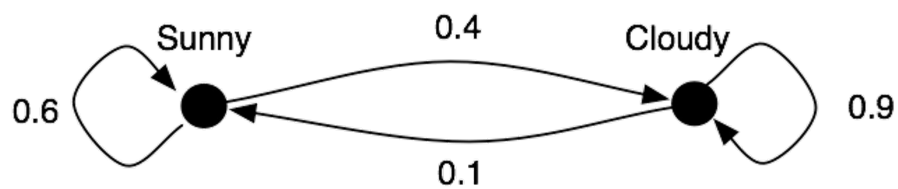
MCMC and Statistical learning

Announcements

- A2 posted, sign up and create your teams
- Midterm exam 10/26 6:30pm-7:45pm
  - Mostly multiple choice
  - Review questions posted
  - Online using Canvas
- Final exam
  - Friday 12/16 7:40pm-9:40pm
  - Online using Canvas
- Don't forget the quiz

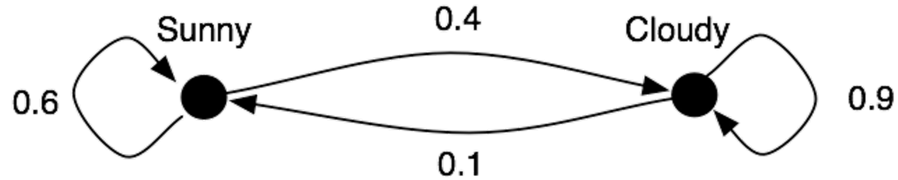
Markov chains



- Suppose there's an 80% chance of sun on day 0. What is the probability of sun on day 3?

$$
\begin{aligned}
P(Q_3 = \text{☀}) &= P(Q_3 = \text{☀} | Q_2 = \text{☀}) P(Q_2 = \text{☀}) + P(Q_3 = \text{☀} | Q_2 = \text{🌧}) P(Q_2 = \text{🌧}) \\
&= 0.6 P(Q_2 = \text{☀}) + 0.1 P(Q_2 = \text{🌧}) \\
&= 0.6(0.6 P(Q_1 = \text{☀}) + 0.1 P(Q_1 = \text{🌧})) + 0.1(0.4 P(Q_1 = \text{☀}) + 0.9 P(Q_1 = \text{🌧})) \\
&= 0.6(0.6(0.6 P(Q_0 = \text{☀}) + 0.1 P(Q_0 = \text{🌧})) + 0.1(0.4 P(Q_0 = \text{☀}) + 0.9 P(Q_0 = \text{🌧}))) \\
&+ 0.1(0.4(0.4 P(Q_0 = \text{☀}) + 0.1 P(Q_0 = \text{🌧})) + 0.9(0.4 P(Q_0 = \text{☀}) + 0.9 P(Q_0 = \text{🌧}))) \\
&= 0.6(0.6(0.6(0.8) + 0.1(0.2)) + 0.1(0.4(0.8) + 0.9(0.2))) \\
&+ 0.1(0.4(0.6(0.8) + 0.1(0.2)) + 0.9(0.4(0.8) + 0.9(0.2))) \\
&= 0.275
\end{aligned}
$$

Markov chains



- Suppose there's an 80% chance of sun on day 0. What is the probability of sun on day 3?

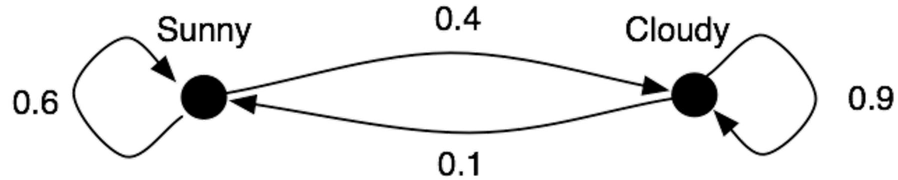
$$B = \begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix} \qquad w = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$$

$$(B^T)^3 w = \begin{bmatrix} 0.275 \\ 0.725 \end{bmatrix}$$

P(X_3 = sun)

P(X_3 = cloudy)

Stationary distributions

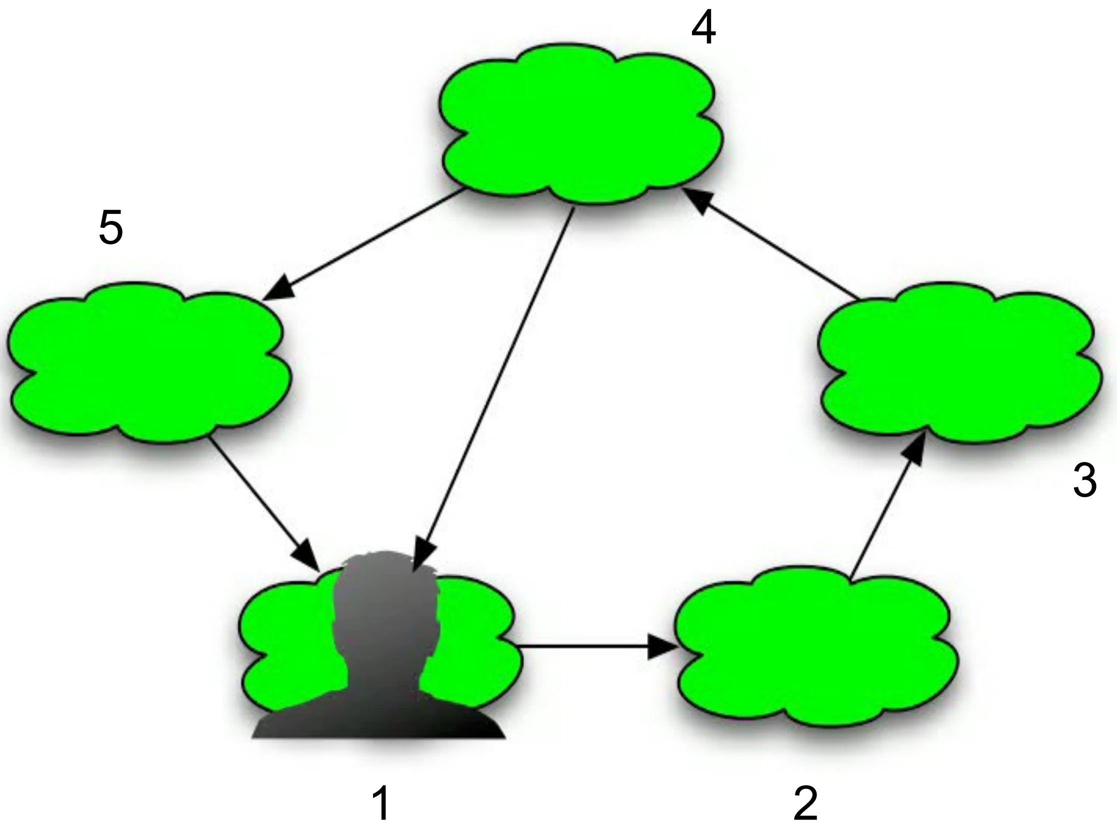


- For an *ergodic* chain, a *stationary distribution* exists
  - ergodic: all states are recurrent and aperiodic
  - stationary distribution: for large *t*, the probability of being in state *i* at time *t* depends *only* on the transition probabilities
  - the stationary distribution π is the vector satisfying

