Missing Data Imputation and Acquisition with Deep Hierarchical Models and Hamiltonian Monte Carlo

Ignacio Peis
Universidad Carlos III de Madrid
ipeis@tsc.uc3m.es

Chao Ma
Microsoft Research Cambridge
University of Cambridge
cm905@cam.ac.uk

José Miguel Hernández-Lobato
University of Cambridge
jmh233@cam.ac.uk

NEURAL INFORMATION PROCESSING SYSTEMS 2022

Challenges

Enhance information acquisition with VAEs

- Discovery of high-value information.
- Bayesian reward function [2] as an expected gain of information:

$$R(i, \mathbf{x}_O) = \mathbb{E}_{\mathbf{x}_i \sim p(\mathbf{x}_i|\mathbf{x}_O)} D_{\mathrm{KL}} \left[p(\mathbf{x}_\phi | \mathbf{x}_i, \mathbf{x}_O) \,\|\, p(\mathbf{x}_\phi | \mathbf{x}_O) \right]$$

- Approximated in [3,4] by transforming into $\mathbf{z}$ space:

$$\begin{aligned}\hat{R}_I(\mathbf{x}_i, \mathbf{x}_O) = \mathbb{E}_{p_\theta(\mathbf{x}_i, \mathbf{z}_i, \mathbf{z}_O | \mathbf{x}_O)} \{ &\mathbb{KL} \left[q_\lambda(\mathbf{h}|\mathbf{z}_i, \mathbf{z}_O) || q_\lambda(\mathbf{h}|\mathbf{z}_O) \right] - \\ &\mathbb{E}_{p_\theta(\mathbf{x}_\phi, \mathbf{z}_\Phi, | \mathbf{x}_O)} \mathbb{KL} \left[q_\lambda(\mathbf{h}|\mathbf{z}_\Phi, \mathbf{z}_i, \mathbf{z}_O) || q_\lambda(\mathbf{h}|\mathbf{z}_\Phi, \mathbf{z}_O) \right] \}\end{aligned}$$

- These methods are based on **Gaussian** approximations of the true posterior.

[2] Bernardo et at., 1979 [3] Ma et at., 2018 [4] Ma et at., 2020

Challenges

Improve missing data imputation with VAEs

- Imputation under a VAE framework [3,4,5,6]:

$$p(\boldsymbol{x}_U|\boldsymbol{x}_O) = \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}_O)}[p(\boldsymbol{x}_U|\boldsymbol{z})] \approx \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_O)}[p(\boldsymbol{x}_U|\boldsymbol{z})]$$

- Also based on **Gaussian** approximations of the true posterior.

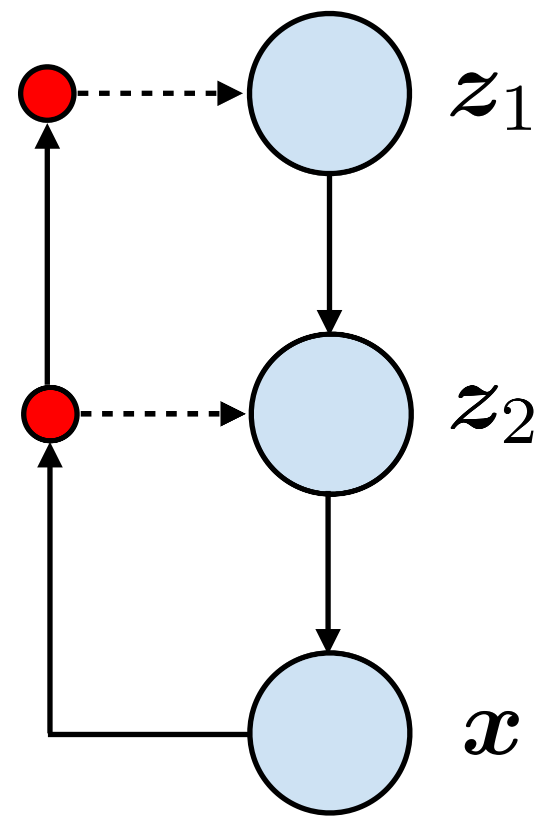
[3] Ma et at., 2018 [4] Ma et at., 2020 [5] Nazabal et at., 2020 [6] Mattei et at., 2020

Challenges

Jointly increase flexibility and improve inference

- Hierarchical VAEs are successful at modeling flexibility.

$$p(\boldsymbol{z}_1)\prod_{l=2}^{L} p(\boldsymbol{z}_l \mid \boldsymbol{z}_{l-1})$$



- Complicated posteriors -> **Delicate inference** (posterior collapse).

- Gaussian approximate inference worsens.

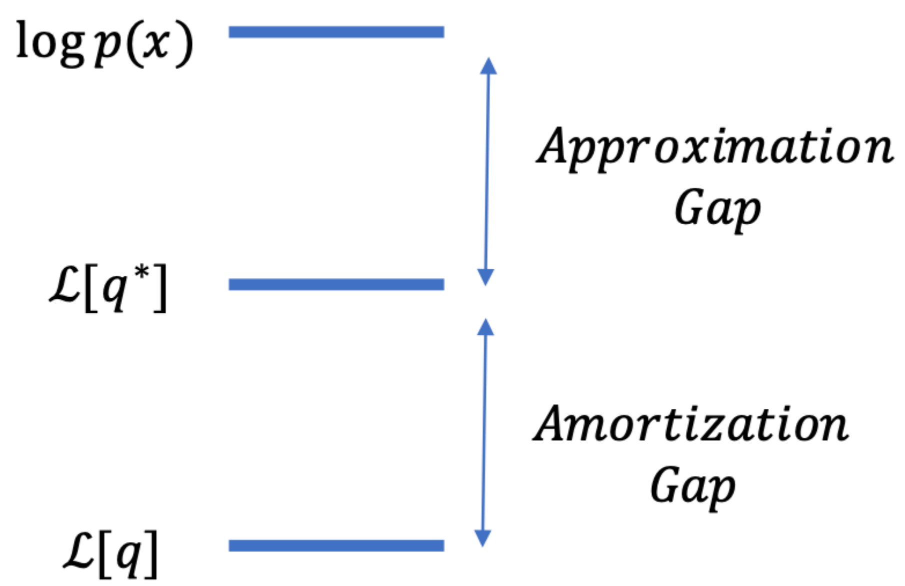
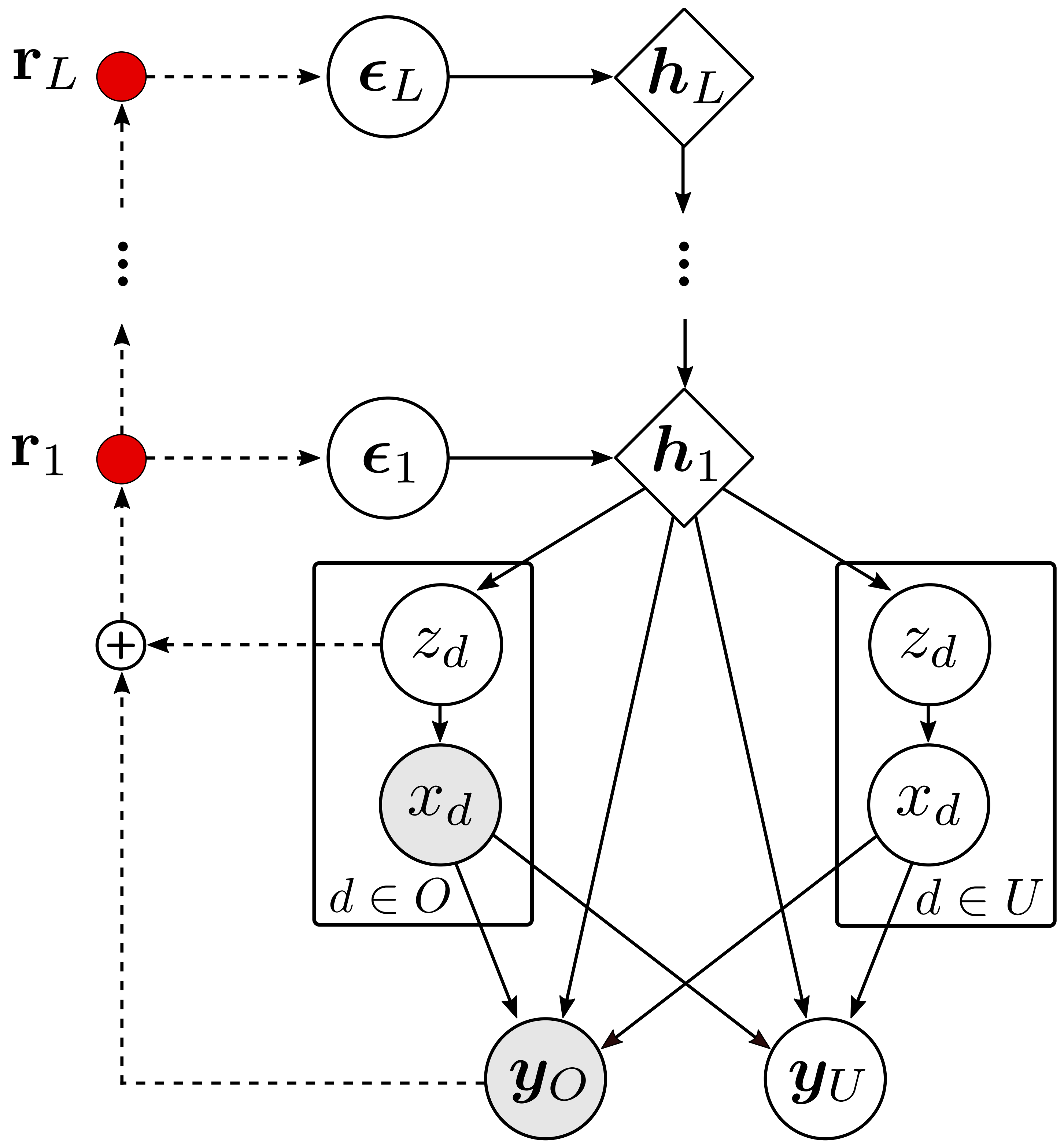


Figure 1. Gaps in Inference [6]

Contributions

Hierarchical Hamiltonian VAE for mixed-type incomplete data (HH-VAEM)

- Increased flexibility by using hierarchical latent space.
- Improved inference by means of automatically tuned HMC.
- Reparameterization for well-posed HMC on relaxed posterior.
- More accurate imputation and prediction.
- More effective information acquisition.



