Stochastic variational inference
Structured stochastic variational inference

CIS 620 paper discussion

Simeng Sun

Oct. 9th, 2018

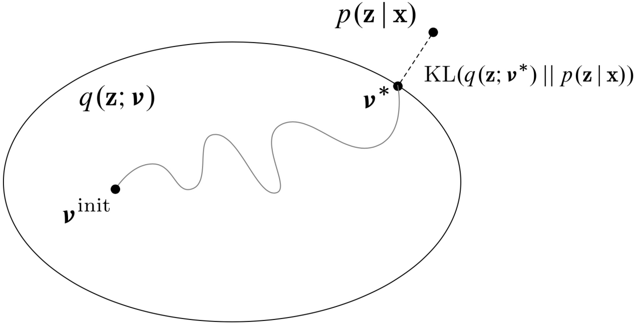
Outline

- A quick review of mean-field variational inference
- Traditional coordinate ascent algorithm
- Stochastic variational inference(SVI)
- Structured SVI

Outline

- A quick review of mean-field variational inference
- Traditional coordinate ascent algorithm
- Stochastic variational inference(SVI)
- Structured SVI

Variational Inference



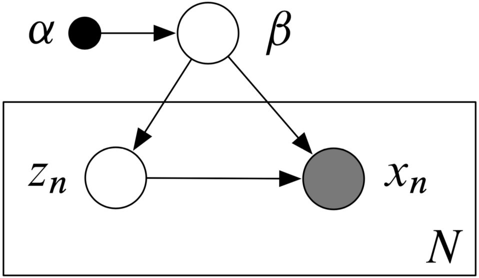
$p(\mathbf{z}|\mathbf{x})$ true posterior distribution

$q(\mathbf{z}; \mathbf{v})$ tractable variational posterior distribution

Minimize the Kullback–Leibler divergence

[1] pic from https://media.nips.cc/Conferences/2016/Slides/6199-Slides.pdf

Variational Inference - setup

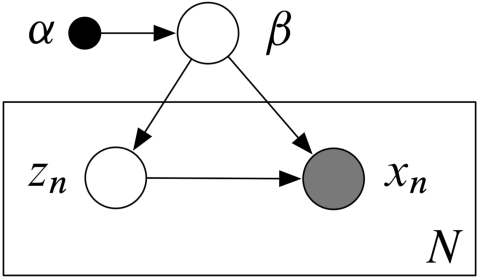


Variables

α: fixed parameters
β: global hidden variables
x_n: n^{th} observation
z_n: context of n^{th} observation ($z_{n,1:J}$)
$z_{n,1:J}$: set of J local hidden variables

[1] pic from http://www.columbia.edu/ jwp2128/Papers/HoffmanBleiWangPaisley2013.pdf

Variational Inference - setup



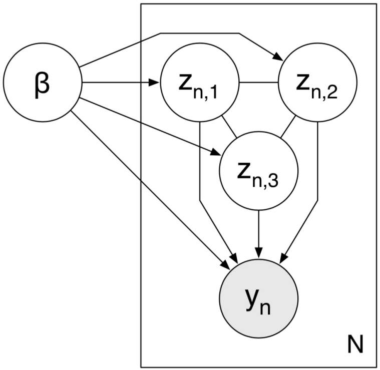
Assumptions

- Independence of hidden variables (**Mean-Field**)

[1] pic from http://www.columbia.edu/ jwp2128/Papers/HoffmanBleiWangPaisley2013.pdf

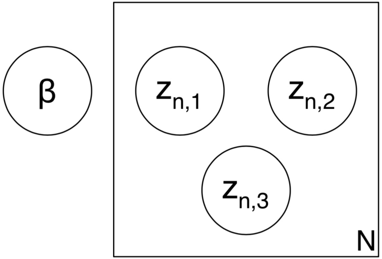
Mean-Field Variational Inference

Model dependencies



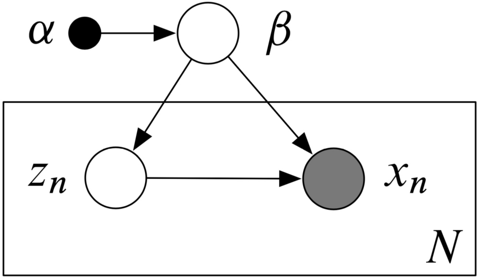
(Naive) Mean-Field dependencies

Intractable posterior distribution becomes a distribution where all variables are independent



[1] pic from http://proceedings.mlr.press/v38/hoffman15.pdf

Variational Inference - setup



Assumptions

- Independence of hidden variables (**Mean-Field**)
- **Complete conditionals** are from exponential families
  - Conditional distribution of a hidden variable given the other hidden variables and observations.
  - $p(\beta \mid x, z)$
  - $p(z_{nj} \mid x_n, z_{n,\setminus j}, \beta)$

