Here, the final summation is over all potential edges (i, j) linked to node i. Updates for assignment distributions depend on expectations of log community assignment probabilities:

$$\mathbb{E}_q[\log(w_k)] = \psi(\lambda_{ka}) - \psi(\lambda_{ka} + \lambda_{kb}), \qquad \mathbb{E}_q[\log(1 - w_k)] = \psi(\lambda_{kb}) - \psi(\lambda_{ka} + \lambda_{kb}), \quad (8)$$

$$\tilde{\pi}_{ik} \triangleq \exp\{\mathbb{E}_q[\log(\pi_{ik})]\} = \exp\{\psi(\theta_{ik}) - \psi(\textstyle\sum_{\ell=1}^{K+1} \theta_{i\ell})\}, \qquad \tilde{\pi}_i \triangleq \textstyle\sum_{k=1}^{K} \tilde{\pi}_{ik}. \quad (9)$$

Given these sufficient statistics, the assignment distributions can be updated as follows:

$$\phi_{ijkk} \propto \tilde{\pi}_{ik}\tilde{\pi}_{jk} f(w_k, y_{ij}), \quad (10)$$

$$\phi_{ijk\ell} \propto \tilde{\pi}_{ik}\tilde{\pi}_{j\ell} f(\epsilon, y_{ij}), \quad \ell \neq k. \quad (11)$$

Here, $f(w_k, y_{ij}) = \exp\{y_{ij}\mathbb{E}_q[\log(w_k)] + (1 - y_{ij})\mathbb{E}_q[\log(1 - w_k)]\}$. More detailed derivations of related updates have been developed for the MMSB [1].

A naive implementation of these updates would require $\mathcal{O}(K^2)$ computation and storage for each assignment distribution $q(e_{ij} \mid \phi_{ij})$. Note, however, that the updates for $q(w_k \mid \lambda_k)$ in Eq. (6) depend only on the K probabilities ϕ_{ijkk} that nodes select the same community. Using the updates for $\phi_{ijk\ell}$ from Eq. (11), the update of $q(\pi_i \mid \theta_i)$ in Eq. (7) can be expanded as follows:

$$\begin{aligned}\theta_{ik} &= \alpha\beta_k + \textstyle\sum_{(i,j)\in E} \phi_{ijkk} + \frac{1}{Z_{ij}} \sum_{\ell\neq k} \tilde{\pi}_{ik}\tilde{\pi}_{j\ell} f(\epsilon, y_{ij}) \\ &= \alpha\beta_k + \textstyle\sum_{(i,j)\in E} \phi_{ijkk} + \frac{1}{Z_{ij}} \tilde{\pi}_{ik} f(\epsilon, y_{ij})(\tilde{\pi}_j - \tilde{\pi}_{jk}).\end{aligned} \quad (12)$$

Note that $\tilde{\pi}_j$ need only be computed once, in $\mathcal{O}(K)$ operations. The normalization constant Z_{ij}, which is defined so that ϕ_{ij} is a valid categorical distribution, can also be computed in linear time:

$$Z_{ij} = \tilde{\pi}_i\tilde{\pi}_j f(\epsilon, y_{ij}) + \textstyle\sum_{k=1}^{K} \tilde{\pi}_{ik}\tilde{\pi}_{jk}(f(w_k, y_{ij}) - f(\epsilon, y_{ij})). \quad (13)$$

Finally, to evaluate our variational bound and assess algorithm convergence, we still need to calculate the likelihood and entropy terms dependent on $\phi_{ijk\ell}$. However, we can compute part of our bound by caching our partition function Z_{ij} in linear time. See $\ddagger A.2$ for details regarding the full derivation of this ELBO and its extensions.

3.3 Stochastic Variational Inference

Standard variational batch updates become computationally intractable when N becomes very large. Recent advancements in applying stochastic optimization techniques within variational inference [8] showed that if our variational mean-field family of distributions are members of the exponential family, we can derive a simple stochastic natural gradient update for our global parameters λ, θ, v. These gradients can be calculated from only a subset of the data and are noisy approximations of the true natural gradient for the variational objective, but represent an unbiased estimate of that gradient.

To accomplish this, we define a new variational objective with respect to our current set of observations. This function, in expectation, is equivalent to our true ELBO. By taking natural gradients with respect to our new variational objective for our global variables λ, θ, we have

$$\nabla\lambda_{ka}^{*} = \tfrac{1}{g(i,j)}\phi_{ijkk}y_{ij} + \tau_a - \lambda_{ka}; \quad (14)$$

$$\nabla\theta_{ik}^{*} = \tfrac{1}{g(i,j)} \textstyle\sum_{(i,j)\in E} \sum_{\ell=1}^{K} \phi_{ijk\ell} + \alpha\beta_k - \theta_{ik}, \quad (15)$$

where the natural gradient for $\nabla\lambda_{kb}^{*}$ is symmetric to $\nabla\lambda_{ka}^{*}$ and where y_{ij} in Eq. (14) is replaced by $(1 - y_{ij})$. Note that $\sum_{(i,j)\in E} \sum_{\ell=1}^{K} \phi_{ijk\ell}$ was shown in the previous section to be computable in $\mathcal{O}(K)$. The scaling term $g(i, j)$ is needed for an unbiased update to our expectation. If $g(i, j) = 2/N(N-1)$, then this would represent a uniform distribution over possible edge selections in our undirected graphs. In general, $g(i, j)$ can be an arbitrary distribution over possible edge selections such as a distribution over sets of edges as long as the expectation with respect to this distribution is equivalent to the original ELBO [6]. When referring to the scaling constant associated with sets, we consider the notation of $h(T)$ instead of $g(i, j)$.

We optimize this ELBO with a Robbins-Monro algorithm which iteratively steps along the direction of this noisy gradient. We specify a learning rate $\rho_t \triangleq (\mu_0 + t)^{-\kappa}$ at time t where $\kappa \in (.5, 1]$ and $\mu_0 \geq 0$ downweights the influence of earlier updates. With the requirement that $\sum_t \rho_t^2 < \infty$ and