10-708: Probabilistic Graphical Models 10-708, Spring 2014

15: Mean Field Approximation and Topic Models

Lecturer: Eric P. Xing Scribes: Jingwei Shen (Mean Field Approximation), Jiwei Li (Topic Models)

1 Mean Field Approximation

1.1 Notations and Exponential Family

Recall that many density functions can be written in the exponential family form :

$$p_\theta(x_1, \cdots, x_m) = \exp(\theta^T \phi(x) - A(\theta))$$

Where θ is called the canonical parameters, $\phi(x)$ is the sufficient statistics of $x_1, \cdots, x_m$, and $A(\theta)$ is the log partition function. We often require that $A(\theta) < +\infty$ and the space of such θ is called effective canonical parameters :

$$\Omega := \{\theta \in \mathbb{R}^d | A(\theta) < +\infty\}$$

The mean parameter μ_α associated with a sufficient ϕ_α is defined by the expectation

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \text{for } \alpha \in \mathcal{I}$$

We then define the set

$$\mathcal{M} := \{\mu \in \mathbb{R}^d | \exists p \ s.t. \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha, \forall \alpha \in \mathcal{I}\}$$

corresponding to all realizable mean parameters. Further, for an exponential family with sufficient statistics ϕ defined on graph G, the set of realizable mean parameter set is :

$$\mathcal{M}(G; \phi) := \{\mu \in \mathbb{R}^d | \exists p \ s.t. \mathbb{E}_p[\phi(X)] = \mu\}$$

More generally, consider an exponential family with a collection $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ of sufficient statistics associated with the cliques of $G = (V, E)$. Given a subgraph F, let $\mathcal{I}(F)$ be the subset of sufficient statistics associated with subgraph F. Then the set of all distributions associated with F is a sub-family of full ϕ-exponential family. It is parameterized by the subspace of canonical parameters:

$$\Omega(F) := \{\theta \in \Omega | \theta_\alpha = 0, \forall \alpha \in \mathcal{I} - \ \mathcal{I}(F)\}$$

1.2 Mean Field Method

The exact variational formulation of log partition function is

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^T \mu - A^*(\mu)\}$$

where $\mathcal{M}$ is the marginal polytope which is difficult to characterize and A^* is the conjugate dual function of $A(\theta)$.

Mean field method uses non-convex inner product and exact form of entropy. Instead of inference under $\mathcal{M}$, mean field method uses a tractable subgraph.

For a given tractable subgraph F, a subset of canonical parameters is

$$\mathcal{M}(F;\phi) := \{\tau \in \mathbb{R}^d | \tau = \mathbb{E}_\theta[\phi(X)] \text{ for some } \theta \in \Omega(F)\}$$

Since F is a subgraph, $\mathcal{M}(F;\phi) \subset \mathcal{M}(G,\phi)$, which is called an inner approximation. Then mean field method solves the relaxed optimization problem :

$$\max_{\tau \in \mathcal{M}_F(G)} \{\tau^T\theta - A_F^*(\tau)\}$$

Here, $A_F^* = A^*|\mathcal{M}_F(G)$ is the exact dual function restricted to $\mathcal{M}_F(G)$, ϕ is the set of potentials assigned to the graph.

1.3 Naive Mean Field for Ising model

The joint probability of Ising model can be represented as

$$p(x) \propto \exp(\sum_{s\in V} x_s\theta_s + \sum_{(s,t)\in E} x_s x_t \theta_{st})$$

Then mean parameters we are interested in are :

$$\mu_s = E_p(x_s) = P(X_s = 1), \forall s \in V$$
$$\mu_{st} = E_p(x_s x_t) = P[(x_s, x_t) = (1,1)], \forall (s,t) \in E$$

It is difficult for inference in Ising model since there are many loops in the model. We consider a fully disconnected graph where there are no edges connecting each pair of nodes. For fully disconnected graph F,

$$\mathcal{M}_F(G) := \{\tau \in \mathbb{R}^{|V|+|E|} | 0 \le \tau_S \le 1, \forall s \in V, \tau_{st} = \tau_s\tau_t, \forall (s,t) \in E\}$$

