ASSIGNMENT-3

Q1] Importance Sampling estimate of $E_p(f(x))$ is

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)\,p(x_i)}{q(x_i)}, \quad x_i \sim q$$

, where q is a probability distribution such that $q(x) > 0$ whereever $p(x)f(x) \neq 0$.

$$E_q[\hat{\mu}_q] = E_q\left[\frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)\,f(x_i)}{q(x_i)}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E_q\left(\frac{p(x_i)f(x_i)}{q(x_i)}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int \frac{p(x_i)\,f(x_i)}{\cancel{q(x_i)}}\,\cancel{q(x_i)}\,d(x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E_p[f(x)]$$

$$= E_p[f(x)]$$

$$\text{Bias} = E_p[f(x)] - E_q[\hat{\mu}_q]$$

$$= \underline{\underline{0}}$$

$$\text{Variance} \Rightarrow \mathrm{Var}_q\left[\frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)\ f(x_i)}{q(x_i)}\right]$$

$$= \frac{1}{n^2}\mathrm{Var}_q\left[\sum_{i=1}^{n}\frac{p(x_i)\,f(x_i)}{q(x_i)}\right]$$

$$= \frac{1}{n^2}\left(E_q\left[\left(\sum_{i=1}^{n}\frac{p(x_i)\ f(x_i)}{q(x_i)}\right)^2\right] - E_q\left[\sum_{i=1}^{n}\frac{p(x_i)\ f(x_i)}{q(x_i)}\right]^2\right)$$

$$= \frac{1}{n^2}\left(\left(\sum_{i=1}^{n} E_q\left[\frac{p(x_i)^2 \cdot f(x_i)^2}{q(x_i)^2}\right]\right) - E_p[f(x)]^2\right]$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\int \frac{p(x_i)^2 f(x_i)^2}{q(x_i)^2}\,q(x_i)\,d(x_i) - \mu^2\right)\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \left( \int \frac{p(x_i)^2 f(x_i)^2}{q(x_i)} d(x_i) - \mu^2 \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \left( \int \frac{p(x_i) f(x_i)^2}{q(x_i)} p(x_i) dx_i - \mu^2 \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \left[ E_p \left[ \frac{p(x) f(x)^2}{q(x)} \right] - \mu^2 \right]$$

$$= \frac{1}{n} \left[ E_p \left[ \frac{p(x) f(x)^2}{q(x)} \right] - \mu^2 \right]$$

**Q2]**

Since Cumulative distribution function $F_x$ is monotonic we can say $\Rightarrow$

$$P(F_x^{-1}(U) \le x_0)$$

$$= P(F_x(F_x^{-1}(U)) \le F_x(x_0))$$

$$= P(U \le F_x(x_0)) \qquad \text{[Acc to def of Inverse]}$$

$$= P[0 \le U \le F_x(x_0)]$$

$$= F_x(x_0) - 0$$

$$= F_x(x_0)$$

Hence $F_x^{-1}(U)$ has density $p$

**Q3]**

$X$ is sampled from $q$ distribution

$U$ is sampled from $U_{[0,1]}$

Let $Z$ be the sampled Random variable using the algorithm

$$F_Z(z) = P(Z < z) = P\left[X < z \,\middle|\, U < \frac{p(x)}{q(x)}\right) = \frac{P\left(X < z \,;\, U < \frac{p(x)}{q(x)}\right)}{P\left(U < \frac{p(x)}{q(x)}\right)}$$

$$P\left(X < z, U < \frac{p(x)}{q(x)}\right) = \int_{-\infty}^{\infty} P\left((x < z) \cap \left(U < \frac{p(x)}{q(x)}\right) \Big| X = x\right) q(x)\,dx$$

$$= \int_{-\infty}^{z} P\left(U < \frac{p(x)}{q(x)}\right) q(x)\,dx$$

$$= \int_{-\infty}^{z} \frac{p(x)}{q(x)} q(x)\,dx = P_X(z)$$

$$P\left(U < \frac{p(x)}{q(x)}\right) = \int_{-\infty}^{\infty} P\left(U < \frac{p(x)}{q(x)} \Big| X = x\right) q(x)\,dx$$

$$= \int_{-\infty}^{\infty} \frac{p(x)}{q(x)} q(x)\,dx = \int_{-\infty}^{\infty} p(x)\,dx = 1$$

$$\therefore F_Z(z) = P_X(z)$$

Hence $\underline{Z \sim q}$

**Q4]**

$p(x) = \frac{\tilde{p}(x)}{Z}$, $\tilde{p}(x)$ is unnormallzed.

