Nonparametric Deep Generative Models with Stick-Breaking Priors

Eric Nalisnick
University of California, Irvine

In collaboration with

Padhraic Smyth

Motivation: The Variational Inference Pipeline

Motivation: The Variational Inference Pipeline

1. Write Model in Terms of the Exponential Family

Motivation: The Variational Inference Pipeline

1. Write Model in Terms of the Exponential Family

2. Derive Coordinate Ascent Updates

Motivation: The Variational Inference Pipeline

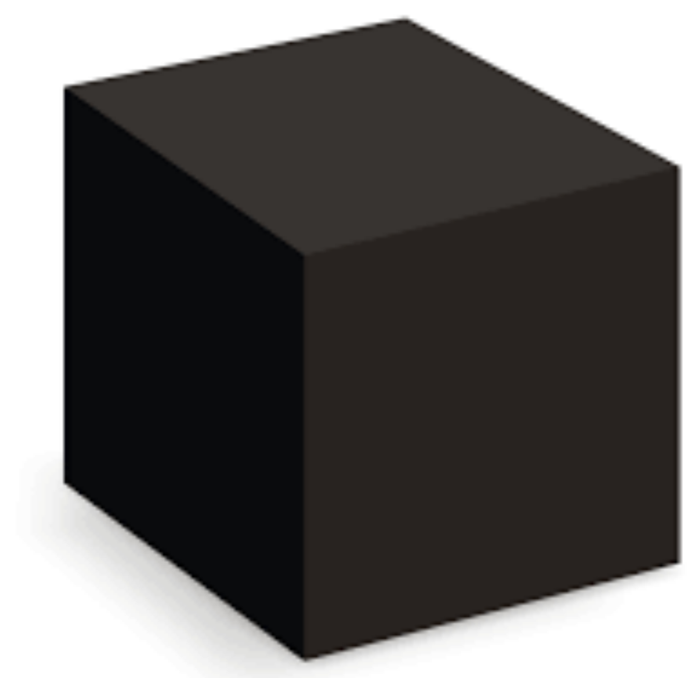
1. Write Model in Terms of the Exponential Family
2. Derive Coordinate Ascent Updates
3. Write Conference Paper

Motivation: The Variational Inference Pipeline

1. Write Model in Terms of the Exponential Family
2. Derive Coordinate Ascent Updates
3. Write Conference Paper
4. Repeat

Motivation: The Variational Inference Pipeline

1. ~~Write Model in Terms of the Exponential Family~~
2. ~~Derive Coordinate Ascent Updates~~
3. ~~Write Conference Paper~~
4. ~~Repeat~~



Software: Stan, Edward…

Black Box Variational Inference

Latent variable for which we want posterior

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \mathbb{E}_q[\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)] - KL(q_{\boldsymbol{\phi}}(\mathbf{z}_i)||p(\mathbf{z}_i))$$

Black Box Variational Inference

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \mathbb{E}_q[\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)] - KL(q_{\boldsymbol{\phi}}(\mathbf{z}_i)||p(\mathbf{z}_i))$$

Stochastic Gradient Variational Bayes (SGVB) Estimator:

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)

$$\approx \frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_{i,s}) - KL(q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))$$

Black Box Variational Inference

$$\log p_{\theta}(\mathbf{x}_i) \geq \mathbb{E}_q[\log p_{\theta}(\mathbf{x}_i|\mathbf{z}_i)] - KL(q_{\phi}(\mathbf{z}_i)||p(\mathbf{z}_i))$$

Stochastic Gradient Variational Bayes (SGVB) Estimator:

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)

$$\approx \underbrace{\frac{1}{S}\sum_{s=1}^{S} \log p_{\theta}(\mathbf{x}_i|\mathbf{z}_{i,s})}_{\text{Monte Carlo Expectation (relieves conjugacy constraints)}} - KL(q_{\phi}(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))$$

Black Box Variational Inference

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \mathbb{E}_q[\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)] - KL(q_{\boldsymbol{\phi}}(\mathbf{z}_i)||p(\mathbf{z}_i))$$

Stochastic Gradient Variational Bayes (SGVB) Estimator:

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)

$$\approx \underbrace{\frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_{i,s})}_{\text{Monte Carlo Expectation (relieves conjugacy constraints)}} - KL(q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))$$

Gradients can be taken through MC samples into **z**'s parameters via a non-centered representation

Black Box Variational Inference

$$\log p_{\theta}(\mathbf{x}_i) \geq \mathbb{E}_q[\log p_{\theta}(\mathbf{x}_i|\mathbf{z}_i)] - KL(q_{\phi}(\mathbf{z}_i)||p(\mathbf{z}_i))$$

Stochastic Gradient Variational Bayes (SGVB) Estimator:

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)

