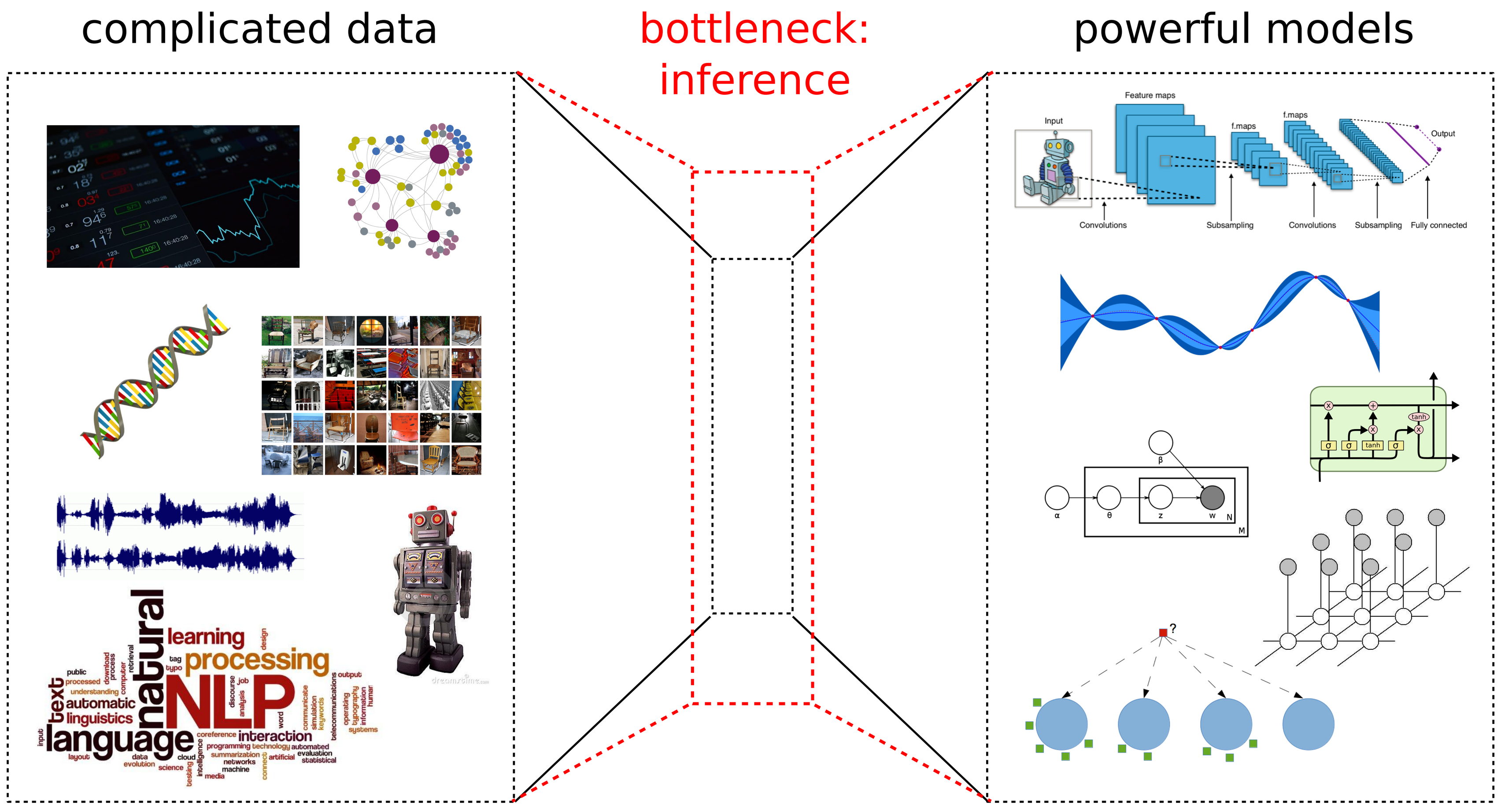
WILD VARIATIONAL APPROXIMATIONS

YINGZHEN LI[1] AND QIANG LIU[2]
[1]UNIVERSITY OF CAMBRIDGE [2]DARTMOUTH COLLEGE

BOTTLENECK OF MODELLING



We would like to overcome the bottleneck but still get fast inference!

