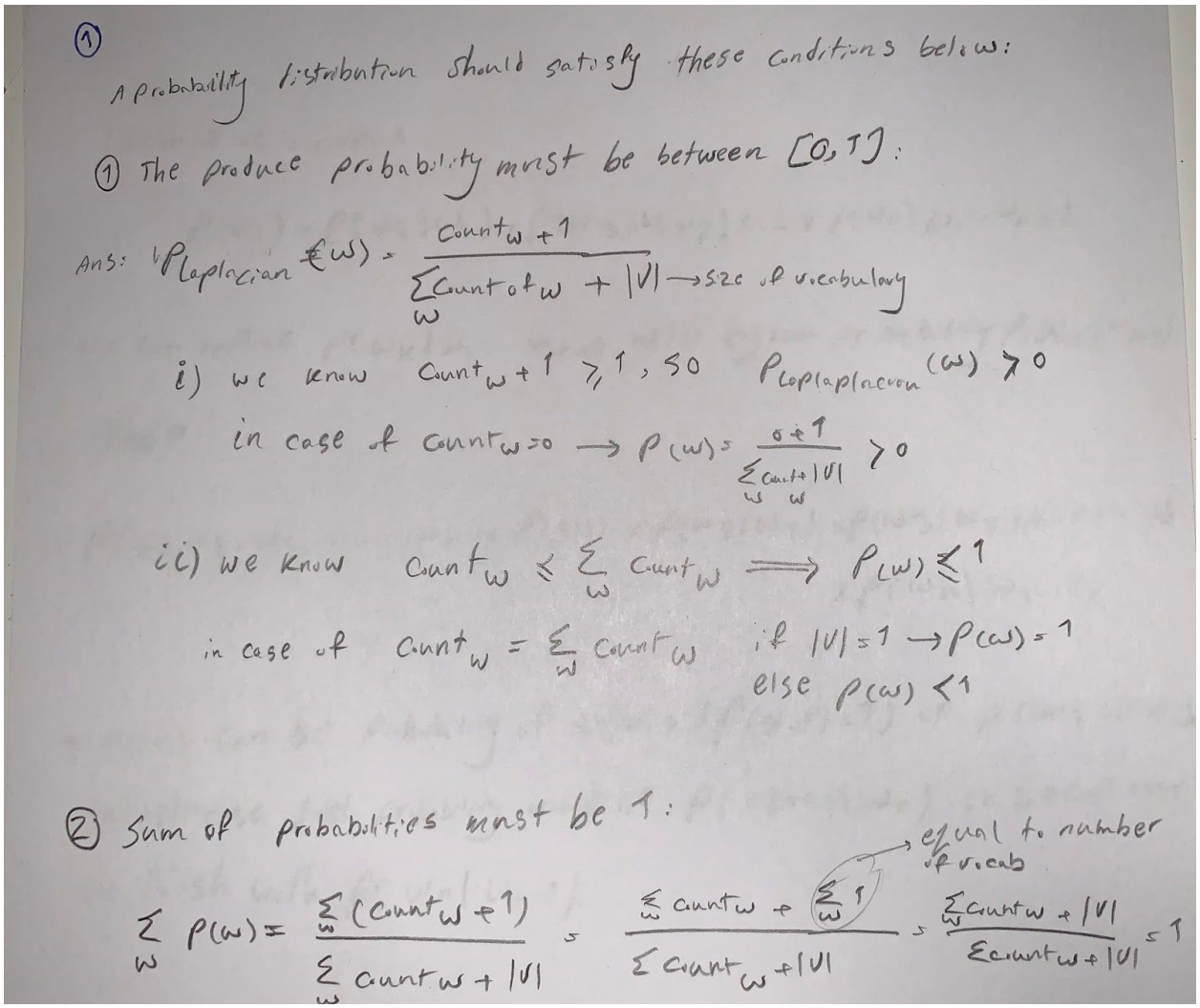

① A probability distribution should satisfy these conditions below:

① The produce probability must be between [0, 1]:

Ans: $P_{Laplacian}(w) = \frac{Count_w + 1}{\sum_w Count\ of\ w + |V|}$ → size of vocabulary

i) we know $Count_w + 1 \geq 1$, so $P_{Laplacian}(w) > 0$

in case of $Count_w = 0 \rightarrow P(w) = \frac{0+1}{\sum_w Count_w + |V|} > 0$

ii) we know $Count_w \leq \sum_w Count_w \Longrightarrow P(w) \leq 1$

in case of $Count_w = \sum_w Count_w$ if $|V| = 1 \rightarrow P(w) = 1$
else $P(w) < 1$

② Sum of probabilities must be 1:

$$\sum_w P(w) = \frac{\sum_w (Count_w + 1)}{\sum_w Count_w + |V|} = \frac{\sum_w Count_w + \sum_w 1}{\sum Count_w + |V|} = \frac{\sum_w Count_w + |V|}{\sum Count_w + |V|} = 1$$

(equal to number of vocab)

(2)

likelihood of a corpus :

$$P(w_1) \times P(w_2|w_1) \times P(w_3|w_1 w_2) \times \cdots \times P(w_n|w_1 \cdots w_{n-1})$$

we can replace $P(w_k|w_1 \cdots w_{k-1})$ with bigram probability $P(w_k|w_{k-1})$

then:

$$L(w_1, w_2, w_3, \cdots, w_n|\theta) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times \cdots \times P(w_n|w_{n-1})$$

$\theta = \{P(w_i), P(w_i|w_j) : w \in V, i : 1 \cdots n, j = 1 \cdots n, i \neq j\}$

* $P(w_1)$ can be probability of <start> ($P(start) = 1$) or $P(w_1|<start>)$. in the first case last probability would be $P(<Eos>|w_n)$. In second case we finish with $P(w_n|w_{n-1})$.

** $P(w_i|w_{i-1}) = \frac{Count_{w_i, w_{i-1}} + 1}{Count_{w_{i-1}} + |V|}$

Amin Hosseiny Marani

---

(3)

$$f(x,y) = x^2 - y^2$$

$$\nabla_{(x,y)} f(x,y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ -2y \end{bmatrix}$$

(4) $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\eta = 0.1$

$$x_{new} = x_{old} - \eta \nabla_x f(x,y) = 2 - 0.1(2x)\Big|_{x=2} = 2 - 0.1 \times 4 = 1.6$$

$$y_{new} = y_{old} - \eta \nabla_y f(x,y) = 1 - 0.1 \times (-2y)\Big|_{y=1} = 1 + 0.2 = 1.2$$

$$f_{old}(x,y) = x^2 - y^2 \Big|_{\substack{x=2 \\ y=1}} = 4 - 1 = 3$$

$$f_{new}(x,y) = x^2 - y^2 \Big|_{\substack{x=1.6 \\ y=0.8}} = (1.6)^2 - (1.2)^2 = 2.56 - 1.44 = 1.12$$

$f_{old} > f_{new}$ ⟶ we know that $\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} 0 \\ \max y \end{bmatrix}$ and gradient descent will merge to those values eventually by decreasing x, and $f(x,y)$ and also increasing y