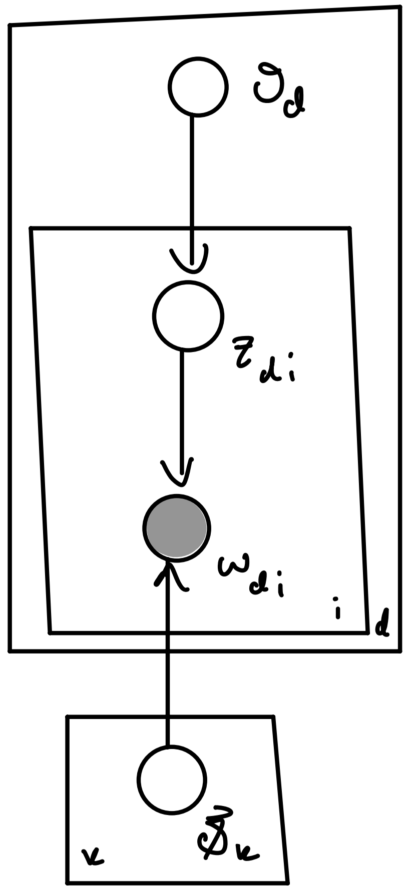
Latent Dirichlet Allocation (LDA)



Data: $w_{d1} \cdots w_{dN_d}$ "tokens" in document d

Model:

for topic $k = 1 \ldots K$:

$\vec{\phi}_k \sim \mathrm{Dir}(\beta_1 \cdots \beta_V)$

for doc $d = 1 \cdots D$:

$\vec{\theta}_d \sim \mathrm{Dir}(\alpha_1 \cdots \alpha_K)$

for token i in $1 \ldots N_d$:

$z_{di} \sim \mathrm{Cat}(\vec{\theta}_d)$

$w_{di} \sim \mathrm{Cat}(\vec{\phi}_{z_{di}})$

Complete conditionals

$$p(z_{di} = k \mid w_{di} = v, \cdots) \propto p(z_{di} = k)\, p(w_{di} = v \mid z_{di} = k, \cdots)$$

$$\propto \theta_{dk}\, \phi_{kv}$$

$$= \frac{\theta_{dk}\, \phi_{kv}}{\sum_j \theta_{dj}\, \phi_{jv}}$$

$$p(\vec{\theta}_d \mid -) \propto \mathrm{Dir}(\vec{\theta}_d ; \vec{\alpha}) \prod_i \mathrm{Cat}(z_{di} ; \vec{\theta}_d)$$

$$= \mathrm{Dir}(\vec{\theta}_d ; \tilde{\alpha}_{d1} \cdots \tilde{\alpha}_{dK})$$

where $\tilde{\alpha}_{dj} = \alpha_k + \overbrace{\sum_{i=1}^{N_d} \mathbb{1}(z_{di} = k)}^{= y_{dk}}$

$$p(\vec{\phi}_k \mid -) \propto \mathrm{Dir}(\vec{\phi}_k; \vec{\beta}) \prod_d \prod_i \left[\mathrm{Cat}(w_{di}; \vec{\phi}_k)\right]^{1(z_{di}=k)}$$

$$\hookrightarrow = \mathrm{Dir}(\vec{\phi}_k; \tilde{\beta}_{k1} \cdots \tilde{\beta}_{kV})$$

where $\tilde{\beta}_{kv} = \beta_v + \underbrace{\sum_d \sum_i 1(w_{di}=v)\, 1(z_{di}=k)}_{\equiv y_{kv}}$

Gibbs sampling

B "burn-in" iterations

S posterior samples

for iter $s = -B \ldots -1, 0, 1 \cdots ST$:

"thinning" every T samples

① for doc d:

for token i:

$$z_{di}^s \sim \mathrm{Cat}\left(\frac{\theta_{d1}\phi_{1w_{di}}}{\sum_j \theta_{dj}\phi_{jw_{di}}} \cdots \frac{\theta_{dK}\phi_{Kw_{di}}}{\sum_j \theta_{dj}\phi_{jw_{di}}}\right)$$

for k, v:

$$y_{dkv}^s = \sum_{i=1}^{N_d} 1(z_{di}^s = k)\, 1(w_{di} = v)$$

② for doc d:

$$\vec{\theta}_d^s \sim \mathrm{Dir}\left(\alpha_1 + \sum_v y_{d1v}^s \cdots \alpha_K + \sum_v y_{dKv}^s\right)$$

③ for topic k:

$$\vec{\phi}_k^s \sim \mathrm{Dir}\left(\beta_1 + \sum_d y_{dk1}^s \cdots \beta_V + \sum_d y_{dkV}^s\right)$$

Notice that the sufficient statistics y_{dkv} are invariant to the token ordering $i = 1 \dots N_d$.

