UAI 2017 Australia

Tutorial on

Deep Generative Models

Shakir Mohamed and Danilo Rezende

@shakir_za @deepspiker

DeepMind

Abstract

This tutorial will be a review of recent advances in deep generative models. Generative models have a long history at UAI and recent methods have combined the generality of probabilistic reasoning with the scalability of deep learning to develop learning algorithms that have been applied to a wide variety of problems giving state-of-the-art results in image generation, text-to-speech synthesis, and image captioning, amongst many others. Advances in deep generative models are at the forefront of deep learning research because of the promise they offer for allowing data-efficient learning, and for model-based reinforcement learning. At the end of this tutorial, audience member will have a full understanding of the latest advances in generative modelling covering three of the active types of models: Markov models, latent variable models and implicit models, and how these models can be scaled to high-dimensional data. The tutorial will expose many questions that remain in this area, and for which there remains a great deal of opportunity from members of the UAI community.

Beyond Classification

Move beyond associating inputs to outputs

Recognise objects in the world and their factors of variation

Establish concepts as useful for reasoning and decision making

Understand and imagine how the world evolves

Detect surprising events in the world

Anticipate and generate rich plans for the future

What is a Generative Model?

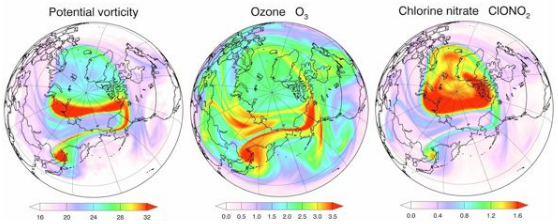
A model that allows us to learn a simulator of data

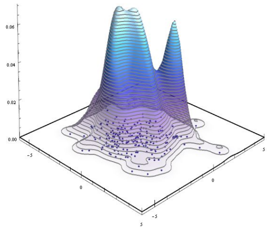
Models that allow for (conditional) density estimation

Approaches for unsupervised learning of data

Characteristics are:

- **Probabilistic** models of data that allow for uncertainty to be captured.
- **Data distribution p(x)** is targeted.
- **High-dimensiona**l outputs.



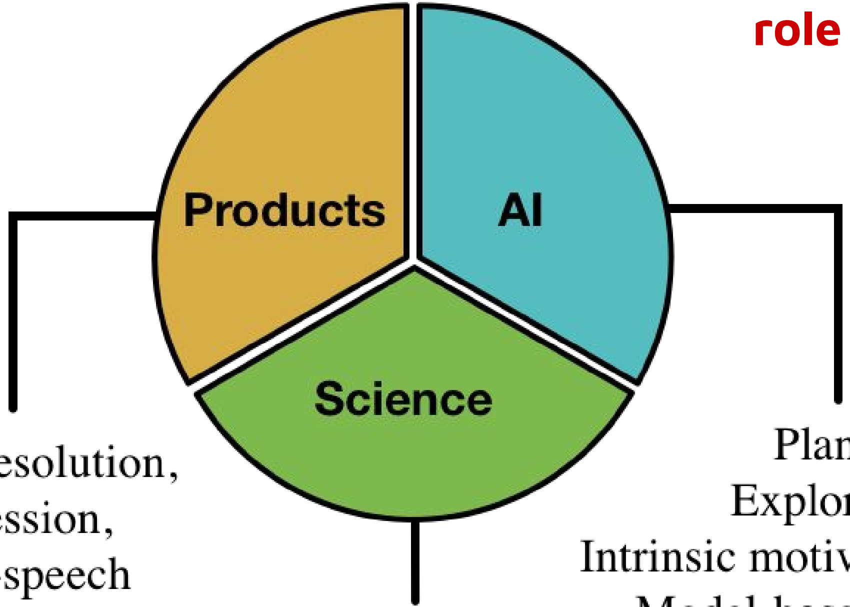


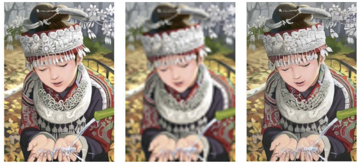
NO LABELS

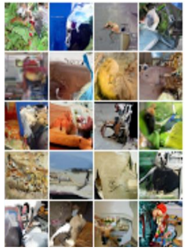
Why Generative Models?

Why Generative Models

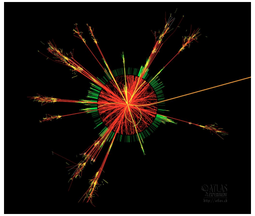
Generative models have a role in many problems.





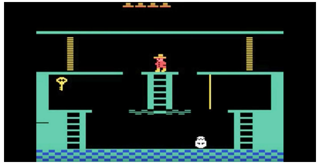


Super-resolution,
Compression,
Text-to-speech



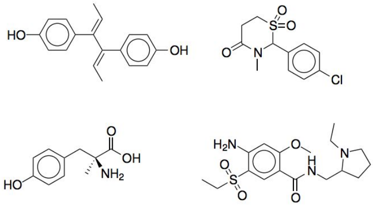
Proteomics,
Drug Discovery,
Astronomy,
High-energy physics

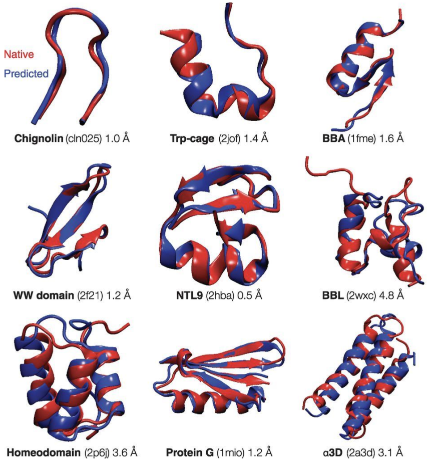
Planning,
Exploration
Intrinsic motivation
Model-based RL

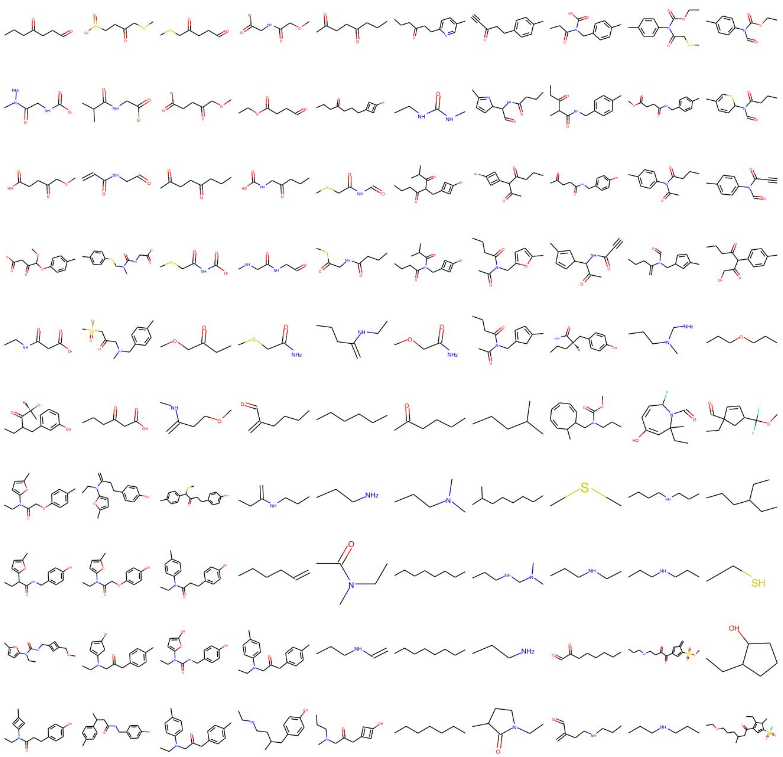


Drug Design and Response Prediction

Proposing candidate molecules and for improving prediction through semi-supervised learning.



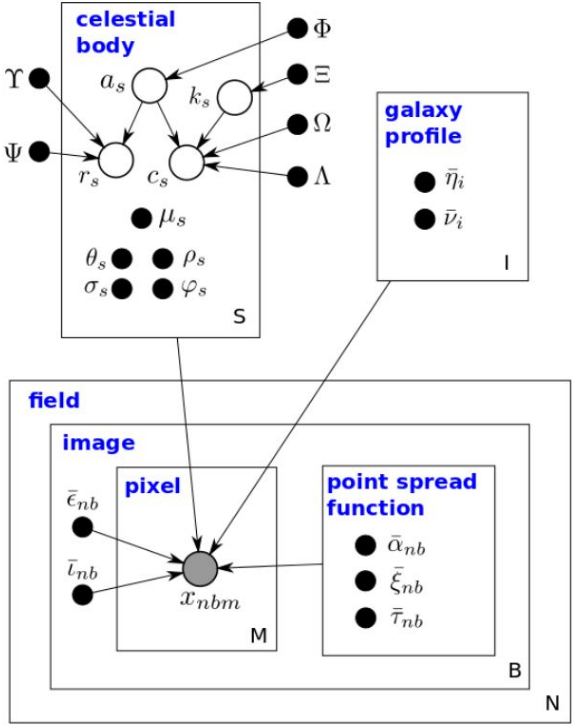




Gómez-Bombarelli, et al. 2016

Locating Celestial Bodies

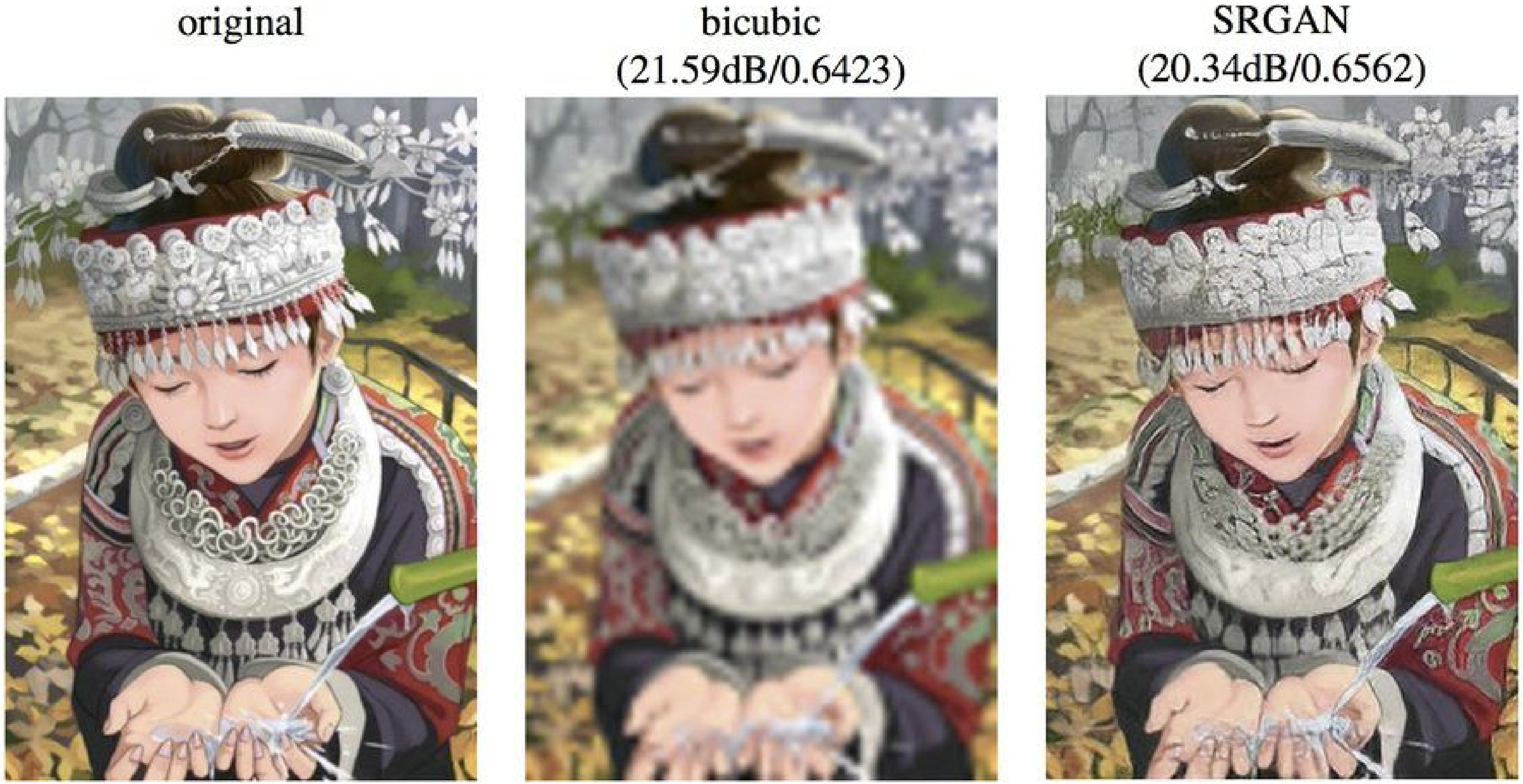
Generative models for applications in astronomy and high-energy physics.



Regier et al., 2015

Image super-resolution

Photo-realistic single image super-resolution



original

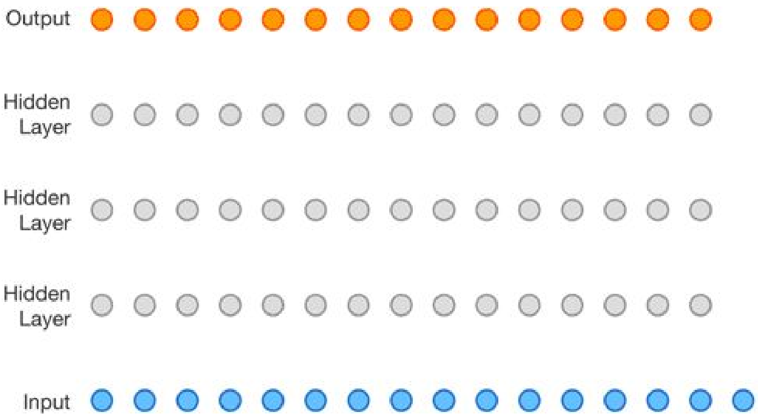
bicubic
(21.59dB/0.6423)

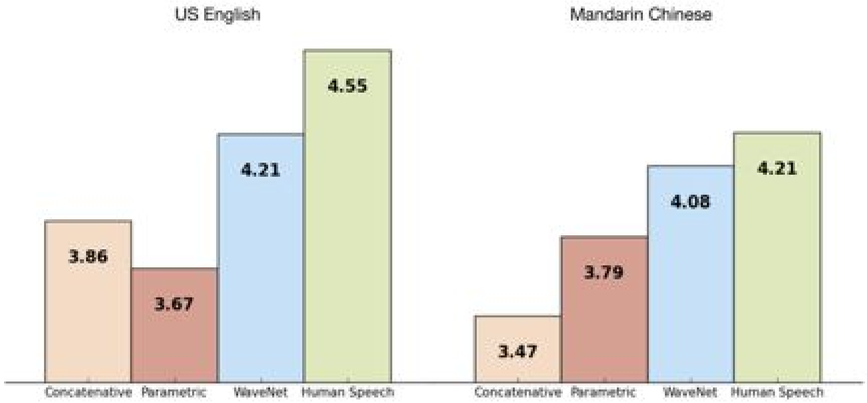
SRGAN
(20.34dB/0.6562)

Ledig et al., 2016

Text-to-speech Synthesis

Generating audio conditioned on text



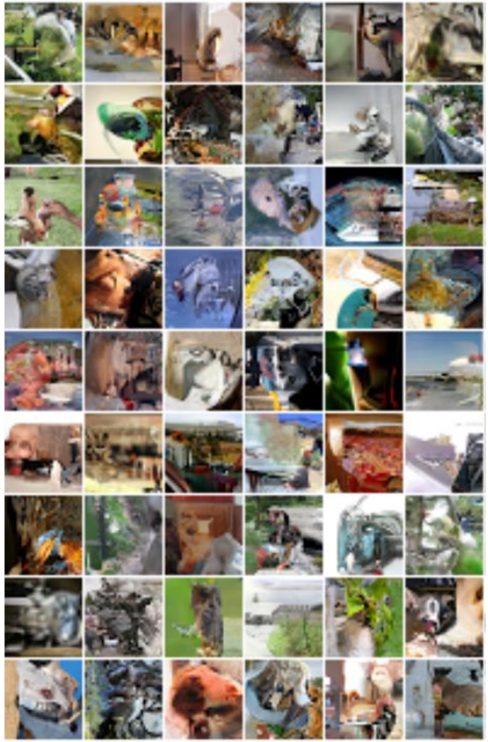


Oord et al., 2016

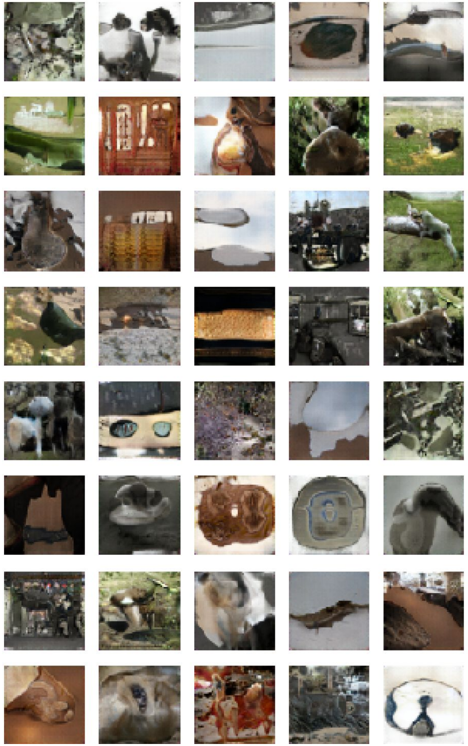
Image and Content Generation

Generating images and video content.

DRAW



Pixel RNN



ALI

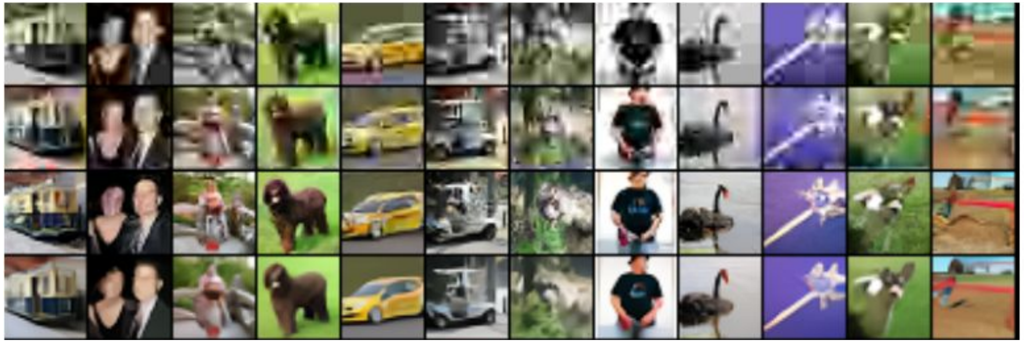
Gregor et al., 2015, Oord et al., 2016, Dumoulin et al., 2016

Communication and Compression

Hierarchical compression of images and other data.

