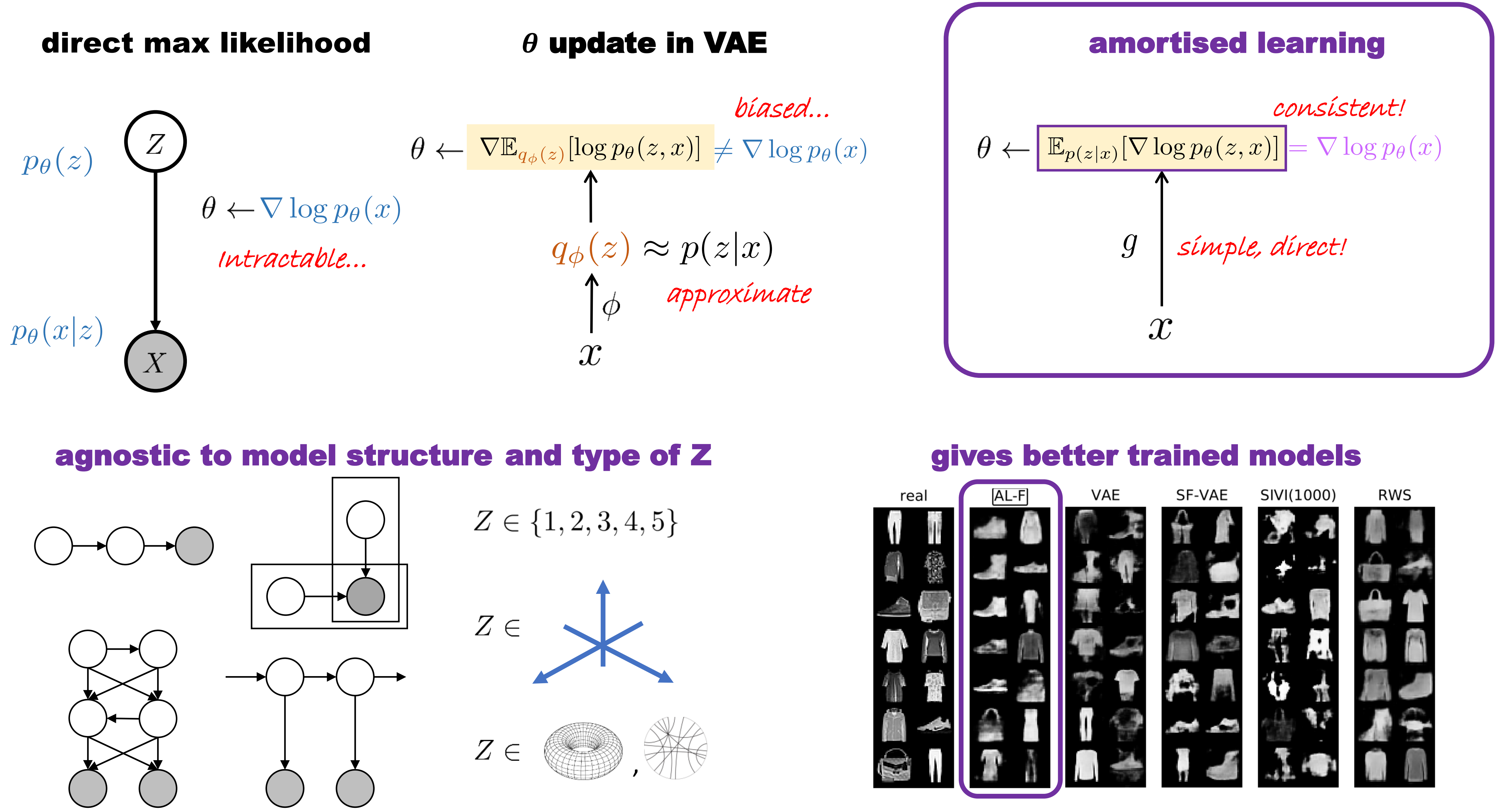
Amortised learning by wake-sleep

