Approximate Expectation Propagation for Bayesian Inference on Large-scale Problems

Yuan Qi, Tommi S. Jaakkola, and David K. Gifford,
MIT Computer Science and Artificial Intelligence Laboratory

October, 2005

1 Introduction

In this report, we present a novel approach for approximate Bayesian inference on large-scale networks. Speficifically, we consider the following model.

First, we write down the likelihood function of the data as

$$p(\mathbf{y}|\mathbf{b},\mathbf{s}) = \prod_k \prod_i p(y_i^k|\mathbf{b},\mathbf{s}) \tag{1}$$

$$= \prod_k \prod_i \mathcal{N}(y_i^k \mid \sum_{j:a_{|i-j|}>0} a_{|i-j|} s_j b_j, \sigma_i). \tag{2}$$

where k indexes experimental replicates, i indexes the probe positions, j indexes the binding positions, and $\mathcal{N}(\cdot| \sum_j a_{|i-j|} s_j b_j, \sigma_i)$ represents the probability density function of a Gaussian distribution with mean $\sum_j a_{|i-j|} s_j b_j$ and variance σ_i.

We assign prior distributions on the binding event b_j and the binding strength s_j:

$$p(b_j|\pi_j) = \pi_j^{b_j}(1-\pi_j)^{1-b_j} \tag{3}$$

$$p_0(s_j) = \mathrm{Gamma}(s_j|c_0, d_0) \tag{4}$$

where $\mathrm{Gamma}(\cdot|c_0, d_0)$ stands for the probability density functions of Gamma distributions with hyperparameters c_0 and d_0.

We assign a hyperprior distribution on the binding probability π_j as:

$$p_0(\pi_j) = \mathrm{Beta}(\pi_j|\alpha_0, \beta_0) \tag{5}$$

2 Approximate Expectation Propagation for Bayesian inference

First, given the data likelihood (2), the prior distributions (3) and (4) on the binding event $\mathbf{b}$ and strength $\mathbf{s}$, and the hyperprior distribution (5) on the binding probability $\boldsymbol{\pi}$, the posterior distribution $p(\mathbf{b},\mathbf{s},\boldsymbol{\pi}|\mathbf{y})$ is proportional to the joint distribution $p(\mathbf{b},\mathbf{s},\boldsymbol{\pi},\mathbf{y})$:

$$p(\mathbf{b},\mathbf{s},\boldsymbol{\pi}|\mathbf{y}) \propto p(\mathbf{b},\mathbf{s},\boldsymbol{\pi},\mathbf{y}) = \prod_i g_i(\mathbf{b},\mathbf{s}) \prod_j p_0(\pi_j) f_j(b_j,\pi_j) p_0(s_j)$$

where i indexes probe positions, j indexes binding positions, $f_j(b_j, \pi_j) = p(b_j|\pi_j)$ is the prior for b_j, $g_i(\mathbf{b}, \mathbf{s}) = \mathcal{N}(y_i|\sum_j a_{|i-j|} s_j b_j, \sigma_i)$ is the likelihood for the observation at the i^{th} probe position, $p_0(\pi_j)$ is the hyperprior distribution of π_j, and $p(b_j|\pi_j)$ and $p_0(s_j)$ are the prior distributions of b_j and s_j, respectively. For simplicity and clarity, here we drop the superscript k, which indexes replicates, and only consider the case of one replicate. Since the posterior distribution $p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi}|\mathbf{y})$ cannot be computed in a closed form, we use EP to approximate this complicated posterior distribution by a distribution in the exponential family.

EP exploits the fact that the posterior is a product of simple terms. EP iteratively refines the approximation of each term to improve the approximation of the posterior. Mathematically, EP approximates $p(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi}|\mathbf{y})$ as $q(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi})$:

$$q(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi}) = \prod_j q(b_j, s_j) q(\pi_j) = \prod_i \prod_{j: a_{|i-j|} > 0} \tilde{g}_i(b_j, s_j) \prod_j p_0(\pi_j) p_0(s_j) \tilde{f}_j(b_j) \tilde{f}_j(\pi_j) \tag{6}$$

where $g_i(\mathbf{b}, \mathbf{s}) = \prod_{j: a_{|i-j|} > 0} \tilde{g}_i(b_j, s_j)$ is the approximation term corresponding to the likelihood term $g_i(\mathbf{b}, \mathbf{s})$, and $\tilde{f}_j(b_j)\tilde{f}_j(\pi_j)$ is the approximation term corresponding to the prior term $f_j(b_j, \pi_j)$. For simplicity, we denote $f_j(b_j, \pi_j)$ and $\tilde{f}_j(b_j)\tilde{f}_j(\pi_j)$ as $f(b_j, \pi_j)$ and $\tilde{f}(b_j)\tilde{f}(\pi_j)$, respectively. We use a mixture of Gamma distributions to model $q(b_j, s_j)$, i.e.,

$$q(b_j, s_j) = q(b_j) q(s_j|b_j)$$

where $q(b_j)$ is a binomial distribution and $q(s_j|b_j)$ is a Gamma distribution conditional on the binding event b_j. Note that $q(b_j, s_j)$ is still in the exponential family though it is a mixture model.