Original images

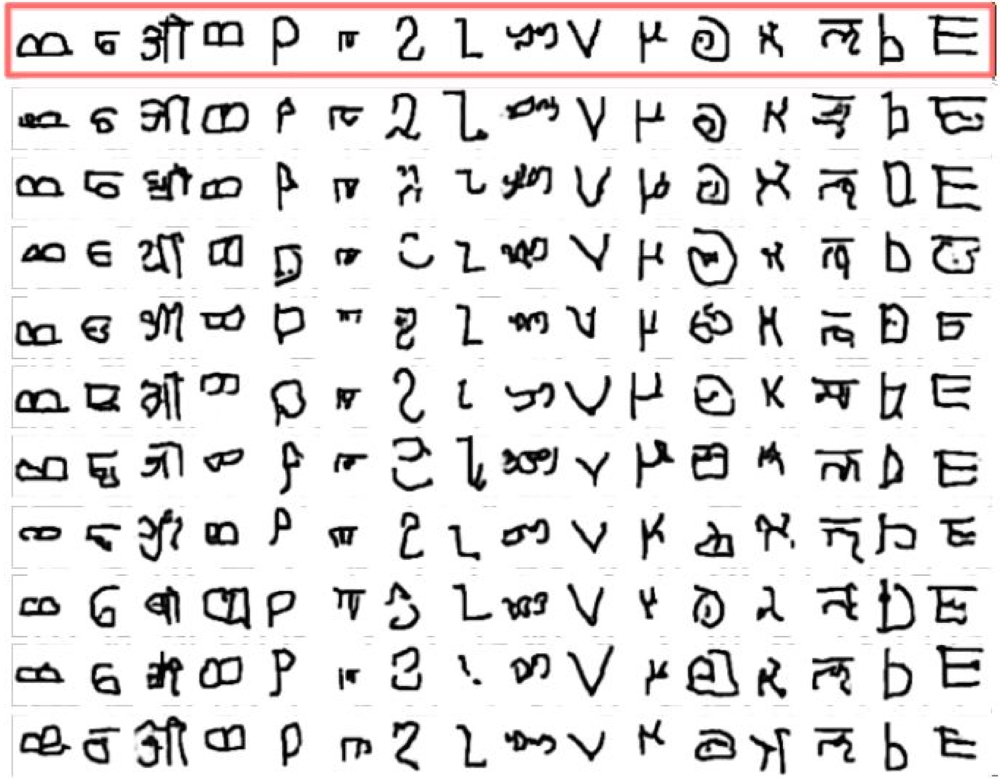
Compression rate: 0.2bits/dimension

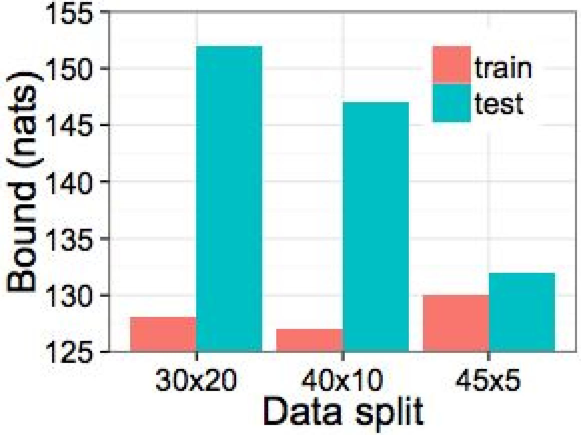


Gregor et al., 2016

One-shot Generalisation

Rapid generalisation of novel concepts

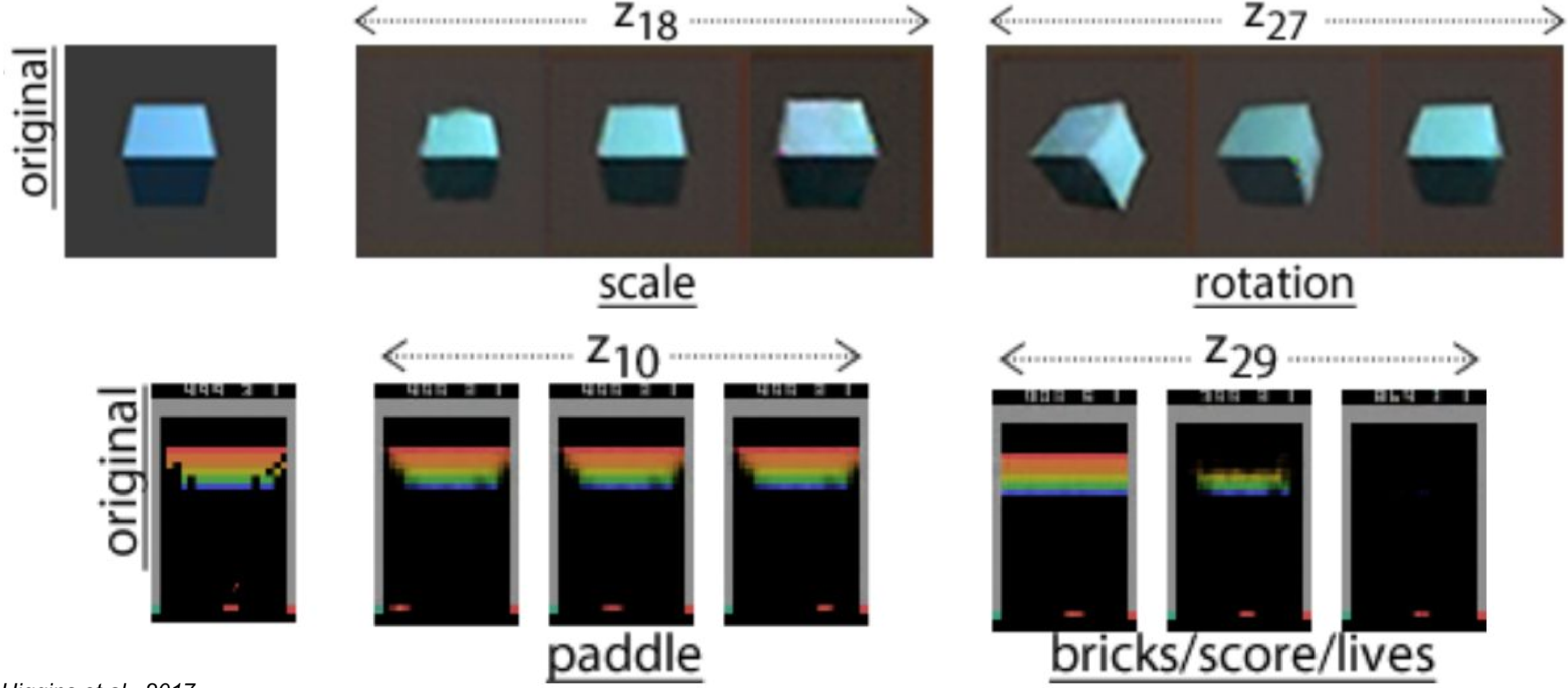




Rezende et al., 2016

Visual Concept Learning

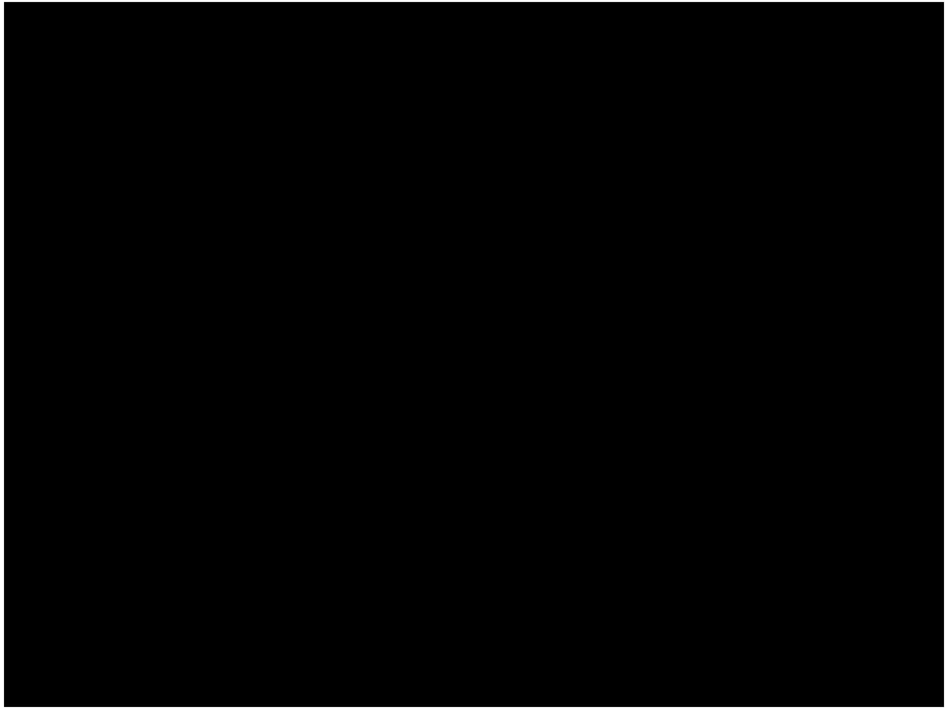
Understanding the factors of variation and invariances.



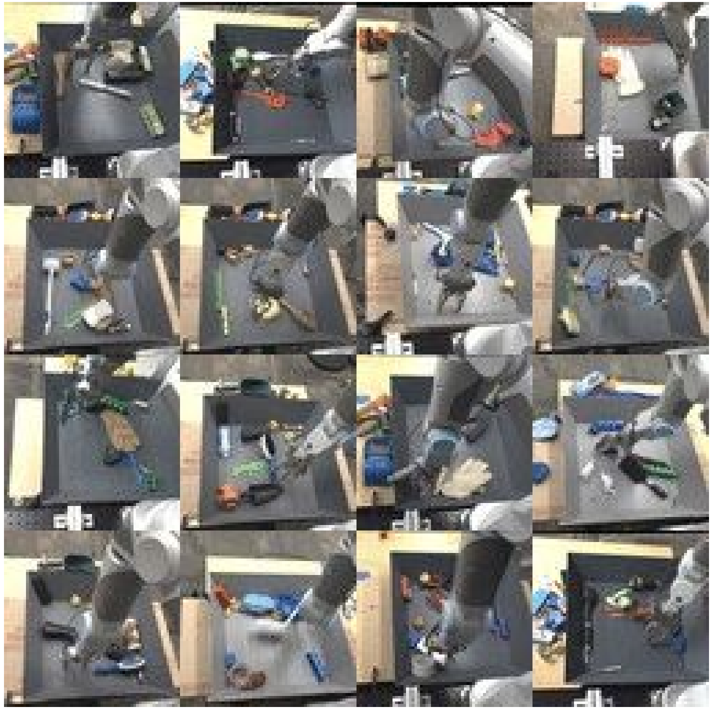
Higgins et al., 2017

Future Simulation

Simulate future trajectories of environments based on actions for planning



Atari simulation

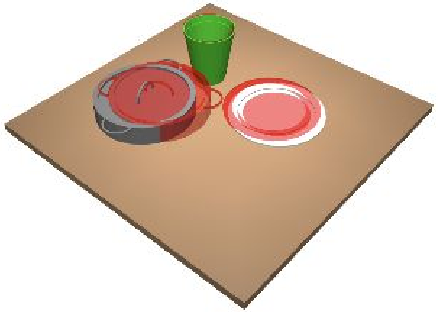


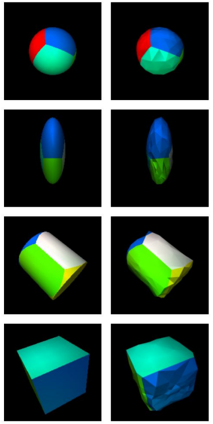
Robot arm simulation

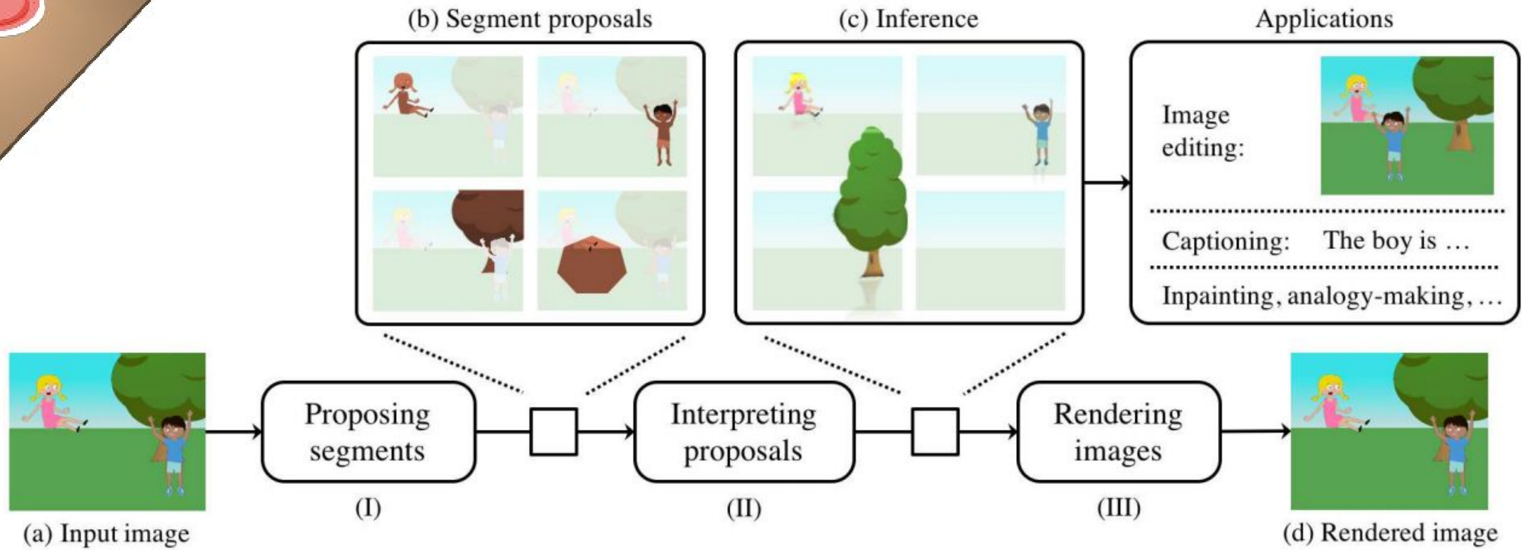
Chiappa et al, 2017, Kalchbrenner et al., 2017

Scene Understanding

Understanding the components of scenes and their interactions



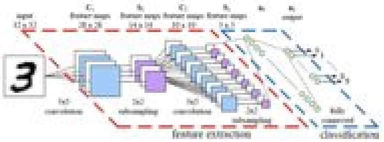




Wu et al., 2017

Probabilistic Deep Learning

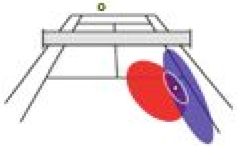
Two Streams of Machine Learning



Deep Learning

- + Rich non-linear models for classification and sequence prediction.
- + Scalable learning using stochastic approximation and conceptually simple.
- + Easily composable with other gradient-based methods.

- - Only point estimates.
- - Hard to score models, do selection and complexity penalisation.



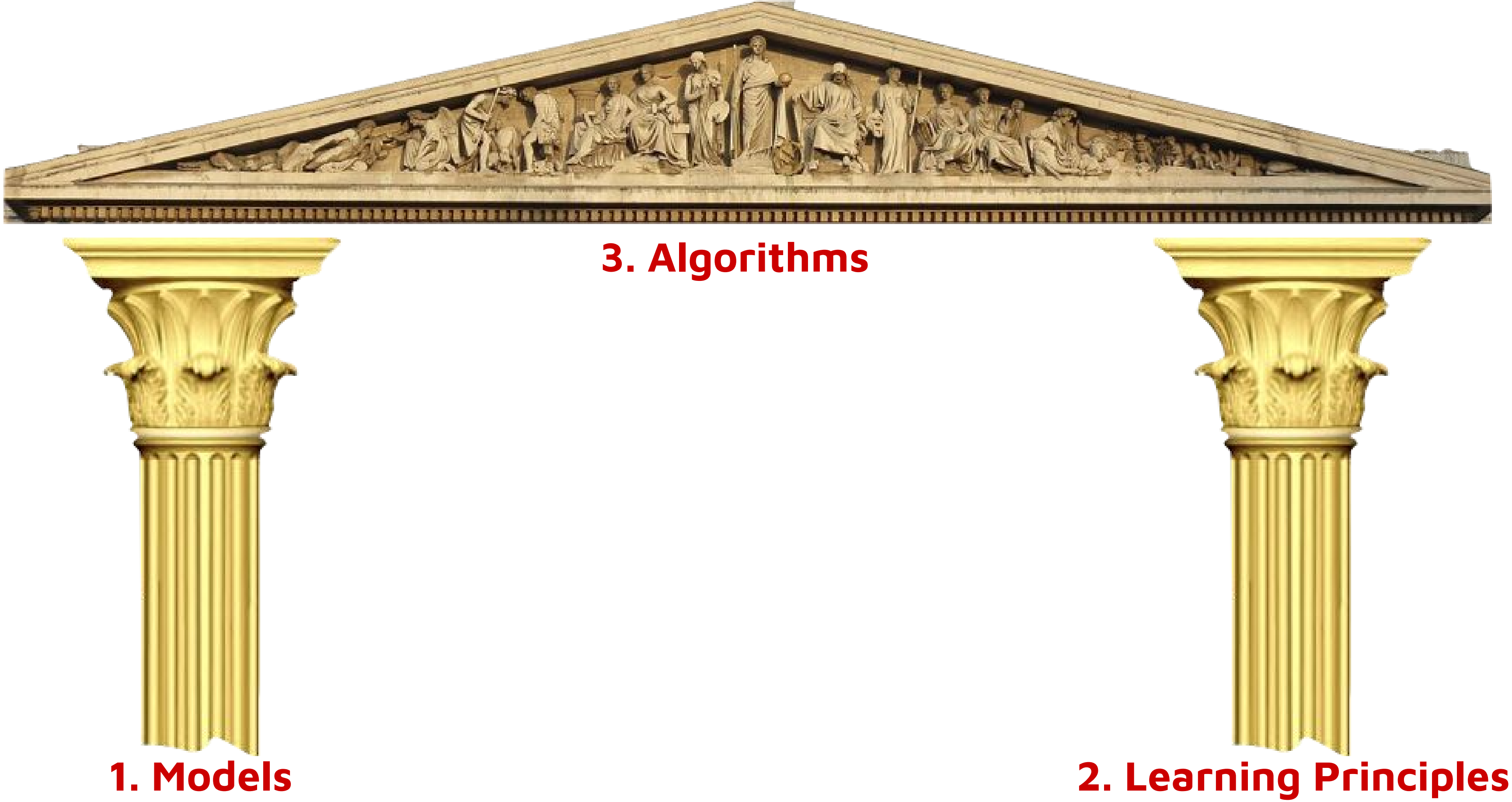
Probabilistic Reasoning

- - Mainly conjugate and linear models.
- - Potentially intractable inference, computationally expensive or long simulation time.

- + Unified framework for model building, inference, prediction and decision making.
- + Explicit accounting for uncertainty and variability of outcomes.
- + Robust to overfitting; tools for model selection and composition.

Complementary strengths, making it natural to combine them

Thinking about Machine Learning

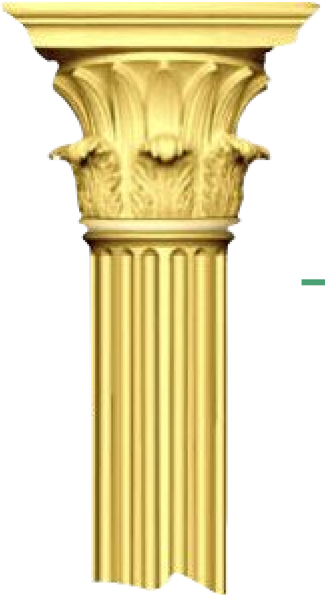


3. Algorithms

1. Models

2. Learning Principles

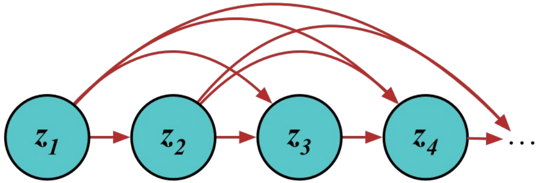
Types of Generative Models

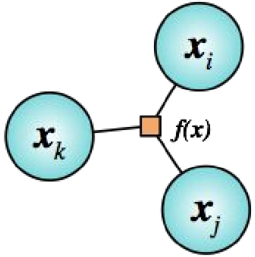


1. Models

Fully-observed models

Model observed data directly without introducing any new unobserved local variables.

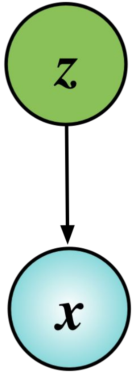


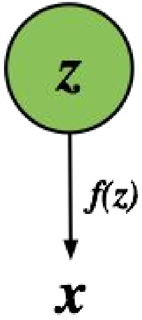


Latent Variable Models

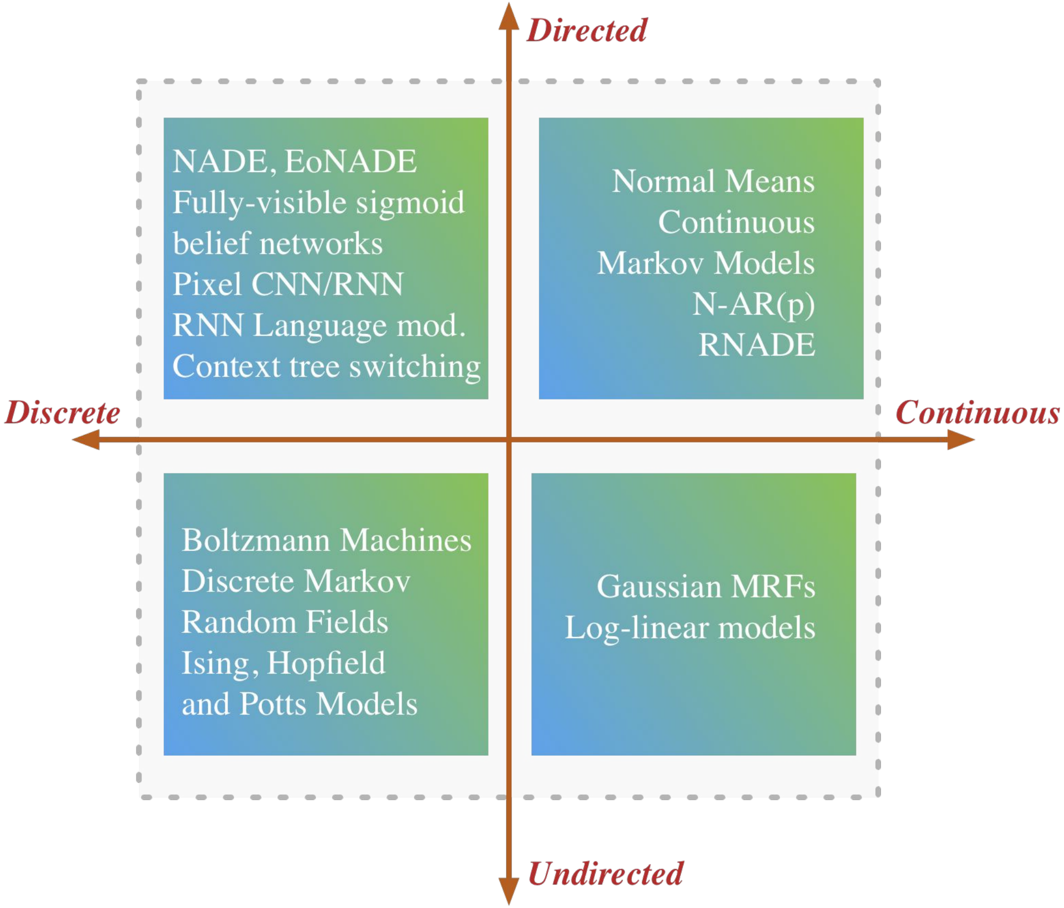
Introduce an unobserved random variable for every observed data point to explain hidden causes.