An equivalent way to sample is:

for doc $d = 1 \dots D$
for word $v = 1 \dots V$:

$$(y_{dv1} \cdots y_{dvk}) \sim \text{Mult}\left(y_{dv};\ \frac{\theta_{d1}\phi_{1v}}{\sum_j \theta_{dj}\phi_{jv}} \cdots \frac{\theta_{dk}\phi_{kv}}{\sum_j \theta_{dj}\phi_{jv}}\right)$$

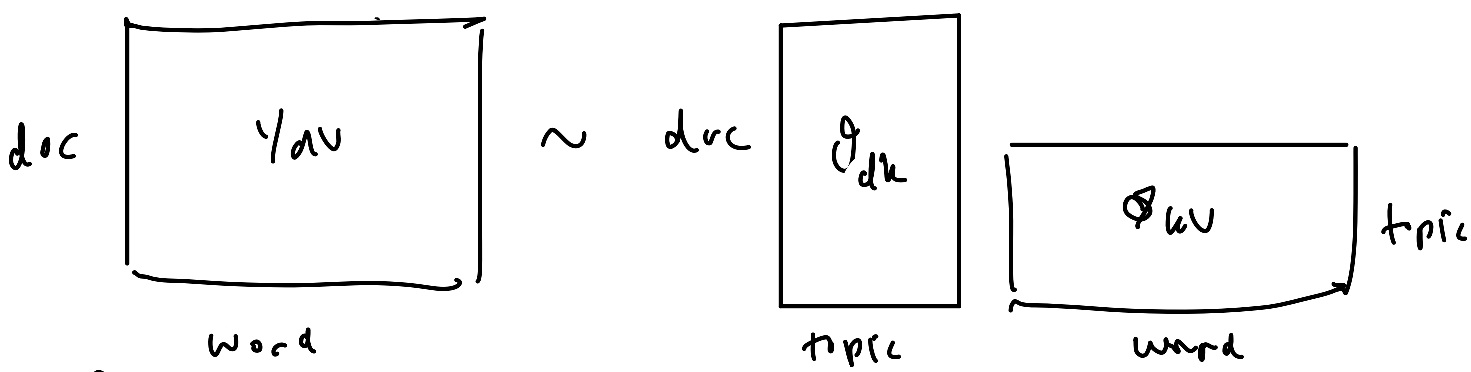
$$y_{dv} \equiv \sum_i \mathbb{1}(w_{di} = v)$$

(observed)

Notice that this is exactly the same auxiliary variable sampling step we derived for Poisson matrix factorization

$$y_{dv} \sim \text{Pois}\left(\sum_j \theta_{dj}\phi_{jv}\right)$$

LDA is "just" a (very) special case of NMF.



each doc is a "bag of words".

Another way to view LDA is as an admixture or mixed membership model. Each document is part of multiple mixture components ("topics").

If $\vec{\theta}_d$ were a "one-hot" vector, we would recover the mixture model.

Label switching

- In LDA and other forms of AMF, there is no inherent ordering of the topics $1 \dots k$:

$$\sum_k \theta_{dk} \phi_{kv} = \sum_k \theta_{d\sigma(k)} \phi_{\sigma(k)v}$$

where $\sigma(k)$ is a permutation of indices $1 \dots k$

- During MCMC, the "labels" can switch
- True for most (ad)mixtures
- This means you should not average any quantity indexed by k across posterior samples:

$$\text{e.g.} \quad \underbrace{\frac{1}{S}\sum_s \phi^s_{kv}}_{\text{bad}} \quad \text{vs.} \quad \underbrace{\frac{1}{S}\sum_s \sum_k \theta^s_{dk} \phi^s_{kv}}_{\text{good}}$$

Variational Inference (VI)

setting:

data: $x_1 \dots x_n \equiv X$

latents: $z_1 \dots z_D$

posterior: $p(z \mid x)$

(intractable)

We have seen one way to approximate the posterior using Gibbs sampling:

$$z_d^s \sim p(z_d \mid z_{\neg d}, x)$$

VI is another way...

