Learning Deep Sigmoid Belief Networks with Data Augmentation: Supplemental Material

Zhe Gan **Ricardo Henao** **David Carlson** **Lawrence Carin**
Department of Electrical and Computer Engineering, Duke University, Durham NC 27708, USA

A Graphical Model

Figure 1 shows the graphical model for the deep SBN with autoregressive structure. $\mathbf{S}^{(\ell)}$ and $\mathbf{U}$ contain the autoregressive weights within layers, while $\mathbf{W}^{(\ell)}$ is utilized to capture the dependencies between different layers.

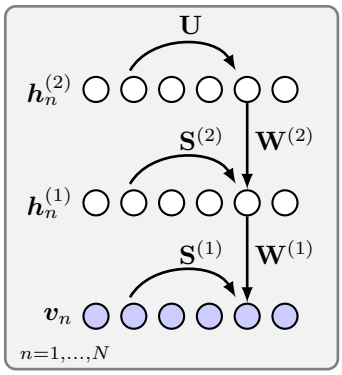

Figure 1: Graphical model for the deep SBN with autoregressive structure.

B Properties of Pólya-Gamma distribution

A random variable X has a Pólya-Gamma distribution (Polson et al., 2013) with parameters $b > 0$ and $c \in \mathbb{R}$, denoted $X \sim \text{PG}(b, c)$, if

$$X = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)}\,, \qquad (1)$$

where each $g_k \sim \text{Ga}(b, 1)$ is an independent gamma random variable. We have

$$\mathbb{E}[X] = \frac{b}{2c}\tanh(c/2) = \frac{b}{2c}\left(\frac{e^c - 1}{e^c + 1}\right). \qquad (2)$$

A key observation is that binomial likelihoods parametrized by log-odds can be written as mixtures of Gaussians with respect to a Pólya-Gamma distribution. Specifically, if $\gamma \sim \text{PG}(b, 0)$, $b > 0$, then

$$\frac{(e^\psi)^a}{(1+e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\gamma\psi^2/2} p(\gamma) d\gamma\,, \qquad (3)$$

where $\kappa = a - b/2$. And we have $\gamma|\psi \sim \text{PG}(b, \psi)$. Proof is given in Polson et al. (2013), Section 3.

The generation of the Pólya-Gamma variables is detailed in Polson et al. (2013), Section 4. Other approximate methods for generation are discussed in the supplemental material of Zhou et al. (2012) and Chen et al. (2013).

C Inference details on ARSBN

We consider the one-hidden-layered ARSBN model defined as (see Section 2.2)

$$p(v_{jn} = 1|\boldsymbol{h}_n, \boldsymbol{v}_{<j,n}) = \sigma(\boldsymbol{w}_j^\top \boldsymbol{h}_n + \boldsymbol{s}_{j,<j}^\top \boldsymbol{v}_{<j,n} + c_j)\,,$$
$$p(h_{kn} = 1|\boldsymbol{h}_{<k,n}) = \sigma(\boldsymbol{u}_{k,<k}^\top \boldsymbol{h}_{<k,n} + b_k)\,. \qquad (4)$$

Assume isotropic normal priors are imposed on the $\boldsymbol{s}_{j,<j}$ and $\boldsymbol{u}_{k,<k}$, the other prior settings are the same as in SBN. The conditional posterior distributions used in the Gibbs sampling are as follows.

For $\gamma^{(0)}, \gamma^{(1)}$: The conditional distribution of $\gamma^{(0)}$ is $p(\gamma_{jn}^{(0)}|-) = \text{PG}(1, \boldsymbol{w}_j^\top \boldsymbol{h}_n + \boldsymbol{s}_{j,<j}^\top \boldsymbol{v}_{<j,n} + c_j)$. Similarly, $p(\gamma_{kn}^{(1)}|-) = \text{PG}(1, \boldsymbol{u}_{k,<k}^\top \boldsymbol{h}_{<k,n} + b_k)$.

