Unit 1

Function Estimation

Probability theory

Bayes' Nets

Hypotheses

- A hypothesis is a guess at what an appropriate action (or set of actions) is
  - Some actions are obvious
  - Some require experimentation to find the best fit
- Traffic light hypothesis?
- In this course, we will talk about different ways to make hypotheses

Pick a number

- Let’s play a game…

  Consistent

- What are the variables?
- What is your hypothesis about the game function?

Evaluating hypotheses

- It is important to validate your hypothesis
- May develop hypothesis by training on some data
  - Very important to have different testing data
- Quantitative evaluation
  - I got x% of the test scenarios correct
  - The average car takes x seconds to go down the street
- Qualitative evaluation
  - My hypothesis explains such-and-such phenomenon, where the competing hypothesis doesn’t

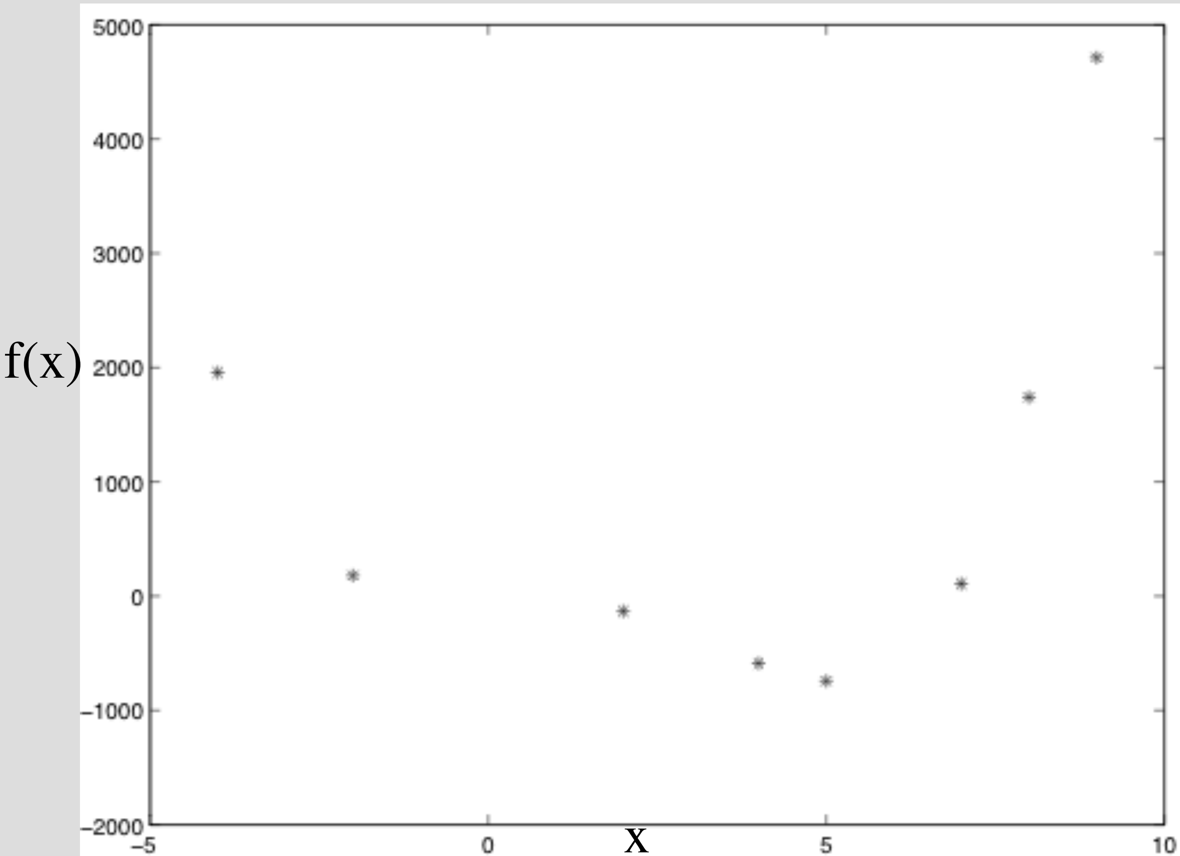
Inductive Learning

- A form of supervised learning
- Hypothesizes a function mapping inputs to correct outputs
- Presented with example Tuples (x,f(x)):
  - x is input
  - f(x) is output of function applied to x

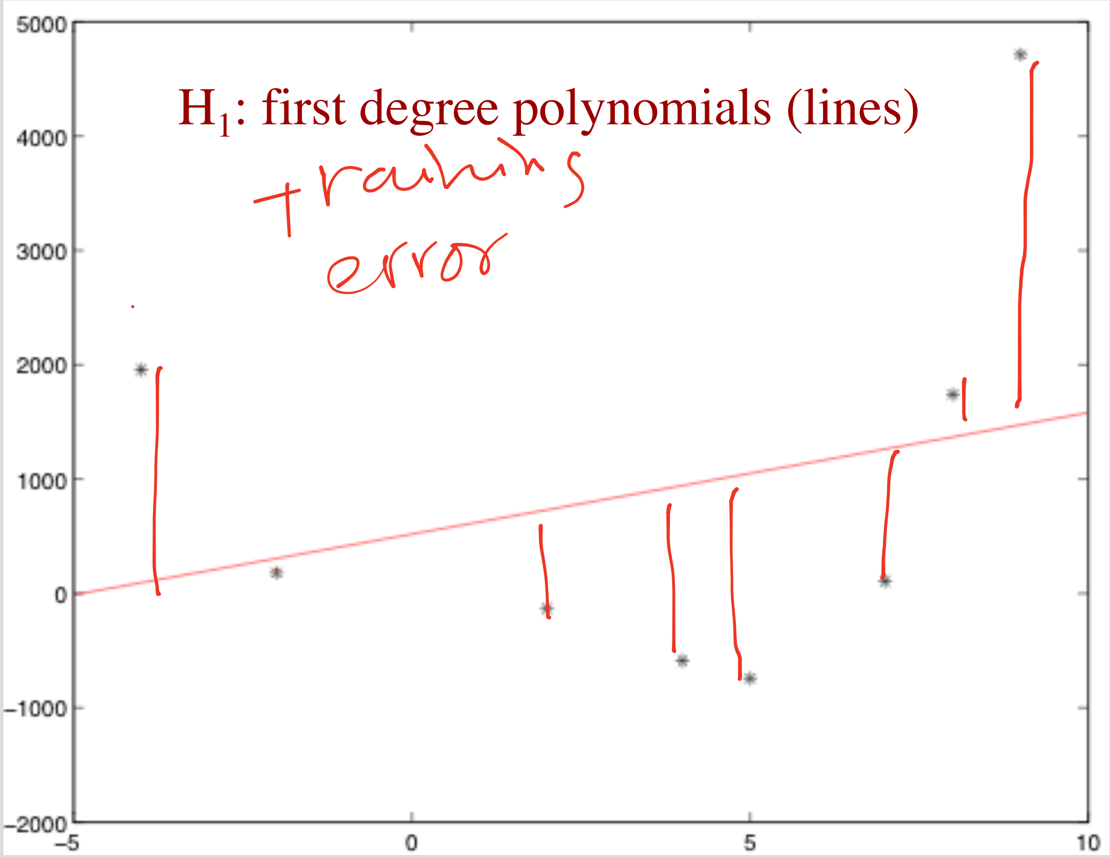
Hypotheses

- A hypothesis h is an approximation of the true function f that you are trying to learn
- Pure inductive inference
  - Given {(x,f(x))}, return h(x) which approximates f(x)
- The space of all hypothesis functions is **H**
  - This is chosen by the person designing the learning algorithm

