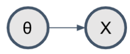
Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Ajita Shree
Roll Number: 20111262
Date: May 14, 2021

QUESTION 1

This is Bayes' rule!

Given observations $x_1, x_2, ...x_N$, drawn i.i.d from the observations model $p(x/\theta)$ and prior distribution $p(\theta)$



To Prove: solving the below equation is equivalent to finding the posterior distribution of **Eqn 1:** $\theta\ argmin_{q(\theta)} - \sum_{n \in N}[\int q(\theta) log(x_n/\theta) d\theta)] + KL(q(\theta||p(\theta)))$

- **Explanation:** The above objective function is ELBO expression, i.e. expected lower bound optimization.
- Maximizing ELBO will give an approximating distribution $q(\theta)$ which explains data well i.e. give it large probability i.e. large exp log-likelihood i.e. $E_q[log(p(X/\theta))]$
- KL term however, takes care of that $q(\theta)$ is close to the prior distribution $p(\theta)$, this will act as simple regularizer so that it doesn't over-fit the data much.
- MAP, posterior estimation of $\theta, p(\theta/x) \propto p(x/\theta)p(\theta)$
- Here also, there are two major objectives
  - Maximizing the likelihood of the data
  - Regularizer $p(\theta)$ will ensure there is not much overfitting.
- As we have seen, both approaches are for optimizing same objectives and hence, minimizing the -ve of ELBO is equivalent to maximizing the posterior prediction distribution.
- **Formal proof is as follows:** Let us optimize Eqn 1 wrt $q(\theta)$ and find out the optimal $q(\theta)$ value; Differentiating and equating it to 0 will give
- $d/d(q(\theta))\Big\{ - \sum_{n \in N} \Big[\int q(\theta) log(x_n/\theta) d\theta\Big] + \int q(\theta) log(q(\theta)/p(\theta/X)) d\theta\Big\}$
- $\sum_{n \in N} \int log(x_n/\theta) d\theta + \int log(q(\theta)/p(\theta)) = 0$
- $\int log(q(\theta/p(\theta))) = \sum_{n \in N} \int log(x_n/\theta) d\theta$
- Rearranging terms and removing integrals, we have
- $log(q(\theta)) = \sum_{n \in N} log(x_n/\theta) + log(p(\theta))$
- $q(\theta) = \prod_{n \in N} p(x_n/\theta).p(\theta)$
- **Hence proved, best possible value of $q(\theta)$ is nothing but the posterior distribution** $p(\theta/X)$

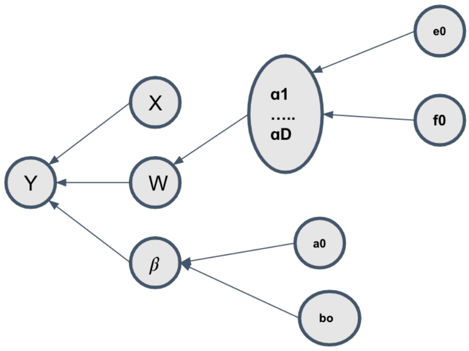
Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Ajita Shree
Roll Number: 20111262
Date: May 14, 2021

QUESTION 2

Mean Field VI for Sparse Bayesian Linear Regression

The probability distributions of all rvs and their respective graphical model is as follows.



- **Given** $y_n = N(y_n/w^T x_n, \beta^{-1})$
- $p(w) = N(w/0, diag(\alpha_1^{-1}, ...\alpha_D^{-1}))$
- $p(\beta) = Gamma(\beta/a_0, b_0)$
- $\alpha_d = Gamma(\alpha_d/e_0, f_0)$
- To Derive: $q(w, \beta, \alpha_1 ... \alpha_D) = q(w)q(\beta)q(\alpha_1)...q(\alpha_D) = p(w, \beta, \alpha_1, \alpha_2, ...\alpha_D/Y, X)$
- $Note : Gamma(\eta/\tau_1, \tau_2) = \frac{\tau_2^{\tau_1}}{T(\tau_1)}\eta^{\tau_1 - 1}exp(-\tau_2\eta)$
- **Solution:**
- **Joint distribution** $p(y, w, \beta, \alpha_1, \alpha_2, ...\alpha_D/X) = \prod_{n\in N} p(y_n/w^T x_n, \beta) * p(w/\alpha_1, ...\alpha_D) * p(\beta) * \prod_{d\in D} p(\alpha_D)$
- Taking log of joint distribution will give
- $log(p(y, w, \beta, \alpha_1, ...\alpha_D)/X) = \sum_{n\in N} log(p(y_n/w, x_n, \beta)) + log(p(w/\alpha_1...\alpha_D)) + log(\beta) + \sum_{d\in D} log(p(\alpha_d))$
- $log(p(y, w, \beta, \alpha_1, ...\alpha_D)/X) = \sum_{n\in N} logN(y_n/w^T x_n, \beta^{-1}) + logN(w/0, diag(\alpha_1^{-1}, ...\alpha_D^{-1})) + logGamma(\beta/a_0, b_0) + \sum_{d\in D} logGamma(\alpha_d/e_0, f_0)$
- Expanding the probability distributions, we will get
- $\sum_{n\in N} log(\sqrt{(\frac{\beta}{2*\pi}}exp(\frac{-\beta}{2} * (y_n - w^T x_n)^2))) + log(\sqrt{(\frac{\alpha_1...\alpha_D}{(2\pi)^D})}exp(\frac{-w^T\Sigma w}{2}))$