After initializing $q(b_j = 1) = 0.5$, $q(s_j|b_j) = p_0(s_j)$, and $q(\pi_j) = p_0(\pi_j)$, EP iteratively performs the following two phases, each of which has three steps, to refine the approximate posterior $q(\mathbf{b}, \mathbf{s}, \boldsymbol{\pi})$, until reaching convergence or the maximal number of iterations:

1. Process each likelihood term as follows:

   (a) Deletion: we compute the "leave-one-out" approximate posterior $q^{\setminus i}(\mathbf{b}, \mathbf{s})$, by dividing the current approximate posterior $q(\mathbf{b})q(\mathbf{s}|\mathbf{b})$ by the old approximation term $g_i(\mathbf{b}, \mathbf{s}) = \prod_{j: a_{|i-j|} > 0} \tilde{g}_i(b_j, s_j)$.

   $$q^{\setminus i}(\mathbf{b}, \mathbf{s}) \propto \frac{q(\mathbf{b}, \mathbf{s})}{\prod_j \tilde{g}_i(b_j, s_j)} \tag{7}$$

   (b) Projection: given the "leave-one-out" approximate posterior $q^{\setminus i}(\mathbf{b}, \mathbf{s})$, we compute

   $$\hat{q}(\mathbf{b}, \mathbf{s}) = \underset{\acute{q}(\mathbf{b}, \mathbf{s})}{\operatorname{argmin}} KL(q^{\setminus i}(\mathbf{b}, \mathbf{s}) g_i(\mathbf{b}, \mathbf{s}) || \acute{q}(\mathbf{b}, \mathbf{s}))$$

   This step can be interpreted as projecting a complicated distribution $q^{\setminus i}(\mathbf{b}, \mathbf{s}) g_i(\mathbf{b}, \mathbf{s})$ into a simpler distribution $\hat{q}(\mathbf{b}, \mathbf{s})$, through the above KL minimization. This minimization can be achieved by matching the moments of $\hat{q}(\mathbf{b},\mathbf{s})$ to those of $q^{\setminus i}(\mathbf{b},\mathbf{s})g_i(\mathbf{b},\mathbf{s})$. However, the computation of the moments of $q^{\setminus i}(\mathbf{b},\mathbf{s})g_i(\mathbf{b},\mathbf{s})$ is not tractable since the likelihood term $g_i(\mathbf{b},\mathbf{s})$ involves many latent variables b_j and s_j, leading to a high-dimensional integration over s_j and summation over b_j. To address this problem, we approximate the needed integrations based on Gaussian approximation and quadratures (See Section 2.1).

Given $\hat{q}(\mathbf{b},\mathbf{s})$, we update $\tilde{g}_i^{new}(\mathbf{b},\mathbf{s})$ as follows:

$$\tilde{g}_i^{\text{new}}(\mathbf{b},\mathbf{s}) \propto \left(\hat{q}(\mathbf{b},\mathbf{s})/q^{\setminus i}(\mathbf{b},\mathbf{s})\right)^{\lambda}\tilde{g}_i(\mathbf{b},\mathbf{s})^{1-\lambda},$$

where λ is a step size that controls the update speed.

(c) Inclusion: we replace the old approximation term $\tilde{g}_i(\mathbf{b},\mathbf{s})$ with a new one to obtain $q^{\text{new}}(\mathbf{b},\mathbf{s})$.

$$q^{\text{new}}(\mathbf{b},\mathbf{s}) = q(\mathbf{b},\mathbf{s})\frac{\tilde{g}_i^{\text{new}}(\mathbf{b},\mathbf{s})}{\tilde{g}_i(\mathbf{b},\mathbf{s})} = q^{\setminus i}(\mathbf{b},\mathbf{s})\tilde{g}_i^{\text{new}}(\mathbf{b},\mathbf{s}) \quad (8)$$

2. Process each prior term $f(b_j,\pi_j)$ as follows:

(a) Deletion: we compute the "leave-one-out" approximate posterior $q^{\setminus i}(b_j,\pi_j)$, by dividing the current approximate posterior $q(b_j)q(\pi_j)$ by the old approximation term $\tilde{f}(b_j)\tilde{f}(\pi_j)$.

