Small Variance Asymptotics(SVA) for Nonparametric Latent Feature Relational Model

Anupreet Porwal 12817143

Avani Samdariya 13173

Kanupriya Agarwal 13338

1 Abstract

Non parametric latent feature relational model(LFRM) was introduced in [7] which uses nonparametric Bayesian approach to simultaneously infer the number of features as well as learn the entities which have that feature. Despite better flexibility, Nonparametric Bayesian (NPB) methods like LFRM face a lot of criticism due to the difficulty in implementing the sampling algorithms/ variational inference techniques, which limits their scalability. In this project, we apply small variance asymptotics (SVA) for non-parametric LFRM. We utilise the connection between exponential families and Bregman divergence, discussed in [5], to derive the scaled version of the Bernoulli likelihood of the model. We then using MAP- based asymptotics derivation discussed in [3] where we apply the SVA directly to the posterior of Bayesian non parametric model to obtain a K-means like objective function. We propose an iterative greedy algorithm to optimize the objective function and support the performance using some empirical results.

2 Literature Survey

The approach of using SVA to devise simple algorithms is inspired from the connection between k-means and mixtures of Gaussians, i.e. the k-means algorithm may be viewed as a limit of the expectation-maximization (EM) algorithm. SVA allows us to asymptotically reduce these NPB clustering techniques to a variety of novel algorithms which have not been discovered in the k-means literature.

There has been some work done in designing scalable hard clustering algorithms from a Bayesian nonparametric viewpoint using SVA for Dirichlet Process Mixture models (see [6]), generalised exponential family mixture models (see [5]) and hidden Markov models (see [10]). Although SVA has been applied to vector valued i.i.d. data and sequential data (see [10]), SVA has not been applied to models relevant to relational data.

Large and complex networks with various patterns of connections between different kind of elements are abound everywhere around us e.g. protein-protein interaction connects proteins by physical binding, social networks, scientific literature connects papers by citations etc. To detect the local connectivity and collective behaviour, one needs to be able to analyse large relational datasets and form meaningful clusters.

One important difference in modelling relational data is that its observations are dependent because of the way they are connected. Thus, they violate the classical independence or exchangeability assumptions made in machine learning and statistics.

Various algorithms exists that uses observed relational network to learn latent structure generating the same. The Stochastic Block Model (SBM), see [9], partition the nodes into non-overlapping communities. However, in general, a node may belong to multiple communities and its community assignment may vary depending upon whom it is interacting with. Mixed Membership Stochastic Block model (MMSBM), see [1], which allows a node to possess multiple features and infer feature indicators for nodes, irrespective of whether edge exist. In this project, we focus at the non-

parameteric latent feature relational model, see [7] which allows to simultaneously learn the number of features and infer the entities which have those features.

3 Model Description

3.1 Likelihood for Non-parametric LFRM

Assume that we are given a graph between N entities, along with its adjacency matrix Y where $y_{ij} = 1$ when a link is observed between i and j entity and $y_{ij} = 0$ if we do not observe a link for $i, j = 1, 2, \ldots N$. In our basic model, we assume that each entity is associated with a set of underlying latent binary features. We also assume that the probability of having a link between two entities depends on the pairwise interaction between the two entities as a part of various features they possess. Formally, given that the number of features is K^+, we define Z as a $N \times K^+$ matrix, where $z_{ik} = 1$ denotes that i^{th} feature possesses feature k. Thus, z_i^T denotes a binary row vector indicating the features shown by the entity i. Let, W denote a $K^+ \times K^+$ matrix where each entry $w_{kk'}$ denotes the weight affecting the probability of link between entity i with feature k and entity j having feature k'. Thus, the conditional likelihood of the data can be written as

$$P(Y|Z, W) = \prod_{i,j=1}^{N} P(y_{ij}|z_i, z_j, W)$$

Given weight matrix W and latent feature matrix Z we can define the the probability of link between the two entities as

$$P(y_{ij}|z_i, z_j, W) = (\sigma(z_i^T W z_j))^{y_{ij}}(1 - \sigma(z_i^T W z_j))^{1-y_{ij}}$$

where $\sigma(.)$ denotes the sigmoid function $\sigma(x) = \frac{1}{1+exp(-x)}$. Thus, the conditional likelihood of the data can be written as:

$$P(Y|Z, W) = \prod_{i,j=1}^{N} p_{ij}^{y_{ij}}(1 - p_{ij})^{1-y_{ij}}$$

where $p_{ij} = \sigma(z_i^T W z_j) = \sigma(\sum_{k,k'=1}^{K^+} z_{ik} z_{jk'} w_{kk'})$

3.2 Prior for feature matrix Z and weight matrix W

Now that we have found the likelihood of the observed data Y, since we wish to infer the posterior over weight matrix W and feature matrix Z, we need to define the priors over the entries of the same.