- **Prescribed models**: Use observer likelihoods and assume observation noise.
- **Implicit models**: Likelihood-free models.





Spectrum of Fully-observed Models



Directed

NADE, EoNADE
Fully-visible sigmoid belief networks
Pixel CNN/RNN
RNN Language mod.
Context tree switching

Normal Means
Continuous Markov Models
N-AR(p)
RNADE

Discrete

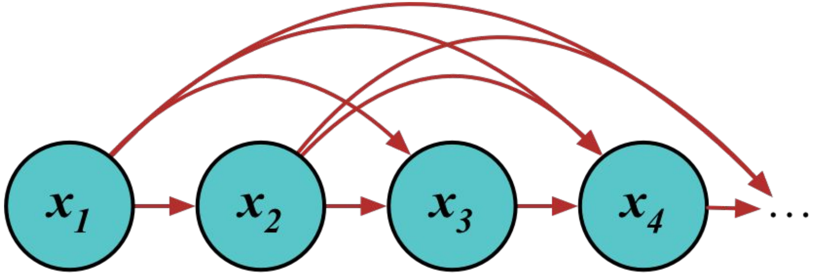
Continuous

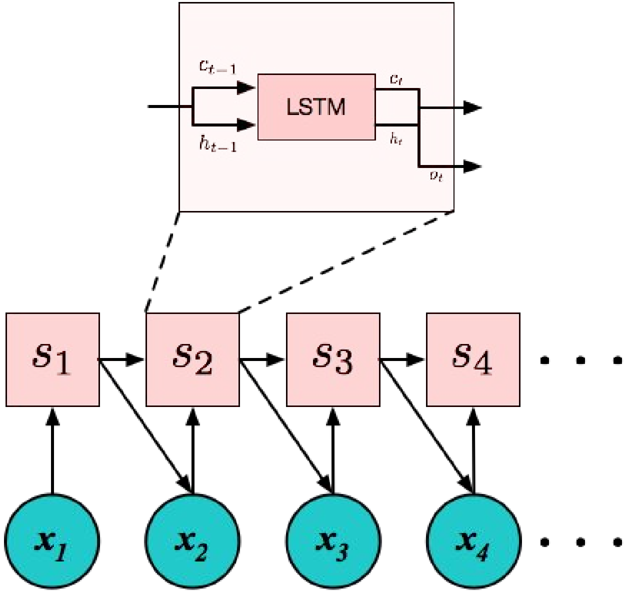
Boltzmann Machines
Discrete Markov Random Fields
Ising, Hopfield and Potts Models

Gaussian MRFs
Log-linear models

Undirected

Building Generative Models





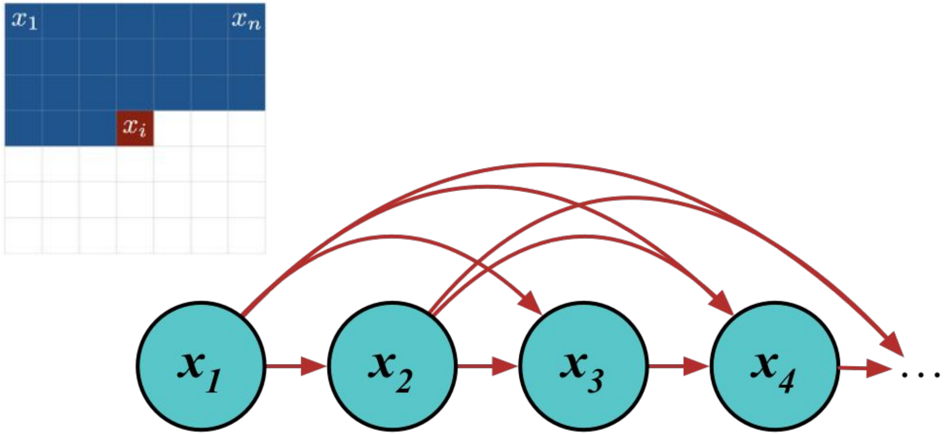
$$p(x_{1,\ldots,N}) = \prod_{i=1}^{N} p(x_i | x_{1,\ldots,(i-1)})$$

$$p(x_{1,\ldots,N}) = \prod_{i=1}^{N} p(x_i | s_i(s_{i-1}, x_{i-1}))$$

Equivalent ways of representing the same DAG

Fully-observed Models

$$p(x_{1,\ldots,N}) = \prod_{i=1}^{N} p(x_i | x_{1,\ldots,(i-1)})$$



All conditional probabilities described by deep networks.

- + Can directly encode how observed points are related.
- + Any data type can be used
- + For directed graphical models: Parameter learning simple
- + Log-likelihood is directly computable, no approximation needed.
- + Easy to scale-up to large models, many optimisation tools available.

- - Order sensitive.
- - For undirected models, parameter learning difficult: Need to compute normalising constants.
- - Generation can be slow: iterate through elements sequentially, or using a Markov chain.

Spectrum of Latent Variable Models

Cascaded Indian Buffet process
Hierarchical Dirichlet process

Deep Nonparametric Discrete

Sigmoid Belief Net
Deep auto-regressive networks (DARN)

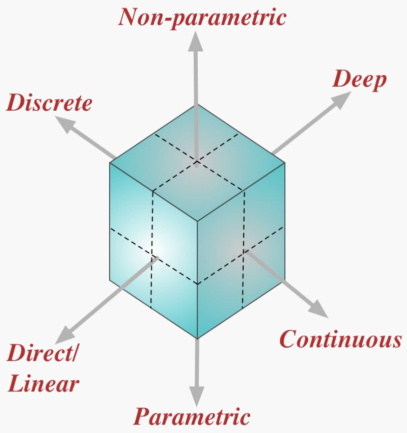
Deep Parametric Discrete

Indian buffet process
Dirichlet process mixture

Direct Nonparametric Discrete

Hidden Markov Model
Discrete LVM
Sparse LVMs

Linear Parametric Discrete



Deep Gaussian processes
Recurrent Gaussian Process
GP State space model

Deep Nonparametric Continuous

Nonlinear factor analysis
Nonlinear Gaussian belief network
Deep Latent Gaussian (VAE, DRAW)

Deep Parametric Continuous

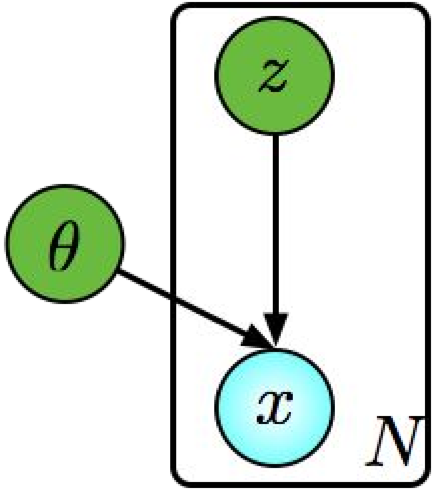
PCA, factor analysis
Independent components analysis
Gaussian LDS
Latent Gauss Field

Linear Parametric Continuous

Gaussian process LVM

Direct Nonparametric Continuous

Building Generative Models

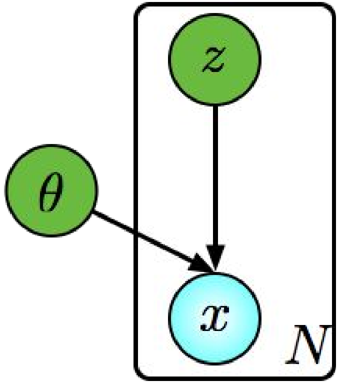


$$p(x, z, \theta) = \rho(\theta) \prod_{i=1}^{N} p(x_i | z_i, \theta) \pi(z_i)$$

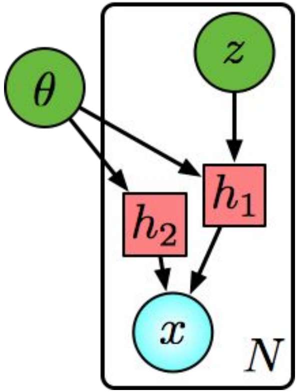
$$\begin{aligned} \pi(z) &= \mathcal{N}(0, \mathbb{I}_{d_z}) \\ \rho(\theta) &= \mathcal{N}(0, \kappa^2 \mathbb{I}_{d_\theta}) \\ p(x|z, \theta) &= \mathcal{N}(\theta_0 + \theta_1 z, \exp(\theta_2)) \\ \theta &= \{\theta_0 \in \mathbb{R}^{d_x}, \theta_1 \in \mathbb{R}^{d_x \times d_z}, \theta_2 \in \mathbb{R}^{d_x}\} \end{aligned}$$

Building Generative Models

Graphical Models + Computational Graphs (aka NNets)

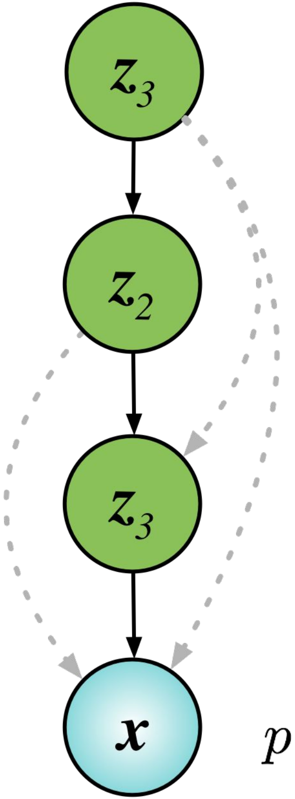


$$\begin{aligned}
\pi(z) &= \mathcal{N}(0, \mathbb{I}_{d_z}) \\
\rho(\theta) &= \mathcal{N}(0, \kappa^2 \mathbb{I}_{d_\theta}) \\
p(x|z,\theta) &= \mathcal{N}(\theta_0 + \theta_1 z, \exp(\theta_2))
\end{aligned}$$



$$\begin{aligned}
\pi(z) &= \mathcal{N}(0, \mathbb{I}_{d_z}) \\
\rho(\theta) &= \mathcal{N}(0, \kappa^2 \mathbb{I}_{d_\theta}) \\
h_1 &= \theta_0 + \theta_1 z \\
h_2 &= \exp(\theta_2) \\
p(x|z,\theta) &= \mathcal{N}(h_1, h_2)
\end{aligned}$$

Latent Variable Models



- Inversion process to determine latents corresponding to a input is difficult in general
- Difficult to compute marginalised likelihood requiring approximations.
- Not easy to specify rich approximations for latent posterior distribution.

$$p(x, z, \theta) = \rho(\theta) \prod_{i=1}^{N} p(x_i|z_i, \theta)\pi(z_i)$$

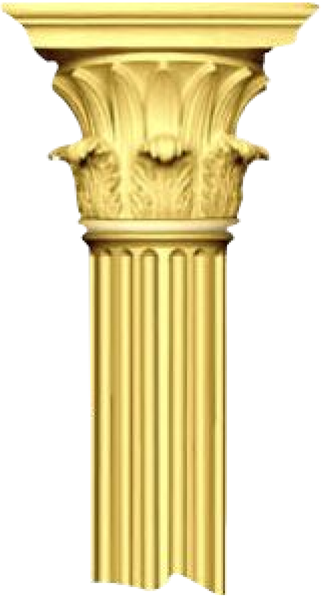
+ Easy sampling.
+ Easy way to include hierarchy and depth.
+ Easy to encode structure
+ Avoids order dependency assumptions: marginalisation induces dependencies.
+ Provide compression and representation.
+ Scoring, model comparison and selection possible using the marginalised likelihood.

Introduce an unobserved local random variables that represents hidden causes.

Choice of Learning Principles

For a given model, there are many competing inference methods.

- Exact methods (conjugacy, enumeration)
- Numerical integration (Quadrature)
- Generalised method of moments
- **Maximum likelihood (ML)**
- Maximum a posteriori (MAP)
- Laplace approximation
- Integrated nested Laplace approximations (INLA)
- **Expectation Maximisation (EM)**
- Monte Carlo methods (MCMC, SMC, ABC)
- Contrastive estimation (NCE)
- Cavity Methods (EP)
- **Variational methods**



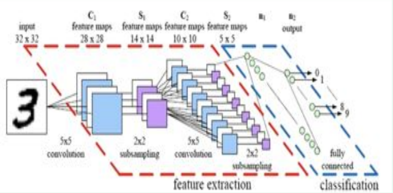
2. Learning Principles

Combining Models and Inference

3. Algorithms

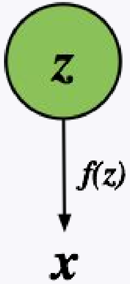
A given model and learning principle can be implemented in many ways.

Convolutional neural network + penalised maximum likelihood



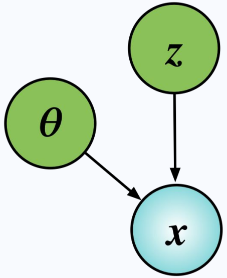
- Optimisation methods (SGD, Adagrad)
- Regularisation (L1, L2, batchnorm, dropout)

Implicit Generative Model + Two-sample testing



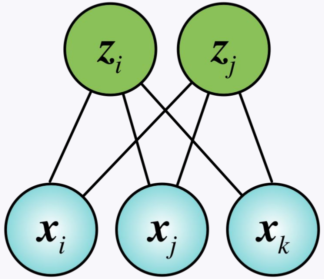
- Method-of-moments
- Approximate Bayesian Computation (ABC)
- Generative adversarial network (GAN)

Latent variable model + variational inference



- VEM algorithm
- Expectation propagation
- Approximate message passing
- Variational auto-encoders (VAE)

Restricted Boltzmann Machine + maximum likelihood



- Contrastive Divergence
- Persistent CD
- Parallel Tempering
- Natural gradients

Inference Questions?

Objective	Quantity of Interest
Prediction	$p(x_{(t+1),\ldots,\infty} \vert x_{-\infty,\ldots,t})$
Planning	$J = \mathbb{E}_p\left[\int_0^\infty dt C(x_t)\middle\vert x_0, u\right]$
Parameter estimation	$p(\theta \vert x_{0,\ldots,N})$
Experimental Design	$\mathrm{EIG} = \mathrm{D}[p(f(x_{t,\ldots,\infty})\vert u); p(f(x_{-\infty,\ldots,t}))]$
Hypothesis testing	$\frac{p(f(x_{-\infty,\ldots,t})\vert H_0)}{p(f(x_{-\infty,\ldots,t})\vert H_1)}$

Approximate Inference

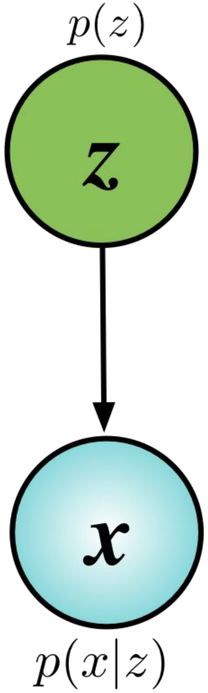
Latent Variable Models

$$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$$

$$\mathcal{D} = \{x_i\} \quad i \in \{1, \ldots, N\}$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz = \log \mathbb{E}_{p(z)}[p_\theta(x|z)]$$

$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)]$$



Methods for Approximate Inference

- **Laplace approximations**
- **Importance sampling**
- **Variational approximations**
- **Perturbative corrections**
- Other methods: MCMC, Langevin, HMC, Adaptive MCMC

Laplace Approximation

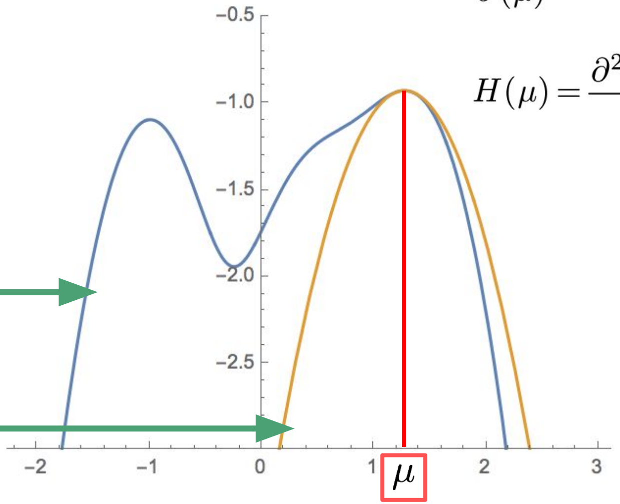
$$J(\mu) = \frac{\partial u(x,z)}{\partial z}\Big|_{z=\mu} = 0$$

$$H(\mu) = \frac{\partial^2 u(x,z)}{\partial z \partial z}\Big|_{z=\mu}$$

$$\begin{aligned}\log \mathbb{E}_{p(z)}[p_\theta(x|z)] &= \log \int p_\theta(x|z)p(z)dz \\ &= \log \int e^{-u(x,z)}dz\end{aligned}$$

$$u(x,z) = -\log p_\theta(x|z)p(z)$$

$$u(x,z) \approx u(x,\mu) + \frac{1}{2}(z-\mu)^T H(\mu)(z-\mu)$$



$$\begin{aligned}\log \mathbb{E}_{p(z)}[p_\theta(x|z)] &\approx \log \int e^{-u(x,\mu) - \frac{1}{2}(z-\mu)^T H(\mu)(z-\mu)}dz \\ &= -u(x,\mu) - \frac{1}{2}\ln \det(2\pi H^{-1}(\mu))\end{aligned}$$

Other names

Saddle-point approximation, Delta-method

Importance Sampling

$$\begin{aligned}
\log p(x_i) &= \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)] \\
&= \log \mathbb{E}_{q_\phi(z|x_i)}\left[\frac{p_\theta(x_i|z)\boxed{p(z)}}{\boxed{q_\phi(z|x_i)}}\right] \\
&= \log \mathbb{E}_{q_\phi(z|x_i)}[e^{-\mathcal{F}(x_i,z)}] \\
&\approx \log \sum_{k=1}^{K} e^{-\mathcal{F}(x_i,z_k)} - \log K
\end{aligned}$$

