Machine Learning 2020

Volker Roth

Department of Mathematics & Computer Science
University of Basel

10th March 2020

Section 2

Generative models for discrete data

Bayesian concept learning

- Consider how a child learns the meaning of the word *dog*.
- Presumably from positive examples, like *"look at the cute dog!"*
- Negative examples much less likely, *"look at that non-dog"* (?)
- Psychological research has shown that people can learn concepts from positive examples alone.
- Learning meaning of a word = concept learning = binary classification: $f(x) = 1$ if x is example of concept C, and 0 otherwise.
- Standard classification requires positive and negative examples...
  Bayesian concept learning uses positive examples alone.
- Example: the *number game*: I choose some arithmetical concept C, such as "prime number". I give you a (random) series of positive examples $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn from C.
  Question: does new $\tilde{x}$ belong to C?

The number game

- Consider integers in $[1, 100]$. I tell you 16 is a positive example. Other positive examples? Difficult with only one example...
- Intuition: numbers similar to 16 are more likely.
- But what means *similar*? 17 (close by), 6 (one digit in common), 32 (also even and a power of 2), etc.
- Represent this as a probability distribution, $p(\tilde{x}|\mathcal{D})$: probability that $\tilde{x} \in C$ given $\mathcal{D}$.
  $\rightsquigarrow$ **posterior predictive distribution.**
- After seeing $\mathcal{D} = \{16, 8, 2, 64\}$, you may guess that the concept is "powers of two".
- ...if instead I tell you $\mathcal{D} = \{16, 23, 19, 20\}$...
- How can we explain this behavior and emulate it in a machine?
- Suppose we have a **hypothesis space** of concepts, $\mathcal{H}$.