DEFINITION OF WVA

Given the exact posterior $p(\boldsymbol{z}|\boldsymbol{x})$, we want to construct a *wild variational approximation* $q(\boldsymbol{z}|\boldsymbol{x})$ such that:

- It is fitted to the $p(\boldsymbol{z}|\boldsymbol{x})$ using an optimisation-based method;
- Inference with $q(\boldsymbol{z}|\boldsymbol{x})$ is comparatively easier:
  - for the function $F(\boldsymbol{z})$ in interest, it is easier to compute (or estimate with MC methods) $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[F(\boldsymbol{z})]$ than $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[F(\boldsymbol{z})]$;
- Its density is intractable, or difficult to compute in a fast way.

HOW DO WE FIT A WVA?

(Should use different approximation method for different q!)

Idea 1: Energy Approximation (e.g. for VFE)

- Approximate $\log q$ or $\mathbb{H}[q]$ (e.g. using density estimation);
- Approximate $\mathrm{KL}[q||p_0]$ with density ratio estimation;
- Might require solving a minimax optimisation problem!

Idea 2: Direct Gradient Approximation

- Directly fit a model to the gradient by optimisation;
- Example: using Kernel Ridge regression (Sasaki et al. 2015)
- Use Stein's Identity?

Idea 3: Other Objective Functions

- Stein Discrepancy: with $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[(\mathcal{T}\boldsymbol{g})(\boldsymbol{z})] = 0$ for $\boldsymbol{g} \in \mathcal{G}$,

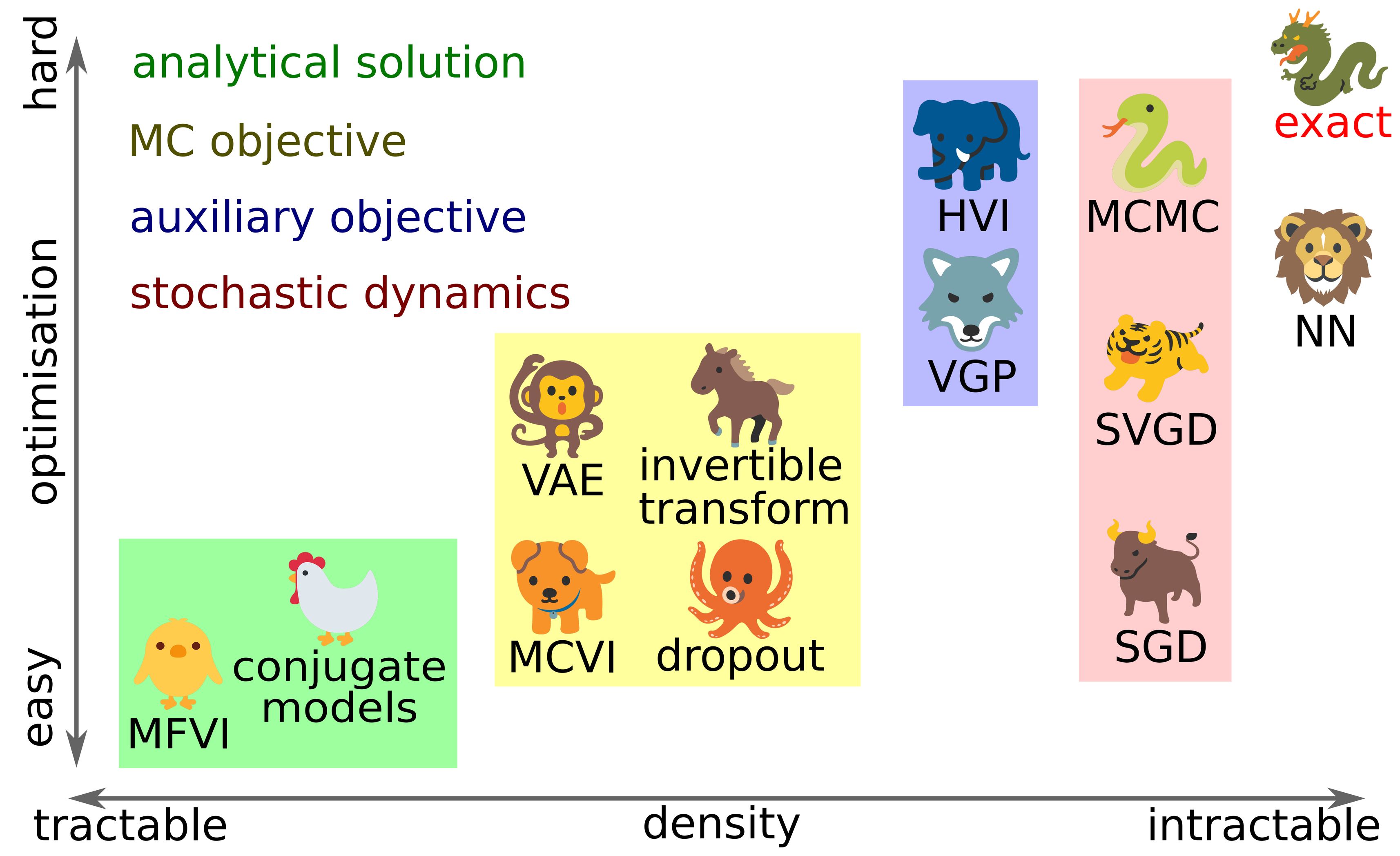
$$\min_q \mathcal{S}[q||p] = \min_q \sup_{\boldsymbol{g}\in\mathcal{G}} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[(\mathcal{T}\boldsymbol{g})(\boldsymbol{z})].$$

