CS340

Bayesian concept learning cont'd

Kevin Murphy

Prior p(h)

- X={60,80,10,30}
- Why prefer "multiples of 10" over "even numbers"?
  - Size principle (likelihood)
- Why prefer "multiples of 10" over "multiples of 10 except 50 and 20"?
  - Prior
- Cannot learn efficiently if we have a uniform prior over all 2^{100} logically possible hypotheses

Need for prior (inductive bias)

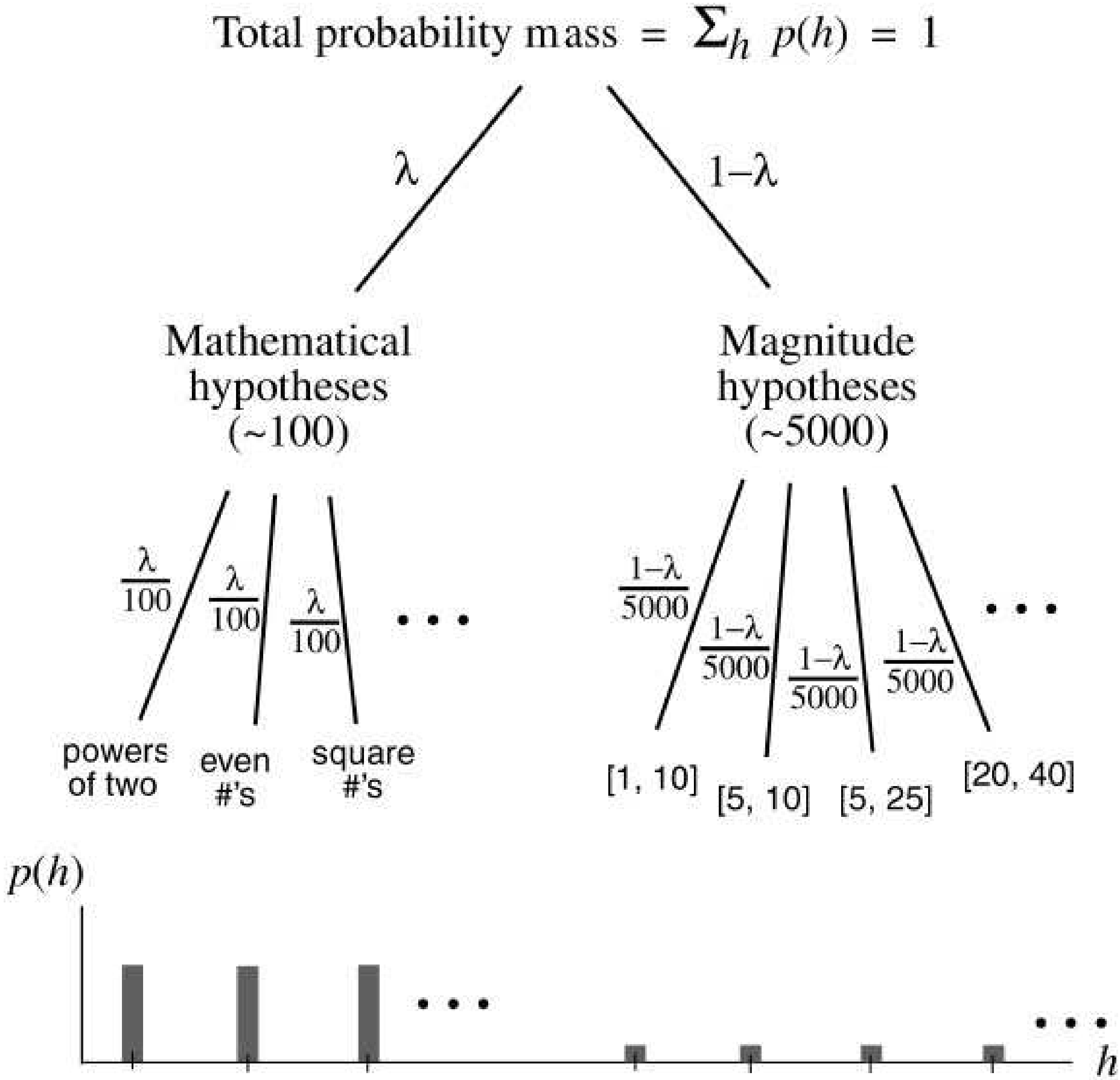
- Consider all $2^{2^2} = 16$ possible binary functions on 2 binary inputs

Boolean functions.

