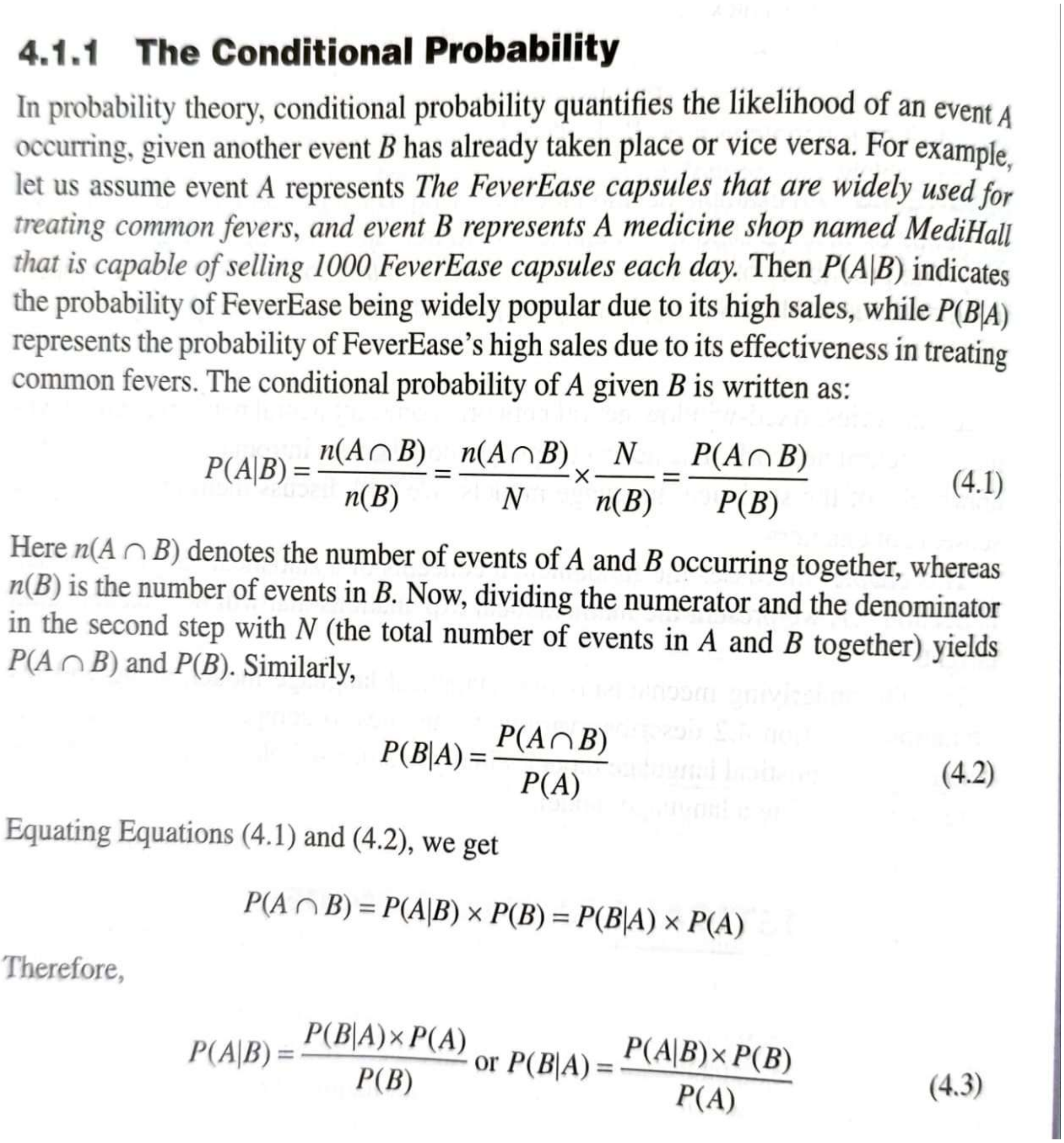
Statistical Language Model

A Statistical Language Model estimates the probability of a sequence of words occurring in a language.
It helps a system understand how likely a sentence is according to learned language patterns.



