Survey on Stochastic Variational Inference

Debojyoti Dey

Indian Institute of Technology Kanpur

debojyot@cse.iitk.ac.in

April 28, 2017

Overview

1. Variational Inference
   - Definition
   - Mean-field VB
   - Conjugate model
   - Coordinate ascent
2. Natural gradient
3. Stochastic Variational Inference
4. Black Box Variational Inference
   - Definition
   - Controlling Variance

Approximating Posterior

Aim is to compute posterior

$$p(z|x) = \frac{p(x,z)}{p(x)}$$

Oftentimes, marginal likelihood $p(x)$ in not available in closed form or computing takes exponential time.

Role of VB: Approximate posterior with *exact conditional* $q(z) \in \mathcal{Q}$, where $\mathcal{Q}$ is a family of pdf over latent variables.

$$q^\star(z) = \underset{q(z)\in\mathcal{Q}}{\arg\min}\, KL(q(z)||p(z|x))$$

Equivalent to maximizing *Variational Lower bound* given by,

$$ELBO(q) = \mathbb{E}_{q(z)}\left[\log \frac{p(x,z)}{q(z)}\right]$$

Mean Field Variational Family

Latent variables are assumed to be independent.

$$q(\bar{z}) = \prod_{j=1}^{m} q_j(z_j)$$

Doesn't take correlation into account.

Example Model: Hierarchical model with global(β) and local(z) latent variables.

$$p(\beta, z, x) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

$$q(\beta, z) = q(\beta) \prod_{i} q(z_i)$$

Complete Data Conjugate Model

Complete conditionals over global parameters are assumed to be from exponential family,

$$p(\beta|x, z, \alpha) \propto \exp\{\langle \eta_g(x, z, \alpha), t(\beta)\rangle - a_g(\eta_g(x, z, \alpha))\}$$

$$p(z_{nj}|x_n, z_{n,-j}, \beta) \propto \exp\{\langle \eta_l(x_n, z_{n,-j}, \beta), t(z_{nj})\rangle - a_l(\eta_l(x_n, z_{n,-j}, \beta))\}$$

This implies conjugacy relationship between global variable β and local context (z_n, x_n).

Variational family,

$$q(\beta|\lambda) \propto \exp\{\langle \lambda, t(\beta)\rangle - a_g(\lambda)\}$$

$$q(z_{nj}|\phi_{nj}) \propto \exp\{\langle \phi_{nj}, t(z_{nj})\rangle - a_l(\phi_{nj})\}$$

Evidence Lower bound and Coordinate ascent

We need to maximize ELBO,

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)] \tag{1}$$

By virtue of mean-field assumption i.e. independence between variational parameters,

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta|x, z)] - \mathbb{E}_q[\log q(\beta)] + const$$

Replacing $\mathbb{E}_q[t(\beta)] = \nabla_\lambda a_g(\lambda)$ and differentiating,

$$\nabla_\lambda \mathcal{L} = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda) \tag{2}$$

Similarly,

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla^2_{\phi_{nj}} a_g(\phi_{nj})(\mathbb{E}_q[\eta_l(x_n, z_{n,-j}, \beta)] - \phi_{nj}) \tag{3}$$

Coordinate Ascent Variational Inference(CAVI)

Algorithm 1 Coordinate Ascent VI

```
Initialize λ^0 randomly
repeat
    for each local variational parameter φ_nj do
        Update φ_nj^(t): φ_nj^(t) = E_{q^(t-1)}[η_l(x_n, z_{n,-j}, β)]
    end for
    Update the global variational parameters, λ^(t) = E_{q^(t)}[η_g(z_{1:N}, x_{1:N})]
until the ELBO converges
```

Initialize λ^0 randomly
repeat
 for each local variational parameter ϕ_{nj} **do**
  Update $\phi_{nj}^{(t)}$: $\phi_{nj}^{(t)} = \mathbb{E}_{q^{(t-1)}}[\eta_l(x_n, z_{n,-j}, \beta)]$
 end for
 Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{q^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$
until the ELBO converges

Note: Convergence using threshold of change in ELBO

Problems:

- Random initialization of λ
- Updating all local variational parameter based on the random initialization.

Natural Gradient of ELBO

Gradient update

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho \nabla_\lambda f(\lambda^{(t)})$$

gives steepest ascent in Euclidean space of λ.

Realistic distance between pdf: Symmetrized KL

$$D_{KL}^{sym}(\lambda, \lambda') = \mathbb{E}_\lambda \left[\log \frac{q(\beta|\lambda)}{q(\beta|\lambda')} \right] + \mathbb{E}'_\lambda \left[\log \frac{q(\beta|\lambda')}{q(\beta|\lambda)} \right]$$

can be expressed in terms of Riemannian metric $G(\lambda)$ giving linear transformation to λ, Euclidean distance between λ and $\lambda + d\lambda$ in transformed space

$$d\lambda^T G(\lambda) d\lambda = D_{KL}^{sym}(\lambda, \lambda + d\lambda)$$

Natural Gradient of ELBO

Natural gradient of $f(\lambda)$ is shown to be,

$$\hat{\nabla}_\lambda f(\lambda) \triangleq G(\lambda)^{-1} \nabla_\lambda f(\lambda)$$