- $+log(\frac{b_0^{a_0}\beta^{a_0-1}}{T(a_0)}exp(-b_0\beta)) + \sum_{d\in D} log(\frac{f_0^{e_0}\alpha^{e_0-1}}{T(e_0)}exp(-f_0\alpha_d))$ where Σ is diagonal matrix with α_d at dth row, col.
- **Eq1: Joint dist** $log(p(y,w,\beta,\alpha_1,...\alpha_D)/X) \propto (N/2)log\beta - (\beta/2)\sum_{n\in N}(y_n - w^T x_n)^2 + (1/2)\sum_{d\in D} log\alpha_d - (1/2)w^T\Sigma w + (a_0-1)log\beta - b_0\beta + (e_0-1)\sum_{d\in D} log\alpha_d - f_0\sum_{d\in D}\alpha_d$
- **Note 1:** We know that standard mean field VI algorithm, the approximate distribution for an rv $z_j, q_j(z_j) \propto [E_{i!=j}(logp(X,Z))]$, where X is the true data and z's are the rvs on which data is dependent.
- Using the fact in **Note 1**, we can write $logq_w{}^*(w), logq_\beta^*(\beta)$ and $logq_{\alpha_d}^*(\alpha_d)$ as follows:
- **Eq2: For w**
  1. $log(q_w{}^*(w)) = E_{\beta,\alpha_1,...\alpha_D}[logp(y,w,\beta,\alpha_1,..\alpha_D)/X]$
  2. Taking terms containing w from equation 1, we have
  3. $E_{q,\beta,\alpha_1,...\alpha_D}\big[\frac{-\beta}{2}\sum_{n\in N}(y_n - w^T x_n)^2 - \frac{1}{2}w^T\Sigma w\big] + const.$
  4. $(-1/2)\big[E[\beta]w^T(\sum_{n\in N} x_n x_n{}^T)w - E[\beta]2w^T\sum_{n\in N} y_n x_n + w^T diag(E[\alpha_1]...E[\alpha_d]w)\big]$
  5. **For w's update,** $E[\beta], E[\alpha_1]...E[\alpha_D]$ **will be needed**
- **Eq3: For** β
  1. $log(q_\beta{}^*(\beta)) = E_{w,\alpha_1,...\alpha_D}[logp(y,w,\beta,\alpha_1,..\alpha_D)/X]$
  2. $\propto E[(N/2)log\beta - (\beta/2)\sum_{n\in N}(y_n - w^T x_n)^2 + (a_0-1)log\beta - b_0\beta]$
  3. On further solving, we have, $(N/2+a_0-1)log\beta - \beta(\sum_{n\in N}(1/2)E(w^T(\sum_{n\in N} x_n x_n{}^T)w - 2w^T\sum_{n\in N} y_n x_n) + b_0)$
  4. The above form is similar to Gamma form with hyperparamters $N/2 + a_0$ and $(1/2)E(w^T(\sum_{n\in N} x_n x_n{}^T)w - 2w^T\sum_{n\in N} y_n x_n) + b_0)$
  5. **For** β**'s update,** $E[w^T(\sum_{n\in N} x_n x_n{}^T)w]$ **and** $E[w]$ **will be needed**
- **Eq4: For** $\alpha_d, d = 1,2,...d$
  - $log(q_{\alpha d}{}^*(\alpha_d)) = E_{w,\beta}[logp(y,w,\beta,\alpha_1,..\alpha_D)/X]$
  - $\propto E[log(\alpha_d/2 - w_d^2\alpha_d/2 + (e_0-1)log(\alpha_d) - f_0\alpha_d)]$
  - $\propto (1/2 + e_0 - 1)log\alpha_d - \alpha_d(f_0 + E[w_d^2]/2)$
  - The above form is the Gamma distribution with hyper parameters $1/2 + e_0$ and $f_0 + E[w_d^2]/2$; **dependent on** $E[w_d^2]$ **term**