Suppose we are looking at the relation R, then entry (k,k') would actually correspond to weight with which entities possessing feature k are related with entities possessing feature k' through the relation R. Thus, a positive (negative) weight would imply increased (decreased) probability and zero weight would imply no correlation between two features and observed relation. Hence, we choose

$$w_{kk'} \sim N(0, \sigma_w^2)$$

for all features k, k' for which z_k, $z_{k'}$ are non zero.
Since, we do not assume the number of latent features a priori, we use a variation of Indian Buffet Process prior (see [4]). N entities sample a total of K^+ features and $Z_{1:N,1:K^+}$ is the resulting feature allocation matrix. H represents the number of unique values of $Z_{1:N,k}$ across the k features and $\tilde{K_h}!$ is the number of k with hth unique value of this vector. $S_{N,k}$ denotes the count of feature k being one for first N entities which means that nth entity samples feature k with probability $S_{n-1,k}/n$. The form of IBP for our case is given as follows:

$$P(Z) = \frac{\alpha^{K^+}}{\prod_{h=1}^{H} \tilde{K_h}!} \exp(-\sum_{n=1}^{N} \frac{\alpha}{n}) \prod_{k=1}^{K^+} S_{N,k}^{-1} \binom{N}{S_{N,k}}^{-1}$$

3.3 Generative Model for Non parametric LFRM

As discussed in [7], using an IBP prior on Z gives us the following generative model for non parametric latent feature relational model:

$$
\begin{aligned}
&Z \sim IBP(\alpha) \\
&W \sim N(0, \sigma_w^2) \text{ for all k, k' for which } z_k, z_{k'} \text{ are non zero} \\
&y_{ij} \sim \sigma(z_i^T W z_j) \text{ for each observation}
\end{aligned}
$$

4 Bregman Divergence and Scaled Bernoulli Distribution

In this section, we define small variance asymptotics for the above described model, having Bernoulli likelihood (exponential family). We first express bernoulli distribution in its canonical form using a generalised distance using the bijective relationship between Bregman divergences and exponential families. This is to understand how to scale the variance in bernoulli distribution. Once we establish the scaled form of Bernoulli distribution, we then discuss the asymptotic form of the MAP objective function i.e. proportional to the log posterior for the above model. We know that if $X \sim Ber(q)$ then the exponential family representation of X can be written as the following:

$$P(X|\eta, \psi) = \exp\left[x\eta - \psi(\eta) - h_1(x)\right]$$

$$h_1(x) = 0 \qquad \eta = log\left(\frac{q}{1-q}\right) \qquad \psi(\eta) = \log\left(1 + e^{\eta}\right)$$

where, η denotes the natural parameter, $\psi(\eta)$ denotes the log partition function and x is the sufficient statistics associated with the family. Using properties of log partition function, we can define mean μ and variance σ^2 as

$$E(x) = \nabla_\eta \psi = \frac{e^n}{1+e^n} = q = \mu$$

$$V(x) = \nabla_\eta^2 \psi = q(1-q) = \sigma^2$$

We define a rescaled version of the Bernoulli likelihood with natural parameter $\tilde{\eta} = \beta\eta$ and the log partition function $\tilde{\psi}(\tilde{\eta}) = \beta\psi(\frac{\tilde{\eta}}{\beta})$, where $\beta > 0$. Using the Lemma 3.1 of [5], we infer that the mean $\tilde{\mu}$ and variance $\tilde{\sigma^2}$ of the scaled distribution $\tilde{p}(.)$ will be related to μ and σ^2 as

$$
\begin{aligned}
&\tilde{\mu} = \mu = q \\
&\tilde{\sigma}^2 = \nabla_{\tilde{\eta}}^2 \tilde{\psi}(\tilde{\eta}) = \frac{\sigma^2}{\beta} = \frac{q(1-q)}{\beta}
\end{aligned}
$$

