Variational Auto-Encoder:
Deep Learning meets Generative Modeling

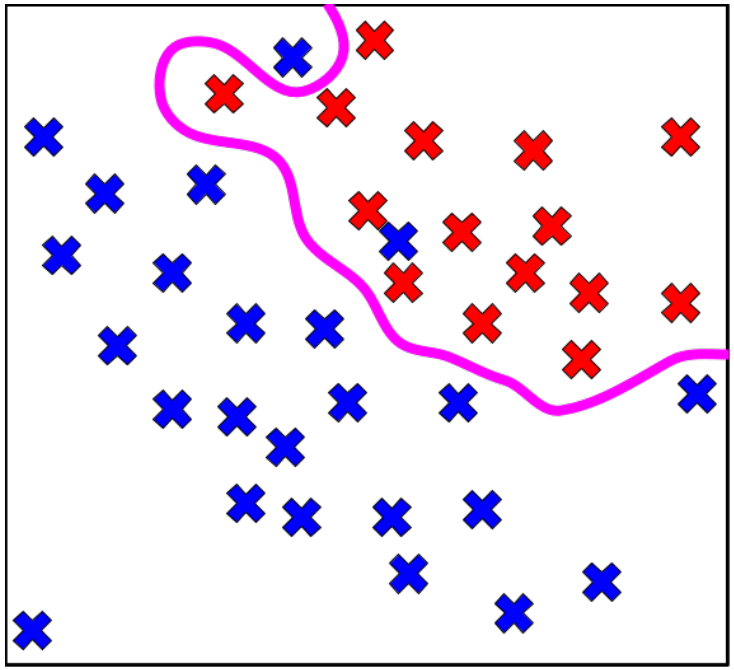
Jakub M. Tomczak
AMLAB, Universiteit van Amsterdam

Eindhoven 2017

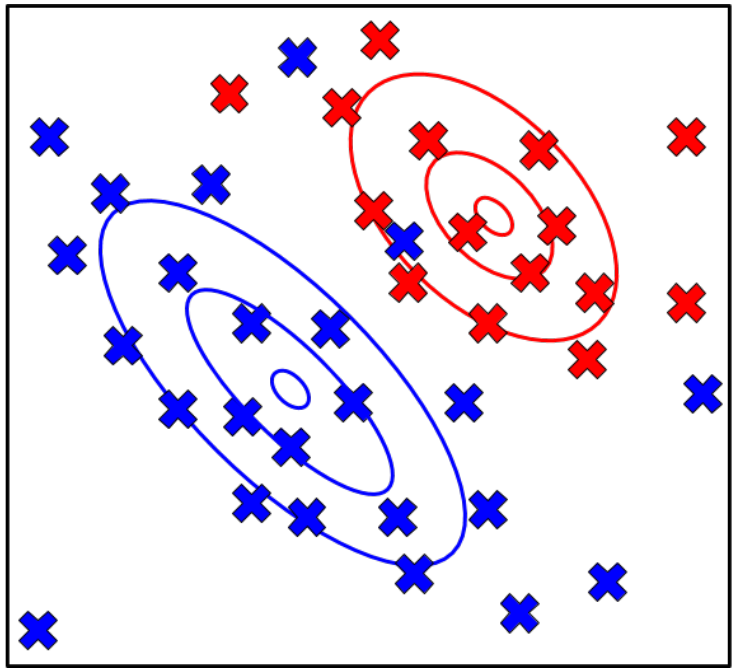
The presented research was funded by the European Commission within the Marie Skłodowska-Curie Individual Fellowship (Grant No. 702666, "*Deep learning and Bayesian inference for medical imaging*").

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Generative Modeling



$p(y|\mathbf{x})$



$p(\mathbf{x}, y)$

Generative Modeling

- Providing decision is not enough. *How to evaluate uncertainty? Distribution of y is only a part of the story.*

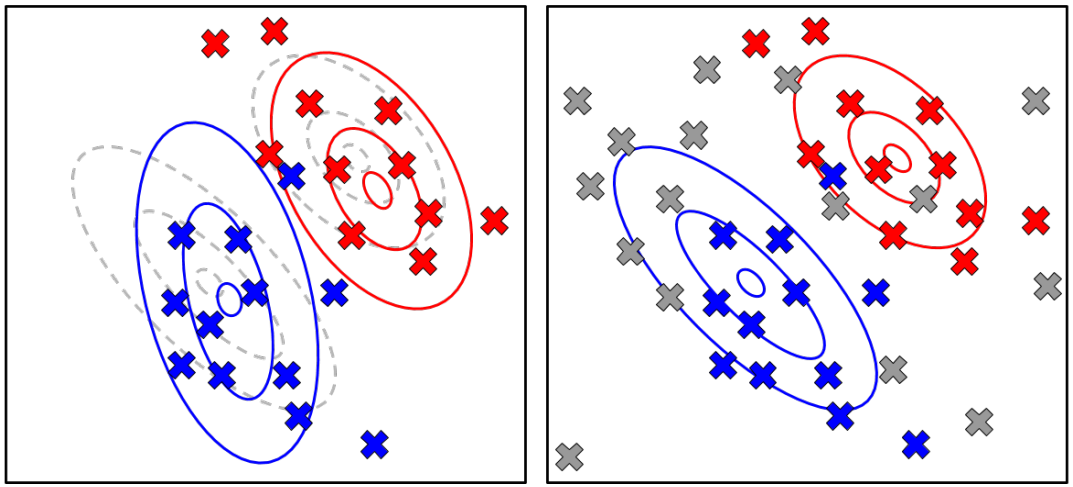
- Generalization problem. *Without knowing about distribution of* $\mathbf{x}$ *how we can generalize to new data?*

- Understanding the problem is crucial ("**What I cannot create, I do not understand**", Richard P. Feynman). *Knowing higher order statistics is essential to make better decisions.*

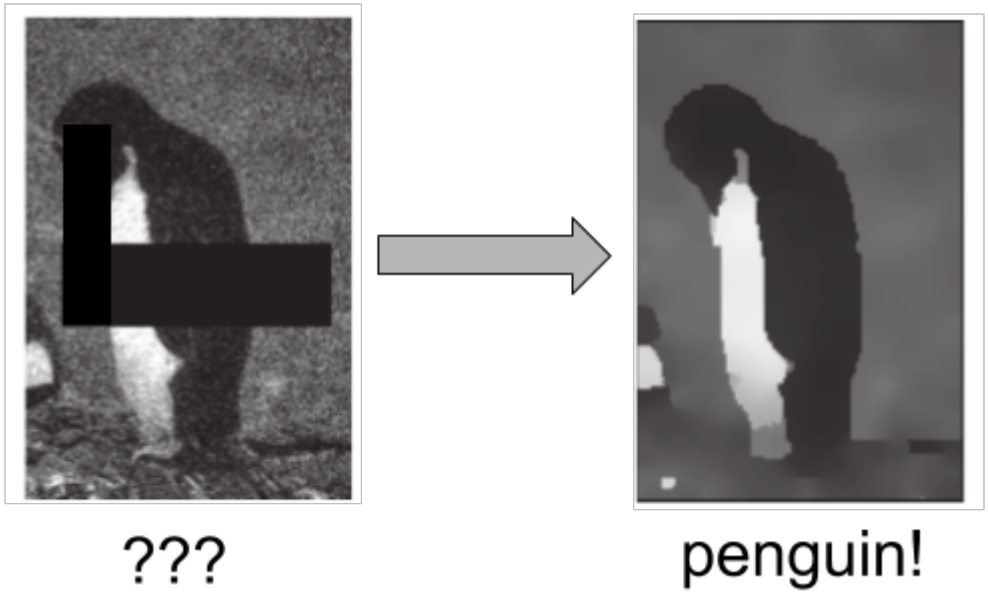
RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Generative Modeling

- Semi-supervised learning.
  Use unlabeled data to train a better classifier.

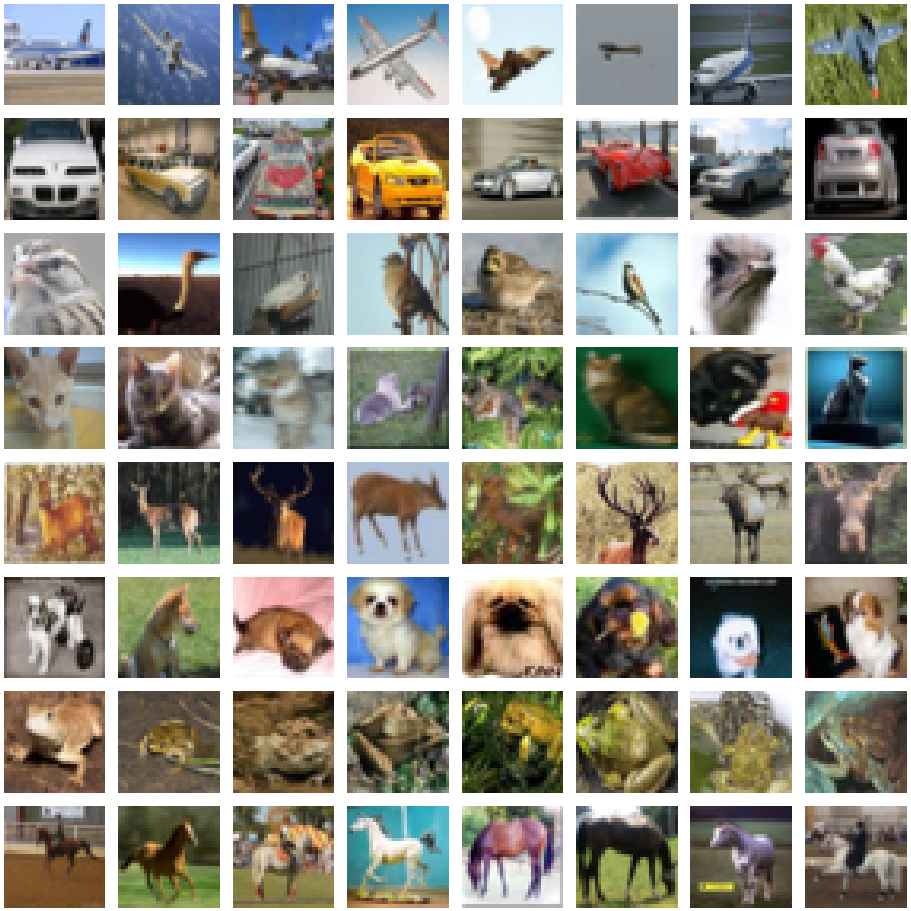


- Handling missing or distorted data.
  Reconstruct and/or denoise data.

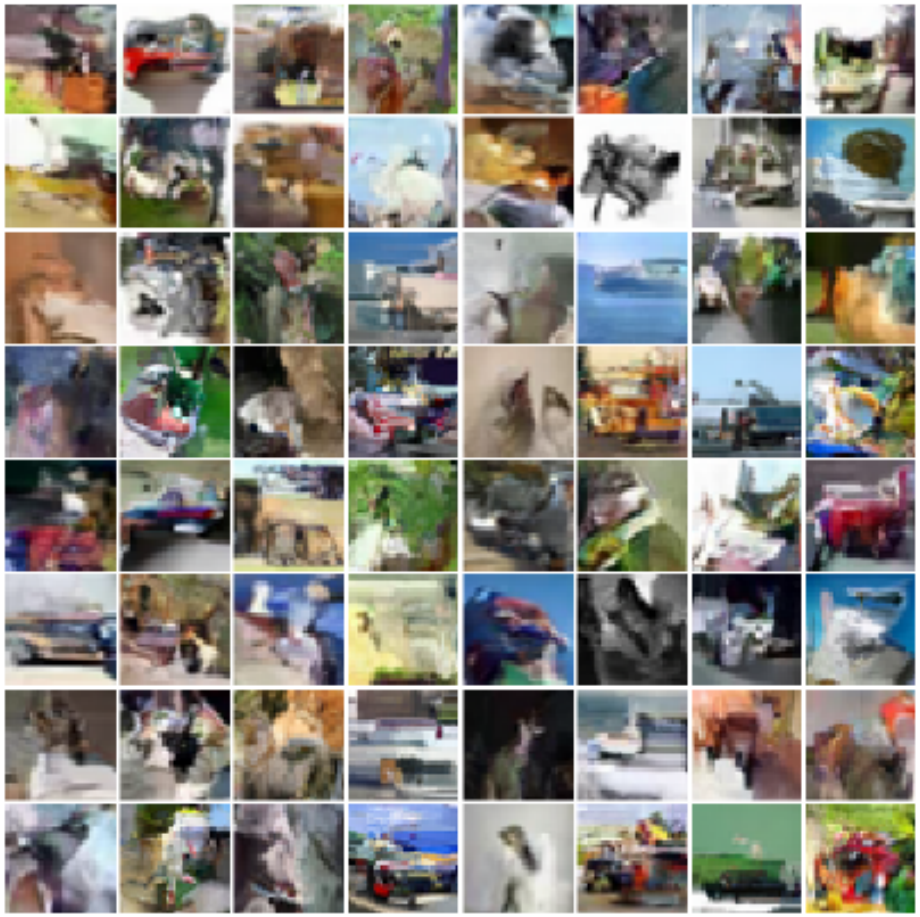


Generative Modeling

Image generation



Real



Generated

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Generative Modeling

Sequence generation

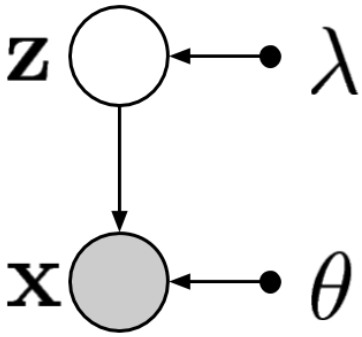
he had been unable to conceal the fact that there was a logical explanation for his inability to alter the fact that they were supposed to be on the other side of the house .
with a variety of pots strewn scattered across the vast expanse of the high ceiling , a vase of colorful flowers adorned the tops of the rose petals littered the floor and littered the floor .
atop the circular dais perched atop the gleaming marble columns began to emerge from atop the stone dais, perched atop the dais .

