- Forward Sampling

```
def forward-sample(B: Bayesian Network):
        let X_1 ... X_n be a topological ordering of X
        for i = 1, ... n:
                u_i = Parents(X_i)
                sample x_i from P(X_i | u_i)
        return (x_1 ... x_n)
```

from a set of particles $D = \{ \xi[1], \cdots \xi[M] \}$ generated via forward sampling

can estimate the expectation of any f as

$$\hat{\mathbb{E}}_D(f) = 1/M \sum_{m=1}^{M} f(\xi[m])$$

in case of computing $P(y)$

$$\hat{P}_D(y) = 1/M \sum_{m=1}^{M} \mathbb{1}\{\xi[m]\langle Y \rangle = y\}$$

// $\xi[m]\langle Y \rangle$ is the assignment to variables Y of partical $\xi[m]$

Consider the quality of the estimate for a particular event $Y=y$

define a new random variable $\mathbb{1}\{Y=y\}$,

M particles in D define M independent Bernoulli trials, each with probability $P(Y=y) = p$

let $Z_i = \mathbb{1}\{\xi[i]\langle Y \rangle = y\}$, $P(Z_i = 1) = p = P(Y=y)$ ($Z_i \sim \text{Bernoulli}(p)$, iid)

$$P(\hat{P}_D(y) \notin [P_D(y) - \epsilon, P_D(y) + \epsilon]) = P_D(\hat{P}_D(y) - P(y) \geq \epsilon) + P_D(\hat{P}_D(y) - P(y) \leq \epsilon)$$

$$= P_D\left(\frac{1}{M}\sum_i Z_i - \frac{1}{M}\sum_i E[Z_i] \geq \epsilon\right) + \cdots$$

$$= P_D\left(\frac{1}{M}\sum_i (Z_i - E[Z_i]) \geq \epsilon\right) + P_D\left(\frac{1}{M}\sum_i (Z_i - E[Z_i]) \leq -\epsilon\right)$$

$$\leq 2 \cdot \exp(-2M\epsilon^2)$$

$$2\exp(-2M\epsilon^2) \leq \delta$$

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Hoeffding's Inequality:

let $Z_1 \ldots Z_n$ be independent bounded $-\infty < a \leq Z_i \leq b \leq +\infty$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i]) \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i]) \leq -t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

$$P(\hat{P}_D(y) \notin [P_D(y)(1-\epsilon), P_D(y)(1+\epsilon)]) \leq 2e^{-MP(y)\epsilon^2/3} \leq \delta$$

$$M \geq \frac{3\ln(2/\delta)}{P(y)\epsilon^2}$$

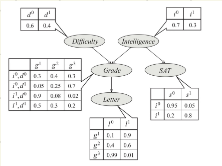
when querying CPDs $P(Y=y \mid E=e)$, we can simply perform forward sampling

and reject those particles incompatible with $E=e$.

if $P(E=e) = 0.01$, only 100 out of 1w can be used to compute $P(Y=y \mid E=e)$

· Likelihood Weighting

Assume we have evidence $D=d^1, S=s^1$



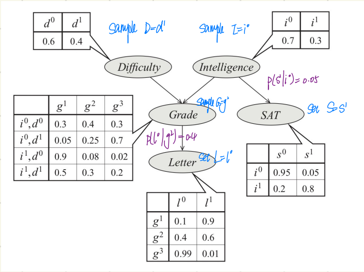
$P(S=s^1 \mid I=i^1) = 80\%$ $\quad P(S=s^1 \mid I=i^0) = 5\%$

To simulate this long-run behavior within a single sample, we should conclude that a sample where $I=i^1$ $S=s^1$ should be worth 80% of a sample
a sample where $I=i^0$ $S=s^1$ should be worth 5% of a sample

When we have multiple observations and we want our sampling process to set all of them to their observed values, we need to consider the probability that each of the observation nodes, had it been sampled using the standard forward sampling, would have resulted in the observed values. The sampling events for each node in forward sampling are independent, hence the weight for each sample should be the product of the weights induced by each evidence node separately

```
def likelihood_weighted_sampling(B: Bayesian Net over X, Z=z)
        Let X_1, ..., X_n be a topological ordering of X
        w=1
        for i=1,...,n
                u_i = Pa(X_i)          // assignment to Pa(X_i)
                if X_i ∉ Z:
                        sample x_i from P(X_i | u_i)
                else:
                        x_i = z_i      // set assignment as observation
                        w = w · P(x_i | u_i)
        return (x_1, ... x_n), w
```

e.g. evidence l^0, s^1



the weight for the sample $(d^1, i^0, g^2, l^0) = 0.4 \times 0.05$

This process generates a weighted particle

can estimate a conditional probability $P(y \mid e)$ with M weight particles $(\xi[1], w[1]), \dots, (\xi[M], w[M])$

$$\hat{P}_B(y \mid e) = \frac{\sum_{m=1}^{M} w[m]\, \mathbb{1}\{y[m]=y\}}{\sum_{m=1}^{M} w[m]}$$

e.g.

