

# Approximate Inference

Jinhwan Suk

Department of Mathematical Science, KAIST

July 30, 2020

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family
  - Stochastic Variational Inference
  - Application

# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family

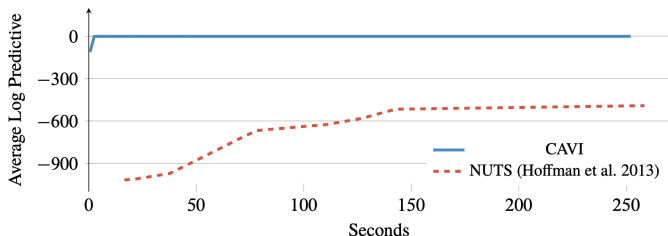
- In general model, it is difficult to compute posterior due to  $p(x)$
- **Variational inference** is widely used to approximate posterior for Bayesian models
- Compared to **MCMC**, variational inference tends to be faster and easier to scale to large data.
- Rather than using sampling, variational inference uses optimization

$$q^*(z) = \operatorname{argmin}_{q \in \mathcal{F}} \mathcal{D}_{KL}(q(z) || p(z|x))$$

- $q^*(z)$  serves as a proxy for  $p(z|x)$

# When MCMC, When Variational Inference?

- MCMC is computationally intensive, but provides guarantees
- Variational inference doesn't provide such guarantees, but suited to large data.
- Multi-modal?



# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family

# Variational Inference

## The evidence lower bound

**Goal** : solve the following optimization problem,

$$q^*(z) = \operatorname{argmin}_{q \in \mathcal{F}} \mathcal{D}_{KL}(q(z) || p(z|x))$$

Recall that

$$\begin{aligned} \mathcal{D}_{KL}(q(z) || p(z|x)) &= \mathbb{E}_{z \sim q} [\log q(z) - \log p(z|x)] \\ &= \mathbb{E}_{z \sim q} [\log q(z) - \log p(x, z)] + \log p(x) \end{aligned}$$

Thus, we cannot compute  $\mathcal{D}_{KL}$ . We optimize an alternative objective that is equivalent to  $\mathcal{D}_{KL}$ ,

$$ELBO(q) = \mathbb{E}_{z \sim q} [\log p(x, z) - \log q(z)]$$

# Variational Inference

## The evidence lower bound

- $ELBO(q)$  lower-bounds the evidence,  $\log p(x) \geq ELBO(q)$

$$\log p(x) = \mathcal{D}_{KL}(q(z)||p(z|x)) + ELBO(q)$$

- $ELBO(q)$  also can be used as a good approximation of  $\log p(x)$ .
- Variational Inference vs. EM algorithm
  - Put  $q(z) = p(z|x;\theta_0)$
  - E step : compute  $Q(\theta; x, \theta_0) = \mathbb{E}_{z \sim p(z|x;\theta_0)} \log p(x, z; \theta)$
  - M step :  $\theta' = \operatorname{argmax}_{\theta} Q(\theta; x, \theta_0)$
  - *Variational EM*



# Variational Inference

## The mean-field variational family

- The complexity of a variational family  $\mathcal{F}$  determines the complexity of the optimization
- A generic member of the **mean-field variational family** is

$$q(z) = \prod_j q_j(z_j)$$

- It cannot capture correlation between them.  
→ Structured variational inference

# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable**
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family

