Name: Jikhan Jeong

1.State the complete iterative algorithm for variational inference to approximate a marginal density via a candidate function that factorizes into the product of densities of subvectors of the parameter vector.

Variational inference (VI) turns the inference problem as an optimization problem. When the data set has a lot of latent variable, MCMC may not reasonable due to its computational load; therefore, variational inference may be a good choice with its assumption with mean-field for factorizing the distribution of latent variables.

VI approximate the posterior and it is

- full Bayesian approach (all parameters are random variable)
- By using approximation, solve the complex probability function as a simpler probability function, in here, we can regard probability density function as a set function like defining the probability as a set function with measure approach.
- VI is about approximation of posterior distribution
- Mean-field theory is applied to easier functional form of distribution, even though the assumption may look strong if the distribution of latent variables is correlated (=dependent)
- Minimization of the difference of the approximated posterior distribution and the actual one is the dual problem of maximization of the similarity between approximated posterior distribution and the true one, in this setting, minimization of Kullback-Leibler Divergence(KL) is the same problem with maximization of The evidence lower bound (ELBO)

* Proof. Minimizing the KL = Maximizing the ELBO_____

The KL for the VI is

$\mathrm{KL}(q||p) = E_q\left[log\frac{q(Z)}{p(Z|x)}\right], X\ is\ observed, Z\ is\ unobserved$_____(1)

$$\mathrm{KL}(q(z)||p(z|x)) = E_q\left[log\frac{q(Z)}{p(Z|x)}\right], X\ is\ observed, Z\ is\ unobserved$$

$$= E_q[log\ q(Z)] - E_q[logp(Z|x)]$$

$$= E_q[log\ q(Z)] - E_q[logp(Z,x)] + \log p(x)$$

$$= -(-E_q[log\ q(Z)] + E_q[logp(Z,x)]) + \log p(x) = -\mathbf{ELBO} + \text{constant}$$ _(2)

$\log p(x) = log\int p(x,z)dz = log\int p(x,z)\frac{q(z)}{q(z)}dz = \log\left(E_q\left[\frac{p(x,Z)}{q(z)}\right]\right) \geq E_q[logp(Z,x)] - E_q[log\ q(Z)]$ = **ELBO**___(3)

(3) of inequality happens due to Jensen's Inequality in the case of concave function, because log-function is concave

Jensen's Inequality $f(E[X]) \geq E[f(X)], f(\cdot)\ is\ a\ concave\ function$

-Finding the set of latent variables argmax the similarity between the approximated posterior distribution and the exact one is a simply optimization problem.

The above approach is Jensen's Inequality approach used in the computer science and following is statistical approach in this class based on the paper [1] and I modified a little bit algebra in the slide.

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y,\theta)}{\int p(y,\theta)d\theta}, posterior$$

$p(\theta|y)\ is\ a\ posterior$,

$p(y,\theta)\ is\ a\ joint\ density\ for\ y,\theta$ ,

$p(y) is\ a\ marginal\ density\ for\ y$ , $p(y) = \int p(y,\theta)d\theta\ \ is\ a\ hard\ compute$, so that we approximate it

$$\log p(y) = \log p(y) \times 1 = \log p(y) \times \int q(\theta)d\theta \text{ , where } p(y) = \frac{p(y,\theta)}{q(\theta)} \cdot \frac{q(\theta)}{p(\theta|y)}$$

$$= \int q(\theta)\log(\frac{p(y,\theta)}{q(\theta)})d\theta + \int q(\theta)\log\left(\frac{q(\theta)}{p(\theta|y)}\right)d\theta \text{ , where } \int q(\theta)\log\left(\frac{q(\theta)}{p(\theta|y)}\right)d\theta \geq 0$$

$$\geq \int q(\theta)\log\left(\frac{p(y,\theta)}{q(\theta)}\right)d\theta = ELBO \text{ in (3)} \quad \text{(4)}$$

