01	9.07 \pm .01	-17.70 \pm .01	-11.75 \pm .22	155.69 \pm .14
Glow* (10)	0.42 \pm .01	12.24 \pm .03	-16.99 \pm .02	-10.55 \pm .45	156.95 \pm .28
UMNN-MAF* (5)	0.63 \pm .01	10.89 \pm .70	-13.99 \pm .21	-9.67 \pm .13	157.98 \pm .01
Q-NSF* (10)	0.66 \pm .01	12.91 \pm .01	-14.67 \pm .02	-9.72 \pm .24	157.42 \pm .14
FFJORD* (5-5-10-1-2)	0.46 \pm .01	8.59 \pm .12	-14.92 \pm .08	-10.43 \pm .04	157.40 \pm .19

Table 6.5: Rounded average number of edges (over 3 runs) in the equivalent Bayesian network. *The graphical conditioners lead to sparser BNs compared to autoregressive conditioners.*

Dataset	P	G	H	M	B
Graph.-Aff.	15	26	152	277	471
Graph.-Mon.	15	27	159	265	1594
Coupling	9	16	110	462	992
Autoreg.	15	28	210	903	1953

over, graphical conditioners correspond to BNs whose sparsity is largely greater than for autoregressive conditioners while providing equivalent if not better performance. The depth [?] of the equivalent BN directly limits the number of steps required to inverse the flow. Thus sparser graphs that are inevitably shallower correspond to NFs for which sampling, i.e. computing their inverse, is faster.

6.6 DISCUSSION

Cost of learning the graph structure The attentive reader will notice that learning the topology does not come for free. Indeed, the Lagrangian formulation requires solving a sequence of optimization problems which increases the number of epochs before convergence. In our experiments we observed different overheads depending on the problems, however in general the training time is at least doubled. This does not impede the practical interest of using graphical normalizing flows.

The computation overhead is more striking for non-affine normalizers (e.g. UMMNs) that are computationally heavy. However, we observed that most of the time the topology recovered by affine graphical NFs is relevant. It can thus be used as a prescribed topology for normalizers that are heavier to run, hence alleviating the computation overhead. Moreover, one can always hypothesize on the graph topology but more importantly the graph learned is usually sparser than an autoregressive one while achieving similar if not better results. The sparsity is interesting for two reasons: it can be exploited for speeding up the forward density evaluation; but more importantly it usually corresponds to shallower BNs that can be inverted faster than autoregressive structures.

Bayesian network topology learning Formal BN topology learning has extensively been studied for more than 30 years now and many strong theoretical results on the computational complexity have been obtained. Most of these results however focus on discrete random variables, and how they generalize in the continuous case is yet to be explained. The topic of BN topology learning for discrete variables has been proven to be NP-hard by ?. However, while some greedy algorithms exist, they do not lead in general to a minimal I-map although allowing for an efficient factorization of random discrete vectors distributions in most of the cases. These algorithms are usually separated between the constrained-based family such as the *PC algorithm* [?] or the *incremental association Markov blanket* [?] and the score-based family as used in the present work. Finding the best BN topology for continuous variables has not been proven to be NP-hard however the results for discrete variables suggest that without strong assumptions on the function class the problem is hard.

The recent progress made in the continuous setting relies on the heuristic used in score-based methods. In particular, ? showed that the acyclicity constraint required in BNs can be expressed with NO TEARS, as a continuous function of the adjacency matrix, allowing the Lagrangian formulation to be used. ? proposed DAG-GNN, a follow up work of ? which relies on variational inference and auto-encoders to generalize the method to non-linear structural equation models. Further investigation of continuous DAG learning in the context of causal models was carried out by ?. They use the adjacency matrix of the causal network as a mask over neural networks to design a score which is the log-likelihood of a parameterized normal distribution. The requirement to pre-define a parametric distribution before learning restricts the factors to simple conditional distributions. In contrast, our method combines the constraints given by the BN topology with NFs which are free-form universal density estimators. Remarkably, their method leads to an efficient one-pass computation of the joint density. This neural masking scheme can also be implemented for NF architectures such as already demonstrated by ? and ? for autoregressive conditioners.

Shuffling between transformation steps As already mentioned, consecutive transformation steps are often combined with randomly fixed permutations in order to mitigate the ordering problem. Linear flow steps [?] and 1x1 invertible convolutions [?] generalize these fixed permutations. They are parameterized by a matrix $W = PLU$ where P is the fixed permutation matrix, and L and U are respectively a lower and an upper triangular matrix. Although linear flow improves the simple permutation scheme, they do still rely on an arbitrary permutation. To the best of our knowledge, graphical conditioners are the first attempt to get completely rid of any fixed permutation in NFs.

Inductive bias Graphical conditioners eventually lead to binary masks that model the conditioning components of a factored joint distribution. In this way, the conditioners process their input as they would process the full vector. We show experimentally in Appendix ?? that this effectively leads to good inductive bias for processing images with NFs. In addition, we have shown that normalizing flows built from graphical conditioners combined with monotonic transformations are expressive density estimators. In effect, this means that enforcing some a priori known independencies can be performed thanks to graphical normalizing flows without hurting their modelling capacity. We believe such models could be of high practical interest because they cope well with large datasets and complex distributions while preserving some readability through their equivalent BN.

Close to our work, ? improve amortized inference by prescribing a BN structure between the latent and observed variables into a FFJORD NF, once again showing the interest of using the potential BN knowledge. Similar to our work, ? see causal autoregressive flows as structural equation modelling. They show bivariate autoregressive affine flows can be used to identify the causal direction under mild conditions. Under similar mild conditions, discovering causal relationships with graphical normalizing flows could well be an exciting research direction.

Conclusion We have revisited coupling and autoregressive conditioners for normalizing flows as Bayesian networks. From this new perspective, we proposed the more general graphical conditioner architecture for normalizing flows. We have illustrated the importance of assuming or learning a relevant Bayesian network topology for density estimation. In addition, we have shown that this new architecture compares favorably with autoregressive and coupling conditioners and on par to the most common flow architectures on standard density estimation tasks even without any hypothesized topology. One interesting and straightforward extension of our work would be to combine it with normalizing flows for discrete variables. We also believe that graphical conditioners could be used when the equivalent Bayesian network is required for downstream tasks such as in causal reasoning.

6.7 EPILOGUE

6.7.1 *On the relationship between smoothness and independence*

OUTLINE

Hybrid modelling reduces the misspecification of expert models by combining them with machine learning (ML) components learned from data. Like for many ML algorithms, hybrid model performance guarantees are limited to the training distribution. Leveraging the insight that the expert model is usually valid even outside the training domain, we overcome this limitation by introducing a hybrid data augmentation strategy termed *expert augmentation*. Based on a probabilistic formalization of hybrid modelling, we show why expert augmentation improves generalization. Finally, we validate the practical benefits of augmented hybrid models on a set of controlled experiments, modelling dynamical systems described by ordinary and partial differential equations.