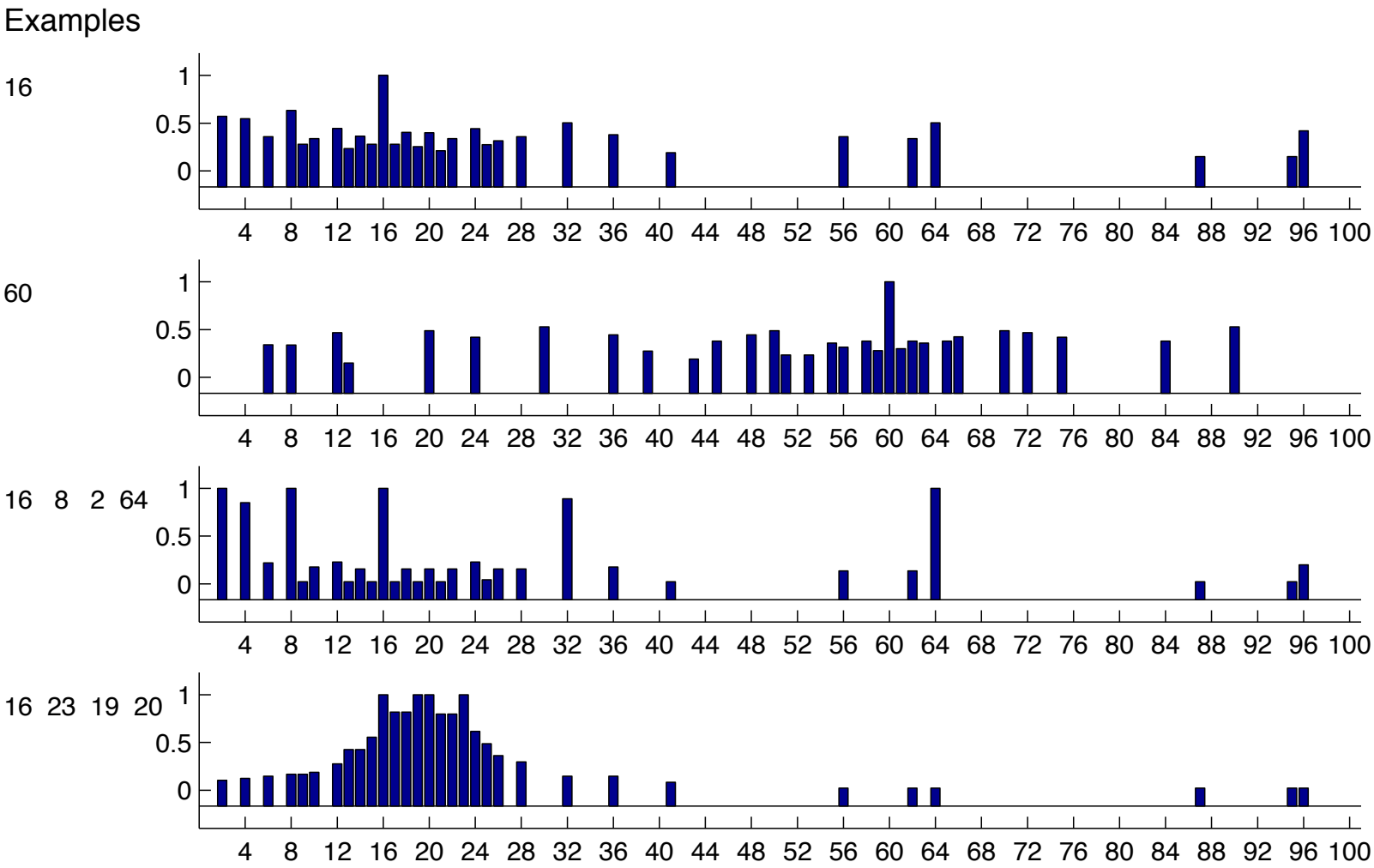


Figure 3.1 in K.Murphy. Empirical predictive distribution averaged over 8 humans in the number game. First two rows: after seeing $\mathcal{D} = \{16\}$ and $\mathcal{D} = \{60\}$. This illustrates diffuse similarity. Third row: after seeing $\mathcal{D} = \{16, 8, 2, 64\}$. This illustrates rule-like behavior (powers of 2). Bottom row: after seeing $\mathcal{D} = \{16, 23, 19, 20\} \rightsquigarrow$ focused similarity (numbers near 20)

The number game

- **Version space:** subset of $\mathcal{H}$ that is consistent with $\mathcal{D}$.
- As we see more examples, the version space shrinks and we become increasingly certain about the concept.
- But: version space is not the whole story:
  - After seeing $\mathcal{D} = \{16\}$, there are many consistent rules; how do you combine them to predict if $\tilde{x} \in C$?
  - Also, after seeing $\mathcal{D} = \{16, 8, 2, 64\}$, why did you choose the rule "powers of two" and not "all even numbers", or "powers of two except for 32", which are equally consistent with the evidence?
- **Bayesian explanation.**

The number game: Likelihood

- Having seen $\mathcal{D} = \{16, 8, 2, 64\}$, we must explain why we chose h_{two} = "powers of two", and not h_{even} = "even numbers".
- Key intuition: want to avoid suspicious coincidences. If the true concept was h_{even}, how come we only saw powers of two?
- Formalization: assume that examples are sampled uniformly at random from the extension of a concept, e.g. $h_{\text{even}} = \{2, 4, 6, \ldots, 100\}$.
- Given this assumption, the probability of independently sampling N items (with replacement) from h is $p(\mathcal{D}|h) = \left[\frac{1}{|h|}\right]^N$.
- **Size principle:** the model favors the simplest hypothesis consistent with the data. Known as **Occam's razor.**
- **William of Ockham** (1287-1347): When presented with **competing hypotheses that make the same predictions**, select the simplest one.

The number game: Likelihood

- Let $\mathcal{D} = \{16\} \rightsquigarrow p(\mathcal{D}|h_{\text{two}}) = 1/6$, since there are 6 powers of two less than 100, but $p(\mathcal{D}|h_{\text{even}}) = 1/50$, since there are 50 even numbers.
- So the likelihood that $h = h_{\text{two}}$ is higher than if $h = h_{\text{even}}$.
- After 4 examples, $p(\mathcal{D}|h_{\text{two}}) = (1/6)^4$, $p(\mathcal{D}|h_{\text{even}}) = (1/50)^4$.
- This is a **likelihood ratio** of almost 5000:1 in favor of h_{two}.
- This quantifies our earlier intuition that $\mathcal{D} = \{16, 8, 2, 64\}$ would be a very suspicious coincidence if generated by h_{even}.

The number game: Prior

- Given $\mathcal{D} = \{16, 8, 2, 64\}$, the concept

  h' = "powers of two except 32"

  is more likely than

  h = "powers of two",

  since h' does not need to explain the coincidence that 32 is missing.
- However, h' seems "conceptually unnatural".
- Capture such intuition by assigning **low prior probability** to "unnatural" concepts.
- Your prior might be different than mine, and this **subjective aspect** of Bayesian reasoning is a source of much controversy.
- But priors are actually quite useful:
  - If you are told the numbers are from some arithmetic rule, then given 1200, 1500, and 900, you may think 400 is likely but 1183 is unlikely.
  - But if you are told that the numbers are examples of healthy cholesterol levels, you would probably think 400 is unlikely and 1183 is likely.

The number game: Prior

- The prior is the mechanism to formalize background knowledge. Without this, rapid learning is impossible.
- Example: use a simple prior which puts uniform probability on 30 simple arithmetical concepts.
- To make things more interesting, we make the concepts “even” and “odd” more likely a priori.
- We also include two “unnatural” concepts, namely “powers of 2, plus 37” and “powers of 2, except 32”, but give them low prior weight.

The number game: Posterior

- The posterior is simply the likelihood times the prior, normalized:

$$p(h|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\mathcal{D}|h)p(h) = \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{I}(\mathcal{D} \in h')/|h|^N},$$

  where $\mathbb{I}(\mathcal{D} \in h) = 1$ iff the data are in extension of hypothesis h.
- After seeing $\mathcal{D} = \{16, 8, 2, 64\}$, the likelihood is much more peaked on the *powers of two* concept, so this dominates the posterior.
- In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the **MAP estimate**

