Laplacian Autoencoder for Stochastic Representation Learning

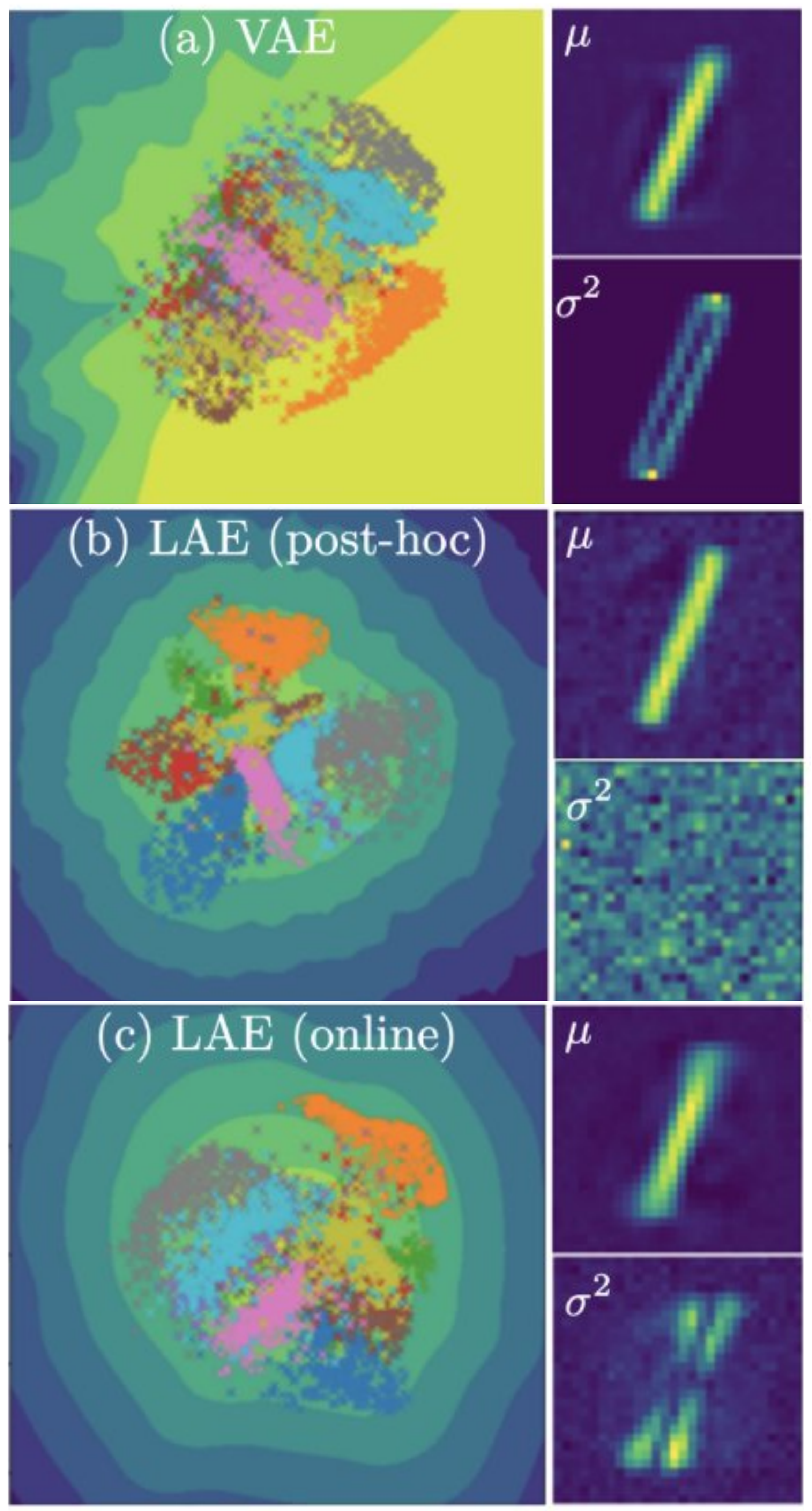
Marco Miani*, Frederik Warburg*, Pablo Moreno-Muñoz, Nicki Skafte Detlefsen, Søren Hauberg

