Artificial Intelligence
EECS 491

Generative Models for Data

Generative Bayesian models

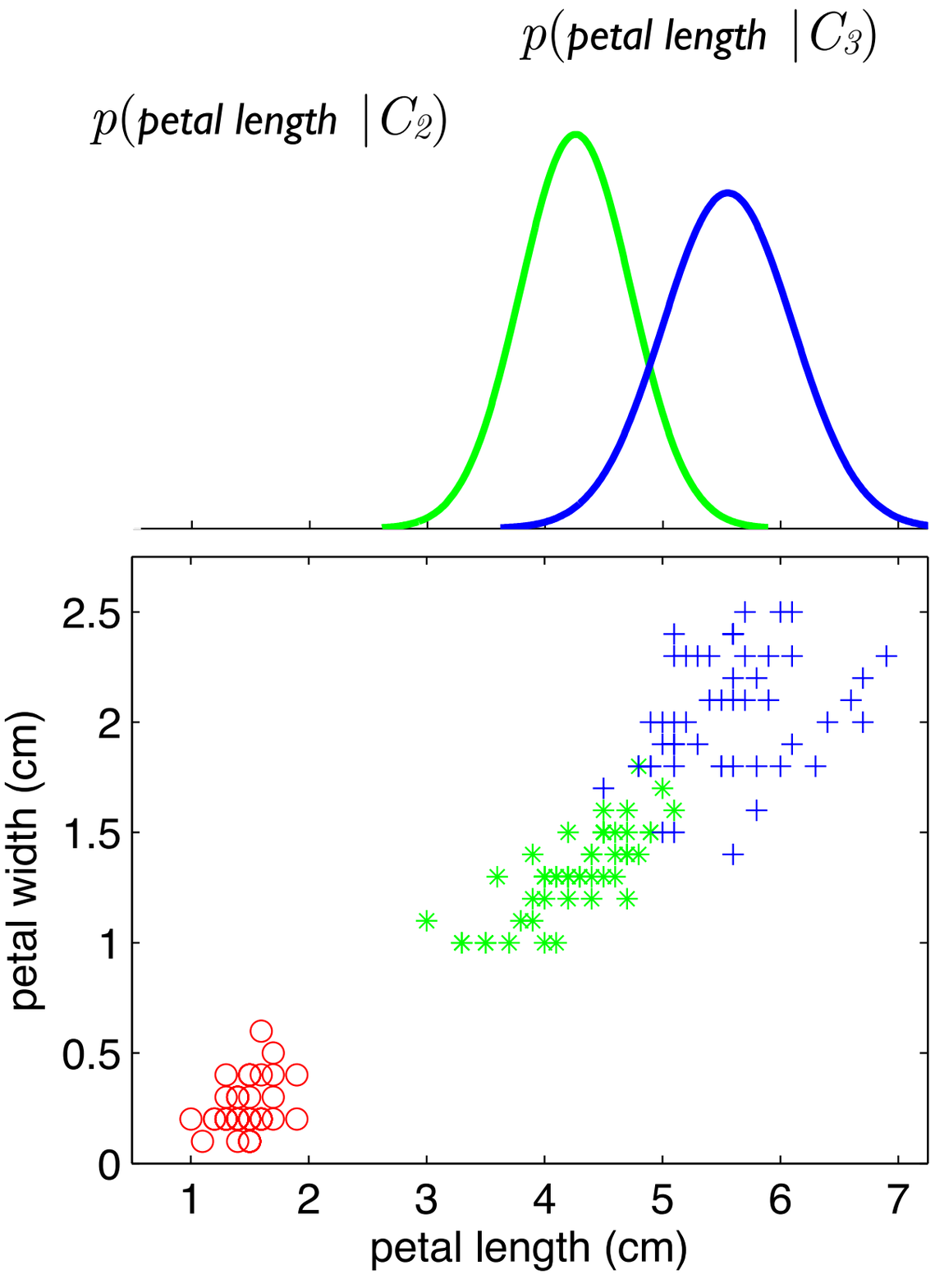
- In the Iris data example we used Bayes rule to a generative classifier

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} \propto p(\mathbf{x}|C_k)p(C_k)$$

- The key to generative models is how they describe the data
- The likelihood models the data

$$p(\mathbf{x}|C_k)$$

- It determines how well the model *generates* data it expects to see.
- Here, the data have two dimensions.
- What if the data were **symbols**?



The numbers game (Tenenbaum, 1999)

- Suppose you are given a simple arithmetic concept, e.g.
  - a prime number
  - numbers between 1 and 10
- You see a series of randomly chosen positive examples of the concept, C

$$\mathcal{D} = \{x_1, \ldots, x_N\}$$

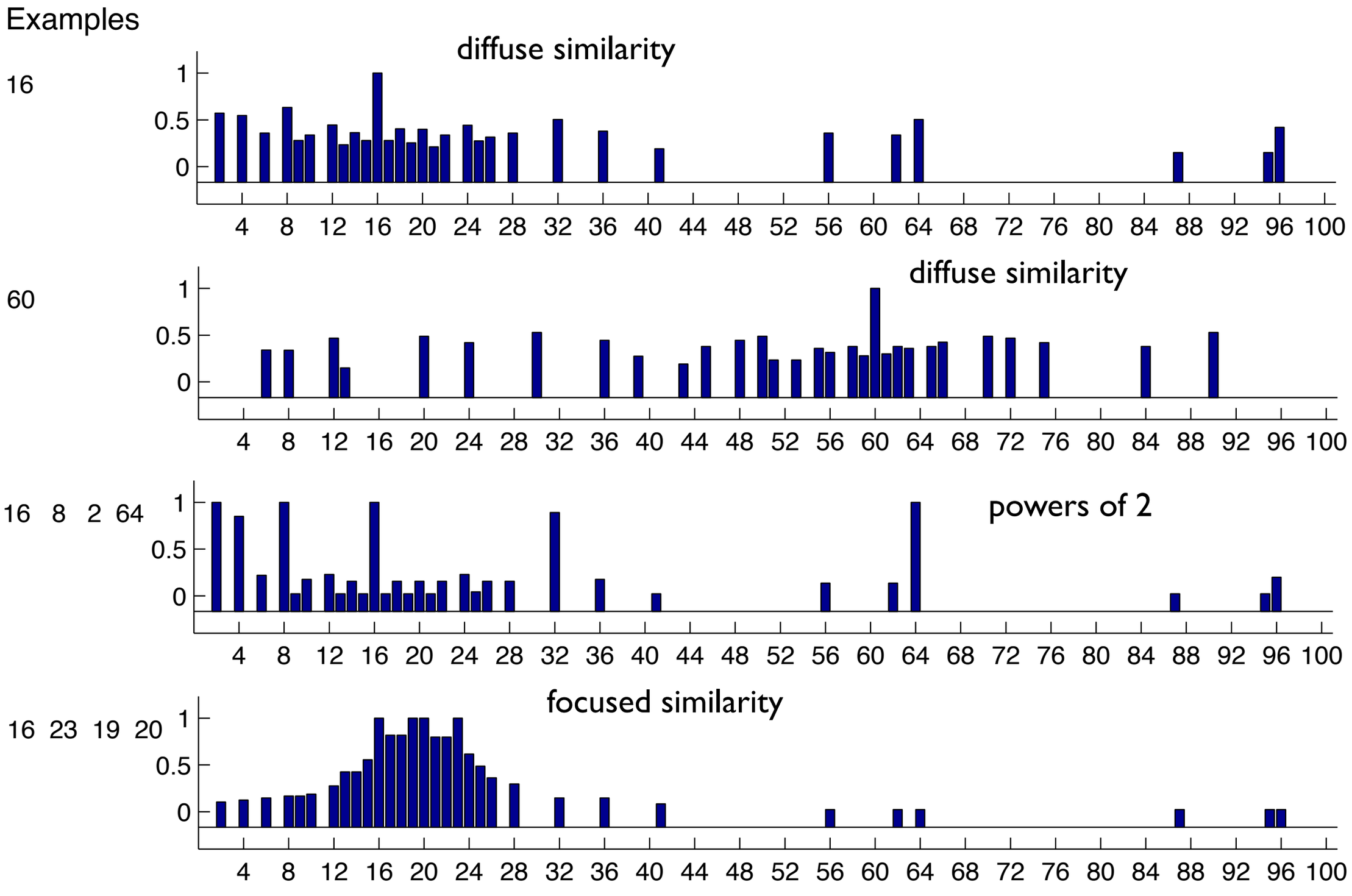
- Now you see a new test case.
- Does it belong to C or not?

Examples

- For simplicity, suppose all numbers are between 0 and 100.
- Here are some examples from the concept
  - 16: 20? 32? 91?
  - 60: 17? 72? 58?
- What about these?
  - 16, 8, 2, 64: 12? 41? 32?
  - 16, 23, 19, 20: 17? 47? 21?

- Notice that your beliefs change depending on what numbers follow.
- You do have some internal hypotheses about the sequence.

Empirical predictive distribution averaged over 8 subjects



Representing concepts

- How can we model this?
  - observed data are the numbers
  - unknown hidden causes or concepts
  - need p(**x** | D)
- Need a hypothesis space:
  - odd numbers
  - even numbers
  - powers of two
  - all numbers ending in j
  - and so on
- But how do you choose among them? Many could consistent with the data.
- Different people (with different experience) would have different priors.

