Deep Generative Models
Lecture 6

Roman Isachenko

Ozon Masters

2021

Gaussian autoregressive model

Consider autoregressive model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i|\mathbf{x}_{1:i-1}, \boldsymbol{\theta}),$$

with conditionals

$$p(x_i|\mathbf{x}_{1:i-1}, \boldsymbol{\theta}) = \mathcal{N}\left(\hat{\mu}_i(\mathbf{x}_{1:i-1}), \hat{\sigma}_i^2(\mathbf{x}_{1:i-1})\right).$$

Forward and inverse

$$x_i = \hat{\sigma}_i(\mathbf{x}_{1:i-1}) \cdot z_i + \hat{\mu}_i(\mathbf{x}_{1:i-1}), \quad z_i \sim \mathcal{N}(0, 1).$$

$$z_i = (x_i - \hat{\mu}_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\hat{\sigma}_i(\mathbf{x}_{1:i-1})}.$$

Gaussian autoregressive model

Forward and inverse

$$\mathbf{x} = g(\mathbf{z}, \boldsymbol{\theta}); \quad x_i = \hat{\sigma}_i(\mathbf{x}_{1:i-1}) \cdot z_i + \hat{\mu}_i(\mathbf{x}_{1:i-1}), \quad z_i \sim \mathcal{N}(0, 1).$$

$$\mathbf{z} = f(\mathbf{x}, \boldsymbol{\theta}); \quad z_i = (x_i - \hat{\mu}_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\hat{\sigma}_i(\mathbf{x}_{1:i-1})}.$$

Jacobian

$$\log \left| \det \left(\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \mathbf{x}} \right) \right| = -\log \left| \det \left(\frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \mathbf{z}} \right) \right| = -\sum_{i=1}^{m} \log \hat{\sigma}_i(\mathbf{x}_{1:i-1}).$$

We get an autoregressive model with tractable (triangular) Jacobian, which is easily invertible. It is a flow!

Inverse autoregressive flow (IAF)

Gaussian autoregressive model ($\mathbf{z} \rightarrow \mathbf{x}$)

$$
\begin{aligned}
x_i &= \hat{\sigma}_i(\mathbf{x}_{1:i-1}) \cdot z_i + \hat{\mu}_i(\mathbf{x}_{1:i-1}). \\
z_i &= (x_i - \hat{\mu}_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\hat{\sigma}_i(\mathbf{x}_{1:i-1})}.
\end{aligned}
$$

This process is sequential.
Let use the following reparametrization: $\boldsymbol{\sigma} = \frac{1}{\hat{\boldsymbol{\sigma}}}; \quad \boldsymbol{\mu} = -\frac{\hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\sigma}}}$.

Inverse transform ($\mathbf{x} \rightarrow \mathbf{z}$)

$$
\begin{aligned}
z_i &= \sigma_i(\mathbf{x}_{1:i-1}) \cdot x_i + \mu_i(\mathbf{x}_{1:i-1}). \\
x_i &= (z_i - \mu_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\sigma_i(\mathbf{x}_{1:i-1})}.
\end{aligned}
$$

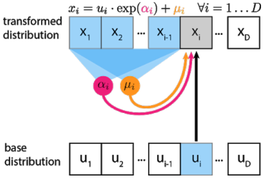
This process is **not** sequential.

https://arxiv.org/pdf/1606.04934.pdf

Inverse autoregressive flow (IAF)

Gaussian autoregressive model

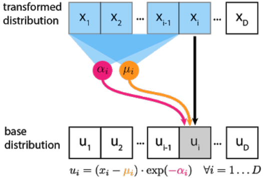
$$x_i = \hat{\sigma}_i(\mathbf{x}_{1:i-1}) \cdot z_i + \hat{\mu}_i(\mathbf{x}_{1:i-1}).$$