Importance weights

Monte-Carlo

$$\mathcal{F}(x,z) = \ln q(z|x) - \ln p(z) - \ln p(x|z)$$

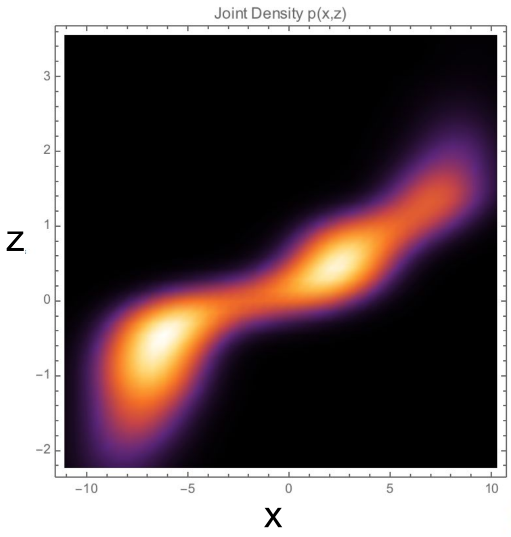
Pointwise Free-energy

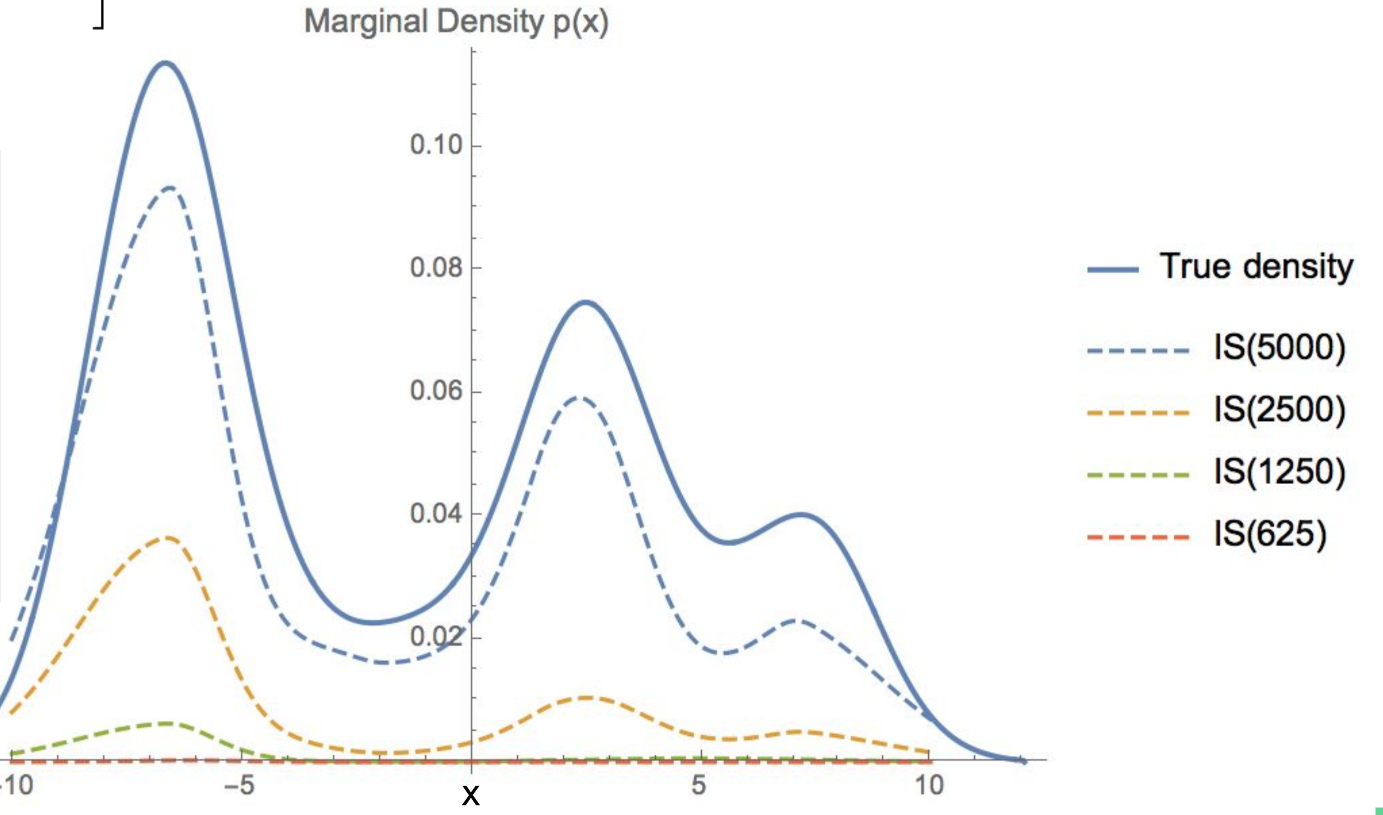
$$\log p(x) \geqslant \mathbb{E}_{q_\phi(z|x_i)}\left[\log \sum_{k=1}^{K} e^{-\mathcal{F}(x_i,z_k)}\right] - \log K$$

Important property

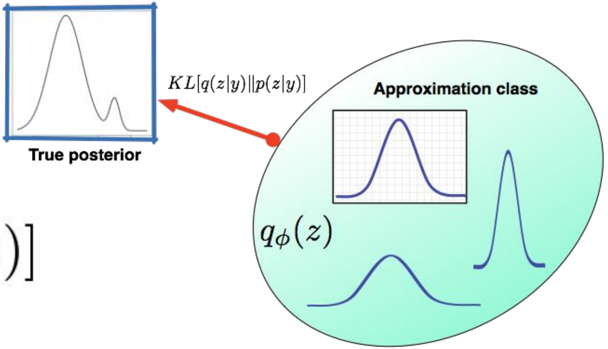
Importance sampling provides a bound in expectation

$$\log p(x) \quad \geqslant \mathbb{E}_{q_\phi(z|x)}\left[\log \sum_{k=1}^{K} e^{-\mathcal{F}(x,z_k)}\right] - \log K$$





Variational Inference



$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)]$$

$$\log \mathbb{E}_{p(z)}[p_\theta(x_i|z)] = \log \mathbb{E}_{q_i(z)}\left[\frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right], \quad \forall q_i > 0$$

$$\log \mathbb{E}_{q_i(z)}\left[\frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right] \geqslant \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

$$\log p_\theta(\mathcal{D}) \geqslant \sum_{i=1}^{N} \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

Variational Inference

$$\log p_\theta(\mathcal{D}) \geqslant \sum_{i=1}^{N} \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

$$\mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right] = \boxed{\mathbb{E}_{q_i(z)}[\log p_\theta(x_i|z)]} - \boxed{\mathrm{KLD}(q_i \| p)}$$

Reconstruction

Regularizer

Perturbative Corrections

$$\begin{aligned}\mathcal{F}(x,z) &= \ln q(z|x) + u(x,z)\\ \mathcal{F}(x) &= \mathbb{E}_{q(z|x)}[\mathcal{F}(x,z)]\\ \Delta &= -\mathcal{F}(x,z) + \mathcal{F}(x)\end{aligned}$$

$$\begin{aligned}\log \mathbb{E}_{p(z)}[p_\theta(x|z)] &= \log \int e^{-u(x,z)} dz\\ &= -\mathcal{F}(x) + \log \mathbb{E}_{q(z|x)}[e^{\Delta(x,z)}]\\ &= -\mathcal{F}(x) + \log \mathbb{E}_{q(z|x)}\left[\sum_{k=0}^{\infty} \frac{\Delta(x,z)^k}{k!}\right]\\ &= -\mathcal{F}(x) + \log \sum_{k=0}^{\infty} \frac{1}{k!} \mathbb{E}_{q(z|x)}[\Delta(x,z)^k]\end{aligned}$$

$$e^y = \sum_{k=0}^{\infty} \frac{y^k}{k!}$$

Design Choices

Choice of Model

Computation graphs, Renderers, simulators and environments

Variational Optimisation

- Variational EM
- Stochastic VEM
- Monte Carlo gradient estimators

Approximate Posteriors

- Mean-field
- Structured approx
- Aux. variable methods

Variational EM Algorithm

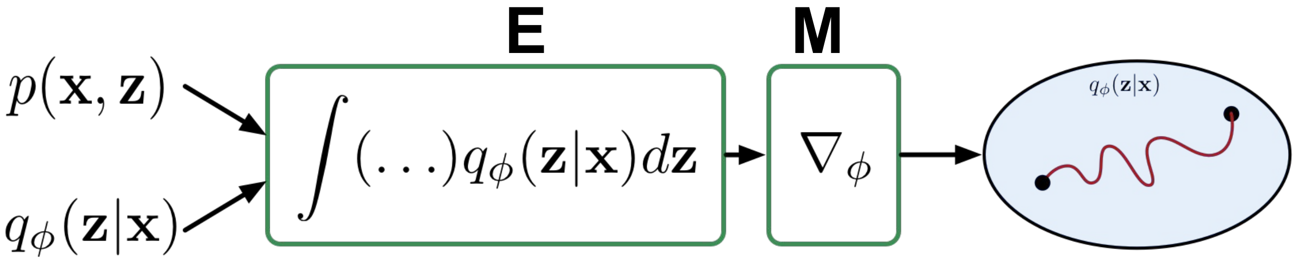
Fixed-point iterations between variational and model parameters

E

$$q_i^\star(z) \;= \operatorname{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i^\star(z)}\right] \Leftrightarrow q_i^\star(z) = \frac{p_\theta(x_i|z)p(z)}{p(x_i)}$$

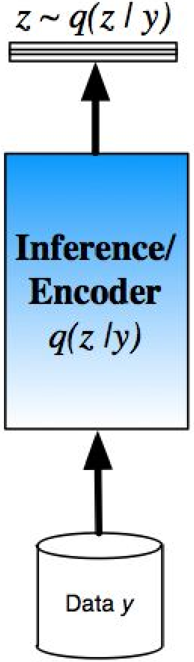
M

$$\theta^\star \;= \operatorname{argmax}_\theta \sum_{i=1}^{N} \mathbb{E}_{q_i^\star(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i^\star(z)}\right]$$



E
M
$p(\mathbf{x}, \mathbf{z})$
$q_\phi(\mathbf{z}|\mathbf{x})$
$\int (\dots) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z}$
∇_ϕ
$q_\phi(\mathbf{z}|\mathbf{x})$

Amortised Inference



$$q_i^\star(z) \quad = \operatorname{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}[-\mathcal{F}(x_i, z)]$$

Introduce a parametric family of conditional densities

$$\operatorname{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}[-\mathcal{F}(x_i, z)] \Rightarrow \operatorname{argmax}_{\phi} \mathbb{E}_{q_\phi(z \mid x)}[-\mathcal{F}_\phi(x_i, z)]$$

Rezende et al., 2015

Variational Auto-encoders

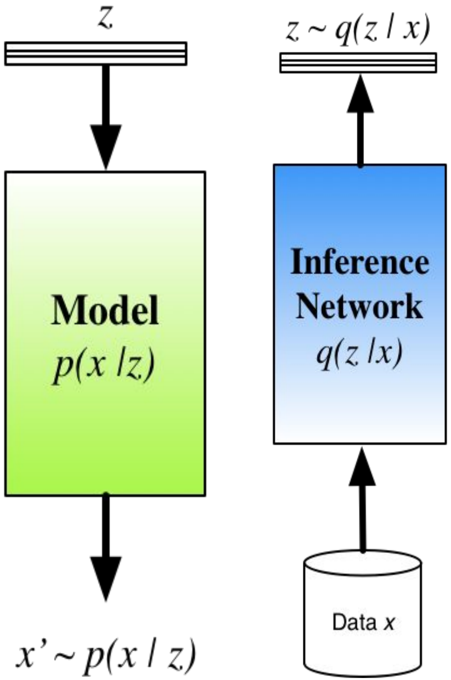
Simplest instantiation of a VAE

Deep Latent Gaussian Model p(x,z)

$$\begin{aligned} \text{prior sample} \quad & z \sim \mathcal{N}(0, \mathbb{I}) \\ \text{data sufficient statistics} \quad & \eta = f_\theta(z) \\ \text{data conditional likelihood} \quad & x \sim \mathcal{N}(\eta) \end{aligned}$$

Gaussian Recognition Model q(z)

$$\begin{aligned} \text{data sample} \quad & x \sim \mathcal{D} \\ \text{latent sufficient statistics} \quad & \eta = f_\phi(x) \\ \text{posterior sample} \quad & z \sim \mathcal{N}(\eta) \end{aligned}$$



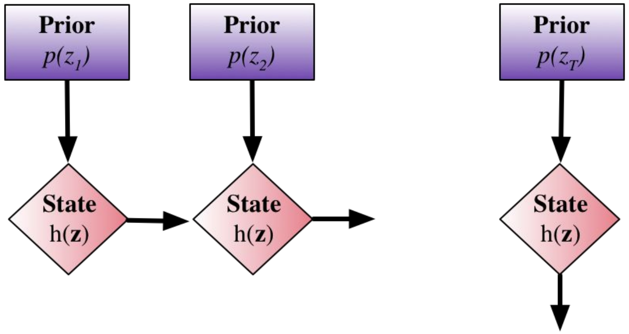
$$\mathbb{E}_{q_i(z)}[\log p_\theta(x_i|z)] - \mathrm{KLD}(q_i\|p)$$

We then optimise the free-energy wrt model and variational parameters

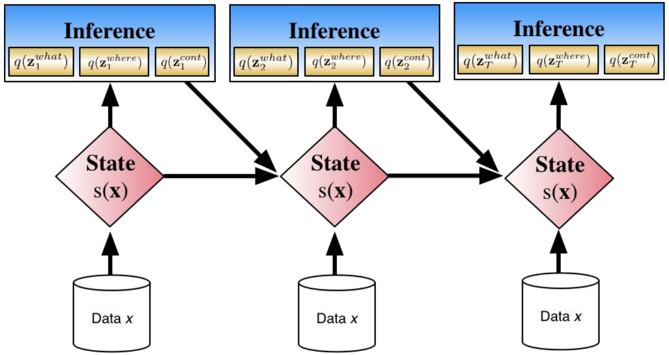
Kingma and Welling, 2014, Rezende et al., 2014

Richer VAES

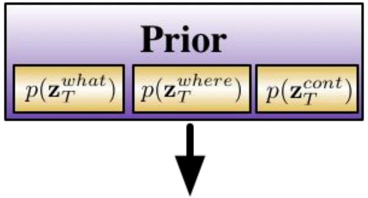
DRAW: Recurrent/Dependent Priors



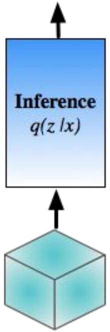
Recurrent/Dependent Inference Networks


Inference
$q(\mathbf{z}_1^{what})$ $q(\mathbf{z}_1^{where})$ $q(\mathbf{z}_1^{cont})$
Inference
$q(\mathbf{z}_2^{what})$ $q(\mathbf{z}_2^{where})$ $q(\mathbf{z}_2^{cont})$
Inference
$q(\mathbf{z}_T^{what})$ $q(\mathbf{z}_T^{where})$ $q(\mathbf{z}_T^{cont})$
State s(**x**)
State s(**x**)
State s(**x**)
Data *x*
Data *x*
Data *x*

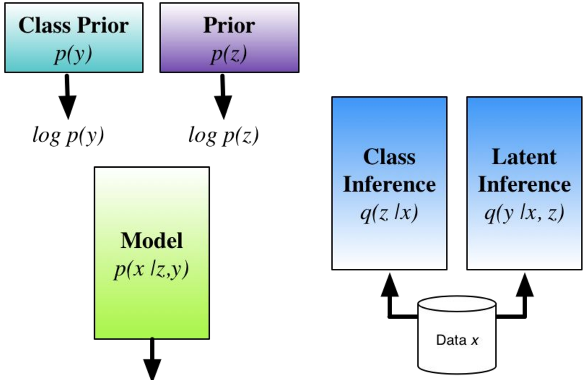
AIR: Structured Priors


Prior
$p(\mathbf{z}_T^{what})$ $p(\mathbf{z}_T^{where})$ $p(\mathbf{z}_T^{cont})$

Volumetric and Sequence data

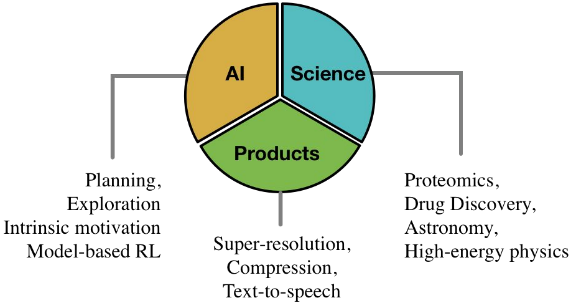


Semi-supervised Learning



Summary so far

Applications of Generative Models

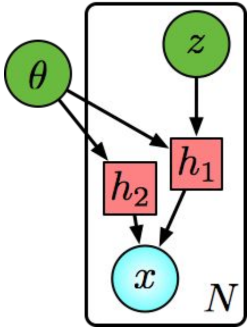


Planning,
Exploration
Intrinsic motivation
Model-based RL

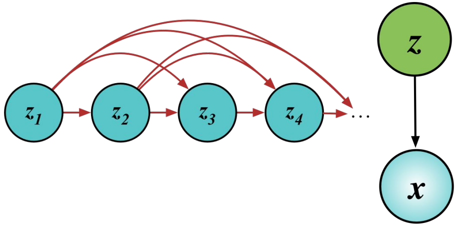
Super-resolution,
Compression,
Text-to-speech

Proteomics,
Drug Discovery,
Astronomy,
High-energy physics

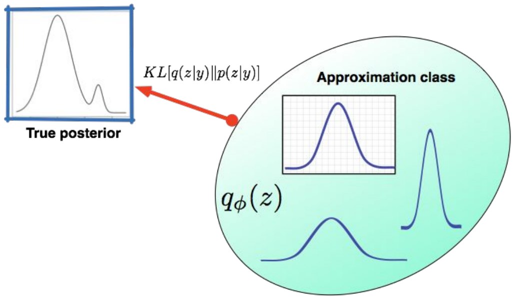
Probabilistic Deep Learning



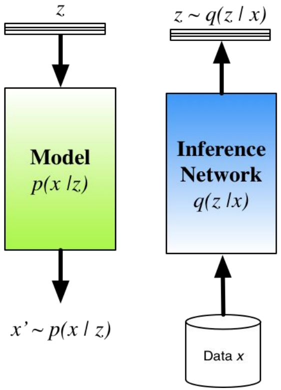
Types of Generative Models



Variational Principles



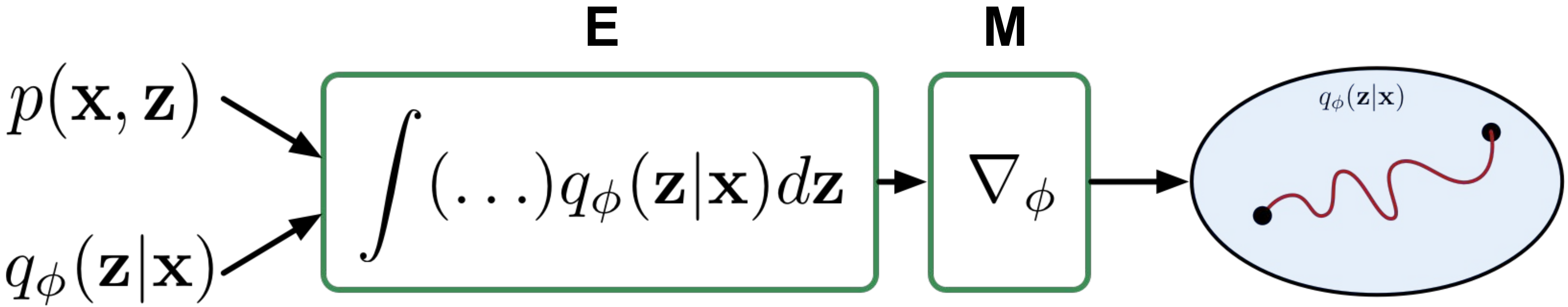
Amortised Inference



END OF FIRST HALF

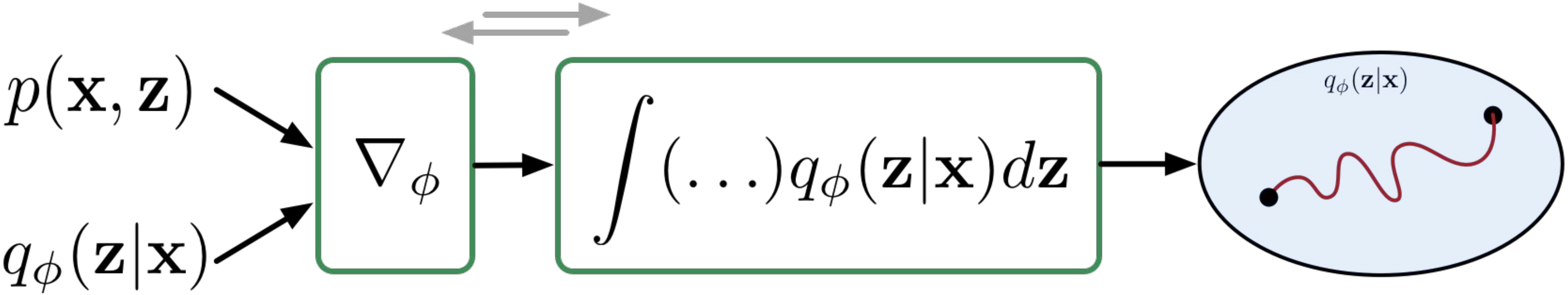
Stochastic Optimisation

Classical Inference Approach


E
M
$p(\mathbf{x}, \mathbf{z})$
$q_\phi(\mathbf{z}|\mathbf{x})$
$\int (\ldots) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z}$
∇_ϕ
$q_\phi(\mathbf{z}|\mathbf{x})$

Compute expectations then M-step gradients

Stochastic Inference Approach


$p(\mathbf{x}, \mathbf{z})$
$q_\phi(\mathbf{z}|\mathbf{x})$
∇_ϕ
$\int (\ldots) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z}$
$q_\phi(\mathbf{z}|\mathbf{x})$

In general, we won't know the expectations.

Gradient is of the parameters of the distribution w.r.t. which the expectation is taken.

