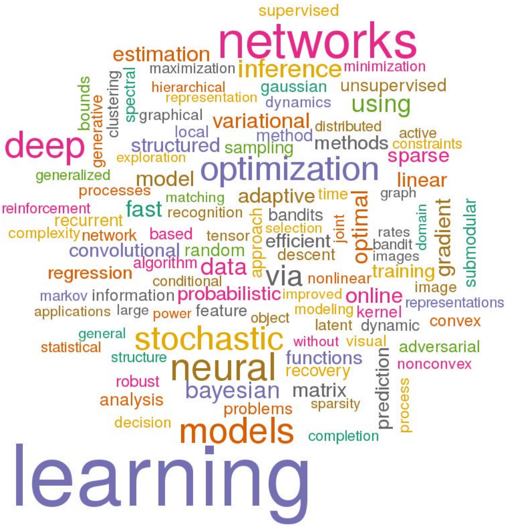
Approximate Inference in Latent Variable Models based on NNs

Pablo M. Olmos, olmos@tsc.uc3m.es

October 3, 2017

uc3m | Universidad Carlos III de Madrid



NIPS 2016 word cloud

Keywords today:

- Latent Variable Models
- Variational Inference
- Neural Networks
- Stochastic Gradient Descend

Index

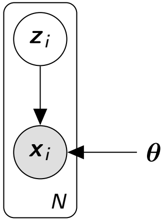
Index

Fitting a distribution using VAEs: Generative Model

Consider a set of i.i.d observations $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, $i = 1, \ldots, N$. Let $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$.

Generative Model using a Latent Space

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$



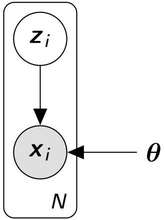
D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes., https://arxiv.org/pdf/1312.6114.pdf

Fitting a distribution using VAEs: Generative Model

Consider a set of i.i.d observations $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, $i = 1, \ldots, N$. Let $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$.

Generative Model using a Latent Space

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$



- $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$ $(k \leq d)$ (low-dimensional embedding)
- $\boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z})))$
- $\mu_{\boldsymbol{\theta}}(\boldsymbol{z})$ and $\log(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))$ are the outputs of a NN $\mathbb{R}^k \to \mathbb{R}^d$ with parameter vector $\boldsymbol{\theta}$ (Decoding network).

D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes., `https://arxiv.org/pdf/1312.6114.pdf`

Index

Fitting a distribution using VAEs: ML objective

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

ML estimation (in general intractable)

$$\begin{aligned}
\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}\in\Omega} \frac{1}{N}\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \\
&\approx \arg\max_{\boldsymbol{\theta}\in\Omega} \int p^*(\boldsymbol{x}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x})d\boldsymbol{x} \\
&= \arg\min_{\boldsymbol{\theta}\in\Omega} \mathrm{KL}\left(p^*(\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{x})\right)
\end{aligned}$$

Recall that $\mathrm{KL}\left(p(\boldsymbol{x})||q(\boldsymbol{x})\right) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \geq 0$

Fitting a distribution using VAEs: Variational objective

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Consider **any** approximation $q(\boldsymbol{z})$ to $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$

Fitting a distribution using VAEs: Variational objective

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Consider **any** approximation $q(\boldsymbol{z})$ to $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathrm{KL}\left(q(\boldsymbol{z})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right) + \mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})$$

where

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}) = \int q(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z} = \mathbb{E}_{q(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \mathrm{KL}\left(q(\boldsymbol{z})||p(\boldsymbol{z})\right)$$

is the Evidence Lower Bound (ELBO).

We will optimize $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ But first we need to select a variational family for $q(\boldsymbol{z})$.

Fitting a distribution using VAEs: Variational objective (II)

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

The Inference Network or Encoding Network

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$$

where $\mu_{\boldsymbol{\eta}}(\boldsymbol{x})$ and $\log(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}))$ are the outputs of a NN $\mathbb{R}^d \to \mathbb{R}^k$ with parameter vector $\boldsymbol{\eta}$.