We now look at bregman divergence representation of Bernoulli. From [2], we define the convex function ϕ that links Bernoulli to corresponding Bregman divergence as follows.
Let,

$$\phi(x) = x\log x + (1-x)\log(1-x)$$

Then, the Bregman divergence between the point x and mean $\mu = q$ can be defined as:

$$
\begin{aligned}
d_\phi(x, \mu) &= x\log\frac{x}{\mu} + (1-x)\log\frac{1-x}{1-\mu} \\
&= x\log\frac{x}{q} + (1-x)\log\frac{1-x}{1-q}
\end{aligned}
$$

The derivation of the same can be found in Appendix A. Using the Bregman divergence $d_\phi(x, \mu)$ defined above, the Bernoulli distribution can be expressed as:

$$P(x|\eta, \psi) = \exp\left[-d_\phi(x, \mu)\right] f_\phi(x)$$

where,

$$\eta = \log\left(\frac{q}{1-q}\right)$$
$$f_\phi(x) = \exp\left(x\log x + (1-x)\log(1-x)\right)$$

The derivation of the above representation is attached in appendix B. The rescaled version can be obtained by replacing d_ϕ by $\beta d_\phi(\eta, \mu)$. Denoting $\tilde{\phi} = \beta\phi$, the Bregman divergence representation of scaled bernoulli can be written as:

$$\tilde{P}(x|\tilde{\eta}, \tilde{\psi}) = \tilde{P}(x|\tilde{\mu}) = \exp\left\{-d_{\tilde{\phi}}(x, \tilde{\mu})\right\} \times f_{\tilde{\phi}}(x) = \exp\left\{-d_{\tilde{\phi}}(x, \mu)\right\} \times f_{\tilde{\phi}}(x)$$

where, $f_{\tilde{\phi}}(x) = (f_\phi(x))^\beta$

5 MAP- based asymptotics for Non-parametric LFRM

The joint posterior can be written as:

$$L(W, Z) = P(Z, W|Y) \propto P(Y|Z, W)P(Z)P(W)$$
$$\implies -\log L(W, Z) = -\log P(Y|W, Z) - \log P(Z) - \log P(W) + constant$$

The scaled Bernoulli likelihood can be written as:

$$P(Y|Z, W) = \prod_{i,j=1}^{N} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$$
$$= \prod_{i,j=1}^{N} \exp\left[-\beta[y_{ij}\log\frac{y_{ij}}{p_{ij}} + (1-y_{ij})\log\left(\frac{1-y_{ij}}{1-p_{ij}}\right)]\right] \times \exp\left[\beta[y_{ij}\log y_{ij} + (1-y_{ij})\log(1-y_{ij})]\right]$$

This expression can be simplified to give the corresponding negative log term as:

$$-\log P(Y|Z, W) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\beta[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})]$$

In the IBP prior for Z, choose $\alpha = \exp(-\beta\lambda^2)$ as one would want the number of features to get smaller as $\beta \to \infty$ to avoid overfitting of data to features. Upon substituting this in the IBP prior we get:

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N}\frac{\exp-(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

The negative log of prior for W gives:

$$-\log P(W) = \sum_{k=1}^{K^+}\sum_{k'=1}^{K^+}\frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Note that the entire expression for $-\log P(W)$ is constant with respect to β.

Therefore, the negative log likelihood for $P(Y|Z, W)$ can be written as:

$$-\log L(W, Z) = -\log P(Y|W, Z) - \log P(Z) - \log P(W)$$
$$= -\sum_{i=1}^{N}\sum_{j=1}^{N}\beta[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + K^+\beta\lambda^2 + \sum_{n=1}^{N}\frac{\exp-(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

Dividing this equation by β gives:

$$-\frac{\log L(W, Z)}{\beta} = K^+\lambda^2 + -\sum_{i=1}^{N}\sum_{j=1}^{N}[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + \frac{\exp-(\beta\lambda^2)}{\beta}\sum_{n=1}^{N}\frac{1}{n} + O(\frac{1}{\beta})$$

As $\beta \to \infty$, $O(\frac{1}{\beta}) \to 0$ and $O(\frac{\exp -(\beta\lambda^2)}{\beta}) \to 0$. This gives the objective function, $Q(W, Z)$, which is to be minimized w.r.t. W and Z, as:

$$\begin{aligned} Q(W,Z) &= \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}\log p_{ij} - (1-y_{ij})\log(1-p_{ij})] + K^+\lambda^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}\log\frac{p_{ij}}{(1-p_{ij})} - \log(1-p_{ij})] + K^+\lambda^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}(z_i^T W z_j) + \log(1+\exp(z_i^T W z_j))] + K^+\lambda^2 \\ where, p_{ij} &= \sigma(z_i^T W z_j) \end{aligned}$$

We note that the above function, has two terms: the first terms is similar to the negative log likelihood of a logistic regression model while the second terms is a penalty term on the current number of features. Also, it should be noted that the same objective function $Q(W, Z) = \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}\log p_{ij} - (1-y_{ij})\log(1-p_{ij})] + K^+\lambda^2$ would work even for some other link function like Bernoulli Poisson link function with IBP prior on feature matrix Z, as discussed in [8]. Such a choice of link function is motivated by the nice property that the computational cost scales only linearly with the number of links present and missing entries in the training set.