$A \rightarrow B$

	a^0	a^1
	0.2	0.8

	b^0	b^1
a^0	0.1	0.9
a_1	0.7	0.3

want $P(a^1 \mid b^0)$

Suppose we generate 2k samples where $A=a^0$ and 8k samples where $A=a^1$ (with forward sampling, $M=10k$)

among those, $\#(a^0, b^0) = 2k \cdot P(b^0 \mid a^0)$ $\quad \#(a^1, b^0) = 8k \cdot P(b^0 \mid a^1)$

$$P(a^1 \mid b^0) = \frac{\#(a^1, b^0)}{\#(a^1, b^0) + \#(a^0, b^0)} = \frac{8k \cdot P(b^0 \mid a^1)}{2k \cdot P(b^0 \mid a^0) + 8k \cdot P(b^0 \mid a^1)} = \frac{\sum_{m=1}^{M} P(b^0 \mid a^1) \cdot \mathbb{1}\{A=a^1\}}{\sum_{m=1}^{M} P(b^0 \mid A[m])}$$

· Importance Sampling

Importance sampling is a general approach for estimating the expectation of a function $f(x)$ relative to some target distribution $P(X)$

generate $X[1] \cdots X[M]$ iid $X[i] \sim P$ $\quad E_P(f) \approx 1/M \, f(X[m])$

Sometime we may want to generate x from a proposal distribution / sampling distribution $Q(X)$, which is different from $P(X)$

requiring that $\forall x$, if $P(x) > 0$, then $Q(x) > 0$ (the support of Q $\{x \mid Q(x) > 0\}$ contains the support of P)

how to obtain estimates of an expectation relative to P by generating samples from a different distribution Q

· Unnormalized Importance Sampling

$$E_{X\sim P}[f(X)] = E_{X\sim Q}\left[f(X)\frac{P(X)}{Q(X)}\right] \qquad E_{X\sim P}[f(X)] = \int_\Omega f(x)\,P(x)\,dx = \int_\Omega Q(x)\cdot f(x)\cdot\frac{P(x)}{Q(x)}\,dx = E_{X\sim Q}\left[f(X)\frac{P(X)}{Q(X)}\right]$$

unnormalized importance sampling estimator
$$\begin{cases} 1^\circ & \text{generate } X[1]\cdots X[M] \text{ iid}, \; X[i] \sim Q(X) \\ 2^\circ & \hat{E}_D(f) = 1/M \cdot \sum_{i=1}^{M} f(X[i]) \cdot \dfrac{P(X[i])}{Q(X[i])} \end{cases}$$

$$\begin{aligned} E_{D\sim Q}[\hat{E}_D(f)] &= E_{D\sim Q}\left[1/M \sum_{i=1}^{M} f(X[i]) \frac{P(X[i])}{Q(X[i])}\right] \\ &= 1/M \sum_{i=1}^{M} E_{D\sim Q}\left[f(X[i]) \frac{P(X[i])}{Q(X[i])}\right] \\ &= 1/M \sum_{i=1}^{M} E_{X\sim P}[f(X)] \\ &= E_{X\sim P}[f(X)] \end{aligned}$$

sample D from Q

the unnormalized importance sampling estimator is an "unbias" estimation for $E_P[f(X)]$

$$\begin{aligned} Var_{D\sim Q}[\hat{E}_D(f)] &= Var_{D\sim Q}\left[\frac{1}{M}\sum_{i=1}^{M} f(X_i)\frac{P(X_i)}{Q(X_i)}\right] \\ &= \frac{1}{M}\cdot Var_{D\sim Q}\left(f(X)\frac{P(X)}{Q(X)}\right) \\ &= \frac{1}{M}\cdot\left(E_{X\sim Q}\left[\left(f(X)\frac{P(X)}{Q(X)}\right)^2\right] - E_{X\sim Q}\left[f(X)\frac{P(X)}{Q(X)}\right]^2\right) \\ &= \frac{1}{M}\left(E_Q\left[\left(f(X)\frac{P(X)}{Q(X)}\right)^2\right] - E_P[f(X)]^2\right) \end{aligned}$$

variance of estimator

// the closer $P(X)$ and $Q(X)$ are, the better performance importance sampling has

eg. $f(\cdot) = 1$

$$var_{D\sim Q}[\hat{E}_D(f)] = \frac{1}{M}\left[E_Q\left[\left(\frac{P(X)}{Q(X)}\right)^2\right] - E_Q\left[\frac{P(X)}{Q(X)}\right]^2\right] = \frac{1}{M}\cdot var\left(\frac{P}{Q}\right)$$

when f is a constant function, the estimator has a higher var when P is different from Q

