Mathematical Results in Generative Stochasitc Networks

University of Montreal
zhangsa@iro.umontreal.ca

Abstract

This report mainly focuses on some mathematical results of generative stochastic networks (GSN) [1].

1 Model Description

The general structure of GSN contains two layers, one is visible (input) layer X corresponding to input domain $\mathcal{X}$, another is hidden layer H corresponding to hidden domain $\mathcal{H}$. GSN equiped with two Markov chain transition operators $P_{\theta_1}(H|X)$ and $P_{\theta_2}(X|H)$, one can start from some (H_0, X_0) and alternatively sample H_1 by $P_{\theta_1}(H_1|X_0)$ and X_1 by $P_{\theta_2}(X_1|H_1)$ and so on to obtain the chain $H_0, X_0, H_1, X_1, ...$ The model distribution over (X, H) is defined as the stationary distribution of that chain (if any). The first transition operator is defined as $P_{\theta_1}(H|X) = f(H, X; \theta_1)$ [1] parameterized by θ_1, where $f(H, X; \theta_1)$ satisfies

$$\int_{\mathcal{H}} f(H, X; \theta_1) dH = \int_{\mathcal{H}} P_{\theta_1}(H|X) dH = 1 \tag{1}$$

The second transition operator is similarly defined as $P_{\theta_2}(X|H) = g(X, H; \theta_2)$ parameterized by θ_2. Define $\theta = \{\theta_1, \theta_2\}$. If we choose these two transition operators properly so that the Markov chain is *ergodic*, then the chain converges to an unique stationary ditribution $\pi_\theta(X, H)$, where $\pi_\theta(X, H)$ satisfies

$$\int_{\mathcal{X}\times\mathcal{H}} \pi_\theta(X, H) P_{\theta_1}(H'|X) P_{\theta_2}(X'|H') dX dH = \pi_\theta(X', H') \tag{2}$$

Here the one-step transition operator $\mathcal{T}_\theta(X', H'|X, H)$ between (X, H) and (X', H') is

$$\mathcal{T}_\theta(X', H'|X, H) = P_{\theta_1}(H'|X) P_{\theta_2}(X'|H') \tag{3}$$

For discrete state space, eq.2 corresponds to $\pi_\theta = \pi_\theta \cdot \mathbf{T}_\theta$, where $\mathbf{T}_\theta$ is the transition matrix equivalent to $\mathcal{T}_\theta(X', H'|X, H)$.

2 Main Results

All the following results are based on the GSN model mentioned in last section.

Thereom 1. *A GSN Markov chain (X_t, H_t) is defined by alternatively sampling H_{t+1} using $P_{\theta_1}(H_{t+1}|X_t)$ and sampling X_{t+1} using $P_{\theta_2}(X_{t+1}|H_{t+1})$. Assume that the chain is ergodic and has a unique stationary distribution $\pi_\theta(X, H)$ where $\theta = \{\theta_1, \theta_2\}$. $\pi_\theta(X|H)$ is the conditional marginalized from $\pi_\theta(X, H)$. Then we have*

$$P_{\theta_2}(X|H) = \pi_\theta(X|H) \tag{4}$$

[1] In a more general case, P_{θ_1} can be $P_{\theta_1}(H'|X, H) = f(H', X, H; \theta_1)$, H' is the new sample computed not only using X but also old H. Further discussion see

Proof. From Eq.2 we have

$$\begin{aligned}
\pi_\theta(X', H') &= \int_{\mathcal{X}\times\mathcal{H}} \pi_\theta(X, H) P_{\theta_1}(H'|X) P_{\theta_2}(X'|H') dX dH \\
&= P_{\theta_2}(X'|H') \int_{\mathcal{X}} \int_{\mathcal{H}} \pi_\theta(X, H) P_{\theta_1}(H'|X) dX dH \\
&= P_{\theta_2}(X'|H') \int_{\mathcal{X}} (\int_{\mathcal{H}} \pi_\theta(X, H) dH) P_{\theta_1}(H'|X) dX \\
&= P_{\theta_2}(X'|H') \int_{\mathcal{X}} \pi_\theta(X) P_{\theta_1}(H'|X) dX \qquad (5)
\end{aligned}$$

