Variational Inference (Contd)

CS772A: Probabilistic Machine Learning

Piyush Rai

Quick Recap: Variational Inference (VI)

Defines a class of distributions parametrized by ϕ

- Approximate the true posterior $p(\boldsymbol{Z}|\boldsymbol{X})$ by an approx. distribution $q(\boldsymbol{Z}|\phi)$ or $q_\phi(\boldsymbol{Z})$

$$\phi^* = \operatorname{argmin}_\phi \text{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X})]$$

Often, we will simply write it as $\operatorname{argmin}_q \text{KL}[q||p_z]$

- Due to the below identity, minimizing the KL is equivalent to maximizing the ELBO

Log-evidence of model $\boldsymbol{m}$

Evidence lower bound (ELBO)

Non-negative

$$\log p(\boldsymbol{X}|m) = \mathcal{L}(q) + \text{KL}(q||p_z)$$

- The ELBO is defined as $\mathcal{L}(q) = \int q(\boldsymbol{Z}) \log \left[\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] d\boldsymbol{Z}$

$$\mathcal{L}(q) = \mathcal{L}(\phi) = \mathbb{E}_q[\log p(\boldsymbol{X},\boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})]$$

Find $\boldsymbol{q}$ such that $\boldsymbol{Z}$ explains data well

$$= \mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{Z})] - \text{KL}[q(\boldsymbol{Z})||p(\boldsymbol{Z})]$$

Find $\boldsymbol{q}$ that is "simple", i.e., is close to the prior

- VI optimizes (maximizes) the above w.r.t. $\boldsymbol{q}$, i.e., w.r.t. its variational parameters ϕ

Quick Recap: Mean-Field VI

- Assume $q(\boldsymbol{Z}|\phi) = \prod_{i=1}^{M} q(\boldsymbol{Z}_i|\phi_i)$. Simplifies ELBO expression/its maximization
- Learning the optimal q then reduces to learning the optimal $q_1, q_2, \ldots, q_M$
- For mean-field VI, each optimal factor q_j is given by

This general expression holds even if there is no local conjugacy

For locally conjugate models, $q_j^*(\boldsymbol{Z}_j)$ will have the same form as prior $p(Z_j)$

$$\boxed{q_j^* = \frac{\exp(\mathbb{E}_{i \neq j}[\log p(\boldsymbol{X}, \boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\log p(\boldsymbol{X}, \boldsymbol{Z})]\, d\boldsymbol{Z}_j}}$$

$\mathbb{E}_{i \neq j}$ denotes the expectation w.r.t. the distribution $\prod_{i \neq j} q_i$

- Updates of optimal $q_1, q_2, \ldots, q_M$ depend on each other because of the expectations
- Therefore, MFVI works by updating the q_j's in a cyclic fashion
  - Leads to the coordinate ascent VI (CAVI) algorithm

Mean-Field VI: A Closer Look

- Since $\log q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const} = \mathbb{E}_{i \neq j}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}_j, \boldsymbol{Z}_{-j})\right] + \text{const}$

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j})] + \text{const}$$

For any model

- Thus opt variational distr $q_j^*(\boldsymbol{Z}_j)$ basically requires expectations of CP $p(\boldsymbol{Z}_j | \boldsymbol{X}, \boldsymbol{Z}_{-j})$
- For locally conjugate models, CP can be easily found and is an exp-fam distr of the form

$$p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j}) = h(\mathbf{Z}_j) \exp\left[\eta(\mathbf{X}, \mathbf{Z}_{-j})^\top \mathbf{Z}_j - A(\eta(\mathbf{X}, \mathbf{Z}_{-j}))\right]$$

Gibbs sampling samples from each CP. MFVI uses each CP to compute the corresponding $\boldsymbol{q}_j$

- Using the above, we can rewrite the optimal variational distribution as follows

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}\left[\log\left(h(\mathbf{Z}_j) \exp\left[\eta(\mathbf{X}, \mathbf{Z}_{-j})^\top \mathbf{Z}_j - A(\eta(\mathbf{X}, \mathbf{Z}_{-j}))\right]\right)\right] + \text{const}$$

$$\Longrightarrow q_j^*(\mathbf{Z}_j) \propto h(\mathbf{Z}_j) \exp\left[\mathbb{E}_{i \neq j}[\eta(\mathbf{X}, \mathbf{Z}_{-j})]^\top \mathbf{Z}_j\right] \qquad \text{(verify)}$$