Normalized Importance Sampling

$\tilde{P}(x)$ is a unnormalized distribution

$$E_{x\sim Q}\left[\frac{\tilde{P}(x)}{Q(x)}\right] = \int_x Q(x)\frac{\tilde{P}(x)}{Q(x)}dx = Z$$

$$E_{x\sim P}[f(x)] = \int_x P(x)f(x)dx = \int_x Q(x)f(x)\frac{P(x)}{Q(x)}dx = \frac{1}{Z}\int_x Q(x)f(x)\frac{\tilde{P}(x)}{Q(x)}dx$$

$$= \frac{1}{Z}E_{x\sim Q}\left[f(x)\frac{\tilde{P}(x)}{Q(x)}\right]$$

$$E_{x\sim P}[f(x)] = \frac{E_{x\sim Q}\left[f(x)\frac{\tilde{P}(x)}{Q(x)}\right]}{E_{x\sim Q}\left[\frac{\tilde{P}(x)}{Q(x)}\right]}$$ can use an empirical estimator for both numerator and denominator

Given M samples $D = \{X[1], \dots, X[M]\}$ $X[i] \sim Q$ iid.

$$\hat{E}_D(f) = \frac{\sum_{i=1}^{M} f(x[i])\frac{\tilde{P}(X[i])}{Q(X[i])}}{\sum_{i=1}^{M}\frac{\tilde{P}(X[i])}{Q(X[i])}}$$ normalized importance sampling estimator

($\hat{E}_D(f)$ does not depend on C)

for finite M, the estimator is biased. the bias goes down as $\frac{1}{M}$

$$\mathrm{var}_P\left[\hat{E}_{D\sim Q}(f(x))\right] \approx \frac{1}{M}\mathrm{Var}_P[f(x)]\left(1+\mathrm{Var}_Q\left[\frac{\tilde{P}(x)}{Q(x)}\right]\right)$$ the variance also goes down as $\frac{1}{M}$

In practice the variance of the normalized estimator is lower that var of unnormalized estimator

the reduction in variance often outweighs the bias term, so the normalized estimator is used even when P is known

when estimate $E_P[f]$ with a standard sampling method

$$\mathrm{Var}_P\left[\hat{E}_{D\sim P}(f)\right] = \mathrm{var}_P\left(\frac{1}{M}\sum_{i=1}^{M} f(x_i)\right) = \frac{1}{M}\mathrm{var}_P(f(x))$$

the ratio between 2 variances is $\frac{1}{1+\mathrm{Var}_Q\left[\frac{\tilde{P}(x)}{Q(x)}\right]}$

$$\frac{M\cdot \mathrm{Var}_P\left[\hat{E}_{D\sim P}(f)\right]}{M\cdot \mathrm{Var}_P\left[\hat{E}_{D\sim Q}(f)\right]} = \frac{1}{1+\mathrm{Var}_Q\left[\frac{\tilde{P}(x)}{Q(x)}\right]}$$

$$M\cdot \mathrm{var}_P\left[\hat{E}_{D\sim P}(f)\right] = \frac{M}{1+\mathrm{Var}_Q\left[\frac{\tilde{P}(x)}{Q(x)}\right]}\mathrm{Var}_P\left[\hat{E}_{D\sim Q}(f)\right]$$

the effective sample size for normalized importance sampling is $\frac{M}{1+\mathrm{Var}_Q\left(\frac{\tilde{P}(x)}{Q(x)}\right)}$

$$\mathrm{var}_Q\left(\frac{\tilde{P}(x)}{Q(x)}\right) = \sum_{i=1}^{M}\left(\frac{\tilde{P}(x_i)}{Q(x_i)}\right)^2 - \left(\sum_{i=1}^{M}\frac{\tilde{P}(x_i)}{Q(x_i)}\right)^2$$

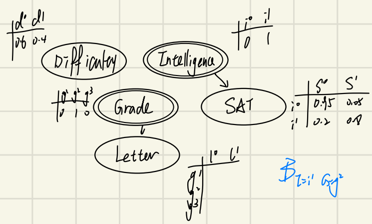
- The mutilated Network Proposal Distribution Importance Sampling

Let $\mathcal{B}$ be a network and $\mathbf{Z}=\mathbf{z}$ be an instantiation of variables

define the mutilated network $\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$ as

each node $Z_i \in \mathbf{Z}$ has no parents in $\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$, and take deterministic CPD $P(Z_i = z_i) = 1$

other nodes and CPDs unchanged.



Let ξ be a sample generated by "likelihood weighting" and w be its weighting

$$w(\xi) = \frac{P_{\mathcal{B}}(\xi)}{P_{\mathcal{B}_{\mathbf{Z}=\mathbf{z}}}(\xi)}$$

can compute the prob $P(\mathbf{Z}=\mathbf{z})$, where the target distribution P is $P_{\mathcal{B}}(x)$

the proposal distribution Q is $P_{\mathcal{B}_{\mathbf{Z}=\mathbf{z}}}(x)$

$$P_{\mathcal{B}}(\mathbf{Z}=\mathbf{z}) = E_{x\sim P}\left[\mathbb{1}\{x\langle \mathbf{Z}\rangle = \mathbf{z}\}\right] = \int_x P(x)\cdot \mathbb{1}\{x\langle \mathbf{Z}\rangle = \mathbf{z}\}\, dx$$

$$= \int_x Q(x)\cdot \frac{P(x)}{Q(x)} \cdot \mathbb{1}\{x\langle \mathbf{Z}\rangle = \mathbf{z}\}\, dx$$

$$= E_{x\sim Q}\left[\frac{P(x)}{Q(x)}\, \mathbb{1}\{x\langle \mathbf{Z}\rangle = \mathbf{z}\}\right]$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} w_i \cdot \mathbb{1}\{x_i\langle \mathbf{Z}\rangle = \mathbf{z}\}$$

