

Deep Generative Models

Lecture 7

Roman Isachenko



Ozon Masters

Spring, 2021

Dequantization

- ▶ Images are discrete data, pixels lie in the $[0, 255]$ integer domain (the model is $P(\mathbf{x}|\theta) = \text{Categorical}(\boldsymbol{\pi}(\theta))$).
- ▶ Flow is a continuous model (it works with continuous data \mathbf{x}).

By fitting a continuous density model to discrete data, one can produce a degenerate solution with all probability mass on discrete values.

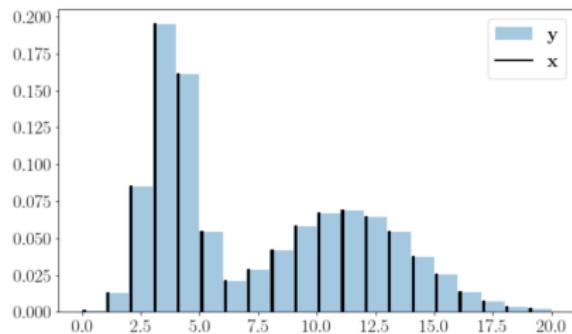
How to convert a discrete data distribution to a continuous one?

Uniform dequantization

$$\mathbf{x} \sim \text{Categorical}(\boldsymbol{\pi})$$

$$\mathbf{u} \sim U[0, 1]$$

$$\mathbf{y} = \mathbf{x} + \mathbf{u} \sim \text{Continuous}$$



Uniform dequantization

Statement

Fitting continuous model $p(\mathbf{y}|\theta)$ on uniformly dequantized data $\mathbf{y} = \mathbf{x} + \mathbf{u}$, $\mathbf{u} \sim U[0, 1]$ is equivalent to maximization of a lower bound on log-likelihood for a discrete model:

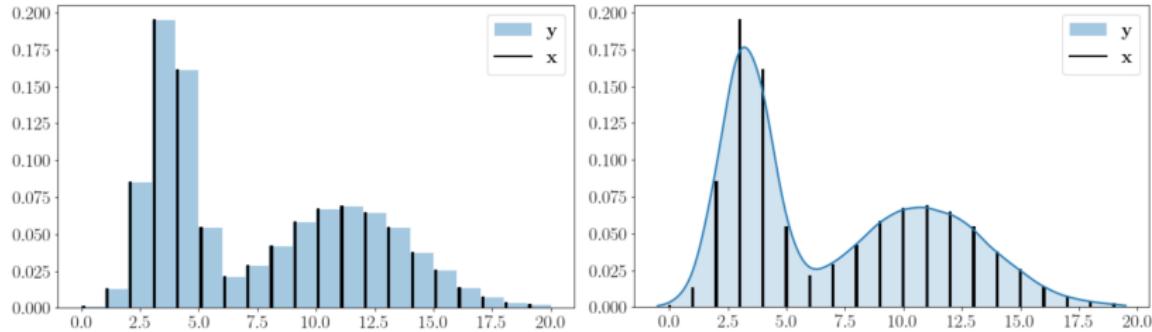
$$P(\mathbf{x}|\theta) = \int_{U[0,1]} p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u}$$

Thus, the maximisation of continuous model log-likelihood on \mathbf{y} can't lead to the a collapse onto the discrete data (the objective is bounded above by the discrete model log-likelihood).

Proof

$$\begin{aligned} \log P(\mathbf{x}|\theta) &= \log \int_{U[0,1]} p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u} \geq \\ &\geq \int_{U[0,1]} \log p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u} = \log p(\mathbf{y}|\theta). \end{aligned}$$

Variational dequantization



- ▶ $p(y|\theta)$ assign uniform density to unit hypercubes $x + U[0, 1]$ (left fig).
- ▶ Neural network density models are smooth function approximators (right fig).
- ▶ Smooth dequantization is more natural.

How to perform the smooth dequantization?

Flow++

Variational dequantization

Introduce variational dequantization noise distribution $q(\mathbf{u}|\mathbf{x})$ and treat it as an approximate posterior.