[1] pic from http://www.columbia.edu/ jwp2128/Papers/HoffmanBleiWangPaisley2013.pdf

Variational distribution q

- Variational distribution q under Mean-Field assumption

$$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj})$$

  - λ: global variational parameters, govern global variables β
  - ϕ_{nj}: local variational parameters, govern j^{th} local variable in the context of n^{th} observation z_{nj}
- Assumption on the form of complete conditionals
  $+$ conjugacy property of exponential family
  $\Rightarrow$ $q(\beta \mid \lambda)$ and $q(z_{nj} \mid \phi_{nj})$ are also from exponential families.

Evidence lower bound (ELBO)

- Variational distributions

$$q(\beta \mid \lambda) = \exp\{\lambda^\top t(\beta) - A_g(\lambda)\}$$

$$q(z_{nj} \mid \phi_{nj}) = \exp\{\phi_{nj}^\top t(z_{nj}) - A_l(\phi_{nj})\}$$

$t(\cdot)$ indicates sufficient statistics
$A_{g/l}(\cdot)$ indicates global/local cumulant function

- **Evidence lowerbound**

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

- Tradeoff between making q as spread out as possible and making q concentrate on one point that maximizes the expected log joint

Outline

- A quick review of mean-field variational inference
- Traditional coordinate ascent algorithm
- Stochastic variational inference(SVI)
- Structured SVI

Coordinate ascent (batch)

- Update of global variational parameters λ

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla^2_\lambda A_g(\lambda)(\mathbb{E}_q[\eta_g(x,z)] - \lambda) = 0$$

$$\Rightarrow \lambda = \mathbb{E}_q[\eta_g(x,z)]$$

- complete conditional of β

$$p(\beta \mid x, z) = \exp\{\eta_g(x,z)^\top t(\beta) - A_g(\eta_g(x,z))\}$$

- $\eta_g(x,z)$ is the canonical parameter of β's complete conditional distribution
- λ is set to the mean parameter of β's complete conditional distribution
- Similarly, for local variational parameter ϕ_{nj}

$$\phi_{nj} = \mathbb{E}_q[\eta_l(x_n, z_{n,\setminus j}, \beta)]$$

Coordinate ascent algorithm for VI

Coordinate ascent mean-field variational inference

- Initialize $\lambda^{(0)}$ randomly
- Repeat until converges
  - for each local parameter ϕ_{nj}
    - set $\phi_{nj}^{(t)}$ to $\mathbb{E}_{q^{(t-1)}}[\eta_l(x_n, z_{n,\setminus j,\beta})]$ (E-step)
  - set global parameter $\lambda^{(t)}$ to $\mathbb{E}_{q^{(t)}}[\eta_g(x, z)]$ (M-step)

Problem of coordinate ascent

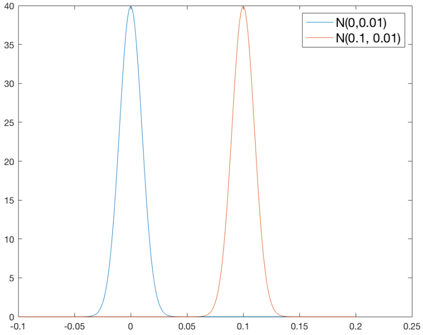
- Global parameters only get updated after updating every local parameters
- Wasteful if we can learn something about global parameters from a subset of data.

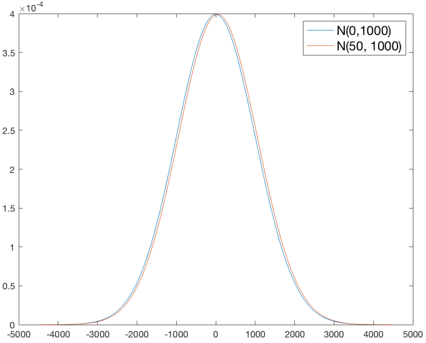
Outline

- A quick review of mean-field variational inference
- Traditional coordinate ascent algorithm
- Stochastic variational inference(SVI)
  - Natural gradient
  - SVI algorithm
  - compare with coordinate ascent
- Structured SVI

Natural gradient

- Stochastic gradient ascent/descent
  - Gradient computed in Euclidean space
- Problem with Euclidean space when minimizing KL divergence





Natural gradient cont.

- **Euclidean gradient**
  The direction of steepest ascent in *Euclidean* space
- **Natural gradient**
  The direction of steepest ascent in *Riemannian* space
  The space where distance is defined by KL divergence rather than $L2$ norm.
- Transform Euclidean gradient to natural gradient by left-multiplying the inverse of Riemannian metric $G(\cdot)^{-1}$

Natural gradient cont.