Information acquisition

Sampling-based method

- **Sampling**-based estimator [11] of the Mutual Information:

$$R(i, \boldsymbol{x}_O) = D_{\text{KL}}\left[p(\boldsymbol{y}, x_i|\boldsymbol{x}_O)||p(\mathbf{y}|\boldsymbol{x}_O)p(x_i|\boldsymbol{x}_O)\right] = \mathcal{I}(\boldsymbol{y}; x_i\ |\boldsymbol{x}_O) =$$

$$= \iint_{x_i, \boldsymbol{y}} p_{x_i,\boldsymbol{y}|\boldsymbol{x}_O}(x_i, \boldsymbol{y}|\boldsymbol{x}_O) \log\left(\frac{p_{x_i,\boldsymbol{y}|\boldsymbol{x}_O}(x_i, \boldsymbol{y}|\boldsymbol{x}_O)}{p_{x_i|\boldsymbol{x}_O}(x_i|\boldsymbol{x}_O)p_{\boldsymbol{y}|\boldsymbol{x}_O}(\boldsymbol{y}|\boldsymbol{x}_O)}\right)$$

$$\hat{I}(\boldsymbol{y}; x_i\ |\boldsymbol{x}_O) \approx \sum_{ij} p_{x_i,\boldsymbol{y}|\boldsymbol{x_O}}(i,j) \log \frac{p_{x_i,\boldsymbol{y}|\boldsymbol{x_O}}(i,j)}{p_{x_i|\boldsymbol{x}_O}(i)p_{\boldsymbol{y}|\boldsymbol{x}_O}(j)}$$

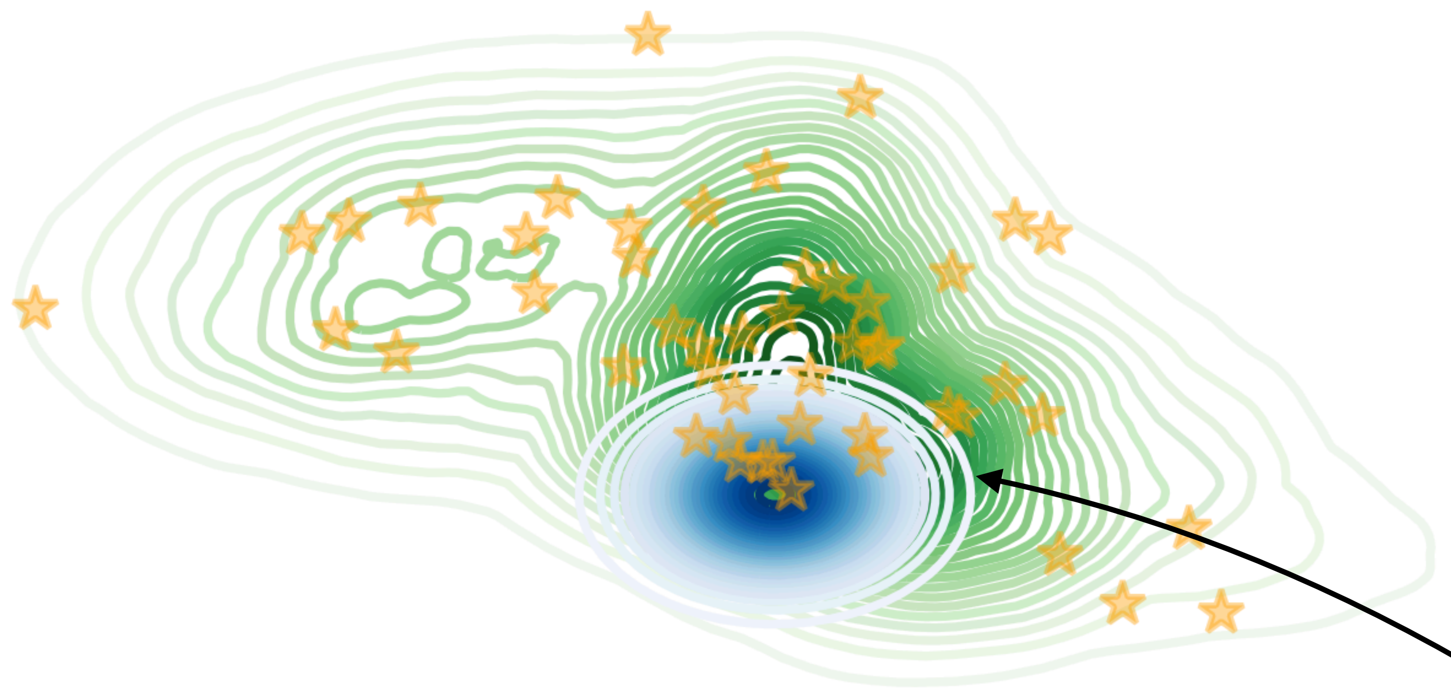
✓ Avoids the Gaussian approximation

✓ Efficient, easy parallelization.

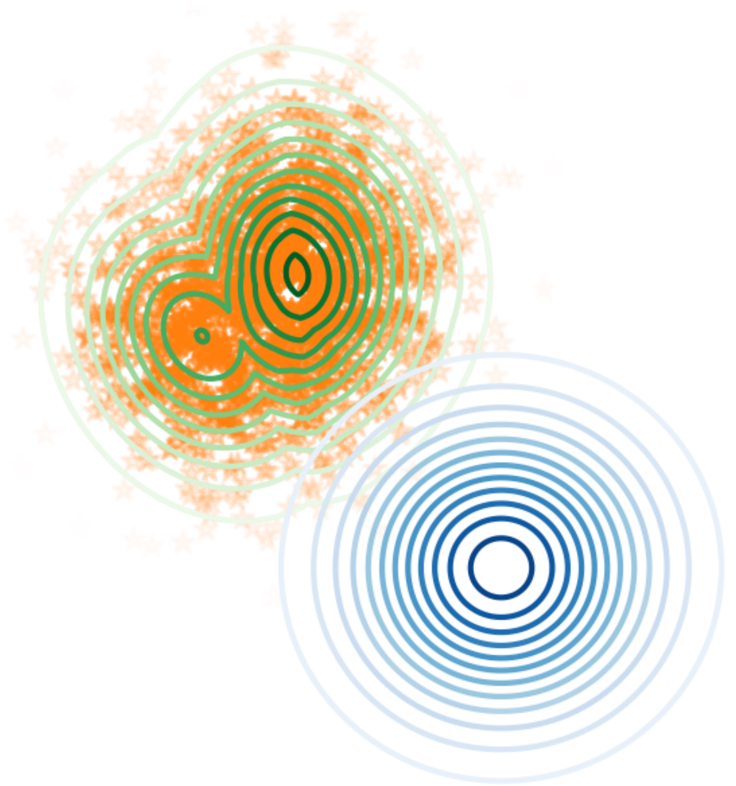
[11] Kraskov et at., 2004

HH-VAEM

Improving inference



[2]HMC samples (orange)



[2]HMC samples (orange)

$$\mathcal{L}(\boldsymbol{x};\theta,\phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] - D_{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \| p(\boldsymbol{z})\right)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}^{(s)}) - D_{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \| p(\boldsymbol{z})\right)$$