$$p(h|\mathcal{D}) \to \delta_{\hat{h}^{\text{MAP}}}(h),$$

  where

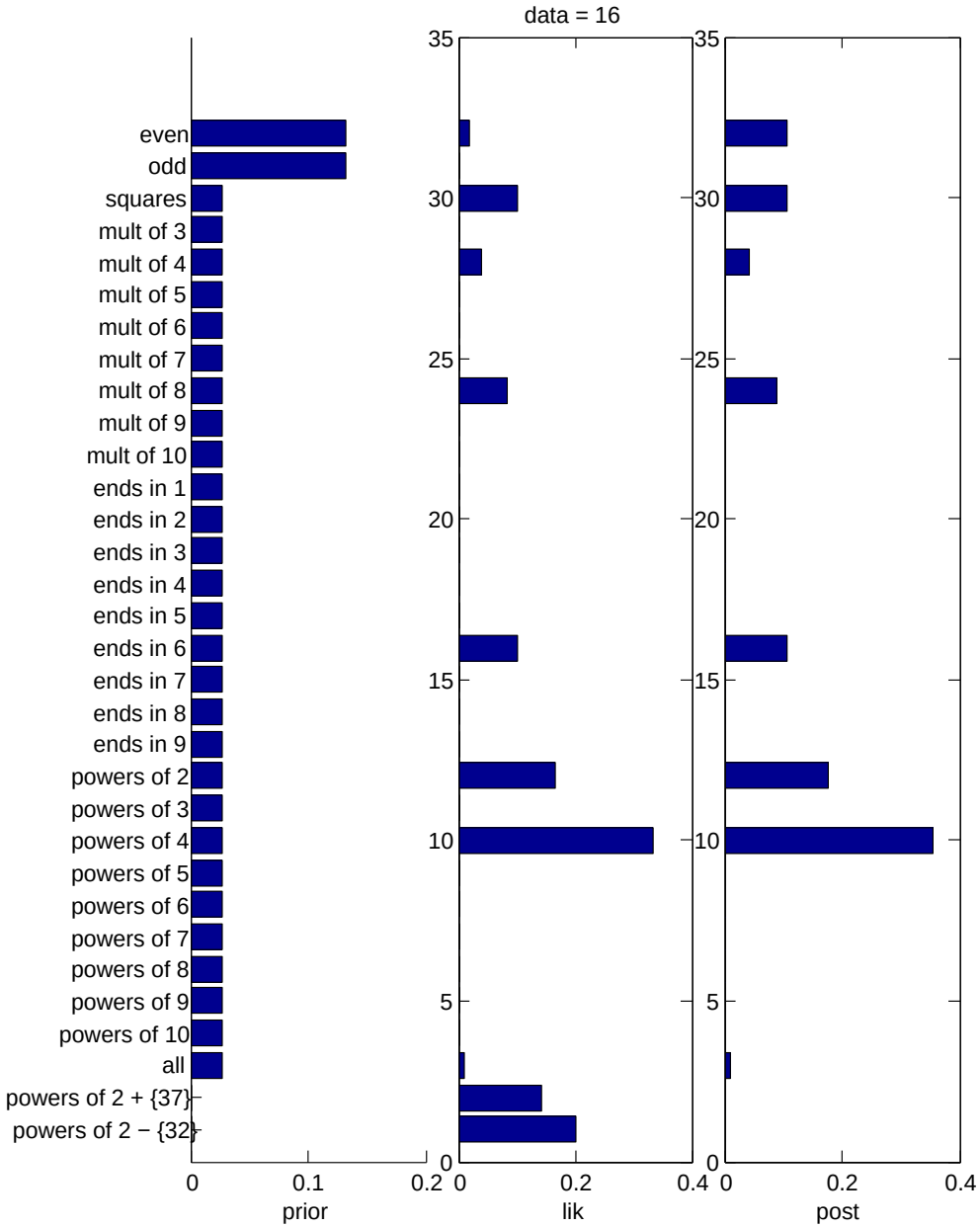
$$\hat{h}^{\text{MAP}} = \arg\max_h p(h|\mathcal{D})$$