Mean Field VI Algorithm

1. **Input:** Model in the form of priors and likelihood
2. **Output:** A variational distribution $q_w, q_\beta, q_{\alpha 1,...D}$
3. **ELBO Expression =** $E_q[logP(Y, w^T X, \beta, \alpha_{1...D})] - E_q[logq(w,\beta,\alpha_{1,2...D})]$
4. While the ELBO is not converged
   - $q_w, q_\beta, q_{\alpha 1,...D}$ are updated in an alternating fashion using ALT-OPT because as shown in above equations, update for $q_w, q_\beta, q_{\alpha 1,...D}$ are dependent on each other.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

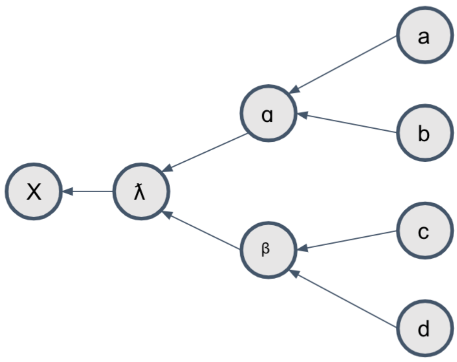
Student Name: Ajita Shree
Roll Number: 20111262
Date: May 14, 2021

QUESTION 3

Gibbs Sampling

Below is the probability distribution for all the random variables and the graphical network.

- $p(x_n/\lambda_n) = Poisson(x_n/\lambda_n)$
- $p(\lambda_n/\alpha, \beta) = Gamma(\lambda_n/\alpha, \beta); n = 1, 2...N$
- $p(\alpha/a, b) = Gamma(\alpha/a, b)$
- $p(\beta/c, d) = Gamma(\beta/c, d)$



Solution

Todo: derive the CP of all variable, if the CP are in closed form or not?

- Joint distribution, $p(X, \lambda, \alpha, \beta/a, b, c, d) = \prod_{n\in N}[p(x_n/\lambda_n)p(\lambda_n/\alpha, \beta)] * p(\alpha/a, b)p(\beta/c, d)$
- $p(\lambda_n/x_n, \alpha, \beta) = Poisson(x_n/\lambda_n) * Gamma(\lambda_n/\alpha, \beta)$
- Using the property of Conjugate priors, $p(\lambda_n/x_n, \alpha, \beta)$ will be a Negative binomial distribution with parameters $\alpha + x_n and \beta + 1$, Hence closed form.
- $p(\beta/\lambda_1...\lambda_n, \alpha,) = \prod_{n\in N} Gamma(\lambda_n/\alpha, \beta) * Gamma(\beta/c, d)$
- Using the property of Conjugate priors, $p(\beta/\lambda_1...\lambda_n, \alpha, \beta)$ will be a CG distribution with parameters $c + na, d + \sum_{n\in N} \lambda_n$ Hence closed form.
- $p(\alpha/\lambda_1...\lambda_n, \beta) = \prod_{n\in N} Gamma(\lambda_n/\alpha, \beta) * Gamma(\alpha/a, b)$
- The Closed form for CP of α is not available because the terms are not conjugate priors of each other.
- Ref: Wikipedia, Conjugate Priors

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 4

Student Name: Ajita Shree
Roll Number: 20111262
Date: May 14, 2021

Using Samples for Prediction

Given a matrix factorization $N*M$, R matrix where $p(r_{ij}/u_i, v_j) = N(r_{ij}/u_i^T v_j, \beta^{-1})$, where u_i and v_j are latent factors of ith and jth column. The PPD of r_{ij} is $p(r_{ij}/R) = p(r_{ij/u_i,v_j})p(u_i, v_j/R)du_i dv_j$. Note we can write each $r_{ij} = u_i^T v_j + \epsilon_{ij}$ where $\epsilon_{ij} = N(\epsilon_{ij}/0, \beta^{-1})$. Also, we are given set of S samples $u^s, v^s{}_{s\in S}$ generated by Gibbs Sampler.