- Thus, with local conj, we just require expectation of nat. params. of CP of $\boldsymbol{Z}_j$

VI by Computing ELBO Gradients

- Can also do VI by computing ELBO's gradient and doing gradient based optimization
- Gradient based approach is broadly applicable, not just for mean-field VI
    1. Assume $q(\boldsymbol{Z})$ to be from some family of distributions with variational parameters $\boldsymbol{\phi}$
    2. Write down the full ELBO expression (will give us a function of var. parameters $\boldsymbol{\phi}$)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

    3. Compute ELBO gradients, i.e., $\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi})$ and use gradient methods to find optimal $\boldsymbol{\phi}$
- Step 2 may be simplified due to the problem structure or the form of $q(\boldsymbol{Z})$
    - i.i.d. observations simplify $\log p(\boldsymbol{X}|\boldsymbol{Z})$; conditionally independent priors simplify $\log p(\boldsymbol{Z})$
    - Locally-conjugate models
    - The mean-field assumption simplifies $q(\boldsymbol{Z})$ as $q = \prod_{i=1}^{M} q_i$
        - Moreover, the last term reduces to sum of entropies of q_i's (which usually has known forms)

Mean-Field VI by Taking ELBO's Gradients

- Mean-field assumption $q(\boldsymbol{Z}|\phi) = \prod_{i=1}^{M} q(\boldsymbol{Z}_i|\phi_i)$ results in following optimal distribution

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})]\, d\boldsymbol{Z}_j}$$

This approach is applicable even if we don't have mean-field assumption

Note that here we do not have to assume the form of this variational distribution. We simply compute the RHS and find what it is (in the locally-conjugate case, it will be the same distribution as the prior)

- Alternatively, we can take ELBO's partial deriv w.r.t. $\phi_1, \phi_2, \ldots, \phi_M$ to find their optimal values
- Consider a Bayesian linear regression model

Likelihood

Prior on $\boldsymbol{w}$

λ assumed fixed

Prior on variance of Gaussian likelihood

$$y_i \sim \mathrm{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \mathrm{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \mathrm{Gamma}(a, b)$$

Needed in ELBO

Joint distribution on data and unknowns

$$p(y, w, \alpha|x) = p(\alpha)p(w)\prod_{i=1}^{N} p(y_i|x_i, w, \alpha)$$

Assumed variational posterior with mean-field assumption

$$q(w, \alpha) = q(\alpha)q(w) = \mathrm{Gamma}(\alpha|a', b')\mathrm{Normal}(w|\mu', \Sigma')$$

Note that in this approach, we have to assume a form for each variational distribution. It is common to assume them to have the same form as the respective priors

- Now doing VI amounts to maximizing ELBO to find the optimal variational params a', b', μ', Σ'

Mean-Field VI by Taking ELBO's Gradients

- The ELBO is

For the Bayesian linear regression model, instead of $p(\boldsymbol{X},\boldsymbol{Z})$, it will be of the form $p(\boldsymbol{y},\boldsymbol{Z}|\boldsymbol{X})$

$$\mathcal{L}(q) = \mathcal{L}(\phi) = \mathbb{E}_q[\log p(\boldsymbol{X},\boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})] = \mathbb{E}_q[\log p(\boldsymbol{Z})] + \mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})]$$
$$= \int q(\boldsymbol{Z})\log p(\boldsymbol{Z})d\boldsymbol{Z} + \int q(\boldsymbol{Z})\log p(\boldsymbol{X}|\boldsymbol{Z})d\boldsymbol{Z} + \int q(\boldsymbol{Z})\log q(\boldsymbol{Z})d\boldsymbol{Z}$$

- Thus the ELBO in the Bayesian linear regression model will be (assuming i.i.d. obs)

$$\mathcal{L}(a',b',\mu',\Sigma') = \int q(\alpha)\ln p(\alpha)d\alpha + \int q(w)\ln p(w)dw + \sum_{i=1}^{N}\int\int q(\alpha)q(w)\ln p(y_i|x_i,w,\alpha)dwd\alpha - \int q(\alpha)\ln q(\alpha)d\alpha - \int q(w)\ln q(w)dw$$

Expectations of the log of the prior

Expectations of the log of the likelihood

Expectations of the log of the var. distributions (= their entropies)

