10-708 PGM (Spring 2020): Homework 2

Andrew ID: [your Andrew ID]
Name: [your first and last name]
Collaborators: [Andrew IDs of all collaborators, if any]

1 Variational Inference [40 points] (Junxian)

In this problem, we are going to work with approximate posterior inference via variational inference for a given topic model.

The standard Latent Dirichlet Allocation model only models the word co-occurrences, without considering temporal information, i.e. the time when a document is generated. However, a large number of subjects in documents change dramatically over time. It is important to interpret the topics in the context of the timestamps of the documents. To address how topics occur and shift over time, Topics on Time (TOT) model was proposed, by explicitly modeling of time jointly with word co-occurrence patterns (Wang and McCallum, 2006). The model is shown in Figure 1.

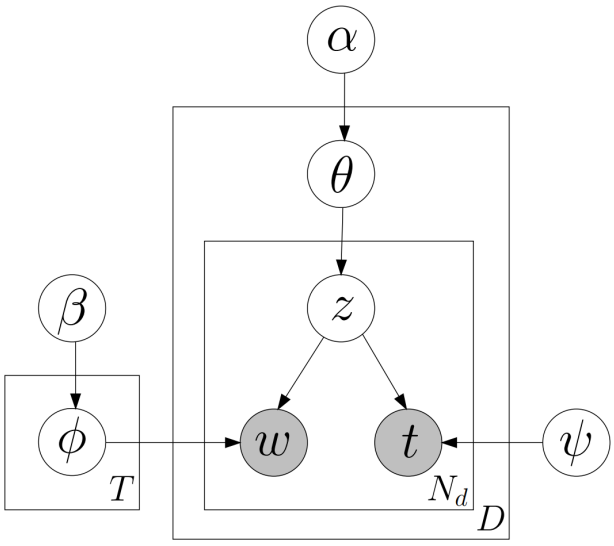

Figure 1: TOT Model

In the model, there are D documents. Each document d contains N_d words $w_{d1}, w_{d2}, ..., w_{dN_d}$. Each word w_{di} has a timestamp $t_{di} \in (0, 1)$, indicating when the document is generated in a relative time scale $(0, 1)$. All words in the same document have the same timestamp. There are K topics (also $T = K$ topics for the notation in the paper and Figure 1) in the document corpora. Each topic follows a multinomial distribution ϕ over the V words in the vocabulary. Each document follows a multinomial distribution θ over the K topics. The prior distribution for ϕ and θ are Dirichlet distributions with parameters β and α respectively. For each topic k, the temporal occurrence follows a Beta distribution $Beta(\psi_{k1}, \psi_{k2})$, where $\psi_k = (\psi_{k1}, \psi_{k2})$ and we

use $\boldsymbol{\psi} \in \mathbb{R}_+^{K\times 2}$ to denote ψ_k for all topics. Each word w_{di} and its timestamp t_{di} are assumed to be generated from a topic, with a topic label $z_{di} \in \{1, ..., K\}$.

The generative process of this model is described as follows.

1. Draw K multinomials ϕ_k from a Dirichlet prior β, one for each topic k.
2. For each document d,
   - Draw a multinomial θ_d from a Dirichlet prior α;
   - For each word w_{di} in document d,
     (a) Sample a topic z_{di} from multinomial θ_d;
     (b) Sample a word w_{di} from multinomial $\phi_{z_{di}}$;
     (c) Sample a timestamp t_{di} from Beta $\psi_{z_{di}}$.

We use variational EM to approximate the posterior of latent variables and learn model parameters. To do this, a mean field variational distribution needs to be defined, which is parameterized by some parameters called variational parameters. The variational EM algorithm iteratively performs two steps: 1) in the E step, variational parameters are updated; 2) in the M step, model parameters are optimized. Same as in the paper, we consider α and β are predefined fixed hyperparameters with no need to update. Therefore, in M step, only the other model parameters are optimized. The pseudo-code for the proposed algorithm is shown in Algorithm 1.