Inverse transform

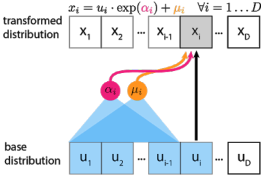
$$z_i = (x_i - \hat{\mu}_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\hat{\sigma}_i(\mathbf{x}_{1:i-1})};$$

$$z_i = \sigma_i(\mathbf{x}_{1:i-1}) \cdot x_i + \mu_i(\mathbf{x}_{1:i-1}).$$



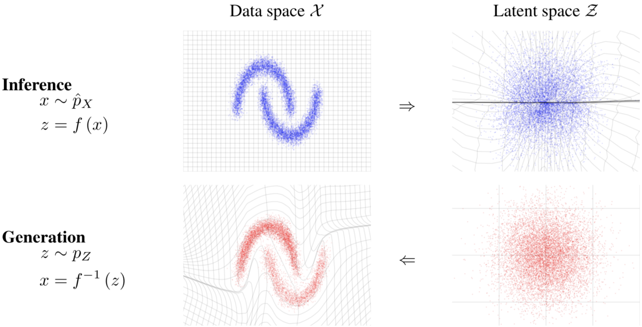
Inverse autoregressive flow

$$x_i = \sigma_i(\mathbf{z}_{1:i-1}) \cdot z_i + \mu_i(\mathbf{z}_{1:i-1}).$$



https://blog.evjang.com/2018/01/nf2.html

Flows



- Inference mode in autoregressive flows is used for density estimation task.
- Generation mode in autoregressive flows (IAF) is used for stochastic variational inference to get more flexible posterior distribution.

https://arxiv.org/pdf/1605.08803.pdf

Inverse autoregressive flow (IAF)

Inverse transform ($\mathbf{x} \to \mathbf{z}$)

$$z_i = \boldsymbol{\sigma}_i(\mathbf{x}_{1:i-1}) \cdot x_i + \boldsymbol{\mu}_i(\mathbf{x}_{1:i-1}).$$

$$x_i = (z_i - \boldsymbol{\mu}_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\boldsymbol{\sigma}_i(\mathbf{x}_{1:i-1})}.$$

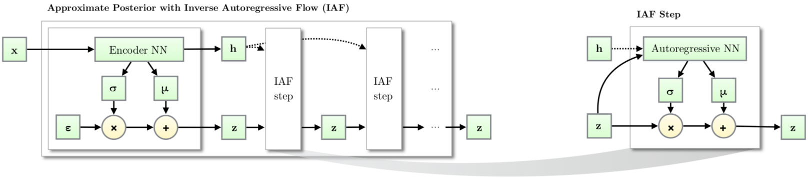
Inverse autoregressive flow use such inverted autoregressive model as a flow in VAE:

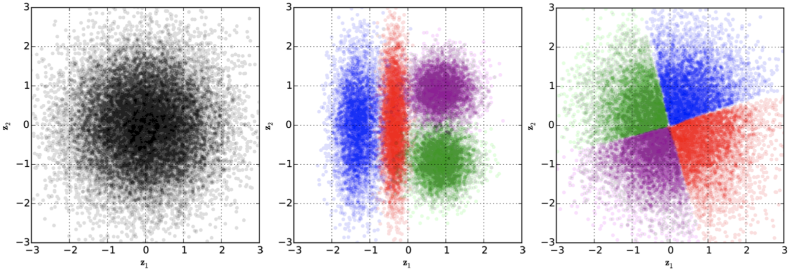
$$\mathbf{z}_0 = \boldsymbol{\sigma}(\mathbf{x}) \cdot \boldsymbol{\epsilon} + \boldsymbol{\mu}(\mathbf{x}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1); \quad \sim q(\mathbf{z}_0 | \mathbf{x}, \phi).$$

$$\mathbf{z}_k = \boldsymbol{\sigma}_k(\mathbf{z}_{k-1}) \cdot \mathbf{z}_{k-1} + \boldsymbol{\mu}_k(\mathbf{z}_{k-1}), \quad k \geq 1; \quad \sim q_k(\mathbf{z}_k | \mathbf{x}, \phi, \{\phi_j\}_{j=1}^k).$$

https://arxiv.org/pdf/1606.04934.pdf

Inverse autoregressive flow (IAF)





(a) Prior distribution (b) Posteriors in standard VAE (c) Posteriors in VAE with IAF