$$B^T \pi = \pi$$

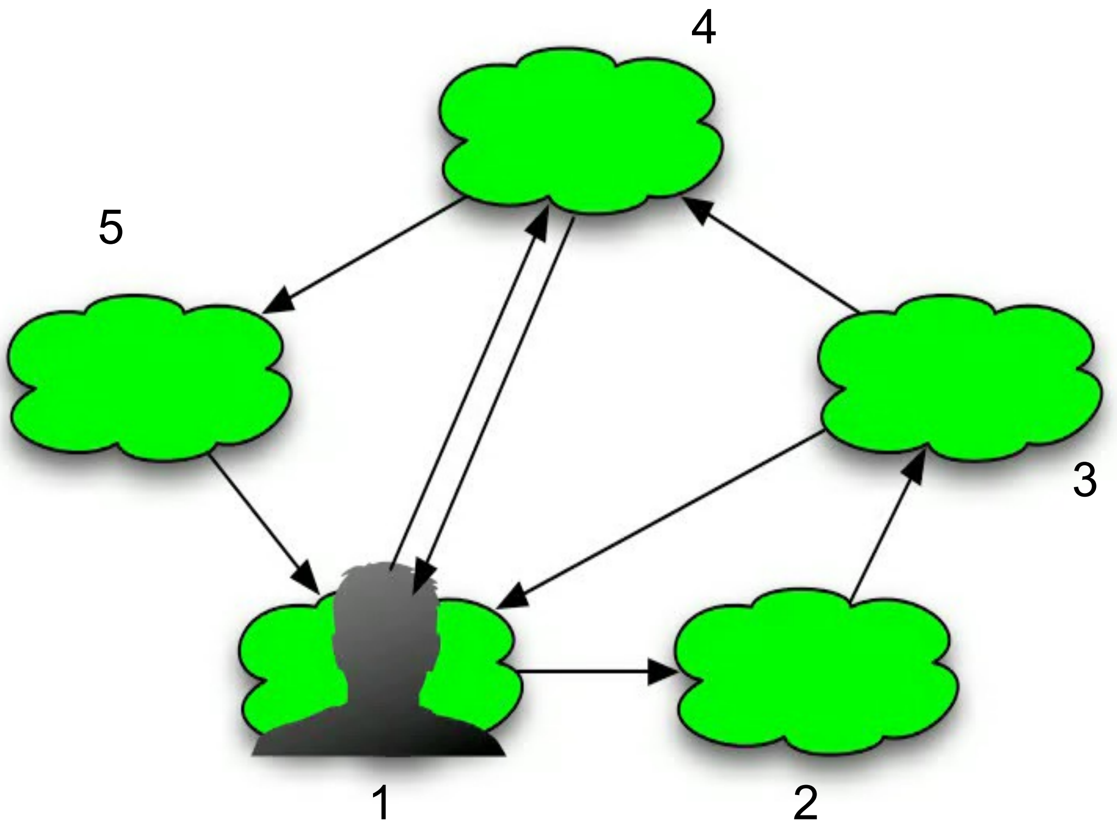
Q: At any moment in time, what's the probability that the frog is on pad 1?

A: P(Pad=1) = ?

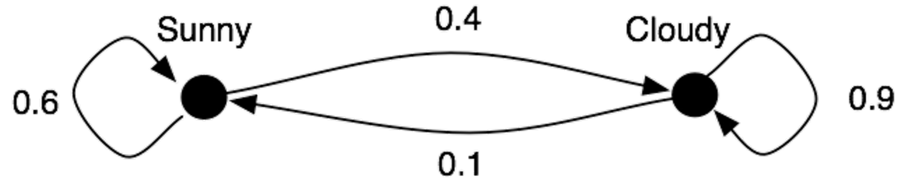


Q: At any moment in time, what's the probability that the frog is on pad 1?

A: P(Pad=1) = ?



Stationary distributions

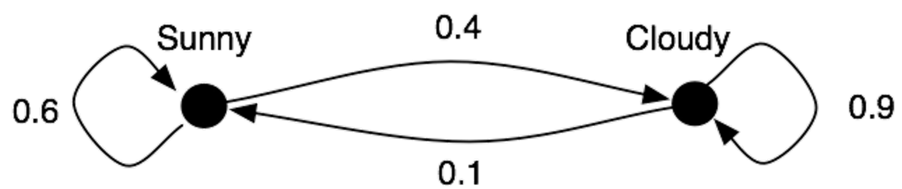


- For an *ergodic* chain, a *stationary distribution* exists
  - ergodic: all states are recurrent and aperiodic
  - stationary distribution: for large *t*, the probability of being in state *i* at time *t* depends *only* on the transition probabilities
  - the stationary distribution π is the vector satisfying

$$B^T \pi = \pi$$

How do we compute π?

Stationary distribution of Markov chain



- What is the stationary distribution of this chain?

```
>> % e.g. in Matlab:

>> [v d]=eigs([0.6 0.4; 0.1 0.9]',1)

v =
  -0.24253562503633
  -0.97014250014533

d =
     1

>> v/sum(v)

ans =

   0.20000000000000
   0.80000000000000
```

$$\pi = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}$$

Back to Markov Chain Monte Carlo (MCMC)

- Recall we want to estimate some distribution P(X)
  - But inference is too hard to compute it directly
- Basic idea: Construct a Markov Chain whose *stationary distribution* is exactly P(X)
  - Then take random walks on the Markov Chain
  - If we walk long enough, sampling from the Markov Chain is exactly equivalent to sampling from P(X)

MCMC in practice

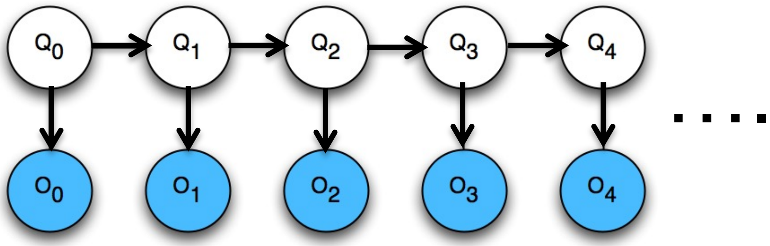
- Might take a long time for samples to be close to the stationary distribution – to “mix”
  - The “burn-in” or “warm-up” time
- The burn-in time depends on the structure of the chain and the transition probabilities
  - When might the burn-in time be particularly long?
- It’s possible to compute bounds on the burn-in time, by spectral analysis of the transition matrix
  - I.e. computing eigenvalues and eigenvectors of the Markov Chains’ transition matrix
  - But this is completely unhelpful in practice – why?

Practical solutions

- Construct a small number of identical Markov chains
  - Take random walks on each of the chains for a large number of time steps, starting from different initial states
  - Run the chains until the samples seem to be coming from the same distribution across all (or most) of the chains
  - Now use each of the chains to generate (estimated) samples from the posterior distribution

Statistical learning

Hidden Markov Models (HMMs)



- More formally, an HMM consists of:
  - Transition probabilities

$$p_{ij} = P(Q_{t+1} = j | Q_t = i)$$

  - Initial state distribution

$$w_i = P(Q_0 = i)$$

  - Emission probabilities

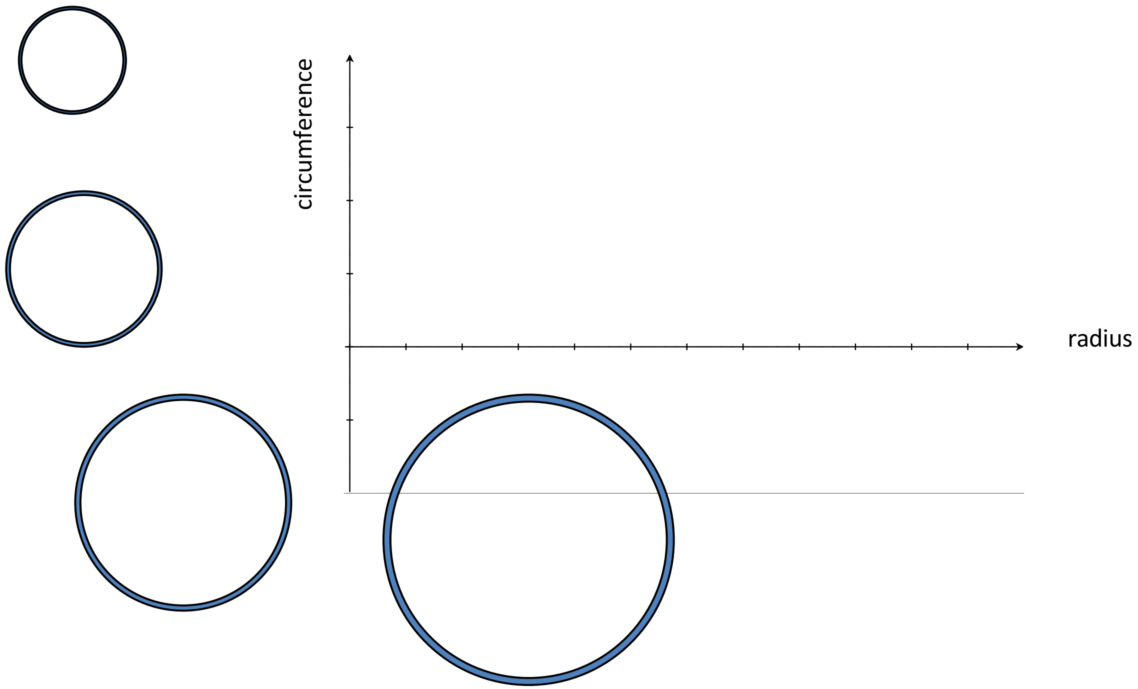
$$e_i(a) = P(O_t = a | Q_t = i)$$

Learning in General

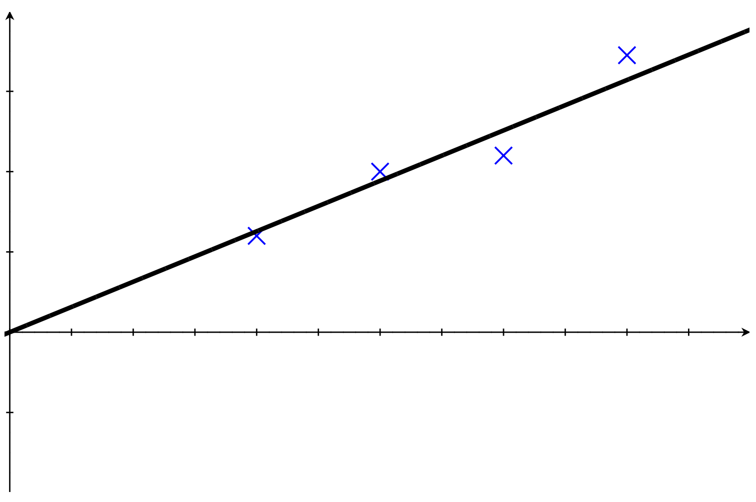
- Agent has made observations (data)
- Now must make sense of it (model hypotheses)
- *Why*?
  - Hypotheses alone may be important (e.g., in basic science)
  - For inference (e.g., forecasting)
  - To take sensible actions (decision making)
- A basic component of economics, social and physical sciences, engineering, …

