Generative modeling

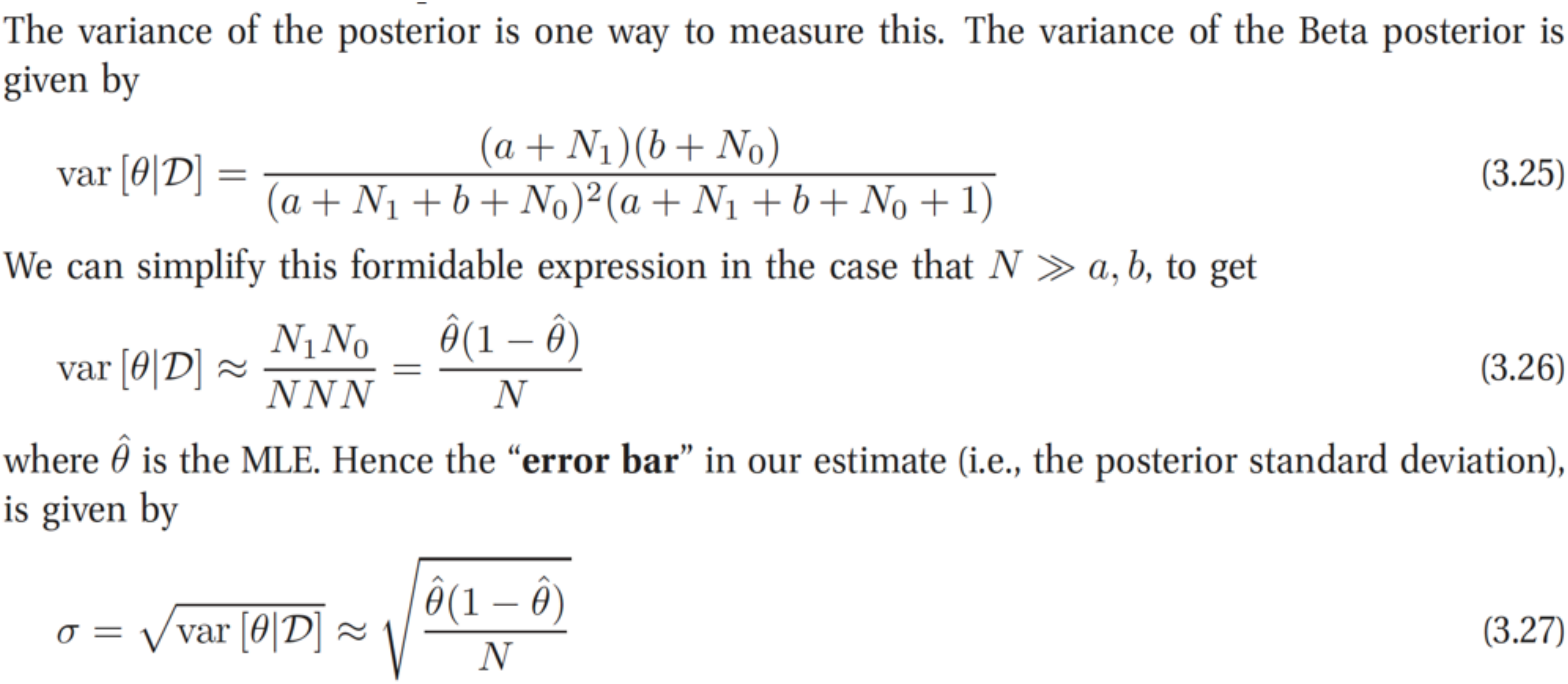
Sep 2022

Murphy chap3

Posterior variance

Can use mean or mode as parameter estimates

But, these are point estimates, it is useful to know how much we can trust them.



The variance of the posterior is one way to measure this. The variance of the Beta posterior is given by

$$\text{var}[\theta|\mathcal{D}] = \frac{(a+N_1)(b+N_0)}{(a+N_1+b+N_0)^2(a+N_1+b+N_0+1)} \tag{3.25}$$

We can simplify this formidable expression in the case that $N \gg a, b$, to get

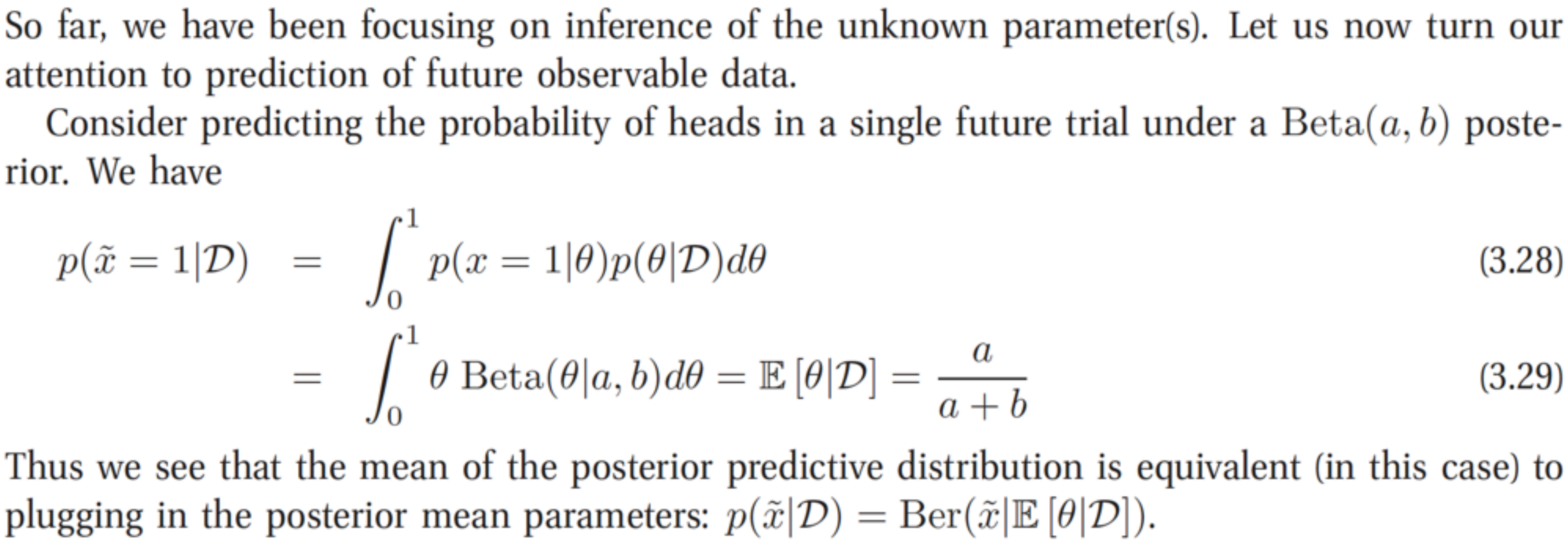
$$\text{var}[\theta|\mathcal{D}] \approx \frac{N_1 N_0}{NNN} = \frac{\hat{\theta}(1-\hat{\theta})}{N} \tag{3.26}$$

where $\hat{\theta}$ is the MLE. Hence the "**error bar**" in our estimate (i.e., the posterior standard deviation), is given by

$$\sigma = \sqrt{\text{var}[\theta|\mathcal{D}]} \approx \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}} \tag{3.27}$$

Posterior predictive distribution

Probability of head in a single future trial ?



So far, we have been focusing on inference of the unknown parameter(s). Let us now turn our attention to prediction of future observable data.

