Appendix for Graph Sequential Neural ODE Process for Link Prediction on Dynamic and Sparse Graphs

LINHAO LUO, Monash University, Australia
GHOLAMREZA HAFFARI, Monash University, Australia
SHIRUI PAN*, Griffith University, Australia

1 DERIVATION OF ELBO LOSS

Given the context data $G_{\leq T}$ and target data $G_{>T}$, the object of GSNOP is to infer the distribution $P(z_{>T}|G_{\leq T})$ from the context data that minimize the prediction loss on the target data. Therefore, the generation process of GSNOP can be written as

$$P(z_{>T}, Y_D|\mathcal{E}_D, \mathcal{G}_{\leq T}) = P(z_{>T}|G_{\leq T}) \prod_{e_D(t)\in\mathcal{E}_D} P\big(y_D|f(e_D(t), z_{>T})\big), \tag{1}$$

where f is the link prediction decoder, and $\mathcal{E}_D = \{e_D(t) \in G_{>T}\}$ denotes the links to be predicted in the target data.

Following the Equation 1, we can rewrite the $logP(Y_D|\mathcal{E}_D, G_{\leq T})$ as

$$logP(Y_D|\mathcal{E}_D, G_{\leq T}) = log\frac{P(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T})}{P(z_{>T}|G_{\leq T})}, \tag{2}$$

$$= logP(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T}) - logP(z_{>T}|G_{\leq T}). \tag{3}$$

Assuming the true distribution of $z_{>T}$ is $Q(z_{>T})$, we can rewrite the Equation 3 as

$$logP(Y_D|\mathcal{E}_D, G_{\leq T}) = log\frac{P(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T})}{Q(z_{>T})} - log\frac{P(z_{>T}|G_{\leq T})}{Q(z_{>T})}. \tag{4}$$

We integrate both side with $Q(z_{>T})$.

$$\int_z Q(z_{>T})logP(Y_D|\mathcal{E}_D, G_{\leq T}) = \int_z Q(z_{>T})log\frac{P(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T})}{Q(z_{>T})} - \int_z Q(z_{>T})log\frac{P(z_{>T}|G_{\leq T})}{Q(z_{>T})}, \tag{5}$$

$$logP(Y_D|\mathcal{E}_D, G_{\leq T}) = \int_z Q(z_{>T})log\frac{P(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T})}{Q(z_{>T})} + KL\big(Q(z_{>T})||P(z_{>T}|G_{\leq T})\big). \tag{6}$$

Because $KL\big(Q(z_{>T})||P(z_{>T}|G_{\leq T})\big) \geq 0$, we can write Equation 6 as

$$logP(Y_D|\mathcal{E}_D, G_{\leq T}) \geq \int_z Q(z_{>T})log\frac{P(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T})}{Q(z_{>T})}, \tag{7}$$

$$= \mathbb{E}_{Q(z_{>T})}log\frac{P(Y_D, z_{>T}|\mathcal{E}_D, G_{\leq T})}{Q(z_{>T})}, \tag{8}$$

$$= \mathbb{E}_{Q(z_{>T})}[logP(Y_D|\mathcal{E}_D, z_{>T}) + log\frac{P(z_{>T}|G_{\leq T})}{Q(z_{>T})}], \tag{9}$$

$$= \mathbb{E}_{Q(z_{>T})}[logP(Y_D|\mathcal{E}_D, z_{>T})] - KL\big(Q(z_{>T})||P(z_{>T}|G_{\leq T})\big), \tag{10}$$

*Corresponding author.

Authors' addresses: Linhao Luo, linhao.luo@monash.edu, Monash University, Australia; Gholamreza Haffari, Gholamreza.Haffari@monash.edu, Monash University, Australia; Shirui Pan, s.pan@griffith.edu.au, Griffith University, Australia.

where $P(Y_D|\mathcal{E}_D, z_{>T})$ is calculated by the link prediction decoder $P_\phi(Y_D|\mathcal{E}_D, z_{>T})$, $Q(z_{>T})$ is approximated by the encoder and RNN aggregator, written as $Q_\psi(z|G_{\leq T}, G_{>T})$, and $P(z_{>T}|G_{\leq T})$ is approximated by the NODE, written as $P_{ode}(z_{>T}|G_{\leq T})$.

Thus, we can obtain the final ELBO loss as

$$\mathcal{L} = \mathbb{E}_{Q_\psi(z_{>T}|G_{\leq T}, G_{>T})}[logP_\phi(Y_D|\mathcal{E}_D, z_{>T})] - KL(Q_\psi(z_{>T}|G_{\leq T}, G_{>T})||P_{ode}(z_{>T}|G_{\leq T})). \tag{11}$$

2 OPTIMIZATION OF ELBO LOSS

To optimize ELBO loss, we need to compute the gradients with respect to the parameters of the model.

In ELBO loss, we encourage the NODE to learn the derivative of distribution and infer the distribution in the future by minimizing the KL divergence between variational posterior $Q_\psi(z_T|G_{\leq T}, G_{>T})$ and prior $P_{ode}(z_{>T}|G_{\leq T})$. Thus, the gradients for ψ and NODE can be directly computed.

However, for the first expectation term, since $z_{>T}$ is sampled from $\mathcal{N}\big(\mu(\mathbf{r'}_{>T}), \sigma(\mathbf{r'}_{>T})\big)$, we adopt the reparameterization trick [2] to compute the gradients and compute the expectation using Monte Carlo methods, written as

$$z^l_{>T} = \mu(\mathbf{r'}_{>T}) + \sigma(\mathbf{r'}_{>T})\epsilon^l, \epsilon^l \sim \mathcal{N}(0, 1), \tag{12}$$

$$\mathbb{E}_{Q(z_{>T})}[logP_\phi(Y_D|\mathcal{E}_D, z_{>T})] \simeq \frac{1}{L}\sum_{l=1}^{L} logP_\phi(Y_D|\mathcal{E}_D, z^l_{>T}). \tag{13}$$

We set $L = 10$ in experiments.

For parameters θ in NODE, we adopt the adjoint sensitivity method [1, 3] to compute the gradients $\frac{d\mathcal{L}}{d\theta}$, written as

$$\frac{d\mathcal{L}}{d\theta} = -\int_{T+\triangle T}^{T} \mathbf{a}(t)^\top \frac{\partial f_{ode}(\mathbf{r}_t, t, \theta)}{\partial \theta}, \tag{14}$$

where $\mathbf{a}(t)$ is denoted as the gradient of the loss depends on the hidden state. Its dynamics are given by another ODE, written as

$$\frac{d\mathbf{a}(t)}{dt} = -\mathbf{a}(t)^\top \frac{\partial f_{ode}(\mathbf{r}_t, t, \theta)}{\partial \mathbf{r}_t}. \tag{15}$$

REFERENCES

[1] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems* 31 (2018).

[2] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[3] Lev Semenovich Pontryagin. 1987. *Mathematical theory of optimal processes.* CRC press.