$X \sim q$ and $U \sim$ uniform $[0, kq(x)]$

Let Sampled R.V be $Z$

$$P(U < \tilde{p}(x)) = \int_{-\infty}^{\infty} P(U < \tilde{p}(x) \mid X = x)\, q(x)\,dx = \int_{-\infty}^{\infty} \frac{\tilde{p}(x)}{k\, q(x)} q(x)\,dx$$

$$= \frac{1}{k} \int_{-\infty}^{\infty} \tilde{p}(x)\,dx$$

we know that $\rightarrow \int_{-\infty}^{\infty} \tilde{p}(x)\,dx = Z$

$$\therefore P(U < \tilde{p}(x)) = \frac{Z}{K}$$

$$F_Z(z) = P(Z<z) = P(X<z \mid U<\tilde{p}(x)) = \frac{P(X<z, U<\tilde{p}(x))}{P(U<\tilde{p}(x))}$$

$$P(X<z, U<\tilde{p}(x)) = \int_{-\infty}^{\infty} p(X<z, U<\tilde{p}(x) \mid X=x)\, q(x)\, dx$$

$$= \int_{-\infty}^{z} p(U<\tilde{p}(x) \mid X=x)\, q(x)\, dx$$

$$= \int_{-\infty}^{z} \frac{\tilde{p}(x)}{k\cancel{q(x)}} \cancel{q(x)}\, dx = \frac{1}{k}\int_{-\infty}^{z} \tilde{p}(x)\, dx$$

$$P(Z<z) = \frac{\frac{1}{k}\int_{-\infty}^{z} \tilde{p}(x)\, dx}{Z/k} = \int_{-\infty}^{z} \frac{\tilde{p}(x)}{Z}\, dx =$$

$$= \int_{-\infty}^{z} p(x)\, dx = F_X(z)$$

Thus $Y$ r.p. gr means same as drawing samples from $p(x)$

Q5] a)

$$P(X_l = x_l \mid X_{-l} = x_{-l}, H=h) = \frac{P(X=x, H=h)}{P(X_{-l}=x_{-l}, H=h)}$$

$$= \frac{P(X=x, H=h)}{\sum_{x_l} P(X=x, H=h)}$$

$$= \frac{\exp\left(\sum_{i=1}^{d}\sum_{j=1}^{m} x_i h_j w_{ij}\right)/\cancel{Z}}{\sum_{x_l} \exp\left(\sum_{i=1}^{d}\sum_{j=1}^{m} x_i h_j w_{ij}\right)/\cancel{Z}}$$

$$= \frac{\prod_{i=1}^{d}\prod_{j=1}^{m} \exp(x_i h_j w_{ij})}{\sum_{x_l}\prod_{i=1}^{d}\prod_{j=1}^{m} \exp(x_i h_j w_{ij})}$$

$$= \frac{\exp\left(\sum_{j=1}^{m} x_l h_j w_{lj}\right) \cancel{\prod_{\substack{i=1\\ i\neq l}}^{d} \prod_{j=1}^{m} \exp(x_i h_j w_{ij})}}{\sum_{x_l} \exp\left(\sum_{j=1}^{m} x_l h_j w_{lj}\right)\left(\cancel{\prod_{\substack{i=1\\ i\neq l}}^{d} \prod_{j=1}^{m} \exp(x_i h_j w_{ij})}\right) \left(\cancel{\exp}\right.}$$

$$\boxed{P(X_l = x_l \mid X_{-l} = x_{-l}, H = h) = \frac{\exp\left(\sum_{j=1}^{m} x_l h_j w_{lj}\right)}{1 + \exp\left(\sum_{j=1}^{m} h_j w_{lj}\right)}}$$

$$P(X_l = 1 \mid X_{-l} = x_{-l}, H = h) = \frac{\exp\left(\sum_{j=1}^{m} h_j w_{lj}\right)}{1 + \exp\left(\sum_{j=1}^{m} h_j w_{lj}\right)} = \sigma\left(\sum_{j=1}^{m} h_j w_{lj}\right)$$