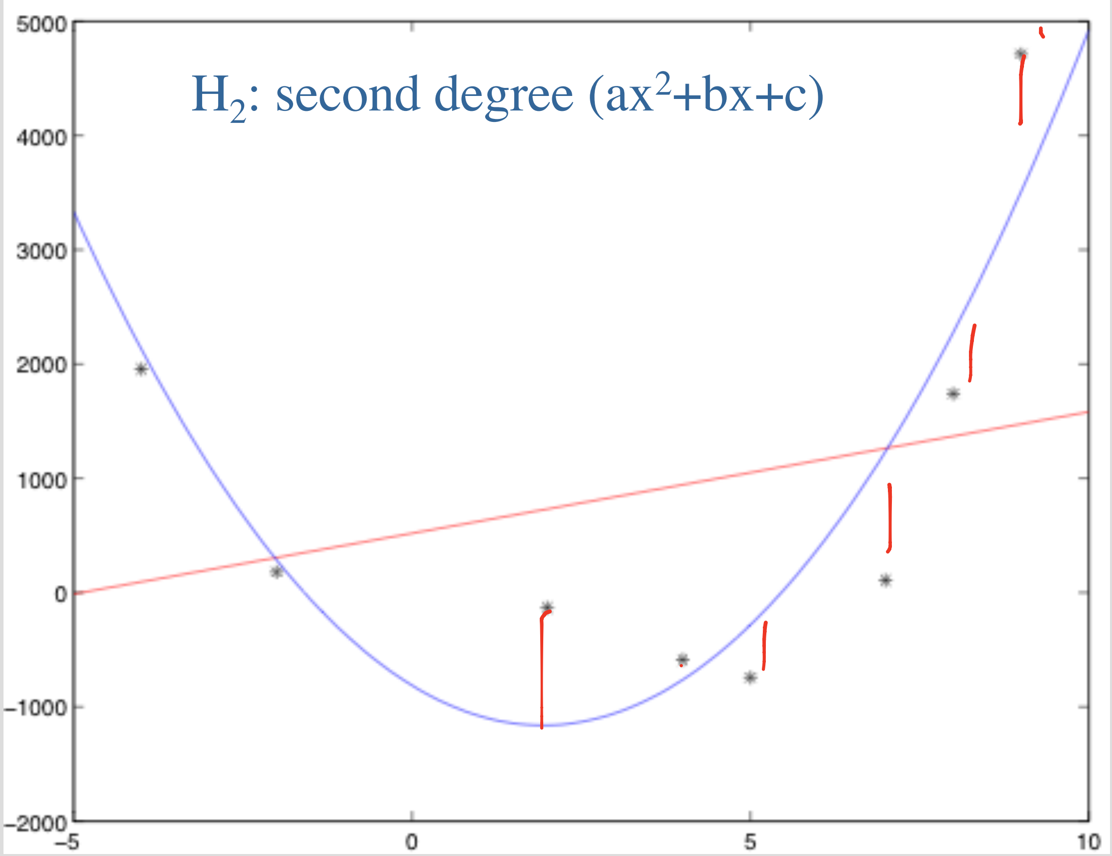
Hypothesis spaces & Inductive learning



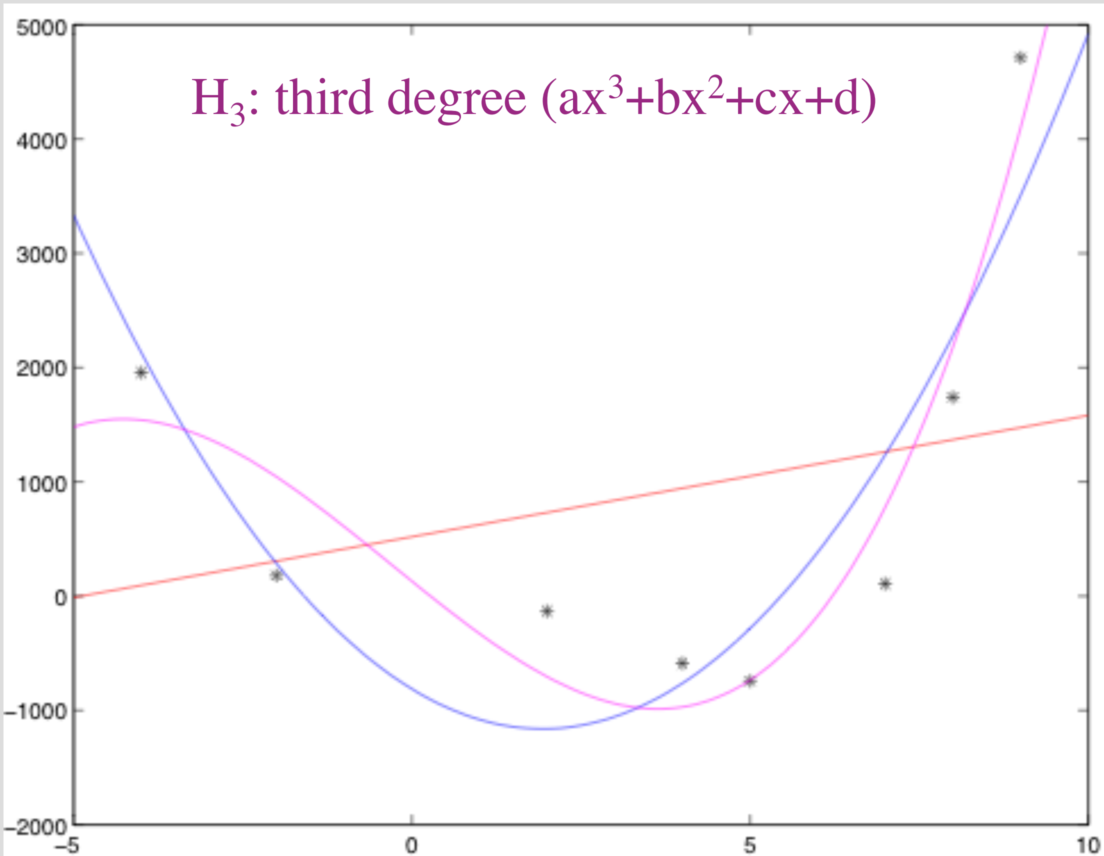
Hypothesis spaces & Inductive learning



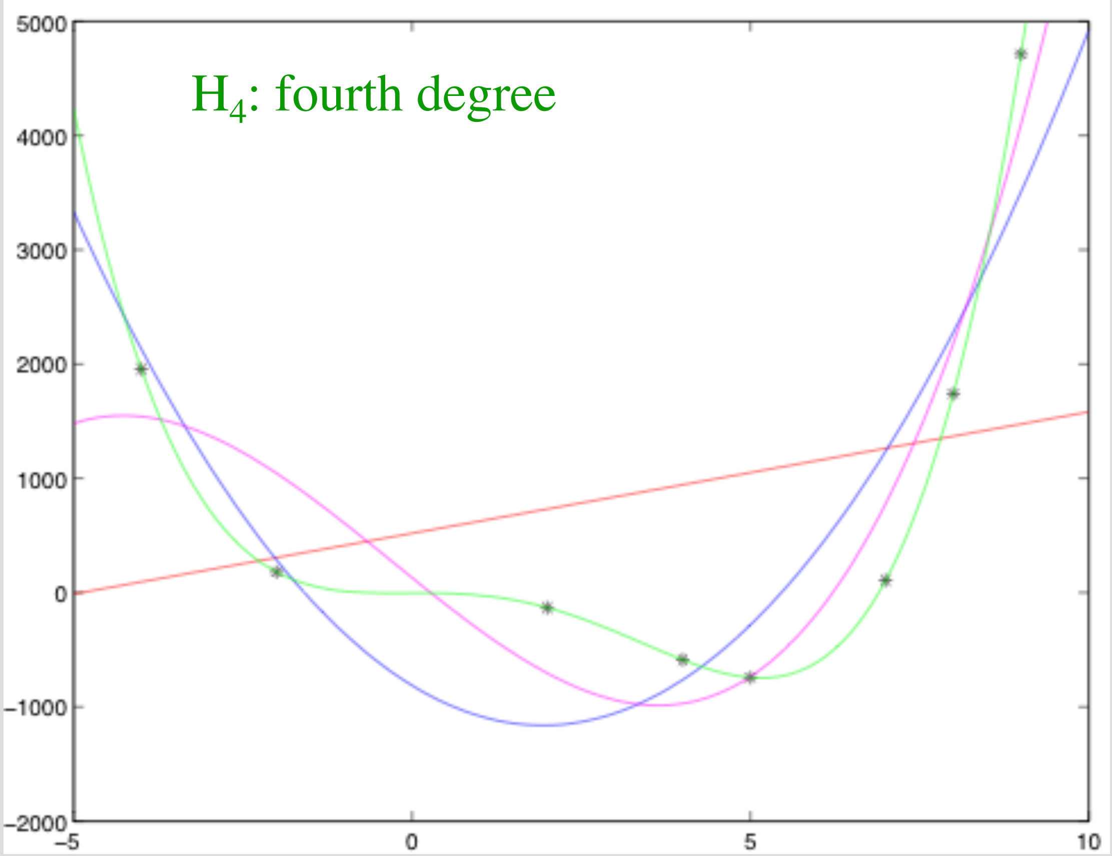
Hypothesis spaces & Inductive learning



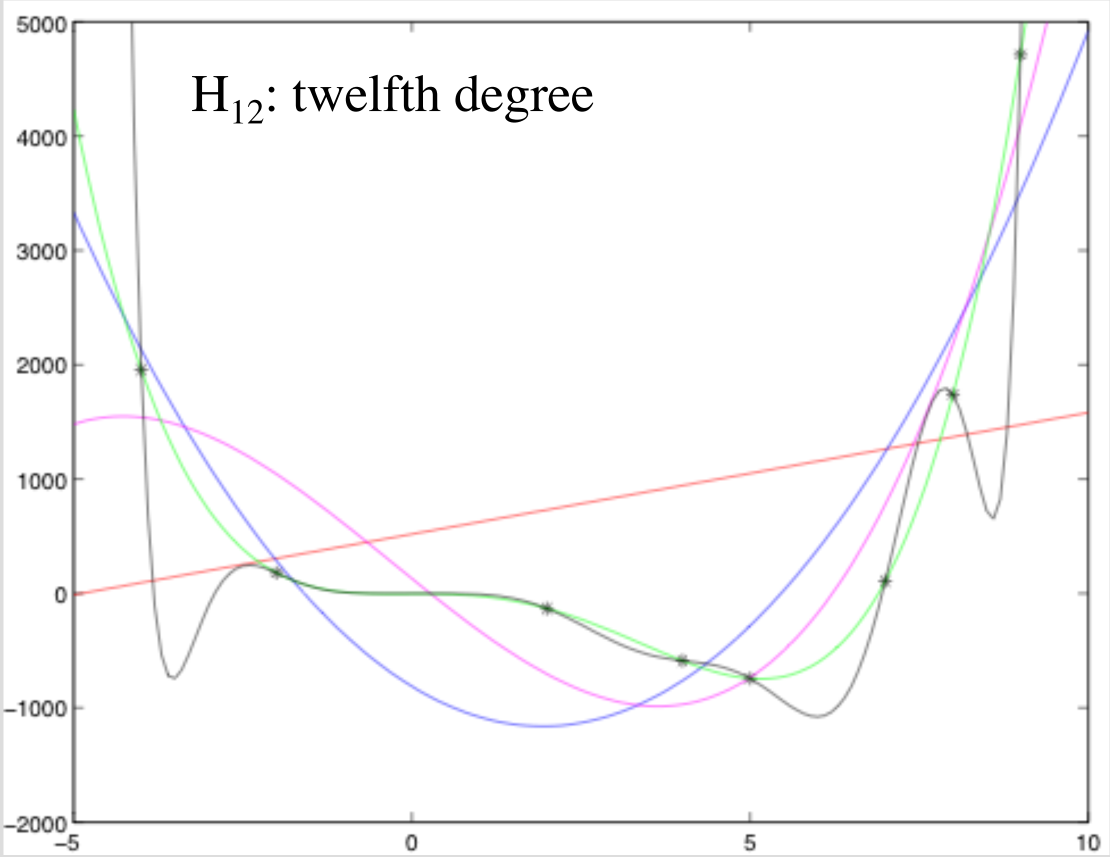