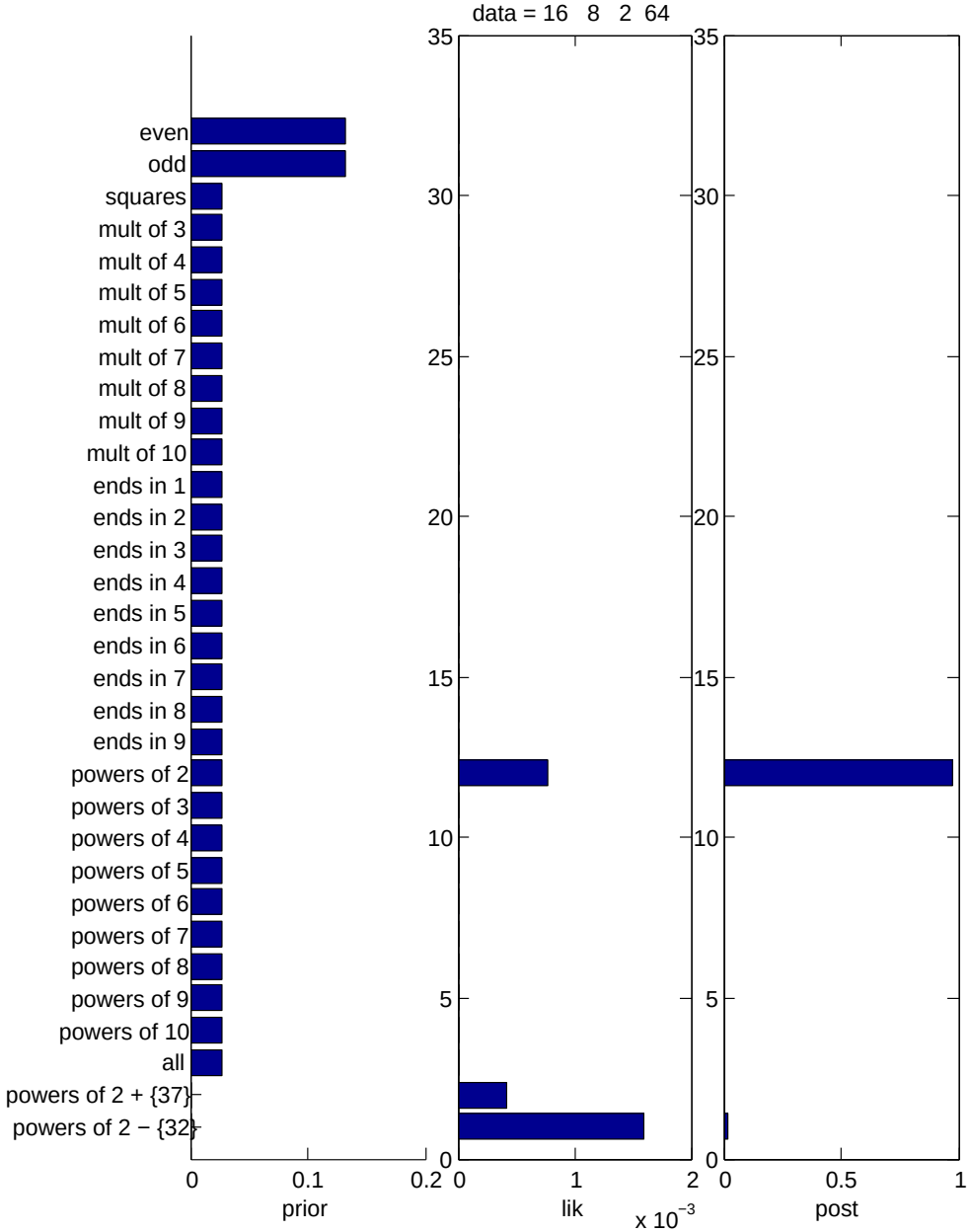
  is the posterior mode, and δ is the Dirac measure

$$\delta_x(A) = \begin{cases} 1 & \text{, if } x \in A, \\ 0 & \text{otherwise} \end{cases}$$

The number game: Posterior

- Note that the MAP estimate can be written as
$$\hat{h}^{\text{MAP}} = \arg\max_h p(h|\mathcal{D}) = \arg\max_h [\log p(\mathcal{D}|h) + \log p(h)]$$
- Likelihood depends exponentially on N, prior stays constant
$\rightsquigarrow$ as we get more data, the MAP estimate converges to the **maximum likelihood estimate** (MLE):
$$\hat{h}^{\text{MLE}} = \arg\max_h p(\mathcal{D}|h) = \arg\max_h \log p(\mathcal{D}|h).$$
$\rightsquigarrow$ **Enough data overwhelms the prior.**
- If the true hypothesis is in the hypothesis space, then the MAP/ ML estimate will converge upon this hypothesis. Thus Bayesian inference (and ML estimation) are **consistent estimators.**
- We also say that the hypothesis space is identifiable in the limit, meaning we can recover the truth in the limit of infinite data.


data = 16
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 − {32}
0
0.1
0.2
prior
0
0.2
0.4
lik
0
0.2
0.4
post
0
5
10
15
20
25
30
35


data = 16 8 2 64
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 − {32}
0
5
10
15
20
25
30
35
0
0.1
0.2
prior
0
1
2
lik
x 10^−3
0
0.5
1
post