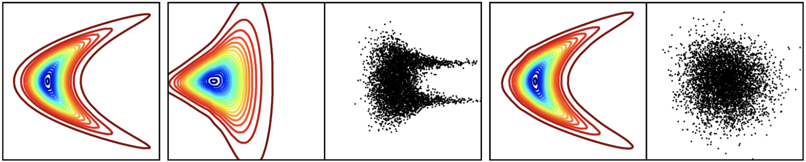
https://arxiv.org/pdf/1606.04934.pdf

Masked autoregressive flow (MAF)

Gaussian autoregressive model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i|\mathbf{x}_{1:i-1}, \boldsymbol{\theta}) = \prod_{i=1}^{m} \mathcal{N}\left(x_i|\mu_i(\mathbf{x}_{1:i-1}), \sigma_i^2(\mathbf{x}_{1:i-1})\right).$$

We could use MADE (masked autoencoder) as conditional model.
The sampling order could be crucial.



(a) Target density (b) MADE with Gaussian conditionals (c) MAF with 5 layers

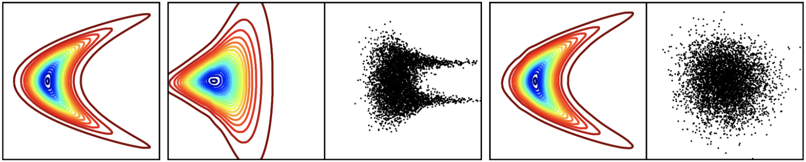
Samples from the base distribution could be an indicator of how good the flow was fitted.

https://arxiv.org/pdf/1705.07057.pdf

Masked autoregressive flow (MAF)

Gaussian autoregressive model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i|\mathbf{x}_{1:i-1}, \boldsymbol{\theta}) = \prod_{i=1}^{m} \mathcal{N}\left(x_i|\mu_i(\mathbf{x}_{1:i-1}), \sigma_i^2(\mathbf{x}_{1:i-1})\right).$$



(a) Target density (b) MADE with Gaussian conditionals (c) MAF with 5 layers

MAF is just a stacked MADE model.

---

https://arxiv.org/pdf/1705.07057.pdf

MAF vs IAF

Sampling and inverse transform in MAF

$$x_i = \hat{\sigma}_i(\mathbf{x}_{1:i-1}) \cdot z_i + \hat{\mu}_i(\mathbf{x}_{1:i-1}).$$

$$z_i = (x_i - \hat{\mu}_i(\mathbf{x}_{1:i-1})) \cdot \frac{1}{\hat{\sigma}_i(\mathbf{x}_{1:i-1})}.$$

- Sampling is slow (sequential).
- Density estimation is fast.

Sampling and inverse transform in IAF

$$x_i = \sigma_i(\mathbf{z}_{1:i-1}) \cdot z_i + \mu_i(\mathbf{z}_{1:i-1}).$$

$$z_i = (x_i - \mu_i(\mathbf{z}_{1:i-1})) \cdot \frac{1}{\sigma_i(\mathbf{z}_{1:i-1})}.$$

- Sampling is fast.
- Density estimation is slow (sequential).

---

https://arxiv.org/pdf/1705.07057.pdf

MAF vs IAF

Theorem

Training a MAF with maximum likelihood corresponds to fitting an implicit IAF with stochastic variational inference where the posterior is taken to be the base density $\pi(\mathbf{z})$:

$$\max_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}) \quad \Leftrightarrow \quad \min_{\boldsymbol{\theta}} KL\left(p(\mathbf{z}|\boldsymbol{\theta})||\pi(\mathbf{z})\right).$$

- $\pi(\mathbf{z})$ is a base distribution; $\pi(\mathbf{x})$ is a data distribution.
- $\mathbf{z} = f(\mathbf{x}, \boldsymbol{\theta})$ – MAF model; $\mathbf{x} = g(\mathbf{z}, \boldsymbol{\theta})$ – IAF model.

$$\log p(\mathbf{z}|\boldsymbol{\theta}) = \log \pi(g(\mathbf{z}, \boldsymbol{\theta})) + \log \left| \det \left(\frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \mathbf{z}} \right) \right|$$

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \pi(f(\mathbf{x}, \boldsymbol{\theta})) + \log \left| \det \left(\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \mathbf{x}} \right) \right|$$

