Improved Variational Inference with Inverse Autoregressive Flow

Conference on Neural Information Processing Systems, 2016

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)

SGVB (Kingma ICLR 2014)

Idea: Maximize variational lower bound $\mathcal{L}$ on data likelihood $p(\boldsymbol{x})$.

$$\log p(\boldsymbol{x}) = \log \int_z p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \quad (1)$$

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

$$= \mathcal{L}(\theta, \phi; \boldsymbol{x}) \quad (3)$$

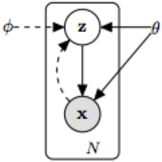


Figure: VAE Model

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

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

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[f_\theta(z)] \leftrightarrow \mathbb{E}_{\mathcal{N}(\epsilon|0,1)}[\nabla_\phi f_\theta(\mu + \sigma\epsilon)] \quad (5)$$

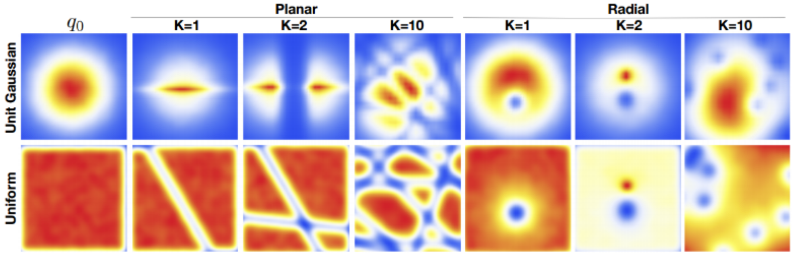
Normalizing Flows (Rezende ICML 2015)

Idea: Transform latent distribution with invertible mappings.

$$\boldsymbol{z}_K = f_K \circ \ldots \circ f_2 \circ f_1(\boldsymbol{z}_0) \tag{6}$$

$$\ln q_K(\boldsymbol{z}_K) = \ln q_0(\boldsymbol{z}_0) - \sum_{k=1}^{K} \ln \det \left| \frac{\partial f_k}{\partial \boldsymbol{z}_k} \right| \tag{7}$$

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.

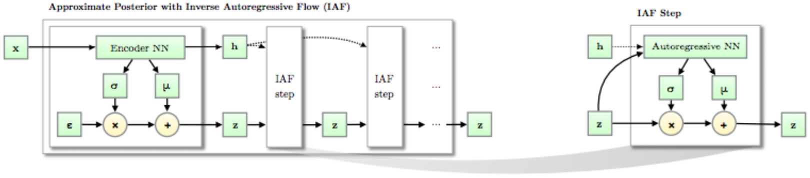
$$\boldsymbol{z} = \boldsymbol{\mu}_0 + \boldsymbol{\sigma}_0 \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0|I) \tag{8}$$

$$\epsilon_i = \frac{z_i - \mu_i(\boldsymbol{z}_{1:i-1})}{\sigma_i(\boldsymbol{z}_{1:i-1})} \tag{9}$$

$$\log \det \left| \frac{d\boldsymbol{\epsilon}}{d\boldsymbol{y}} \right| = \sum_{i=1}^{D} -\log \sigma_i(\boldsymbol{z}) \tag{10}$$

$$\log q(\boldsymbol{z}_T|\boldsymbol{x}) = -\sum_{i=1}^{D} \left(\frac{1}{2}\epsilon_i^2 + \frac{1}{2}\log(2\pi) + \sum_{t=0}^{T} \log \sigma_{t,i} \right) \tag{11}$$

Inverse Autoregressive Flow



Result:

$\mathbf{z}$: a random sample from $q(\mathbf{z}|\mathbf{x})$, the approximate posterior distribution

l: the scalar value of $\log q(\mathbf{z}|\mathbf{x})$, evaluated at sample 'z'

$[\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{h}] \leftarrow \texttt{EncoderNN}(\mathbf{x}; \boldsymbol{\theta})$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$

$\mathbf{z} \leftarrow \boldsymbol{\sigma} \odot \boldsymbol{\epsilon} + \boldsymbol{\mu}$

$l \leftarrow -\text{sum}(\log \boldsymbol{\sigma} + \frac{1}{2}\boldsymbol{\epsilon}^2 + \frac{1}{2}\log(2\pi))$

for $t \leftarrow 1$ **to** T **do**

  $[\mathbf{m}, \mathbf{s}] \leftarrow \texttt{AutoregressiveNN}[t](\mathbf{z}, \mathbf{h}; \boldsymbol{\theta})$

  $\boldsymbol{\sigma} \leftarrow \texttt{sigmoid}(\mathbf{s})$

  $\mathbf{z} \leftarrow \boldsymbol{\sigma} \odot \mathbf{z} + (1 - \boldsymbol{\sigma}) \odot \mathbf{m}$

  $l \leftarrow l - \text{sum}(\log \boldsymbol{\sigma})$

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.

