Stochastic Variational Inference

From Session-6-a, in multinomial mixture model:-

$$q(z_n) = \text{Multinomial}(z_n | \nu z_n)$$

$$q(\theta) = \text{Dirichlet}(\theta | \nu\theta)$$

$$q(\beta_k) = \text{Dirichlet}(\beta_k | \nu\beta_k) \qquad k = 1, \ldots, K$$

ELBO

(A) $$\nu z_n \leftarrow \underset{\nu z_n}{\text{argmax}} \; E_q \ln P(x, z) - E_q \ln q(z)$$

$$\nu z_{n,k} \propto \exp\left(E_{q(\theta)} \ln \theta_k + \sum_{w=1}^{W} m_{n,w} E_{q(\beta_k)} \ln \beta_k \right)$$

(B) $$\nu\theta \leftarrow \underset{\nu\theta}{\text{argmax}} \; E_q \ln P(x, z) - E_q \ln q(z)$$

$$\nu\theta_k = \alpha - 1 + \sum_{n=1}^{N} \nu z_{n,k}$$

(C) $$\nu\beta_k \leftarrow \underset{\nu\beta_k}{\text{argmax}} \; E_q \ln P(x, z) - E_q \ln q(z)$$

$$\nu\beta_{k,w} = \gamma - 1 + \sum_{n=1}^{N} m_{n,w} \nu z_{n,k}$$

* (A), (B) and (C) are solved by

$\nabla_{\nu z_n} ELBO = 0$, $\nabla_{\nu\theta} ELBO = 0$, $\nabla_{\nu\beta_k} ELBO = 0$

respectively !

* This is the Batch variational inference
* To update $\nu\theta$ and $\nu\beta_k$, all the observations have to be analyzed (notice resemblance to Batch EM and (non-collapsed) Gibbs Sampler)
* Update to <u>global</u> parameters is slow !
* Can <u>local</u> parameter (νz_n) update be propagated faster to <u>global</u> parameters ($\nu\theta$, $\nu\beta_k$, $k = 1, \ldots, K$)
* STOCHASTIC OPTIMIZATION !

ALGORITHM :-

Given data set of size N and ρ_t for $t = 1, \ldots, \infty$

$\sum \rho_t = \infty$, $\sum \rho_t^2 < \infty$, $\lim_{t \to \infty} \rho_t = 0$

Initialize global parameters "$q^{(0)}$" randomly.

for $t: 0, \ldots, \infty$

1) Sample an observation

2) Solve the local parameter

3) Construct global parameters $\hat{g}$ from local parameter of step (2), pretending observation in step (1) was observed 'N' times.

4) Update global parameters

$$g^{(t+1)} = g^{(t)} + \rho_t (\hat{g} - g^{(t)})$$

natural gradient

learning rate

$$\left\{ g^{(t+1)} = (1-\rho_t) g^{(t)} + \rho_t \hat{g} \right\}$$ Same as above

running average

SVI for Multinomial Mixture Model :-

Initialize $v\beta_k^{(0)}$, $k = 1, \ldots, K$ and $v\theta^{(0)}$ randomly

for $t = 0, \ldots \infty$

1) $x_n \leftarrow$ uniformly sample a document

2) $\nu z_{n,k} \propto \exp\left(\mathbb{E}_{q(\theta|\nu\theta^{(t)})} \ln \theta_k + \sum_{w=1}^{W} m_{n,w} \mathbb{E}_{q(\beta_k|\nu\beta_k^{(t)})} \ln \beta_k \right)$

3) $\hat{\nu\theta}_k = \alpha - 1 + N \nu z_{n,k}$

$\hat{\nu\beta}_{k,w} = \gamma - 1 + N m_{n,w} \nu z_{n,k}$, $\forall\ k, w$

4) $\nu\theta^{(t+1)} = (1 - \rho_t)\nu\theta^{(t)} + \rho_t \hat{\nu\theta}$

$\nu\beta_k^{(t+1)} = (1 - \rho_t)\nu\beta_k^{(t)} + \rho_t \hat{\nu\beta}_k$, $k = 1, \ldots, K$

Question 1:- Why does data-subsampling work?

Question 2:- What is natural gradient?