Fitting a distribution using VAEs: Variational objective (II)

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

The Inference Network or Encoding Network

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$$

where $\mu_{\boldsymbol{\eta}}(\boldsymbol{x})$ and $\log(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}))$ are the outputs of a NN $\mathbb{R}^d \to \mathbb{R}^k$ with parameter vector $\boldsymbol{\eta}$.

The VAE optimization problem

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)$$

Index

Fitting a distribution using VAEs: Stochastic Optimization

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) = \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)$$

Recall that $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$, thus

$$\mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) = \frac{1}{2} \left(-k + \sum_{j=1}^{k} \sigma_{\boldsymbol{\eta},j}(\boldsymbol{x}) - \log \sigma_{\boldsymbol{\eta},j}(\boldsymbol{x}) + \mu_{\boldsymbol{\eta},j}^2 \right)$$

$\rightarrow$ Differentiable w.r.t. $\boldsymbol{\eta}$

Fitting a distribution using VAEs: Stochastic Optimization

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) = \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)$$

Recall that $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$, thus

$$\mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) = \frac{1}{2} \left(-k + \sum_{j=1}^{k} \sigma_{\boldsymbol{\eta},j}(\boldsymbol{x}) - \log \sigma_{\boldsymbol{\eta},j}(\boldsymbol{x}) + \mu_{\boldsymbol{\eta},j}^2 \right)$$

$\rightarrow$ Differentiable w.r.t. $\boldsymbol{\eta}$

Unbiased gradient estimator

$$\nabla_{\boldsymbol{\eta}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right) \approx \nabla_{\boldsymbol{\eta}} \left(\frac{1}{M} \sum_{i \in \mathcal{M}} \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)$$

where $\mathcal{M}$ represents a M-sized minibatch of data sampled at random from $\mathcal{D}$ ($M << N$).

Fitting a distribution using VAEs: Stochastic Optimization (II)

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) = \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)$$

We also need unbiased gradient estimates for

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right]$$

Fitting a distribution using VAEs: Stochastic Optimization (II)

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) = \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)$$

We also need unbiased gradient estimates for

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right]$$

We use a Monte Carlo sampling estimator

$$\mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] \approx \frac{1}{S} \sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{i,s})$$

where $\boldsymbol{z}^{(s,i)}$, $s = 1, \ldots, S$ are i.i.d. samples from $q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})$.

We typically use a **single** sample, i.e., $S = 1$ (huge estimator variance, but cheap computation).

Fitting a distribution using VAEs: Stochastic Optimization (III)

If $\boldsymbol{z}^{(s,i)}$ are i.i.d. samples from $q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})))$

$$\frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{i,s}) \rightarrow \text{how do we compute gradients w.r.t. } \boldsymbol{\eta}?$$

Fitting a distribution using VAEs: Stochastic Optimization (III)

If $\boldsymbol{z}^{(s,i)}$ are i.i.d. samples from $q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})))$

$$\frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{i,s}) \rightarrow \text{how do we compute gradients w.r.t. } \boldsymbol{\eta}?$$

Reparameterization Trick

Express each sample $\boldsymbol{z}^{(s,i)}$ as a deterministic function of $\boldsymbol{x}^{(i)}$ and some noise vector $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ that is independent of $\boldsymbol{\eta}$. For a Gaussian distribution we have

$$\boldsymbol{z}^{(s,i)} = f_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}, \epsilon) = \mu_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}) + \sqrt{\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})} \cdot \epsilon$$

Fitting a distribution using VAEs: Stochastic Optimization (III)

If $\boldsymbol{z}^{(s,i)}$ are i.i.d. samples from $q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})))$

$$\frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{i,s}) \rightarrow \text{how do we compute gradients w.r.t. } \boldsymbol{\eta}?$$

Reparameterization Trick

Express each sample $\boldsymbol{z}^{(s,i)}$ as a deterministic function of $\boldsymbol{x}^{(i)}$ and some noise vector $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ that is independent of $\boldsymbol{\eta}$. For a Gaussian distribution we have

$$\boldsymbol{z}^{(s,i)} = f_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}, \epsilon) = \mu_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}) + \sqrt{\sigma_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})} \cdot \epsilon$$