x_1	x_2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	~~1~~	~~1~~	~~1~~	~~1~~	0	0	0	0	~~1~~	~~1~~	~~1~~	~~1~~
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

- If we observe (x_1=0, x_2=1, y=0), this removes h_5, h_6, h_7, h_8, h_{13}, h_{14}, h_{15}, h_{16}
- Still leaves exponentially many hypotheses!
- Cannot learn efficiently without assumptions (no free lunch theorem)

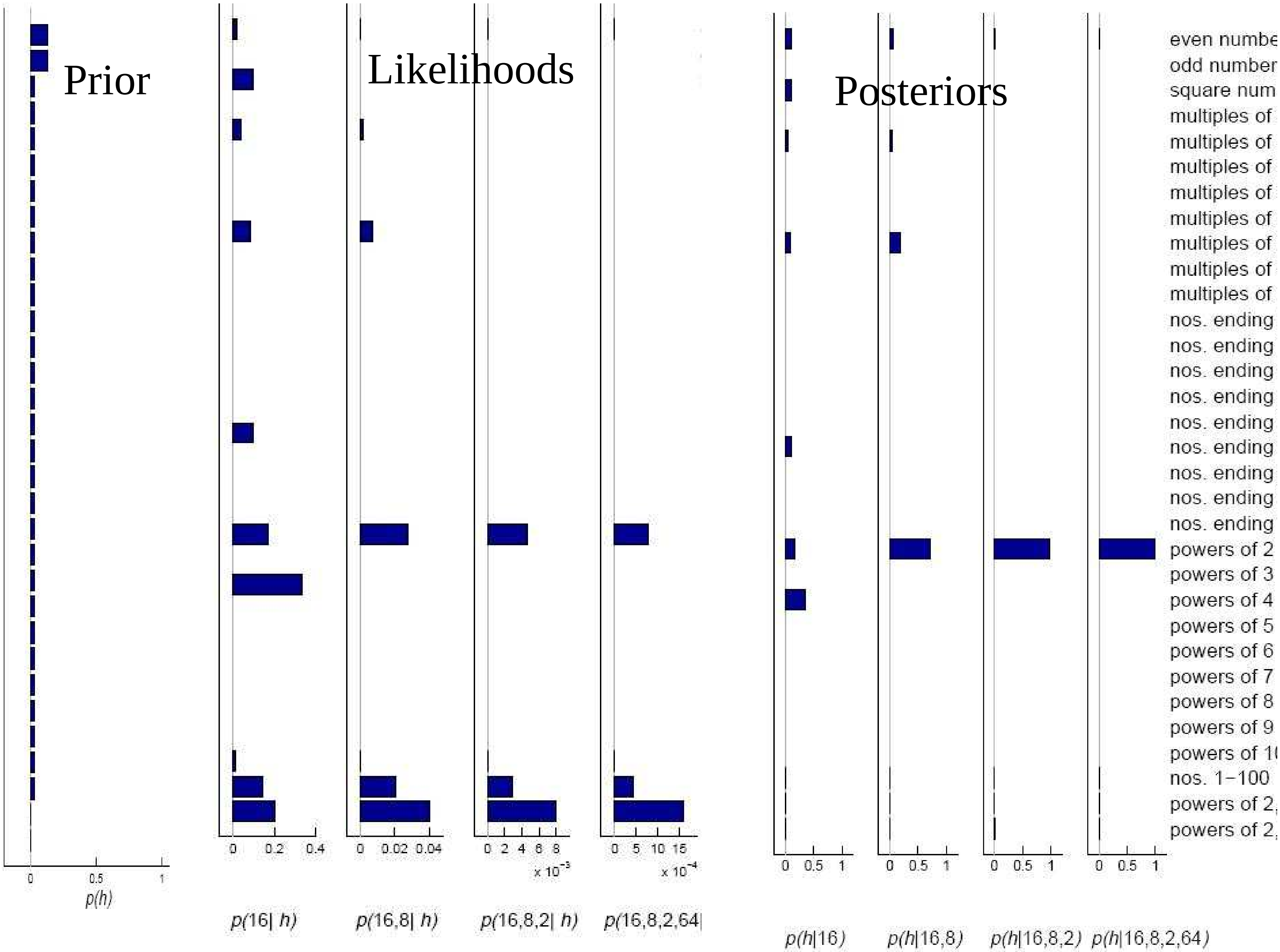
Hierarchical prior



Computing the posterior

- In this talk, we will not worry about computational issues (we will perform brute force enumeration or derive analytical expressions).