Hypothesis spaces & Inductive learning



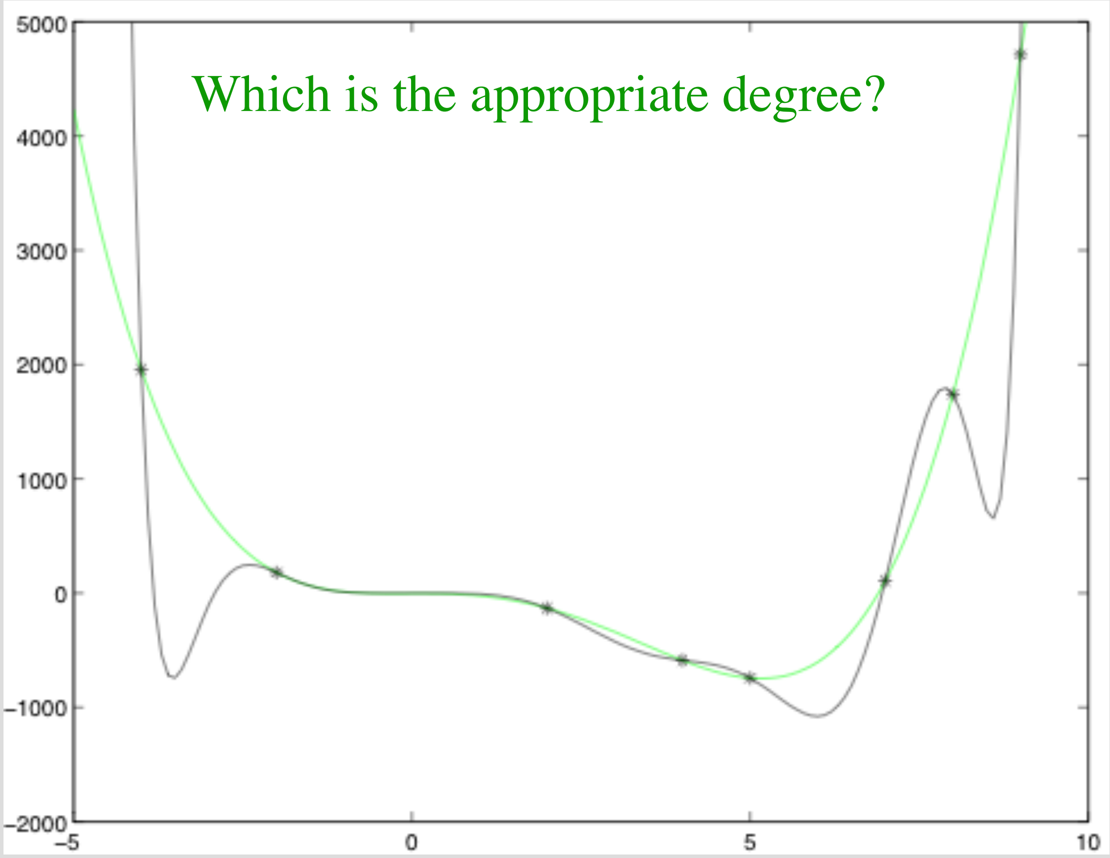
Hypothesis spaces & Inductive learning



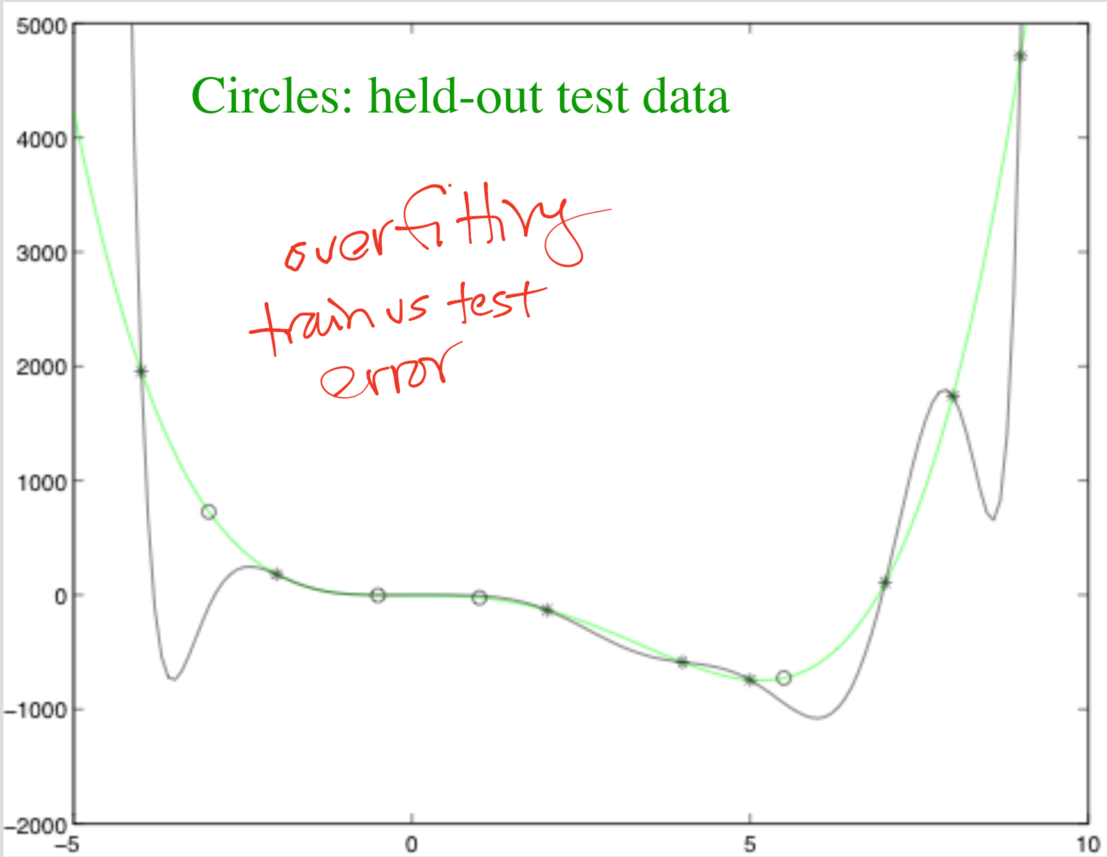
Hypothesis spaces & Inductive learning



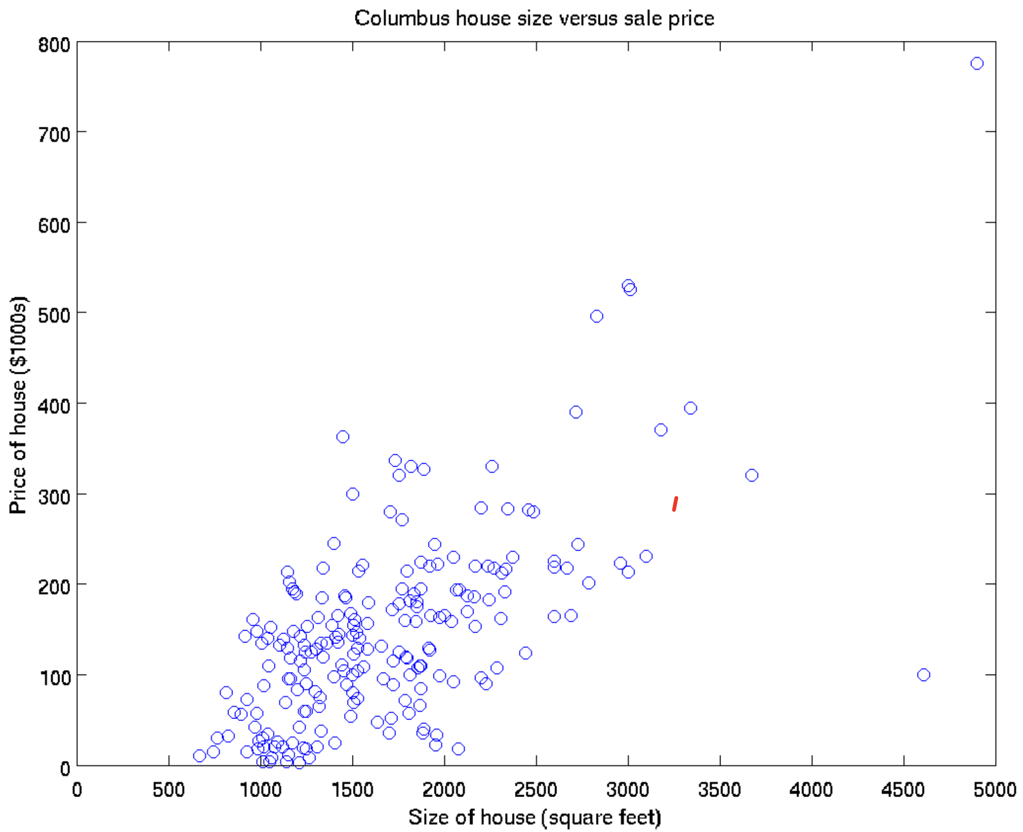
Hypothesis spaces & Inductive learning