6 Optimization Algorithm

As observed above, the local MAP that maximizes the log joint posterior, becomes asymptotically equivalent to

$$\min_{W,Z} Q(W, Z)$$

We propose two algorithms for solving the optimization problem, both of which are greedy approaches inspired by FL means algorithm in [11] and the second one from a greedier approach motivated from [3]. The objective function is to be optimized over two arguments: Z and W. Our algorithms, on similar lines as FL-means, optimize the objective function over W and Z in an iterative manner. The key difference from FL- means algorithm is the fact that our objective function is quadratic in Z and thus optimizing over discrete possibilities can be tricky.

The first algorithm we propose is *K-LAFTER I* which is a brute-force greedy approach as it chooses the Z_n value which minimizes the objective function from all $(2^C - 1)$ possible instances of the row vector Z_n in each iteration.

Algorithm 1 K-LAFTER I(Latent feature learning on relational data) algorithm

1: Set C=1. Initialise Z as a N x C matrix by setting $Z_{n1} = 1$ with probability 0.5 $\forall n = 1 \ldots N$
2: Initialise W as C x C matrix with entries drawn from $\mathcal{N}(0, \sigma^2)$
3: **repeat**
4: $\forall n$, optimize Q with respect to Z_n over all $2^C - 1$ possibilities
5: optimize Q(W,Z) with respect to W for current Z and C values
6: Construct Z' from Z by adding a new feature as (C+1) column with one randomly initialized n having that feature
7: Augment W in a similar manner by drawing entries from $\mathcal{N}(0, \sigma^2)$ to form a C+1 dimensional square matrix W'
8: optimize Q with respect to W' for Z' and C+1 features
9: optimize Q with respect to Z' for current W' and C+1 features
10: If (C+1, W', Z') lowers Q from (C,W,Z), replace latter with former
11: **until** convergence

The second algorithm algorithm, *K-LAFTER II*, takes a greedier approach. In Step 3 of *K-LAFTER I*, the algorithm was optimizing the objective function Q over all the $2^C - 1$ possible values of the

binary vector Z_n. In the corresponding step of this algorithm we "switch" the value of each entry Z_{nc} of the matrix Z and choose the optimal value (0 or 1). The time complexity thus becomes $O(N \times C)$ which is a decent improvement from previous $O(N \times 2^C)$ of *K-LAFTER II*.

Algorithm 2 K-LAFTER II(Latent feature learning on relational data) algorithm

1: Set C=1. Initialise Z as a N x C matrix by setting $Z_{n1} = 1$ with probability 0.5 $\forall n = 1 \ldots N$
2: Initialise W as C x C matrix with entries drawn from $\mathcal{N}(0, \sigma^2)$
3: **repeat**
4: $\forall n, c$, Choose the optimal value(0 or 1) of each Z_{nc}
5: optimize Q(W,Z) with respect to W for current Z and C values
6: Construct Z' from Z by adding a new feature as (C+1) column with one randomly initialised n having that feature
7: Augment W in a similar manner by drawing entries from $\mathcal{N}(0, \sigma^2)$ to form a C+1 dimensional square matrix W'
8: optimize Q with respect to W' for Z' and C+1 features
9: optimize Q with respect to Z' for current W' and C+1 features
10: If (C+1, W', Z') lowers Q from (C,W,Z), replace latter with former
11: **until** convergence

7 Experiments and Results

We tested *K-LAFTER I* and *K-LAFTER II* for Lazega lawyers dataset with $\lambda = 0.5$ for 30 iterations. We understand that K-LAFTER II performance, in terms of area under the curve(AUC) for receiver operating characteristic(ROC), is comparable to performance of K-LAFTER I on the dataset and a significant reduction in time from approximately 6 hrs to 1 hrs is obtained with the more greedy approach.