$\hat{P}_{\mathcal{D}}(\mathbf{Z}=\mathbf{z}) = \frac{1}{M}\sum_{i=1}^{M} w_i$ // sample from $Q=\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$, $\therefore$ $\mathbf{Z}=\mathbf{z}$ holds for all samples

$P(y|e) = \frac{P(y,e)}{P(e)}$ // both numerator and denominator can be computed by imporance sampling

generate $\{(\xi[1], w[1]), \cdots, (\xi[M], w[M])\}$ with likelihood weighting with $Y=y, E=e$

generate $\{(\xi'[1], w'[1]), \cdots, (\xi'[N], w'[N])\}$ with likelihood weighting with $E=e$

$$\hat{P}_{\mathcal{D}}(y|e) = \frac{1/M \sum_{i=1}^{M} w[i]}{1/N \sum_{i=1}^{N} w'[i]}$$

def data_dependent_likelihood_weighting ($\mathcal{B}$: Bayesian net over x, $\mathbf{Z}=\mathbf{z}$, u: upper bound on CPD entry of z

ϵ: desired error bound, δ: desired probability of error):

$\gamma = \frac{4(1+\epsilon)}{\epsilon^2}\ln\frac{2}{\delta}$ $k = |\mathbf{Z}|$ $W=0$ $M=0$

while $W < \gamma u^k$:

ξ, w = likelihood_weighted_sampling($\mathcal{B}$, $\mathbf{Z}=\mathbf{z}$)

W += w

M += 1

return W/M // $\hat{P}_{\mathcal{D}}(\mathbf{z})$ with $P\left((1-\epsilon)P_{\mathcal{B}}(\mathbf{z}) \le \hat{P}_{\mathcal{D}}(\mathbf{z}) \le (1+\epsilon)P_{\mathcal{B}}(\mathbf{z})\right) \ge \delta$

· Normalized Likelihood Weighting

want evaluate $P(Y|e)$, evaluating $P(y|e)$ $\forall$ $y \in Val(Y)$ is too expensive

Apply normalized importance sampling, when the target distribution $\tilde{P}(\mathcal{X})$ is $P_B(\mathcal{X}, e)$, the unnormalized version of P

the proposal distribution $Q(\mathcal{X})$ is $P_{B_{E=e}}(\mathcal{X}, e)$

$$E_{x\sim Q}\left[\frac{\tilde{P}(x)}{Q(x)}\right] = \int_x Q(x) \frac{\tilde{P}(x)}{Q(x)} d(x) = Z$$

$$P_B(Y=y|e) = E_{x\sim P}[\mathbb{1}\{x\langle Y\rangle = y\}] = \int_x P(x) \cdot \mathbb{1}\{x\langle Y\rangle = y\}\, dx$$

$$= \int_x Q(x) \cdot \mathbb{1}\{x\langle Y\rangle = y\} \frac{P(x)}{Q(x)} dx = \frac{1}{Z}\int_x Q(x)\, \mathbb{1}\{x\langle Y\rangle = y\} \frac{\tilde{P}(x)}{Q(x)} dx$$

$$= \frac{1}{Z} E_{x\sim Q}\left[\mathbb{1}\{x\langle Y\rangle = y\} \frac{\tilde{P}(x)}{Q(x)}\right]$$

$$P_B(Y=y|e) = \frac{E_{x\sim Q}\left[\mathbb{1}\{x\langle Y\rangle = y\} \frac{\tilde{P}(x)}{Q(x)}\right]}{E_{x\sim Q}\left[\frac{\tilde{P}(x)}{Q(x)}\right]}$$

$$\begin{cases} \text{sample } D = \{(\xi[1], w[1]), \dots, (\xi[M], w[M])\} \text{ with likelihood weighting } E=e \\ \hat{P}_B(y|e) = \dfrac{\sum_i \mathbb{1}\{\xi[i]\langle Y\rangle = y\}\, w[i]}{\sum_i w[i]} \quad \forall\ y \in Val(Y) \end{cases}$$

The quality of the importance sampling estimator depends largely on how close P and Q.

if all evidences are at roots, $\tilde{P}(x) = Q(x)$, $\hat{P}_B(y|e) = \frac{1}{M}\sum_i \mathbb{1}\{\xi[i]\langle Y\rangle = y\}$ is exactly the posterior

if all evidences are at leaves, $Q(x)$ is the prior $P_B(x)$, $\hat{P}_B(y|e) \approx \frac{1}{M}\sum_i \mathbb{1}\{\xi[i]\langle Y\rangle = y\}$ only when $P_B(x) \approx P_B(x, e)$

