The Variational Predictive Natural Gradient

Da Tang\textsuperscript{1}  Rajesh Ranganath\textsuperscript{2}

\textsuperscript{1}Columbia University
\textsuperscript{2}New York University

June 12, 2019
Latent variable models: $p(x, z; \theta) = p(z)p(x | z; \theta)$. 
Latent variable models: $p(x, z; \theta) = p(z)p(x|z; \theta)$.

Variational inference approximates the posterior through maximizing the ELBO:

$$L(\lambda, \theta) = \mathbb{E}_q[\log p(x|z; \theta)] - \text{KL}(q(z|x; \lambda)||p(z)).$$
Variational Inference

- Latent variable models: \( p(x, z; \theta) = p(z)p(x|z; \theta) \).
- Variational inference approximates the posterior through maximizing the ELBO:

\[
\mathcal{L}(\lambda, \theta) = \mathbb{E}_q [\log p(x|z; \theta)] - \text{KL}(q(z|x; \lambda) \| p(z)).
\]

- \( q \)-Fisher Information \( F_q = \mathbb{E}_q \left[ \nabla_\lambda \log q(z|x; \lambda) \cdot \nabla_\lambda \log q(z|x; \lambda)^\top \right] \) (Hoffman et al., 2013) approximates the negative Hessian of the objective.
- The natural gradient: \( \nabla^\text{NG}_\lambda \mathcal{L}(\lambda) = F_q^{-1} \cdot \nabla_\lambda \mathcal{L}(\lambda) \).
The curvature of the ELBO may be pathological.
The curvature of the ELBO may be pathological.

Example: A bivariate Gaussian model with unknown mean and known covariance

\[
\Sigma = \begin{pmatrix} 1 & 1 - \varepsilon \\ 1 - \varepsilon & 1 \end{pmatrix}, \quad 0 < \varepsilon \ll 1.
\]
Pathological Curvature of the ELBO

- The curvature of the ELBO may be pathological.
- Example: A bivariate Gaussian model with unknown mean and known covariance
  \[ \Sigma = \begin{pmatrix} 1 & 1 - \varepsilon \\ 1 - \varepsilon & 1 \end{pmatrix}, \ 0 < \varepsilon \ll 1. \]

- The natural gradient fails to help.
Limitations of the $q$-Fisher information:

- Approximates the Hessian of the objective well only when $q(z|x; \lambda) \approx p(z|x; \theta)$.

- Ignore the model likelihood $p(x|z; \theta)$ in computations.
Construct a positive definite matrix that resembles the negative Hessian of the expected log-likelihood part $L^\text{ll} = \mathbb{E}_{q(z|x;\lambda)} [\log p(x|z;\theta)]$ of the ELBO.
Construct a \textbf{positive definite} matrix that resembles the negative Hessian of the expected log-likelihood part $L_{\text{ll}} = \mathbb{E}_{q(z|x; \lambda)}[\log p(x|z; \theta)]$ of the ELBO.

Reparameterize the variational distribution $q$:

$$z = g(x, \varepsilon; \lambda) \sim q(z|x; \lambda) \iff \varepsilon \sim s(\varepsilon).$$
Construct a positive definite matrix that resembles the negative Hessian of the expected log-likelihood part $L_{\text{ll}} = \mathbb{E}_{q(z|x;\lambda)} [\log p(x|z;\theta)]$ of the ELBO.

Reparameterize the variational distribution $q$:

$$z = g(x, \epsilon; \lambda) \sim q(z|x;\lambda) \iff \epsilon \sim s(\epsilon).$$

The variational predictive Fisher information:

$$F_r = \mathbb{E}_\epsilon \left[ \mathbb{E}_{p(x'|z=g(x,\epsilon;\lambda);\theta)} \left[ \nabla_{\lambda,\theta} \log p(x'|z=g(x,\epsilon;\lambda);\theta) \cdot \nabla_{\lambda,\theta} \log p(x'|z=g(x,\epsilon;\lambda);\theta)^\top \right] \right],$$

exactly the “expected” Fisher information of the reparameterized predictive distribution $p(x'|z=g(x,\epsilon;\lambda);\theta)$. 

The Variational Predictive Fisher Information
Variational predictive Fisher captures the curvature of variational inference.
Variational predictive Fisher captures the curvature of variational inference.

Matrix spectrum comparison (for the bivariate Gaussian example):

(d) Precision mat $\Sigma^{-1}$  
(e) $q$-Fisher info $F_q$  
(f) Our Fisher info $F_r$
The variational predictive natural gradient (VPNG):

$$\nabla_{\lambda, \theta}^{\text{VPNG}} \mathcal{L} = F_r^{-1} \cdot \nabla_{\lambda, \theta} \mathcal{L}(\lambda, \theta).$$

In practice, use Monte Carlo estimations to approximate $F_r$ and add a small dampening parameter to ensure invertibility.
The variational predictive natural gradient (VPNG):

\[ \nabla_{\lambda, \theta}^{\text{VPNG}} \mathcal{L} = F_r^{-1} \cdot \nabla_{\lambda, \theta} \mathcal{L}(\lambda, \theta). \]

In practice, use Monte Carlo estimations to approximate \( F_r \) and add a small dampening parameter to ensure invertibility.
Experiments: Bayesian Logistic Regression

- Tested on synthetic data with high correlations.
- Empirical results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Train AUC</th>
<th>Test AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>0.734 ± 0.017</td>
<td>0.718 ± 0.022</td>
</tr>
<tr>
<td>NG</td>
<td>0.744 ± 0.043</td>
<td>0.751 ± 0.047</td>
</tr>
<tr>
<td>VPGN</td>
<td><strong>0.972 ± 0.011</strong></td>
<td><strong>0.967 ± 0.011</strong></td>
</tr>
</tbody>
</table>

Table: Bayesian Logistic regression AUC
Figure: Learning curves of variational autoencoders (upper) and variational matrix factorization (lower) on real datasets.
The VPNG corrects for curvature in the objective between the parameters in variational inference.

Future work includes extending to general Bayesian networks with multiple stochastic layers.
Conclusion and Future Work

- The VPNG corrects for curvature in the objective between the parameters in variational inference.

- Future work includes extending to general Bayesian networks with multiple stochastic layers.
Thanks!

Poster #234