Overdispersed Black-Box Variational Inference

Francisco J. R. Ruiz, Michalis K. Titsias, David M. Blei

Columbia University
Athens University of Economics and Business

June 27th, 2016

Overdispersed Black-Box Variational Inference

- General variational inference for any probabilistic model
- Builds on black-box variational inference (BBVI)
- Reduces the variance of the estimator ($\implies$ faster convergence)
- Requires a variational distribution in the exponential family
- Key idea: analyze the optimal *importance sampling proposal*

Notation

- Probabilistic model $p(\mathbf{x}, \mathbf{z})$
  - $\mathbf{x}$: Data
  - $\mathbf{z}$: Latent variables
- Assume the posterior $p(\mathbf{z} \mid \mathbf{x})$ is intractable:

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}}$$

- We wish to *approximate* the posterior using **variational inference**

Variational Inference

- Approximate the posterior with a simpler distribution $q(\mathbf{z}; \boldsymbol{\lambda})$
- Minimize the KL divergence w.r.t. $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda}^\star = \arg\min_{\boldsymbol{\lambda}} D_{\mathrm{KL}}(q(\mathbf{z}; \boldsymbol{\lambda}) || p(\mathbf{z} \mid \mathbf{x}))$$

- Evidence lower bound (ELBO):

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})}\left[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \boldsymbol{\lambda})\right]$$

Variational Inference

- Approximate the posterior with a simpler distribution $q(\mathbf{z};\boldsymbol{\lambda})$
- Minimize the KL divergence w.r.t. $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda}^\star = \arg\min_{\boldsymbol{\lambda}} D_{\mathrm{KL}}(q(\mathbf{z};\boldsymbol{\lambda})||p(\mathbf{z}\,|\,\mathbf{x}))$$

- Evidence lower bound (ELBO):

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}\left[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\boldsymbol{\lambda})\right]$$

- Optimization problem:
  - Conditionally conjugate models: *coordinate ascent*
  - Non-conjugate models: one recent approach is **BBVI**

Examples of Conditionally Non-Conjugate Models

- Time series models
- Probabilistic matrix factorization
- Deep probabilistic models
- Correlated topic models
- ...

Black-Box Variational Inference[1]

- Stochastic optimization
- Builds Monte Carlo estimates of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}$
- Relies on the score function method:

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L} = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}\left[f(\mathbf{z})\right],$$

where

$$f(\mathbf{z}) \triangleq \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z};\boldsymbol{\lambda})\left(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\boldsymbol{\lambda})\right)$$

[1] Ranganath et al. (2014)

Black-Box Variational Inference[1]

- Stochastic optimization
- Builds Monte Carlo estimates of the gradient $\nabla_{\boldsymbol{\lambda}} \mathcal{L}$
- Relies on the score function method:

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})} \left[f(\mathbf{z}) \right],$$

  where

$$f(\mathbf{z}) \triangleq \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z};\boldsymbol{\lambda}) \left(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\boldsymbol{\lambda}) \right)$$

- Algorithm:
  1. Sample $\mathbf{z}^{(s)}$ iid from $q(\mathbf{z};\boldsymbol{\lambda})$
  2. Evaluate $f(\mathbf{z}^{(s)})$ for each sample s
  3. Obtain a Monte Carlo estimate of the gradient
  4. Take a gradient step for $\boldsymbol{\lambda}$

[1] Ranganath et al. (2014)

Controlling the Variance

- The estimator of the gradient may suffer from **high variance**
- This leads to slow convergence
- Methods to reduce the variance:
  - Rao-Blackwellization[2]
  - Control variates[3]
  - Reparameterization trick[4]
  - Local expectations[5]

---

[2] Casella and Robert (1996); Ranganath et al. (2014)

[3] Ross (2002); Paisley et al. (2012); Ranganath et al. (2014); Gu et al. (2016)

[4] Price (1958); Bonnet (1964); Salimans and Knowles (2013); Kingma and Welling (2014); Rezende et al. (2014); Kucukelbir et al. (2015)

[5] Titsias and Lázaro-Gredilla (2015)

Controlling the Variance

- The estimator of the gradient may suffer from **high variance**
- This leads to slow convergence
- Methods to reduce the variance:
  - Rao-Blackwellization[2]
  - Control variates[3]
  - Reparameterization trick[4]
  - Local expectations[5]
- New method: Overdispersed BBVI

[2] Casella and Robert (1996); Ranganath et al. (2014)

[3] Ross (2002); Paisley et al. (2012); Ranganath et al. (2014); Gu et al. (2016)

[4] Price (1958); Bonnet (1964); Salimans and Knowles (2013); Kingma and Welling (2014); Rezende et al. (2014); Kucukelbir et al. (2015)

[5] Titsias and Lázaro-Gredilla (2015)

Overdispersed BBVI

- Builds on BBVI
- Samples from another distribution $r(\mathbf{z}) \neq q(\mathbf{z}; \boldsymbol{\lambda})$

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})} [f(\mathbf{z})] = \mathbb{E}_{r(\mathbf{z})} \left[f(\mathbf{z}) \frac{q(\mathbf{z}; \boldsymbol{\lambda})}{r(\mathbf{z})} \right]$$