Define a family of approximate densities:

$$q(z) \in \mathcal{Q}$$

find the member that minimizes:

$$q^*(z) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\ KL\left(q(z) \,\|\, p(z \mid x)\right)$$

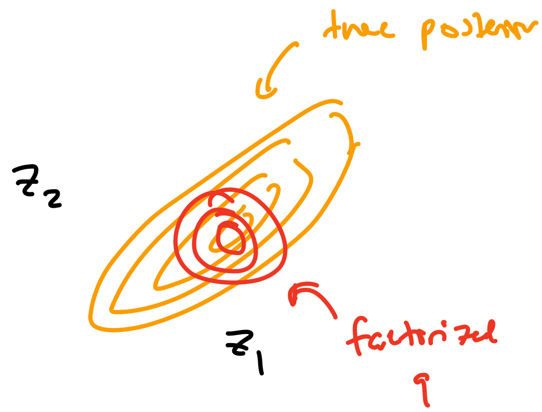
How do you minimize if you cannot compute $p(z \mid x)$?

$$KL(q \,\|\, p) = \mathbb{E}_q\left[\log \frac{q(z)}{p(z \mid x)}\right]$$

$$= \underbrace{\mathbb{E}_q\left[\log \frac{q(z)}{p(z, x)}\right]}_{-\text{ELBO}(q)} + \underbrace{\log p(x)}_{\text{log evidence}}$$

$$q^*(z) = \underset{q}{\operatorname{argmax}}\ \text{ELBO}(q)$$

Define the family Q to be "convenient".



Factorized family:

$$q(z) = \prod_{d=1}^{D} q_d(z_d) = \prod_{d=1}^{D} q(z_d; \lambda_d)$$

"variational parameters"

Coordinate ascent VI (CAVI)

repeat:

for $d = 1 \ldots D$:

$$q^*(z_d) = \underset{q(z_d)}{\arg\max}\ \mathrm{ELBO}(q)$$

Fact (Bishop 2006):

$$q^*(z_d) \propto \exp\left(\mathbb{E}_{\setminus d}\left[\log p(z_d \mid z_{\setminus d}, x)\right]\right)$$

If $p(z_d \mid z_{\setminus d}, x)$ is expfam, this will often be "nice".

Proof: $KL(q(z_d) \,\|\, q^*(z_d))$

$$= \mathbb{E}_{q(z_d)}\left[\log q(z_d) - \log q^*(z_d)\right]$$

$$\propto \mathbb{E}_{q(z_d)}\left[\log q(z_d)\right] - \mathbb{E}_{q(z_d)}\left[\mathbb{E}_{q(z_{\setminus d})}\left[\log p(z_d \mid z_{\setminus d}, x)\right]\right]$$

$$\underset{q(z_d)}{\propto} -H(q) - \mathbb{E}_q\left[\log(z, x)\right]$$

$$\propto -\mathrm{ELBO}$$

Minimizing the KL above maximizes the ELBO.

CAVI for LDA

$$q(\cdots) = \Big[\prod_d \prod_v q(y_{dv} \cdots y_{dvk})\Big] \prod_d q(\vec{\theta}_d) \prod_k q(\vec{\phi}_k)$$

$$q^*(\vec{\theta}_d) \propto \exp\Big(\mathbb{E}_{q_{\setminus \vec{\theta}_d}}\big[\log p(\vec{\theta}_d \mid -)\big]\Big)$$

$$\propto \exp\Big(\mathbb{E}\big[\log \mathrm{Dir}(\vec{\theta}_d ; \alpha_1 + \sum_v y_{dv1} \cdots \alpha_k + \sum_v y_{dvk})\big]\Big)$$

$$\propto_{\theta_d} \exp\Big(\mathbb{E}\Big[\log \prod_k \theta_{dk}^{\alpha_k + \sum_v y_{dvk} - 1}\Big]\Big)$$

$$\propto_{\theta_d} \prod_k \theta_{dk}^{\alpha_k + \sum_v \mathbb{E}[y_{dvk}] - 1}$$

$$\Downarrow$$

$$q^*(\vec{\theta}_d) = \mathrm{Dir}(\vec{\theta}_d ; \lambda_{d1} \cdots \lambda_{dk}) \qquad (1)$$