7.1 INTRODUCTION

Generalization to unseen data is a key property of a useful model. When training and test data are independently and identically distributed (IID), one way to check generalization is by evaluating the model on a held out subset of the training data or with k-fold cross validation. Unfortunately, this setting is often unrealistic because the training scenario is rarely fully representative of the test scenario. This has motivated lot of recent research efforts to focus on the robustness of ML models [???]. Common strategies can be broadly grouped in two categories: The first class of methods aims at aligning specific properties of the model (e.g., invariance, equivariance, monotonicity, etc.) with expertise on the problem of interest [????]. The second category is data focused [????] and leverages variations present in the training data, e.g. by minimizing worst case sub-group performance, to achieve robustness.

The data oriented methods, which include Group-DRO [?] and Invariant Risk Minimization [?, IRM], can be very appealing because they only require implicit specification of invariances via domains or environments. However, these methods' performance is limited to variations present in the training data and the inductive bias of the ML algorithm. This may be insufficient when the modelling problem is too complex or the variations of interest are not present in the training data. On the other hand, methods based on domain-specific expertise do not suffer from such limitations. Embedding expertise into a model can be done via architectural inductive biases [??], data augmentation [?], or a

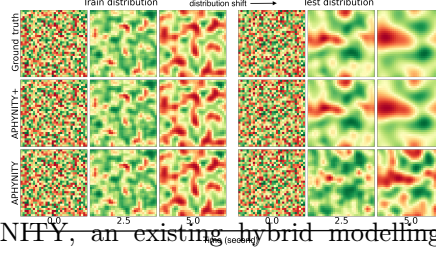


Figure 7.1: APHYNITY, an existing hybrid modelling strategy, is unable to predict accurately the dynamic of a 2D diffusion reaction for a shifted test distribution although it predicts well configuration that follows the training distribution. On the opposite, APHYNITY+, the same model fine-tuned with our data augmentation, generalizes to shifted distributions as expected from the validity of the underlying physics.

learning objective [?] that enforces established symmetries of the problem. As an example, simple data augmentation techniques combined with convolutions lead to excellent performance on natural image problems [?]. Another natural approach to embed expertise in ML models, and the one studied in this paper, is called hybrid learning (HyL). HyL combines an expert model (e.g., physics-motivated equations) with a learned component that improves the expert model so that the combination better fits real-world data. A particularity of HyL is the central role played by the expertise, which is supposed to provide a simple and well-grounded parametric description of the process considered. HyL usually considers the expert model as an analytical function, or as a set of equations, that relates the expert parameters to the quantity of interest. The expert model is often motivated by the underlying physics of the system considered. Hence, we will use the terms *expert* model and *physical* model interchangeably.

In recent work [?????], HyL demonstrated success in complementing partial physical models and improving the inference of the corresponding parameters. However, contrarily to the common belief that HyL achieves better generalization than black box ML models, we argue that hybrid models do not meet their promise regarding robustness. Although HyL achieves strong performance on IID test distributions by exploiting the inductive bias of the expert models, we show that their performance collapses when the test domain is not included in the training domain. This is unsatisfactory as the expert model is typically well-defined for a range of parameters that can correspond to realistic data far outside of the training distribution. A test distribution not covered by the training data, but for which an expert model exists, happens often in the real world. As an example, ? apply HyL to a pharmacological model describing the effect of a COVID-19 treatment for which only a limited quantity of real-world data is available. In this context, although the underlying biochemical dynamic of treatments is well modelled, data is often scarce and biased. Therefore, the hybrid model does not necessarily generalize to configurations that are well modelled by the pharmacological model but unseen during training.

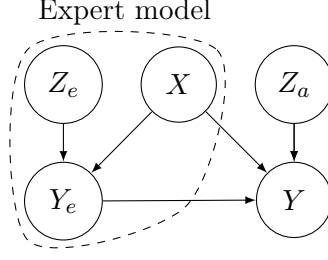


Figure 7.2: A hybrid probabilistic model which describes the relationship between the input X and the output Y for a configuration of the system as defined by the latent variables Z_e and Z_a . The prescribed expert model defines the conditional density $p(y_e|z_e, x)$, where Y_e is an approximation of Y . Hybrid learning aims at learning the conditional distribution $p(y|z_a, y_e, x)$.

We introduce *expert augmentations* for training augmented hybrid models (AHMs), a procedure that extends the range of validity of hybrid models and improves generalization as pictured by Fig. 7.1. Our contribution is to first formalise the HyL problem as: 1) Learning a probabilistic model partially defined by the expert model; 2) Performing inference over this probabilistic hybrid model. In this context, we show that HyL is vulnerable to distribution shifts for which the expert model is well defined (see Figure 7.1, bottom row). Motivated by our analysis, we propose to fine-tune the hybrid model on an expert-augmented dataset that includes distribution shifts (see results of augmentation in Figure 7.1, middle row). These expert augmentations only rely on the hybrid model itself, leveraging that the expert model is also well-defined outside of the training distribution. Our experiments on various controlled HyL problems demonstrate that AHMs achieve multiple orders of magnitude superior generalization in realistic situations and can be applied to any state-of-the-art HyL algorithm.

7.2 HYBRID LEARNING

In order to show that our proposed expert augmentations lead to robust models, we first formalize hybrid learning with the probabilistic model depicted in Fig. 7.2. In this Bayesian network, capital letters denote random variables (e.g., Y) and, in the following, we will use calligraphic letters for the domain of the corresponding realization (e.g., $y \in \mathcal{Y}$). In our formalism, the expert model is a conditional density $p(y_e|x, z_e)$ that describes the distribution of the *expert* response Y_e to an input x together with a parametric description of the system z_e , denoting expert or physical parameters. We augment the expert model with the *interaction model* which is a conditional distribution $p(y|x, y_e, z_a)$ that describes the distribution of the observation Y given the input x , the expert model response y_e , and a parametric description of the interaction model z_a .

Our final goal is to create a robust predictive model $p(y|x, (x_o, y_o))$ of the random variable Y , given the input x together with independent observations (x_o, y_o) of the same system, where the subscript o denotes an observed quantity. As a concrete example, we consider predicting the evolution of a damped pendulum (described in Section 7.4.1) given its initial angle and speed ($x = [\theta, \dot{\theta}]$) and a sequence of observations of the same pendulum. The expert model we assume is able to describe a frictionless pendulum whose dynamic is only characterized by one parameter $z_e := \omega_0$, denoting its fundamental frequency. A perfect description of the system should model the friction with a second parameter $z_a := \alpha$, the damping factor. In this problem, (x_o, y_o) and (x, y) are IID realization of the same pendulum which corresponds, in general terms, to samples from $p(x, y|z_a, z_e)$ for some fixed but unknown values of z_a and z_e . The expert variables z_e (e.g., ω_0) together with z_a (e.g., α) should accurately describe the system that produces Y (e.g., the evolution of the pendulum's angle and speed along time) from X (e.g., the initial pendulum's state). In our setting we assume that we are given a pair (x_o, y_o) (e.g., past observations) from which we can accurately infer the state of the system (z_a, z_e) as described by the interaction and expert models, and then predict the distribution of Y for a given input x (e.g., forecasting future observations) to the same system. Because the interaction between z_e and y is essentially defined by the expert model, it should be possible, and preferable, to learn an accurate predictive model of Y whose accuracy is independent from the training distribution of the expert variables z_e . Provided all probability distributions in Fig. 7.2 are known, the Bayes optimal hybrid predictor p_B can be written as

$$p_B(y|x, (x_o, y_o)) = \mathbb{E}_{p(z_a, z_e|(x_o, y_o))} [p(y|x, z_a, z_e)]. \quad (7.1)$$

We observe that the Bayes optimal predictor explicitly depends on the posterior $p(z_a, z_e|(x_o, y_o))$ which is itself a function of the marginal distribution over z_e . This may preclude the existence of a good predictor that is invariant to shift of $p(z_e)$. However, in the following we will consider that the pair (x_o, y_o) contains enough information about the parameters z_a, z_e . As a consequence, the posterior distribution shrinks around the correct parameters value and the effect of the prior becomes negligible.

