UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

CS 598 BAN
VAE 1

Shengyu Feng

Sep 16, 2021

Outline

Introduction & Related Works

Motivation

Methods

Experimental Results

Outline

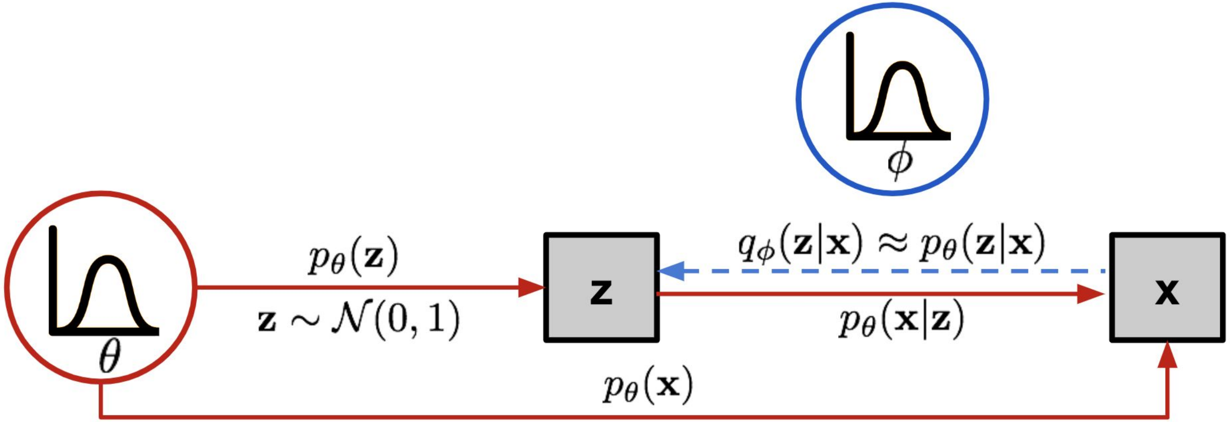
Introduction & Related Works

Motivation

Methods

Experimental Results

VAE review


$q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$
ϕ
$p_\theta(\mathbf{z})$
$\mathbf{z} \sim \mathcal{N}(0,1)$
θ
z
$p_\theta(\mathbf{x}|\mathbf{z})$
x
$p_\theta(\mathbf{x})$

ELBO review

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x)\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{p_\theta(z|x)}\right)\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\frac{q_\phi(z|x)}{p_\theta(z|x)}\right)\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\right)\right]}_{\mathcal{L}_{\theta,\phi}(x)\ \text{(ELBO)}} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{q_\phi(z|x)}{p_\theta(z|x)}\right)\right]}_{D_{KL}(q_\phi(z|x)\|p_\theta(z|x))}
\end{aligned}
$$

$$
\mathcal{L}_{\theta,\phi}(x) = -D_{KL}(q_\phi(z|x)\|p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]
$$

Outline

Introduction & Related Works

Motivation

Methods

Experimental results

Bias in ELBO

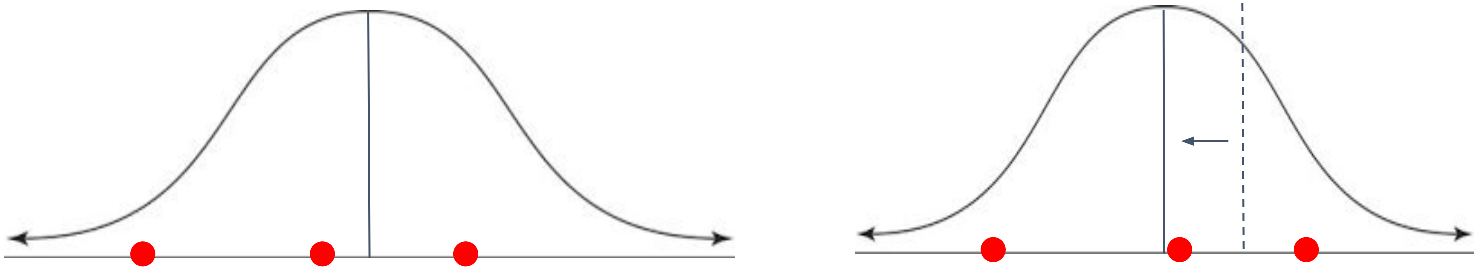
$$\log p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x)]$$
$$= \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{p_\theta(z|x)}\right)\right]$$
$$= \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\frac{q_\phi(z|x)}{p_\theta(z|x)}\right)\right]$$
$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\right)\right]}_{\mathcal{L}_{\theta,\phi}(x)\ \text{(ELBO)}} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{q_\phi(z|x)}{p_\theta(z|x)}\right)\right]}_{D_{KL}(q_\phi(z|x)\|p_\theta(z|x))}$$