Variational lower bound

$$\begin{aligned}\log P(\mathbf{x}|\theta) &= \left[\log \int q(\mathbf{u}|\mathbf{x}) \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u} \right] \geq \\ &\geq \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u} = \mathcal{L}(q, \theta).\end{aligned}$$

Uniform dequantization bound

$$\begin{aligned}\log P(\mathbf{x}|\theta) &= \log \int_{U[0,1]} p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u} \geq \\ &\geq \int_{U[0,1]} \log p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u} = \log p(\mathbf{y}|\theta).\end{aligned}$$

Flow++

Variational lower bound

$$\mathcal{L}(q, \theta) = \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u}.$$

Let $\mathbf{u} = h(\epsilon, \phi)$ is a flow model with base distribution $\epsilon \sim p(\epsilon) = \mathcal{N}(0, \mathbf{I})$:

$$q(\mathbf{u}|\mathbf{x}) = p(h^{-1}(\mathbf{u}, \phi)) \cdot \left| \det \frac{\partial h^{-1}(\mathbf{u}, \phi)}{\partial \mathbf{u}} \right|.$$

Then

$$\log P(\mathbf{x}|\theta) \geq \mathcal{L}(\phi, \theta) = \int p(\epsilon) \log \left(\frac{p(\mathbf{x} + h(\epsilon, \phi)|\theta)}{p(\epsilon) \cdot \left| \det \frac{\partial h(\epsilon, \phi)}{\partial \epsilon} \right|^{-1}} \right) d\epsilon.$$

Variational lower

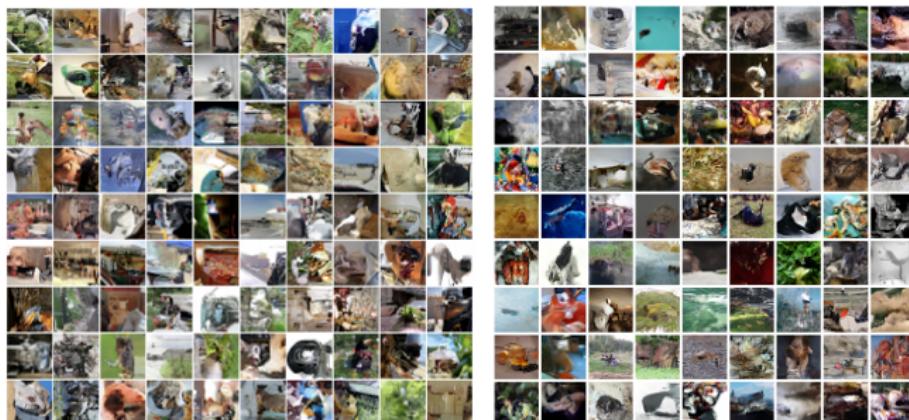
$$\log P(\mathbf{x}|\theta) \geq \int p(\epsilon) \log \left(\frac{p(\mathbf{x} + h(\epsilon, \phi))}{p(\epsilon) \cdot \left| \det \frac{\partial h(\epsilon, \phi)}{\partial \epsilon} \right|^{-1}} \right) d\epsilon.$$

- ▶ If $p(\mathbf{x} + \mathbf{u}|\theta)$ is also a flow model, it is straightforward to calculate stochastic gradient of this ELBO.
- ▶ Uniform dequantization is a special case of variational dequantization ($q(\mathbf{u}|\mathbf{x}) = U[0, 1]$). The gap between $\log P(\mathbf{x}|\theta)$ and the derived ELBO is $KL(q(\mathbf{u}|\mathbf{x})||p(\mathbf{u}|\mathbf{x}))$.
- ▶ In the case of uniform dequantization the model unnaturally places uniform density over each hypercube $\mathbf{x} + U[0, 1]$ due to inexpressive distribution q .

Flow++

Table 1. Unconditional image modeling results in bits/dim

Model family	Model	CIFAR10	ImageNet 32x32	ImageNet 64x64
Non-autoregressive	RealNVP (Dinh et al., 2016)	3.49	4.28	—
	Glow (Kingma & Dhariwal, 2018)	3.35	4.09	3.81
	IAF-VAE (Kingma et al., 2016)	3.11	—	—
	Flow++ (ours)	3.08	3.86	3.69
Autoregressive	Multiscale PixelCNN (Reed et al., 2017)	—	3.95	3.70
	PixelCNN (van den Oord et al., 2016b)	3.14	—	—
	PixelRNN (van den Oord et al., 2016b)	3.00	3.86	3.63
	Gated PixelCNN (van den Oord et al., 2016c)	3.03	3.83	3.57
	PixelCNN++ (Salimans et al., 2017)	2.92	—	—
	Image Transformer (Parmar et al., 2018)	2.90	3.77	—
	PixelSNAIL (Chen et al., 2017)	2.85	3.80	3.52