- Substituting the priors, likelihoods, and variational distributions

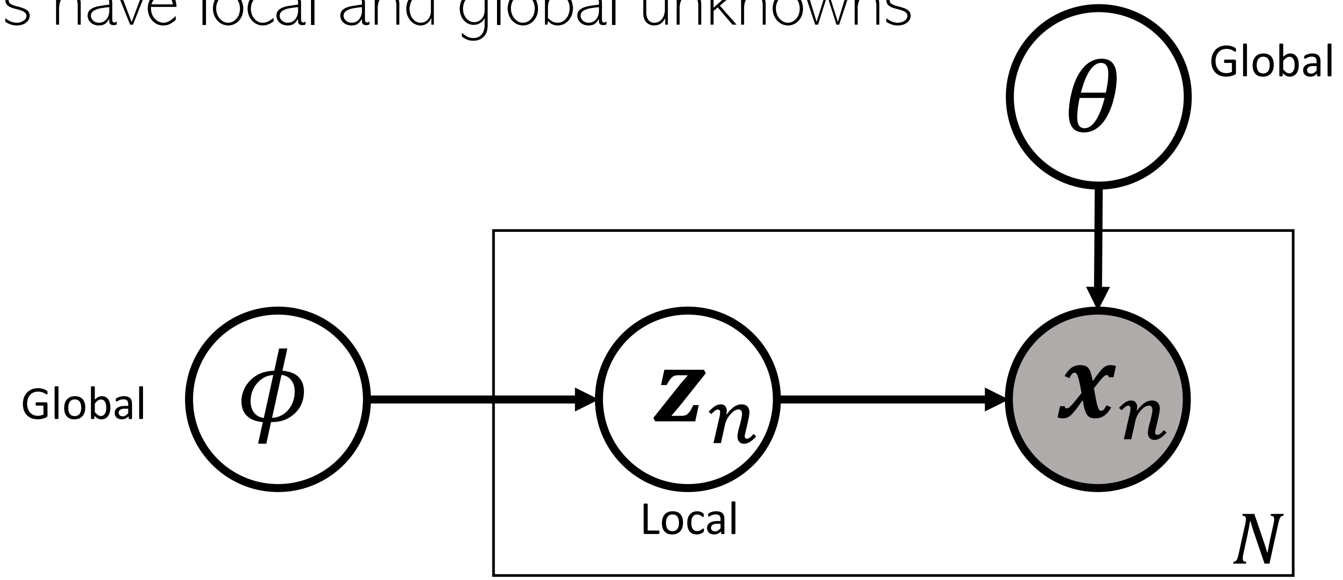
$$\mathcal{L}(a',b',\mu',\Sigma') = (a-1)(\psi(a')-\ln b') - b\frac{a'}{b'} + \text{constant} - \frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} + \frac{N}{2}(\psi(a')-\ln b') - \sum_{i=1}^{N}\frac{1}{2}\frac{a'}{b'}\left((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\right) + \text{constant}$$
$$+a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') + \frac{1}{2}\ln|\Sigma'| + \text{constant}$$

Digamma function (log of gamma function)

- Can now maximize the above ELBO w.r.t. a', b', μ', Σ' in an alternating fashion
- For most models, ELBO or its gradients won't have a simple form (methods like "black-box" variational inference, reparametrization trick etc will be needed in those cases)

MFVI for LVMs with Local and Global Unknowns

- Many LVMs have local and global unknowns

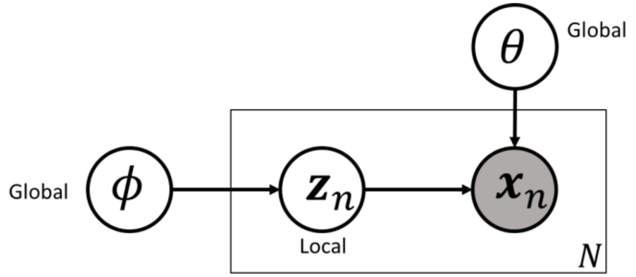


- Examples: Gaussian Mixture Model, Prob. PCA, Variational Autoencoder (VAE), etc
- Denote all local unknowns $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\}$ as $\boldsymbol{Z}$ and global unknown as $\boldsymbol{\beta} = (\theta, \phi)$
- The goal is to infer the posterior $p(\boldsymbol{Z}, \boldsymbol{\beta} | \boldsymbol{X})$ which is intractable in general
- Mean-field VI will approximating this posterior as $p(\boldsymbol{Z}, \boldsymbol{\beta} | \boldsymbol{X}) \approx q(\boldsymbol{Z}, \boldsymbol{\beta}) \approx q(\boldsymbol{Z}) q(\boldsymbol{\beta})$