Bias exists

$$\mathcal{L}_{\theta,\phi}(x) = -D_{KL}(q_\phi(z|x)\|p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$$

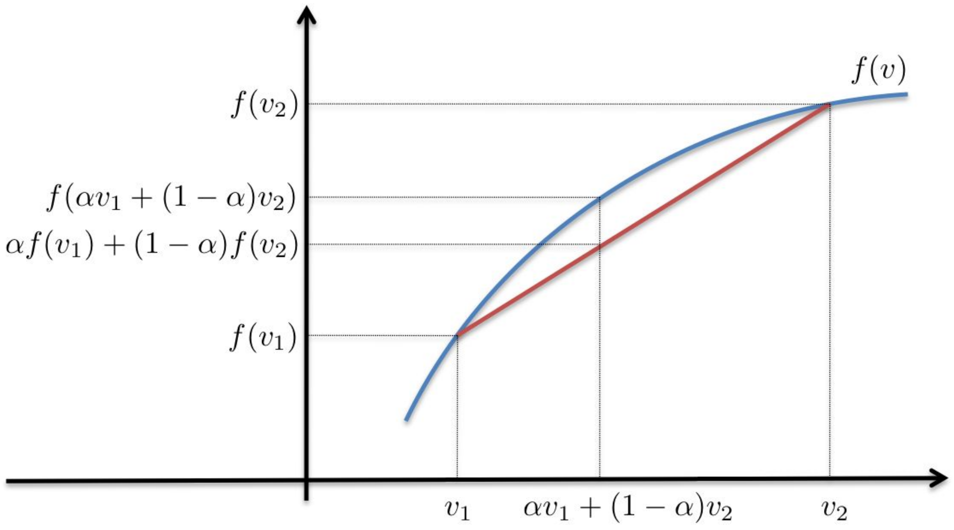
Penalization from the regularization term

$$D_{KL}(q_\phi(z|x)\|p_\theta(z))$$



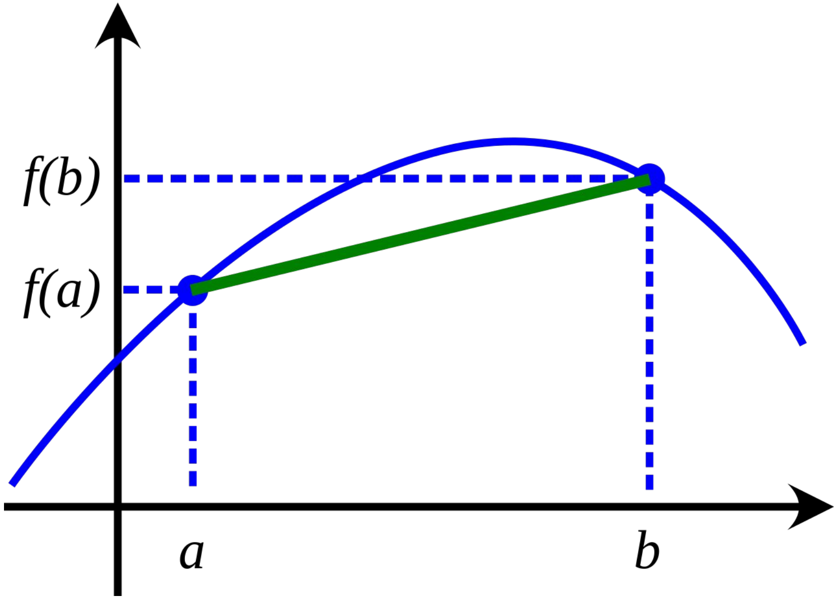
Another derivation: Jensen Inequality

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right] \geq \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right] = \mathcal{L}(\mathbf{x}).$$