Above Eq.5 is equivalent to

$$\frac{\pi_\theta(X', H')}{P_{\theta_2}(X'|H')} = \int_{\mathcal{X}} \pi_\theta(X) P_{\theta_1}(H'|X) dX \qquad (6)$$

Note that the right side of Eq.6 is a function only for H' (because X is integrated over $\mathcal{X}$), so we rewrite the left side of Eq.6 as $k(H')$

$$k(H') = \int_{\mathcal{X}} \pi_\theta(X) P_{\theta_1}(H'|X) dX \qquad (7)$$

Now for $P_{\theta_2}(X'|H')$ we have

$$P_{\theta_2}(X'|H') = \frac{\pi_\theta(X', H')}{k(H')} \qquad (8)$$

From the definition of $P_{\theta_2}(X'|H')$ we have

$$\begin{aligned}
1 &= \int_{\mathcal{X}} P_{\theta_2}(X'|H') dX' \\
&= \int_{\mathcal{X}} \frac{\pi_\theta(X', H')}{k(H')} dX' \\
&= \frac{\pi_\theta(H')}{k(H')} \qquad (9)
\end{aligned}$$

Finally from above we get $\pi_\theta(H') = k(H')$, so we have

$$P_{\theta_2}(X'|H') = \frac{\pi_\theta(X', H')}{k(H')} = \frac{\pi_\theta(X', H')}{\pi_\theta(H')} = \pi_\theta(X'|H') \qquad (10)$$

Obviously if one changes the notion from X', H' to X, H, the same conclusion still holds. □

Corollary 1.1. *For the same GSN Markov chain in Theorem 1, for $P_{\theta_1}(H'|X)$ we have*

$$\int_{\mathcal{X}} \pi_\theta(X) P_{\theta_1}(H|X) dX = \pi_\theta(H) \qquad (11)$$

Proof. Trivial based on Eq.7 and Eq.9 in proof of Theorem 1. □

Thereom 2. *For the same GSN Markov chain (X_t, H_t) in Theorem 1, define $X'_t = X_t$ and $H'_t = H_{t+1}$, then (X'_t, H'_t) is another Markov chain that its stationary distribution is $\pi'_\theta(X, H)$. This new chain is equivalently obtained by alternatively sampling X'_{t+1} using $P_{\theta_2}(X'_{t+1}|H'_t)$ and sampling H'_{t+1} using $P_{\theta_1}(H'_{t+1}|X'_{t+1})$. Furthermore, we have*

$$P_{\theta_1}(H|X) = \pi'_\theta(H|X) \qquad (12)$$

$$\int_{\mathcal{H}} \pi'_\theta(H) P_{\theta_2}(X|H) dH = \pi'_\theta(X) \qquad (13)$$

$$\pi'_\theta(X) = \pi_\theta(X) \qquad (14)$$

$$\pi'_\theta(H) = \pi_\theta(H) \qquad (15)$$

Proof. Suppose that the original chain generated from Theorem 1 is

$$H_0, X_0, H_1, X_1, H_2, \cdots, X_{t-1}, H_t, X_t, \cdots \tag{16}$$

where H_{t+1} is sampled based on $P_{\theta_1}(H_{t+1}|X_t)$ and X_{t+1} is sampled based on $P_{\theta_1}(X_{t+1}|H_{t+1})$. $\pi_\theta(X, H)$ is the stationary distribution of this chain if one groups H_t and X_t to (X_t, H_t). Now by definition we have

$$X'_t = X_t \tag{17}$$

$$H'_t = H_{t+1} \tag{18}$$

If one deletes H_0 from the original chain, then the rest of the original chain (16) is equivalent to the new chain as following