G is Fisher information matrix of $q(\lambda)$,

$$G(\lambda) = \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta|\lambda))(\nabla_\lambda \log q(\beta|\lambda))^T \right] \approx \nabla^2_\lambda a_g(\lambda)$$

So natural gradient has simple form,

$$\hat{\nabla}_\lambda \mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda$$

Similarly follows,

$$\hat{\nabla}_{\phi_{nj}} \mathcal{L}(\phi_{nj}) = \mathbb{E}_q[\eta_l(x_n, z_{n,-j}, \beta)] - \phi_{nj}$$

Stochastic VB

Properties of SVI:

1. Sample a data point $i \sim Unif(1, 2, \ldots, N)$ at random. Optimize its local variational parameters
2. Form intermediate global variational parameters. Traditional coordinate ascent with sampled data point repeated N times in computing ELBO, producing natural gradient

$$\hat{\nabla}_\lambda \mathcal{L}_i = \alpha + N.(\mathbb{E}_{\phi_i(\lambda)}[t(x_i, z_i)], 1) - \lambda$$

$$\hat{\lambda}_t \triangleq \alpha + N.(\mathbb{E}_{\phi_i(\lambda)}[t(x_i, z_i)], 1) \tag{4}$$

3. Update global variational parameter as weighted avg of intermediate and current value.

$$\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}_t \tag{5}$$

SVI Algorithm

Algorithm 2 Stochastic Variational Inference

Initialize λ^0 randomly

Set the step size ρ_t s.t. $\sum \rho_t = \infty$, $\sum \rho_t^2 < \infty$

repeat

  Sample a data point x_i uniformly at random.

  Compute its local variational parameter

$$\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})]$$

  Compute intermediate global parameters with x_i is replicated N times as equation 4

  Update the current estimate of global variational parameters as equation 5

until the ELBO converges

Black Box Variational Inference

General model with observations x and parameter θ having joing distribution

$$p(\theta, x) = p(\theta) \prod_{i=1}^{N} p(x^i | \theta)$$

and variational distribution of θ,

$$q(\theta) = q(\theta|\lambda)$$

with free parameter λ

$$\mathcal{L} = \mathbb{E}_q(\theta) \left[\log \frac{p(\theta, x)}{q(\theta)} \right]$$

$$\nabla_\lambda \mathcal{L} = \nabla_\lambda \int q(\theta) \log \frac{p(\theta, x)}{q(\theta)} d\theta$$

Black Box Variational Inference

Gradient can be obtained using Monte Carlo approximation

$$
\begin{aligned}
\nabla_\lambda \mathcal{L} &= \int \nabla_\lambda \log \frac{p(\theta, x)}{q(\theta)} q(\theta) d\theta + \int \log \frac{p(\theta, x)}{q(\theta)} \nabla_\lambda (q_\theta) d\theta \\
&= 0 + \mathbb{E}_{q(\theta)} \left[\log \frac{p(\theta, x)}{q(\theta)} \nabla_\lambda \log q(\theta) \right] \\
&\approx \frac{1}{|S|} \sum_{\hat{\theta} \in S} \log \frac{p(\hat{\theta}, x)}{q(\hat{\theta})} \nabla_\lambda \log q(\hat{\theta})
\end{aligned}
$$

Score function $\nabla_\lambda \log q(\theta)$. Scalable BBVI with subsampling

$$
\nabla_\lambda \mathcal{L} \approx \frac{1}{|S|} \sum_{\hat{\theta} \in S} \left(\log \frac{p(\hat{\theta}, x)}{q(\hat{\theta})} + N \log p(x_i | \hat{\theta}) \right) \nabla_\lambda \log q(\hat{\theta})
$$

where $i \sim Uniform(1, 2, \ldots, N)$

Reducing the Variance

Because of two levels of sampling, stochastic version of BBVI has gradient with high variance. Small steps will lead to slow convergence.

Variance reduction methods

- Rao Blackwellization: replacing random variable by replacing with conditional expectation wrt subset of variables.
- Control Variates
- Reparameterization gradient instead of score function gradient

References I

- Elaine Angelino, Matthew James Johnson, Ryan P Adams, et al., *Patterns of scalable bayesian inference*, Foundations and Trends® in Machine Learning **9** (2016), no. 2-3, 119–247.
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe, *Variational inference: A review for statisticians*, Journal of the American Statistical Association (2017), no. just-accepted.
- Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley, *Stochastic variational inference.*, Journal of Machine Learning Research **14** (2013), no. 1, 1303–1347.
- James Martens and Roger B Grosse, *Optimizing neural networks with kronecker-factored approximate curvature.*, ICML, 2015, pp. 2408–2417.

References II

Stephan Mandt, Matthew D Hoffman, and David M Blei, *A variational analysis of stochastic gradient algorithms*, arXiv preprint arXiv:1602.02666 (2016).

Rajesh Ranganath, Sean Gerrish, and David Blei, *Black box variational inference*, Artificial Intelligence and Statistics, 2014, pp. 814–822.

The End