Putting all together...

$$\mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})\right] = \mathbb{E}_{\epsilon}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|f_{\boldsymbol{\eta}}(\epsilon, \boldsymbol{x}^{(i)}))\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|f_{\boldsymbol{\eta}}(\epsilon^{s,i}, \boldsymbol{x}^{(i)}))$$

where $\epsilon^{s,i}$, $s = 1, \ldots, S$ are i.i.d. samples from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Fitting a distribution using VAEs: the algorithm

Algorithm 1 The Variational Autoencoder (S=1)

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$, $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}_0$
$\ell \leftarrow 0$
while not converged **do**
  Sample minibatch $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\}$ from $\mathcal{D}$
  Sample $\{\epsilon^1, \ldots, \epsilon^M\}$ from $p(\epsilon) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  Compute noisy gradients:

$$\boldsymbol{g}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\eta}} \leftarrow \frac{1}{M} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}, \boldsymbol{\eta}} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} | f_{\boldsymbol{\eta}}(\epsilon^i, \boldsymbol{x}^{(i)})) - \mathrm{KL}\left(q_{\boldsymbol{\eta}, \boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right]$$

  Perform SGD-updates:

$$\boldsymbol{\theta}_{\ell+1} \leftarrow \boldsymbol{\theta}_{\ell+1} + h_\ell \boldsymbol{g}_{\boldsymbol{\theta}}$$
$$\boldsymbol{\eta}_{\ell+1} \leftarrow \boldsymbol{\eta}_{\ell+1} + h_\ell \boldsymbol{\eta}_{\boldsymbol{\theta}}$$

  $\ell \leftarrow \ell + 1$
end while

Index

- Approximate Inference with Amortised MCMC
  `https://arxiv.org/pdf/1702.08343.pdf`
- Autoencoding Variational Inference for Topic Models
  `https://arxiv.org/pdf/1703.01488.pdf`
- Improved Variational Autoencoders for Text Modeling using Dilated Convolutions
  `http://proceedings.mlr.press/v70/yang17d/yang17d.pdf`
- Variational Sequential Monte Carlo
  `https://arxiv.org/pdf/1705.11140.pdf`
- Stick Breaking Variational Autoencoders
  `https://arxiv.org/pdf/1605.06197.pdf`
- AutoGP: Exploring the Capabilities and Limitations of Gaussian Process Models `http://arxiv.org/abs/1610.05392`

Index

Revisiting the VAE generative model

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))
\end{aligned}
$$

$$
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)
$$

Revisiting the VAE generative model

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))
\end{aligned}
$$

$$
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)
$$

Some critical assumptions (among many others)

- Unimodal posterior approximation.
- $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is known.
- I can find a valid reconstruction $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ model.
- Prior too simple? It does not enforce interpretability in the latent space.
- ...

Revisiting the VAE generative model

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))
\end{aligned}
$$

$$
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{1}{N}\max_{\boldsymbol{\theta},\boldsymbol{\eta}}\left(\sum_{i=1}^{N}\mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})\right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})||p(\boldsymbol{z})\right)\right)
$$

Some critical assumptions (among many others)

- **Unimodal posterior approximation.**
- $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is known.
- I can find a valid reconstruction $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ model.
- Prior too simple? It does not enforce interpretability in the latent space.
- ...

Index

Normalizing flows

Let $\boldsymbol{z} \sim q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})$, where $q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$. Given an invertible, smooth mapping $g : \mathbb{R}^k \to \mathbb{R}^k$, the random variable $\boldsymbol{z}' = g(\boldsymbol{z})$ has a distribution

$$q(\boldsymbol{z}') = q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) \left| \det \frac{\partial g^{-1}}{\partial \boldsymbol{z}'} \right| = q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) \left| \det \frac{\partial g}{\partial \boldsymbol{z}} \right|^{-1}$$

which in general is not unimodal.