$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$



Prior
Likelihoods
Posteriors
p(h)
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64|
p(h|16)
p(h|16,8)
p(h|16,8,2)
p(h|16,8,2,64)
x 10^-3
x 10^-4
even numbe
odd number
square num
multiples of
multiples of
multiples of
multiples of
multiples of
multiples of
multiples of
multiples of
nos. ending
nos. ending
nos. ending
nos. ending
nos. ending
nos. ending
nos. ending
nos. ending
nos. ending
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 1
nos. 1−100
powers of 2,
powers of 2,

Generalizing to new objects

Given $p(h|X)$, how do we compute $p(y \in C \mid X)$, the probability that C applies to some new stimulus y?

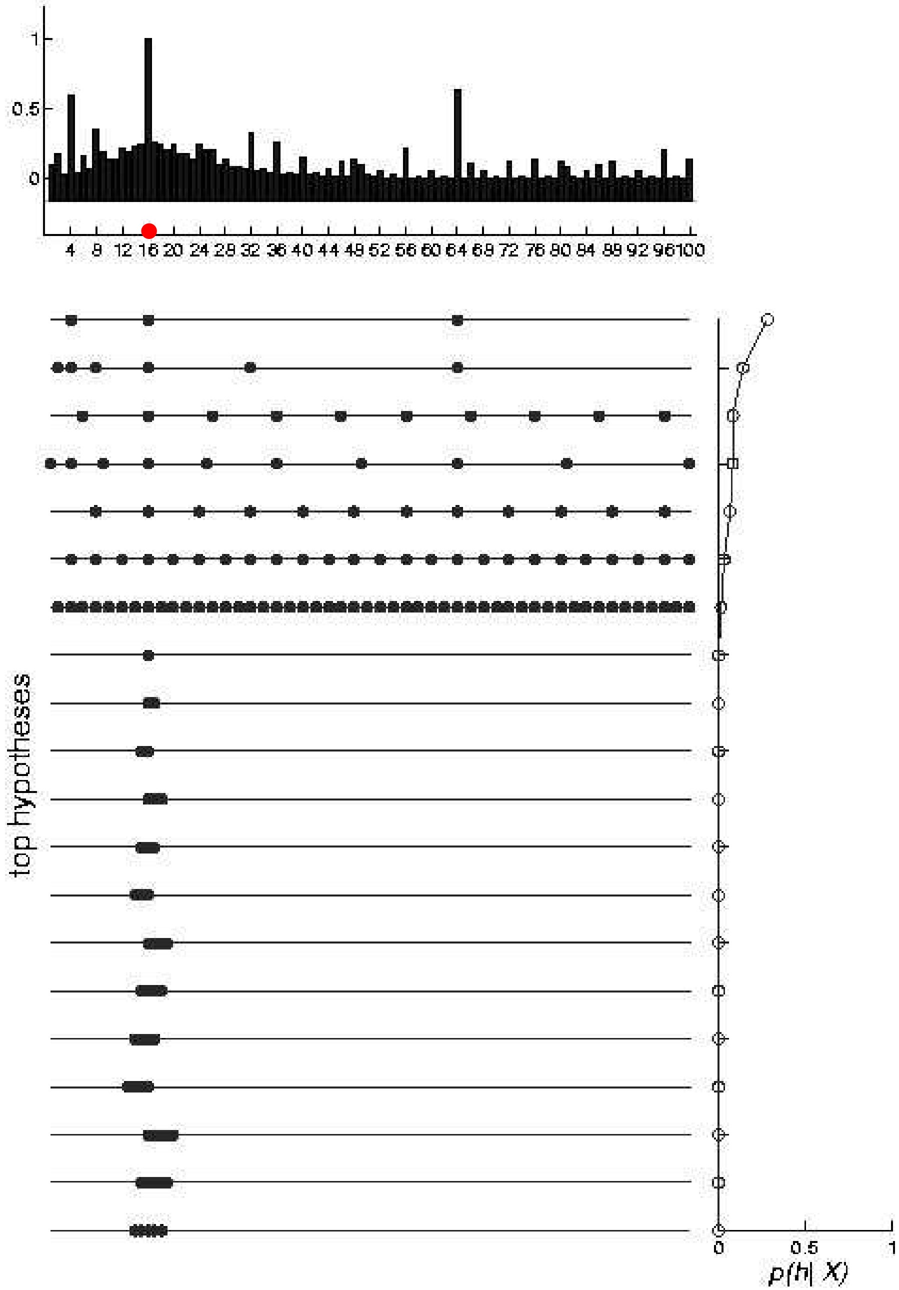
Posterior predictive distribution

Compute the probability that *C* applies to some new object *y* by averaging the predictions of all hypotheses *h*, weighted by $p(h|X)$ (**Bayesian model averaging**):

$$p(y \in C \mid X) = \sum_{h \in H} \underbrace{p(y \in C \mid h)}_{= \begin{cases} 1 \text{ if } y \in h \\ 0 \text{ if } y \notin h \end{cases}} \; p(h \mid X)$$

