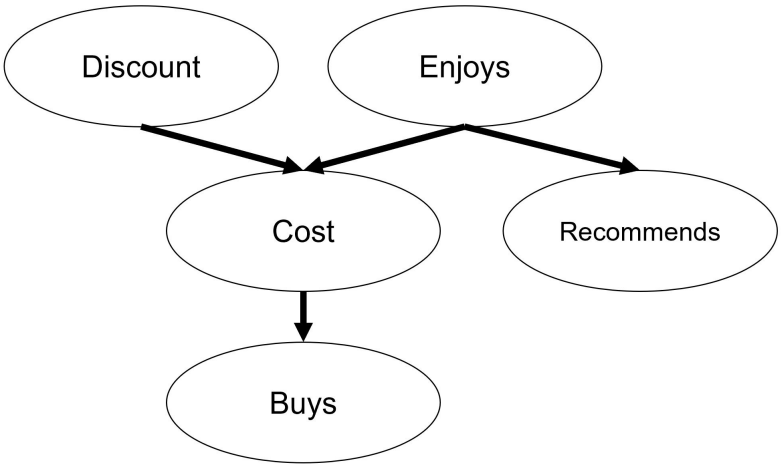
# 1 Inference

Realizing that students aren't particularly fond of reading the textbook, the 281 course staff have developed a software that automatically scans the textbook and outputs key points for each individual chapter. However, since the development of the software requires time and computational resources, the 281 staff decides to offer a free one month trial to students, after which a paid subscription is necessary to keep using the software. The following network and variables are used to represent the problem:

- $Discount(D)$: $+d$ if a discount is offered, $-d$ otherwise
- $Enjoys(E)$: $+e$ if a student enjoys the software, $-e$ otherwise
- $Cost(C)$: $+c$ if the software cost is $< 20$, $-c$ otherwise
- $Recommends(R)$: $+s$ if the student recommends the software to a friend, $-s$ otherwise
- $Buys(B)$: $+b$ if the student buys a software subscription, $-b$ otherwise




(a) How can we represent the probability that a student buys and recommends the software using the conditional probabilities at each node?

$P(+b, +r) = \sum_{c,d,e} P(+b|c)P(c|d,e)P(d)P(e)P(+r|e)$

This sum is equivalent to summing out the hidden variables in the join distribution: $\sum_{c,d,e} P(d, e, c, +r, +b)$.

(b) The staff has surveyed students and collected data on whether the students enjoyed the software or not. With this information, we want to perform a inference on a joint distribution where the query variable is $Buys$ $(B)$.

(i) How can we represent the probability expression in terms of conditional probabilities from the network?

$P(B|E) = \alpha P(B, E) = \alpha \sum_{d,c,r} P(B, E, d, c, r) = P(E)\alpha \sum_{d,c,r} P(d)P(c|d, E)P(B|c)P(r|E)$

Note: Equation 13.9 on page 493 of the TB goes into detail about why we use $\alpha$. In short, when we are calculating conditional probabilities, $\alpha$ acts as a normalization constant. However, we can proceed with calculating the conditional probabilities even without knowing the value of $\alpha$ because relative proportions remain the same without normalization (e.g. relative proportions of $P(+b|E)$ and $P(-b|E)$ remain the same without knowing the exact value of $\alpha = 1/P(E)$).

(ii) What are the hidden and evidence variable(s)?

The hidden variables are $D, C, R$, and the evidence variable is $E$.

(c) Using the probability expression from the previous part, we want to compute the query $B$ given evidence that the student enjoys the software. Assume the variable ordering is in alphabetical order.

(i) How many factors are there, and what are the dimensions of each factor?

Our expression is: $P(B|+e) = \alpha P(+e) \sum_{r,d,c} P(d)P(c|d,+e)P(B|c)P(r|+e)$

$= \alpha P(+e) \sum_r P(r|+e) \sum_d PP(d) \sum_c P(c|d,+e)(B|c)$

Each conditional probability corresponds to an individual factor, so there are 5 factors total. The factor for $P(d)$ and $P(r|+e)$ each have dimension $2 \times 1$, the factors for $P(c|d,+e)$ and $P(B|c)$ each have dimension $2 \times 2$, and the factor for $P(+e)$ is a one-element vector.