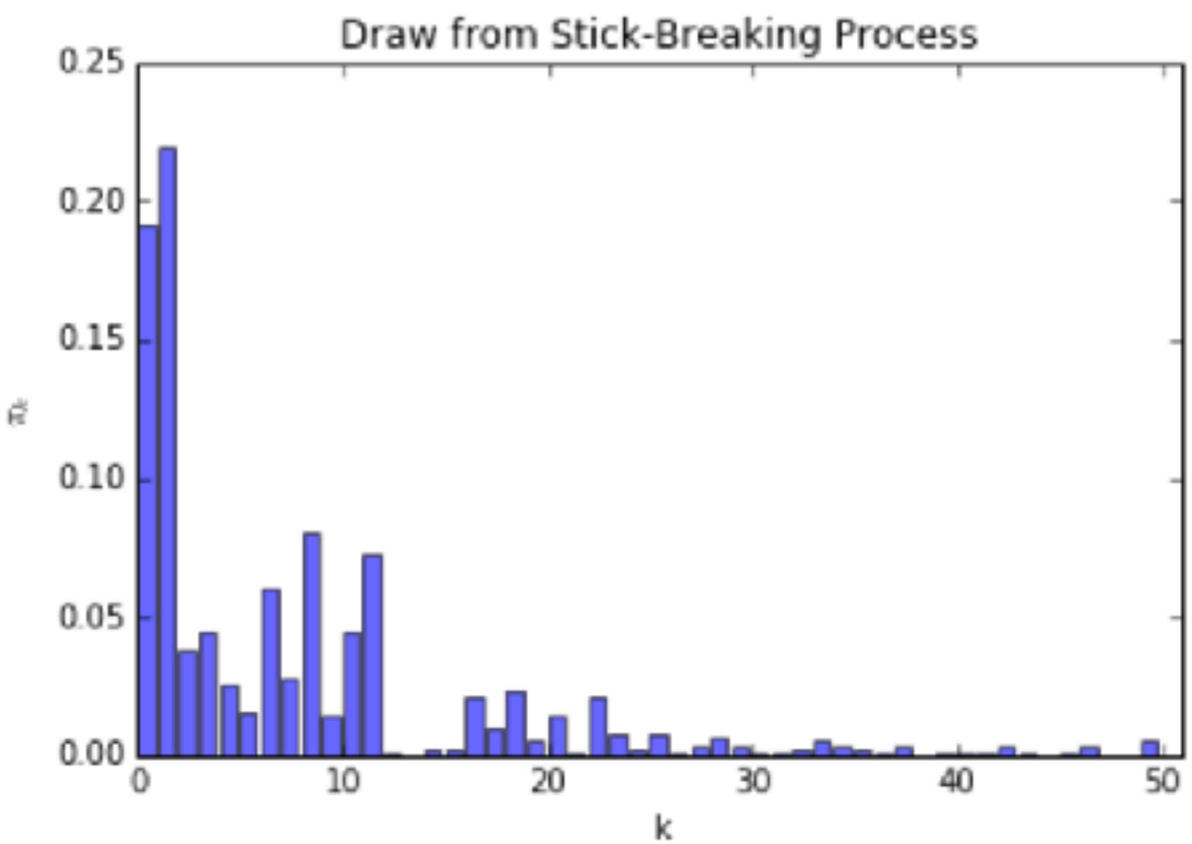
$$\approx \underbrace{\frac{1}{S}\sum_{s=1}^{S} \log p_{\theta}(\mathbf{x}_i|\mathbf{z}_{i,s})}_{\text{Monte Carlo Expectation}} - KL(\underbrace{q_{\phi}(\mathbf{z}_i|\mathbf{x}_i)}_{\text{Variational posterior}}||p(\mathbf{z}_i))$$

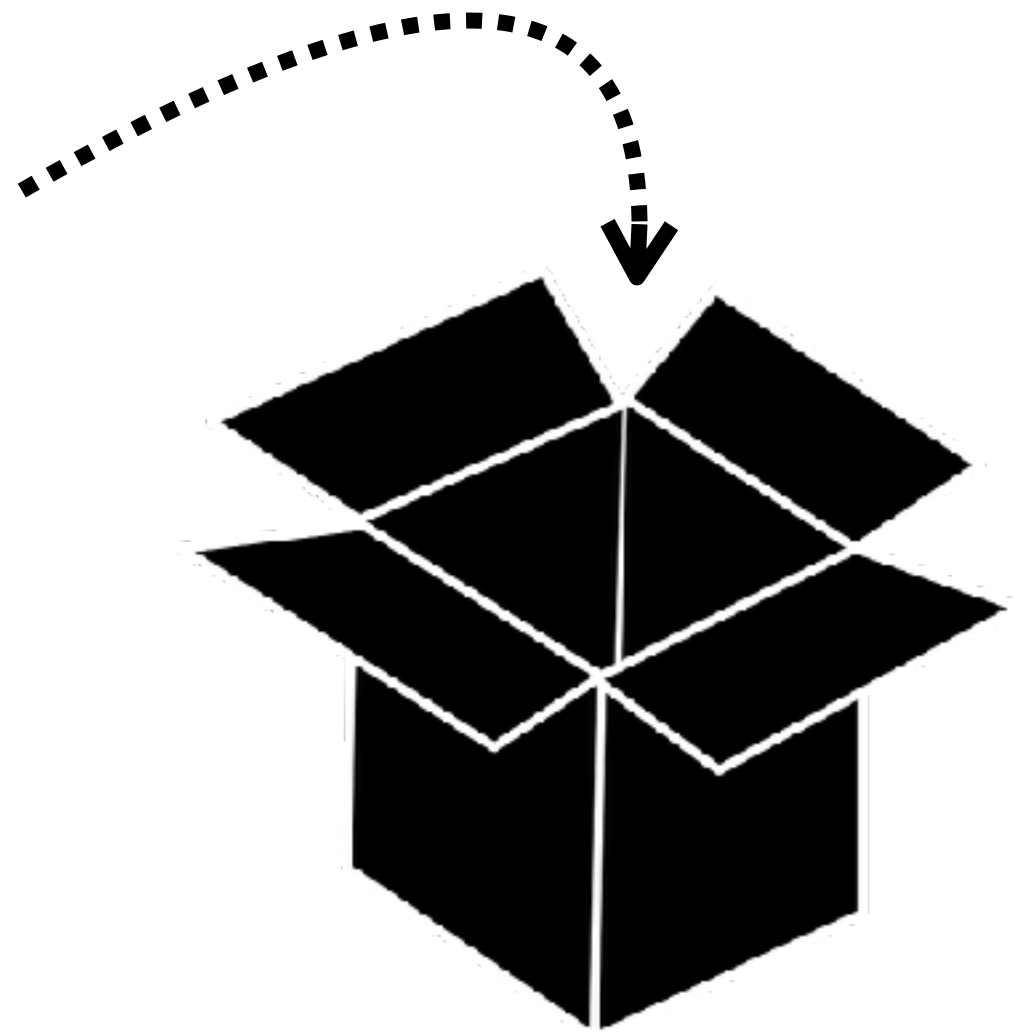
Monte Carlo Expectation
(relieves conjugacy constraints)