Bayesian Occam's razor

- 16, 8, 2, 64, ?
- Why to we choose *h1* = "powers of 2" rather than *h2* = "even numbers" ?
- Key: want to avoid suspicious coincidences -- they're improbable
  - How is it we happened to observe even numbers that were also powers of 2?
- We also need priors.
- Assume numbers are uniformly sampled from the hypothesis space.
  - e.g. "even numbers" are sampled evenly from {2, 4, 6, ..., 98, 100}
- The probability of independently sampling N items from *h* with replacement is:

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^n = \left[\frac{1}{|h|}\right]^n$$

- This embodies the size principle:
  - choose the simplest (or smallest) model that is consistent with the data

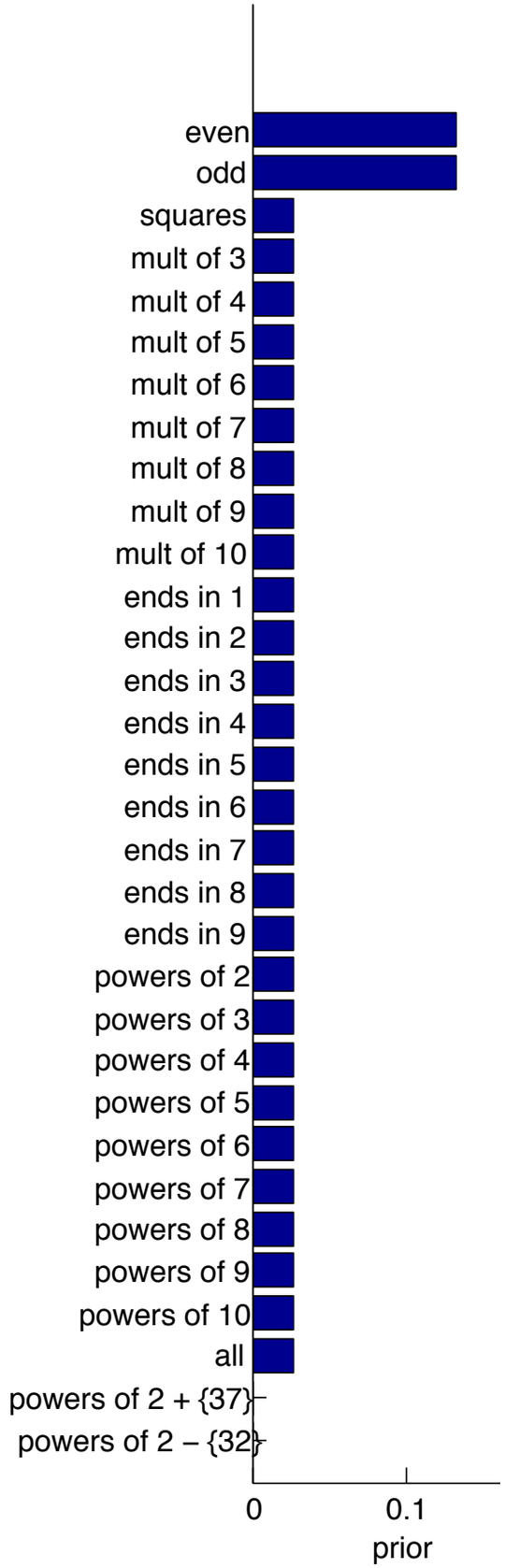
Bayesian Occam's razor

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^n = \left[\frac{1}{|h|}\right]^n$$

- Suppose D={16}
  - powers of two: 2, 4, 8,16, 32, 64
  - p(D|h = powers of two) = 1/6
  - even numbers: 2, 4, 6, 8, 10, …, 100
  - p(D|h = even numbers) = 1/50
- Now we observe D={16, 8, 2, 64}
  - p(D|h = powers of two) = $(1/6)^4 = 1/(7.7 \times 10^{4})$
  - p(D|h = even numbers) = $(1/50)^4 = 1/(1.6 \times 10^{7})$
- The likelihood ratio is 5000:1 in favor of powers of two.
- If it were "h = even", {16, 8, 2, 64} would be a very suspicious coincidence.

Subjective priors

- What about h3 = “powers of two except 32” ?
  - Couldn’t you have anything?
  - Random numbers?
- All you can do is use the data to weight among your prior hypotheses
- Random could be a default hypothesis
  - others provide more probable explanations for non-random sequences
- Murphy’s example uses 30 subjective priors that are simple arithmetic concepts.
- Different people (or situations) could assign different weights.
  - I tell you they are from a simple arithmetic rule
  - I tell you they are cholesterol levels



The posterior distribution over concepts

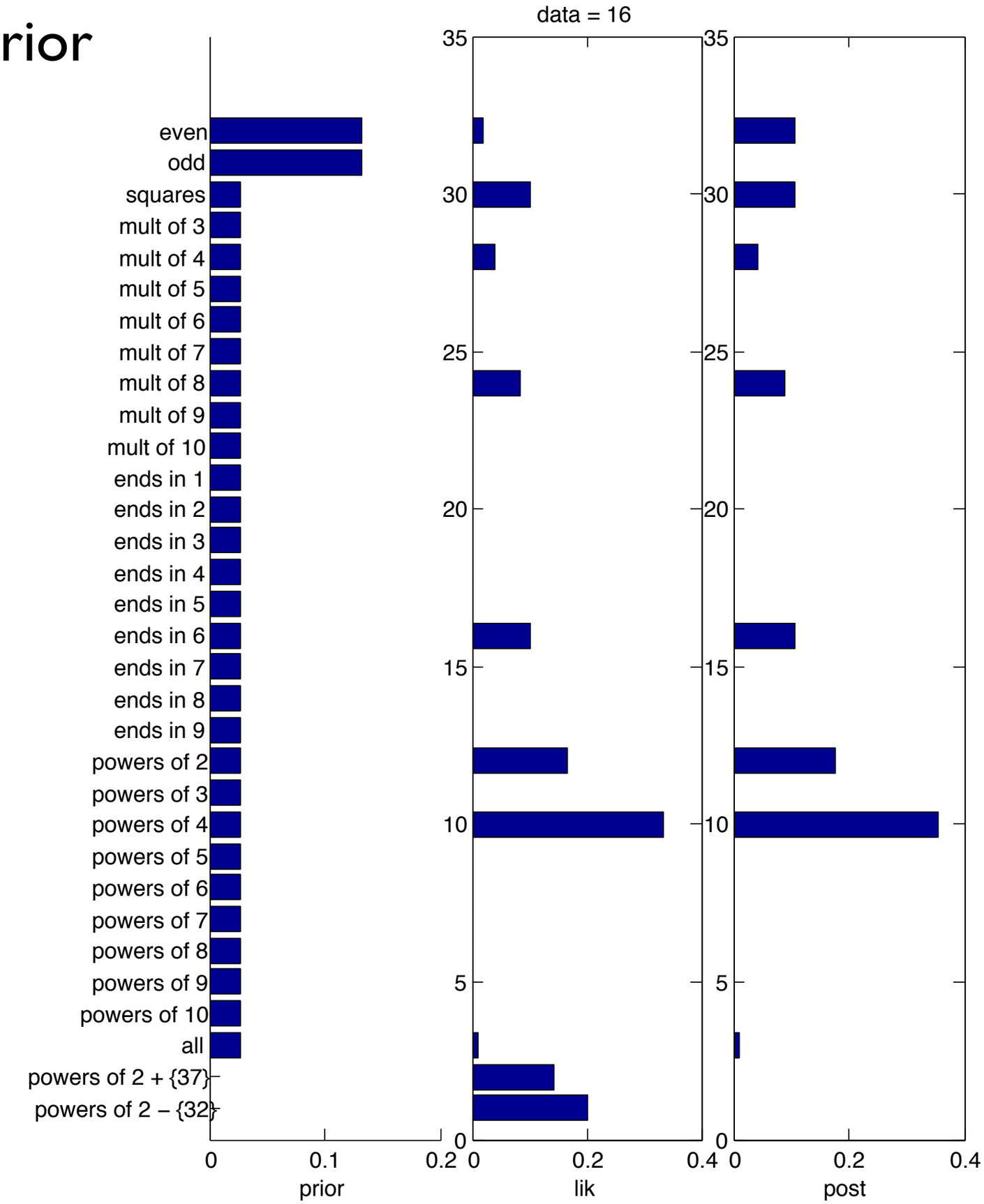
- The posterior is the prior times the likelihood normalized

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

- where $\mathbb{I}(\mathcal{D} \in h)$ is the indicator function is one iff all the data are in hyp *h*.
- When there is enough data, the p(h|D) becomes peaked on a single concept
  - the data overwhelms the prior
  - it converges to the maximum likelihood estimate (MLE)
  - it converges to the truth (if contained in the hypothesis space)
  - or the closest to the truth (if it is not)