Consider predicting the probability of heads in a single future trial under a $\text{Beta}(a, b)$ posterior. We have

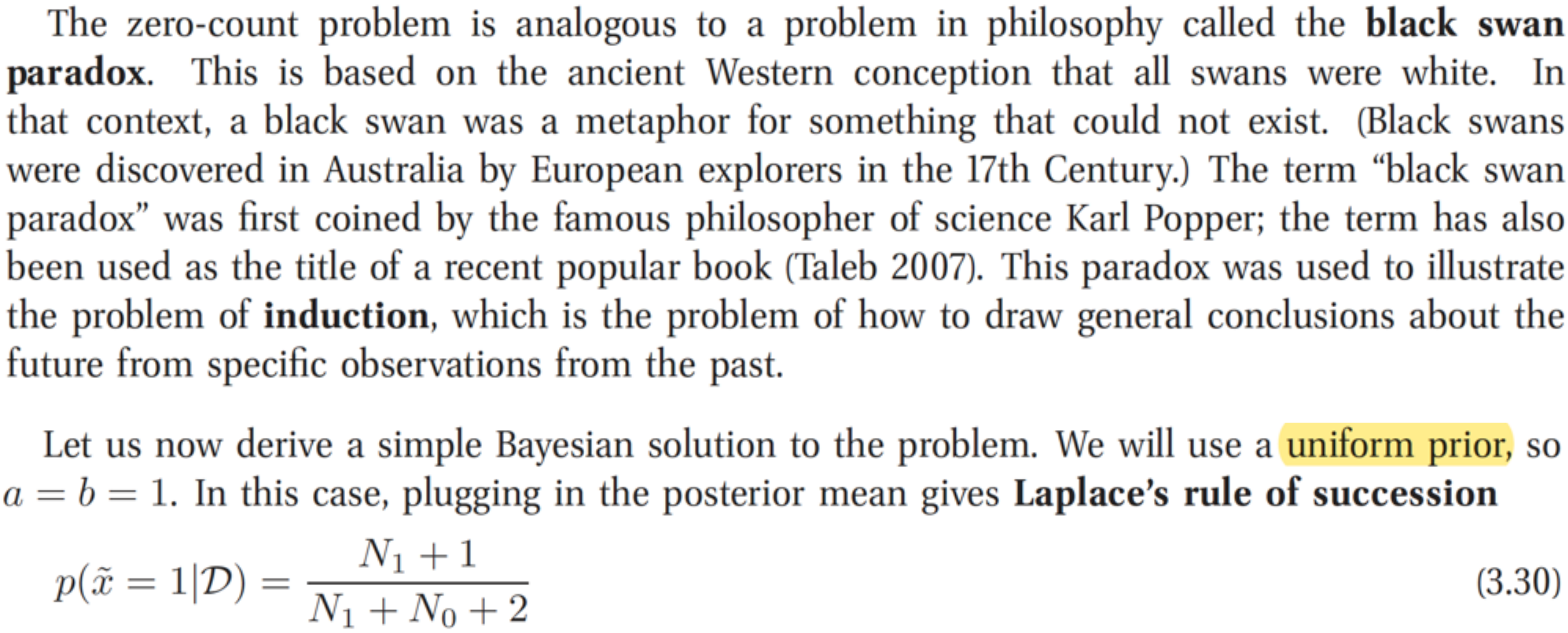
$$p(\tilde{x} = 1|\mathcal{D}) = \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D})d\theta \tag{3.28}$$

$$= \int_0^1 \theta \ \text{Beta}(\theta|a, b)d\theta = \mathbb{E}[\theta|\mathcal{D}] = \frac{a}{a+b} \tag{3.29}$$

Thus we see that the mean of the posterior predictive distribution is equivalent (in this case) to plugging in the posterior mean parameters: $p(\tilde{x}|\mathcal{D}) = \text{Ber}(\tilde{x}|\mathbb{E}[\theta|\mathcal{D}])$.

Overfitting and black swan paradox

Three tails in a row, what's the MLE ?   ZREO !



The zero-count problem is analogous to a problem in philosophy called the **black swan paradox**. This is based on the ancient Western conception that all swans were white. In that context, a black swan was a metaphor for something that could not exist. (Black swans were discovered in Australia by European explorers in the 17th Century.) The term "black swan paradox" was first coined by the famous philosopher of science Karl Popper; the term has also been used as the title of a recent popular book (Taleb 2007). This paradox was used to illustrate the problem of **induction**, which is the problem of how to draw general conclusions about the future from specific observations from the past.

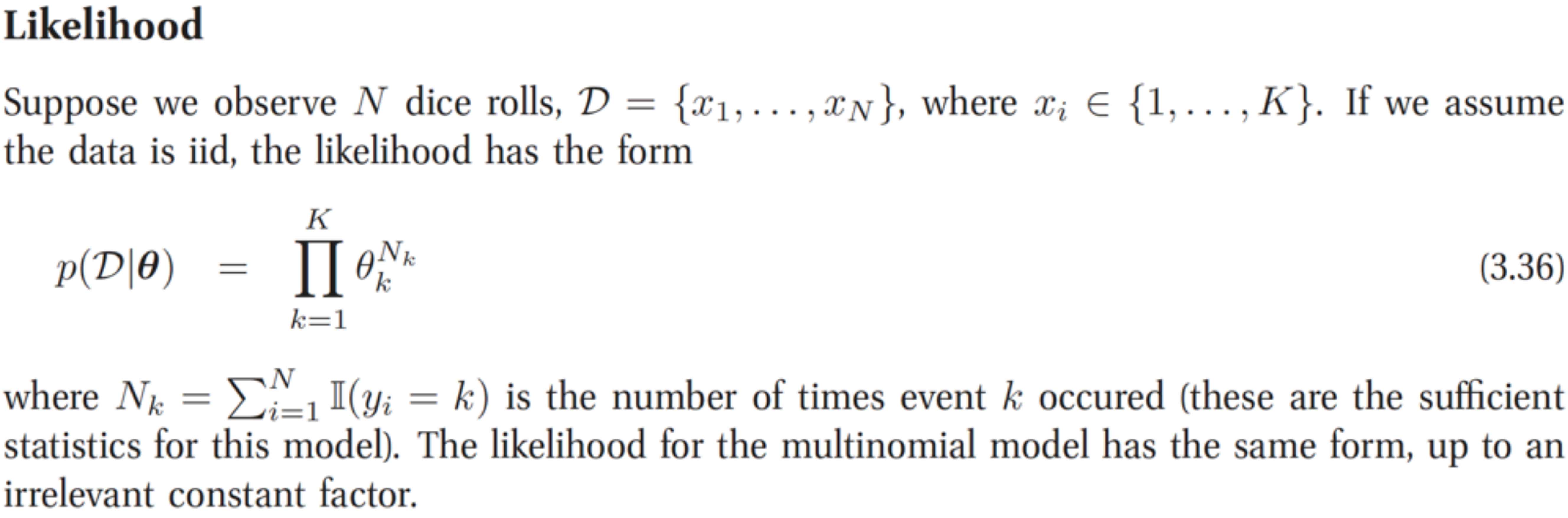
Let us now derive a simple Bayesian solution to the problem. We will use a uniform prior, so $a = b = 1$. In this case, plugging in the posterior mean gives **Laplace's rule of succession**

$$p(\tilde{x} = 1|\mathcal{D}) = \frac{N_1 + 1}{N_1 + N_0 + 2} \tag{3.30}$$

Laplace add-one smoothing !

Dirichlet-multinomial

Replace coin with a K-sided dice, the probability of face k occurring?



Likelihood

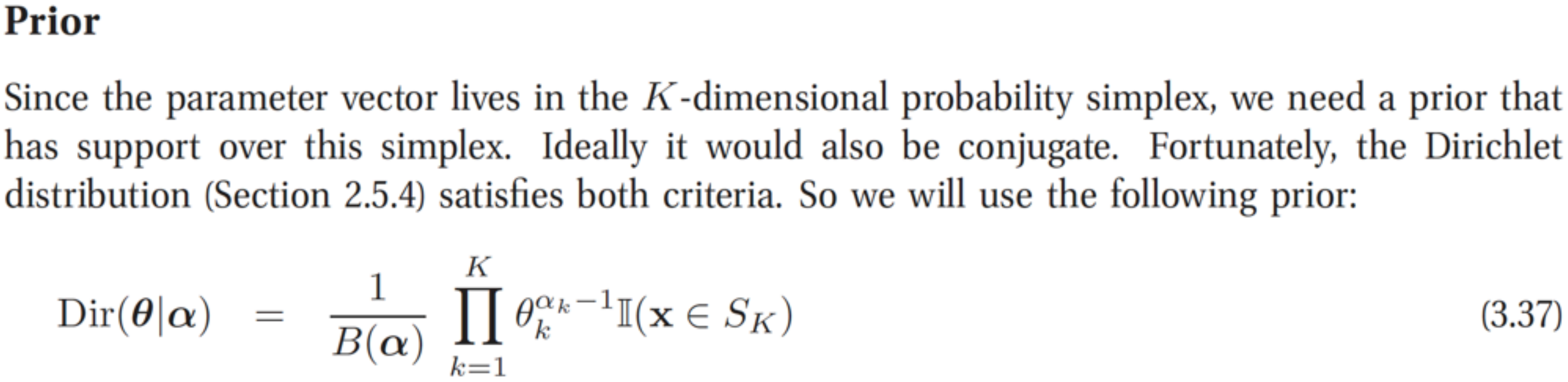
Suppose we observe N dice rolls, $\mathcal{D} = \{x_1, \ldots, x_N\}$, where $x_i \in \{1, \ldots, K\}$. If we assume the data is iid, the likelihood has the form

$$p(\mathcal{D}|\boldsymbol{\theta}) \quad = \quad \prod_{k=1}^{K} \theta_k^{N_k} \qquad (3.36)$$

where $N_k = \sum_{i=1}^{N} \mathbb{I}(y_i = k)$ is the number of times event k occured (these are the sufficient statistics for this model). The likelihood for the multinomial model has the same form, up to an irrelevant constant factor.