Variational posterior is a
globally-parametrized model
('amortized' approach)

Gradients can be taken through MC samples into **z**'s
parameters via a non-centered representation

SGVB for Stick-Breaking Processes





Black-Boxing Bayesian Nonparametrics

Can we use SGVB for the GEM component of stick-breaking priors?

$$G(\cdot) = \sum_{k=1}^{\infty} \underline{\pi_k} \delta_{\zeta_k}$$

$$\pi_k = \begin{cases} v_1 \text{ if } k = 1 \\ v_k \prod_{j<k} (1 - v_j) \text{ for } k > 1 \end{cases} \qquad v_k \sim \text{Beta}(\alpha, \beta)$$

Black-Boxing Bayesian Nonparametrics

Can we use SGVB for the GEM component of stick-breaking priors?

$$G(\cdot) = \sum_{k=1}^{\infty} \underline{\pi_k} \delta_{\zeta_k}$$

$$\pi_k = \begin{cases} v_1 \text{ if } k = 1 \\ v_k \prod_{j<k} (1 - v_j) \text{ for } k > 1 \end{cases} \qquad v_k \sim \text{Beta}(\alpha, \beta)$$

Two Requirements:

1. Need to take gradients through π_k into the var. parameters
2. Analytical KL divergence with Beta (not strict, could try MC approx.)

1. Differentiating Through π_k

Obstacle: The Beta distribution does not have a non-centered parametrization (except in special cases)

1. Differentiating Through π_k

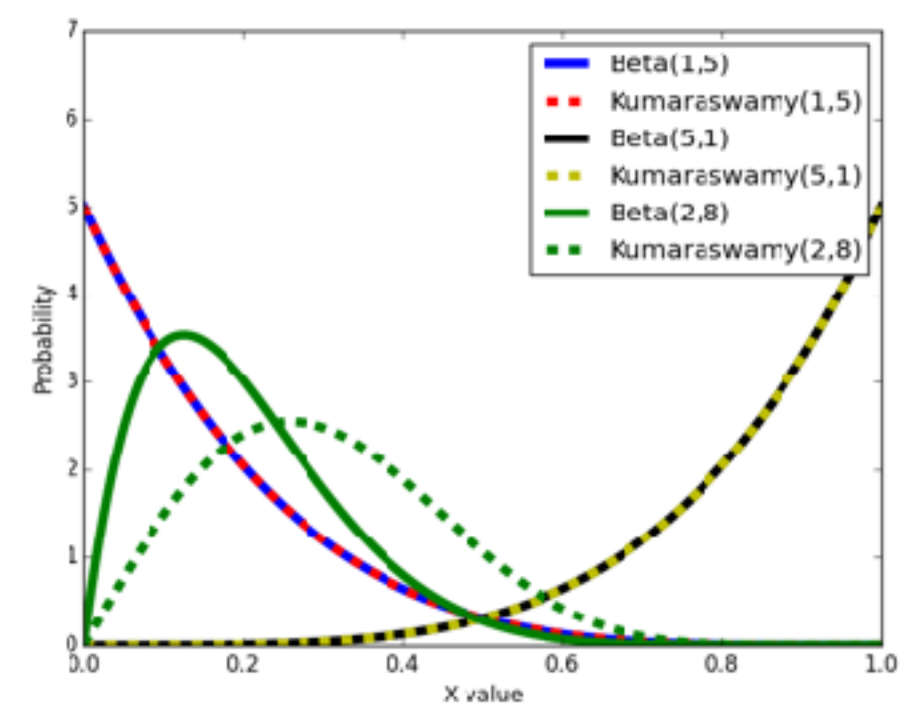
Obstacle: The Beta distribution does not have a non-centered parametrization (except in special cases)

Poondi Kumaraswamy
(1930-1988)

Kumaraswamy Distribution: A Beta-like distribution with a closed-form inverse CDF. Use as variational posterior.

$$\text{Kumaraswamy}(x; a, b) = abx^{a-1}(1 - x^a)^{b-1}$$

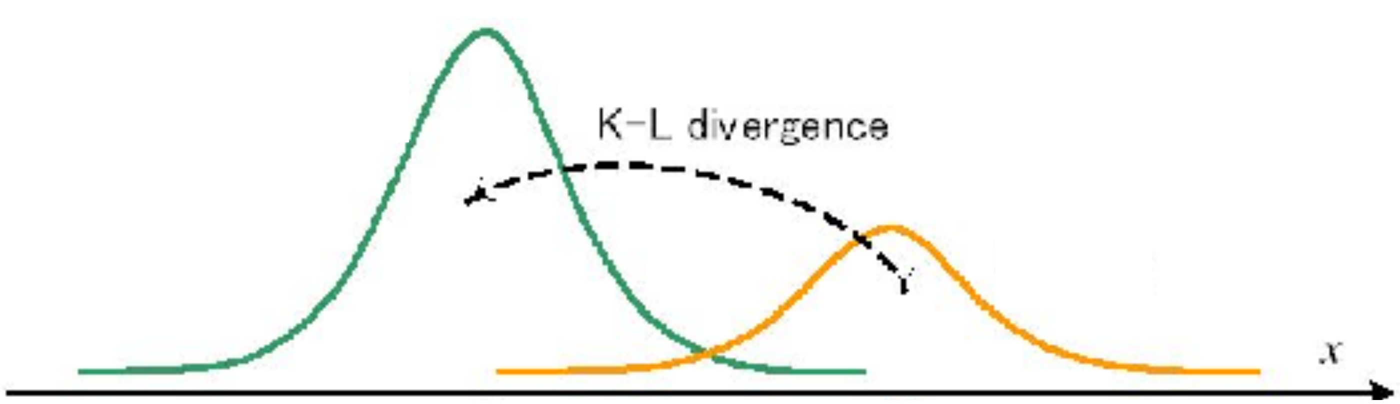
$$x \sim (1 - u^{\frac{1}{b}})^{\frac{1}{a}} \text{ where } u \sim \text{Uniform}(0, 1)$$