D. Jimenez Rezende and S. Mohamed, Variational Inference with Normalizing Flows., https://arxiv.org/pdf/1505.05770.pdf

Normalizing flows

Let $\boldsymbol{z} \sim q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})$, where $q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$. Given an invertible, smooth mapping $g : \mathbb{R}^k \to \mathbb{R}^k$, the random variable $\boldsymbol{z}' = g(\boldsymbol{z})$ has a distribution

$$q(\boldsymbol{z}') = q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) \left| \det \frac{\partial g^{-1}}{\partial \boldsymbol{z}'} \right| = q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) \left| \det \frac{\partial g}{\partial \boldsymbol{z}} \right|^{-1}$$

which in general is not unimodal.

We can construct arbitrarily complex densities by composing several simple maps and applying this result.

$$\boldsymbol{z}_T = g_T \circ g_{T-1} \circ g_{T-2} \ldots \circ g_1(\boldsymbol{z}), \quad \boldsymbol{z} \sim q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})$$

$$\log q_{\boldsymbol{\eta},\boldsymbol{x},T}(\boldsymbol{z}_T) = \log q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) - \sum_{t=1}^{T} \log \left| \det \frac{\partial g_t}{\partial \boldsymbol{z}_{t-1}} \right|,$$

where $\boldsymbol{z}_t$ is the output of the *normalizing flow* after the t-th transformation.

D. Jimenez Rezende and S. Mohamed, Variational Inference with Normalizing Flows., https://arxiv.org/pdf/1505.05770.pdf

Normalizing flows (II)

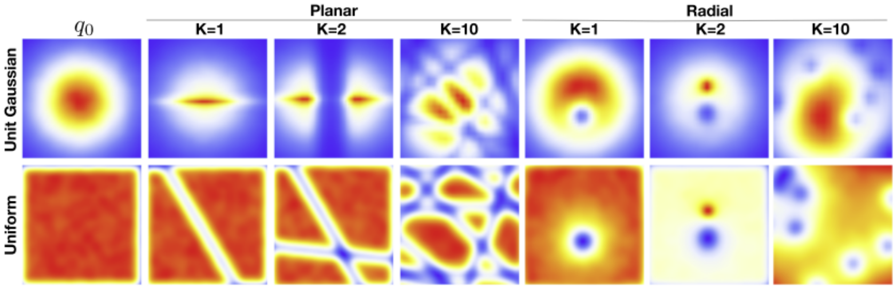


Figure 1. Effect of normalizing flow on two distributions.

D. Jimenez Rezende and S. Mohamed, Variational Inference with Normalizing Flows., https://arxiv.org/pdf/1505.05770.pdf

Normalizing flows (III)

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T)p(\boldsymbol{z}_T)d\boldsymbol{z} \\
p(\boldsymbol{z}_T) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}_T), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}_T))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{x},T}(\boldsymbol{z}_T)
\end{aligned}
$$

Normalizing flows (III)

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T)p(\boldsymbol{z}_T)d\boldsymbol{z} \\
p(\boldsymbol{z}_T) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}_T), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}_T))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{x},T}(\boldsymbol{z}_T)
\end{aligned}
$$

SGD optimization is run using the following expression for the ELBO:

$$
\begin{aligned}
&\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
&\geq \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x},T}(\boldsymbol{z}_T)}\left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T)p(\boldsymbol{z}_T)) - \log q_{\boldsymbol{\eta},\boldsymbol{x},T}(\boldsymbol{z}_T)\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})}\left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T)p(\boldsymbol{z}_T)) - \log q_{\boldsymbol{\eta},\boldsymbol{x},T}(\boldsymbol{z}_T)\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})}\left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_T)p(\boldsymbol{z}_T)) - \log q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})\right] + \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})}\left[\sum_{t=1}^{T} \log \left|\det \frac{\partial g_t}{\partial \boldsymbol{z}_{t-1}}\right|\right]
\end{aligned}
$$

Index

Revisiting the VAE generative model

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))
\end{aligned}
$$

$$
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right)
$$

Some critical assumptions (among many others)

- Unimodal posterior approximation.
- $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is known.
- I can find a valid reconstruction $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ model.
- Prior too simple? It does not enforce interpretability in the latent space.
- ...

