Deep RBM states

Hu Haoyu

September 23, 2019

1 Introduction

1.1 Network state

Hilbert space $\mathcal{H}$ has dimension d^N. d is the dimension of the local Hilbert space and N is the number of the sites

The quantum state is a vector in the $\mathcal{H}$. Use the Fock space representation of the $\mathcal{H}$. A general quantum state is defined as

$$|\psi >= \sum_n p_\psi(n)|n >$$

The neural network with parameter w is defined as the mapping $P : \mathcal{H} \to \mathbb{C}$. Then each network defines an unique quantum state

$$|\psi >= \sum_n \rho(n; w)|n >$$

1.2 Hamiltonian and ground state

The Hamiltonian is defined as the linear mapping $H : \mathcal{H} \to \mathcal{H}$. The eigenstates and eignvalue of H:

$$H|\psi >= E_\psi|\psi >$$

The ground state $|GS >$ is the eigenstate of H with lowest energy E_G

1.3 Observable

The energy of the network state is

$$\begin{aligned}
E_w &= \frac{< \psi|H|\psi >}{< \psi|\psi >} \\
&= \sum_n \bar{\rho}(n; w)\frac{\sum_{n'} H_{n,n'}\rho(n'; w)}{\sum_{n'} |\rho(n', w)|^2} \\
&= \left[\sum_n |\rho(n; w)|^2 \sum_{n'} \frac{H_{n,n'}\rho(n'; w)}{\rho(n; w)}\right] \Big/ \left[\sum_{n'} |\rho(n', w)|^2\right] \\
&= < E(n) >
\end{aligned}$$

where

$$\begin{aligned} E(n) &= \sum_{n'} \frac{H_{n,n'}\rho(n';w)}{\rho(n;w)} \\ H_{n,n'} &= <n|H|n'> \end{aligned}$$

And $< . >$ is the statistical average w.r.t. the probability distribution

$$p(n) = \frac{|\rho(n;w)|^2}{\sum_{n'} |\rho(n',w)|^2}$$

1.4 Minimize energy

To find the ground state, one only need to minimize E_w by varying the parameters of network w.
Using the gradient decent method, consider

$$\begin{aligned} \partial_w E_w &= < E(n)\bar{D}_w(n) + D_w(n)\bar{E}(n) > - < E(n) >< \bar{D}_w(n) > - < D_w(n) >< \bar{E}(n) > \\ &= 2\Re\bigg\{ < E(n)\bar{D}_w(n) > - < E(n) >< \bar{D}_w(n) > \bigg\} \end{aligned}$$

with

$$D_w(n) = \frac{\partial_w \rho(n;w)}{\rho(n;w)}$$

1.5 Algorithm

Optimize algorithm:
1. Initialize w, compute $H_{n,n'}$
2. Compute $E(n)$ and $D_w(n)$ for each n w.r.t. w.
3. Sampling. Calculate $< E(n) >$, $< \bar{D}_w(n) >$ and $< E(n)\bar{D}_w(n) >$
4. Update w, $w \to w - \lambda\partial_w E_w$
5. Go to step 2. Stop until converge.

Time complexity and space complexity. Let $D = d^N$ is the dimension of Hilbert space and L is the time complexity of the neural network.

For each sweep,
1.compute $P(n;w)$, the time complexity $o(DL)$ and space complexity $o(D)$.
2.compute $E(n)$, $D_w(n)$ from $P(n;w)$. Time: $o(D^2)$ and space $o(D)$.
3.Sampling. Use Metropolis or Gibbs

2 Restricted Boltzmann Machine

2.1 Introduction

Visible units n and hidden unites h. n is a D_v dimensional vector(binary) and h is a D_h dimensional vector(binary). Probability distribution is $P(n,h)/Z$

$$P(n,h) = \exp\Bigg[- n^T W h - a^T n - b^T h \Bigg]$$

The distribution of n is $p(n) = \frac{1}{Z} \sum_h P(n,h)$.