$$q^{\setminus j}(b_j,\pi_j) = q^{\setminus j}(b_j)q^{\setminus j}(\pi_j) \propto \frac{q(b_j)q(\pi_j)}{\tilde{f}(b_j)\tilde{f}(\pi_j)} \quad (9)$$

(b) Projection: given the "leave-one-out" approximate posterior $q^{\setminus i}(b_j,\pi_j)$, we compute

$$\hat{q}(b_j)\hat{q}(\pi_j) = \operatorname*{argmin}_{\acute{q}(b_j)\acute{q}(\pi_j)} KL(q^{\setminus j}(b_j,\pi_j)f(b_j,\pi_j)||\acute{q}(b_j)\acute{q}(\pi_j))$$

The above KL divergence can be minimized by moment matching. The details are in Section 2.2. Given $\hat{q}(b_j)\hat{q}(\pi_j)$, we update $\tilde{f}^{new}(b_j)\tilde{f}^{new}(\pi_j)$ as follows:

$$\tilde{f}^{\text{new}}(b_j) \propto \left(\hat{q}(b_j)/q^{\setminus j}(b_j)\right)^{\lambda}\tilde{f}(b_j)^{1-\lambda} \quad (10)$$

$$\tilde{f}^{\text{new}}(\pi_j) \propto \left(\hat{q}(\pi_j)/q^{\setminus j}(\pi_j)\right)^{\lambda}\tilde{f}(\pi_j)^{1-\lambda} \quad (11)$$

where λ is a step size that controls the update speed.

(c) Inclusion: we replace the old approximation term $\tilde{f}(b_j)\tilde{f}(\pi_j)$ with a new one to obtain $q^{\text{new}}(b_j)q^{\text{new}}\pi_j)$.

$$q^{\text{new}}(b_j) = q(b_j)\frac{\tilde{f}^{\text{new}}(b_j)}{\tilde{f}(b_j)} = q^{\setminus j}(b_j)\tilde{f}^{\text{new}}(b_j) \quad (12)$$

$$q^{\text{new}}(\pi_j) = q^{\setminus j}(\pi_j)\tilde{f}^{\text{new}}(\pi_j) \quad (13)$$

2.1 Approximate moment matching for incorporating likelihood terms

This section proposes an efficient way to approximate the needed moments in the projection step when incorporating the likelihood terms $g_i(\mathbf{b}, \mathbf{s})$.

We define the normalization constants Z and Z_{b_k}:

$$Z = \sum_{\mathbf{b}} \int q^{\setminus i}(\mathbf{b}, \mathbf{s}) g_i(\mathbf{b}, \mathbf{s}) \mathrm{d}\mathbf{s} \tag{14}$$

$$= \sum_{\{b_m\}_{m \in J}} \int \mathcal{N}(y_i | \sum_{j \in J} a_{|i-j|} s_j b_j, \sigma_i) \prod_{j \in J} q^{\setminus i}(b_j) \prod_{j \in J} q^{\setminus i}(s_j | b_j) \mathrm{d}\mathbf{s} \tag{15}$$

$$Z_{b_k} = \sum_{\{b_m\}_{m \neq k, m \in J}} \int \mathcal{N}(y_i | \sum_{j \in J} a_{|i-j|} s_j b_j, \sigma_i) \prod_{j \neq k, j \in J} q^{\setminus i}(b_j) \prod_{j \neq k, j \in J} q^{\setminus i}(s_j | b_j) \mathrm{d}\mathbf{s} \tag{16}$$

where J represents the set $\{j : a_{|i-j|} > 0\}$. Given Z and Z_{b_k}, we can easily compute the $q(b_k)$ as follows:

$$\hat{q}(b_k) = q^{\setminus i}(b_k) \frac{Z_{b_k}}{Z} \tag{17}$$

However, a direct and exact calculation of Z and Z_{b_k} is computationally expensive because of the high-dimensional integration over $\mathbf{s}$ and summation over $\{b_m\}_{m \neq k, m \in J}$. Therefore, we propose the following method to approximate Z and Z_{b_k}.

