10-708: Probabilistic Graphical Models 10-708, Spring 2018

16: Mean Field (Contd), Stochastic Gradient Descend, SVI, and Scalability

Lecturer: Kayhan Batmanghelich *Scribes: Yuanyuan Feng, Jiatian Wu*

1 Recap

Variational principle: general family of methods for approximating complicated densities by a simpler class of densities.

1.1 Evidence Lower Bound (ELBO)

$$q^*(z) = \arg\min_{q(z)} KL(q(z)||p(z||\mathcal{D})) \tag{1}$$

$$KL(q(z)||p(z||\mathcal{D})) = \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|\mathcal{D})] \tag{2}$$

$$KL(q(z)||p(z||\mathcal{D})) = \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,\mathcal{D})] + \log p(\mathcal{D}) \tag{3}$$

And,

$$ELBO(q) = \mathbb{E}_q[\log p(z,\mathcal{D})] - \mathbb{E}_q[\log q(z)] \tag{4}$$

1.2 Interpreting the Lower Bound (ELBO)

$$\mathcal{F}(y,q) = \mathbb{E}_{q(z)}[\log p(y|x)] - KL[q(z)||p(z)] \tag{5}$$

- y is data
- Approximate posterior distribution q(z): Best match to true posterior $p(z|y)$, one of the unknown inferential quantities of interest to us.
- $\mathbb{E}_{q(z)}[\log p(y|x)]$ is reconstruction cost, the expected log-likelihood measure how well samples from $q(z)$ are able to explain the data y.
- $KL[q(z)||p(z)]$ is penalty, ensures the the explanation of the data $q(z)$ doesnt deviate too far from your beliefs $p(z)$. A mechanism for realising Okhams razor.

1.3 Mean Field Approach

Mean Field Approach assumes the posterior is fully factorizable.

$$q(x;\phi) = \prod_i q_i(x_i;\phi_i) \tag{6}$$

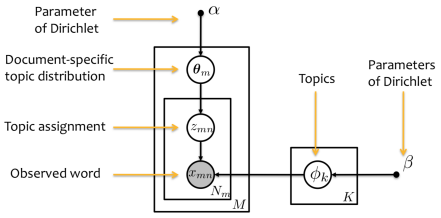

Figure 1: LDA

Let's focus on q_j, holding all other terms constant,

$$L(q_j) = \sum_x \prod_i q_i(x)[\log \tilde{p}(x) - \sum_k \log q_k(x_k)] \tag{7}$$

$$= \sum_{x_j} \log f_j(x_j) - \sum_{x_j} q_j(x_j) \log q_j(x_j) + const \tag{8}$$

$$\tag{9}$$

where

$$\log f_j(x_j) = \sum_{x_j} q_j(x_j) \sum_{x_j} \prod_{i \neq j} q_i(x_i) \log \tilde{p}(x) \tag{10}$$

thus,

$$L(q_j) = \mathbb{E}_{q_j}[\mathbb{E}_{q_j}[\log \tilde{p}(x)]] + H(q_j) \tag{11}$$

$$\frac{\delta L(q_j)}{\delta q_j} = \mathbb{E}_{q_j}[\log \tilde{p}(x)] - \log_{q_j} -1 = 0 \tag{12}$$

$$\tag{13}$$

Finally we can get

$$g_j^* \propto exp \mathbb{E}_{q_j}[\log \tilde{p}(x)] \tag{14}$$

2 LDA

Latent Dirichlet allocation (LDA) is a conditionally conjugate topic model. It treats documents as containing multiple topics, where a topic is a distribution over words in a vocabulary. Let us first refer to Figure 1 to get a general idea about this model.