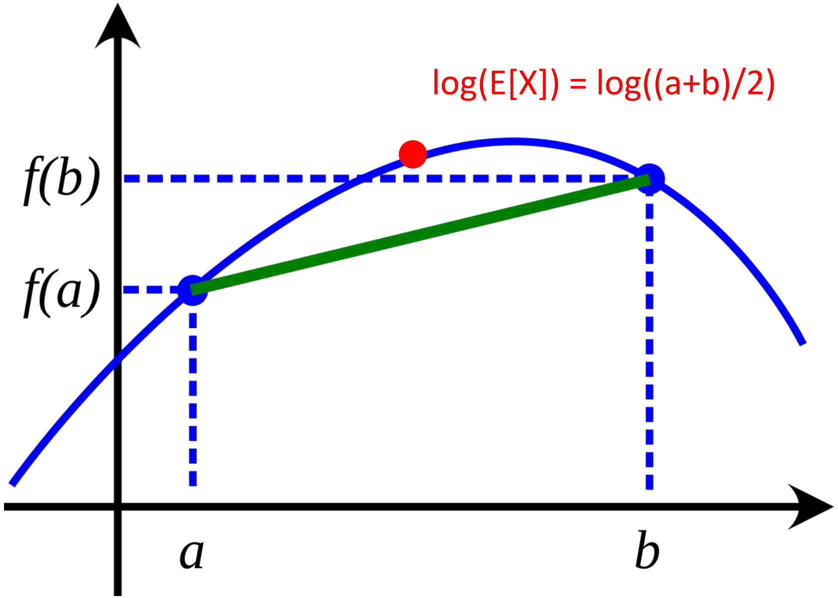
Example: log expectation estimation

Consider a simple case: X sampled from {a, b}, want to estimate log(E[X])



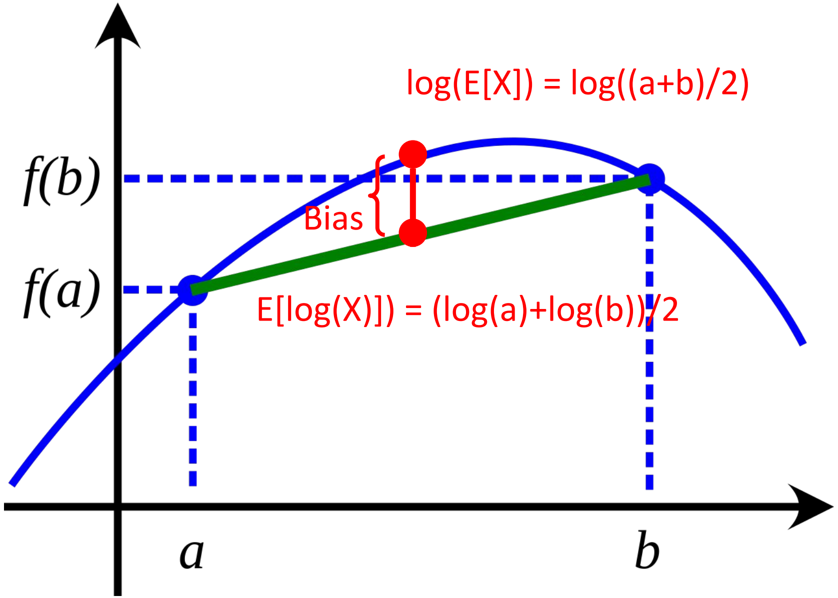
Example: log expectation estimation

Consider a simple case: X sampled from {a, b}, want to estimate log(E[X])