Stochastic Gradient Estimators

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int \boxed{q_\phi(\mathbf{z})}\boxed{f_\theta(\mathbf{z})} d\mathbf{z}$$

Score-function estimator:
Differentiate the density *q(**z**|**x**)*

Pathwise gradient estimator:
Differentiate the function *f(**z**)*

Typical problem areas:

- Generative models and inference
- Reinforcement learning and control
- Operations research and inventory control
- Monte Carlo simulation
- Finance and asset pricing
- Sensitivity estimation

Fu, 2006

Score Function Estimators

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int \boxed{q_\phi(\mathbf{z})} f_\theta(\mathbf{z}) d\mathbf{z}$$

$$\boxed{= \mathbb{E}_{q(z)}[f_\theta(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}))]}$$

Gradient reweighted by the value of the function

Other names:

- Likelihood-ratio trick
- Radon-Nikodym derivative
- REINFORCE and policy gradients
- Automated inference
- Black-box inference

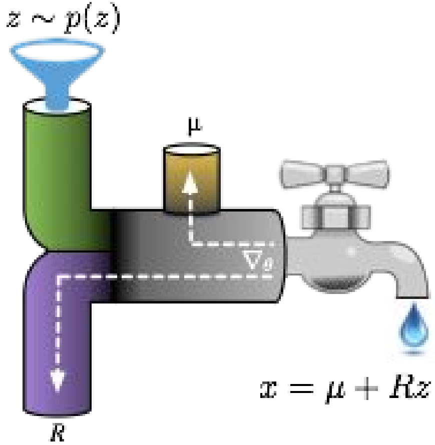
When to use:

- Function is not differentiable.
- Distribution *q* is easy to sample from.
- Density *q* is known and differentiable.

Reparameterisation

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int q_\phi(\mathbf{z}) \boxed{f_\theta(\mathbf{z})} d\mathbf{z}$$

Find an invertible function *g(.)* that expresses z as a transformation of a base distribution .

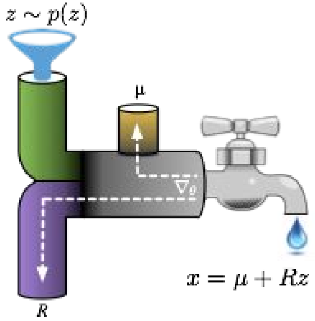


$$\mathbf{z} = g_\phi(\boldsymbol{\epsilon}) \qquad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$$

$$\mathbb{E}_{q_\phi(z|x)}[f(z)] = \mathbb{E}_{p(\epsilon)}[f(g_\phi(x, \epsilon))]$$

Kingma and Welling, 2014, Rezende et al., 2014

Pathwise Derivative Estimator



$$\mathbf{z} = g(\epsilon, \phi) \quad \epsilon \sim p(\epsilon)$$

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int q_\phi(\mathbf{z}) \boxed{f_\theta(\mathbf{z})} d\mathbf{z}$$

$$\boxed{= \mathbb{E}_{p(\epsilon)}[\nabla_\phi f_\theta(g(\epsilon, \phi))]}$$

Other names:

- Reparameterisation trick
- Stochastic backpropagation
- Perturbation analysis
- Affine-independent inference
- Doubly stochastic estimation
- Hierarchical non-centred parameterisations.

When to use

- Function *f* is differentiable
- Density *q* can be described using a simpler base distribution: inverse CDF, location-scale transform, or other co-ordinate transform.
- Easy to sample from base distribution.

Gaussian Stochastic Gradients

$$\nabla_\phi \mathbb{E}_{\mathcal{N}(\mu, CC^\top)}[f_\theta(\mathbf{z})]$$

First-order Gradient

$$p(\epsilon) = \mathcal{N}(0,1) \quad g(\epsilon, \phi) = \mu_\phi(x) + C_\phi(x)\epsilon$$

$$\mathbb{E}_{p(\epsilon)}[J^\top(\nabla_\phi \mu_\phi + \nabla_\phi C_\phi^\top \epsilon)]$$

Second-order Gradient

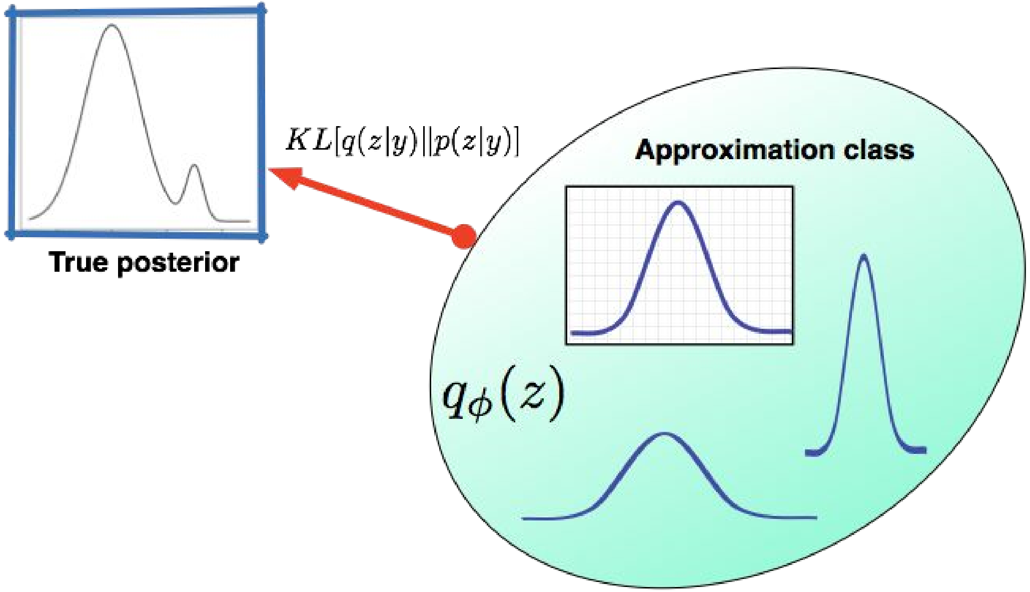
$$\mathbb{E}_{q(z)}[J^\top \nabla_\phi \mu_\phi + Tr[HC_\phi \nabla_\phi C_\phi]]$$

We can develop low-variance estimators by exploiting knowledge of the distributions involved when we know them

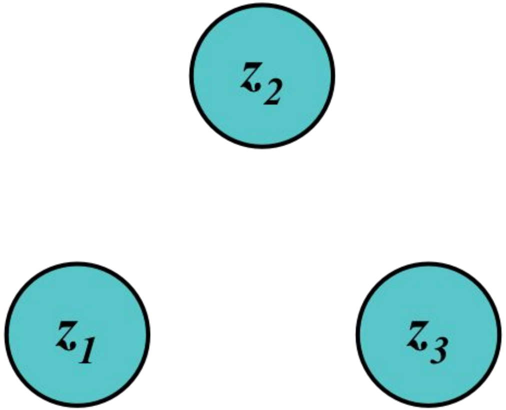
Rezende et al., 2014

Beyond the Mean Field

Mean Field Approximations



Fully-factorised



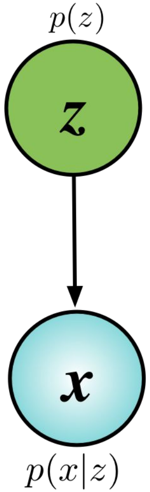
$$q_{MF}(z|x) = \prod_k q(z_k)$$

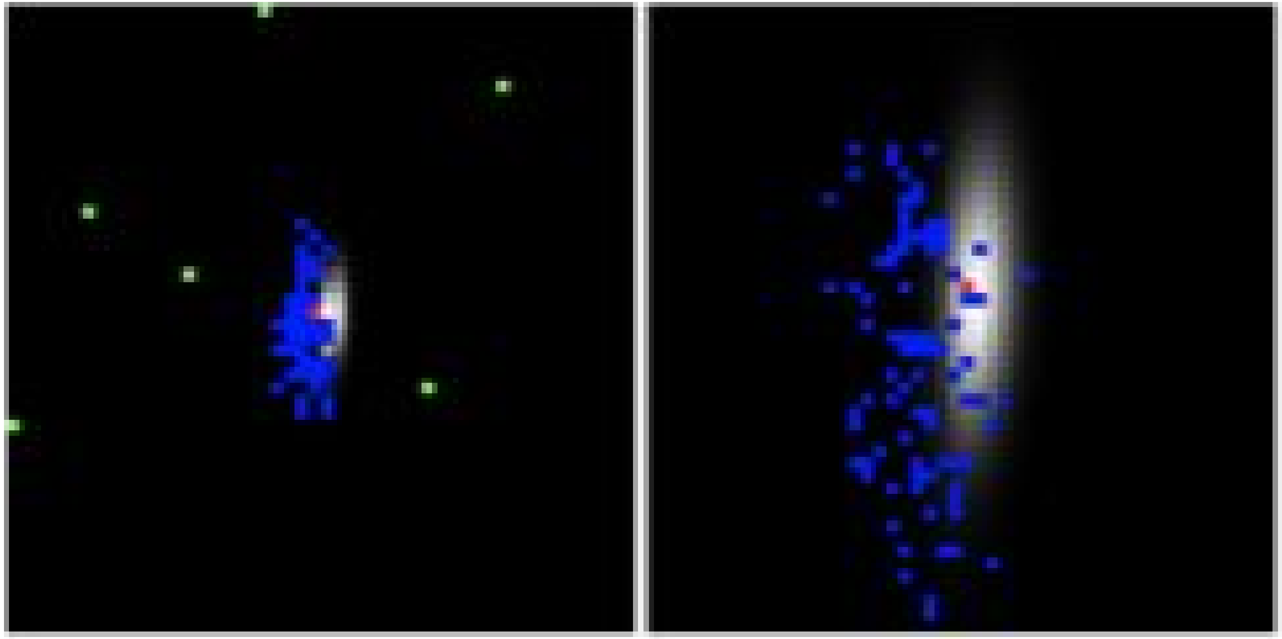
Key part of variational inference is choice of approximate posterior distribution q.

$$\mathcal{F}(q, \boldsymbol{\theta}) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x}|)||p(\mathbf{z})]$$

Mean-Field Posterior Approximations

Deep Latent Gaussian Model

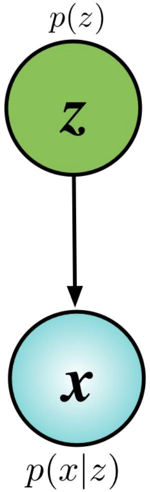


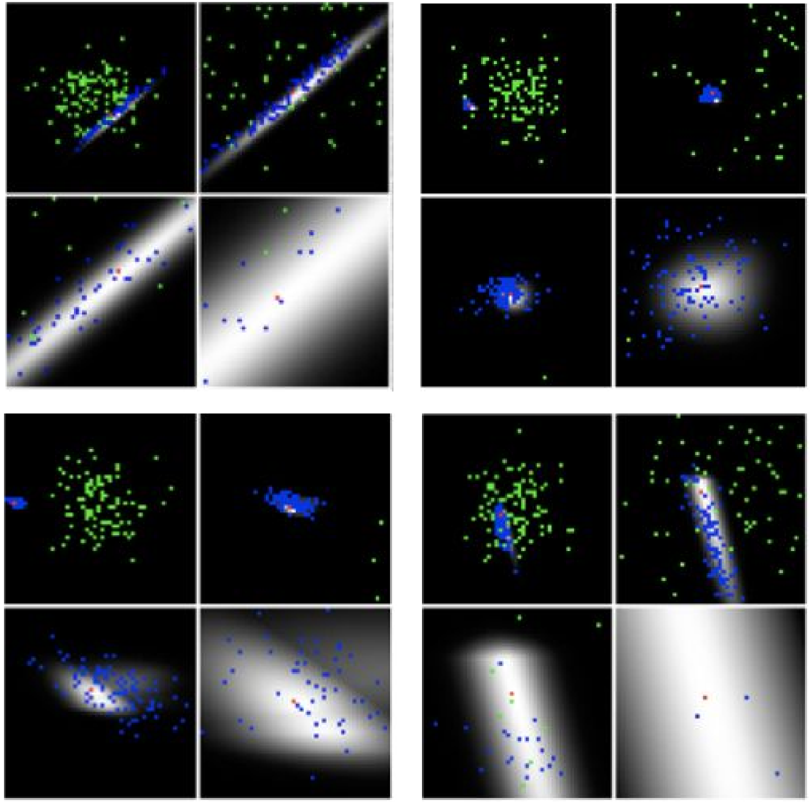


Mean-field or fully-factorised posterior is usually not sufficient

Real-world Posterior Distributions

Deep Latent Gaussian Model



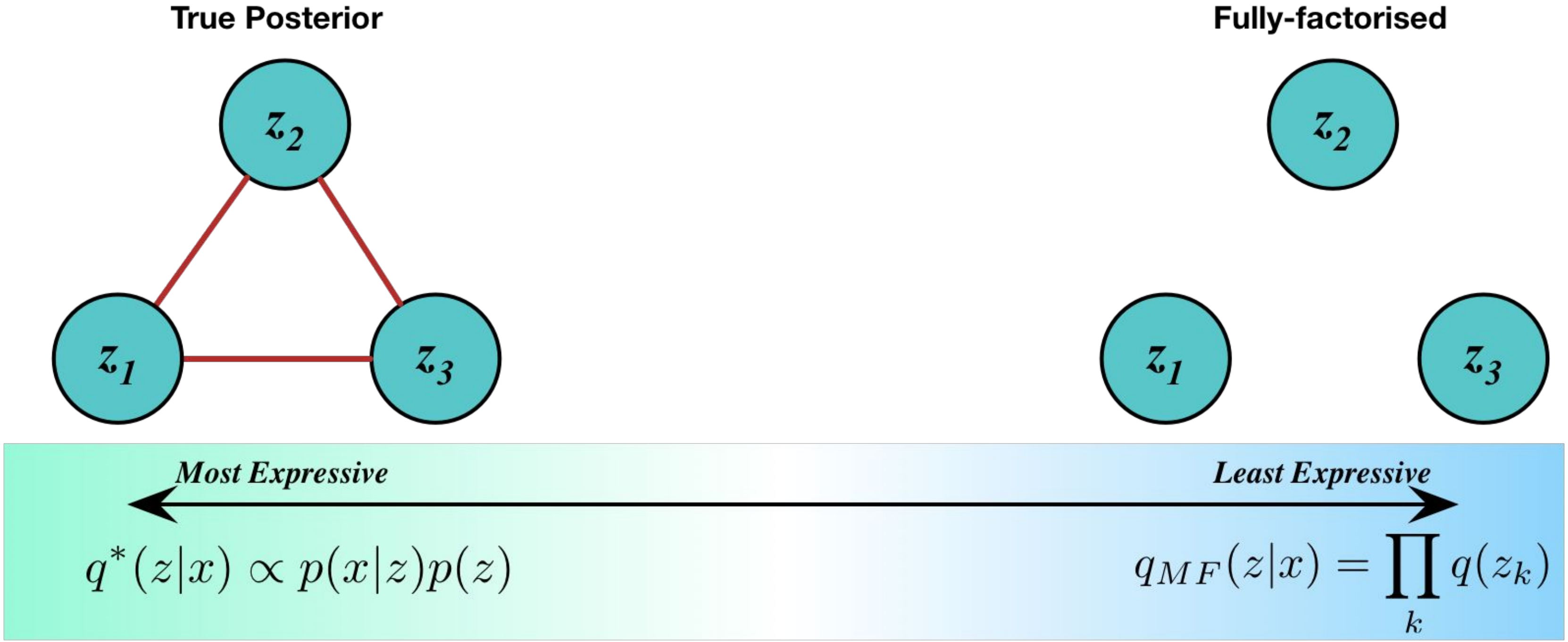


Complex dependencies · Non-Gaussian distributions · Multiple modes

Richer Families of Posteriors

Two high-level goals:

- Build richer approximate posterior distributions.
- Maintain computational efficiency and scalability.



True Posterior

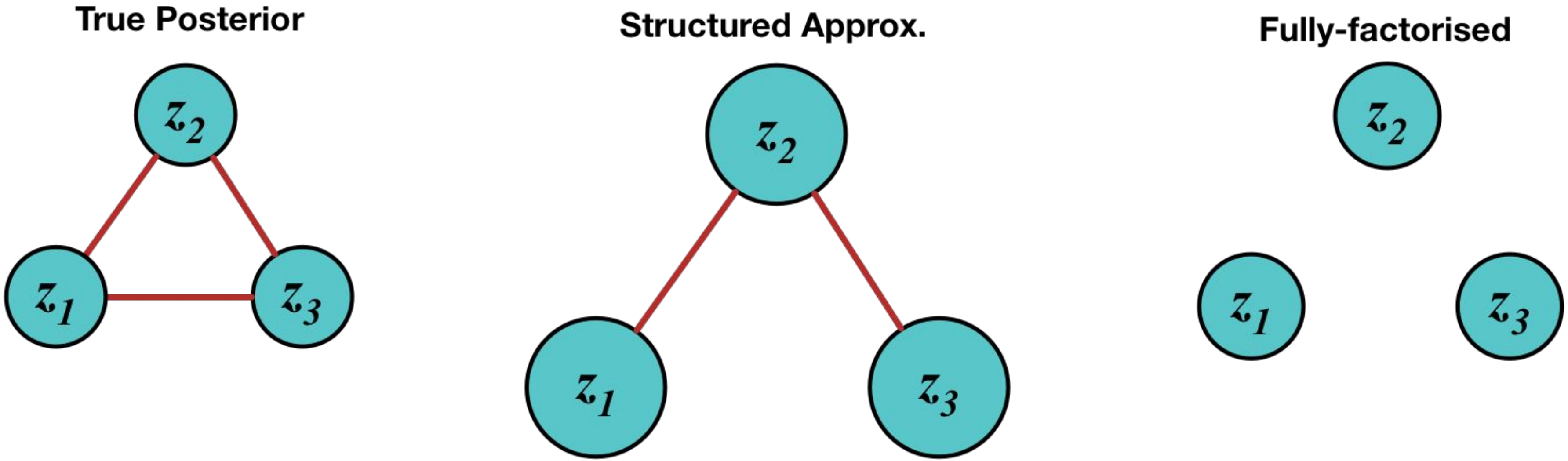
Fully-factorised

$$q^*(z|x) \propto p(x|z)p(z)$$

$$q_{MF}(z|x) = \prod_k q(z_k)$$

Same as the problem of specifying a model of the data itself

Structured Approximations



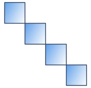
Most Expressive ⟷ *Least Expressive*

$$q^*(z|x) \propto p(x|z)p(z) \qquad q(z) = \prod_k q_k(z_k|\{z_j\}_{j\neq k}) \qquad q_{MF}(z|x) = \prod_k q(z_k)$$

Families of Approximate Posteriors

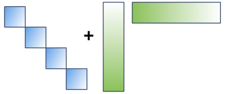
Covariance Models

$\mathrm{diag}(\alpha_1, \ldots, \alpha_K)$



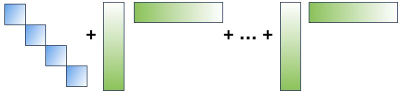
Mean-field

$\mathrm{diag}(\alpha_1, \ldots, \alpha_K)$
$+\mathbf{u}\mathbf{u}^\top$



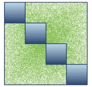
Rank-1

$\mathrm{diag}(\alpha_1, \ldots, \alpha_K)$
$+\sum_j \mathbf{u_j}\mathbf{u_j}^\top$



Rank-J

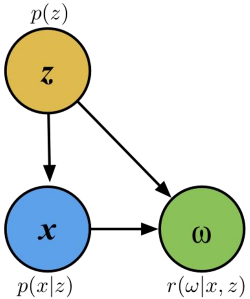
$\mathbf{U}\mathbf{U}^\top$



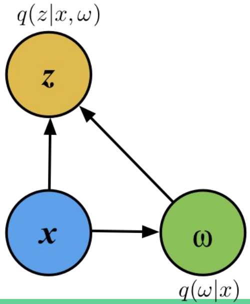
Full

Auxiliary Variable Models

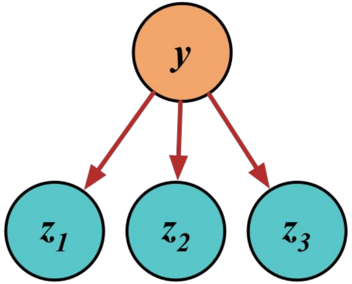
Auxiliary latent variable model p(x,z,ω)



Inference model q(z,ω)

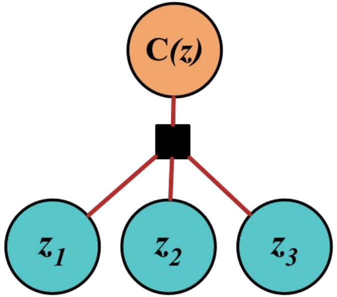


Mixture model



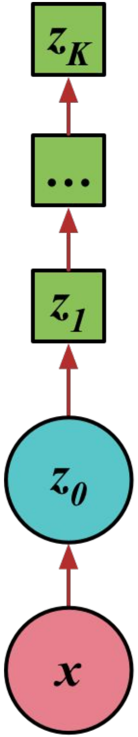
$$q_{mm}(\mathbf{z};\boldsymbol{\nu}) = \sum_r \rho_r q_r(\mathbf{z}_r|\boldsymbol{\nu}_r)$$