2. KL Divergence

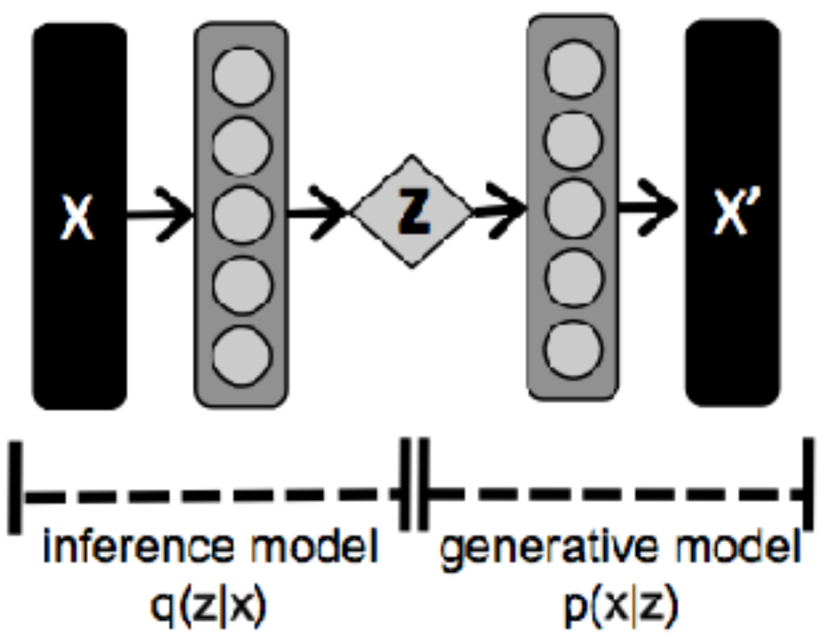
$$\mathbb{E}_q[\log q(v_{i,k})] - \mathbb{E}_q[\log p(v_{i,k})] =$$

$$\frac{a_\phi - \alpha}{a_\phi}\left(-\gamma - \Psi(b_\phi) - \frac{1}{b_\phi}\right) + \log a_\phi b_\phi + \log B(\alpha, \beta)$$
$$- \frac{b_\phi - 1}{b_\phi} + (\beta - 1) b_\phi \sum_{m=1}^{\infty} \frac{1}{m + a_\phi b_\phi} B\left(\frac{m}{a_\phi}, b_\phi\right)$$



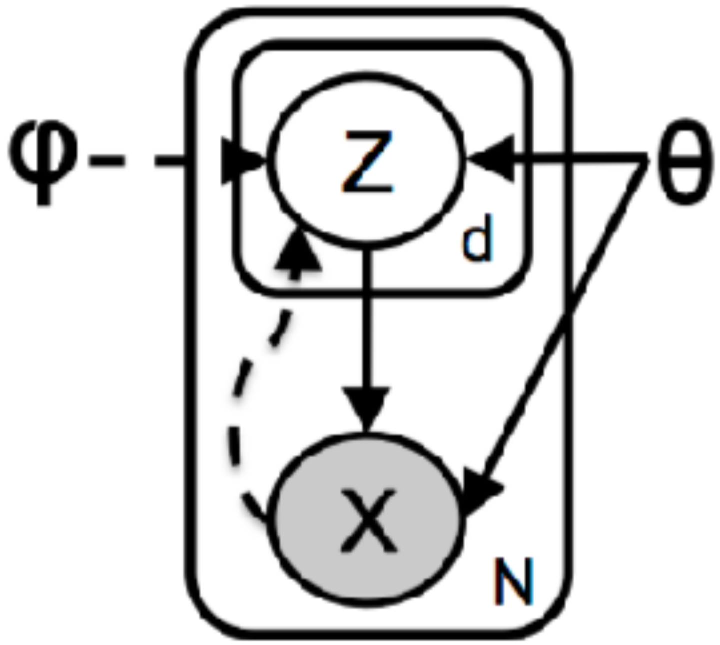
Application to Deep Generative Models

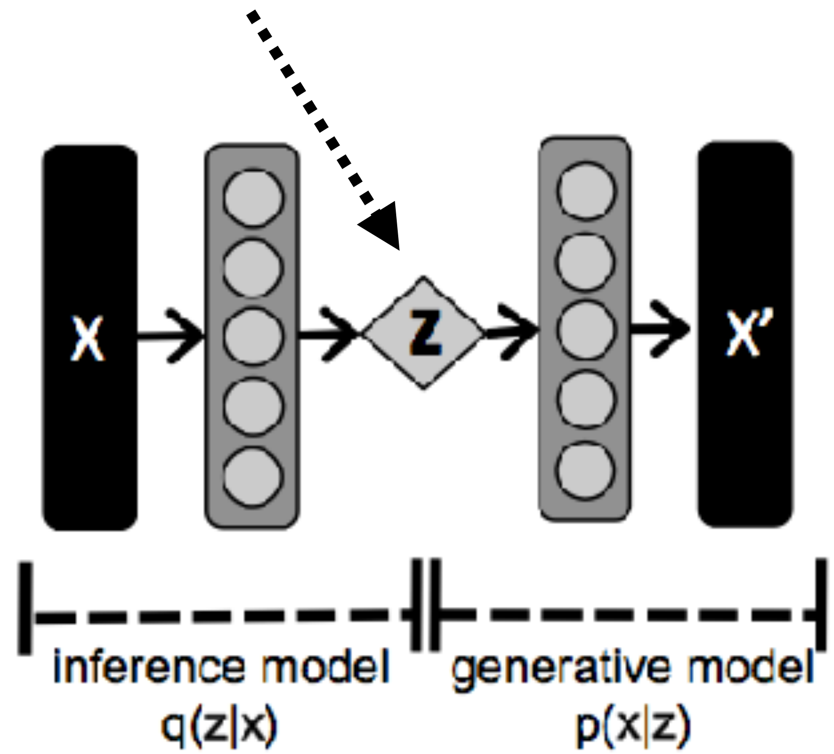
*Applicable to just about every VAE-based model, including the *Neural Statistician*