- Example: $(\mathcal{T}g)(\boldsymbol{z}) = \nabla_{\boldsymbol{z}} \log p(\boldsymbol{x}, \boldsymbol{z})^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{z}) + \nabla_{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{z})$;
- OPVI (Ranganath et al. 2016) uses parametric test functions $\mathcal{G} = \{\boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{z})\}$; (inefficient: it uses Hessian info for update)
- Avoid minimax problems: use kernels;
- Other objective function choices, e.g. MMD?

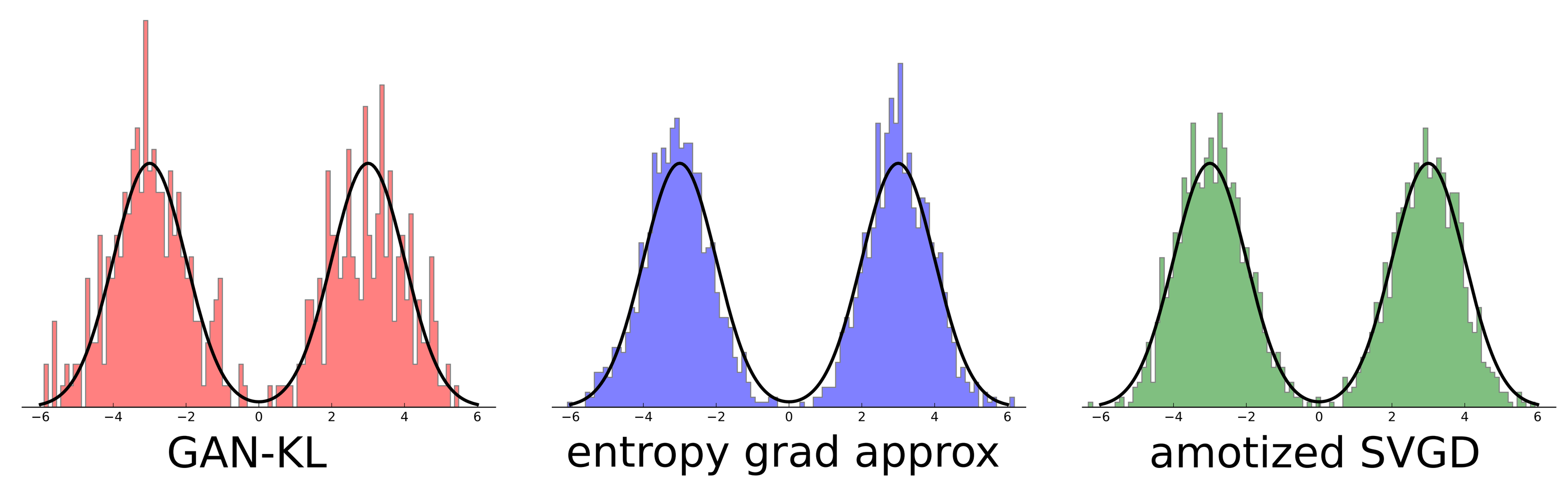
Idea 4: Amortize Stochastic Dynamics

- Sample $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x})$;
- Compute $\boldsymbol{z}'$ by running T-step stochastic dynamics;
- Update $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + (\boldsymbol{z}' - \boldsymbol{z})^{\mathrm{T}} \nabla_{\boldsymbol{\phi}} \boldsymbol{f}$; (one step gradient descent with L_2 measure $||z' - z||_2^2$)
- Already applied to energy-based models (Wang and Liu 2016);
- Can use other measures to chain the gradients.