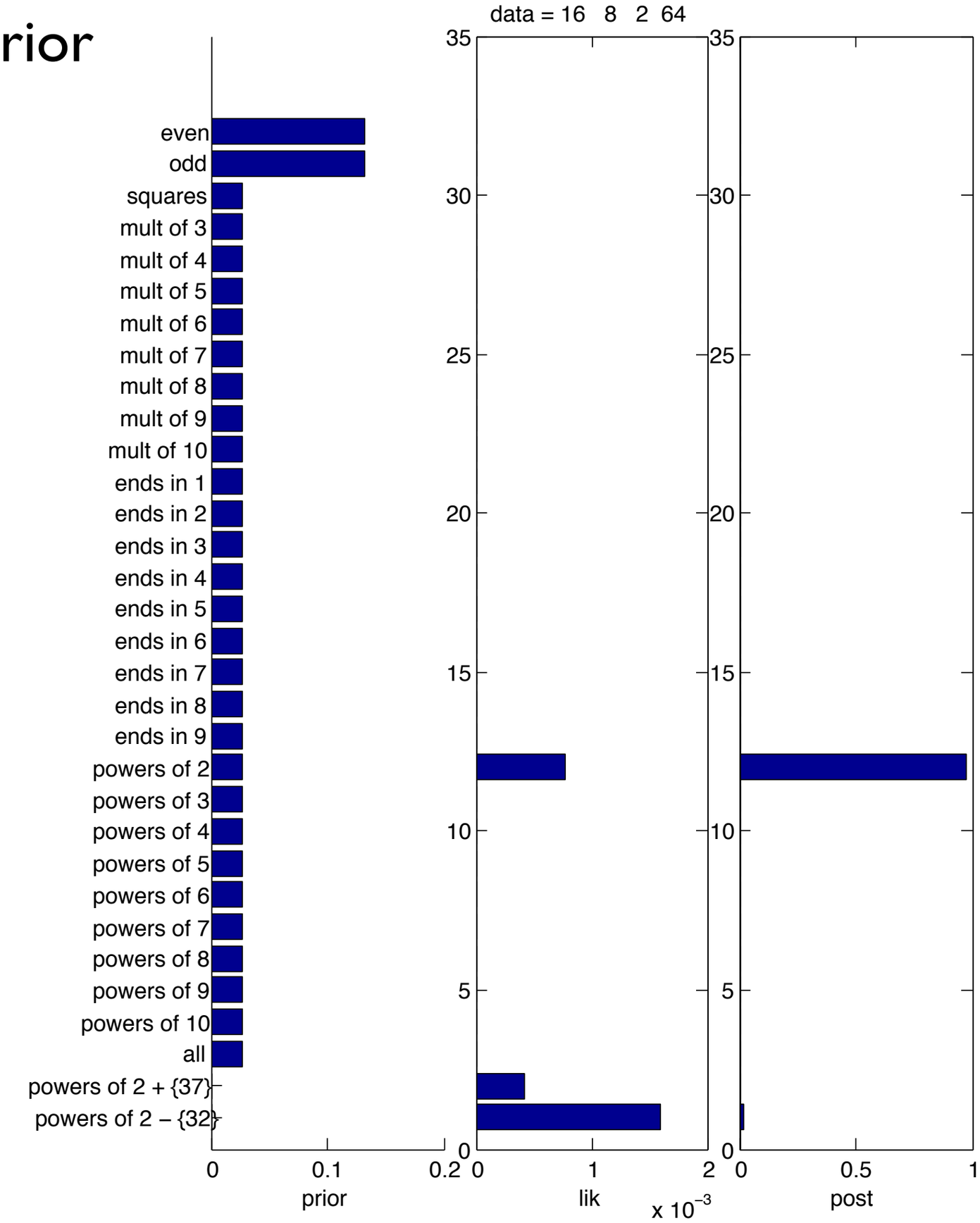
Prior, likelihood, and posterior

- data = 16



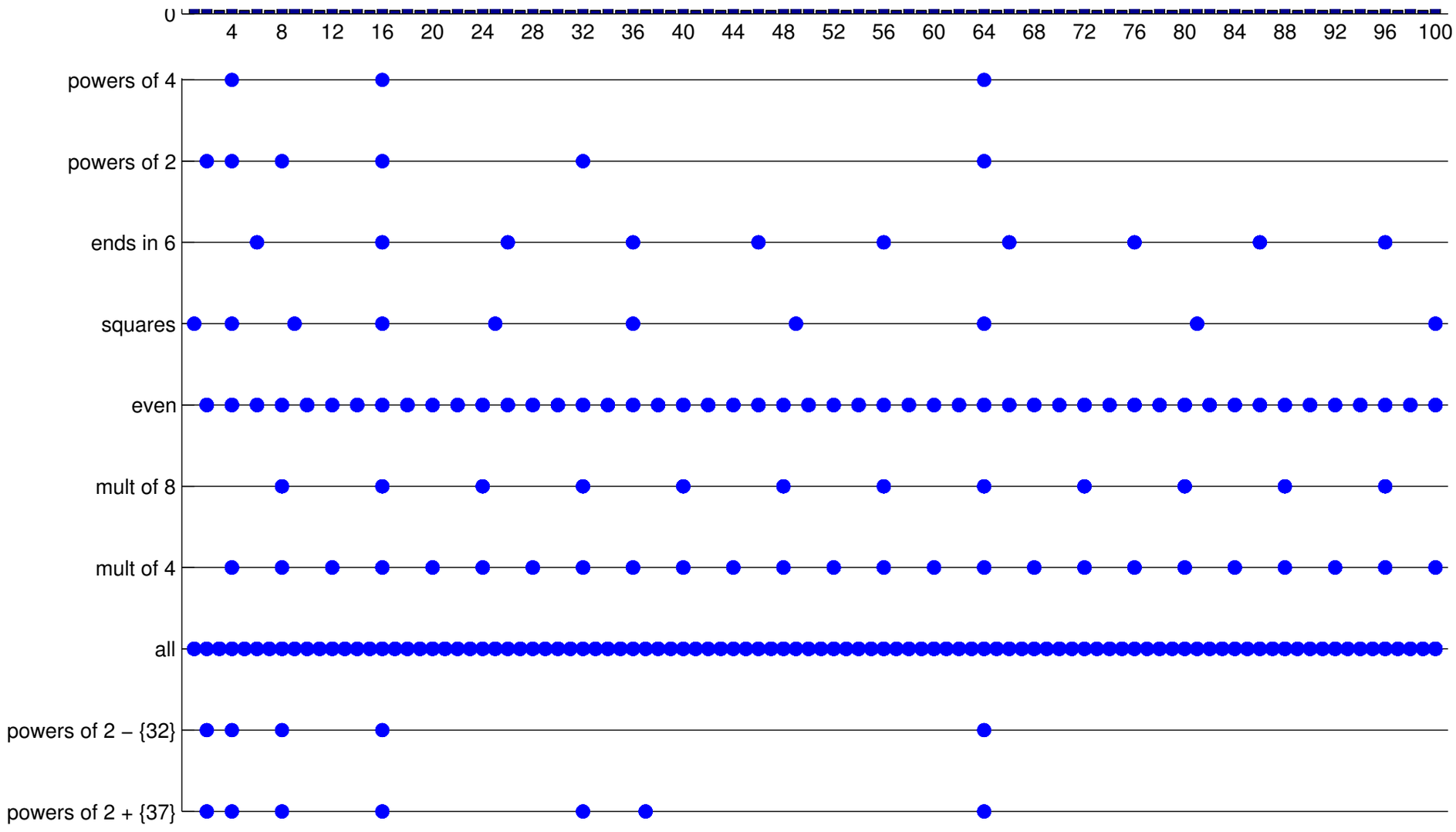
Prior, likelihood, and posterior

- data = 16, 8, 2, 64



Given our observations can we predict the next number?

- What probabilities would you assign?
- How?
- Different hypotheses make different predictions



Bayesian model comparison

- Let M_K represents a model with K classes. (Here we will assume that we can choose the best among all such models, but this assumption is not necessary). How do we evaluate the probability of model M_K? Bayes rule again.

- We start with our existing model, but add a term to represent the model itself. Also, we marginalize out the dependency on the parameters, because we want the result independent of any specific value. Letting $\mathcal{X} = x^{(1:\mathrm{N})}$, we have

$$p(M_K|\mathcal{X}) = \frac{p(\mathcal{X}|M_K)p(M_K)}{p(\mathcal{X})}$$

- The denominator is constant across models, and if we assume all models are equally probable a priori, the only data-dependent term is $\mathcal{X} = x^{(1:\mathrm{N})}$.

- How do we compute $p(\mathcal{X}|M_K)$? We've encountered this before.

Evaluating the model evidence

- $p(\mathcal{X}|M_K)$ is just the normalizing constant for the posterior for parameters

$$p(\theta_K|\mathcal{X}, M_K) = \frac{p(\mathcal{X}|\theta_K, M_K)p(M_K)}{p(\mathcal{X}|M_K)}$$

- (slight change of notation: θ_K represents all parameters for model K.)

- The normalizing constant here is evaluated just like before by marginalization

$$p(\mathcal{X}|M_K) = \int p(\mathcal{X}|\theta_K, M_K)p(M_K)d\theta_K$$

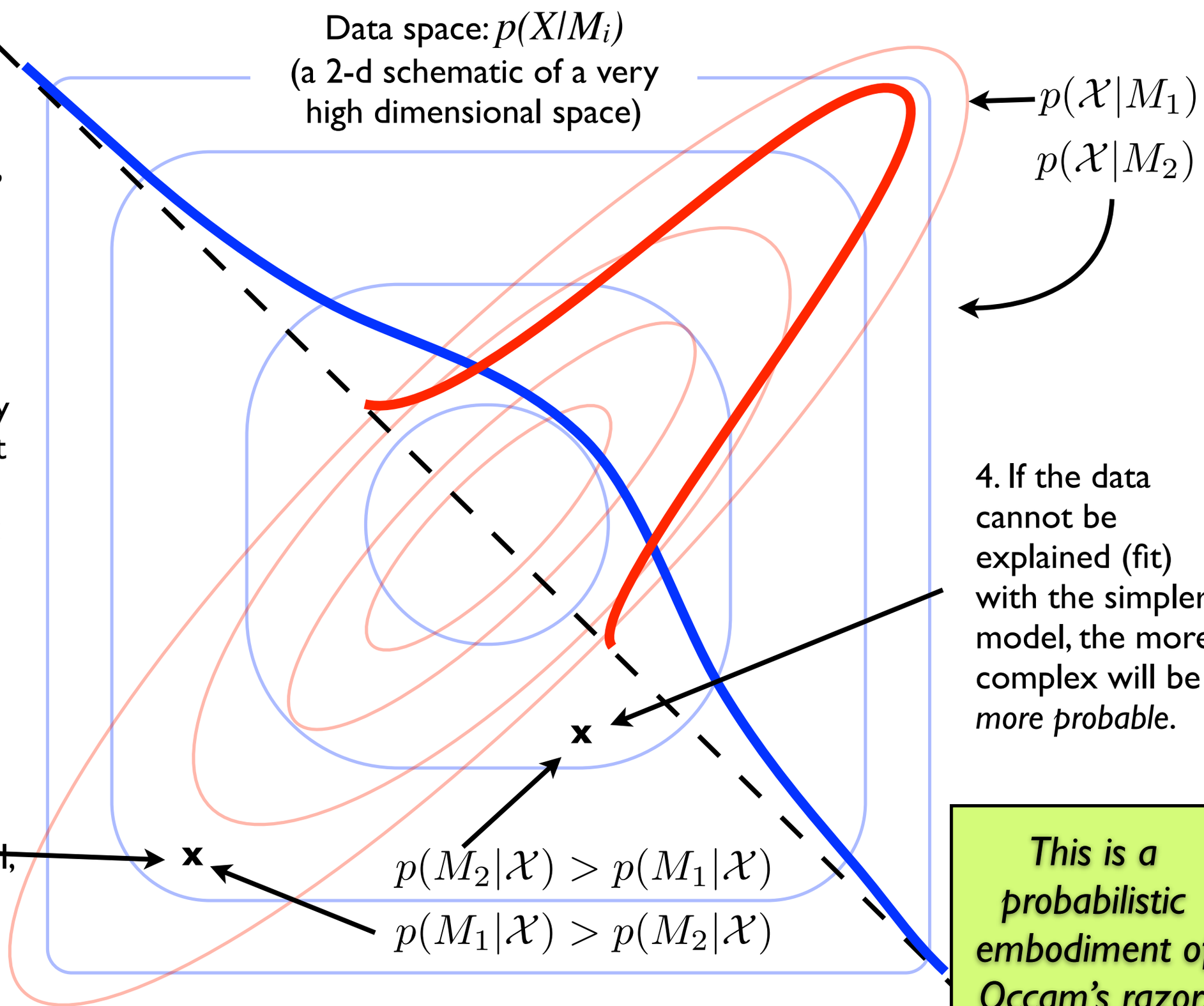
- Evaluating this term is practically a whole subfield of probability theory: Laplace's method, monte carlo integration, variational approximation, etc.