Dirichlet-multinomial

Distribution over K-simplex? Dirichlet distribution !

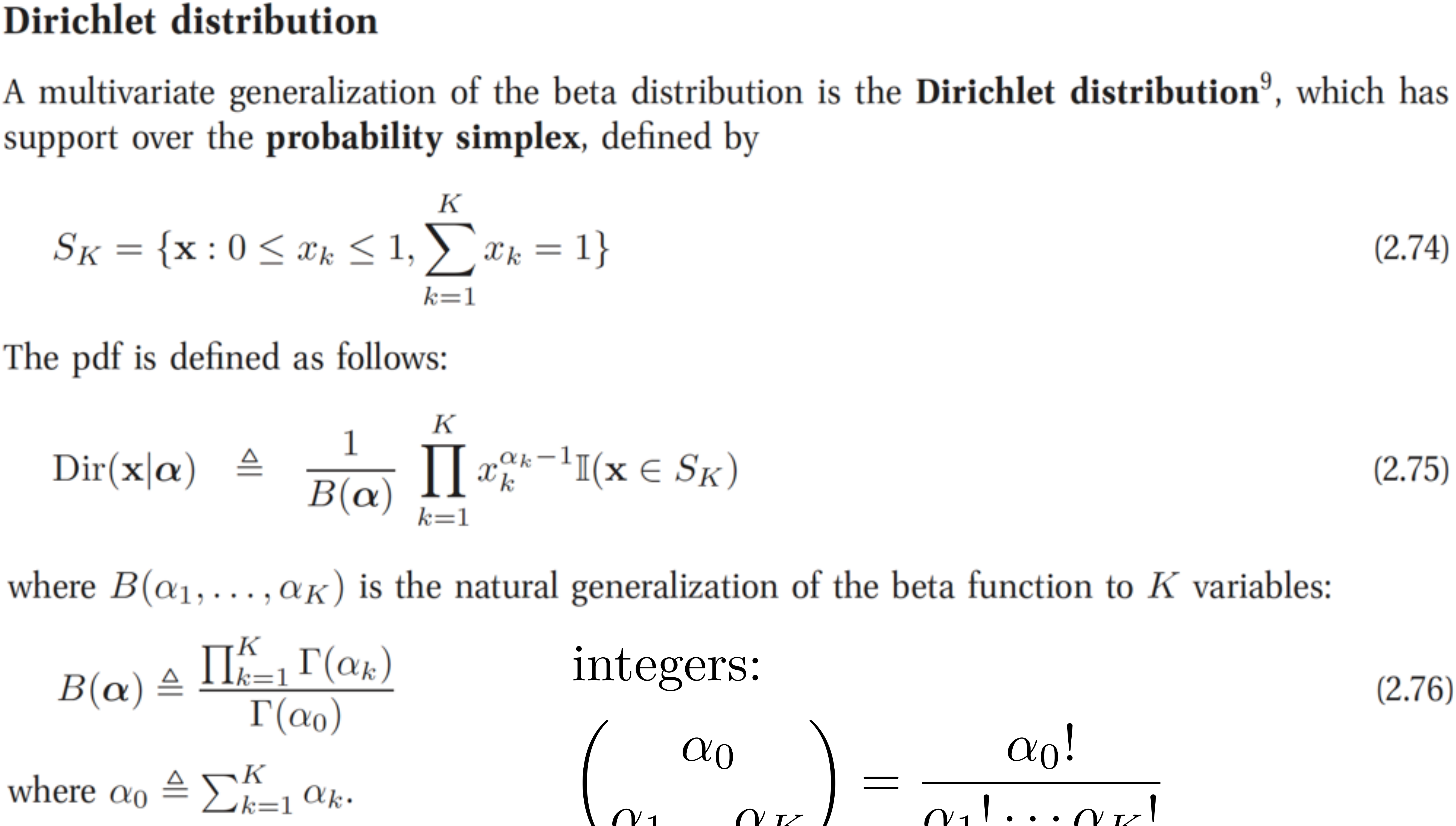


Prior

Since the parameter vector lives in the K-dimensional probability simplex, we need a prior that has support over this simplex. Ideally it would also be conjugate. Fortunately, the Dirichlet distribution (Section 2.5.4) satisfies both criteria. So we will use the following prior:

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K) \tag{3.37}$$

Dirichlet-multinomial



Dirichlet distribution

A multivariate generalization of the beta distribution is the **Dirichlet distribution**[9], which has support over the **probability simplex**, defined by

$$S_K = \{\mathbf{x} : 0 \le x_k \le 1, \sum_{k=1}^{K} x_k = 1\} \tag{2.74}$$

The pdf is defined as follows:

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K) \tag{2.75}$$

where $B(\alpha_1, \ldots, \alpha_K)$ is the natural generalization of the beta function to K variables:

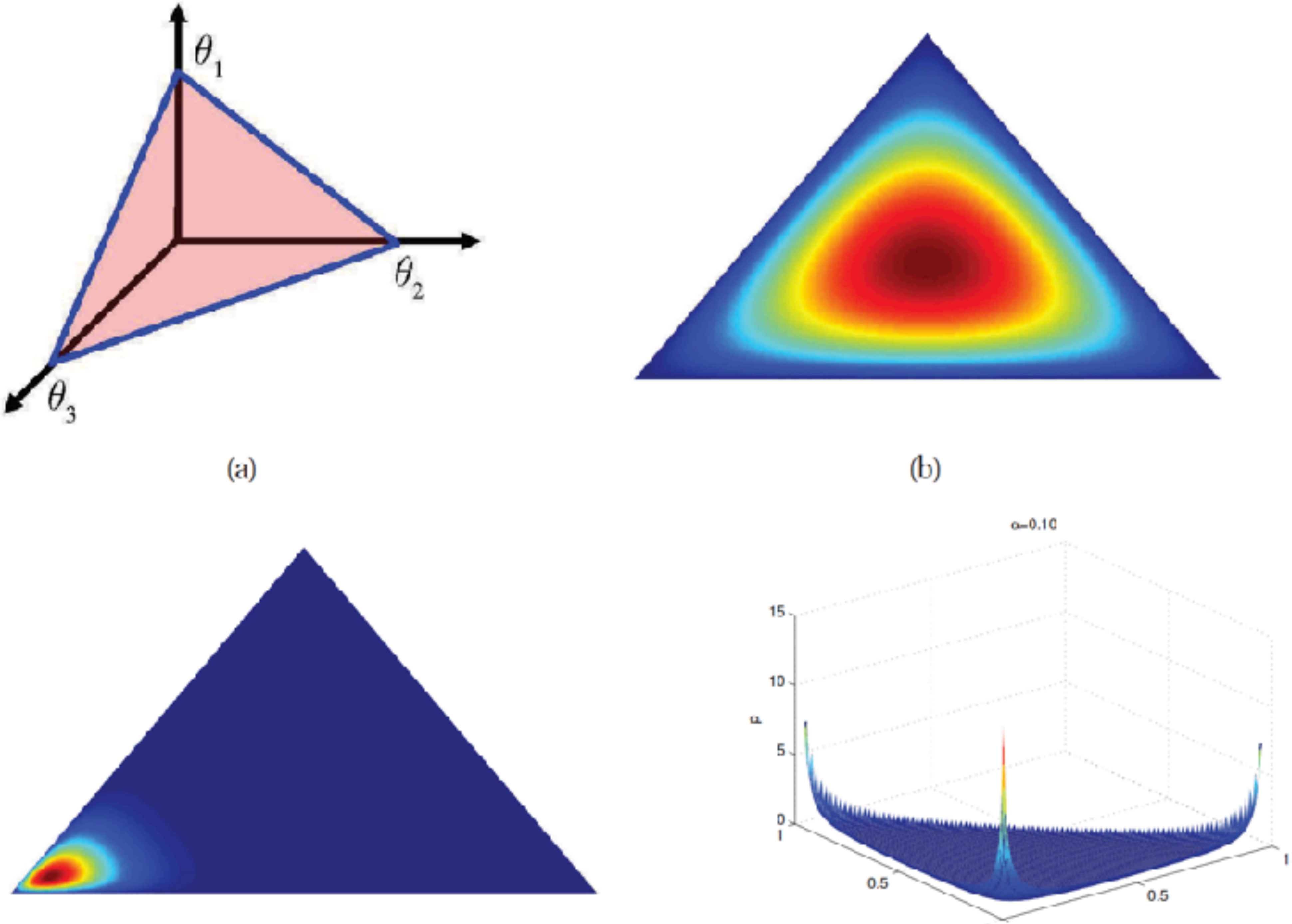
$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \tag{2.76}$$