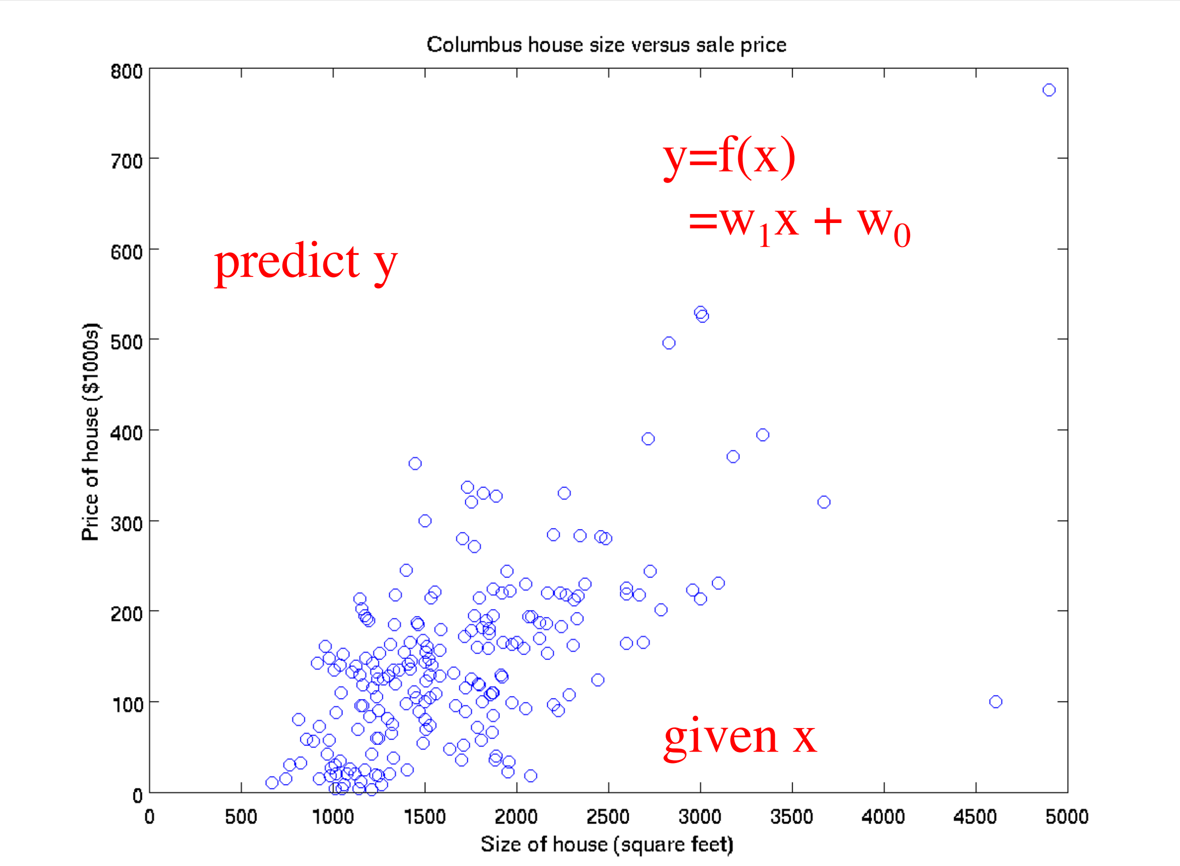
Hypothesis spaces & Inductive learning



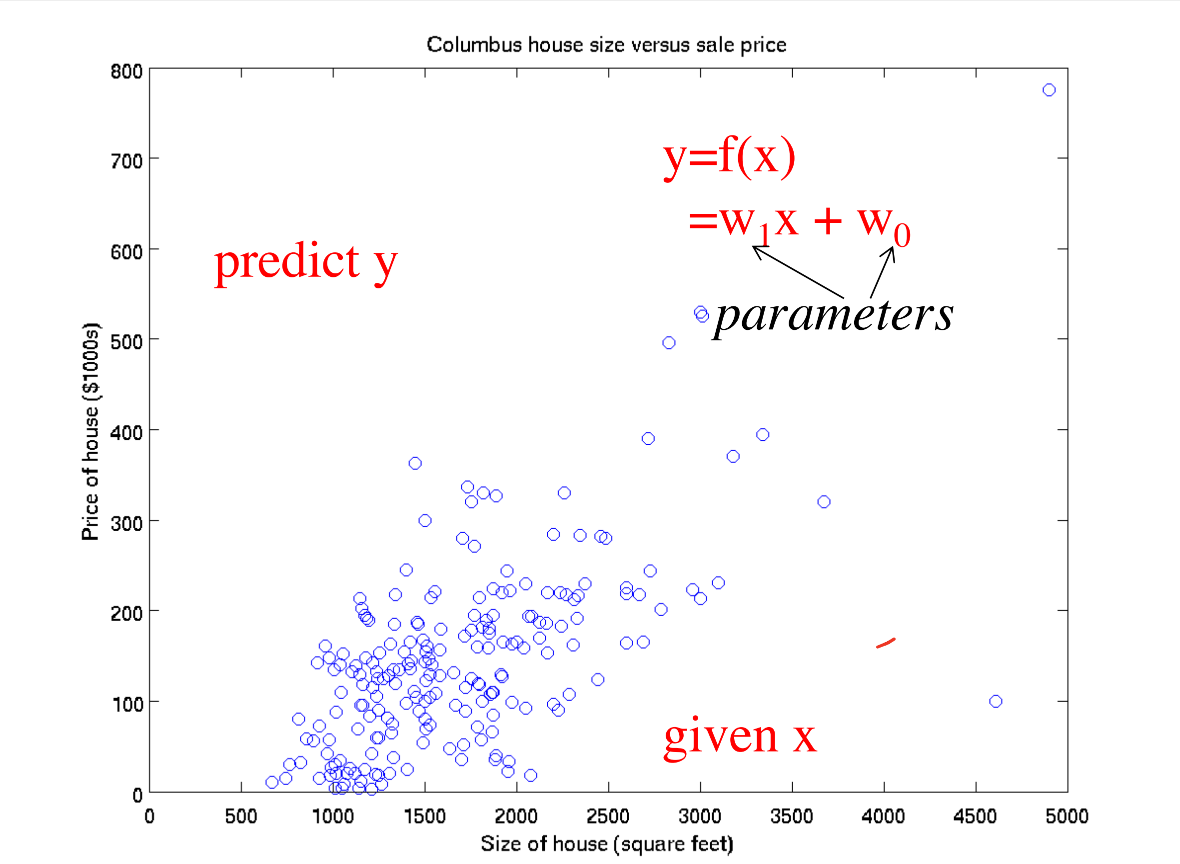
A first learning algorithm: Linear regression



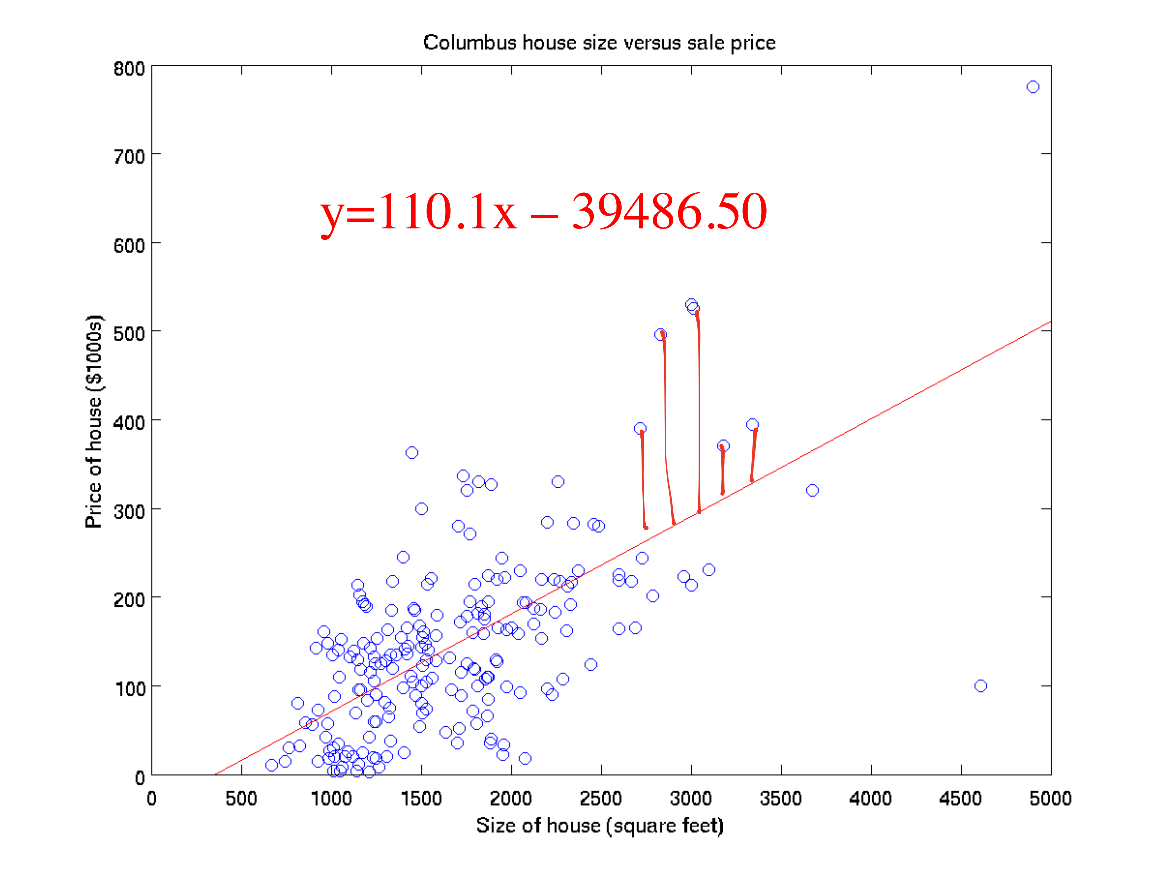
A first learning algorithm: Linear regression



