Efficient Gradient-Based Inference Through Transformations Between Bayes Nets and Neural Nets

Diederik P (Durk) Kingma, Max Welling
University of Amsterdam
Ph.D. Candidate, advised by Max

Durk Kingma
Max Welling
Neural Nets and Deep Generative Models (1)

- Directed latent variables models
  - Can model complex (multimodal) distributions over observed variables

- Feedforward neural nets:
  - Typically used for learning a single conditional: \( \log p(y|x) = f(x,y) \)
    - e.g. Classification/regression with deep net
  - But: not good for modelling complex distributions, e.g. multi-modal
Neural Nets and Deep Generative Models (2)

• Natural idea: neural nets as components of deep latent-variable models:
  - Learn extremely complex multimodal distributions

• Inner loop of learning requires **posterior inference**
  - Exact inference is intractable
  - But we can still efficiently compute gradients using backpropagation:

\[
p_{\theta}(z|x) = p_{\theta}(x, z)/p_{\theta}(x)
\]

\[
p_{\theta}(z|x) \propto p_{\theta}(x, z)
\]

\[
\nabla_z \log p_{\theta}(z|x) = \nabla_z \log p_{\theta}(x, z)
\]

• Efficient gradient-based approximate inference methods:
  - sampling-based methods (unbiased but slow), e.g. HMC
  - Stochastic Variational inference (biased but fast)
    - [Hoffman & Blei 2012] [Kingma & Welling 2013] [Deepmind 2014]

Can we increase the efficiency of gradient-based posterior inference?
Idea: Reparameterizations (Gaussian example)

**Centered Form (CP)**

\[ p_\theta(z | p_a) = \mathcal{N}(z; \mu, \sigma^2 I) \]
\[ \mu = f_\theta(p_a) \]
\[ \text{(e.g. neural net)} \]

**Differentiable Non-centered Form (DNCP)**

\[ \epsilon \sim \mathcal{N}(0, I) \]
\[ \tilde{z} = f_\theta(p_a) + \sigma \epsilon \]

Neural net perspective: hidden unit with injected noise
Reparameterizations (General form)

**Centered Form (CP)**

\[ p_\theta(z_j | \text{pa}_j) \]

**Differentiable Non-centered Form (DNCP)**

\[ \epsilon \sim p(\epsilon_j) \]
\[ z_j = g_j(\text{pa}_j, \epsilon_j, \theta) \]
Reparameterizations

Can be performed for very broad class of distributions

<table>
<thead>
<tr>
<th>Example</th>
<th>( q_\varphi(z) )</th>
<th>( p(\varepsilon) )</th>
<th>( g(\varphi, \varepsilon) )</th>
<th>Also...</th>
</tr>
</thead>
</table>
| Normal dist.| \( z \sim N(\mu, \sigma) \) | \( \varepsilon \sim N(0,1) \) | \( z = \mu + \sigma \cdot \varepsilon \) | Location-scale familie: Laplace, Elliptical, Student’s t, Logistic, Uniform, Triangular, ...
| Exponential | \( z \sim \exp(\lambda) \) | \( \varepsilon \sim U(0,1) \) | \( z = -\log(1 - \varepsilon)/\lambda \) | Invertible CDF: Cauchy, Logistic, Rayleigh, Pareto, Weibull, Reciprocal, Gompertz, Gumbel and Erlan, ...
| Other       | \( z \sim \log\text{N}(\mu, \sigma) \) | \( \varepsilon \sim N(0,1) \) | \( z = \exp(\mu + \sigma \cdot \varepsilon) \) | Gamma, Dirichlet, Beta, Chi-Squared, and F distributions |

Which form is better for inference?
Poster correlation analysis (1)

Squared posterior correlation:
second-order metric of posterior dependency between 2 variables

\[ C = \begin{pmatrix} \sigma_A^2 & \sigma_{AB}^2 \\ \sigma_{AB}^2 & \sigma_B^2 \end{pmatrix} = -H^{-1} = \frac{1}{\det(H)} \begin{pmatrix} -H_B & H_{AB} \\ H_{AB} & -H_A \end{pmatrix} \]

\[ \rho^2 = \frac{(\sigma_{AB}^2)^2}{(\sigma_A^2 \sigma_B^2)} \\
= \frac{(H_{AB}/\det(H))^2}{((-H_A/\det(H))(-H_B/\det(H)))} \\
= \frac{H_{AB}^2}{H_A H_B} \]
Poster correlation analysis (2)

