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

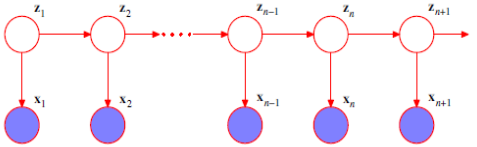
Durk Kingma

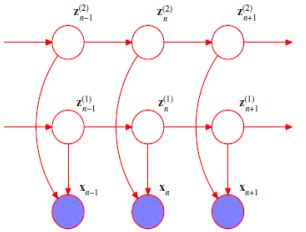
Max Welling

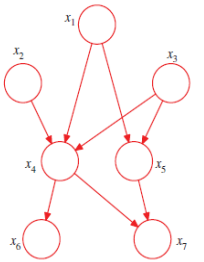
UNIVERSITEIT VAN AMSTERDAM

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:

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) / p_{\boldsymbol{\theta}}(\mathbf{x})$$

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) \propto p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$$

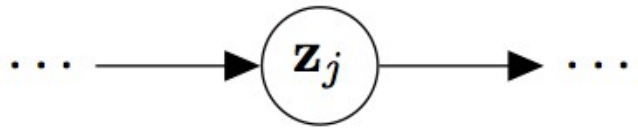
$$\nabla_{\mathbf{z}} \log p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \nabla_{\mathbf{z}} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$$

- Efficient gradient-based approximate inference methods:
  - sampling-based methods (
  - Stochastic Variational infer
    [Hoffman & Blei 2012] [King

Can we increase the efficiency
of gradient-based posterior inference?

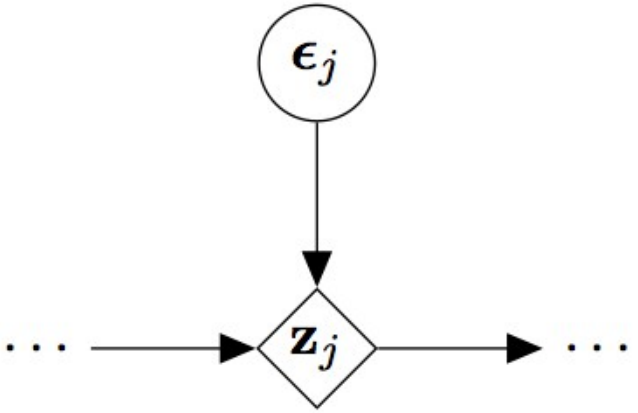
Idea: Reparameterizations (Gaussian example)

Centered Form (CP)



$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{pa}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \sigma^2 I)$$
$$\mu = f_{\boldsymbol{\theta}}(\mathbf{pa})$$
$$(\text{e.g. neural net})$$

Differentiable Non-centered Form (DNCP)

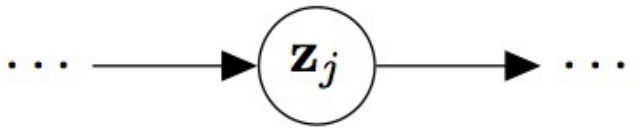


$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$$
$$\widetilde{\mathbf{z}} = f_{\boldsymbol{\theta}}(\mathbf{pa}) + \sigma\boldsymbol{\epsilon}$$

Neural net perspective:
hidden unit with injected noise

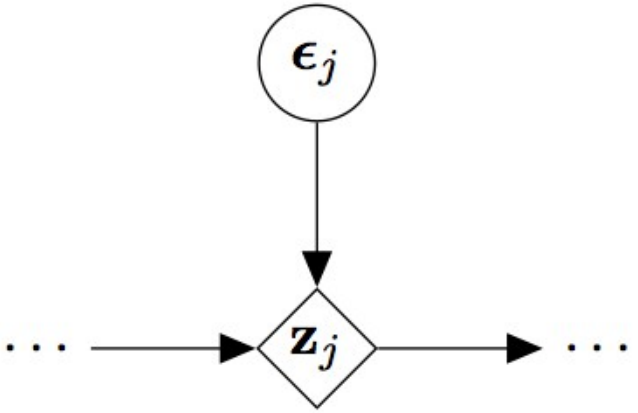
Reparameterizations (General form)

Centered Form (CP)



$$p_{\boldsymbol{\theta}}(\mathbf{z}_j|\mathbf{pa}_j)$$

Differentiable Non-centered Form (DNCP)



$$\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}_j)$$
$$\mathbf{z}_j = g_j(\mathbf{pa}_j, \boldsymbol{\epsilon}_j, \boldsymbol{\theta})$$

Reparameterizations

Can be performed for very broad class of distributions

Example	$q_\varphi(z)$	$p(\varepsilon)$	$g(\varphi, \varepsilon)$	Also...
Normal dist.	$z \sim N(\mu,\sigma)$	$\varepsilon \sim N(0,1)$	$z = \mu + \sigma * \varepsilon$	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	$z \sim \log N(\mu,\sigma)$	$\varepsilon \sim N(0,1)$	$z = \exp(\mu + \sigma * \varepsilon)$	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

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

$$\begin{aligned} \rho^2 &= (\sigma_{AB}^2)^2 / (\sigma_A^2 \sigma_B^2) \\ &= (H_{AB}/det(\mathbf{H}))^2 / ((-H_A/det(\mathbf{H}))(-H_B/det(\mathbf{H})) \\ &= H_{AB}^2 / (H_A H_B) \end{aligned}$$