Generated

Latent Generative Models

- Latent model:

$$p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$



- If $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mathbf{b}, \Psi)$ and $p_\lambda(\mathbf{z}) = \mathcal{N}(\mu_0, \Sigma_0)$, then $\rightarrow$ **Factor Analysis**.

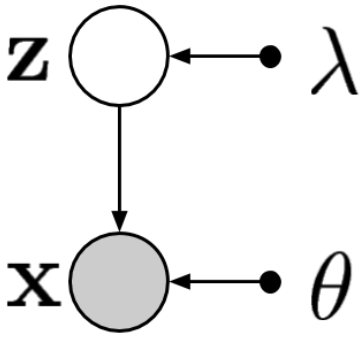
- What if we take a **non-linear transformation** of z?
  $\rightarrow$ ***an infinite mixture of Gaussians.***

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Latent Generative Models

- Latent model:

$$p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$



- If $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mathbf{b}, \Psi)$ and $p_\lambda(\mathbf{z}) = \mathcal{N}(\mu_0, \Sigma_0)$, then → **Factor Analysis**.

Convenient but limiting!

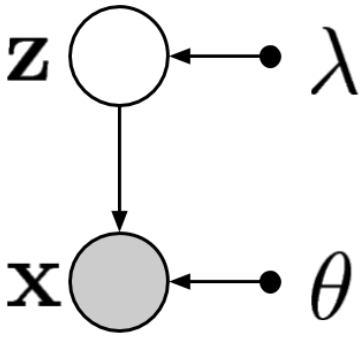
- What if we take a **non-linear transformation** of z? → *an infinite mixture of Gaussians.*

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Latent Generative Models

- Latent model:

$$p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$



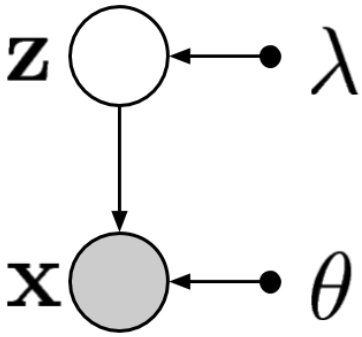
- If $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mathbf{b}, \Psi)$ and $p_\lambda(\mathbf{z}) = \mathcal{N}(\mu_0, \Sigma_0)$, then $\rightarrow$ **Factor Analysis**.

- What if we take a **non-linear transformation** of z? $\rightarrow$ *an infinite mixture of Gaussians.*

Latent Generative Models

- Latent model:

$$p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

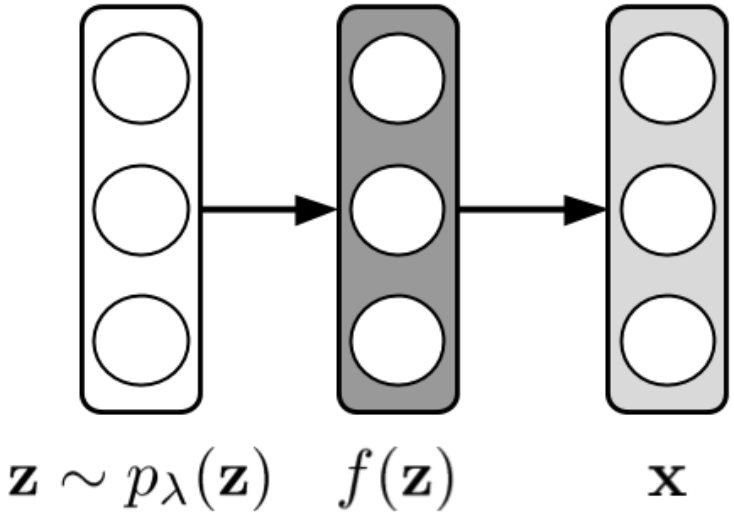


- If $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mathbf{b}, \Psi)$ and $p_\lambda(\mathbf{z}) = \mathcal{N}(\mu_0, \Sigma_0)$, then → **Factor Analysis**.

- What if we take a **non-linear transformation** of z?
  → *an infinite mixture of Gaussians.*

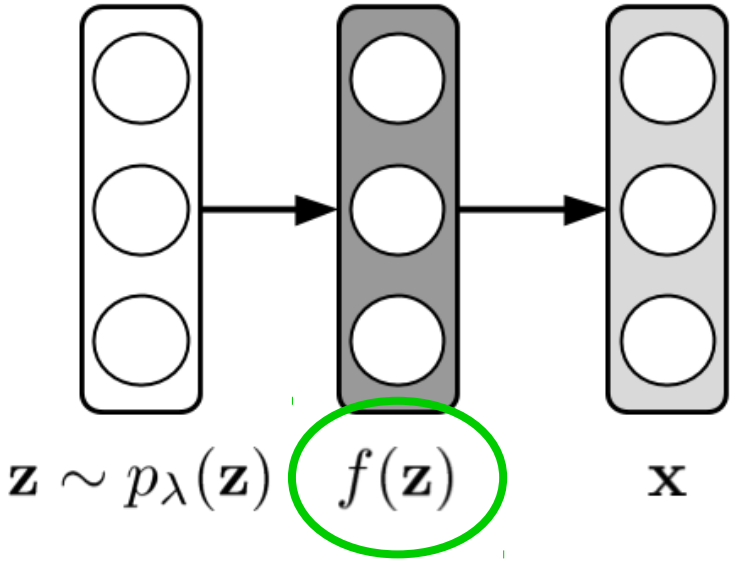
Neural network

Deep Generative Models (DGM): Density Network



MacKay, D. J., & Gibbs, M. N. (1999). *Density networks*. Statistics and neural networks: advances at the interface. Oxford University Press, Oxford, 129-144.

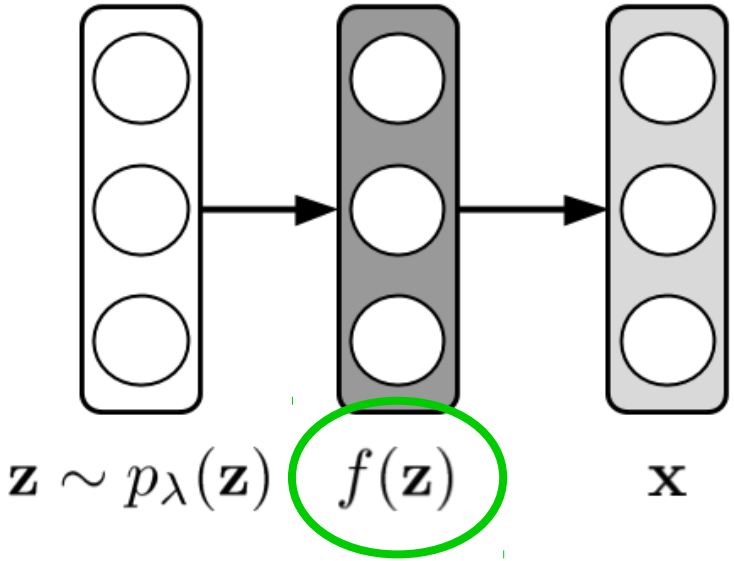
DGM: Density Network



Neural Network

MacKay, D. J., & Gibbs, M. N. (1999). *Density networks*. Statistics and neural networks: advances at the interface. Oxford University Press, Oxford, 129-144.

DGM: Density Network



Neural Network

How to train this model?!

MacKay, D. J., & Gibbs, M. N. (1999). *Density networks*. Statistics and neural networks: advances at the interface. Oxford University Press, Oxford, 129-144.

DGM: Density Network

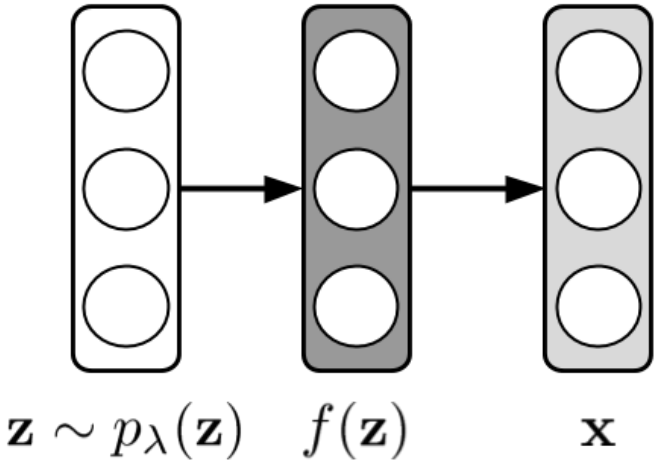
- MC approximation:

$$\log p(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$\approx \log \frac{1}{S} \sum_{s=1}^{S} \exp\Big(\log p_\theta(\mathbf{x}|\mathbf{z}_s)\Big)$$

where:

$$\mathbf{z}_s \sim p_\lambda(\mathbf{z})$$



MacKay, D. J., & Gibbs, M. N. (1999). *Density networks*. Statistics and neural networks: advances at the interface. Oxford University Press, Oxford, 129-144.

DGM: Density Network

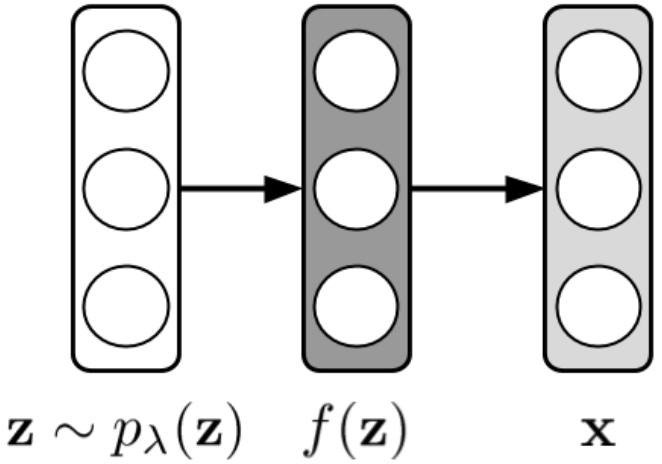
- MC approximation:

$$\log p(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$\approx \log \frac{1}{S} \sum_{s=1}^{S} \exp\left(\log p_\theta(\mathbf{x}|\mathbf{z}_s) \right)$$

where:

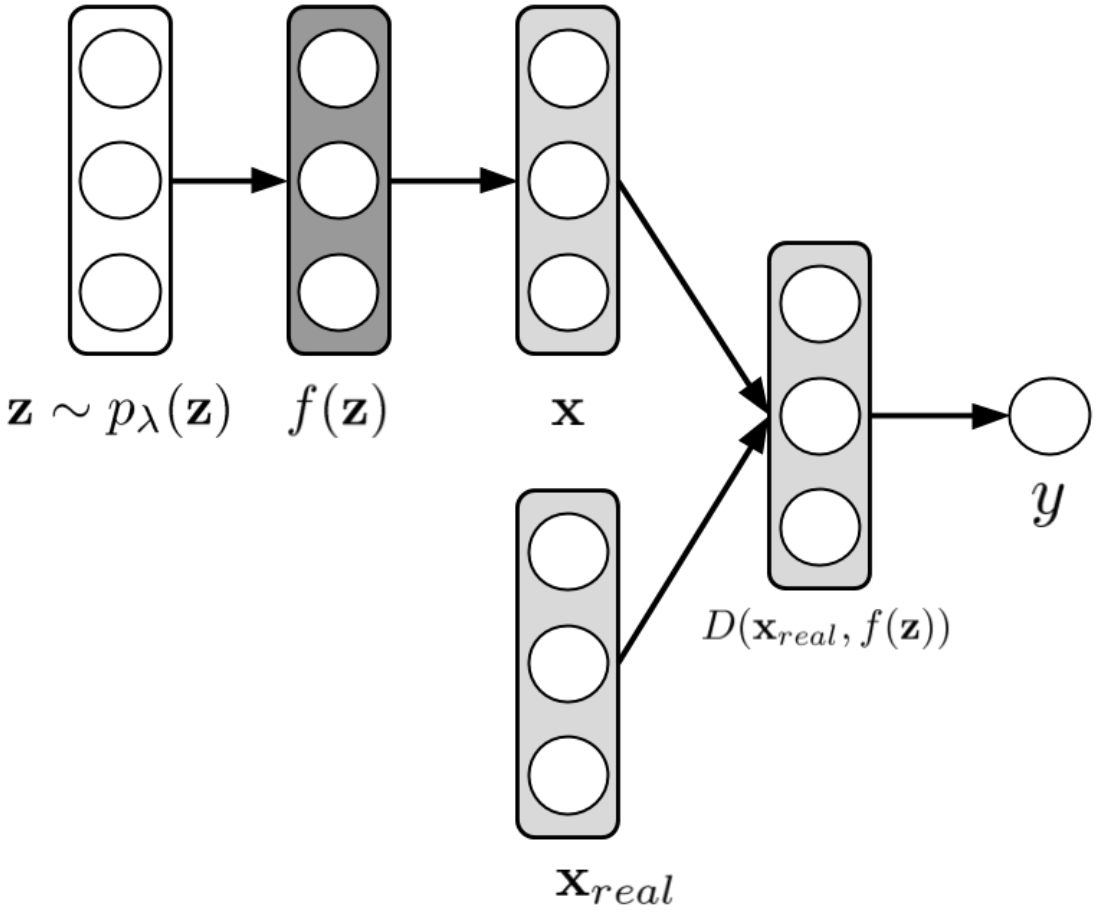
$$\mathbf{z}_s \sim p_\lambda(\mathbf{z})$$



Works only in a low-dimensional case!