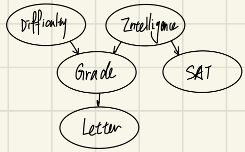
· Gibbs Sampling

```
def gibbs_sample(X: set of variables to be sampled, Φ: factors defining P_Φ, P^(0)(X): initial state distribution, T: time steps):
    sample x^(0) from P^(0)(X)
    for t = 1, ..., T:                    // e.g. P^(0)(X) can be the distribution induced by mutilated network
        x^(t) = x^(t-1)
        for each X_i ∈ X:
            sample x_i^(t) from P_Φ(X_i | x_{-i})
            change X_i in x^(t)
    return x^(0), ..., x^(T)
```

eg. having the observation $S=s^1, L=l^0$

the set of reduced factor Φ is

$$\{P(I), P(D), P(G|I,D), P(s^1|I), P(l^0|G)\}$$



begin: sample from mutilated network

$d^{(0)} = d^1 \quad i^{(0)} = i^0 \quad g^{(0)} = g^2$

1st iteration: sample in order (say G I D)

sample $g^{(1)} \sim P_\Phi(G|d^1, i^0)$

$$P_\Phi(G|d^1, i^0) = \frac{P_\Phi(G, d^1, i^0)}{P_\Phi(d^1, i^0)} \propto P_\Phi(G, d^1, i^0) = P(i^0)\, P(d^1) \cdot P(G|i^0, d^1)\, P(l^0|G)\, P(s^1|i^0)$$

(multiply all factors that contains G, with other variables instantiated, and renormalize over G.)

$g^{(1)} = g^{(3)}$

sample $i^{(1)} \sim P_\Phi(I|d^1, g^3) \qquad i^{(1)} = i^1$

sample $d^{(1)} \sim P_\Phi(D|i^1, g^3)$

The sampling process for G takes into consideration the downstream evidence at child L

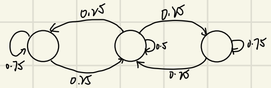
$\therefore$ is arguably closer to the posterior than forward sampling

As we repeat the sampling process, the distribution from which we generate each sample gets closer and closer to the posterior $P_\Phi(X) = P(X|e)$

· Markov chain

A Markov chain is defined via

- a state space $Val(X)$
- a model defines a next-state distribution over $Val(X)$ $\quad T(x \to x') = P(x'|x)$



$$P^{(t+1)}(X^{(t+1)} = x') = \sum_{x \in Val(X)} P(X^{(t)} = x) \cdot T(x \to x')$$

Markov chain Monte Carlo (MCMC) sampling is a process that mirrors the dynamic of the markov chain

```
def MCMC_sample( P^(0)(X): initial state distribution , T: Markov chain transition model , T: Time steps ):
        Sample x^(0) from P^(0)(X)
        for t=1, ..., T:
                sample x^(t) from T(x^(t-1) -> X)
        return x^(0), ... x^(T)
```

As the process converges, we would expect that $P^{(t+1)}$ to be close to $P^{(t)}$

(Markov matrix has at least eigenvalue $\lambda=1$ and $\lambda_i \le 1$ for all evs).

the resulting distribution $\pi(x)$ being an equilibrium / steady state

A distribution $\pi(x)$ is a stationary distribution for a Markov chain T if

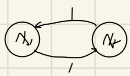
$$\pi(X=x') = \sum_{x \in Val(X)} \pi(X=x)\, T(x \to x')$$

$$\begin{bmatrix} \pi \end{bmatrix} \qquad \begin{bmatrix} T \end{bmatrix}\begin{bmatrix} \pi \end{bmatrix}$$

$T\pi = \pi$

$\pi(x)$ is an eigenvector of T associated with eigenvalue $\lambda=1$

eg. periodic markov chains $T(x^1 \to x^2)=1$, $T(x^2 \to x^1)=1$

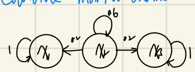


$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \{(\lambda, v)\} = \left\{ \left(1, \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right) \right\}$$

only one eigenvector, no convergence to a stationary distribution

there is no guarantee that MCMC sampling process converges to a stationary distribution

eg. reducible markov chains



$$T = \begin{bmatrix} 1 & 0.2 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0.2 & 1 \end{bmatrix} \quad \{(\lambda, v)\} = \left\{ \left(1, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right), \left(1, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right), \left(0.6, \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}\right) \right\}$$

$$P^{(t)}(x) = c_1 \cdot 1^t \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + c_2 \cdot 1^t \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + c_3 \cdot 0.6^t \cdot \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}$$

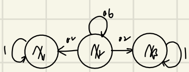
where $P^{(0)}(x) = c_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + c_3 \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}$

there is no guarantee that the stationary distribution is unique

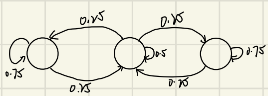
A Markov chain is regular if $\exists\, k$ s.t.

