Improved Variational Inference with Inverse Autoregressive Flow


Diederik P. Kingma    Tim Salimans    Rafal Jozefowicz
Xi Chen    Ilya Sutskever    Max Welling

OpenAI, San Francisco

August 7th, 2017

Presented by: Dan Salo
Outline

1. Inverse Autoregressive Flow (4 slides)
2. ResNet VAE (3 slides)
3. Results and Conclusion (3 slides)
Idea: Maximize variational lower bound $\mathcal{L}$ on data likelihood $p(x)$.

$$\log p(x) = \log \int_z p_\theta(x|z)p(z)dz$$  \hspace{1cm} (1)

$$\geq \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)} \right]$$ \hspace{1cm} (2)

$$= \mathcal{L}(\theta, \phi; x)$$ \hspace{1cm} (3)

Limit: Distributions must have differentiable non-centered parametrization. Has resulted in simpler latent distributions.

$$z \sim \mathcal{N}(z|\mu, \sigma^2) \leftrightarrow z = \mu + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$ \hspace{1cm} (4)

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[f_\theta(z)] \leftrightarrow \mathbb{E}_{\mathcal{N}(\epsilon|0,1)}[\nabla_\phi f_\theta(\mu + \sigma \epsilon)]$$ \hspace{1cm} (5)
Normalizing Flows (Rezende ICML 2015)

**Idea:** Transform latent distribution with invertible mappings.

\[
     z_K = f_K \circ \ldots \circ f_2 \circ f_1(z_0) \tag{6}
\]

\[
     \ln q_K(z_K) = \ln q_0(z_0) - \sum_{k=1}^{K} \ln \det \left| \frac{\partial f_k}{\partial z_k} \right| \tag{7}
\]

**Limit:** Only implements linear-time flows with MLPs.
Inverse Autoregressive Flow

**Idea:** Replace MLP with a RNN in Normalizing Flow framework. Triangular Jacobian yields easy determinant computation.

\[ z = \mu_0 + \sigma_0 \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0|I) \] (8)

\[ \epsilon_i = \frac{z_i - \mu_i(z_{1:i-1})}{\sigma_i(z_{1:i-1})} \] (9)

\[ \log \det \left| \frac{d\epsilon}{dy} \right| = \sum_{i=1}^{D} - \log \sigma_i(z) \] (10)

\[ \log q(z_T|x) = - \sum_{i=1}^{D} \left( \frac{1}{2} \epsilon_i^2 + \frac{1}{2} \log(2\pi) + \sum_{t=0}^{T} \log \sigma_{t,i} \right) \] (11)
Inverse Autoregressive Flow

Result:

- $z$: a random sample from $q(z|x)$, the approximate posterior distribution
- $l$: the scalar value of $\log q(z|x)$, evaluated at sample $z$

\[
[\mu, \sigma, h] \leftarrow \text{EncoderNN}(x; \theta)
\]
\[
e \sim \mathcal{N}(0, I)
\]
\[
z \leftarrow \sigma \odot e + \mu
\]
\[
l \leftarrow -\text{sum}(\log \sigma + \frac{1}{2}e^2 + \frac{1}{2} \log(2\pi))
\]
\[
\text{for } t \leftarrow 1 \text{ to } T \text{ do}
\]
\[
\quad [m, s] \leftarrow \text{AutoregressiveNN}[t](z, h; \theta)
\]
\[
\quad \sigma \leftarrow \text{sigmoid}(s)
\]
\[
\quad z \leftarrow \sigma \odot z + (1 - \sigma) \odot m
\]
\[
\quad l \leftarrow l - \text{sum}(\log \sigma)
\]
end
Ladder VAE (Sonderby NIPS 2016)

**Idea:** Multiple stochastic layers with lateral connections to prune away information in higher layers. Generative model serves as prior.

\[
m_{q,i}, m_{p,i} = \text{enc/dec feature maps}
\]

\[
q_\phi(z|x) = q_\phi(z_L|x) \prod_{i=1}^{L-1} q_\phi(z_i|z_{i+1}, x)
\]

\[
q_\phi(z_i|\cdot) = \mathcal{N}(z_i|\mu_i, \sigma_i^2)
\]

\[
\mu_i = f_\mu(m_{q,i}, m_{p,i})
\]

\[
\sigma_i^2 = f_\sigma^2(m_{q,i}, m_{p,i})
\]

**Limit:** Restrictive and deterministic lateral connections \([f_\mu, f_\sigma^2]\).
**Bottom-Up ResNet VAE**

### MNIST NLL

<table>
<thead>
<tr>
<th>Model</th>
<th>VLB</th>
<th>$\log p(x) \approx$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional VAE + HVI [1]</td>
<td>-83.49</td>
<td>-81.94</td>
</tr>
<tr>
<td>DLGM 2hl + IWAE [2]</td>
<td></td>
<td>-82.90</td>
</tr>
<tr>
<td>LVAE [3]</td>
<td></td>
<td>-81.74</td>
</tr>
<tr>
<td><strong>Diagonal covariance</strong></td>
<td>-84.08 (± 0.10)</td>
<td>-81.08 (± 0.08)</td>
</tr>
<tr>
<td>IAF (Depth = 2, Width = 320)</td>
<td>-82.02 (± 0.08)</td>
<td>-79.77 (± 0.06)</td>
</tr>
<tr>
<td>IAF (Depth = 2, Width = 1920)</td>
<td>-81.17 (± 0.08)</td>
<td>-79.30 (± 0.08)</td>
</tr>
<tr>
<td>IAF (Depth = 4, Width = 1920)</td>
<td>-80.93 (± 0.09)</td>
<td>-79.17 (± 0.08)</td>
</tr>
<tr>
<td>IAF (Depth = 8, Width = 1920)</td>
<td>-80.80 (± 0.07)</td>
<td><strong>-79.10</strong> (± 0.07)</td>
</tr>
<tr>
<td>Method</td>
<td>bits/dim ≤</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td><strong>Results with tractable likelihood models:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform distribution (van den Oord et al., 2016b)</td>
<td>8.00</td>
<td></td>
</tr>
<tr>
<td>Multivariate Gaussian (van den Oord et al., 2016b)</td>
<td>4.70</td>
<td></td>
</tr>
<tr>
<td>NICE (Dinh et al., 2014)</td>
<td>4.48</td>
<td></td>
</tr>
<tr>
<td>Deep GMMs (van den Oord and Schrauwen, 2014)</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>Real NVP (Dinh et al., 2016)</td>
<td>3.49</td>
<td></td>
</tr>
<tr>
<td>PixelRNN (van den Oord et al., 2016b)</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>Gated PixelCNN (van den Oord et al., 2016c)</td>
<td>3.03</td>
<td></td>
</tr>
<tr>
<td><strong>Results with variationally trained latent-variable models:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep Diffusion (Sohl-Dickstein et al., 2015)</td>
<td>5.40</td>
<td></td>
</tr>
<tr>
<td>Convolutional DRAWR (Gregor et al., 2016)</td>
<td>3.58</td>
<td></td>
</tr>
<tr>
<td>ResNet VAE with IAF (Ours)</td>
<td>3.11</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Inverse Autoregressive Flow extends NF for more expressive posteriors without sacrificing computation or speed.
- ResNet VAE incorporates the ladder structure into a more principled probabilistic framework.
- Competitive with PixelCNNs for image generation tasks at a fraction of the time.
- While autoregressive neural networks are powerful generators, they may be overexpressive for semi-supervised learning and distribution estimation. This is explored in the ”Variational Lossy Autoencoder” (OpenAI ICLR 2017).