4.1.1 The Conditional Probability

In probability theory, conditional probability quantifies the likelihood of an event A occurring, given another event B has already taken place or vice versa. For example, let us assume event A represents *The FeverEase capsules that are widely used for treating common fevers, and event B represents A medicine shop named MediHall that is capable of selling 1000 FeverEase capsules each day.* Then $P(A|B)$ indicates the probability of FeverEase being widely popular due to its high sales, while $P(B|A)$ represents the probability of FeverEase's high sales due to its effectiveness in treating common fevers. The conditional probability of A given B is written as:

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B)}{N} \times \frac{N}{n(B)} = \frac{P(A \cap B)}{P(B)} \tag{4.1}$$

Here $n(A \cap B)$ denotes the number of events of A and B occurring together, whereas $n(B)$ is the number of events in B. Now, dividing the numerator and the denominator in the second step with N (the total number of events in A and B together) yields $P(A \cap B)$ and $P(B)$. Similarly,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4.2}$$

Equating Equations (4.1) and (4.2), we get

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

Therefore,

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \text{ or } P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \tag{4.3}$$

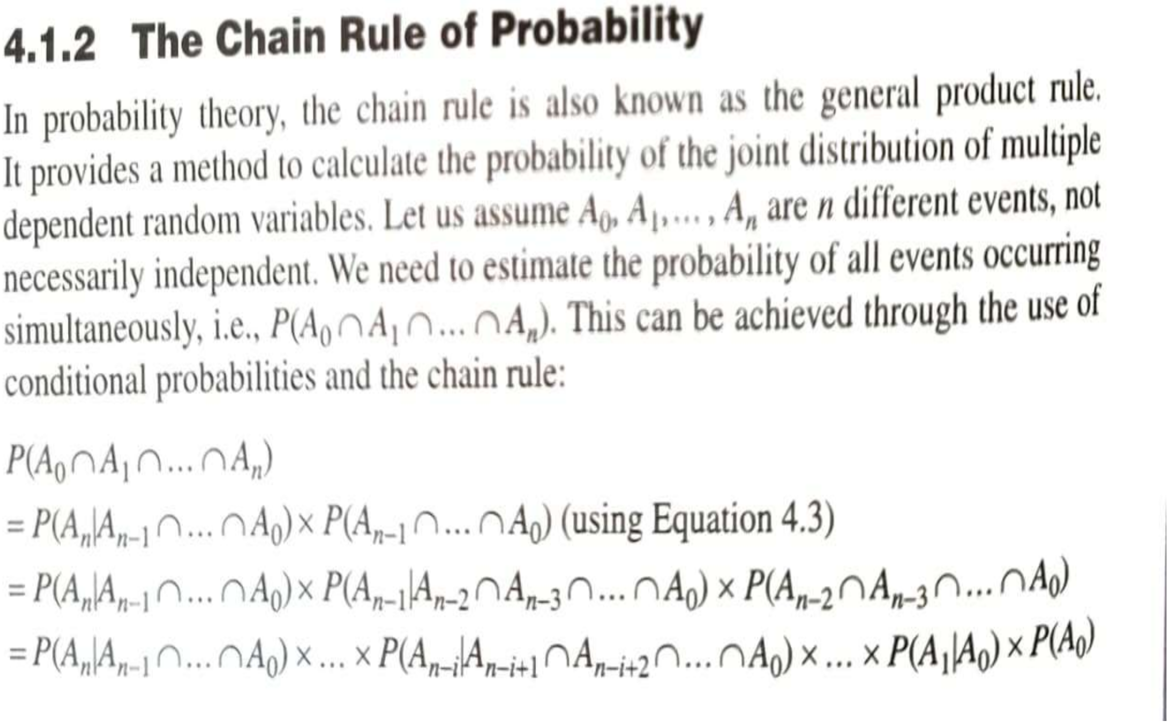
Chain Rule

The chain rule of probability (also known as the probability multiplication rule) allows us to express the joint probability of multiple random variables in terms of conditional probabilities.