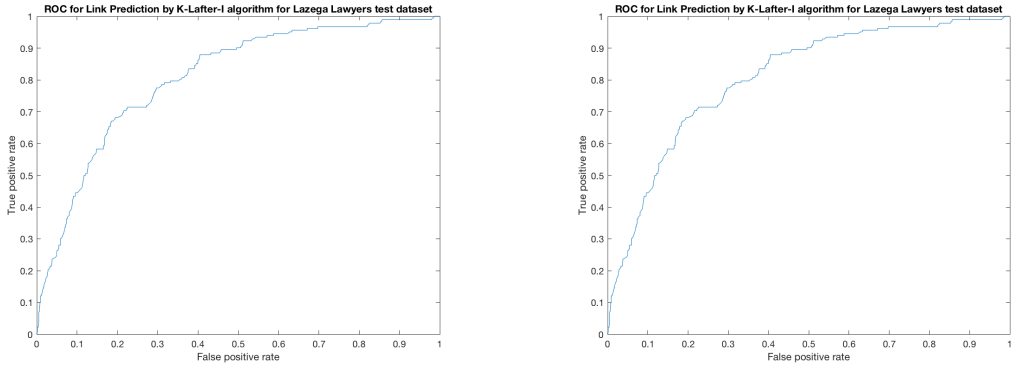

Figure 1: (a)-(b): The graph above shows the performance of K-LAFTER I and II Algorithm for link prediction for Lazega lawyers dataset. Fig (a)-AUC score-0.80; Time taken-5.9Hrs. Fig (b)-AUC score-0.79; Time taken- 0.9Hrs

Since the performance of the faster algorithm i.e. K-LAFTER II is comparable to other graph algorithms for link prediction like LFRM for Lazega dataset, we then apply the algorithm for NIPS234 dataset. Even the faster and better performing K-LAFTER II algorithm takes approximately 3 hrs to converge and manages to get an AUC score of 0.77 in a 80-20% train-test split. We understand that this AUC score is a bit less than the AUC scores obtained by other popular algorithms like LFRM where the score is around 0.95.

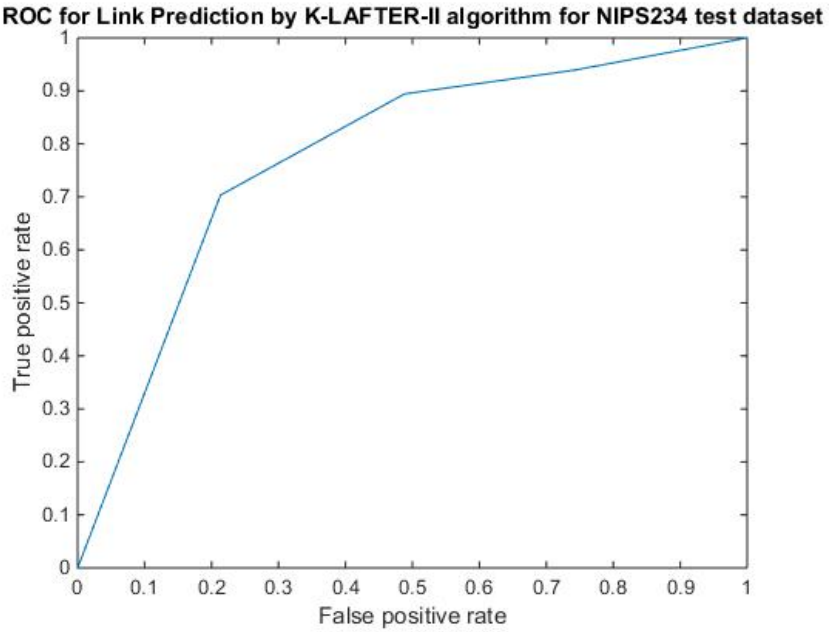

Figure 2: The graphs above shows the performance of K-LAFTER II Algorithm for link prediction for NIPS Co-authorship(NIPS234) dataset .AUC = 0.77; Time Taken = 3Hrs

8 Discussion

As discussed in the previous section, the AUC score for Lazega lawyers dataset for both *K-LAFTER I* and *K-LAFTER II*, are comparable to other graph clustering algorithms such as LFRM. However, when using *K-LAFTER II* for NIPS co-authorship dataset i.e. NIPS234 which is a higher dimensional sparse dataset, the AUC scores (0.8) don't compare well with other algorithms like LFRM(0.95).