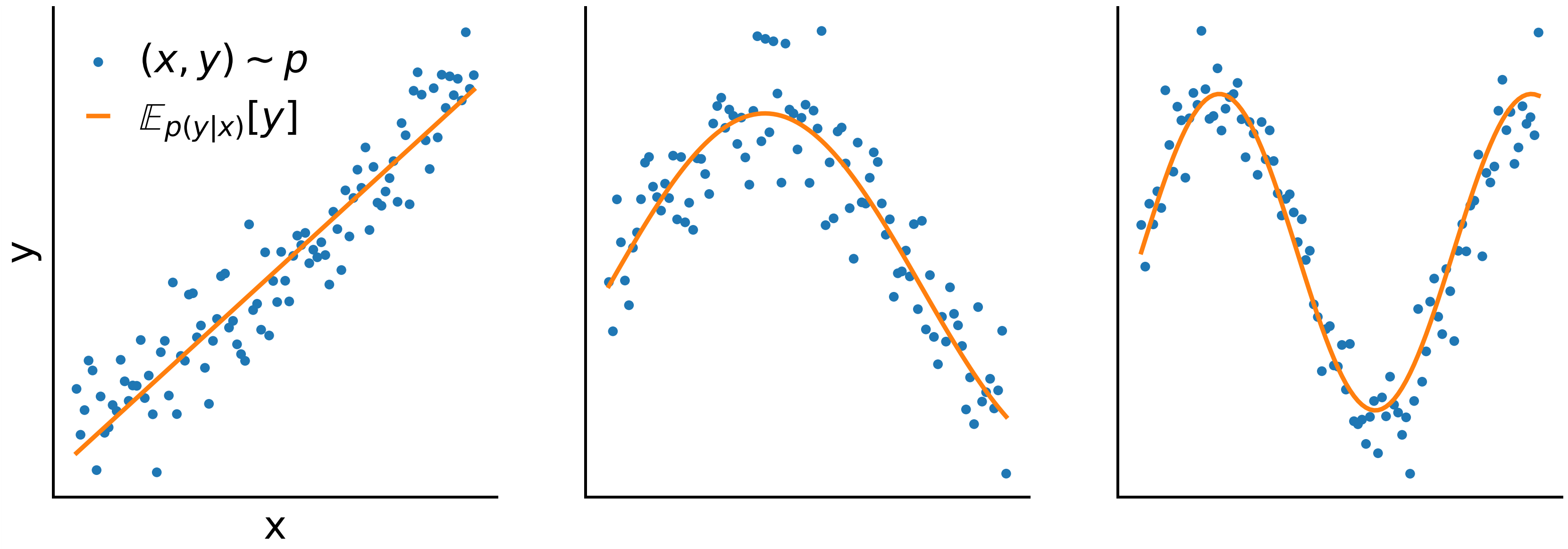
Li Kevin Wenliang, Ted Moskovitz, Heishiro Kanagawa, Maneesh Sahani

