CSE 150A/250A - Homework 4

Due: *Mon May 5* (by 11:59 PM, Pacific Time, via Gradescope)
Grace period: 24 hours

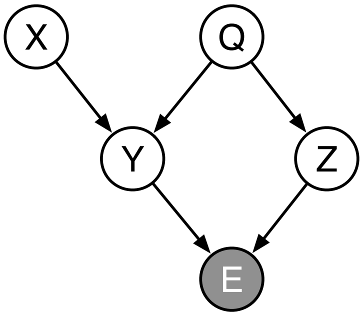
General Guidelines:
You should submit your homework assignments via Gradescope. Typesetting LaTeXis preferred, but neatly handwritten solutions are also accepted. Upload a PDF of your answers to Homework 4 on Gradescope, and carefully select the pages corresponding to each question using the submission interface. If you are submitting handwritten answers, please scan them and create a PDF for upload. Here is a primer on submitting PDF homework via Gradescope:
`https://tinyurl.com/gradescope-guide`
If you have not done this before, please allow some extra time to familiarize yourself with this process.

Late Day Policy: There will be no penalty for turning in any of the homework assignments up to 24 hours late. However, we cannot guarantee that late assignments will be graded in a timely fashion. Beyond the 24-hour grace period, we will not accept late homework.

Collaboration: We strongly encourage collaboration (**but NOT copying**) on the homework assignments. You may talk to anyone in the course about how to solve the problems, and you may even compare your solutions. However, you must write up your solutions yourself, and you may not copy them.

4.1 Likelihood weighting

(a) **Single node of evidence**



Suppose that T samples $\{q_t, x_t, y_t, z_t\}_{t=1}^{T}$ are drawn from the CPTs of the belief network shown above (with fixed evidence $E = e$). Show how to estimate $P(Q\!=\!q|E\!=\!e)$ from these samples using the method of likelihood weighting. Express your answer in terms of sums over indicator functions, such as:

$$I(q, q') = \begin{cases} 1 & \text{if} \quad q = q' \\ 0 & \text{otherwise} \end{cases}$$

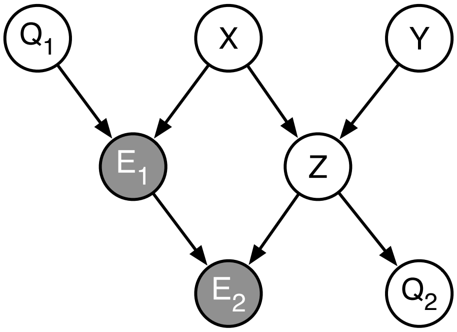
In addition, all probabilities in your answer should be expressed in terms of CPTs of the belief network (i.e., probabilities that do not require any additional computation).

(b) **Multiple nodes of evidence**

Suppose that T samples $\{q_{1t}, q_{2t}, x_t, y_t, z_t\}_{t=1}^{T}$ are drawn from the CPTs of the network shown below (with fixed evidence $E_1\!=\!e_1$ and $E_2\!=\!e_2$). Show how to estimate

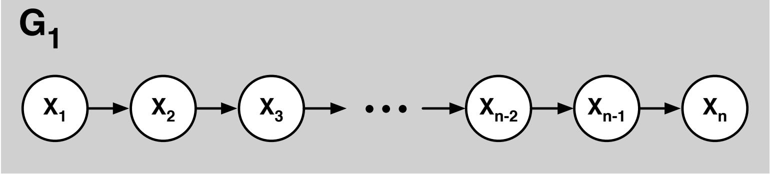
$$P(Q_1\!=\!q_1, Q_2\!=\!q_2|E_1\!=\!e_1, E_2\!=\!e_2)$$

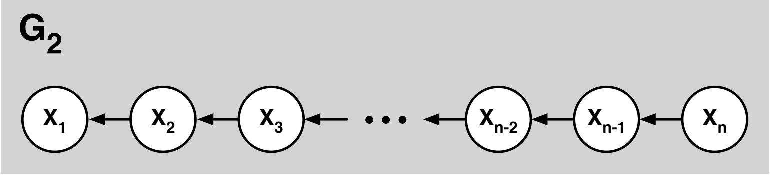
from these samples using the method of likelihood weighting. Express your answer in terms of indicator functions and CPTs of the belief network.



