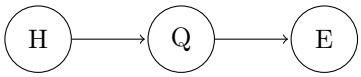
1 Inference Conceptual Review

1. When would we want to use inference?

2. Suppose we are given binary random variables Q, H, E (query, hidden, evidence).
   We want to query $P(q \mid e)$.



   (a) **Enumeration**
   Perform inference on a joint distribution. Use the Bayes net above to break down joint into CPT factors.

   Note: You may use a proportionality constant α in your answer.

   (b) **Variable Elimination**
   Rewrite your answer to enumeration by moving summations inwards as far as possible.

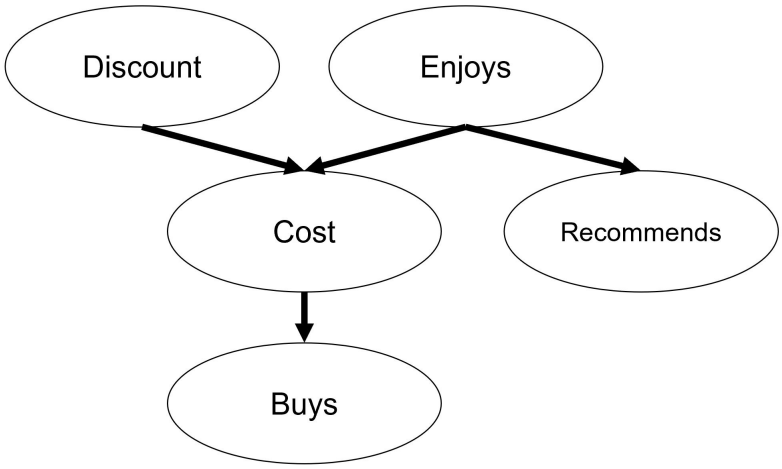
   Note: You may use a proportionality constant α in your answer.

   (c) Based on 2a and 2b, why is variable elimination more efficient than enumeration?

2 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



D	$P(D)$
$+d$	0.7
$-d$	0.3

E	$P(E)$
$+e$	0.1
$-e$	0.9

C	D	E	$P(C \mid D, E)$
$+c$	$+d$	$+e$	0.2
$-c$	$+d$	$+e$	0.8
$+c$	$+d$	$-e$	0.3
$-c$	$+d$	$-e$	0.7
$+c$	$-d$	$+e$	0.9
$-c$	$-d$	$+e$	0.1
$+c$	$-d$	$-e$	0.5
$-c$	$-d$	$-e$	0.5

R	E	$P(R \mid E)$
$+r$	$+e$	0.8
$-r$	$+e$	0.2
$+r$	$-e$	0.5
$-r$	$-e$	0.5

B	C	$P(B \mid C)$
$+b$	$+c$	0.7
$-b$	$+c$	0.3
$+b$	$-c$	0.5
$-b$	$-c$	0.5

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

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

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

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

3. 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.

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

   (b) Run the variable elimination algorithm to eliminate repeated computations for the expression $P(B| + e)$, and fill in the factor table as necessary for each variable eliminated.

   Eliminating C:

D	B	$f_1(D, B)$
$+d$	$+b$	
$-d$	$+b$	
$+d$	$-b$	
$-d$	$-b$	

   Eliminating D:

B	$f_2(B)$
$+b$	
$-b$	

(c) How does the resulting expression change if the variable ordering is instead in reverse alphabetical order? Similarly, fill in the factor table as necessary for each variable eliminated.

Eliminating D:

C	$f_2(C)$
$+c$	
$-c$	

Eliminating C:

B	$f_3(B)$
$+b$	
$-b$	

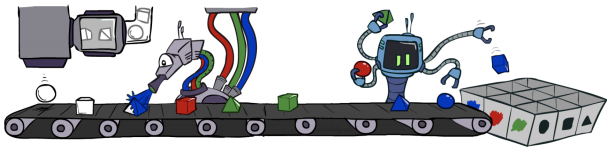
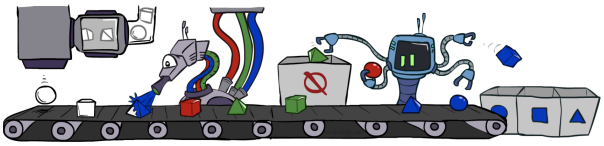
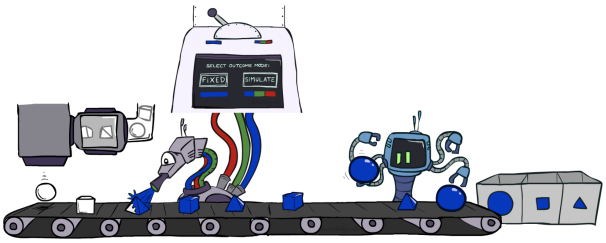
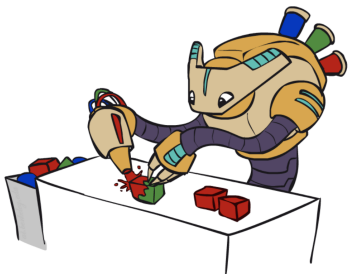
(d) How do the two orderings compare with respect to time and space complexity?

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

3 Candy Sampling

1. In the year 2020, the Oompa Loompas at Charlie's Chocolate Factory have decided that they want to try a new automated way of sampling their candies for quality assurance. However, they have spilled chocolate sauce on their only copy of the user manual! Help out the Oompa Loompas by filling in the blanks below with the names of the four different types of sampling methods we've discussed in lecture, and then match each one to the corresponding image, probability distribution, and algorithm from the tables below.

Sampling Method Name	Image and Distribution (A-D)	Algorithm (1-4):

Name:	Images and Corresponding Probability Distributions:
(A)	 **P(Q, E)**
(B)	 **P(Q | e)**
(C)	 **P(Q, e)**
(D)	 **P(Q | e)**

Name:	Algorithms:

(1)	`def _____(evidence instantiation):`
	`  w = 1.0`
	`  for i = 1,2,...n:`
	`  if(X_i is an evidence variable):`
	`  X_i = observation x_i for X_i`
	`  w = w * P(x_i | Parents(X_i))`
	`  else:`
	`  Sample x_i from P(X_i | Parents(X_i))`
	`  return ((x_1, x_2,...x_n), w)`

(2)	**function** _____(bn) **returns** an event sampled from the prior specified by bn **inputs**: bn, a Bayesian network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$ $\mathbf{x} \leftarrow$ an event with n elements **foreach** variable X_i **in** $X_1, \ldots, X_n$ **do** $\mathbf{x}[i] \leftarrow$ a random sample from $\mathbf{P}(X_i \mid parents(X_i))$ **return x**

(3)	**function** _____(X, **e**, bn, N) **returns** an estimate of $\mathbf{P}(X\|\mathbf{e})$
	local variables: **N**, a vector of counts for each value of X, initially zero
	Z, the nonevidence variables in bn
	x, the current state of the network, initially copied from **e**
	
	initialize **x** with random values for the variables in **Z**
	for j = 1 to N **do**
	for each Z_i in **Z do**
	set the value of Z_i in **x** by sampling from $\mathbf{P}(Z_i\|mb(Z_i))$
	$\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$ where x is the value of X in **x**
	return NORMALIZE(**N**)

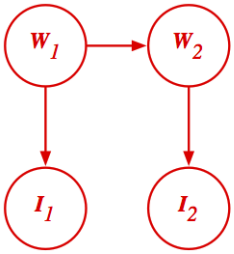
(4)	**function** _____(X, **e**, bn, N) **returns** an estimate of $\mathbf{P}(X\|\mathbf{e})$
	inputs: X, the query variable
	e, observed values for variables **E**
	bn, a Bayesian network
	N, the total number of samples to be generated
	local variables: **N**, a vector of counts for each value of X, initially zero
	
	for j = 1 to N **do**
	$\mathbf{x} \leftarrow$ PRIOR-SAMPLE(bn)
	if x is consistent with **e then**
	$\mathbf{N}[x] \leftarrow \mathbf{N}[x]$+1 where x is the value of X in **x**
	return NORMALIZE(**N**)

4 Sampling Practice

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

2. How does rejection sampling work on a high level, and what is its biggest/immediate weakness?

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 $I_2 = f$ using **Likelihood Weighted Sampling**:

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

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		

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

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

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

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

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