Homework 2 — Answer Sheet and Marking Scheme

1.1 Markov equivalence

Consider a joint distribution that can be factorized with respect to the $A \leftarrow B \rightarrow C$ graph as $P(A \mid B)P(C \mid B)P(B)$. The chain rule tells us that $P(A \mid B)P(B) = P(B \mid A)P(A)$, so we can re-write the joint distribution as $P(C \mid B)P(B \mid A)P(A)$. But this is the factorization of the $A \rightarrow B \rightarrow C$ graph. So any distribution that factorizes according to the first graph factorizes according to the second graph, and vice-versa. Therefore the two graphs are Markov-equivalent. ■

1.2 Markov non-equivalence

Consider a joint distribution $P(A, B)$ where $P(A \mid B) \neq P(A)$. This distribution can be factorized as $P(A \mid B)P(B)$, which is consistent with the first DAG. But it cannot be factorized as $P(A)P(B)$, so it is not consistent with the second DAG. Since there is at least one distribution that can be factorized with respect to one DAG but not the other, the two DAGs are not Markov-equivalent. ■

2.1 Prior probabilities

The grammar is

$$H \rightarrow A \quad (0.6), \qquad H \rightarrow (H \wedge H) \quad (0.25), \qquad H \rightarrow (H \vee H) \quad (0.15),$$

and

$$A \rightarrow \texttt{red}, \texttt{blue}, \texttt{round}, \texttt{square}, \texttt{soft}, \texttt{hard} \quad \text{each with probability } \frac{1}{6}.$$

Hypothesis $h_1 = \texttt{red}$. Production rules used:

$$H \rightarrow A, \qquad A \rightarrow \texttt{red}.$$

Therefore

$$P(h_1) = 0.6 \cdot \frac{1}{6} = 0.1.$$

Hypothesis $h_2 = (\texttt{red} \wedge \texttt{soft})$. Production rules used:

$$H \rightarrow (H \wedge H),$$

then for the left child:

$$H \rightarrow A, \qquad A \rightarrow \texttt{red},$$

and for the right child:

$$H \rightarrow A, \qquad A \rightarrow \texttt{soft}.$$

Therefore

$$P(h_2) = 0.25 \cdot \left(0.6 \cdot \frac{1}{6}\right) \cdot \left(0.6 \cdot \frac{1}{6}\right)$$
$$= 0.25 \cdot 0.1 \cdot 0.1 = 0.0025.$$

Hypothesis $h_3 = (\texttt{red} \vee \texttt{round})$**.** Production rules used:

$$H \to (H \vee H),$$

then for the left child:

$$H \to A, \qquad A \to \texttt{red},$$

and for the right child:

$$H \to A, \qquad A \to \texttt{round}.$$

Therefore

$$P(h_3) = 0.15 \cdot \left(0.6 \cdot \frac{1}{6}\right) \cdot \left(0.6 \cdot \frac{1}{6}\right)$$
$$= 0.15 \cdot 0.1 \cdot 0.1 = 0.0015.$$

Hypothesis $h_4 = ((\texttt{red} \wedge \texttt{soft}) \vee \texttt{round})$**.** Production rules used:

$$H \to (H \vee H).$$

The left child is same as h_2.
For the right child:

$$H \to A, \quad A \to \texttt{round}.$$

Therefore

$$P(h_4) = 0.15 \cdot P(h_2) \cdot \left(0.6 \cdot \frac{1}{6}\right)$$
$$= 0.15 \cdot 0.0025 \cdot 0.1$$
$$= 0.0000375.$$

2.2 Comparison and explanation

The probabilities are ordered as

$$P(h_1) > P(h_2) > P(h_3) > P(h_4).$$

Numerically,

$$0.1 > 0.0025 > 0.0015 > 0.0000375.$$

The atomic hypothesis $h_1 = \texttt{red}$ has the largest prior probability because it uses only one atomic predicate and no binary connectives.

The hypotheses h_2 and h_3 both contain one binary connective and two atomic predicates. They are less probable than h_1 because their derivations require more production rules, each of which contributes a probability factor less than 1.

Between h_2 and h_3, we have

$$P(h_2) > P(h_3),$$

because the grammar assigns

$$P(H \to (H \wedge H)) = 0.25$$

but

$$P(H \to (H \vee H)) = 0.15.$$

Thus conjunctions are more probable than disjunctions under this PCFG.

The hypothesis h_4 has the smallest prior probability because it has two binary connectives and three atomic predicates. Its derivation is longer and multiplies together more probability factors less than 1.

2.3 Likelihoods

Recall that $P(D \mid H) = P(o_1 \mid H)P(o_2 \mid H)P(o_3 \mid H)P(o_4 \mid H)$, where each factor is 0.9 if H makes a correct prediction and 0.1 if it makes an incorrect prediction.

Hypothesis $h_1 = \texttt{red}$.