Poster correlation analysis (2)

- Squared correlations for CP:

$$\rho^2_{pa(z)_i,z} = \frac{(H_{pa(z)_i z})^2}{H_{pa(z)_i pa(z)_i} H_{zz}} = \frac{w_i^2/\sigma^4}{(\alpha - w_i^2/\sigma^2)(\sigma^{-2}_{z|ch(z)} - 1/\sigma^2)}$$

- Squared correlations for DNCP:

$$\rho^2_{pa(z)_i,\epsilon} = \frac{(H_{pa(z)_i \epsilon})^2}{H_{pa(z)_i pa(z)_i} H_{\epsilon\epsilon}} = \frac{\sigma^2 w_i^2 (\sigma^{-2}_{z|ch(z)})^2}{(\alpha + w_i^2(\sigma^{-2}_{z|ch(z)}))(\sigma^2 \sigma^{-2}_{z|ch(z)} - 1)}$$

Poster correlation analysis (3)

Inequality: all terms cancel with very simple result:

$$\rho^2_{pa(z)_i,z} > \rho^2_{pa(z)_i,\epsilon}$$

$$\Leftrightarrow$$

$$\sigma^2_{z|pa(z)} < \sigma^2_{z|ch(z)}$$

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:

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

„beauty-and-beast pair“

Robust HMC sampler

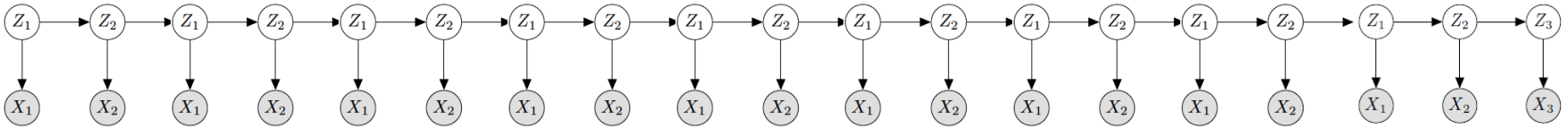
Proposal distribution is mix between two proposal distributi

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

Where we use ρ = 0.5 in experiments

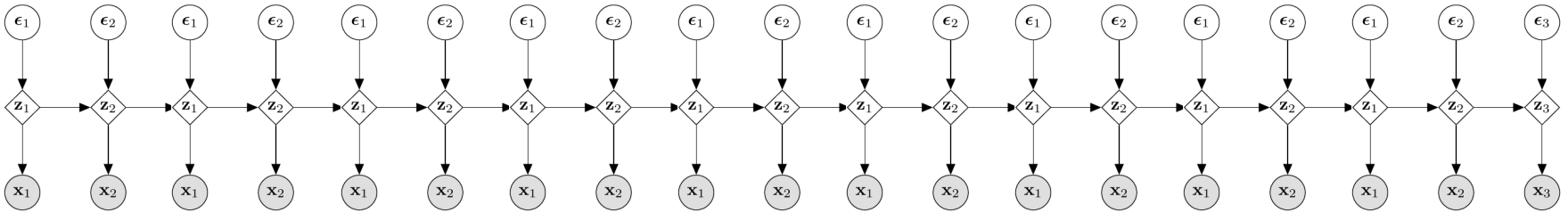
Experiment: HMC sampling with Dynamic Bayesian network (DBN) (1)

Dynamic Bayesian Network



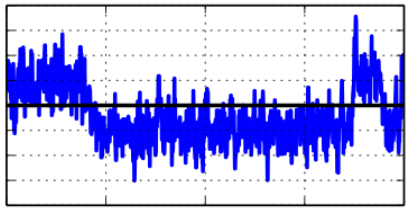
$$p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}; \tanh(W\mathbf{z}_{t-1}), \sigma_z^2 I)$$
$$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}; \tanh(W\mathbf{z}_t), \sigma_x^2 I)$$

Equivalent recurrent neural net with injected noise

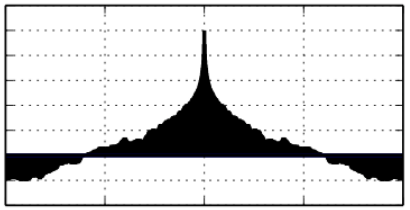


Autocorrelation results on DBN:

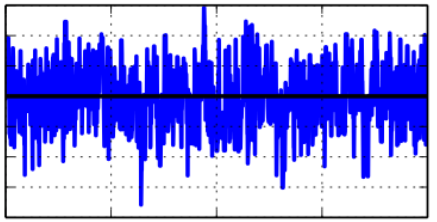
$\log \sigma_z$	CP	DNCP	ROBUST
-5	2	305	640
-4.5	26	348	498
-4	10	570	686
-3.5	225	417	624
-3	386	569	596
-2.5	542	608	900
-2	406	972	935
-1.5	672	1078	918
-1	1460	1600	1082



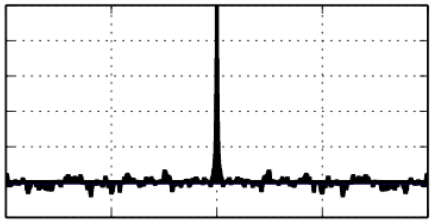
Samples



Autocorrelation



Samples



Autocorrelation

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