STOCHASTIC VARIATIONAL INFERENCE

1. Review variational inference (mean-field)
2. natural gradient
3. stochastic variational inference

1. VI. Evidence lower bound.

$$\log p(x) \geq E_q[\log p(x,\theta)] - E_q[\log q(\theta)]$$

mean-field assumption. $\log p(x) \geq E_q[\log p(x,\theta)] - \sum_i E_{q_i}[\log q_i(\theta_i)]$

(co-ordinat ascent. fixed point update $q_i^*(\theta_i) \propto \exp\{E_{q_{\neg i}}[\log p(\theta_i \mid \theta_{\neg i}, x)]\}$ $H(q_i)$

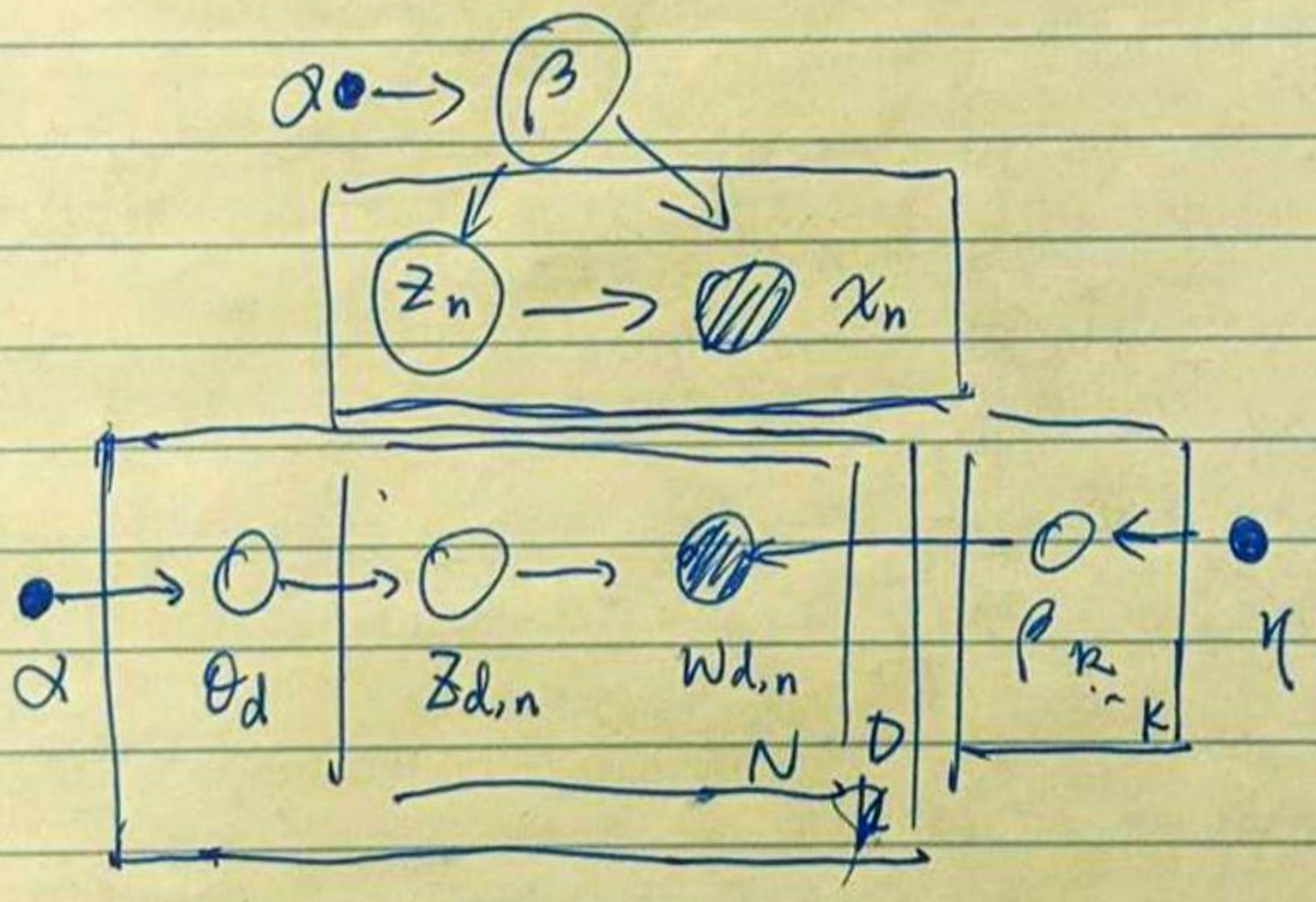
complete conditions

θ — global hidden variables β.