where $\lambda_{dk} = \alpha_k + \sum_v \mathbb{E}[y_{dvk}]$

Similarly:

$$q^*(\vec{\phi}_k) = \mathrm{Dir}(\vec{\phi}_k ; \lambda_{k1} \cdots \lambda_{kV}) \qquad (2)$$

where $\lambda_{kv} = \beta_v + \sum_d \mathbb{E}[y_{dvk}]$

$$q^*(y_{dv1} \cdots y_{dvk}) \propto \exp\left(\log E_Q[\mathrm{Mult}(\cdots)]\right)$$

$$\propto \exp\left(E\left[\log \delta(\cdots) \frac{y_{dv}!}{\prod_k y_{dvk}!} \prod_k (\cdots)^{y_{dvk}}\right]\right)$$

$$\propto \delta(\cdots) \frac{y_{dv}!}{\prod_k y_{dvk}!} \exp\left(E\left[\log \prod_k \left(\frac{\theta_{dk}\phi_{kv}}{\theta_d^T \phi_v}\right)^{y_{dvk}}\right]\right)$$

$$\propto \delta(\cdots) \frac{y_{dv}!}{\prod_k y_{dvk}!} \prod_k \underbrace{\exp\left(E\left[\log(\theta_{dk}\phi_k)\right]\right)}^{y_{dvk}}$$

$$\equiv G_Q[\theta_{dk}\phi_k] = G_Q(\theta_{dk})\,G_Q(\phi_{kv})$$

$$\boxed{\begin{array}{l}\underline{\text{Def}^n} \\ \vec{x} \sim \mathrm{Dir}(\alpha_1 \cdots \alpha_n) \\ E[\ln x_v] = \psi(x_v) - \psi\left(\sum_v x_v\right)\end{array}}$$

digamma function

$$\hookrightarrow q^*(\vec{y}_{dv}) = \mathrm{Mult}(y_{dv};\ \vec{p}_{dv})$$

where $$p_{dvk} = \frac{G(\theta_{dk})\,G(\phi_{kv})}{\sum_j G(\theta_{dj})\,G(\phi_{jv})}$$ (3)

CAVI for LDA

initialize variational parameters λ_{dk}, λ_{kv}, p_{dvk} and calculate initial expectations $\mathbb{E}_q[y_{dvk}]$, $G[\theta_{dk}]$, $G[\phi_{kv}]$

repeat until convergence:

① update $q(\vec{y}_{dv}) = \text{Mult}(y_{dv}; Y_{dv}, \vec{p}_{dv})$ only need to do this for $Y_{dv} > 0$

$$p_{dvk} = \frac{G_q[\theta_{dk}]\, G_q[\phi_{kv}]}{\sum_{j=1}^{K} G_q[\theta_{dj}]\, G_q[\phi_{jv}]}$$

$$\mathbb{E}_q[y_{dvk}] = Y_{dv}\, p_{dvk}$$

② update $q(\vec{\theta}_d) = \text{Dir}(\vec{\theta}_d; \vec{\lambda}_d)$

$$\lambda_{dk} = \alpha_k + \sum_v \mathbb{E}_q[y_{dvk}]$$

$$G_q[\theta_{dk}] = \exp\left(\psi(\lambda_{dk}) - \psi\left(\sum_j \lambda_{dj}\right)\right)$$

③ update $q(\vec{\phi}_k) = \text{Dir}(\vec{\phi}_k; \vec{\lambda}_k)$

$$\lambda_{kv} = \beta_v + \sum_d \mathbb{E}_q[y_{dvk}]$$

$$G_q[\phi_{kv}] = \ldots$$

④ Calculate $\text{ELBO}(q)$; assess convergence

ELBO:

$$\mathbb{E}_q\left[\log \frac{P(Y, \vec{y}, \theta, \phi)}{q(\vec{y}, \theta, \phi)}\right]$$

prior

$$= \mathbb{E}_q[\log P(y, \vec{y} \mid \theta, \phi)] + \mathbb{E}_q[\log P(\theta, \phi)]$$