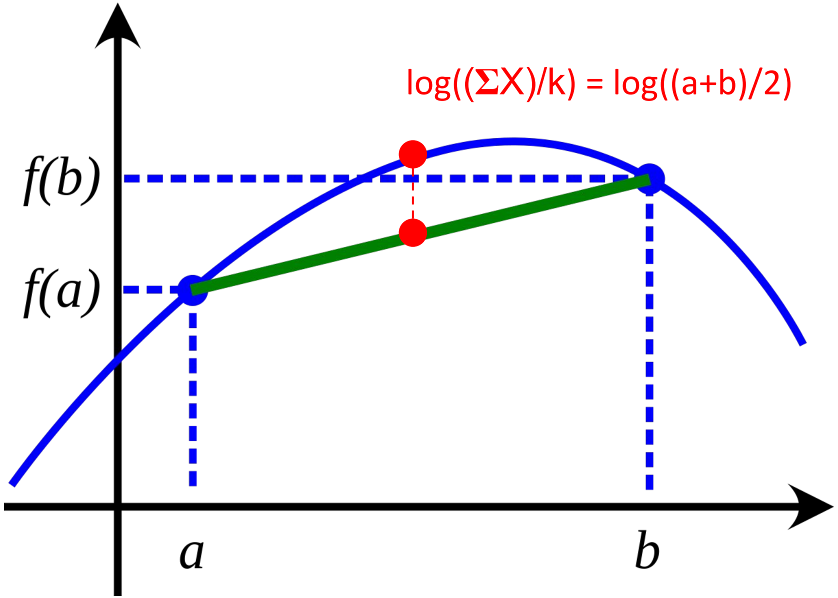
Example: log expectation estimation

Using E[log(X)] to estimate log(E[X]), assuming convergence



Example: log expectation estimation

Using E[log((ΣX)/k)] to estimate log(E[X]), assuming convergence

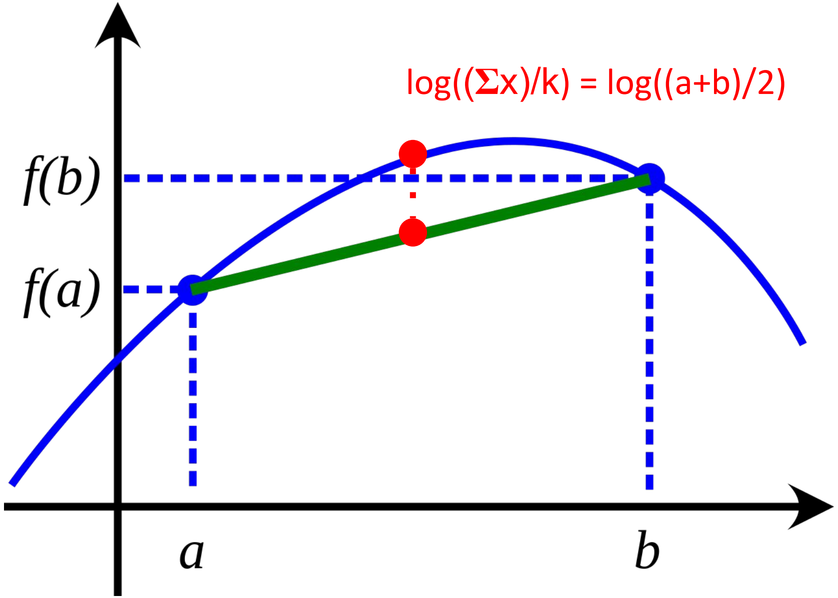


Example: log expectation estimation

Theorem 1. *For all k, the lower bounds satisfy*

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k.$$

Moreover, if $p(\mathbf{h}, \mathbf{x})/q(\mathbf{h}|\mathbf{x})$ is bounded, then $\mathcal{L}_k$ approaches $\log p(\mathbf{x})$ as k goes to infinity.



Outline

Introduction & Related Works

Motivation

Methods

Experimental Results

Multiple stochastic layers

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{h}^1,\ldots,\mathbf{h}^L} p(\mathbf{h}^L|\boldsymbol{\theta})p(\mathbf{h}^{L-1}|\mathbf{h}^L,\boldsymbol{\theta})\cdots p(\mathbf{x}|\mathbf{h}^1,\boldsymbol{\theta}).$$

$$q(\mathbf{h}|\mathbf{x}) = q(\mathbf{h}^1|\mathbf{x})q(\mathbf{h}^2|\mathbf{h}^1)\cdots q(\mathbf{h}^L|\mathbf{h}^{L-1}),$$

Reparameterization trick

$$\mathbf{h}^\ell(\boldsymbol{\epsilon}^\ell,\mathbf{h}^{\ell-1},\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\mathbf{h}^{\ell-1},\boldsymbol{\theta})^{1/2}\boldsymbol{\epsilon}^\ell + \boldsymbol{\mu}(\mathbf{h}^{\ell-1},\boldsymbol{\theta}).$$