An example

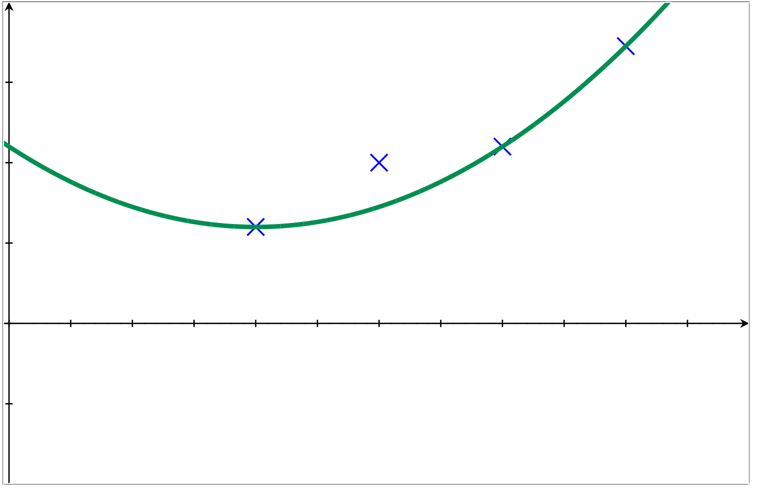
- Suppose we want to learn how to calculate the circumference of a circle from its radius.



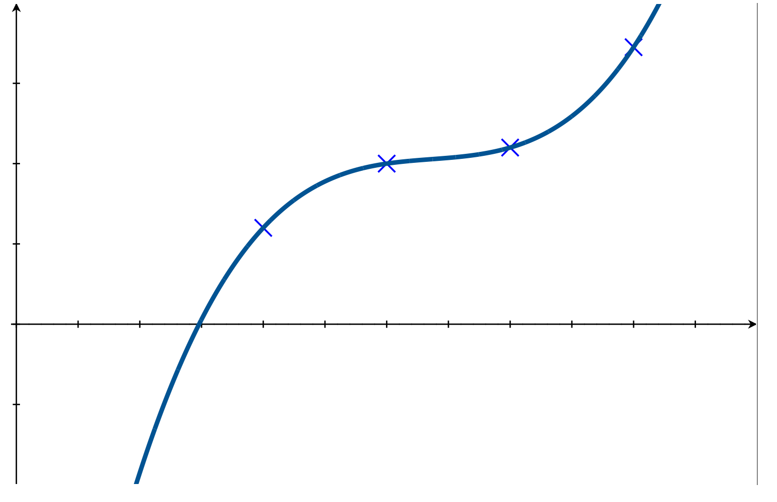
An example



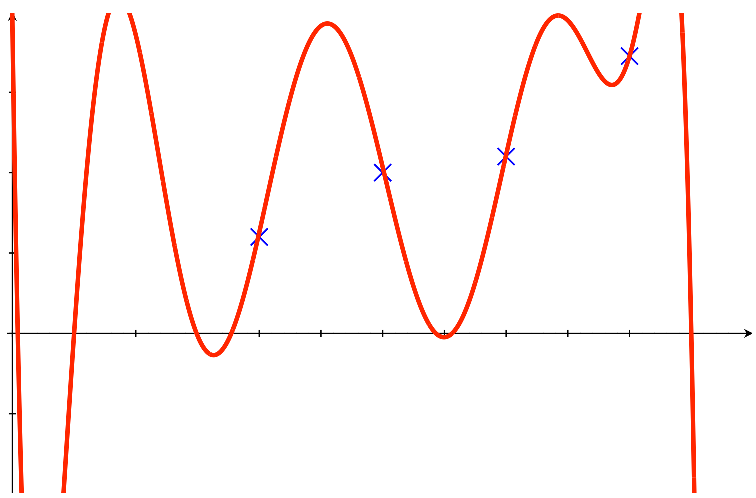
An example



An example

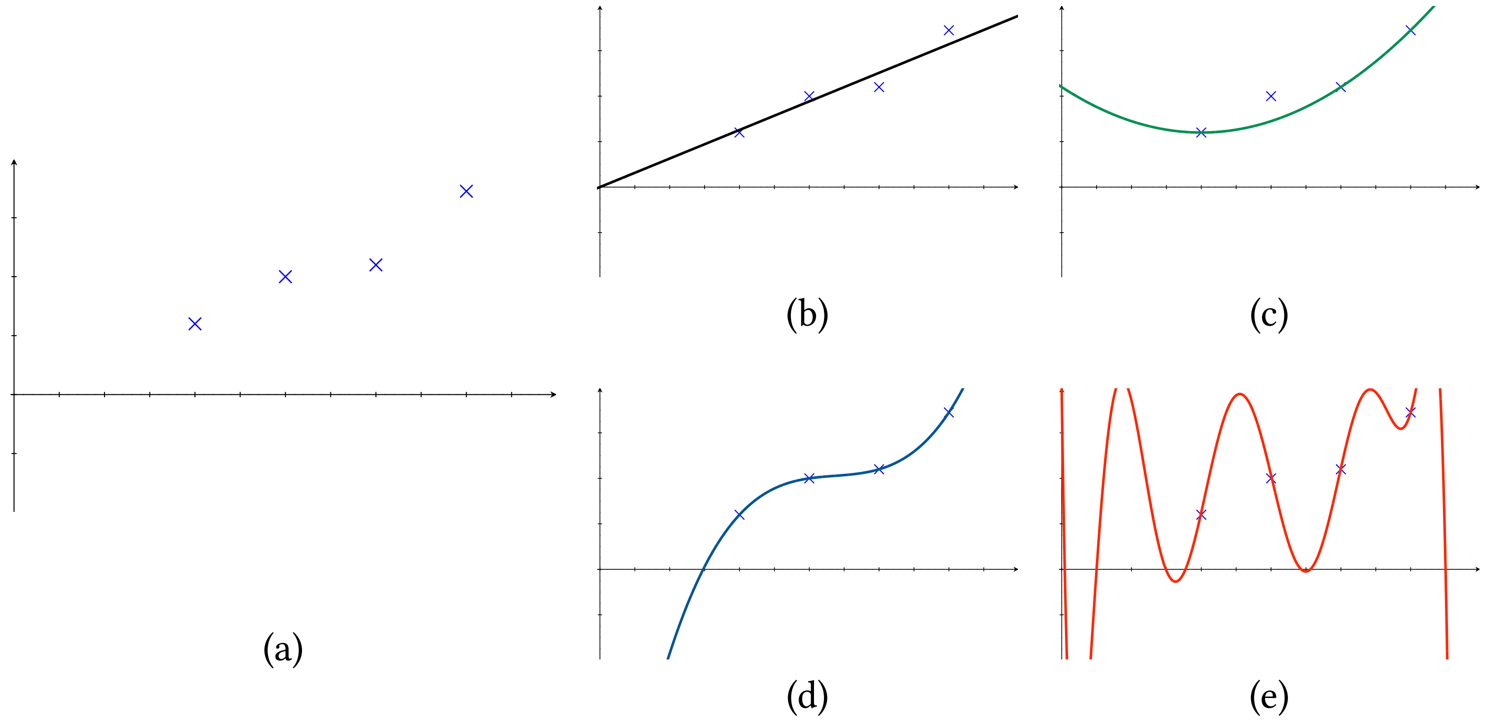


An example



An example

- Which is the best model?!

