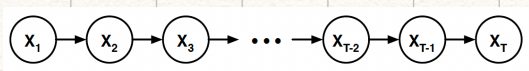
3.1 Inference in a chain



We consider the BN shown above with random variables $X_t \in \{1, 2, \dots, m\}$.

We suppose that the CPT at each non-root node is given by the same $m \times m$ matrix; that is, $\forall t \geq 1$, we have:

$$A_{ij} = P(X_{t+1} = j \mid X_t = i)$$

a) Prove that $P(X_{t+1} = j \mid X_1 = i) = [A^t]_{ij}$, where A^t is the t^{th} power of the matrix A. (Hint: Induction)

$t = 1$:

By definition, $A_{ij} = P(X_2 = j \mid X_1 = i)$

$t = n$:

We assume $P(X_{n+1} = j \mid X_1 = i) = [A^n]_{ij}$

$t = n+1$

$$P(X_{n+2} = j \mid X_1 = i) \overset{PR}{=} \frac{P(X_{n+2} = j, X_1 = i)}{P(X_1 = i)}$$

$$\overset{marg}{=} \frac{\sum_k P(X_{n+2} = j, X_{n+1} = k, X_1 = i)}{P(X_1 = i)}$$

$$= \sum_k \underbrace{P(X_{n+2} = j \mid X_{n+1} = k)}_{A_{kj}} \underbrace{P(X_{n+1} = k \mid X_1 = i)}_{[A^n]_{ik}}$$

$$= \sum_k A_{kj} \cdot [A^n]_{ik}$$

$$= [A^{n+1}]_{ij}$$ □

b) Consider the computational complexity of this inference. Devise a simple algorithm, based on matrix-vec multiplication, that scales as $O(m^2 t)$.

Because we have a square matrix with dimensions $m \times m$, and we have a vector with m elements, matrix-vector multiplication would take $O(m^2)$ time. Now assuming that we have t such multiplications, we would end up with a time complexity of $O(m^2 t)$.

c) Show alternatively that the inference can also be done in $O(m^3 \log_2 t)$.

Let's suppose we want to find the t^{th} power of a matrix. We can express t in binary form to only have to do $\log(t)$ calculations at most, so the method of expressing the powers in binary form has time complexity $O(\log t)$. Then we need to multiply the two matrices together, which has a time complexity of $O(m^3)$ for two $m \times m$ matrices. Thus, this approach has a complexity of $O(m^3 \log t)$. QED

Suppose that the transition matrix A_{ij} is sparse with at most $s \ll m$ non-zero elements per row.

d) Show that in this case the inference can be done in $O(smt)$.

Say we have our sparse A matrix. Then, when doing the matrix-vector multiplication, there will be at most s non-zero elements for each row in the matrix $\Rightarrow$ s multiplications for the elements in the vector.

Thus, when doing the calculation for t matrices, we will have the complexity $O(smt)$. QED

e) Show how to compute the posterior probability $P(X_1 = i \mid X_{T+1} = j)$ in terms of A and $P(X_1 = i)$. Hint: Bayes and result from a)

We remember that we have $[A^T]_{ij} = P(X_{T+1} = j \mid X_1 = i)$ from a).

$$P(X_1 = i \mid X_{T+1} = j) \overset{\text{Bayes}}{=} \frac{P(X_{T+1} = j \mid X_1 = i) P(X_1 = i)}{P(X_{T+1} = j)}$$

$$\overset{\text{from a)}}{=} \frac{[A^T]_{ij} P(X_1 = i)}{P(X_{T+1} = j)}$$

$$\overset{\text{marg}}{=} \frac{[A^T]_{ij} P(X_1 = i)}{\sum_k P(X_{T+1} = j, X_T = k)}$$

$$\overset{\text{PR}}{=} \frac{[A^T]_{ij} P(X_1 = i)}{\sum_k \underbrace{P(X_{T+1} = j \mid X_T = k)}_{A_{kj}} P(X_T = k)}$$

$$= \frac{[A^T]_{ij} P(X_1 = i)}{\sum_k A_{kj} P(X_T = k)}$$

$$\overset{\text{marg}}{=} \frac{[A^T]_{ij} P(X_1 = i)}{\sum_k \sum_l A_{kj} P(X_T = k, X_{T-1} = l)}$$

$$\overset{\text{PR}}{=} \frac{[A^T]_{ij} P(X_1 = i)}{\sum_k \sum_l A_{kj} \underbrace{P(X_T = k \mid X_{T-1} = l)}_{A_{lk}} P(X_{T-1} = l)}$$