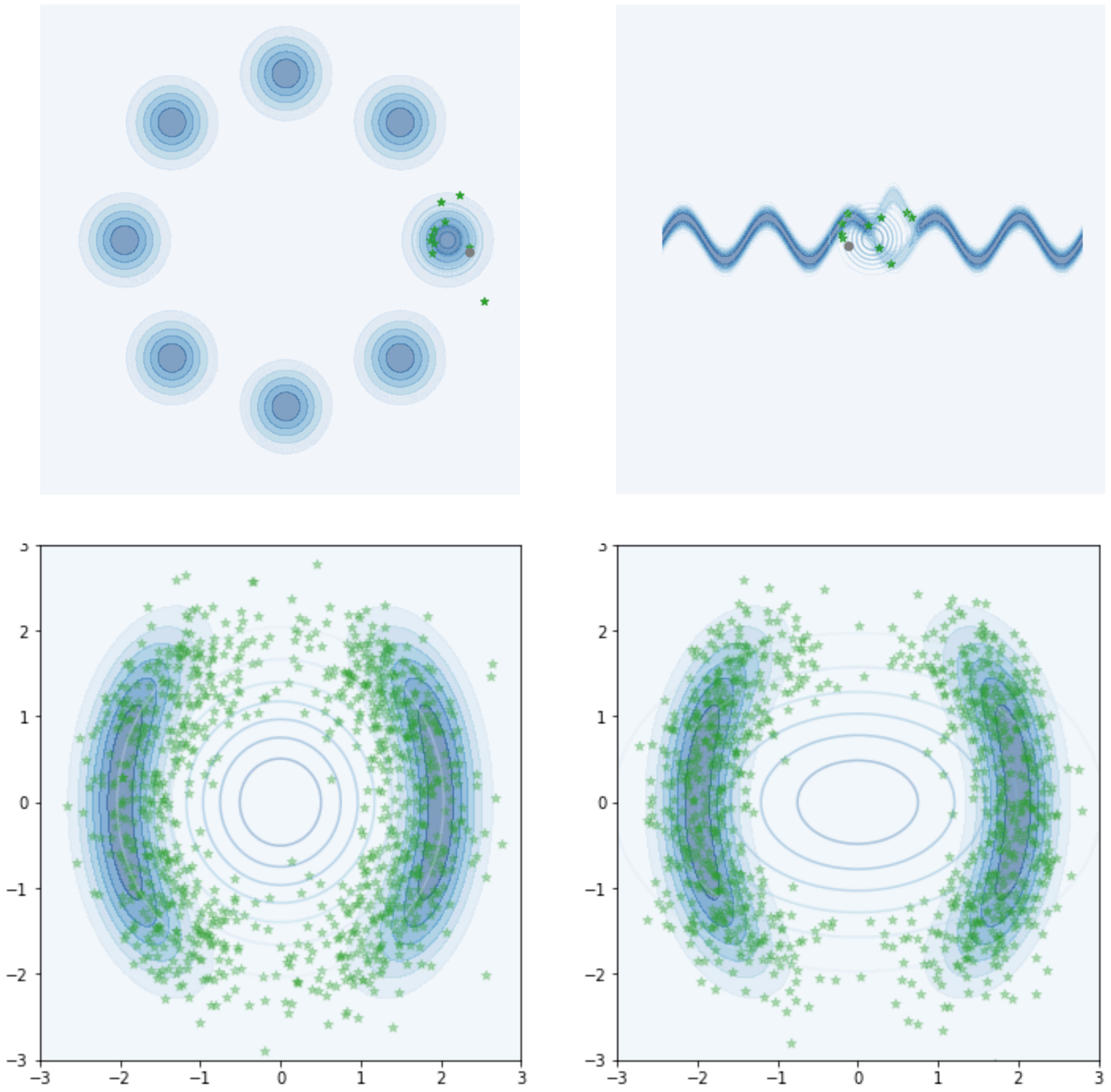
Hamiltonian Monte Carlo

Improving inference

- Discrete trajectories (*chains*) of T updates, ending in:

$$q^{(T)}(\boldsymbol{z}|\boldsymbol{x}) \approx p(\boldsymbol{z}|\boldsymbol{x})$$

- Target: true posterior density.
- Needed:
  - 1. Good initial proposal (encoder).
  - 2. Well-defined hyperparameters.



Random hyperparameters Tuned hyperparameters

Hamiltonian Monte Carlo

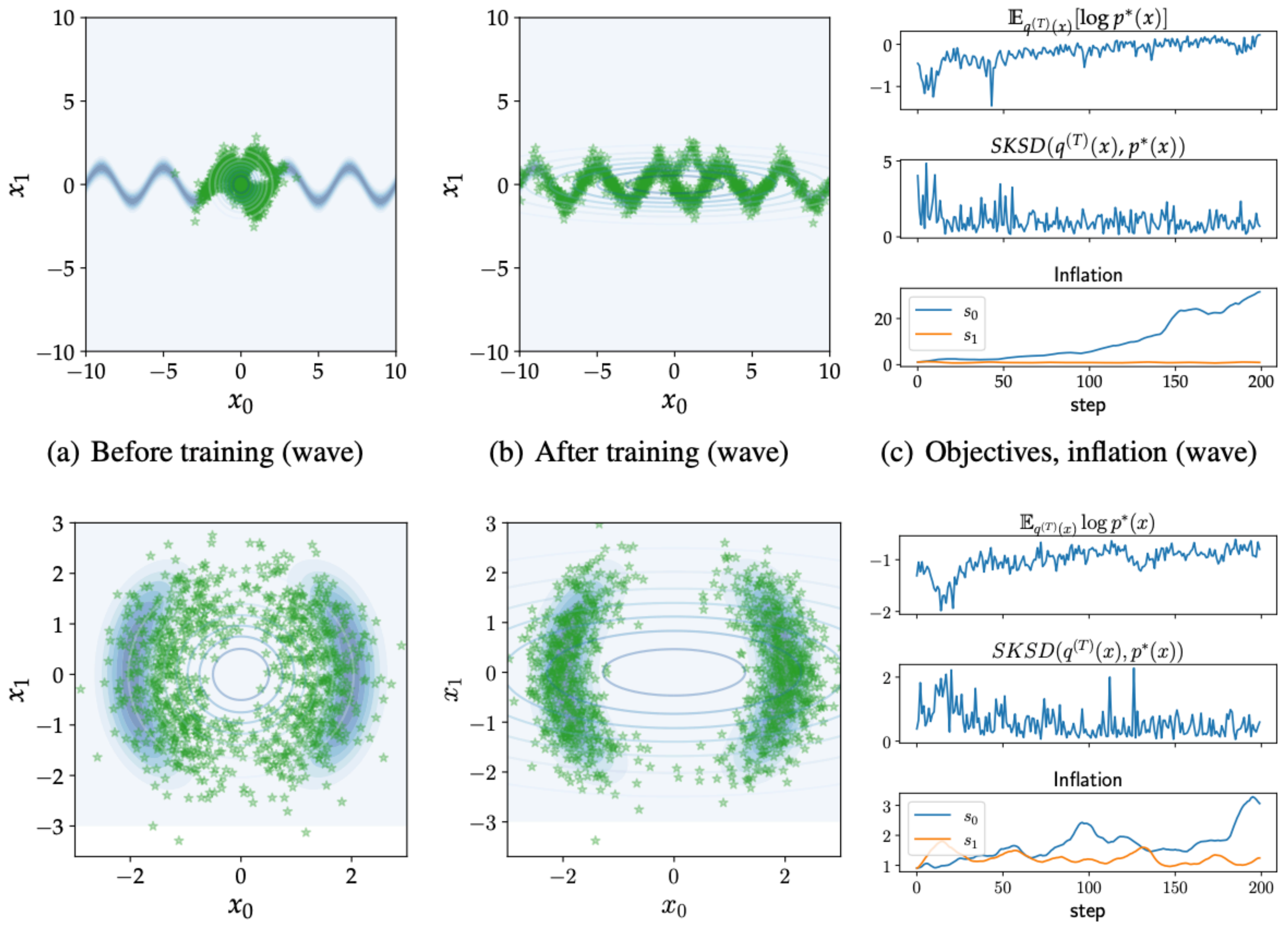
Hyperparameter tuning [7]

- Tuning the hyperparameters via Variational Inference:

$$\phi^* = \underset{\phi}{\operatorname{argmax}} \ \mathbb{E}_{q_\phi^{(T)}(\boldsymbol{z})}\left[\log p^*(\boldsymbol{z})\right] + H\left[q_\phi^{(T)}(\boldsymbol{z})\right]$$

- Add an **inflation** parameter for scaling the proposal [8]

$$\boldsymbol{s}^* = \underset{s}{\operatorname{argmin}} \ \mathrm{SKSD}(\boldsymbol{z}^{(T)}, \nabla_z \log p^*(\boldsymbol{z}))$$



(a) Before training (wave) (b) After training (wave) (c) Objectives, inflation (wave)
(d) Before training (dual moon) (e) After training (dual moon) (f) Objectives, inflation (dual moon)