From (4) using exponential transformation and ELBO means the ELBO in (4) from here

$$P(y) \geq \tilde{p}(y;q) = \exp(\text{ELBO}) \quad \text{(5)}$$

$$\underset{q\in Q}{\operatorname{argmax}}\, \tilde{p}(y;q) = \underset{q\in Q}{\operatorname{argmax}}\, ELBO \quad \text{(6)}$$

From (6), there are two restrictions including the mean-field theory [1] as follows

(a) $\mathbf{q}(\boldsymbol{\theta})$ *factorizes into* $\prod_{i=1}^{M} q_i(\theta_i)$, *for some partition* $\{\theta_1, \theta_2, \theta_3 \ldots\ldots, \theta_M\}$ *of* $\theta$

(b) q is a member of a parametric family of density functions

Again, from (4) and the above restriction (a)

$$\log p(y) \geq \log \tilde{p}(y;q) = \int \boldsymbol{q}(\boldsymbol{\theta})\log\left(\frac{p(y,\theta)}{\boldsymbol{q}(\boldsymbol{\theta})}\right)d\theta = ELBO, \text{where } (a) \text{ applied to } \boldsymbol{q}(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\theta_i)$$

$$= \int \prod_{i=1}^{M} q_i(\theta_i), \{\log p(y,\theta) - \sum_{i=1}^{M} \log q_i(\theta_i)\}d\theta_1 \cdots d\theta_M \quad \text{(7)}$$

From (7) dividing the equation into two groups include relating the partition $d\theta_1$ and others as follows

$$= \int q_1(\theta_1)\{\int \prod_{i\neq 1}^{M} q_i(\theta_i) \log p(y,\theta)d\theta_2 \cdots d\theta_M\}d\theta_1 - \int q_1(\theta_1)\log q_1(\theta_1)d\theta_1 + others\ (\theta_{i\neq 1})$$

By logarithm transformation from (4),

$$\tilde{p}(y;\theta_1) \propto \exp[\int \prod_{i\neq 1}^{M} q_i(\theta_i) \log p(y,\theta)d\theta_2 \cdots d\theta_M] \text{ such that } \int \tilde{p}(y;\theta_1)d\theta_1 dy = 1 \quad \text{(8)}$$

The right side of (8) is not depends on $\theta_1$ due to the mean-field theory.

Then, from (7)

$$\log \tilde{p}(y;\theta_1) = \int q(\theta_1) \log\left(\frac{p(y,\theta_1)}{q(\theta_1)}\right) d\theta_1 + others\ (\theta_{i\neq 1}) \quad \text{(9)}$$

By applying the algorithm 1 in the paper [1], applying the optimization, the we can get

$$q_1^*(\theta_1) = \tilde{p}(\theta_1|y = data) \text{ [5]} \quad \text{(10)}$$

By doing this optimization under mean-field assumption, we can make the prior as close as the posterior of interest.

2. State the main motivation and innovation of stochastic variational inference and the key difference between stochastic variational inference and its non-stochastic counterpart.

In Markov chain Monte Carlo (MCMC) sampling, we constructes a Markov chain over the unobserved hidden variables and run this Markov chain until it has reached equilibrium and gather samples to approximate the posterior. [6] However, the case of large latent variables such as shopping data from millions of goods may not plausible to use MCMC due to its computational load. In this big-data situation, VI may be a good candidate to apply. The conventional VI requires batch process so using all dataset before updating the variational parameters, this can still require big computation load and time so that stochastic variational inference (SVI) is developed. By applying stochastic optimization, VI can be more efficient and it called as a stochastic variational inference (SVI). SVI uses noisy estimates of the gradient(G) by subsampling the data and computing a scaled G on the subsample. The expectation of this noisy G is the same with the true G if sampling is independent. [6]

In order to understand the conventional VI in Q1 and SVI in Q2, we will compare its algorithm based on the paper [7] as follows:

**[A] Conventional VI: Coordinate ascent mean-field variational inference.**

$q_j^*(z_j) \propto \exp\{E_{-j}[\log p(z_j|z_{-j}, x)]\}, where\ z\ is\ latent\ variable$ _____(11)

$p(z_j|z_{-j}, x)$ is the complete conditional of $z_j$, using the conditional in (11) is similar with Gibbs sampler which iteratively sampling from each variables' complete conditional.