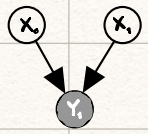
$$= \frac{[A^T]_{ij} P(X_1 = i)}{\sum_k \sum_l A_{kj} A_{lk} P(X_{T-1} = l)}$$

$$= \frac{[A^T]_{ij} P(X_1 = i)}{\underbrace{\sum_l P(X_{T-1} = l) A^2_{kj}}_{[A^{T-1}]_{ij}}}$$

$$= \frac{[A^T]_{ij} P(X_1 = i)}{[A^{T-1}]_{ij}}$$

$$= A_{ij} P(X_1 = i)$$

3.2 More inference in a chain



We consider the simple BN shown above, with nodes X_0, X_1 and Y_1.

To compute the posterior probability $P(X_1|Y_1)$, we can use Bayes rule: $P(X_1|Y_1) = \frac{P(Y_1|X_1)P(X_1)}{P(Y_1)}$

a) Show how to compute the conditional probability $P(Y_1|X_1)$ that appears in the numerator of Bayes rule from the CPTs of this BN.

$$P(Y_1|X_1) \overset{PR}{=} \frac{P(X_1,Y_1)}{P(X_1)}$$

$$\overset{marg}{=} \frac{\sum_{x_0} P(X_0=x_0, X_1, Y_1)}{P(X_1)}$$

$$\overset{P.R.}{=} \frac{\sum_{x_0} P(X_0=x_0)\cancel{P(X_1)}P(Y_1|X_0=x_0,X_1)}{\cancel{P(X_1)}}$$

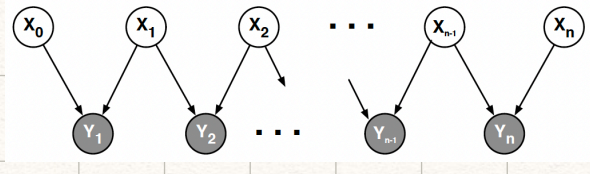
$$= \underline{\underline{\sum_{x_0} P(X_0=x_0)\,P(Y_1|X_0=x_0,X_1)}}$$

b) Show how to compute the conditional probability $P(Y_1)$ that appears in the denominator of Bayes rule from the CPTs of this BN.

$$P(Y_1) \overset{marg}{=} \sum_{x_0}\sum_{x_1} P(X_0=x_0, X_1=x_1, Y_1)$$

$$\overset{PR\ \&\ CI}{=} \underline{\underline{\sum_{x_0}\sum_{x_1} P(X_0=x_0)P(X_1=x_1)P(Y_1|X_0=x_0,X_1=x_1)}}$$

Next, we will show how to generalize these computations when the basic structure of this DAG is repeated to form a chain.



We will consider how to efficiently compute $P(X_n|Y_1,Y_2,\ldots,Y_n)$ in the BN. One approach is to derive a recursion from the conditionalized form of Bayes rule $P(X_n|Y_1,Y_2,\ldots,Y_n) = \frac{P(Y_n|X_n,Y_1,Y_2,\ldots,Y_{n-1})P(X_n|Y_1,Y_2,\ldots,Y_{n-1})}{P(Y_n|Y_1,Y_2,\ldots,Y_{n-1})}$

where the nodes $Y_1,\ldots,Y_{n-1}$ are evidence. We will express the conditional probabilities on the RHS in terms of the CPTs and $P(X_{n-1}=x|Y_1,Y_2,\ldots,Y_{n-1})$. Answers from a) and b) should be helpful.

c) Simplify the term $P(X_n | Y_1, Y_2, \ldots, Y_{n-1})$ that appears in the numerator of Bayes rule.

Because Y_n is not given as evidence, d-separation condition 3 says that X_n and $\{Y_1, Y_2, \ldots, Y_{n-1}\}$ are independent.

Thus, $\underline{P(X_n | Y_1, Y_2, \ldots, Y_{n-1}) = P(X_n)}$.

d) Show how to compute $P(Y_n | X_n, Y_1, \ldots, Y_{n-1})$ in the numerator of Bayes rule in terms of the CPTs and $P(X_{n-1} = x | Y_1, Y_2, \ldots, Y_{n-1})$.