[7] Campbell et at., 2021 [8] Gong et at., 2020

Hierarchical latent space

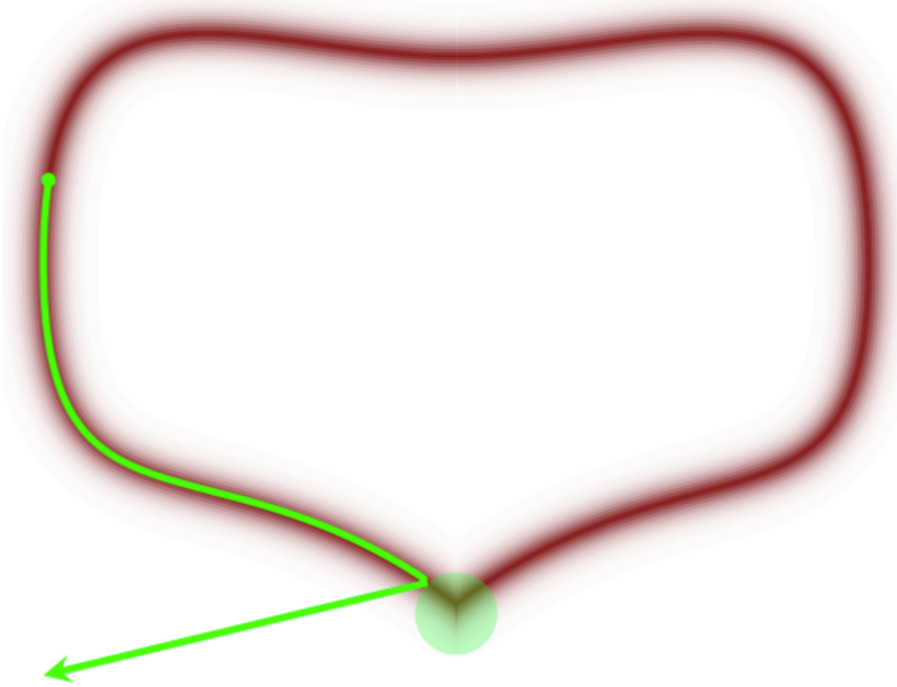
Ill-posed for HMC

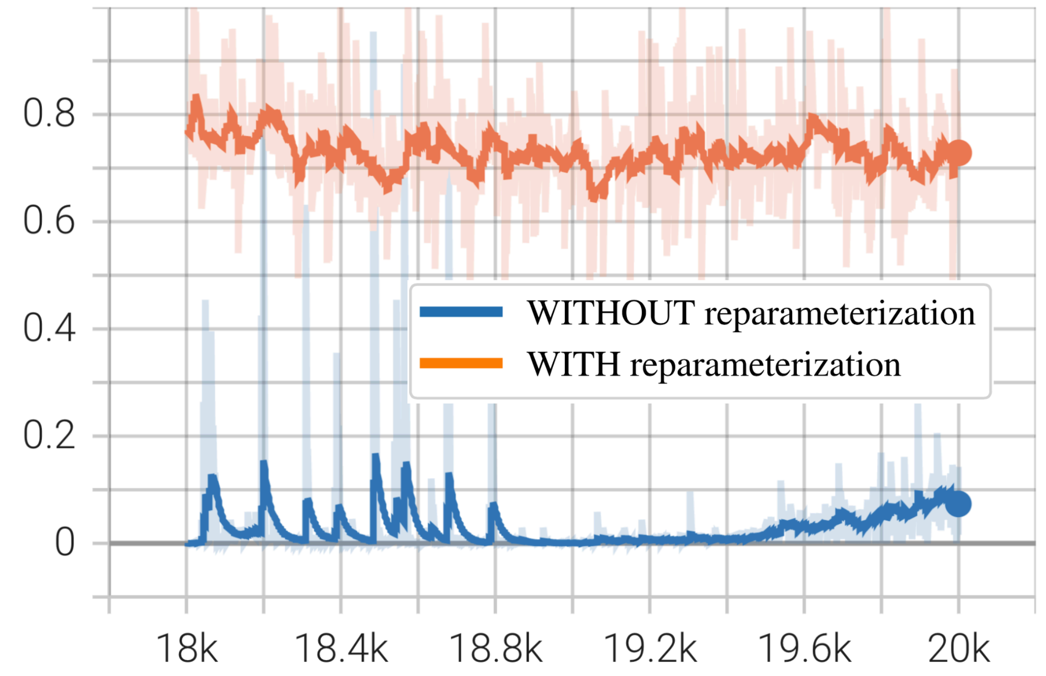
- Hierarchical dependencies lead to **huge gradients [11, 12]**

$$\boldsymbol{r}_{l+\frac{1}{2}} = \boldsymbol{r}_l + \frac{1}{2}\phi \odot \nabla_{z_l} \log p^*(\boldsymbol{z}_l)\,,$$

$$\boldsymbol{z}_{l+1} = \boldsymbol{z}_k + \boldsymbol{r}_{l+\frac{1}{2}} \odot \phi \odot \frac{1}{\boldsymbol{M}}\,,$$

$$\boldsymbol{r}_{l+1} = \boldsymbol{r}_{l+\frac{1}{2}} + \frac{1}{2}\phi \odot \nabla_{z_{l+1}} \log p^*(\boldsymbol{z}_{l+1}),$$





(a) Mean acceptance rate $\bar{p}_a$

[9] Betancourt et at., 2017 [10] Betancourt et at., 2015

Hierarchical latent space

Ill-posed for HMC

- **Solution**:

    ✓ Reparameterization for relaxed posterior:

$$\boldsymbol{h}_l = f_{\mu_l}(\boldsymbol{h}_{l+1}) + f_{\sigma_l}(\boldsymbol{h}_{l+1}) \cdot \boldsymbol{\epsilon}_l$$

NNs with parameters $\theta_{\mu_l} \rightarrow f_{\mu_l}, \theta_{\sigma_l} \rightarrow f_{\sigma_l}$

(a) AR hierarchy (b) Reparameterization

Hierarchical latent space

Ill-posed for HMC

- **Solution**:

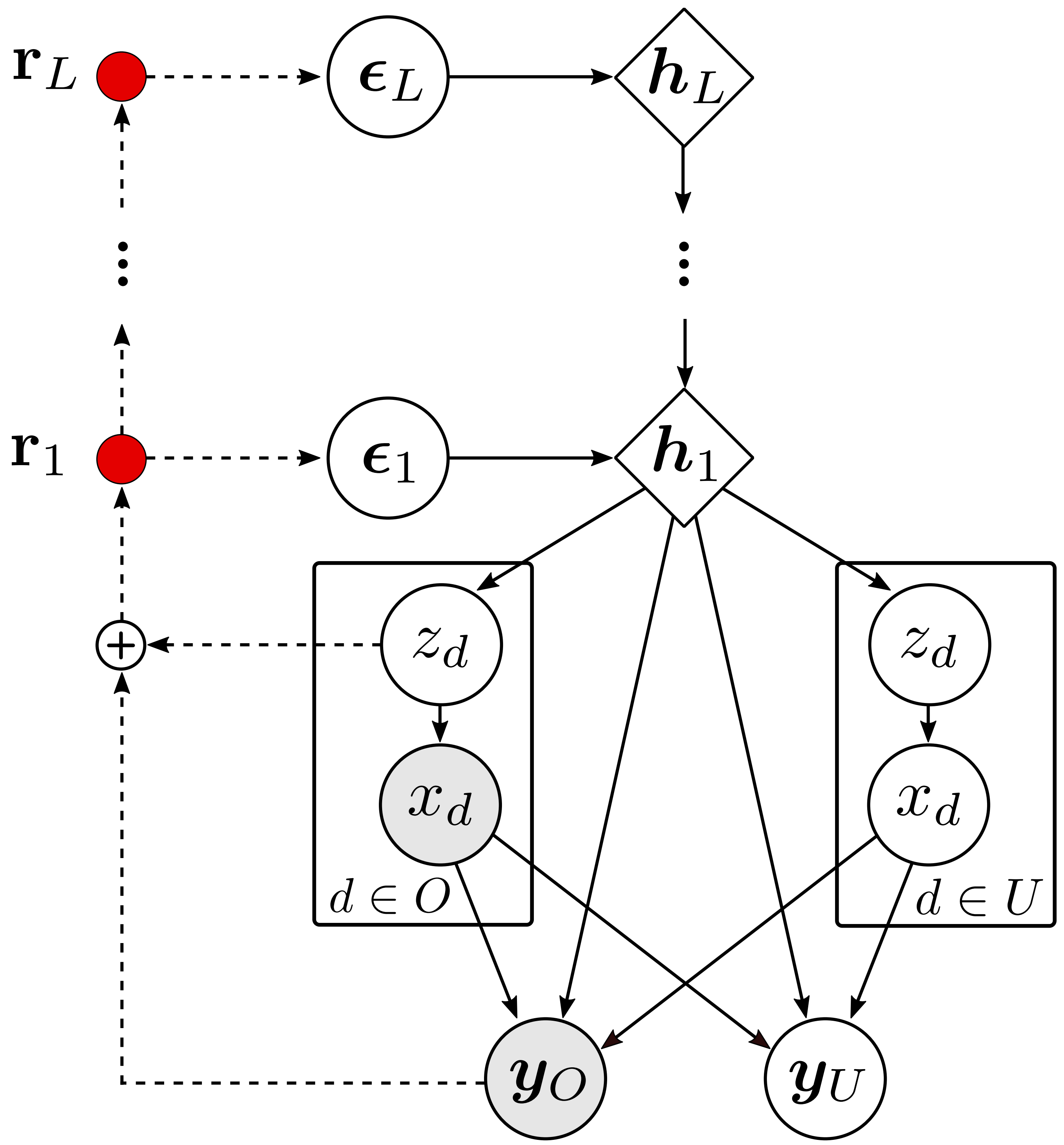
  ✓ Reparameterization for relaxed posterior:

  $$\boldsymbol{h}_l = f_{\mu_l}(\boldsymbol{h}_{l+1}) + f_{\sigma_l}(\boldsymbol{h}_{l+1}) \cdot \boldsymbol{\epsilon}_l$$

  NNs with parameters $\theta_{\mu_l} \rightarrow f_{\mu_l}, \theta_{\sigma_l} \rightarrow f_{\sigma_l}$

  ✓ Perform inference on $\epsilon = \{\epsilon_1, \ldots, \epsilon_1\}$ with standard Gaussian prior.