$$X'_0, H'_0, X'_1, H'_1, \cdots, X'_{t-1}, H'_{t-1}, X'_t, \cdots \tag{19}$$

This new chain can be generated by alternatively sampling X'_{t+1} by $P_{\theta_2}(X'_{t+1}|H'_t)$ and sampling H'_{t+1} by $P_{\theta_1}(H'_{t+1}|X'_{t+1})$. To see this, on can check that in original chain X_{t+1} is generated by $P_{\theta_2}(X_{t+1}|H_{t+1})$, while we have $X'_{t+1} = X_{t+1}$ and $H_{t+1} = H'_t$, so $P_{\theta_2}(X_{t+1}|H_{t+1})$ is equivalent to $P_{\theta_2}(X'_{t+1}|H'_t)$ which means that $X'_{t+1} = X_{t+1}$ is sampled using $P_{\theta_2}(X'_{t+1}|H'_t)$. Following the same way one can illustrate that H'_{t+1} is sampled using $P_{\theta_1}(H'_{t+1}|X'_{t+1})$. By definition the new chain's stationary distribution is $\pi'_\theta(X, H)$, then Eq.12 and Eq.13 are direct conclusions from Theorem 1 and Corollary 1.1. When t goes to infinity, the ditribution over X'_∞ can be marginalized from $\pi'_\theta(X, H)$, which is $\pi'_\theta(X)$. And we also have that the distribution over X_∞ is $\pi_\theta(X)$. Because $X'_t = X_t$, so $\pi'_\theta(X)$ is the same distribution as $\pi_\theta(X)$. Similar results for $\pi'_\theta(H)$ and $\pi_\theta(H)$. □

Remark. Theorem 1 and Theorem 2 prove that for a GSN Markov chain, there are two different stationary distributions over $\mathcal{X} \times \mathcal{H}$, depending on grouping X and H by (H_t, X_t) or (X_t, H_{t+1}). At the same time, these two distributions have the same marginal for X and H. If these two stationary distributions become equal, then we are sampling by Gibbs sampling.

Lemma 3.1 *(Sampling from Conditional) Assume we have a distribution $P_\theta(X) = f(X; \theta), X \in \mathcal{X}$ which we can directly sample X from. Then there is always a way to directly sample from any of its conditionals like $P_\theta(X|X \in \mathcal{S}), \mathcal{S} \subseteq \mathcal{X}$.*

Proof. $\mathcal{S} = \mathcal{X}$ is trivial. When $\mathcal{S} \subset \mathcal{X}$, we still directly sample X using $f(X; \theta)$, however, we accept this X only when $X \in \mathcal{S}$ satisfies, otherwise we reject this X and directly sample X again using $f(X; \theta)$. By this process, one can easily check that all the accepted Xs are equivalently sampled from condtional $P_\theta(X|X \in \mathcal{S}), \mathcal{S} \subseteq \mathcal{X}$. □

Thoerem 3. *(Clamping Theorem) Assume we have an ergodic GSN Markov chain equiped with tranisition operator $P_{\theta_1}(H|X)$ and $P_{\theta_2}(X|H)$, its unique stationary distribution is $\pi_\theta(X, H)$. Suppose that the chain starts from (H_0, X_0) where $X_0 \in \mathcal{S}$, $\mathcal{S} \subseteq \mathcal{X}$ ($\mathcal{S}$ can be considered as constraint over X). If $P_{\theta_1}(H|X)$ satisfies*

$$\int_{\mathcal{S}} \pi_\theta(X|X \in \mathcal{S}) P_{\theta_1}(H'|X) dX = \pi_\theta(H'|X \in \mathcal{S}) \tag{20}$$

where $\pi_\theta(X|X \in \mathcal{S})$ and $\pi_\theta(H'|X \in \mathcal{S})$ are conditionals that

$$\pi_\theta(X|X \in \mathcal{S}) = \frac{\pi_\theta(X)}{\int_{\mathcal{S}} \pi_\theta(X') dX'}, \quad \pi_\theta(H'|X \in \mathcal{S}) = \frac{\int_{\mathcal{S}} \pi_\theta(X, H') dX}{\int_{\mathcal{S} \times \mathcal{H}} \pi_\theta(X, H) dX dH} \tag{21}$$