(a) PixelCNN

(b) Flow++

VAE limitations

- ▶ Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x})).$$

- ▶ Poor prior distribution

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

- ▶ Poor probabilistic model (decoder)

$$p(\mathbf{x}|\mathbf{z}, \theta) = \mathcal{N}(\mathbf{x}|\mu_\theta(\mathbf{z}), \sigma_\theta^2(\mathbf{z})).$$

- ▶ Loose lower bound

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

Importance Sampling

Generative model

$$\begin{aligned} p(\mathbf{x}|\theta) &= \int p(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z} = \int \left[\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] q(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\ &= \int f(\mathbf{x}, \mathbf{z}) q(\mathbf{z}|\mathbf{x}) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} f(\mathbf{x}, \mathbf{z}) \end{aligned}$$

Here $f(\mathbf{x}, \mathbf{z}) = \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})}$.

ELBO

$$\begin{aligned} \log p(\mathbf{x}|\theta) &= \log \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} f(\mathbf{x}, \mathbf{z}) \geq \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \log f(\mathbf{x}, \mathbf{z}) = \\ &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} = \mathcal{L}(q, \theta). \end{aligned}$$

Could we choose better $f(\mathbf{x}, \mathbf{z})$?

IWAE

Let define

$$f(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_K) = \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \theta)}{q(\mathbf{z}_k | \mathbf{x})}$$

$$\mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} f(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_K) = p(\mathbf{x} | \theta)$$

ELBO

$$\begin{aligned} \log p(\mathbf{x} | \theta) &= \log \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} f(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_K) \geq \\ &\geq \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} \log f(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_K) = \\ &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} \log \left[\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \theta)}{q(\mathbf{z}_k | \mathbf{x})} \right] = \mathcal{L}_K(q, \theta). \end{aligned}$$

IWAE

VAE objective

$$\log p(\mathbf{x}|\theta) \geq \mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \rightarrow \max_{\phi, \theta}$$

$$\mathcal{L}(q, \theta) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} \left(\frac{1}{K} \sum_{k=1}^K \log \frac{p(\mathbf{x}, \mathbf{z}_k|\theta)}{q(\mathbf{z}_k|\mathbf{x})} \right) \rightarrow \max_{q, \theta}.$$

IWAE objective

$$\mathcal{L}_K(q, \theta) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} \log \left(\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\theta)}{q(\mathbf{z}_k|\mathbf{x})} \right) \rightarrow \max_{q, \theta}.$$

If $K = 1$, these objectives coincide.

IWAE

Theorem

1. $\log p(\mathbf{x}|\theta) \geq \mathcal{L}_K(q, \theta) \geq \mathcal{L}_M(q, \theta)$, for $K \geq M$;
2. $\log p(\mathbf{x}|\theta) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \theta)$ if $\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})}$ is bounded.

Proof of 1.

$$\begin{aligned}\mathcal{L}_K(q, \theta) &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \log \left(\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \theta)}{q(\mathbf{z}_k | \mathbf{x})} \right) = \\ &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \log \mathbb{E}_{k_1, \dots, k_M} \left(\frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_{k_m} | \theta)}{q(\mathbf{z}_{k_m} | \mathbf{x})} \right) \geq \\ &\geq \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \mathbb{E}_{k_1, \dots, k_M} \log \left(\frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_{k_m} | \theta)}{q(\mathbf{z}_{k_m} | \mathbf{x})} \right) = \\ &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_M} \log \left(\frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_m | \theta)}{q(\mathbf{z}_m | \mathbf{x})} \right) = \mathcal{L}_M(q, \theta)\end{aligned}$$

$$\frac{a_1 + \dots + a_K}{K} = \mathbb{E}_{i_1, \dots, i_M} \frac{a_{i_1} + \dots + a_{i_M}}{M}, \quad i_1, \dots, i_M \sim U[1, K]$$

Theorem

1. $\log p(\mathbf{x}|\theta) \geq \mathcal{L}_K(q, \theta) \geq \mathcal{L}_M(q, \theta)$, for $K \geq M$;
2. $\log p(\mathbf{x}|\theta) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \theta)$ if $\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})}$ is bounded.

Proof of 2.

Consider r.v. $\xi_K = \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\theta)}{q(\mathbf{z}_k|\mathbf{x})}$.