where $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$.

integers:

$$\binom{\alpha_0}{\alpha_1, \ldots \alpha_K} = \frac{\alpha_0!}{\alpha_1! \cdots \alpha_K!}$$

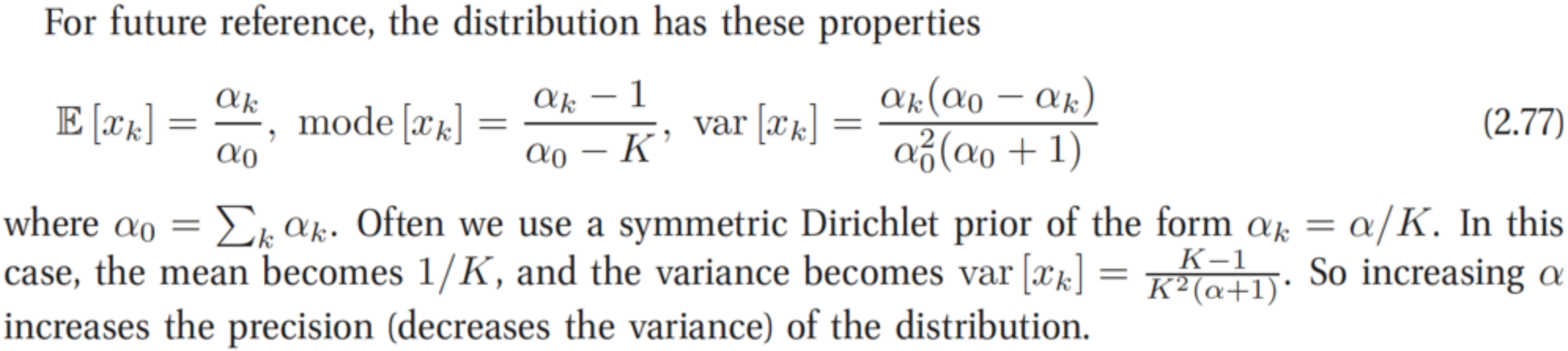
Dirichlet-multinomial



(a) (b)

(c) (d)

Dirichlet-multinomial



For future reference, the distribution has these properties

$$\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \ \text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \ \text{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.77}$$

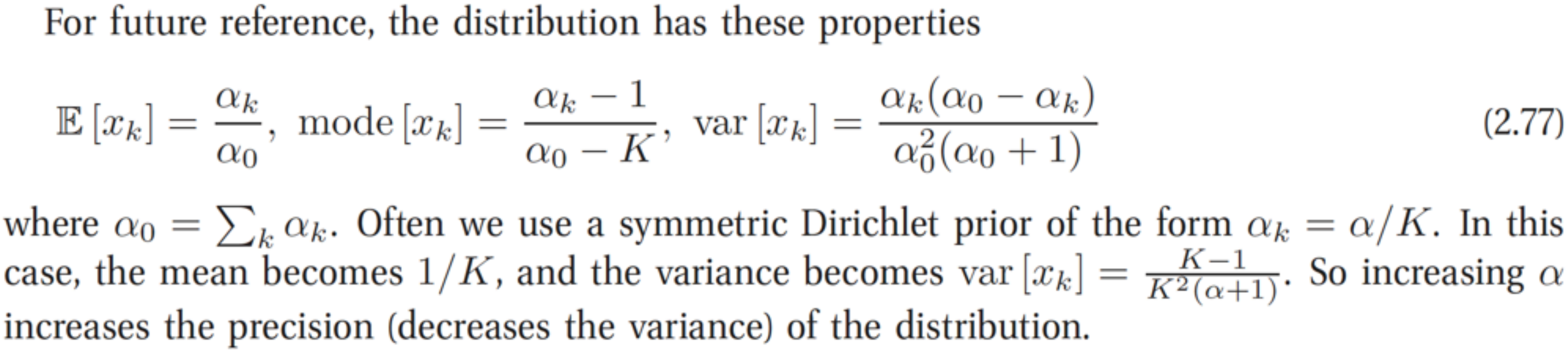
where $\alpha_0 = \sum_k \alpha_k$. Often we use a symmetric Dirichlet prior of the form $\alpha_k = \alpha/K$. In this case, the mean becomes $1/K$, and the variance becomes $\text{var}[x_k] = \frac{K-1}{K^2(\alpha+1)}$. So increasing α increases the precision (decreases the variance) of the distribution.

marginal:

$$x_k \sim Beta(\alpha_k, \alpha_0 - \alpha_k)$$

Bishop 2.2

Dirichlet-multinomial



For future reference, the distribution has these properties

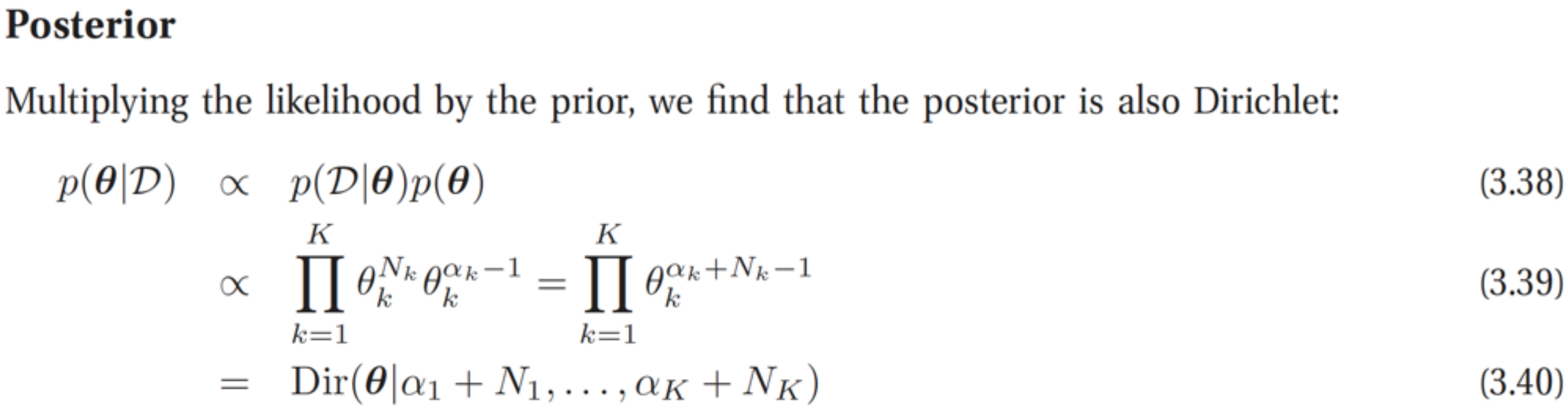
$$\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \text{ mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \text{ var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.77}$$

where $\alpha_0 = \sum_k \alpha_k$. Often we use a symmetric Dirichlet prior of the form $\alpha_k = \alpha/K$. In this case, the mean becomes $1/K$, and the variance becomes $\text{var}[x_k] = \frac{K-1}{K^2(\alpha+1)}$. So increasing α increases the precision (decreases the variance) of the distribution.