7.2.1 Hybrid generative modelling

We consider expert models that are deterministic; that is, for which $p_\theta(y_e|x, z_e)$ is a Dirac distribution. The expert model describes the system as a function $f_e : \mathcal{X} \times \mathcal{Z}_e \rightarrow \mathcal{Y}_e$ that computes the response y_e to an input x , parameterized by expert variables z_e . The goal of hybrid modelling is to augment the expert model with a learned component from data as depicted in Fig. 7.2. Formally, given a dataset

$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ of N IID samples, we aim to learn the interaction model $p_\theta(y|x, y_e, z_a)$ that fits the data well but is close to the expert model. For example, we could define closeness via a small L2-distance between expert and hybrid outputs or via a small Kullback-Leibler (KL) divergence between the marginal distributions of Y and Y_e .

Learning a model that is close to the expert model and fits the training data well is a hard problem. However, the APHYNITY algorithm [?] and the Hybrid-VAE [?, HVAE] are two recent approaches that offer promising solutions to this problem. We now briefly describe these two methods and how they can be used to approximate the Bayes optimal predictor of (7.1). Our augmentation strategy is compatible (and effective) with both approaches.

APHYNITY. [?] formulate hybrid learning in a context where the expert model is an ordinary differential equation (ODE). They consider an additive hybrid model that should perfectly fit the data, which is equivalent to assuming the conditional distribution $p_\theta(y|x, y_e, z_a)$ is a Dirac distribution. Formally, they solve the optimization problem

$$\begin{aligned} \min_{z_e, F_a} \|F_a\| \quad \text{s.t.} \quad \forall (x, y) \in \mathcal{D}, \forall t, \frac{dy_t}{dt} = (F_e + F_a)(y_t) \\ \text{with } y_0 := x, \end{aligned} \quad (7.2)$$

where $\|\cdot\|$ is a norm operator on the function space, $F_a : \mathcal{Y}_t \times \mathcal{Z}_a \rightarrow \mathcal{Y}_t$ is a learned function, $F_e : \mathcal{Y}_t \times \mathcal{Z}_e \rightarrow \mathcal{Y}_t$ defines the expert model and \mathcal{D} is a dataset of initial states $x := y_0$ and sequences $y \in \mathcal{Y} := (\mathcal{Y}_t)^k$, where k is the number of observed timesteps. APHYNITY solves this problem with Lagrangian optimization and Neural ODEs [?] to compute derivatives. In the context of ODEs, the random variable X is the initial state of the system at t_0 and Y is the observed sequence of k states between t_0 and t_1 .

This formulation only considers learning a missing dynamic for one realization of the system described by Fig. 7.2, for a single z_a and z_e . However, we are interested in learning a hybrid model that works for the full set of systems described by Fig. 7.2. As suggested in [?], we use an encoder network $g_\psi(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}_a \times \mathcal{Z}_e$ that corresponds to a Dirac distribution located at g_ψ as the approximate posterior $q_\psi(z_a, z_e|x, y)$. The interaction model is a product of Dirac distributions whose locations correspond to the solution of the ODE

$$\frac{dy_t}{dt} = F_e(y_t, z_e) + F_a(y_t, z_a; \theta), \quad y_0 := x. \quad (7.3)$$

Hence the corresponding approximate Bayes predictor replaces the parameters (z_a, z_e) in (7.3) with the prediction of g_ψ and predicts a product of Dirac distributions.

Hybrid-VAE (HVAE). In contrast to APHYNITY, the model proposed by [?] is not limited to additive interactions between the expert

model and the ML model, nor to ODEs. Instead, their goal is to learn the generative model described by Fig. 7.2. They achieve this with a variational auto-encoder (VAE) where the decoder specifically follows Fig. 7.2. Similarly to the amortized APHYNITY model, the encoder $g_\psi(x, y)$ predicts a posterior distribution over z_a and z_e , and the model is trained with the classical Evidence Lower Bound on the likelihood (ELBO). We observe that relying only on an architectural inductive bias and maximum likelihood training is not enough to ground the generative model to the expert equations. They propose to add three regularizers R_{PPC} , $R_{DA,1}$, and $R_{DA,2}$ that encourage the generative model to rely on the expert model. The final objective is

$$\begin{aligned} \max_{\theta, \psi} \mathbb{E}_{\mathcal{D}} [\text{ELBO}((x, y); \psi, \theta)] + \alpha R_{PPC} + \beta R_{DA,1} \\ + \gamma R_{DA,2}. \end{aligned} \quad (7.4)$$

The first regularizer, R_{PPC} , encourages the marginal distribution of samples generated by the complete model to be close to the marginal distribution that would be only generated by the physical model. The two other regularizers specifically require the encoder network for z_e to be made of two sub-networks. The first network filters the observations to keep only what can be generated by the expert model alone, and the second should map the filtered observations to the posterior distribution over z_e . $R_{DA,1}$ penalizes the objective if the observations generated by the expert model are not close to the filtered observations. Finally, $R_{DA,2}$ relies on data augmentation with the expert model to enforce that the second sub-network correctly identifies the expert variables z_e when the observations are correctly filtered. We refer the reader to [?] for more details on HVAE. For HVAE, the approximate predictor takes the form described by (7.1) where $p(z_a, z_e | (x_o, y_o))$ is approximated by the encoder $q_\psi(z_a, z_e | x, y)$ and $p(y | x, z_e, z_a)$ by the learned hybrid generative model.

7.3 ROBUST HYBRID LEARNING

We now formalize our definition of out of distribution (OOD) and robustness. In general, a test scenario is OOD if the joint test distribution $\tilde{p}(x, y)$ is different from the training distribution $p(x, y)$, that is $d(\tilde{p}, p) > 0$ for any properly defined divergence or distance d . In the following, we reduce our discussion to a sub-class of distribution shifts for which the marginal train and test distributions over z_e may be different, $d(p(z_e), \tilde{p}(z_e)) > 0$, but the marginals of z_a and x are constant. As a consequence, the joint distribution of (x, y) pairs is also shifted. Formally, the training and test distributions are respectively defined as

$$\begin{aligned} p(x, y) &:= \mathbb{E}_{p(z_e)p(z_a)p(y_e|z_e,x)} [p(x)p(y|z_a, x, y_e)], \\ \tilde{p}(x, y) &:= \mathbb{E}_{\tilde{p}(z_e)p(z_a)p(y_e|z_e,x)} [p(x)p(y|z_a, x, y_e)]. \end{aligned}$$