If summands are bounded, then (from the strong law of large numbers)

$$\xi_K \xrightarrow[K \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} = p(\mathbf{x}|\theta).$$

Hence $\mathcal{L}_K(q, \theta) = \mathbb{E} \log \xi_K$ converges to $\log p(\mathbf{x}|\theta)$ as $K \rightarrow \infty$.

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}_K(q, \boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

If $K > 1$ the bound could be tighter.

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})};$$

$$\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} \log \left(\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x})} \right).$$

- ▶ $\mathcal{L}_1(q, \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta})$;
- ▶ $\mathcal{L}_\infty(q, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})$.

Which $q(\mathbf{z}|\mathbf{x})$ gives $\mathcal{L}(q, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})$?

Which $q(\mathbf{z}|\mathbf{x})$ gives $\mathcal{L}(q, \boldsymbol{\theta}) = \mathcal{L}_K(q, \boldsymbol{\theta})$?

IWAE

Theorem

The VAE objective is equal to IWAE objective

$$\mathcal{L}(q_{EW}, \theta) = \mathcal{L}_K(q, \theta)$$

for the following variational distribution

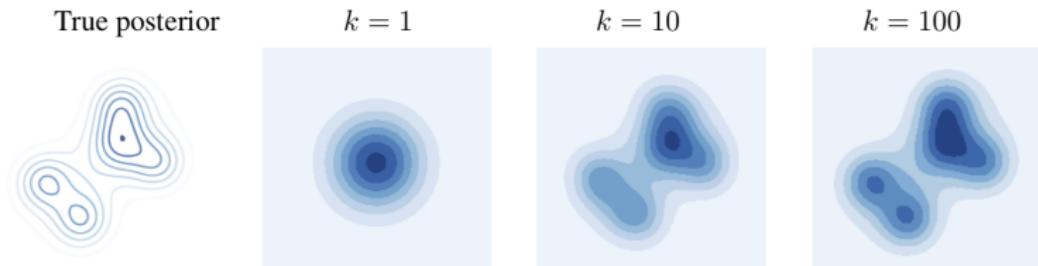
$$q_{EW}(\mathbf{z}|\mathbf{x}) = \mathbb{E}_{\mathbf{z}_2, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} q_{IW}(\mathbf{z}|\mathbf{x}, \mathbf{z}_{2:K}),$$

where

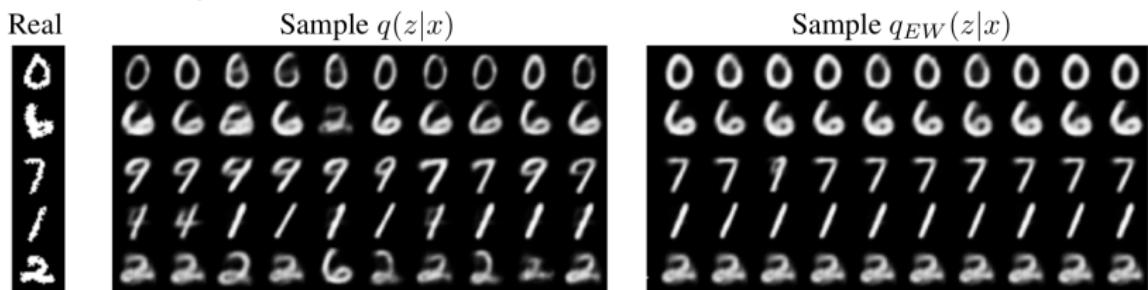
$$q_{IW}(\mathbf{z}|\mathbf{x}, \mathbf{z}_{2:K}) = \frac{\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})}}{\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})}} q(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{\frac{1}{K} \left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \sum_{k=2}^K \frac{p(\mathbf{x}, \mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})} \right)}.$$

IWAE

IWAE posterior



IWAE samples



IWAE

Objective

$$\mathcal{L}_K(q, \theta) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \phi)} \log \left(\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \theta)}{q(\mathbf{z}_k | \mathbf{x}, \phi)} \right) \rightarrow \max_{\phi, \theta} .$$

Gradient

$$\Delta_K = \nabla_{\theta, \phi} \log \left(\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \theta)}{q(\mathbf{z}_k | \mathbf{x}, \phi)} \right), \quad \mathbf{z}_k \sim q(\mathbf{z} | \mathbf{x}, \phi).$$