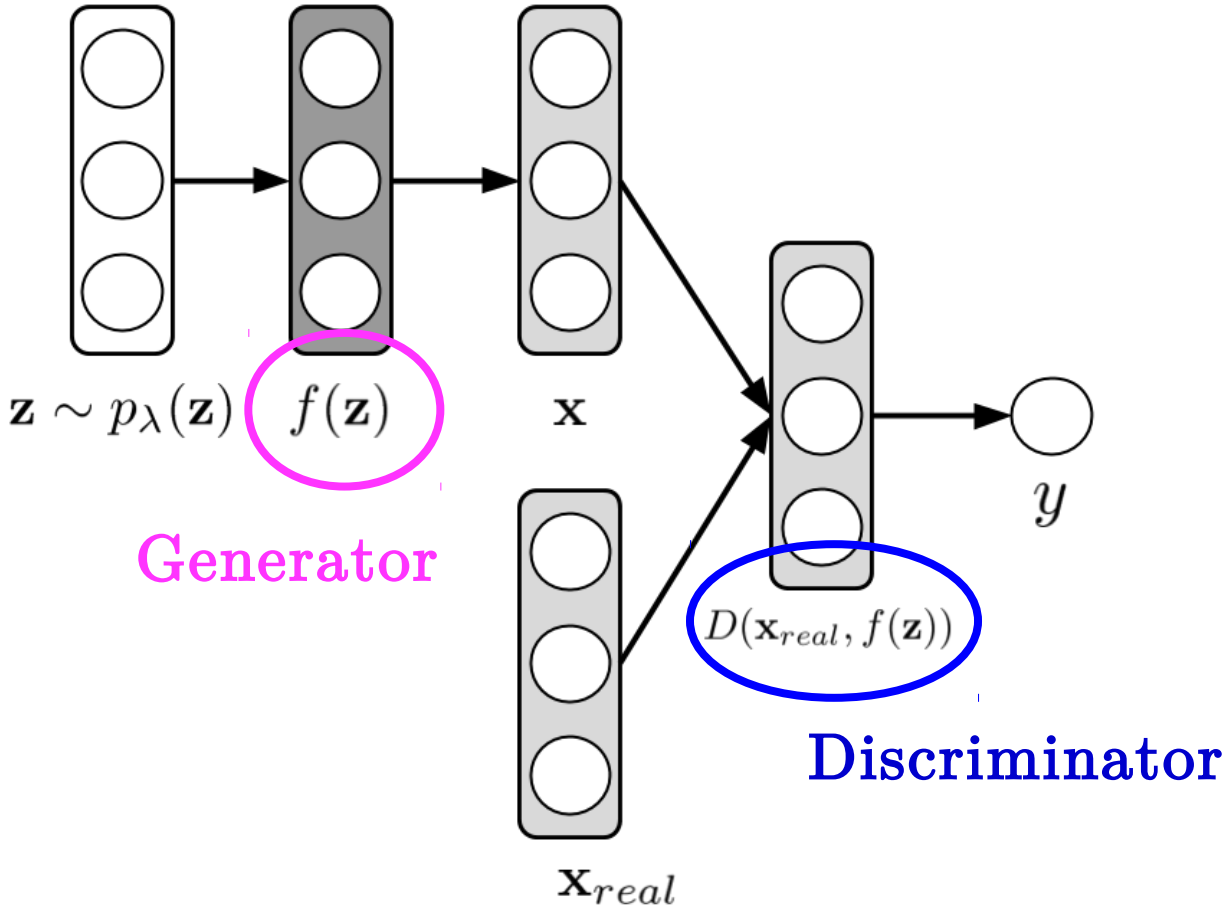
MacKay, D. J., & Gibbs, M. N. (1999). *Density networks*. Statistics and neural networks: advances at the interface. Oxford University Press, Oxford, 129-144.

DGM: Generative Adversarial Nets



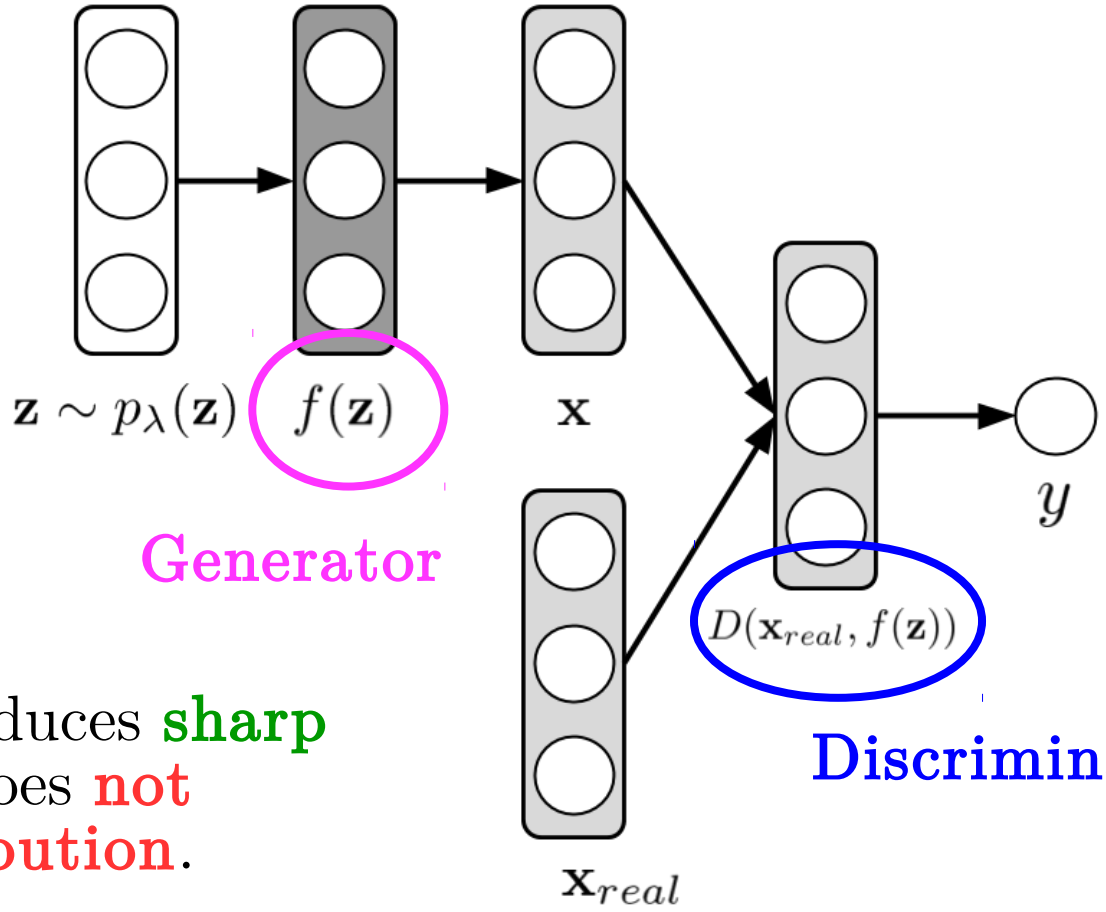
Goodfellow, I., et al. (2014). *Generative adversarial nets*. In Advances in neural information processing systems

DGM: Generative Adversarial Nets



Goodfellow, I., et al. (2014). *Generative adversarial nets*. In Advances in neural information processing systems

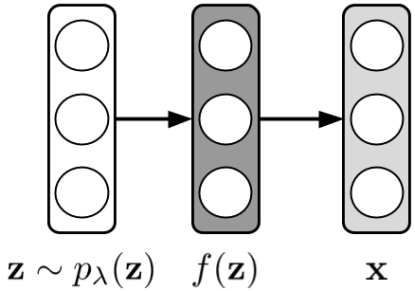
DGM: Generative Adversarial Nets



Scalable, produces sharp images but does not train a distribution.

Goodfellow, I., et al. (2014). *Generative adversarial nets*. In Advances in neural information processing systems

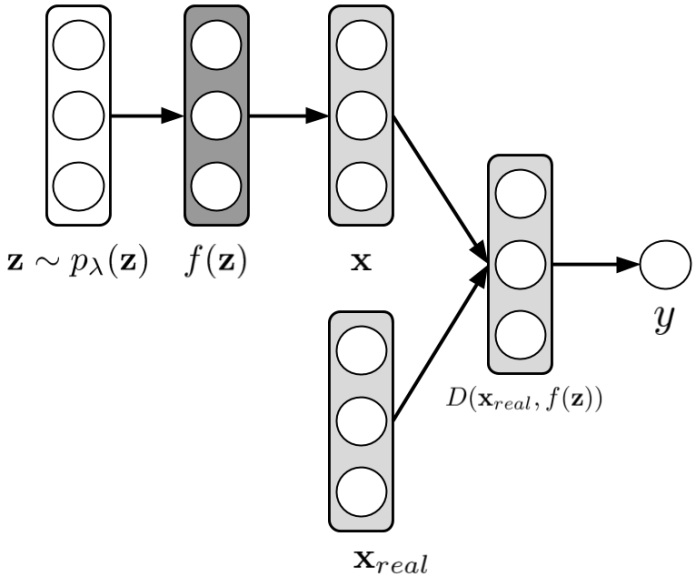
DGM: so far we have



Density Network

Works only for low dimensions.

Doesn't train x → z.



Generative Adversarial Net

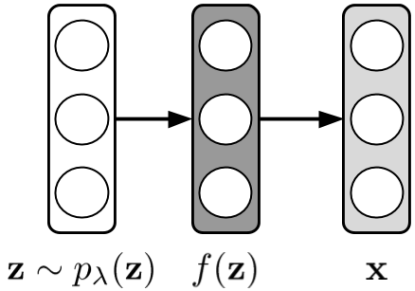
Works for high dimensions.

Doesn't train a distribution...

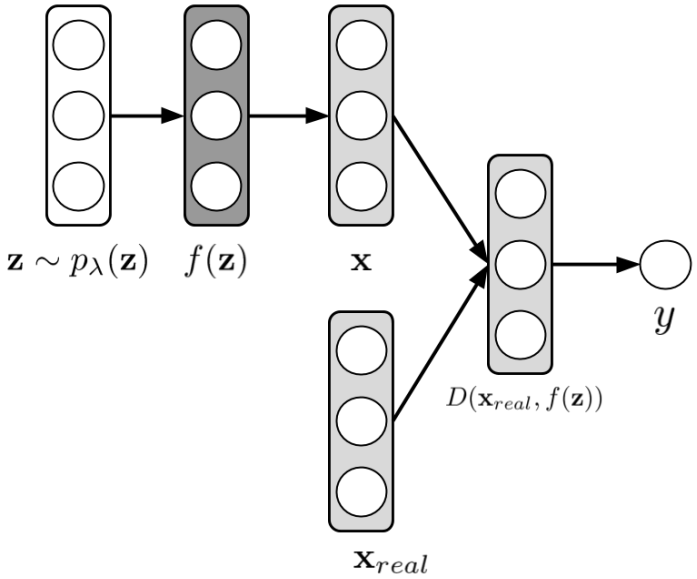
Doesn't train x → z.

