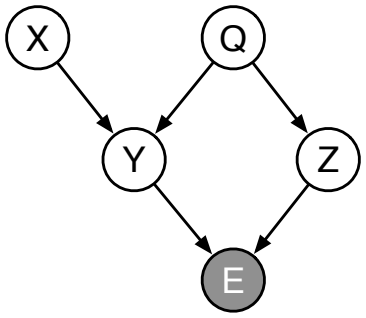
CSE 150A/250A - Homework 4 Solutions (Spring 2025)

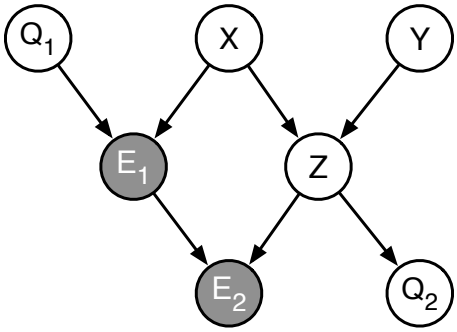
4.1 Likelihood weighting *(5 pts)*

(a) **Single node of evidence (2 pts)**



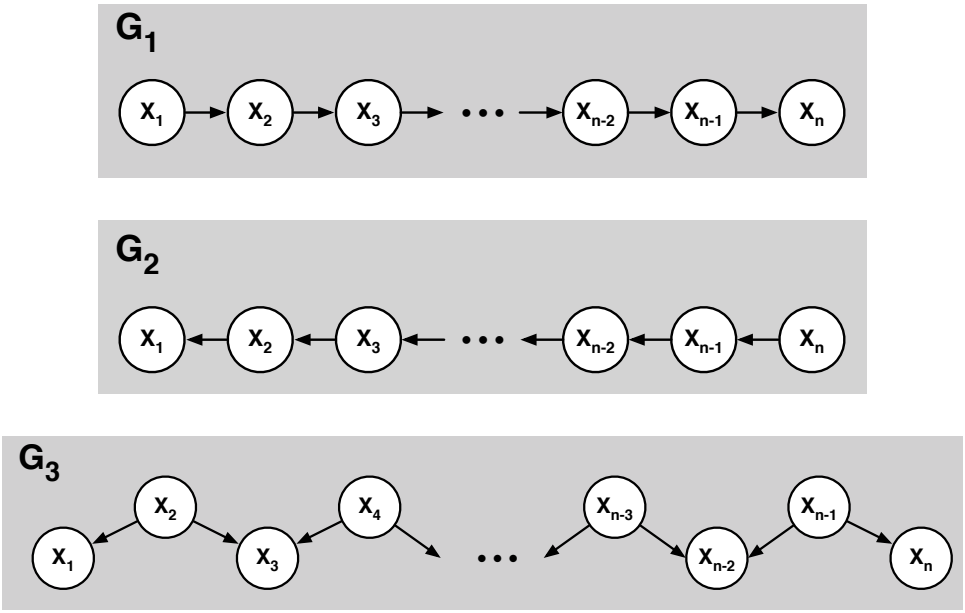
$$\boxed{P(Q = q \mid E = e) \approx \frac{\sum_{t=1}^{T} I(q,q_t)\, P(e|y_t,z_t)}{\sum_{t=1}^{T} P(e|y_t,z_t)}}$$

(b) **Multiple nodes of evidence (3 pts)**



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

4.2 Maximum likelihood estimation in belief networks (*8 pts - 2pts per part*)



(a) Maximum likelihood estimation in DAG G_1:

$$P^{\text{ML}}(X_1 = x) = \frac{\text{COUNT}_1(x)}{T}$$

$$P^{\text{ML}}(X_{i+1} = x' \mid X_i = x) = \frac{\text{COUNT}_i(x, x')}{\text{COUNT}_i(x)}$$

(b) Maximum likelihood estimation in DAG G_2:

$$P^{\text{ML}}(X_n = x) = \frac{\text{COUNT}_n(x)}{T}$$

$$P^{\text{ML}}(X_i = x \mid X_{i+1} = x') = \frac{\text{COUNT}_i(x, x')}{\text{COUNT}_{i+1}(x')}$$

(c) Over the DAG G_1, we estimate:

$$
\begin{aligned}
P_1^{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= P^{\text{ML}}(X_1 = x_1) \prod_{i=1}^{n-1} P^{\text{ML}}(X_{i+1} = x_{i+1} \mid X_i = x_i) \\
&= \frac{\text{COUNT}_1(x_1)}{T} \prod_{i=1}^{n-1} \frac{\text{COUNT}_i(x_i, x_{i+1})}{\text{COUNT}_i(x_i)} \\
&= \frac{1}{T} \left(\prod_{i=1}^{n-1} \text{COUNT}_i(x_i, x_{i+1}) \right) \left(\prod_{i=2}^{n-1} \frac{1}{\text{COUNT}_i(x_i)} \right)
\end{aligned}
$$