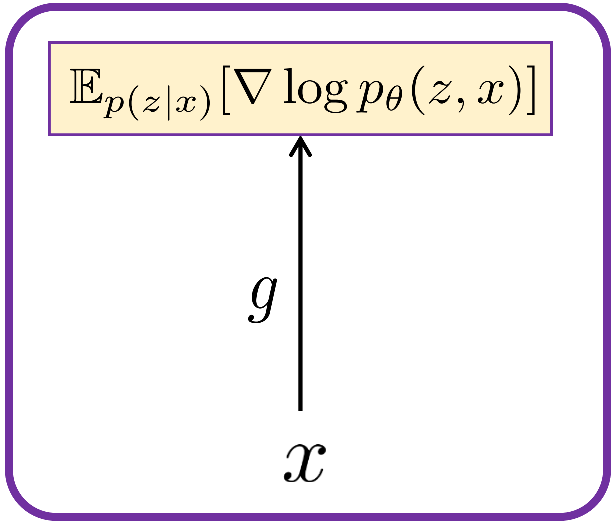
Gatsby Unit, University College London


direct max likelihood
$p_\theta(z)$
Z
$p_\theta(x|z)$
X
$\theta \leftarrow \nabla \log p_\theta(x)$
intractable...
θ update in VAE
biased...
$\theta \leftarrow \nabla \mathbb{E}_{q_\phi(z)}[\log p_\theta(z,x)] \neq \nabla \log p_\theta(x)$
$q_\phi(z) \approx p(z|x)$
ϕ
approximate
x
amortised learning
consistent!
$\theta \leftarrow \boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]} = \nabla \log p_\theta(x)$
g
simple, direct!
x
agnostic to model structure and type of Z
$Z \in \{1,2,3,4,5\}$
$Z \in$
$Z \in$,
gives better trained models
real
AL-F
VAE
SF-VAE
SIVI(1000)
RWS

Least square regression gives conditional expectation

$$\mathbb{E}_{p(y|x)}[y] = \arg\min_g \mathbb{E}_{p(x,y)}[\|y - g(x)\|_2^2]$$





How to estimate $\boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]}$?

- define

$$\ell_\theta(z,x) = \log p_\theta(z,x) \qquad \nabla\ell(z,x) = \nabla_\theta \log p_\theta(z,x)$$

- then

$$\mathbb{E}_{p(z|x)}[\nabla\ell(z,x)] = \arg\min_g \mathbb{E}_{p(z,x)}[\|\nabla\ell(z,x) - g(x)\|_2^2]$$

- In practice, draws $z_n, x_n \sim p_\theta$ and solve

$$\hat{g} = \arg\min_g \sum_{n=1}^{N} \|\nabla\ell(z_n,x_n) - g(x_n)\|_2^2$$

$$\begin{array}{c} \boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]} \\ \Big\uparrow g \\ x \end{array}$$

Algorithm:

1. *$z_n, x_n \sim p_\theta$* } sleep
2. *find $\hat{g}$ by regression*
3. *$x_m \sim \mathcal{D}$* } wake
4. *update θ by $\hat{g}(x_m)$*

Issues:

- $\nabla\ell(z,x) = \nabla_\theta \log p_\theta(z,x)$ is high dimensional
- computing $\nabla\ell(z_n,x_n)$ for all sleep samples can be slow

How to estimate $\boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]}$ **more efficiently**?

- define

$$\ell_\theta(z,x) = \log p_\theta(z,x) \qquad \nabla \ell(z,x) = \nabla_\theta \log p_\theta(z,x)$$

$$\begin{array}{c} \boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]} \\ \uparrow g \\ x \end{array}$$

- suppose we estimate $\mathbb{E}_{p(z|x)}[\ell_\theta(z,x)]$ with kernel ridge regression, then

$$\hat{f}_\theta(x) = \begin{bmatrix} \ell_\theta(z_1,x_1) & \cdots & \ell_\theta(z_N,x_N) \end{bmatrix} \begin{bmatrix} \alpha_1(x) \\ \vdots \\ \alpha_N(x) \end{bmatrix}$$

auto-diff

$$\nabla_\theta \hat{f}_\theta(x) = \begin{bmatrix} \nabla_\theta \ell_\theta(z_1,x_1) & \cdots & \nabla_\theta \ell_\theta(z_N,x_N) \end{bmatrix} \begin{bmatrix} \alpha_1(x) \\ \vdots \\ \alpha_N(x) \end{bmatrix} = \hat{g}(x)$$

is an estimator of $\mathbb{E}_{p(z|x)}[\nabla \ell(z,x)]$ by kernel ridge regression

Theorem: if $\boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]} \in \mathcal{L}_p^2$ and the kernel is rich, then $\nabla_\theta \hat{f}_\theta(x)$ is a consistent estimator of $\boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]}$

Amortised learning by wake-sleep

1. $z_n, x_n \sim p_\theta$
2. *kernel ridge regression*

$$\hat{f}_\theta = \underset{f \in \mathcal{H}}{\arg\min} \sum_{n=1}^{N} \| \log p_\theta(z_n, x_n) - f(x_n) \|_2^2$$

3. $x_m \sim \mathcal{D}$
4. *update* θ *by* $g(x_m) = \nabla_\theta \hat{f}_\theta(x_m)$

consistent!

