Variational Inference (Contd) and Some Recent Advances

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Plan

- Some properties of VI
- VI for non-conjugate models
- Some recent advances in VI

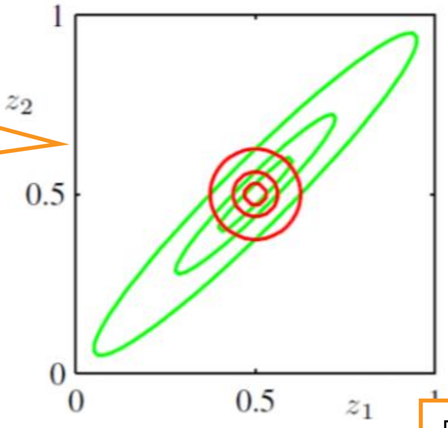
Some Properties of VB

- Recall that VB approximates a posterior p by finding q that minimizes $\mathrm{KL}(q||p)$

$$\mathrm{KL}(q||p) = \int q(\mathbf{Z}) \log \left[\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \right]$$

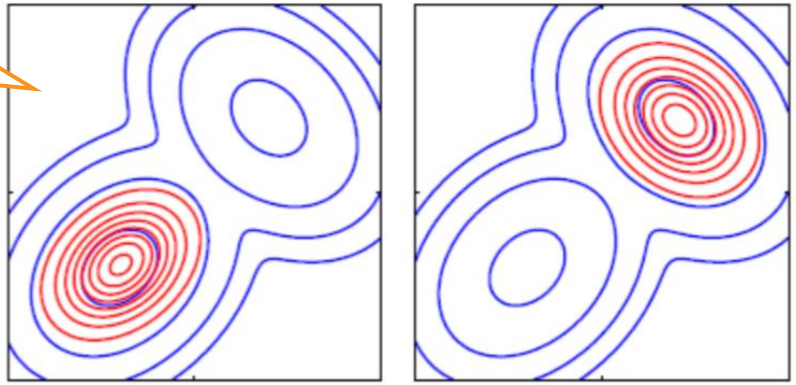
- $q(\boldsymbol{Z})$ will be small where $p(\boldsymbol{Z}|\boldsymbol{X})$ is small otherwise KL will blow up
- Thus $q(\boldsymbol{Z})$ avoids low-probability regions of the true posterior

q (red) avoids regions of p (green) where the latter has low values



q (red) concentrated on one of the modes of p (blue)

For q to also capture the other mode, it will require crossing the low-probability region of p, thereby blowing up the KL



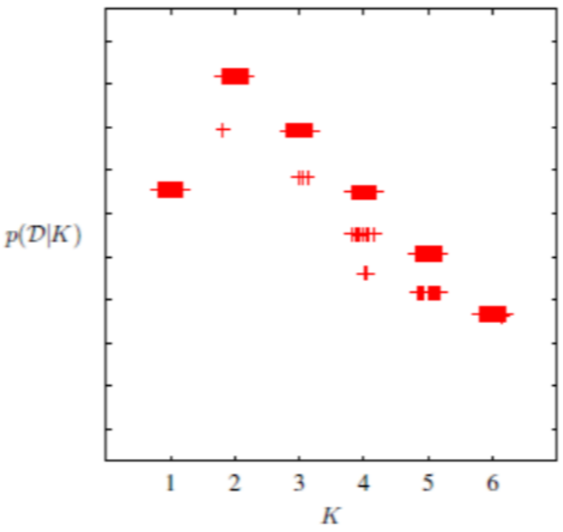
EP minimizes $\mathrm{KL}(p||q)$

- Some methods, e.g., Expectation Propagation (EP), can avoid this behavior

ELBO for Model Selection

- Recall that ELBO is a <u>lower bound</u> on log of model evidence $\log p(\boldsymbol{X}|\boldsymbol{m})$
- Can compute ELBO for each model $\boldsymbol{m}$ and choose the one with largest ELBO

Plot of the variational lower bound $\mathcal{L}$ versus the number K of components in the Gaussian mixture model, for the Old Faithful data, showing a distinct peak at $K = 2$ components. For each value of K, the model is trained from 100 different random starts, and the results shown as '+' symbols plotted with small random horizontal perturbations so that they can be distinguished. Note that some solutions find suboptimal local maxima, but that this happens infrequently.



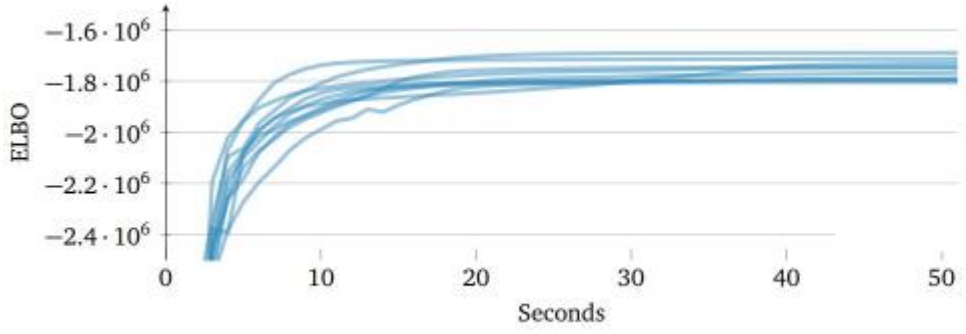
Each value of K represents a different model