Likewise, over the DAG G_2, we estimate:

$$
\begin{aligned}
P_2^{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= P^{\text{ML}}(X_n = x_n) \prod_{i=1}^{n-1} P^{\text{ML}}(X_i = x_i \mid X_{i+1} = x_{i+1}) \\
&= \frac{\text{COUNT}_n(x_n)}{T} \prod_{i=1}^{n-1} \frac{\text{COUNT}_{i+1}(x_i, x_{i+1})}{\text{COUNT}_{i+1}(x_{i+1})} \\
&= \frac{1}{T} \left(\prod_{i=1}^{n-1} \text{COUNT}_i(x_i, x_{i+1}) \right) \left(\prod_{i=2}^{n-1} \frac{1}{\text{COUNT}_i(x_i)} \right) \\
&= P_1^{\text{ML}}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\end{aligned}
$$

(d) The joint distributions are no longer guaranteed to be the same. Note that in G_3 the node X_{n-2} has two parents. Hence $P(X_{n-1} \mid X_{n-2})$ is not necessarily equal to $P(X_{n-1} \mid X_{n-2}, X_{n-3})$ in G_3 (due to the effect of explaining away). On the other hand, $P(X_{n-1} \mid X_{n-2}) = P(X_{n-1} \mid X_{n-2}, X_n = x_n)$ always holds in G_1 and G_2. In short, G_3 does not imply all the same statements of conditional independence as G_1 and G_2.

4.3 Statistical language modeling (*15 pts*)

(a) **Unigram probabilities** ***(1 pt)***

Here are the unigram probabilities of words starting with the letter M.

word w	$P_u(w)$	word w	$P_u(w)$
MILLION	0.002073	MONDAY	0.000382
MORE	0.001709	MAJOR	0.000371
MR.	0.001442	MILITARY	0.000352
MOST	0.000788	MEMBERS	0.000336
MARKET	0.000780	MIGHT	0.000274
MAY	0.000730	MEETING	0.000266
M.	0.000703	MUST	0.000267
MANY	0.000697	ME	0.000264
MADE	0.000560	MARCH	0.000260
MUCH	0.000515	MAN	0.000253
MAKE	0.000514	MS.	0.000239
MONTH	0.000445	MINISTER	0.000240
MONEY	0.000437	MAKING	0.000212
MONTHS	0.000406	MOVE	0.000210
MY	0.000400	MILES	0.000206

(b) **Bigram probabilities** ***(1 pt)***

Here are the *ten* most likely words to follow the word THE.

next word w	$P_b(w \mid \text{THE})$	next word w	$P_b(w \mid \text{THE})$
(UNK)	0.615020	UNITED	0.008672
U.	0.013372	GOVERNMENT	0.006803
FIRST	0.011720	NINETEEN	0.006651
COMPANY	0.011659	SAME	0.006287
NEW	0.009451	TWO	0.006161

(c) **Sentence log-likelihood** ***(3 pts)***

The sentence is [THE STOCK MARKET FELL BY ONE HUNDRED POINTS LAST WEEK]. Let L_u denote the log-likelihood of the unigram model, and L_b the log-likelihood of the bigram model. Then:

$$\begin{aligned} L_u &= -64.51 \quad (1 \text{ pt}) \\ L_b &= -40.92 \quad (2 \text{ pts}) \end{aligned}$$

Since $L_u < L_b$, the bigram model yields the higher log-likelihood.

(d) **Sentence log-likelihood** ***(3 pts)***

The sentence is [THE SIXTEEN OFFICIALS SOLD FIRE INSURANCE]. Let L_u denote the log-likelihood of the unigram model, and L_b the log-likelihood of the bigram model. Then:

$$\begin{aligned} L_u &= -44.291934 \quad (1 \text{ pt}) \\ L_b &= -\infty \quad (2 \text{ pts}) \end{aligned}$$