$\forall\, x, x' \in Val(X)$, the probability of getting from x to x' in exactly k steps is >0

eg.



not regular



regular, $k=2$

If a finite state Markov chain T is regular, then it has a unique stationary distribution

(if $0 \le T_{ij} < 1$ and $\sum_j T_{ij} = 1$, then T has a ev $\lambda=1$ and other evs $\lambda_i < 1$) (perron-frobenius theorem)

· Markov Chain : Multiple Transition Models

decompose the transition model, states of which consists of several vars.
consider transitions that update only a single variable at a time

define a set of transition models, each with its own dynamics.
each such transition model $\mathcal{T}_i$ is called a kernel

In the case of graphical models, can define a multikernel chain
have a kernel $\mathcal{T}_i$ for each $X_i \in \mathcal{X}$, model $\mathcal{T}_i$ takes a state (x_{-i}, x_i) to a state (x_{-i}, x_i')

· Gibbs Sampling via Markov chain

wish to generate samples from posterior $P(X|E=e)$
thus need to define a chain for which $P(X|E=e)$ is the stationary distribution

Gibbs sampling is based on the multikernel chain

$$\mathcal{T}_i\big((x_{-i}, x_i) \to (x_{-i}, x_i')\big) = P(x_i' | x_{-i})$$

the transition probability does not depend on the current value x_i,
depends only on x_{-i}, $P(X|e)$ is a stationary distribution of Gibbs sampling

let $P^{(t)}(x) = P(X|e)$, Gibbs sampling updates X_i at this step.

$$\begin{aligned}
P^{(t+1)}(x_i', x_{-i}) &= \sum_{x \in Val(X)} P^{(t)}(x_i, x_{-i}) \cdot \mathcal{T}\big((x_i, x_{-i}) \to (x_i', x_{-i})\big) \\
&= \sum_{x_i \in Val(X_i)} P^{(t)}(x_i, x_{-i}) \cdot \mathcal{T}_i\big((x_{-i}, x_i) \to (x_i', x_{-i})\big) \\
&= \sum_{x_i \in Val(X_i)} P(x_i, x_{-i} | e) \cdot P(x_i' | x_{-i}, e) \\
&= \sum_{x_i \in Val(X_i)} P(x_i, x_{-i} | e) \cdot \frac{P(x_i', x_{-i} | e)}{P(x_{-i} | e)} \\
&= \sum_{x_i \in Val(X_i)} P(x_i, x_{-i} | e) \cdot \frac{P(x_i', x_{-i} | e)}{\sum_{x_i''} P(x_i'', x_{-i} | e)} \\
&= P(x_i', x_{-i} | e)
\end{aligned}$$

when applying Gibbs sampling with Markov chain $\mathcal{T}_i\big((x_i, x_{-i}) \to (x_i', x_{-i})\big) = P(x_i' | x_{-i})$
if $(x_i, x_{-i}) \sim P(X|e)$
then $(x_i', x_{-i}') \sim P(X|e)$
$\therefore$ $P(X|e)$ is a stationary distribution of Gibbs Sampling

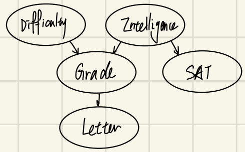
Calculate the transition probability

$$\begin{aligned}
P_{\Phi}(X) &= \frac{1}{Z} \prod_j \phi_j(D_j) \\
&= \frac{1}{Z} \prod_{j: X_i \in D_j} \phi_j(D_j) \cdot \prod_{j: X_i \notin D_j} \phi_j(D_j)
\end{aligned}$$

$$\begin{aligned}
P(x_i' | x_{-i}) &= \frac{P(x_i', x_{-i})}{P(x_{-i})} = \frac{P(x_i', x_{-i})}{\sum_{x_i} P(x_i, x_{-i})} \\
&= \frac{\frac{1}{Z}\left(\prod_{X_i \in D_j} \phi_j\right)(x_i', x_{-i}) \cdot \left(\prod_{X_i \notin D_j} \phi_j\right)(x_{-i})}{\frac{1}{Z}\sum_{x_i}\left(\prod_{X_i \in D_j} \phi_j\right)(x_i, x_{-i}) \cdot \left(\prod_{X_i \notin D_j} \phi_j\right)(x_{-i})} \\
&= \underset{X_i}{\text{renormalize}}\left[\left(\prod_{X_i \in D_j} \phi_j\right)(X_i, x_{-i})\right]
\end{aligned}$$

use only the factors involving X_i, depends only on the instantiation of X_i's Markov Blanket eg $E=\{S=s^1, L=l^0\}$

$T((i,g,d,s^1,l^0)\to(i,g',d,s^1,l^0)) = P(g'|i,d,s^1,l^0)$



$$P(g'|i,d,s^1,l^0) = \frac{P(g',i,d,s^1,l^0)}{P(i,d,s^1,l^0)}$$

$$= \frac{1}{Z}\cdot \phi_D(D)\cdot\phi_I(I)\,\phi_G(g',D,I)\cdot\phi_S(s^1,I)\cdot\phi_L(l^0,g')$$

$$= \text{renormalize}(\phi_G(g',d,i)\cdot\phi_L(l^0,g'))$$

Let $\mathcal{H}$ be a Markov network s.t. all of the clique potentials are strictly positive.
then the gibbs-sampling markov chain is regular