---

https://arxiv.org/pdf/1705.07057.pdf

MAF vs IAF

Theorem

Training a MAF with maximum likelihood corresponds to fitting an implicit IAF with stochastic variational inference where the posterior is taken to be the base density $\pi(\mathbf{z})$:

$$\max_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}) \quad \Leftrightarrow \quad \min_{\boldsymbol{\theta}} KL\left(p(\mathbf{z}|\boldsymbol{\theta})||\pi(\mathbf{z})\right).$$

Proof

$$KL\left(p(\mathbf{z}|\boldsymbol{\theta})||\pi(\mathbf{z})\right) = \mathbb{E}_{p(\mathbf{z}|\boldsymbol{\theta})}\left[\log p(\mathbf{z}|\boldsymbol{\theta}) - \log \pi(\mathbf{z})\right] =$$

$$= \mathbb{E}_{p(\mathbf{z}|\boldsymbol{\theta})}\left[\log \pi(g(\mathbf{z}, \boldsymbol{\theta})) + \log \left|\det\left(\frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \mathbf{z}}\right)\right| - \log \pi(\mathbf{z})\right] =$$

$$= \mathbb{E}_{\pi(\mathbf{x})}\left[\log \pi(\mathbf{x}) - \log \left|\det\left(\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \mathbf{x}}\right)\right| - \log \pi(f(\mathbf{x}, \boldsymbol{\theta}))\right].$$

https://arxiv.org/pdf/1705.07057.pdf

MAF vs IAF

Proof (continued)

$$KL\left(p(\mathbf{z}|\boldsymbol{\theta})||\pi(\mathbf{z})\right) =$$

$$= \mathbb{E}_{\pi(\mathbf{x})}\left[\log \pi(\mathbf{x}) - \log \left|\det\left(\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \mathbf{x}}\right)\right| - \log \pi(f(\mathbf{x}, \boldsymbol{\theta}))\right] =$$

$$= \mathbb{E}_{\pi(\mathbf{x})}\big[\log \pi(\mathbf{x}) - \log p(\mathbf{x}|\boldsymbol{\theta})\big] = KL(\pi(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\theta})).$$

$$\arg\min_{\boldsymbol{\theta}} KL(\pi(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\theta})) = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\pi(\mathbf{x})}\left[\log \pi(\mathbf{x}) - \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

$$= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\pi(\mathbf{x})} \log p(\mathbf{x}|\boldsymbol{\theta})$$

Unbiased estimator is MLE:

$$\mathbb{E}_{\pi(\mathbf{x})} \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}).$$

---

https://arxiv.org/pdf/1705.07057.pdf

MAF vs IAF vs RealNVP

MAF

$$\mathbf{x} = \hat{\boldsymbol{\sigma}}(\mathbf{x}) \odot \mathbf{z} + \hat{\boldsymbol{\mu}}(\mathbf{x}).$$

- Calculating the density $p(\mathbf{x}|\boldsymbol{\theta})$ - 1 pass.
- Sampling - m passes.

IAF

$$\mathbf{x} = \boldsymbol{\sigma}(\mathbf{z}) \odot \mathbf{z} + \boldsymbol{\mu}(\mathbf{z}).$$

- Calculating the density $p(\mathbf{x}|\boldsymbol{\theta})$ - m passes.
- Sampling - 1 pass.

RealNVP

$$\mathbf{x}_{1:d} = \mathbf{z}_{1:d};$$
$$\mathbf{x}_{d:m} = \mathbf{z}_{d:m} \odot \exp\left(c_1(\mathbf{z}_{1:d}, \boldsymbol{\theta})\right) + c_2(\mathbf{x}_{1:d}, \boldsymbol{\theta}).$$

---

https://arxiv.org/pdf/1705.07057.pdf

MAF vs IAF vs RealNVP

RealNVP

$$
\begin{aligned}
\mathbf{x}_{1:d} &= \mathbf{z}_{1:d}; \\
\mathbf{x}_{d:m} &= \mathbf{z}_{d:m} \odot \exp\left(c_1(\mathbf{z}_{1:d}, \boldsymbol{\theta})\right) + c_2(\mathbf{x}_{1:d}, \boldsymbol{\theta}).
\end{aligned}
$$

- Calculating the density $p(\mathbf{x}|\boldsymbol{\theta})$ - 1 pass.
- Sampling - 1 pass.