4.2 Maximum likelihood estimation in belief networks (*8 pts*)

Consider the two DAGs shown below, G_1 and G_2, over the same nodes $\{X_1, X_2, \ldots, X_n\}$, that differ only in the direction of their edges.





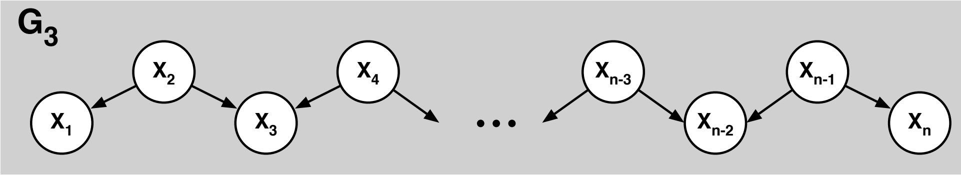
Suppose that we have a (fully observed) data set $\{x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}\}_{t=1}^{T}$ in which each example provides a complete instantiation of the nodes in these DAGs. Let $\textsc{count}_i(x)$ denote the number of examples in which $X_i = x$, and let $\textsc{count}_i(x, x')$ denote the number of examples in which $X_i = x$ and $X_{i+1} = x'$.

(a) Express the maximum likelihood estimates for the CPTs in G_1 in terms of these counts. (*2 pts*)

(b) Express the maximum likelihood estimates for the CPTs in G_2 in terms of these counts. (*2 pts*)

(c) Using your answers from parts (a) and (b), show that the maximum likelihood CPTs for G_1 and G_2 from this data set give rise to the same joint distribution over the nodes $\{X_1, X_2, \ldots, X_n\}$. (*2 pts*)

(d) Suppose that some but not all of the edges in these DAGs were reversed, as in the graph G_3 shown below. Would the maximum likelihood CPTs for G_3 also give rise to the same joint distribution? (*Hint*: does G_3 imply all the same statements of conditional independence as G_1 and G_2? Look in particular at nodes X_3 and X_{n-2}.) (*2 pts*)



4.3 Statistical language modeling (*15 pts*)

In this problem, you will explore some simple statistical models of English text. Download and examine the data files on Canvas for this assignment. (Start with the `hw4_readme.txt` file.) These files contain unigram and bigram counts for 500 frequently occurring tokens in English text. These tokens include actual words as well as punctuation symbols and other textual markers. In addition, an "unknown" token is used to represent all words that occur outside this basic vocabulary. For this problem, as usual, you may program in the language of your choice.

(a) Compute the maximum likelihood estimate of the unigram distribution $P_u(w)$ over words w. Print out a table of all the tokens (i.e., words) that start with the letter "M", along with their numerical unigram *probabilities* (not counts). (You do not need to print out the unigram probabilities for all 500 tokens.) (*1 pt*)

(b) Compute the maximum likelihood estimate of the bigram distribution $P_b(w'|w)$. Print out a table of the ten most likely words to follow the word "THE", along with their numerical bigram probabilities. (*1 pt*)

(c) Consider the sentence **"The stock market fell by one hundred points last week."** Ignoring punctuation, compute and compare the log-likelihoods of this sentence under the unigram and bigram models:

$$\begin{aligned}\mathcal{L}_u &= \log\Big[P_u(\textbf{the})\, P_u(\textbf{stock})\, P_u(\textbf{market})\, \ldots\, P_u(\textbf{points})\, P_u(\textbf{last}) P_u(\textbf{week})\Big]\\ \mathcal{L}_b &= \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\, P_b(\textbf{stock}|\textbf{the})\, P_b(\textbf{market}|\textbf{stock})\, \ldots\, P_b(\textbf{last}|\textbf{points})\, P_b(\textbf{week}|\textbf{last})\,\Big]\end{aligned}$$

In the equation for the bigram log-likelihood, the token $\langle s\rangle$ is used to mark the beginning of a sentence. Which model yields the highest log-likelihood? (*3 pts*)