We have the empirical evidence of convergence of proposed optimization algorithms. Using this, it is safe to claim that the algorithms converge to at least the local minima, if not global minima. However, we are yet to provide a theoretical proof of convergence, which is one of our future plans for the project.

Another extension yet to be tested is using Bernoulli-Poisson Link function[8] with IBP prior on Z to devise an algorithm that scales up computationally only linearly with number of links present in contrast to sigmoid link function where computation scales quadratically with the number of vertices.

In future, we would also try to look at the SVA for a recently proposed model called Infinite Edge Partition model (IEPM), which was discussed in [12], that provided notable computational savings on sparse network by partitioning only the observed edges and hence the nodes.

9 Conclusion

In this project, we derive the non-probabilistic counterpart of non-parameteric LFRM model by applying SVA to the posterior of the same using link between Bregman divergence and exponential family for scaling the covariance of Bernoulli distribution. We propose a greedy algorithm K-LAFTER II for optimizing the K-means like objective function obtained using MAD- Bayes approach defined above. While AUC scores are comparable to other popular algorithms for small dataset like Lazega lawyers, the code need to be more efficient to improve its performance and speed on more sparse dataset like NIPS234. However, this work definitely helps in bridging the connection between flexible non-parametric Bayesian models with scalable k-means like optimization problems

and open a vast pool of problems where "trick" of small variance asymptotics can be explored to obtain non-probabilistic counterparts of complex non-parametric Bayesian models.

Appendix

Appendix A

Here, we show the dertivation of the Bregman divergence associated with the Bernoulli random variable. From [2], we define the convex function ϕ that links Bernoulli to corresponding Bregman divergence is as follows. Let,

$$\phi(x) = x \log x + (1 - x) \log (1 - x)$$

then,

$$\begin{aligned} \nabla\phi(\mu) &= 1 + \log \mu - 1 - \log (1 - \mu) \\ &= \log \frac{\mu}{1 - \mu} \end{aligned}$$

Using the derivation information and definition of Bregman divergence,

$$\begin{aligned} d_\phi(x, \mu) &= \phi(x) - \phi(\mu) - (x - \mu)\nabla\phi(\mu) \\ &= x \log x + (1 - x) \log (1 - x) - \mu \log \mu - (1 - \mu) \log (1 - \mu) - (x - \mu) \log \frac{\mu}{1 - \mu} \\ &= x \log x - x \log \mu + (1 - x) \log \frac{1 - x}{1 - \mu} \\ d_\phi(x, \mu) &= x \log \frac{x}{\mu} + (1 - x) \log \frac{1 - x}{1 - \mu} \end{aligned}$$

Appendix B

In this subsection, we show the Bregman divergence representation of Bernoulli distribution.

$$\begin{aligned} P(x|\eta, \psi) &= \exp (x \log q + (1 - x) \log (1 - q)) \\ &= \exp (-x \log x + x \log q - (1 - x) \log (1 - x) + (1 - x) \log (1 - q)) \quad \times \\ &\quad \exp (x \log x + (1 - x) \log (1 - x)) \\ &= \exp [-d_\phi(x, q)] f_\phi(x) \end{aligned}$$

Acknowledgments

We want to express our sincere gratitude towards our instructor Prof. Piyush Rai whose inspiring guidance, unfailing support and continuous efforts and encouragement helped us to learn novel stuff and overcome the difficulties faced till now in the project. We are gratefully indebted to him for helping us acquire a decent knowledge in the concerned topic.

References

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[3] Tamara Broderick, Brian Kulis, and Michael I Jordan. Mad-bayes: Map-based asymptotic derivations from bayes. In *ICML (3)*, pages 226–234, 2013.

[4] Thomas L Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *NIPS*, volume 18, pages 475–482, 2005.

[5] Ke Jiang, Brian Kulis, and Michael I Jordan. Small-variance asymptotics for exponential family dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 3158–3166, 2012.

[6] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*, 2011.

[7] Kurt Miller, Michael I Jordan, and Thomas L Griffiths. Non-parametric latent feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284, 2009.

[8] Morten Mørup, Mikkel N Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. In *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, pages 1–6. IEEE, 2011.

[9] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[10] Anirban Roychowdhury, Ke Jiang, and Brian Kulis. Small-variance asymptotics for hidden markov models. In *Advances in Neural Information Processing Systems*, pages 2103–2111, 2013.

[11] Yanxun Xu, Peter Müller, Yuan Yuan, Kamalakar Gulukota, and Yuan Ji. Mad bayes for tumor heterogeneity - feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514, 2015.

[12] Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, 2015.