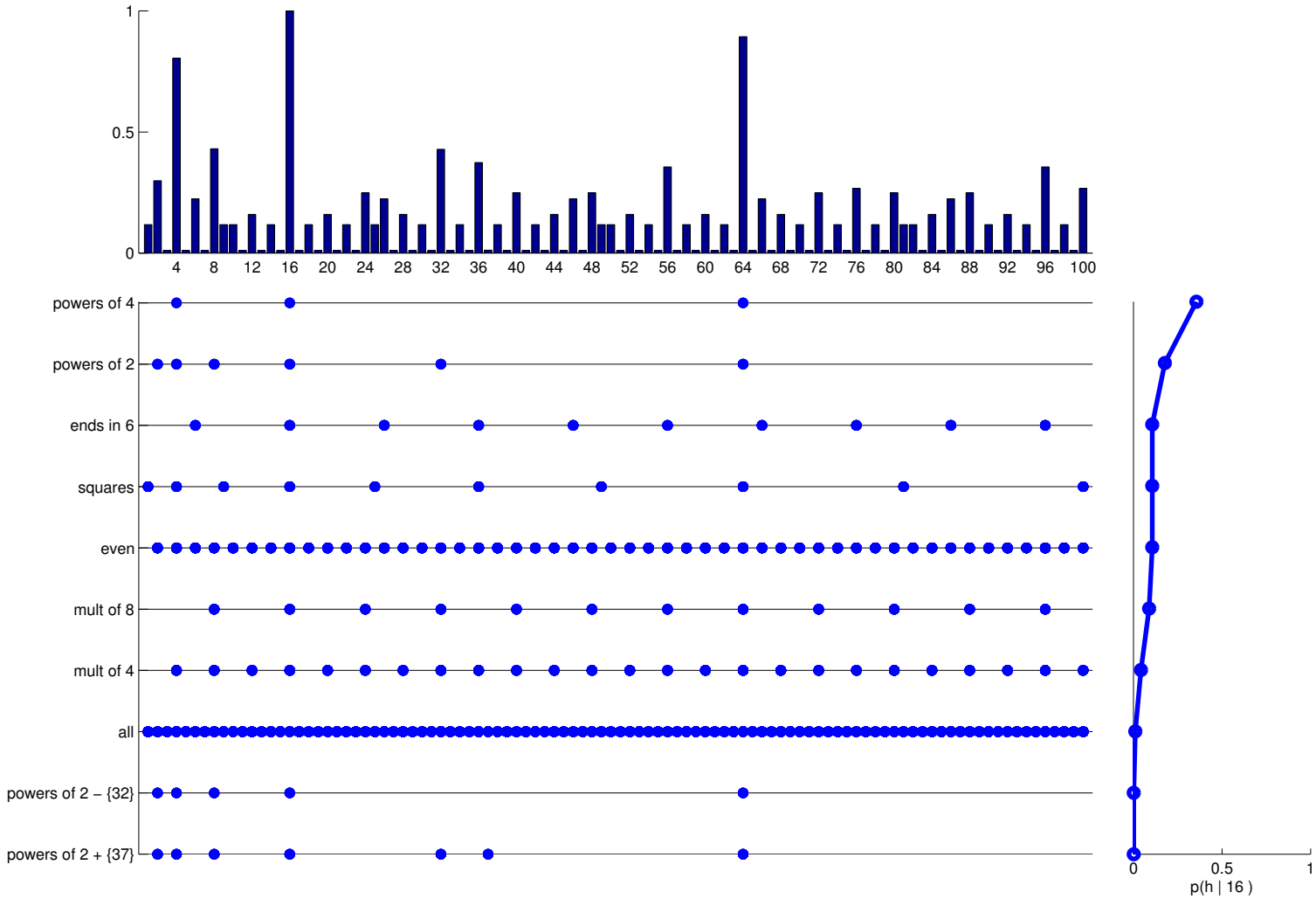


Figure 3.4 in K.Murphy. Posterior over hypotheses and predictive distribution after seeing $\mathcal{D} = \{16\}$. A dot means this number is consistent with h. Right: $p(h|\mathcal{D})$. Weighed sum of dots $\rightsquigarrow$ $p(\tilde{x} \in C|\mathcal{D})$ (top).

The number game: Posterior predictive distribution

- Posterior = internal belief state about the world.
  Test these beliefs by making predictions.
- The **posterior predictive distribution** is given by

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)p(h|\mathcal{D})$$

  $\rightsquigarrow$ weighted average of the predictions of each hypothesis
  $\rightsquigarrow$ **Bayes model averaging.**
- Small dataset $\rightsquigarrow$ vague posterior $p(h|\mathcal{D})$ $\rightsquigarrow$ broad predictive distribution.
- Once we have "figured things out", posterior becomes a delta function centered at the MAP estimate:

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)\delta_{\hat{h}^{\text{MAP}}}(h) = p(\tilde{x}|\hat{h})$$

  $\rightsquigarrow$ **Plug-in approximation.** In general, under-represents uncertainty!
- Typically, predictions by plug-in and Bayesian approach quite different for small N although they converge to same answer as $N \to \infty$.