From a), we have $P(Y_1 | X_1) = \sum_{x_0} P(X_0 = x_0) P(Y_1 | X_0 = x_0, X_1)$. Similarly, here we get

$$P(Y_n | X_n, Y_1, Y_2, \ldots, Y_{n-1}) = \sum_{x_{n-1}} P(X_0 = x_0 | Y_1, Y_2, \ldots, Y_{n-1}) \, P(Y_n | X_{n-1} = x_{n-1}, X_n, Y_1, Y_2, \ldots, Y_{n-1})$$

d-sep condition 2 as X_{n-1} makes Y_n C.I from $Y_1, \ldots, Y_{n-1}$

$$\underline{= \sum_{x_{n-1}} P(X_0 = x_0 | Y_1, Y_2, \ldots, Y_{n-1}) \, P(Y_n | X_{n-1} = x_{n-1}, X_n)}$$

e) Show how to compute $P(Y_n | Y_1, \ldots, Y_{n-1})$ in the denominator of Bayes rule in terms of the CPTs and $P(X_{n-1} = x | Y_1, Y_2, \ldots, Y_{n-1})$.

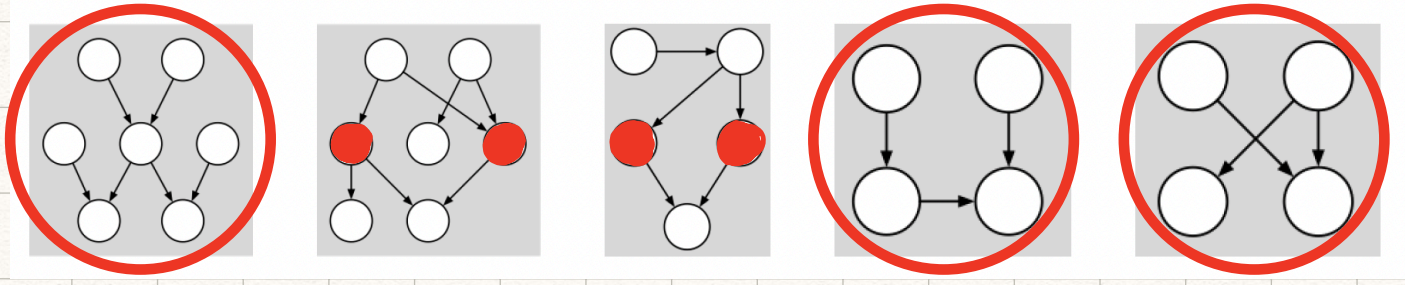
From b) we have $P(Y_1) = \sum_{x_0} \sum_{x_1} P(X_0 = x_0) P(X_1 = x_1) P(Y_1 | X_0 = x_0, X_1 = x_1)$. Similarly, here we get

$$P(Y_n | Y_1, \ldots, Y_{n-1}) = \sum_{x_{n-1}} \sum_{x_n} P(X_{n-1} = x_{n-1} | Y_1, Y_2, \ldots, Y_{n-1}) \underbrace{P(X_n = x_n | Y_1, Y_2, \ldots, Y_{n-1})}_{= P(X_n = x_n) \text{ because } Y_n \text{ is not evidence} \Rightarrow \text{d-sep 3}} \underbrace{P(Y_n | X_{n-1} = x_{n-1}, X_n = x_n, Y_1, Y_2, \ldots, Y_{n-1})}_{\text{d-sep 2 gives that } Y_n \text{ is independent of all other } Y\text{'s.}}$$

$$\underline{P(Y_n | Y_1, \ldots, Y_{n-1}) = \sum_{x_{n-1}} \sum_{x_n} P(X_{n-1} = x_{n-1} | Y_1, Y_2, \ldots, Y_{n-1}) \, P(X_n = x_n) \, P(Y_n | X_{n-1} = x_{n-1}, X_n = x_n)}$$

3.3 Node clustering and polytrees

In the figure, circle the DAGs that are polytrees. In the other DAGs, shade two nodes that could be clustered so that the resulting DAG is a polytree.