Technical University of Denmark

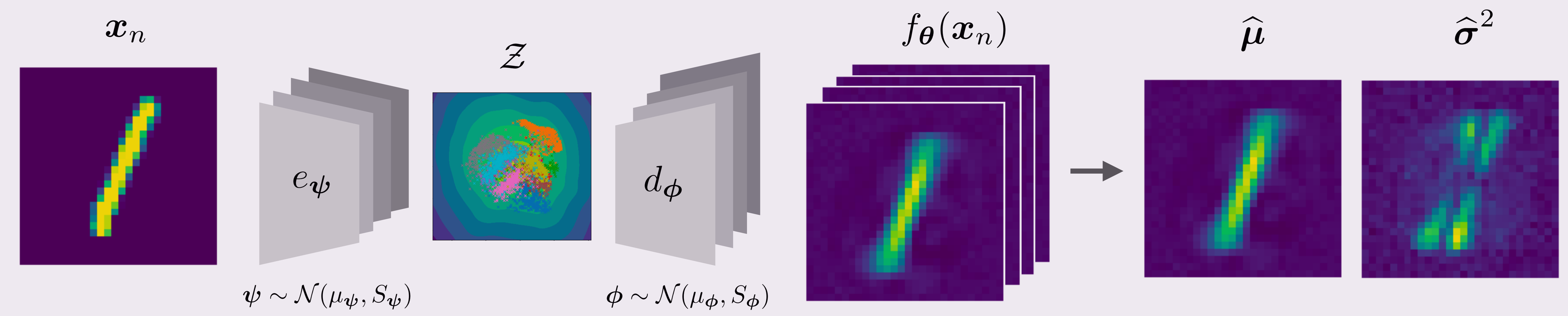
Abstract

AEs and VAEs produce none or poorly calibrated uncertainty estimates making it hard to evaluate if learned representations are stable and reliable.

1. We present a **Bayesian autoencoder** for unsupervised representation learning, which is trained using a novel variational lower-bound of the autoencoder evidence.
2. This is maximized using **Monte Carlo EM** with a variational distribution that takes the shape of a Laplace approximation.
3. We develop a new **Hessian approximation** that scales linearly with data size allowing us to model high-dimensional data.