- Natural gradient of global variational parameters (similarly for local variational parameters)

$$\hat{\nabla}_\lambda \mathcal{L}(\lambda) = G(\lambda)^{-1} \nabla_\lambda \mathcal{L}(\lambda)$$

  $G(\lambda)$ is the second derivative of the log partition of $q(\beta \mid \lambda)$

$$G(\lambda) = \nabla_\lambda^2 A_g(\lambda)$$

- In the coordinate ascent section, we have derived the gradient of ELBO w.r.t. λ

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda^2 A_g(\lambda)(\mathbb{E}_q[\eta_g(x, z)] - \lambda)$$

- The natural gradient of λ is thus

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

Stochastic variational inference

- Rewrite ELBO to global term and sum of local terms

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)] \\ &= \mathbb{E}_q[\log \frac{p(\beta)}{q(\beta)}] + \sum_{n=1}^{N} \mathbb{E}_q[\log \frac{p(z_n, x_n \mid \beta)}{q(z_n)}]\end{aligned}$$

- Sample one x_i uniformly from the data set and duplicate N times as the sum of local terms

$$\mathcal{L}_I(q) = \mathbb{E}_q[\log \frac{p(\beta)}{q(\beta)}] + N\mathbb{E}_q[\log \frac{p(z_i, x_i \mid \beta)}{q(z_i)}]$$

- Natural gradient of $\mathcal{L}_I$ w.r.t λ is noisy but unbiased, because

$$\mathbb{E}[\mathcal{L}_I(\lambda)] = \mathcal{L}(\lambda)$$

Stochastic variational inference

- Let $\{x_i^{(N)}, z_i^{(N)}\}$ be a set of N replicates of x_i and z_i
- Noisy natural gradient of ELBO

$$\hat{\nabla}_\lambda \mathcal{L}_I(\lambda) = \mathbb{E}_q[\eta_g(x_i^{(N)}, z_i^{(N)})] - \lambda$$

- Define intermediate global parameter $\hat{\lambda}_i$

$$\hat{\nabla}_\lambda \mathcal{L}_I(\lambda) = 0$$

$$\hat{\lambda}_i = \mathbb{E}_q[\eta_g(x_i^{(N)}, z_i^{(N)})] = N\mathbb{E}_q[\eta_g(x_i, z_i)]$$

A noisy estimate of λ using only one point, easy to compute

Stochastic variational inference

- $\lambda^{(t-1)}$: estimate of global parameters in previous step
  $\hat{\lambda}_t$: intermediate global estimate of current step
- Let ρ_t be the step size at time t

$$\begin{aligned}\lambda^{(t)} &= \lambda^{(t-1)} + \rho_t(\hat{\lambda}_t - \lambda^{(t-1)}) \\ &= (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}_t\end{aligned}$$

  - since

$$\mathbb{E}[\hat{\lambda}_t - \lambda^{(t-1)}] = \mathbb{E}[\hat{\nabla}_\lambda \mathcal{L}_I(\lambda)] = \nabla_\lambda \mathcal{L}(\lambda)$$

- Global parameter $\lambda^{(t)}$ is the weighted average between previous $\lambda^{(t-1)}$ and estimate of λ if the sampled data point was replicated N times.

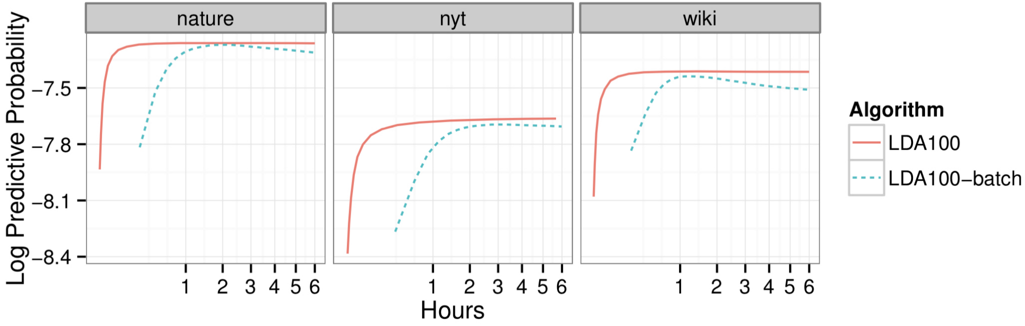
Stochastic variational Inference

Steps:

Repeat until converges

1. Sample data point x_i uniformly from the data set
2. Update local variational parameters ϕ_{ij} to $\mathbb{E}_{q^{(t-1)}}[\eta_l(x_i, \beta, z_{i,\setminus j})]$
3. Compute intermediate estimate of global variational parameter by duplicating N times of a point $\hat{\lambda}_t = \mathbb{E}_q[\eta_g(x_i^{(N)}, z_i^{(N)})]$
4. Update global parameter using weighted average $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}_t$

Stochastic variational Inference - LDA experiment



- Per-word predictive log likelihood for 100-topic LDA model on 3 large corpora.
- Split each document to observed words (obs) and held out words (new); compute $p(w_{new} \mid w_{obs}, \mathcal{D})$
- SVI converges faster and to a better place

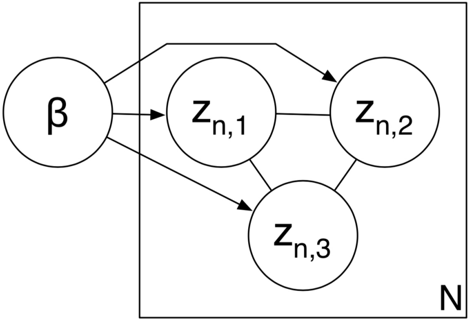
Outline

- A quick review of mean-field variational inference
- Traditional coordinate ascent algorithm
- Stochastic variational inference(SVI)
- Structured SVI

Structured SVI

- Mean-Field assumption of SVI
  Hidden variables are all independent of each other
  Although easier to compute, it becomes less likely to approximate the correct $p(z, \beta \mid x)$
- **Structured stochastic variational inference**
  Allow arbitrary dependencies between global and local variables

Structured SVI



Hidden structure

$$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} q(z_n \mid \gamma_n(\beta))$$

$\gamma_n(\beta)$ is a vector-valued function, represents any possible dependencies between z_n and β

Structured SVI

- Require $p(\beta)$ and $p(x_n, z_n \mid \beta)$ are from exponential families and having the form

$$p(\beta) = \exp\{\eta^\top t(\beta) - A_g(\eta)\}$$

$$p(x_n, z_n \mid \beta) = \exp\{\eta_n(x_n, z_n)^\top t(\beta) + g_n(x_n, z_n)\}$$

  $\eta_n(x_n, z_n)$ is a vector-valued function
- Weakened assumption: do not require exponential form or tractability of the complete conditionals of local hidden variables

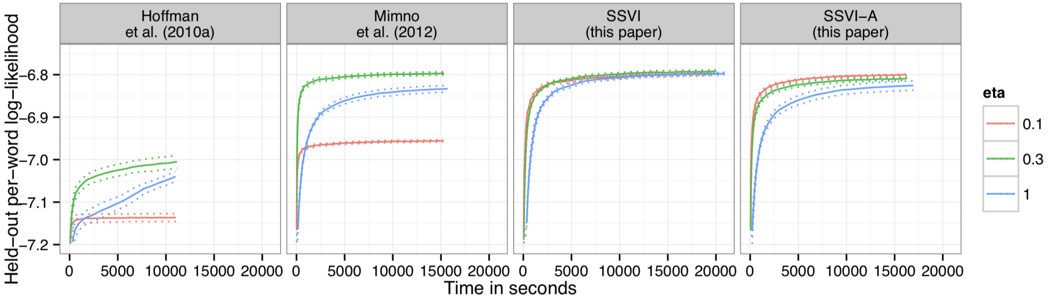
Structured SVI

SSVI steps:

Repeat until converges

1. Sample global variable $\beta^{(t)}$ from $q(\beta \mid \lambda^{(t-1)})$
2. Compute local variational parameters $\gamma_n(\beta^{(t)})$ which maximizes local ELBO $\mathbb{E}_q[\log p(x_n, z_n \mid \beta) - \log q(z_n \mid \beta)]$
   - *analogous to updating local variational parameter ϕ_{nj} in SVI*
3. Update $\hat{\eta}_n$ to be $\mathbb{E}[\eta_n(x_n, z)] = \sum_{z_n} q(z_n \mid \gamma_n(\beta^{(t)}))\eta_n(x_n, z_n)$
   - contribution of n^{th} local context to the update of global parameters
   - *analogous to computing noisy estimate of global parameter in SVI*
4. Update $\lambda^{(t)}$ to be the weighted average of previous $\lambda^{(t-1)}$ and noisy estimate
   - *Standard Robbins-Monro algorithm*

Structured SVI



- Wiki copora, per-word log likelihood
- SSVI-A: same procedure, simplfied version of noisy estimate
- SSVI converge to better place than SVI with mean field assumption
- SSVI is less sensitive to the chosen hyper-parameters

QA

Questions?