From (11),

$q_j^*(z_j) \propto \exp\{E_{-j}[\log p(z_j, z_{-j}, x)]\}, where\ z\ is\ latent\ variable$ _____(12)

By using (12), updating prior $q_j(z_j)$ to make it close to the true posterior as follows:

---

Algorithm 1: Coordinate ascent variational inference (CAVI), [7] page 10

---

**Input**: A model p(x,z), a data set x

**Output**: A variational density q(z) = $\prod_{j=1}^{m} q_j(z_j)$

**Initialize**: Variational factor $q_j(z_j)$

**While** the ELBO has not converged do (= until no improvement of ELBO)

**for** j $\in \{1, \dots, m\}$ *do*

Set $q_j(z_j) \propto \exp\{E_{-j}[\log p(z_j|z_{-j}, x)]\}$

**end**

Compute ELBO(q) = E[log p(z,x)] – E[logq(z)]

**end**

**return** q(z)

---

ELBO is a concave function because it consists with log-function, therefore, by iteration of Algorithm 1 will eventually, goes to local optimal. However, as we can see in above algorithm, it require batch process at each updating, because update after while commend.

**[B] SVI with the natural gradient of the ELBO and stochastic optimization of the ELBO**

By applying natural gradient and stochastic optimization, SVI can be scalable and computationally efficiency. The main focus of SVI is optimizing the global variational parameter $\lambda$ of a conditional conjugate model as following steps:

1. Subsampling is conducted in the total data set.
2. Using current global parameters to compute the optimal local parameters for the subsample data set.
3. Adjust the current global parameters.

**The natural gradient**

The natural gradients warp the latent variable space so that moving the same distance in different directions amount to equal change in symmetrized KL divergence.

In this setting, we assume the exponential family and conditional conjugate model.

The natural gradient $g(\lambda) = E_{\varphi}[\hat{\alpha}] - \lambda$ (13)

$\varphi$ is a local variational parameter and $\alpha$ is a natural parameter of prior in the exponential family.

We can use (13) in a gradient based updating of the global parameter at each iteration as follows:

$\lambda_t = \lambda_{t-1} + \epsilon_t g(\lambda_{t-1}) = (1-\epsilon_t)\lambda_{t-1} + \epsilon_t E_{\varphi}[\hat{\alpha}]$ (14)

$\epsilon_t$ is a step size

**Stochastic optimization**

By modifying (13),

$g(\lambda) = \alpha + [\sum_{i=1}^{n} E_{\varphi_t^*}[t(z_t, x_t)], n]' - \lambda, where\ t\ is\ a\ sufficient\ statics$ (15)

$\varphi_t^*$ means optimized local variational parameters with fixed current global parameter $\lambda$

By using (15), applying a noisy gradient by sampling an index from the data and modifying the second term as follows:

$t \sim U(1,\ldots,n)$

$\hat{g}(\lambda) = \alpha + n[\sum_{i=1}^{n} E_{\varphi_t^*}[t(z_t, x_t)], 1]' - \lambda$, where $E(\hat{g}(\lambda)) = g(\lambda)$ (16)

This is an easy and cheap to compute compared to that of VI. Only a single sampled data point and one set of optimized local parameters are used for the gradient as you can see the following Algorithm 2.

Algorithm 2: SVI for conditionally conjugate models [7] page 23

---

**Input**: Model p(x ,z), data x, and set size sequence $\epsilon_t$

**Output**: A variational density $q_\lambda(\beta)$, $\beta$ is a vector of global latent variables

**Initialize**: Variational parameters $\lambda_0$

**While** True **do**

Choose a data point uniformly at random, t~ U(1,….,n)

Optimize its local variational parameters, $\varphi_t^* = E_\lambda[\eta(\beta, x_t)]$