Schematic of Bayesian model comparison

1. M2 is has more degrees of freedom than M1, and therefore a larger set of data as higher probability

2. Both probability distributions must integrate to one, so more complex models are necessarily more spread out

3. If the data falls within the plausible range of the simpler model, it will be *more probable*.



4. If the data cannot be explained (fit) with the simpler model, the more complex will be *more probable*.

This is a probabilistic embodiment of Occam's razor.

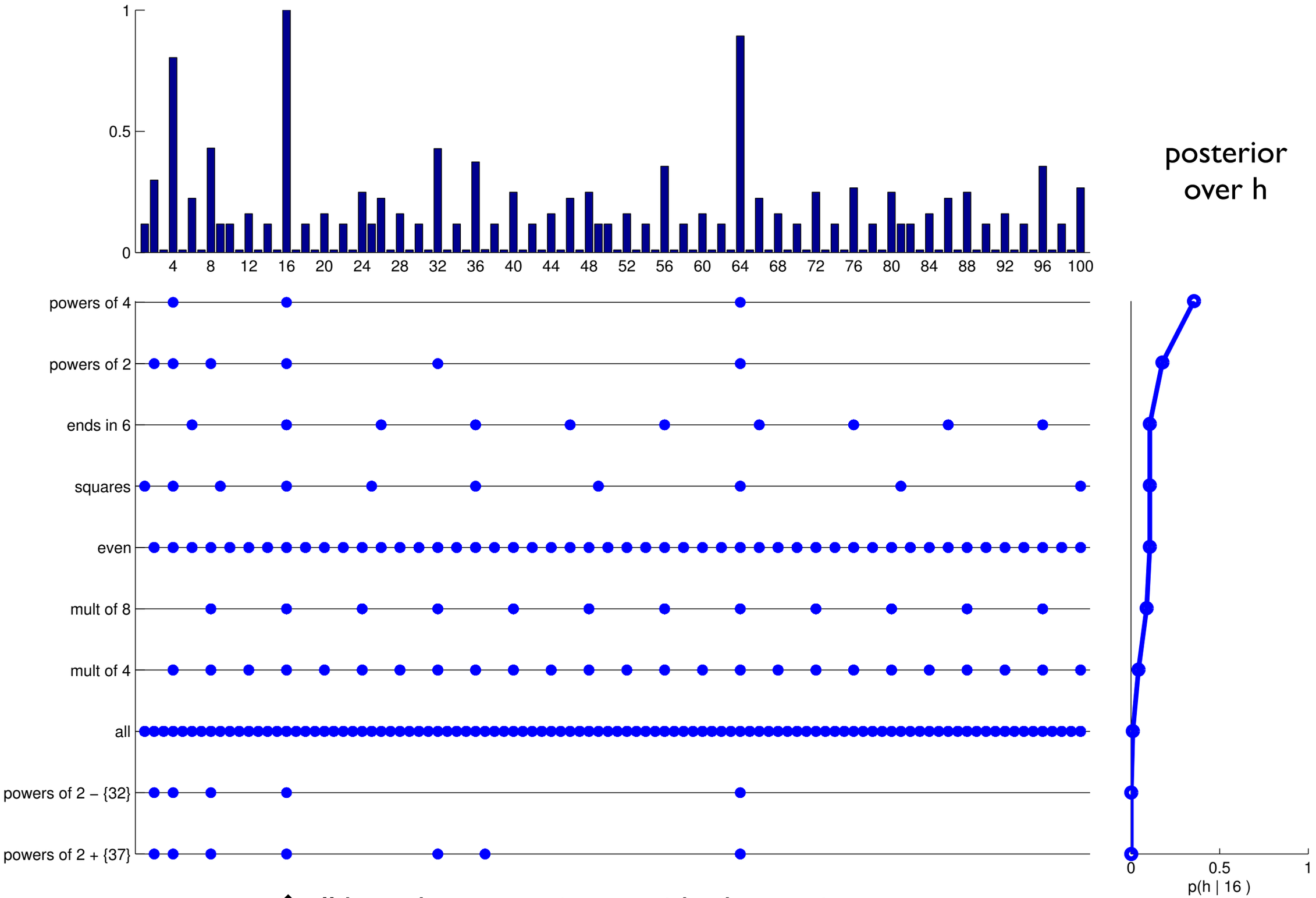
The posterior predictive distribution

- The posterior predictive marginalizes out the hypothesis.
- It's a weighted average of the predictions of each hypothesis

$$p(y = 1|\tilde{x}, \mathcal{D}) = \sum_h p(y = 1|\tilde{x}, h)p(h|\mathcal{D})$$

- This is also called Bayesian model averaging.

Posterior predictive distribution given 16



↑ all hypotheses consistent with observation

The Beta-Binomial for the coin flip example

- The general expression for the posterior distribution is given by

$$p(\theta|y,n) = \frac{p(y|\theta,n)p(\theta|n)}{p(y|n)}$$

- In the coin-flipping example, the probability of heads (or tails) is governed by the Bernoulli distribution

$$X_i \sim \text{Ber}(\theta), \text{ so } X_i \in \{0,1\}.$$

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

where we have $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ tails.

- the conjugate prior for the Bernoulli is the Beta distribution:

$$\text{Beta}(\theta|\alpha_1,\alpha_2) \propto \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

- therefore the posterior distribution has the same form:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &\propto [\theta^{N_1}(1-\theta)^{N_2}][\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}] \\ &= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1} \\ &\propto \text{Beta}(\theta|N_1+\alpha_1, N_2+\alpha_2) \end{aligned}$$

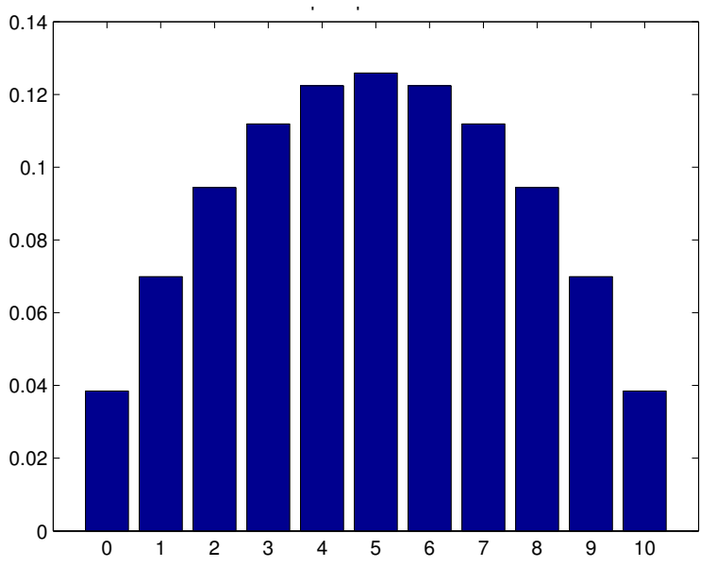
Posterior predictive distribution for the Binomial problem

- The posterior predictive density for a single future trial is

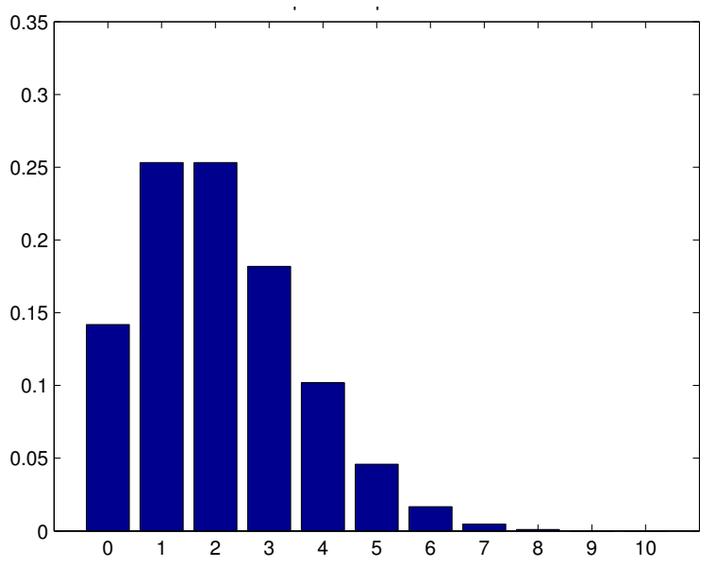
$$
\begin{aligned}
p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(x = 1|\theta) p(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta \, \mathrm{Beta}(\theta|\alpha_1', \alpha_2') d\theta = \mathbb{E}\left[\theta|\mathcal{D}\right] = \frac{\alpha_1'}{\alpha_1' + \alpha_2'}
\end{aligned}
$$

- This is a weighted average of the hypotheses over the posterior.

prior predictive Be(2,2), 10 trials



posterior predictive (for 10 trials)
after 3 heads, 17 tails



Language modeling: bag of words

- Assume the ith word, Xi in {1, ..., K} is sampled independently of the others.
- Given the previous words, can we predict what word comes next?
- Given
  - Mary had a little lamb, little lamb, little lamb
    Mary had a little lamb, its fleece was white as snow
- Suppose our vocabulary is:
  mary lamb little big fleece white black snow rain *unknown*
  1 2 3 4 5 6 7 8 9 10
- To encode the rhyme, we "normalize" it and replace each word by its index:
  - strip words like: a, as, the, etc
  - perform stemming, *raining* becomes *rain*.
  - Encoded data:
    1 10 3 2 3 2 3 2
    1 10 3 2 10 5 6 8
- These counts can modeled this with a Dirichlet-multinomial model.
  Only the word count matters, not the order.