The dual decomposes into sum, one for each node

$$A_F^*(\tau) = \sum_{s\in V}[\tau_s \log \tau_s + (1-\tau_s)\log(1-\tau_s)]$$

Then the relaxed optimization problem becomes

$$\max_{\mu\in[0,1]^m} \{\sum_{s\in V}\mu_s\theta_s + \sum_{(s,t)\in E}\theta_{st}\mu_s\mu_t + \sum_{s\in V} H_s(\mu_s)\}$$

Taking gradient w.r.t μ_s and let it be zero we have

$$\theta_s + \sum_{(s,t)\in E}\theta_{st}\mu_t + \log\mu_s - \log(1-\mu_s) = 0$$

The update rule is

$$\mu_s \leftarrow \sigma(\theta_s + \sum_{(s,t)\in E}\theta_{st}\mu_t)$$

where $\sigma(.)$ is the sigmoid function.

1.4 Geometry of Mean Field

Mean field optimization is always non-convex for any exponential family in which the state space $\mathcal{X}^m$ is finite. The marginal polytope $M(G)$ is a convex hull. If $M_F(G)$ is a strictly subset then it must be non-convex since it contains all the extreme points.

For example, in the two-node Ising model,

$$\mathcal{M}_F(G) = \{0 \leq \tau_1 \leq 1, 0 \leq \tau_2 \leq 1 \tau_{12} = \tau_1 \tau_2\}$$

We can easily check that it is not a convex set.

1.5 Cluster-based Approximation for the Gibbs Free Energy

When the inference for the entire graph is intractable, we divide the graph into small clusters which can be inferred by exact inference algorithms each.

Given a disjoint clustering , $\{C_1, C_2, \cdots, C_l\}$, of all variables. Let

$$q(X) = \prod_i q_i(X_{C_i})$$

The mean field free energy is

$$G_{MF} = \sum_i \sum_{X_{C_i}} \prod q_i(X_{C_i}) E(X_{C_i}) + \sum_i \sum_{X_{C_i}} q(X_{C_i}) \log q_i(X_{C_i})$$

will never equal to the exact Gibbs free energy no matter how what clustering is used, however it always defines a lower bound of the likelihood.

1.6 Generalized Mean Field Algorithm

Theorem: The optimum GMF approximation to the marginal cluster marginal is isomorphic to the cluster posterior of the original distribution given internal evidence and its generalized mean fields.

Theorem: The GMF algorithm is guaranteed to converge to a local optimum and provides a lower bound for the likelihood of the evidence or the partition function in the model.

The GMF algorithm iterates over each clique q_i for the optimization.

The accuracy increases as the size of clusters grows while the computation cost for each cluster also increases. The extreme case is that there is only one cluster : the original graph, then it is exactly the true inference but it is often intractable. So there is a trade off between the computation cost and the inference accuracy.

1.7 The Naive Mean Field Approximation

The idea is to approximate $p(X)$ by fully factorized $q(X) = P_i q_i(X_i)$. For example, for Boltzmann distribution, it is

$$p(X) = \exp(\sum_{i<j} q_{ij} X_i X_j + q_{iO} X_i)/Z$$

The mean field equation is

$$\begin{aligned} q_i(X_i) &= \exp(\theta_{iO} X_i + \sum_{j \in \mathcal{N}_i} \theta_{ij} X_i < X_j >_{q_j} + A_i) \\ &= p(X_i | \{< X_j >_{q_j} : j \in \mathcal{N}_i\} \end{aligned}$$

where $< X_j >_{q_j}$ resembles a message sent from node j to i, $\mathcal{N}_i$ is the neighbor of X_i.

2 Probabilistic Topic Models

2.1 Latent Dirichlet Allocation (LDA)

In LDA [1], each document d is represented as a sequence of its containing word $d = \{w_1, w_2, ..., w_{N_d}\}$, where N_d denotes the number of words in d. The word w is defined to be an item from a vocabulary indexed by $\{1, 2, ..., V\}$, where V is the vocabulary size. w can also be represented by an $1 \times V$ vector with its correspondent element 1 and others zero.