Conditional probabilities of the RBM have the following nice structure

$$\begin{aligned} P(n|h) &= \Pi_i P(n_i|h) \\ P(h|n) &= \Pi_i P(h_i|n) \end{aligned}$$

For the binary vector

$$\begin{aligned} P(n_i = 1|h) = sigmoid(-b_i - \sum_j W_{ij} h_j) \\ P(h_i = 1|n) = sigmoid(-a_i - \sum_j n_j W_{ji}) \end{aligned}$$

The distribution of n is

$$\begin{aligned} p(n) &= \frac{P(n)}{Z} \\ P(n) &= e^{-a^T n} \Pi_j \Bigg[\exp(-\sum_i n_i W_{ij} - a_j) + 1 \Bigg] \end{aligned}$$

Derivative

$$\begin{aligned} \partial_{W_{ij}} P(n) &= \sum_h (-n_i h_j) P(n,h) \\ \partial_{a_i} P(n) &= \sum_h (-n_i) P(n,h) \\ \partial_{b_i} P(n) &= \sum_h (-h_i) P(n,h) \end{aligned}$$

2.2 RBM state

We use the Boltzmann machine to define the RBM quantum state

$$|RBM >= \sum_n \sum_h p(n,h) |n >$$

Unlike the usual RBM, the parameters of the RBM here are complex variables.

Similarly, we can define the deep RBM (DRB) as

$$|DRB >= \sum_n \sum_h p(n,h) \sum_d q(h,d)|n >$$

where hs are the hidden units of the 1st layer and ds are the hidden units for the deep layer.

2.3 Observable

For the observable of $\hat{O}$

$$\begin{aligned} < \hat{O} > &= \frac{< DRB|\hat{O}|DRB >}{< DRB|DRB >} \\ &= \frac{\sum_{n,n',h,h',d,d'} O_{n,n'} \bar{p}(n,h,d) p(n',h',d')}{\sum_{n,n',h,h',d,d'} \bar{p}(n,h,d) p(n',h',d')} \\ &= \frac{F[O]}{F[I_{nn'}]} \end{aligned}$$

Then

$$\begin{aligned} F[O] &= \sum_{n,n',h,h',d,d'} |p(n,h,d)|^2 |p(n',h',d')|^2 \frac{O_{n,n'}}{p(n,h,d)\bar{p}(n',h',d')} \\ &= Z^2 \sum_{n,h,d,n',h',d'} P(n,h,d) P(n',h',d') \tilde{O}_{n,h,d;n',h',d'} \end{aligned}$$

where

$$P(n,h,d) = \frac{1}{Z} |p(n,h,d)|^2$$

is the joint probability distribution of $n, h, d.$ and

$$\tilde{O}_{n,h,d;n',h',d'} = \frac{O_{n,n'}}{p(n,h,d)\bar{p}(n',h',d')}$$

Similarly

$$F[I] = Z^2 \sum_{n,h,d,n',h',d'} P(n,h,d) P(n',h',d') \tilde{I}_{n,h,d;n',h',d'}$$

with

$$\tilde{I}_{n,h,d;n',h',d'} = \frac{\delta_{n,n'}}{p(n,h,d)\bar{p}(n',h',d')}$$

2.4 parameter of the DRB

For the general DRB, the energy function and "probability" distribution is

$$\begin{aligned} E(n,h,d) &= n^T W h + h^T M d + a^T n + b^T h + c^T d \\ p(n,h,d) &= \exp(-E(n,h,d)) \end{aligned}$$

The joint probability distribution we need to sample with is

$$P(n,h,d) = \frac{1}{Z} \exp(-E(n,h,d) - \bar{E}(n,h,d))$$

Note tha $p(n,h,d)$ are complex numbers but $P(n,h,d)$ are real numbers and the distribution $P(n,h,d)$ still has the form of DRB.