A first learning algorithm: Linear regression



A first learning algorithm: Linear regression



Deriving Linear Regression

- Want to predict y=f(x)
  - Hypothesis $h_W(x)$ is parameterized by w
- What is the error?

$$\sum_i y_i - f(x_i) \times$$

$$\sum_i |y_i - f(x_i)| \quad L_1 \text{ dist.}$$

$$\sum_i \frac{|y_i - f(x_i)|}{y_i}$$

$$\frac{1}{N} \sum_i (y_i - f(x_i))^2$$

Loss function (error)

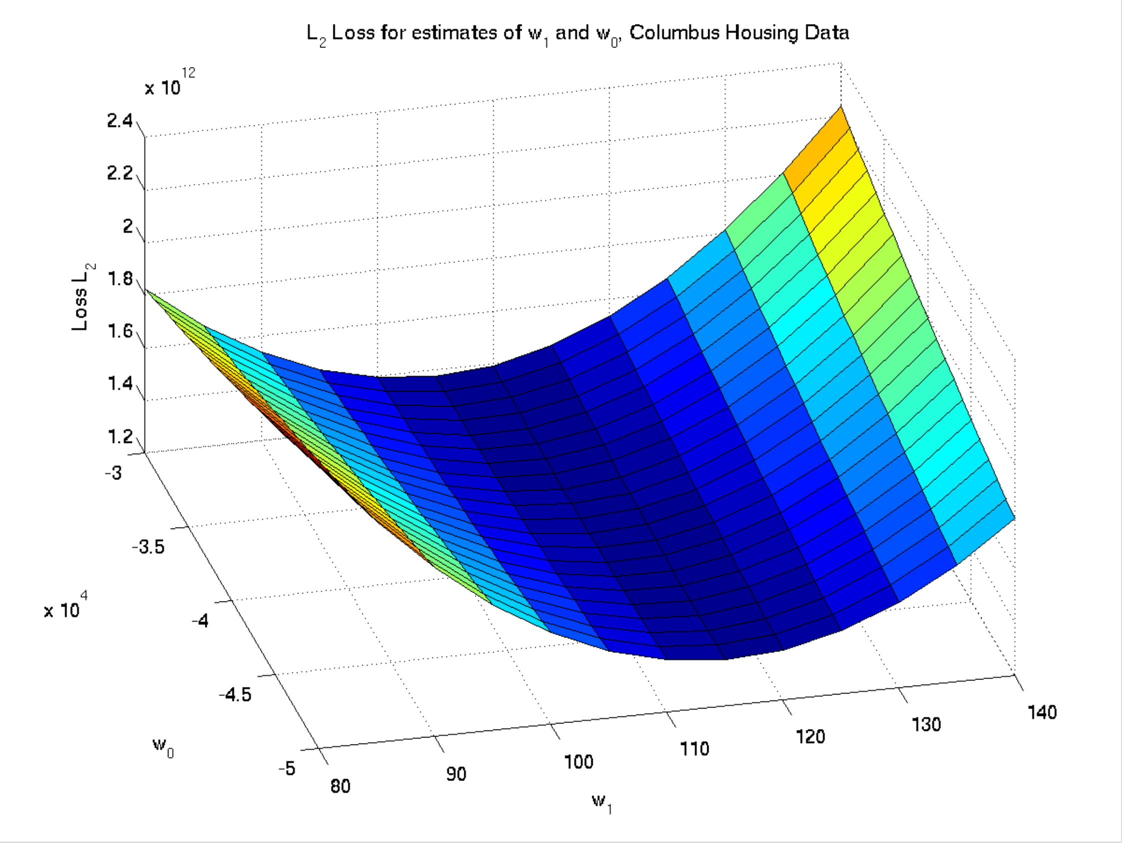
- For linear regression

$$Loss(h_w) = \sum_{j=1}^{N} L_2(y_j, h_w(x_j))$$

$$= \sum_{j=1}^{N} (y_j - h_w(x_j))^2$$

$$= \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2$$

Searching for parameter settings: L_2 Loss



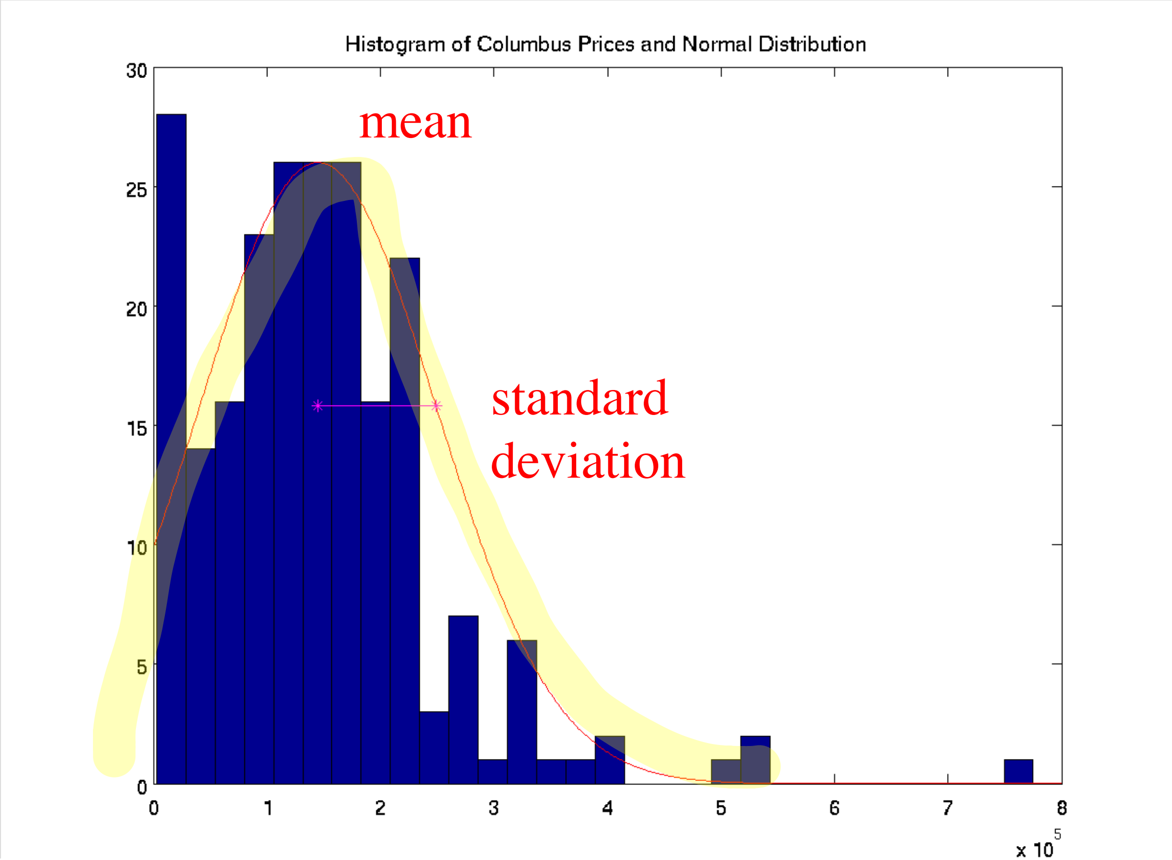
Convex optimization

- Since L_2 loss is convex, it has a global minimum
  - This can be solved analytically in this case

$$\frac{\partial}{\partial w_1}\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))^2=0,\frac{\partial}{\partial w_0}\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))^2=0$$

$$w_1=\frac{N\sum x_jy_j-\left(\sum x_j\right)\left(\sum y_j\right)}{N\left(\sum x_j^2\right)-\left(\sum x_j\right)^2},w_0=\frac{\left(\sum y_j-w_1\left(\sum x_j\right)\right)}{N}$$

