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

$$\log p_{\theta}(x_i) \geq \mathbb{E}_q[\log p_{\theta}(x_i|z_i)] - KL(q_{\phi}(z_i)\|p(z_i))$$

Latent variable for which we want posterior
Black Box Variational Inference

\[ \log p_\theta(x_i) \geq \mathbb{E}_q[\log p_\theta(x_i|z_i)] - KL(q_\phi(z_i)||p(z_i)) \]

\[ \approx \frac{1}{S} \sum_{s=1}^{S} \log p_\theta(x_i|z_{i,s}) - KL(q_\phi(z_i|x_i)||p(z_i)) \]

Stochastic Gradient Variational Bayes (SGVB) Estimator:

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)
Black Box Variational Inference

\[
\log p_\theta(x_i) \geq \mathbb{E}_q[\log p_\theta(x_i|z_i)] - KL(q_\phi(z_i)||p(z_i))
\]

Stochastic Gradient Variational Bayes (SGVB) Estimator:

\[
\approx \frac{1}{S} \sum_{s=1}^{S} \log p_\theta(x_i|z_{i,s}) - KL(q_\phi(z_i|x_i)||p(z_i))
\]

Monte Carlo Expectation
(relieves conjugacy constraints)

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)
Black Box Variational Inference

\[ \log p_\theta(x_i) \geq \mathbb{E}_q[\log p_\theta(x_i|z_i)] - KL(q_\phi(z_i)||p(z_i)) \]

Stochastic Gradient Variational Bayes (SGVB) Estimator:

\[ \approx \frac{1}{S} \sum_{s=1}^{S} \log p_\theta(x_i|z_{i,s}) - KL(q_\phi(z_i|x_i)||p(z_i)) \]

Monte Carlo Expectation (relieves conjugacy constraints)

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

(Kingma & Welling, ICLR 2014; Rezende et al, ICML 2014)
Monte Carlo Expectation (relieves conjugacy constraints)

Stochastic Gradient Variational Bayes (SGVB) Estimator:

\[
\log p_\theta(x_i) \geq \mathbb{E}_q[\log p_\theta(x_i|z_i)] - KL(q_\phi(z_i)||p(z_i))
\]

\[
\approx \frac{1}{S} \sum_{s=1}^S \log p_\theta(x_i|z_{i,s}) - KL(q_\phi(z_i|x_i)||p(z_i))
\]

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

Monte Carlo Expectation (relieves conjugacy constraints)

Gradients can be taken through MC samples into $z$’s parameters via a non-centered representation
SGVB for Stick-Breaking Processes
Can we use SGVB for the GEM component of stick-breaking priors?

\[ G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\xi_k} \]

\[ \pi_k = \begin{cases} v_1 & \text{if } k = 1 \\ v_k \prod_{j<k}(1 - v_j) & \text{for } k > 1 \end{cases} \quad v_k \sim \text{Beta}(\alpha, \beta) \]
Can we use SGVB for the GEM component of stick-breaking priors?

\[ G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\xi_k} \]

\[ \pi_k = \begin{cases} 
  \nu_1 & \text{if } k = 1 \\
  \nu_k \prod_{j<k} (1 - \nu_j) & \text{for } k > 1
\end{cases} \]

\[ \nu_k \sim \text{Beta}(\alpha, \beta) \]

Two Requirements:
1. Need to take gradients through \( \pi_k \) into the var. parameters
2. Analytical KL divergence with Beta (not strict, could try MC approx.)
1. Differentiating Through $\pi_k$

**Obstacle:** The Beta distribution does not have a non-centered parametrization (except in special cases)
1. Differentiating Through $\pi_k$

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

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

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

$x \sim (1 - u^b)^{1/a}$ 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)

\[
\tilde{L}(\theta, \phi; x_i) = \frac{1}{S} \sum_{s=1}^{S} \log p_\theta(x_i | z_{i,s}) - KL(q_\phi(z_i | x_i) || p(z_i))
\]
Stick-Breaking Variational Autoencoder

Kumaraswamy Samples

Truncated posterior; not necessary but learns faster

\[
\hat{L}(\theta, \phi; x_i) = \frac{1}{S} \sum_{s=1}^{S} \log p_{\theta}(x_i | \pi_{i,s}) - KL(q_{\phi}(\pi_i | x_i) \| p(\pi_i; \alpha_0))
\]
Quantitative Results

Unsupervised

MNIST: Dirichlet Process Latent Space (t-SNE)
MNIST: Gaussian Latent Space (t-SNE)

(a) Frey Faces

(b) MNIST

MNIST: kNN Classifier on Latent Space

(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

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

Semi-Supervised

<table>
<thead>
<tr>
<th>MNIST (N=45,000)</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB-DGM</td>
<td>4.86±.14</td>
<td>5.29±.39</td>
<td>7.34±.47</td>
</tr>
<tr>
<td>Gauss-DGM</td>
<td>3.95±.15</td>
<td>4.74±.43</td>
<td>11.55±.28</td>
</tr>
<tr>
<td>kNN</td>
<td>6.13±.13</td>
<td>7.66±.10</td>
<td>15.27±.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SVHN (N=65,000)</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB-DGM</td>
<td>32.08±4.00</td>
<td>37.07±5.22</td>
<td>61.37±3.60</td>
</tr>
<tr>
<td>Gauss-DGM</td>
<td>36.08±1.49</td>
<td>48.75±1.47</td>
<td>69.58±1.64</td>
</tr>
<tr>
<td>kNN</td>
<td>64.81±.34</td>
<td>68.94±.47</td>
<td>76.64±.54</td>
</tr>
</tbody>
</table>
Samples from Generative Model

Stick-Breaking VAE

Gaussian VAE

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

50 dimensions, N(0,1) Prior
Thank you. Questions?
Appendix
(a) Gauss VAE

(b) Stick-Breaking VAE