"complete data" likelihood

$$+ H_q[q(\cdots)]$$

entropy of variational dist

e.g.
$$\mathbb{E}_q[\log P(\theta)] = \sum_d \mathbb{E}_q[\log P(\vec{\theta}_d)]$$

$$\mathbb{E}_q[\log \mathrm{Dir}(\vec{\theta}_d; \vec{\alpha})] = \mathbb{E}_q\left[\log \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_{dk}^{\alpha_k - 1}\right]$$

$$= \log \frac{\Gamma(\cdots)}{\prod_k \Gamma(\cdots)} + \sum_k (\alpha_k - 1)\, \mathbb{E}_q[\log \theta_{dk}]$$

Recall that $q(\vec{\theta}_d) = \mathrm{Dir}(\vec{\theta}_d; \vec{\lambda}_d)$ and that

$$\mathbb{E}_q[\ln \theta_{dk}] = \psi(\lambda_{dk}) - \psi(\sum_j \lambda_{dj})$$

e.g.
$$H_q[q(\cdots)] = \sum_d H[q(\vec{\theta}_d)] + \cdots$$

entropy of $\mathrm{Dir}(\vec{\theta}_d; \vec{\lambda}_d)$

The entire ELBO can be (carefully) derived and calculated.

CAVI for conditionally conjugate models

$$p(\eta, z, x) = p(\beta) \prod_i p(z_i, x_i \mid \eta)$$

- η are <u>global</u> params
- z are <u>local</u> latents
- x are data

complete data likelihod:

$$p(z_i, x_i \mid \eta) = h_\ell(z_i, x_i) \exp\left(\eta^T t_\ell(z_i, x_i) - A_\ell(\eta)\right)$$

$$\eta \equiv \eta(\theta)$$

(natural parameterization)

$$p(\eta \mid \alpha) = F(\eta ; \alpha) \propto_\eta h_c(\eta) \exp\left(\alpha^T t_c(\eta)\right)$$

$$t_c(\eta) = [\eta, -A_\ell(\eta)]$$

$$p(\eta \mid -) \propto h_c(\eta) \exp\left(\hat{\alpha}^T t_c(\eta)\right)$$

$$\hat{\alpha} = \begin{bmatrix} \alpha_1 + \sum_{i=1}^{n} t_\ell(z_i, x_i) \\ \alpha_2 + n \end{bmatrix}$$

↳ natural parameter of complete conditional

$$= F(\eta ; \hat{\alpha})$$ (same family)

Optimal Variational family:

$$q^*(\eta) \propto_\eta \exp\left(\mathbb{E}_q [\log p(\eta \mid -)] \right)$$

$$\propto_\eta h_c(\eta) \exp\left(\mathbb{E}_q[\hat{\alpha}]^T t_c(\eta) \right)$$

$$= F(\eta ; \mathbb{E}_q[\hat{\alpha}])$$

variational natural parameter λ

So $q^*(\eta ; \lambda)$ is the same family as the prior $F(\eta ; \alpha)$ and complete cond. $F(\eta ; \hat{\alpha})$.

Taking gradients:

$$\nabla_\lambda \text{ELBO} = \underbrace{A_c''(\lambda)}_{\text{Hessian of the log normalizer}} \left(\mathbb{E}_q[\hat{\alpha}] - \lambda \right)$$

i.e. $A_c''(\lambda)_{ij} = \dfrac{\partial^2 A_c(\lambda)}{\partial \lambda_i \, \partial \lambda_j}$

So $\lambda = \mathbb{E}_q[\hat{\alpha}]$ maximizes ELBO.

Stochastic Variational Inference

setting: large n

stochastic optimization:

$$\left\| \begin{array}{l} \text{for iter } t: \\ \quad \lambda_t = \lambda_{t-1} + \epsilon_t \widehat{\nabla_\lambda \text{ELBO}} \end{array} \right.$$

if $\widehat{\nabla_\lambda \text{ELBO}}$ is an unbiased estimate of the gradient:

$$\mathbb{E}\left[\widehat{\nabla_\lambda \text{ELBO}}\right] = \nabla_\lambda \text{ELBO}$$

and if the steps follow the Robins-Munroe conditions:

$$\sum_t \epsilon_t = \infty \quad ; \quad \sum_t \epsilon_t^2 < \infty$$

$$\text{e.g. } \epsilon_t = t^{-k}, \quad k \in (1/2, 1]$$

then λ_t will reach a (local) optimum of ELBO.