Algorithm 1 Pseudo-code of variational EM algorithm for TOT model

Input: Observations, Topic number K, MaxIter, and other optional parameters
Output: Posterior distributions for latent variables, optimized model parameters
Initialize parameters;
Compute and record ELBO with initial parameters
for $k \leftarrow 1$ to $MaxIter$ **do**
    Update variational variables ▷ Stage 1: E-Step
    Update $\boldsymbol{\psi}$ with projected Newton method ▷ Stage 2: M-Step
    Compute and record ELBO
end for

In the TOT model, $\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\phi}$ are latent variables and $\boldsymbol{\psi}$ is the model parameter to be learned. As a start, we use mean-field variational inference and the variational distribution has the form:

$$q(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \prod_{k=1}^{K} q(\phi_k|\lambda_k) \prod_{d=1}^{D} \big[q(\theta_d|\gamma_d) \prod_{n=1}^{N_d} q(z_{dn}|\pi_{dn})\big], \tag{1}$$

where $\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\pi}$ are variational parameters that need to be updated in E-step.

Next, we write out the joint distribution of latent and observed variables:

$$p(\mathbf{x}, \mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z}|\alpha, \beta, \boldsymbol{\psi}) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} [p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(x_{dn}|z_{dn}, \boldsymbol{\phi}) p(t_{dn}|z_{dn}, \boldsymbol{\psi})] \tag{2}$$

Given Eq. 1 and 2, we can write out ELBO with:

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z})}[\log p(\mathbf{x}, \mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z}) - \log q(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z})]. \tag{3}$$

Variational EM basically maximizes ELBO w.r.t. variational parameters and model parameters in E- and M-step respectively.

Questions:

1. **[10 points]** Update variational parameters
   Derive the update equations of variational parameters, and also specify their distributions. Here you can directly use the conclusion below for the derivation.

$$q_j^* \propto \exp\{\mathbb{E}_{q_{-j}}[\log p(\mathbf{x}, \mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z})]\},$$

   where $\mathbb{E}_{q_{-j}}$ denotes expectation over all latent variables excluding variable j.

2. **[10 points]**Update model parameters
   Derive the update equations of model parameters, as mentioned before, there is no need to update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For the updating rule of $\boldsymbol{\psi}$, please be careful that $\boldsymbol{\psi}$ should be constrained as positive. Hint: For a problem with positive solution ($x > 0$), a projected Newton method could be applied:

$$y = (x - H^{-1}g)_+$$

$$x^+ = x + \lambda(y - x)$$

   where x is the current variable, y is the projected update, g and H are gradient and Hessian matrix respectively, λ is the step size and x^+ is the updated variable. $(\cdot)_+$ is defined as $s_+ := \max(0, s)$.

3. **[20 points]** Detailed variational lower bound
   Based on the variational distributions, expand Eq. 3 to obtain detailed variational lower bound. The result should be as specific as possible, that is, it can be directly used in the implementation.

Hint: the problem is designed based on the paper (Wang and McCallum, 2006). In the paper, Gibbs sampling was used for posterior inference, and here we are working with variational inference. You may gain better understanding of the model and get some ideas of how to solve the problem by reading the paper.

2 Monte Carlo [20 points] (Ben)

Given a random distribution $p(x)$ on $x = [x_1, ..., x_D]^T \in \mathbb{R}^D$. Suppose we want to perform inference $\mathbb{E}_{p(x)}[f(x)]$ using importance sampling, with $q(x)$ as the proposal distribution. According to importance sampling, we draw L i.i.d. samples $x^{(1)}, ..., x^{(L)}$ from $q(x)$, and we have

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{\sum_{i=1}^{L} u_i} \sum_{i=1}^{L} f(x^{(i)}) u_i$$

where the (unnormalized) importance weights $u_i = \frac{p(x^{(i)})}{q(x^{(i)})}$.

1. **[5 points]** Find the mean and variance of the unnormalized importance weights $\mathbb{E}_{q(x)}[u_i]$ and $\mathrm{Var}_{q(x)}[u_i]$.

2. **[5 points]** Prove the following lemma: $\mathbb{E}_{p(x)}\left[\frac{p(x)}{q(x)}\right] \geq 1$, and the equality holds only when $p = q$.