MFVI for LVMs with Local and Global Unknowns

- Assuming independence, the joint distribution of data $\boldsymbol{X}$ and unknowns $\boldsymbol{\beta} = (\theta, \phi)$

$$p(\mathbf{X}, \mathbf{Z}, \beta) = p(\beta) \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{z}_n, \beta) p(\mathbf{z}_n | \beta) = p(\beta) \prod_{n=1}^{N} p(\mathbf{x}_n, \mathbf{z}_n | \beta)$$



- Assume the joint dist. of $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$ to be an exp-fam dist with natural params $\boldsymbol{\beta}$

$$p(\mathbf{x}_n, \mathbf{z}_n | \beta) = h(\mathbf{x}_n, \mathbf{z}_n) \exp\left[\beta^\top \overbrace{t(\mathbf{x}_n, \mathbf{z}_n)}^{\text{Sufficient statistics}} - A(\beta)\right]$$

- Assume a prior on $\boldsymbol{\beta}$, that is conjugate to the above exp-fam dist

$$p(\beta | \boldsymbol{\alpha}) = h(\beta) \exp\left[\boldsymbol{\alpha}^\top [\beta, -A(\beta)] - A(\boldsymbol{\alpha})\right]$$

where $\alpha = [\alpha_1, \alpha_2]^\top$ are the the hyperparamers of the prior $p(\beta)$ and $[\beta, -A(\beta)]$ is the sufficient statistics vector for this exp-family distribution

MFVI for LVMs with Local and Global Vars

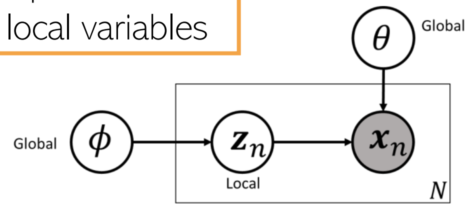
- Recall that mean-field VB can be obtained using CP of each unknown
- Optimal var. distribution for each unknown requires expec. of nat. params of its CP

$$p(\beta|\alpha) = h(\beta)\exp\left[\alpha^\top[\beta, -A(\beta)] - A(\alpha)\right]$$

- Due to conj, CP of global vars $\boldsymbol{\beta} = (\boldsymbol{\theta}, \boldsymbol{\phi})$, will have the same form as prior $\boldsymbol{p(\beta|\alpha)}$

Updates to the natural parameters requires a summing suff-stats over all data and local variables

$$p(\beta|\mathbf{X},\mathbf{Z}) = p(\beta|\hat{\alpha}) \quad \text{where} \quad \hat{\alpha} = \left[\alpha_1 + \sum_{n=1}^{N} t(\mathbf{x}_n, \mathbf{z}_n), \alpha_2 + N\right]$$



- Likewise, CP of each local variable $\boldsymbol{z_n}$

Due to the independence structure

Assuming CP is an exp-fam distribution (will be the case if the prior $p(z_n|\phi)$ and likelihood $p(x_n|z_n,\theta)$ are exp-family and conjugate to each other)

$$p(\mathbf{z}_n|\mathbf{Z}_{-n}, \mathbf{X}, \beta) = p(\mathbf{z}_n|\mathbf{x}_n, \beta) = h(\mathbf{z}_n)\exp\left[\eta(\mathbf{x}_n, \beta)^\top \mathbf{z}_n - A(\eta(\mathbf{x}_n, \beta))\right]$$

Nat. params depends on data x_n and global var β

- Having these CPs, we can compute the mean-field updates for $\boldsymbol{q(\beta)}$ and $\boldsymbol{q(z_n)}$

MFVI for LVMs with Local and Global Vars

- Let's assume our mean-field approximation to be of the form

$$q(\beta, \mathbf{Z}) = q(\beta|\lambda)\prod_{n=1}^{N} q(\mathbf{z}_n|\phi_n)$$

- CPs are exp-fam, so optimal q's depend on expected suff-stats of CP's nat. params
- The optimal variational dist. for local vars $\mathbf{z}_n$ will be $q(\mathbf{z}_n|\phi_n)$ with