marginal:

$$x_k \sim Beta(\alpha_k, \alpha_0 - \alpha_k)$$

Dirichlet-multinomial



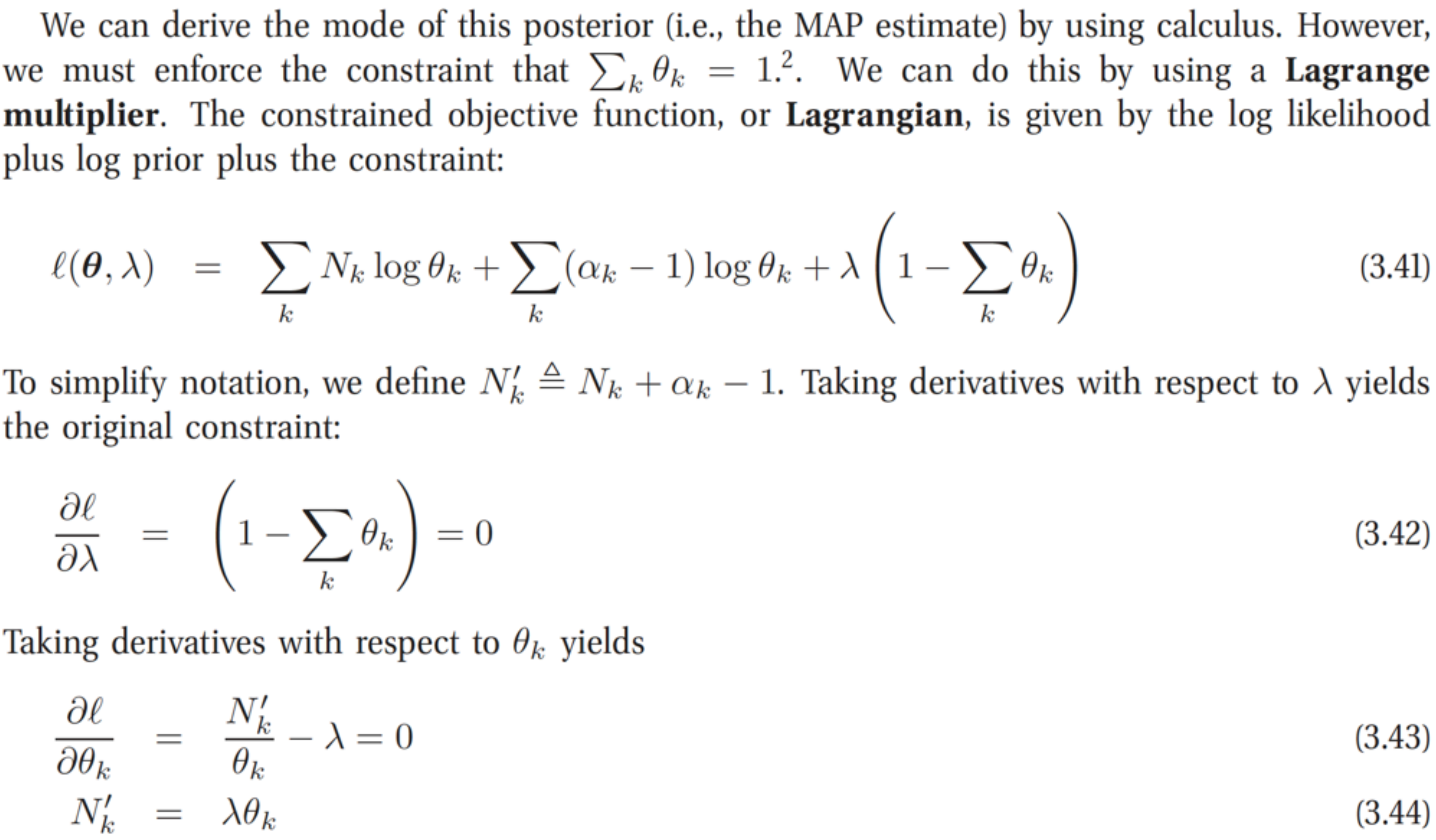
Posterior

Multiplying the likelihood by the prior, we find that the posterior is also Dirichlet:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) && (3.38)\\ &\propto \prod_{k=1}^{K} \theta_k^{N_k}\theta_k^{\alpha_k-1} = \prod_{k=1}^{K}\theta_k^{\alpha_k+N_k-1} && (3.39)\\ &= \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1+N_1,\ldots,\alpha_K+N_K) && (3.40)\end{aligned}$$

MAP estimate ?

Dirichlet-multinomial



We can derive the mode of this posterior (i.e., the MAP estimate) by using calculus. However, we must enforce the constraint that $\sum_k \theta_k = 1$.[2] We can do this by using a **Lagrange multiplier**. The constrained objective function, or **Lagrangian**, is given by the log likelihood plus log prior plus the constraint:

$$\ell(\boldsymbol{\theta}, \lambda) = \sum_k N_k \log \theta_k + \sum_k (\alpha_k - 1) \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right) \tag{3.41}$$

To simplify notation, we define $N'_k \triangleq N_k + \alpha_k - 1$. Taking derivatives with respect to λ yields the original constraint:

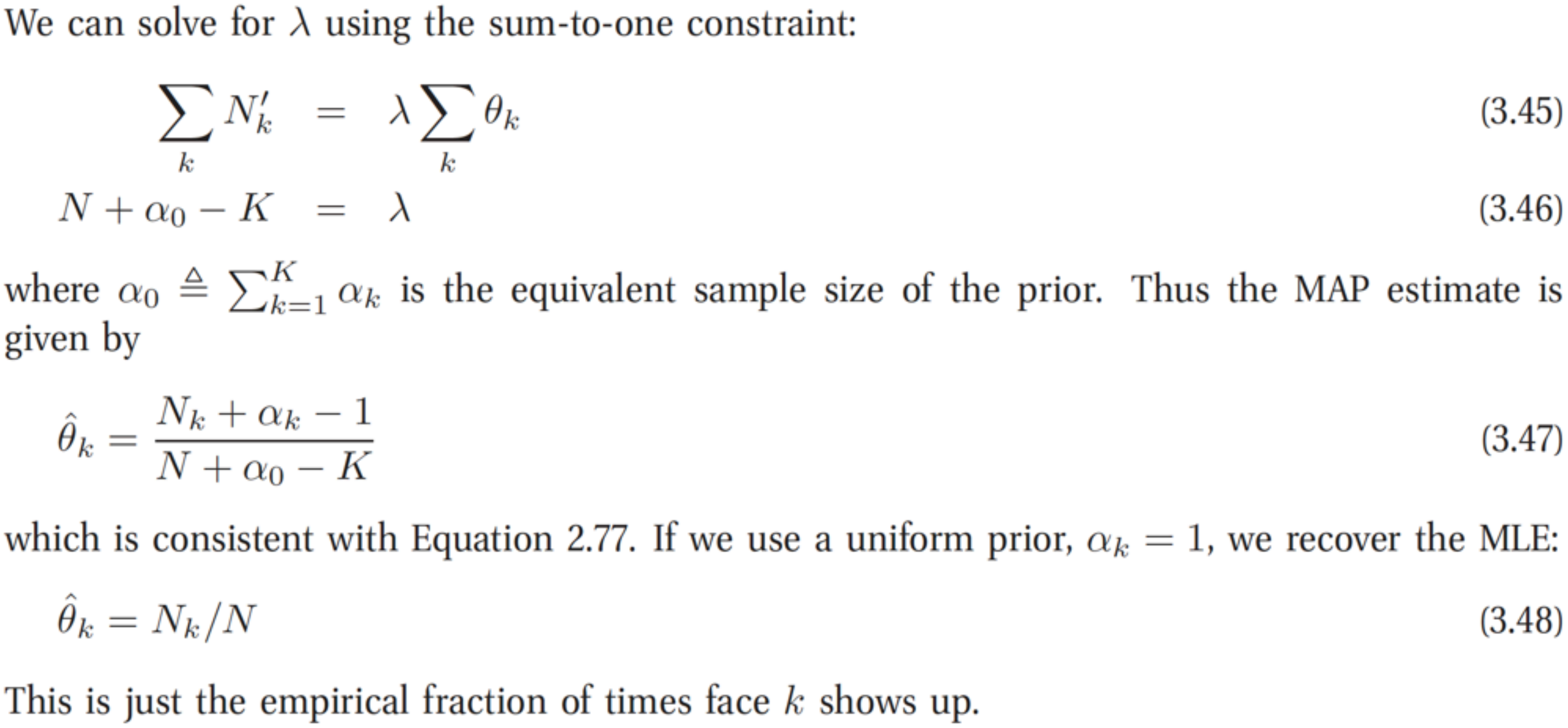
$$\frac{\partial \ell}{\partial \lambda} = \left(1 - \sum_k \theta_k\right) = 0 \tag{3.42}$$