and we sample (X_{t+1}, H_{t+1}) by first sample H_{t+1} with $P_{\theta_1}(H_{t+1}|X_t)$ and then sample X_{t+1} with "Sampling from Conditional" method in Lemma 3.1 on $P_{\theta_2}(X_{t+1}|H_{t+1}, X_{t+1} \in \mathcal{S})$. If the new chain is ergodic, then the unique stationary distribution of this new chain is

$$\pi_\theta(X, H|X \in \mathcal{S}) \tag{22}$$

In another word, if we marginalize over H, when the chain converges we are just sampling X from conditional

$$\pi_\theta(X|X \in \mathcal{S}) \tag{23}$$

Proof. Suppose that the stationary distribution of this new chain is $\pi_{\mathcal{S}}(X, H)$, then according to the definition of stationary distribution, $\pi_{\mathcal{S}}(X, H)$ is the unique distribution that satisfies

$$\int_{\mathcal{S}\times\mathcal{H}} \pi_{\mathcal{S}}(X, H) P_{\theta_1}(H'|X) P_{\theta_2}(X'|H', X' \in \mathcal{S}) dX dH = \pi_{\mathcal{S}}(X', H') \tag{24}$$

Now let us check if $\pi_\theta(X, H|X \in \mathcal{S})$ satisfies the equation above. According to Theorem 1 and Lemma 3.1, we have

$$P_{\theta_2}(X'|H', X' \in \mathcal{S}) = \pi_\theta(X'|H', X' \in \mathcal{S}) \tag{25}$$

If we use $\pi_\theta(X, H|X \in \mathcal{S})$ to substitute $\pi_{\mathcal{S}}(X, H)$ in Eq.24, the left side of Eq.24 becomes

$$\begin{aligned}
& \int_{\mathcal{S}\times\mathcal{H}} \pi_\theta(X, H|X \in \mathcal{S}) P_{\theta_1}(H'|X) \pi_\theta(X'|H', X' \in \mathcal{S}) dX dH \\
= \quad & \pi_\theta(X'|H', X' \in \mathcal{S}) \int_{\mathcal{S}} (\int_{\mathcal{H}} \pi_\theta(X, H|X \in \mathcal{S}) dH) P_{\theta_1}(H'|X) dX \\
= \quad & \pi_\theta(X'|H', X' \in \mathcal{S}) \int_{\mathcal{S}} \pi_\theta(X|X \in \mathcal{S}) P_{\theta_1}(H'|X) dX \\
= \quad & \pi_\theta(X'|H', X' \in \mathcal{S}) \pi_\theta(H'|X \in \mathcal{S}) & (26) \\
= \quad & \pi_\theta(X'|H', X' \in \mathcal{S}) \pi_\theta(H'|X' \in \mathcal{S}) \\
= \quad & \pi_\theta(X', H'|X' \in \mathcal{S}) & (27)
\end{aligned}$$

Here we use Eq.20 to get Eq.26. Equations above show that $\pi_\theta(X, H|X \in \mathcal{S})$ satisfies

$$\int_{\mathcal{S}\times\mathcal{H}} \pi_\theta(X, H|X \in \mathcal{S}) P_{\theta_1}(H'|X) P_{\theta_2}(X'|H', X' \in \mathcal{S}) dX dH = \pi_\theta(X', H'|X' \in \mathcal{S}) \tag{28}$$

while the distribution $\pi_{\mathcal{S}}(X, H)$ satisfied equations above is unique, so we have $\pi_{\mathcal{S}}(X, H) = \pi_\theta(X, H|X \in \mathcal{S})$. $\square$

References

[1] Bengio, Y., Thibodeau-Laufer, E., Alain, G., & Yosinski, J. (2013). Deep generative stochastic networks trainable by backprop. arXiv preprint arXiv:1306.1091.