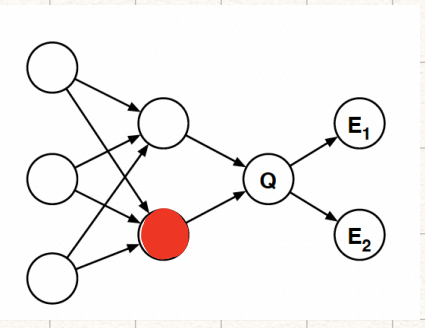
We can't have cycles in a polytree, so the nodes in DAG 2 and 3 need to be clustered.

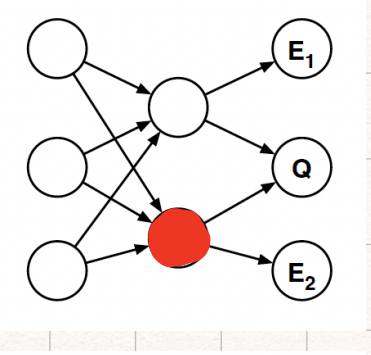
3.4 Cutsets and polytrees

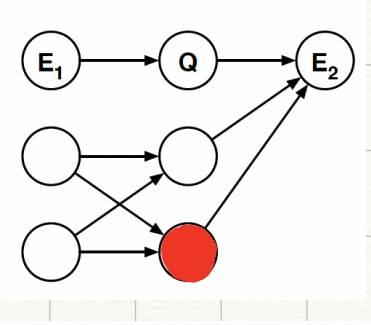
For each of the five loopy BNs in the problem, we will consider how to compute $P(Q|E_1, E_2)$.

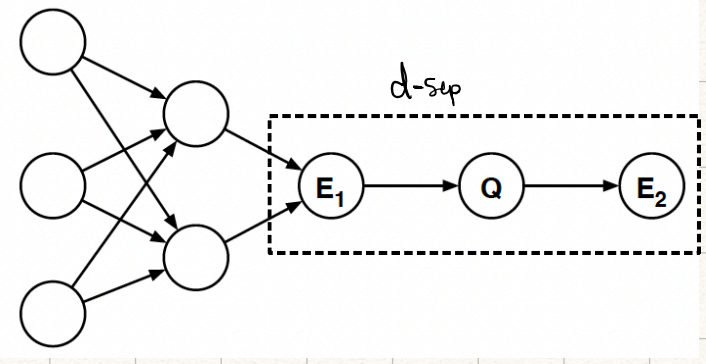
If the inference can be performed by running the polytree algorithm on a subgraph, enclose the subgraph by a dotted line.

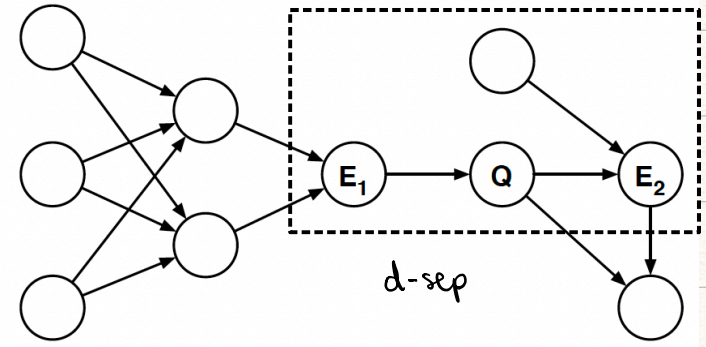
Otherwise, shade **one** node in the BN that can be instantiated to induce a polytree by the method of cutset conditioning.





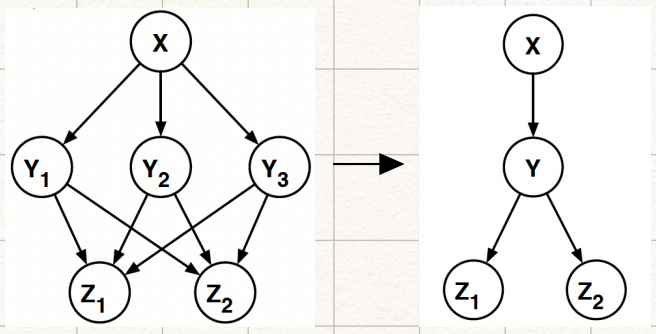






3.5 Node clustering

We consider the following BN that can be transformed into a polytree by clustering Y_1, Y_2 and Y_3 into Y:



From the CPTs in the original BN, fill in the missing elements of the CPTs for the polytree.