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

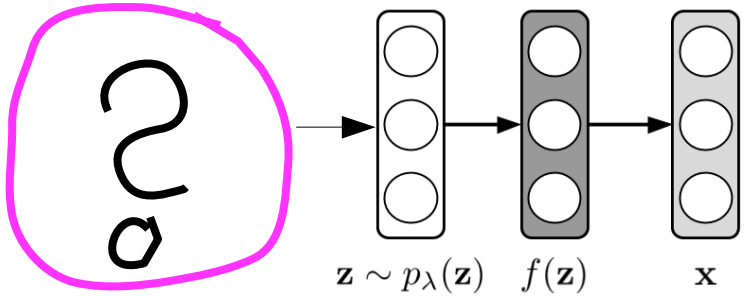
DGM: so far we have



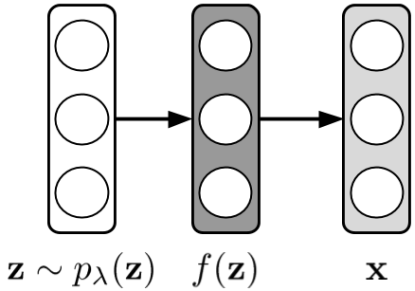
Density Network



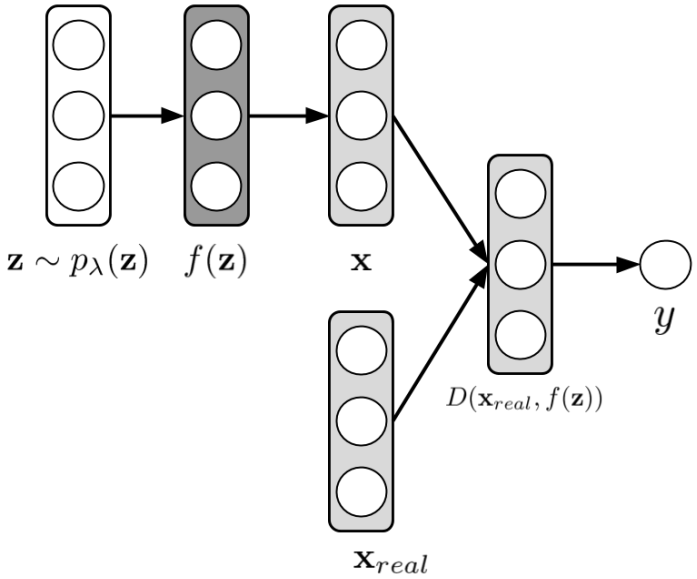
Generative Adversarial Net



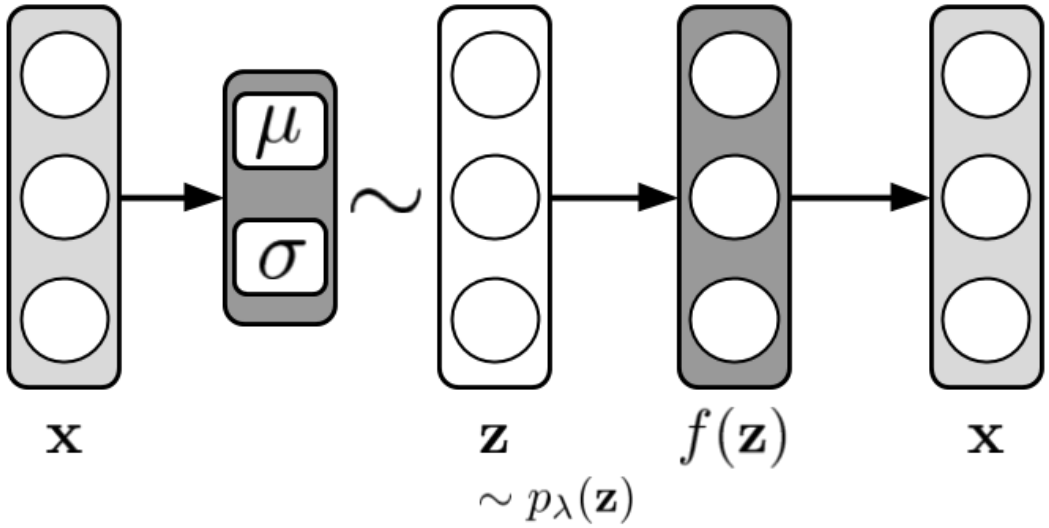
DGM: so far we have



Density Network

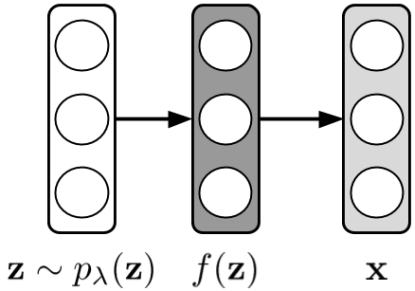


Generative Adversarial Net

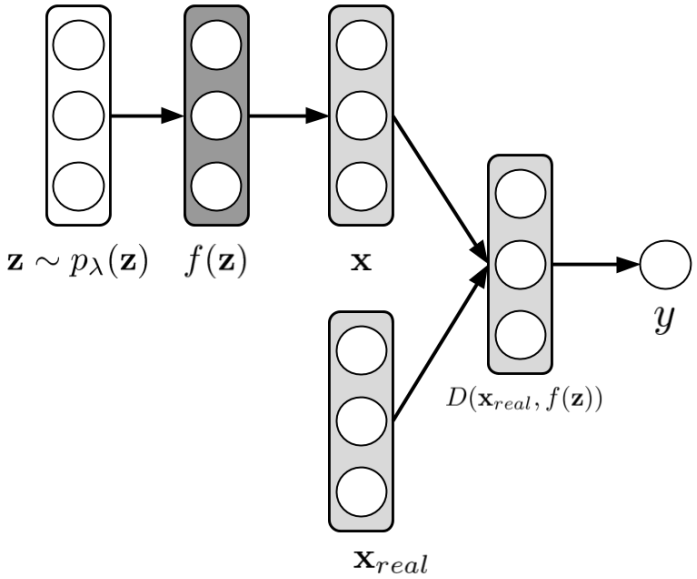


Variational Auto-Encoder

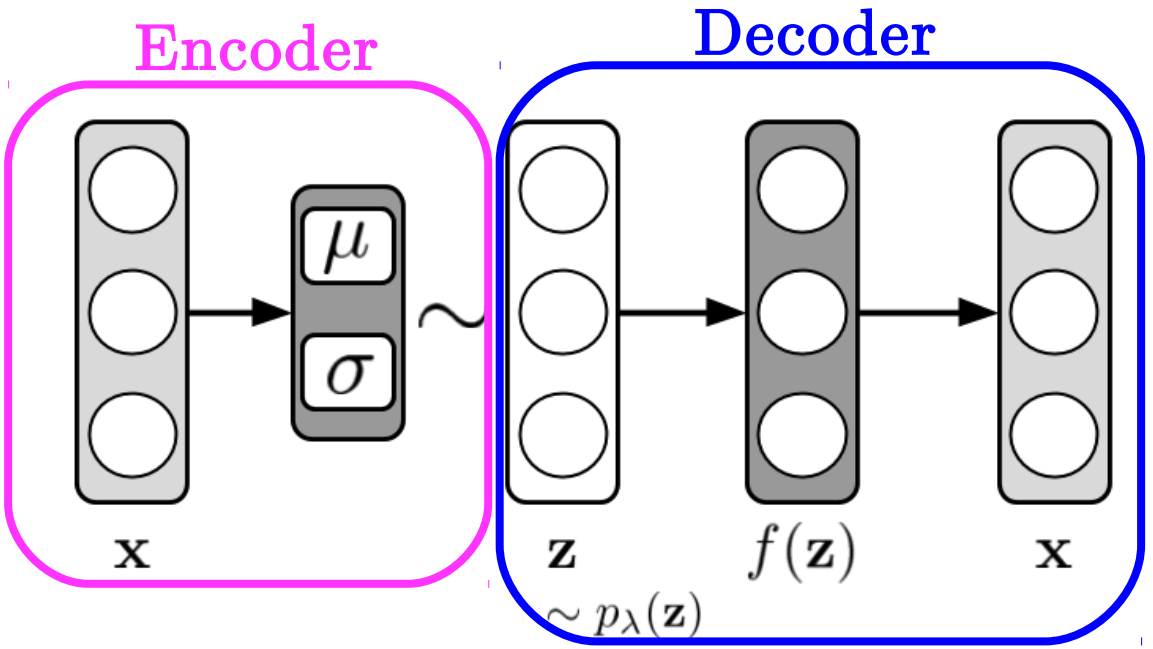
DGM: so far we have



Density Network

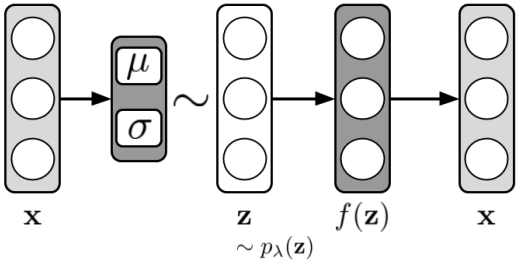


Generative Adversarial Net



Variational Auto-Encoder

DGM: Variational Auto-Encoder



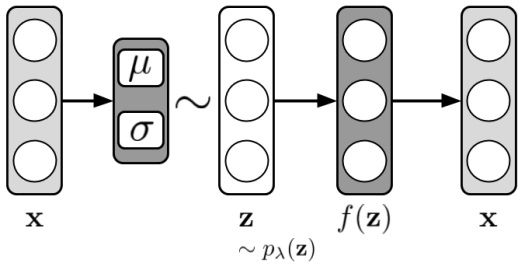
$$\log p(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]$$

DGM: Variational Auto-Encoder



$$\log p(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

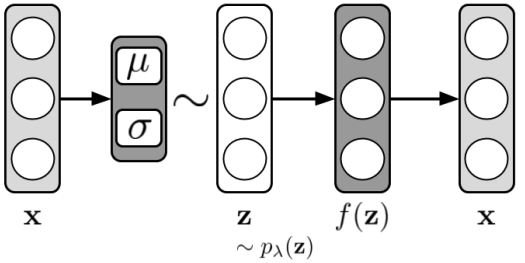
Variational posterior

$$= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]$$

DGM: Variational Auto-Encoder



$$\log p(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z}$$

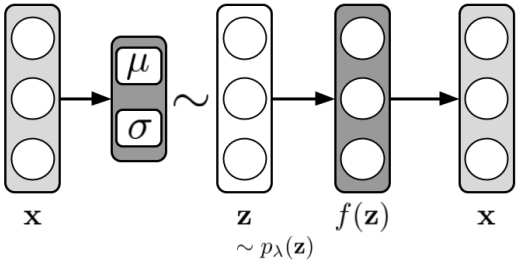
$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]$$

Reconstruction error

Regularization

DGM: Variational Auto-Encoder

We approximate the objective (**the evidence lower bound**) using samples of $\mathbf{z}$.



PROBLEM: calculating gradient wrt parameters of the variational posterior (*i.e.*, sampling process).

SOLUTION: use non-centered parameterization (a.k.a. ***reparameterization trick***).