3. **[9 points]** A measure of the variability of two components in vector $u = [u_1, ..., u_L]^T$ is given by $\mathbb{E}_{q(x)}\left[(u_i - u_j)^2\right]$. Assume that both p and q can be factorized, i.e. $p(x) = \prod_{i=1}^{D} p_i(x_i)$, and $q(x) = \prod_{i=1}^{D} q_i(x_i)$. Show that $\mathbb{E}_{q(x)}\left[(u_i - u_j)^2\right]$ has exponential growth with respect to D.

4. **[1 points]** Use the conclusion in (c) to explain why the standard importance sampling does not scale well with dimensionality and would blow up in high-dimensional cases.

3 MCMC [40 points] (Xiang)

3.1 Multiple Choice [10 points]

1. **[5 points]** Which of the following statements is true for the acceptance probability

$$A(x'|x) = \min(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)})$$

of Metropolis-Hastings algorithm?

   A. It satisfies detailed balance.

   B. We can just evaluate $P(x')$ and $P(x)$ up to a normalization constant.

   C. It ensures that the MH algorithm eventually converges to the true distribution.

   D. All of the above.

2. **[5 points]** Which of the following statements is true for Hamiltonian Monte Carlo in comparison with vanilla MCMC?

   A. It can improve acceptance rate and give better mixing.

   B. Stochastic variants can be used to improve performance in large dataset scenarios.

   C. It may not be used for discrete variable.

   D. All of the above.

3.2 Modeling with Markov Chain Monte Carlo [30 points]

We are going to use the data from the 2013-2014 Premier League PL1 to build a predictive model on the number of goals scored in a single game by the two opponents. Bayesian hierarchical model is a good candidate for this kind of modeling task. We model each team's strength (both attacking and defending) as latent variables. Then in each game, the goals scored by the home team is a random variable conditioned on the attacking strength of the home team and the defending strength of the away team. Similarly, the goals scored by the away team is a random variable conditioned on the attack strength of the away team and the defense strength of the home team. Therefore, the distribution of the scoreline of a specific game is dependent on the relative strength between the home team A and the away team B, which also depends on the relative strength between those teams with their other opponents.

Table 1: 2013-2014 Premier League teams

Index	0	1	2	3	4
Team	Arsenal	Aston Villa	Cardiff City	Chelsea	Crystal Palace
Index	5	6	7	8	9
Team	Everton	Fulham	Hull City	Liverpool	Manchester City
Index	10	11	12	13	14
Team	Manchester United	Newcastle United	Norwich City	Southampton	Stoke City
Index	15	16	17	18	19
Team	Sunderland	Swansea City	Tottenham Hotspur	West Bromwich Albion	West Ham United

Here we consider using the same model as described by Baio and Blangiardo (2010). The Premier League has 20 teams, and we index them as in Table 1. Each team would play 38 matches every season (playing each of the other 19 teams home and away), which totals 380 games in the entire season. For the g-th game, assume that the index of home team is $h(g)$ and the index of the away team is $a(g)$. the observed number of goals is:

$$y_{gj} \mid \theta_{gj} = \text{Poisson}(\theta_{gj})$$

where the $\theta = (\theta_{g1}, \theta_{g2})$ represent the scoring intensity in the g-th game for the team playing at home ($j = 1$) and away ($j = 2$), respectively. We put a log-linear model for the θs:

$$\log \theta_{g1} = home + att_{h(g)} - def_{a(g)}$$
$$\log \theta_{g2} = att_{a(g)} - def_{h(g)}$$

Note that team strength is broken into attacking and defending strength. And *home* represents home-team advantage, and in this model is assumed to be constant across teams. The prior on the home is a normal distribution

$$home \sim \mathcal{N}(0, \tau_0^{-1})$$

where the precision $\tau_0 = 0.0001$.

The team-specific attacking and defending effects are modeled as exchangeable:

$$att_t \sim \mathcal{N}(\mu_{att}, \tau_{att}^{-1})$$

$$def_t \sim \mathcal{N}(\mu_{def}, \tau_{def}^{-1})$$

We use conjugate priors as the hyper-priors on the attack and defense means and precisions:

$$\mu_{att} \sim \mathcal{N}(0, \tau_1^{-1})$$
$$\mu_{def} \sim \mathcal{N}(0, \tau_1^{-1})$$
$$\tau_{att} \sim \text{Gamma}(\alpha, \beta)$$
$$\tau_{def} \sim \text{Gamma}(\alpha, \beta)$$

where the precision $\tau_1 = 0.0001$, and we set parameters $\alpha = \beta = 0.1$.