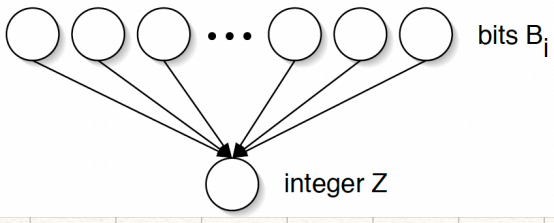
Y_1	Y_2	Y_3	Y	$P(Y\|X=0)$	$P(Y\|X=1)$	$P(Z_1=1\|Y)$	$P(Z_2=1\|Y)$
0	0	0	1	0.09375	0.09375	0.9	0.1
1	0	0	2	0.28125	0.09375	0.8	0.2
0	1	0	3	0.09375	0.03125	0.7	0.3
0	0	1	4	0.03125	0.28125	0.6	0.4
1	1	0	5	0.28125	0.03125	0.5	0.5
1	0	1	6	0.09375	0.28125	0.4	0.6
0	1	1	7	0.03125	0.09375	0.3	0.7
1	1	1	8	0.09375	0.09375	0.2	0.8

3.6 Likelihood weighting

Consider the belief network shown below, with n binary random variables $B_i \in \{0,1\}$ and an *integer* random variable Z. Let $f(B) = \sum_{i=1}^{n} 2^{i-1} B_i$ denote the nonnegative integer whose binary representation is given by $B_n B_{n-1} \ldots B_2 B_1$. Suppose that each bit has prior probability $P(B_i = 1) = \frac{1}{2}$, and that

$$P(Z|B_1, B_2, \ldots, B_n) = \left(\frac{1-\alpha}{1+\alpha}\right) \alpha^{|Z - f(B)|}$$

where $0 < \alpha < 1$ is a parameter measuring the amount of noise in the conversion from binary to decimal. (Larger values of α indicate greater levels of noise.)



a) Show that the conditional distribution for binary to decimal conversion is normalised; $\sum_z P(Z = z|B_1, B_2, \ldots, B_n) = 1$, $z \in [-\infty, \infty]$.

$$\sum_z P(Z = z|B_1, B_2, \ldots, B_n) = \frac{1-\alpha}{1+\alpha}\left[\sum_{z=-\infty}^{f(B)} \alpha^{f(B)-z} + \sum_{z=f(B)+1}^{\infty} \alpha^{z-f(B)}\right]$$

$$= \frac{1-\alpha}{1+\alpha}\left[\sum_{z=-\infty}^{f(B)} \alpha^{f(B)} \alpha^{-z} + \sum_{z=f(B)+1}^{\infty} \alpha^{z} \alpha^{-f(B)}\right]$$

$$= \frac{1-\alpha}{1+\alpha}\left[\alpha^{f(B)} \sum_{z=-\infty}^{f(B)} \alpha^{-z} + \alpha^{-f(B)} \sum_{z=f(B)+1}^{\infty} \alpha^{z}\right]$$

$$= \frac{1-\alpha}{1+\alpha}\left[\alpha^{f(B)} \frac{\alpha^{-f(B)} - \alpha^{\overbrace{\infty+1}^{=0}}}{1-\alpha} + \alpha^{-f(B)} \frac{\alpha^{f(B)+1} - \alpha^{\overbrace{\infty+1}^{=0}}}{1-\alpha}\right]$$

$$= \frac{1-\alpha}{1+\alpha}\left[\frac{\alpha^{f(B)-f(B)}}{1-\alpha} + \frac{\alpha^{-f(B)+f(B)+1}}{1-\alpha}\right]$$

$$= \frac{(1-\alpha)(1+\alpha)}{(1+\alpha)(1-\alpha)}$$

$$\sum_z P(Z = z|B_1, B_2, \ldots, B_n) = 1 \qquad \text{QED}$$

b) Consider a network with $n = 10$ bits and a noise level $\alpha = 0.1$. Use likelihood weighting to estimate $P(B_i = 1|Z = 128)$ for $i \in \{2, 5, 8, 10\}$