Based on the graph, we could write the joint distribution as

$$p(\cdot) = \Pi_m^M p(\theta_m|\alpha) \Pi_n^{N_m} p(x_{mn}|z_{mn}, \{\phi_k\}_{k=1}^K) p(z_{mn}|\theta_m) \Pi_k^K p(\phi_k|\beta), \tag{15}$$

where $\{\phi_k\}, \{theta_m\}, \{z_{mn}\}$ are latent variables. Also we could write the posterior distribution as

$$q(\cdot) = \Pi_k p(\phi_k) \Pi_m p(\theta_m) \Pi_n p(z_{mn}). \tag{16}$$

Now let us study the updates. First of all, we have

$$q(\theta_m) \propto \exp\left[\mathbb{E}_{\Pi_n q(z_{mn})} \log p(\theta_m|\alpha) + \sum_n \log p(z_{mn|\theta_m})\right].$$

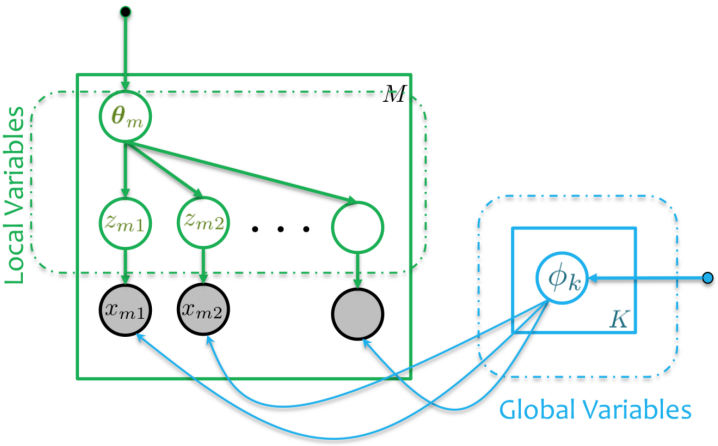

Figure 2: Re-draw LDA model

In LDA:

$$\textit{Dirichlet}: \qquad p(\theta_m|\alpha) \propto \exp\Big[\sum_k (\alpha_k - 1)\log\theta_{mk}\Big],$$
$$\textit{Categorical}: p(z_{mn}|\theta_m) \propto \exp\Big[\sum_k I(z_{mn=k})\log\theta_{mk}\Big].$$

Combining those two above, we further get

$$q(\theta_m) \propto \exp\Big[\sum_k \Big(\sum_n q(z_{mn}=k) + \alpha_k - 1\Big)\log\theta_{mn}\Big], \tag{17}$$

which is also a Dirichlet.

We could also re-draw LDA model as Figure 2, in which it clearly shows local and global variables.

Moreover, we generate such model in Figure 3

Based on the graph, we have a few observations.

$$p(x, z, \beta|\alpha) = p(\beta|\alpha)\Pi_n p(x_n, z_n|\beta) \tag{18}$$
$$p(x_n, z_n|x_{n-1}, z_{-n}, \beta, \alpha) = p(x_n, z_n|\beta, \alpha). \tag{19}$$

It is reasonable to give two further assumptions:

- Exponential family,
- conjugacy.

To be specific, we assume

$$p(\beta|x, z, \alpha) = h(\beta)\exp\{\eta_g(x, z, \alpha)^T t(\beta) - a_g(\eta_g(x, z, \alpha))\}, \tag{20}$$
$$p(z_{nj}|x_n, z_{n,-j}, \beta) = h(z_{nj})\exp\{\eta_l(x_n, z_{n,-j}, \beta)^T t(z_{nj}) - \alpha_l(\eta_l(x_n, z_{n,-j}, \beta))\}. \tag{21}$$

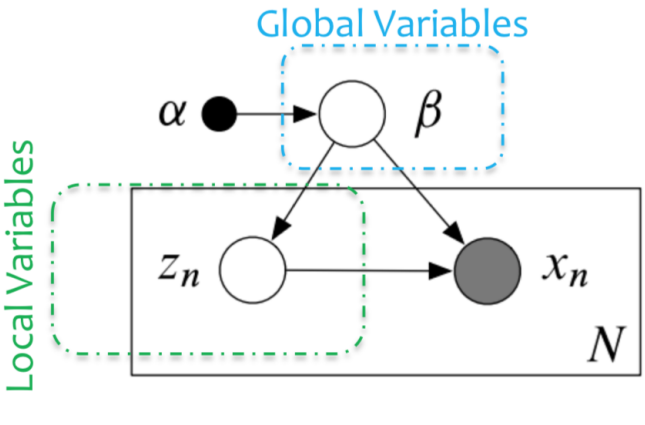

Figure 3: Generate LDA model

Hence we get the approximate posterior (mean field):

$$q(z, \beta) = q(\beta|\lambda)\Pi_n\Pi_j q(z_{nj}|\phi_{nj}),$$

where λ are global variables. The algorithm for global parameters' updates can be written as