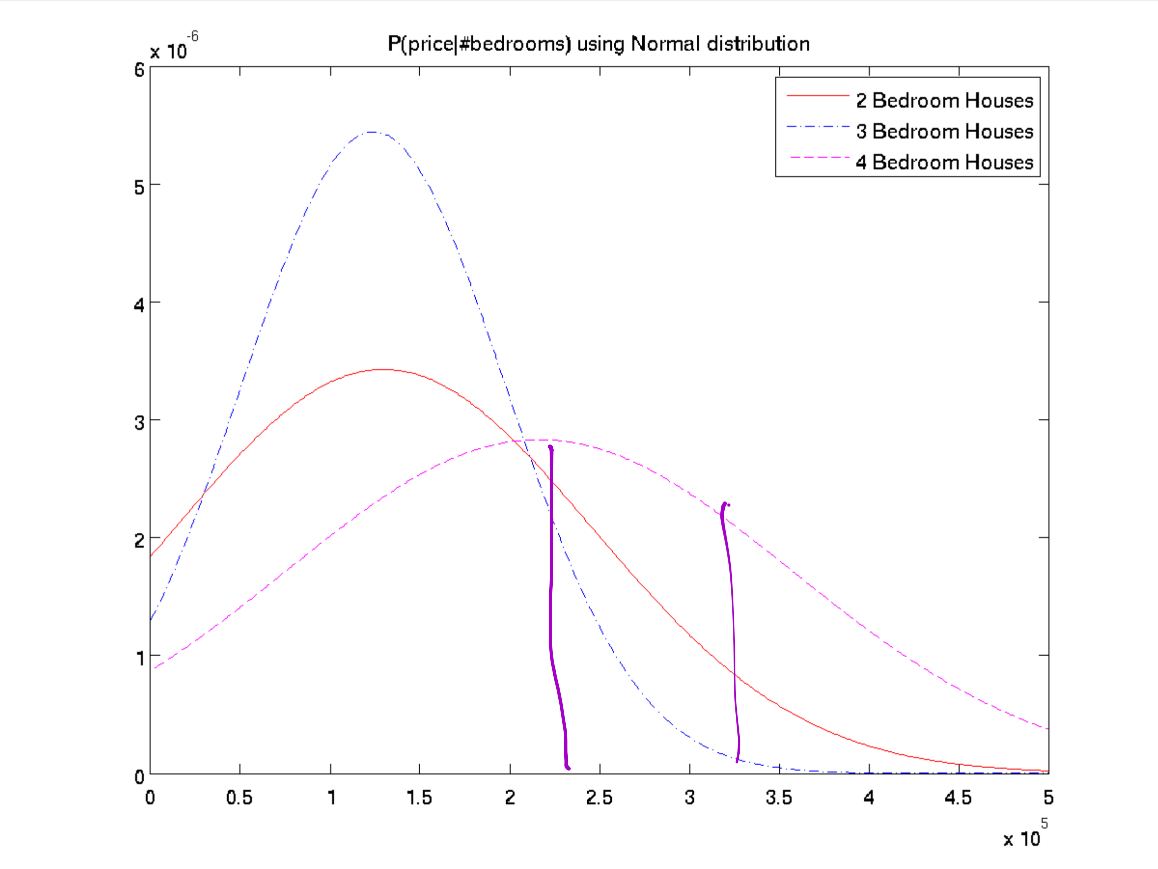
Function Approximation #2: Approximating histograms



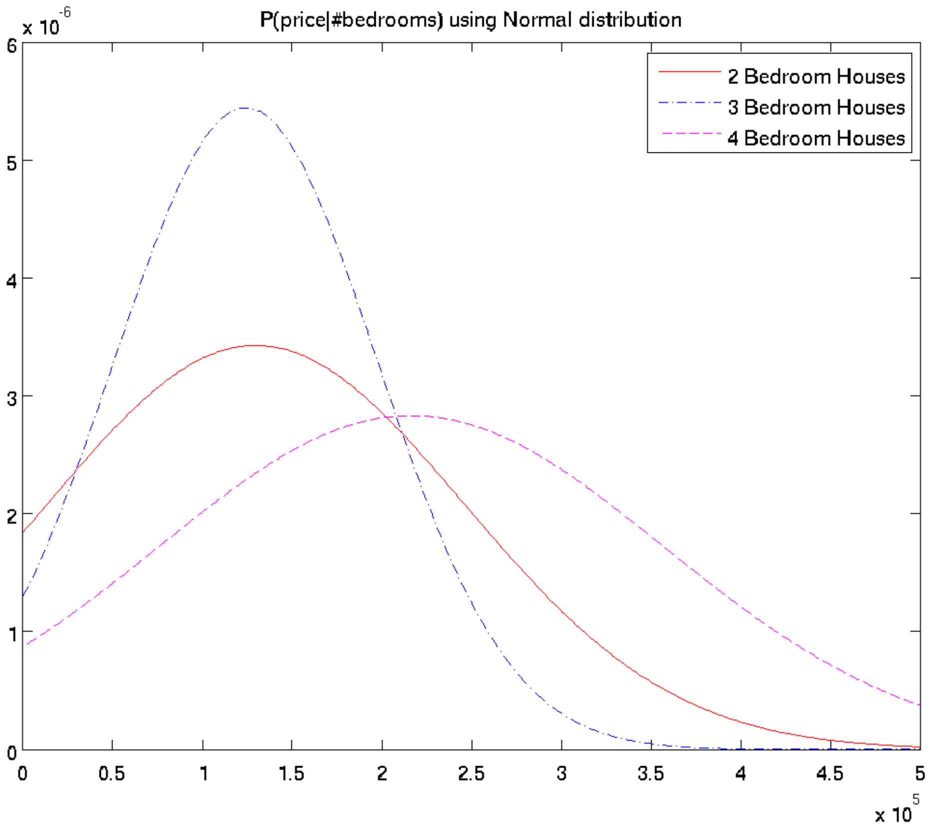
Probability distributions

- Probability theory drives much of machine learning
- Learning == function approximation
  - What are parameters?
  - What is loss function?

Predicting bedrooms from price

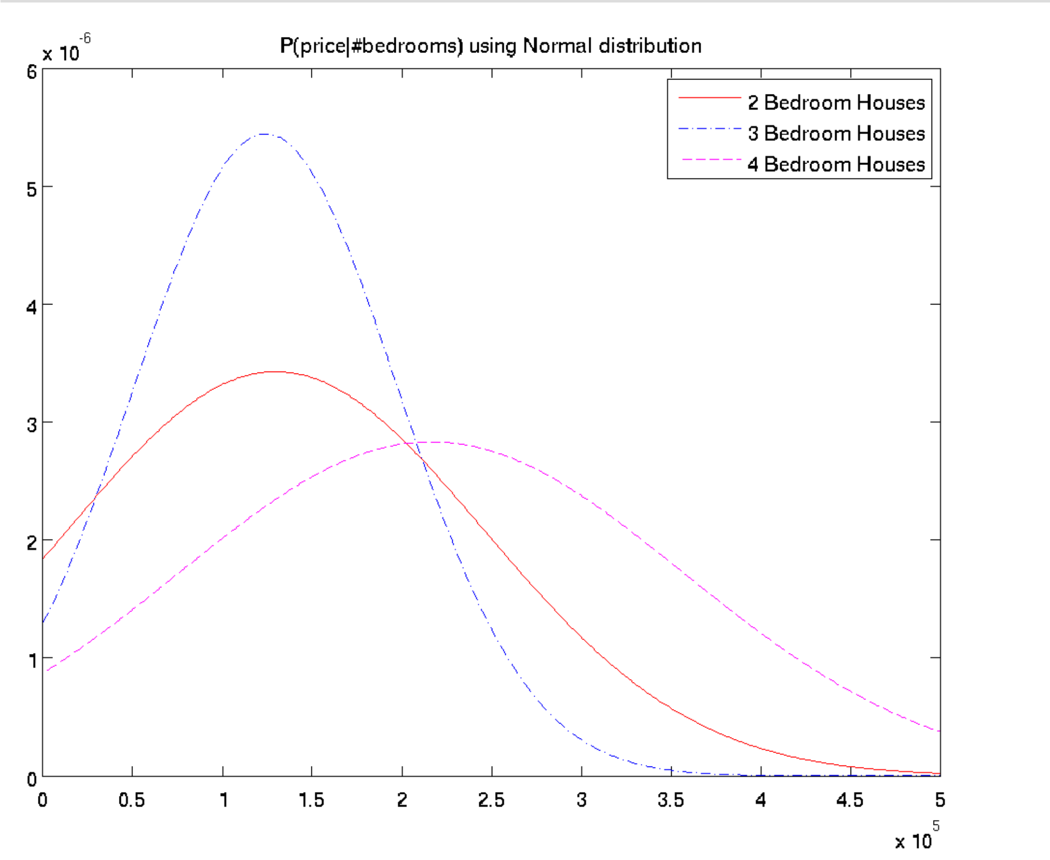


Simplest predictor:
most likely #bedroom given price



Why is this not right?

A real classifier: Incorporate *prior* information



Transition: Probability Theory

- We can approach problems
  - by positing a hypothesis function class,
  - estimating parameters of the hypothesis function,
  - using a loss function to guide us
    - *Do the right thing!*
- First example of regression, classification
- Now we explore topics in probability theory for making classification decisions

Logical reasoning

- In 630/5521, we talked about logic

Mary loves John		
Betty loves John		
John loves John		
Q: Does everyone love John?		