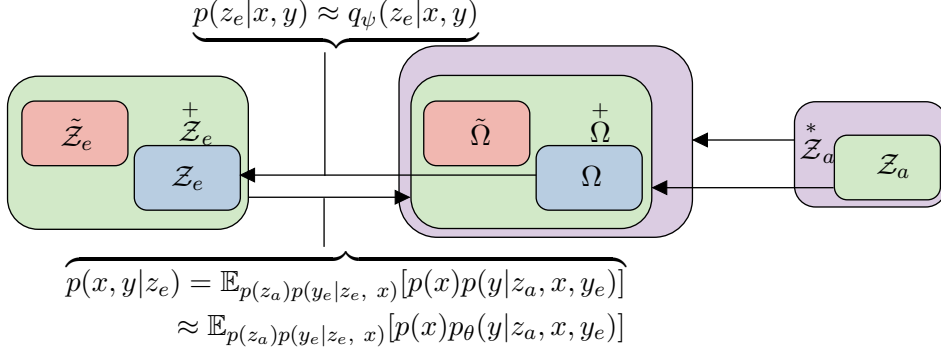


Figure 7.3: Visualization of the distribution shifts considered in this work. The train support Ω of (x, y) results from $(z_a, z_e) \in \mathcal{Z}_e \times \mathcal{Z}_a$. The test supports (in red) are denoted with a tilde symbols as \tilde{Z}_e for z_e and $\tilde{\Omega}$ for (x, y) . The augmented support $\tilde{\Omega}^+$ (in green) includes both train and test scenarios and corresponds to $(z_a, z_e) \in \tilde{\mathcal{Z}}_e \times \mathcal{Z}_a$. The outer violet domain that includes $\tilde{\Omega}^+$ depicts one of our experiment in which the domain of z_a is also shifted. Hybrid modelling algorithms alone may learn a mapping $p_\theta : \tilde{\mathcal{Z}}_e \rightarrow \tilde{\Omega}^+$ but augmentation is necessary to learn the inverse mapping $q_\psi : \tilde{\Omega}^+ \rightarrow \tilde{\mathcal{Z}}_e$.

In this context, we demonstrate, theoretically and empirically, that classical hybrid models fail. To address this failure, we introduce *augmented hybrid models* and show that, under some assumptions, they achieve optimal performance on both the train and test distributions.

Our goal is to learn a predictive model

$$p_{\theta, \psi}(y|x, (x_o, y_o)) = \mathbb{E}_{q_\psi(z_a, z_e|x_o, y_o)} [p_\theta(y|y_e, x, z_a)]$$

$p(y_e|z_e, x)$

that is *exact* on both the train and test domains when they follow the aforementioned training and testing distribution shifts. We say that a learned predictive model $\hat{p}(a|b)$ is \mathcal{E} -*exact*, or *exact* on the sample space \mathcal{E} , if $\hat{p}(a|b) = p(a|b) \quad \forall (a, b) \in \mathcal{E}$. Here we qualify a predictive model as *robust* to a test scenario if its *exactness* on the training domain is sufficient to ensure exactness on the test domain.

We now define an augmented distribution $\tilde{p}^+(z_e)$ over the expert variables whose support $\tilde{\mathcal{Z}}_e^+$ includes the joint support $\mathcal{Z}_e \cup \tilde{\mathcal{Z}}_e$ between the train and test distribution of the physical parameters. As depicted in Fig. 7.3, we denote the corresponding support over the observation space $\mathcal{X} \times \mathcal{Y}$ as $\tilde{\Omega}^+, \Omega$, and $\tilde{\Omega}$, respectively. In this context, and with A1-, we may demonstrate that even under perfect learning, classical hybrid learning algorithms do not produce an $\tilde{\Omega}$ -*exact* predictor while our augmentation strategy does.

Assumption A1-. *Hybrid modelling learns an interaction model $p_\theta(y|y_e, x, z_a)$ that is $\tilde{\Omega}^+$ -exact.*

Although strong, [A1-](#) is consistent with the recent literature on hybrid modelling, which assumes that $p(y_e|x, z_e)$ is an accurate description of the system, thereby $p_\theta(y|y_e, x, z_a)$ should not be overly complex. As an example, we consider an additive interaction model in our experiments for which extrapolation to unseen y_e holds if this assumption is correct. That said, we still notice that the exactness of the interaction model p_θ on Ω^+ is insufficient to prove that the predictive model $p_{\theta,\psi}$ is Ω^+ -exact. Indeed, the encoder q_ψ is only trained on the training data and cannot rely on a strong inductive bias in contrast to p_θ . Thus, even if the encoder is exact on the training distribution, the corresponding predictive model does not achieve exactness outside Ω .

7.3.1 Expert augmentation

We propose a data augmentation strategy to improve the robustness of hybrid models to unseen test scenarios. Once trained, the hybrid model is composed of an encoder q_ψ and an interaction model p_θ that are respectively Ω - and Ω^+ -exact. We may create a new training distribution with a support over Ω^+ by sampling physical parameters z_e from a distribution that covers \mathcal{Z}_e^+ . We can then train the encoder q_ψ on Ω^+ , under perfect training the corresponding predictive model $p_{\theta,\psi}(y|x, (x_o, y_o))$ is Ω^+ -exact, hence exact on both train and test domains.

Our learning strategy is grounded in existing hybrid modelling algorithms. Here, we focus on APHYNITY and HVAE, but our approach is applicable to other HyL algorithms. We first train an encoder q_ψ and a decoder p_θ with a HyL algorithm. Together with experts we then decide on a realistic distribution $\bar{p}(z_e)$ and create a new dataset $\bar{\mathcal{D}}$ by sampling from the hybrid generative model defined by Fig. 7.2 and the interaction model p_θ . A notable difference between the augmented training set $\bar{\mathcal{D}}$ and the original training set \mathcal{D} is that the former contains ground truth values for the expert’s variables z_e . As we assume that the interaction model is Ω^+ -exact, we freeze it and only fine-tune the encoder q_ψ on $\bar{\mathcal{D}}$. We use a combination of the loss function ℓ of the original HyL algorithm (e.g., (7.4) for HVAE, and the Lagrangian of (7.2) for APHYNITY) and a supervision on the latent variable objective to learn a decoder that solves

$$\psi = \arg \min_{\psi} \mathbb{E}_{\bar{\mathcal{D}}} [\ell(x, y; \theta, \psi) - \log q_\psi(z_e|x, y)].$$

In our experiments we chose a Gaussian model for the posterior, which is equivalent to a mean square error (MSE) loss on the physical parameters. We provide a detailed description of the expert augmentation scheme in Appendix ??.

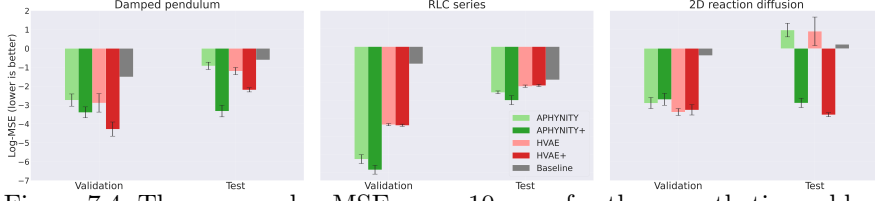


Figure 7.4: The average log-MSEs over 10 runs for three synthetic problems on the validation and test sets. We compare HVAE (in red), in light colours, to their expert augmented versions HVAE+ and APHYNITY+, in darker colours. *On the test sets, AHMs outperform the original models, and by a large margin on the pendulum and diffusion problems. Moreover, augmentation conserves the relatively good performance on the validation set (IID w.r.t. the training set).*

As a side note, we would like to emphasize the difference between the data augmentation proposed in this paper and the one from ?. While HVAE also requires to sample new physical parameters z_e , it is only to ensure that a sub-part of the encoder is able to infer correctly z_e given y_e . This augmentation does not contribute to robustness distribution shifts on y in contrast to ours.

