Improving Variational Auto-Encoders using Volume-Preserving Flows

A preliminary study

Jakub M. Tomczak

Joint work with Max Welling

AMLAB Seminar
14/02/2017

The presented research was funded by the European Commission within the Marie Skłodowska-Curie Individual Fellowship (Grant No. 702666, "Deep learning and Bayesian inference for medical imaging").
Motivation

- **Generative models** are appealing
  → generating images
  → data augmentation
  → ...

- Two components:
  - **Encoder** $p(z|x)$
  - **Decoder** $p(x|z)$
Variational Auto-Encoders (VAE)

- How to represent distributions?
  \[ \rightarrow \text{neural network} \text{ outputs parameters} \]

- Encoder approximated by \textit{variational posterior}:
  \[ q(z|x) = \mathcal{N}(\mu(x), \text{diag}\{\sigma^2(x)\}) \]

- Training \textit{objective}:
  \[ \ln p(x) \geq \mathbb{E}_{q(z|x)}[\ln p(x|z)] - \text{KL}(q(z|x) \| p(z)) \]
Variational Auto-Encoders (VAE)

- How to represent distributions?
  → neural network outputs parameters

- Encoder approximated by variational posterior:
  \[ q(z|x) = \mathcal{N}(\mu(x), \text{diag}\{\sigma^2(x)\}) \]

- Training objective:
  \[ \ln p(x) \geq \mathbb{E}_{q(z|x)}[\ln p(x|z)] - \text{KL}(q(z|x)\|p(z)) \]

  + reparameterization trick
Normalizing flows

- Diagonal posterior - ***insufficient*** and ***inflexible***.

- How to get more flexible posterior?

  → apply a series of $T$ invertible transformations $f^{(t)}$ to $z^{(0)} \sim q(z|x)$.

- New objective:

  \[
  \ln p(x) \geq \mathbb{E}_{q(z^{(0)}|x)} \left[ \ln p(x|z^{(T)}) + \sum_{t=1}^{T} \ln \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right| \right] - KL(q(z^{(0)}|x) || p(z^{(T)})).
  \]
Normalizing flows

• Diagonal posterior - **insufficient** and **inflexible**.

• How to get more flexible posterior?
  → apply a series of $T$ **invertible transformations** $f^{(t)}$ to $z^{(0)} \sim q(z|x)$.

• New objective:

$$\ln p(x) \geq \mathbb{E}_{q(z^{(0)}|x)} \left[ \ln p(x|z^{(T)}) + \sum_{t=1}^{T} \ln \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right| \right] - KL(q(z^{(0)}|x) || p(z^{(T)})),$$
Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?
  → apply a series of $T$ **invertible transformations**
  
  $$f^{(t)} \text{ to } z^{(0)} \sim q(z|x).$$

- New objective:
  $$\ln p(x) \geq \mathbb{E}_{q(z^{(0)}|x)} \left[ \ln p(x|z^{(T)}) \right]$$

  $$q(z^{(T)}|x) = q(z^{(0)}|x) \prod_{t=1}^{T} \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right|^{-1}.$$
Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?
  → apply a series of $T$ **invertible transformations**

  $f^{(t)}$ to $z^{(0)} \sim q(z|x)$.

- **New objective:**

  $$\ln p(x) \geq \mathbb{E}_{q(z^{(0)}|x)} \left[ \ln p(x|z^{(T)}) + \sum_{t=1}^{T} \ln \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right| \right] - KL(q(z^{(0)}|x) || p(z^{(T)}))$$

Law of Unconscious Statistician (LOTUS):

$$\mathbb{E}_{q(z^{(T)} | x)}[h(z^{(T)})] = \mathbb{E}_{q(z^{(0)} | x)}[h(f^{(T)} \circ \ldots \circ f^{(1)}(z^{(0)}))]$$

gives us the following:

$$\mathbb{E}_{q(z^{(T)} | x)}[\ln p(x, z^{(T)}) - \ln q(z^{(T)} | x)] = \mathbb{E}_{q(z^{(0)} | x)}[\ln p(x, z^{(T)}) - \ln q(z^{(0)} | x) \prod_{t=1}^{T} \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right|^{-1}]$$

$$= \mathbb{E}_{q(z^{(T)} | x)}[\ln p(x, z^{(T)}) - \ln q(z^{(0)} | x) - \ln \prod_{t=1}^{T} \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right|^{-1}]$$

**New objective:**

$$\ln p(x) \geq \mathbb{E}_{q(z^{(0)} | x)}[\ln p(x | z^{(T)}) + \sum_{t=1}^{T} \ln \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right|] - \text{KL}(q(z^{(0)} | x) || p(z^{(T)}))$$

Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?
  → apply a series of $T$ invertible transformations $f^{(t)}$ to $z^{(0)} \sim q(z|x)$.

- **New objective:**

$$\ln p(x) \geq \mathbb{E}_{q(z^{(0)}|x)} \left[ \ln p(x|z^{(T)}) + \sum_{t=1}^{T} \ln \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right| \right] - \text{KL}(q(z^{(0)}|x)||p(z^{(T)})).$$

Normalizing flows

- Diagonal posterior - **insufficient** and **inflexible**.

- How to get more flexible posterior?
  → apply a series of $T$ **invertible transformations** $f^{(t)}$ to $z^{(0)} \sim q(z|x)$.

- **New objective:**

$$
\ln p(x) \geq \mathbb{E}_{q(z^{(0)}|x)} \left[ \ln p(x|z^{(T)}) + \sum_{t=1}^{T} \ln \left| \det \frac{\partial f^{(t)}}{\partial z^{(t-1)}} \right| \right] - \text{KL}(q(z^{(0)}|x) || p(z^{(T)}))
$$

**Jacobian-determinant:** (i) general normalizing flow ($|\det J|$ is **easy** to compute)

(ii) **volume-preserving flow**, i.e., $|\det J|=1$
Volume-preserving flows

• How to obtain more flexible posterior and preserve $|\det J|=1$?