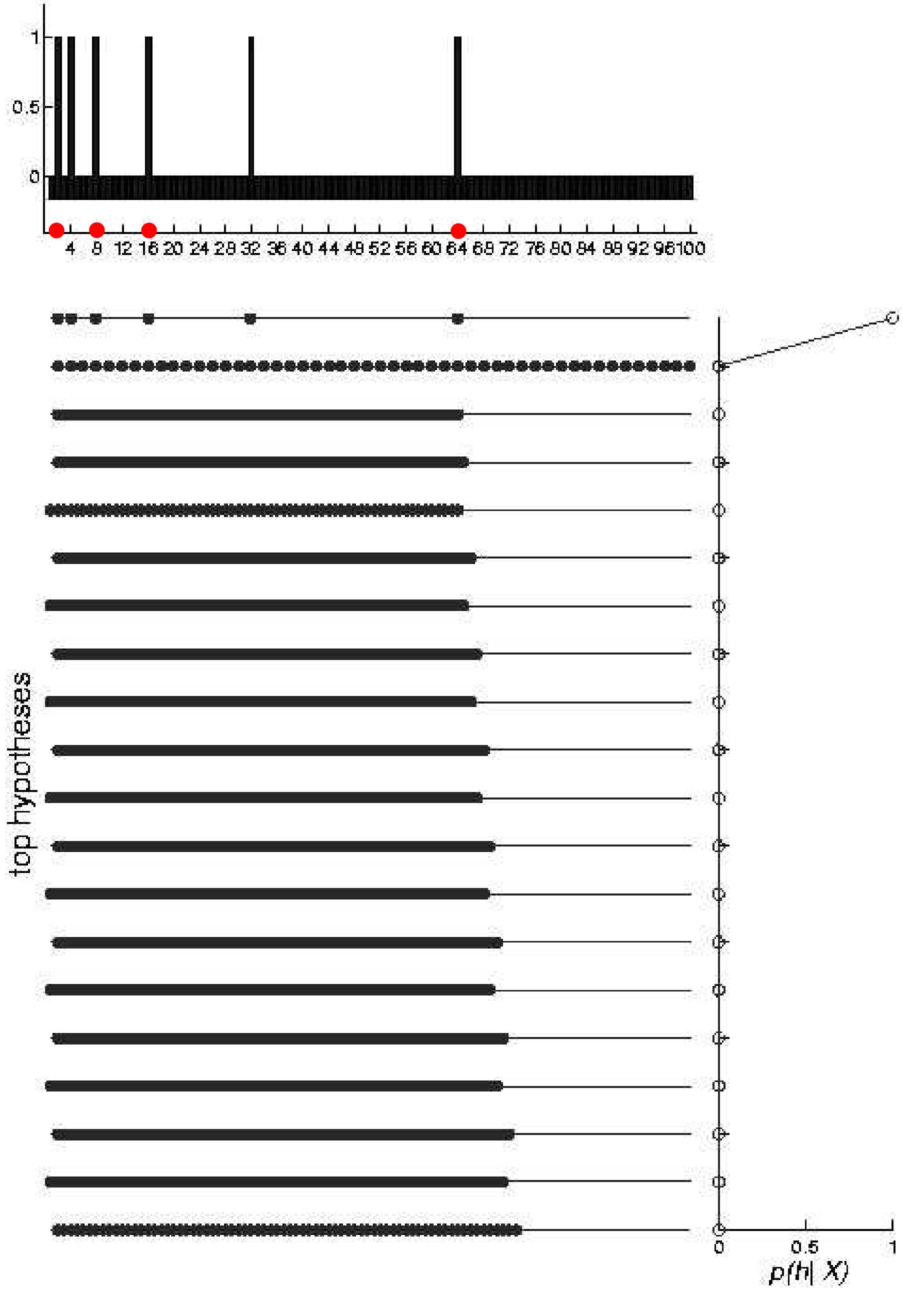
$$= \sum_{h \supset \{y, X\}} p(h \mid X)$$