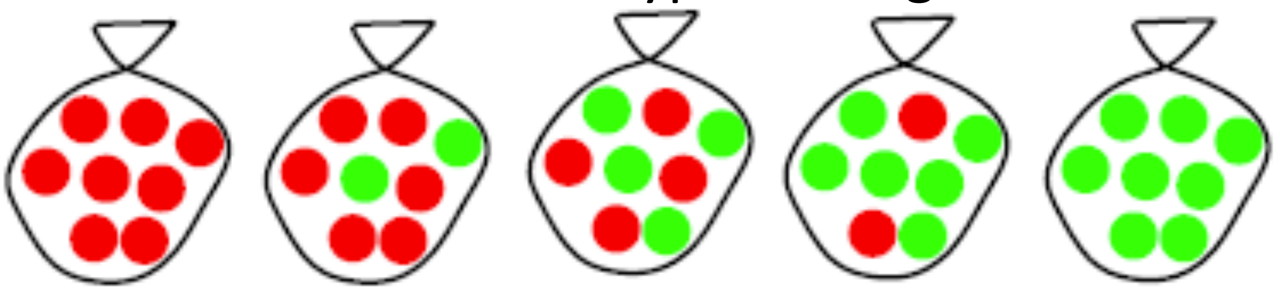


Machine Learning vs. Statistics

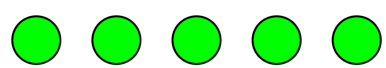
- Machine Learning ≈ automated statistics
- Today
  - Statistical learning (aka Bayesian learning)
  - Maximum likelihood (ML) learning
  - Maximum a posteriori (MAP) learning
  - Learning Bayes Nets (R&N 20.1-3)
- Future lectures try to do more with even less data
  - Neural nets
  - Support vector machines
  - …

Candy Example

- Candy comes in 2 flavors, cherry and lime
- Manufacturer makes 5 types of bags:

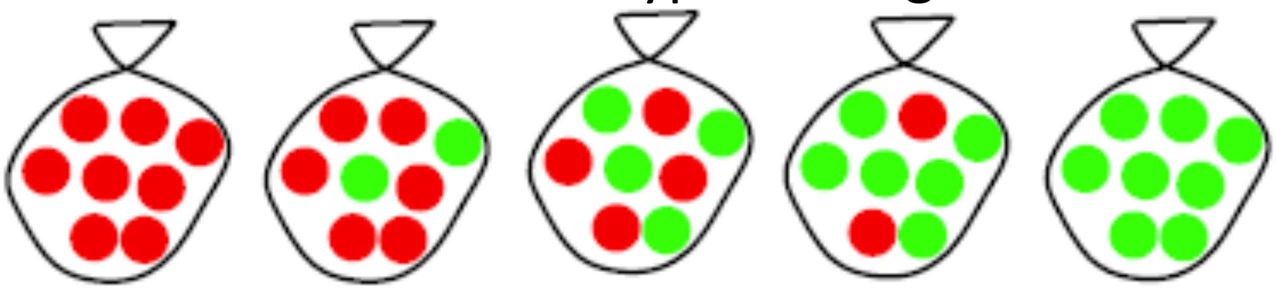


h1	h2	h3	h4	h5
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

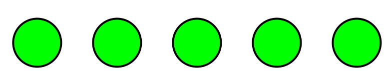
- h1 and h5 are equally common. h2 is twice as common as h1, h4 is twice as common as h5, and h3 is twice as common as h2.
- Suppose we draw 

Candy Example

- Candy comes in 2 flavors, cherry and lime
- Manufacturer makes 5 types of bags:

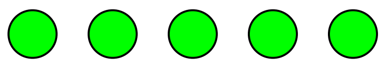


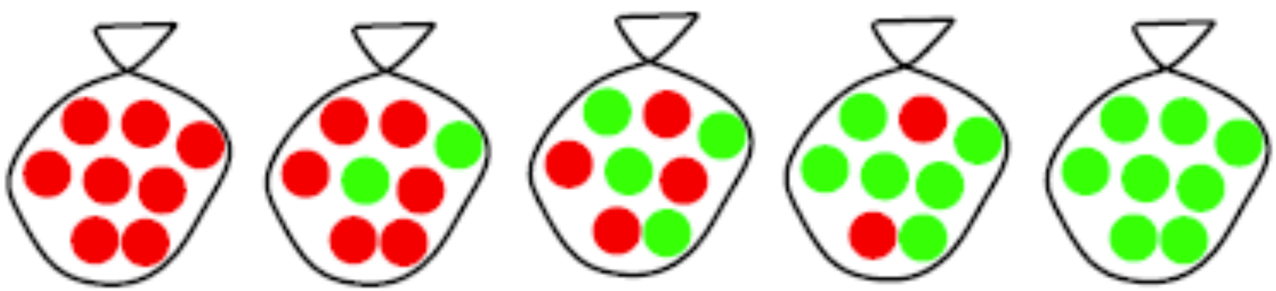
h1	**h2**	**h3**	**h4**	**h5**
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

- h1 and h5 are equally common. h2 is twice as common as h1, h4 is twice as common as h5, and h3 is twice as common as h2.
- Suppose we draw 

P(h1) = p(h5)
P(h2) = 2p(h1)
P(h4) = 2p(h5)
P(h3) = 2p(h2)
P(h1) + 2p(h1) + 4p(h1) + 2p(h1) + p(h1) = 1 => p(h1) = 0.1, p(h2) = 0.2…

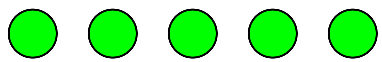
Bayesian Learning

- Main idea: Compute the probability of each hypothesis, given the data
- Data **d**: 
- Hypotheses: $h_1, \ldots, h_5$



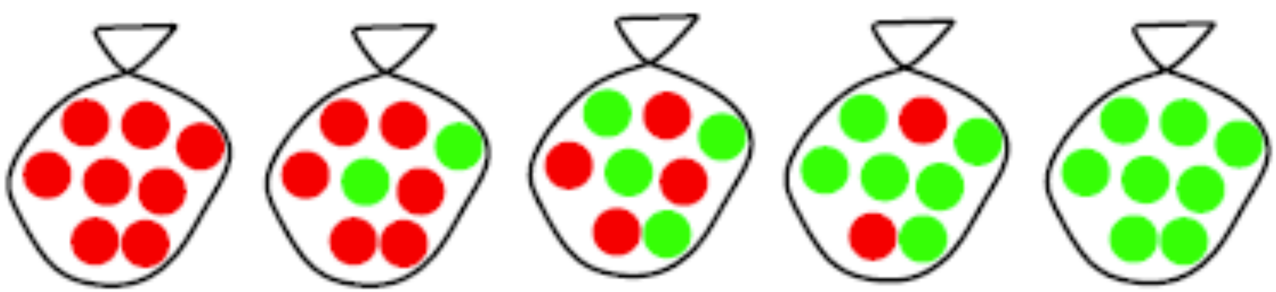
h1	**h2**	**h3**	**h4**	**h5**
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

Bayesian Learning

- Main idea: Compute the probability of each hypothesis, given the data
- Data **d**: 
- Hypotheses: $h_1,\ldots,h_5$

$P(h_i|d)$ We want this…

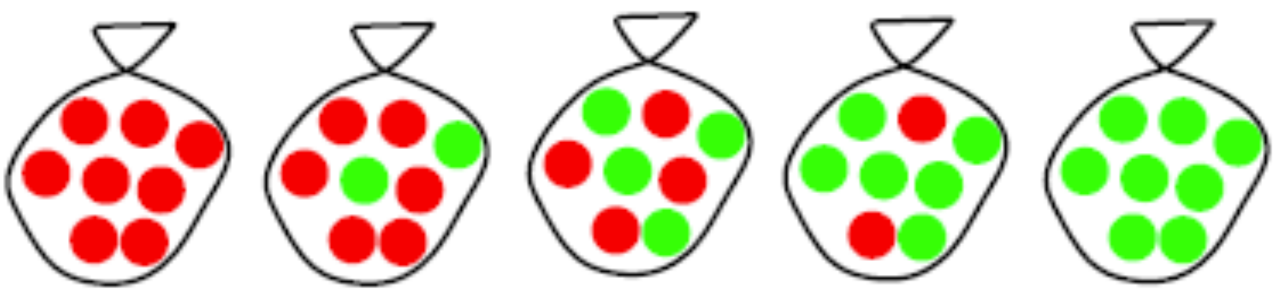
$P(d|h_i)$ But all we have is this!



h1	**h2**	**h3**	**h4**	**h5**
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

