Supplementary Material: Deep Generative Models for Relational Data with Side Information

1. Proof of Lemma 1

We can compute $\mathbb{E}[\mathbf{I}\{A_{ij}=0\}]$ as

$$
\begin{aligned}
&\mathbb{E}[\mathbf{I}\{A_{ij}=0\}] = p(X_{ij}=0)\\
=\ & \mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j,\Lambda}\left[\prod_{k_1,k_2}^{K} p(X_{ij}=0|z_{ik_1},z_{jk_2},\Lambda_{k_1k_2})\right]\\
=\ & \mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j,\Lambda}\left[\prod_{k_1,k_2=1}^{K} \exp(-\Lambda_{k_1k_2}z_{ik_1}z_{jk_2})\right]\\
\geq\ & \exp\left(\mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j,\Lambda}\left[\log\prod_{k_1,k_2=1}^{K} \exp(-\Lambda_{k_1k_2}z_{ik_1}z_{jk_2})\right]\right)\\
=\ & \mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j,\Lambda}\left[-\sum_{k_1,k_2=1}^{K} \Lambda_{k_1k_2}z_{ik_1}z_{jk_2}\right] \qquad (1)
\end{aligned}
$$

where the inequality step follows from Jensen's inequality. Following Lemma 1 in (Zhou, 2015), we have $\mathbb{E}\left[\sum_{k_1,k_2=1}^{K}\Lambda_{k_1k_2}\right] = \frac{\zeta\gamma_c}{\gamma_b c_{k_1k_2}} + \frac{\gamma_a^2}{\gamma_b^2 c_{k_1k_2}}$. Then the last line in Equation (1) can be written as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j,\Lambda}\left[-\sum_{k_1,k_2=1}^{K} \Lambda_{k_1k_2}z_{ik_1}z_{jk_2}\right] \qquad (2)\\
&= \exp\left(-\left[\frac{\zeta\gamma_c}{\gamma_b c_{k_1k_2}} + \frac{\gamma_a^2}{\gamma_b^2 c_{k_1k_2}}\right]\mathbb{E}_{\boldsymbol{z}_i^{(1)}\boldsymbol{z}_j^{(1)}}\left[z_{ik_1}^{(1)}z_{jk_2}^{(1)}\right]\right)
\end{aligned}
$$

Based on Equation (1) and (3), the expected number of zeros in $\mathbf{A}$ is lower bounded by

$$
\begin{aligned}
&\mathbb{E}[\sum_{i,j=1}^{N}\mathbf{I}\{A_{ij}=0\}] \geq N^2\mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j,\Lambda}\left[-\sum_{k_1,k_2=1}^{K}\Lambda_{k_1k_2}z_{ik_1}z_{jk_2}\right]\\
&= N^2\exp\left(-\left[\frac{\zeta\gamma_c}{\gamma_b c_{k_1k_2}} + \frac{\gamma_a^2}{\gamma_b^2 c_{k_1k_2}}\right]\mathbb{E}_{\boldsymbol{z}_i^{(1)}\boldsymbol{z}_j^{(1)}}\left[z_{ik_1}^{(1)}z_{jk_2}^{(1)}\right]\right) \qquad (3)
\end{aligned}
$$

2. HYPERPARAMETER INFERENCE

We sample $\boldsymbol{w}_k^{(\ell)}$, $b_k^{(\ell)}$ and $\boldsymbol{m}_k$ leveraging the Pólya-Gamma augmentation (Polson et al., 2013). This enables us to derive the Gibbs sampler updates for the hyper-parameters γ_{k_1}, ξ, $\boldsymbol{\Gamma}_{k,\ell}^{(\boldsymbol{w})}$ and $\boldsymbol{\Gamma}_k^{(\boldsymbol{m})}$, in closed form.

. Correspondence to: Changwei Hu <changweih@yahoo-inc.com>, Piyush Rai <piyush@cse.iitk.ac.in>, Lawrence Carin <lcarin@duke.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Sample $\boldsymbol{w}_k^{(\ell)}$ and $b_k^{(\ell)}$: We consider the update of layer-1 weights $\boldsymbol{w}_k^{(1)}$ as an example, and assume the side information is available (which is the more general case). Weights for the other layers can be sampled in a similar manner.