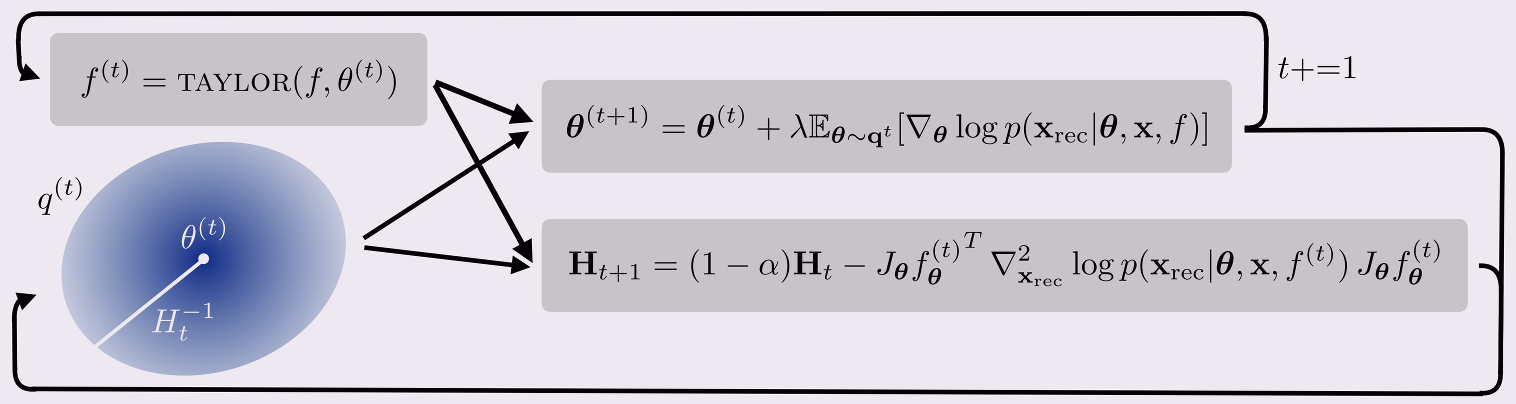


1. Bayesian Autoencoder



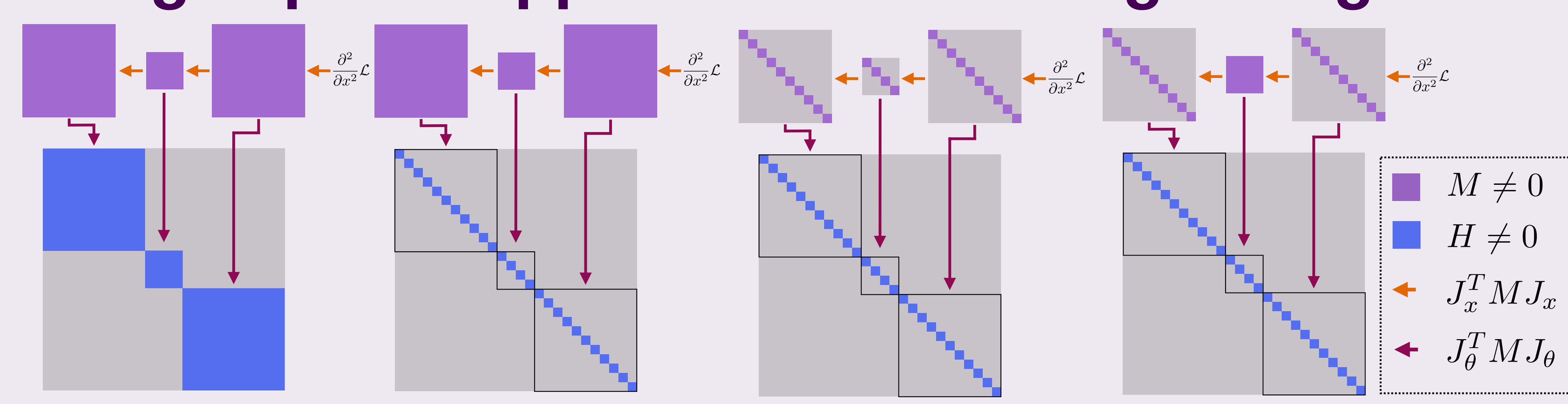
Model overview. We learn a distribution over parameters such that we can sample encoders and decoders. This allow us to compute the empirical mean and variance in both the latent space and the output space.

2. Iterative Learning (Monte Carlo EM)



Iterative training procedure. Given a distribution q over parameters, and a linearized function f, compute first and second-order derivatives to update the distribution on parameters.

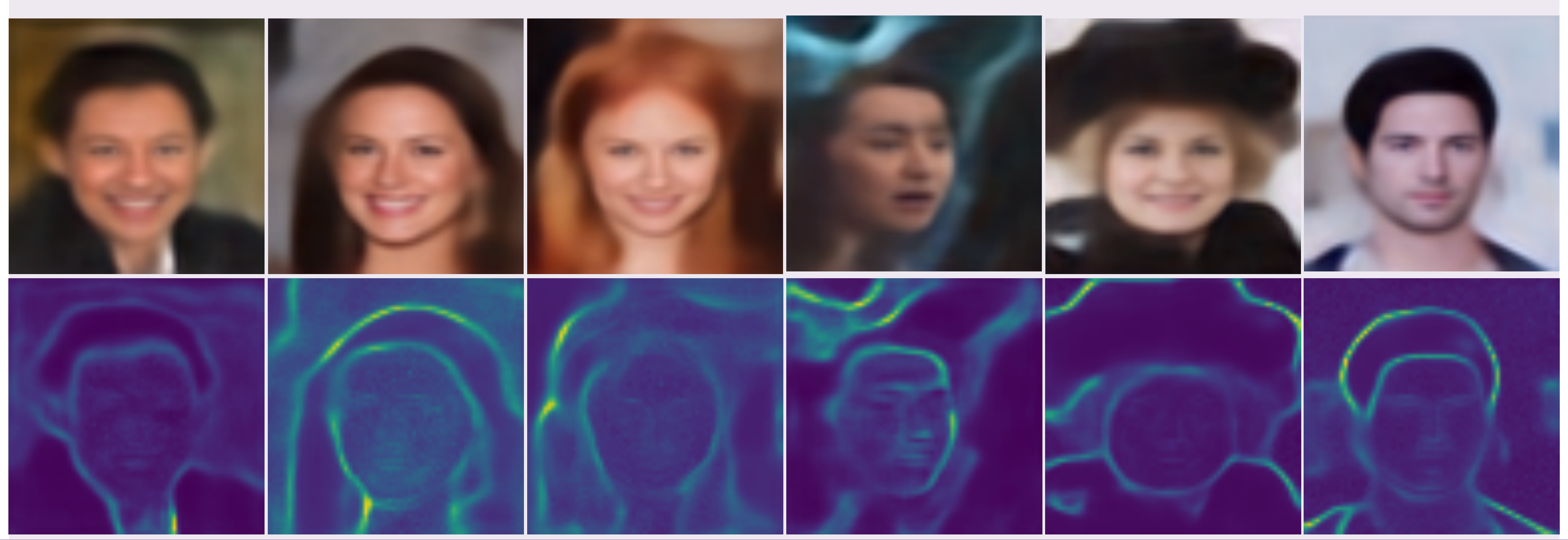
3. Scaling Laplace Approximation to Large Images