Theorem

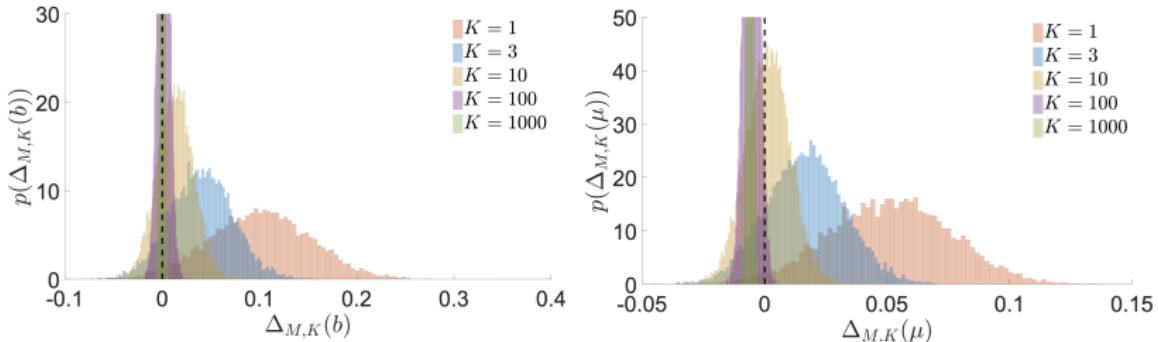
$$\text{SNR}_K = \frac{\mathbb{E}[\Delta_K]}{\sigma(\Delta_K)}; \quad \text{SNR}_K(\theta) = O(\sqrt{K}); \quad \text{SNR}_K(\phi) = O\left(\sqrt{\frac{1}{K}}\right).$$

Hence, increasing K vanishes gradient signal of inference network $q(\mathbf{z} | \mathbf{x}, \phi)$.

IWAE

Theorem

$$\text{SNR}_K = \frac{\mathbb{E}[\Delta_K]}{\sigma(\Delta_K)}; \quad \text{SNR}_K(\theta) = O(\sqrt{K}); \quad \text{SNR}_K(\phi) = O\left(\sqrt{\frac{1}{K}}\right).$$



- ▶ IWAE makes the variational bound tighter and extends the class of variational distributions.
- ▶ Gradient signal becomes really small, training is complicated.
- ▶ IWAE is very popular technique as a quality measure for VAE models.

VAE limitations

- ▶ Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x})).$$

- ▶ Poor prior distribution

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

- ▶ Poor probabilistic model (decoder)

$$p(\mathbf{x}|\mathbf{z}, \theta) = \mathcal{N}(\mathbf{x}|\mu_\theta(\mathbf{z}), \sigma_\theta^2(\mathbf{z})).$$

- ▶ Loose lower bound

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

ELBO interpretations

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) + KL(q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})) \geq \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}).$$

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

- ▶ Evidence minus posterior KL

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

- ▶ Average negative energy plus entropy

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} [\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) + \mathbb{H}[q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})].\end{aligned}$$

- ▶ Average reconstruction minus KL to prior

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} [\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - KL(q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})||p(\mathbf{z})).\end{aligned}$$

ELBO surgery

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

Theorem

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

- ▶ $q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i)$ – **aggregated** posterior distribution.
- ▶ $\mathbb{I}_q[\mathbf{x}, \mathbf{z}]$ – mutual information between \mathbf{x} and \mathbf{z} under empirical data distribution and distribution $q(\mathbf{z}|\mathbf{x})$.
- ▶ First term pushes $q(\mathbf{z})$ towards the prior $p(z)$.
- ▶ Second term reduces the amount of information about \mathbf{x} stored in \mathbf{z} .

ELBO surgery

Theorem

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

Proof

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

ELBO surgery

Theorem

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

Proof (continued)

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

It could be shown (exercise):

$$\mathbb{I}_q[\mathbf{x}, \mathbf{z}] = \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || q(\mathbf{z})) \in [0, \log n].$$

ELBO surgery

Theorem

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

ELBO revisiting

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) &= \frac{1}{n} \sum_{i=1}^n [\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}_i) || p(\mathbf{z}))] = \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta)}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_q[\mathbf{x}, \mathbf{z}]}_{0 \leq \text{MI} \leq \log n} - \underbrace{KL(q(\mathbf{z}) || p(\mathbf{z}))}_{\text{Marginal KL}}\end{aligned}$$

Prior distribution $p(\mathbf{z})$ is only in the last term.