Basically requires expectation over the $q(\beta|\lambda)$ distribution

$$\phi_n = \mathbb{E}_\lambda\left[\eta(\mathbf{x}_n, \beta)\right] \qquad \forall n$$

- The optimal variational dist. for global vars β will be $q(\beta|\lambda)$ with

Basically requires expectation over the $q(z_n|\phi_n)$ distribution

$$\lambda = \left[\alpha_1 + \sum_{n=1}^{N}\mathbb{E}_{\phi_n}[t(\mathbf{x}_n, \mathbf{z}_n)], \alpha_2 + N\right]^\top$$

- Mean-Field updates alternate between estimating ϕ_n's and λ until convergence
- Potential bottleneck: Updating λ requires waiting for all ϕ_n's to be updated (thus slow)
  - Can be handled using online VI (stochastic VI)

Stochastic Variational Inference (SVI)

- An "online" algorithm† to speed-up VI for LVMs with local and global variables
- Recall the mean-field VI updates $(q(\boldsymbol{\beta}, \mathbf{Z}) = q(\boldsymbol{\beta}|\lambda) \prod_{n=1}^{N} q(\mathbf{z}_n|\phi_n))$ for such models

Local var. params

Nat. param of CP of $\boldsymbol{z}_n$

Global var. params

Slow; requires all local var params $\boldsymbol{\phi}_n$'s to be computed already

Nat. param of CP of $\boldsymbol{\beta}$

$$\phi_n = \mathbb{E}_{\lambda}\left[\eta(\mathbf{x}_n, \boldsymbol{\beta})\right] \quad \forall n \quad \text{and} \quad \lambda = \left[\alpha_1 + \sum_{n=1}^{N} \mathbb{E}_{\phi_n}[t(\mathbf{x}_n, \mathbf{z}_n)], \alpha_2 + N\right]^\top = \mathbb{E}_{\phi}[\hat{\boldsymbol{\alpha}}(\mathbf{X}, \mathbf{Z})]$$

- SVI uses minibatches to make the global param $\boldsymbol{\lambda}$ updates more efficient
  1. Initialize $\boldsymbol{\lambda}$ randomly as $\boldsymbol{\lambda}^{(0)}$ and set current iteration number as $\boldsymbol{i} = 1$
  2. Set the learning rate (decaying as) as $\boldsymbol{\epsilon_i} = (\boldsymbol{i} + 1)^{-\boldsymbol{\kappa}}$ where $\boldsymbol{\kappa} \in (0.5,1]$
  3. Choose a data point $\boldsymbol{n}$ uniformly randomly, i.e., $\boldsymbol{n} \sim \mathbf{Uniform}(1,2, \ldots, N)$ Assuming minibatch size = 1
  4. Compute local var. param $\boldsymbol{\phi_n}$ for data point $\boldsymbol{x_n}$ as $\phi_n = \mathbb{E}_{\lambda^{(i-1)}}\left[\eta(\mathbf{x}_n, \boldsymbol{\beta})\right]$
  5. Update $\boldsymbol{\lambda}$ as $\lambda^{(i)} = (1-\epsilon_i)\lambda^{(i-1)} + \epsilon_i \lambda_n$ where $\lambda_n = [\alpha_1 + \mathbb{E}_{\phi_n}[t(\mathbf{x}_n, \mathbf{z}_n)], \alpha_2 + 1]^\top = \mathbb{E}_{\phi_n}[\hat{\boldsymbol{\alpha}}(\mathbf{x}_n, \mathbf{z}_n)]$
  6. Set $\boldsymbol{i} = \boldsymbol{i} + \mathbf{1}$. If ELBO not converged, go to Step 2

†Stochastic Variational Inference (Hoffman et al, 2013)

What is SVI Doing?

- SVI updates the global var params $\boldsymbol{\lambda}$ using stochastic optimization† of the ELBO
- However, instead of usual gradient of ELBO w.r.t. $\boldsymbol{\lambda}$, SVI uses the natural gradient
- Denoting the double derivative of the log-partition function of CP of $\boldsymbol{\beta}$ as A''

$$\text{Usual gradient:} \quad \nabla_\lambda \text{ELBO} = A''(\lambda)(\mathbb{E}_\phi[\hat{\alpha}(\mathbf{X}, \mathbf{Z})] - \lambda)$$