RealNVP is a special case of MAF and IAF:

MAF

$$
\begin{cases}
\hat{\mu}_i = \hat{\sigma}_i = 0, \; i = 1, \dots, d; \\
\hat{\mu}_i, \hat{\sigma}_i \text{ – functions of } \mathbf{x}_{1:d}, \; i = d+1, \dots, m.
\end{cases}
$$

IAF

$$
\begin{cases}
\mu_i = \sigma_i = 0, \; i = 1, \dots, d; \\
\mu_i, \sigma_i \text{ – functions of } \mathbf{z}_{1:d}, \; i = d+1, \dots, m.
\end{cases}
$$

---

https://arxiv.org/pdf/1705.07057.pdf

MAF/IAF pros and cons

MAF

- Sampling is slow.
- Likelihood evaluation is fast.

IAF

- Sampling is fast.
- Likelihood evaluation is slow.

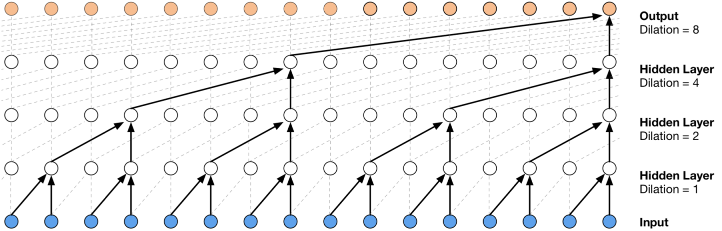
How to take the best of both worlds?

WaveNet (2016)

Autoregressive model for raw audio waveforms generation

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(x_t|\mathbf{x}_{1:t-1}, \boldsymbol{\theta}).$$

The model uses causal dilated convolutions.



https://arxiv.org/pdf/1609.03499.pdf

Parallel WaveNet, 2017

Previous WaveNet model

- raw audio is high-dimensional (e.g. 16000 samples per second for 16kHz audio);
- WaveNet encodes 8-bit signal with 256-way categorical distribution.

Goal

- improved fidelity (24kHz instead of 16kHz) $\rightarrow$ increase dilated convolution filter size from 2 to 3;
- 16-bit signals $\rightarrow$ mixture of logistics instead of categorical distribution.

---

https://arxiv.org/pdf/1711.10433.pdf

Parallel WaveNet, 2017

Probability density distillation

1. Train usual WaveNet (MAF) via MLE (teacher network).
2. Train IAF WaveNet model (student network), which attempts to match the probability of its own samples under the distribution learned by the teacher.

Student objective

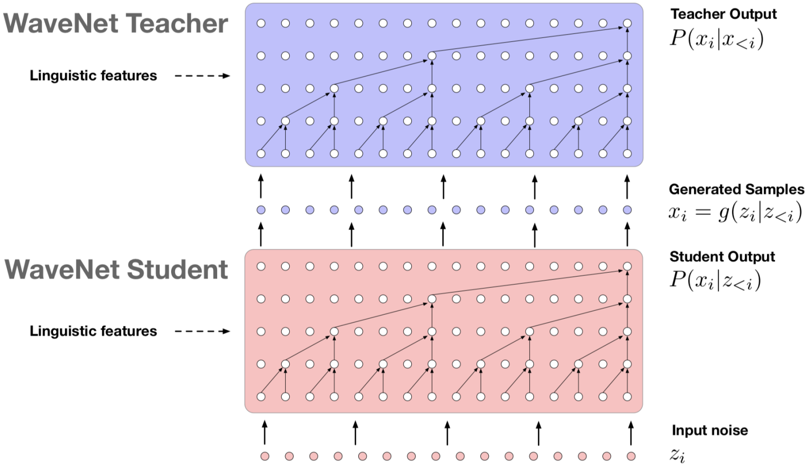
$$KL(p_s||p_t) = H(p_s, p_t) - H(p_s).$$

More than 1000x speed-up relative to original WaveNet!