Variational Autoencoder

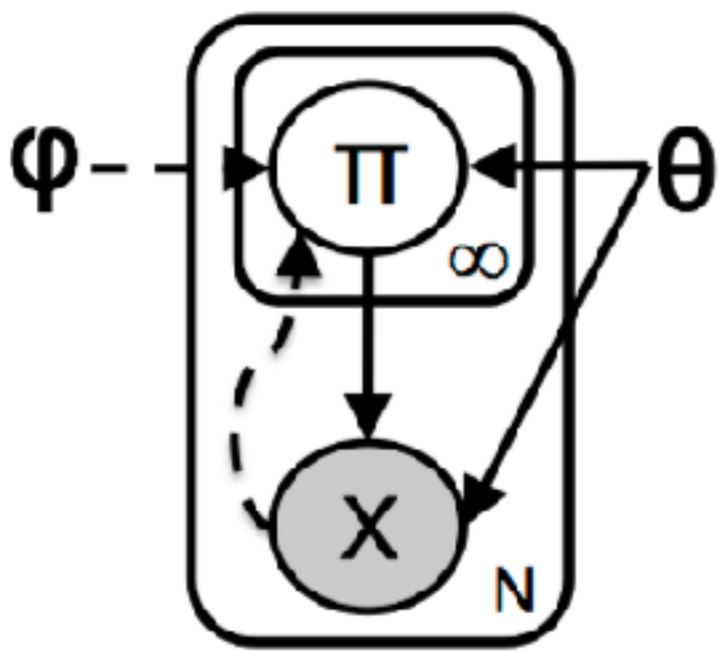
(Kingma & Welling, ICLR 2014)



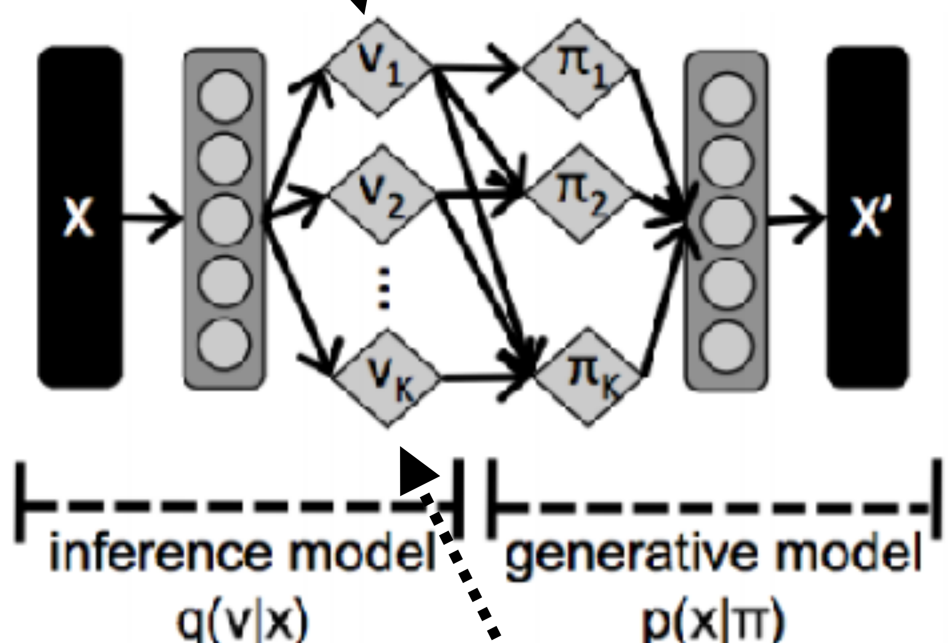


$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \phi; \mathbf{x}_i) = \frac{1}{S} \sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i | \mathbf{z}_{i,s}) - KL(q_{\boldsymbol{\phi}}(\mathbf{z}_i | \mathbf{x}_i) || p(\mathbf{z}_i))$$

Stick-Breaking Variational Autoencoder



Kumaraswamy Samples

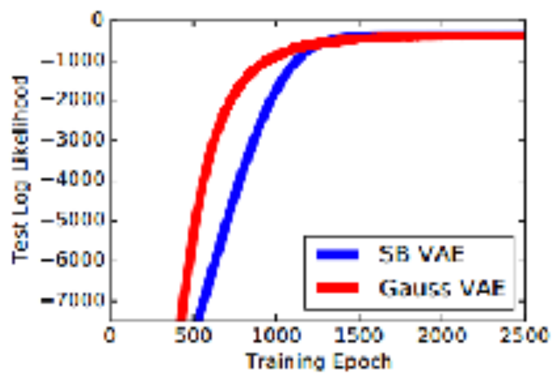


Truncated posterior;
not necessary but learns faster

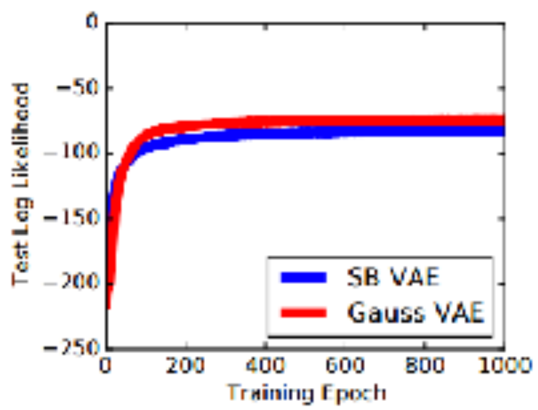
$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \phi; \mathbf{x}_i) = \frac{1}{S}\sum_{s=1}^{S} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i | \boldsymbol{\pi}_{i,s}) - KL(q_{\boldsymbol{\phi}}(\boldsymbol{\pi}_i | \mathbf{x}_i) || p(\boldsymbol{\pi}_i; \boldsymbol{\alpha}_0))$$