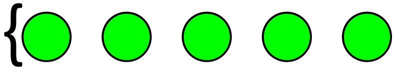
Using Bayes' Rule

- $P(h_i|\mathbf{d}) = \alpha\, P(\mathbf{d}|h_i)\, P(h_i)$ is the **posterior**
  - (Recall, $1/\alpha = P(\mathbf{d}) = \sum_i P(\mathbf{d}|h_i)\, P(h_i)$)
- $P(\mathbf{d}|h_i)$ is the **likelihood**
- $P(h_i)$ is the **hypothesis prior**



h1	**h2**	**h3**	**h4**	**h5**
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

Computing the Posterior

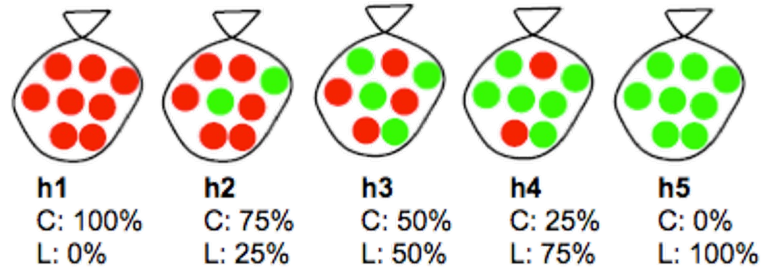
- Assume draws are independent
- Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$
- $\mathbf{d} = \{$  $\}$

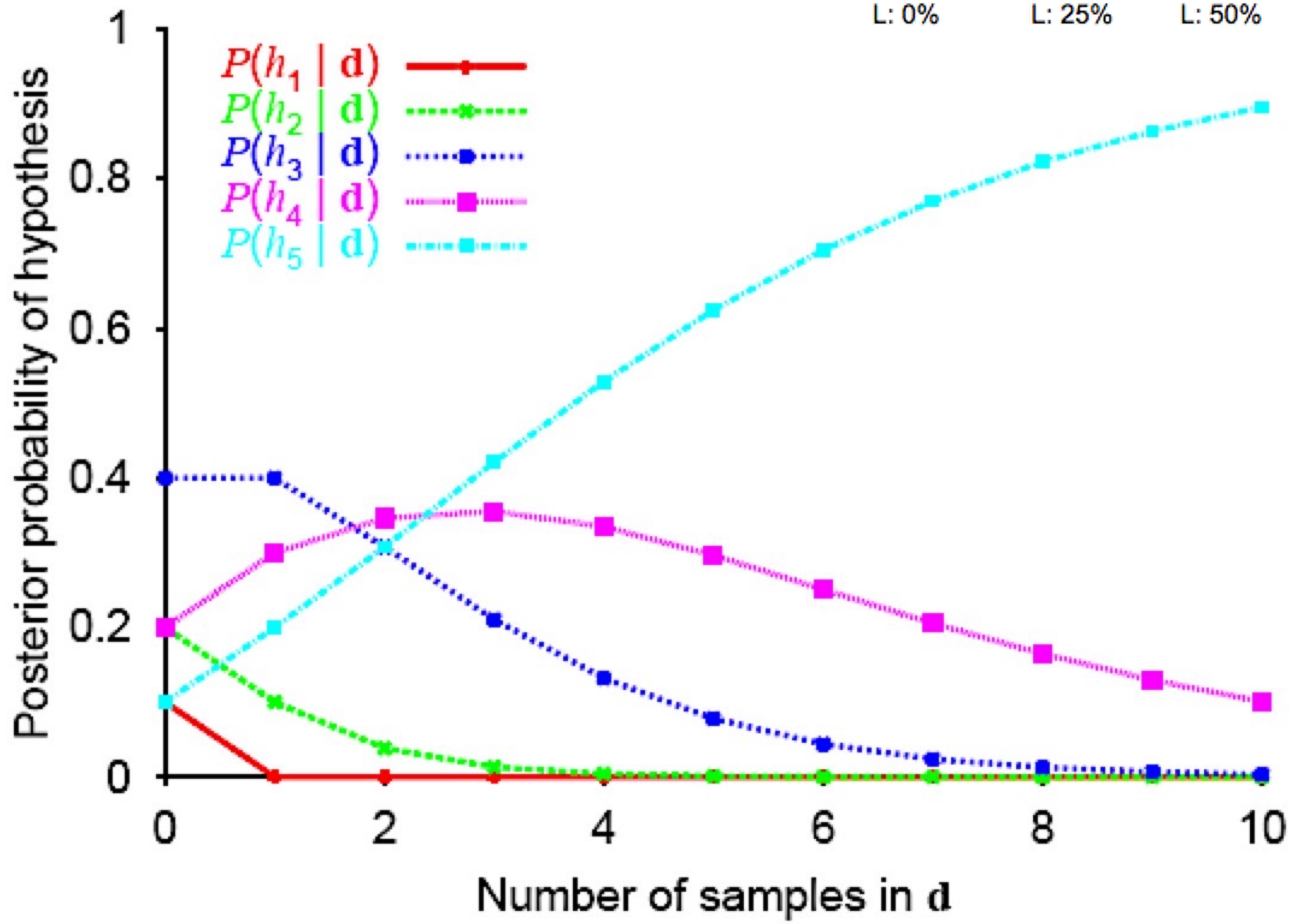
$P(d\|h_1) = 0$	$P(\mathbf{d}\|h_1)P(h_1) \approx 0$	$P(h_1\|\mathbf{d}) = 0$
$P(d\|h_2) = 0.25^5$	$P(\mathbf{d}\|h_2)P(h_2) \approx 1.9\text{e-}4$	$P(h_2\|\mathbf{d}) \approx 1.2\text{e-}3$
$P(d\|h_3) = 0.5^5$	$P(\mathbf{d}\|h_3)P(h_3) \approx 1.2\text{e-}2$	$P(h_3\|\mathbf{d}) \approx 0.078$
$P(d\|h_4) = 0.75^5$	$P(\mathbf{d}\|h_4)P(h_4) \approx 4.7\text{e-}2$	$P(h_4\|\mathbf{d}) \approx 0.29$
$P(d\|h_5) = 1^5$	$P(\mathbf{d}\|h_5)P(h_5) \approx 0.1$	$P(h_5\|\mathbf{d}) \approx 0.62$
	$P(\mathbf{d}) \approx 0.16$	

Posterior Hypotheses

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

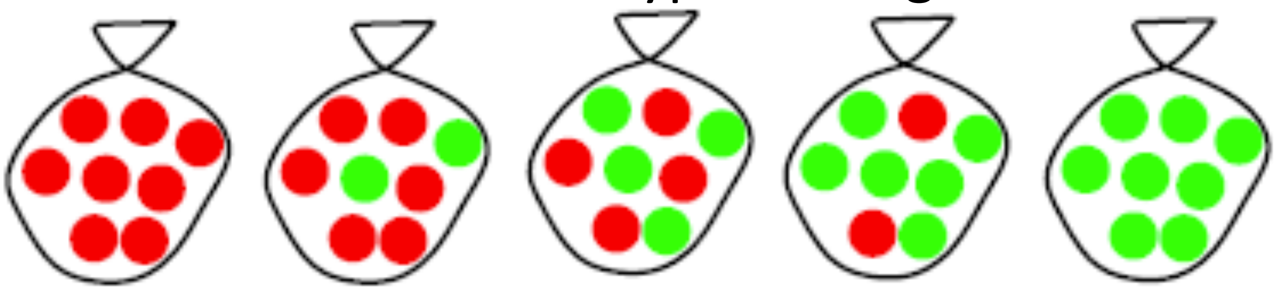
Data: All our samples are limes.