Logical reasoning

- Here, the “variables” are logic sentences

Mary loves John	S1: loves(M,J)	
Betty loves John	S2: loves(B,J)	
John loves John	S3: loves(J,J)	
Q: Does everyone love John?	S1^S2^S3=>∀ x loves(x,J)	

Logical reasoning

- Values are true, false

Mary loves John	S1: loves(M,J)	true
Betty loves John	S2: loves(B,J)	true
John loves John	S3: loves(J,J)	true
Q: Does everyone love John?	$S1 \wedge S2 \wedge S3 \Rightarrow \forall x \text{ loves}(x,J)$	true (if world only consists of M,B,J)

Reasoning with uncertainty

- What if we don't know the answers?

It's likely Mary loves John	P(loves(M,J))	
I'm not sure if Betty loves John	P(loves(B,J))	
John almost certainly loves John	P(loves(J,J))	
Q: Does everyone love John?	P(∀x loves(x,J))	

Reasoning with uncertainty

- Can give the probability of values

It's likely Mary loves John	P(loves(M,J)=true)=0.8
I'm not sure if Betty loves John	P(loves(B,J)=true)=0.5
John almost certainly loves John	P(loves(J,J)=true)=0.95
Q: Does everyone love John?	P(∀x loves(x,J)=true)= 0.8*0.5*0.95=0.38 (assuming above are independent)

Where do probabilities come from?

- From life experience
- From guessing
- From controlled sample pools
- The quality of the judgments made using this data will depend on the sample that the probabilities came from
  - How well does the source match the test conditions?
  - Language statistics from newswire applied to childrens books

What are probabilities in terms of logic?

- Probabilities describe the degree of belief in a particular proposition
  - No longer just true or false
  - “The chance of rain today is 10%”
    P(rain) = .1
  - “80% of the time, squealing indicates bad brakes”
    … means that we believe 80% of the time
    Squeal => BadBrakes
- It is not that the proposition is x% true
  - P(rain)=.1 does not mean it is raining 10%

Random variables

- In order to determine the probability of events, we have to know how many different possibilities there are
- A random variable takes on one or more values
  - 6-sided die roll: Roll=1, Roll=2, …, Roll=6
  - Squealing: Squeal=true, Squeal=false
- Random variables have three components:
  - The name of the variable
  - The range of its elements
  - A probability associated with each element
    - This is called a probability distribution

Random variables

- Typically written with a capital letter (particularly in R&N)
- 3 types, depending on domain
  - boolean: <true, false>
    - Logical propositions
    - Can abbreviate P(Rain=true)=P(rain), P(Rain=false)=P(~rain)
  - discrete: <a,b,c,d>
  - continuous: [0,1]
- Examples of each?

Unconditional probabilities

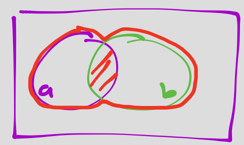
- What you think of as "regular" probabilities
- Gives the probability of a variable taking a particular value without any conditions
  - P(Roll=2)=1/6
  - P(rainToday)=0.1
- Probabilities must sum to 1 over all values of the variable
  - Implies that P(~rainToday)=0.9
  - P(Roll=1vRoll=3vRoll=4vRoll=5vRoll=6)=5/6
- Also called the "prior"

Unconditional probabilities

- Often, we want to write out a whole distribution
  - P(Cavity)=<0.1,0.9>
  - P(Weather)=<0.2,0.1,0.6,0.1>
    - Weather: <sun,rain,cloudy,snow>
  - P(Temperature=x)=U[50,70](x)
    - Continuous distribution: we'll come back to this

The Axioms of Probability

- All probabilities are between 0 and 1
  - $0 \leq P(a) \leq 1 \ \forall\ a$
- A proposition which is true has probability 1, false has probability 0
  - $P(true) = 1, \quad P(false) = 0$
- The probability of a disjunction (a v b):
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



Joint probabilities

- We can consider probabilities of more than one variable
  - E.g. P(rainToday ^ windToday)
- This is a joint probability table (JPT):

	rainToday	~rainToday
windToday	0.08	0.2
~windToday	0.02	0.7

0.28

0.72

marginalization

0.1

0.9

Joint probabilities

- We can consider probabilities of more than one variable
  - E.g. RainToday & WindToday
- This is a joint probability table:

	rainToday	~rainToday
windToday	0.08	0.2
~windToday	0.02	0.7

All entries must sum to 1

Marginalization

	rainToday	~rainToday
windToday	0.08	0.2
~windToday	0.02	0.7

- We can calculate the probability of individual variables by summing the columns or rows
  - P(windToday)=0.28
  - P(~rainToday)=0.9

Inference examples (from book)

- What is P(cavity) given the following JPT?
  - To answer this, sum wherever cavity is true

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Inference examples (from book)

- What is P(cavity) given the following JPT?
  - To answer this, sum wherever cavity is true
    $P(cavity)=0.108+0.012+0.072+0.008=0.20$

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Inference examples (from book)

- What is P(cavity v toothache)?

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Inference examples (from book)

- What is P(cavity v toothache)?
  - P(cavity v toothache) = 0.108+0.012+0.072+0.008+0.016+0.064 = 0.28

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Inference examples (from book)

- What is P(cavity ^ toothache)?
  - P(cavity ^ toothache) = 0.108 + 0.012 = 0.12

  $= \sum_{c \in Catch} P(cavity, toothache, c)$

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Inference examples (from book)

- How many parameters (distinct numbers that can't be calculated) are needed to specify JPT with binary variables?

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Conditional Probabilities

- Sometimes we know some things to be true, and want to find out how that affects other variables
  - I know my brakes are squealing, are they bad?
  - I know I have a toothache, do I have a cavity?
  - I know I got an 80 on the exam, will I pass the class?

Conditional Probabilities

- We write P(alb) to mean “what’s the probability of A being true given that B is true”
  - I know my brakes are squealing, are they bad?
    - P(badBrakes l squeal)
  - I know I have a toothache, do I have a cavity?
    - P(cavity l toothache)
  - I know I got an 80 on the exam, will I pass the class?
    - P(pass l Score=80)

Conditional probabilities

- The conditional probability $P(A|B)$ is related to the joint $P(A,B)$ and the prior $P(B)$:

$$P(A|B) = P(A,B) / P(B)$$

Conditional probabilities

- What's P(rainToday|~windToday)?

	rainToday	~rainToday
windToday	0.08	0.2
~windToday	0.02	0.7

- P(rainToday,~windToday)/P(~windToday)
  - 0.02/0.72 ~= 0.027

Conditional probabilities

- What's P(rainToday|windToday)?

	rainToday	~rainToday
windToday	0.08	0.2
~windToday	0.02	0.7

- P(rainToday,windToday)/P(windToday)
  - 0.08/0.28 ~= 0.285

Marginalization (again)

- In general, if you have P(X,Y,Z) and you want P(X,Y):
  - $P(X,Y) = \sum_{z \in Z} P(X,Y,z)$
- Similarly, want P(X|Z) and have P(X,Y|Z)
  - $P(X|Z) = \sum_{y \in Y} P(X,y|Z)$

$P(X) = \sum_z P(X, Z=z)$

$= \sum_z P(X \mid Z=z) P(Z=z)$

- Moral: if you want to remove a variable, sum over it

$P(X) \neq \sum_{z \in Z} P(X|z)$

Derivation of Bayes' Rule

- We defined conditional probability as:
  - P(A|B) = P(A,B)/P(B)
- Which means:
  - P(A|B) P(B) = P(A,B)
- But:
  - P(A,B)=P(B,A)
- So:
  - P(A|B) P(B) = P(B|A) P(A)
  - (P(A|B)P(B))/P(A) = P(B|A)   *Bayes' Rule*

$P(A,B) = P(A|B)\,P(B)$

$P(B,A) = P(A,B)$

$P(A,B) = P(B|A)\,P(A)$

$P(A|B)\,P(B) = P(B|A)\,P(A)$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayes' rule in action

- Suppose a patient came in with a stiff neck; what's the probability of meningitis?
  - What's $P(m|s)$?
  - Could estimate this directly, but might have other evidence that makes it easier
  - $P(s|m) = 0.5$, $P(m)=1/50000$, $P(s)=1/20$
  - $P(m|s) = (P(s|m)P(m))/P(s) = 0.0002$
- What if there's a meningitis epidemic?