For H: The conditional distribution of h_{kn} is $p(h_{kn}|-) = \text{Ber}(\sigma(d_{kn}))$, where

$$d_{kn} = b_k + \boldsymbol{w}_k^\top \boldsymbol{v}_n + \boldsymbol{u}_{k,<k}^\top \boldsymbol{h}_{<k,n} - \frac{1}{2}\sum_{j=1}^{J}\left(w_{jk} + \gamma_{jn}^{(0)}(2\psi_{jn}^{\setminus k} w_{jk} + w_{jk}^2)\right), \qquad (5)$$

and $\psi_{jn}^{\setminus k} = \boldsymbol{w}_j^\top \boldsymbol{h}_n - w_{jk} h_{kn} + \boldsymbol{s}_{j,<j}^\top \boldsymbol{v}_{<j,n} + c_j$.

For W: The conditional distribution of $\boldsymbol{w}_j$ is

$p(\boldsymbol{w}_j|-) = N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where

$$\boldsymbol{\Sigma}_j = \left[\sum_{n=1}^{N} \gamma_{jn}^{(0)} \boldsymbol{h}_n \boldsymbol{h}_n^\top + \mathrm{diag}(\boldsymbol{\zeta}_j^{-1})\right]^{-1}, \quad (6)$$

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \left[\sum_{n=1}^{N} \left(v_{jn} - \frac{1}{2} - \gamma_{jn}^{(0)}(\boldsymbol{s}_{j,<j}^\top \boldsymbol{v}_{<j,n} + c_j)\right) \boldsymbol{h}_n\right].$$

For S: The conditional distribution of $\boldsymbol{s}_{j,<j}$ is $p(\boldsymbol{s}_{j,<j}|-) = N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where

$$\boldsymbol{\Sigma}_j = \left[\sum_{n=1}^{N} \gamma_{jn}^{(0)} \boldsymbol{v}_{<j,n} \boldsymbol{v}_{<j,n}^\top + \mathbf{I}_{j-1}\right]^{-1}, \quad (7)$$

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \left[\sum_{n=1}^{N} \left(v_{jn} - \frac{1}{2} - \gamma_{jn}^{(0)}(\boldsymbol{w}_j^\top \boldsymbol{h}_n + c_j)\right) \boldsymbol{v}_{<j,n}\right].$$

For U: The conditional distribution of $\boldsymbol{u}_{k,<k}$ is $p(\boldsymbol{u}_{k,<k}|-) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where

$$\boldsymbol{\Sigma}_k = \left[\sum_{n=1}^{N} \gamma_{kn}^{(1)} \boldsymbol{h}_{<k,n} \boldsymbol{h}_{<k,n}^\top + \mathbf{I}_{k-1}\right]^{-1}, \quad (8)$$

$$\boldsymbol{\mu}_k = \boldsymbol{\Sigma}_k \left[\sum_{n=1}^{N} \left(h_{kn} - \frac{1}{2} - \gamma_{kn}^{(1)} b_k\right) \boldsymbol{h}_{<k,n}\right].$$

D Inference details on VB SBN

D.1 Derivation of the Lower Bound

The lower bound in Equation (16), Section 3.2 is derived below. First,

$$\langle \log p(v_{jn}|-)\rangle = -\log 2 + (v_{jn} - 1/2)\langle \psi_{jn}\rangle + \left\langle \log \int \exp\left(-\frac{\gamma_{jn}^{(0)}\psi_{jn}^2}{2}\right) p(\gamma_{jn}^{(0)}) \frac{q(\gamma_{jn}^{(0)})}{q(\gamma_{jn}^{(0)})} d\gamma_{jn}^{(0)} \right\rangle. \quad (9)$$