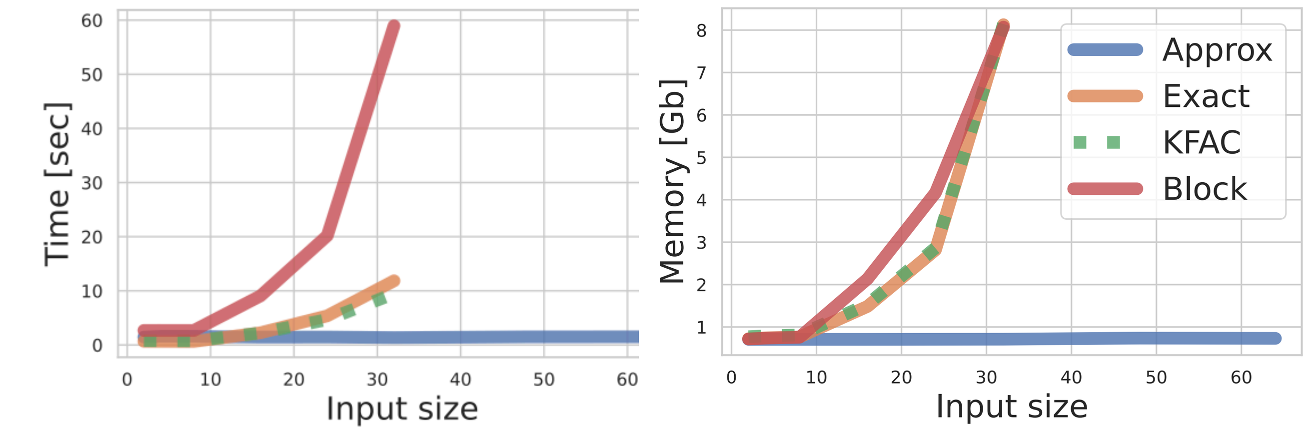
(a) Block diagonal (b) Exact diagonal (c) Approx. diagonal (d) Mixed diagonal

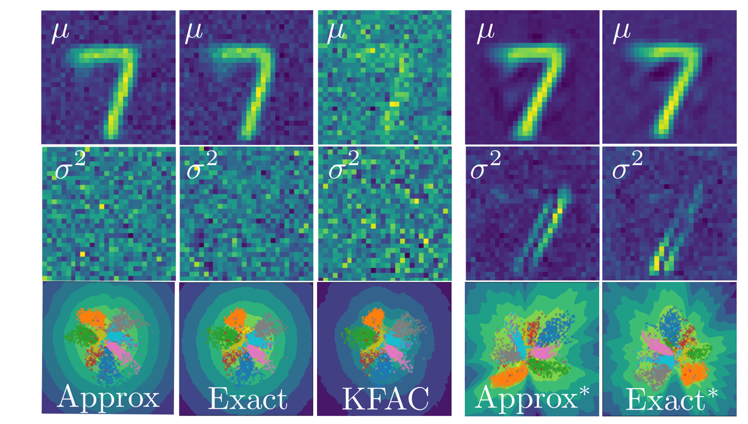
Comparison of Hessian approximation methods. Common approximations (a–b) scale quadratically with the output resolution. Our proposed approximate and mixed diagonal Hessians (c–d) scale linearly with the resolution. This is essential for scaling the LAE to large images.