If interested in the proof, can see the derivation in the SVI paper

$$\text{Natural gradient:} \quad g(\lambda) = A''(\lambda)^{-1} \times \nabla_\lambda \text{ELBO} = \mathbb{E}_\phi[\hat{\alpha}(\mathbf{X}, \mathbf{Z})] - \lambda$$

Note: $A''(\lambda)$ is cov. of suff-stats of CP of β and $A''(\lambda)^{-1}$ is the Fisher information matrix

- Using the natural gradient has some nice advantages
  - Nat. grad. based updates of $\boldsymbol{\lambda}$ have simple form + easy to compute (no need to compute A'')

$$\lambda^{(i)} = \lambda^{(i-1)} + \epsilon_i g(\lambda)|_{\lambda^{(i-1)}} = (1 - \epsilon_i)\lambda^{(i-1)} + \epsilon_i \mathbb{E}_\phi[\hat{\alpha}(\mathbf{X}, \mathbf{Z})] \quad \text{(assuming full batch)}$$

  - Natural grad. are more intuitive/meaningful: Euclidean distance isn't often meaningful when used to compute distance between parameters of probability distributions, e.g., $q(\beta|\lambda)$ and $q(\beta|\lambda')$

†Stochastic Variational Inference (Hoffman et al, 2013)

SVI: Some Comments

- Often operates on minibatches: For iteration i minibatch $\mathcal{B}_i$, update λ as follows

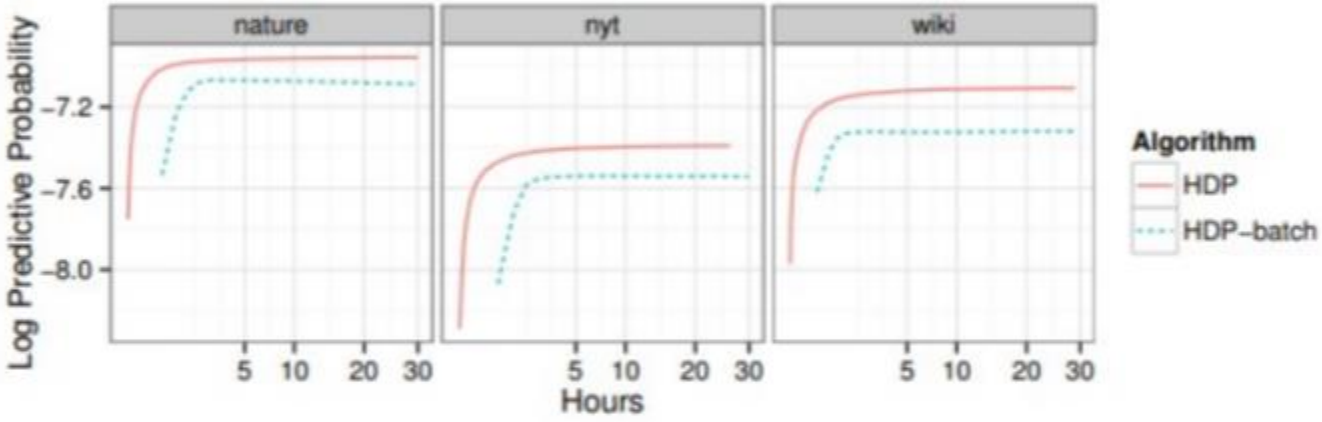
Global var. param computed on this minibatch

$$\hat{\lambda} = \frac{1}{|\mathcal{B}_i|}\sum_{n\in\mathcal{B}_i}\lambda_n$$

Now blending with the older estimate of λ from iteration $i-1$

$$\lambda^{(i)} = (1-\epsilon_i)\lambda^{(i-1)} + \epsilon_i\hat{\lambda}$$

- Decaying learning rate ϵ_i is necessary for convergence (need $\sum_i \epsilon_i = \infty$ and $\sum_i \epsilon_i^2 < \infty$)
- SVI successfully used on many large-scale problems (topic modeling, citation network analysis, etc). Much faster convergence (and better results) compared to batch VI



SVI vs Batch VI on a nonparametric Bayesian Topic Model (Hierarchical Dirichlet Process)

†Stochastic Variational Inference (Hoffman et al, 2013)

Coming Up Next

- VI for non-conjugate models
  - Black-box VI (BBVI) for general purpose VI
  - Reparametrization Trick for general purpose VI
  - Some model-specific tricks

- Other recent advances in VI
  - Amortized VI
  - Structured VI