The Dirichlet-multinomial model

- A generalization of the Bernoulli model for coin-tosses is the multinomial model
- This describes the probability of N rolls of a k-sided die

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k} \qquad \mathcal{D} = \{x_1, \ldots, x_N\} \text{ where } x_i \in \{1, \ldots, K\}$$

- Nk is the number of times "event" k occurred, i.e. a particular face came up.
- The conjugate prior for the multinomial model is the Dirichlet

$$p(\boldsymbol{\theta}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathbb{I}(\boldsymbol{\theta} \in S_K)$$

- The posterior has the same form

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

The posterior is obtained by adding the pseudocounts α_i to the empirical counts N_k

The Dirichlet-multinomial model

- The posterior predictive distribution is

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

- This gives the probability of each word given the observed data.
  - Mary had a little lamb, little lamb, little lamb
    Mary had a little lamb, its fleece was white as snow
- Assume $\alpha_i = 1$
- word: mary lamb little big fleece white black snow rain *unknown*

index:	1	2	3	4	5	6	7	8	9	10
count:	2	4	4	0	1	1	0	1	0	4
P(Xi=j|D)	3/27	5/27	5/27	0	2/27	2/27	0	2/27	0	5/27

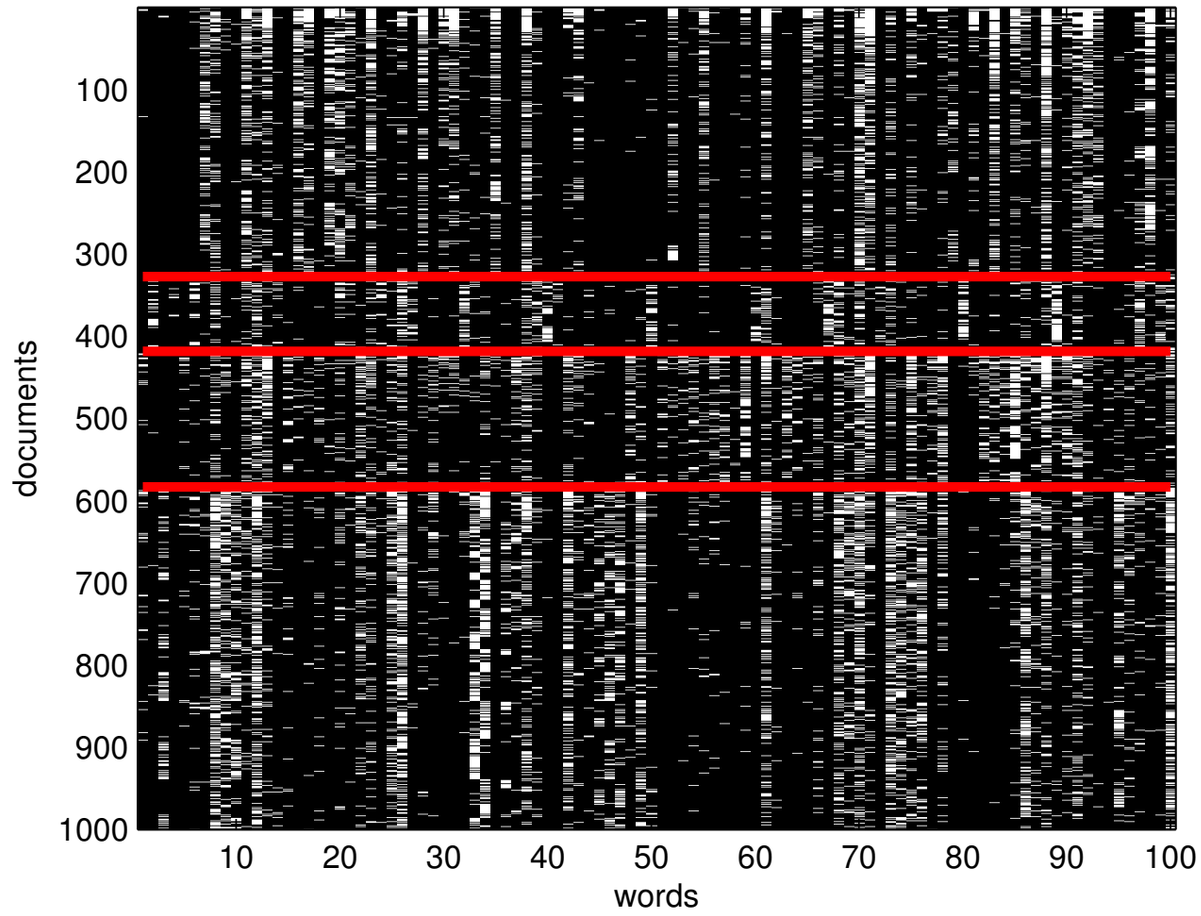
Text classification with the bag of words model

- Each row is a document represented as a bag-of-words vector
- each word is either *present or absent*
- The different classes are different newsgroups.
- The differences in word frequencies are readily apparent.
- We can use mixture models and naïve Bayes to classify the documents

$$p(C_k|\mathbf{x}) \quad = \quad \frac{p(C_k)\prod_n p(x_n|C_k)}{\sum_k p(C_k)\prod_n p(x_n|C_k)}$$

We only replace the data likelihood with our bag-of-words model.

- This is a common way to build a spam filter.



Simplifying with "Naïve" Bayes

- What if we assume the features are independent?

$$\begin{aligned} p(\mathbf{x}|C_k) &= p(x_1, \ldots, x_N|C_k) \\ &= \prod_{n=1}^{N} p(x_n|C_k) \end{aligned}$$

- We know that's not precisely true, but it might make a good approximation.
- Now we only need to specify N different likelihoods:

$$p(x_i = v_i|C_k = k) = \frac{\text{Count}(x_i = v_i \wedge C_k = k)}{\text{Count}(C_k = k)}$$

- Huge savings in number of of parameters

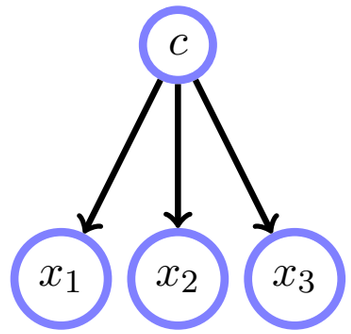
Inference with Naïve Bayes

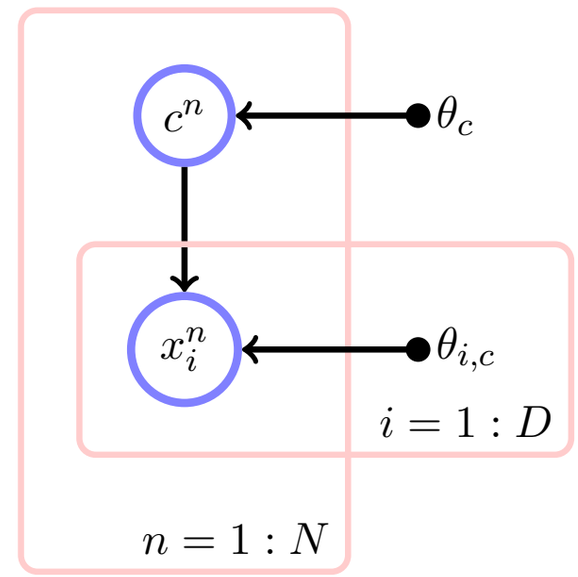
- Inference is just like before, but with the independence approximation:

$$p(\mathbf{x}, c) = p(\mathbf{x}|c)p(c)$$

$$= p(c)p(\mathbf{x}^*|c) = p(c)\prod_{i=1}^{D} p(x_i|c)$$

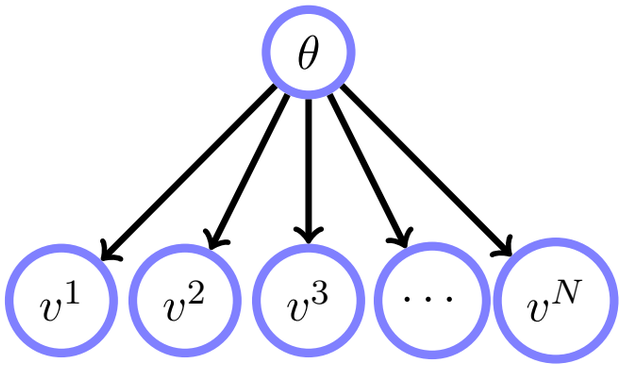
$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

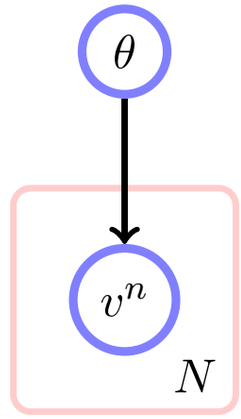




- where θ_c parameterizes p(c) and $\theta_{i,c}$ parameterizes $p(x_i|c)$.
- ***Plate notation*** representation of coin-tossing example

$$p(v^1, \ldots, v^N, \theta) = p(\theta)\prod_{n=1}^{N} p(v^n|\theta)$$





Naïve Bayes: Barber's English-Scottish example

- vector of binary attributes: (shortbread, lager, whiskey, porridge, football)
- binary vector of nationalities: (scottish, english)
- Two tables of attributes for (a) 6 English people and (b) 7 Scottish people

x_1	0	1	1	1	0	0
x_2	0	0	1	1	1	0
x_3	1	1	0	0	0	0
x_4	1	1	0	0	0	1
x_5	1	0	1	0	1	0

(a) English

1	1	1	1	1	1	1
0	1	1	1	1	0	0
0	0	1	0	0	1	1
1	0	1	1	1	1	0
1	1	0	0	1	0	0

(b) Scottish

- Using naïve Bayes, we can calculate $p(x_i|c)$ (Note this is an MLE estimation)

$$
\begin{array}{llll}
p(x_1 = 1|\text{english}) & = 1/2 & p(x_1 = 1|\text{scottish}) & = 1 \\
p(x_2 = 1|\text{english}) & = 1/2 & p(x_2 = 1|\text{scottish}) & = 4/7 \\
p(x_3 = 1|\text{english}) & = 1/3 & p(x_3 = 1|\text{scottish}) & = 3/7 \\
p(x_4 = 1|\text{english}) & = 1/2 & p(x_4 = 1|\text{scottish}) & = 5/7 \\
p(x_5 = 1|\text{english}) & = 1/2 & p(x_5 = 1|\text{scottish}) & = 3/7
\end{array}
$$