- **Binary sparse coding model :**  $q_i(z_i = 1) = \hat{z}_i$

$$p(z_i = 1) = \sigma(b_i), \quad p(x|z) = \mathcal{N}(Wz; \beta^{-1}I)$$

- Make a mean field approximation

$$q(z) = \prod_{i=1}^m q_i(z_i)$$

- Solve the fixed-point equation,

$$\frac{\partial}{\partial \hat{z}_i} ELBO(\hat{z}_1, \dots, \hat{z}_m) = 0$$

$$\begin{aligned} ELBO &= \mathbb{E}_{z \sim q} [\log p(z) + \log p(x|z) - \log q(z)] \\ &= \mathbb{E}_{z \sim q} \left[ \sum_{i=1}^m \log p(z_i) + \sum_{i=1}^n \log p(x_i|z) - \sum_{i=1}^m \log q_i(z_i) \right] \\ &= \sum_{i=1}^m \mathbb{E}_{z_i \sim q_i} [\log p(z_i) - \log q_i(z_i)] + \mathbb{E}_{z \sim q} \left[ \sum_{i=1}^n \log p(x_i|z) \right] \\ &= \sum_{i=1}^m \left[ \hat{z}_i (\log \sigma(b_i) - \log \hat{z}_i) + (1 - \hat{z}_i) (\log \sigma(-b_i) - \log(1 - \hat{h}_i)) \right] \\ &\quad + \frac{1}{2} \sum_{i=1}^n \left[ \log \frac{\beta_i}{2\pi} - \beta_i \left( x_i^2 - 2x_i W_{i,:} \hat{z} + \sum_j \left[ W_{ij}^2 \hat{z}_j + \sum_{k \neq j} W_{ij} W_{ik} \hat{z}_j \hat{z}_k \right] \right) \right] \end{aligned}$$

$$\begin{aligned} & \frac{\partial}{\partial \hat{z}_i} ELBO \\ &= b_i - \log \hat{z}_i + \log(1 - \hat{z}_i) + x^T \beta W_{:,i} - \frac{1}{2} W_{:,i}^T \beta W_{:,i} - \sum_{j \neq i} W_{:,j}^T \beta W_{:,i} \hat{z}_j \\ &= 0 \end{aligned}$$

We solve for the  $\hat{z}_i$  :

$$\hat{z}_i = \sigma \left( b_i + x^T \beta W_{:,i} - \frac{1}{2} W_{:,i}^T \beta W_{:,i} - \sum_{j \neq i} W_{:,j}^T \beta W_{:,i} \hat{z}_j \right)$$

Repeat the cycle until we satisfy a converge criterion

# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable**
- 5 Variational Inference with Exponential Family

# Continuous Latent Variable

Coordinate ascent mean-field variational inference (Bishop, 2006)

- Same assumption : Mean-field family
- **CAVI**(coordinate ascent variational inference) is one of the most commonly used algorithms for solving the optimization problem.
- Fix the other variational factors  $q_\ell(z_\ell)$ ,  $\ell \neq j$

$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{-j}[\log p(z_j|z_{-j}, x)]\right)$$

# Continuous Latent Variable

Coordinate ascent mean-field variational inference (Bishop, 2006)

---

## Algorithm 1: Coordinate Ascent Variational Inference(CAVI)

---

**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 factors  $q_j(z_j)$ ;

**while** the *ELBO* has not converged **do**

    Set  $q_j(z_j) \propto \exp\left(\mathbb{E}_{-j}[\log p(z_j|z_{-j}, x)]\right)$

    Compute  $ELBO(q) = \mathbb{E}_{z \sim q} [\log p(x, z) - \log q(z)]$

**end**

---

- *Gibbs sampler* maintains a realization of the latent variables
- *CAVI* takes the expected log



# Continuous Latent Variable

Coordinate ascent mean-field variational inference (Bishop, 2006)

**Derivation :**

$$\begin{aligned} ELBO(q; q_{-j}) &= \mathbb{E}_{z \sim q} [\log p(x, z) - \log q(z)] \\ &= \mathbb{E}_{z_j \sim q_j} \mathbb{E}_{z_{-j} \sim q_{-j}} [\log p(x, z) - \log q(z)] \\ &= \mathbb{E}_j [\mathbb{E}_{-j} [\log p(x, z)] - \log q_j(z_j)] - \mathbb{E}_{-j} [\log q_{-j}(z_{-j})] \\ &= -\mathcal{D}_{KL}(q_j \| \frac{q_j^*}{Z}) + Const. \end{aligned}$$

So,  $ELBO_j$  is minimized when  $q_j = \frac{q_j^*}{Z}$  (Calculus of Variation)

# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family
  - Stochastic Variational Inference
  - Application

# Variational Inference with Exponential Family

## Complete conditionals in the exponential family

- Suppose each **complete conditional** is in the exponential family :

$$p(z_j|z_{-j}, x) = h(z_j) \exp(\eta_j^T z_j - \alpha(\eta_j)), \quad \eta_j = \eta_j(z_{-j}, x)$$