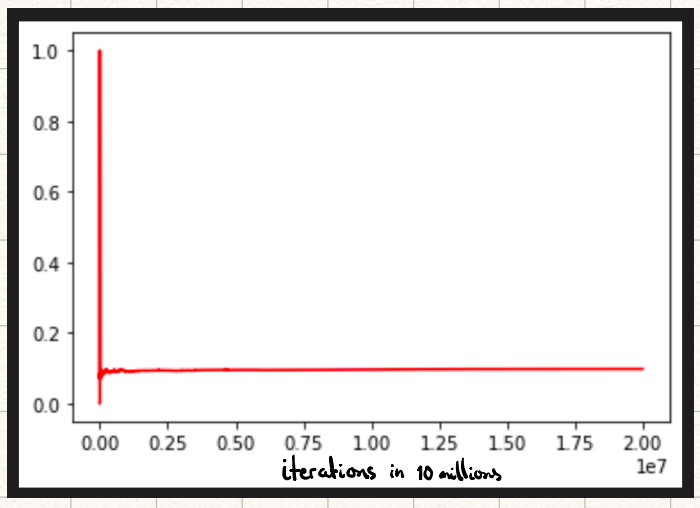
```
P(B=2|z=128)      = 0.09954341596462604
P(B=5|z=128)      = 0.09097414115022234
P(B=8|z=128)      = 0.9096063262998437
P(B=10|z=128)     = 0.0
```

This is after 20 million iterations.

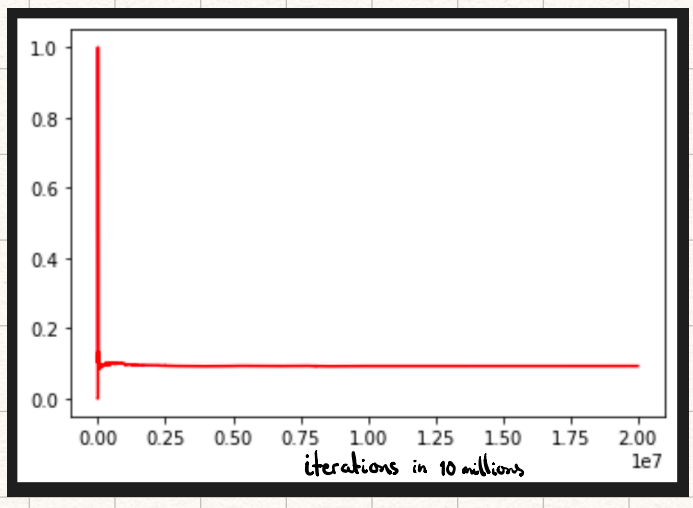
c) Plot your estimates from b) as a function of the number of samples.

Plots for 20 million samples. x-axis is iterations and y-axis is $P(B_i = 1 \mid z = 128)$ with $i \in \{2,5,8,10\}$