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu, \sigma^2)$$

$$\mathbf{z}_s = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

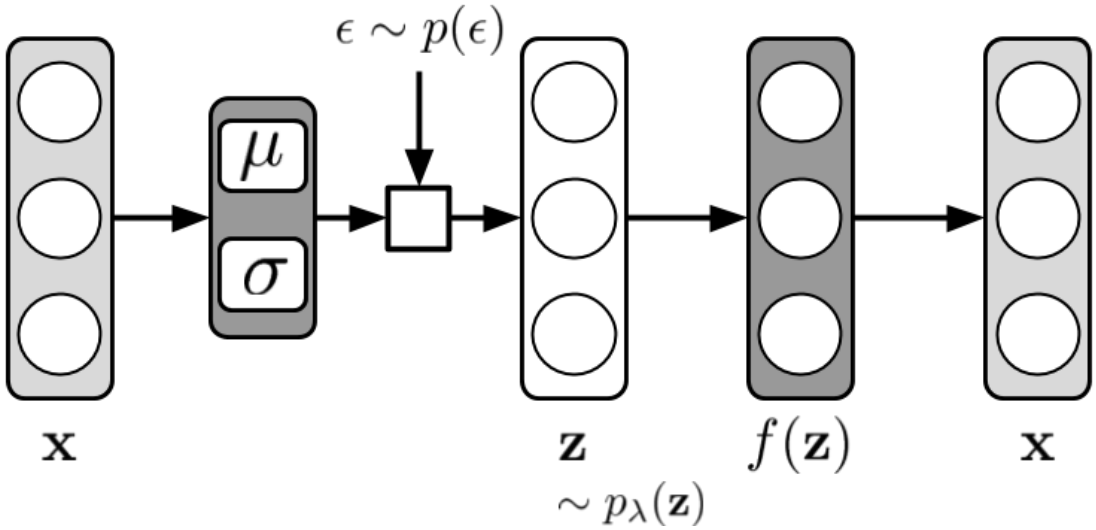
DGM: Variational Auto-Encoder

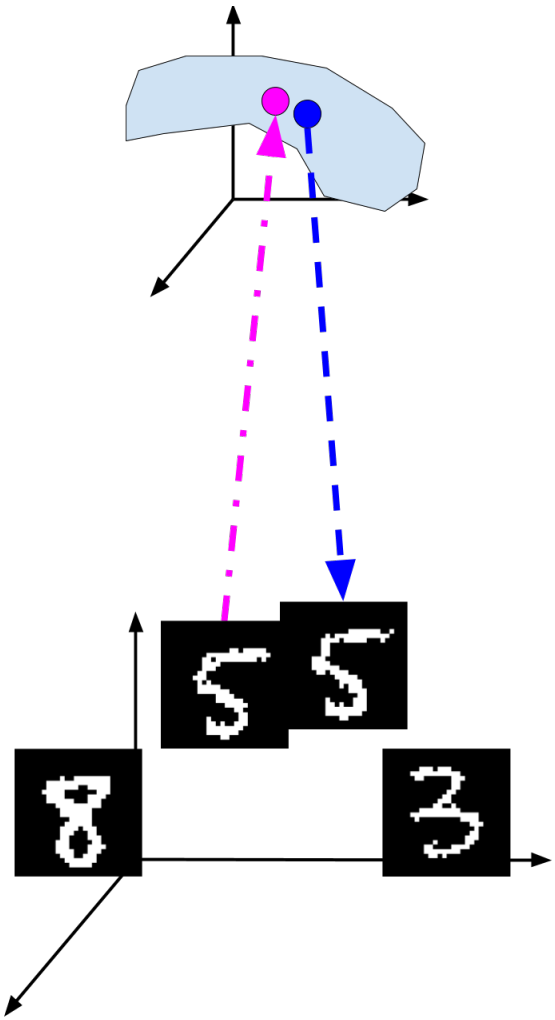
We approximate the objective (**the evidence lower bound**) using samples of $\mathbf{z}$.

$$\log p(\mathbf{x}) \approx \frac{1}{S} \sum_{s=1}^{S} \Big(\log p_\theta(\mathbf{x}|\mathbf{z}_s) - \log q_\phi(\mathbf{z}_s|\mathbf{x}) + \log p(\mathbf{z}_s) \Big)$$

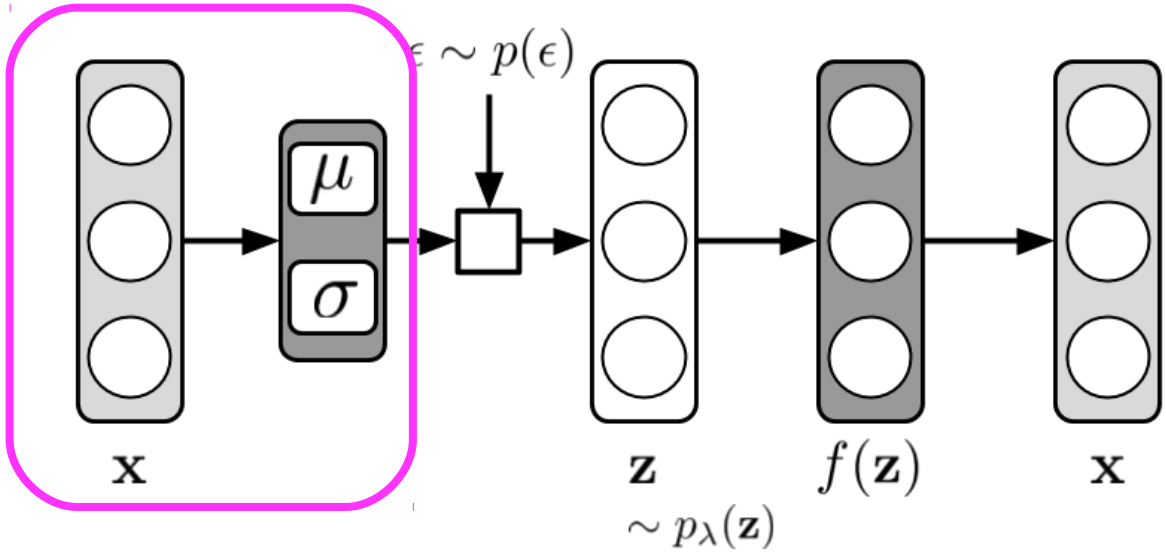
$$\text{where: } \mathbf{z}_s = \mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \epsilon_s$$

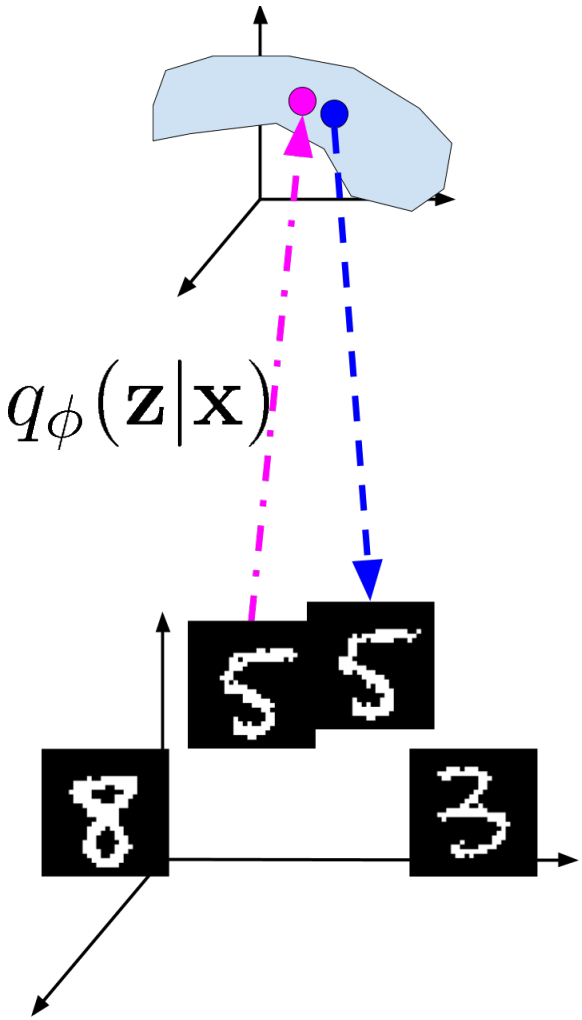
DGM: Variational Auto-Encoder





DGM: Variational Auto-Encoder

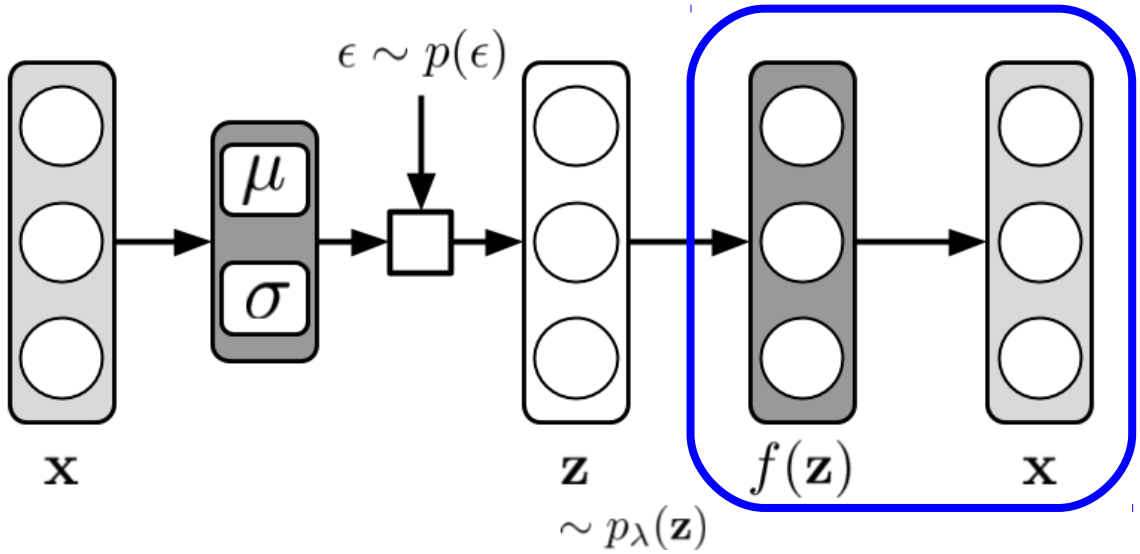


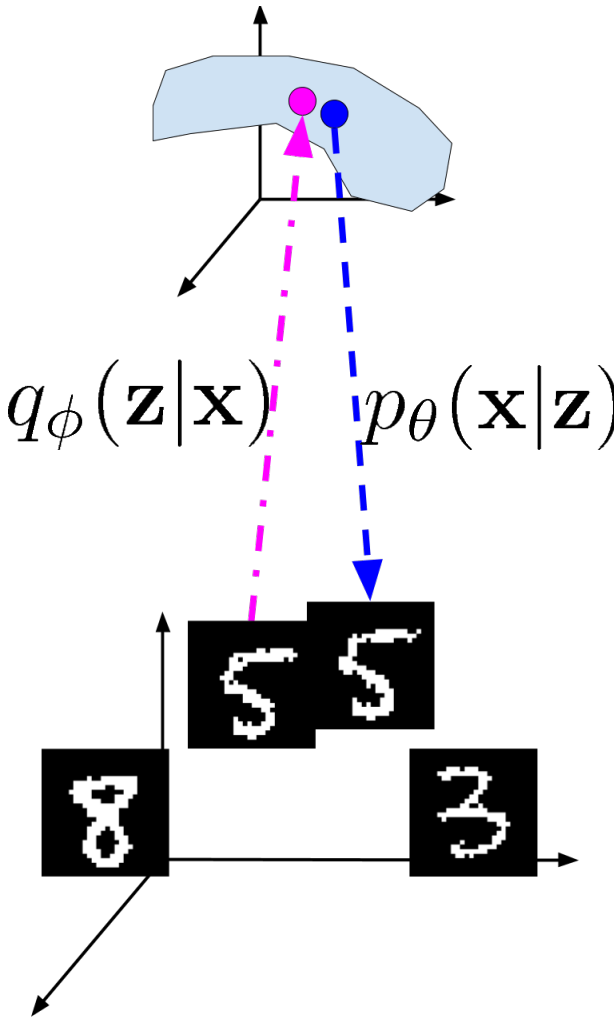


A **deep neural net** that outputs parameters of the variational posterior (**encoder**):

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu(\mathbf{x}), \text{diag}\{\sigma^2(\mathbf{x})\})$$