(e.g) Gaussian Mixture Model

- CAVI is simplified by

$$\begin{aligned} q(z_j)^* &\propto \exp\left(\mathbb{E}_{-j}[\log p(z_j|z_{-j}, x)]\right) \\ &= \exp\left[\log h(z_j) + \mathbb{E}_{-j}[\eta_j]^T z_j - \mathbb{E}_{-j}\alpha(\eta_j)\right] \\ &\propto h(z_j) \exp(\mathbb{E}_{-j}[\eta_j]^T z_j) \\ &\quad (\nu_j = \mathbb{E}_{-j}[\eta_j]^T) \end{aligned}$$

$\nu_j$  is the variational parameter for local latent variable  $z_j$

# Variational Inference with Exponential Family

## Conditional conjugacy and Bayesian models

- Let  $\beta$  be a vector of *global latent variables*
- Let  $z$  be a vector of *local latent variables*
- Then, the joint density is

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

- Assume that  $p(z_i, x_i | \beta)$  has an exponential family form.

$$p(z_i, x_i | \beta) = h(z_i, x_i) \exp(\beta^T t(z_i, x_i) - a(\beta))$$

- Take the prior

$$p(\beta) = h(\beta) \exp(\alpha^T [\beta, -a(\beta)] - a(\alpha))$$

$$\alpha = [\alpha_1, \alpha_2]^T$$

# Variational Inference with Exponential Family

## Conditional conjugacy and Bayesian models

- With the conjugate prior,

$$p(\beta | z, x) = h(\beta) \exp(\hat{\alpha}^T [\beta, -a(\beta)] - a(\hat{\alpha}))$$

$$\hat{\alpha} = [\alpha_1 + \sum_{i=1}^n t(z_i, x_i), \alpha_2 + n]^T$$

- Assume that

$$p(z_i | x_i, \beta, z_{-i}, x_{-i}) = p(z_i | x_i, \beta)$$

- Then, local variational update is

$$v_i \leftarrow \mathbb{E}_{-i} \eta_i(\beta, z_{-i}, x) = \mathbb{E}_{-i} \eta_i(\beta, x_i)$$

# Variational Inference with Exponential Family

## Variational inference in conditionally conjugate models

- $q(\beta|\lambda)$  : variational posterior approximation on  $\beta$
- $q(z_i|\phi_i)$  : variational posterior approximation on  $z_i$
- Then, local variational update is

$$v_i \leftarrow \mathbb{E}_{\lambda, \phi_{-i}} \eta_i(\beta, z_{-i}, x) = \mathbb{E}_{\lambda} \eta_i(\beta, x_i)$$

- global variational update is

$$\lambda \leftarrow [\alpha_1 + \sum_{i=1}^n \mathbb{E}_{\phi_i} t(z_i, x_i), \alpha_2 + n]^T$$

- When updating global variational parameter, the algorithm requires iterating through the **entire data set**.

# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family
  - Stochastic Variational Inference
  - Application

# Variational Inference with Exponential Family

## Stochastic Variational Inference

- Most posterior inference algorithms do not easily scale(MCMC)
- CAVI is no exception

$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{-j}[\log p(z_j|z_{-j}, \mathbf{x})]\right)$$

- An alternative to CAVI is gradient-based optimization.

$$\nabla_{\lambda} ELBO = a''(\lambda)(\mathbb{E}_{\phi}[\hat{\alpha}] - \lambda), \quad g(\lambda) = \mathbb{E}_{\phi}[\hat{\alpha}] - \lambda$$

(e.g.) Neural Net



# Variational Inference with Exponential Family

## Stochastic Variational Inference

- global variational update is

$$\lambda_t = \lambda_{t-1} + \varepsilon_t g(\lambda_{t-1}) = (1 - \varepsilon_t) \lambda_{t-1} + \varepsilon_t \mathbb{E}_\phi[\hat{\alpha}]$$

- The noisy unbiased estimator for  $g(\lambda)$  is

$$j \sim \text{Unif}(1, 2, \dots, n)$$

$$\hat{g}(\lambda) = \alpha + n \left[ \mathbb{E}_{\phi_j} t(z_j, x_j), 1 \right]^T - \lambda$$