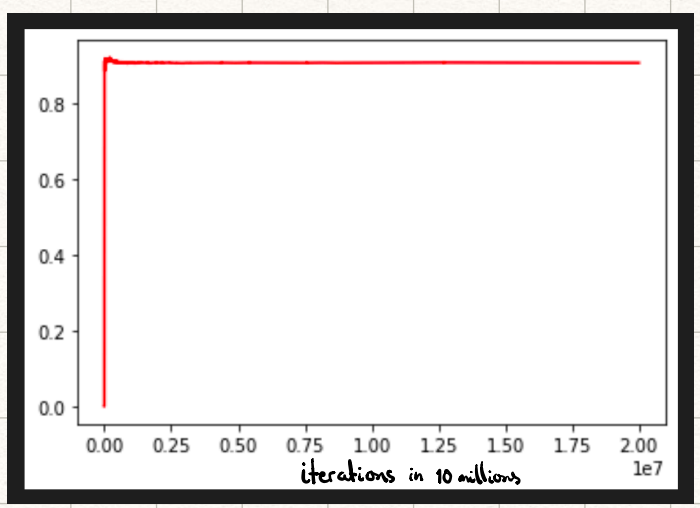
$P(B_2 = 1 \mid z = 128)$



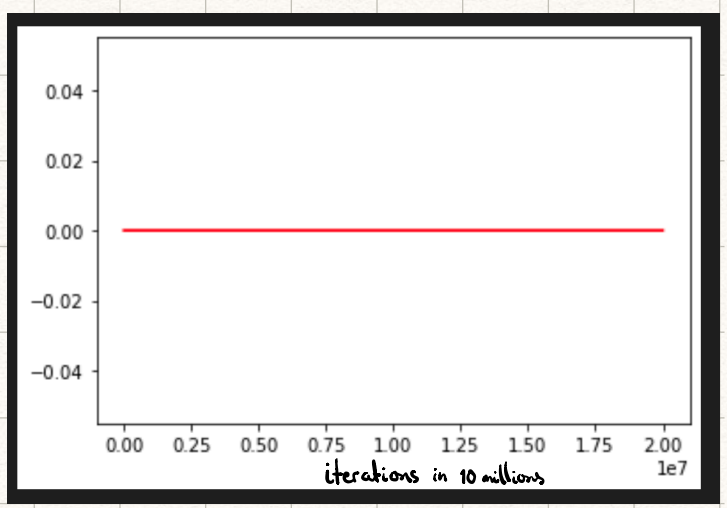
$P(B_5 = 1 \mid z = 128)$



$P(B_8 = 1 \mid z = 128)$



$P(B_{10} = 1 \mid z = 128)$



The end values are the ones answered in b). We observe that the values are very close to their convergence values.

d) Submit hard copy of code.

```
# Imports
import numpy as np
from matplotlib import pyplot as plt
from numpy import random
```

✓ 0.1s

```
# Constants
N = 20000000
alpha = 0.1
num_bits = 10
z = 128
target_bits = [2,5,8,10]
```

✓ 0.5s

```
def likelihood(z, alpha, f):
    return ((1-alpha)/(1+alpha))*np.power(alpha,np.absolute(z-f))
```

✓ 0.3s

```
def indicator(q,q_):
    if q&q_ == 0:
        return 0
    else:
        return 1
```

✓ 0.3s

```
def estimate(bitI):
    bit = np.power(2,bitI-1)
    numerator = 0.0
    denominator = 0.0
    estimates = np.zeros(shape=(N))
    for i in range(N):
        randomBits = random.randint(2,np.power(2,num_bits)-1)
        weight = likelihood(z,alpha,randomBits)
        denominator += weight
        indi = indicator(bit,randomBits)
        numerator += weight * indi
        estimates[i] = numerator / denominator
    return estimates
```

✓ 0.6s

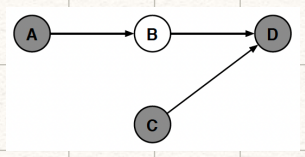
```
for i in target_bits:
    probs = estimate(i)
    final_prob = probs[-1]
    print(f"P(B={i}|z=128) \t = {final_prob} \n")
```

```
iterations = np.arange(0,N)
for i in target_bits:
    probs = estimate(i)
    plt.plot(iterations,probs,'r-')
    plt.show()
```

3.7 Even more inference

We will show how to perform the desired inference in each of the belief networks.

a) Markov blanket Show how to compute $P(B|A,B,C)$ in terms of the CPTs of the BN ($P(A)$, $P(B|A)$, $P(C)$ and $P(D|B,C)$).



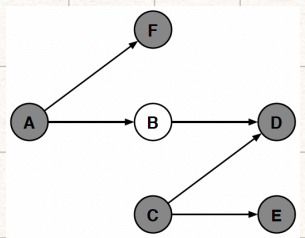
$$P(B|A,C,D) \overset{\text{P.R}}{=} \frac{P(A,B,C,D)}{P(A,C,D)}$$

$$\overset{\text{marg}}{=} \frac{P(A,B,C,D)}{\sum_b P(A,B=b,C,D)}$$

$$\overset{\text{P.R \& d-sep}}{=} \frac{\cancel{P(A)}\,P(B|A)\,\cancel{P(C)}\,P(D|B,C)}{\sum_b \cancel{P(A)}\,P(B=b|A)\,\cancel{P(C)}\,P(D|B=b,C)}$$

$$= \frac{P(B|A)\,P(D|B,C)}{\sum_b P(B=b|A)\,P(D|B=b,C)}$$

b) Conditional independence Show how to compute $P(B|A,C,D,E,F)$ in terms of $P(F|A)$, $P(E|C)$ and the answer from a).