If the gradient had the following structure:

$$\nabla_\lambda \text{ELBO} = \sum_{i=1}^{n} f(x_i, \cdots)$$

then a sub-sampled estimate would be unbiased:

$$i \sim \text{Uniform}(1 \ldots n)$$

$$\widehat{\nabla_\lambda \text{ELBO}} = n\, f(x_i \cdots)$$

Unfortunately the Euclidean gradient ∇_λ ELBO does not.

Hoffman et al. (2013) showed:

$$\nabla_\lambda \text{ELBO} = A_c''(\lambda) \Big(\underbrace{\mathbb{E}_q[\hat{\alpha}] - \lambda}_{\triangleq\, g(\lambda)} \Big)$$

natural gradient of the ELBO

$$g(\lambda) = [A_c''(\lambda)]^{-1} \nabla_\lambda \text{ELBO}$$

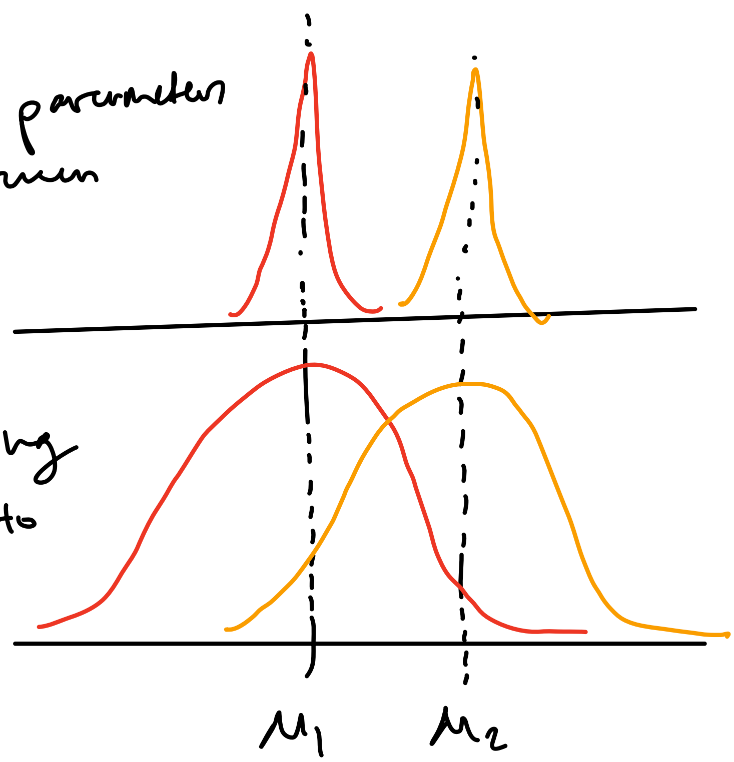
the Euclidean gradient preconditioned by inverse $A_c''(\lambda)$.

$$= \mathbb{E}_q[\hat{\alpha}] - \lambda$$

$$= \begin{bmatrix} \alpha_1 + \sum_i \mathbb{E}_q[t(x_i, z_i)] \\ \alpha_2 + n \end{bmatrix} - \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}$$

Euclidean distance between parameters $\neq$ statistical distance between distributions.

The natural gradient is defined by a local rescaling of the Euclidean gradient to account for this.



SVI

$$\lambda_t = \lambda_{t-1} + \varepsilon_t\, g(\lambda_{t-1})$$

$$= \lambda_{t-1} + \varepsilon_t \left(\mathbb{E}_q[\hat{\alpha}] - \lambda_{t-1} \right)$$

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

replace this with an unbiased estimate

$i \sim \text{Uniform}(1 \dots n)$

$$\hat{\mathbb{E}}_q[\hat{\alpha}] = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} n\, \mathbb{E}_q[t(z_i, x_i)] \\ n \end{bmatrix}$$

Note that $\mathbb{E}_q[t(z_i, x_i)] \equiv \mathbb{E}_{q(z_i)}[t(z_i, x_i)]$

SVI :

for iter t :

$i \sim \text{Uniform}(1 \dots n)$

1. update local latent

$$q^*(z_i) = \dots$$

2. update global param

$$\lambda_t = (1-\varepsilon_t)\lambda_{t-1} + \varepsilon_t \left(\alpha + \begin{bmatrix} n\, \mathbb{E}_q[t(z_i, x_i)] \\ n \end{bmatrix} \right)$$