This hierarchical Bayesian model can be represented using a directed acyclic graph as shown in Figure 2. Where the goals of each game $\mathbf{y} = \{y_{gj} | g = 0, ..., 379, j = 1, 2\}$ are 760 observed variables, and parameters $\boldsymbol{\theta} = (home, att_0, ..., att_{19}, def_0, ..., def_{19})$ and hyper-parameters $\boldsymbol{\eta} = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$ are unobserved variables that we need to make inference. To ensure identifiability, we enforce a corner constraint on the parameters (pinning one team's parameters to 0,0). Here we use the first team as reference and assign its attacking and defending strength to be 0:

$$att_0 = def_0 = 0$$

In this question, we want to estimate the posterior mean of the attacking and defending strength for each team, i.e. $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[att_i]$, $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[def_i]$, and $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[home]$.

1. **[10 points]** Find the joint likelihood $p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta})$.
2. **[10 points]** Write down the Metropolis-Hastings algorithm for sampling from posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y})$, and derive the acceptance function for a proposal distribution of your choice (e.g. isotropic Gaussian).

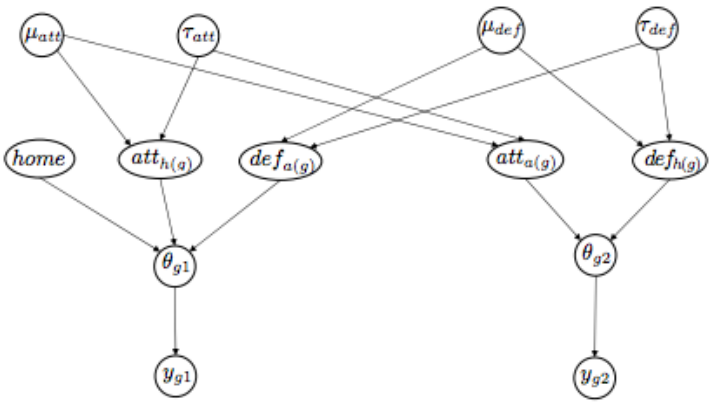

Figure 2: The DAG representation of the hierarchical Bayesian model

3. **[10 points]** Implement the M-H algorithm to inference the posterior distribution. The data can be found from `premier_league_2013_2014.dat`, which contains a 380×4 matrix. The first column is the number of goals y_{g1} scored by the home team, the second column is the number of goals y_{g2} scored by the away team, the third column is the index for the home team $h(g)$, and the fourth column is the index for the away team $a(g)$. Use isotropic Gaussian proposal distribution, $\mathcal{N}(0, \sigma^2 I)$ and 0 as the starting point. Run the MCMC chain for 5000 steps to burn in and then collect 5000 samples with t steps in between (i.e., run M-H for $5000t$ steps and collect only each t-th sample). This is called *thinning*, which reduces the autocorrelation of the MCMC samples introduced by the Markovian process. The parameter sets are $\sigma = 0.05$, and $t = 5$. Plot the trace plot of the burning phase and the MCMC samples for the latent variable *home* using the proposed distribution.

4. **[Bonus, 20 points]** Set the parameters as $\sigma = 0.005, 0.05, 0.5$ and $t = 1, 5, 20, 50$, and:
   - Plot the trace plot of the burning phase and the MCMC samples for the latent variable *home* using proposal distributions with different σ and t.
   - Estimate the rejection ratio for each parameter setting, report your results in a table.
   - Comment on the results. Which parameter setting worked the best for the algorithm?
   - Use the results from the optimal parameter setting
     - plot the posterior histogram of *home* from the MCMC samples
     - plot the estimated attacking strength $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[att_i]$ against the estimated defending strength $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[def_i]$ for each the team in one scatter plot. Please make sure to identify the team index of each point on your scatter plot.

You are NOT allowed to use any existing implementations of M-H in this problem. Please include all the required results (figures + tables) in your writeup PDF submission.

References

2013-14 premier league. https://en.wikipedia.org/wiki/2013%E2%80%9314_Premier_League. Accessed: 2017-03-31.

G. Baio and M. Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.

X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. 2006.