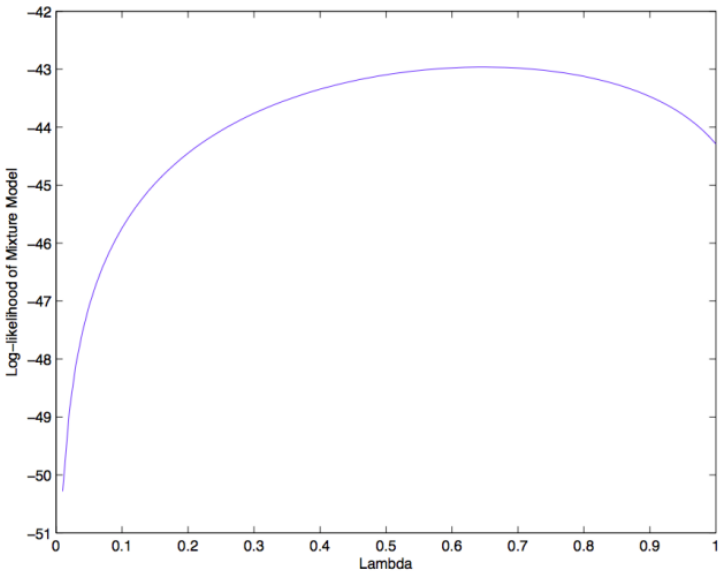
Since $L_u > L_b$, the unigram model yields highest log-likelihood. And we have the following probabilities for each bigram in the sentence:

$$\begin{aligned} P_b(\text{THE} \mid \langle \text{s} \rangle) &= 0.158653 \\ P_b(\text{SIXTEEN} \mid \text{THE}) &= 0.000229 \\ P_b(\text{OFFICIALS} \mid \text{SIXTEEN}) &= 0 \\ P_b(\text{SOLD} \mid \text{OFFICIALS}) &= 0.000092 \\ P_b(\text{FIRE} \mid \text{SOLD}) &= 0 \\ P_b(\text{INSURANCE} \mid \text{FIRE}) &= 0.003052 \end{aligned}$$

The bigrams "SIXTEEN OFFICIALS" and "SOLD FIRE" are not observed in the training set. The missing bigrams cause the sentence to be assigned zero probability under the bigram model.

(e) **Mixture model** ***(3 pts – 2 pts for the graph, 1 pt for the maximum)***

The following plot shows the log-likelihood L_m of the mixture model as a function of the interpolation parameter $\lambda \in [0, 1]$. The optimal value is $\lambda = 0.65$, which yields a log-likelihood of $L_m = -42.96$.



(*) **Source code** ***(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
$P_1(\tau)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

(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.

$P_2(\tau'\|\tau)$	τ' = a	τ' = b	τ' = c	τ' = d
τ = a	$\frac{1}{2}$	$\frac{1}{4}$	0	$\frac{1}{4}$
τ = b	0	$\frac{3}{4}$	$\frac{1}{4}$	0
τ = c	0	0	$\frac{2}{3}$	$\frac{1}{3}$
τ = d	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$

(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 (underlined) that are not in the training sequence $\mathcal{S}$.

$$\begin{aligned}
\mathcal{S} &= \text{“a a b b b b c c d d a a d d c c”} \\
\mathcal{T}_1 &= \text{“a d a d a d a d a d a d a d a d”} \\
\mathcal{T}_2 &= \text{“a a a a d d d d c c c c b b b b”} \\
\mathcal{T}_3 &= \text{“b d b d b d b d b d b d b d b d”}
\end{aligned}$$

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.

$$\begin{array}{lcl}
P_U(\mathcal{S}) & \boxed{=} & P_U(\mathcal{T}_1) \\
P_U(\mathcal{S}) & \boxed{=} & P_U(\mathcal{T}_2) \\
P_U(\mathcal{S}) & \boxed{=} & P_U(\mathcal{T}_3)
\end{array}$$

$$\begin{array}{lcl}
P_B(\mathcal{T}_1) & \boxed{<} & P_B(\mathcal{S}) \\
P_B(\mathcal{T}_2) & \boxed{<} & P_B(\mathcal{S}) \\
P_B(\mathcal{T}_3) & \boxed{=} & P_B(\mathcal{T}_2)
\end{array}$$

$$\begin{array}{lcl}
P_U(\mathcal{S}) & \boxed{<} & P_B(\mathcal{S}) \\
P_U(\mathcal{T}_1) & \boxed{=} & P_B(\mathcal{T}_1) \\
P_U(\mathcal{T}_2) & \boxed{>} & P_B(\mathcal{T}_2) \\
P_U(\mathcal{T}_3) & \boxed{>} & P_B(\mathcal{T}_3)
\end{array}$$

(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?)

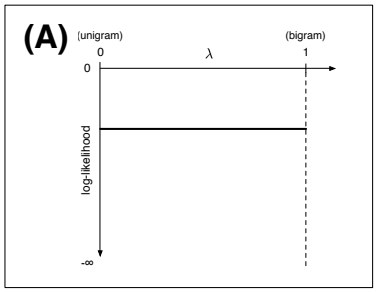
The plots below illustrate four possible behaviors of the mixture model's log-likelihood as a function of $\lambda \in [0,1]$. For each sequence below, indicate the one plot (either A, B, C, or D) that sketches the correct qualitative behavior. (Note that these graphs are not exactly what they would 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.))