- The optimal importance sampling proposal[6] is

$$r_n^{\star}(\mathbf{z}) \propto q(\mathbf{z}; \boldsymbol{\lambda}) |f_n(\mathbf{z})|$$

- The optimal proposal is intractable
- O-BBVI searches for another proposal $r(\mathbf{z})$

[6] Robert and Casella (2005); Owen (2013)

Overdispersed BBVI

- Assume an exponential family variational distribution

$$q(\mathbf{z}; \boldsymbol{\lambda}) \propto \exp\{\boldsymbol{\lambda}^\top t(\mathbf{z}) - A(\boldsymbol{\lambda})\}$$

- Recall the optimal proposal:

$$r_n^\star(\mathbf{z}) \propto q(\mathbf{z}; \boldsymbol{\lambda}) |f_n(\mathbf{z})|$$
$$f_n(\mathbf{z}) = \nabla_{\lambda_n} \log q(\mathbf{z}; \boldsymbol{\lambda}) \left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \boldsymbol{\lambda})\right)$$

[7] Jørgensen (1987)

Overdispersed BBVI

- Assume an exponential family variational distribution

$$q(\mathbf{z}; \boldsymbol{\lambda}) \propto \exp\{\boldsymbol{\lambda}^\top t(\mathbf{z}) - A(\boldsymbol{\lambda})\}$$

- Recall the optimal proposal:

$$r_n^\star(\mathbf{z}) \propto q(\mathbf{z}; \boldsymbol{\lambda})|f_n(\mathbf{z})|$$
$$f_n(\mathbf{z}) = \nabla_{\lambda_n} \log q(\mathbf{z}; \boldsymbol{\lambda}) \left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \boldsymbol{\lambda})\right)$$

- The optimal proposal assigns higher mass in the *tails* of $q(\mathbf{z}; \boldsymbol{\lambda})$

[7] Jørgensen (1987)

Overdispersed BBVI

- Assume an exponential family variational distribution

$$q(\mathbf{z}; \boldsymbol{\lambda}) \propto \exp\{\boldsymbol{\lambda}^\top t(\mathbf{z}) - A(\boldsymbol{\lambda})\}$$

- Recall the optimal proposal:

$$r_n^\star(\mathbf{z}) \propto q(\mathbf{z}; \boldsymbol{\lambda}) |f_n(\mathbf{z})|$$
$$f_n(\mathbf{z}) = \nabla_{\lambda_n} \log q(\mathbf{z}; \boldsymbol{\lambda}) \left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \boldsymbol{\lambda})\right)$$

- The optimal proposal assigns higher mass in the *tails* of $q(\mathbf{z}; \boldsymbol{\lambda})$
- We use an **overdispersed distribution**[7]

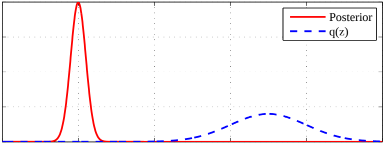
$$r(\mathbf{z}; \boldsymbol{\lambda}, \tau) \propto \exp\left\{\frac{\boldsymbol{\lambda}^\top t(\mathbf{z}) - A(\boldsymbol{\lambda})}{\tau}\right\}$$

[7] Jørgensen (1987)

Heavier Tails

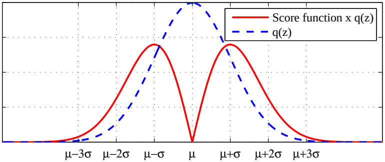
$$r_n^\star(\mathbf{z}) \propto q(\mathbf{z};\boldsymbol{\lambda})|f_n(\mathbf{z})|$$
$$f_n(\mathbf{z}) = \nabla_{\lambda_n} \log q(\mathbf{z};\boldsymbol{\lambda})\left(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\boldsymbol{\lambda})\right)$$

- Through the model $p(\mathbf{x},\mathbf{z})$



- Through the score function

$$\nabla_{\lambda_n} \log q(\mathbf{z};\boldsymbol{\lambda}) = t_n(\mathbf{z}) - \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}\left[t_n(\mathbf{z})\right]$$



Implementation

- Importance sampling fails in high dimensionality settings
  $\rightarrow$ We use local expectations[8]
  $\rightarrow$ A proposal distribution per latent variable

$$\begin{aligned}
\nabla_{\lambda_n}\mathcal{L} &= \mathbb{E}_{q(z_n;\lambda_n)}\left[\mathbb{E}_{q(\mathbf{z}_{\neg n};\boldsymbol{\lambda}_{\neg n})}\left[f_n(\mathbf{z})\right]\right] \\
&= \mathbb{E}_{r(z_n;\lambda_n,\tau_n)}\left[\frac{q(z_n;\lambda_n)}{r(z_n;\lambda_n,\tau_n)}\mathbb{E}_{q(\mathbf{z}_{\neg n};\boldsymbol{\lambda}_{\neg n})}\left[f_n(\mathbf{z})\right]\right]
\end{aligned}$$

[8] Titsias and Lázaro-Gredilla (2015)