Define $y_{\setminus k} = \sum_{j \neq k} a_{|i-j|} s_j b_j + n_i$, where $n_i \sim \mathcal{N}(0, \sigma_i)$. The probability distribution of y_k is a mixture of independent distributions of $s_j b_j$ and n_i. We approximate $y_{\setminus k} = \sum_{j \neq k} a_{|i-j|} s_j b_j + n_i$ by a Gaussian distribution with mean m_k and variance v_k based on moment matching, such that the approximate distribution $N(y_{\setminus k} | m_k, v_k)$ has the same mean and variance as the exact distribution:

$$m_k = \sum_{j \neq k, j \in J} \sum_{b_j} \int b_j s_j q^{\setminus i}(b_j) q^{\setminus i}(s_j | b_j) \mathrm{d}s_m \tag{18}$$

$$= \sum_{j \neq k, j \in J} q^{\setminus i}(b_j = 1) < s_j > \tag{19}$$

$$v_k = \sigma_i^2 + \sum_{j \neq k, j \in J} \left(q^{\setminus i}(b_j = 1) v_{s_j} + q^{\setminus i}(b_j = 1)(1 - q^{\setminus i}(b_j = 1)) < s_j >^2 \right) \tag{20}$$

where $< s_j >$ and v_{s_j} are, respectively, the mean and the variance of $q^{\setminus i}(s_j | b_j = 1)$. This approximation can be justified by the central limit theorem: the distribution of the summation of many similar independent variables converges to a Gaussian distribution. Having obtained m_k and z_k, we can rewrite Z and Z_{b_k} as follows:

$$Z_{b_k} \approx \int \mathcal{N}(y_i - a_{|i-k|} s_k b_k | m_k, v_k) q^{\setminus i}(b_k) q^{\setminus i}(s_k | b_k) \mathrm{d}s_k \tag{21}$$

$$Z = \sum_{b_k} q^{\setminus i}(b_k) Z_{b_k} \tag{22}$$

where $q^{\backslash i}(s_k^g|b_k) = \mathrm{Gamma}(s_k^g|c_{b_k}^{\backslash i}, d_{b_k}^{\backslash i})$. Then, we use the Hermite-Gauss quadrature to approximate the integration in equation (21). The Hermite-Gauss quadrature is a numerical integration technique. It approximates an integration as a weighted sum of integrants evaluated at quadrature nodes. As with importance sampling, it is crucial to have a good proposal distribution to draw the quadrature nodes. Ideally the Gaussian proposal distribution should be similar to the distribution $\hat{q}(s_k|b_j)$, which is proportional to $\int \mathcal{N}(y_i - a_{|i-k|}s_k b_k|m_k, v_k)q^{\backslash i}(s_k|b_j)\mathrm{d}s_k$. Therefore, we want to use a Gaussian distribution that has the same moments as $\hat{q}(s_k|b_j)$. Since we have not obtained the new approximate posterior $\hat{q}(s_k|b_j)$ yet, we match the moments of the Gaussian proposal distribution with those of $q(s_k|b_k)$:

$$Z_{b_k} \approx \int \frac{\mathcal{N}(y_i - a_{|i-k|}s_k b_k|m_k, v_k)q^{\backslash i}(s_k|b_k)}{\mathcal{N}(s_k|\mu_k, \lambda_k)}\mathcal{N}(s_k|\mu_k, \lambda_k)\mathrm{d}s_k \tag{23}$$

$$\approx \sum_g w^g \frac{\mathcal{N}(y_i - a_{|i-k|}s_k^g b_k|m_k, v_k)q^{\backslash i}(s_k^g|b_k)}{\mathcal{N}(s_k^g|\mu_k, \lambda_k)} \tag{24}$$

where μ_k and λ_k are, respectively, the mean and variance of $q(s_k|b_k)$, and s^g and w^g are, respectively, the Gaussian-Hermite quadrature node and the corresponding weight from $\mathcal{N}(s_k|\mu_k, \lambda_k)$. Note that from equation (21), we can directly compute $Z_{b_k=0}$ without using any approximation:

$$Z_{b_k=0} = \mathcal{N}(y_i|m_k, v_k) \tag{25}$$

Similarly, we can compute the new mean and the new variance of $\hat{q}(s_k|b_k)$ as follows:

$$\mu_{b_k} = \frac{1}{Z_{b_k}} \int s_k \mathcal{N}(y_i|\sum_{j\in J} a_{|i-j|}s_j b_j, \sigma_i) \prod_{j\in J} q^{\backslash i}(b_j) \prod_{j\in J} q^{\backslash i}(s_j|b_j)\mathrm{d}\mathbf{s} \tag{26}$$

$$= \frac{1}{Z} \sum_{b_k} q^{\backslash i}(b_k) \sum_g \frac{w^g s_k^g \mathcal{N}(y_i - a_k s_k^g b_k|m_k, v_k)q^{\backslash i}(s_k^g|b_k)}{\mathcal{N}(s_k^g|\mu_k, \lambda_k)} \tag{27}$$