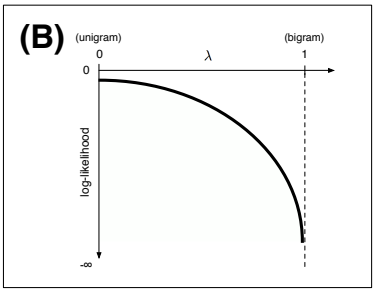
$\mathcal{S}$ = "a a b b b b c c d d a a d d c c" D

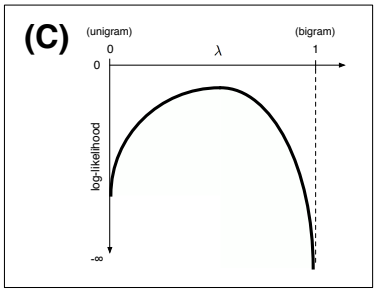
$\mathcal{T}_1$ = "a d a d a d a d a d a d a d a d" A

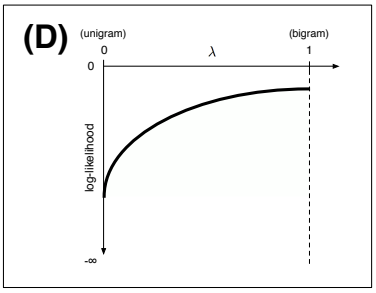
$\mathcal{T}_2$ = "a a a a d d d d c c c c b b b b" C

$\mathcal{T}_3$ = "b d b d b d b d b d b d b d b d" B


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

$$\mathcal{L} = \log P(\text{data}) = \log \prod_{d=1}^{2D} P(X = d)^{C_d} = \sum_{d=1}^{2D} C_d \log p_d$$

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

(0.5 pts) Introduce a Lagrange multiplier for the constraint:

$$F(p, \mu) = \sum_{d=1}^{2D} C_d \log p_d + \mu \left(1 - \sum_{d=1}^{2D} p_d\right).$$

(0.5 pts) Enforcing $\partial F / \partial p_d = 0$ gives:

$$C_d / p_d = \mu.$$

(0.5 pts) Multiplying both sides by p_d, summing over n, and enforcing the constraint gives:

$$\sum_{d=1}^{2D} C_d = \mu.$$

(0.5 pts) Finally, substituting this value for μ gives:

$$p_d = \frac{C_d}{\sum_{k=1}^{2D} C_k}.$$

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

Probabilities add for the union of mutually exclusive events. Hence:

$$P(X \text{ is even}) = p_2 + p_4 + \cdots + p_{2D-2} + p_{2D},$$
$$P(X \text{ is odd}) = p_1 + p_3 + \cdots + p_{2D-3} + p_{2D-1}.$$

Suppose that these probabilities are equal. Then subtracting the bottom equation from the top equation gives the desired constraint:

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

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

(1 pt) Introduce Lagrange multipliers for the constraints:

$$G(p, \mu, \nu) = \sum_{d=1}^{2D} C_d \log p_d + \mu \left(1 - \sum_{d=1}^{2D} p_d\right) + \nu \left(\sum_{d=1}^{2D} (-1)^d p_d\right).$$

(1 pt) Enforcing $\partial G / \partial p_d = 0$ gives:

$$\frac{C_d}{p_d} = \mu + \nu(-1)^{d+1}.$$

(1 pt) Multiplying both sides by p_d, then summing over n and enforcing the constraints gives:

$$\sum_{d=1}^{2D} C_d = 1 \cdot \mu - 0 \cdot \nu = \mu.$$

(1 pt) Similarly, multiplying by $(-1)^d p_d$ and summing over d gives:

$$\sum_{d=1}^{2D} (-1)^d C_d = 0 \cdot \mu - 1 \cdot \nu = -\nu.$$

(1 pt) Substituting these solutions for μ and ν gives:

$$p_d = \frac{C_d}{\sum_{k=1}^{2D} \left[C_k + (-1)^{k+d} C_k\right]}.$$

After some simplification, this gives the more intuitive result:

$$p_{2k-1} = \frac{1}{2}\left(\frac{C_{2k-1}}{\sum_{d=1}^{D} C_{2d-1}}\right) \qquad p_{2k} = \frac{1}{2}\left(\frac{C_{2k}}{\sum_{d=1}^{D} C_{2d}}\right)$$

for the odd and even faces of the die, respectively. Note that $P(X \text{ is odd}) = P(X \text{ is even}) = \frac{1}{2}$ as required.