(ii) Run the variable elimination algorithm to eliminate repeated computations for the expression $P(B|+e)$.

All factors: $P(D), P(+e), P(C|D,+e), P(B|C), P(R|+e)$

- Choose C: The relevant factors are $P(C|D,+e), P(B|C)$. We sum out $C$ to get $f_1(D,B) = \sum_c P(C=c|D,+e)P(B|C=c)$.

  Expression: $P(B|+e) = \alpha P(+e) \sum_{d,r} P(D=d)P(R=r|+e) \times f_1(D,B)$
- Choose D: We sum out the relevant factors $P(D)$ and $f_1(D,B)$ to get $f_2(B) = \sum_d f_1(D=d,B)P(d)$.

  Expression: $P(B|+e) = \alpha P(+e) \sum_r P(R=r|+e) \times f_2(B)$
- Choose R: We sum out the relevant factor $P(R|+e)$ to get $f_3(R) = \sum_r P(R=r|+e)$.

  Expression: $P(B|+e) = \alpha P(+e) f_3(R) \times f_2(B)$

(iii) How does the resulting expression change if the variable ordering is instead in reverse alphabetical order?

- Choose R: We sum out relevant factor $P(R|+e)$ to get $\sum_r P(R=r|+e) = 1$. We can discard this variable since it is irrelevant.
- Choose D: We sum out relevant factors $P(D)$, $P(C|D,+e)$ to get $f_2(C) = \sum_d P(C|D=d,+e) \times P(D=d)$.

  Expression: $P(B|+e) = \alpha P(+e) \sum_c P(B|c) \times f_2(C=c)$
- Choose C: We sum out relevant factors $P(B|c)$ and $f_2(C)$ to get $f_3(B) = \sum_c P(B|C=c) \times f_2(C=c)$.

  $P(B|+e) = \alpha P(+e) f_3(B)$

(iv) How do the two orderings compare with respect to time and space complexity?

When the terms were ordered in alphabetical order, the largest factor had 2 variables. When the terms were ordered in reverse alphabetical order, the largest factor had 1 variable. Since the size of the largest factor determines the space/time complexity, the second ordering performs better.

(v) Describe a heuristic that could be useful in determining a variable ordering to minimize the size of the largest factor.

Potential ideas:

- Eliminate whichever variable minimizes the size of the next factor to be constructed.
- Eliminate the variable with the fewest dependent variables

# 2 Sampling

(a) Compared to other sampling methods (rejection, likelihood weighting, Gibbs), what kind of information can prior sampling not use (that other methods can)?

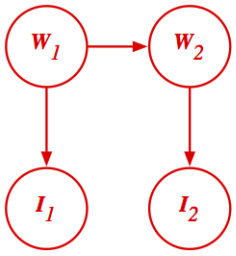
Other methods can compute probabilities with fixed evidence, while the network for prior sampling has no evidence associated.

(b) How does reject sampling work on a high level, and what is its biggest/immediate weakness?

It generates samples from the given prior distribution, rejects all samples that do not match the evidence, and then derives the probability (# times the desired value appears in the remaining samples).

Its biggest weakness is potential inefficiency when evidence is rare. Most samples would then be rejected, so all this information would be thrown away despite being calculated.