# Contents

- 1 Motivation
- 2 Variational Inference
- 3 Discrete Latent Variable
- 4 Continuous Latent Variable
- 5 Variational Inference with Exponential Family
  - Stochastic Variational Inference
  - Application

# Variational Inference with Exponential Family

## Application : Probabilistic Topic Models

### Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

### Documents

#### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions**

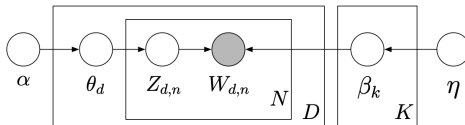
\* Genome Mapping and Sequencing. Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a geneticist at Uppsala University in Sweden. "We arrived at the 800 number. But coming up with a consensus answer may be more than just a **scientific** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

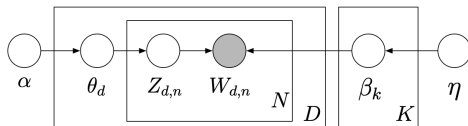
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

### Topic proportions and assignments



# Variational Inference with Exponential Family

## Application : Probabilistic Topic Models



- 1 For each topic in  $k = 1, 2, \dots, K$ ,
  - 1 draw a distribution over words  $\beta_k \sim \text{Dir}_V(\eta)$
- 2 For each document in  $d = 1, 2, \dots, D$ ,
  - 1 draw a vector of topic proportions  $\theta_d \sim \text{Dir}_K(\alpha)$
  - 2 For each word in  $n = 1, 2, \dots, N_d$ 
    - 1 draw a topic assignment  $z_{dn} \sim \text{Mult}(\theta_d)$
    - 2 draw a word  $w_{dn} \sim \text{Mult}(\beta_{z_{dn}})$

# Variational Inference with Exponential Family

Application : Probabilistic Topic Models

- Posit a mean-field variational family

$$q(\beta, \theta, z) = \prod_{k=1}^K q(\beta_k; \lambda_k) \prod_{d=1}^D \left( q(\theta_d; \gamma_d) \prod_{n=1}^{N_d} q(z_{dn}; \phi_{dn}) \right)$$

- $p(\theta_d | z_d) = \text{Dir}_K(\alpha + \sum_{n=1}^{N_d} z_{dn}) \leftarrow q(\theta_d; \gamma_d)$
- $p(\beta_k | z, w) = \text{Dir}_V(\eta + \sum_{d,n} z_{dn}^k w_{dn}) \leftarrow q(\beta_k; \lambda_k)$

# Variational Inference with Exponential Family

## Application : Probabilistic Topic Models

- CAVI update rule :

$$\phi_{dn}^k \propto \exp \left( \Psi(\gamma_{dk}) + \Gamma(\lambda_{k,w_{dn}}) - \Gamma(\sum_v \lambda_{kv}) \right)$$

$$\gamma_d = \alpha + \sum_{n=1}^{N_d} \phi_{dn}$$

$$\lambda_k = \eta + \sum_{d,n} \phi_{dn}^k w_{dn}$$

- (Hoffman, 2013) Stochastic variational inference

# Conclusion

## Theory

- (You et al. 2014) Variational posterior for Bayesian linear model.
- (Hall et al. 2011) Poisson mixed-effects model
- (Westling and McCormick, 2015)  
Consistency of VI through a connection to M-estimation
- (Wang and Titterton, 2006) VI for mixtures of Gaussians
  - CAVI converges to a local optimum
  - VI estimate and MLE approach each other at a rate of  $\mathcal{O}(1/n)$

- $\mathcal{D}_{KL}(q(z)||p(z|x))$ 
  - (Minka, 2001)  $\mathcal{D}_{KL}(p(z|x)||q(z))$
  - (Barber and de van Laar, 1999) Tighter lower bounds than ELBO
- Mean-field approximation
  - help with scalable optimization, but limit the expressibility
  - Structured variational inference
- Interface between MCMC and variational inference
  - (Freitas et al., 2001) proposal distribution
  - (Salimans et al., 2015) variational approximation + MCMC chain
- **(Reference) Variational Inference : A Review for Statisticians**