  ✓ No residual distributions needed.



Hierarchical latent space

Ill-posed for HMC

- **Solution**:

  ✓ Reparameterization for relaxed posterior:

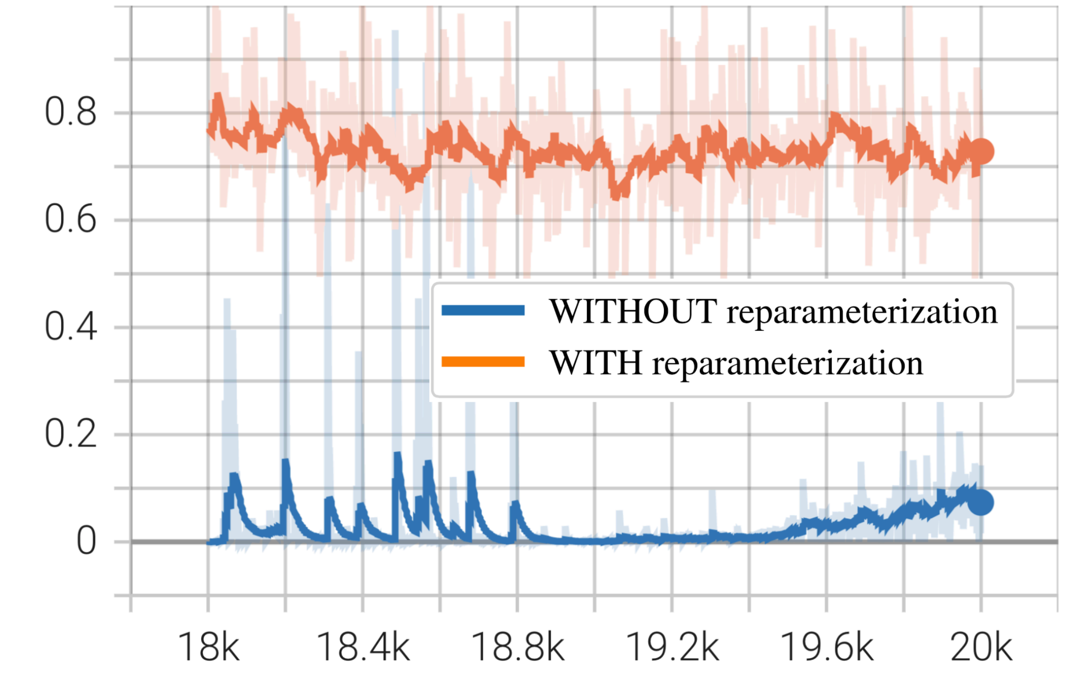
$$\boldsymbol{h}_l = f_{\mu_l}(\boldsymbol{h}_{l+1}) + f_{\sigma_l}(\boldsymbol{h}_{l+1}) \cdot \boldsymbol{\epsilon}_l$$

  NNs with parameters $\theta_{\mu_l} \rightarrow f_{\mu_l}, \theta_{\sigma_l} \rightarrow f_{\sigma_l}$

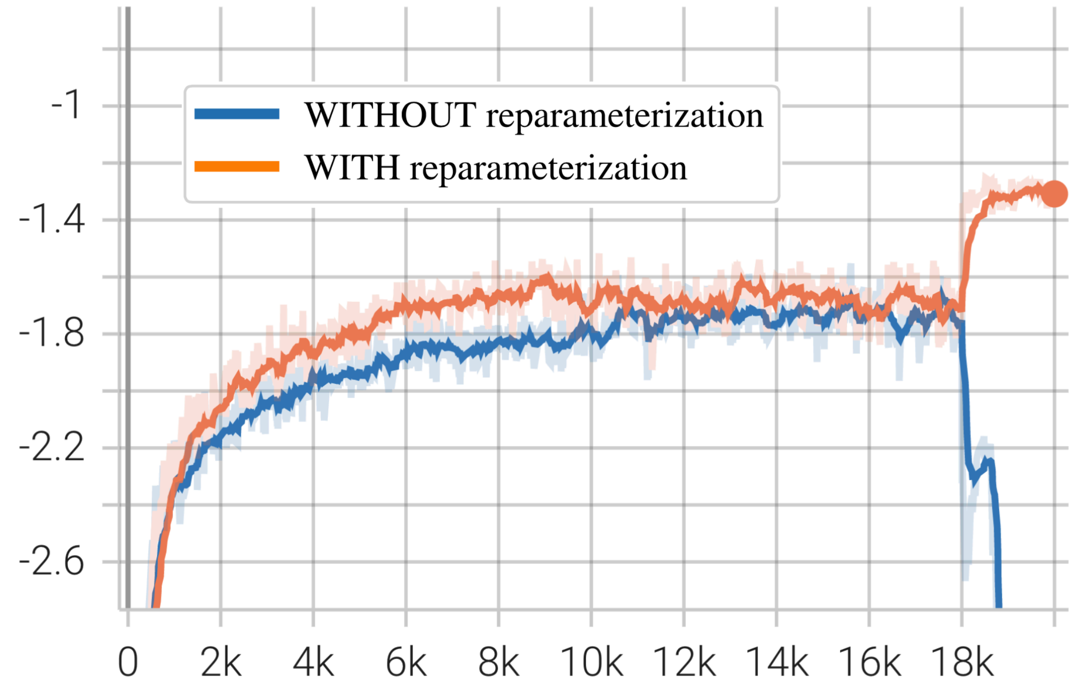
  ✓ Perform inference on $\epsilon = \{\epsilon_1, \ldots, \epsilon_1\}$ with standard Gaussian prior.

  ✓ No residual distributions needed.

  ✓ No need to increase complexity of the HMC method.



(a) Mean acceptance rate $\bar{p}_a$



(b) $\log p(\boldsymbol{x}_U|\boldsymbol{x}_O)$

HH-VAEM

Optimization algorithm

Algorithm 1 Training algorithm for HH-VAEM

Input: data $\left(\boldsymbol{x}_O^{(1:N)}, \boldsymbol{y}_O^{(1:N)}\right)$, steps: T_d, T_{VI}, T_{HMC}
Parameters: γ, θ, ψ, ϕ, s
STAGE 1: MARGINAL VAEs
for $d = 1$ **to** D **do**
  Initialize marginal VAE $\{\theta_d, \gamma_d\}$
  for $t = 1$ **to** T_d **do**
    $\gamma_d^{t+1}, \theta_d^{t+1} \leftarrow \text{Adam}_{\gamma_d^t, \theta_d^t}(\mathcal{L}_d)$
  end for
end for

- 1. Train marginal VAEs using:

$$\mathcal{L}_d\left(x_d; \{\theta_d, \gamma_d\}\right) = \mathbb{I}\left(x_d \in \boldsymbol{x}_O\right) \mathbb{E}_{q_{\gamma_d}(z_d \mid x_d)} \log \frac{p_{\theta_d}\left(x_d, z_d\right)}{q_{\gamma_d}\left(z_d \mid x_d\right)}$$

HH-VAEM

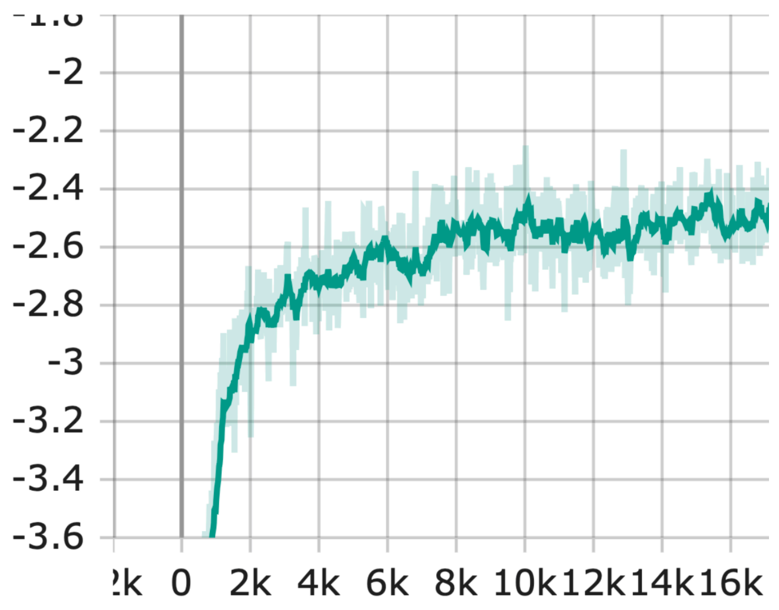
Optimization algorithm

Algorithm 1 Training algorithm for HH-VAEM

Input: data $\left(\boldsymbol{x}_O^{(1:N)}, \boldsymbol{y}_O^{(1:N)}\right)$, steps: T_d, T_{VI}, T_{HMC}
Parameters: $\gamma, \theta, \psi, \phi, s$
STAGE 1: MARGINAL VAEs
for $d = 1$ **to** D **do**
  Initialize marginal VAE $\{\theta_d, \gamma_d\}$
  for $t = 1$ **to** T_d **do**
    $\gamma_d^{t+1}, \theta_d^{t+1} \leftarrow \text{Adam}_{\gamma_d^t, \theta_d^t}(\mathcal{L}_d)$
  end for
end for
STAGE 2: DEPENDENCY VAE
for $t = 1$ **to** T_{VAE} **do**
  $\theta^{t+1}, \psi^{t+1} \leftarrow \text{Adam}_{\theta^t, \psi^t}(\mathcal{L}_{VI})$
end for