(d) Consider the sentence **"The sixteen officials sold fire insurance."** Ignoring punctuation, compute and compare the log-likelihoods of this sentence under the unigram and bigram models:

$$\begin{aligned}\mathcal{L}_u &= \log\Big[P_u(\textbf{the})\, P_u(\textbf{sixteen})\, P_u(\textbf{officials})\, \ldots\, P_u(\textbf{sold})\, P_u(\textbf{fire})\, P_u(\textbf{insurance})\Big]\\ \mathcal{L}_b &= \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\, P_b(\textbf{sixteen}|\textbf{the})\, P_b(\textbf{officials}|\textbf{sixteen})\, \ldots\, P_b(\textbf{fire}|\textbf{sold})\, P_b(\textbf{insurance}|\textbf{fire})\,\Big]\end{aligned}$$

Which pairs of adjacent words in this sentence are not observed in the training corpus? What effect does this have on the log-likelihood from the bigram model? (*3 pts*)

(e) Consider the so-called *mixture* model that predicts words from a weighted interpolation of the unigram and bigram models:

$$P_m(w'|w) = \lambda P_u(w') + (1-\lambda) P_b(w'|w),$$

where $\lambda \in [0,1]$ determines how much weight is attached to each prediction. Under this mixture model, the log-likelihood of the sentence from part (d) is given by:

$$\mathcal{L}_m = \log\Big[P_m(\textbf{the}|\langle\textbf{s}\rangle)\, P_m(\textbf{sixteen}|\textbf{the})\, P_m(\textbf{officials}|\textbf{sixteen})\, \ldots\, P_m(\textbf{fire}|\textbf{sold})\, P_m(\textbf{insurance}|\textbf{fire})\,\Big].$$

Compute and plot the value of this log-likelihood $\mathcal{L}_m$ as a function of the parameter $\lambda \in [0,1]$. From your results, deduce the optimal value of λ to two significant digits. (*3 pts*)

(f) Submit a printout of your source code for the previous parts of this problem. (*4 pts*)

4.4 Markov modeling (12 pts)

In this problem, you will construct and compare unigram and bigram models defined over the four-letter alphabet $\mathcal{A} = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}\}$. Consider the following 16-token sequence $\mathcal{S}$:

$$\mathcal{S} = \text{“}\texttt{a a b b b b c c d d a a d d c c}\text{”}$$

(a) **Unigram model** ***(1 pt)***

Let τ_ℓ denote the ℓth token of this sequence, and let $L = 16$ denote the total sequence length. The overall likelihood of this sequence under a unigram model is given by:

$$P_U(\mathcal{S}) = \prod_{\ell=1}^{L} P_1(\tau_\ell),$$

where $P_1(\tau)$ is the unigram probability for the token $\tau \in \mathcal{A}$. Compute the maximum likelihood estimates of these unigram probabilities on the training sequence $\mathcal{S}$. Complete the table with your answers.

τ	a	b	c	d
$P_1(\tau)$				

(b) **Bigram model** ***(2 pts)***

Suppose that the overall likelihood of the sequence $\mathcal{S}$ under a bigram model is computed by:

$$P_B(\mathcal{S}) = P_1(\tau_1) \prod_{\ell=2}^{L} P_2(\tau_\ell | \tau_{\ell-1}),$$

where $P_2(\tau'|\tau)$ is the bigram probability that token $\tau \in \mathcal{A}$ is followed by token $\tau' \in \mathcal{A}$. Compute the maximum likelihood estimates of these bigram probabilities on the training sequence $\mathcal{S}$. Complete the table with your answers.