The diagram below describes a person's ice-cream eating habits based on the weather. The nodes $W_i$ stand for the weather on a day $i$, which can either be **s** (sunny) or **r** (rainy). The nodes $I_i$ represent whether the person ate ice-cream on day $i$, which can either be **t** (true) or **f** (false).




| $W_1$ | $P(W_1)$ |
|---|---|
| s | 0.6 |
| r | 0.4 |

| $I$ | $W$ | $P(I\|W)$ |
|---|---|---|
| t | s | 0.9 |
| f | s | 0.1 |
| t | r | 0.2 |
| f | r | 0.8 |

| $W_2$ | $W_1$ | $P(W_2\|W_1)$ |
|---|---|---|
| s | s | 0.7 |
| r | s | 0.3 |
| s | r | 0.5 |
| r | r | 0.5 |

Assume we generate the following six samples given the evidence $I_1 = t$ and $I_2 = f$:

$$(W_1, I_1, W_2, I_2) = < s, t, r, f >, < r, t, r, f >, < s, t, r, f >, < s, t, s, f >, < s, t, s, f >, < r, t, s, f >$$

Using these samples, we will complete the following table:

| $(W_1, I_1, W_2, I_2)$ | Count/N | $w$ | Joint |
|---|---|---|---|
| s, t, s, f | 2/6 | 0.09 | 0.03 |
| s, t, r, f | /6 | | |
| r, t, s, f | /6 | | |
| r, t, r, f | /6 | | |

(c) What is the weight of the sample (s, t, r, f) above? Recall that the weight given to a sample in likelihood weighting is:

$$w = \prod_{\text{Evidence variables } e} P(e|\text{Parents}(e)).$$

In this case, the evidence is $I_1 = t, I_2 = f$. The weight of the first sample is therefore

$w = Pr(I_1 = t|W_1 = s) \cdot Pr(I_2 = f|W_2 = r) = 0.9 \cdot 0.8 = 0.72$

(d) What is the estimate of P(s, t, r, f) given the samples?

The estimate of the joint probability is simply $Count/N * w = 2/6 * 0.72 = 0.24$.

(e) Compute the rest of the entries in the table. Use the estimated joint probabilities to estimate $P(W_2|I_1 = t, I_2 = f)$.

| $(W_1, I_1, W_2, I_2)$ | Count/N | $w$ | Joint |
|---|---|---|---|
| s, t, s, f | 2/6 | 0.09 | 0.03 |
| s, t, r, f | 2/6 | 0.72 | 0.24 |
| r, t, s, f | 1/6 | 0.02 | 0.003 |
| r, t, r, f | 1/6 | 0.16 | 0.027 |

To compute the probabilities, we sum out variables as usual:

$P(W_2 = r|I_1 = t, I_2 = f) = P(I_1 = t, W_2 = r, I_2 = f) \ / \ P(I_1 = t, I_2 = f)$

We sum over $W_1$ using the rows from the table:

$P(W_2 = r, I_1 = t, I_2 = f) = \sum_{w_1} P(W_1 = w_1, I_1 = t, W_2 = r, I_2 = f) = 0.24 + 0.027 = 0.267$

Since all the rows in the table have $I_1 = t, I_2 = f$, the probability is just the sum of all the joint probabilities.

$P(I_1 = t, I_2 = f) = 0.03 + 0.24 + 0.003 + 0.027 = 0.3$

So $P(W_2 = r|I_1 = t, I_2 = f) = 0.267 \ / \ 0.3 = 0.89$.

(f) What is a weakness of likelihood weighing sampling? How does Gibbs sampling work, and how does it address this limitation?

Likelihood weighing only conditions on upstream evidence (so evidence only influences the choice of downstream variables).

Gibbs sampling starts with an arbitrary instantiation of a complete sample (consistent with evidence), and then samples on one variable at a time, conditioned on all the rest, while keeping evidence consistent. This way, both upstream and downstream variables condition on evidence.

# 3 HMMs: Tracking a Jabberwock

You have been put in charge of a Jabberwock for your friend Lewis. The Jabberwock is kept in a large tugley wood which is conveniently divided into an $N \times N$ grid. It wanders freely around the $N^2$ possible cells. At each time step $t = 1, 2, 3, \ldots$, the Jabberwock is in some cell $X_t \in \{1, \ldots, N\}^2$, and it moves to cell $X_{t+1}$ randomly as follows: with probability $1 - \epsilon$, it chooses one of the (up to 4) valid neighboring cells uniformly at random; with probability $\epsilon$, it uses its magical powers to teleport to a random cell uniformly at random among the $N^2$ possibilities (it might teleport to the same cell). Suppose $\epsilon = \frac{1}{2}$, $N = 10$ and that the Jabberwock always starts in $X_1 = (1, 1)$.