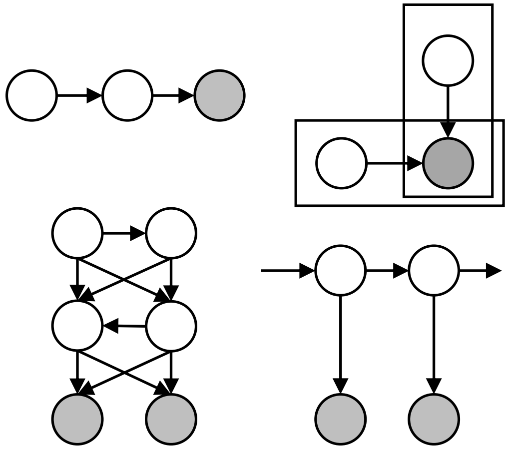
$$\boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z, x)]} = \nabla \log p_\theta(x)$$

$$x \xrightarrow{g} \mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z, x)]$$

simple, direct!

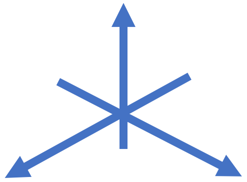
Assumptions:
- easy to sample from p_θ
- $\nabla_\theta \log p_\theta(x, z)$ exists
- true gradient is $\mathcal{L}_p^2$

Non-assumptions:
- posterior
- structure of p_θ
- type of Z

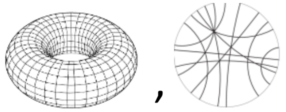


$Z \in \{1, 2, 3, 4, 5\}$

$Z \in$



$Z \in$



Experiments

- **Log likelihood gradient estimation**
- **Non-Euclidean latent**
- **Dynamical models**
- **Image generation**
- **Non-negative matrix factorisation**
- Hierarchical models
- Independent component analysis
- Neural processes

consistent!

$$\begin{tikzcd} \boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z,x)]} = \nabla \log p_\theta(x) \\ x \arrow[u, "g"] \end{tikzcd}$$

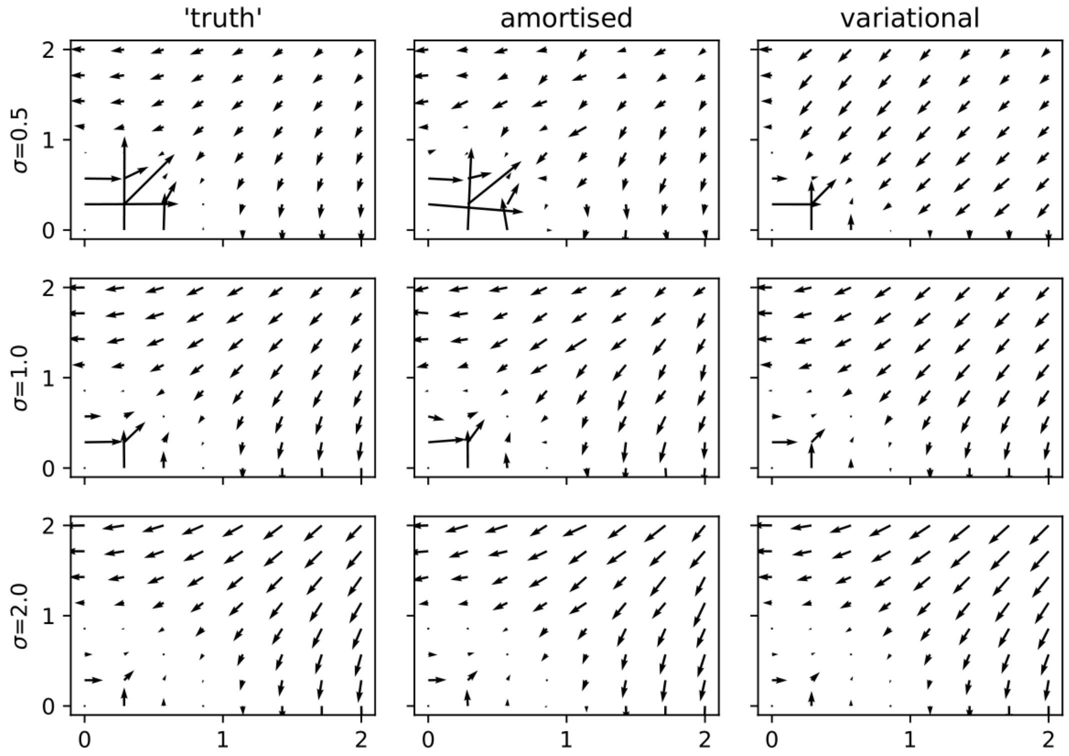
simple, direct!