τ'

	$P_2(\tau'\|\tau)$	a	b	c	d
	a				
τ	b				
	c	0	0	$\frac{2}{3}$	$\frac{1}{3}$
	d				

(c) **Likelihoods *(5 pts)***

Consider again the training sequence $\mathcal{S}$, as well as three test sequences $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ of the same length, shown below. Note that $\mathcal{T}_2$ and $\mathcal{T}_3$ contain bigrams ($\underline{\text{underlined}}$) that are not in the training sequence $\mathcal{S}$.

$$\begin{aligned}
\mathcal{S} &= \text{“}\texttt{aabbbbccddaaddcc}\text{”}\\
\mathcal{T}_1 &= \text{“}\texttt{adadadadadadadad}\text{”}\\
\mathcal{T}_2 &= \text{“}\texttt{aaaaddddccc}\underline{\texttt{cb}}\texttt{bbb}\text{”}\\
\mathcal{T}_3 &= \text{“}\underline{\texttt{bdbdbdbdbdbdbdbd}}\text{”}
\end{aligned}$$

Consider the probabilities of these sequences under the unigram and bigram models from parts (a) and (b) of this problem (i.e., the models that you estimated from the training sequence $\mathcal{S}$). For each of the following, indicate whether the probability on the left is equal ($=$), greater ($>$), or less ($<$) than the probability on the right.

Note: you can (and should) answer these questions without explicitly computing the numerical values of the expressions on the left and right hand sides.

$P_U(\mathcal{S})$ □ $P_U(\mathcal{T}_1)$

$P_U(\mathcal{S})$ □ $P_U(\mathcal{T}_2)$

$P_U(\mathcal{S})$ □ $P_U(\mathcal{T}_3)$

$P_B(\mathcal{T}_1)$ □ $P_B(\mathcal{S})$

$P_B(\mathcal{T}_2)$ □ $P_B(\mathcal{S})$

$P_B(\mathcal{T}_3)$ □ $P_B(\mathcal{T}_2)$

$P_U(\mathcal{S})$ □ $P_B(\mathcal{S})$

$P_U(\mathcal{T}_1)$ □ $P_B(\mathcal{T}_1)$

$P_U(\mathcal{T}_2)$ □ $P_B(\mathcal{T}_2)$

$P_U(\mathcal{T}_3)$ □ $P_B(\mathcal{T}_3)$

(d) **Likelihoods *(4 pts)***

Consider the model obtained by linear interpolation (or mixing) of the unigram and bigram models estimated in part (a) and (b) of this problem:

$$P_M(\tau'|\tau) \;=\; (1-\lambda)P_1(\tau') + \lambda P_2(\tau'|\tau),$$

with mixing coefficient $\lambda \in [0,1]$. For a sequence of tokens of length L, the mixture model computes the log-likelihood as:

$$\mathcal{L} \;=\; \log P_1(\tau_1) + \sum_{\ell=2}^{L} \log P_M(\tau_\ell|\tau_{\ell-1}).$$

Naturally, this value varies as a function of the coefficient λ. For λ near zero, it is close to the log-likelihood of the unigram model; for λ near one, it is close to that of the bigram model. This last part of this problem asks you to consider, for each of the sequences below, the *qualitative* behavior of the mixture model's log-likelihood as a function of $\lambda \in [0,1]$. (For instance, is this function constant, or if not, where do its maximum and minimum occur?)

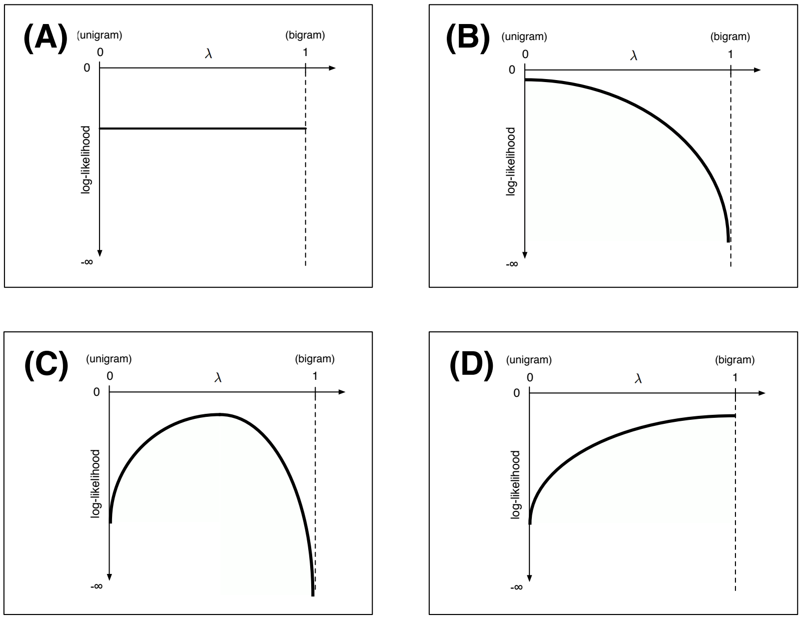
The plots below illustrate four possible behaviors of the mixture model's log-likelihood as a function of $\lambda \in [0,1]$. For each sequence below, indicate the one plot (either A, B, C, or D) that sketches the correct qualitative behavior. (Note that these graphs are not exactly what they would look like if we were to graph the function; we are only asking about the general behavior of the function (e.g. increasing as λ goes to 1.))

