Deep Learning 2024011303 Liu Hanzuo

HW2

True or False

P1 True, we can set part of the parameters to the input photo and take another neuron's output as the result of classification to classify; and initialize one parameter by the prompt and generate other nueorns as a picture to generation.

P2 True, we shell approximate Z for a faster training.

P3 False, MCMC is used to solve the optimization problem of $\arg\max_\theta \mathbb{P}(\theta|X)$

QA

P4 **1.** For noisy image's denoise tasks, we can initialize the EBM by the parameters of the picture. Do the process and it will converge to the true distribution.
For the image's masked task, we could initialize the unmasked part by the picture and zero for the masked part, it will converge to the true distribution.
2. From Hebbian we can denote that

$$W_{i,j} = \frac{1}{N}\sum_{p=1}^{N} y_p^i y_p^j = \frac{1}{N} Y {Y^T}_{i,j}$$

Thus for any y we have

$$E = -\sum_{i<j} W_{i,j} y^i y^j = 0$$

P5 let

$$Z = \sum_{v,h} \exp\left(y^T W y\right)$$

then we have

$$\nabla_W L(W) = -\frac{1}{|P|}\sum_{v\in P} \nabla_W P(v) \cdot \frac{1}{P(v)}$$

$$= -\frac{1}{|P|}\sum_{v\in P} \frac{1}{P(v)} \sum_{h} P(v,h) \nabla_W \log P(v,h)$$

Note that

$$\nabla_W \log P(v,h) = y^T y - \frac{\sum_{y'} \exp\left(y'^T W y'\right) y'^T y'}{\sum_{y'} \exp\left(y'^T W y'\right)}$$

Thus we have

$$\nabla_W L(W) = -\frac{1}{|P|}\sum_{v\in P}\left(\mathbb{E}_{h|v}[yy^T] - \sum_{y'}\frac{\exp\left(y'^T W y'\right)}{\sum_{y''}\exp\left(y''^T W y''\right)} y'^T y'\right)$$

$$= -\frac{1}{|P|}\sum_{v\in P}\left(\mathbb{E}_{h|v}[yy^T] - \mathbb{E}_{y'}[y'y'^T]\right)$$

P6 Note that

$$P(v,h) = \frac{1}{Z}\exp\left(-E_{W,b}(v,h)\right)$$

$$P(v|h) = \frac{P(v,h)}{\int_h P(v,h)dh} = \frac{\exp\left(-\frac{1}{2}(v-b)^T(v-b) + v^T W h\right)}{\int_h \exp\left(-\frac{1}{2}(v-b)^T(v-b) + v^T W h\right) dh}$$

For calculation part, let $x = v - b$,$c = Wh$

$$\int \exp\left(-\frac{1}{2}x^T x + x^T c\right) dx = \int\int\cdots\int \exp\left(-\frac{1}{2}\sum_{i=1}^{N_v} x_i^2 + \sum_{i=1}^{N_v} x_i c_i\right) dx_1 dx_2 \cdots dx_{N_v}$$

$$= \exp\left(\frac{1}{2}\sum_{i=1}^{N_v} c_i^2\right)\cdot(2\pi)^{\frac{N_v}{2}}.$$

Thus, we have

$$P(v|h) = (2\pi)^{-\frac{N_v}{2}}\exp\left(-\frac{1}{2}h^T W^T W h - \frac{1}{2}(v-b)^T(v-b) + (v-b)^T W h\right)$$

$$= (2\pi)^{-\frac{N_v}{2}}\exp\left(-\frac{1}{2}\|v - b - Wh\|^2\right).$$

P7 **1.**Note that

$$P(D|F) = \int_{A,B,C,E}\frac{1}{Z}f_{AD}(A,D)f_{AC}(A,C)f_{AE}(A,E)f_{BC}(B,C)f_{EF}(E,F)dAdBdCdE$$

$$= \int_{A,E}\frac{1}{Z'}f_{AD}(A,D)f_{EF}(E,F)$$

Note that we also have

$$P(D) = \int_A \frac{1}{Z''}f_{AD}(A,D)$$

When $f_{EF}(E,F)$ is not same(for instance, be a one-hot vector), thus

$$P(D|F) \neq P(D)$$

Which indicates that D and F are not independent!

2.

$$P(B,E|A) = \int_{C,D,F} \frac{1}{Z} f_{AD}(A,D) f_{AC}(A,C) f_{AE}(A,E) f_{BC}(B,C) f_{EF}(E,F) dC dD dE$$

$$= \frac{1}{Z} f_{AE}(A,E) \cdot u(A,B) \cdot v(A) \cdot h(E)$$

Notation:

$$u(A,B) = \int_C f_{AC}(A,C) f_{BC}(B,C) dC$$

To conclude, B and E are independent given A.(no entanglement between B and E).