Implementation

- Importance sampling fails in high dimensionality settings
  $\rightarrow$ We use local expectations[8]
  $\rightarrow$ A proposal distribution per latent variable

$$\nabla_{\lambda_n}\mathcal{L} = \mathbb{E}_{q(z_n;\lambda_n)}\left[\mathbb{E}_{q(\mathbf{z}_{\neg n};\boldsymbol{\lambda}_{\neg n})}\left[f_n(\mathbf{z})\right]\right]$$
$$= \mathbb{E}_{r(z_n;\lambda_n,\tau_n)}\left[\frac{q(z_n;\lambda_n)}{r(z_n;\lambda_n,\tau_n)}\mathbb{E}_{q(\mathbf{z}_{\neg n};\boldsymbol{\lambda}_{\neg n})}\left[f_n(\mathbf{z})\right]\right]$$

- Algorithm:
  1. Sample $\mathbf{z}^{(0)} \sim q(\mathbf{z};\boldsymbol{\lambda})$
  2. For each n, sample $z_n^s \sim r(z_n;\lambda_n,\tau_n)$, for $s = 1,\ldots,S$
  3. For each n, obtain a Monte Carlo estimate of $\nabla_{\lambda_n}\mathcal{L}$
  4. Take a gradient step for $\boldsymbol{\lambda}$

[8] Titsias and Lázaro-Gredilla (2015)

Implementation

- Need to choose the dispersion coefficients τ_n
  $\rightarrow$ Gradient steps for τ_n to minimize the variance
  $\rightarrow$ Monte Carlo estimator with little extra overhead

[9] Veach and Guibas (1995)

Implementation

- Need to choose the dispersion coefficients τ_n
  $\rightarrow$ Gradient steps for τ_n to minimize the variance
  $\rightarrow$ Monte Carlo estimator with little extra overhead
- High variance of the importance weights
  $\rightarrow$ Multiple importance sampling[9]
  $\rightarrow$ The proposal $r(z_n; \lambda_n, \tau_{n1}, \tau_{n2})$ can be a mixture

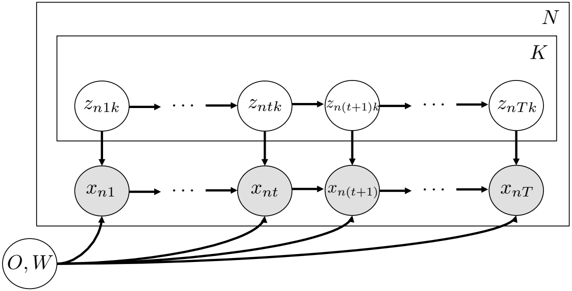
[9] Veach and Guibas (1995)

Full Algorithm

- Control variates
- Rao-Blackwellization
- O-BBVI with
  - Local expectations
  - Adaptation of the dispersion coefficients
  - Multiple proposals

Experiments: GN-TS Model

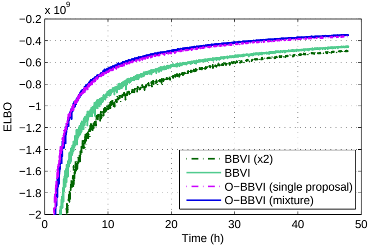
Gamma-Normal Time Series[10] Model

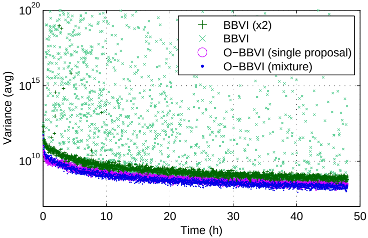


[10] Ranganath et al. (2014)

Experiments: GN-TS Model

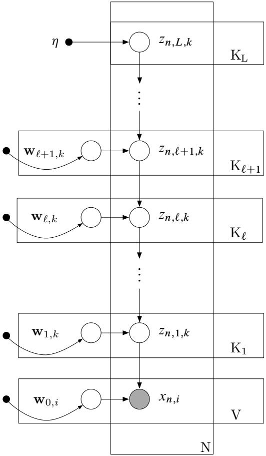
Dataset: Synthetic





Experiments: DEF

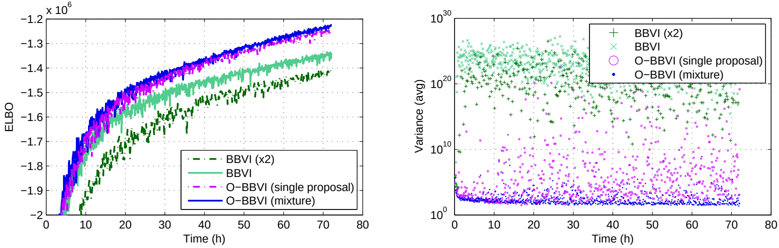
Poisson Deep Exponential Family[11]



[11] Ranganath et al. (2015)

Experiments: DEF

Dataset: Papers in NIPS'11 conference



Summary: O-BBVI

- Unconventional application of importance sampling to general VI
- Reduce the variance of the gradient estimator
- Lower variance than BBVI with $2\times$ Monte Carlo samples
- Faster convergence

Thank you for your attention!