Quantitative Results

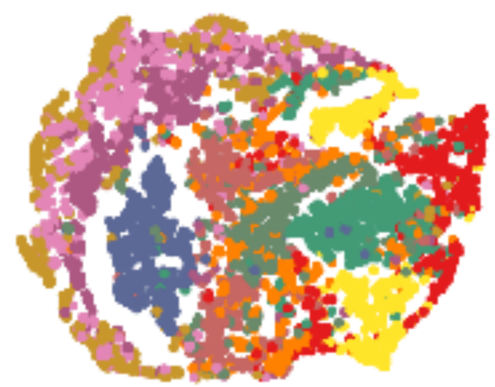
Unsupervised



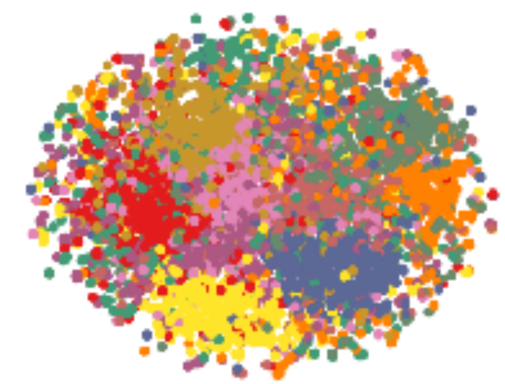
(a) Frey Faces



(b) MNIST



MNIST: Dirichlet Process Latent Space (t-SNE)



MNIST: Gaussian Latent Space (t-SNE)

	k=3	k=5	k=10
SB-VAE	9.34	8.65	8.90
Gauss-VAE	28.4	20.96	15.33
Raw Pixels	2.95	3.12	3.35