Taking derivatives with respect to θ_k yields

$$\frac{\partial \ell}{\partial \theta_k} = \frac{N'_k}{\theta_k} - \lambda = 0 \tag{3.43}$$

$$N'_k = \lambda \theta_k \tag{3.44}$$

Dirichlet-multinomial



We can solve for λ using the sum-to-one constraint:

$$\sum_k N'_k = \lambda \sum_k \theta_k \tag{3.45}$$

$$N + \alpha_0 - K = \lambda \tag{3.46}$$

where $\alpha_0 \triangleq \sum_{k=1}^K \alpha_k$ is the equivalent sample size of the prior. Thus the MAP estimate is given by

$$\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K} \tag{3.47}$$

which is consistent with Equation 2.77. If we use a uniform prior, $\alpha_k = 1$, we recover the MLE:

$$\hat{\theta}_k = N_k / N \tag{3.48}$$

This is just the empirical fraction of times face k shows up.

Exercise: posterior predictive distribution

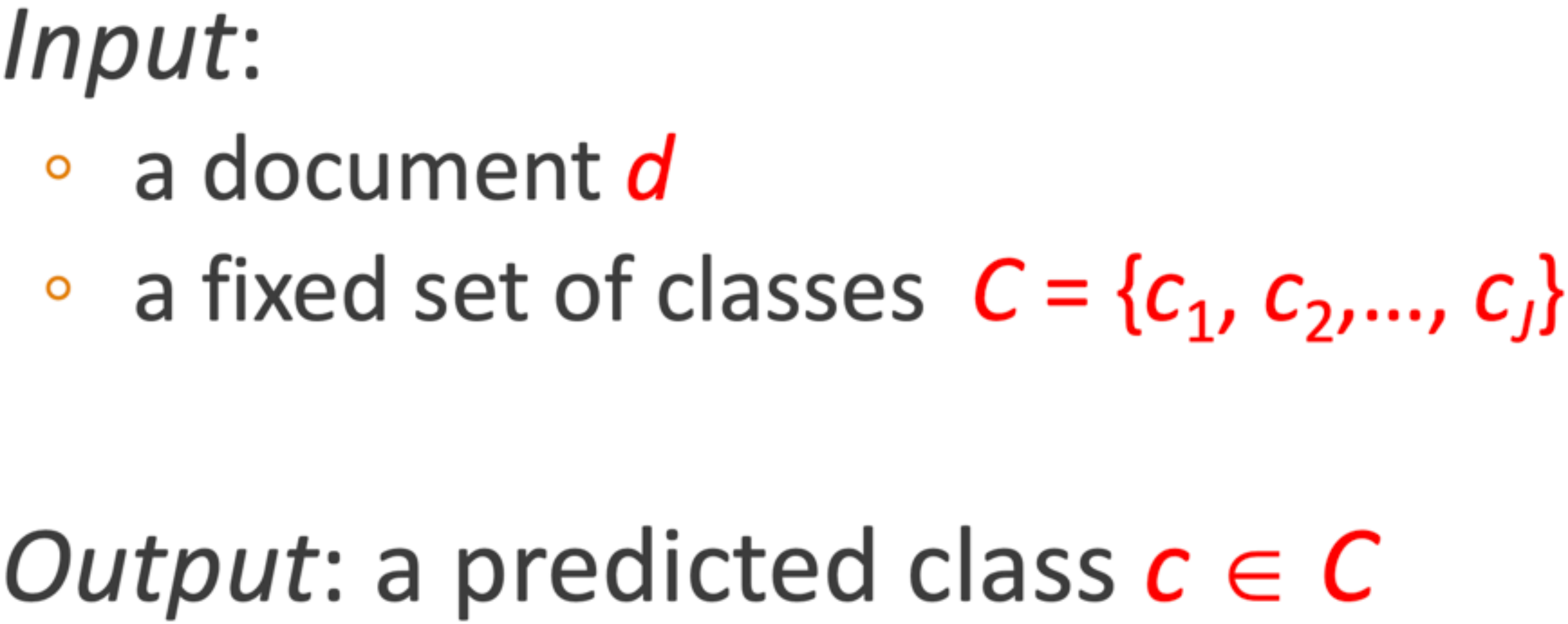
Multinomial Naive Bayesian for LM

Based on Jurafsky and Murphy

Spam classification, subject classification, sentiment classification

Arts
Science
Fashion
......