$$P(B|A,C,D,E,F) \overset{\text{P.R.}}{=} \frac{P(A,B,C,D,E,F)}{P(A,C,D,E,F)}$$

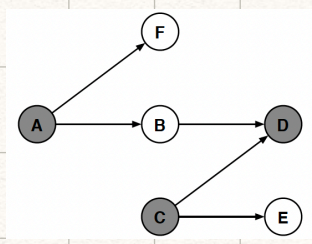
$$\overset{\text{marg}}{=} \frac{P(A,B,C,D,E,F)}{\sum_b P(A,B=b,C,D,E,F)}$$

$$\overset{\text{P.R \& d-sep}}{=} \frac{\cancel{P(A)}\,\cancel{P(F|A)}\,P(B|A)\,P(D|B,C)\,\cancel{P(C)}\,\cancel{P(E|C)}}{\sum_b \cancel{P(A)}\,\cancel{P(F|A)}\,P(B=b|A)\,P(D|B=b,C)\,\cancel{P(C)}\,\cancel{P(E|C)}}$$

(d-sep: D indep of A given B)

$$= \frac{P(B|A)\,P(D|B,C)}{\sum_b P(B=b|A)\,P(D|B=b,C)} = \text{answer from a)}$$

b) More conditional independence Show how to compute $P(B,E,F|A,C,D)$ in terms of the CPTs of the BN and the answer from a) & b).



$$P(B,E,F|A,C,D) \overset{\text{P.R}}{=} \frac{P(A,B,C,D,E,F)}{P(A,C,D)}$$

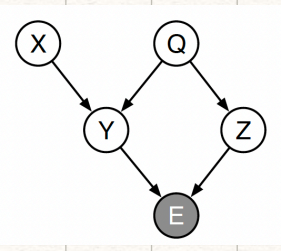
$$\overset{\text{marg.}}{=} \frac{P(A,B,C,D,E,F)}{\sum_b \sum_e \sum_f P(A,B=b,C,D,E=e,F=f)}$$

$$\overset{\text{P.R. \& d-sep}}{=} \frac{\cancel{P(A)}\,P(F|A)\,P(B|A)\,P(D|B,C)\,\cancel{P(C)}\,P(E|C)}{\sum_b \sum_e \sum_f \cancel{P(A)}\,P(F=f|A)\,P(B=b|A)\,P(D|B=b,C)\,\cancel{P(C)}\,P(E=e|C)}$$

$$= \frac{P(F|A)\,P(B|A)\,P(D|B,C)\;P(E|C)}{\sum_b \sum_e \sum_f P(F=f|A)\,P(B=b|A)\,P(D|B=b,C)\,P(E=e|C)}$$

3.8 More likelihood weighting

a) Single node of evidence



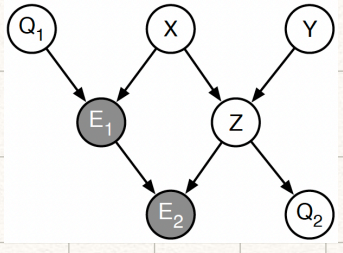
We suppose that T samples $\{q_t, x_t, y_t, z_t\}_{t=1}^{T}$ are drawn from the CPTs of the BN above (with fixed evidence $E=e$).

Show how to estimate $P(Q=q \mid E=e)$ using these samples using the method of likelihood weighting

$$P(Q=q \mid E=e) \overset{LW}{\approx} \frac{\sum_{t=1}^{T} I(q,q_t)\, P(E=e \mid Y,Z)}{\sum_{t=1}^{T} P(E=e \mid Y,Z)}$$

$$= \frac{\sum_{t=1}^{T} I(q,q_t)\, P(E=e \mid Y=y_t, Z=z_t)}{\sum_{t=1}^{T} P(E=e \mid Y=y_t, Z=z_t)}$$

b) Multiple nodes of evidence



We suppose that T samples $\{q_{1t}, q_{2t}, x_t, y_t, z_t\}_{t=1}^{T}$ are drawn from the CPTs of the BN above (with fixed evidence $E_1=e_1$ and $E_2=e_2$)

Show how to estimate $P(Q_1=q_1, Q_2=q_2 \mid E_1=e_1, E_2=e_2)$ using these samples using the method of likelihood weighting

$$P(Q_1=q_1, Q_2=q_2 \mid E_1=e_1, E_2=e_2) = \frac{\sum_{t=1}^{T} I(q_1,q_{1t}) I(q_2,q_{2t}) P(E_1=e_1 \mid Q_1=q_{1t}, X=x_t) P(E_2=e_2 \mid E_1=e_1, Z=z_t)}{\sum_{t=1}^{T} P(E_1=e_1 \mid Q_1=q_{1t}, X=x_t) P(E_2=e_2 \mid E_1=e_1, Z=z_t)}$$