- 2. Training Hierarchical VAE using the ELBO:

$$\mathcal{L}_{VI}\left(\boldsymbol{x}_O, \boldsymbol{y}_O; \{\theta, \psi\}\right) = \mathbb{E}_{q_\psi}\left[\log p_\theta\left(\boldsymbol{z}_O \mid \boldsymbol{h}_1\right) + \log p_\theta\left(\boldsymbol{y}_O \mid \hat{\boldsymbol{x}}, \boldsymbol{h}_1\right)\right] - \sum_{l=1}^{L} D_{\mathrm{KL}}\left(q_\psi\left(\boldsymbol{\epsilon}_l \mid \boldsymbol{x}_O, \boldsymbol{y}_O\right) \| p\left(\boldsymbol{\epsilon}_l\right)\right)$$



(a) $\log p(\boldsymbol{x}_U | \boldsymbol{x}_O)$

HH-VAEM

Optimization algorithm

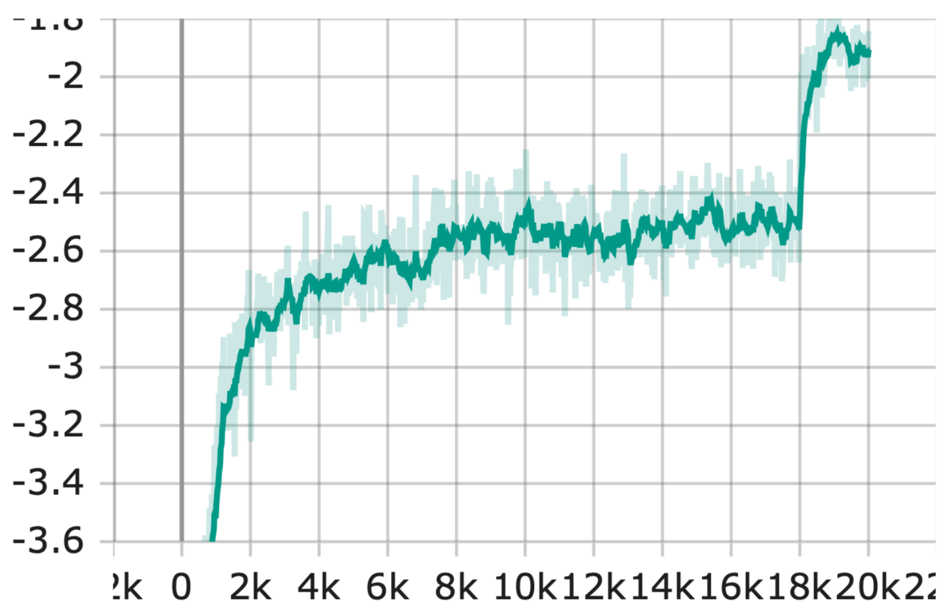
Algorithm 1 Training algorithm for HH-VAEM

Input: data $\left(\boldsymbol{x}_O^{(1:N)}, \boldsymbol{y}_O^{(1:N)}\right)$, steps: T_d, T_{VI}, T_{HMC}
Parameters: γ, θ, ψ, ϕ, s
STAGE 1: MARGINAL VAEs
for $d = 1$ **to** D **do**
  Initialize marginal VAE $\{\theta_d, \gamma_d\}$
  for $t = 1$ **to** T_d **do**
    $\gamma_d^{t+1}, \theta_d^{t+1} \leftarrow \text{Adam}_{\gamma_d^t, \theta_d^t}(\mathcal{L}_d)$
  end for
end for
STAGE 2: DEPENDENCY VAE
for $t = 1$ **to** T_{VAE} **do**
  $\theta^{t+1}, \psi^{t+1} \leftarrow \text{Adam}_{\theta^t, \psi^t}(\mathcal{L}_{VI})$
end for
STAGE 3: JOINTLY OPTIMIZING VAE + HMC
for $t = 1$ **to** T_{HMC} **do**
  $\psi^{t+1} \leftarrow \text{Adam}_{\psi^t}(\mathcal{L}_{VI})$
  $\theta^{t+1}, \phi^{t+1} \leftarrow \text{Adam}_{\theta^t, \phi^t}(\mathcal{L}_{HMC})$
  $s^{t+1} \leftarrow \text{Adam}_{s^t}(\mathcal{L}_{SKSD})$
end for

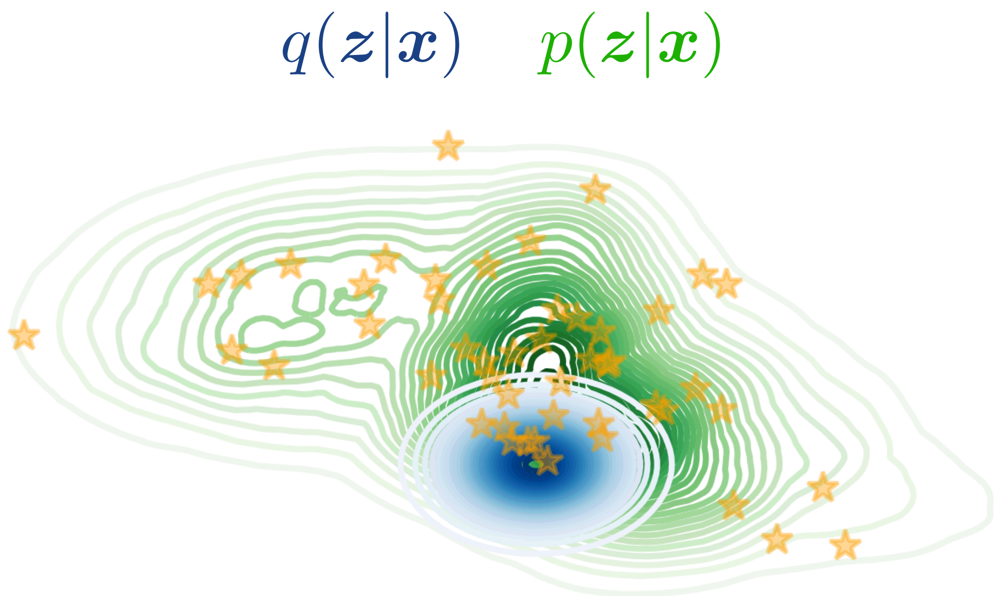
3. Train a) encoder using ELBO, b) HMC hyperparams, decoder and predictor parameters using HMC objective and c) scale using SKSD.

$$\mathcal{L}_{HMC}\left(\boldsymbol{z}_O, \boldsymbol{y}_O; \{\theta, \psi, \phi\}\right) = \mathbb{E}_{q_\phi^{(T)}(\boldsymbol{\epsilon})}\left[\log p_\theta\left(\boldsymbol{z}_O \mid \boldsymbol{h}_1\right) + \log p_\theta\left(\boldsymbol{y}_O \mid \hat{\boldsymbol{x}}, \boldsymbol{h}_1\right) + \sum_{l=1}^{L} p(\boldsymbol{\epsilon}_l^{(T)})\right]$$

$$\mathcal{L}_{SKSD}(\boldsymbol{x}_O, \boldsymbol{y}_O; \boldsymbol{s}) = \text{SKSD}\left(q_\phi^{(T)}(\boldsymbol{\epsilon}|\boldsymbol{z}_O, \boldsymbol{x}_O, \boldsymbol{y}_O; \boldsymbol{s}),\, p(\boldsymbol{\epsilon}|\boldsymbol{z}_O, \boldsymbol{x}_O, \boldsymbol{y}_O)\right)$$



(a) $\log p(\boldsymbol{x}_U|\boldsymbol{x}_O)$



[2]HMC samples (orange)

Experiments

Mixed-type data

$$\log p(\boldsymbol{x}_U|\boldsymbol{x}_O) = \log \mathbb{E}_{\boldsymbol{\epsilon}\sim q^{(T)}(\boldsymbol{\epsilon}|\boldsymbol{x}_O)}\left[p(\boldsymbol{x}_U|\boldsymbol{\epsilon})\right] \approx \log \frac{1}{k}\sum_i^k p(\boldsymbol{x}_U|\boldsymbol{\epsilon}_i)$$