Observation	Satisfies h_1?	Prediction	Correct?
o_1 (red, round, soft, yes)	yes	blicket	✓
o_2 (red, square, soft, yes)	yes	blicket	✓
o_3 (blue, round, hard, no)	no	not blicket	✓
o_4 (blue, square, soft, no)	no	not blicket	✓

All four predictions are correct, so

$$P(D \mid h_1) = 0.9^4 \approx 0.6561.$$

Hypothesis $h_2 = (\texttt{red} \wedge \texttt{soft})$.

Observation	Satisfies h_2?	Prediction	Correct?
o_1 (red, round, soft, yes)	yes	blicket	✓
o_2 (red, square, soft, yes)	yes	blicket	✓
o_3 (blue, round, hard, no)	no	not blicket	✓
o_4 (blue, square, soft, no)	no	not blicket	✓

All four predictions are correct, so

$$P(D \mid h_2) = 0.9^4 \approx 0.6561.$$

Hypothesis $h_3 = (\texttt{red} \vee \texttt{round})$.

Observation	Satisfies h_3?	Prediction	Correct?
o_1 (red, round, soft, yes)	yes	blicket	✓
o_2 (red, square, soft, yes)	yes	blicket	✓
o_3 (blue, round, hard, no)	yes	blicket	×
o_4 (blue, square, soft, no)	no	not blicket	✓

o_3 is blue and round, so h_3 predicts it is a blicket — but it is not. One prediction is incorrect, so

$$P(D \mid h_3) = 0.9^3 \times 0.1 = 0.0729.$$

Hypothesis $h_4 = ((\texttt{red} \wedge \texttt{soft}) \vee \texttt{round})$**.**

Observation	Satisfies h_4?	Prediction	Correct?
o_1 (red, round, soft, yes)	yes	blicket	✓
o_2 (red, square, soft, yes)	yes	blicket	✓
o_3 (blue, round, hard, no)	yes	blicket	×
o_4 (blue, square, soft, no)	no	not blicket	✓

o_3 is round, so the disjunct `round` is satisfied and h_4 predicts it is a blicket — but it is not. One prediction is incorrect, so

$$P(D \mid h_4) = 0.9^3 \times 0.1 = 0.0729.$$

2.4 Posterior probabilities

By Bayes' rule,

$$P(h_i \mid D) = \frac{P(D \mid h_i)\, P(h_i)}{P(D)}.$$

Since $P(D)$ is the same for all hypotheses, we can rank the posteriors by comparing the unnormalized quantities $P(D \mid h_i)\, P(h_i)$:

Hypothesis	$P(D \mid h_i)$	$P(h_i)$	$P(D \mid h_i)\, P(h_i)$
h_1	$0.9^4 \approx 0.6561$	0.1	≈ 0.06561
h_2	$0.9^4 \approx 0.6561$	0.0025	≈ 0.00164
h_3	$0.9^3 \times 0.1 \approx 0.0729$	0.0015	≈ 0.000109
h_4	$0.9^3 \times 0.1 \approx 0.0729$	0.0000375	≈ 0.00000274

Therefore the posterior ranking is

$$P(h_1 \mid D) > P(h_2 \mid D) > P(h_3 \mid D) > P(h_4 \mid D).$$

Parsimony bias

3.1 Show that $n_A(h) = n_B(h) + 1$

Each complete hypothesis generated by the grammar can be represented as a derivation tree.

Atomic predicates are leaves of the tree. Binary connectives, such as $\wedge$ and $\vee$, are internal nodes with exactly two children.

Thus every complete hypothesis corresponds to a full binary tree: every internal node has exactly two children.

For any full binary tree, the number of leaves equals the number of internal nodes plus one. Here, the leaves are atomic predicates and the internal nodes are binary connectives. Therefore,

$$n_A(h) = n_B(h) + 1.$$

Alternatively, this can be shown by induction.

Base case. If the hypothesis is atomic, for example

$$h = \texttt{red},$$

then

$$n_A(h) = 1$$

and

$$n_B(h) = 0.$$

So

$$n_A(h) = n_B(h) + 1.$$

Inductive step. Suppose two hypotheses h_L and h_R satisfy

$$n_A(h_L) = n_B(h_L) + 1$$

and

$$n_A(h_R) = n_B(h_R) + 1.$$

Now form

$$h = (h_L \wedge h_R)$$

or

$$h = (h_L \vee h_R).$$

Then

$$n_A(h) = n_A(h_L) + n_A(h_R),$$

while

$$n_B(h) = n_B(h_L) + n_B(h_R) + 1.$$

Using the induction hypothesis,

$$n_A(h) = \big(n_B(h_L) + 1\big) + \big(n_B(h_R) + 1\big).$$

Therefore

$$n_A(h) = n_B(h_L) + n_B(h_R) + 2.$$

But

$$n_B(h) + 1 = \big(n_B(h_L) + n_B(h_R) + 1\big) + 1 = n_B(h_L) + n_B(h_R) + 2.$$

Hence

$$n_A(h) = n_B(h) + 1.$$

3.2 Why PCFG creates a parsimony bias

A PCFG creates a parsimony bias because shorter and simpler hypotheses require fewer production rules. Since each production rule has probability less than 1, multiplying more rule probabilities tends to make the prior probability smaller.