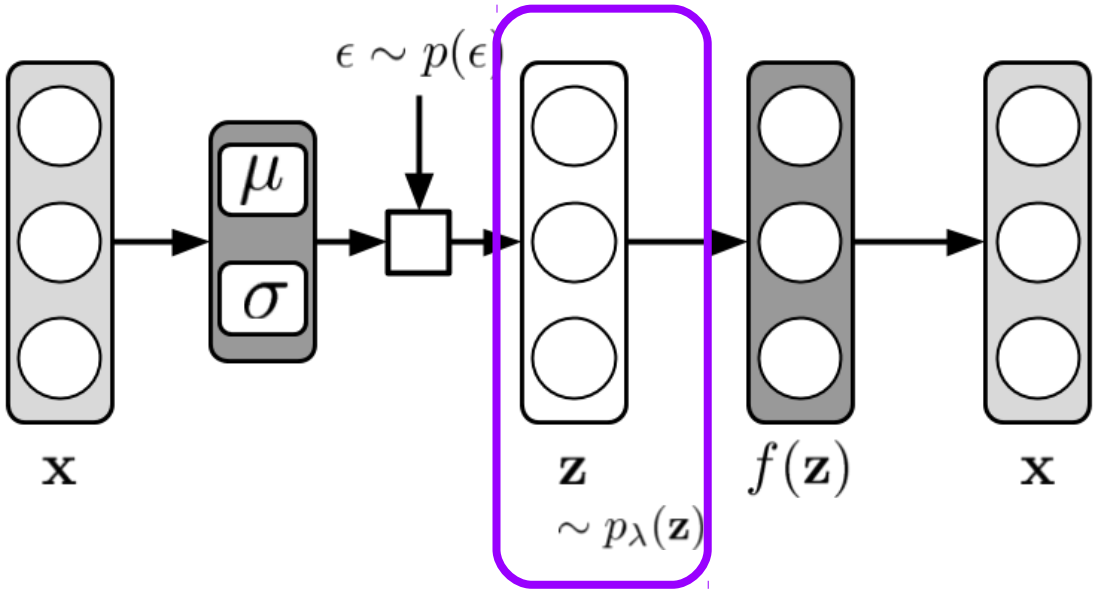
DGM: Variational Auto-Encoder

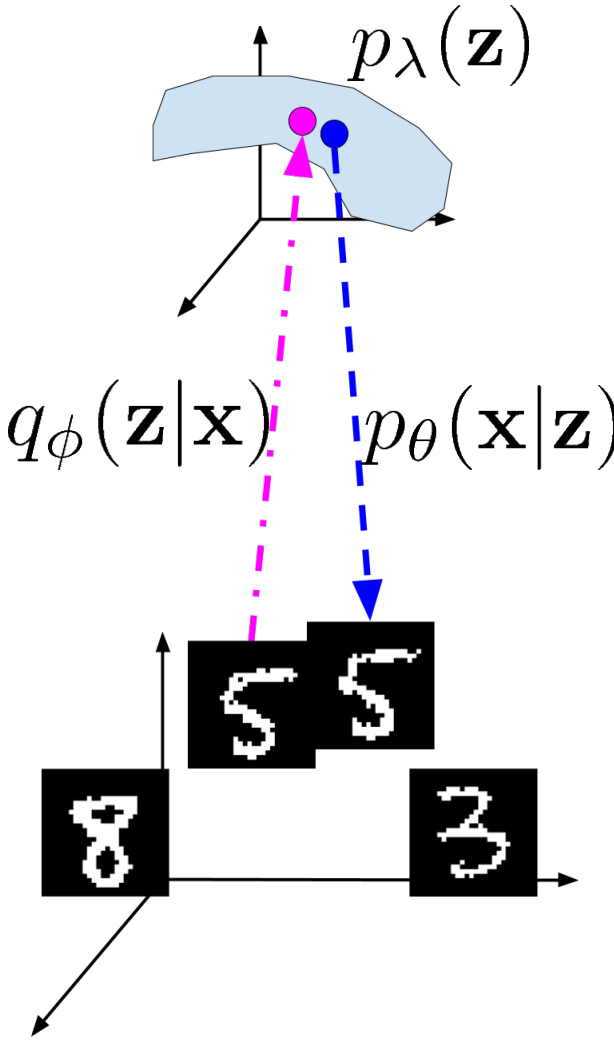




A **deep neural net** that outputs parameters of the generator (**decoder**), *e.g.*, a normal distribution or Bernoulli distribution.

DGM: Variational Auto-Encoder



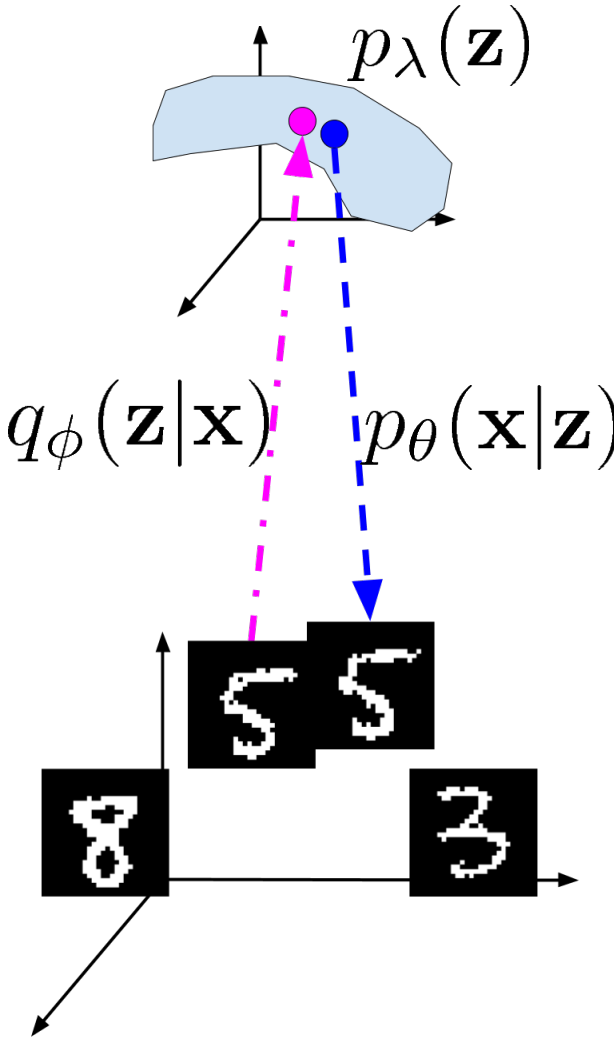


A **prior** that **regularizes** the encoder and takes part in the **generative process**.

$$p_\lambda(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I})$$

DGM: Variational Auto-Encoder

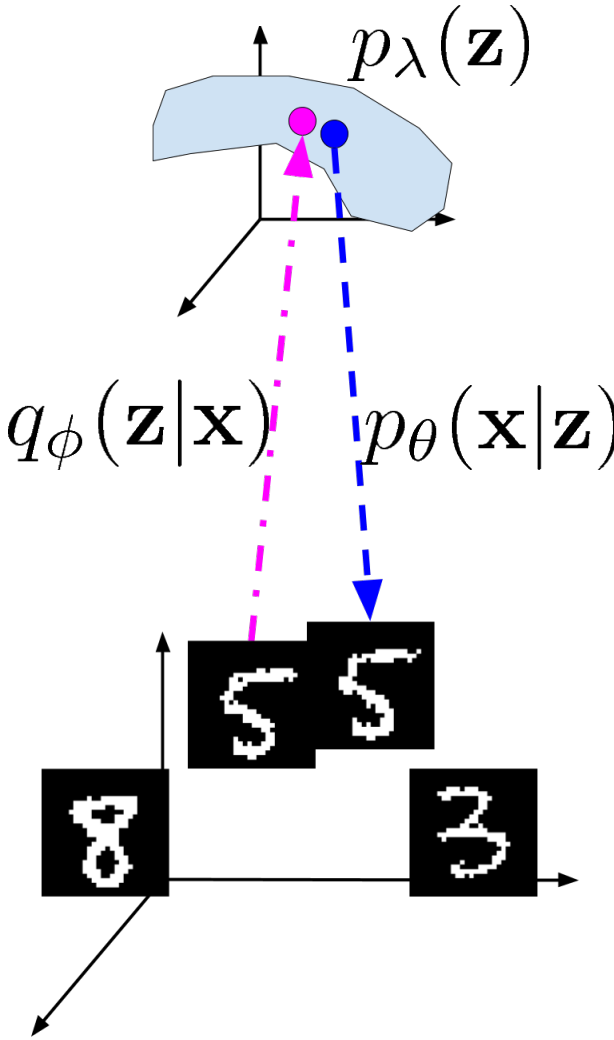
$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})$$



DGM: Variational Auto-Encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Feedforward nets
Convolutional nets
PixelCNN
Gated PixelCNN

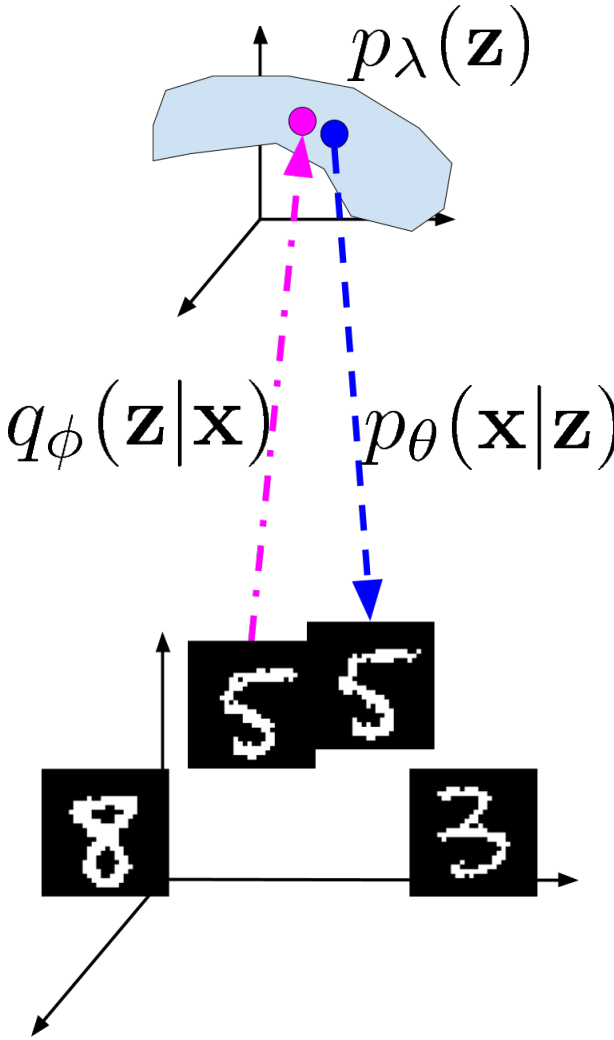


DGM: Variational Auto-Encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows
Gaussian processes
Stein Particle Descent
Operator VI

Feedforward nets
Convolutional nets
PixelCNN
Gated PixelCNN



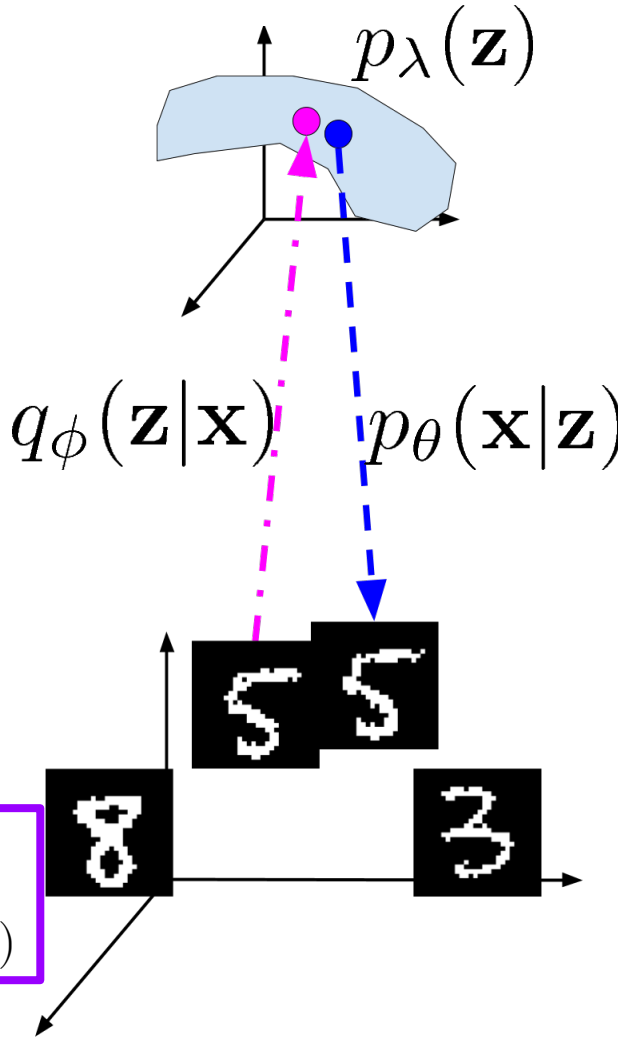
DGM: Variational Auto-Encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows
Gaussian processes
Stein Particle Descent
Operator VI

Feedforward nets
Convolutional nets
PixelCNN
Gated PixelCNN

Auto-regressive Prior (Chen et al.)
Objective Prior (Nalisnick & Smyth)
Stick-Breaking Prior (Nalisnick & Smyth)



DGM: Variational Auto-Encoder

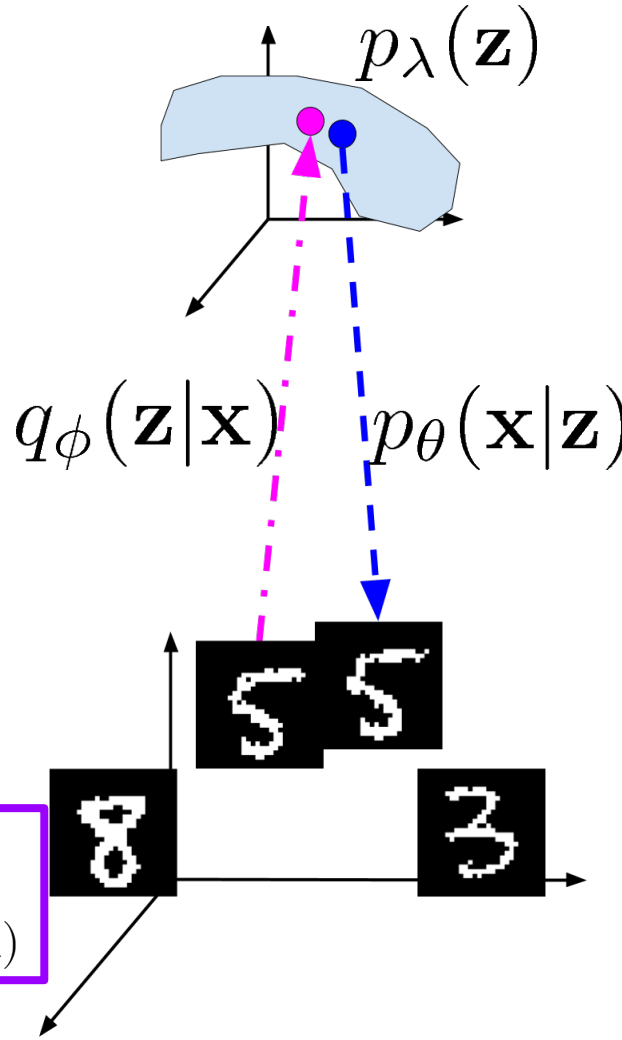
$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows
Gaussian processes
Stein Particle Descent
Operator VI