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \log \mathbb{E}_{\mathbf{h}\sim q(\mathbf{h}|\mathbf{x},\boldsymbol{\theta})}\left[\frac{p(\mathbf{x},\mathbf{h}|\boldsymbol{\theta})}{q(\mathbf{h}|\mathbf{x},\boldsymbol{\theta})}\right] &= \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\epsilon}^1,\ldots,\boldsymbol{\epsilon}^L\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\log\frac{p(\mathbf{x},\mathbf{h}(\boldsymbol{\epsilon},\mathbf{x},\boldsymbol{\theta})|\boldsymbol{\theta})}{q(\mathbf{h}(\boldsymbol{\epsilon},\mathbf{x},\boldsymbol{\theta})|\mathbf{x},\boldsymbol{\theta})}\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}^1,\ldots,\boldsymbol{\epsilon}^L\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\nabla_{\boldsymbol{\theta}}\log\frac{p(\mathbf{x},\mathbf{h}(\boldsymbol{\epsilon},\mathbf{x},\boldsymbol{\theta})|\boldsymbol{\theta})}{q(\mathbf{h}(\boldsymbol{\epsilon},\mathbf{x},\boldsymbol{\theta})|\mathbf{x},\boldsymbol{\theta})}\right].\end{aligned}$$

Methods

Importance weighted ELBO:

$$\mathcal{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})}\left[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(\mathbf{x},\mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{x})}\right].$$

$$\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right]$$

ELBO in VAE

Gradient calculation:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_k(\mathbf{x}) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k}\left[\log \frac{1}{k}\sum_{i=1}^{k} w_i\right] &= \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k}\left[\log \frac{1}{k}\sum_{i=1}^{k} w(\mathbf{x},\mathbf{h}(\mathbf{x},\boldsymbol{\epsilon}_i,\boldsymbol{\theta}),\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k}\left[\nabla_{\boldsymbol{\theta}} \log \frac{1}{k}\sum_{i=1}^{k} w(\mathbf{x},\mathbf{h}(\mathbf{x},\boldsymbol{\epsilon}_i,\boldsymbol{\theta}),\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k}\left[\sum_{i=1}^{k} \widetilde{w}_i \nabla_{\boldsymbol{\theta}} \log w(\mathbf{x},\mathbf{h}(\mathbf{x},\boldsymbol{\epsilon}_i,\boldsymbol{\theta}),\boldsymbol{\theta})\right],\end{aligned}$$

Importance weights, 1 in VAE $\tilde{w}_i = w_i / \sum_{j=1}^{k} w_j$

Comparison of the objectives

Importance weighted ELBO$_{IWAE}$

$$\mathcal{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})}\left[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p(\mathbf{x},\mathbf{h}_i)}{q(\mathbf{h}_i \mid \mathbf{x})}\right]$$

Mini batch ELBO (ELBO$_{VAE}$)

$$\mathcal{L}_{minibatch}(\mathbf{x}) = \frac{1}{\boxed{k}}\sum_{i=1}^{k}\left[\log \frac{p(\mathbf{x},\mathbf{h}_i)}{q(\mathbf{h}_i \mid \mathbf{x})}\right], \qquad \mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h} \mid \mathbf{x})$$

M in MIWAE, CIWAE and PIWAE

Combination of both

- **MIWAE (M>1, K>1)**

$$\Delta_{M,K} := \frac{1}{M} \sum_{m=1}^{M} \nabla_{\theta,\phi} \log \frac{1}{K} \sum_{k=1}^{K} w_{m,k},$$

- **CIWAE**

$$\mathrm{ELBO}_{\mathrm{CIWAE}} = \beta\, \mathrm{ELBO}_{\mathrm{VAE}} + (1-\beta)\, \mathrm{ELBO}_{\mathrm{IWAE}}$$

- **PIWAE:**

$$\Delta^{\mathrm{C}}_{K,\beta}(\theta) = \nabla_{\theta} \log \frac{1}{K} \sum_{k=1}^{K} w_k$$
$$\Delta^{\mathrm{C}}_{M,K,\beta}(\phi) = \frac{1}{M} \sum_{m=1}^{M} \nabla_{\phi} \log \frac{1}{L} \sum_{\ell=1}^{L} w_{m,\ell}$$