Multinomial Naive Bayesian for LM

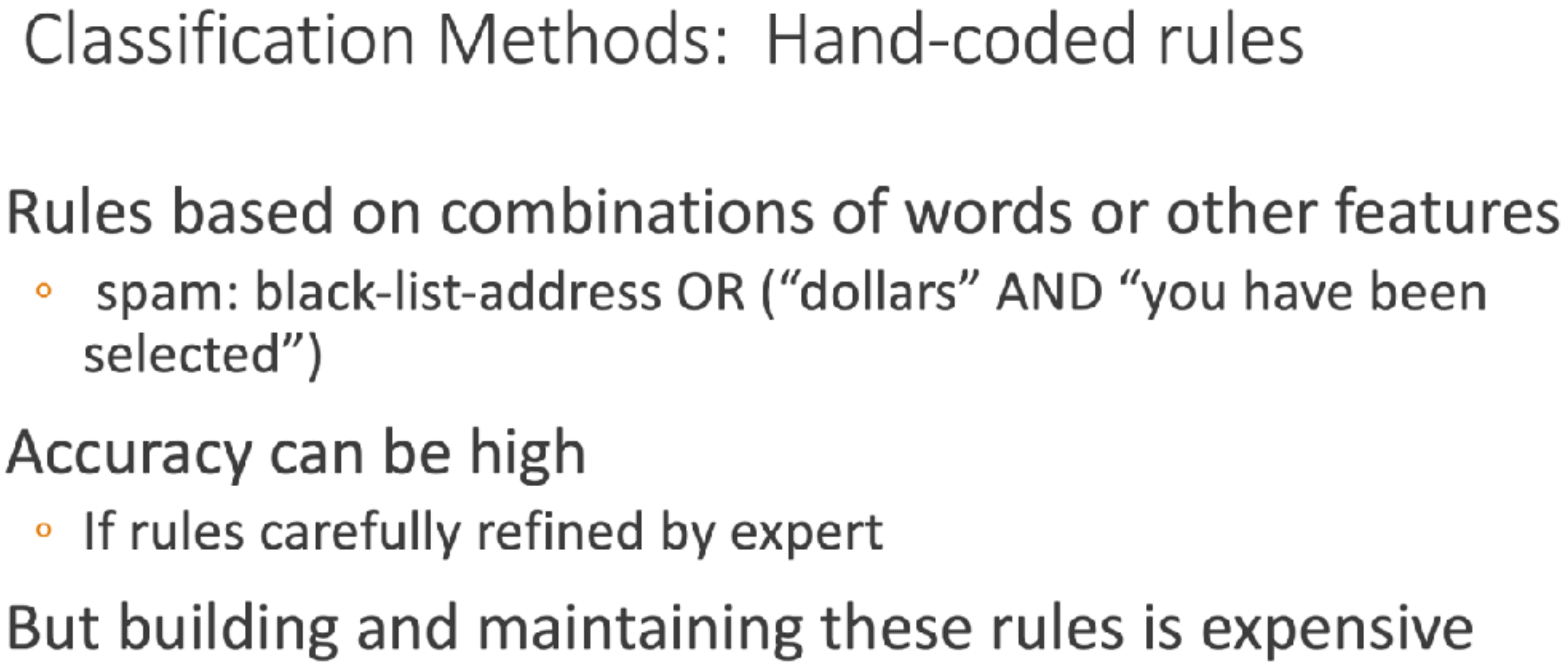


Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

Output: a predicted class $c \in C$

Multinomial Naive Bayesian for LM



Classification Methods: Hand-coded rules

Rules based on combinations of words or other features

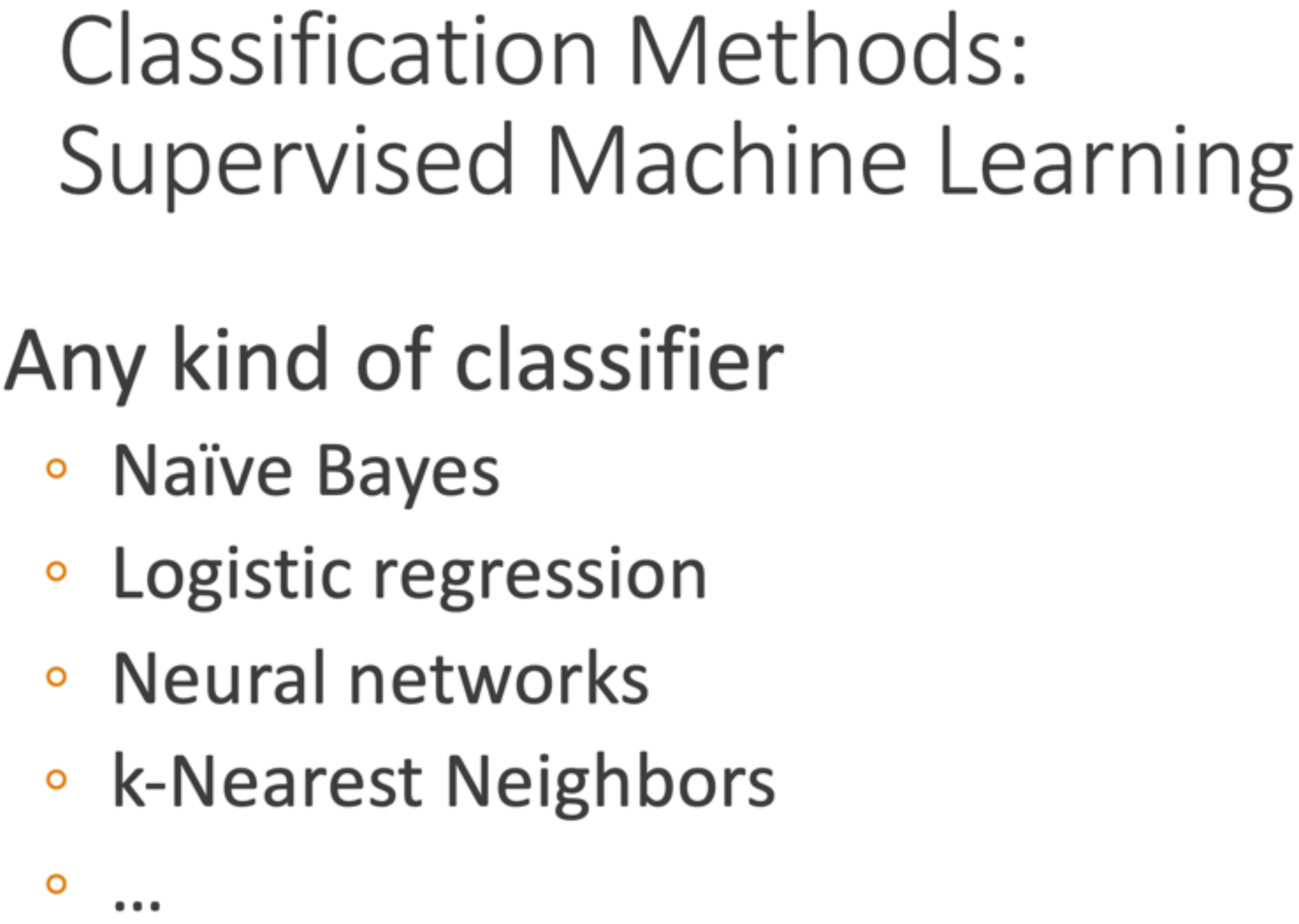
- spam: black-list-address OR ("dollars" AND "you have been selected")

Accuracy can be high

- If rules carefully refined by expert

But building and maintaining these rules is expensive

Multinomial Naive Bayesian for LM



Classification Methods: Supervised Machine Learning

Any kind of classifier

- Naïve Bayes
- Logistic regression
- Neural networks
- k-Nearest Neighbors
- …

Multinomial Naive Bayesian for LM

Features ?

Multinomial Naive Bayesian for LM

Features ?

Tokens, PoS, NE, Embeddings, …

Multinomial Naive Bayesian for LM

Features ?

Tokens, PoS, NE, Embeddings, ...

bag of words model.

Multinomial Naive Bayesian for LM

Features ?

Tokens, PoS, NE, Embeddings, ...

bag of words model.

For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Multinomial Naive Bayesian for LM

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c) P(c)}{P(d)}$$

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

Multinomial Naive Bayesian for LM

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

x_1,...x_n are document features

Need independence assumptions to simplify computations!

Multinomial Naive Bayesian for LM

Bag of Words assumption: Assume position doesn't matter

Conditional Independence: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class *c.*

$$P(x_1,\ldots,x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Multinomial Naive Bayesian for LM

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

$$c_{\mathrm{NB}} = \underset{c_j \in C}{\operatorname{argmax}} \left[\log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i | c_j) \right]$$

Next ? Learning the models

Multinomial Naive Bayesian for LM

- simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Multinomial Naive Bayesian for LM

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

fraction of times word w_i appears among all words in documents of topic c_j

Create mega-document for topic *j* by concatenating all docs in this topic

- Use frequency of *w* in mega-document

Multinomial Naive Bayesian for LM

What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive** (***thumbs-up)***?

$$\hat{P}(\text{"fantastic"}\ |\text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

Zero probabilities cannot be conditioned away, no matter the other evidence!

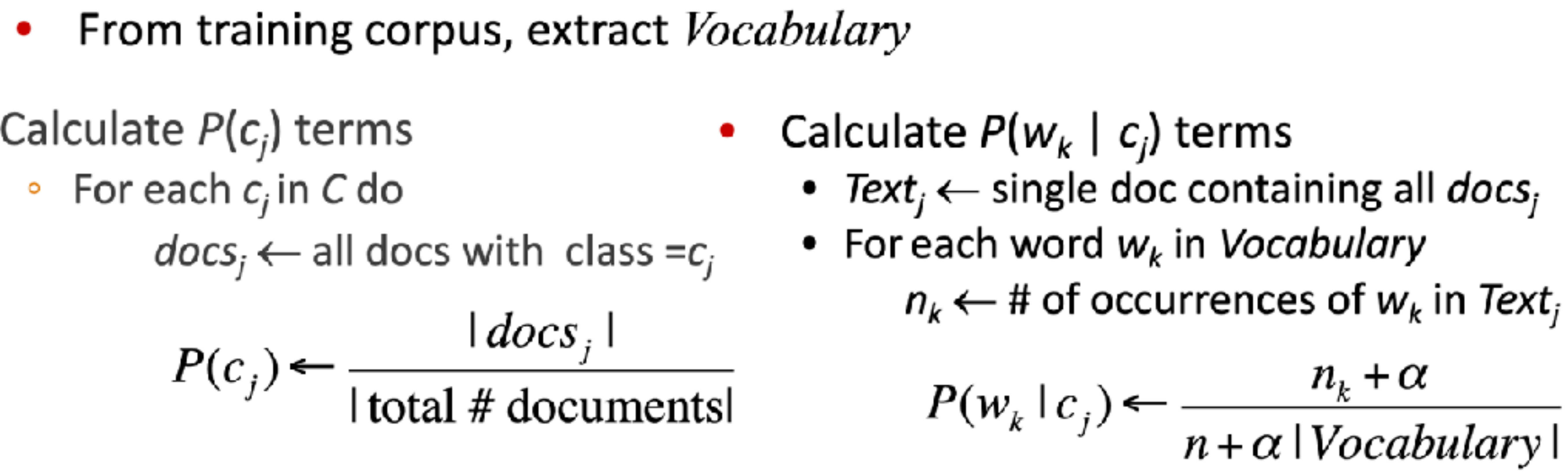
$$c_{MAP} = \text{argmax}_c \, \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Multinomial Naive Bayesian for LM