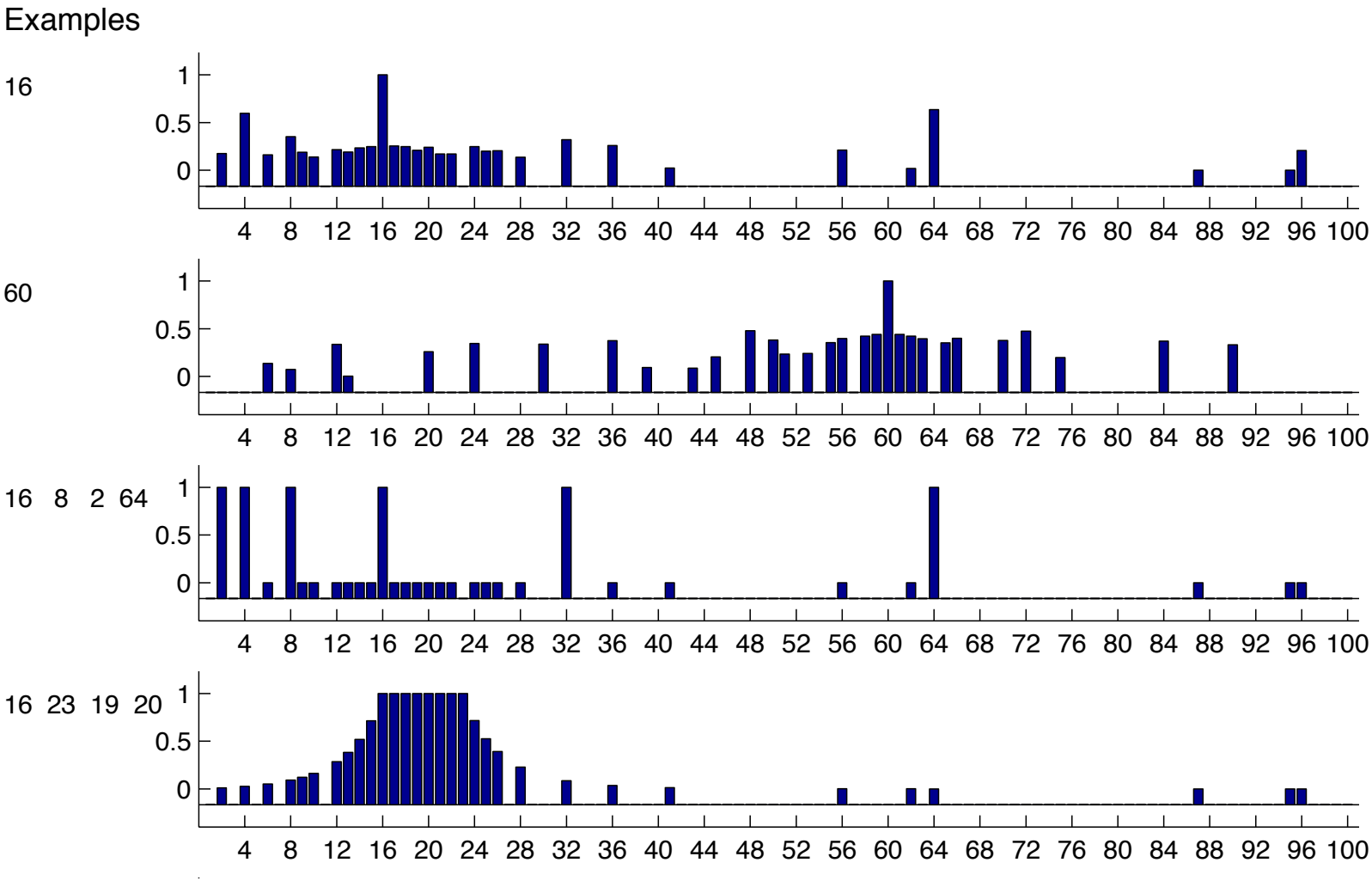


Figure 3.5 in K. Murphy. Predictive distributions for the model using the full hypothesis space.

The beta-binomial model

- Number game: inferring a distribution of a **discrete variable** drawn from a **finite hypothesis space**, $h \in \mathcal{H}$, given a **series of discrete observations.**
- This made the computations simple: just needed to sum, multiply and divide.
- Often, the K unknown parameters are **continuous**, so the hypothesis space is (some subset) of $\mathbb{R}^K$.
- This complicates mathematics (replace sums with integrals), but the basic ideas are the same.
- Example: inferring the probability that a coin shows up heads, given a series of observed coin tosses.

The beta-binomial model: Likelihood

- Suppose $X_i \sim \text{Ber}(\theta)$, where $X_i = 1$ represents "heads", and $\theta \in [0, 1]$ is the probability of heads.
- Assuming iid data, the likelihood is

$$p(\mathcal{D}|\theta) = \theta_1^N (1-\theta)_0^N, \quad N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1) \text{ heads},$$

$$N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0) \text{ tails}.$$

- $\{N_1, N_0\}$ are a **sufficient statistics of the data:** all we need to know to infer θ.
- Formally: $s(\mathcal{D})$ is a sufficient statistic for $\mathcal{D}$ if $p(\theta|\mathcal{D}) = p(\theta|s(\mathcal{D}))$.
- Two datasets with same sufficient statistics $\rightsquigarrow$ same estimated value for θ.

The beta-binomial model: Likelihood

- Suppose we observe the count of the number of heads N_1 in a fixed number $N = N_1 + N_0$ of trials: $N_1 \sim \mathrm{Bin}(N_1|N, \theta)$, where

$$\mathrm{Bin}(k, n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

- The factor $\binom{n}{k}$ is independent of θ
  $\rightsquigarrow$ **likelihood for binomial sampling = Bernoulli likelihood.**
- Any inferences we make about θ will be the same whether we observe the counts, $\mathcal{D} = (N_1, N)$, or a sequence of trials, $\mathcal{D} = \{x_1, \ldots, x_N\}$.