Copula Methods



$$q_{lm}(\mathbf{z};\boldsymbol{\nu}) = \left(\prod_k q_k(z_k|\boldsymbol{\nu}_k)\right) C(\mathbf{z};\boldsymbol{\nu}_{k+1})$$

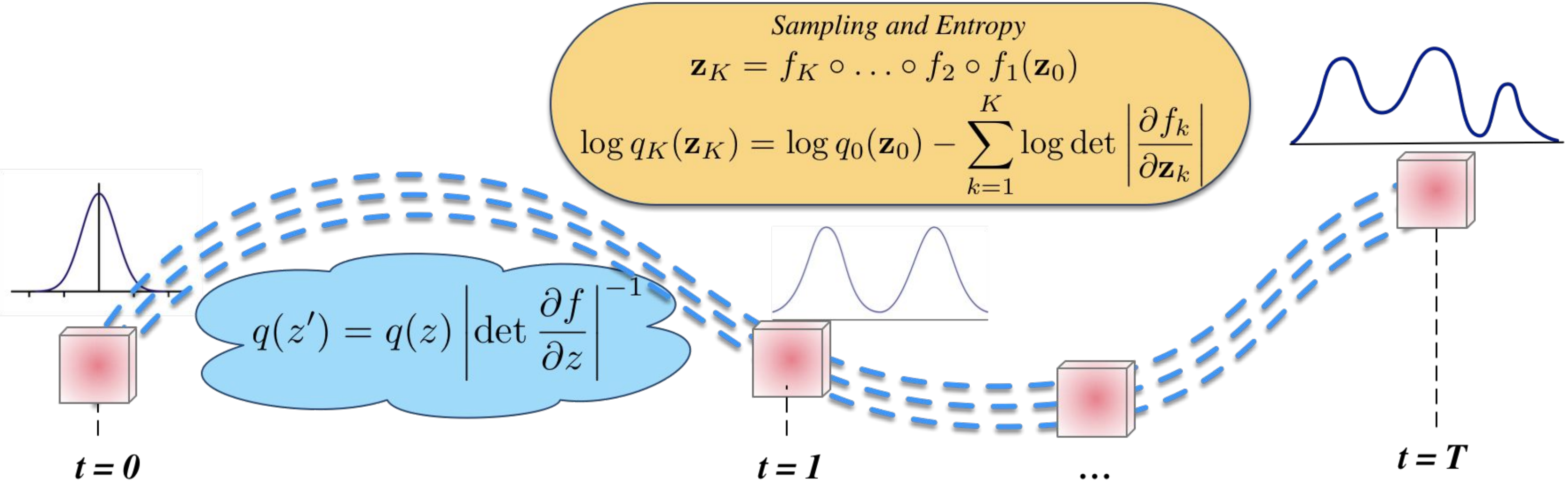
Normalising Flows



Normalising Flows

Exploit the rule for change of variables:

- Begin with an initial distribution
- Apply a sequence of K invertible transforms

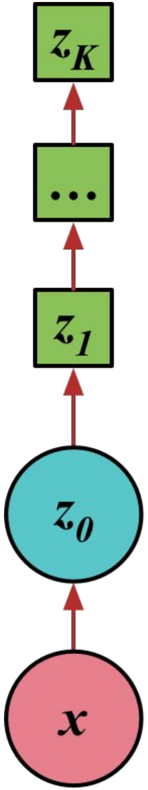


Sampling and Entropy

$$\mathbf{z}_K = f_K \circ \ldots \circ f_2 \circ f_1(\mathbf{z}_0)$$

$$\log q_K(\mathbf{z}_K) = \log q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \log \det \left| \frac{\partial f_k}{\partial \mathbf{z}_k} \right|$$

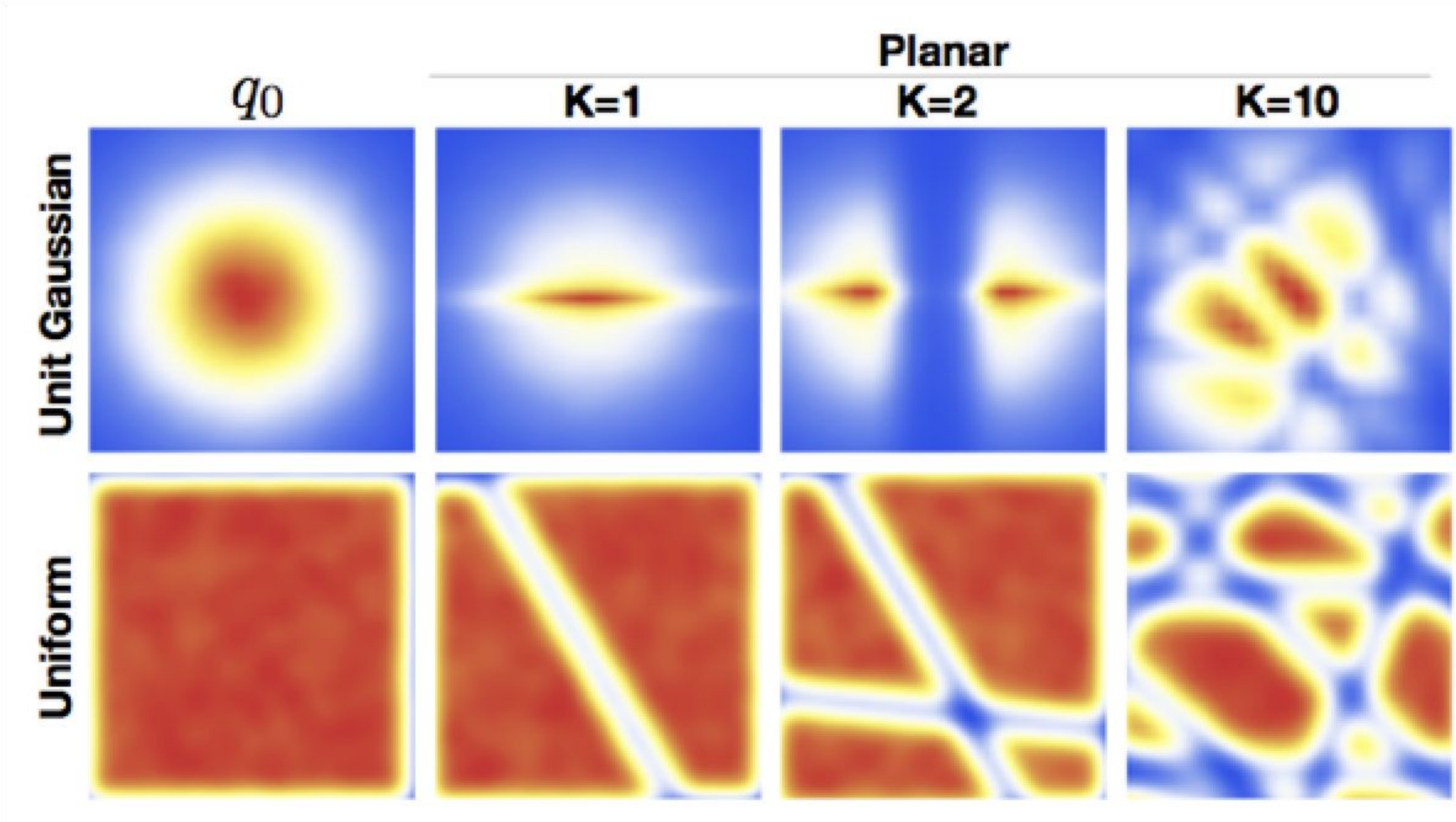
$$q(z') = q(z) \left| \det \frac{\partial f}{\partial z} \right|^{-1}$$



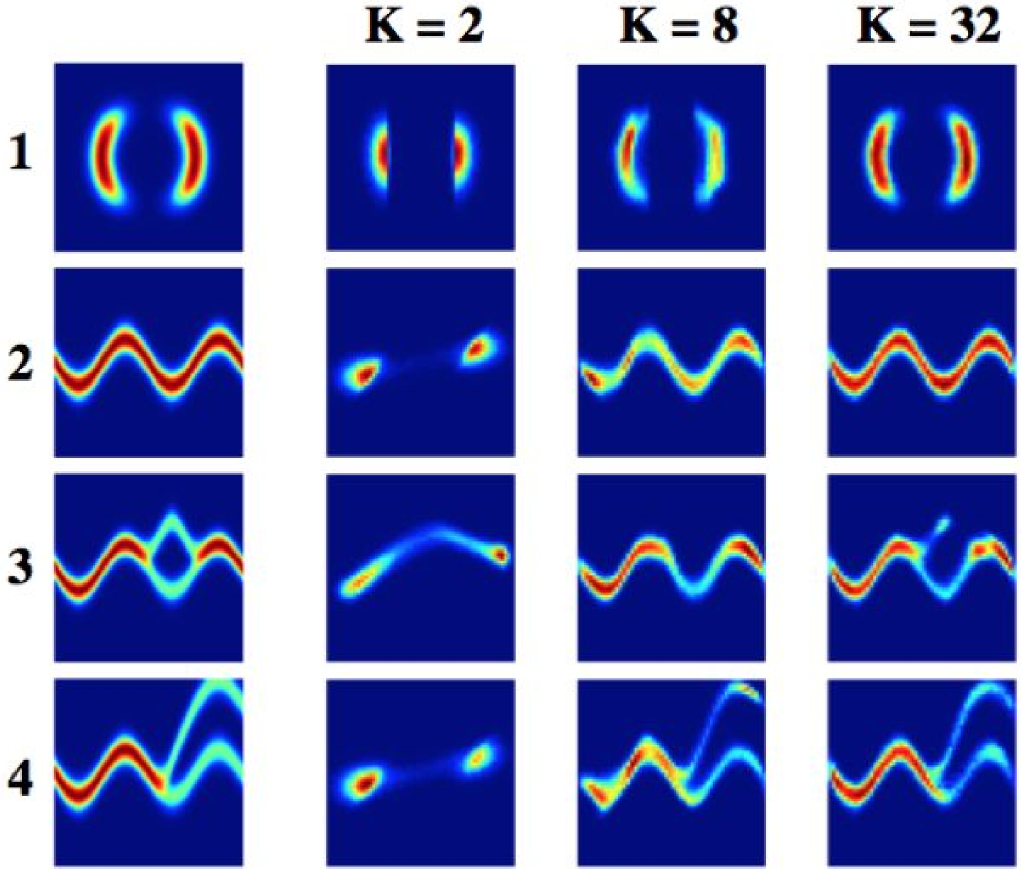
Distribution flows through a sequence of invertible transforms

Rezende and Mohamed, 2015

Normalising Flows



Normalising Flows

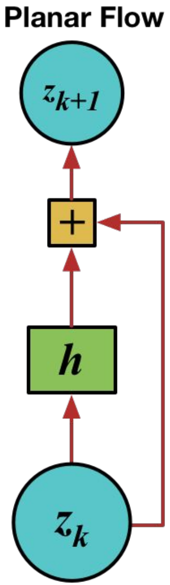


Choice of Transformation

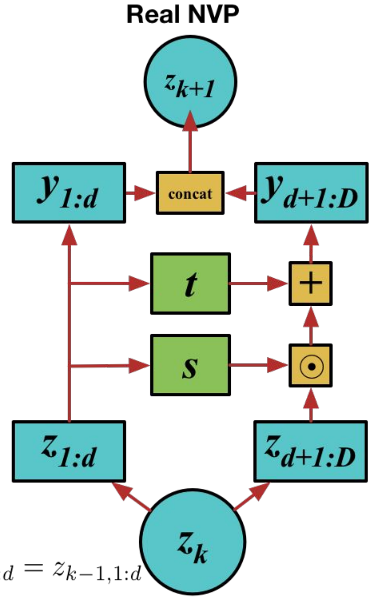
$$\mathcal{L} = \mathbb{E}_{q_0(\mathbf{z}_0)}[\log p(\mathbf{x}, \mathbf{z}_K)] - \mathbb{E}_{q_0(\mathbf{z}_0)}[\log q_0(\mathbf{z}_0)] - \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\sum_{k=1}^{K} \log \det \left| \frac{\partial f_k}{\partial \mathbf{z}_k} \right| \right]$$

Begin with a fully-factorised Gaussian and improve by change of variables.

Triangular Jacobians allow for computational efficiency.

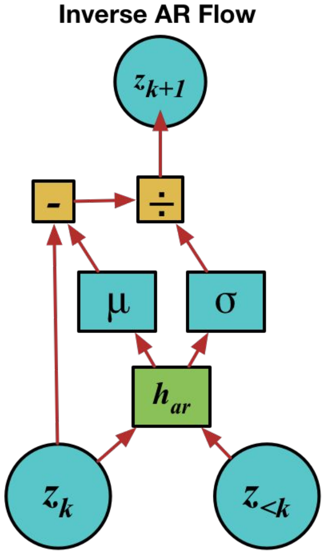


$z_k = z_{k-1} + uh(w^\top z_{k-1} + b)$



$y_{1:d} = z_{k-1,1:d}$

$y_{d+1:D} = t(z_{k-1,1:d}) + z_{d+1:D} \odot \exp(s(z_{k-1,1:d}))$

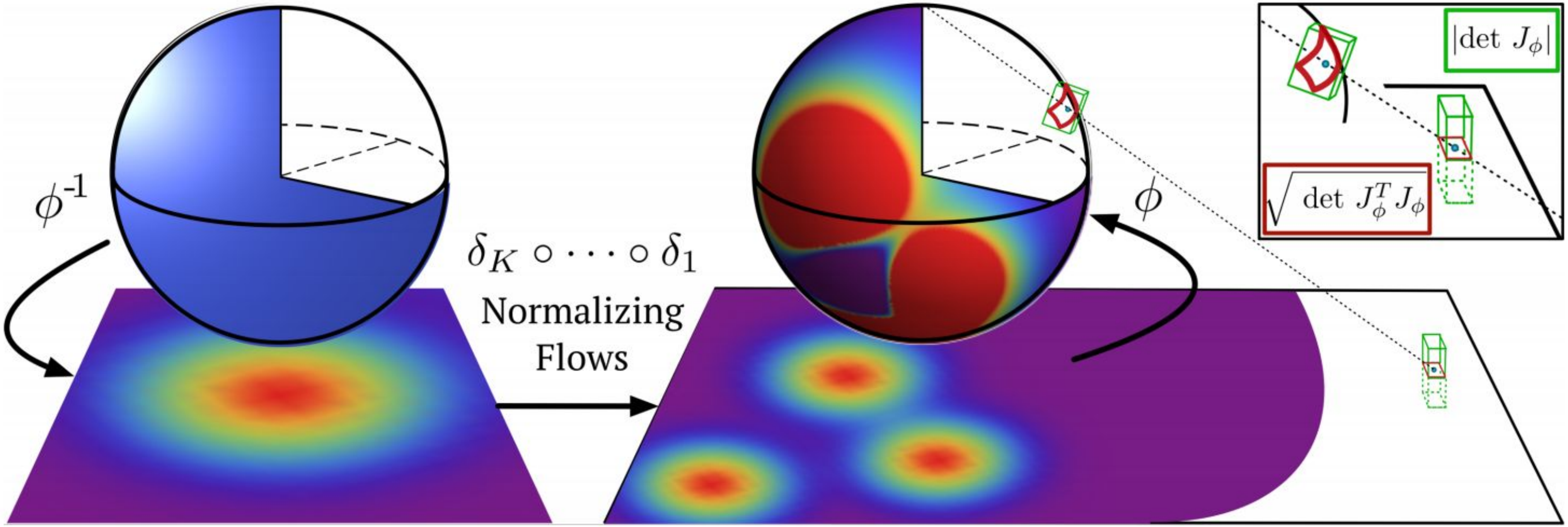


$z_k = \frac{z_{k-1} - \mu_k(z_{<k}, x)}{\sigma_k(z_{<k}, x)}$

Linear time computation of the determinant and its gradient.

Rezende and Mohamed, 2015; Dinh et al, 2016, Kingma et al, 2016

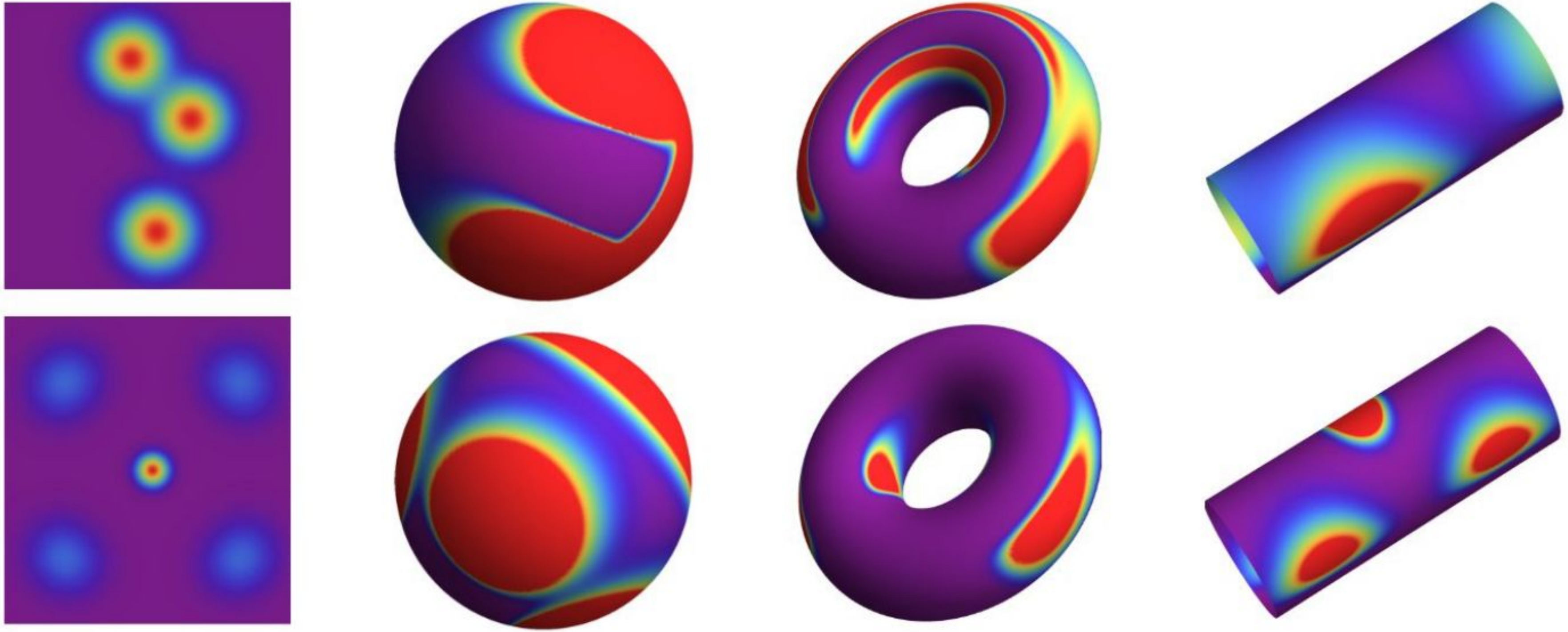
Normalising Flows on Non-Euclidean Manifolds


ϕ^{-1}
$\delta_K \circ \cdots \circ \delta_1$
Normalizing
Flows
ϕ
$|\det J_\phi|$
$\sqrt{\det J_\phi^T J_\phi}$

$$\log q_K(\mathbf{z}_K) = \log q_0(\mathbf{z}_0) - \frac{1}{2}\sum_{k=1}^{K} \log\det\left|\mathbf{J}_\phi{}^\top \mathbf{J}_\phi\right|$$

Gemici et al., 2016

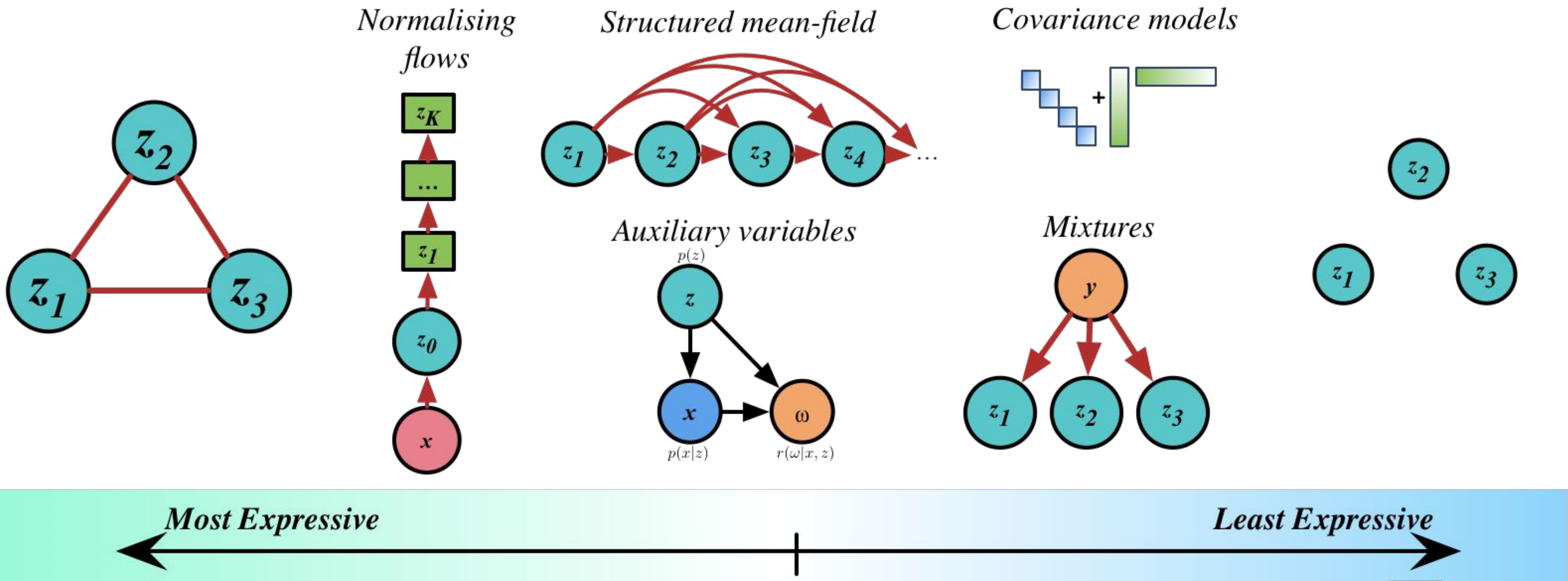
Normalising Flows on non-Euclidean Manifolds