(a) Compute the probability that the Jabberwock will be in $X_2 = (2, 1)$ at time step 2. What about $P(X_2 = (4, 4))$?

$P(X_2 = (2,1)) = 1/2 \cdot 1/2 + 1/2 \cdot 1/100 = 0.255$
$P(X_2 = (4,4)) = 1/2 \cdot 1/100 = 0.005$

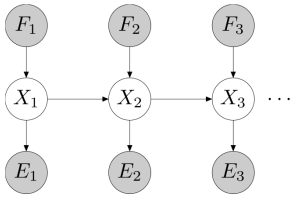
At each time step $t$, you don't see $X_t$ but see $E_t$, which is the row that the Jabberwock is in; that is, if $X_t = (r, c)$, then $E_t = r$. You still know that $X_1 = (1, 1)$.

(b) Suppose we see that $E_1 = 1$, $E_2 = 2$. Fill in the following table with the distribution over $X_t$ after each time step, taking into consideration the evidence. Your answer should be concise. Hint: you should not need to do any heavy calculations.

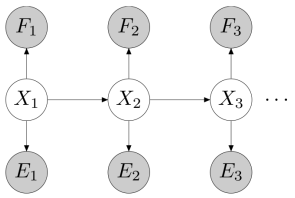
| $t$ | $P(X_t \mid e_{1:t-1}, X_1 = (1,1))$ | $P(X_t \mid e_{1:t}, X_1 = (1,1))$ |
|---|---|---|
| 1 | | |
| 2 | | |

| $t$ | $P(X_t \mid e_{1:t-1}, X_1 = (1,1))$ | $P(X_t \mid e_{1:t}, X_1 = (1,1))$ |
|---|---|---|
| 1 | $X_1$ \| $P(X_1 \mid X_1 = (1,1))$: (1, 1) \| 1; all other values \| 0 | $X_1$ \| $P(X_1 \mid e_1, X_1 = (1,1))$: (1, 1) \| 1; all other values \| 0 |
| 2 | $X_2$ \| $P(X_2 \mid e_1, X_1 = (1,1))$: (1, 2) \| 51/200; (2, 1) \| 51/200; all other values \| 1/200 | $X_2$ \| $P(X_2 \mid e_{1:2}, X_1 = (1,1))$: (2, 1) \| 51/60; $(2, a)$ $(\forall a, a > 1)$ \| 1/60; all other values \| 0 |

You are a bit unsatisfied that you can't pinpoint the Jabberwock exactly. But then you remembered Lewis told you that the Jabberwock teleports only because it is frumious on that time step, and it becomes frumious independently of anything else. Let us introduce a variable $F_t \in \{0, 1\}$ to denote whether it will teleport at time $t$. We want to to add these frumious variables to the HMM. Consider the two candidates:




(A)




(B)

| (A) | (B) |
|---|---|
| $X_1 \perp X_3 \mid X_2$ | $X_1 \perp X_3 \mid X_2$ |
| $X_1 \perp E_2 \mid X_2$ | $X_1 \perp E_2 \mid X_2$ |
| $X_1 \perp F_2 \mid X_2$ | $X_1 \perp F_2 \mid X_2$ |
| $X_1 \perp E_4 \mid X_2$ | $X_1 \perp E_4 \mid X_2$ |
| $X_1 \perp F_4 \mid X_2$ | $X_1 \perp F_4 \mid X_2$ |
| $E_3 \perp F_3 \mid X_3$ | $E_3 \perp F_3 \mid X_3$ |
| $E_1 \perp F_2 \mid X_2$ | $E_1 \perp F_2 \mid X_2$ |
| $E_1 \perp F_2 \mid E_2$ | $E_1 \perp F_2 \mid E_2$ |

(c) For each model, circle the conditional independence assumptions above which are true in that model.