· Broader Class of Markov Chain

In continuous models, the conditional prob $P(X_i|x_{-i})$ may not have a parametric form that allows sampling
the Gibbs sampling is not applicable

the Gibbs sampling updates one var at a time, if vars are highly correlated convergence may be slow

A finite-state Markov chain T is reversible if
$\exists$ a unique distribution π st $\forall\, x, x' \in Val(X)$
$\pi(x)\cdot T(x\to x') = \pi(x')\cdot T(x'\to x)$ detailed balance

$\pi(x)\cdot T(x\to x')$ = prob of a transition $x\to x'$
detailed balance: transition $x\to x'$ is equally likely as $x'\to x$

reversibility implies that π is "a" stationary distribution of T

$$\sum_x \pi(x)\cdot T(x\to x') = \sum_x \pi(x')\cdot T(x'\to x) = \pi(x')\cdot\sum_x T(x'\to x) = \pi(x')$$

$$\pi(x') = \sum_x \pi(x)\cdot T(x\to x')$$

(not necessarily that the chain will converge to π)

if a Markov chain is regular and satisfies the detailed balance equation relative to π
then π is the unique stationary distribution of T

in the case of multiple kernel, if each kernel T_i satisfies the detailed balance equation
then so does the mixture transition model T

· Metropolis - Hasting Algorithm

a general construction that allows to build a reversible Markov chain with a particular stationary distribution

The proposal distribution $\mathcal{T}^Q$ defines a transition model over state space

At each transition, we can $\begin{cases} \text{either accept the proposal and transition to } x' \\ \text{or reject the proposal and stay and } x \end{cases}$

for each pair of states (x, x'), have a acceptance probability $\mathcal{A}(x \to x')$

The actual transition model of the Markov chain is

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x') \cdot \mathcal{A}(x \to x') \qquad \text{// probability of make transition}$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x') \cdot [1 - \mathcal{A}(x \to x')] \qquad \text{// prob of staying} = p(\text{stay}) + \sum p(\text{trans}) \cdot p(\text{reject})$$

Given a proposal distribution $\mathcal{T}^Q$, can use the detailed balance equation to select the acceptance probability

$$\pi(x)\, \mathcal{T}^Q(x \to x') \cdot \mathcal{A}(x \to x') = \pi(x')\, \mathcal{T}^Q(x' \to x) \cdot \mathcal{A}(x' \to x)$$

Given any proposal distribution $\mathcal{T}^Q$, consider the markov chain $\begin{cases} \mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\, \mathcal{A}(x \to x') \\ \mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')\, [1 - \mathcal{A}(x \to x')] \end{cases}$

where $\mathcal{A}(x \to x') = \min\left[1, \dfrac{\pi(x')\, \mathcal{T}^Q(x' \to x)}{\pi(x)\, \mathcal{T}^Q(x \to x')}\right]$

if the markov chain is regular, then it has the stationary distribution π

$$\text{if } \frac{\pi(x')\, \mathcal{T}^Q(x' \to x)}{\pi(x)\, \mathcal{T}^Q(x \to x')} \leq 1 \quad \mathcal{A}(x \to x') = \frac{\pi(x')\, \mathcal{T}^Q(x' \to x)}{\pi(x) \cdot \mathcal{T}^Q(x \to x')} \quad \mathcal{A}(x' \to x) = 1$$

$$\pi(x) \cdot \mathcal{T}^Q(x \to x') \cdot \mathcal{A}(x \to x') = \pi(x) \cdot \mathcal{T}^Q(x \to x') \cdot \frac{\pi(x')\, \mathcal{T}^Q(x' \to x)}{\pi(x) \cdot \mathcal{T}^Q(x \to x')} = \pi(x') \cdot \mathcal{T}^Q(x' \to x) \cdot \mathcal{A}(x' \to x)$$

$$\pi(x) \cdot \mathcal{T}(x \to x') = \pi(x') \cdot \mathcal{T}(x' \to x)$$

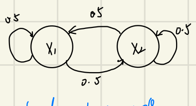
$\therefore$ the markov chain satisfies the detailed balance equation

$\therefore$ π is a stationary distribution

$\because$ $\mathcal{T}$ is regular

$\therefore$ π is the unique stationary distribution

eg. $\mathcal{T}^Q$:



The proposal distribution $\mathcal{T}^Q$ is defined by the chain

want to sample from a stationary distribution $\pi(x_1) = 0.25$ $\pi(x_2) = 0.75$

$$\mathcal{A}(x_1 \to x_2) = \min\left[1, \frac{\pi(x') \cdot \mathcal{T}^Q(x' \to x)}{\pi(x) \cdot \mathcal{T}^Q(x \to x')}\right] = \min\left[1, \frac{3/4 \cdot 1/2}{1/4 \cdot 1/2}\right] = 1$$

$$\mathcal{A}(x_2 \to x_1) = 1/3$$

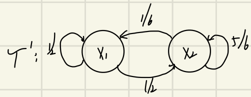
$$\mathcal{T}'(x_1 \to x_2) = \mathcal{T}^Q(x_1 \to x_2) \cdot \mathcal{A}(x_1 \to x_2) = 1/2$$