The beta-binomial model: Prior

- Need a prior over the interval $[0, 1]$. Would be convenient if the prior had the same form as the likelihood: $p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}$.
- Then, the posterior would be

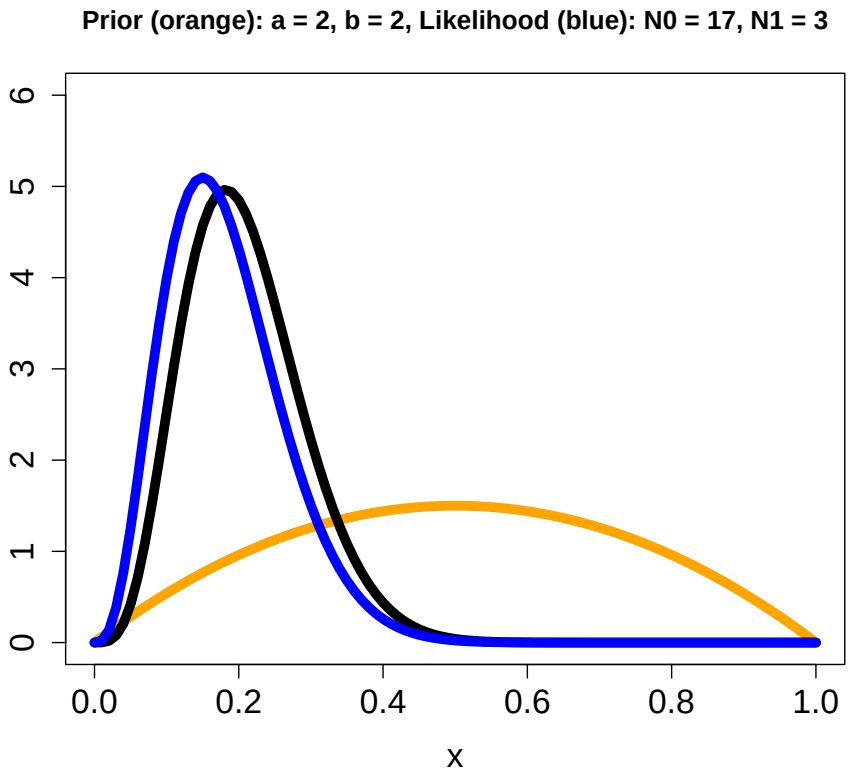
  $$p(\theta|\mathcal{D}) \propto \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_2}.$$

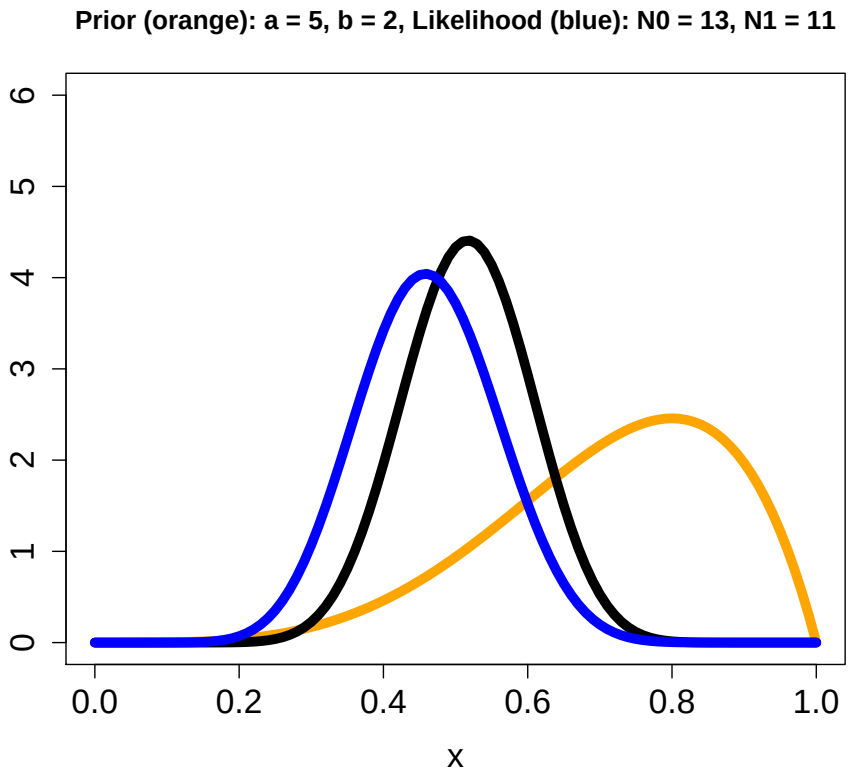
  Prior and posterior have the same form $\rightsquigarrow$ **conjugate prior.**
- In the case of the Bernoulli likelihood, the conjugate prior is the **beta distribution:**

  $$\text{Beta}(\theta|a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

- The parameters of the prior are called **hyper-parameters.**
  We can set them to **encode our prior beliefs.**
- If we know "nothing" about θ, we can use a uniform prior.
  Can be represented by a beta distribution with $a = b = 1$.

The beta-binomial model





a) Updating a Beta$(2, 2)$ prior with a Binomial likelihood with sufficient statistics $N_1 = 3, N_0 = 17$ to yield a Beta$(5, 19)$ posterior. (b) Updating a Beta$(5, 2)$ prior with a Binomial likelihood with sufficient statistics $N_1 = 11, N_0 = 13$ to yield a Beta$(16, 15)$ posterior.