In topic models, each document d is represented as a mixture of topics, characterized by the document-specific vector θ_d. It is a bit tricky to interpret the concept of *topic*, where here are to be viewed as particular distributions over vocabularies, denoted as β. For example, the topic of sports[1] tend to give higher probability to sports related words than entertainment related ones, or in other words, sports words are more likely generated from the topic of sports. Table 1 gives an illustration of what *topics* are like in topic models.

	basketball	ball	score	rebound	...	movie	spiderman	actor
topic-sports	0.08	0.09	0.05	0.06	...	0.00001	0.00001	0.00002
topic-entertainments	0.0002	0.00006	0.0002	0.00004	...	0.07	0.06	0.12

Table 1: Illustration of *topics* in topic models. The value corresponds to the probability that particular word is generated by the topic.

LDA is a generative model and its generative story can be interpreted as follows: when a writer wants to write something in document d, he has to first decide which topics he wishes to cover in this particular document, as he will choose from document-topic distribution θ_d. Specifically, the topic z will be chosen from the multinomial distribution $z \sim Multi(\theta)$. Once the topic z is settled, he would choose a word w to fill in the position from the topic distribution β_z. As we just discussed in Table 1, if the writer decides to write something about sports, words such as *basketball* and *rebound* are more likely to be chosen than *actor* or *spiderman*. Word is similarly chosen from the multinomial distribution $w \sim Multi(\beta_z)$. Such decision process iterates for each position until the end of the document. The generative story is given in Figure 2.

A Dirichlet prior is commonly given to θ_d, as $\theta_d \sim Dir(\alpha)$ for the facilitation of calculation due to the conjugate property of Dirichlet prior and multinomial distribution. Similarly, β follows the Dirichlet prior parameterized by η.

2.2 Variational Inference for LDA

In this subsection, we get down to the Variational Inference for LDA, the key point of which is trying the minimizing the KL divergence between the variational distribution $q(\theta, z|\gamma, \phi)$ and the actual posterior

[1] Topic models do not offer a name for the mined topics. These names are usually manually identified according to word distributions or top words.

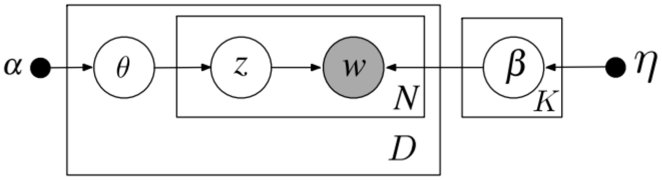

Figure 1: Graphical Model for LDA (taken from lecture15, page 41, Eric Xing).

1. For each document m
   Draw a document proportion vector $\theta_m|\alpha \sim Dir(\alpha)$
2. For each word in $w \in m$
   (a)draw topic assignment $z_w|\theta \sim Multi(\theta_{z_m})$
   (b)draw word $w|z_w, \beta \sim Multi(\beta_{z_w})$

Figure 2: Generative Story for LDA topic model.

distribution $p(\theta, z|w, \alpha, \beta)$, where γ and ϕ are variational parameters involved in q. Specifically, $q(\theta|\gamma)$ follows a Dirichlet distribution parameterized by γ and $q(z_n|\phi_n)$ is the multinomial distribution parameterized by ϕ_n. $q(\theta, z|\gamma, \phi)$ is factorized as follows:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_n q(z_n|\phi_n) \tag{1}$$

$$(\gamma^*, \phi^*) = \operatorname*{argmin}_{\gamma,\phi} D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) \tag{2}$$

$$\begin{aligned} & KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) = E_q(\theta, z|\gamma, \phi) \log \frac{q(\theta, z|\gamma, \phi)}{p(\theta, z|w, \alpha, \beta))} \\ & = E_q \log q(\theta, z|\gamma, \phi) - E_q p(\theta, z|w, \alpha, \beta) \\ & = E_q \log q(\theta, z|\gamma, \phi) - E_q p(\theta, z, w|\alpha, \beta) + E_q p(w|\alpha, \beta) \end{aligned} \tag{3}$$

Let $L(\gamma, \phi : \alpha, \beta) = -E_q \log q(\theta, z|\gamma, \phi) + E_q p(\theta, z, w|\alpha, \beta)$, we have

$$p(w|\alpha, \beta) = L(\gamma, \phi : \alpha, \beta) + KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) \tag{4}$$