For two events $A$and B:

$$P(A, B) = P(A) \cdot P(B \mid A)$$

This means the probability that **both A and B** occur equals the probability that **A** occurs multiplied by the probability that **B** occurs given that **A** has already happened.



4.1.2 The Chain Rule of Probability

In probability theory, the chain rule is also known as the general product rule. It provides a method to calculate the probability of the joint distribution of multiple dependent random variables. Let us assume $A_0, A_1, \ldots, A_n$ are n different events, not necessarily independent. We need to estimate the probability of all events occurring simultaneously, i.e., $P(A_0 \cap A_1 \cap \ldots \cap A_n)$. This can be achieved through the use of conditional probabilities and the chain rule:

$$\begin{aligned}
&P(A_0 \cap A_1 \cap \ldots \cap A_n) \\
&= P(A_n | A_{n-1} \cap \ldots \cap A_0) \times P(A_{n-1} \cap \ldots \cap A_0) \text{ (using Equation 4.3)} \\
&= P(A_n | A_{n-1} \cap \ldots \cap A_0) \times P(A_{n-1} | A_{n-2} \cap A_{n-3} \cap \ldots \cap A_0) \times P(A_{n-2} \cap A_{n-3} \cap \ldots \cap A_0) \\
&= P(A_n | A_{n-1} \cap \ldots \cap A_0) \times \ldots \times P(A_{n-i} | A_{n-i+1} \cap A_{n-i+2} \cap \ldots \cap A_0) \times \ldots \times P(A_1 | A_0) \times P(A_0)
\end{aligned}$$

or rewriting the previous step in the reverse order,

$$= P(A_0) \times P(A_1|A_0) \times P(A_2|A_1 \cap A_0) \times P(A_3|A_2 \cap A_1 \cap A_0) \times \ldots \times P(A_n|A_{n-1} \cap \ldots \cap A_0)$$

$$= \prod_{i=0}^{n} P(A_i|A_{i-1} \cap \ldots \cap A_0)$$

Therefore,

$$P(A_0 \cap A_1 \cap \ldots \cap A_n) = \prod_{i=0}^{n} P(A_i|A_{i-1} \cap \ldots \cap A_0) = \prod_{i=0}^{n} P(A_i|A_{i-1}, \ldots, A_0) \quad (4.4)$$

Following the above-mentioned generalised chain rule of probability, Equation (4.4), the joint probabilistic distribution over a sequence of words $\{W_0, W_1, W_2, \ldots, W_n\}$ can be expressed as follows:

$$P(W_0 W_1 W_2 \ldots W_n) = \prod_{i=0}^{n} P(W_i|W_{i-1}, \ldots, W_0) = \prod_{i=0}^{n} P(W_i|W_0 : W_{i-1}) \quad (4.5)$$

Here, $P(W_i|W_0, W_1, \ldots, W_{i-1})$ can be calculated using maximum likelihood estimate as $\frac{\text{count}(W_0, W_1, \ldots, W_i)}{\text{count}(W_0, W_1, \ldots, W_{i-1})}$, where count $(W_0, W_1, \ldots, W_i)$ = the number of occurrences of the sequence $(W_0, W_1, \ldots, W_i)$ in the training corpus.

Example 4.1

Find the expression for the joint distribution of the words in the sentence, '*The monsoon season has begun*'.

Solution

Here, we will see each token at ith position, i.e., W_i using $W_{0:i-1}$ as its context in the sentence.

$$\begin{aligned} &P(\text{The monsoon season has begun}) \\ &= P(\text{The}) \times P(\text{monsoon}|\text{The}) \times P(\text{season}|\text{The monsoon}) \\ &\quad \times P(\text{has}|\text{The monsoon season}) \times P(\text{begun}|\text{The monsoon season has}) \end{aligned} \quad (4.6)$$