---

https://arxiv.org/pdf/1711.10433.pdf

Parallel WaveNet, 2017



https://arxiv.org/pdf/1711.10433.pdf

Summary

- Flows is a continuous model. To use it for discrete distribution, the data should be dequantized.
- Original VAE model has lot of limitations. One of them is a restricted class of variational posteriors.
- Using flows in a latent space of VAE could give more flexible posterior distribution.
- Gaussian autoregressive model is a special type of flow (RealNVP model is a special type of this autoregressive model)
- MAF is an example of such model which is suitable for density estimation tasks.
- IAF used the inverse autoregressive transformation for variational inference task.

VAE limitations

- Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})).$$

- Poor prior distribution

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

- Poor probabilistic model (decoder)

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{z})).$$

- Loose lower bound

$$p(\mathbf{x}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = (?).$$

ELBO surgery, 2016

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}|\mathbf{X}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}.$$

ELBO interpretations

- Evidence minus posterior KL

$$\mathcal{L}(q, \boldsymbol{\theta}) = \log p(\mathbf{X}|\boldsymbol{\theta}) - KL(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})).$$

- Average negative energy plus entropy

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) + \mathbb{H}\left[q(\mathbf{Z}|\mathbf{X})\right].$$

- Average term-by-term reconstruction minus KL to prior

$$\mathcal{L}(q, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) - KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))\right].$$

http://approximateinference.org/accepted/HoffmanJohnson2016.pdf

ELBO surgery, 2016

$$\mathcal{L}(q, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) - KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) \right].$$

Theorem

$$\frac{1}{n} \sum_{i=1}^{n} KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = KL(q(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}],$$

where i is treated as random variable:

$$q(i, \mathbf{z}) = q(i)q(\mathbf{z}|i); \quad p(i, \mathbf{z}) = p(i)p(\mathbf{z}); \quad q(i) = p(i) = \frac{1}{n}; \quad q(\mathbf{z}|i) = q(\mathbf{z}|\mathbf{x}_i).$$

$$q(\mathbf{z}) = \sum_{i=1}^{n} q(i, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i); \quad \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}] = \mathbb{E}_{q(i,\mathbf{z})} \log \frac{q(i, \mathbf{z})}{q(i)q(\mathbf{z})}.$$

---

http://approximateinference.org/accepted/HoffmanJohnson2016.pdf

ELBO surgery, 2016

Theorem

$$\frac{1}{n}\sum_{i=1}^{n} KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = KL(q(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}].$$

Proof

$$\frac{1}{n}\sum_{i=1}^{n} KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = \sum_{i=1}^{n}\int q(i)q(\mathbf{z}|i)\log\frac{q(\mathbf{z}|i)}{p(\mathbf{z})}d\mathbf{z} =$$

$$= \sum_{i=1}^{n}\int q(i,\mathbf{z})\log\frac{q(i,\mathbf{z})}{p(\mathbf{z})p(i)}d\mathbf{z} = \int\sum_{i=1}^{n} q(i,\mathbf{z})\log\frac{q(\mathbf{z})q(i|\mathbf{z})}{p(\mathbf{z})p(i)}d\mathbf{z} =$$

$$= \int q(\mathbf{z})\log\frac{q(\mathbf{z})}{p(\mathbf{z})}d\mathbf{z} + \int\sum_{i=1}^{n} q(i|\mathbf{z})q(\mathbf{z})\log\frac{q(i|\mathbf{z})}{p(i)}d\mathbf{z} =$$

$$= KL(q(\mathbf{z})||p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})}\mathbb{H}\left[q(i|\mathbf{z})\right] + \log n.$$

---

http://approximateinference.org/accepted/HoffmanJohnson2016.pdf

ELBO surgery, 2016

Theorem

$$\frac{1}{n}\sum_{i=1}^{n} KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = KL(q(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}].$$

Proof (continued)

$$\frac{1}{n}\sum_{i=1}^{n} KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = KL(q(\mathbf{z})||p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})}\mathbb{H}\left[q(i|\mathbf{z})\right] + \log n$$

$$\mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}] = \mathbb{E}_{q(i,\mathbf{z})} \log \frac{q(i,\mathbf{z})}{q(i)q(\mathbf{z})} = \mathbb{E}_{q(\mathbf{z})}\mathbb{E}_{q(i|\mathbf{z})} \log \frac{q(i|\mathbf{z})q(\mathbf{z})}{q(i)q(\mathbf{z})} =$$

$$= \mathbb{E}_{q(\mathbf{z})}\mathbb{E}_{q(i|\mathbf{z})} \log \frac{q(i|\mathbf{z})}{q(i)} = -\mathbb{E}_{q(\mathbf{z})}\mathbb{H}\left[q(i|\mathbf{z})\right] + \log n.$$