Experiment 1: gradient estimation

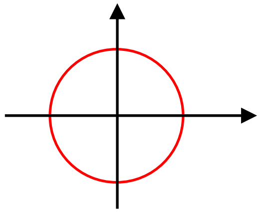
Generative model

$$z_1, z_2 \sim \mathcal{N}(0,1),\ x|\boldsymbol{z} \sim \mathcal{N}(\mathrm{softplus}(\boldsymbol{b}\cdot\boldsymbol{z}) - \|\boldsymbol{b}\|_2^2, \sigma_x^2)$$

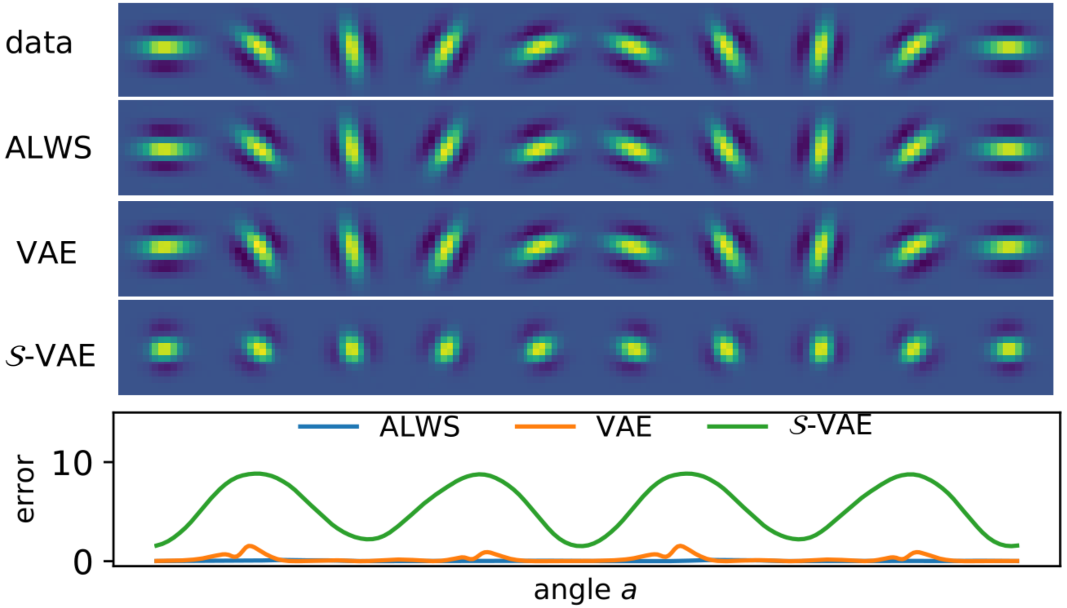
Task: estimate $\nabla_{\boldsymbol{b}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ for different $\boldsymbol{b}$ and σ



Experiment II: prior on the unit circle

Model: $z \in$  $$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathrm{NN}_{\boldsymbol{w}}(\boldsymbol{z}), \sigma_x^2\mathbf{I})$$

Task: generate Gabor filters of uniformly distributed orientations (no special reparameterisation)