ELBO surgery

ELBO revisiting

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

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

The optimal prior distribution $p(\mathbf{z})$ is aggregated posterior $q(\mathbf{z})$.

	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

$$\log n \approx 11.0021$$

VAE prior

ELBO revisiting

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

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

- ▶ SG: $p(\mathbf{z}) = \mathcal{N}(0, I)$ \Rightarrow over-regularization;
- ▶ MoG: $p(\mathbf{z}|\lambda) = \frac{1}{K} \sum_{k=1}^K \mathcal{N}(\mu_k, \sigma_k^2)$ \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.

Note that now we consider cases of parametrized priors.

(*) Dilokthanakul N. et al. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders, 2016

(**) Nalisnick E., Hertel L., Smyth P. Approximate Inference for Deep Latent Gaussian Mixtures, 2016

VampPrior

Variational Mixture of posteriors

$$p(\mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{K} \sum_{k=1}^K q(\mathbf{z}|\mathbf{u}_k),$$

where $\boldsymbol{\lambda} = \{\mathbf{u}_1, \dots, \mathbf{u}_K\}$ are trainable pseudo-inputs.

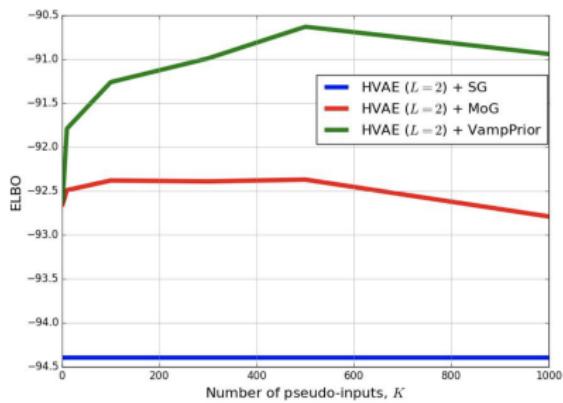
- ▶ Multimodal \Rightarrow prevents over-regularization;.
- ▶ $K \ll n \Rightarrow$ prevents from potential overfitting + less expensive to train.
- ▶ Pseudo-inputs are prior hyperparameters \Rightarrow connection to the Empirical Bayes.

VampPrior

- ▶ Do we equally need the multimodal prior?
- ▶ Is it beneficial to couple the prior with the variational posterior or MoG is enough?

MNIST results

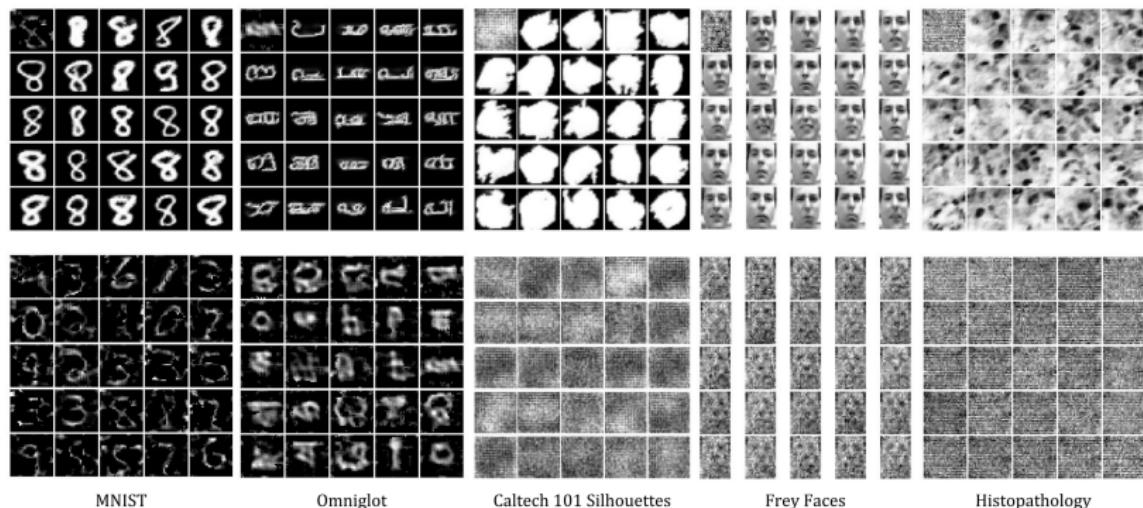
MODEL	LL
VAE ($L = 1$) + NF [32]	-85.10
VAE ($L = 2$) [6]	-87.86
IWAE ($L = 2$) [6]	-85.32
HVAE ($L = 2$) + SG	-85.89
HVAE ($L = 2$) + MoG	-85.07
HVAE ($L = 2$) + VAMPPIOR data	-85.71
HVAE ($L = 2$) + VAMPPIOR	-83.19
AVB + AC ($L = 1$) [28]	-80.20
VLAЕ [7]	-79.03
VAE + IAF [18]	-79.88
CONVHVAE ($L = 2$) + VAMPPIOR	-81.09
PIXELHVAE ($L = 2$) + VAMPPIOR	-79.78