7.4 EXPERIMENTS

7.4.1 Problem description

We assess the benefits of expert augmentation on three controlled problems described and simulated by the ODE

$$\frac{dy_t}{dt} = F_e(y_t; z_e) + F_a(y_t; z_a), \quad (7.5)$$

where $F_e : \mathcal{Y}_t \times \mathcal{Z}_e \rightarrow \mathcal{Y}_t$ is the expert model and $F_a : \mathcal{Y}_t \times \mathcal{Z}_a \rightarrow \mathcal{Y}_t$ complements it. In our notation X is the initial state y_0 and the response Y is the sequence of states $y_{1:t_1} := [y_{i\Delta t}]_{i=1}^{t_1/\Delta t}$. For all experiments we train the models to maximize $p_{\theta,\psi}(y = y_{1:t_1} | x = y_0)$ on the training data. We validate and test the models on the predictive distribution $p(y = y_{1:t_2} | x = y_0, x_o = y_0, y_o = y_{1:t_1})$, where $t_2 > t_1$ assesses the generalization over time. A brief description of the different problems is provided below.

The damped pendulum is often used as an example in the hybrid modelling literature [??]. The system’s state at time t is $y_t = [\theta_t \quad \dot{\theta}_t]^T$, where θ_t is the angle of the pendulum at time t and $\dot{\theta}_t$ its angular speed. The evolution of the state over time is described by (7.5), where $z_e := \omega$, $z_a = \alpha$ and

$$F_e := \begin{bmatrix} \dot{\theta} & -\omega_0^2 \sin \theta \end{bmatrix}^T \quad \text{and} \quad F_a := \begin{bmatrix} 0 & -\alpha \dot{\theta} \end{bmatrix}^T. \quad (7.6)$$

The corresponding systems are defined by the damping factor α and ω_0 , the fundamental frequency of the pendulum.

The RLC series circuits are electrical circuits made of 3 electrical components that may model a large range of transfer functions. These models are often used in biology (e.g., the Hodgkin-Huxley class of models [?], in photoplethysmography [?]) and in electrical engineering to model the dynamics of various systems. The system’s state at time t is $y_t = \begin{bmatrix} U_t & I_t \end{bmatrix}^T$, where U_t is the voltage around the capacitance and I_t the current in the circuit. The evolution of the state over time is described by (7.5), where $z_e := \{L, C\}$, $z_a = \{R\}$ and

$$F_e := \begin{bmatrix} \frac{I_t}{C} \\ \frac{1}{L}(V(t) - U_t) \end{bmatrix} \quad \text{and} \quad F_a := \begin{bmatrix} 0 \\ -\frac{R}{C}I_t \end{bmatrix}. \quad (7.7)$$

The dynamics described by the RLC circuit is more diverse than for the pendulum and the system can be hard to identify. This system is characterised by the resistance R , capacitance C , and inductance L , provided $V(t)$ is known.

The 2D reaction diffusion was used by ? to assess the quality of APHYNITY. It is a 2D FitzHugh-Nagumo on a 32×32 grid. The system’s state at time t is a $2 \times 32 \times 32$ tensor $y_t = \begin{bmatrix} u_t & v_t \end{bmatrix}^T$. The evolution of the state over time is described by (7.5), where $z_e := \{a, b\}$, $z_a = \{k\}$ and

$$F_e := \begin{bmatrix} a\Delta u_t \\ b\Delta v_t \end{bmatrix} \quad \text{and} \quad F_a := \begin{bmatrix} R_u(u_t, v_t; k) \\ R_v(u_t, v_t) \end{bmatrix}, \quad (7.8)$$

where Δ is the Laplace operator, the local reaction terms are $R_u(u, v; k) = u - u^3 - k - v$ and $R_v(u, v) = u - v$. This model is interesting to study as it considers a state space for which neural architectures may have a real advantage compared to other ML models.

In the following experiments we analyze the effect of our data augmentation strategy on APHYNITY and HVAE. All models explicitly use the assumption that the interaction model follows the structure of (7.5). For each problem the validation and test sets are respectively IID and OOD with respect to the training distribution. The best models are always selected based on validation performance, that is with samples from Ω . We provide additional details on the different expert models, dataset creation, and neural networks architectures in Appendix ??.

7.4.2 Results

Performance gain from augmentation. *This experiment demonstrates that HVAE and APHYNITY are not robust to OOD test scenarios in opposition to the corresponding AHMs, as shown in Fig. 7.1 for the 2D diffusion problem and in Appendix ?? for the two other problems. We emphasize that our intention is not to declare a winner between HVAE*

Dataset		APH.	HVAE	APH.+	HVAE+
Pendulum	Valid.	$6_{\pm 2}$	$3_{\pm 1}$	$6_{\pm 2}$	$2_{\pm 1}$
	Test	$66_{\pm 9}$	$117_{\pm 10}$	$10_{\pm 4}$	$11_{\pm 2}$
RLC	Valid.	$6_{\pm 3}$	$38_{\pm 2}$	$7_{\pm 5}$	$28_{\pm 1}$
	Test	$17_{\pm 3}$	$25_{\pm 2}$	$5_{\pm 2}$	$12_{\pm 1}$
Diffusion	Valid.	$2_{\pm 0}$	$2_{\pm 0}$	$2_{\pm 0}$	$2_{\pm 0}$
	Test	$27_{\pm 2}$	$32_{\pm 10}$	$3_{\pm 1}$	$2_{\pm 0}$

Table 7.1: Comparison of mean relative precision (in %, \pm indicates one standard deviation) over 10 runs of predicted physical parameters of different hybrid modelling strategies in validation and OOD test settings. Augmented versions are denoted with a +. *While the accuracy of APHYNITY and HVAE is good on the validation set, it collapses on the OOD test set. On the opposite, the augmented versions perform well on both validation and test sets.*

and APHYNITY. Indeed, both algorithms have already demonstrated superior performance than black box ML models. Hence, we only report a very simple baseline that is the mean value of the signals. We want to compare performance in OOD settings and empirically validate the benefit of AHMs. We compare the predictive performance in Fig. 7.4 (see ?? for the raw numbers). Although classical hybrid learning strategies do very well on the IID validation set, they exhibit poor generalization on OOD test sets for all three problems. We also observe some disparity between APHYNITY and HVAE. In addition to different learning strategies, this is probably due to differences in the networks’ architectures as they were respectively inspired from the corresponding pendulum experiment in each paper. However, even if one method may outperform the other for some problems, they both benefit from our augmentation strategy (APHYNITY+, HVAE+). Overall, the effect of augmentation goes up to dividing the test error by a factor of $e^{4.6} \approx 100$ in some cases.

