The Variational Predictive Natural Gradient

Da Tang[1] Rajesh Ranganath[2]

[1]Columbia University

[2]New York University

June 12, 2019

Variational Inference

- Latent variable models: $p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})$.

Variational Inference

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

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

Variational Inference

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

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

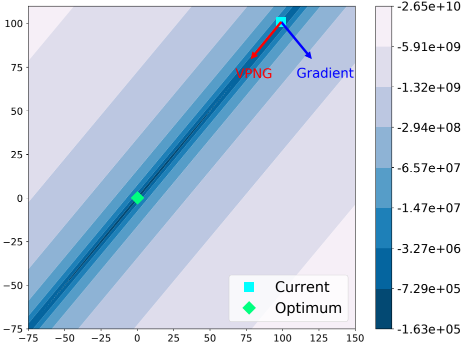
- q-Fisher Information $F_q = \mathbb{E}_q \left[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda}) \cdot \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda})^\top\right]$ (Hoffman et al., 2013) approximates the negative Hessian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

Pathological Curvature of the ELBO

- The curvature of the ELBO may be pathological.

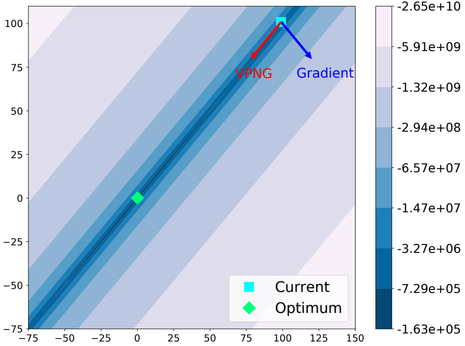
Pathological Curvature of the ELBO

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



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.

The Natural Gradient is Insufficient

Limitations of the q-Fisher information:

- Approximates the Hessian of the objective well only when $q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda}) \approx p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})$.
- Ignore the model likelihood $p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$ in computations.

The Variational Predictive Fisher Information

- Construct a positive definite matrix that resembles the negative Hessian of the expected log-likelihood part $\mathcal{L}^{\mathrm{ll}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})\right]$ of the ELBO.

The Variational Predictive Fisher Information

- Construct a positive definite matrix that resembles the negative Hessian of the expected log-likelihood part $\mathcal{L}^{\mathrm{ll}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})\right]$ of the ELBO.
- Reparameterize the variational distribution q:

$$\boldsymbol{z} = g(\boldsymbol{x}, \varepsilon; \boldsymbol{\lambda}) \sim q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda}) \iff \varepsilon \sim s(\varepsilon).$$

The Variational Predictive Fisher Information

- Construct a positive definite matrix that resembles the negative Hessian of the expected log-likelihood part $\mathcal{L}^{\mathrm{ll}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda})}[\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})]$ of the ELBO.
- Reparameterize the variational distribution q:

$$\boldsymbol{z} = g(\boldsymbol{x}, \varepsilon; \boldsymbol{\lambda}) \sim q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda}) \iff \varepsilon \sim s(\varepsilon).$$

- The variational predictive Fisher information:

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

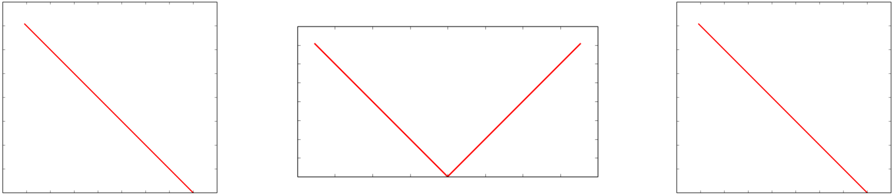
exactly the "expected" Fisher information of the *reparameterized predictive distribution* $p(\boldsymbol{x}'|\boldsymbol{z} = g(\boldsymbol{x}, \varepsilon; \boldsymbol{\lambda}); \boldsymbol{\theta})$.

The Variational Predictive Fisher Information

- Variational predictive Fisher captures the curvature of variational inference.

The Variational Predictive Fisher Information

- Variational predictive Fisher captures the curvature of variational inference.
- Matrix spectrum comparison (for the bivariate Gaussian example):



(d) Precision mat Σ^{-1} (e) q-Fisher info F_q (f) Our Fisher info F_r

The Variational Predictive Natural Gradient

- The variational predictive natural gradient (VPNG):

$$\nabla^{\mathrm{VPNG}}_{\boldsymbol{\lambda},\boldsymbol{\theta}}\mathcal{L} \;=\; F_r^{-1} \cdot \nabla_{\boldsymbol{\lambda},\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta}).$$

The Variational Predictive Natural Gradient

- The variational predictive natural gradient (VPNG):

$$\nabla^{\mathrm{VPNG}}_{\boldsymbol{\lambda},\boldsymbol{\theta}} \mathcal{L} \;=\; F_r^{-1} \cdot \nabla_{\boldsymbol{\lambda},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta}).$$

- 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:

Method	**Train AUC**	**Test AUC**
Gradient	0.734 ± 0.017	0.718 ± 0.022
NG	0.744 ± 0.043	0.751 ± 0.047
VPNG	$\mathbf{0.972 \pm 0.011}$	$\mathbf{0.967 \pm 0.011}$

Table: Bayesian Logistic regression AUC

Experiments: VAE and VMF

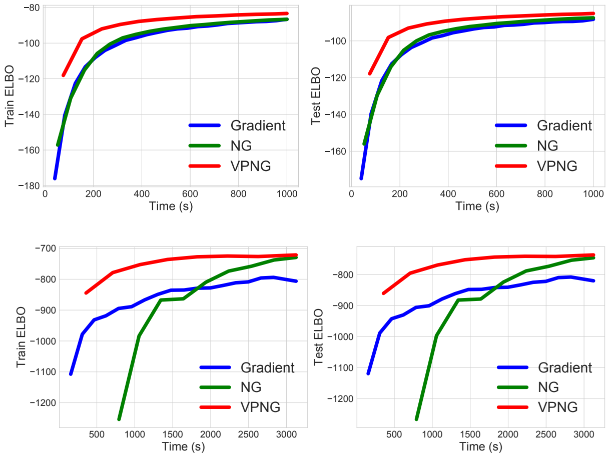


Figure: Learning curves of variational autoencoders (upper) and variational matrix factorization (lower) on real datasets.

Conclusion and Future Work

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

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

Code available at https://github.com/datang1992/VPNG.