- For $\mathbf{x} = (1, 0, 1, 1, 0)^T$ **yes**: shortbread, whiskey, porridge; **no**: lager, football

$$
p(\text{scottish}|\mathbf{x}) = \frac{1 \times \frac{3}{7} \times \frac{3}{7} \times \frac{5}{7} \times \frac{4}{7} \times \frac{7}{13}}{1 \times \frac{3}{7} \times \frac{3}{7} \times \frac{5}{7} \times \frac{4}{7} \times \frac{7}{13} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{6}{13}} = 0.8076
$$

Maximum-likelihood estimation

- The likelihood for the naïve Bayes

$$p(\mathbf{x}|c) = \prod_{i=1}^{D} p(x_i|c) = \prod_{i=1}^{D} (\theta_i^c)^{x_i} (1-\theta_i^c)^{1-x_i}$$

- where xi are binary variables (raising to the power of x_i or (1-x_i) selects just one)
- To derive the maximum likelihood estimate, we start with the log likelihood

$$L = \sum_n \log p(\mathbf{x}^n, c^n) = \sum_n \log p(c^n) \prod_i p(x_i^n|c^n)$$

$$= \left\{ \sum_{i,n} x_i^n \log \theta_i^{c^n} + (1-x_i^n) \log(1-\theta_i^{c^n}) \right\} + n_0 \log p(c=0) + n_1 \log p(c=1)$$

Maximum-likelihood estimation

- The log likelihood for the naïve Bayes

$$L = \sum_n \log p(\mathbf{x}^n, c^n) = \sum_n \log p(c^n) \prod_i p(x_i^n | c^n)$$

$$= \left\{ \sum_{i,n} x_i^n \log \theta_i^{c^n} + (1 - x_i^n) \log(1 - \theta_i^{c^n}) \right\} + n_0 \log p(c = 0) + n_1 \log p(c = 1)$$

- can be written as

$$L = \sum_{i,n} \left\{ \mathbb{I}\left[x_i^n = 1, c^n = 0\right] \log \theta_i^0 + \mathbb{I}\left[x_i^n = 0, c^n = 0\right] \log(1 - \theta_i^0) + \mathbb{I}\left[x_i^n = 1, c^n = 1\right] \log \theta_i^1 \right.$$

$$\left. + \mathbb{I}\left[x_i^n = 0, c^n = 1\right] \log(1 - \theta_i^1) \right\} + n_0 \log p(c = 0) + n_1 \log p(c = 1)$$

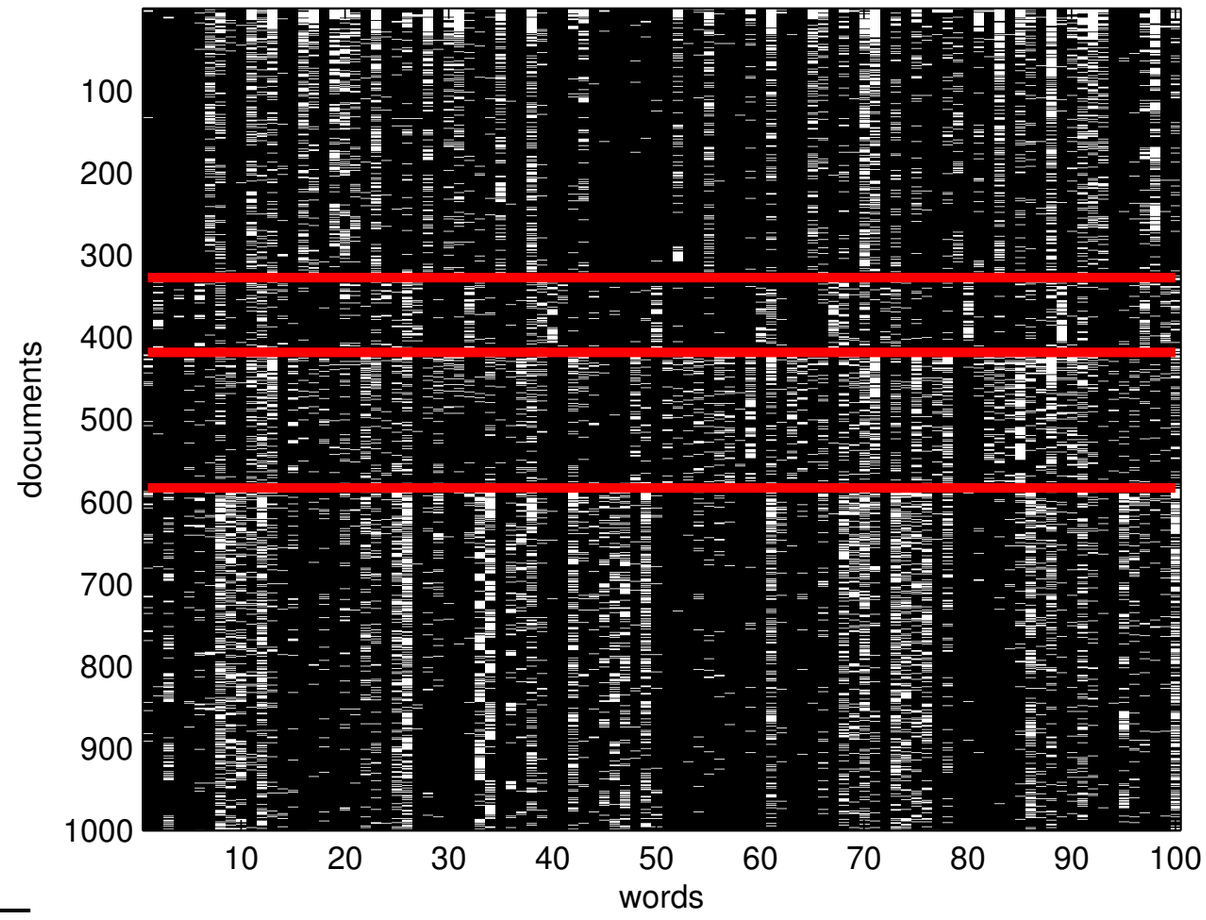
- differentiating wrt θ_c^i and solving for the optimum yields

$$\theta_i^c = p(x_i = 1 | c) = \frac{\sum_n \mathbb{I}\left[x_i^n = 1, c^n = c\right]}{\sum_n \mathbb{I}\left[x_i^n = 0, c^n = c\right] + \mathbb{I}\left[x_i^n = 1, c^n = c\right]}$$

$$= \frac{\text{number of times } x_i = 1 \text{ for class c}}{\text{number of datapoints in class c}}$$

Text classification with the bag of words model

- Each row is a document represented as a bag-of-words vector.
- The different classes are different newsgroups.
- The differences in word frequencies are readily apparent.
- We can use mixture models and naïve Bayes to classify the documents



$$p(C_k|\mathbf{x}) \quad = \quad \frac{p(C_k)\prod_n p(x_n|C_k)}{\sum_k p(C_k)\prod_n p(x_n|C_k)}$$

- We only replace the data likelihood with our bag-of-words model.
- This is a common way to build a spam filter or classify web pages.

Bayesian Naïve Bayes

- MLE for NB can fail for small counts
- binary vector of nationalities: (scottish, english)
- Two tables of attributes for (a) 6 English people and (b) 7 Scottish people

x_1	0	1	1	1	0	0
x_2	0	0	1	1	1	0
x_3	1	1	0	0	0	0
x_4	1	1	0	0	0	1
x_5	1	0	1	0	1	0

(a) English

1	1	1	1	1	1	1
0	1	1	1	1	0	0
0	0	1	0	0	1	1
1	0	1	1	1	1	0
1	1	0	0	1	0	0

(b) Scottish

- Using naïve Bayes, we can calculate $p(x_i|c)$ (Note this is an MLE estimation)

$$
\begin{aligned}
p(x_1 = 1|\text{english}) &= 1/2 & p(x_1 = 1|\text{scottish}) &= 1 \\
p(x_2 = 1|\text{english}) &= 1/2 & p(x_2 = 1|\text{scottish}) &= 4/7 \\
p(x_3 = 1|\text{english}) &= 1/3 & p(x_3 = 1|\text{scottish}) &= \boxed{3/7 \rightarrow 0/7} \\
p(x_4 = 1|\text{english}) &= 1/2 & p(x_4 = 1|\text{scottish}) &= 5/7 \\
p(x_5 = 1|\text{english}) &= 1/2 & p(x_5 = 1|\text{scottish}) &= 3/7
\end{aligned}
$$

classification is over confident

- For x = (1, 0, 1, 1, 0)^T **yes**: shortbread, whiskey, porridge; **no**: lager, football

$$
p(\text{scottish}|\mathbf{x}) = \frac{1 \times \boxed{\frac{0}{7}} \times \frac{3}{7} \times \frac{5}{7} \times \frac{4}{7} \times \frac{7}{13}}{1 \times \frac{3}{7} \times \frac{3}{7} \times \frac{5}{7} \times \frac{4}{7} \times \frac{7}{13} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{6}{13}} = \boxed{0}
$$

Bayesian Naïve Bayes

- MLE does not define a prior - this leads to poor estimates for small numbers
- A Dirichlet prior is a natural prior for a categorical distribution:

$$p(\boldsymbol{\theta}) = \prod_{i,c} p(\boldsymbol{\theta}^i(c))$$

$$p(\boldsymbol{\theta}^i(c)) = \text{Dirichlet}(\boldsymbol{\theta}^i(c)|\boldsymbol{u}^i(c))$$

- Note: Barber uses $\boldsymbol{u}^i(c))$ rather than $\boldsymbol{\alpha}^i(c))$ for the category priors.
- The posterior distribution for the parameters factorizes

$$p(\theta(c^*)|\mathcal{D}) = \prod_i p(\boldsymbol{\theta}^i(c^*)|\mathcal{D})$$

- because the prior is conjugate, the posterior distribution is also Dirichlet

$$p(\boldsymbol{\theta}^i(c^*)|\mathcal{D}) = \text{Dirichlet}\left(\boldsymbol{\theta}^i(c^*)|\hat{\mathbf{u}}^i(c^*)\right)$$

- where the Dirichlet posterior parameters are defined as

$$\left[\hat{\mathbf{u}}^i(c^*)\right]_s = u_s^i(c^*) + \sum_{n:c^n=c^*} \mathbb{I}\left[x_i^n = s\right]$$

The parameters $u_s^i(c^*)$ are "ghost" observations.