3. We have

$$\log P(A,B,C,D,E,F) = -E(A,B,C,D,E,F) + C_2,$$

where C_2 is a constant independent of $A, B, \ldots, F$. Thus, we can have

$$E(A,B,C,D,E,F) = C_1 - \log f_{AD}(A,D) - \log f_{AC}(A,C) - \log f_{AE}(A,E) - \log f_{BC}(B,C) - \log f_{EF}(E,F)$$

$$= E_{AD}(A,D) + E_{AC}(A,C) + E_{AE}(A,E) + E_{BC}(B,C) + E_{EF}(E,F).$$

P8 Note that we have

$$Var\left(\frac{p(x)}{q(x)} f(x)\right) = \mathbb{E}_{x\sim q}\left((\frac{p(x)}{q(x)})^2 f(x)^2 \right) - \left(\mathbb{E}_{x\sim q} \frac{p(x)}{q(x)} f(x) \right)^2$$

the later term equals to

$$\left(\mathbb{E}_{x\sim p} f(x)\right)^2$$

which is a constant. Here our problem is simplified to minimize

$$\int_x \frac{p(x)^2 f(x)^2}{q(x)} dx$$

with constraint of

$$\int_x q(x) = 1$$

From Cauchy-Schwarz inequality, we have

$$\int_x \frac{p(x)^2 f(x)^2}{q(x)} dx \cdot \int_x q(x) dx \geq \left(\int_x p(x) f(x) dx \right)^2$$

The equation holds when

$$q(x) \propto p(x)|f(x)|$$

Thus our statement is proved.

P9 **1.** Note that

$$T(s \to s') = q(s'|s)\alpha(s' \to s)$$

$$\alpha(s' \to s) = \min\left(1, \frac{p(s')q(s|s')}{p(s)q(s'|s)}\right) = \min\left(1, \frac{p(s')}{p(s)}\right)$$

- if $\pi(s) \geq \pi(s')$, then we have

$$\pi(s)T(s \to s') = \pi(s)q(s'|s)\frac{\pi(s')}{\pi(s)} = \pi(s')q(s'|s)\alpha(s' \to s) = \pi(s')T(s' \to s)$$

- if $\pi(s) < \pi(s')$, then we have

$$\pi(s)T(s \to s') = \pi(s)q(s'|s) = \pi(s')q(s'|s)\alpha(s' \to s) = \pi(s')T(s' \to s)$$

Thus detailed balance is satisfied.

$$\min_z \min_{z':\pi(z')>0} \frac{T(z \to z')}{\pi(z')} = \min_z \min_{z':\pi(z')>0} \frac{q(z'|z)\alpha(z' \to z)}{\pi(z')} = \min_z \min_{z':\pi(z')>0} q(z'|z) \min\left(\frac{1}{\pi(z')}, \frac{1}{\pi(z)}\right) > 0$$

Thus the ergotic condition is satisfied.

2.

$$q(s'|s) = \frac{1}{d}p(s'_i|s_{-i}) \cdot \delta(s'_{-i} = s_{-i})$$

where δ term means that except for updating i, others keep unchanged.

Note that

$$\frac{\pi(s')q(s|s')}{\pi(s)q(s'|s)} = \frac{\pi(s')p(s_i|s_{-i})}{\pi(s)p(s'_i|s_{-i})}$$

From

$$\pi(s) = p(s_i|s_{-i})\pi(s_{-i})$$

We have

$$\frac{\pi(s')q(s|s')}{\pi(s)q(s'|s)} = 1$$

Thus a Gibbs Sampling is equivalent to a MH Sampling that $\alpha(s \to s') = 1$.

3. Same to the mark we used in the above proof, we have (totally d dimensions)

$$K_{cyclic} = K_d \circ K_{d-1} \circ \cdots \circ K_1$$

$$K_{random} = \frac{1}{d}\sum_{i=1}^{d} K_i$$

Theorem 1 each update kernel K_i will keep the stationary distribution π unchanged. To be specific,

$$\int p(s)K_i(s \to s')ds = p(s')$$

Proof. Note that we have

$$p(s) = p(s_i|s_{-i})p(s_{-i})$$

Thus

$$\begin{aligned}\int p(s)K_i(s \to s')ds &= \int p(s_i|s_{-i})p(s_{-i})p(s'_i|s_{-i})\delta(s'_i = s_{-i})ds \\ &= p(s_{-i})p(s'_i|s_{-i}) = p(s')\end{aligned}$$

□

To conclude, the two kernels have the same stationary distribution.