A ZOO OF INFERENCE ENGINES



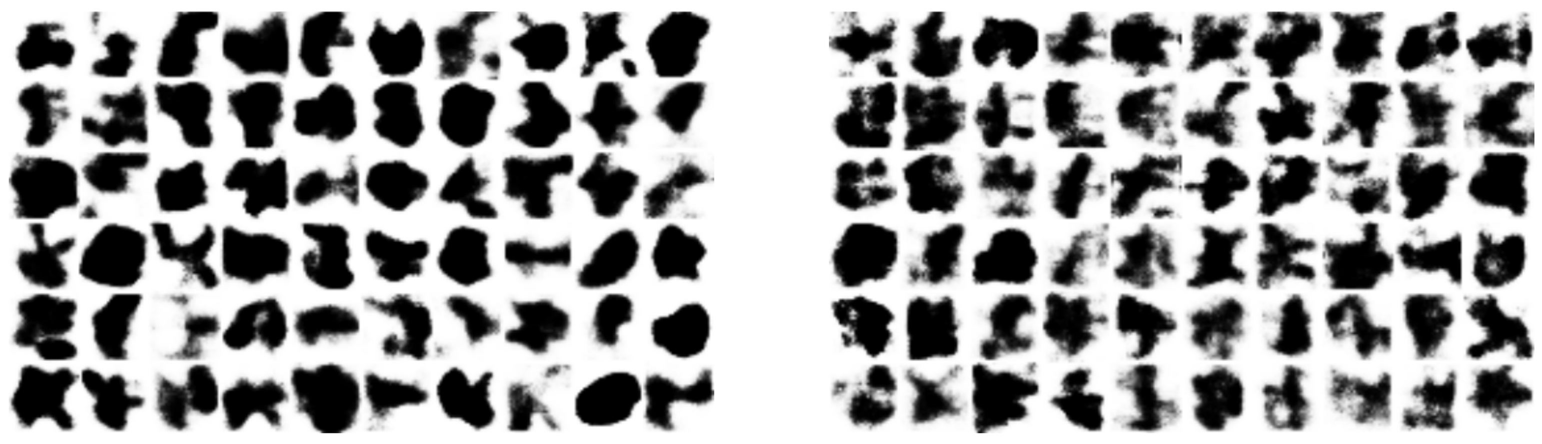
EXAMPLE: MIXTURE OF GAUSSIANS



- True density: $p(z) = 0.5\mathcal{N}(z; -3, 1) + 0.5\mathcal{N}(z; 3, 1)$;
- Approximation q is specified by the following procedure: $\epsilon_i \sim \mathcal{N}(0, 1), z = (\epsilon_3 \geq 0)R(\epsilon_1; \phi_1) - (\epsilon_3 \leq 0)R(\epsilon_2; \phi_2)$, with $R(\epsilon; \phi)$ defined by a one-hidden layer NN;
- GAN-KL: we discriminate between samples from $q(z)$ and $\tilde{p}(z) = \mathcal{N}(z; 0, 2)$ (using f-GAN (Nowozin et al. 2016) objective);
- Entropy Grad Approx: $\nabla_{\boldsymbol{\phi}} \mathbb{H}[q] \approx \nabla_{\boldsymbol{\phi}} \boldsymbol{z}(\boldsymbol{K} + \eta \boldsymbol{I})^{-1} \nabla_z \boldsymbol{K}$

EXAMPLE: GENERATIVE MODELLING

- Goal: compare with the benchmark (Gaussian VAE): in this case the density of q is tractable, but we will test methods which do not require $\log q$.
- Model: 2-hidden layer MLP (500 units), latent dimension 20;
- WVA: NN with input size $D_{in} + 100$ (same hidden layers);
- All used $K = 50$ samples during training;



Gaussian+VAE NN+SVGD

Dataset	Gaussian+VAE	Gaussian+SVGD	NN+SVGD
Caltech 101	-123.50	-134.03	-129.27
MNIST	-89.95	-106.40	-92.14

FUTURE WORK

- Visualise more toy examples;
- Test the GAN-KL type method on (deep) generative models;
- Develop amortized MCMC methods; (currently testing SGLD + rejection step)
- How to improve sample efficiency? (You can train a Gaussian VAE with only one sample!)