---

http://approximateinference.org/accepted/HoffmanJohnson2016.pdf

ELBO surgery, 2016

Theorem

$$\frac{1}{n}\sum_{i=1}^{n} KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = KL(q(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}].$$

ELBO revisiting

$$\mathcal{L}(q, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left[\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) - KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) - \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}] - KL(q(\mathbf{z})||p(\mathbf{z})) =$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})}_{\text{Reconstruction loss}} - \underbrace{\left(\log n - \mathbb{E}_{q(\mathbf{z})}\mathbb{H}\left[q(i|\mathbf{z})\right]\right)}_{0 \leq \text{Mutual info} \leq \log n} - \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

---

http://approximateinference.org/accepted/HoffmanJohnson2016.pdf

ELBO surgery, 2016

ELBO revisiting

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})}_{\text{Reconstruction loss}} - \underbrace{\left(\log n - \mathbb{E}_{q(\mathbf{z})}\mathbb{H}\left[q(i|\mathbf{z})\right]\right)}_{0 \leq \text{Mutual info} \leq \log n} - \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

$$KL(q(\mathbf{z})||p(\mathbf{z})) = 0 \quad \Leftrightarrow \quad p(\mathbf{z}) = q(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i).$$

	ELBO	Avg. KL	Mutual info. 2	Marg. KL 3
2D latents	-129.63	7.41	7.20	0.21
10D latents	-88.95	19.17	10.82	8.35
20D latents	-87.45	20.2	10.67	9.53

$$\log n \approx 11.0021$$

http://approximateinference.org/accepted/HoffmanJohnson2016.pdf

VAE prior

ELBO revisiting

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})}_{\text{Reconstruction loss}} - \underbrace{\left(\log n - \mathbb{E}_{q(\mathbf{z})} \mathbb{H}\left[q(i|\mathbf{z})\right]\right)}_{0 \leq \text{Mutual info} \leq \log N} - \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

How to choose the optimal $p(\mathbf{z})$?

- SG: $p(\mathbf{z}) = \mathcal{N}(0, I) \Rightarrow$ over-regularization;
- MoG: $p(\mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) \Rightarrow$ (*), (**);
- $p(\mathbf{z}) = q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i) \Rightarrow$ overfitting and highly expensive.

(*) https://arxiv.org/abs/1611.02648

(**) https://pdfs.semanticscholar.org/f6fe/5e8e25994c188ba6a124462e2cc55f2c5a67.pdf

References

- **IAF**: Improving Variational Inference with Inverse Autoregressive Flow
  https://arxiv.org/abs/1606.04934
  Summary: Introduce inverse autoregressive flow (IAF). Models each autoregressive conditional as gaussian with autoregressive means and covariances. Inverse transformation allows to parallelize sampling.
- **MAF**: Masked Autoregressive Flow for Density Estimation
  https://arxiv.org/pdf/1705.07057.pdf
  Summary: Similar to IAF. Give comprehensive overview with link to IAF and RealNVP. MAF is suitable for density estimation, IAF as a recognition network.
- **Parallel WaveNet**: Fast High-Fidelity Speech Synthesis
  https://arxiv.org/pdf/1711.10433.pdf
  Summary: WaveNet is MAF (sequential generation). To exploit IAF fast sampling, knowledge distillation used. Teacher network is large WaveNet, student - is a IAF small WaveNet (generate samples from noise is parallel). The loss is KL divergence between student and teacher distributions. The additional perceptual, contrastive and power losses used to create more natural sounds.
- **ELBO surgery:** yet another way to carve up the variational evidence lower bound
  http://approximateinference.org/accepted/HoffmanJohnson2016.pdf
  Summary: Propose the decomposition of standard ELBO into 3 terms. The prior distribution should be close to average posterior. Show empirically that weak prior has a significant impact on ELBO value.