Feedforward nets
Convolutional nets
PixelCNN
Gated PixelCNN

Auto-regressive Prior (Chen et al.)
Objective Prior (Nalisnick & Smyth)
Stick-Breaking Prior (Nalisnick & Smyth)

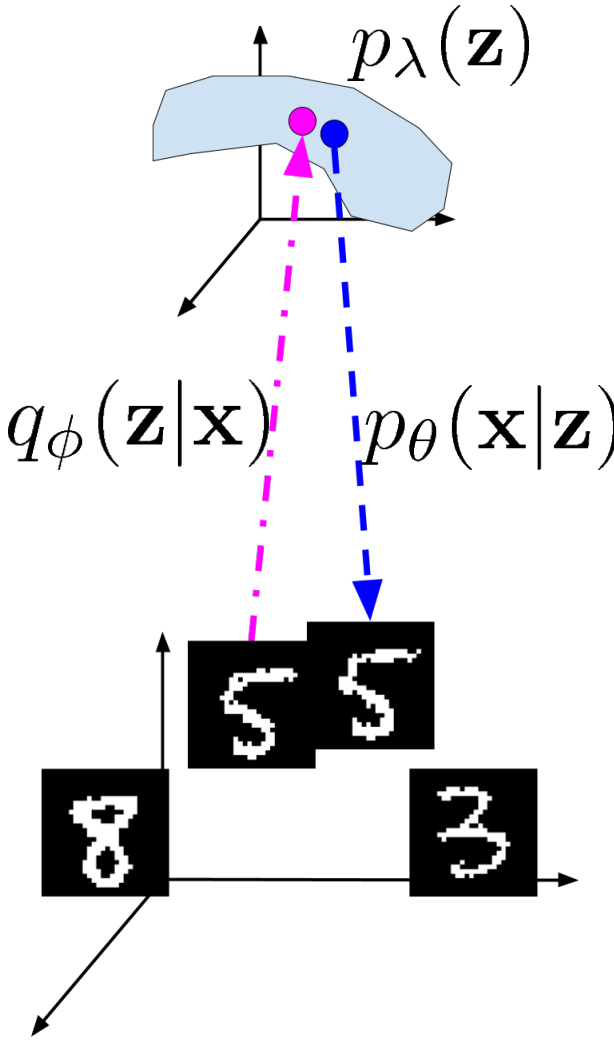
Importance Weighted AE
Renyi Divergence
Stein Divergence



RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Improving the posterior

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})$$



Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?

  $\rightarrow$ apply a series of T **invertible transformations**

  $\mathbf{f}^{(t)}$ to $\mathbf{z}^{(0)} \sim q(\mathbf{z}|\mathbf{x})$.

- **New objective:**

$$\ln p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}^{(0)}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z}^{(T)}) + \sum_{t=1}^{T} \ln \left|\det \frac{\partial \mathbf{f}^{(t)}}{\partial \mathbf{z}^{(t-1)}}\right|\right] - \mathrm{KL}\left(q(\mathbf{z}^{(0)}|\mathbf{x})||p(\mathbf{z}^{(T)})\right).$$

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?

  → apply a series of T invertible transformations $\mathbf{f}^{(t)}$ to $\mathbf{z}^{(0)} \sim q(\mathbf{z}|\mathbf{x})$.

- New objective:

$$\ln p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}^{(0)}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z}^{(T)}\right.$$

$$q(\mathbf{z}^{(T)}|\mathbf{x}) = q(\mathbf{z}^{(0)}|\mathbf{x}) \prod_{t=1}^{T} \left| \det \frac{\partial f^{(t)}}{\partial \mathbf{z}^{(t-1)}} \right|^{-1}.$$

Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?

  $\rightarrow$ apply a series of T **invertible transformations** $\mathbf{f}^{(t)}$ to $\mathbf{z}^{(0)} \sim q(\mathbf{z}|\mathbf{x})$.

- **New objective**:

$$\ln p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}^{(0)}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z}^{(T)}) + \sum_{t=1}^{T} \ln \left|\det \frac{\partial \mathbf{f}^{(t)}}{\partial \mathbf{z}^{(t-1)}}\right|\right] - \mathrm{KL}\left(q(\mathbf{z}^{(0)}|\mathbf{x})||p(\mathbf{z}^{(T)})\right).$$

Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770. ICML 2015

Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?

  $\rightarrow$ apply a series of T **invertible transformations** $\mathbf{f}^{(t)}$ to $\mathbf{z}^{(0)} \sim q(\mathbf{z}|\mathbf{x})$.

- **New objective**:

$$\ln p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}^{(0)}|\mathbf{x})}\Big[\ln p(\mathbf{x}|\mathbf{z}^{(T)}) + \sum_{t=1}^{T} \ln \left|\det \frac{\partial \mathbf{f}^{(t)}}{\partial \mathbf{z}^{(t-1)}}\right|\Big] - \mathrm{KL}\big(q(\mathbf{z}^{(0)}|\mathbf{x})||p(\mathbf{z}^{(T)})\big).$$

Jacobian-determinant: (i) general normalizing flow (|det J| is ***easy*** to compute)

(ii) **volume-preserving flow**, *i.e.*, |det J|=1

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Volume-preserving flows

- How to obtain more flexible posterior and preserve $|\det J|=1$?

- Model **full-covariance matrix**:

  - using ***orthogonal matrices*** → Householder flow

    Tomczak, J. M., & Welling, M. (2016). Improving Variational Inference with Householder Flow. arXiv preprint arXiv:1611.09630. NIPS Workshop on Bayesian Deep Learning 2016

  - using ***lower-triangular matrix*** (with ones on the diagonal) → **linear Inverse Autoregressive Flow**

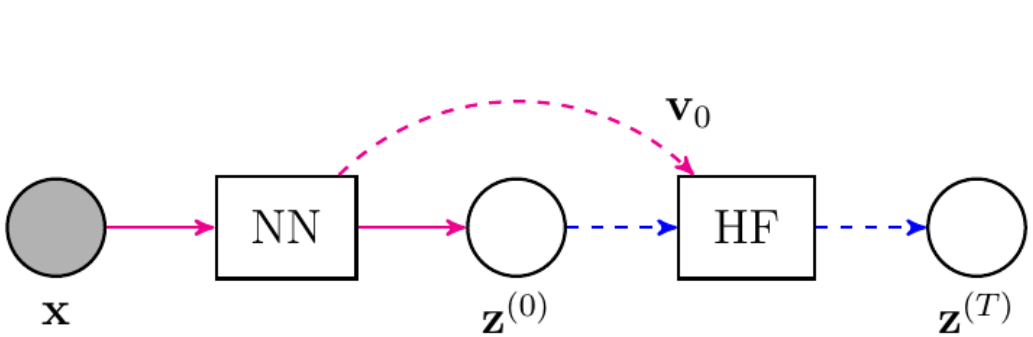
    Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improving Variational Inference with Inverse Autoregressive Flow. NIPS 2016

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

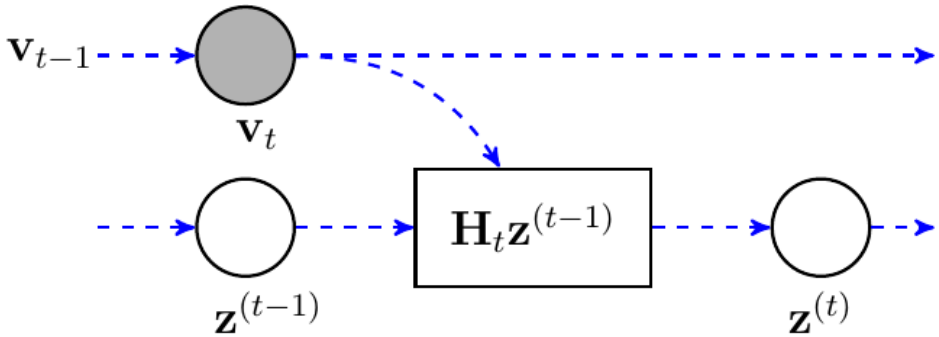
Householder Flow

- In the Householder transformation we reflect a vector around a hyperplane defined by a ***Householder vector*** $\mathbf{v}_t \in \mathbb{R}^M$:

$$\mathbf{z}^{(t)} = \underbrace{\left(\mathbf{I} - 2\frac{\mathbf{v}_t\mathbf{v}_t^\top}{||\mathbf{v}_t||^2}\right)}_{Householder\ matrix} \mathbf{z}^{(t-1)} = \mathbf{H}_t\ \mathbf{z}^{(t-1)}.$$



(a) encoder network + Householder Flow



(b) one step of the Householder Flow

Tomczak, J. M., & Welling, M. (2016). Improving Variational Inference with Householder Flow. NIPS Workshop on Bayesian Deep Learning 2016

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

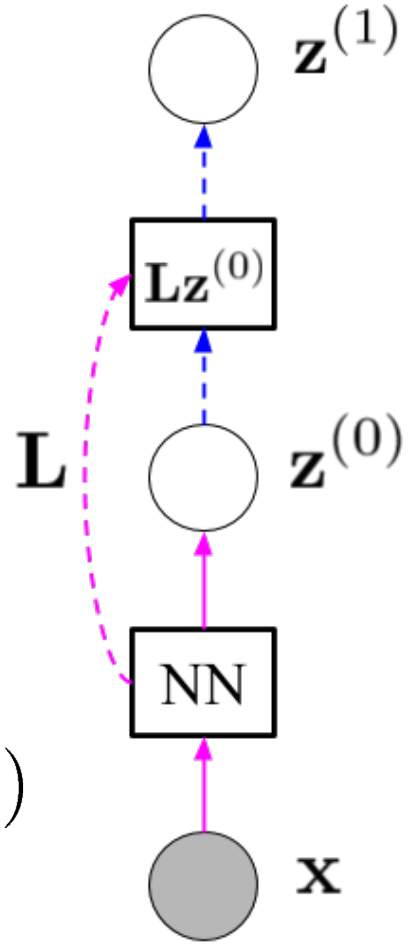
Linear Inverse Autoregressive Flow

- Model the full-covariance matrix ***directly*** using a **lower-triangular matrix with ones on the diagonal** $\mathbf{L}(\mathbf{x})$:

$$\mathbf{z}^{(1)}(\mathbf{x}) = \mathbf{L}(\mathbf{x})\ \mathbf{z}^{(0)}(\mathbf{x}).$$

Then:

$$\mathbf{z}^{(1)} \sim \mathcal{N}(\mathbf{L}(\mathbf{x})\mu(\mathbf{x}), \mathbf{L}(\mathbf{x})diag(\sigma^2(\mathbf{x}))\mathbf{L}(\mathbf{x})^\top)$$



Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improving Variational Inference with Inverse Autoregressive Flow. NIPS 2016

Convex combination of lower triangular matrices in linear IAF

- Can we further improve on linear IAF?
  → several $\mathbf{L}$'s
  → convex combination of $\mathbf{L}$'s

- Encoder returns also: $\mathbf{y}(\mathbf{x}) = \texttt{softmax}(\mathbf{h})$

$$\mathbf{L}_k(\mathbf{x}),\ k = 1, \ldots, K$$

- This is still volume-preserving flow:

$$\mathbf{z}^{(1)} = \left(\sum_{k=1}^{K} \mathbf{y}_k\ \mathbf{L}_k \right)\ \mathbf{z}^{(0)}$$

- Intuition: taking into account small variations among images.