	Bank	Insurance	Avocado	Naval	Yatch	Diabetes	Concrete	Wine	Energy	Boston
VAEM	2.84 ± 0.07	1.81 ± 0.03	1.89 ± 0.01	0.55 ± 0.05	3.15 ± 0.28	2.78 ± 0.16	2.45 ± 0.26	3.01 ± 0.61	2.09 ± 0.10	2.01 ± 0.23
MIWAEM	2.74 ± 0.05	1.88 ± 0.04	1.92 ± 0.04	0.57 ± 0.03	2.66 ± 0.11	2.55 ± 0.09	2.34 ± 0.51	2.76 ± 0.48	2.06 ± 0.14	1.94 ± 0.23
H-VAEM	2.82 ± 0.06	1.80 ± 0.04	1.89 ± 0.01	0.48 ± 0.06	3.06 ± 0.31	2.74 ± 0.09	2.42 ± 0.21	2.85 ± 0.56	1.72 ± 0.11	1.89 ± 0.24
HMC-VAEM	2.69 ± 0.05	1.77 ± 0.06	1.89 ± 0.02	0.49 ± 0.07	$\mathbf{2.21 \pm 0.24}$	2.72 ± 0.20	2.28 ± 0.29	2.83 ± 0.46	1.73 ± 0.05	1.83 ± 0.16
HH-VAEM	$\mathbf{2.63 \pm 0.04}$	$\mathbf{1.75 \pm 0.03}$	$\mathbf{1.88 \pm 0.05}$	$\mathbf{0.40 \pm 0.05}$	2.47 ± 0.27	$\mathbf{2.54 \pm 0.13}$	$\mathbf{2.28 \pm 0.09}$	$\mathbf{1.90 \pm 0.17}$	$\mathbf{1.71 \pm 0.04}$	$\mathbf{1.83 \pm 0.11}$

Table 1: Test negative log likelihood of the unobserved features for our model and baselines.

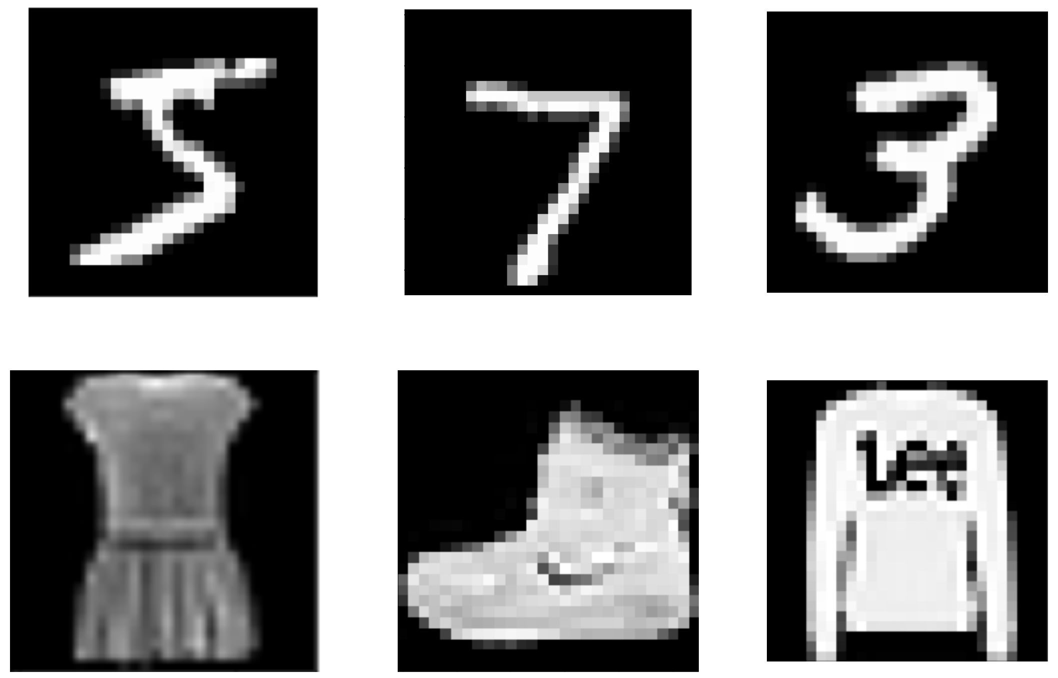
$$\log p(\boldsymbol{y}|\boldsymbol{x}_O) = \log \mathbb{E}_{\boldsymbol{\epsilon}\sim q^{(T)}(\boldsymbol{\epsilon}|\boldsymbol{x}_O)}\left[p(\boldsymbol{y}|\boldsymbol{\epsilon})\right] \approx \log \frac{1}{k}\sum_i^k p(\boldsymbol{y}|\boldsymbol{\epsilon}_i),$$

	Bank	Insurance	Avocado	Naval	Yatch	Diabetes	Concrete	Wine	Energy	Boston
VAEM	0.56 ± 0.06	1.20 ± 0.03	1.18 ± 0.02	2.69 ± 0.01	0.61 ± 0.02	1.59 ± 0.19	1.07 ± 0.09	0.28 ± 0.09	0.61 ± 0.14	0.85 ± 0.21
MIWAEM	0.51 ± 0.03	1.15 ± 0.03	1.15 ± 0.03	2.70 ± 0.01	0.60 ± 0.03	$\mathbf{1.36 \pm 0.10}$	0.95 ± 0.22	0.28 ± 0.13	0.54 ± 0.12	0.80 ± 0.21
H-VAEM	0.50 ± 0.03	1.06 ± 0.02	1.18 ± 0.02	2.68 ± 0.01	0.60 ± 0.02	1.71 ± 0.14	1.02 ± 0.09	0.26 ± 0.11	0.46 ± 0.14	0.90 ± 0.22
HMC-VAEM	0.52 ± 0.02	1.00 ± 0.03	1.12 ± 0.03	2.71 ± 0.01	$\mathbf{0.52 \pm 0.15}$	1.55 ± 0.29	0.95 ± 0.26	0.28 ± 0.09	0.41 ± 0.07	0.71 ± 0.13
HH-VAEM	$\mathbf{0.49 \pm 0.03}$	$\mathbf{0.93 \pm 0.06}$	$\mathbf{1.10 \pm 0.01}$	$\mathbf{2.62 \pm 0.01}$	0.56 ± 0.02	1.38 ± 0.18	$\mathbf{0.95 \pm 0.08}$	$\mathbf{0.20 \pm 0.04}$	$\mathbf{0.32 \pm 0.05}$	$\mathbf{0.55 \pm 0.04}$

Table 2: Test negative log likelihood of the predicted target for our model and baselines.

Experiments

MNIST datasets



	VAE	MIWAE	H-VAE	HMC-VAE	HH-VAE
MNIST	0.124 ± 0.001	0.121 ± 0.001	0.119 ± 0.001	0.101 ± 0.004	$\mathbf{0.094 \pm 0.003}$
F-MNIST	0.162 ± 0.002	0.160 ± 0.002	0.156 ± 0.002	0.150 ± 0.002	$\mathbf{0.144 \pm 0.002}$

Table 3: Test negative log likelihood of the unobserved features for the MNIST datasets.

	VAE	MIWAE	H-VAE	HMC-VAE	HH-VAE
MNIST	0.153 ± 0.009	0.151 ± 0.007	0.146 ± 0.006	0.067 ± 0.007	$\mathbf{0.056 \pm 0.019}$
F-MNIST	0.501 ± 0.012	0.496 ± 0.008	0.494 ± 0.007	0.357 ± 0.060	$\mathbf{0.337 \pm 0.069}$

Table 4: Test negative log likelihood of the predicted target for the MNIST datasets.

	VAE	MIWAE	H-VAE	HMC-VAE	HH-VAE
MNIST	0.953 ± 0.004	0.953 ± 0.003	0.953 ± 0.003	0.978 ± 0.003	$\mathbf{0.981 \pm 0.005}$
F-MNIST	0.824 ± 0.005	0.824 ± 0.004	0.824 ± 0.004	0.869 ± 0.015	$\mathbf{0.876 \pm 0.017}$

Table 5: Test accuracy of the predicted digits for the MNIST datasets.

Experiments

Sequential Active Information Acquisition (SAIA)

- Sequentially acquiring high-value information by selecting features that maximize

$$\hat{I}\left(\boldsymbol{y}; x_i \mid \boldsymbol{x}_O\right) \approx \sum_{ij} p_{x_i,\boldsymbol{y}|\boldsymbol{x}_O}(i,j) \log \frac{p_{x_i,\boldsymbol{y}|\boldsymbol{x}_O}(i,j)}{p_{x_i|\boldsymbol{x}_O}(i)p_{\boldsymbol{y}|\boldsymbol{x}_O}(j)}$$