Therefore, hypotheses with fewer connectives and fewer atomic predicates usually receive higher prior probability than longer, more complex hypotheses.

This implements a version of Occam's razor: simpler explanations are preferred before looking at the data. More complex explanations are not impossible, but they are penalized by the prior.

(In)Tractability

4.1 Three valid hypotheses with exactly one binary connective

Examples include:

$$(\texttt{red} \wedge \texttt{soft}),$$

$$(\texttt{blue} \vee \texttt{round}),$$

$$(\texttt{square} \wedge \texttt{hard}).$$

Each has two atomic predicates and exactly one binary connective.

4.2 Three valid hypotheses with exactly two binary connectives

Examples include:

$$((\texttt{red} \wedge \texttt{soft}) \vee \texttt{round}),$$

$$(\texttt{blue} \wedge (\texttt{square} \vee \texttt{hard})),$$

$$((\texttt{red} \vee \texttt{round}) \wedge \texttt{soft}).$$

Each has three atomic predicates and exactly two binary connectives.

4.3 Why the number of hypotheses grows quickly

The number of hypotheses grows quickly because at each binary connective there are multiple choices.

First, one must choose whether the connective is

$$\wedge$$

or

$$\vee.$$

Second, each leaf can be any of the six atomic predicates:

$$\texttt{red}, \texttt{blue}, \texttt{round}, \texttt{square}, \texttt{soft}, \texttt{hard}.$$

Third, as the number of connectives increases, there are also more possible tree shapes and parenthesizations. For example,

$$((a \wedge b) \vee c)$$

and

$$(a \wedge (b \vee c))$$

have different syntactic structures.

Thus increasing the number of connectives creates many more choices of tree shape, connective labels, and atomic predicates. This leads to rapid combinatorial growth.

4.4 Why computing the posterior normalizer exactly is difficult

The posterior normalizer is

$$P(D) = \sum_{h \in \mathcal{H}} P(D \mid h) P(h).$$

Computing this exactly is difficult because the hypothesis space $\mathcal{H}$ is infinite. The grammar is recursive, so it can generate hypotheses with arbitrarily many connectives.

Even if we tried to enumerate hypotheses in order of size, the number of hypotheses at each size grows very quickly. Therefore, the sum contains infinitely many terms, and each term may require evaluating how well the corresponding hypothesis explains the data.

As a result, exact computation is usually intractable, and approximate inference methods are needed.

4.5 Why we cannot always ignore all long hypotheses

Although each individual long hypothesis may have small prior probability, there may be very many long hypotheses. Their total probability mass can still be important.

Also, some long hypotheses may fit the data much better than short hypotheses. In Bayesian inference, the posterior depends on both the prior and the likelihood:

$$P(h \mid D) \propto P(D \mid h) P(h).$$

So a long hypothesis with low prior probability may still receive significant posterior probability if it explains the data extremely well.

Therefore, we cannot always ignore all long hypotheses simply because they are individually unlikely under the prior.

Marking Scheme

Question	Criteria	Marks
1.1	At marker's discretion	0.5
1.2	At marker's discretion	0.5
2.1	Correctly computes $P(h_1)$ and shows rules $H \to A$, $A \to \texttt{red}$.	0.5
2.1	Correctly computes $P(h_2)$ and shows all required productions.	0.5
2.1	Correctly computes $P(h_3)$ and shows all required productions.	0.5
2.1	Correctly computes $P(h_4)$ and shows all required productions.	0.5
2.2	Correct comparison of the four probabilities.	0.5
2.2	Clear explanation that additional connectives/atoms require additional probability factors less than 1, and that $\wedge$ is more probable than $\vee$.	0.5
2.3	At marker's discretion	0.25
2.4	At marker's discretion	0.25
3.1	Correct argument that the derivation tree is a full binary tree.	0.5
3.1	Correctly concludes or proves that $n_A(h) = n_B(h) + 1$.	0.5
3.2	Explains that each connective and atom contributes probability factors less than 1.	0.5
3.2	Connects increasing n_B to increasing n_A and therefore more factors in the derivation probability.	0.5
4.1	Gives three valid hypotheses with exactly one binary connective.	0.75
4.2	Gives three valid hypotheses with exactly two binary connectives.	0.75
4.3	Explains rapid growth due to choices of atoms, connectives, and tree shapes/parenthesizations.	0.75
4.4	Explains that exact computation of $P(D)$ is difficult because $\mathcal{H}$ is infinite and grows combinatorially.	0.75
4.5	Explains why long hypotheses cannot always be ignored: large total mass and/or high likelihood can matter.	0.5
	Total	10