4.1.4 Unigram Language Model

As the term *unigram* suggests, the probabilistic value of a word, W_i, in a sequence of words $\{W_0, W_1, W_2, W_3, \ldots, W_n\}$ depends on itself only. This is also known as the zero-order Markov process. A unigram language model considers no previous context for W_i. The unigram probability of a word W_i can be estimated as the fraction of times it has appeared in the training corpus, w.r.t. the total number of words available in the corpus,

$$P(W_i) = \frac{\text{count}(W_i)}{N}$$

where N = total number of words in the training corpus. Also, the probability of a sequence of words or a sentence can be expressed as a multiplication of the probabilistic estimation of each word, i.e.,

$$P(W_0 W_1 W_2 \ldots W_n) = P(W_0) \times P(W_1) \times P(W_2) \times \cdots \times P(W_i) \times \cdots \times P(W_n) = \prod_{i=0}^{n} P(W_i)$$

Example 4.3 Find the probability of the sentence, '*The lazy cat sells sea shells*' using a unigram language model and the training data (case-insensitive) given below.

Sentence 1: *The cat sat on the mat*
Sentence 2: *A quick brown fox jumps over the lazy dog*
Sentence 3: *She sells sea shells by the sea shore*
Sentence 4: *He reads books every evening before bed*
Sentence 5: *The sun rises in the east and sets in the west*

Solution

There are a total of 41 words (N) in the given corpus. Let us calculate the probability of individual words, *The*, *lazy*, *cat*, *sells*, *sea*, and *shells* of the given input sentence.

$$P(The) = \text{count}(The)/N = 7/41,\ P(lazy) = \text{count}(lazy)/N = 1/41$$
$$P(cat) = \text{count}(cat)/N = 1/41,\ P(sells) = \text{count}(sells)/N = 1/41$$
$$P(sea) = \text{count}(sea)/N = 2/41,\ P(shells) = \text{count}(shells)/N = 1/41$$

Now, we can use the above probabilities to estimate the probability of the input sentence:

$$\begin{aligned} &P(The\ lazy\ cat\ sells\ sea\ shells) \\ &= P(The) \times P(lazy) \times P(cat) \times P(sells) \times P(sea) \times P(shells) \\ &= \frac{7}{41} \times \frac{1}{41} \times \frac{1}{41} \times \frac{1}{41} \times \frac{2}{41} \times \frac{1}{41} = \frac{14}{41^6} \approx 2.947 \times 10^{-9} \end{aligned}$$

4.1.5 Bigram Language Model

As the term *bigram* suggests, the probability of a word, W_i, in a sequence $\{W_0, W_1, W_2, W_3, \ldots, W_n\}$ depends only on the previous word, i.e., W_{i-1}. In short, a bigram language model considers a previous context of length one. The probabilistic value of a word, W_i, can be expressed as,

$$P(W_i|W_{i-k}, W_{i-k+1}, \ldots, W_{i-1}) \approx P(W_i|W_{i-1})$$

and

$$P(W_i|W_{i-1}) = \frac{\text{count}(W_{i-1}, W_i)}{\text{count}(W_{i-1})}$$

Example 4.4

Find the probability of the sentence, '*The cat sells sea shells*' using a bigram language model and the training data (case-insensitive) given below.

Sentence 1: *The cat sat on the mat*
Sentence 2: *The cat sells a sea shell to the lazy dog*
Sentence 3: *She sells sea shells by the sea shore*
Sentence 4: *He reads books every evening before bed*
Sentence 5: *The sun rises in the east and sets in the west*

Solution

Let us introduce the start token '*<START>*' at the beginning of each sentence. Therefore, all the sentences will appear as

Sentence 1: *<START> The cat sat on the mat*
Sentence 2: *<START> The cat sells a sea shell to the lazy dog*
Sentence 3: *<START> He sells sea shells by the sea shore*
Sentence 4: *<START> He reads books every evening before bed*
Sentence 5: *<START> The sun rises in the east and sets in the west*

***Target input* :** *<START> The cat sells sea shells*

Now, we need to calculate the likelihood of each bigram token for the input sentence, '*<START> The cat sells sea shells*'. Note that the likelihood of the start token is 1.