Answer 1 :-

Consider :- $z \sim p(z|\theta)$

$$\mathbb{E}_{p(z|\theta)}[z] = \int z\, p(z|\theta)\, dz$$

$$= f(\theta)$$

find θ^* such that $f(\theta^*) = 0$

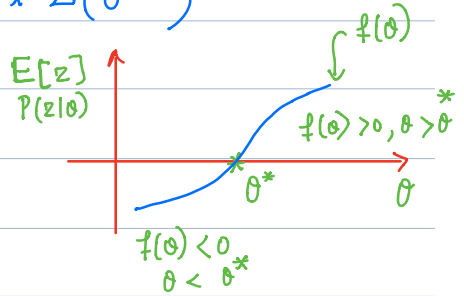
i.e $\mathbb{E}_{p(z|\theta^*)}[z] = \int z\, p(z|\theta^*)\, dz = 0$

then iterate in this fashion :-

$$\theta^{(t)} = \theta^{(t-1)} - \rho_t * z(\theta^{t-1})$$

where $z(\theta)^{t-1} \sim p(z|\theta^{(t-1)})$

and $\sum \rho_t = \infty$, $\sum \rho_t^2 < \infty$, $\lim_{t \to \infty} \rho_t = 0$



* This root finding procedure (1951, ROBBINS-MONRO) forms the basis of stochastic optimization

Specifically, lets say

$$\nabla_\theta g(\theta) = f(\theta) = \underset{P(z|\theta)}{E}[z] = \int z p(z|\theta) \, d\theta$$

* then 'z' has the interpretation of samples of gradient and the above iterative procedure gives the optima of $g(\theta)$

* In many optimization procedures, $\nabla_\theta g(\theta)$ appears as a sum over data-points.

i.e $$\nabla_\theta g(\theta) = \sum_n \underbrace{[\nabla_\theta g(\theta)]_n}_{\text{contribution of nth-datapoint to gradient}}$$

Under uniform sampling of data point (Step-1 of SVI), we have

$$\underset{i \sim \text{uniform}(1,N)}{\mathbb{E}} [\nabla_\theta g(\theta)]_i = \sum_{n=1}^{N} \left(\frac{1}{N}\right) [\nabla_\theta g(\theta)]_n$$

$$= \frac{1}{N} \sum_{n=1}^{N} [\nabla_\theta g(\theta)]_n$$

$$= \frac{1}{N} \nabla_\theta g(\theta)$$

likelihood thickening factor

Hence $N[\nabla_\theta g(\theta)]_i$ where $i \sim \text{uniform}(1, N)$ is an unbiased estimate of gradient.

$\Longrightarrow$ We can follow $N[\nabla_\theta g(\theta)]_i$ instead of batch gradient

Answer 2 :- (What is natural gradient?)

$$F(\lambda)^{-1} \nabla_\lambda \text{ELBO}$$

$F(\lambda)^{-1}$: Fisher information Matrix.

∇_λ ELBO: gradient of ELBO w.r.t variational parameter λ

Two Views:-
1) Optimization & 2) Information theoretic.

View 1) Optimization :-

Let $g(\theta)$ be some objective function,

Goal :- optimize $g(\theta)$ using iterative procedure

* Instead of following direction of regular gradient $\nabla_\theta g(\theta)$, follow instead :-

$$-\left[H(\theta)\right]^{-1} \nabla_\theta g(\theta)$$

where $H(\theta)$ is the "Hessian" of objective function

defined as $H(\theta)[i,j]$ = i,j entry of Hessian matrix

$$= \frac{\partial^2}{\partial\theta_i \partial\theta_j} g(\theta)$$

(Well-known Newton method)

* Fisher-information matrix plays the same role in optimizing parameters of some probability distribution (e.g $v\theta$)

i-e $F(\lambda)[i,j] = -E_{q(\beta|\lambda)} \frac{\partial^2}{\partial\lambda_i \partial\lambda_j} \ln q(\beta|\lambda)$

for exponential families. → true $= E_{q(\beta|\lambda)} \frac{\partial}{\partial\lambda_i} \ln q(\beta|\lambda) \cdot \frac{\partial}{\partial\lambda_j} \ln q(\beta|\lambda)$

Hence use $F(\lambda)^{-1} \nabla_\lambda \text{ELBO}$

where λ is some variational parameter.