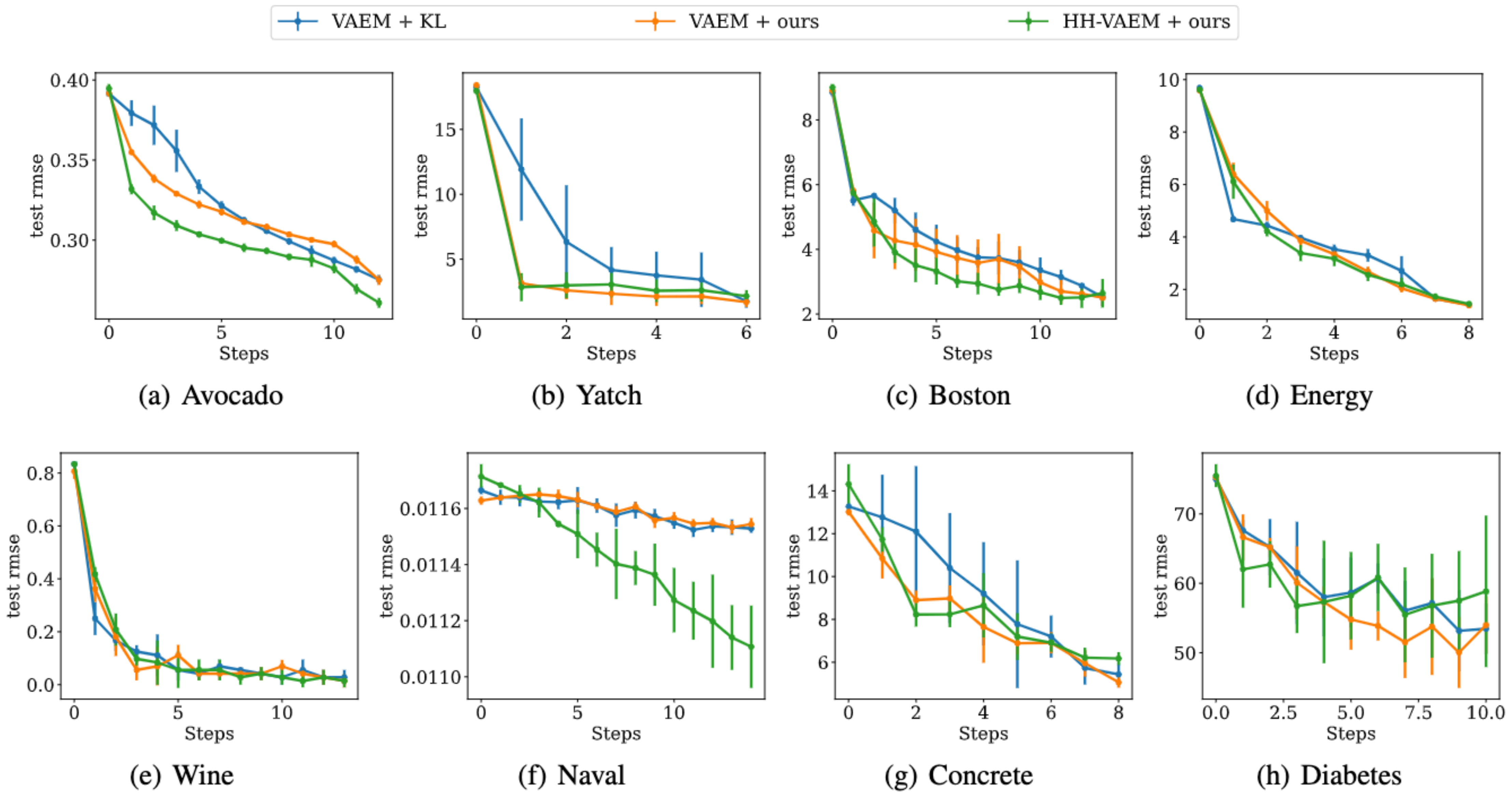
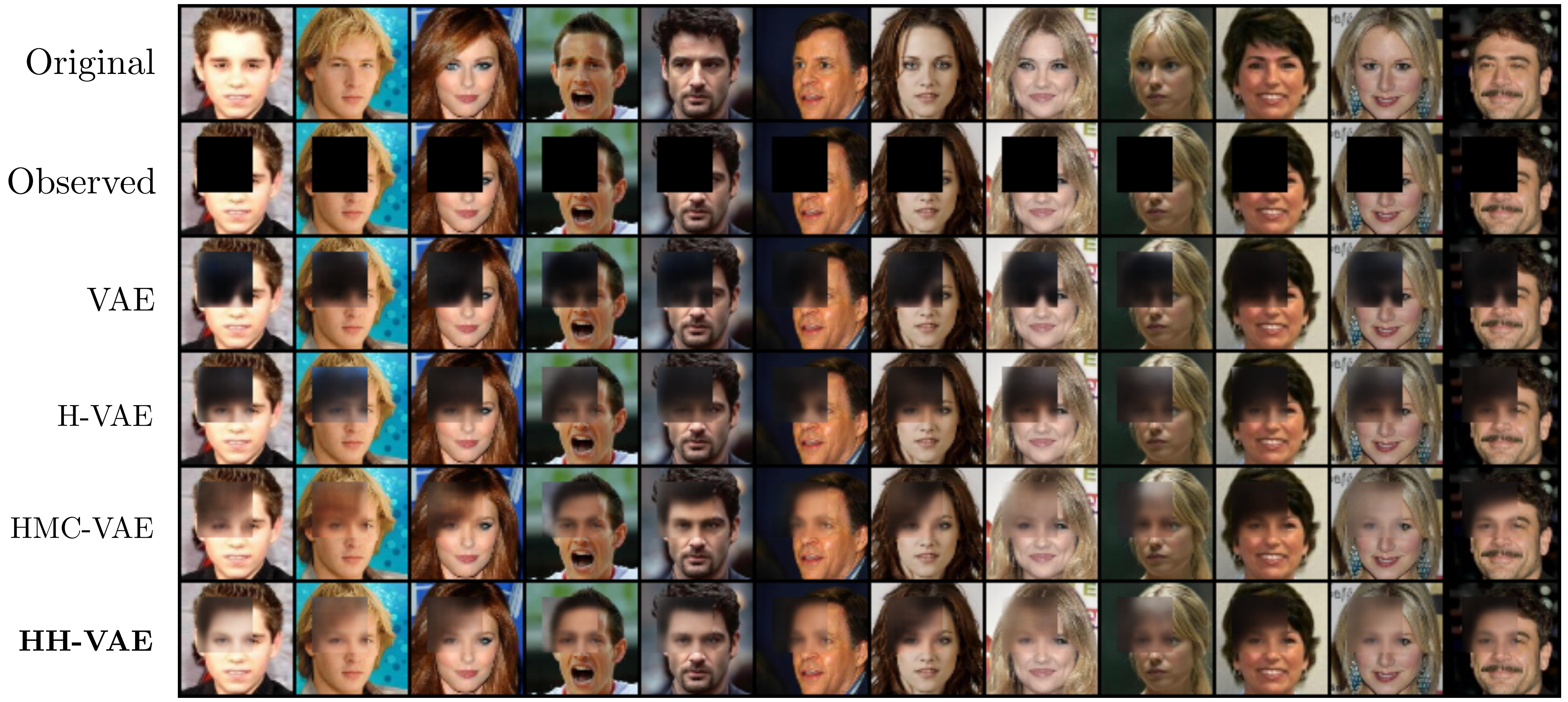


Figure 5: SAIA metric curves. Horizontal axis shows acquisition steps (number of discovered features). Vertical axis is the RMSE.

Experiments

Conditional image inpainting

1. Encode to $q_\phi^{(0)}\left(\boldsymbol{\epsilon} \mid \boldsymbol{z}_O, \boldsymbol{x}_O, \boldsymbol{y}_O\right)$
2. Using HMC, sample from $q_\phi^{(T)}\left(\boldsymbol{\epsilon} \mid \boldsymbol{z}_O, \boldsymbol{x}_O, \boldsymbol{y}_O\right)$
3. Decode to $p\left(\boldsymbol{x}_U \mid \boldsymbol{\epsilon}^{(T)}\right)$



Conclusion

- We presented:

  1. **HH-VAEM**: novel Hierarchical VAE improved with HMC with automatic hyperparameter optimization.

  2. Novel **sampling-based technique** based on the Mutual Information estimation for efficient information acquisition.

- Based on the provided experiments, we demonstrate that our methods:

  ✓ Improve approximate inference in hierarchical VAEs wrt to the Gaussian approximation.

  ✓ Improve missing data imputation task.

  ✓ Improve prediction task.

  ✓ Improve active information acquisition task.

References

- [1] Peis, I., Ma, C., & Hernández-Lobato, J. M. (2022). Missing Data Imputation and Acquisition with Deep Hierarchical Models and Hamiltonian Monte Carlo. arXiv preprint arXiv:2202.04599.
- [2] Bernardo, J. M. (1979). Expected information as expected utility. *the Annals of Statistics*, 686-690.
- [3] Ma, C., Tschiatschek, S., Palla, K., Hernández-Lobato, J. M., Nowozin, S., & Zhang, C. (2018). Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*.
- [4] Ma, C., Tschiatschek, S., Turner, R., Hernández-Lobato, J. M., & Zhang, C. (2020). VAEM: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, *33*, 11237-11247.
- [5] Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, *107*, 107501.
- [6] Mattei, P. A., & Frellsen, J. (2019, May). MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning* (pp. 4413-4423). PMLR.

References

- [7] Campbell, A., Chen, W., Stimper, V., Hernandez-Lobato, J. M., & Zhang, Y. (2021, July). A gradient based strategy for hamiltonian monte carlo hyperparameter optimization. In *International Conference on Machine Learning* (pp. 1238-1248). PMLR.

- [8] Gong, W., Li, Y., & Hernández-Lobato, J. M. (2020). Sliced kernelized Stein discrepancy. *arXiv preprint arXiv:2006.16531*.

- [9] Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

- [10] Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, *79*(30), 2-4.

- [11] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.