Outline

Introduction & Related Works

Motivation

Methods

Experimental Results

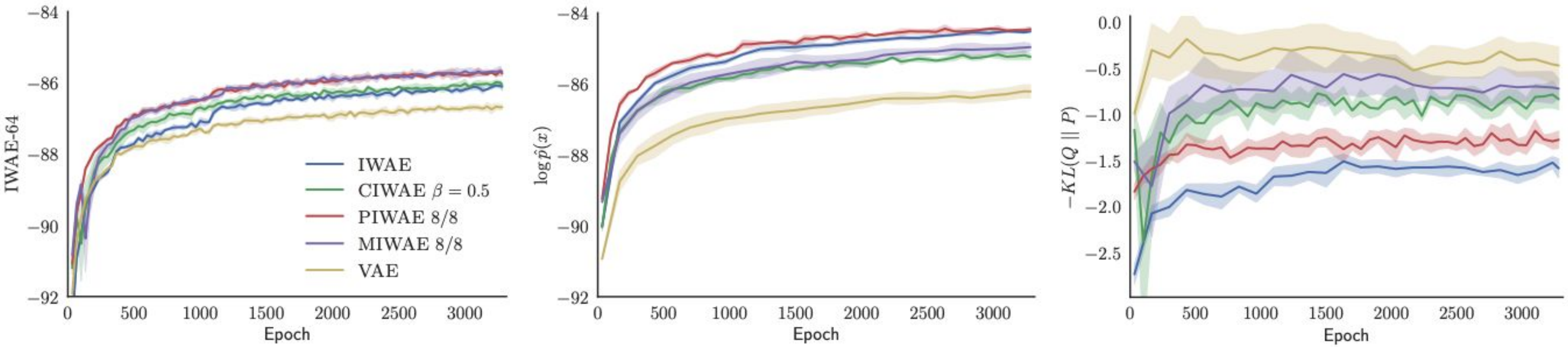
Experimental results for density estimation

		MNIST				OMNIGLOT			
		VAE		IWAE		VAE		IWAE	
# stoch. layers	k	NLL	active units	NLL	active units	NLL	active units	NLL	active units
1	1	86.76	19	86.76	19	108.11	28	108.11	28
	5	86.47	20	85.54	22	107.62	28	106.12	34
	50	86.35	20	84.78	25	107.80	28	104.67	41
2	1	85.33	16+5	85.33	16+5	107.58	28+4	107.56	30+5
	5	85.01	17+5	83.89	21+5	106.31	30+5	104.79	38+6
	50	84.78	17+5	82.90	26+7	106.30	30+5	103.38	44+7

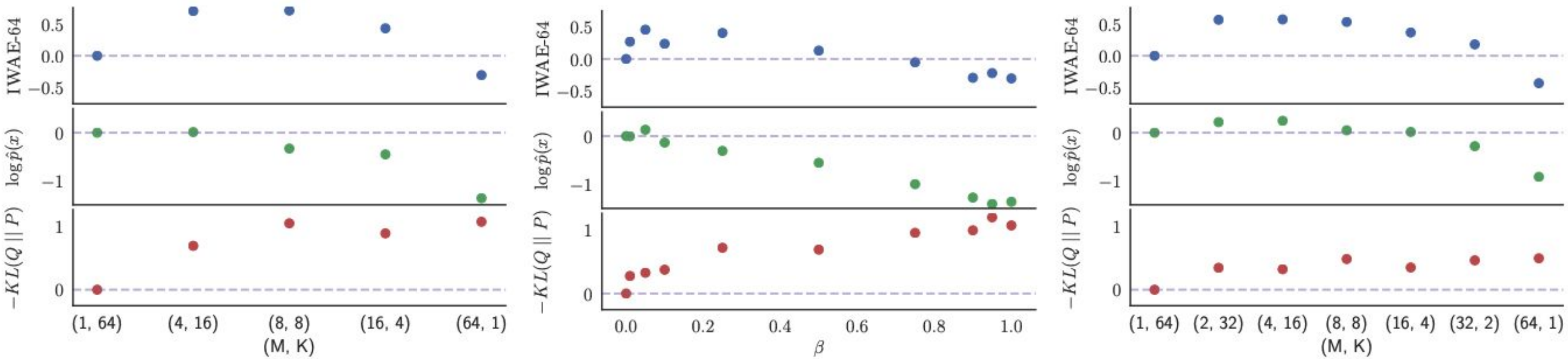
Different objectives in different stages

	First stage			Second stage		
	trained as	NLL	active units	trained as	NLL	active units
Experiment 1	VAE	86.76	19	IWAE, $k = 50$	84.88	22
Experiment 2	IWAE, $k = 50$	84.78	25	VAE	86.02	23