Tomczak, J. M., & Welling, M. (2017). Improving Variational Inference with convex combination linear Inverse Autoregressive Flow. Benelearn 2017

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Other volume-preserving flows

- **NICE**: Non-linear independent components estimation

  Dinh, L., Krueger, D., & Bengio, Y. (2014). *NICE: Non-linear independent components estimation*. arXiv preprint arXiv:1410.8516. Workshop paper at ICLR 2015

- **HVI**: Hamiltonian Variational Inference

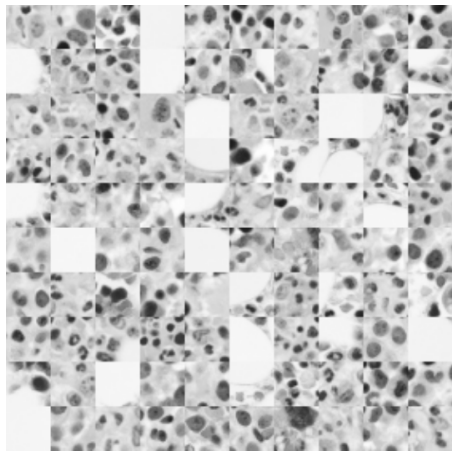
  Salimans, T., Kingma, D. P., & Welling, M. (2015). *Markov Chain Monte Carlo and Variational Inference: Bridging the Gap*. ICML 2015 (Vol. 37, pp. 1218-1226)

Experiments: static MNIST

Method	Test Log-Likelihood
VAE	-93.9
VAE+HF(T=1)	-88.1
VAE+HF(T=10)	-87.8
VAE+linIAF	-86.7
VAE+cc linIAF(K=5)	-86.1
VAE+NICE(T=10)	-88.6
VAE+NICE(T=80)	-87.2
VAE+HVI(T=1)	-91.7
VAE+HVI(T=8)	-88.3
VAE+NF(T=10)	-87.5
VAE+NF(T=80)	*-85.1*

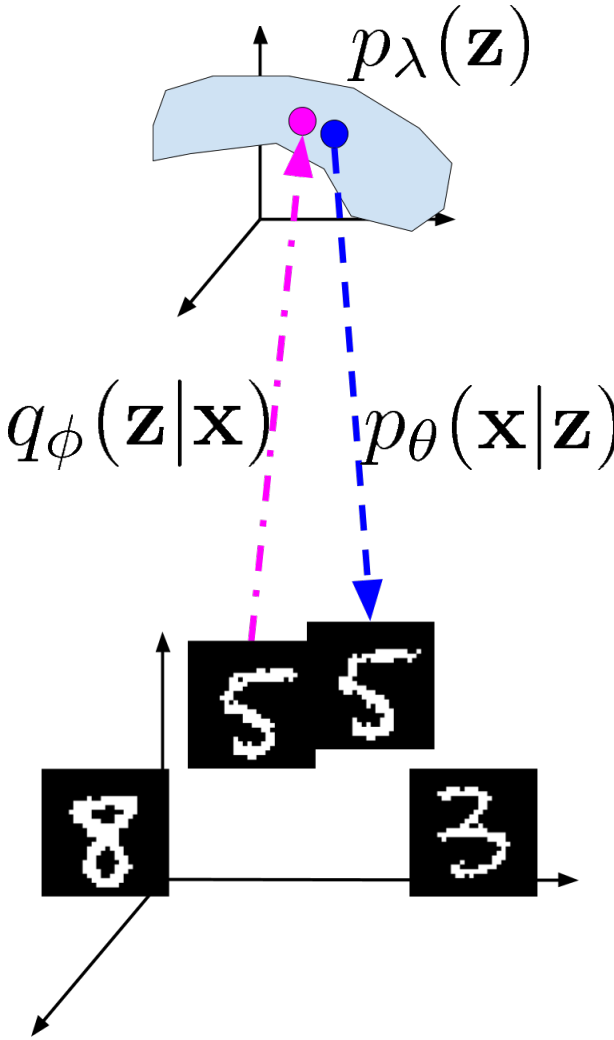
Experiments: Histopathology data

Method	ELBO
VAE	1371.4 +/- 32.1
VAE+HF(T=1)	1388.0 +/- 22.1
VAE+HF(T=10)	1397.0 +/- 15.2
VAE+HF(T=20)	1398.3 +/- 8.1
VAE+linIAF	1388.6 +/- 7.1
VAE+cc linIAF(K=5)	1413.8 +/- 22.9



Improving the prior

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})$$



RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Re-writing the ELBO

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] &\geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\lambda(\mathbf{z})] \\ &= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log q_\theta(\mathbf{z}|\mathbf{x})]] + \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\lambda(\mathbf{z})]\end{aligned}$$

Further we get:

$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[-\log q_\theta(\mathbf{z}|\mathbf{x})]] - \int \left(\int q(\mathbf{x})q(\mathbf{z}|\mathbf{x})\mathrm{d}\mathbf{x} \right) [-\log p_\lambda(\mathbf{z})]\mathrm{d}\mathbf{z}$$

$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[q_\theta(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})|p_\lambda(\mathbf{z})]$$

Re-writing the ELBO

Empirical distribution

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\lambda(\mathbf{z})]$$
$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log q_\theta(\mathbf{z}|\mathbf{x})]] + \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\lambda(\mathbf{z})]$$

Further we get:

$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[-\log q_\theta(\mathbf{z}|\mathbf{x})]] - \int \left(\int q(\mathbf{x})q(\mathbf{z}|\mathbf{x})\mathrm{d}\mathbf{x} \right) [-\log p_\lambda(\mathbf{z})]\mathrm{d}\mathbf{z}$$
$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[q_\theta(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})|p_\lambda(\mathbf{z})]$$

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\lambda(\mathbf{z})]$$
$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log q_\theta(\mathbf{z}|\mathbf{x})]] + \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\lambda(\mathbf{z})]$$

Further we get:

$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[-\log q_\theta(\mathbf{z}|\mathbf{x})]] - \int \Big(\int q(\mathbf{x})q(\mathbf{z}|\mathbf{x})\mathrm{d}\mathbf{x} \Big) [-\log p_\lambda(\mathbf{z})]\mathrm{d}\mathbf{z}$$

$$= \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[q_\theta(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})|p_\lambda(\mathbf{z})]$$

Aggregated encoding distribution

$$= \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)$$

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Forces $\mathbf{z}$ to be peaked for given $\mathbf{x}$

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Forces $\mathbf{z}$ to be peaked for given $\mathbf{x}$

Forces $\mathbf{z}$ to have high variance

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Forces $\mathbf{z}$ to be peaked for given $\mathbf{x}$

Forces $\mathbf{z}$ to have high variance

Forces ***encoder*** to match the ***prior***.

Minimal CE
$\rightarrow$ tighter bound!

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Forces $\mathbf{z}$ to be peaked for given $\mathbf{x}$

Forces $\mathbf{z}$ to have high variance

Forces ***encoder*** to match the ***prior***.

Minimal CE
$\rightarrow$ tighter bound!

For what prior is it minimized?

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Forces $\mathbf{z}$ to be peaked for given $\mathbf{x}$

Forces $\mathbf{z}$ to have high variance

Forces ***encoder*** to match the ***prior***.

Minimal CE
$\rightarrow$ tighter bound!

For what prior is it minimized?

$$p_\lambda(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)$$

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Re-writing the ELBO

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\log p(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{q(\mathbf{x})}[\mathbb{H}[\log q_\phi(\mathbf{z}|\mathbf{x})]] - \mathrm{CE}[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$

Forces $\mathbf{z}$ to be peaked for given $\mathbf{x}$

Forces $\mathbf{z}$ to have high variance

Forces ***encoder*** to match the ***prior***.

Minimal CE
$\rightarrow$ tighter bound!

For what prior is it minimized?

$$p_\lambda(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)$$

Infeasible

$$\approx \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

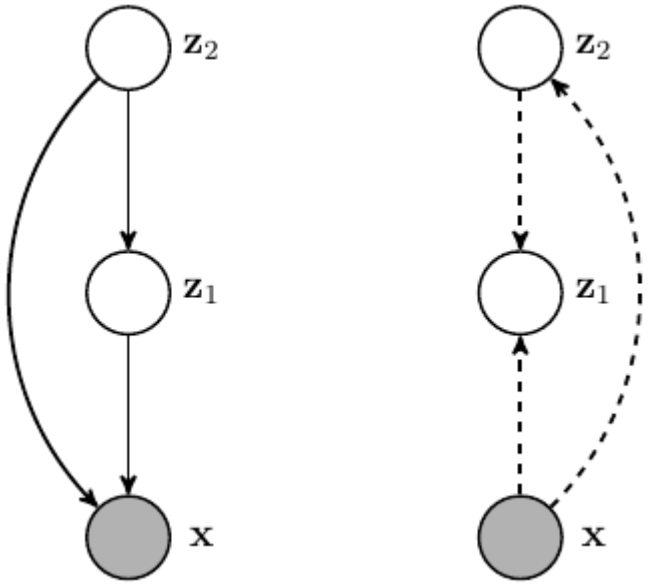
Learnable pseudo-inputs

New prior and two-level VAE

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

Variational **M**ixture of **P**osteriors **Prior**

Hierarchical VampPrior Variational Auto-Encoder:



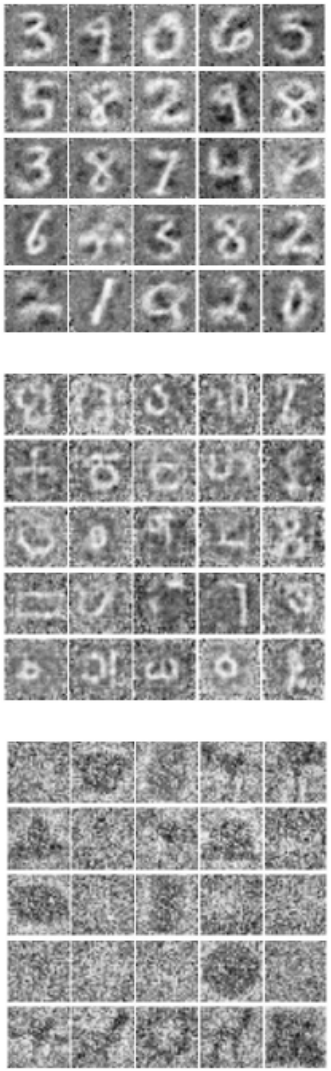
$$\begin{aligned}
p(\mathbf{z}_2) &= \frac{1}{K}\sum_{k=1}^{N} q_\psi(\mathbf{z}_2|\mathbf{u}_k), \\
p_\lambda(\mathbf{z}_1|\mathbf{z}_2) &= \mathcal{N}\big(\mathbf{z}_1|\mu_\lambda(\mathbf{z}_2), \mathrm{diag}(\sigma^2_\lambda(\mathbf{z}_2))\big), \\
q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{z}_2) &= \mathcal{N}\big(\mathbf{z}_1|\mu_\phi(\mathbf{x},\mathbf{z}_2), \mathrm{diag}(\sigma^2_\phi(\mathbf{x},\mathbf{z}_2))\big), \\
q_\psi(\mathbf{z}_2|\mathbf{x}) &= \mathcal{N}\big(\mathbf{z}_2|\mu_\psi(\mathbf{x}), \mathrm{diag}(\sigma^2_\psi(\mathbf{x}))\big).
\end{aligned}$$

Tomczak, J. M., & Welling, M. (2017). VAE with a VampPrior. arXiv:1705.07120

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

New prior and two-level VAE

Generations for a pseudo-input



Pseudo-inputs

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Some extensions and applications

- Semi-supervised learning with VAE.

  Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). *Semi-supervised learning with deep generative models*. NIPS

- VAE for sequences.

  Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2015). *Generating sentences from a continuous space*. arXiv preprint arXiv:1511.06349.

  Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., & Bengio, Y. (2015). *A recurrent latent variable model for sequential data*. NIPS

- More powerful decoders (using PixelCNN).

  Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., & Courville, A. (2016). *PixelVAE: A latent variable model for natural images*. arXiv preprint arXiv:1611.05013.

  Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., ... & Abbeel, P. (2016). *Variational lossy autoencoder*. arXiv preprint arXiv:1611.02731.

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions

Some extensions and applications

- Applications: graph data

Kipf, T. N., & Welling, M. (2016). *Variational Graph Auto-Encoders*. arXiv preprint arXiv:1611.07308. NIPS Workshop

Berg, R. V. D., Kipf, T. N., & Welling, M. (2017). *Graph Convolutional Matrix Completion*. arXiv preprint arXiv:1706.02263.

- Applications: drug response prediction

Rampasek, L., & Goldenberg, A. (2017). *Dr.VAE: Drug Response Variational Autoencoder*. arXiv preprint arXiv:1706.08203.

- Applications: text generation

Yang, Z., Hu, Z., Salakhutdinov, R., & Berg-Kirkpatrick, T. (2017). *Improved Variational Autoencoders for Text Modeling using Dilated Convolutions*. arXiv preprint arXiv:1702.08139.

Web-page:
https://jmtomczak.github.io

Code on github:
https://github.com/jmtomczak

Contact:
J.M.Tomczak@uva.nl
jakubmkt@gmail.com

The presented research was funded by the European Commission within the Marie Skłodowska-Curie Individual Fellowship (Grant No. 702666, "*Deep learning and Bayesian inference for medical imaging*").

RESEARCH & INNOVATION
Marie Skłodowska-Curie actions