| (A) | (B) |
|---|---|
| $X_1 \perp X_3 \mid X_2$ + | $X_1 \perp X_3 \mid X_2$ + |
| $X_1 \perp E_2 \mid X_2$ + | $X_1 \perp E_2 \mid X_2$ + |
| $X_1 \perp F_2 \mid X_2$ | $X_1 \perp F_2 \mid X_2$ + |
| $X_1 \perp E_4 \mid X_2$ + | $X_1 \perp E_4 \mid X_2$ + |
| $X_1 \perp F_4 \mid X_2$ + | $X_1 \perp F_4 \mid X_2$ + |
| $E_3 \perp F_3 \mid X_3$ + | $E_3 \perp F_3 \mid X_3$ + |
| $E_1 \perp F_2 \mid X_2$ | $E_1 \perp F_2 \mid X_2$ + |
| $E_1 \perp F_2 \mid E_2$ | $E_1 \perp F_2 \mid E_2$ |

(d) Which Bayes net is more appropriate for the problem domain here, (A) or (B)? Justify your answer.

(A) because the choice of X depends on F in the problem description.

For the following questions, your answers should be fully general for models of the structure shown above, not specific to the teleporting Jabberwock.

(e) For (A), express $P(X_{t+1}, e_{1:t+1}, f_{1:t+1})$ in terms of $P(X_t, e_{1:t}, f_{1:t})$ and the conditional probability tables used to define the network. Assume the $E$ and $F$ nodes are all observed.

$$P(x_{t+1}, e_{1:t+1}, f_{1:t+1}) = P(e_{t+1}|x_{t+1})P(f_{t+1}) \sum_{x_t} P(x_{t+1}|x_t, f_{t+1})P(x_t, e_{1:t}, f_{1:t}).$$

We're already provided with $P(x_t, e_{1:t}, f_{1:t})$. To get $P(x_t + 1, e_{1:t}, f_{1:t})$, we can sum over all $x_t$ and multiply by $P(x_{t+1} \mid x_t, f_{t+1})$, the conditional probability table of $x_{t+1}$.
Then, to get the joint probability $P(x_t + 1, e_{1:t+1}, f_{1:t+1})$, we multiply the above quantity with the emission probability ($P(e_{t+1} \mid x_{t+1})$) and $P(f_{t+1})$, the CPT of $P(f_{t+1})$.

(f) For (B), express $P(X_{t+1}, e_{1:t+1}, f_{1:t+1})$ in terms of $P(X_t, e_{1:t}, f_{1:t})$ and the CPTs used to define the network. Assume the $E$ and $F$ nodes are all observed.

$$P(x_{t+1}, e_{1:t+1}, f_{1:t+1}) = P(e_{t+1}|x_{t+1})P(f_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t)P(x_t, e_{1:t}, f_{1:t}).$$

Similar idea as above, except this time we multiply the joint probability by $P(x_{t+1}|x_t)$, since $x_{t+1}$ now no longer depends on $f_{t+1}$).

Suppose that we don't actually observe the $F_t$s.

(g) For (A), express $P(X_{t+1}, e_{1:t+1})$ in terms of $P(X_t, e_{1:t})$ and the CPTs used to define the network.

$$P(x_{t+1}, e_{1:t+1}) = P(e_{t+1}|x_{t+1}) \sum_{f_{t+1}} P(f_{t+1}) \sum_{x_t} P(x_{t+1}|x_t, f_{t+1})P(x_t, e_{1:t}).$$

(h) For (B), express $P(X_{t+1}, e_{1:t+1})$ in terms of $P(X_t, e_{1:t})$ and the CPTs used to define the network.

$$P(x_{t+1}, e_{1:t+1}) = P(e_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t)P(x_t, e_{1:t}).$$

For (g) and (h), we essentially use the same logic as (e) and (f). However, we no longer need the $F_t$s in the joint probability - so for any probability values that are conditioned on an $f_t$, we multiply by $P(f_t)$ and sum over all possible $f_t$ values. If not (i.e., for graph (B)), we simply drop that term when computing the joint probability.