Given the Pólya-Gamma auxiliary variables $\boldsymbol{\alpha}_k^{(1)}$, the posterior for $\boldsymbol{w}_k^{(1)}$ will be $\boldsymbol{w}_k^{(1)} \sim \mathcal{N}(\boldsymbol{\mu}_k^{(\boldsymbol{w})}, \mathbf{V}_k^{(\boldsymbol{w})})$, where

$$
\begin{aligned}
\boldsymbol{\mu}_k^{(\boldsymbol{w})} &= \mathbf{V}_k^{(\boldsymbol{w})}(\mathbf{Z}^{(2)})^T(\boldsymbol{z}_k^{(2)} - \frac{1}{2}\mathbf{1}_N - \text{diag}(\boldsymbol{\alpha}_k^{(1)})(\mathbf{S}\boldsymbol{m}_k + b_k^{(1)}\mathbf{1}_N))\\
\mathbf{V}_k^{(\boldsymbol{w})} &= ((\mathbf{Z}^{(2)})^T\text{diag}(\boldsymbol{\alpha}_k^{(1)})\mathbf{Z}^{(2)} + (\boldsymbol{\Gamma}_{k,\ell}^{(\boldsymbol{w})})^{-1})^{-1}
\end{aligned}
$$

In the above, $\mathbf{1}_N$ is a vector of length N with all entries being 1, and $\boldsymbol{\alpha}_k^{(1)} \in \mathbb{R}_+^N$, each entry $\alpha_{ik}^{(1)}$ is drawn from the Pólya-Gamma distribution

$$\alpha_{ik}^{(1)} \sim \text{PG}(1, \boldsymbol{m}_k^T\boldsymbol{s}_i + (\boldsymbol{w}_k^{(1)})^\top\boldsymbol{z}_i^{(2)} + b_k^{(1)})$$

Conditioned on these PG variables, the posterior over $b_k^{(\ell)}$ will also be a Gaussian.

Sample $\boldsymbol{m}_k$: Akin to the way we sample $\boldsymbol{w}_k^{(\ell)}$, the side information based regression weights $\boldsymbol{m}_k$ can also be sampled using the Pólya-Gamma scheme (using the layer 1 PG variables $\boldsymbol{\alpha}_k^{(1)}$). The posterior will be a Gaussian $\boldsymbol{m}_k \sim \mathcal{N}(\boldsymbol{\mu}_k^{(\boldsymbol{m})}, \mathbf{V}_k^{(\boldsymbol{m})})$, where

$$
\begin{aligned}
\boldsymbol{\mu}_k^{(\boldsymbol{m})} &= \mathbf{V}_k^{(\boldsymbol{m})}\mathbf{S}^T(\boldsymbol{z}_k^{(2)} - \frac{1}{2}\mathbf{1}_N - \text{diag}(\boldsymbol{\alpha}_k^{(1)})(\mathbf{Z}^{(2)}\boldsymbol{w}_k^{(1)} + b_k^{(1)}\mathbf{1}_N))\\
\mathbf{V}_k^{(\boldsymbol{m})} &= ((\mathbf{S}^T\text{diag}(\boldsymbol{\alpha}_k^{(1)})\mathbf{S} + (\boldsymbol{\Gamma}_k^{(\boldsymbol{m})})^{-1})^{-1}
\end{aligned}
$$

Sample γ_{k_1}: γ_{k_1} can be sampled as

$$\gamma_{k_1} \sim \text{Gamma}(\gamma_a + \ell_{k_1k_2}, \frac{1}{\gamma_b - \sum_{k_2}\xi^{\delta_{k_1k_2}}\gamma_{k_2}^{1-\delta_{k_1k_2}}\ln(\frac{c_{k_1k_2}}{Q_{k_1k_2}+c_{k_1k_2}})})$$

where $\ell_{k_1\cdot} = \sum_{k_2}\ell_{k_1k_2}$ with $\ell_{k_1k_2}$ drawn from the Chinese Restaurant Table (CRT) distribution (Zhou, 2015)

$$\ell_{k_1k_2} \sim \text{CRT}(X_{\cdot\cdot k_1k_2}, g_{k_1k_2})$$

Sample ξ: The hyperparameter ξ can be sampled as

$$\xi \sim \text{Gamma}(\xi_a + \sum_k \ell_{kk}, \frac{1}{\xi_b - \sum_k\gamma_k\ln(\frac{c_{kk}}{Q_{kk}+c_{kk}})})$$

Sample $\boldsymbol{\Gamma}_{k,\ell}^{(\boldsymbol{w})}$, $\boldsymbol{\Gamma}_{k}^{(\boldsymbol{m})}$: Each diagonal entry of the precision matrix $\boldsymbol{\Gamma}_{k,\ell}^{(\boldsymbol{w})}$ is sampled as

$$\boldsymbol{\Gamma}_{k,\ell}^{(\boldsymbol{w})} \sim \text{Gamma}(a + \frac{K_{\ell+1}}{2}, \frac{1}{\text{diag}((b + 0.5(\boldsymbol{w}_k^{(\ell)})^T \boldsymbol{w}_k^{(\ell)}) \mathbf{1}_{K_{\ell+1}})})$$

where a and b are the scale and rate parameters for the prior of $\boldsymbol{\Gamma}_{k,\ell}^{(\boldsymbol{w})}$ respectively. $\boldsymbol{\Gamma}_{k}^{(\boldsymbol{m})}$ can be sampled similarly.

References

Polson, Nicholas G, Scott, James, and Windle, Jesse. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

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