Implicit posterior distribution

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = f_{\boldsymbol{\eta}}(\boldsymbol{x}, \epsilon) \rightarrow \text{We can only sample from it}
\end{aligned}
$$

$$
\begin{aligned}
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) &= \frac{1}{N}\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}) \right] - \mathrm{KL}\left(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z}) || p(\boldsymbol{z}) \right) \right) \\
&= \frac{1}{N}\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})p(\boldsymbol{z})) - \log(q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})) \right] \right)
\end{aligned}
$$

$q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})$ cannot appear explicitly in the objective function!

Discriminative Classifiers for log-likelihood ratio estimation

Consider the following two joint distributions:

$$\left\{ \begin{array}{c} q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})p^*(\boldsymbol{x}) \\ p(\boldsymbol{z})p^*(\boldsymbol{x}) \end{array} \right. \rightarrow p^*(\boldsymbol{x}) \text{ is the real distribution of the data.}$$

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin, 2001

Discriminative Classifiers for log-likelihood ratio estimation

Consider the following two joint distributions:

$$\begin{cases} q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})p^*(\boldsymbol{x}) \\ p(\boldsymbol{z})p^*(\boldsymbol{x}) \end{cases} \rightarrow p^*(\boldsymbol{x}) \text{ is the real distribution of the data.}$$

Let $T(\boldsymbol{z}, \boldsymbol{x})$ be a classifier network that is trained to discriminate between samples coming from either $q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})p^*(\boldsymbol{x})$ (+1 class), or $p(\boldsymbol{z})p^*(\boldsymbol{x})$ (0 class). Given a $(\boldsymbol{z}', \boldsymbol{x}')$ sample

$$\mathbb{P}\left((\boldsymbol{z}', \boldsymbol{x}')\text{is drawn from } p(\boldsymbol{z})p^*(\boldsymbol{x})\right) = \frac{1}{1 + e^{-T(\boldsymbol{z}', \boldsymbol{x}')}} \triangleq \text{sigm}(T(\boldsymbol{z}', \boldsymbol{x}'))$$

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin, 2001

Discriminative Classifiers for log-likelihood ratio estimation (II)

The optimal discriminator $T^*(\boldsymbol{z}, \boldsymbol{x})$ according to the cross entropy loss function

$$\arg\max_{T(\boldsymbol{z},\boldsymbol{x})} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})p^*(\boldsymbol{x})} [\log \mathrm{sigm}(T(\boldsymbol{z}, \boldsymbol{x}))] + \mathbb{E}_{p(\boldsymbol{z})p^*(\boldsymbol{x})} [\log(1 - \mathrm{sigm}(T(\boldsymbol{z}, \boldsymbol{x})))]$$

is

$$T^*(\boldsymbol{z}, \boldsymbol{x}) = \log \frac{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})}{p(\boldsymbol{z})}$$

Discriminative Classifiers for log-likelihood ratio estimation (II)

The optimal discriminator $T^*(\boldsymbol{z}, \boldsymbol{x})$ according to the cross entropy loss function

$$\arg\max_{T(\boldsymbol{z},\boldsymbol{x})} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})p^*(\boldsymbol{x})}\left[\log \mathrm{sigm}(T(\boldsymbol{z},\boldsymbol{x}))\right] + \mathbb{E}_{p(\boldsymbol{z})p^*(\boldsymbol{x})}\left[\log(1-\mathrm{sigm}(T(\boldsymbol{z},\boldsymbol{x})))\right]$$

is

$$T^*(\boldsymbol{z}, \boldsymbol{x}) = \log \frac{q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})}{p(\boldsymbol{z})}$$

In practice, we approximate $T^*(\boldsymbol{z}, \boldsymbol{x})$ using a deep NN with parameters $\boldsymbol{\psi}$ that is trained using samples of both distributions:

$$\boldsymbol{\psi}^* = \arg\max_{\boldsymbol{\psi}} \sum_{i=1}^{N} \frac{1}{2} \log \mathrm{sigm}(T(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)})) + \frac{1}{2}\log(1-\mathrm{sigm}(T(\boldsymbol{x}^{(i)}, \tilde{\boldsymbol{z}}^{(i)})))$$

where $(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)})$, $i = 1, \ldots, N$ are i.i.d. samples from $q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z})p^*(\boldsymbol{x})$ and $(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)})$, $i = 1, \ldots, N$ are i.i.d. samples from $p(\boldsymbol{z})p^*(\boldsymbol{x})$.