Examples:
16



1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
top hypotheses
0
0.5
1
p(h| X)

Examples:
16
8
2
64



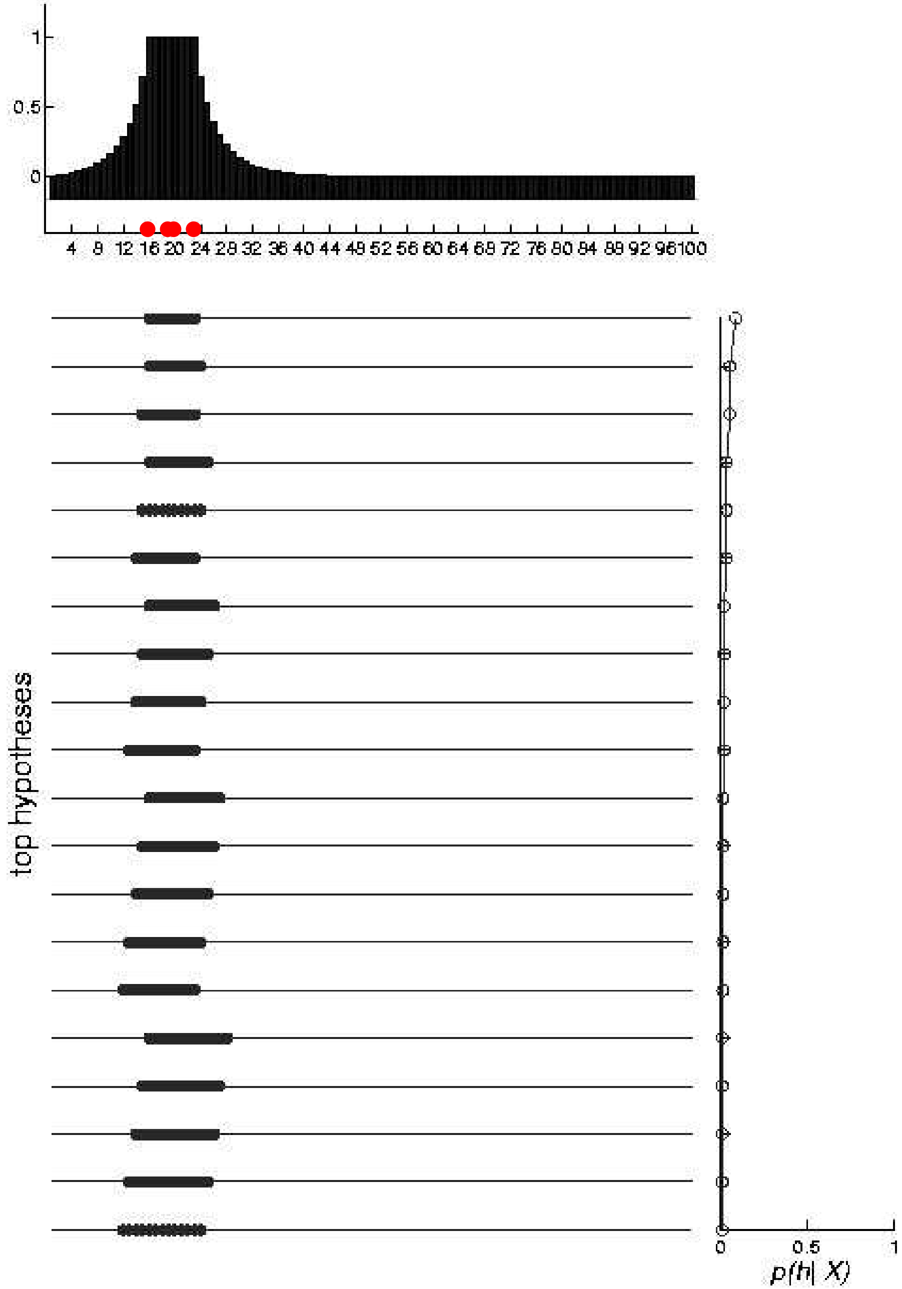
Examples:

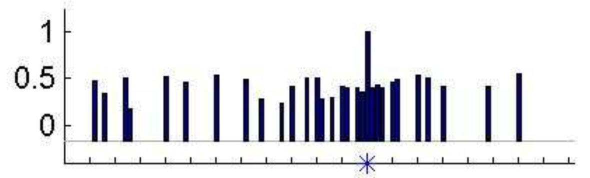
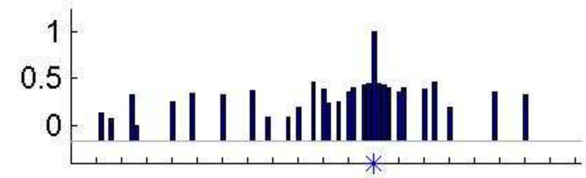
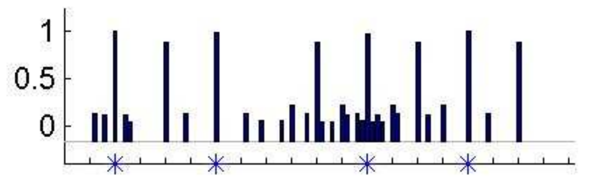
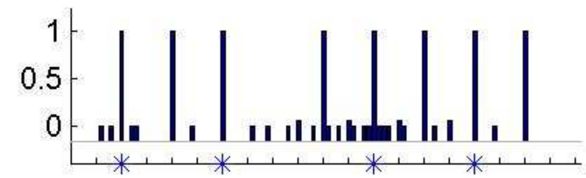
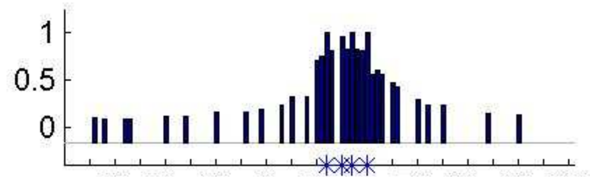
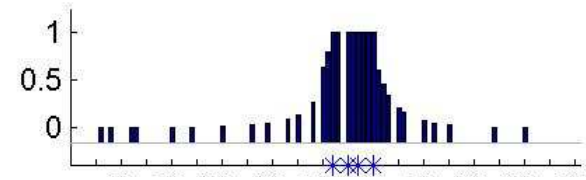
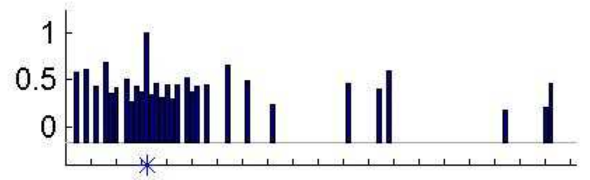
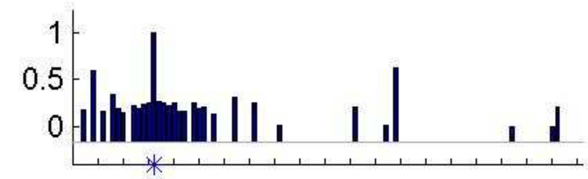
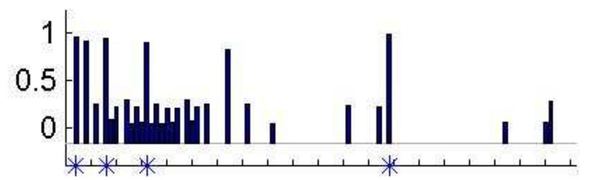
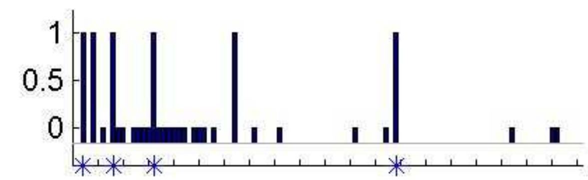
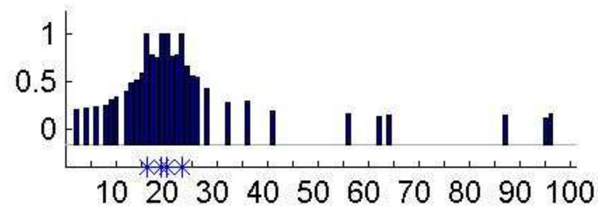
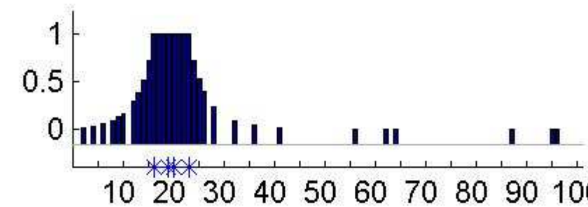
16