A more flexible prior: Beta-Binomial model

- The general expression for the posterior distribution is given by

$$p(\theta|y,n) = \frac{p(y|\theta,n)p(\theta|n)}{p(y|n)}$$

- In the coin-flipping example, the probability of heads (or tails) is governed by the Bernoulli distribution

$$X_i \sim \text{Ber}(\theta), \text{ so } X_i \in \{0,1\}.$$

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

where we have $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ tails.

- the ***conjugate prior*** for the Bernoulli is the Beta distribution:

$$\text{Beta}(\theta|\alpha_1,\alpha_2) \propto \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

- therefore the posterior distribution *has the same form*:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &\propto [\theta^{N_1}(1-\theta)^{N_2}][\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}] \\ &= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1} \\ &\propto \text{Beta}(\theta|N_1+\alpha_1, N_2+\alpha_2) \end{aligned}$$

Beta distribution

Definition 8.23 (Beta Distribution)**.**

$$p(x|\alpha,\beta) = B(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1 \tag{8.3.17}$$

where the Beta function is defined as

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{8.3.18}$$

and $\Gamma(x)$ is the Gamma function. Note that the distribution can be flipped by interchanging x for $1-x$, which is equivalent to interchanging α and β. See fig(8.4).

The mean and variance are given by

$$\langle x \rangle = \frac{\alpha}{\alpha+\beta}, \qquad \text{var}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{8.3.19}$$

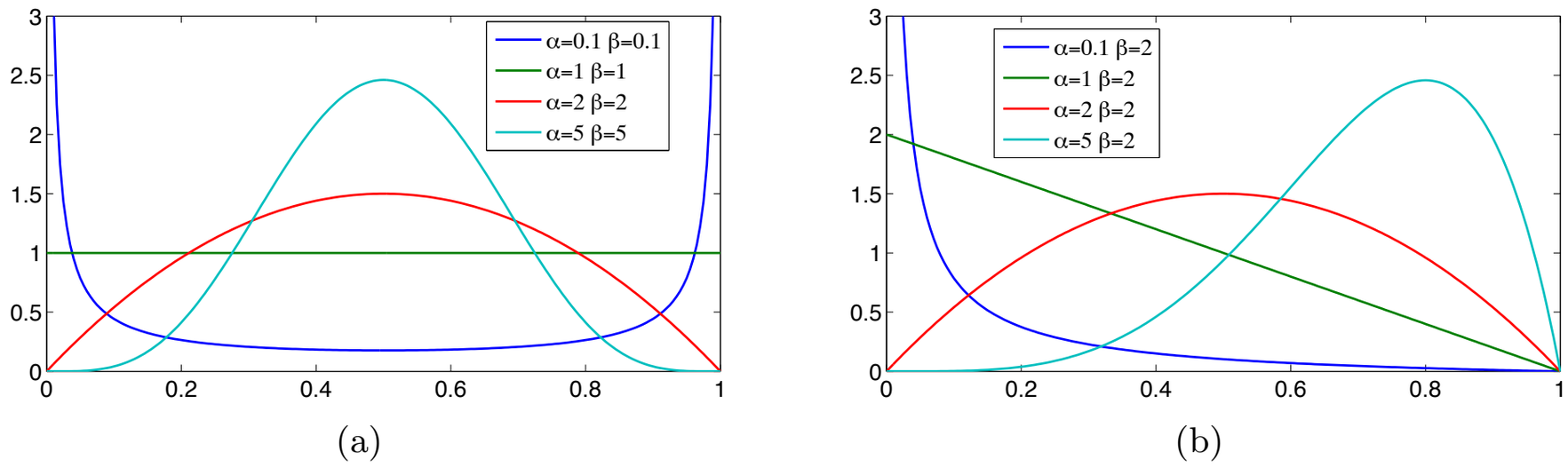


Figure 8.4: Beta distribution. The parameters α and β can also be written in terms of the mean and variance, leading to an alternative parameterisation, see exercise(8.16).

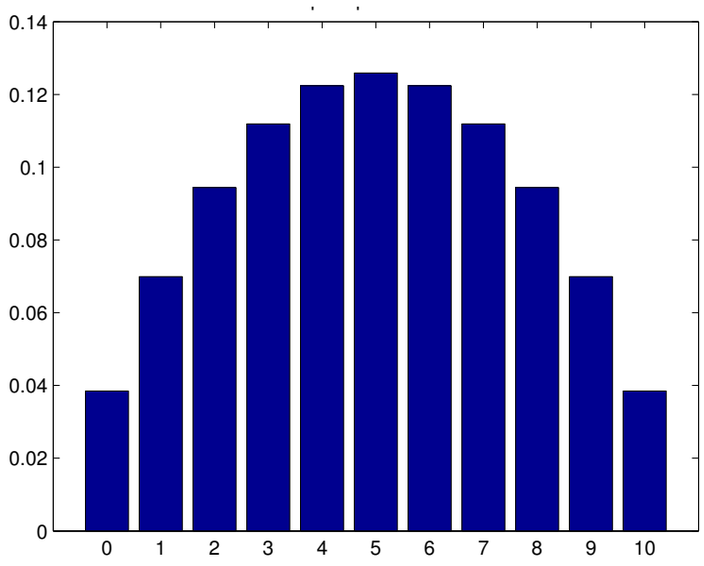
Posterior predictive distribution for the Binomial problem

- The posterior predictive density for a single future trial is

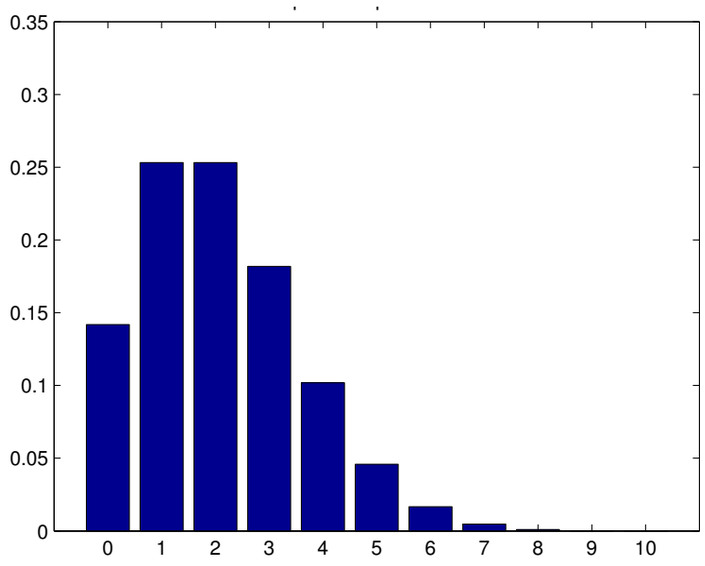
$$
\begin{aligned}
p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(x = 1|\theta) p(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta \, \mathrm{Beta}(\theta|\alpha_1', \alpha_2') d\theta = \mathbb{E}\left[\theta|\mathcal{D}\right] = \frac{\alpha_1'}{\alpha_1' + \alpha_2'}
\end{aligned}
$$

- This is a weighted average of the hypotheses over the posterior.

prior predictive Be(2,2), 10 trials



posterior predictive (for 10 trials)
after 3 heads, 17 tails



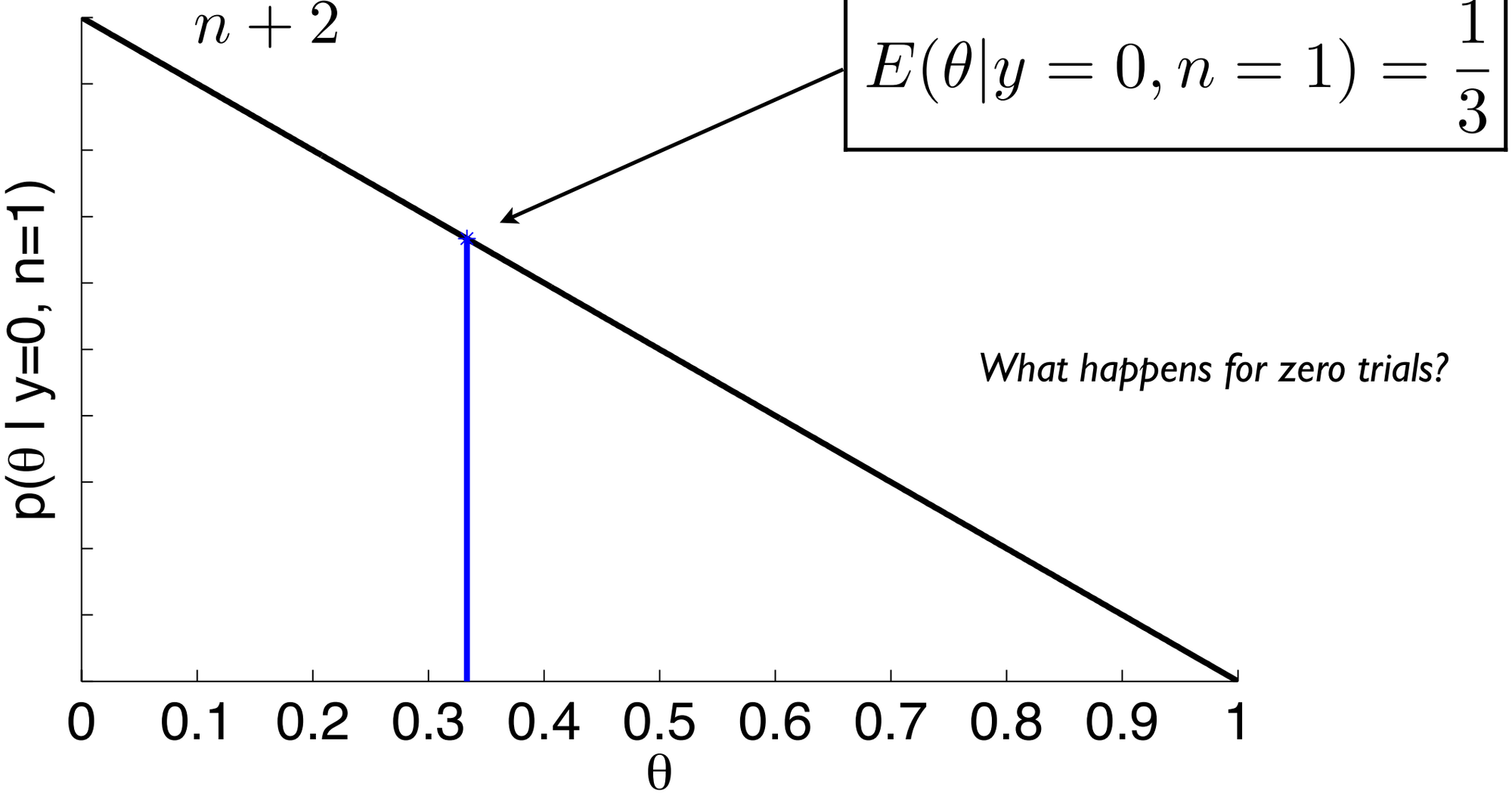
Recall the expected value estimate for a uniform

- The expected value of a pdf is:

$$E(\theta|y,n) = \int_0^1 \theta p(\theta|y,n) d\theta$$

$$= \frac{y+1}{n+2}$$

$$E(\theta|y=0,n=1) = \frac{1}{3}$$



What happens for zero trials?

