Inference Suboptimality in Variational Autoencoders

Chris Cremer, Xuechen Li, David Duvenaud
VAE Objective

\[
\log p(x) = \mathbb{E}_{z \sim q(z|x)} \left[ \log \left( \frac{p(x, z)}{q(z|x)} \right) \right] + \text{KL} (q(z|x) \| p(z|x))
\]

ELBO        Inference Gap
Inference Gaps

• Approximation Gap
  • Inability of the variational distribution to model the true posterior

• Amortization Gap
  • Limited capacity of the recognition network to generalize inference over all datapoints
Posterior Visualizations

\[ p(z|x) \]
\[ q_{\text{FFG}}(z|x) \]
\[ q_{\text{FFG}}^*(z|x) \]

Datapoints

A

B

C

D

Amortization

Gap

Approximation

Gap
Flexible Approximations

**Flow Transformation**

\[ v' = v \circ \sigma_1(z) + \mu_1(z) \]
\[ z' = z \circ \sigma_2(v') + \mu_2(v') \]

**Auxiliary Variable**

\[ q \downarrow \text{Flow} (z|\mathbf{x}) \]

\[ q \downarrow \text{AF} (z|\mathbf{x}) \]
Posterior Visualizations

\[ p(z|x) \]
\[ q_{\downarrow \text{FFG}}(z|x) \]
\[ q_{\downarrow \text{FFG}}^{\star}(z|x) \]
\[ q_{\downarrow \text{Flow}}^{\star}(z|x) \]

Datapoints

A

B

C

D

True Posterior

Amortized Factorized Gaussian

Optimal Factorized Gaussian

Optimal Flow
Estimating the Gaps

$$\max \ (\mathcal{L}_{\text{IWAE}}[q^*], \mathcal{L}_{\text{AIS}}) \approx \log p(x)$$

Lower bound with optimal $q$ within its variational family
For every datapoint, optimize its variational parameters

$$\approx \mathcal{L}[q^*]$$

Lower bound with the amortized $q$

$$\approx \mathcal{L}[q]$$

Approximation Gap

Amortization Gap
Amortization vs Approximation

For these model choices:

- Amortization is generally larger than approximation gap

Can we reduce the amortization gap by increasing encoder capacity?
Larger Encoder Reduces Amortization Error

<table>
<thead>
<tr>
<th>Encoder Type</th>
<th>MNIST</th>
<th>Fashion-MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$q_{FFG}$</td>
<td>$q_{AF}$</td>
</tr>
<tr>
<td>Regular Encoder</td>
<td>1.34</td>
<td>1.41</td>
</tr>
<tr>
<td>Larger Encoder</td>
<td>1.11</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Parameters of Flow Reduce Amortization Gap

• Common reasoning for flow: reduce approximation gap
  • Could improvements also be due to reduction in amortization gap?

• Experiment:
  • Trained a VAE on MNIST
  • Retrained new encoders on the fixed decoder
    • Encoders differ only in their variational distribution

<table>
<thead>
<tr>
<th></th>
<th>$q\downarrow FFG$</th>
<th>$q\downarrow Flow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximation</td>
<td>1.91</td>
<td>0.43</td>
</tr>
<tr>
<td>Amortization</td>
<td>43.22</td>
<td>12.86</td>
</tr>
</tbody>
</table>
Generator Learns to Accommodate the Approximation

How much does $p_z x$ fit to $q(z|x)$?
How Gaussian is $p_z x$ when trained with $q_{\downarrow FFG}$ vs $q_{\downarrow AF}$?

<table>
<thead>
<tr>
<th>Generator Trained With</th>
<th>$q_{\downarrow FFG}$</th>
<th>$q_{\downarrow AF}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$KL(q_{\downarrow FFG} \downarrow z x | p(z</td>
<td>x))$</td>
<td>1.43</td>
</tr>
<tr>
<td>$KL(q_{\downarrow AF} \downarrow z x | p(z</td>
<td>x))$</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Larger Decoder Capacity Reduces Approximation Gap

• Does increasing decoder capacity decrease the approximation gap?
  • Does a more powerful decoder make the true posterior easier to model with the choice of approximation?

• Experiment:
  • Train VAEs with decoders that have 0, 2, 4 hidden layers
  • Compute the approximation gaps (i.e. How Gaussian is $p_{Z \mid X}$?)

<table>
<thead>
<tr>
<th>Generator Hidden Layers</th>
<th>0</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximation Gap</td>
<td>3.90</td>
<td>1.83</td>
<td>1.63</td>
</tr>
</tbody>
</table>
Inference Generalization

- From training to validation set:
  - amortization gap increases
  - approximation gap constant

- Increased capacity: more prone to overfitting but better inference
- Flow improves model while overfitting less
Summary

- Inference Gap: Amortization vs Approximation
- Amortization > Approximation
- Generator accommodates approximation
- Inform model design choices

Poster: Hall B #176
Thanks
Experiments

- Encoder Capacity
- Decoder Capacity
- Variational Distribution

<table>
<thead>
<tr>
<th>Dataset Models</th>
<th>MNIST/Fashion</th>
<th>3-BIT CIFAR-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>784-200-200-50</td>
<td></td>
<td>Conv-Conv-Conv-FC</td>
</tr>
<tr>
<td>50-200-200-784</td>
<td></td>
<td>FC-ConvT-ConvT-ConvT</td>
</tr>
</tbody>
</table>
VAE

- Latent Variable Model
- Amortized Inference

- Inference Suboptimality

\[
\log p(x) = \mathbb{E}_{z \sim q(z|x)} \left[ \log \left( \frac{p(x, z)}{q(z|x)} \right) \right] + \text{KL} (q(z|x) \| p(z|x))
\]

ELBO

Inference Gap