$$\lambda_{b_k} = \frac{1}{Z_{b_k}} \int s_k^2 \mathcal{N}(y_i|\sum_{j\in J} a_{|i-j|}s_j b_j, \sigma_i) \prod_{j\in J} q^{\backslash i}(b_j) \prod_{j\in J} q^{\backslash i}(s_j|b_j)\mathrm{d}\mathbf{s} - \mu_{b_k}^2 \tag{28}$$

$$= \frac{1}{Z} \sum_{b_k} q^{\backslash i}(b_k) \sum_g \frac{w^g (s_k^g)^2 \mathcal{N}(y_i - a_k s_k^g b_k|m_k, v_k)q^{\backslash i}(s_k^g|b_k)}{\mathcal{N}(s_k^g|\mu_k, \lambda_k)} - \mu_{b_k}^2 \tag{29}$$

It is not difficult to see that $\hat{q}(s_k|b_k = 0) \approx q(s_k|b_k = 0)$. Therefore, we only need to use equations (28) and (29) to compute the new mean and the new variance of $\hat{q}(s_k|b_k = 1)$.

Finally, we convert the moment parameters into the natural parameters $\hat{c}_{b_k}$ and $\hat{d}_{b_k}$ of the Gamma distributions for $\hat{q}(s_k|b_k)$:

$$\hat{c}_{b_k} = \mu_{b_k}/\lambda_{b_k} \tag{30}$$

$$\hat{d}_{b_k} = \mu_{b_k}\beta_{b_k} \tag{31}$$

Once having the natural parameters of $\hat{q}(s_k|b_k)$ and $\hat{q}(b_k)$, it is straightforward to compute $\hat{q}(s_k, b_k) = \hat{q}(s_k|b_k)\hat{q}(b_k)$.

2.2 Moment matching for incorporating prior terms

This section presents moment matching when incorporating the prior terms $f_j(b_j, \pi_j) = \pi_j^{b_j}(1-\pi_j)^{1-b_j}$. Given $q^{\setminus j}(\pi_j) \propto \pi_j^{\alpha_j^{\setminus j}-1}(1-\pi_j)^{\beta_j^{\setminus j}-1}$, we compute the normalization constants when processing these terms:

$$Z_{b_j} = \int q^{\setminus j}(\pi_j)\pi_j^{b_j}(1-\pi_j)^{1-b_j}\mathrm{d}\pi_j \tag{32}$$

$$\propto \frac{\Gamma(\alpha_{b_j})\Gamma(\beta_{b_j})}{\Gamma(\alpha_{b_j}+\beta_{b_j})} \tag{33}$$

$$Z = \sum_{b_j} q^{\setminus j}(\pi_j) Z_{b_j} \tag{34}$$

where $\Gamma(\cdot)$ is a Gamma function, and

$$\alpha_{b_j} = \alpha_j^{\setminus j} + b_j \tag{35}$$

$$\beta_{b_j} = \beta_j^{\setminus j} - b_j + 1, \tag{36}$$

Having Z_{b_j} and Z, we can compute $\hat{q}(b_j)$ easily:

$$\hat{q}(b_j) = q^{\setminus j}(b_j)\frac{Z_{b_j}}{Z} \tag{37}$$

Then we compute the mean m_{π_j} and variance v_{π_j} of $\hat{q}(\pi_j)$ as follows:

$$m_{\pi_j|b_j} = \frac{\alpha_{b_j}}{\alpha_{b_j}+\beta_{b_j}} \tag{38}$$

$$v_{\pi_j|b_j} = \frac{m_{\pi_j|b_j}\beta_{b_j}}{(\alpha_{b_j}+\beta_{b_j})(\alpha_{b_j}+\beta_{b_j}+1)} \tag{39}$$

$$m_{\pi_j} = \sum_{b_j}\hat{q}(b_j)m_{\pi_j|b_j} \tag{40}$$

$$v_{\pi_j} = \sum_{b_j}\hat{q}(b_j)(m_{\pi_j|b_j}^2 + v_{\pi_j|b_j}) - m_{\pi_j}^2 \tag{41}$$

Finally, we convert the moment parameters into the natural parameters $\hat{\alpha}_j$ and $\hat{\beta}_j$ of the Beta distribution for $\hat{q}(\pi_j)$:

$$\hat{\alpha}_j = (1-m_{\pi_j})\frac{m_{\pi_j}^2}{v_{\pi_j}} - m_{\pi_j} \tag{42}$$

$$\hat{\beta}_j = \alpha_j(\frac{1}{m_{\pi_j}} - 1) \tag{43}$$