VampPrior

Top row: generated images by PixelHVAE + VampPrior for chosen pseudo-input in the left top corner.

Bottom row: pseudo-inputs for different datasets.



MNIST

Omniglot

Caltech 101 Silhouettes

Frey Faces

Histopathology

Flow prior in VAE

ELBO revisiting

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

VampPrior

$$p(\mathbf{z}|\lambda) = \frac{1}{K} \sum_{k=1}^K q(\mathbf{z}|\mathbf{u}_k),$$

where $\mathbf{u}_1, \dots, \mathbf{u}_K$ is trainable preudo-inputs.

Autoregressive flow prior

$$\log p(\mathbf{z}|\lambda) = \log p(\epsilon) + \log \det \left| \frac{d\epsilon}{d\mathbf{z}} \right|$$

$$\mathbf{z} = g(\epsilon, \lambda) = f^{-1}(\epsilon, \lambda)$$

Variational Lossy AutoEncoder, 2016

Theorem

VAE with the AF prior for latent code \mathbf{z} is equivalent to using the IAF posterior for latent code ϵ .

Proof

$$\begin{aligned}\mathcal{L}(q, \theta) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(\mathbf{z}, \lambda) - q(\mathbf{z}|\mathbf{x})] \\ &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \underbrace{\left(\log p(f(\mathbf{z}, \lambda)) + \log \left| \det \frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right| \right)}_{\text{AF prior}} - q(\mathbf{z}|\mathbf{x}) \right] \\ &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(f(\mathbf{z}, \lambda)) - \underbrace{\left(q(\mathbf{z}|\mathbf{x}) - \log \left| \det \frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right| \right)}_{\text{IAF posterior}} \right]\end{aligned}$$

Flows in VAE

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(\mathbf{z}_0|\mathbf{x}, \phi)} \left[\log p(\mathbf{x}, \mathbf{z}_K | \theta) - \log q(\mathbf{z}_0 | \mathbf{x}, \phi) + \log \left| \det \left(\frac{\partial g(\mathbf{z}_0, \phi_*)}{\partial \mathbf{z}_0} \right) \right| \right].$$

Variational Lossy AutoEncoder, 2016

Autoregressive flow prior

$$\begin{aligned}\mathcal{L}(q, \theta) &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \underbrace{\left(\log p(f(\mathbf{z}, \lambda)) + \log \left| \det \frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right| \right)}_{\text{AF prior}} - q(\mathbf{z}|\mathbf{x}) \right] \\ &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(f(\mathbf{z}, \lambda)) - \underbrace{\left(q(\mathbf{z}|\mathbf{x}) - \log \left| \det \frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right| \right)}_{\text{IAF posterior}} \right]\end{aligned}$$

- ▶ IAF posterior decoder path: $p(\mathbf{x}|\mathbf{z}, \theta)$, $\mathbf{z} \sim p(\mathbf{z})$.
- ▶ AF prior decoder path: $p(\mathbf{x}|\mathbf{z}, \theta)$, $\mathbf{z} = g(\epsilon, \lambda)$, $\epsilon \sim p(\epsilon)$.

The AF prior and the IAF posterior have the same computation cost, so using the AF prior makes the model more expressive at no training time cost.

Summary

- ▶ Uniform dequantization is the simplest form of dequantization. Variational dequantization is a more natural type that was proposed in Flow++ model.
- ▶ The importance sampling could get the tighter lower bound to the likelihood.
- ▶ The ELBO surgery reveals insights about a prior distribution in VAE. The optimal prior is the aggregated posterior.
- ▶ VampPrior proposes to use a variational mixture of posteriors as the prior to approximate the aggregated posterior.
- ▶ The autoregressive flows could be used as the prior. This is equivalent to the use of the IAF posterior.