Stability for non-exact models. The empirical results from Fig. 7.4 are very important as they show that even when the decoder is not Ω -exact (and hence not $\overset{+}{\Omega}$ -exact), augmentation is still useful. In particular, Table 7.1 shows that the encoder does not predict the physical parameters perfectly. This indicates that the encoder is not Ω -exact and neither should be the decoder. This table shows the relative error on the physical parameters computed as $\sum_{i=1}^k \frac{1}{k} \left| \frac{z_e^i - \mu_{\theta}^i}{z_e^i} \right|$, where μ_{θ}^i is the estimated most likely value of the i^{th} component of the physical parameters. We first notice that APHYNITY and HVAE perform differently and their performance depends on the specific problem. While APHYNITY accurately estimates the physical parameters on the IID validation set for the 3 problems, HVAE’s performance are mixed on the RLC problem as it makes prediction that are 38% away from the nominal parameter value on average whereas APHYNITY reduces this

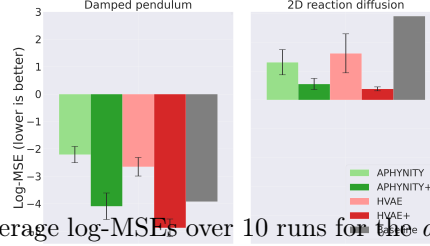


Figure 7.5: The average log-MSEs over 10 runs for the *damped pendulum* and *2D reaction diffusion* problems on a test distribution for which z_a , in addition to z_e , is also shifted. *AHM achieves better performance than stand HyL algorithms even when the test distribution support z_a differs from the training.*

error to 6%. Interestingly, we observe that the proposed augmentation strategies improve the encoder such that it accurately estimates the physical parameters also on the OOD test set even for HVAE on the RLC problem. This confirms that the augmentation strategy is helpful even when the hybrid model is not Ω -exact. As a conclusion, augmented hybrid learning outperforms classical hybrid learning both on the predictive accuracy and at inferring the expert variables.

Effect of out of expertise shift. *This experiment shows that our augmentation strategy may remain beneficial even when the train and test supports of z_a are not identical.* This scenario corresponds to samples (x, y) generated by $(z_a, z_e) \in (\mathcal{Z}_a^* \setminus \mathcal{Z}_a) \times \tilde{\mathcal{Z}}_e$ depicted by the violet domains in Fig. 7.3. In Fig. 7.5 we observe the log-MSE of augmented and non-augmented hybrid models trained for $(z_a, z_e) \in \mathcal{Z}_a \times \mathcal{Z}_e$ on test data that are generated with $(z_a, z_e) \in \tilde{\mathcal{Z}}_a \times \tilde{\mathcal{Z}}_e$. For the pendulum, the support over $z_a = \alpha$ is $[0, 0.3]$ in train and $[0.3, 0.6]$ in test; For the 2D reaction diffusion, $z_a = k$ is $[0.003, 0.005]$ in train and $[0.005, 0.008]$ in test. We observe that augmented models outperform the original models by a large margin. These results suggest that augmentation could be very valuable in practice, even when the distribution shift is also caused by non expert variables. However, if the shift on z_a becomes the dominant effect, augmented models also eventually becomes vulnerable to shifts on z_e as demonstrated by supplementary experiments in Appendix ??.

7.5 RELATED WORK

7.5.1 Hybrid modelling

Hybrid Learning (HyL), or gray box modelling as called in its early days in the 90's [?????], has been an appropriate method to learn models that are both expressive and interpretable, while also allowing them to be learnt on fewer data. The interest for HyL [??????] has greatly renewed since the outbreak of recent neural network architectures that simplify the combination of physical equations within ML models. As an

example, Neural ODE [?] and convolutional neural networks [?, CNN] are privileged architectures to work with dynamical systems described by ODEs or PDEs. While most of the HyL’s literature focus on the predictive performance of hybrid models, recent work have also showed that HyL may help to infer the physical parameter accurately [??]. This is aligned with ? (see Section 40.2.2.2) which observe that inference on incomplete models results in a *systematic bias*. Similar to HyL, they extend the model with *nuisance* parameters in order to improve its fidelity, and to reduce the systematic bias.

In this work, we decided to study ? and ? for two reasons that distinguish them from the rest of the HyL literature. First, these are notable examples of HyL algorithms that can be applied to a broad class of problems in contrast to papers that focus on specific applications [??]. Second, those methods also learn a reliable inference model for the physical parameters, suggesting that the expert model is used properly in the generative model, which is a key assumption for our augmentation. While ? claim to achieve robustness with HyL, we argue that this statement is incomplete as HVAE fails in OOD settings. In particular, their approach is only able to generalize with respect to unseen time or initial state if the model correctly identifies the latent variables z_a, z_e .

7.5.2 Combining hybrid modelling and data augmentation

Close to our idea is the one proposed in ? where they train a GAN model that improves the realism of a simulated image while conserving its semantic content (e.g. eyes colour) as modeled by the simulation parameters. The generated data with their annotations may then be used for a downstream task, such as inferring the properties of real images that corresponds to simulation parameters. The GAN objective from ? requires that the two distributions induced by the semantic content of real and simulated data are identical. On the opposite, we consider training data that corresponds to expert parameters with limited diversity, and overcome this scarcity with expert augmentation. Another line of work similar to ours is Sim2Real, which considers the task of transferring a model trained on simulated data to real world [???]. Robust HyL, as a way to enhance simulations, could be used for Sim2Real.

7.5.3 Robust ML and Invariant Learning

Various statistical methods have been introduced to ensure models generalize under distribution shift. Domain-adversarial objectives aimed at learning (conditionally) invariant predictors [???], GroupDRO [?] optimizing for worst-case loss over multiple domains and IRM [?] as well as sub-group calibration [?] aiming to satisfy calibration or sufficiency constraints to learn features invariant across domains. Extensions, able to infer domain labels from training data have been proposed as well

[??], partially inspired by fairness objectives [??]. In contrast to AHM, all of these methods rely on the variation of interest being present in the training data.

7.6 DISCUSSION

We now examine the assumptions we made to derive our augmentation strategy and discuss potential limitations.

Erroneous interaction model. The exactness of the hybrid component $p_\theta(y|x, y_e, z_a)$ is a critical assumption underlying our expert-based augmentation strategy. Unfortunately, this component is learned from training data only, hence we cannot prove its exactness on the test domain, which corresponds to a different domain \mathcal{Y}_e . However, we argue that soft assumptions on the class of interaction model may alleviate this problem. As an example, when we consider an additive hybrid model, as in APHYNITY [?], and embed this hypothesis into the interaction model, generalization to unseen y_e follows. When this assumption is too strong, we could still expect generalization of $p_\theta(y|x, y_e, z_a)$ because HyL drives y samples from p_θ to be close to y_e . It implies that the corresponding function approximator is smooth, which helps generalization to unseen scenarios. This contrasts with the encoder q_ψ for which a good inductive bias usually is not available.

Diagnostic. While crucial, we cannot guarantee the exactness of the decoder p_θ in general because we only evaluate the encoder and the decoder jointly on data points (x, y, x_o, y_o) . However, in some cases we can detect model misspecification by observing that the predictive model $p_{\theta, \psi}(y|x, x_o, y_o)$ is imperfect. Making this observation is not always simple as it requires prior knowledge on the expected accuracy of an exact model. However, when the system is deterministically identifiable, we may argue that the accuracy should be only limited by the intrinsic noise between x and y given z_a and z_e .