- Initialize λ^0 randomly.
- **Repeat**
- **for** each local variational parameter ϕ_{nj} **do**
- Update ϕ_{nj}, $\phi_{nj}^t = \mathbb{E}_{q^{t-1}}[\eta_{lj}(x_n, z_{n,-j}, \beta)]$.
- **end for**
- Update the global variational parameters, $\lambda^t = \mathbb{E}_{q^t}[\eta_g(z_{1:N}, x_{1:N})]$.
- **until** the ELBO converges

3 Optimization View

3.1 Decoder Encoder View

Instead of making assumption that the posterior has closed form, we can switch view and take gradient go on. Recall the original problem:

$$\max_{\phi,\theta} F(y, q_\phi) = \mathbb{E}_{q_\phi(z)}[\log p_\theta(y|z)] - KL[q_\phi(z)||p(z)] \tag{22}$$

In an encoder decoder view, encoder is the variational distribution $q_\phi(z|y)$ and decoder is likelihood $p_\theta(y|z)$, where z is all latent variables that encode data to learn dimensional space, and posterior decode to reconstruct. Thus, recall our original problem, we can view the first term as reconstruction as data code-length, and the second term as hypothesis code.

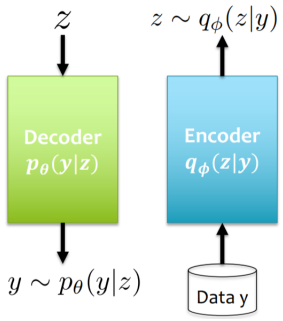

Figure 4: Encoder decoder view

Alternative optimization for the variational parameters and then model parameters (VEM) follows two modules:

- **Repeat:**
- **E-Step:**
- **For i=1,...,N**
- $\phi_n \propto \nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(y_n|z_n)] - \nabla_\phi KL[q_\phi(z_n)||p(z_n)]$
- **M-Step:**
- $\theta \propto \frac{1}{N}\sum_n \mathbb{E}_{q_\phi(z)}[\nabla_\phi \log p_\theta(y_n|z_n)]$

To compute $\nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(y_n|z_n)]$, we can follow: suppose $f(z) = \log p_\theta(y_n|z_n)$,

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(y_n|z_n)] = \nabla_\phi \mathbb{E}_{q_\phi(z)}[f(z)] \tag{23}$$

$$\int f(z)\nabla_\phi q_(z) d_z = \int f(z)\frac{\nabla_\phi q_{(z)}}{q_\phi(z)} q_\phi(z) d_z \tag{24}$$

$$\int f(z)\frac{\nabla_\phi q_{(z)}}{q_\phi(z)} q_\phi(z) d_z = \mathbb{E}_{q_\phi(z)}[f(z)\nabla_\phi \log q_\phi(z)] \tag{25}$$

3.2 Controlling the Variance

Monte Carlo gradient approximation suffers from variance. To deal with problem, we introduce the general idea control variates. Control variates are random quantities that are used to reduce the variance of a statistical estimator without introducing any bias by injecting information into the estimator. Lets say you want to estimate $\mathbb{E}[\hat{g}] = g$,the idea is to use another random variable $\tilde{g}$ with known mean $\tilde{m}$ that is correlated with $\hat{g}$.