23

19

20

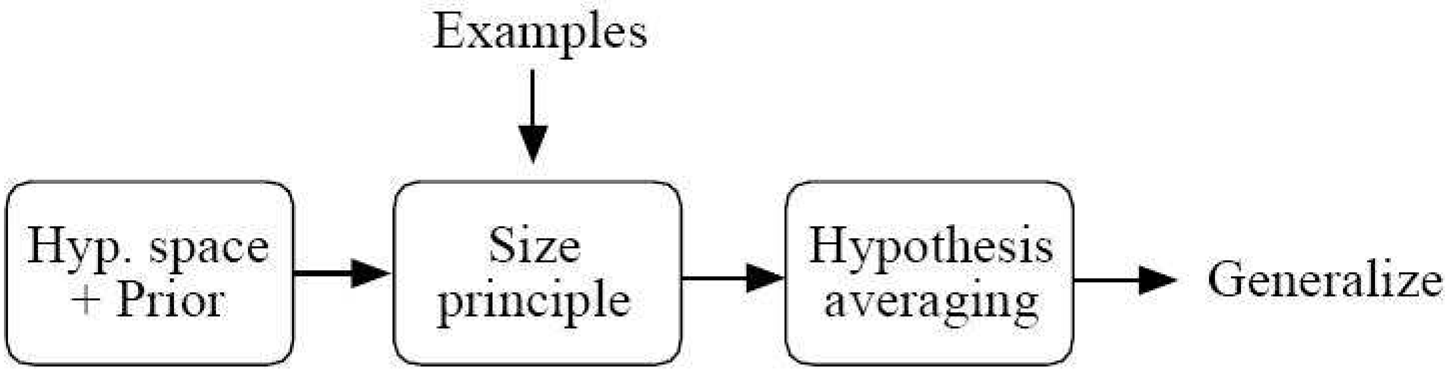


+ Examples	Human generalization	Bayesian Model
60		
60 80 10 30		
60 52 57 55		
16		
16 8 2 64		
16 23 19 20		

Rules and exemplars in the number game

- Hyp. space is a mixture of sparse (mathematical concepts) and dense (intervals) hypotheses.
- If data supports mathematical rule (eg X={16,8,2,64}), we rapidly learn a rule (“aha!” moment), otherwise (eg X={6,23,19,20}) we learn by similarity, and need many examples to get sharp boundary.

Summary of the Bayesian approach



1. Constrained hypothesis space H
2. Prior p(h)
3. Likelihood p(X|h)
4. Hypothesis (model) averaging:

$$p(y \in C\ \ X) = \sum_h p(y \in C|h)p(h|X)$$

MAP (maximum a posterior) learning

- Instead of Bayes model averaging, we can find the mode of the posterior, and use it as a plug-in.

$$
\begin{aligned}
\hat{h} &= \arg\max_h p(h|X) = \arg\max_h p(X|h)p(h) \\
p(y \in C|X) &= p(y \in C|\hat{h})
\end{aligned}
$$

- As N $\rightarrow \infty$, the posterior peaks around the mode, so MAP and BMA converge

$$p(y \in C|X) = \sum_h p(y \in C|h)p(h|X) \rightarrow \sum_h p(y \in C|h)\delta(h, \hat{h})) = p(y \in C|\hat{h})$$

- Cannot explain transition from similarity-based (broad posterior) to rule-based (narrow posterior)

Relation between MAP and MDL

- MAP (penalized likelihood) estimation:

$$P(h \mid X) \propto P(X \mid h)\, P(h)$$

- Minimum description length (MDL):