2.5 Gradient

In order to optimize the variable, we need to take the gradient of the $< \hat{O} >$

$$\begin{aligned} \partial < O > &= \frac{\partial F[O]F[I] - F[O]\partial F[I]}{F[I]^2} \\ &= \frac{\mathbb{E}[[\tilde{O}_{l,l'} \cdot D_{l,l'}]]\mathbb{E}[[\tilde{I}_{l,l'}]] - \mathbb{E}[[\tilde{O}_{l,l'}]]\mathbb{E}[[\tilde{I}_{l,l'} \cdot D_{l,l'}]]}{\mathbb{E}[[\tilde{I}_{l,l'}]]^2} \\ < O > &= \frac{\mathbb{E}[[\tilde{O}_{l,l'}]]}{\mathbb{E}[[\tilde{I}_{l,l'}]]} \end{aligned}$$

where the expectation value is taken w.r.t. the joint distribution $\Pi_{l,l'} = P(n,h,d)P(n',h',d')$ and $l = (n,h,d)$. Also

$$D_{l,l'} = \frac{\partial \bar{p}(n,h,d)}{\bar{p}(n,h,d)} + \frac{\partial p(n',h',d')}{p(n',h',d')}$$

Note that $\frac{\partial p}{p}$ usually has a simple form

2.5.1 Gradient for 3-layer RBM

$$p(n,h,d) = 1/\exp(n^T W h + h^T M d + a^T n + b^T h + c^T d)$$

The gradient

$$\begin{aligned}
\frac{\partial_{W_{ij}} p(l)}{p(l)} &= -n_i h_j \\
\frac{\partial_{M_{ij}} p(l)}{p(l)} &= -h_i d_j \\
\frac{\partial_{a_i} p(l)}{p(l)} &= -n_i \\
\frac{\partial_{b_i} p(l)}{p(l)} &= -h_i \\
\frac{\partial_{v_i} p(l)}{p(l)} &= -d_i
\end{aligned}$$

2.6 Translational Invariance

For RBM

$$\begin{aligned}
p(n,h) &= \exp(-E) \\
E &= \sum_i \sum_T \left[n_{T(i)} W_{ij} h_j + a_i n_{T(i)} + b_i h_i \right] \\
&= \sum_i \left[n_i \tilde{W}_{ij} h_j + \tilde{a}_i n_i + b_i h_i \right]
\end{aligned}$$

where T is the translational operator and $\tilde{W}_{ij} = \sum_T W_{T(i),j}, \tilde{a}_i = \sum_T a_{T(i)}$. We can see to make the RBM translational invariant, we only need to make the parameter of the RBM to be translational invariant w.r.t. the index of visible units

Note that, when we enforce the translational invariance, we can actually reduce the number of the parameters.

3 Sampling

To compute the expectation value, we can use Gibbs sampling or Metropolis sampling

3.1 Gibbs Sampling

Sample $(x_1, x_2, ..., x_n)$ w.r.t. $p(x_1, x_2, x_3, ..., x_n)$.

1. Initialize x_i^0.
2. for $i = 0, i < n$

sample x_i^{t+1} w.r.t. $p(x_i|x_0^{t+1}, ..., x_{i-1}^{t+1}, x_{i+1}^t, ..., x_n^t)$.
3. Go to 2.
4. Get the sequence $\{\mathbf{x}^t\}_{t=0}^n$

3.2 Gibbs sampling for DRB

Aim is to get the Markov chain variables $\{n_t, h_t, d_t\}|_t$

1. sample n_{t+1} w.r.t. $P(n|h_t, d_t)$
2. sample h_{t+1} w.r.t. $P(h|n_{t+1}, d_t)$
3. sample d_{t+1} w.r.t. $P(d|n_{t+1}, h_{t+1})$
4. t=t+1, go to 1

where the probability distributions are

$$\begin{aligned}
P(n_i = 0|h, d) &= sigmoid((W_{ij} + \bar{W}_{ij})h_j + a_i + \bar{a}_i) \\
P(h_i = 0|n, d) &= sigmoid(n_j(W_{ji} + \bar{W}_{ji}) + (M_{ij} + \bar{M}_{ij})d_j + b_i + \bar{b}_i) \\
P(d_i = 0|n, h) &= sigmoid(h_j(M_{ji} + \bar{M}_{ji}) + c_i + \bar{c}_i)
\end{aligned}$$