Language modeling: bag of words

- Assume the ith word, Xi in {1, ..., K} is sampled independently of the others.
- Given the previous words, can we predict what word comes next?
- Given
  - Mary had a little lamb, little lamb, little lamb
    Mary had a little lamb, its fleece was white as snow
- Suppose our vocabulary is:
  Mary lamb little big fleece white black snow rain *unknown*
  1 2 3 4 5 6 7 8 9 10
- To encode the rhyme, we "normalize" it and replace each word by its index:
  - strip words like: a, as, the, etc
  - perform stemming, *raining* becomes *rain*.
  - Encoded data:
    1 10 3 2 3 2 3 2
    1 10 3 2 10 5 6 8
- These counts can modeled this with a Dirichlet-multinomial model.
  Only the word count matters, not the order.

The Dirichlet-multinomial model

- A generalization of the Bernoulli model for coin-tosses is the multinomial model
- This describes the probability of N rolls of a k-sided die

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k} \qquad \mathcal{D} = \{x_1, \ldots, x_N\} \text{ where } x_i \in \{1, \ldots, K\}$$

- Nk is the number of times "event" k occurred, i.e. a particular face came up.
- The conjugate prior for the multinomial model is the Dirichlet

$$p(\boldsymbol{\theta}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathbb{I}(\boldsymbol{\theta} \in S_K)$$

- The posterior has the same form

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

The posterior is obtained by adding the pseudocounts α_i to the empirical counts N_k

Definition 8.27 (Dirichlet Distribution)**.** The Dirichlet distribution is a distribution on probability distributions, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$, $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$:

$$p(\boldsymbol{\alpha}) = \frac{1}{Z(\mathbf{u})} \delta \left(\sum_{i=1}^{Q} \alpha_i - 1 \right) \prod_{q=1}^{Q} \alpha_q^{u_q - 1} \mathbb{I}\left[\alpha_q \geq 0\right] \tag{8.3.29}$$

where

$$Z(\mathbf{u}) = \frac{\prod_{q=1}^{Q} \Gamma(u_q)}{\Gamma\left(\sum_{q=1}^{Q} u_q\right)} \tag{8.3.30}$$

It is conventional to denote the distribution as

$$\text{Dirichlet}\left(\boldsymbol{\alpha}|\mathbf{u}\right) \tag{8.3.31}$$

The parameter $\mathbf{u}$ controls how strongly the mass of the distribution is pushed to the corners of the simplex. Setting $u_q = 1$ for all q corresponds to a uniform distribution, fig(8.6). In the binary case $Q = 2$, this is equivalent to a Beta distribution.

The product of two Dirichlet distributions is another Dirichlet distribution

$$\text{Dirichlet}\left(\boldsymbol{\theta}|\mathbf{u}_1\right) \text{Dirichlet}\left(\boldsymbol{\theta}|\mathbf{u}_2\right) = \text{Dirichlet}\left(\boldsymbol{\theta}|\mathbf{u}_1 + \mathbf{u}_2\right) \tag{8.3.32}$$

The marginal of a Dirichlet is also Dirichlet:

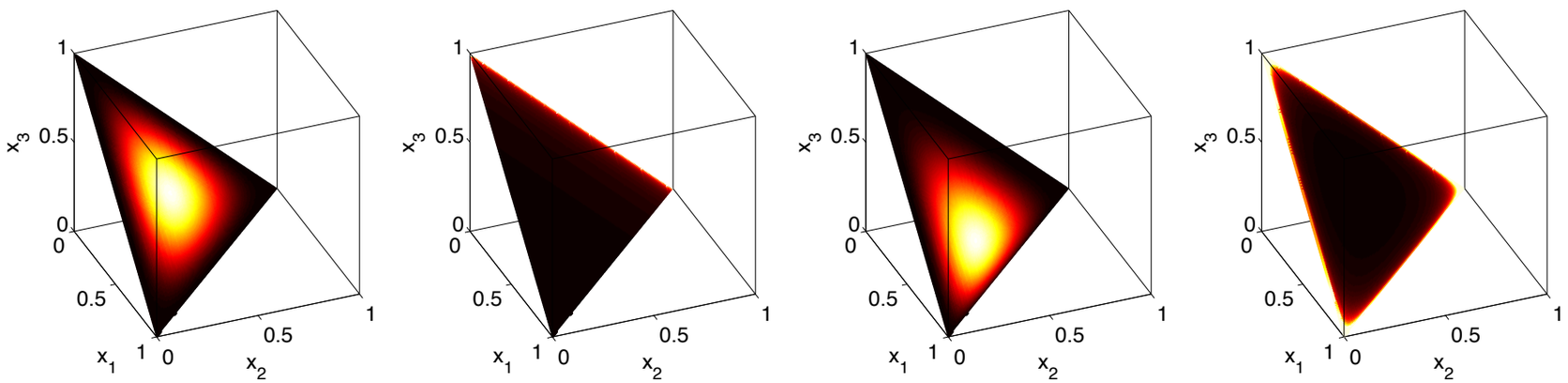
$$\int_{\theta_j} \text{Dirichlet}\left(\boldsymbol{\theta}|\mathbf{u}\right) = \text{Dirichlet}\left(\boldsymbol{\theta}_{\setminus j}|\mathbf{u}_{\setminus j}\right) \tag{8.3.33}$$

The marginal of a single component θ_i is a Beta distribution:

$$p(\theta_i) = B\left(\theta_i | u_i, \sum_{j \neq i} u_j\right) \tag{8.3.34}$$

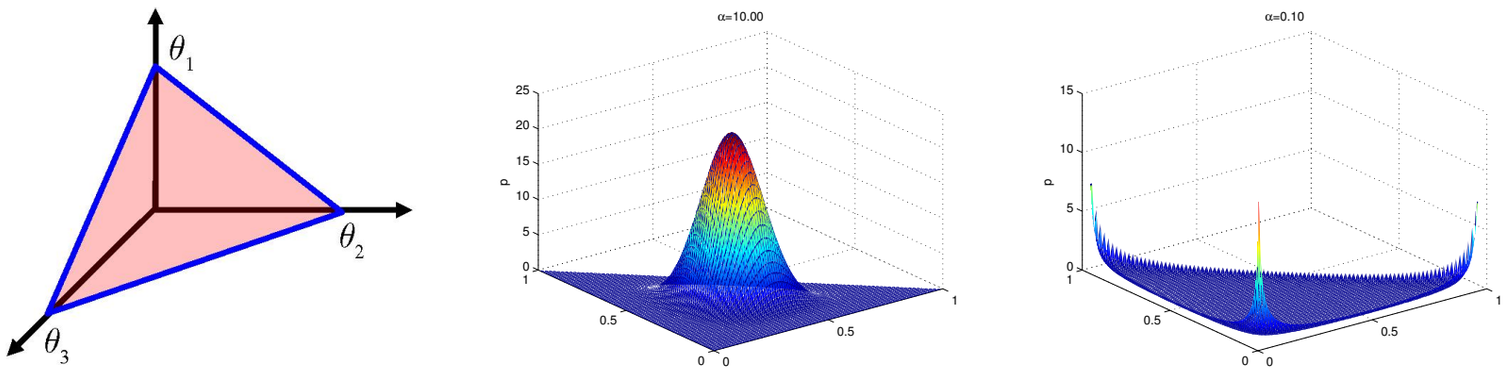
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$, $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$:

$$p(\boldsymbol{\alpha}) = \frac{1}{Z(\mathbf{u})} \delta\left(\sum_{i=1}^{Q} \alpha_i - 1\right) \prod_{q=1}^{Q} \alpha_q^{u_q - 1} \mathbb{I}\left[\alpha_q \geq 0\right]$$



(a) (b) (c) (d)

Figure 8.6: Dirichlet distribution with parameter (u_1, u_2, u_3) displayed on the simplex $x_1, x_2, x_3 \geq 0, x_1 + x_2 + x_3 = 1$. Black denotes low probability and white high probability. **(a)**: $(3, 3, 3)$ **(b)**: $(0.1, 1, 1)$. **(c)**: $(4, 3, 2)$. **(d)**: $(0.05, 0.05, 0.05)$.



(a) (b) (c)

$$p(\mathbf{x}) = \mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

$$B(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ and $B(\alpha_1, ..$
require $0 \leq x_k \leq 1$ and $\sum_{k=1}^{K} x_k = 1$

Figure 2.14: (a) The Dirichlet distribution when $K = 3$ defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{3} \theta_k = 1$. Based on Figure 2.4 of [Bis06a]. (b) Plot of the Dirichlet density when $\alpha_k = 10$. (c) Plot of the Dirichlet density when $\alpha_k = 0.1$. (The comb-like structure on the edges is a plotting artefact.) Based on Figure 2.5 of [Bis06a]. Produced by `dirichlet3dPlot`. (See also `visDirichletGui` by Jonathan Huang.)

Language modeling with the Dirichlet-multinomial model

- The posterior predictive distribution is

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

- This gives the probability of each word given the observed data.
  - Mary had a little lamb, little lamb, little lamb
    Mary had a little lamb, its fleece was white as snow
- Assume $\alpha_i = 1$
- word: mary lamb little big fleece white black snow rain *unknown*

index:	1	2	3	4	5	6	7	8	9	10
count:	2	4	4	0	1	1	0	1	0	4
P(Xi=j|D)	3/27	5/27	5/27	0	2/27	2/27	0	2/27	0	5/27