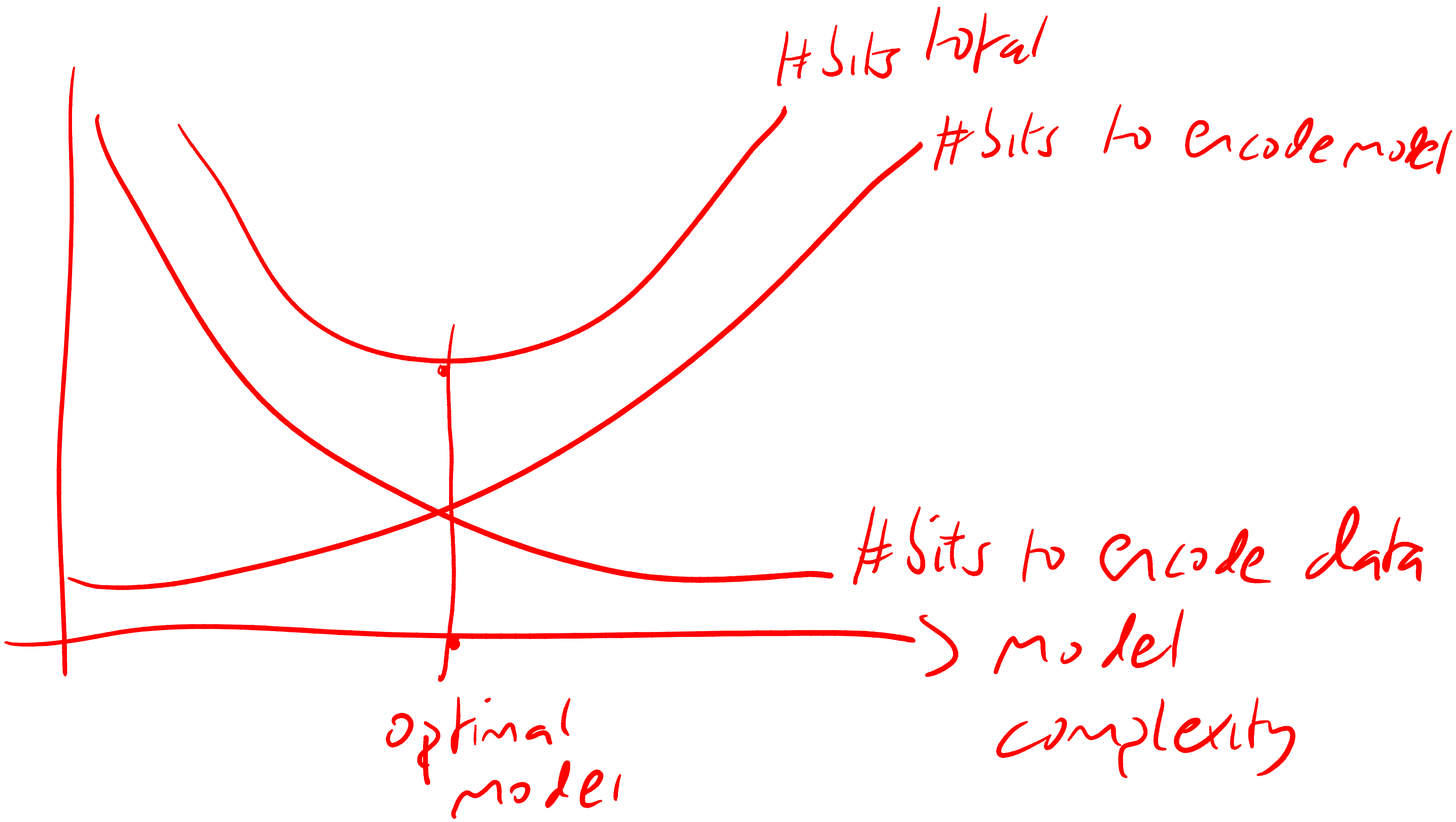
$$-\log P(h \mid X) = -\log P(X \mid h) \quad + \quad -\log P(h) \quad + \quad \text{Const}$$

Total encoding cost

Cost to encode the data given the hypothesis

Cost to encode the hypothesis

Model selection using MDL



Bayesian Occam's Razor

- Which hypothesis is better supported by the examples {54, 6, 22}?
  - "even numbers"
  - "numbers between 6 and 54"
- Intuition: simpler hypotheses come from smaller (more constrained) hypothesis spaces.
  - "Entities should not be multiplied without necessity"
  - Prefer models with fewer free parameters.
- Both prior and likelihood contribute to this, since $p(h|X) \propto p(h)\, p(X|h)$

Maximum likelihood

- ML = no prior, no averaging.
- Plugs-in the MLE for prediction:

$$
\begin{aligned}
\hat{h} &= \arg\max_h p(X|h) \\
p(y \in C|X) &= p(y \in C|\hat{h})
\end{aligned}
$$

- X={16} -> h= "powers of 4"
  X={16,8,2,64} -> h= "powers of 2".
- So predictive distribution gets broader as we get more data, in contrast to Bayes.
- ML is initially very conservative.

Large sample size behavior

- As the amount of data goes to ∞, ML, MAP and BMA all converge to the same solution, since the likelihood overwhelms the prior, since $p(X|h)$ grows with N, but $p(h)$ is constant.
- If truth is in the hypothesis class, all methods will find it; thus they are consistent estimators.