3.3 Metropolis Sampling

Generate Markov chain $\{n_t, h_t, d_t\}|_t$ by Metropolis Algorithm

1. Initialize
2. Generate $\tilde{l}$ w.r.t. the probability $g(\tilde{l}|l)$
3. Acceptance ratio $A_{l,\tilde{l}} = \min\{1, \frac{P(\tilde{l})g(l|\tilde{l})}{P(l)g(\tilde{l}|l)}\}$
4. Accept or reject. Goto 2

3.4 Others

We can also trace some or all of the hidden units to get the probability distribution of the remaining variables in the close form. Then we also need to modify the $\tilde{O}, \tilde{I}$ correspondingly. Then we can sample w.r.t. the new probability distribution and new $\tilde{O}, \tilde{I}$. This procedure may potentially increase the accuracy of the algorithm. The reason is that tracing out the variable is equivalent to integrating out the variable with the given probability. The sampling with such variables is an approximation of the original integral.

Works when only have one hidden layer. In this case we can directly trace out the hidden layer.

For the three layer case(1 visible, 2 hidden). We can trace out the visible layer. (check)

4 Alternative Formulation

4.1 RBM state

We use the Boltzmann machine to define the RBM quantum state

$$|RBM >= \sum_n \sum_h p(n,h)|n >$$

Unlike the usual RBM, the parameters of the RBM here are complex variables.

Similarly, we can define the deep RBM (DRB) as

$$|DRB >= \sum_n \sum_h p(n,h) \sum_d q(h,d)|n >$$

where hs are the hidden units of the 1st layer and ds are the hidden units for the deep layer.

4.2 Observable

For the observable of $\hat{O}$

$$\begin{aligned} < \hat{O} > &= \frac{< DRB|\hat{O}|DRB >}{< DRB|DRB >} \\ &= \frac{\sum_{n,n',h,h',d,d'} O_{n,n'}\bar{p}(n,h,d)p(n',h',d')}{\sum_{n,h,h',d,d'} \bar{p}(n,h,d)p(n,h',d')} \\ &= \frac{F[O]}{F[I_{nn'}]} \end{aligned}$$

Then

$$\begin{aligned} F[O] &= \sum_{n,h,d} |p(n,h,d)|^2 \left[\sum_{n',h',d'} p(n',h',d') \frac{O_{n,n'}}{p(n,h,d)} \right] \\ &= Z \sum_{n,h,d} P(n,h,d)P(n',h',d')\tilde{O}_{n,h,d} \end{aligned}$$

where

$$P(n,h,d) = \frac{1}{Z}|p(n,h,d)|^2$$

is the joint probability distribution of $n, h, d.$ and

$$\tilde{O}_{n,h,d} = \sum_{n',h',d'} O_{n,n'} \frac{p(n',h',d')}{p(n,h,d)}$$

Similarly

$$F[I] = Z \sum_{n,h,d} P(n,h,d)\tilde{I}_{n,h,d}$$

with

$$\tilde{I}_{n,h,d} = \sum_{n',h',d'} \delta_{n,n'} \frac{p(n',h',d')}{p(n,h,d)}$$

4.3 parameter of the DRB

For the general DRB, the energy function and "probability" distribution is

$$\begin{aligned} E(n,h,d) &= n^T W h + h^T M d + a^T n + b^T h + c^T d \\ p(n,h,d) &= \exp(-E(n,h,d)) \end{aligned}$$

The joint probability distribution we need to sample with is

$$P(n,h,d) = \frac{1}{Z} \exp(-E(n,h,d) - \bar{E}(n,h,d))$$

Note tha $p(n,h,d)$ are complex numbers but $P(n,h,d)$ are real numbers and the distribution $P(n,h,d)$ still has the form of DRB.