$P(<START>\ The\ cat\ sells\ sea\ shells)$
$= P(<START>) \times P(The|<START>) \times P(cat|The) \times P(sells|cat) \times P(sea|sells)$
$\times P(shells|sea)$

$$P(<\text{START}>) = 1$$

$$P(\text{The} \mid <\text{START}>) = \frac{\langle <START>, The\rangle \text{ appears three times}}{\text{'}<START>\text{' appears five times}} = \frac{3}{5}$$

$$P(\text{cat} \mid \text{The}) = \frac{\langle The, cat\rangle \text{ appears two times}}{\text{'}The\text{' appears eight times}} = \frac{1}{4}$$

$$P(\text{sells} \mid \text{cat}) = \frac{\langle cat, sells\rangle \text{ appears one times}}{\text{'}cat\text{' appears two times}} = \frac{1}{2}$$

$$P(\text{sea} \mid \text{sells}) = \frac{\langle sells, sea\rangle \text{ appears one times}}{\text{'}sells\text{' appears two times}} = \frac{1}{2}$$

$$P(\text{shells} \mid \text{sea}) = \frac{\langle sea, shells\rangle \text{ appears one times}}{\text{'}sea\text{' appears 2 times}} = \frac{1}{2}$$

$$\Rightarrow P(The\ cat\ sells\ sea\ shells) = 1 \times \frac{3}{5} \times \frac{1}{4} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{160} = 0.01875$$

Note that the likelihood of the initial term, w_0 of a sentence s, $P(w_0)$, is different from the w_0's unigram probability. It shall be indifferent with or without the start token *<START>*.

For example, we have already seen the likelihood of the phrase $P(<START>\ The)$ as $P(<START>) \times P(The|<START>)$, which is $1 \times \frac{3}{5} = \frac{3}{5}$. Similarly, $P(The)$ can be expressed as the number of occurrences of the initial word out of all sentences, i.e., 3/5.

4.2.2 Smoothing

The smoothing technique allows us to avoid the zero probability mentioned earlier by 'stealing' some probability mass from the tokens having high probability and distributing it among those tokens which do not have sufficient probability (or probability of zero). The simple smoothing technique is *add-1* smoothing or *Laplace* smoothing, in which the original count of an n-gram is increased by one. In particular, the conditional probability after smoothing can be written as:

$$P_{add-1}(W_i|W_{i-1}) = \frac{\text{count}(W_{i-1}, W_i)+1}{\text{count}(W_{i-1})+|V|}$$

For an unseen token, the count(W_{i-1}, W_i) is 0, but due to adding 1 to the numerator, it does not reduce to zero. The denominator is also increased by $|V|$, the size of the vocabulary, to make sure that $\Sigma_{v\in V} P_{add-1}(W_v|W_{i-1}) = 1$.

4.6 Find the probability of the sentence, '*The furry cat sells sea shells*' using a bigram language model and the training data.

Sentence 1: *The cat sat on the mat*
Sentence 2: *A quick brown fox jumps over the lazy dog*
Sentence 3: *She sells sea shells by the sea shore*
Sentence 4: *He reads books every evening before bed*
Sentence 5: *The sun rises in the east and sets in the west*

Solution

As we have seen earlier,

$P(\textit{The furry cat sells sea shells})$
$= P(\textit{furry}|\textit{The}) \times P(\textit{cat}|\textit{furry}) \times P(\textit{sells}|\textit{cat}) \times P(\textit{sea}|\textit{sells}) \times P(\textit{shells}|\textit{sells})$

As already noticed, the term '*furry*' is an unseen word. It sets the value of $P(\textit{furry}|\textit{The})$ and $P(\textit{cat}|\textit{furry})$ to zero. But if add-1 smoothing is used, assuming $|V|$ to be 46, we will get

$$P_{add-1}(\textit{furry}|\textit{The}) = \frac{\text{count}(\textit{The furry})+1}{\text{count}(\textit{The})+46} = \frac{0+1}{7+46} = \frac{1}{53}$$

$$P_{add-1}(\textit{cat}|\textit{furry}) = \frac{\text{count}(\textit{furry cat})+1}{\text{count}(\textit{furry})+46} = \frac{0+1}{0+41} = \frac{1}{41}$$