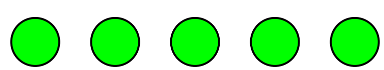


Candy Example

- Candy comes in 2 flavors, cherry and lime
- Manufacturer makes 5 types of bags:

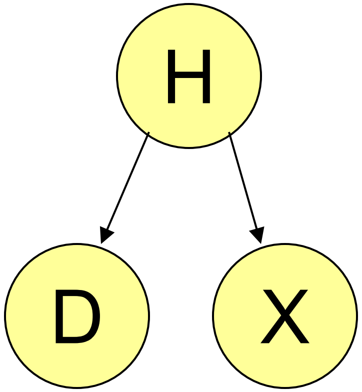


h1	**h2**	**h3**	**h4**	**h5**
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

- h1 and h5 are equally common. h2 is twice as common as h1, h4 is twice as common as h5, and h3 is twice as common as h2.
- Suppose we draw 
- Which bag are we holding? **Which flavor will we draw next?**

Predicting the Next Draw

- $P(X|\mathbf{d}) = \sum_i P(X|h_i,\mathbf{d})P(h_i|\mathbf{d})$
  $= \sum_i P(X|h_i)P(h_i|\mathbf{d})$



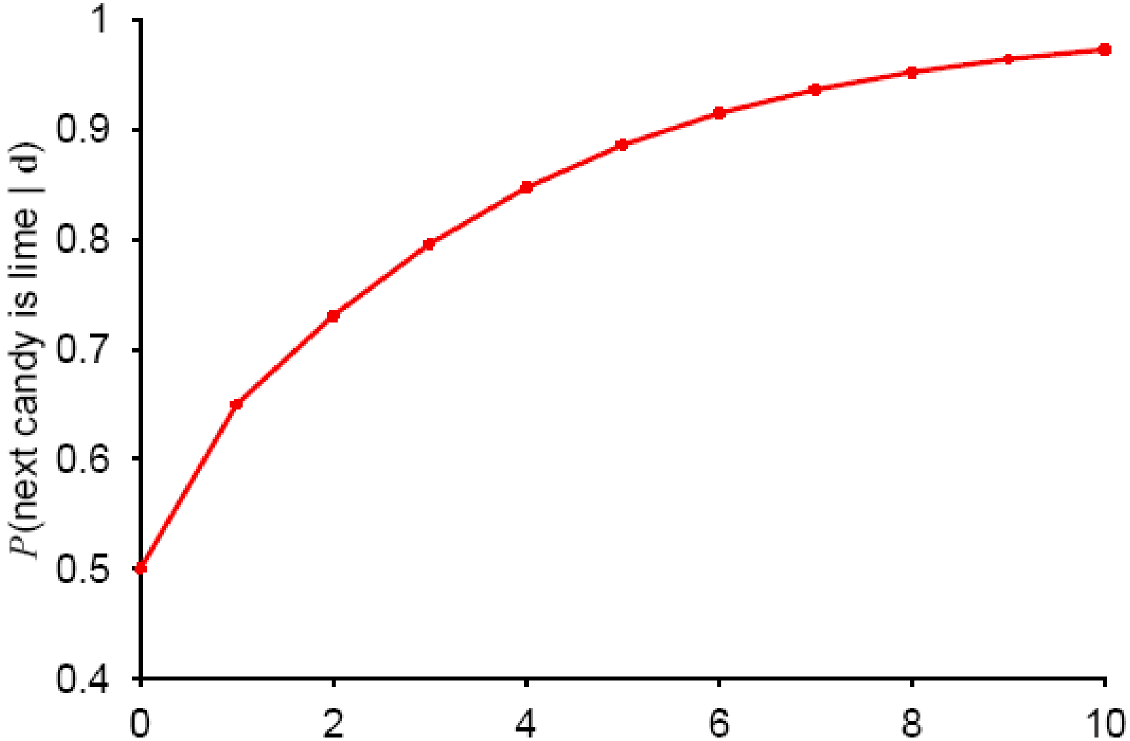
Probability that next candy drawn is a lime

$$\left.\begin{array}{ll} P(h_1|\mathbf{d}) = 0 & P(X|h_1) = 0 \\ P(h_2|\mathbf{d}) \approx 1.2\text{e-}3 & P(X|h_2) = 0.25 \\ P(h_3|\mathbf{d}) \approx 0.078 & P(X|h_3) = 0.5 \\ P(h_4|\mathbf{d}) \approx 0.29 & P(X|h_4) = 0.75 \\ P(h_5|\mathbf{d}) \approx 0.62 & P(X|h_5) = 1 \end{array}\right\} \quad P(X|\mathbf{d}) \approx 0.890$$

P(Next Candy is Lime | d)

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

Data: All our samples are limes.



Number of candies drawn so far (all limes)

Properties of Bayesian Learning

- If exactly one hypothesis is correct, then the posterior probability of the correct hypothesis will tend toward 1 as more data is observed
- The effect of the prior distribution decreases as more data is observed

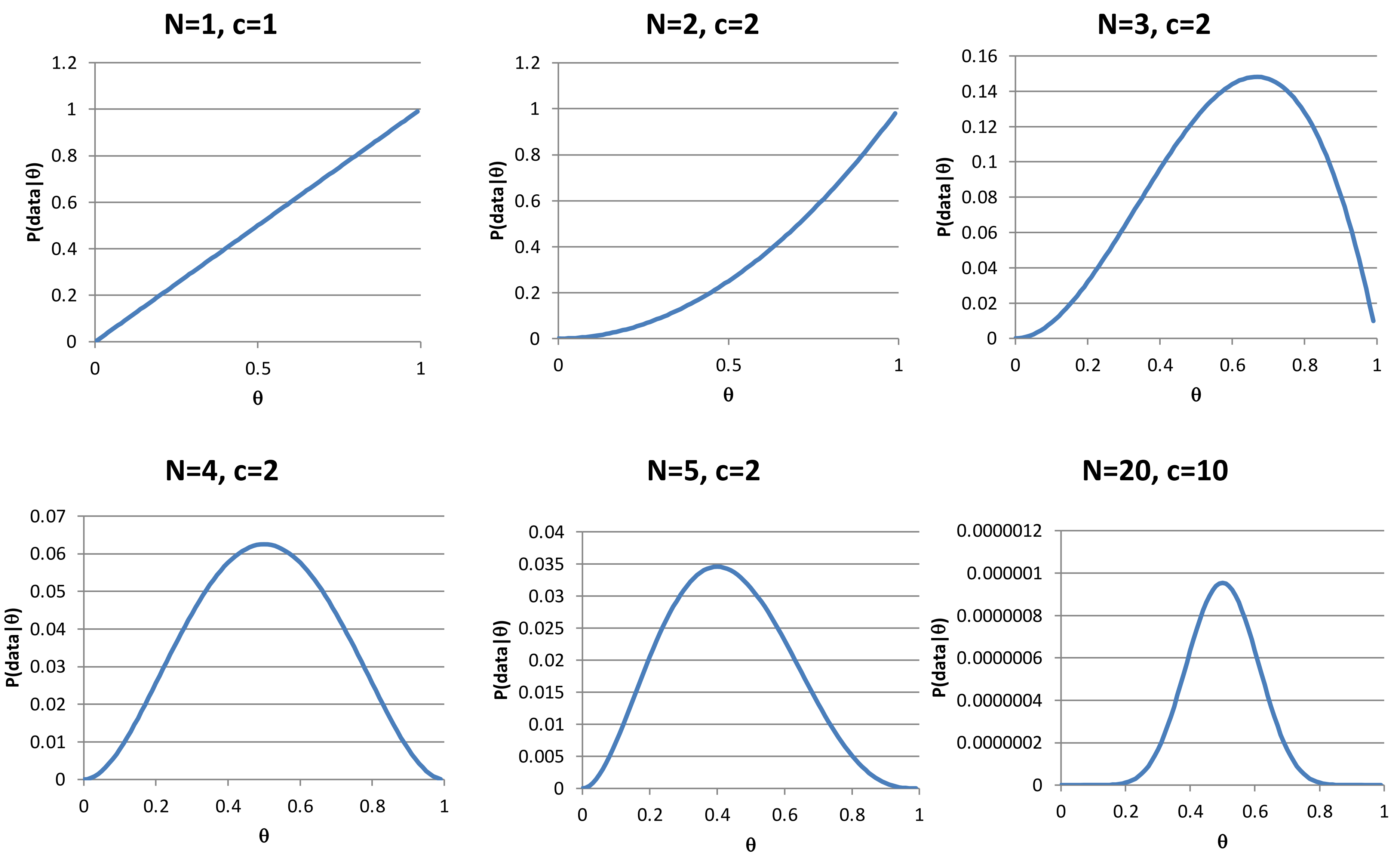
Hypothesis Spaces often Intractable

- But: a hypothesis is a joint probability table over state variables
- What if we don't have a reasonable prior?
- In practice, we'll need to use additional information about the structure of the model
  - E.g. conditional independent assumptions
  - Some parametric form for the hypotheses