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Multinomial Naive Bayesian for LM



- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $=c_j$

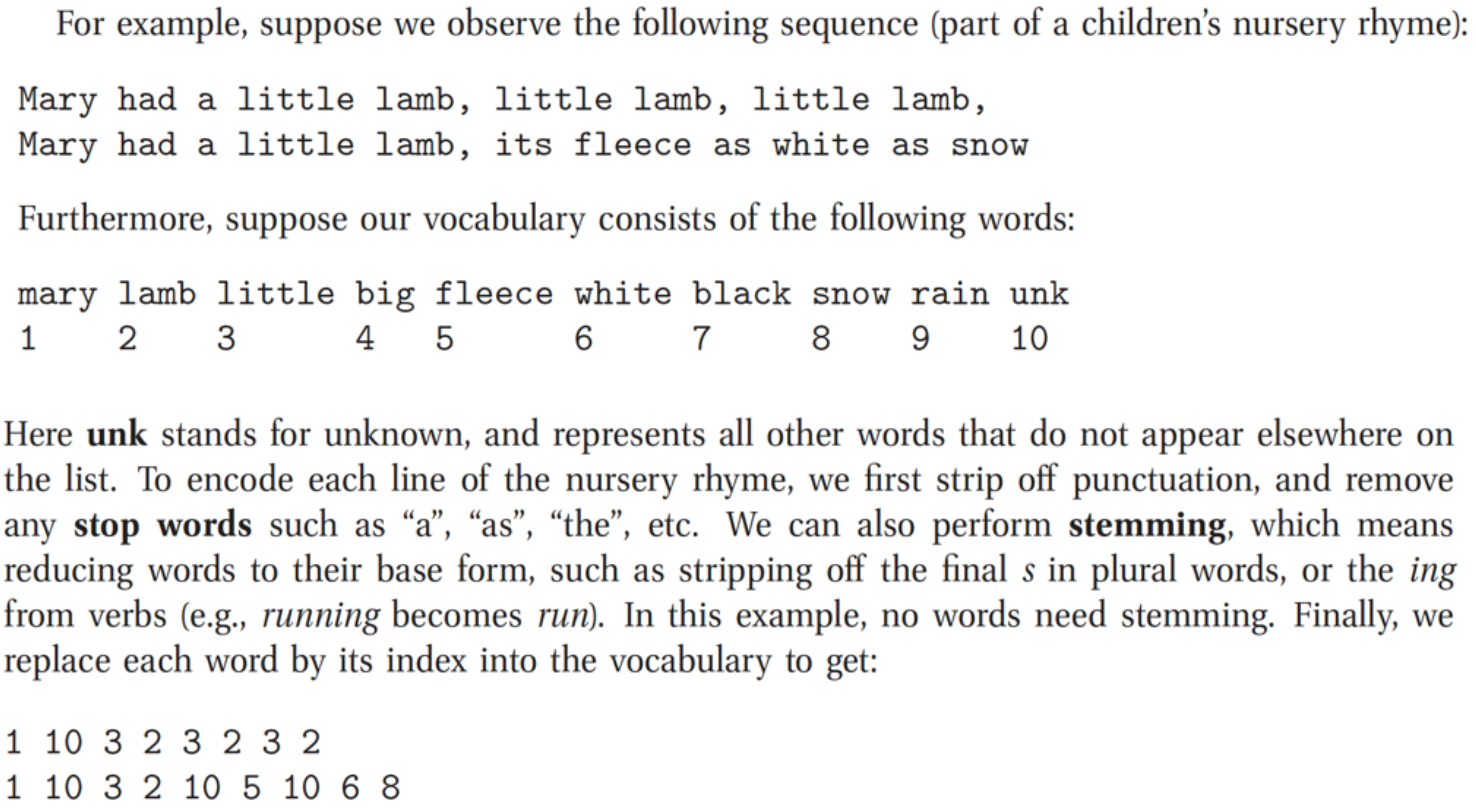
    $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

    $$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Multinomial Naive Bayesian for LM



For example, suppose we observe the following sequence (part of a children's nursery rhyme):

```
Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow
```

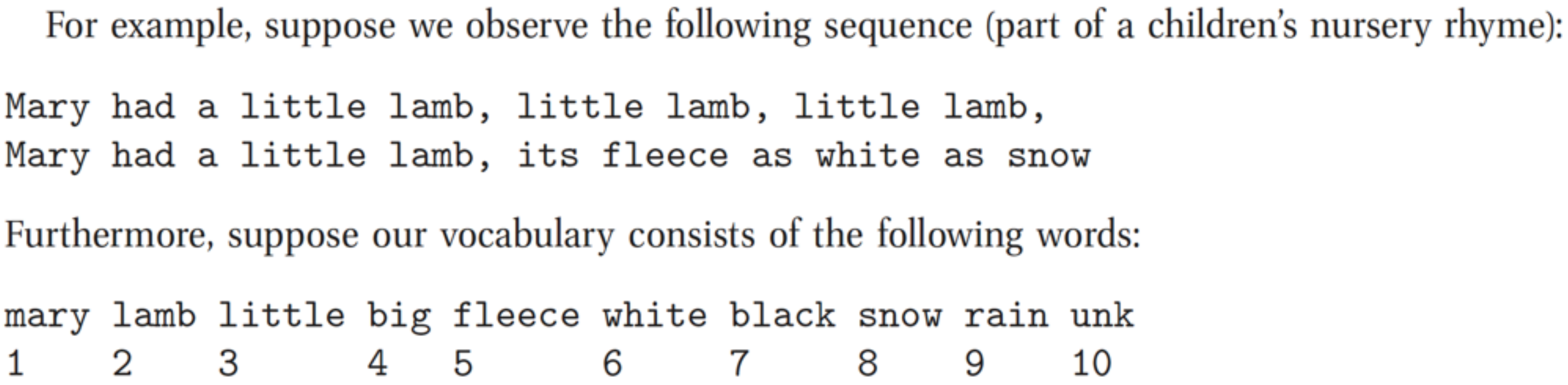
Furthermore, suppose our vocabulary consists of the following words:

mary	lamb	little	big	fleece	white	black	snow	rain	unk
1	2	3	4	5	6	7	8	9	10

Here **unk** stands for unknown, and represents all other words that do not appear elsewhere on the list. To encode each line of the nursery rhyme, we first strip off punctuation, and remove any **stop words** such as "a", "as", "the", etc. We can also perform **stemming**, which means reducing words to their base form, such as stripping off the final *s* in plural words, or the *ing* from verbs (e.g., *running* becomes *run*). In this example, no words need stemming. Finally, we replace each word by its index into the vocabulary to get:

```
1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8
```

Multinomial Naive Bayesian for LM

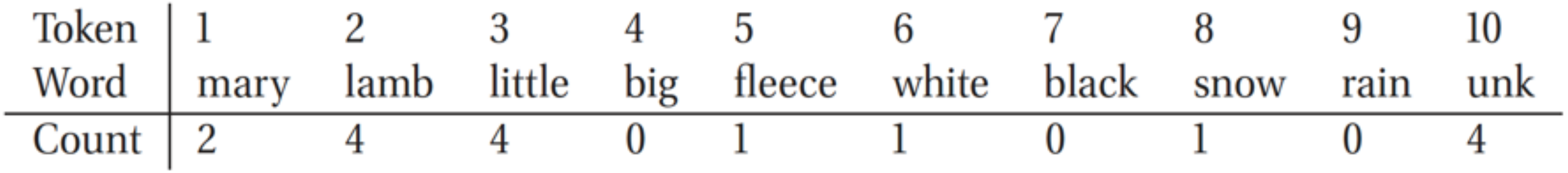


For example, suppose we observe the following sequence (part of a children's nursery rhyme):

```
Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow
```

Furthermore, suppose our vocabulary consists of the following words:

```
mary lamb little big fleece white black snow rain unk
1    2    3      4   5      6     7     8    9    10
```



Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4