Can thus measure the trade-off between model's fit to the data vs the model's complexity

- For GMM, unlike likelihood, ELBO doesn't monotonically increase with K
  - Also true for other models; increasing complexity increases likelihood monotonically but not ELBO
- Some criticism since we are using a lower-bound but often works well in practice

VI and Convergence

- VI is guaranteed to converge to a local optima (just like EM)
- Therefore proper initialization is important (just like EM)
  - Can sometimes run multiple times with different initializations and choose the best run



Different initializations may lead to different optima

- ELBO increases monotonically with iterations
  - Can thus monitor the ELBO to assess convergence

Variational Inference and Expectation Maximization

- VI can be seen as a generalization of the EM algorithm
- In VI, there is no distinction between parameters $\boldsymbol{\Theta}$ and latent variables $\boldsymbol{Z}$
  - Also recall that EM finds CP of $\boldsymbol{Z}$ and point estimate for $\boldsymbol{\Theta}$
  - VI treats all unknowns identically and infers posterior for all
- VI can be used within an EM algorithm if the E step is intractable
- E step is intractable if the CP of latent variables given params is intractable
- This version of EM is known as Variational EM (VEM)
- If we only care about point estimates of the parameters, VEM is widely used if the CP of latent variables is intractable

Scalable VI using Mean-field

+

Local Conjugacy

+

Stochastic Optimization

(Stochastic Variational Inference)

Stochastic Variational Inference (SVI)

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

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

- 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 $\boldsymbol{\epsilon}_i = (i+1)^{-\kappa}$ where $\boldsymbol{\kappa} \in (0.5, 1]$
  3. Choose a data point $\boldsymbol{n}$ uniformly randomly, i.e., $\boldsymbol{n} \sim \text{Uniform}(1,2,\ldots,N)$ Assuming minibatch size = 1
  4. Compute local var. param $\boldsymbol{\phi}_n$ for data point $\boldsymbol{x}_n$ as $\phi_n = \mathbb{E}_{\lambda^{(i-1)}}\left[\eta(\boldsymbol{x}_n, \boldsymbol{\beta})\right]$
  5. Update $\boldsymbol{\lambda}$ as $\lambda^{(i)} = (1-\epsilon_i)\lambda^{(i-1)} + \epsilon_i \lambda_n$ where $\lambda_n = [\alpha_1 + \mathbb{E}_{\phi_n}[t(\boldsymbol{x}_n, \boldsymbol{z}_n)], \alpha_2 + 1]^\top = \mathbb{E}_{\phi_n}[\hat{\boldsymbol{\alpha}}(\boldsymbol{x}_n, \boldsymbol{z}_n)]$
  6. Set $\boldsymbol{i} = \boldsymbol{i} + 1$. 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''

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

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

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

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

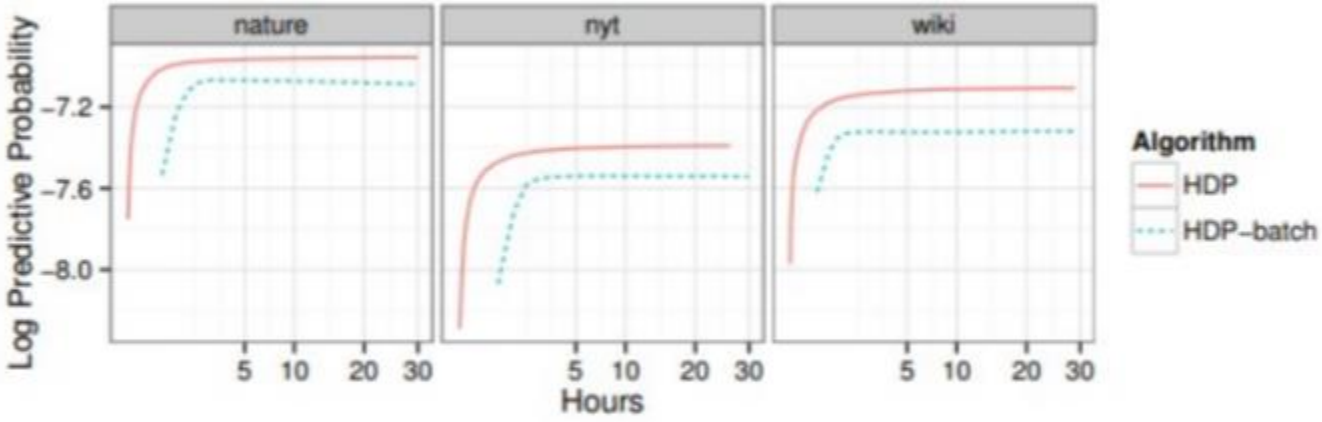
Global var. param computed on this minibatch

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

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
  - Some model-specific tricks
  - Black-box VI (BBVI) for general purpose VI
  - Reparametrization Trick for general purpose VI
- Other recent advances in VI
  - Automatic Differentiation VI
  - Amortized VI
  - Structured VI