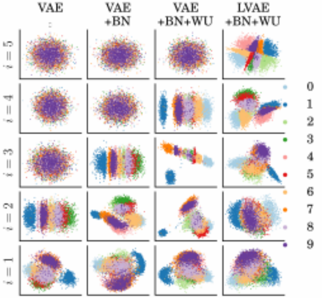
$$m_{q,i}, m_{p,i} = \text{ enc/dec feature maps}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_L|\boldsymbol{x}) \prod_{i=1}^{L-1} q_\phi(\boldsymbol{z}_i|\boldsymbol{z}_{i+1}, \boldsymbol{x})$$

$$q_\phi(\boldsymbol{z}_i|\cdot) = \mathcal{N}(\boldsymbol{z}_i|\mu_i, \sigma_i^2)$$

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

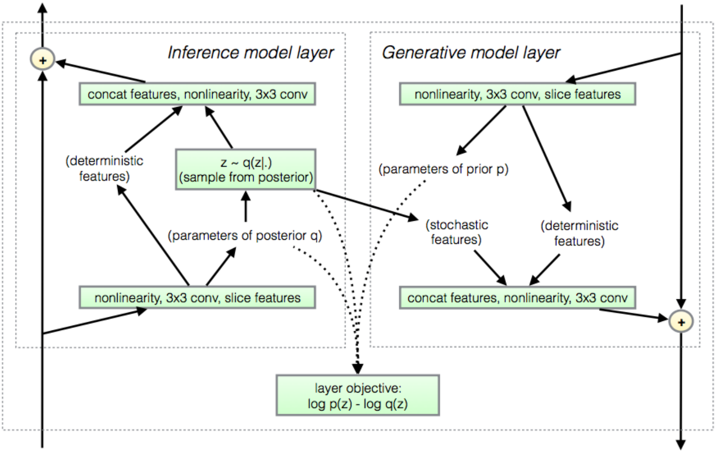
$$\sigma_i^2 = f_{\sigma^2}(m_{q,i}, m_{p,i})$$



Limit: Restrictive and deterministic lateral connections $[f_\mu, f_{\sigma^2}]$.

Bottom-Up ResNet VAE

Idea: Lateral connections expressed by neural networks.
Deterministic features computed by residual networks.



MNIST NLL

Model	VLB	$\log p(\mathbf{x}) \approx$
Convolutional VAE + HVI [1]	-83.49	-81.94
DLGM 2hl + IWAE [2]		-82.90
LVAE [3]		-81.74
DRAW + VGP [4]	-79.88	
Diagonal covariance	-84.08 ($\pm$ 0.10)	-81.08 ($\pm$ 0.08)
IAF (Depth = 2, Width = 320)	-82.02 ($\pm$ 0.08)	-79.77 ($\pm$ 0.06)
IAF (Depth = 2, Width = 1920)	-81.17 ($\pm$ 0.08)	-79.30 ($\pm$ 0.08)
IAF (Depth = 4, Width = 1920)	-80.93 ($\pm$ 0.09)	-79.17 ($\pm$ 0.08)
IAF (Depth = 8, Width = 1920)	-80.80 ($\pm$ 0.07)	**-79.10** ($\pm$ 0.07)

CIFAR-10 bits/dim

Method	bits/dim ≤
Results with tractable likelihood models:	
Uniform distribution (van den Oord et al., 2016b)	8.00
Multivariate Gaussian (van den Oord et al., 2016b)	4.70
NICE (Dinh et al., 2014)	4.48
Deep GMMs (van den Oord and Schrauwen, 2014)	4.00
Real NVP (Dinh et al., 2016)	3.49
PixelRNN (van den Oord et al., 2016b)	**3.00**
Gated PixelCNN (van den Oord et al., 2016c)	**3.03**
Results with variationally trained latent-variable models:	
Deep Diffusion (Sohl-Dickstein et al., 2015)	5.40
Convolutional DRAW (Gregor et al., 2016)	3.58
ResNet VAE with IAF (Ours)	**3.11**

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).