MNIST: kNN Classifier on Latent Space

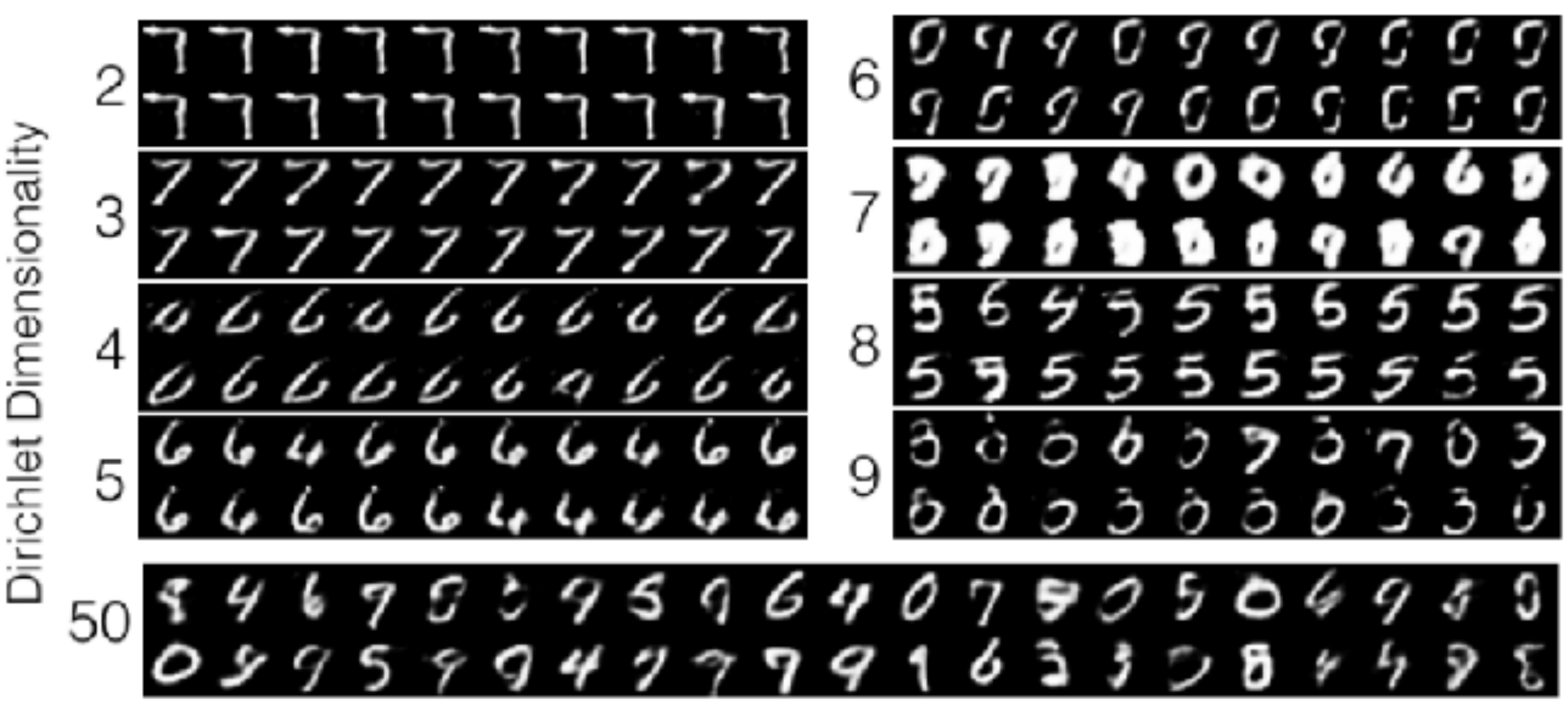
Semi-Supervised

Nonparametric version of (Kingma et al., NIPS 2014)'s M2 model

	MNIST (N=45,000)			**SVHN** (N=65,000)		
	10%	5%	1%	10%	5%	1%
SB-DGM	4.86$_{\pm.14}$	5.29$_{\pm.39}$	**7.34**$_{\pm.47}$	**32.08**$_{\pm4.00}$	**37.07**$_{\pm5.22}$	**61.37**$_{\pm3.60}$
Gauss-DGM	**3.95**$_{\pm.15}$	**4.74**$_{\pm.43}$	11.55$_{\pm2.28}$	36.08$_{\pm1.49}$	48.75$_{\pm1.47}$	69.58$_{\pm1.64}$
kNN	6.13$_{\pm.13}$	7.66$_{\pm.10}$	15.27$_{\pm.76}$	64.81$_{\pm.34}$	68.94$_{\pm.47}$	76.64$_{\pm.54}$