The beta-binomial model: Posterior

- Multiplying with the beta prior we get the following posterior:

$$p(\theta|\mathcal{D}) \propto \text{Bin}(N_1|N,\theta)\text{Beta}(\theta|a,b) \propto \text{Beta}(\theta|N_1+a, N_0+b)$$

- Posterior is obtained by adding the prior hyper-parameters to the empirical counts
  $\rightsquigarrow$ hyper-parameters are known as **pseudo counts.**
- The strength of the prior, also known as the **equivalent sample size**, is the sum of the pseudo counts, $\alpha_0 = a + b$.
- Plays a role analogous to the data set size, $N_1 + N_0 = N$.

The beta-binomial model: Posterior predictive distribution

- So far: focus on **inference of 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 $\text{Beta}(N_1 + a, N_0 + b)$ posterior
  $\rightsquigarrow$ **posterior predictive distribution:**

$$
\begin{aligned}
p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(\tilde{x} = 1|\theta) p(\theta|\mathcal{D})\, d\theta \\
&= \int_0^1 \theta \text{Beta}(\theta|N_1 + a, N_0 + b)\, d\theta \\
&= E[\theta|\mathcal{D}] = \frac{N_1 + a}{N_1 + N_0 + a + b}
\end{aligned}
$$

Overfitting and the black swan paradox

- Suppose that we plug-in the MLE, i.e., we use $p(\tilde{x}|\mathcal{D}) \approx \text{Ber}(\tilde{x}|\hat{\theta}_{\text{MLE}})$.
- Can perform quite poorly when the sample size is small: suppose we have seen $N = 3$ tails $\rightsquigarrow \hat{\theta}_{\text{MLE}} = 0/3 = 0$
  $\rightsquigarrow$ **heads seem to be impossible.**
- This is called the **zero count problem** or sparse data problem.
- Even highly relevant in the era of "big data": think about partitioning (patient) data based on (personalized) criteria.
- Analogous to a problem in philosophy called **black swan paradox:** A black swan was a metaphor for something that could not exist.
- Bayesian solution: use a uniform prior: $a = b = 1$.
- Plugging in the posterior gives **Laplace's rule of succession**

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

Justifies common practice of adding 1 to empirical counts.

The Dirichlet-multinomial model

- So far: inferring the probability that a coin comes up heads.
- Generalization: probability that a die with K sides comes up as face k.
- Likelihood: observe N dice rolls $\mathcal{D} = \{x_1, \ldots, x_N\}$, $x_i \in \{1, \ldots, K\}$.

  $$\text{iid assumption} \;\rightsquigarrow\; p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k},$$

  where N_k is the number of times event k occurred (these are the sufficient statistics for this model).
- Prior: $\boldsymbol{\theta}$ lives in K-dim probability simplex. Conjugate prior with this property: **Dirichlet distribution**

  $$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_i - 1}.$$

The Dirichlet-multinomial model

- Posterior:

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

- Note that we (again) add pseudo-counts α_k to empirical counts N_k.

The Dirichlet-multinomial model

- Posterior predictive:

$$
\begin{aligned}
p(\tilde{X} = j|\mathcal{D}) &= \int p(\tilde{X} = j|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta} \\
&= \int p(\tilde{X} = j|\theta_j) \left[\int p(\boldsymbol{\theta}_{-j}\theta_j|\mathcal{D})\, d\boldsymbol{\theta}_{-j} \right] d\theta_j \\
&= \int \theta_j p(\theta_j|\mathcal{D})\, d\theta_j \\
&= E[\theta_j|\mathcal{D}] = \frac{N_j + a_j}{\sum_k (N_k + a_k)}
\end{aligned}
$$

- Note: This **Bayesian smoothing** avoids the zero-count problem. Even more important in the multinomial case, since we partition the data into many categories.

Example: Simple language model

- Goal: predict which words might occur next in a sequence.
- **Bag of words model:** assume that i'th word $X_i \in \{1, \ldots, K\}$ is sampled independently from other words using $\text{Cat}(\theta)$ distribution.
- Suppose we observe the following sequence (children's nursery rhyme)

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

- 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
```

unk stands for unknown (all other words)

- Standard procedure: strip off punctuation, and remove any stop words such as "a", "as", "the", etc.

Example: Simple language model

- 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 6 8

- Count how often each word occurred $\rightsquigarrow$ histogram of word counts:

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	3

Example: Simple language model

- Denote above counts by N_j, use a Dir($\boldsymbol{\alpha}$) prior
  $\rightsquigarrow$ posterior predictive

$$p(\tilde{X} = j|\mathcal{D}) = E[\theta_j|\mathcal{D}] = \frac{N_j + a_j}{\sum_k (N_k + a_k)}$$

  If we set $\alpha_j = 1$, we get

$$p(\tilde{X} = j|\mathcal{D}) = \left(\frac{3}{27}, \frac{5}{27}, \frac{5}{27}, \frac{1}{27}, \frac{2}{27}, \frac{2}{27}, \frac{1}{27}, \frac{2}{27}, \frac{1}{27}, \frac{5}{27}\right)$$

- Peaks at $X = 2$ ("lamb"), $X = 3$ ("little") and $X = 10$ ("unk").
- Note that the words "big", "black" and "rain" are predicted to occur with non-zero probability in the future, even though they have never been seen before.