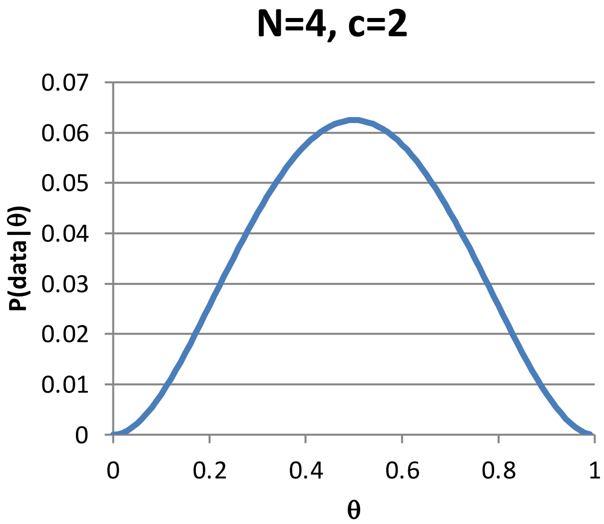
Learning the bias of a coin

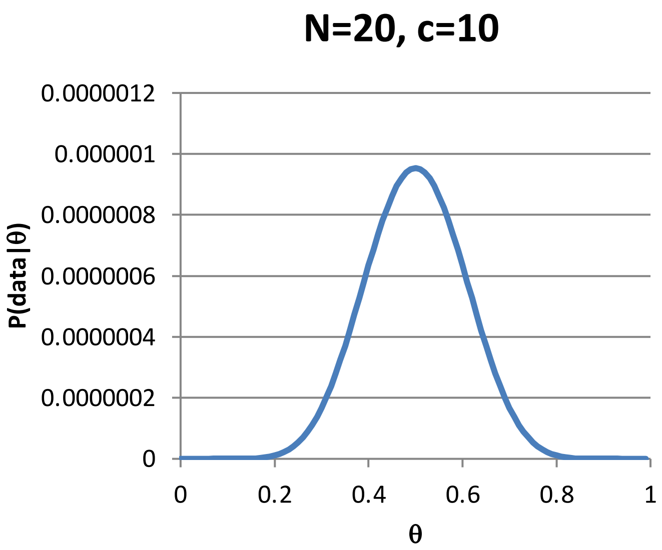
- We have a coin with unknown P(H)
- Let P(H) be θ (hypothesis)
- Flips are i.i.d. – independent, identically distributed
- We flip N times to get a sequence of outcomes D. Of these, c flips are heads (data).
- Consider $P(D \mid \theta)$, the *likelihood of the data given the model*.

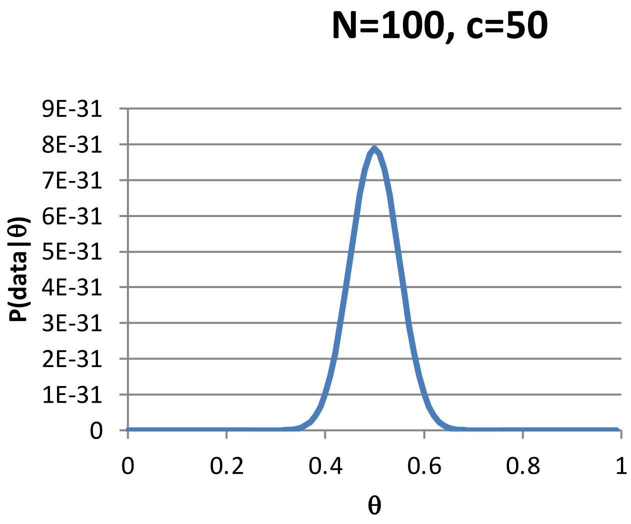
We flip coin N times, c of these are heads. Coin has unknown θ=P(H).



We flip coin N times, c of these are heads. Coin has unknown θ=P(H).







Maximum Likelihood Estimation (MLE)

- Say we have a set of training samples $D=(x_1,\ldots,x_N)$
- We've decided on a parametric model for the distribution, having parameters θ
- We define a likelihood function measuring the probability of the data for a given model θ,

$$P(D|\theta) = \prod_{i}^{N} P(D_i|\theta)$$

- In MLE, we want to find a model that *maximizes the probability of the observed data given the model,*

$$\theta^* = \arg\max_{\theta} P(D|\theta)$$

Other Closed-Form MLE results

- **Multi-valued variables:** take fraction of counts for each value
- **Continuous Gaussian distributions:** take average value as mean, standard deviation of data as standard deviation

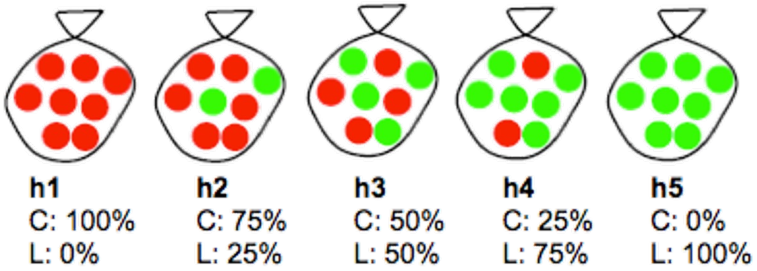
Maximum Likelihood Properties

- As the number of data points approaches infinity, the MLE approaches the true estimate
- With little data, MLEs can vary wildly

MLE in candy example

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

Data: All our samples are limes.

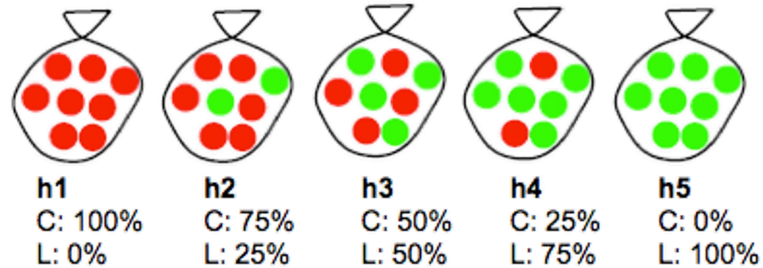


- What is the MLE?

MLE in candy example

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

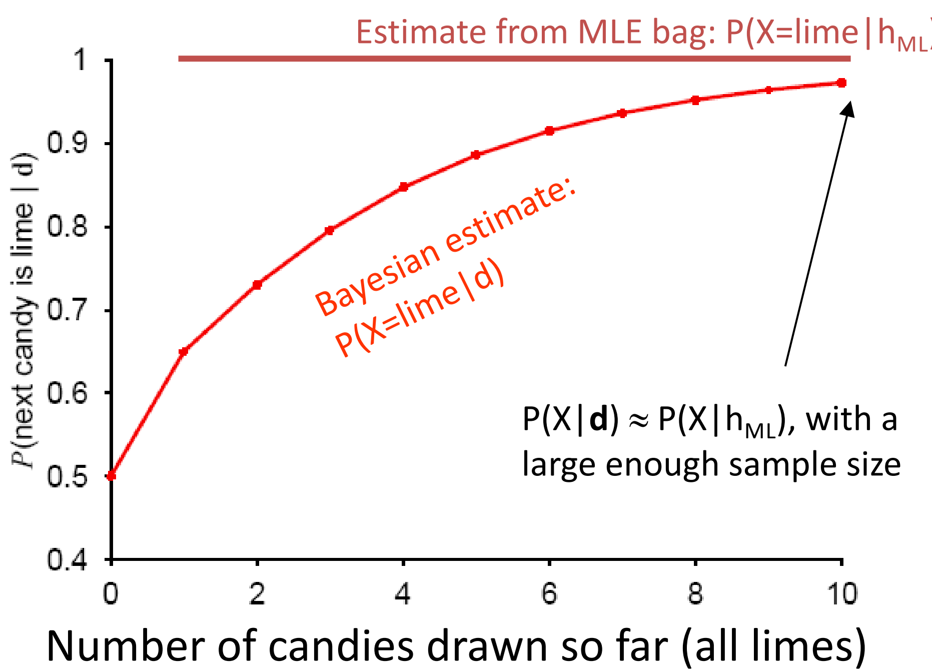
Data: All our samples are limes.



- What is the MLE?

$$h_{ML} = \operatorname{argmax}_i P(\mathbf{d} | h_i)$$

What is probability next candy is lime?