$\mathcal{S}$ = “$\texttt{a a b b b b c c d d a a d d c c}$” □

$\mathcal{T}_1$ = “$\texttt{a d a d a d a d a d a d a d a d}$” □

$\mathcal{T}_2$ = “$\texttt{a a a a d d d d c c c }\underline{\texttt{c b}}\texttt{ b b b}$” □

$\mathcal{T}_3$ = “$\underline{\texttt{b d b d b d b d b d b d b d b d}}$” □


(A)
(unigram)
(bigram)
0
λ
1
0
log-likelihood
-∞
(B)
(unigram)
(bigram)
0
λ
1
0
log-likelihood
-∞
(C)
(unigram)
(bigram)
0
λ
1
0
log-likelihood
-∞
(D)
(unigram)
(bigram)
0
λ
1
0
log-likelihood
-∞

4.5 Maximum likelihood estimation of a multinomial distribution (CSE250A ONLY)

A $2D$-sided die is tossed many times, and the results of each toss are recorded as data. Suppose that in the course of the experiment, the d^{th} side of the die is observed C_d times. For this problem, you should assume that the tosses are identically, independent distributed (i.i.d.) according to the probabilities of the die.

(a) **Log-likelihood *(1 pt)***

Let $X \in \{1, 2, 3, \ldots, 2D\}$ denote the outcome of a toss, and let $p_d = P(X = d)$ denote the probabilities of the die. Express the log-likelihood $\mathcal{L} = \log P(\text{data})$ of the observed results in terms of the probabilities p_d and the counts C_d.

(b) **Maximum likelihood estimate *(2 pts)***

Derive the maximum likelihood estimates of the die's probabilities p_d. Specifically, maximize your expression for the log-likelihood $\mathcal{L}$ in part (a) subject to the constraints

$$\sum_{d=1}^{2D} p_d = 1, \qquad p_d \geq 0.$$

You should use a Lagrange multiplier to enforce the linear equality constraint, but it is sufficient to observe that the resulting solution is nonnegative.

(c) **Even versus odd *(2 pts)***

Compute the probability $P(X \in \{2, 4, 6, \ldots, 2D\})$ that the roll of a die is *even* and also the probability $P(X \in \{1, 3, 5, \ldots, 2D-1\})$ that the roll of a die is *odd*. Show that these two probabilities are equal when

$$\sum_{d=1}^{2D} (-1)^d p_d = 0.$$

(d) **Maximum likelihood estimate *(5 pts)***

Suppose it is known a priori that the probability of an *even* toss is equal to that of an *odd* toss. Derive the maximum likelihood estimates of the die's probabilities p_d subject to this *additional* constraint. Specifically, maximize your expression for the log-likelihood $\mathcal{L}$ in part (a) subject to the constraints

$$\sum_{d=1}^{2D} p_d = 1, \qquad \sum_{d=1}^{2D} (-1)^d p_d = 0, \qquad p_d \geq 0.$$

Hint 1: introduce two Lagrange multipliers, one for each linear equality constraint.
Hint 2: it may simplify your final answer to let $C_{even} = \sum_{d=1}^{d} C_{2d}$ and $C_{odd} = \sum_{d=1}^{D} C_{2d-1}$ denote the sums of even and odd counts.