$$\mathcal{T}'(x_2 \to x_1) = \mathcal{T}^Q(x_2 \to x_1) \cdot \mathcal{A}(x_2 \to x_1) = 1/6$$

$$\mathcal{T}'(x^1 \to x^1) = \mathcal{T}^Q(x_1 \to x_1) + \mathcal{T}^Q(x_1 \to x^2) \cdot (1 - \mathcal{A}(x_1 \to x^2)) = 1/2$$

$$\mathcal{T}'(x^2 \to x^2) = \mathcal{T}^Q(x^2 \to x^2) + \mathcal{T}^Q(x_2 \to x^1) \cdot (1 - \mathcal{A}(x_2 \to x^1)) = 1/2 + 1/2 \cdot \frac{2}{3} = \frac{5}{6}$$

the resulting chain $\mathcal{T}'$ is given by



$$\begin{bmatrix} 1/2 & 1/6 \\ 1/2 & 5/6 \end{bmatrix} \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}$$

$\therefore \pi$ is a eigenvector / stationary distribution of $\mathcal{T}'$

In the context of Graphical model, each local transition model $\mathcal{T}_i$ is defined via an associated proposal distribution $\mathcal{T}_i^{Q_i}$
(P_Φ is the target distribution, which should be the stationary distribution of the resulting chain

$$A(x_{-i}, x_i \to x_{-i}, x_i') = \min\left[1, \frac{\pi(x_{-i}, x_i') \cdot \mathcal{T}_i^{Q_i}((x_{-i}, x_i') \to (x_{-i}, x_i))}{\pi(x_{-i}, x_i) \cdot \mathcal{T}_i^{Q_i}((x_{-i}, x_i) \to (x_{-i}, x_i'))}\right]$$

$$= \min\left[1, \frac{P_\Phi(x_{-i}, x_i') \cdot \mathcal{T}_i^{Q_i}((x_{-i}, x_i') \to (x_{-i}, x_i))}{P_\Phi(x_{-i}, x_i) \cdot \mathcal{T}_i^{Q_i}((x_{-i}, x_i) \to (x_{-i}, x_i'))}\right]$$

where $\frac{P_\Phi(x_i', x_{-i})}{P_\Phi(x_i, x_{-i})} = \frac{P_\Phi(x_i' | x_{-i}) \cdot P_\Phi(x_{-i})}{P_\Phi(x_i | x_{-i}) \cdot P_\Phi(x_{-i})} = \frac{P_\Phi(x_i' | x_{-i})}{P_\Phi(x_i | x_{-i})} = \frac{P_\Phi(x_i' | u_i)}{P_\Phi(x_i | u_i)}$

$\frac{P_\Phi(x_i', x_{-i})}{P_\Phi(x_i, x_{-i})} = \frac{P_\Phi(x_i' | u_i)}{P_\Phi(x_i | u_i)}$ and u_i is the Markov Blanket of x_i

$P_\Phi(x_i | u_i)$ can be efficiently computed based only on local parametrization

def metropolis_hasting(X: set of variables to be sampled, Φ: factors defining P_Φ, $\mathcal{T}_i^{Q_i}$: set of proposal distribution):

initialize $x^{(0)}$ for $t=0$

for $t = 1, \ldots T$:

sample $x^{(t)}$ from $\mathcal{T}^{Q}(x^{(t-1)} \to x^{(t)})$

for each $X_i \in X$:

propose a transition $(x_{-i}, x_i) \to (x_{-i}, x_i')$ via $\mathcal{T}_i^{Q_i}$

compute the acceptance prob $A((x_{-i}, x_i) \to (x_{-i}, x_i')) = \min\left[1, \frac{P_\Phi(x_{-i}, x_i') \cdot \mathcal{T}_i^{Q_i}((x_{-i}, x_i') \to (x_{-i}, x_i))}{P_\Phi(x_{-i}, x_i) \cdot \mathcal{T}_i^{Q_i}((x_{-i}, x_i) \to (x_{-i}, x_i'))}\right]$

change / keep X_i in $x^{(t)}$ with prob $A((x_{-i}, x_i) \to (x_{-i}, x_i'))$

return $x^{(1)}, \ldots x^{(T)}$

Gibbs sampling is a special case of metropolis-hasting when the proposal $\mathcal{T}_i^{Q_i}((x_{-i}, x_i) \to (x_{-i}, x_i')) = P_\Phi(x_i' | x_{-i})$

$$\frac{P_\Phi(x_{-i}, x_i') \cdot P_\Phi(x_i | x_{-i})}{P_\Phi(x_{-i}, x_i) \cdot P_\Phi(x_i' | x_{-i})} = \frac{P_\Phi(x_i' | x_{-i}) \cdot P_\Phi(x_{-i}) \cdot P(x_i | x_{-i})}{P_\Phi(x_i | x_{-i}) \cdot P_\Phi(x_{-i}) \cdot P_\Phi(x_i' | x_{-i})} = 1$$

the acceptance prob simplifies to 1