Disadvantages of ML Estimation

- Tells us nothing about the certainty of our estimates
  - Note that we get exactly the same answer whether we flip 1 head out of 3, or 1,000 out of 3,000
- If we have no observations, can't estimate anything
  - Or: if we have a large number of variables, some values will never be seen and we can conclude nothing about them
- No way to incorporate prior evidence
  - E.g. that most coins are unbiased (or not very biased)

Maximum A Posteriori Estimation

- **Maximum a posteriori** (MAP) estimation
- Idea: use the hypothesis prior to get a better initial estimate than ML, without full Bayesian estimation
  - "Most coins I've seen have been fair coins, so I won't let the first few tails sway my estimate much"
  - "Now that I've seen 100 tails in a row, I'm pretty sure it's not a fair coin anymore"

Maximum A Posteriori

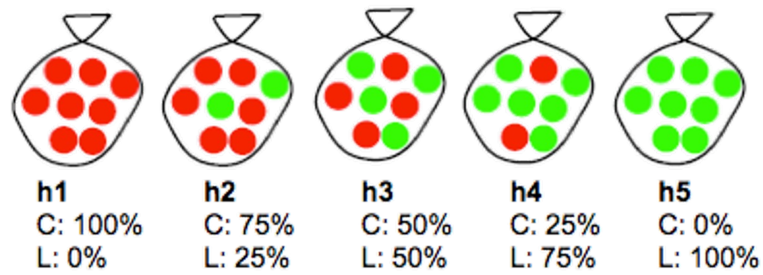
- $P(\theta|\mathbf{d})$ is the posterior probability of the hypothesis, given the data
- $\text{argmax}_\theta\, P(\theta|\mathbf{d})$ is known as the **maximum a posteriori** (MAP) estimate
- Posterior of hypothesis θ given data $\mathbf{d}=\{d_1,\ldots,d_N\}$
  - $P(\theta|\mathbf{d}) = 1/\alpha\; P(d|\theta)\; P(\theta)$
  - Max over θ doesn't affect α
  - So MAP estimate is $\text{argmax}_\theta\, P(d|\theta)\; P(\theta)$

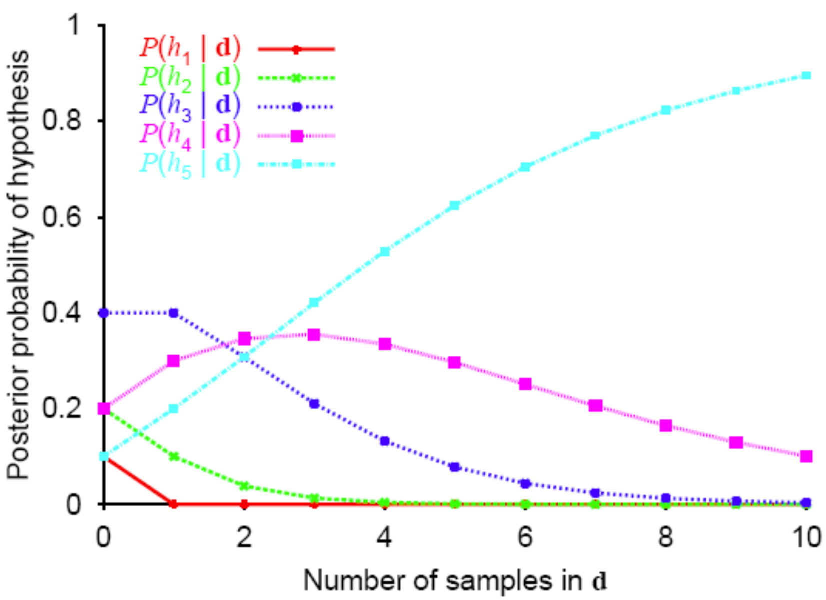
Maximum a Posteriori

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

Data: All our samples are limes.

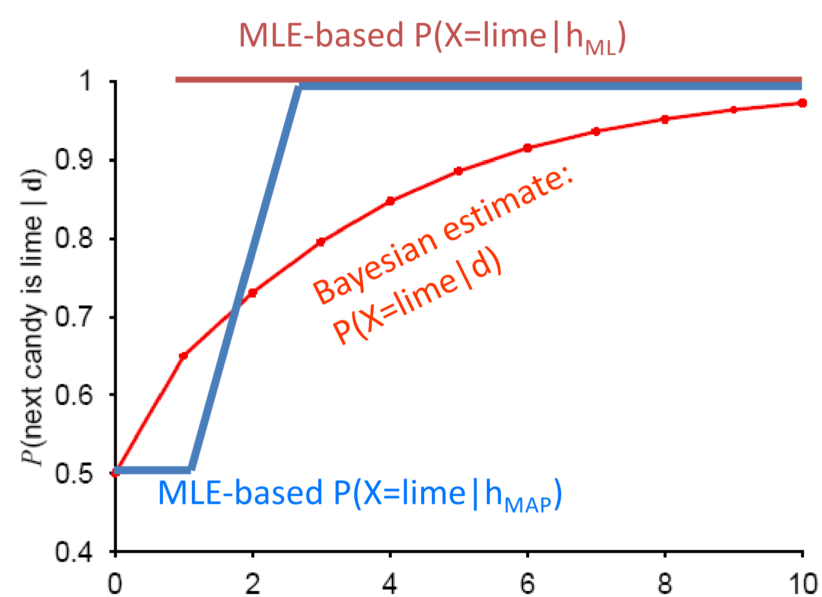
$h_{MAP} = \text{argmax}_i\ P(h_i|\mathbf{d})$ $h_{ML} = \text{argmax}_i\ P(\mathbf{d}|h_i)$





# samples:	0	1	2	3	4	5	6	7	8	9	10
h_{MAP}	h3	h3	h4	h5	h5	h5	h5	h5	h5	h5	h5
h_{ML}		h5	h5	h5	h5	h5	h5	h5	h5	h5	h5

What is probability next candy is lime?



Advantages of MAP and MLE over Bayesian estimation

- Involves an *optimization* rather than a large summation
  - Local search techniques
- For some types of distributions, there are *closed-form* solutions that are easily computed

Next class

- Morea bout ML and MAP

MLE in Bayes Networks

- The likelihood function can be factored into a product over variables and over exemplars,

$$P(D|\theta) = \prod_i^N P(D_i|\theta) \qquad \text{where} \qquad P(D|\theta_j) = \prod_i^N P(d_j^i|\mathrm{Pa}(X_j))$$

where $\theta=(\theta_1, \ldots, \theta_M)$, θ_j is the CPD for variable X_j, $D=(d_1, d_2, \ldots, d_N)$, and each $d_i=(d_i^1, d_i^2, \ldots d_i^M)$

- If all variables can be observed, each of these factors can be maximized individually
  - So we can estimate parameters of each variable's probability distribution individually
- **Key result: ML estimate of Bayes Net parameters is given by the set of ML estimates of CPDs of each variable.**

Maximum Likelihood for BN

- For any BN, the ML parameters of any CPT can be derived by the fraction of observed values in the data, conditioned on matched parent values

N=1000

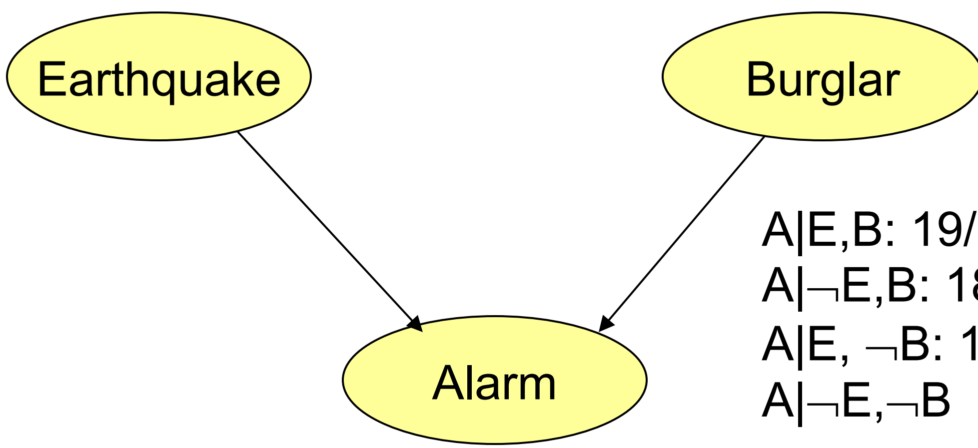
E: 500

P(E) = 0.5

B: 200

P(B) = 0.2

E	B	P(A|E,B)
T	T	0.95
F	T	0.95
T	F	0.34
F	F	0.003



A|E,B: 19/20
A|¬E,B: 188/200
A|E, ¬B: 170/500
A|¬E,¬B : 1/380

Bayes Net ML Algorithm

- Input: BN with nodes $X_1,\ldots,X_n$, dataset $\mathbf{D}=(\mathbf{d}_1,\ldots,\mathbf{d}_N)$
  - Each $d_i=(d_i^1, d_i^2, \ldots d_i^M)$ is a sample with values for <u>all</u> variables
- For each node X with parents $Y_1,\ldots,Y_k$:
  - For all $y_1 \in Val(Y_1),\ldots, y_k \in Val(Y_k)$
    - For all $x \in Val(X)$
      - Count the number of times $(X=x, Y_1=y_1,\ldots, Y_k=y_k)$ is observed in **D**. Let this be m_x
    - Count the number of times $(Y_1=y_1,\ldots, Y_k=y_k)$ is observed in **D**. Let this be m. (note $m=\Sigma_x m_x$)
    - Set $P(x|y_1,\ldots,y_k) = m_x / m$ for all $x \in Val(X)$