True Posterior

Families of Posterior Approximations

Fully-factorised



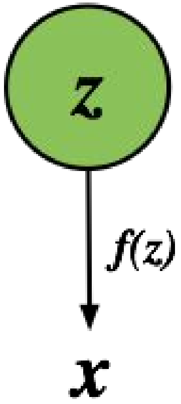
$$q^*(z|x) \propto p(x|z)p(z)$$

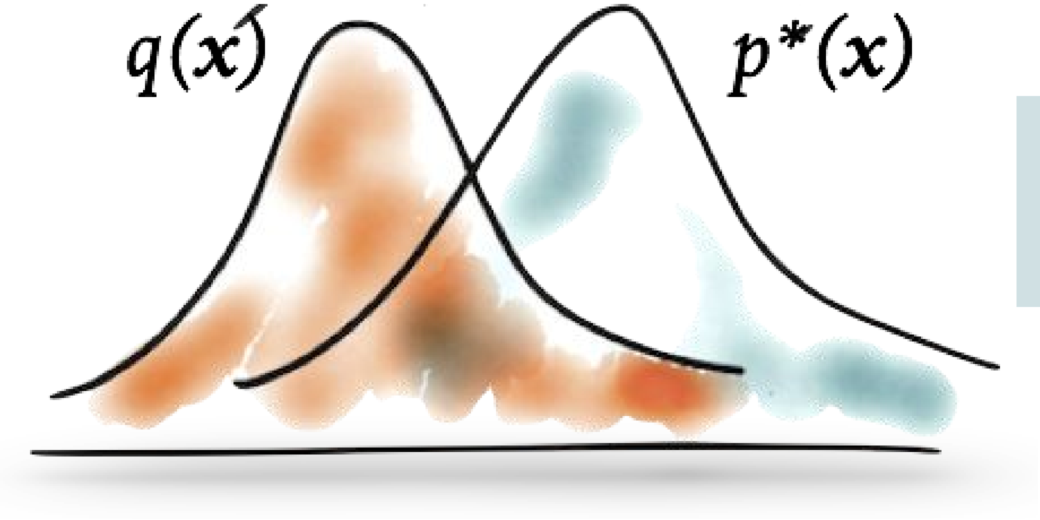
$$q_{MF}(z|x) = \prod_k q(z_k)$$

Learning in
Implicit Generative Models

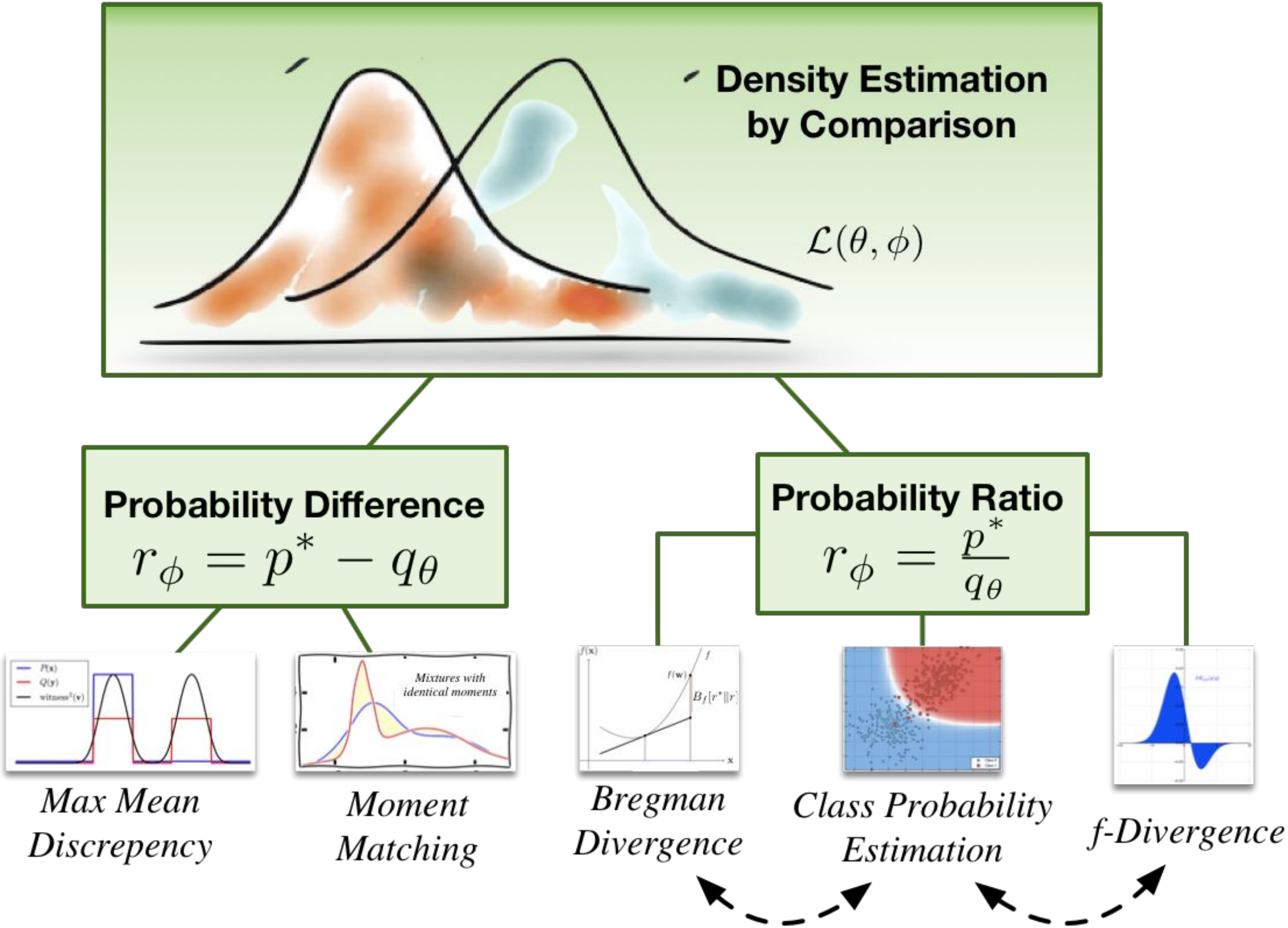
Learning by Comparison

For some models, we only have access to an unnormalised probability, partial knowledge of the distribution, or a simulator of data.





We compare the estimated distribution q(x) to the true distribution p*(x) using samples.



Density Estimation by Comparison

$\mathcal{L}(\theta, \phi)$

Probability Difference

$$r_\phi = p^* - q_\theta$$

Probability Ratio

$$r_\phi = \frac{p^*}{q_\theta}$$

Max Mean Discrepency

Moment Matching

Bregman Divergence

Class Probability Estimation

f-Divergence

$$f(u) = u \log u - (u+1)\log(u+1)$$

Mohamed and Lakshminarayanan, 2017.

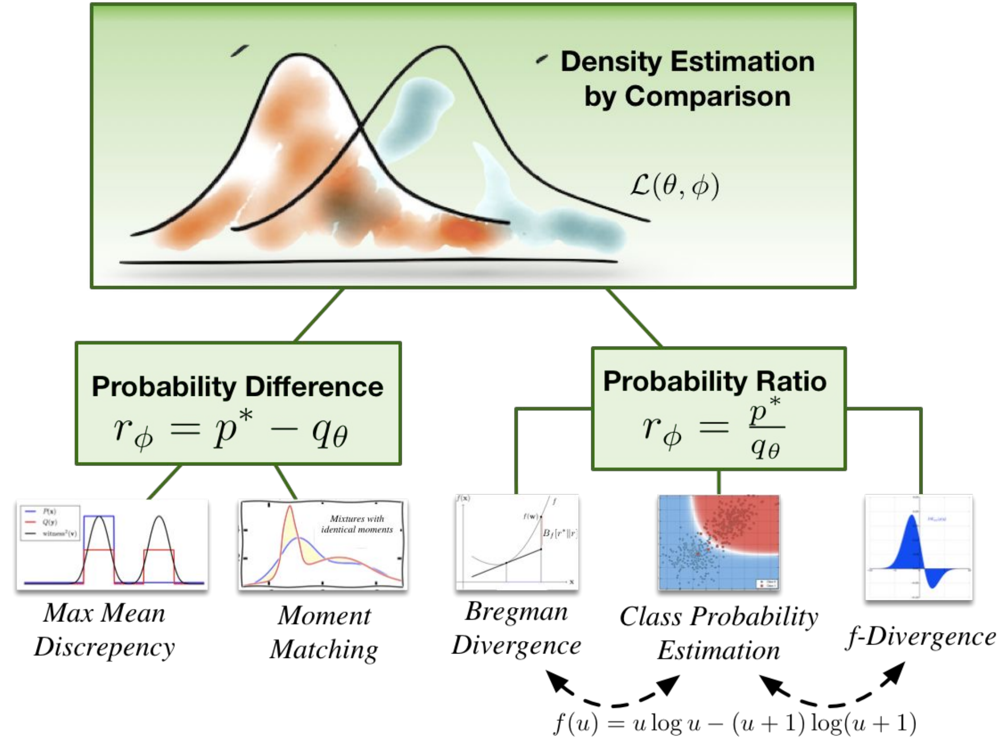
Learning by Comparison

Comparison

Use a hypothesis ***test or comparison*** to build an auxiliary model to indicate how data simulated from the model differs from observed data.

Estimation

Adjust model parameters to better match the data distribution using the comparison.



Density Estimation by Comparison

$\mathcal{L}(\theta, \phi)$

Probability Difference

$r_\phi = p^* - q_\theta$

Probability Ratio

$r_\phi = \frac{p^*}{q_\theta}$

Max Mean Discrepency

Moment Matching

Bregman Divergence

Class Probability Estimation

f-Divergence

$f(u) = u \log u - (u+1) \log(u+1)$

Density Ratios and Classification

Density Ratio $\frac{p^*(\mathbf{x})}{q(\mathbf{x})}$

Bayes' Rule $p(\mathbf{x}|y) = \frac{p(y|\mathbf{x})p(\mathbf{x})}{p(y)}$

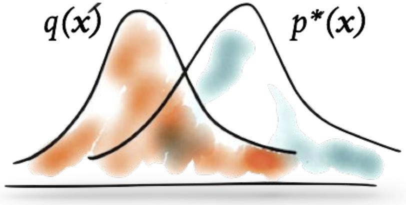
Real Data

Simulated Data

Combine data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{\hat{n}}, \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{\tilde{n}}\}$

Assign labels $\{y_1, \ldots, y_N\} = \{+1, \ldots, +1, -1, \ldots, -1\}$

Equivalence $p^*(\mathbf{x}) = p(\mathbf{x}|y=1) \qquad q(\mathbf{x}) = p(\mathbf{x}|y=-1)$



Sugiyama et al, 2012

Density Ratios and Classification

Conditional

$$\frac{p^*(\mathbf{x})}{q(\mathbf{x})} = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=-1)}$$

Bayes’ substitution

$$= \frac{p(y=+1|\mathbf{x})p(\mathbf{x})}{p(y=+1)} \bigg/ \frac{p(y=-1|\mathbf{x})p(\mathbf{x})}{p(y=-1)}$$

Class probability

$$\frac{p^*(\mathbf{x})}{q(\mathbf{x})} = \frac{p(y=1|\mathbf{x})}{p(y=-1|\mathbf{x})}$$

Computing a density ratio is equivalent to class probability estimation.

Unsupervised-as-Supervised Learning

Scoring Function

$$p(y = +1|\mathbf{x}) = D_\theta(\mathbf{x}) \qquad p(y = -1|\mathbf{x}) = 1 - D_\theta(\mathbf{x})$$

Bernoulli Loss

$$\mathcal{F}(\mathbf{x}, \theta, \phi) = \mathbb{E}_{p^*(x)}[\log D_\theta(\mathbf{x})] + \mathbb{E}_{q_\phi(x)}[\log(1 - D_\theta(\mathbf{x})]$$

Alternating optimisation

$$\min_\phi \max_\theta \mathcal{F}(\mathbf{x}, \theta, \phi)$$

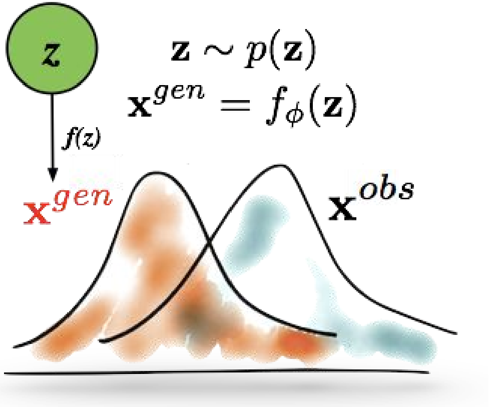
- Use when we have differentiable simulators and models
- Can form the loss using any proper scoring rule.

Other names and places:

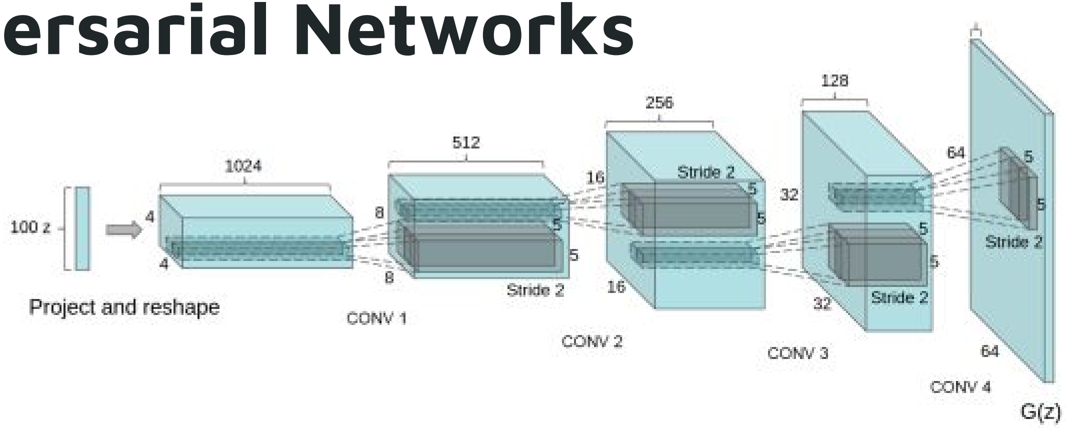
- Unsupervised and supervised learning
- Continuously updating inference
- Classifier ABC
- Generative Adversarial Networks

Friedman et al. 2001

Generative Adversarial Networks



$$\mathbf{z} \sim p(\mathbf{z})$$
$$\mathbf{x}^{gen} = f_\phi(\mathbf{z})$$



$$\mathcal{F}(\mathbf{x}, \theta, \phi) = \mathbb{E}_{p^*(x)}[\log D_\theta(\mathbf{x})] + \mathbb{E}_{q_\phi(x)}[\log(1 - D_\theta(\mathbf{x})]$$

Alternating optimisation $\min_\phi \max_\theta \mathcal{F}(\mathbf{x}, \theta, \phi)$

Comparison loss $\theta \propto \nabla_\theta \mathbb{E}_{p^*(x)}[\log D_\theta(\mathbf{x})] + \nabla_\theta \mathbb{E}_{q_\phi(x)}[\log(1 - D_\theta(\mathbf{x})]$

(Alt) Generative loss $\phi \propto -\nabla_\phi \mathbb{E}_{q(z)}[\log D_\theta(f_\phi(\mathbf{z}))]$

Goodfellow et al. 2014

Integral Probability Metrics

$$\boxed{\mathcal{M}_f(p,q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{p(x)}[f] - \mathbb{E}_{q_\theta(x)}[f]\right|}$$

f sometimes referred to as a
test function, witness function or a critic.

Many choices of *f* available: classifiers or functions in specified spaces.

$$\|f\|_L < 1$$

Wasserstein

$$\|f\|_\infty < 1$$

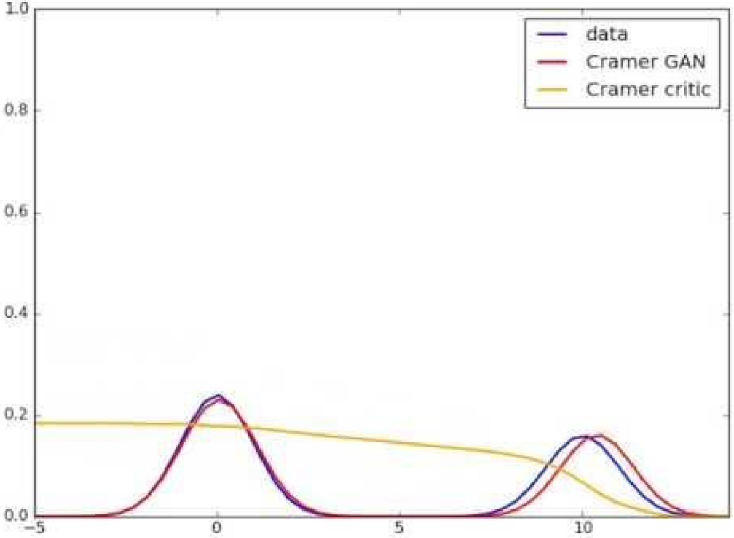
Total Variation

$$\|f\|_{\mathcal{H}} < 1$$

Max Mean Discrepancy

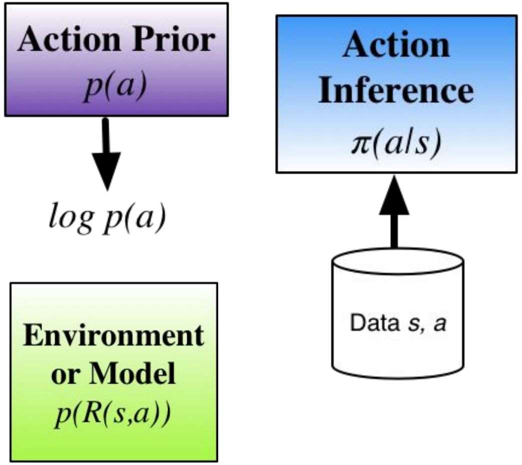
$$\left\|\frac{df}{dx}\right\|_L < 1$$

Cramer



Generative Models and RL

Probabilistic Policy Learning



$$u(s,a) \sim \text{Environment}(a) \qquad p(R(s)|a) \propto \exp(u(s,a))$$

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[R(s,a)] - \text{KL}[\pi_\theta(\mathbf{a}|s) \| p(\mathbf{a})]$$

Policy gradient update:

- Uniform prior on actions
- Score-function gradient estimator (aka Reinforce)

$$\nabla_\theta \mathcal{F}(\theta) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[(R(s,a) - c)\nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})] + \nabla_\theta \mathbb{H}[\pi_\theta(\mathbf{a}|s)]$$

Other algorithms:

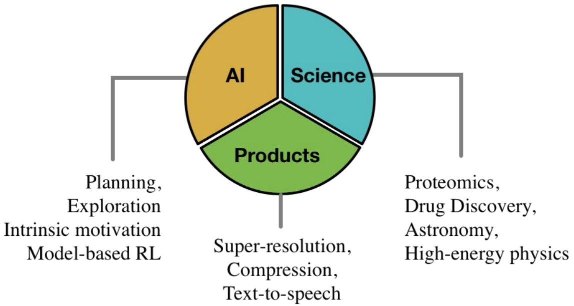
- Relative entropy policy search
- Generative adversarial imitation learning
- Reinforced variational inference

Other names and instantiations:

- Planning-as-inference
- Variational MDPs
- Path-integral control

The Future

Applications of Generative Models

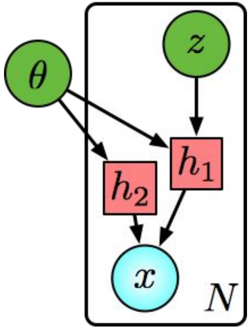


Planning,
Exploration
Intrinsic motivation
Model-based RL

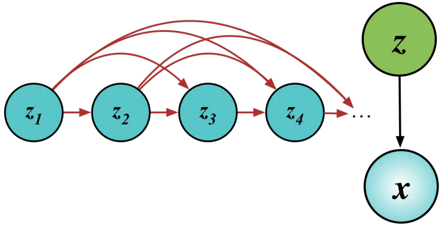
Super-resolution,
Compression,
Text-to-speech

Proteomics,
Drug Discovery,
Astronomy,
High-energy physics

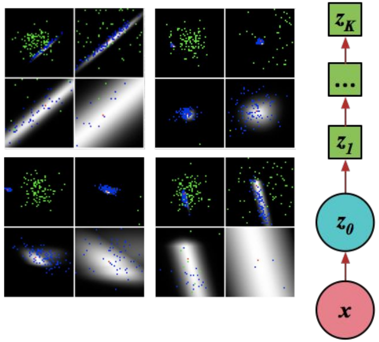
Probabilistic Deep Learning



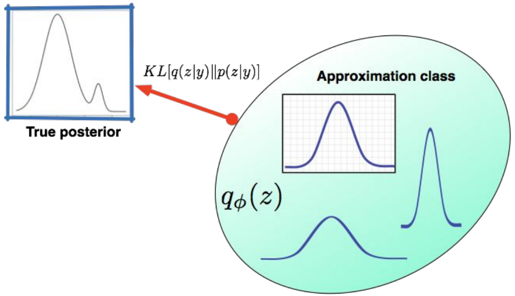
Types of Generative Models



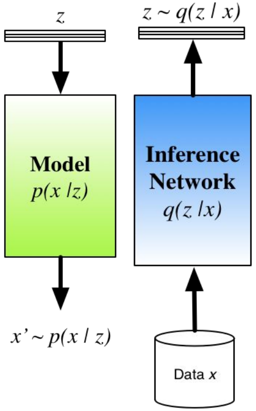
Rich Distributions



Variational Principles



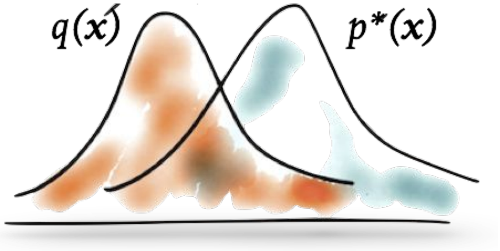
Amortised Inference



Stochastic Optimisation

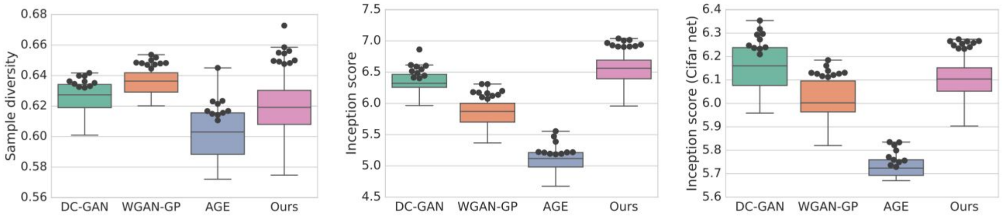
$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int \boxed{q_\phi(\mathbf{z})}\boxed{f_\theta(\mathbf{z})} d\mathbf{z}$$

Learning by Comparison



Challenges

- Scalability to large images, videos, multiple data modalities.
- Evaluation of generative models.
- Robust conditional models.
- Discrete latent variables.
- Support-coverage in models, mode-collapse.
- Calibration.
- Parameter uncertainty.
- Principles of likelihood-free inference.



(a) CelebA (b) Inception score (ImageNet) (c) Inception score (CIFAR)

UAI 2017 Australia

Tutorial on

Deep Generative Models

Shakir Mohamed and Danilo Rezende

@shakir_za @deepspiker

DeepMind

References: Applications

- Frey, Brendan J., and Geoffrey E. Hinton. "Variational learning in nonlinear Gaussian belief networks." Neural Computation 11, no. 1 (1999): 193-213. □
- Eslami, S. M., Heess, N., Weber, T., Tassa, Y., Kavukcuoglu, K., and Hinton, G. E. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. NIPS (2016). □
- Rezende, Danilo Jimenez, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. "One-Shot Generalization in Deep Generative Models." ICML (2016). □
- Kingma, Diederik P., Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. "Semi-supervised learning with deep generative models." In Advances in Neural Information Processing Systems, pp. 3581-3589. 2014.
- Higgins, Irina, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. "Early Visual Concept Learning with Unsupervised Deep Learning." arXiv preprint arXiv:1606.05579 (2016).
- Bellemare, Marc G., Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. "Unifying Count-Based Exploration and Intrinsic Motivation." arXiv preprint arXiv:1606.01868 (2016).
- Odena, Augustus. "Semi-Supervised Learning with Generative Adversarial Networks." arXiv preprint arXiv:1606.01583 (2016).
- Springenberg, Jost Tobias. "Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks." arXiv preprint arXiv:1511.06390 (2015).
- Alexander (Sasha) Vezhnevets, Mnih, Volodymyr, John Agapiou, Simon Osindero, Alex Graves, Oriol Vinyals, and Koray Kavukcuoglu. "Strategic Attentive Writer for Learning Macro-Actions." arXiv preprint arXiv:1606.04695 (2016).
- Gregor, Karol, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. "DRAW: A recurrent neural network for image generation." arXiv preprint arXiv:1502.04623 (2015).

References: Applications

- Gómez-Bombarelli R, Duvenaud D, Hernández-Lobato JM, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. arXiv preprint arXiv:1610.02415. 2016.
- Rampasek L, Goldenberg A. Dr. VAE: Drug Response Variational Autoencoder. arXiv preprint arXiv:1706.08203. 2017 Jun 26.
- Regier J, Miller A, McAuliffe J, Adams R, Hoffman M, Lang D, Schlegel D, Prabhat M. Celeste: Variational inference for a generative model of astronomical images. In International Conference on Machine Learning 2015 Jun 1 (pp. 2095-2103).
- Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W. Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802. 2016 Sep 15.
- Oord AV, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. 2016 Sep 12.
- Dumoulin V, Belghazi I, Poole B, Lamb A, Arjovsky M, Mastropietro O, Courville A. Adversarially learned inference. arXiv preprint arXiv:1606.00704. 2016 Jun 2.
- Gregor K, Besse F, Rezende DJ, Danihelka I, Wierstra D. Towards conceptual compression. In Advances In Neural Information Processing Systems 2016 (pp. 3549-3557).
- Higgins I, Matthey L, Glorot X, Pal A, Uria B, Blundell C, Mohamed S, Lerchner A. Early visual concept learning with unsupervised deep learning. arXiv preprint arXiv:1606.05579. 2016 Jun 17.
- Chiappa S, Racaniere S, Wierstra D, Mohamed S. Recurrent Environment Simulators. arXiv preprint arXiv:1704.02254. 2017 Apr 7.
- Kalchbrenner N, Oord AV, Simonyan K, Danihelka I, Vinyals O, Graves A, Kavukcuoglu K. Video pixel networks. arXiv preprint arXiv:1610.00527. 2016 Oct 3.
- Wu J, Tenenbaum JB, Kohli P. Neural Scene De-rendering., CVPR 2017

References: Fully-observed Models

- Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu. "Pixel recurrent neural networks." arXiv preprint arXiv:1601.06759 (2016).
- Larochelle, Hugo, and Iain Murray. "The Neural Autoregressive Distribution Estimator." In AISTATS, vol. 1, p. 2. 2011.
- Uria, Benigno, Iain Murray, and Hugo Larochelle. "A Deep and Tractable Density Estimator." In ICML, pp. 467-475. 2014.
- Veness, Joel, Kee Siong Ng, Marcus Hutter, and Michael Bowling. "Context tree switching." In 2012 Data Compression Conference, pp. 327-336. IEEE, 2012.
- Rue, Havard, and Leonhard Held. Gaussian Markov random fields: theory and applications. CRC Press, 2005.
- Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." Foundations and Trends® in Machine Learning 1, no. 1-2 (2008): 1-305.

References: Latent Variable Models

- Hyvärinen, A., Karhunen, J., & Oja, E. (2004). Independent component analysis (Vol. 46). John Wiley & Sons.
- Gregor, Karol, Ivo Danihelka, Andriy Mnih, Charles Blundell, and Daan Wierstra. "Deep autoregressive networks." arXiv preprint arXiv:1310.8499 (2013).
- Ghahramani, Zoubin, and Thomas L. Griffiths. "Infinite latent feature models and the Indian buffet process." In Advances in neural information processing systems, pp. 475-482. 2005.
- Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical dirichlet processes." Journal of the american statistical association (2012).
- Adams, Ryan Prescott, Hanna M. Wallach, and Zoubin Ghahramani. "Learning the Structure of Deep Sparse Graphical Models." In AISTATS, pp. 1-8. 2010.
- Lawrence, Neil D. "Gaussian process latent variable models for visualisation of high dimensional data." Advances in neural information processing systems 16.3 (2004): 329-336.
- Damianou, Andreas C., and Neil D. Lawrence. "Deep Gaussian Processes." In AISTATS, pp. 207-215. 2013.
- Mattos, César Lincoln C., Zhenwen Dai, Andreas Damianou, Jeremy Forth, Guilherme A. Barreto, and Neil D. Lawrence. "Recurrent Gaussian Processes." arXiv preprint arXiv:1511.06644 (2015).
- Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann machines for collaborative filtering." In Proceedings of the 24th international conference on Machine learning, pp. 791-798. ACM, 2007.
- Saul, Lawrence K., Tommi Jaakkola, and Michael I. Jordan. "Mean field theory for sigmoid belief networks." Journal of artificial intelligence research 4, no. 1 (1996): 61-76.
- Frey, Brendan J., and Geoffrey E. Hinton. "Variational learning in nonlinear Gaussian belief networks." Neural Computation 11, no. 1 (1999): 193-213.
- Durk Kingma and Max Welling. "Auto-encoding Variational Bayes." ICLR (2014).
- Burda Y, Grosse R, Salakhutdinov R. Importance weighted autoencoders. arXiv preprint arXiv:1509.00519. 2015 Sep 1.

References: Latent Variable Models (cont)

- Ranganath, Rajesh, Sean Gerrish, and David M. Blei. "Black Box Variational Inference." In AISTATS, pp. 814-822. 2014.
- Mnih, Andriy, and Karol Gregor. "Neural variational inference and learning in belief networks." arXiv preprint arXiv:1402.0030 (2014).
- Lázaro-Gredilla, Miguel. "Doubly stochastic variational Bayes for non-conjugate inference." (2014).
- Wingate, David, and Theophane Weber. "Automated variational inference in probabilistic programming." arXiv preprint arXiv:1301.1299 (2013).
- Paisley, John, David Blei, and Michael Jordan. "Variational Bayesian inference with stochastic search." arXiv preprint arXiv:1206.6430 (2012).
- Barber D, de van Laar P. Variational cumulant expansions for intractable distributions. Journal of Artificial Intelligence Research. 1999;10:435-55.

References: Stochastic Gradients

- Pierre L’Ecuyer, Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators, Management Science, 1995 □
- Peter W Glynn, Likelihood ratio gradient estimation for stochastic systems, Communications of the ACM, 1990 □
- Michael C Fu, Gradient estimation, Handbooks in operations research and management science, 2006 □
- Ronald J Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine learning, 1992 □
- Paul Glasserman, Monte Carlo methods in financial engineering, 2003 □
- Omiros Papaspiliopoulos, Gareth O Roberts, Martin Skold, A general framework for the parametrization of hierarchical models, Statistical Science, 2007 □
- Michael C Fu, Gradient estimation, Handbooks in operations research and management science, 2006 □
- Rajesh Ranganath, Sean Gerrish, and David M. Blei. "Black Box Variational Inference." In AISTATS, pp. 814-822. 2014. □
- Andriy Mnih, and Karol Gregor. "Neural variational inference and learning in belief networks." arXiv preprint arXiv:1402.0030 (2014).
- Michalis Titsias and Miguel Lázaro-Gredilla. "Doubly stochastic variational Bayes for non-conjugate inference." (2014). □
- David Wingate and Theophane Weber. "Automated variational inference in probabilistic programming." arXiv preprint arXiv:1301.1299 (2013). □
- John Paisley, David Blei, and Michael Jordan. "Variational Bayesian inference with stochastic search." arXiv preprint arXiv:1206.6430 (2012). □
- Durk Kingma and Max Welling. "Auto-encoding Variational Bayes." ICLR (2014). □
- Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models." ICML (2014).
- Papaspiliopoulos O, Roberts GO, Sköld M. A general framework for the parametrization of hierarchical models. Statistical Science. 2007 Feb 1:59-73.
- Fan K, Wang Z, Beck J, Kwok J, Heller KA. Fast second order stochastic backpropagation for variational inference. InAdvances in Neural Information Processing Systems 2015 (pp. 1387-1395).

References: Amortised Inference

- Dayan, Peter, Geoffrey E. Hinton, Radford M. Neal, and Richard S. Zemel. "The helmholtz machine." Neural computation 7, no. 5 (1995): 889-904. □
- Gershman, Samuel J., and Noah D. Goodman. "Amortized inference in probabilistic reasoning." In Proceedings of the 36th Annual Conference of the Cognitive Science Society. 2014. □
- Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models." ICML (2014).
- Durk Kingma and Max Welling. "Auto-encoding Variational Bayes." ICLR (2014). □ \
- Heess, Nicolas, Daniel Tarlow, and John Winn. "Learning to pass expectation propagation messages." In Advances in Neural Information Processing Systems, pp. 3219-3227. 2013. □
- Jitkrittum, Wittawat, Arthur Gretton, Nicolas Heess, S. M. Eslami, Balaji Lakshminarayanan, Dino Sejdinovic, and ZoltÃ ¿an SzabÃ¸s. "Kernel-based just-in-time learning for passing expectation propagation messages." arXiv preprint arXiv:1503.02551 (2015). □
- Korattikara, Anoop, Vivek Rathod, Kevin Murphy, and Max Welling. "Bayesian dark knowledge." arXiv preprint arXiv:1506.04416 (2015).

References: Structured Mean Field

- Jaakkola, T. S., and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In Learning in graphical models (pp. 163-173). Springer Netherlands. □
- Saul, L.K. and Jordan, M.I., 1996. Exploiting tractable substructures in intractable networks. Advances in neural information processing systems, pp.486-492. □
- Gregor, Karol, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. "DRAW: A recurrent neural network for image generation." ICML (2015). □
- Gershman, S., Hoffman, M. and Blei, D., 2012. Nonparametric variational inference. arXiv preprint arXiv:1206.4665.
- Felix V. Agakov, and David Barber. "An auxiliary variational method." NIPS (2004). □ Rajesh Ranganath, Dustin Tran, and David M. Blei. "Hierarchical Variational Models." ICML (2016). □ Lars Maaløe et al. "Auxiliary Deep Generative Models." ICML (2016). □ Tim Salimans, Durk Kingma, Max Welling. "Markov chain Monte Carlo and variational inference: Bridging the gap. In International Conference on Machine Learning." ICML (2015).
- Maaløe L, Sønderby CK, Sønderby SK, Winther O. Auxiliary deep generative models. arXiv preprint arXiv:1602.05473. 2016 Feb 17.

References: Normalising Flows

- Tabak, E. G., and Cristina V. Turner. "A family of nonparametric density estimation algorithms." Communications on Pure and Applied Mathematics 66, no. 2 (2013): 145-164. □
- Rezende, Danilo Jimenez, and Shakir Mohamed. "Variational inference with normalizing flows." ICML (2015). □
- Kingma, D.P., Salimans, T. and Welling, M., 2016. Improving variational inference with inverse autoregressive flow. arXiv preprint arXiv:1606.04934. □
- Dinh, L., Sohl-Dickstein, J. and Bengio, S., 2016. Density estimation using Real NVP. arXiv preprint arXiv:1605.08803.

References: Other Variational Objectives

- Yuri Burda, Roger Grosse, Ruslan Salakhutidinov. "Importance weighted autoencoders." ICLR (2015). □
- Yingzhen Li, Richard E. Turner. "Rényi divergence variational inference." NIPS (2016). □
- Guillaume and Balaji Lakshminarayanan. "Approximate Inference with the Variational Holder Bound." ArXiv (2015). □
- José Miguel Hernández-Lobato, Yingzhen Li, Daniel Hernández-Lobato, Thang Bui, and Richard E. Turner. Black-box α-divergence Minimization. ICML (2016). □
- Rajesh Ranganath, Jaan Altosaar, Dustin Tran, David M. Blei. Operator Variational Inference. NIPS (2016).

References: Discrete Latent Variable Models

- □Radford Neal. "Learning stochastic feedforward networks." Tech. Rep. CRG-TR-90-7: Department of Computer Science, University of Toronto (1990). □
- Lawrence K. Saul, Tommi Jaakkola, and Michael I. Jordan. "Mean field theory for sigmoid belief networks." Journal of artificial intelligence research 4, no. 1 (1996): 61-76. □
- Karol Gregor, Ivo Danihelka, Andriy Mnih, Charles Blundell, and Daan Wierstra. "Deep autoregressive networks." ICML (2014). □
- Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M. Blei. "Deep Exponential Families." AISTATS (2015). □
- Rajesh Ranganath, Dustin Tran, and David M. Blei. "Hierarchical Variational Models." ICML (2016).

References: Implicit Generative Models

- Borgwardt, Karsten M., and Zoubin Ghahramani. "Bayesian two-sample tests." arXiv preprint arXiv:0906.4032 (2009).
- Gutmann, Michael, and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." AISTATS. Vol. 1. No. 2. 2010.
- Tsuboi, Yuta, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. "Direct Density Ratio Estimation for Large-scale Covariate Shift Adaptation." Information and Media Technologies 4, no. 2 (2009): 529-546.
- Sugiyama, Masashi, Taiji Suzuki, and Takafumi Kanamori. Density ratio estimation in machine learning. Cambridge University Press, 2012.
- Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in Neural Information Processing Systems, pp. 2672-2680. 2014.
- Verrelst, Herman, Johan Suykens, Joos Vandewalle, and Bart De Moor. "Bayesian learning and the Fokker-Planck machine." In Proceedings of the International Workshop on Advanced Black-box Techniques for Nonlinear Modeling, Leuven, Belgium, pp. 55-61. 1998.
- Devroye, Luc. "Random variate generation in one line of code." In Proceedings of the 28th conference on Winter simulation, pp. 265-272. IEEE Computer Society, 1996.
- Mohamed S, Lakshminarayanan B. Learning in implicit generative models. arXiv preprint arXiv:1610.03483. 2016 Oct 11.
- Gutmann MU, Dutta R, Kaski S, Corander J. Likelihood-free inference via classification. Statistics and Computing. 2017 Mar 13:1-5.
- Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002 Dec 1;162(4):2025-35.
- Arjovsky M, Chintala S, Bottou L. Wasserstein gan. ICML 2017.
- Nowozin S, Cseke B, Tomioka R. f-gan: Training generative neural samplers using variational divergence minimization. In Advances in Neural Information Processing Systems 2016 (pp. 271-279).
- Bellemare MG, Danihelka I, Dabney W, Mohamed S, Lakshminarayanan B, Hoyer S, Munos R. The Cramer Distance as a Solution to Biased Wasserstein Gradients. arXiv preprint arXiv:1705.10743. 2017 May 30.
- Dumoulin V, Belghazi I, Poole B, Lamb A, Arjovsky M, Mastropietro O, Courville A. Adversarially learned inference.
- Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics; 2001.

References: Prob. Reinforcement Learning

- Kappen HJ. Path integrals and symmetry breaking for optimal control theory. Journal of statistical mechanics: theory and experiment. 2005 Nov 30;2005(11):P11011.
- Rawlik K, Toussaint M, Vijayakumar S. Approximate inference and stochastic optimal control. arXiv preprint arXiv:1009.3958. 2010 Sep 20.
- Toussaint M. Robot trajectory optimization using approximate inference. InProceedings of the 26th annual international conference on machine learning 2009 Jun 14 (pp. 1049-1056). ACM.
- Weber T, Heess N, Eslami A, Schulman J, Wingate D, Silver D. Reinforced variational inference. In Advances in Neural Information Processing Systems (NIPS) Workshops 2015.
- Rajeswaran A, Lowrey K, Todorov E, Kakade S. Towards Generalization and Simplicity in Continuous Control.
- Peters J, Mülling K, Altun Y. Relative Entropy Policy Search. In AAAI 2010 Jul 11 (pp. 1607-1612).
- Furmston T, Barber D. Variational methods for reinforcement learning. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics 2010 Mar 31 (pp. 241-248).
- Ho J, Ermon S. Generative adversarial imitation learning. In Advances in Neural Information Processing Systems 2016 (pp. 4565-4573