MOST IMPORTANT :-

$\nabla_\lambda \text{ELBO} = F(\lambda)(\lambda - E_q(\text{complete conditional natural parameters}))$

Hence $F(\lambda)^{-1}$ is not needed to be compuled

PROOF :-

X :- data, and Z, θ :- two latent variables.

$$\left.\begin{aligned} Pr(z \mid h_z) &= a(z)\, b(h_z) \exp(h_z^T t(z)) \\ Pr(\theta \mid h_\theta) &= a(\theta)\, b(h_\theta) \exp(h_\theta^T t(\theta)) \end{aligned}\right\} \text{Priors}$$

$$Pr(\theta \mid h_\theta, x, z) = a(\theta)\, b(\eta(h_\theta, x, z)) \exp(\eta(h_\theta, x, z)^T t(\theta)) \quad \text{— (A)}$$

$$Pr(z \mid h_z, x, \theta) = a(z)\, b(\eta(h_z, x, \theta)) \exp(\eta(h_z, x, \theta)^T t(z)) \quad \text{— (B)}$$

(A) and (B) are Pr(latent variable | everything else besides latent-variable)

Let $q(z|\nu_z)$ and $q(\theta|\nu_\theta)$ be variational distributions for z and θ in MEAN-FIELD Approx.

$$q(z|\nu_z) = a(z)\, b(\nu_z) \exp(\nu_z^T t(z))$$

$$q(\theta|\nu_\theta) = a(\theta)\, b(\nu_\theta) \exp(\nu_\theta^T t(\theta))$$

To find ν_z and ν_θ, we maximize ELBO

$$L(\nu_z, \nu_\theta) = E_q \ln P(x, z, \theta | h_z, h_\theta) - E_q \ln q(z, \theta)$$

$$= E_q \ln P(x, z | h_z, h_\theta) + E_q \ln P(\theta | x, z, h_\theta)$$

$$- E_q \ln q(z|\nu_z) - E_q \ln q(\theta|\nu_\theta)$$

optimizing w.r.t ν_θ, let's write ELBO in terms of "ν_θ" dependent terms.

$$L(\nu_\theta) = E_q \ln P(\theta | x, z, h_\theta) - E_q \ln q(\theta|\nu_\theta)$$

using (A)

$$= E_q\Big[\ln a(\theta) + \ln b(\eta(x,z,h_\theta)) + \eta^T(x,z,h_\theta)\, t(\theta) - \ln a(\theta) - \ln b(v_\theta) - v_\theta^T t(\theta)\Big]$$

ignore as no v_θ involved

$$= E_q\big[\ln b(\eta(x,z,h_\theta))\big] + E_{q(z)}\big[\eta(x,z,h_\theta)\big]^T E_{q(\theta)}\, t(\theta) - \ln b(v_\theta) - (v_\theta)^T E_q(t(\theta))$$

$$L(v_\theta) = -E_{q(z)}\big[\eta(x,z,h_\theta)\big]^T \nabla_{v_\theta} \ln b(v_\theta) + (v_\theta)^T \nabla_{v_\theta} \ln b(v_\theta) - \ln b(v_\theta)$$

$$\nabla_{v_\theta} L(v_\theta) = -\big[\nabla^2_{v_\theta} \ln b(v_\theta)\big] E_{q(z)}\big[\eta(x,z,h_\theta)\big] + \big[\nabla^2_{v_\theta} \ln b(v_\theta)\big] v_\theta + \nabla_{v_\theta} \ln b(v_\theta) - \nabla_{v_\theta} \ln b(v_\theta)$$

$$\nabla_{v_\theta} L(v_\theta) = -\big[\nabla^2_{v_\theta} \ln b(v_\theta)\big]\big(E_{q(z)}\big[\eta(x,z,h_\theta)\big] - v_\theta\big)$$

$\nabla^2_{v_\theta} \ln b(v_\theta)$: for exponential family, this is $F(v_\theta)$

2nd View :- information theoretic

steepest descent direction $\leftarrow \underset{d\theta}{\operatorname{argmax}}\ f(\theta + d\theta)$

such that $\|d\theta\|_2 < \epsilon$

steepest descent direction in euclidean. $\leftarrow \frac{f'(\theta)}{\|f'(\theta)\|_2}\epsilon$ = direction of gradient at θ

natural gradient $\leftarrow \underset{d\theta}{\operatorname{argmax}}\ f(\theta + d\theta)$

such that $KL(p(\cdot|\theta) \,||\, p(\cdot|\theta + d\theta)) < \epsilon$