Question 1

We would like to ensure that for all t, u, v, $\sum_w q_{BO}(w|t,u,v) = 1$. Note that the "missing" probability mass is

$$1 - \sum_{w \in \mathcal{A}(t,u,v)} q_{BO}(w|t,u,v) = 1 - \sum_{w \in \mathcal{A}(t,u,v)} \frac{c^*(t,u,v,w)}{c(t,u,v)}$$

If we set $\alpha(t,u,v) = 1 - \sum_{w \in \mathcal{A}(t,u,v)} \frac{c^*(t,u,v,w)}{c(t,u,v)}$ we can verify that $\sum_w q_{BO}(w|t,u,v) = 1$: for any t, u, v

$$\begin{aligned}
&\sum_w q_{BO}(w|t,u,v) \\
&= \sum_{w \in \mathcal{A}(t,u,v)} q_{BO}(w|t,u,v) + \sum_{w \in \mathcal{B}(t,u,v)} q_{BO}(w|t,u,v) \\
&= \sum_{w \in \mathcal{A}(t,u,v)} \frac{c^*(t,u,v,w)}{c(t,u,v)} + \sum_{w \in \mathcal{B}(t,u,v)} \frac{\alpha(t,u,v) \times q_{BO}(w|u,v)}{\sum_{w \in \mathcal{B}(t,u,v)} q_{BO}(w|u,v)} \\
&= \sum_{w \in \mathcal{A}(t,u,v)} \frac{c^*(t,u,v,w)}{c(t,u,v)} + \alpha(t,u,v) = 1
\end{aligned}$$

Questions 2a, 2b

Maximum value of perplexity: if for any sentence $x^{(i)}$, we have $p(x^{(i)}) = 0$, then $l = -\infty$, and $2^{-l} = \infty$. Thus the maximum possible value is ∞.

Minimum value: if for all sentences $x^{(i)}$ we have $p(x^{(i)}) = 1$, then $l = 0$, and $2^{-l} = 1$. Thus the minimum possible value is 1.

Question 2c

An example that gives the maximum possible value for perplexity:
Training corpus conists of the single sentence

the a STOP

Test corpus consists of the single sentence

a the STOP

It can be verified that a bigram language model as described in the question, trained on the single sentence *the a STOP*, gives probablity 0 to the sentence *a the STOP*, and hence has infinite perplexity on this test corpus.

Question 2d

An example that gives the maximum possible value for perplexity:
Training corpus conists of the single sentence

the a STOP

Test corpus consists of the single sentence

the a STOP

It can be verified that a bigram language model as described in the question, trained on the single sentence *the a STOP*, gives probablity 1 to the sentence *the a STOP*, and hence has perplexity equal to one on this test corpus.

Question 3a

Rearranging terms slightly, we have

$$\begin{aligned} q(w|u,v) &= \alpha \times q_{ML}(w|u,v) \\ &+(1-\alpha) \times \beta \times q_{ML}(w|v) \\ &+(1-\alpha) \times (1-\beta) \times q_{ML}(w) \end{aligned}$$

Hence we have $\lambda_1 = \alpha = 0.5$, $\lambda_2 = (1-\alpha) \times \beta = 0.25$, and $\lambda_3 = (1-\alpha) \times (1-\beta) = 0.25$.

Question 3b

We have an interpolated model with $\lambda_1(u,v) = \alpha(u,v)$,
$\lambda_2(u,v) = (1-\alpha(u,v)) \times \beta(u)$, and
$\lambda_3(u,v) = (1-\alpha(u,v)) \times (1-\beta(u))$.
Define $\mathcal{V}' = \mathcal{V} \cup \{\text{STOP}\}$.

$$\sum_{w\in\mathcal{V}'} q(w \mid u,v)$$

$$= \sum_{w\in\mathcal{V}'} [\lambda_1(u,v) \times q_{ML}(w \mid u,v) + \lambda_2(u,v) \times q_{ML}(w \mid v)$$

$$+ \lambda_3(u,v) \times q_{ML}(w)]$$

$$= \lambda_1(u,v) \sum_w q_{ML}(w \mid u,v) + \lambda_2(u,v) \sum_w q_{ML}(w \mid v) + \lambda_3(u,v) \sum_w q_{ML}(w)$$

$$= \lambda_1(u,v) + \lambda_2(u,v) + \lambda_3(u,v)$$

$$= \alpha(u,v) + (1-\alpha(u,v)) \times \beta(u) + (1-\alpha(u,v)) \times (1-\beta(u))$$

$$= 1$$

Question 3c

As Count(u, v) increases, $\alpha(u, v)$ gets closer to 1, reflecting the intuition that as Count(u, v) increases, the estimate $q_{ML}(w|u, v)$ becomes more reliable, and more weight should be put on it.

A similar argument applies to $\beta(v)$ and Count(v).

The constants C_1 and C_2 dictate how quickly $\alpha(u, v)$ and $\beta(v)$ approach 1 respectively. They can be set by optimization of the perplexity on a held-out corpus.

Question 3d

Under the assumptions of the question
$q_{ML}(w) = \text{Count}(w)/N > 0$.

We have

$$\begin{aligned} q(w|u,v) &= \alpha(u,v) \times q_{ML}(w|u,v) \\ &\quad +(1-\alpha(u,v)) \times \beta(u) \times q_{ML}(w|v) \\ &\quad +(1-\alpha(u,v)) \times (1-\beta(u)) \times q_{ML}(w) \end{aligned}$$

Hence

$$q(w|u,v) \geq (1-\alpha(u,v)) \times (1-\beta(u)) \times q_{ML}(w)$$

It can be verified that $1-\alpha(u,v) > 0$, $1-\beta(u) > 0$, and $q_{ML}(w) > 0$. Hence for all u, v, w, $q_{ML}(w|u,v) > 0$. It follows that for any sentence in the test data $x^{(i)}$, $p(x^{(i)}) > 0$. It follows that the perplexity on the test data cannot be infinite.

Question 4

First consider the statement *"for all bigrams v, w, we have $q_{BO}(w|v) \geq 0$"*. For any v, w such that $\text{Count}(v, w) = 1$, we have

$$w \in \mathcal{A}(v)$$

and in addition

$$\text{Count}^*(v, w) = 1 - 1.5 = -0.5$$

It follows that

$$q_{BO}(w|v) = \frac{-0.5}{\text{Count}(v)} < 0$$

So the statement is **false**.

Question 4 (continued)

Now consider the second statement, *for all unigrams v we have $\sum_w q_{BO}(w|v) = 1$.*
We have for all u, v,

$$\sum_w q_{BO}(w|v) = \sum_{w \in \mathcal{A}(v)} q_{BO}(w|v) + \sum_{w \in \mathcal{B}(v)} q_{BO}(w|v)$$

$$= \sum_{w \in \mathcal{A}(v)} \frac{\text{Count}^*(v, w)}{\text{Count}(v)} + \sum_{w \in \mathcal{B}(v)} \frac{\alpha(v) \times q_{ML}(w)}{\sum_w q_{ML}(w)}$$

$$= \sum_{w \in \mathcal{A}(v)} \frac{\text{Count}^*(v, w)}{\text{Count}(v)} + \alpha(v)$$

$$= \sum_{w \in \mathcal{A}(v)} \frac{\text{Count}^*(v, w)}{\text{Count}(v)} + 1 - \sum_{w \in \mathcal{A}(v)} \frac{\text{Count}^*(v, w)}{\text{Count}(v)} +$$

$$= 1$$

Note that this holds even though some values for Count* may be negative. Hence the statement is **true**.