- Squared correlations for CP:

\[
\rho_{pa(z),z}^2 = \frac{(H_{pa(z)}z)^2}{H_{pa(z)}pa(z)H_{zz}} = \frac{w_i^2/\sigma^4}{(\alpha - w_i^2/\sigma^2)(\sigma_{z|ch(z)}^{-2} - 1/\sigma^2)}
\]

- Squared correlations for DNCP:

\[
\rho_{pa(z),\epsilon}^2 = \frac{(H_{pa(z)}\epsilon)^2}{H_{pa(z)}pa(z)H_{\epsilon\epsilon}} = \frac{\sigma^2 w_i^2 (\sigma_{z|ch(z)}^{-2})^2}{(\alpha + w_i^2 (\sigma_{z|ch(z)}^{-2})) (\sigma^2 \sigma_{z|ch(z)}^{-2} - 1)}
\]
Poster correlation analysis (3)

Inequality: all terms cancel with very simple result:

\[ \rho_{pa(z)_{i},z}^2 > \rho_{pa(z)_{i},\epsilon}^2 \]

\[ \Leftrightarrow \]

\[ \sigma_{z|pa(z)}^2 < \sigma_{z|ch(z)}^2 \]

DNCP leads to more efficient inference when latent variable 'z' is more strongly bound to its parents then to its children.
Analysis (4)

Correlation in the limit of parameters:

| $\lim \sigma_{z|pa(z)} \to 0$ | 1 | $\rho_{y_i,z}^2$ (CP) | $\rho_{u_i,e}^2$ (DNCP) $\sigma_{z|pa(z)}^2$ |
| $\lim \sigma_{z|pa(z)} \to \infty$ | 0 | $\frac{\sigma_{z|ch(z)}^2 w_i^2}{\sigma_{z|ch(z)}^2 w_i^2 + \alpha}$ |
| $\lim \sigma_{z|ch(z)} \to 0$ | 0 | $\frac{w_i^2}{w_i^2 - \alpha \sigma_{z|pa(z)}^2}$ |
| $\lim \sigma_{z|ch(z)} \to \infty$ | 0 | 1 |

„beauty-and-beast pair“
Robust HMC sampler

Proposal distribution is mix between two proposal distributions:

\[ Q(z'|z) = \rho \cdot Q_{CP}(z'|z) + (1 - \rho) \cdot Q_{DNCP}(z'|z) \]

Where we use \( \rho = 0.5 \) in experiments
Experiment: HMC sampling with Dynamic Bayesian network (DBN) (1)

\[ p_\theta(z_t | z_{t-1}) = \mathcal{N}(z; \tanh(W z_{t-1}), \sigma_z^2 I) \]
\[ p_\theta(x_t | z_t) = \mathcal{N}(z; \tanh(W z_t), \sigma_x^2 I) \]

Equivalent recurrent neural net with injected noise
Autocorrelation results on DBN:

<table>
<thead>
<tr>
<th>log $\sigma_z$</th>
<th>CP</th>
<th>DNCP</th>
<th>ROBUST</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>2</td>
<td>305</td>
<td>640</td>
</tr>
<tr>
<td>-4.5</td>
<td>26</td>
<td>348</td>
<td>498</td>
</tr>
<tr>
<td>-4</td>
<td>10</td>
<td>570</td>
<td>686</td>
</tr>
<tr>
<td>-3.5</td>
<td>225</td>
<td>417</td>
<td>624</td>
</tr>
<tr>
<td>-3</td>
<td>386</td>
<td>569</td>
<td>596</td>
</tr>
<tr>
<td>-2.5</td>
<td>542</td>
<td>608</td>
<td>900</td>
</tr>
<tr>
<td>-2</td>
<td>406</td>
<td>972</td>
<td>935</td>
</tr>
<tr>
<td>-1.5</td>
<td>672</td>
<td>1078</td>
<td>918</td>
</tr>
<tr>
<td>-1</td>
<td>1460</td>
<td>1600</td>
<td>1082</td>
</tr>
</tbody>
</table>

CP (Terribly inefficient)  DNCP (Very efficient)
Conclusion

• Centered versus Non-centered parameterizations
  – Can make a huge difference in efficiency of sampling, dependent on model parameters
• Robust HMC sampler
• More theoretical and experimental results: See poster (this evening)
• Questions