However, it has been observed that adding a 1 to the numerator can cause a significant shift in the probabilistic distribution, especially when the underlying training corpus is small or medium in size. To address this issue, the add-1 smoothing formula is further modified, resulting in the add-k smoothing. The formula for add-k smoothing is

$$P_{add-k}(W_i|W_{i-1}) = \frac{\text{count}(W_{i-1}, W_i)+k}{\text{count}(W_{i-1})+kV}, \text{ where } k \in (0, 1)$$

The drawbacks of this approach are that it assigns the same probability to all unseen words in the test case, and it is challenging to balance the distributional shift. The issue arises due to the involvement of a longer context of W_i, i.e., $\{W_{i-k}, \ldots, W_{i-2}, W_{i-1}\}$. So, if a shorter context is considered instead of using the entire one, it might result in a match. Next, we will discuss two plausible approaches, which opt for a shorter context if the original context causes a no-match.

4.2.3 Back-Off

This algorithm employs recursion to shorten the context size in order to address the unseen tokens. Initially, the algorithm begins with a higher-order n-gram to calculate $P(W_i|W_{i-1},\ldots, W_{i-k+1})$. If it encounters any challenges, it falls back to lower-order n-grams. For example, assume that $P(fox|A\ quick\ brown)$ is zero. Then, the back-off algorithm will search for $P(fox|quick\ brown)$, and in case of a failure, it will again look for $P(fox|brown)$, i.e., at every step, the algorithm reduces the context size by one. The back-off algorithm can be expressed as,

$$P(W_i|W_{i-k+1}^{i-1}) = \begin{cases} \dfrac{\text{count}(W_{i-k+1}^{i})}{\text{count}(W_{i-k+1}^{i-1})}, & \text{if count}(W_{i-k+1}^{i}) > 0 \\ \alpha \times P(W_i|W_{i-k+2}^{i-1}), & \text{otherwise, where } \alpha \in (0, 1) \end{cases}$$

and

$$P(W_i) = \frac{\text{count}(W_i)}{N}, \text{ where } |V| \text{ is the length of the vocabulary.}$$

Here, α is a penalty for considering a shorter context.

Example 4.7

Find the probability of the sentence '*The furry cat sells sea shells*' w.r.t. the training data and a bigram language model. Assume that $\alpha = 0.4$.

Sentence 1: *The cat sat on the mat*

Sentence 2: *A furry brown fox jumps over the lazy dog*

Sentence 3: *She sells sea shells by the sea shore*

Sentence 4: *He reads books every evening before bed*

Sentence 5: *The sun rises in the east and sets in the west*

Solution

$$P(The\ furry\ cat\ sells\ sea\ shells)$$
$$= P(The) \times P(furry|The) \times P(cat|furry) \times P(sells|cat) \times P(sea|sells) \times P(shells|sells)$$

Notice that $P(The)$ indicates the probability of the word '*The*' at the starting position. It means $P(The)$ is not a unigram probability but conditioned on being an initial word, which is equivalent to having the <*START*> token for all the sentences.

$$P(The) = \frac{2}{5}$$

$$P(furry|The) = \alpha \times P(furry) = 0.4 \times \frac{1}{41} = \frac{4}{410}$$

$$P(cat|furry) = \alpha \times P(cat) = 0.4 \times \frac{1}{41} = \frac{4}{410},$$

$$P(sells|cat) = \alpha \times P(sells) = 0.4 \times \frac{1}{41} = \frac{4}{410}$$

$$P(sea|sells) = \frac{1}{1} = 1,$$

$$P(shells|sea) = \frac{1}{2}$$

Now, replacing the bigram probabilities with corresponding values, we get

$$P(\textit{The furry cat sells sea shells}) = \frac{2}{5} \times \frac{4}{410} \times \frac{4}{410} \times \frac{4}{410} \times 1 \times \frac{1}{2} \approx 1.85 \times 10^{-7}.$$