$$
\begin{aligned}
\nabla^2_{\theta_l} \mathcal{L}(f_\theta) &= J_{\theta_l} f_\theta^\top \cdot \nabla^2_f \mathcal{L} \cdot J_{\theta_l} f_\theta \\
&= J_{\theta_l} f_{\theta_l}^\top \cdot J_x f_{\theta_{l+1}}^\top \cdots J_x f_{\theta_L}^\top \cdot \nabla^2_f \mathcal{L} \cdot J_x f_{\theta_L} \cdots J_x f_{\theta_{l+1}} \cdot J_{\theta_l} f_{\theta_l}
\end{aligned}
$$



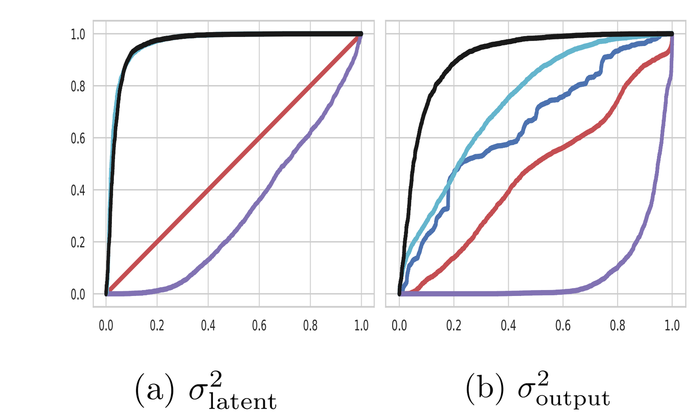
Hessian approximation





Current approximations scales quadratically with image resolution, where as our approximation scales linearly. This enables us to scale to large images.

Out-of-Distribution Detection



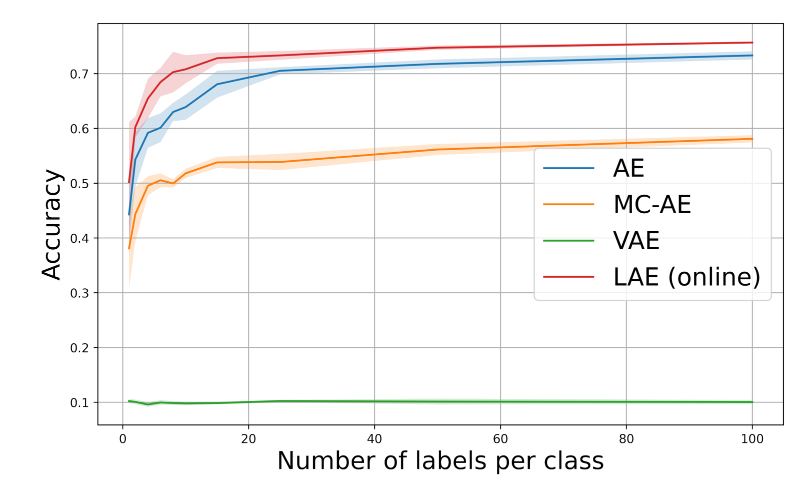
(a) σ^2_{latent} (b) σ^2_{output}

Method	$logp(x)$	Typicality	σ^2_{output}	σ^2_{latent}
AE MSE	**0.98**	**0.98**		
AE likelihood	0.66	0.70	0.65	
VAE	0.54	0.62	0.46	0.5
MC-AE	0.97	0.97	0.08	0.31
LAE (post-hoc)	**0.98**	**0.98**	0.74	**0.96**
LAE (online)	**0.98**	**0.98**	**0.91**	**0.96**

(c) AUROC (↑) for multiple statistics.

Our online training procedure produces reliable uncertainties in both latent and output space, which are useful for out-of-distribution detection.

Semi Supervised Learning



Attribute	AE	VAE	MC-AE	LAE (online)
Arched Eyebrows	0.50	0.52	0.55	0.60
Attractive	0.52	0.50	0.49	0.53
Bald	0.98	0.98	0.98	0.98
Wearing Lipstick	0.52	0.49	0.50	0.54
Heavy Makeup	0.45	0.52	0.49	0.56
Overall	0.73	0.72	0.73	**0.74**

In semi-supervised learning we have plentiful of unlabelled data (on which we train our model), and only few labelled data points. We can augment these with a stochastic feature representation, which leads to improve classification accuracy.

Data Imputation & Generative Capabilities

The VAE suffers from mode collapse, and does not capture the ambiguity in the input data. In contrast the LAE (online) correctly finds the multiple modes the input data could origin from.

(a) VAE (b) LAE (post-hoc) (c) LAE (online)