Relaxing exactness. Even with a strong inductive bias on the decoder, achieving exactness is hopeless in practical settings. However, our experiments demonstrate that expert-augmentation works in practice. We can explain this by taking a look at Fig. 7.3. If the generative model that maps x and (z_a, z_e) is incorrect, the mapping from \mathcal{Z}_a and \mathcal{Z}_e could be slightly off from $\bar{\Omega}^+$. However, this does not preclude the set of augmented samples to be closer to $\bar{\Omega}^+$ than Ω and to induce a better predictive model on $\bar{\Omega}^+$ than the original model trained only on Ω .

Limitations We considered expert models that are parameterized by a small number of parameters, which can be covered densely via sampling.

Covering densely a higher dimensional parameter space with the augmentation strategy becomes quickly impossible, hence a smarter sampling strategy would be required, such as worst-case sampling. Another difficulty is to choose a plausible range of parameters that contains both the train and the test support, this will often require a human expert in the loop. Finally, we assume that the train distribution of z_a should be representative of the test distribution, we empirically observed that a softer version of this assumption could be enough. However, performance will eventually decline as the support of the test distribution for z_a is far from the training domain.

7.7 CONCLUSION

In this work, we describe HyL with a probabilistic model in which one component of the latent process, denoted the expert model, is known. In this context, we establish that state-of-the-art HyL algorithms are vulnerable to distribution shifts even when the expert model is well defined for such configurations. Grounded in this formalisation, we derive that expert augmentations induce robustness to OOD settings. We discuss how our assumptions can transfer to real-world settings and describe how to diagnose potential shortcomings. Finally, empirical evidence asserts that expert augmentations may be beneficial even when one of our assumptions on the class of distribution shift is violated.

Our augmentation is applicable to a large class of hybrid models, hence it should benefit from future progress in HyL. Thus, we believe research in HyL and formally defining its targeted objectives is an important direction for further improving the robustness of hybrid models. As an example, the minimal description length principle [?] could be a great resource to investigate the balance between the model’s capacity and robustness. Finally, robust ML models must eventually translate to real-world applications, hence a next step would be to apply AHMs to real-world data. Paving the way to future research combining AHM with robust ML methods.

Part IV

CONCLUSION AND OPPORTUNITIES

CONCLUSION

OUTLINE

???

Part V

APPENDIX

NOTATIONS

\mathcal{A}	A supervised learning algorithm??
$\mathcal{A}(\theta, \mathcal{L})$	The model $\varphi_{\mathcal{L}}$ produced by algorithm \mathcal{A} over \mathcal{L} and hyper-parameters θ??
α_s	The proportion of samples in a random patch??
α_f	The proportion of features in a random patch??
b_l	The l -th value of a categorical variable??
B	A subset $B \subseteq V$ of variables??
c_k	The k -th class??
C_p^k	The number of k -combinations from a set of p elements ??
$C(N)$	The time complexity for splitting N samples??
\mathbb{E}	Expectation
$\overline{E}(\varphi_{\mathcal{L}}, \mathcal{L}')$	The average prediction error of $\varphi_{\mathcal{L}}$ over \mathcal{L}'??
$Err(\varphi_{\mathcal{L}})$	The generalization error of $\varphi_{\mathcal{L}}$??, ??
$H(X)$	The Shannon entropy of X??
$H(X Y)$	The Shannon entropy of X conditional to Y??
\mathcal{H}	The space of candidate models??
$i(t)$	The impurity of node t??, ??
$i_R(t)$	The impurity of node t based on the local resubstitution estimate??, ??
$i_H(t)$	The entropy impurity of node t??
$i_G(t)$	The Gini impurity of node t??
$\Delta i(s, t)$	The impurity decrease of the split s at node t??
$I(X; Y)$	The mutual information between X and Y??
$\text{Imp}(X_j)$	The variable importance of X_j??, ??
J	The number of classes??
K	The number of folds in cross-validation?? The number of input variables drawn at each node for finding a split??
$K(\mathbf{x}_i, \mathbf{x}_j)$	The kernel of \mathbf{x}_i and \mathbf{x}_j??, ??
L	A loss function?? The number of values of a categorical variable??
\mathcal{L}	A learning set (\mathbf{X}, \mathbf{y})??
\mathcal{L}^m	The m -th bootstrap replicate of \mathcal{L}??

\mathcal{L}_t	The subset of node samples falling into node t ??
M	The number of base models in an ensemble ??
$\mu_{\mathcal{L}, \theta_m}(\mathbf{x})$	The mean prediction at $X = \mathbf{x}$ of $\varphi_{\mathcal{L}, \theta_m}$??
N	The number of input samples ??
N_t	The number of node samples in node t ??
N_{ct}	The number of node samples of class c in node t ... ??
Ω	The universe, or population, from which cases are sampled ??
p	The number of input variables ??
p_L	The proportion of node samples going to t_L ??
p_R	The proportion of node samples going to t_R ??
$p(t)$	The estimated probability $p(X \in \mathcal{X}_t) = \frac{N_t}{N}$??
$p(c t)$	The empirical probability estimate $p(Y = c X \in \mathcal{X}_t) = \frac{N_{ct}}{N_t}$ of class c at node t ??
$\hat{p}_{\mathcal{L}}$	An empirical probability estimate computed from the learning set \mathcal{L} ??
$P(X, Y)$	The joint probability distribution of the input variables $X = (X_1, \dots, X_p)$ and the output variable Y ??
$\mathcal{P}_k(V)$	The set of subsets of V of size k ??
φ	A model or function $\mathcal{X} \mapsto \mathcal{Y}$?? A single decision tree ??
$\tilde{\varphi}$	The set of terminal nodes in φ ??
$\varphi(\mathbf{x})$	The prediction of φ for the sample \mathbf{x} ??
$\varphi_{\mathcal{L}}$	A model built from \mathcal{L} ??
$\varphi_{\mathcal{L}, \theta}$	A model built from \mathcal{L} with random seed θ ??
φ_B	A Bayes model ??
$\psi_{\mathcal{L}, \theta_1, \dots, \theta_M}$	An ensemble of M models built from \mathcal{L} and random seeds $\theta_1, \dots, \theta_M$??
\mathcal{Q}	A set $\mathcal{Q} \subseteq \mathcal{S}$ of splits of restricted structure ??, ??
$\mathcal{Q}(X_j)$	The set $\mathcal{Q}(X_j) \subseteq \mathcal{Q}$ of univariate binary splits that can be defined on variable X_j ??, ??
$\rho(\mathbf{x})$	The correlation coefficient between the predictions at $X = \mathbf{x}$ of two randomized models ??
s	A split ??, ??
s^*	The best split ??, ??
s_j^*	The best binary split defined on variable X_j ??, ??
s_j^v	The binary split $(\{\mathbf{x} x_j \leq v\}, \{\mathbf{x} > v\})$ defined on variable X_j with discretization threshold v ??
s_t	The split labeling node t ??
\tilde{s}_t^j	The best surrogate split for s_t defined from X_j ??