$$P(H_l = h_l \mid X = x, H_{-l} = h_{-l}) = \frac{P(X = x, H = h)}{P(H_{-l} = h_{-l}, X = x)}$$

$$= \frac{P(X = x, H = h)}{\sum_{h_l} P(H = h, X = x)}$$

$$= \frac{\prod_{i=1}^{d} \prod_{j=1}^{m} \exp(x_i h_j w_{ij})}{\sum_{h_l} \prod_{i=1}^{d} \prod_{j=1}^{m} \exp(x_i h_j w_{ij})}$$

$$= \frac{\exp\left(\sum_{i=1}^{d} (x_i h_l w_{il})\right) \cancel{\prod_{i=1}^{d} \prod_{\substack{j=1\\ j\neq l}}^{m} \exp(x_i h_j w_{ij})}}{\sum_{h_l} \exp\left(\sum_{i=1}^{d} x_i h_l w_{il}\right) \cancel{\prod_{i=1}^{d} \prod_{\substack{j=1\\ j\neq l}}^{m} \exp(x_i h_j w_{ij})}}$$

$$= \frac{\exp\left(\sum_{i=1}^{d} x_i h_l w_{il}\right)}{1 + \exp\left(\sum_{i=1}^{d} x_i w_{il}\right)}$$

$$P(H_l = 1 \mid X = x, H_{-l} = h_{-l}) = \frac{\exp\left(\sum_{i=1}^{d} x_i w_{il}\right)}{1 + \exp\left(\sum_{i=1}^{d} x_i w_{il}\right)} = \sigma\left(\sum_{i=1}^{d} x_i w_{il}\right)$$

# Gibbs_RBM

April 20, 2017

In [40]:
```
import numpy as np

d,m = 10,20

x = (np.random.rand(d,1) > 0.5)
h = (np.random.rand(m,1) > 0.5)
W = np.random.rand(d,m) - 0.5

max_iter = 10000
for i in range(max_iter):
    ph = 1 / (1 + np.exp(-(W.T @ x)))
    h = ph > 0.5
    px = 1 / (1 + np.exp(-(W @ h)))
    x = px > 0.5
```

Q6)

## LSTM

LSTMs were designed to solve Vanishing gradient problem of Recurrent Neural Networks, because learning long-term dependencies. They combat this using gating mechanism.

Each LSTM Unit consists of

⇒ Input (i), output (o) and forget (f) gates.

(i) <u>Input Gate</u>: Defines how much of the newly computed state for the current input we want to let through

$$i = \sigma(x_t U^i + h_{t-1} W^i)$$

(f) <u>Forget gate</u>: Defines how much of the previous state we want to let through

$$f = \sigma(x_t U^f + h_{t-1} W^f)$$

(o) <u>Output Gate</u>: Defines how much of the internal state has to be exposed to higher layers and next time step

$$o = \sigma(x_t U^o + h_{t-1} W^o)$$

(g) <u>Candidate Hidden State</u>: Computed based on the current input and previous hidden state.

$$g = \tanh(x_t U^g + h_{t-1} W^g)$$

($c_t$) <u>Internal memory</u>: Combination of previous memory $C_{t-1}$ multiplied by forget gate and g.

$$C_t = C_{t-1} \circ f + g \circ i$$

($h_t$) <u>Hidden state</u>:

$$h_t = \tanh(c_t) \circ O$$

# GRUs

GRU has 2 gates

**Reset Gate (r)**: Determines how to combine new inputs with previous memory.

$$r = \sigma(x_t U^r + h_{t-1} W^r)$$

**Update Gate (z)**: Determines how much of the previous memory to keep around

$$z = \sigma(x_t U^z + h_{t-1} W^z)$$

$$g = \tanh(x_t U^g + (h_{t-1} \circ r) W^g)$$

$$h_t = (1-z) \circ g + z \circ h_{t-1}$$

## DIFFERENCE

1. GRU has two gates, LSTM has three gates
2. GRU doesn't possess internal memory and doesn't have output gate that is present in LSTMs.
3. Input and output gates are coupled by an update gate z & other gate r and is applied directly to the previous hidden state
4. We don't apply second nonlinearity while computing output in GRUs