Solution

- **Derivation of sample based approximation of the mean**
- We know, for rv x with probability distribution $p(x), E(x)= \int xp(x)dx$
- $E(r_{ij}) = \int r_{ij}p(r_{ij}/R)dr$, Computing it is intractable, but we have samples $U^s, V^s{}_{s\in S}$ generated by Gibbs Sampler for $p(u_i, v_i/R)$ and we also know that $r_{ij} = u_i^T v_j + \epsilon_{ij}$
- Expectation, $E[r_{ij}] = E[u_i^T v_j] + E[\epsilon_{ij}] = \int u_i^T v_j p(u_i, v_j/R)du_i dv_j + 0; (E[\epsilon_{ij}] = 0)$
- $E[r_{ij}] = \int u_i^T v_j \{u_i^s, v_j^s\}_{s\in S} du_i dv_j = \dfrac{\sum_{s\in S} u_i^{s^T} v_j^s}{S}$
- Hence done
- **Derivation of sample based approximation of variance**
- $Var[r_{ij}] = E[r_{ij}^2] - E[r_{ij}]^2$
- Based on LOTUS, for a rv x with probability distribution p(x) and its function g(x), $E(g(x)) = \int g(x)p(x)$
- $E[r_{ij}^2] = E[(u_i^T v_j + \epsilon_{ij})^2] = E[(u_i^T v_j)^2] + E[\epsilon_{ij}^2] + E[2\epsilon_{ij}(u_i^T v_j)]$
- Using the fact $E[\epsilon_{ij}^2] = Var[\epsilon_{ij}] + E[\epsilon_{ij}]$, above eqn can be written as
- $E[r_{ij}^2] = \int (u_i^T v_j)^2 \{u_i^s, v_j^s\}_{s\in S} du_i dv_j + \beta^{-1} + 2E[\epsilon_{ij} u_i^T v_j]$
- Term 3: $E[\epsilon_{ij} u_i^T v_j] = \int \epsilon_{ij} u_i^T v_j \{u_i^s, v_j^s\}_{s\in S} du_i dv_j N(\epsilon_{ij}, 0, \beta^{-1}) d\epsilon_{ij}$
- Above exp can be solved with identity,$\int xe^{x^2} = e^{x2}/2 + C$
- Finally, $E[r_{ij}^2] = \dfrac{(\sum_{s\in S} u_i^{s^T} v_j^s)^2}{S} + \beta^{-1} - 2\beta^{-1} \dfrac{\sum_{s\in S} u_i^{s^T} v_j^s}{S}$
- $Var[\epsilon_{ij}] = \dfrac{(\sum_{s\in S} u_i^{s^T} v_j^s)^2}{S} + \beta^{-1} - 2\beta^{-1} \dfrac{\sum_{s\in S} u_i^{s^T} v_j^s}{S} - \Big(\dfrac{(\sum_{s\in S} u_i^{s^T} v_j^s)^2}{S}\Big)^2$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Ajita Shree
Roll Number: 20111262
Date: May 14, 2021

QUESTION
5

Rejection Sampling

Consider a distribution p(x) $\propto$ exp(sin(x)) for $-\pi <= x <= \pi$, use proposal distribution $q(x) = N(x/0, \sigma^2)$

- The expression for M w.r.t. $Mq(x) >= p'(x)$ is as follows:
- $M.N(x/0, \sigma^2) >= exp(sin(x))$
- $M\dfrac{exp(-0.5(\dfrac{x}{\sigma})^2)}{\sigma\sqrt{(2\pi)}} >= exp(sin(x))$
- Rearranging terms, we will have $M >= \sigma\sqrt{2\pi}exp(Sin(x) + 0.5 * (\dfrac{x}{\sigma})^2)$
- Considering range of $x \in [-\pi, \pi]$, $M >= \sigma\sqrt{2\pi}exp(1 + 0.5 * (\dfrac{\pi}{\sigma})^2)$
- **Rejection Sampling Algorithm :**
  - sample x* from q(x)
  - sample a uniform rv u [0, Mq(x*)]
  - if u <= p(z*), accept else reject
  - All accepted z*'s are random samples from p(x)

Below plot is corresponding to the above algorithm and the M expression and sigma $value = 1.5, 10000$ samples are drawn from the distribution p(x) and plotted below in the histogram with bin size 500. The M value calculated to be 91.36.