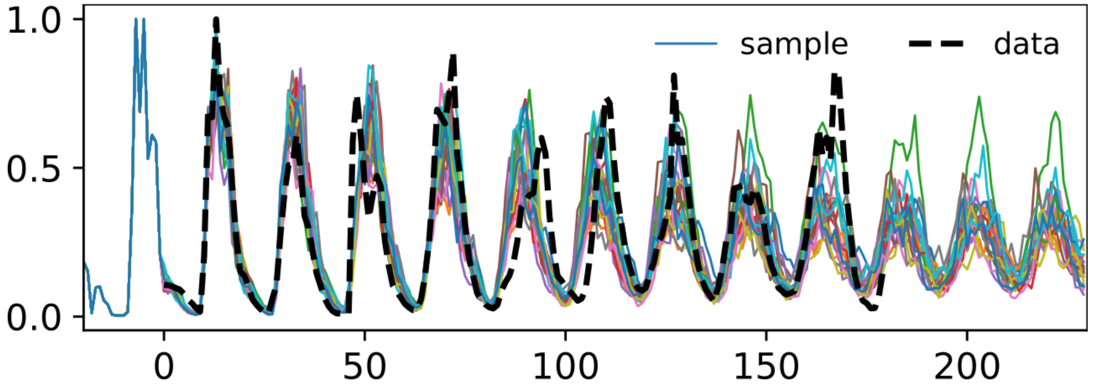


Experiment III: dynamical model

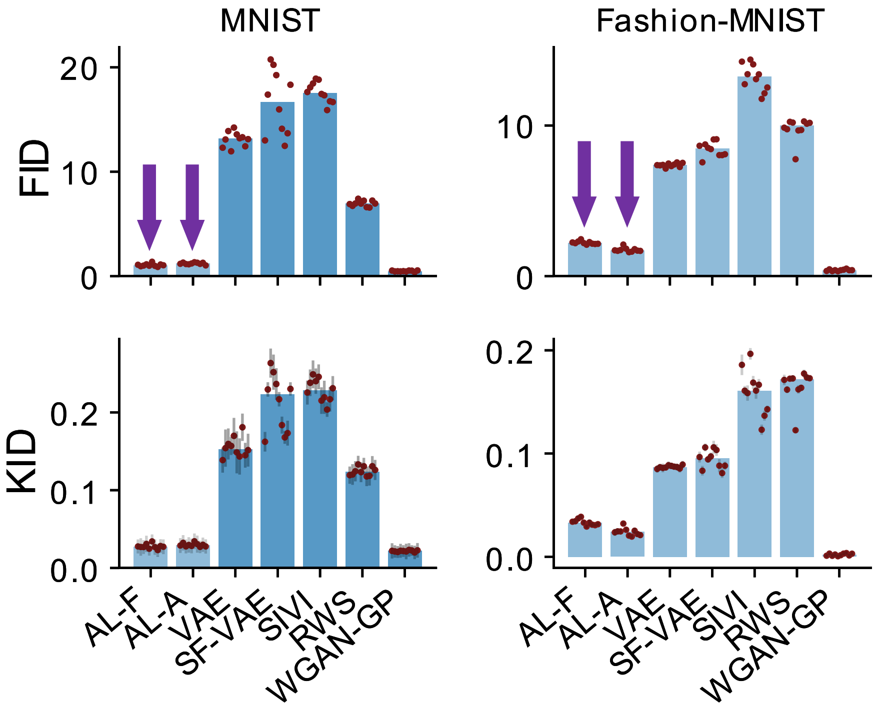
Model

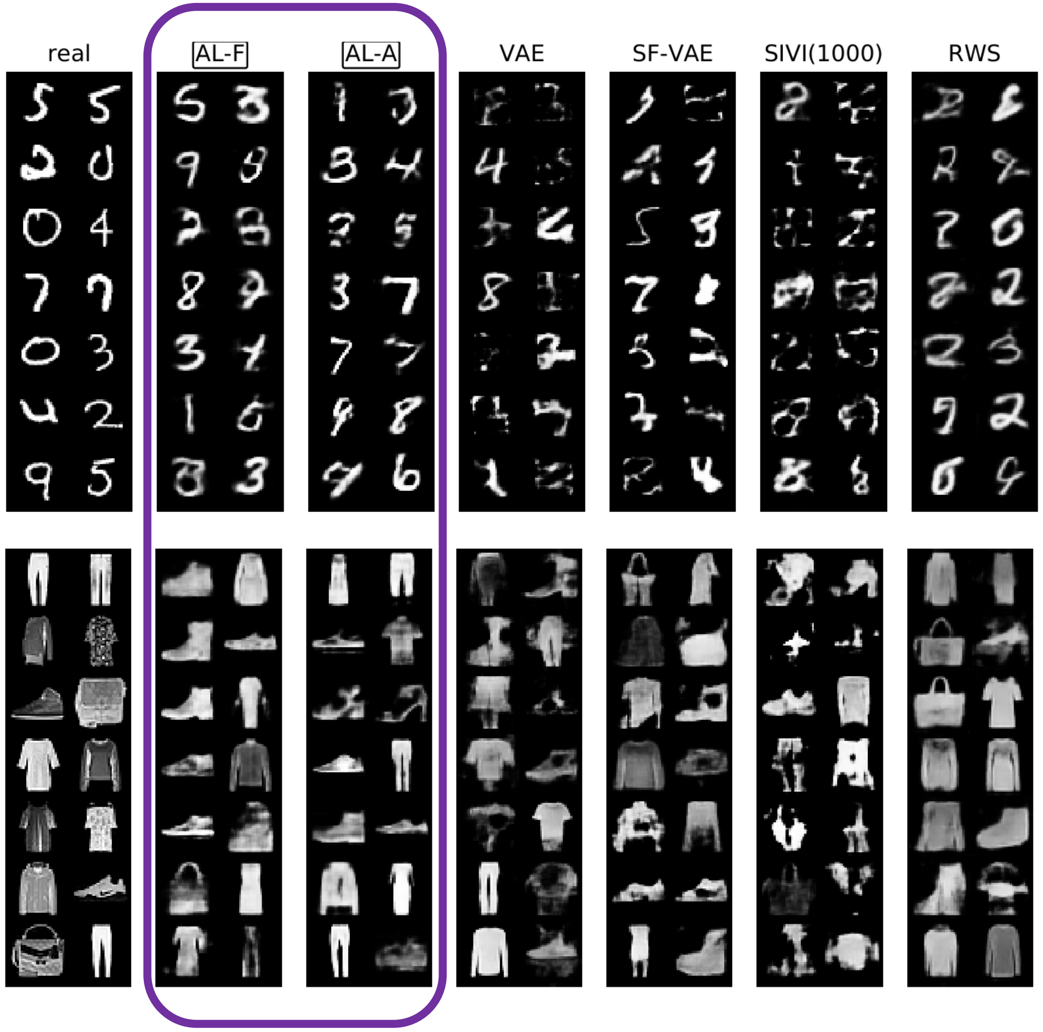
$$\tau \sim \mathrm{Categorical}(\boldsymbol{m}), \tau \in \{1, \ldots, 20\}, \quad e_t \sim \mathrm{Gamma}(\frac{1}{\sigma_p^2}, \sigma_p^2), \quad \epsilon_t \sim \mathrm{Gamma}(\frac{1}{\sigma_d^2}, \sigma_d^2),$$

$$z_t = P x_{t-\tau} \exp(-\frac{x_{t-\tau}}{N_0}) + x_t \exp(-\delta \epsilon_t), \quad p(x_t | z_t) = \mathrm{LogNormal}(\log(z_t), \sigma_n^2)$$