So we have

$$\begin{aligned} (\gamma^*, \phi^*) &= \operatorname*{argmin}_{\gamma,\phi} KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) \\ &= \operatorname*{argmax}_{\gamma,\phi} L(\gamma, \phi : \alpha, \beta) \end{aligned} \tag{5}$$

$$L(\gamma, \phi : \alpha, \beta) = E_q[\log p(\theta|\alpha)] + E_q[\log(z|\theta)] + E_q[\log p(w|z, \beta)] - E_q[\log q(\theta)] - E_q[\log(q(z))] \tag{6}$$

The optimization of Equation 6 is performed in framework called Variational EM, which is so-called as the optimization algorithm maximizes a lower bound with respect to the variational parameters γ and ϕ in E step, and maximizes the lower bound with respect to the model parameters for fixed values of the variational parameters in M step. The algorithm is given in Figure 4 and the details can be found in [1].

```
E step: For each document $d$, find the optimizing values of the variational parameters $\gamma_d$ and $\phi_d^n$
S step: Maximize the lower bound with respect to $\alpha$ and $\beta$.
```

Figure 3: Variational Algorithm for LDA.

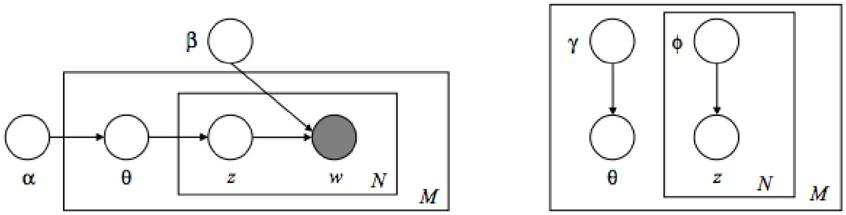

Figure 4: : (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA. Figures borrowed from [1].).

2.3 Gibbs Sampling for LDA

Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) simulation and yield relatively simple algorithms for approximate inference in LDA [2]. In Gibbs sampling for LDA, latent variables in the graphical model are sampled iteratively given the rest based on the conditional distribution. A more commonly applied approach is the collapsed Gibbs sampling, where we do not have to sample all parameters involved, as θ and β can be integrated out.

Let $\mathbf{z}$ denote the concatenation of z for all words and $\mathbf{z_{-n}}$ denotes the topic assignments of all words except w_n. The conditional probability that w_n is assigned to topic index k given all other variables is given by:

$$\begin{aligned} p(z_i = k|\mathbf{z_{-n}}, \mathbf{w}, \alpha, \eta) &\propto \frac{p(\mathbf{w}, \mathbf{z}|\alpha, \eta)}{p(\mathbf{w}, \mathbf{z_{-n}}|\alpha, \eta)} = \frac{p(\mathbf{w}|\mathbf{z}, \eta)}{p(\mathbf{w}|\mathbf{z_{-n}}, \eta)} \frac{p(\mathbf{z}|\alpha)}{p(\mathbf{z_{-n}}|\alpha)} \\ = \int p(\mathbf{z}|\theta) p(\theta|\alpha) d\theta &\propto \frac{n_k^w + \eta_k}{\sum_{k'} n_{k'}^w + \eta_k} (n_d^k + \alpha_k) \end{aligned} \tag{7}$$

where n_k^w denotes the number of times word w appearing in topic k. n_d^k denotes the number of words in document d assigned to topic k. The calculation of $p(\mathbf{z}|\alpha)$ is performed by integrating our parameter θ The details of computation can be found in [2]. The Gibbs Sampling algorithm is given in Figure 5

```
For each document $m$
    For each word $w \in m$
        sample topic $z_w$ according to Equation 7.
```

Figure 5: Gibbs Sampling for LDA.

The estimation for parameters β and θ is given by:

$$\begin{aligned} \theta_d^m &= \frac{n_d^k + \alpha_k}{\sum_{k'} n_d^{k'} + \alpha_{k'}} \\ \beta_k^w &= \frac{n_k^w + \beta_w}{\sum_w' n_k^{w'} + \beta_{w'}} \end{aligned} \tag{8}$$

References

[1] David Blei, Andrew Ng and Michael Jordan. Latent dirichlet allocation. *the Journal of machine Learning research.* 2003.

[2] Gregor Heinrich. Parameter estimation for text analysis. Technical report.