local hidden variables z

fixed param α.

$$p(x,z,\beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta)$$



$$p(\beta \mid x,z,a) = h(\beta)\exp\{\eta_g(x,z,\alpha)^T t(\beta) - a_g(\eta_g(x,z,\alpha)\}$$

$$p(z_{nj} \mid x_n, z_{n,\neg j}, \beta) = h(z_{nj})\exp\{\eta_\ell(x_n, z_{n,\neg j}, \beta)^T t(z_{nj}) - a_\ell(\eta_\ell)\}$$

$$p(x_n, z_n \mid \beta) = \langle \beta^T, a_\ell(\beta)\rangle \qquad p(\beta) = \langle \alpha^T, a_g(\alpha)\rangle$$

$$\eta_g(x,z,\alpha) = (\alpha_1 + \sum_{n=1}^{N} t(z_n, x_n), \alpha_2 + N)$$

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

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

$$q(\beta|\lambda) = h(\beta) \exp\{\lambda^T t(\beta) - a_g(\lambda)\}$$
$$q(z_{nj}|\phi_{nj}) = h(z_{nj}) \exp\{\phi_{nj}^T t(z_{nj}) - a_l(\phi_{nj})\}$$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log(p(x,z) \cdot p(\beta|x,z))] - \left(\mathbb{E}_\lambda[\log q(\beta)] + \sum_{n=1}^{N}\sum_{j=1}^{J} \mathbb{E}_{\phi_{nj}}[\log q(z_{nj})]\right)$$

~~$= \mathbb{E}_q \log p(x,z)$~~ const

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

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

$$\nabla \mathcal{L} = 0 \longrightarrow \lambda^* = \mathbb{E}_q[\eta_g(x,z,\alpha)]$$

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

Gradient find a maximum $f(\lambda)$

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

↑ learning rate $\mathcal{N}(0, 10000)$ 10

$\mathcal{N}(0,0.1)$, $\mathcal{N}(10, 10000)$

$\mathcal{N}(1,0.1)$

symmetrized KL divergence

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

~~$= KL$~~

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

↑ Riemannian metric

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

$G(\lambda) \triangleq$ fisher information matrix of $q(\lambda)$
$=$ second drivative $= \log$ normalizer $= \nabla^2_\lambda a_g(\lambda)$

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

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

intermediate global params $\hat{\lambda} = \mathbb{E}_q[\eta_g(x_i^{(N)}, z_i^{(N)}, \alpha)]$.

$\lambda^{(t)} = \lambda^{(t-1)} + \rho_t \hat{\nabla}_\lambda \mathcal{L}_i$

$= \lambda^{(t-1)} + \rho_t \left(\mathbb{E}_q[\eta_g(\# x_i^{(N)}, z_i^{(N)}, \alpha)] - \lambda^{(t-1)}\right)$

$= (1-\rho_t)\lambda^{(t-1)} + \rho_t \mathbb{E}_q[\eta_g(x_i^{(N)}, z_i^{(N)}, \alpha)]$

$\rho_t = (t+\tau)^{-k}$ ← forgating gate, $\in (0.5, 1]$

mini batch. $\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \frac{\rho_t}{S}\sum_s \hat{\lambda}_s$

LDA example.

$p(z_{dn}=k \mid \theta_d, \beta_{1:K}, w_{dn}) \propto \alpha \exp\{\log \theta_{dk} + \log \beta_{k,w_{dn}}\}$

$q(z_{dn}) \sim$ Multinomial (ϕ_{dn})

$p(\theta_d \mid z_d) =$ Dirichlet $(\alpha + \sum_{n=1}^{N} z_{dn})$

$q(\theta_d) \sim$ Dirichlet (γ_d)

local update. $\phi_{dn}^k \propto \gamma_d$.

$\phi_{dn}^k \propto \exp\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}\}, \quad k \in \{1 \cdots k\}$

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

global upadate.

$p(\beta_k \mid z, w) =$ Dirichlet $(\eta + \sum_d \sum_n z_{dn}^k w_{dn})$

$q(\beta_k) =$ Dirichlet (λ_k)

update: ~~$\lambda_k = \eta + \sum_d \sum$~~ $\hat{\lambda}_k = \eta + D\sum_n \phi_{dn}^k w_{dn}$.

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