\mathcal{S}	The set of all possible splits s??
$\sigma_{\mathcal{L}, \theta_m}^2(\mathbf{x})$	The prediction variance at $X = \mathbf{x}$ of $\varphi_{\mathcal{L}, \theta_m}$??
t	A node in a decision tree??
t_L	The left child of node t??, ??
t_R	The right child of node t??, ??
θ	A vector of hyper-parameter values??
	A random seed??
θ^*	The optimal hyper-parameters??
$\hat{\theta}^*$	The approximately optimal hyper-parameters??
θ_m	The seed of the m -th model in an ensemble??
v	A discretization threshold in a binary split??
v_k	The k -th value of an ordered variable, when node samples are in sorted order??
v'_k	The mid-cut point between v_k and v_{k+1}??
V	The set $\{X_1, \dots, X_p\}$ of input variables??
V^{-j}	$V \setminus \{X_j\}$??
\mathbb{V}	Variance??
\mathbf{x}	A case, sample or input vector (x_1, \dots, x_p)??
\mathbf{x}_i	The i -th input sample in \mathcal{L}??
x_j	The value of variable X_j for the sample \mathbf{x}??
\mathbf{X}	The $N \times p$ matrix representing the values of all N samples for all p input variables??
X_j	The j -th input variable or feature??, ??
X	The random vector (X_1, \dots, X_p)??
\mathcal{X}_j	The domain or space of variable X_j??
\mathcal{X}	The input space $\mathcal{X}_1 \times \dots \times \mathcal{X}_p$??
\mathcal{X}_t	The subspace $\mathcal{X}_t \subseteq \mathcal{X}$ represented by node t??
y	A value of the output variable Y??
\hat{y}_t	The value labelling node t??
\hat{y}_t^*	The optimal value labelling node t??
\mathbf{y}	The output values (y_1, \dots, y_N)??
Y	The output or response variable Y??
\mathcal{Y}	The domain or space of variable Y??

REFERENCES

-
- N. P. Archer and S. Wang. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. 24(1):60–75, 1993-01. URL <https://doi.org/10.1111/j.1540-5915.1993.tb00462.x>. Publisher: Wiley.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- CERN. CERN data centre passes the 200-petabyte milestone. URL <https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone>.
- R. T. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. 2018a.
- T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583, 2018b.
- X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. 2017.
- N. De Cao, W. Aziz, and I. Titov. Block neural autoregressive flow. In *Uncertainty in Artificial Intelligence*, pages 1263–1273. PMLR, 2020.
- A. Delaunoy, A. Wehenkel, T. Hinderer, S. Nissanke, C. Weniger, A. R. Williamson, and G. Louppe. Lightning-fast gravitational wave parameter inference through neural amortization. *arXiv preprint arXiv:2010.12931*, 2020.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.
- L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. In *International Conference in Learning Representations*, 2017.

- J. Dumas, C. Cointe, A. Wehenkel, A. Sutura, X. Fettweis, and B. Cornélusse. A probabilistic forecast-driven strategy for a risk-aware participation in the capacity firming market. *IEEE Transactions on Sustainable Energy*, 2021.
- J. Dumas, A. Wehenkel, D. Lanaspeze, B. Cornélusse, and A. Sutura. A deep generative model for probabilistic energy forecasting in power systems: normalizing flows. *Applied Energy*, 305:117871, 2022.
- D. Geiger, T. Verma, and J. Pearl. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier, 1990.
- M. Germain, K. Gregor, I. Murray, and H. Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015.
- M. Gillon, A. H. Triaud, B.-O. Demory, E. Jehin, E. Agol, K. M. Deck, S. M. Lederer, J. De Wit, A. Burdanov, J. G. Ingalls, et al. Seven temperate terrestrial planets around the nearby ultracool dwarf star trappist-1. *Nature*, 542(7642):456–460, 2017.
- W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Machine Learning*, 2018.
- M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. Van Esbroeck. Monotonic calibrated interpolated look-up tables. 17(1):3790–3836, 2016. Publisher: JMLR.org.
- A. Habibian, T. v. Rozendaal, J. M. Tomczak, and T. S. Cohen. Video compression with rate-distortion autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7033–7042, 2019.
- J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, and G. Louppe. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021.
- I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2017, 2016.
- J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. 2020.
- M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, page 2, 2016.

- C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2083–2092, 2018.
- P. Jaini, K. A. Selby, and Y. Yu. Sum-of-squares polynomial flow. 2019.
- T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2017.
- T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.
- D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2013.
- D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.
- A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.
- A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. 2021.
- J. Oliva, A. Dubey, M. Zaheer, B. Poczos, R. Salakhutdinov, E. Xing, and J. Schneider. Transformation autoregressive networks. In *International Conference on Machine Learning*, pages 3895–3904, 2018.
- A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. 2016.
- G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- G. Papamakarios, D. C. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.
- J. Pearl. Bayesian networks. 2011.

- A. Pesah, A. Wehenkel, and G. Louppe. Recurrent machines for likelihood-free inference. *arXiv preprint arXiv:1811.12932*, 2018.
- A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.
- J. Sill. Monotonic networks. In *Advances in neural information processing systems*, pages 661–667, 1998.
- J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder variational autoencoders. *arXiv preprint arXiv:1602.02282*, 2016.
- Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2019.
- Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- J. Tomczak and M. Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.
- A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. 2020.
- A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. 2021.
- J.-W. van de Meent, B. Paige, H. Yang, and F. Wood. An introduction to probabilistic programming. 2018.
- A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125, 2016.
- A. Van Den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, and others. Parallel WaveNet: Fast high-fidelity speech synthesis.

- In *International Conference on Machine Learning*, pages 3915–3923, 2018.
- M. Vandegar, M. Kagan, A. Wehenkel, and G. Louppe. Neural empirical bayes: Source distribution estimation and its applications to simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 2107–2115. PMLR, 2021.
- N. Vecoven, D. Ernst, A. Wehenkel, and G. Drion. Introducing neuromodulation in deep neural networks to learn adaptive behaviours. *PloS one*, 15(1):e0227922, 2020.
- A. Wehenkel and G. Louppe. Unconstrained monotonic neural networks. In *Advances in Neural Information Processing Systems*, pages 1543–1553, 2019.
- A. Wehenkel and G. Louppe. You say normalizing flows i see bayesian networks. 2020.
- A. Wehenkel and G. Louppe. Diffusion priors in variational autoencoders. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021a.
- A. Wehenkel and G. Louppe. Graphical normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 37–45. PMLR, 2021b.
- A. Wehenkel, A. Mukhopadhyay, J.-Y. Le Boudec, and M. Paolone. Parameter estimation of three-phase untransposed short transmission lines from synchrophasor measurements. *IEEE Transactions on Instrumentation and Measurement*, 69(9):6143–6154, 2020.
- A. Wehenkel, J. Behrmann, H. Hsu, G. Sapiro, G. Louppe, and J.-H. Jacobsen. Robust hybrid learning with expert augmentation. *arXiv preprint arXiv:2202.03881*, 2022.
- S. You, D. Ding, K. Canini, J. Pfeifer, and M. Gupta. Deep lattice networks and partial monotonic functions. In *Advances in Neural Information Processing Systems*, pages 2981–2989, 2017.