• Model full-covariance matrix:
  
  – using orthogonal matrices $\rightarrow$ Householder flow
  
  – using lower-triangular matrix (with ones on the diagonal) $\rightarrow$ linear Inverse Autoregressive Flow
Householder Flow: motivation

- The eigendecomposition of a full-covariance matrix:

\[ \Sigma = UDU^\top \]

- **Proposition**: Apply

\[ z^{(1)} = Uz^{(0)}, \quad z^{(1)} \sim \mathcal{N}(U\mu, U \text{ diag}(\sigma^2) U^\top) \]

and since $U$ is orthogonal, the Jacobian-determinant would be 1.

- **Question**: *Is it possible to model the orthogonal matrix $U$ in a principled manner? YES!*
Householder Flow: motivation

Theorem
Any orthogonal matrix with the basis acting on the $K$-dimensional subspace can be expressed as a product of exactly $K$ Householder transformations.


• **Question**: Is it possible to model the orthogonal matrix $U$ in a principled manner? **YES!**
Householder Flow: formulation

- In the **Householder transformation** we reflect a vector around a hyperplane defined by a **Householder vector** \( \mathbf{v}_t \in \mathbb{R}^M \):

\[
\mathbf{z}^{(t)} = \left( \mathbf{I} - 2 \frac{\mathbf{v}_t \mathbf{v}_t^\top}{\|\mathbf{v}_t\|^2} \right) \mathbf{z}^{(t-1)} = \mathbf{H}_t \mathbf{z}^{(t-1)}.
\]

**Householder matrix**
Linear Inverse Autoregressive Flow

- A different approach: 
  \[ \rightarrow \text{model the full-covariance matrix directly using a lower-triangular matrix with ones on the diagonal } L: \]

\[ z^{(1)} = L \cdot z^{(0)}. \]

- This corresponds to the simplest case of \textbf{Inverse Autoregressive Flow}.

Convex combination of lower triangular matrices in linear IAF

- Can we further improve on linear IAF?
  → several $L$’s
  → convex combination of $L$’s

- Encoder returns: $y = \text{softmax}(h), \ L_1, \ldots, L_K$

- This is still volume-preserving flow:

$$z^{(1)} = \left( \sum_{k=1}^{K} y_k \ L_k \right) \ z^{(0)}$$

- **Intuition**: taking into account small variations among images.
Other volume-preserving flows

- **NICE**: Non-linear independent components estimation

- **HVI**: Hamiltonian Variational Inference
Experiments

- **Encoder**: 784 – 300 – 300 – 40
- **Decoder**: 40 – 300 – 300 – 784
- **Activation function**: gate units

\[
A = W_l h_{l-1} + b_l \quad \sigma = \sigma(V_l h_{l-1} + c_l)
\]

- **Warm-up for 200 epochs** (annealing KL from 0 to 1)
## Experiments: MNIST

<table>
<thead>
<tr>
<th>Method</th>
<th>ELBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>-93.9</td>
</tr>
<tr>
<td>VAE+HF(T=1)</td>
<td>-87.8</td>
</tr>
<tr>
<td>VAE+HF(T=10)</td>
<td>-87.7</td>
</tr>
<tr>
<td>VAE+linIAF</td>
<td>-85.8</td>
</tr>
<tr>
<td>VAE+cc linIAF(K=5)</td>
<td>-85.3</td>
</tr>
<tr>
<td>VAE+NICE(T=10)</td>
<td>-88.6</td>
</tr>
<tr>
<td>VAE+NICE(T=80)</td>
<td>-87.2</td>
</tr>
<tr>
<td>VAE+HVI(T=1)</td>
<td>-91.7</td>
</tr>
<tr>
<td>VAE+HVI(T=8)</td>
<td>-88.3</td>
</tr>
<tr>
<td>VAE+NF(T=10)</td>
<td>-87.5</td>
</tr>
<tr>
<td>VAE+NF(T=80)</td>
<td>-85.1</td>
</tr>
</tbody>
</table>
Experiments: Histopathology data

- A dataset of histopathological image patches (28x28).
- 6,800: training, 2,000: validation, 2,000: test
- Source: http://www.enjoypath.com
## Experiments: Histopathology data

<table>
<thead>
<tr>
<th>Method</th>
<th>ELBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>1371.4 +/- 32.1</td>
</tr>
<tr>
<td>VAE+HF(T=1)</td>
<td>1388.0 +/- 22.1</td>
</tr>
<tr>
<td>VAE+HF(T=10)</td>
<td>1397.0 +/- 15.2</td>
</tr>
<tr>
<td>VAE+HF(T=20)</td>
<td>1398.3 +/- 8.1</td>
</tr>
<tr>
<td>VAE+linIAF</td>
<td>1388.6 +/- 7.1</td>
</tr>
<tr>
<td>VAE+cc linIAF(K=5)</td>
<td>1413.8 +/- 22.9</td>
</tr>
</tbody>
</table>
Conclusion

- Normalizing flows: it is a nice idea to enrich the posterior.

- Volume-preserving flows simplify the objective and allow to obtain competitive results.

- Two new normalizing flows were proposed.

- Gate units – they improve learning!
Reference:

Code on github:
https://github.com/jmtomczak/vae_householder_flow

Contact:
J.M.Tomczak@uva.nl

The presented research was funded by the European Commission within the Marie Skłodowska-Curie Individual Fellowship (Grant No. 702666, "Deep learning and Bayesian inference for medical imaging").