The third term in (9) can be lower-bounded as

$$\begin{aligned} &\left\langle \log \int \exp\left(-\frac{\gamma_{jn}^{(0)}\psi_{jn}^2}{2}\right) p(\gamma_{jn}^{(0)}) \frac{q(\gamma_{jn}^{(0)})}{q(\gamma_{jn}^{(0)})} d\gamma_{jn}^{(0)} \right\rangle \\ =& \left\langle \log \left\langle \exp\left(-\frac{\gamma_{jn}^{(0)}\psi_{jn}^2}{2}\right) \frac{p(\gamma_{jn}^{(0)})}{q(\gamma_{jn}^{(0)})} \right\rangle \right\rangle \quad (10) \\ \geq& -\frac{1}{2}\langle \gamma_{jn}^{(0)}\rangle \langle \psi_{jn}^2\rangle + \langle \log p_0(\gamma_{jn}^{(0)})\rangle - \langle \log q(\gamma_{jn}^{(0)})\rangle. \end{aligned}$$

Substituting (10) into (9), we obtain

$$\begin{aligned} \langle \log p(v_{jn}|-)\rangle \geq\ & -\log 2 + (v_{jn} - 1/2)\langle \psi_{jn}\rangle \\ -\frac{1}{2}\langle \gamma_{jn}^{(0)}\rangle \langle \psi_{jn}^2\rangle &+ \langle \log p_0(\gamma_{jn}^{(0)})\rangle - \langle \log q(\gamma_{jn}^{(0)})\rangle. \quad (11) \end{aligned}$$

D.2 VB update equations

The VB update equations for the SBN model are listed below.

For $\gamma^{(0)}, \gamma^{(1)}$:

$$q(\gamma_{jn}^{(0)}) = \mathrm{PG}\left(1, \sqrt{\langle (\boldsymbol{w}_j^\top \boldsymbol{h}_n + c_j)^2\rangle}\right), \quad (12)$$

$$q(\gamma_k^{(1)}) = \mathrm{PG}\left(1, \sqrt{\langle b_k^2\rangle}\right). \quad (13)$$

For H: $q(h_{kn}) = \mathrm{Ber}(\sigma(d_{kn}))$, where

$$\begin{aligned} d_{kn} =& \langle b_k\rangle + \langle \boldsymbol{w}_k^\top \boldsymbol{v}_n\rangle \quad (14) \\ &- \frac{1}{2}\sum_{j=1}^{J}\left(\langle w_{jk}\rangle + \langle \gamma_{jn}^{(0)}\rangle (2\langle \psi_{jn}^{\setminus k} w_{jk}\rangle + \langle w_{jk}^2\rangle)\right), \end{aligned}$$

where $\psi_{jn}^{\setminus k} = \boldsymbol{w}_j^\top \boldsymbol{h}_n - w_{jk} h_{kn} + c_j$.

For W: $q(\boldsymbol{w}_j) = N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where

$$\boldsymbol{\Sigma}_j = \left[\sum_{n=1}^{N} \langle \gamma_{jn}^{(0)}\rangle \langle \boldsymbol{h}_n \boldsymbol{h}_n^\top\rangle + \mathrm{diag}(\langle \boldsymbol{\zeta}_j^{-1}\rangle)\right]^{-1}, \quad (15)$$

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \left[\sum_{n=1}^{N} (v_{jn} - \frac{1}{2} - \langle c_j\rangle \langle \gamma_{jn}^{(0)}\rangle)\langle \boldsymbol{h}_n\rangle\right]. \quad (16)$$

For TPBN shrinkage:

$$q(\zeta_{jk}) = \mathcal{GIG}(0, 2\langle \xi_{jk}\rangle, \langle W_{jk}^2\rangle), \quad (17)$$

$$q(\xi_{jk}) = \mathrm{Gamma}(1, \langle \zeta_{jk}\rangle + \langle \phi_k\rangle), \quad (18)$$

$$q(\phi_k) = \mathrm{Gamma}\left(\frac{1}{2}J + \frac{1}{2}, \langle \omega\rangle + \sum_{j=1}^{J}\langle \xi_{jk}\rangle\right), \quad (19)$$

$$q(\omega) = \mathrm{Gamma}\left(\frac{1}{2}K + \frac{1}{2}, 1 + \sum_{k=1}^{K}\langle \phi_k\rangle\right). \quad (20)$$