Compute the coordinate update as though $x_t$ was repeated n times,

$\hat{\lambda} = \alpha + nE_{\varphi_t^*}[f(z_t, x_t)]$

Update the global variational parameter, $\lambda_t = (1 - \epsilon_t)\lambda_t + \epsilon_t\hat{\lambda}_t$

**end**

**return** $\lambda$

---

3. State the algorithms for approximate Bayesian computation (ABC) for discrete variables and continuous variables, respectively

The basic idea of ABC from the lecture note [5], when the form of likelihood is hard to know, using approximation approach without guessing the likelihood form with ABC and this method is close related to the likelihood free MCMC [8]. Laplace approximations require additional knowledge of the posterior distribution and VI replace the true model with another psedo-model based on mean-field theory. ABC is a one way to using simulation based inference as follows[3]:

1. Given a belief on the distribution of the parameter $\pi(\theta)$ and observed data Y.
2. Generate the parameter based on the prior of the distribution of the parameter $\theta' \sim \pi(\theta)$
3. Applying $\theta'$ into likelihood function $p(y|\theta')$ and generating y' from $p(y|\theta')$
4. Y in the actual sample and generated y', then applying rejection process if the distance between y and y' is more than tolerance level, not close enough.

[C] Algorithm 3. ABC for the discrete random variable

1. Generate $\theta'$ from the prior of the parameter distribution $\pi(\theta)$
2. Generate the simulated y' from the likelihood f(y'|$\theta'$)
3. Repeat until y'=y, then $\theta_i = \theta'$ , i = number of iterations
4. Using $\theta_i$ to derive the joint distribution g(y, $\theta_i$) as follows:

   $f(\theta_i) = \sum_{z \in Z} \pi(\theta_i) f(y'|\theta_i) 1_{\{y'=y\}} = \pi(\theta_i) f(y'|\theta_i)$ = g(y, $\theta_i$)
5. Using g(y, $\theta_i$) to estimate the posterior $\pi(\theta_i|y)$

The above algorithm 3 is an exact rejection sampling so if not y'=y, it rejects so in the case of continuous y variable, the computation load is high and rejection rate is also high. Therefore, in the case of continuous variable, we need to relax the tolerance level.

[D] Algorithm 4. ABC for the continuous random variable [3]

**For** i=1 to N do

**Repeat**

Generate $\theta'$ from the prior of the parameter distribution $\pi(\theta)$

Generate the simulated y' from the likelihood f(y'|$\theta'$)

**Until** $\rho\{\eta(y'), \eta(y)\} \leq \varepsilon$, where ρ is a distance on η, η a function on y defining a statistics

Set $\theta_i = \theta'$

**End for**

Sampling from the marginal in y' of joint distribution

$\pi_\varepsilon(\theta, y'|y) = \frac{\pi(\theta) f(y'|\theta) I_{A_{\varepsilon,y}(y')}}{\int_{A_{\varepsilon,y} \times \theta} \pi(\theta) f(y'|\theta) dy d\theta}$ , $I_A$ denotes the function of the set B

$A_{\varepsilon,y}$ = {y' ∈ Y| $\rho\{\eta(y'), \eta(y)\} \leq \varepsilon$}

With above setting, the goal of ABC is making a better approximate posterior distribution.

$\pi_\varepsilon(\theta|y) = \int \pi_\varepsilon(\theta, y'|y) dy \approx \pi(\theta|y)$

Reference


[1] Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. The American Statistician, 64(2), 140-153.

[2] Gopalan, P., Hao, W., Blei, D. M., & Storey, J. D. (2016). Scaling probabilistic models of genetic variation to millions of humans. Nature genetics, 48(12), 1587.

[3] Marin, J. M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. Statistics and Computing, 22(6), 1167-1180.

[4] Ruiz, F. J., Athey, S., & Blei, D. M. (2017). Shopper: A probabilistic model of consumer choice with substitutes and complements. *arXiv preprint arXiv:1711.03560*.

[5] WSU Stat 435 Lecture Note (2019)

[6] Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, *14*(1), 1303-1347.

[7] Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877.

[8] https://darrenjw.wordpress.com/2013/03/31/introduction-to-approximate-bayesian-computation-abc/