Experiment IV:sample quality



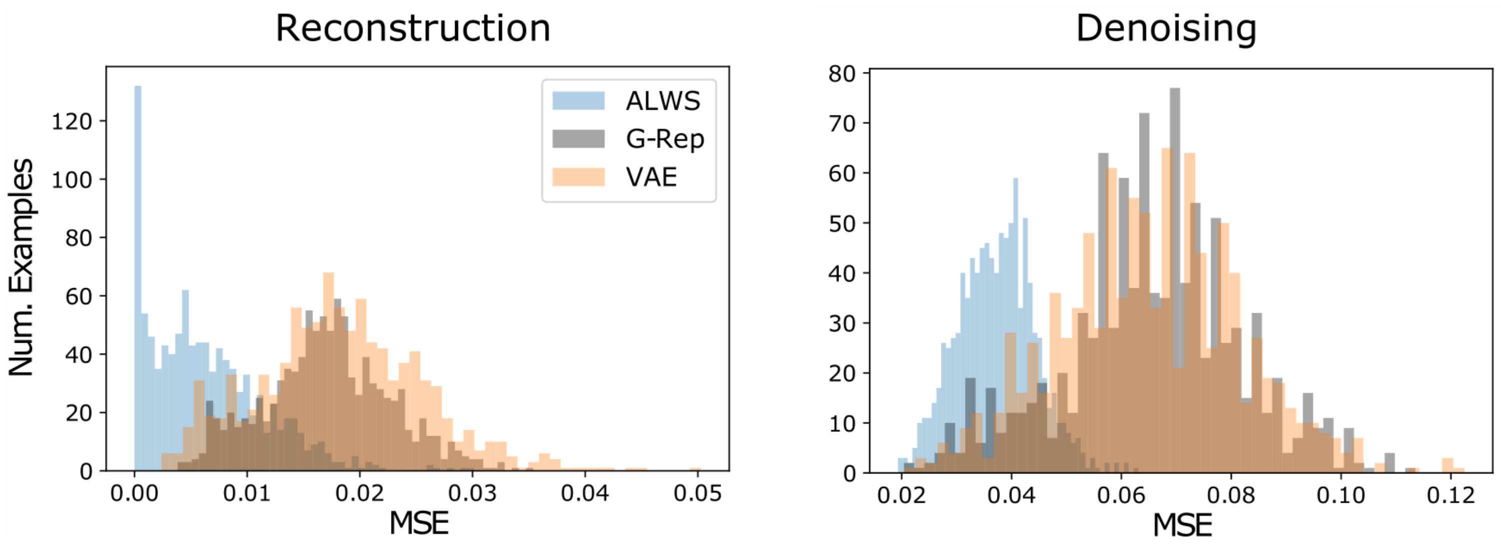


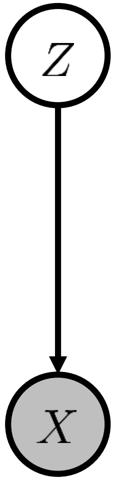
Experiment IV: downstream tasks

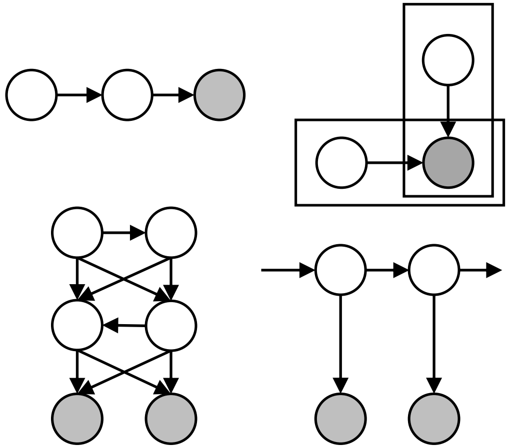
Model:

$$p(z_i) = \mathcal{U}(z_i; 0, 1), \quad p(x_i|\boldsymbol{z}) = \text{Bernoulli}\left(x_i; \bar{x}_i\right), \quad \bar{x}_i = \text{sigmoid}\left(\boldsymbol{w}_i \cdot \text{logit}(\boldsymbol{z}) + b_i\right)$$

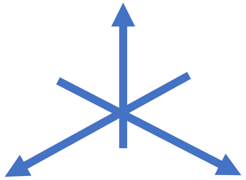
Task: reconstruct or denoise images after training

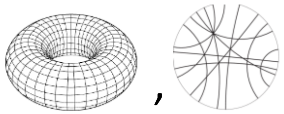






$Z \in \{1, 2, 3, 4, 5\}$

$Z \in$ 

$Z \in$ 

amortised learning

consistent!

$$\theta \leftarrow \boxed{\mathbb{E}_{p(z|x)}[\nabla \log p_\theta(z, x)]} = \nabla \log p_\theta(x)$$

$$\Big\uparrow g$$

simple, direct!

$$x$$

Thank you!