Samples from Generative Model

Stick-Breaking VAE

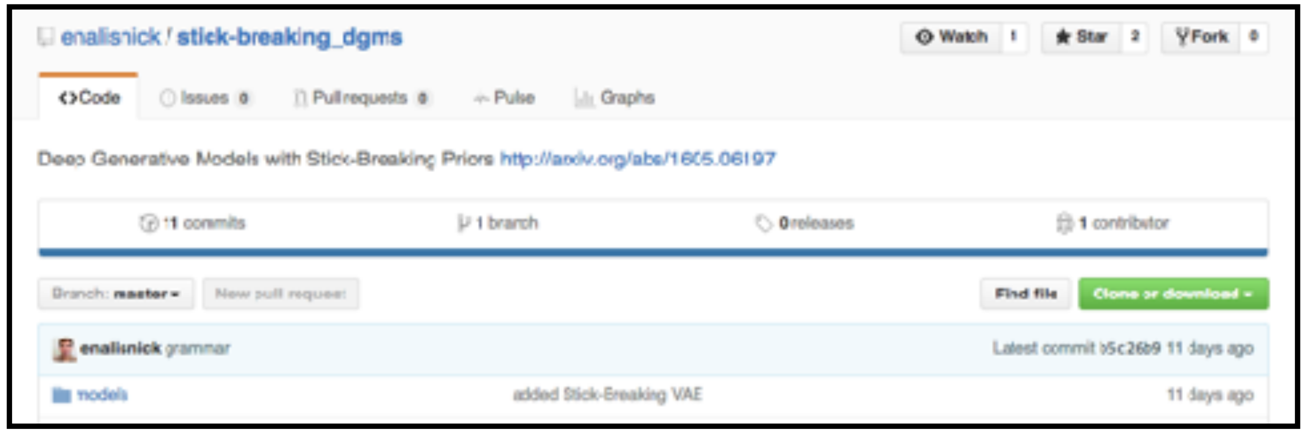


Truncation level of 50 dimensions, Beta(1,5) Prior

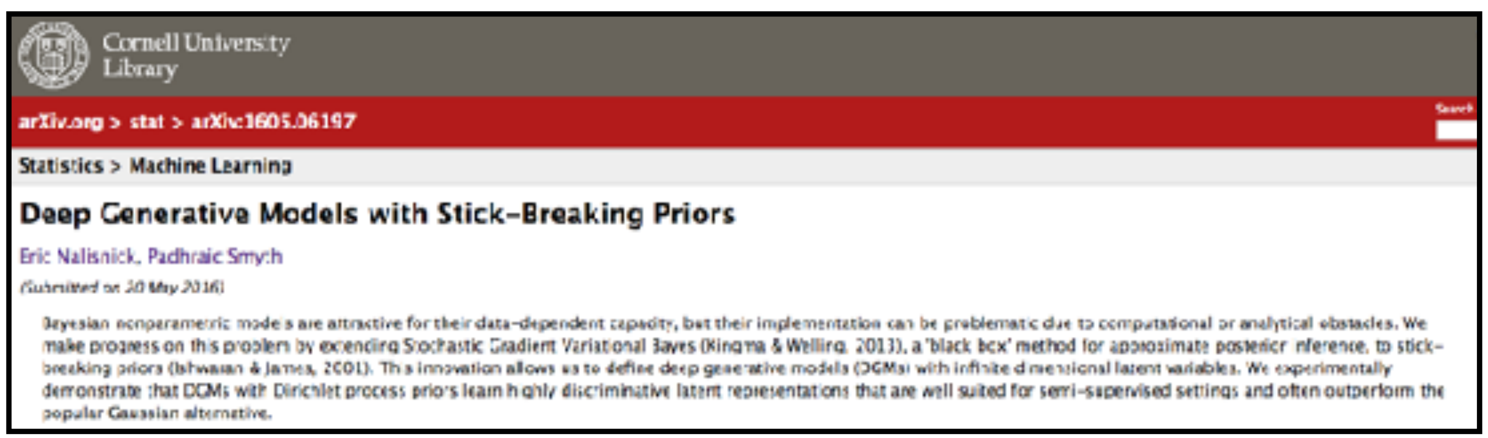
Gaussian VAE

50 dimensions, N(0,1) Prior

Theano code at: `github.com/enalisnick/stick-breaking_dgms`

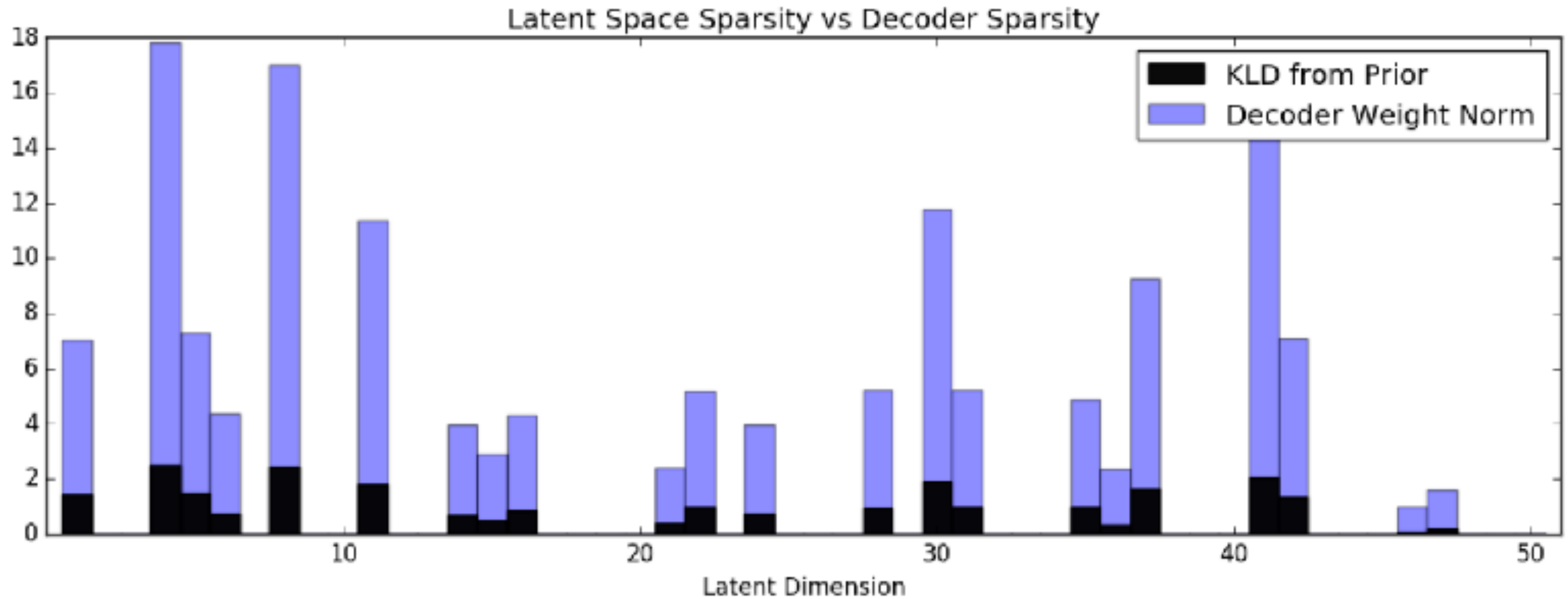


Full paper at: `arxiv.org/abs/1605.06197`

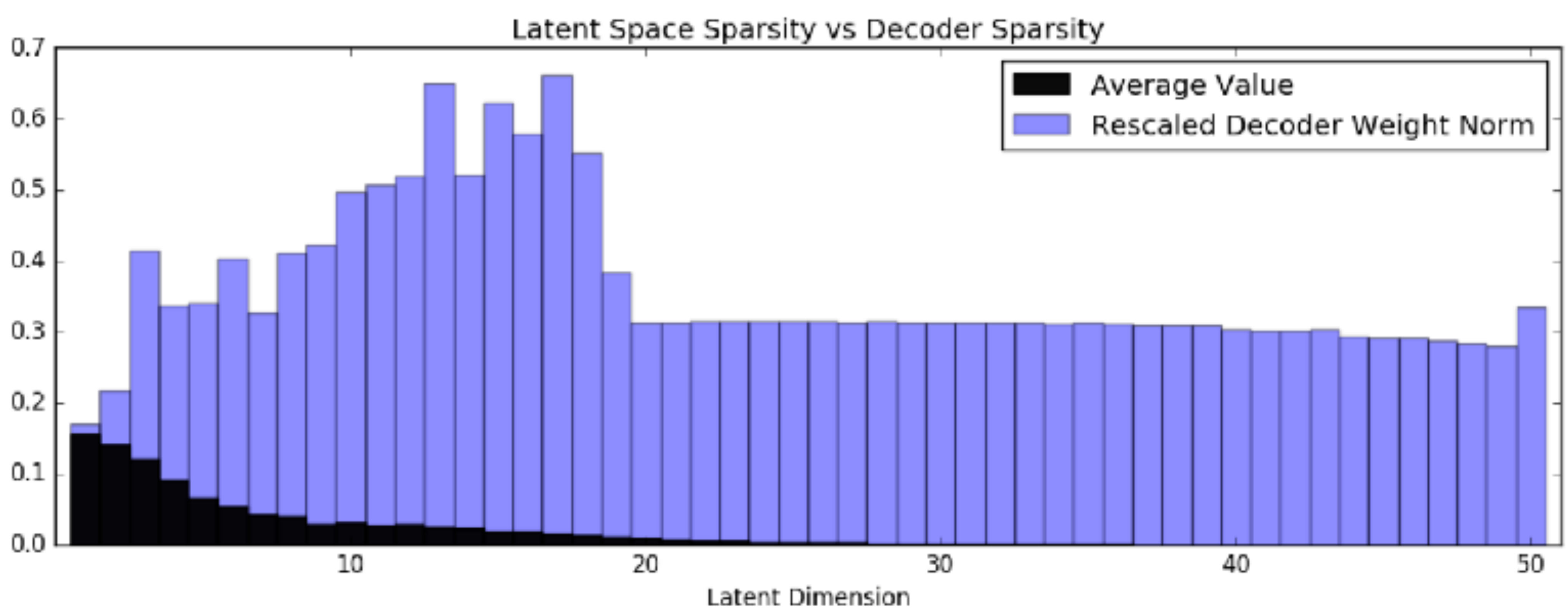


Thank you. Questions?

Appendix



(a) Gauss VAE



(b) Stick-Breaking VAE