The aversarial VAE with implicit posterior approximation

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = f_{\boldsymbol{\eta}}(\boldsymbol{x}, \epsilon) \rightarrow \text{We can only sample from it}
\end{aligned}
$$

$$
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})) + \log(\frac{p(\boldsymbol{z})}{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}) \right] \right)
$$

L. Mescheder , S. Nowozin, and A. Geiger, Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks.

F. Huszár, Variational Inference using Implicit Distributions.

The aversarial VAE with implicit posterior approximation

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{z}), \mathrm{diag}(\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}))) \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = f_{\boldsymbol{\eta}}(\boldsymbol{x}, \epsilon) \rightarrow \text{We can only sample from it}
\end{aligned}
$$

$$
\begin{aligned}
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})) + \log(\frac{p(\boldsymbol{z})}{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}) \right] \right) \\
&= \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})) - T_{\boldsymbol{\psi}^*}(\boldsymbol{z}, \boldsymbol{x}^{(i)}) \right] \right)
\end{aligned}
$$

L. Mescheder , S. Nowozin, and A. Geiger, Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks.

F. Huszár, Variational Inference using Implicit Distributions.

The aversarial VAE with implicit posterior approximation

(a) Training data

(b) Random samples

L. Mescheder , S. Nowozin, and A. Geiger, Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks.

The adversarial VAE with implicit likelihood distribution

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon) \rightarrow \text{We can only sample from it}$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))$$

$$\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})) + \log(\frac{p(\boldsymbol{z})}{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}) \right] \right)$$

F. Huszár, Variational Inference using Implicit Distributions.

The adversarial VAE with implicit likelihood distribution

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
p(\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k) \\
p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) &= f_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon) \rightarrow \text{We can only sample from it} \\
p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) &\approx q_{\boldsymbol{\eta},\boldsymbol{x}}(\boldsymbol{z}) = \mathcal{N}(\mu_{\boldsymbol{\eta}}(\boldsymbol{x}), \mathrm{diag}(\sigma_{\boldsymbol{\eta}}(\boldsymbol{x})))
\end{aligned}
$$

$$
\begin{aligned}
\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})) + \log(\frac{p(\boldsymbol{z})}{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}) \right] \right) \\
&= \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[\log(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})p(\boldsymbol{z})}{p^*(\boldsymbol{x})q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})}) \right] \right) \\
&= \frac{1}{N} \max_{\boldsymbol{\theta},\boldsymbol{\eta}} \left(\sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\eta},\boldsymbol{x}^{(i)}}(\boldsymbol{z})} \left[-T_{\boldsymbol{\psi}^*}(\boldsymbol{z}, \boldsymbol{x}^{(i)}) \right] \right)
\end{aligned}
$$

F. Huszár, Variational Inference using Implicit Distributions.

Algorithm	Implicit			VI
	$p_\theta(z)$	$p_\theta(x\|z)$	$q_\psi(z\|x)$	
VAE (Kingma & Welling, 2014) NF (Rezende & Mohamed, 2015)				✓
PC-Adv, Algorithm 1 AffGAN† (Sønderby et al., 2017) AVB (Mescheder et al., 2017)	I		✓	✓
OPVI (Ranganath et al., 2016)	I		✓	✓
PC-Den, Algorithm 3	I		✓	✓
JC-Adv, Algorithm 2	I	I	✓	✓
JC-Den	I	I	✓	✓
JC-Adv-RMD‡	✓	✓	✓	✓
AAE (Makhzani et al., 2016)	I		✓	
DeePSiM (Dosovitskiy & Brox, 2016)	I		✓	
ALI (Dumoulin et al., 2017) BiGAN (Donahue et al., 2017)	✓	✓	✓	

F. Huszár, Variational Inference using Implicit Distributions.