4.4 Gradient

In order to optimize the variable, we need to take the gradient of the $< \hat{O} >$

$$\begin{aligned} < O > &= \frac{\mathbb{E}[E_l]}{\mathbb{E}[I_l]} \\ \partial < O > &= \frac{\partial F[O] F[I] - F[O] \partial F[I]}{F[I]^2} \\ &= \frac{\mathbb{E}[I_l]\mathbb{E}[\bar{D}_l H_l + h.c.] - \mathbb{E}[H_l]\mathbb{E}[\bar{D}_l I_l + h.c.]}{(\mathbb{E}[I_l])^2} \\ &= \frac{\mathbb{E}[I_l]\mathbb{E}[2\Re[\bar{D}_l H_l]] - \mathbb{E}[H_l]\mathbb{E}[2\Re[\bar{D}_l I_l]]}{(\mathbb{E}[I_l])^2} \end{aligned}$$

where the expectation value is taken w.r.t. the distribution P_l and $l = (n,h,d)$. Also

$$\begin{aligned} O_l &= \sum_{l'} O_{l,l'} \frac{p_{l'}}{p_l} \\ D_l &= \frac{\partial p_l}{p_l} \end{aligned}$$

5 Alternative Formulation- 2

The idea is integrate out the physical state and sample the states of the hidden units.

$$\begin{aligned} < O > &= \frac{\mathbb{E}[\tilde{O}]}{\mathbb{E}[\tilde{I}]} \\ O^M_{h,h'} &= \sum_{n,n'} O_{n,n'} \bar{p}_{n,h} p_{n',h'} \\ \tilde{O}_{h,h'} &= \frac{O^M_{h,h'}}{p_{h,d} \bar{p}_{h',d'}} \end{aligned}$$

The expectation value is taken w.r.t. the distribution $|p_{h,d}|^2$

5.1 Partial derivative of the hidden unit

partial derivative for the parameters of the hidden unit:

$$\partial < O >= \frac{\mathbb{E}[[\tilde{O}_{h,h'} \cdot D_{h,d,h',d'}]]\mathbb{E}[[\tilde{I}_{h,h'}]] - \mathbb{E}[[\tilde{O}_{h,h'}]]\mathbb{E}[[\tilde{I}_{h,h'} \cdot D_{h,d,h',d'}]]}{\mathbb{E}[[\tilde{I}_{h,h'}]]^2}$$

where the expectation value is taken w.r.t. the joint distribution $\Pi_{h,d,h',d'} = P(h,d)P(h',d')$

$$D_{h,d,h',d'} = \frac{\partial \bar{p}(h,d)}{\bar{p}(h,d)} + \frac{\partial p(h',d')}{p(h',d')}$$

5.2 Partial derivative of the visible unit

$$\begin{aligned} \partial < O > &= \frac{\mathbb{E}[\partial \tilde{O}]\mathbb{E}[\tilde{I}] - \mathbb{E}[\tilde{O}]\mathbb{E}[\partial \tilde{I}]}{\mathbb{E}[\tilde{I}]^2} \\ \partial \tilde{O} &= \frac{\partial O^M_{h,h'}}{p_{h,d} \bar{p}_{h',d'}} \end{aligned}$$

6 Entanglement Entropy

Renyi Entropy

$$S_A = \frac{1}{1-\alpha} \log(Tr[\rho_A^\alpha])$$

Take $\alpha = 2$

$$S_A = -\log(Tr[\rho_A^2])$$

The neural network state

$$|\phi >= \sum_h p(n_A|h) p(n_B|h) p(h) |n_A > |n_B >$$

$$
\begin{aligned}
Tr[\rho_A^2] \quad &= \quad \frac{1}{A^2} \sum_{h_1,h_2,h_3,h_4} p(h_1)p(h_3)\bar{p}(h_2)\bar{p}(h_4) \\
& \prod_{n_A} p(n_A|h_1)\bar{p}(n_A|h_4) \\
& \prod_{n'_A} \bar{p}(n'_A|h_2)p(n'_A|h_3) \\
& \prod_{n_B} p(n_B|h_1)\bar{p}(n_B|h_2) \\
& \prod_{n'_B} p(n'_B|h_3)\bar{p}(n'_B|h_4)
\end{aligned}
$$

Normalization

$$
\begin{aligned}
A =< \phi|\phi > \quad &= \quad \sum_{h_1,h_2} \sum_n p(n|h_1)\bar{p}(n|h_2)p(h_1)\bar{p}(h_2) \\
&= \quad \sum_{h_1,h_2} p(h_1)\bar{p}(h_2) \prod_n p(n|h_1)\bar{p}(n|h_2)
\end{aligned}
$$