$$g^{cv} = \hat{g} - \dot{c}(\tilde{g} - \tilde{m}) \tag{26}$$

$$\mathbb{V}(g^{(cv)}) = \mathbb{E}[(\hat{g} - \dot{\tilde{c}g})^2] - \mathbb{E}[\hat{g}^2] \tag{27}$$

$$= \mathbb{E}[\hat{g}^2 + c^2\tilde{g}^2 - 2c\hat{g}\tilde{g}] - \mathbb{E}[\hat{g}]^2 \tag{28}$$

$$= \mathbb{E}[\hat{g}^2] + c^2\mathbb{E}[\tilde{g}^2] - 2c\mathbb{E}[\hat{g}\tilde{g}] - \mathbb{E}[\hat{g}]^2 \tag{29}$$

suppose $c^* = \frac{\mathbb{E}[\hat{g}\tilde{g}]}{\mathbb{E}[\tilde{g}^2]} = \frac{\mathbb{C}(\hat{g},\tilde{g})}{\mathbb{V}(\tilde{g})}$,

$$\mathbb{V}(g^{(cv)}) = (1-\rho^2)\mathbb{V}(\hat{g}) \tag{30}$$

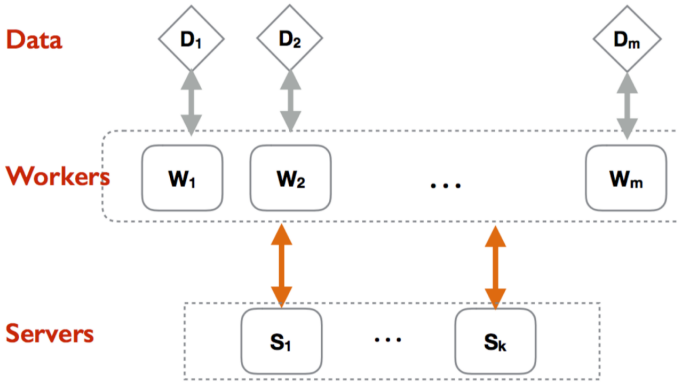

Figure 5: D-SGD framework

3.3 Scalability of Variational EM

Back to the above Variational EM problem, to scale to large numbers of data, we can use stochastic gradient and do parallelization.

3.3.1 Stochastic Gradient Descent

Recall the algorithm of variational EM:

- **Repeat:**
- **E-Step:**
- **For i=1,...,N**
- $\phi_n \propto \nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(y_n|z_n)] - \nabla_\phi KL[q_\phi(z_n)||p(z_n)]$
- **M-Step:**
- $\theta \propto \frac{1}{N}\sum_n \mathbb{E}_{q_\phi(z)}[\nabla_\phi \log p_\theta(y_n|z_n)]$

We can implement SGD, where N is a mini-batch-sampled with replacement from the full data set or received online.It is scalable because it only need to operate on a small batch at a time, thus can operate on large data.

3.3.2 Parallelization

D-SGD is a widely used(centralized) design choice to do parallelization. As shown in Figure 5, worker nodes solve compute intensive subproblems and servers perform simple aggregation.

3.3.3 Note on Implementation

- Stochastic gradient descent and other preconditioned optimization.

- Probabilistic models are modular, can easily be combined.
- Same code can run on both GPUs or on distributed clusters.

4 Summary

Advantages of VI:

- Applicable to almost all probabilistic models: non-linear, non-conjugate, high-dimensional, directed and undirected.
- Can be faster to converge than competing methods.
- Easy convergence assessment.
- Numerically stable.
- Can be used on modern computing architectures (CPUs and GPUs).
- Principled and scalable approach for model selection.

Disadvantages of VI:

- An approximate posterior only - not always
- Difficulty in optimisation can get stuck in guaranteed to find exact posterior in the limit. local minima.
- Typically under-estimates the variance of the posterior and can bias maximum likelihood parameter estimates.
- Limited theory and guarantees for variational methods.

Mean field vs LBP:

- LBP minimizes the Bethe energy while MF maximizes the ELBO.
- LBP is exact for trees whereas MF is not, suggesting LBP will in general.
- LBP optimizes over node and edge marginals, whereas nave MF only optimizes over node marginals, again suggesting LBP will be more accurate.
- MF objective has many more local optima than the LBP objective, so optimizing the MF objective seems to be harder.
- MF tends to be more overconfident than BP
- The advantage of MF is that it gives a lower bound on the partition function while for LBP we dont know the relationship.
- MF is easier to extend to other distributions besides discrete and Gaussian.