Experimental results for IWAE variants

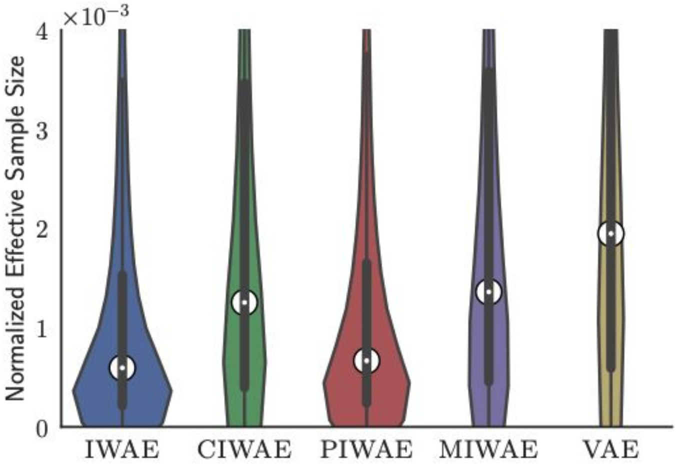


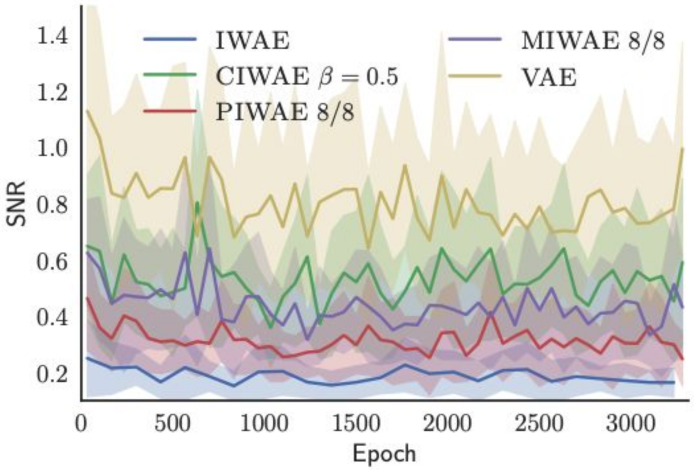
(a) IWAE_{64} (b) $\log \hat{p}(x)$ (c) $-\mathrm{KL}(Q_\phi(z|x)||P_\theta(z|x))$



Experimental results for IWAE variants

Metric	IWAE	PIWAE $(4, 16)$	PIWAE $(8, 8)$	MIWAE $(4, 16)$	MIWAE $(8, 8)$	CIWAE $\beta = 0.05$	CIWAE $\beta = 0.5$	VAE
IWAE-64	-86.11 ± 0.10	$\mathbf{-85.68 \pm 0.06}$	$\mathbf{-85.74 \pm 0.07}$	$\mathbf{-85.60 \pm 0.07}$	$\mathbf{-85.69 \pm 0.04}$	-85.91 ± 0.11	-86.08 ± 0.08	-86.69 ± 0.08
$\log \hat{p}(x)$	$\mathbf{-84.52 \pm 0.02}$	$\mathbf{-84.40 \pm 0.17}$	$\mathbf{-84.46 \pm 0.06}$	$\mathbf{-84.56 \pm 0.05}$	-84.97 ± 0.10	$\mathbf{-84.57 \pm 0.09}$	-85.24 ± 0.08	-86.21 ± 0.19
$-\mathrm{KL}(Q\|\|P)$	-1.59 ± 0.10	-1.27 ± 0.18	-1.28 ± 0.09	$\mathbf{-1.04 \pm 0.08}$	$\mathbf{-0.72 \pm 0.11}$	-1.34 ± 0.14	$\mathbf{-0.84 \pm 0.11}$	$\mathbf{-0.47 \pm 0.20}$





References

- 2014 (ICLR): D. Kingma, M. Welling, Auto-Encoding Variational Bayes, ICLR, 2014.
- 2016 (ICLR): Y. Burda, R. Grosse, R. Salakhutdinov. Importance Weighted Autoencoders. ICLR, 2016.
- 2018 (ICML): T. Rainforth, A. Kosiorek, T. Le, C. Maddison, M. Igl, F. Wood, Y. Teh, Tighter Variational Bounds are Not Necessarily Better. ICML, 2018.
- Lil'Log: From Autoencoder to Beta-VAE
- Dustin Tran: Importance weighted autoencoders