Independence

- Random variables can influence each other, or be independent
- We say a and b are absolutely independent if $P(a,b) = P(a)P(b)$
  - $P(heads,tails)$ in 2 coin flips is 0.25
  - $P(heads)=0.5$, $P(tails)=0.5$
  - Coin flips are independent
- Can also say $P(a|b)=P(a)$

Conditional independence

- Things are not always completely independent:
  - toothache and catch are not independent (they are both more likely if you have a cavity)
  - HOWEVER, you can say that they are independent GIVEN that you have a cavity:
    - P(toothache ^ catch | cavity) =
      P(toothache | cavity) P(catch | cavity)

Combining evidence with conditional independence

- Want P(cavity | toothache, catch)
  - = P(toothache, catch | cavity) P(cavity) / P(toothache, catch)
  - = P(toothache | cavity) P(catch | cavity) P(cavity) / P(toothache, catch)
    conditional independence
- Ignore P(toothache, catch) for a minute
- We can combine causal information from different sources to give a probability of diagnosis

Alpha-normalization

- Notice in Bayes' rule, we need 3 distributions to describe 1 other:
  - $P(A|b) = P(b|A)\ P(A)\ /\ P(b)$
- When we have a single value on conditioning side (e.g., b), P(b) just normalizes the other two distributions
- Can re-write this as
  - $P(A|b) = \alpha\ P(b|A)\ P(A)$
  - or $P(A|b) = \alpha\ P(b,A)$

$\alpha = 1/P(b)$

Alpha normalization

- $P(Toothache|cavity) = \langle .6, .4 \rangle$
  $P(Toothache|\sim cavity) = \langle .1, .9 \rangle$
  $P(Cavity) = \langle .2, .8 \rangle$

- $P(Cavity|toothache) =$
  $\alpha \, P(toothache, Cavity) =$
  $\alpha \, (\langle .6*.2, .1*.8 \rangle) = \alpha \, (\langle .12, .08 \rangle) =$
  $\langle .12/.20, .08/.20 \rangle = \langle .6, .4 \rangle$

Alpha normalization

- P(Toothache|cavity)=<.6,.4>
  P(Toothache|~cavity)=<.1,.9>
  P(Cavity)=<.2,.8>

Note difference
in probability
given new information

- P(Cavity|toothache) =
  α P(toothache,Cavity) =
  α (<.6*.2,.1*.8>) = α (<.12,.08>) =
  <.12/.20,.08/.20> = <.6,.4>

Combining evidence (again)

- The bank is closed, and I didn't receive mail today. What is the probability that today is a holiday?
- Assume:
  - Banks are always closed on holidays
  - No mail on holidays
  - 62 holidays a year
  - Bank is closed on 1 non-holiday/year
  - I don't receive mail 18 (non-holiday) days a year

Combining evidence (again)

- The bank is closed, and I didn't receive mail today. What is the probability that today is a holiday?

- P(H| ~bank , ~mail) =
  α P(~bank , ~mail | H) P(H) =
  α P(~bank|H) P(~mail | H) P(H) =
  $\alpha < 1 * 1 * 62/365,\ 1/303 * 18/303 * 303/365 > =$
  $\alpha <0.1699, 0.000162> = <0.999, 0.001>$

Problems with Joint Probability Tables

- A joint probability table expresses all of the possible situations given a set of variables
- When # of variables gets large, the JPT is often too big to express
  - Remember: binary case is 2^k-1
- Need more efficient mechanism for larger problems

Back to conditional probabilities…

- The joint distribution can always be specified as a product of conditional distributions

$$P(X_1,X_2,\ldots X_n)=P(X_1|X_2,\ldots X_n)P(X_2,\ldots,X_n)$$

$$=P(X_1|X_2,\ldots,X_n)P(X_2|X_3,\ldots X_n)P(X_3,\ldots,X_n)$$

$$=\Pi_{i=1:n}\, P(X_i|X_{i+1}\ldots X_n)$$

$$=\Pi_{i=1:n}\, P(X_i|X_1\ldots X_{i-1})$$

Conditional probability tables

- Given conditioning variables, what is probability of other variable?
  - $P(X|e_1,e_2)$, where X, e_1, and e_2 are binary

e_1	e_2	X=true	X=false
T	T	a	1-a
T	F	b	1-b
F	T	c	1-c
F	F	d	1-d

Conditional probability tables

- Given conditioning variables, what is probability of other variable?
  - $P(X|e_1,e_2)$, where X, e_1, and e_2 are binary

e_1	e_2	X=true	~~X=false~~
T	T	a	~~1-a~~
T	F	b	~~1-b~~
F	T	c	~~1-c~~
F	F	d	~~1-d~~

Another CPT example

- Block is either square or round
- Block is either red, blue, or yellow

Shape	P(Color=red)	P(Color=blue)
round	1/3	0
square	3/7	3/7

Why all the hoopla?

- Remember conditional independence:
  - If A is conditionally independent of B given C:
  - $P(A,B|C) = P(A|C)\ P(B|C)$
  - $P(A|B,C) = P(A|C)$
- So?

Why all the hoopla?

- Look at what happens to CPT: $P(A|B,C) = P(A|C)$

B	C	P(A|B,C)
T	T	a
T	F	b
F	T	a
F	F	b

C	P(A|C)
T	a
F	b

What CI assumption can be made here?

W	X	Y	P(Z|W,X,Y)
T	T	T	0.3
T	T	F	0.4
T	F	T	0.3
T	F	F	0.4
F	T	T	0.7
F	T	F	0.3
F	F	T	0.7
F	F	F	0.3

Summary: counting parameters

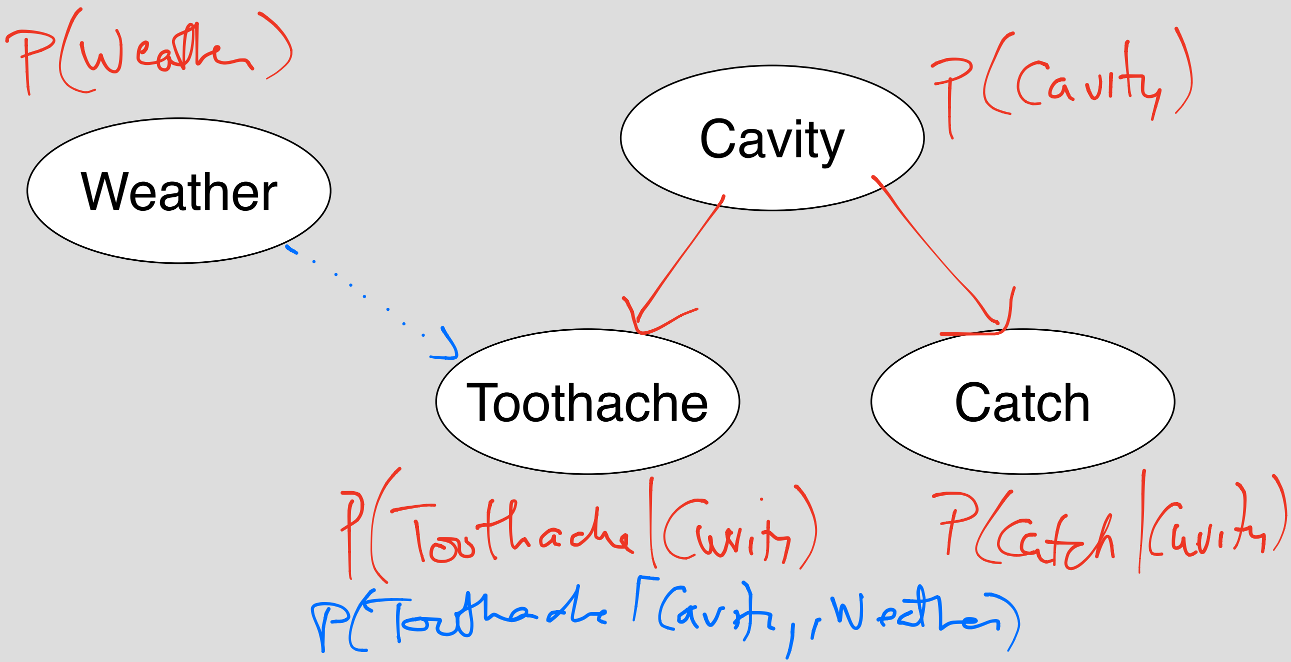
- Sometimes we are concerned with the number of parameters (probabilities)
- Numbers that we don’t have to specify explicitly don’t count
- Full joint distribution has $n_1*n_2*..n_k-1$ parameters (why?)
  - Boolean variables: 2^k-1 parameters
- Independence assumptions can reduce the number of parameters
- How does this relate to Occam’s Razor?

Bayesian Networks

- Graphical formalism to help keep track of independence assumptions
- Each Bayes' Net has the following properties
  - Contains a set of random variables, each represented by the nodes of a network
  - Contains a set of directed links between nodes
    - If X->Y, then X is the *parent* of Y
  - Each node has a CPT associated with it
    - P(X|Parents(X))
  - The links have no directed cycles
    - Bayes' nets are directed, acyclic graphs (DAGs)

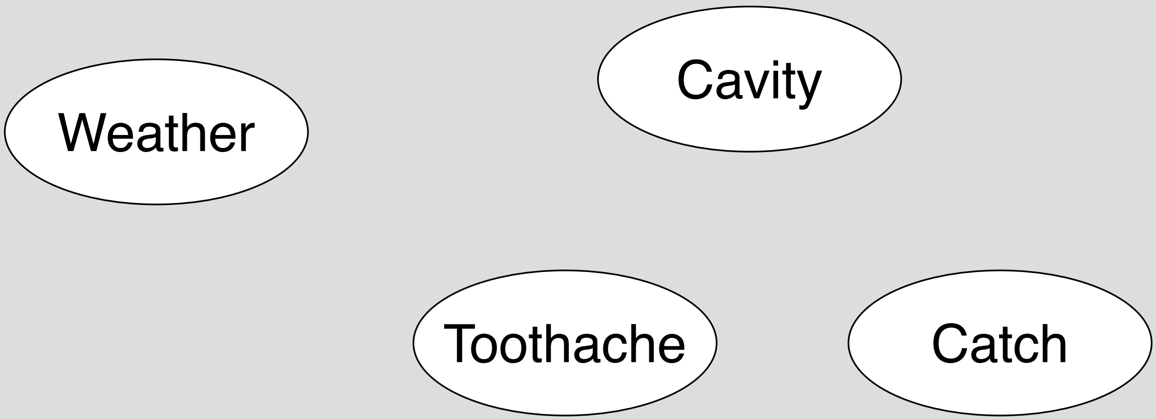
Example from book

- Cavity example: 4 variables
  - Weather, Cavity, Toothache, Catch



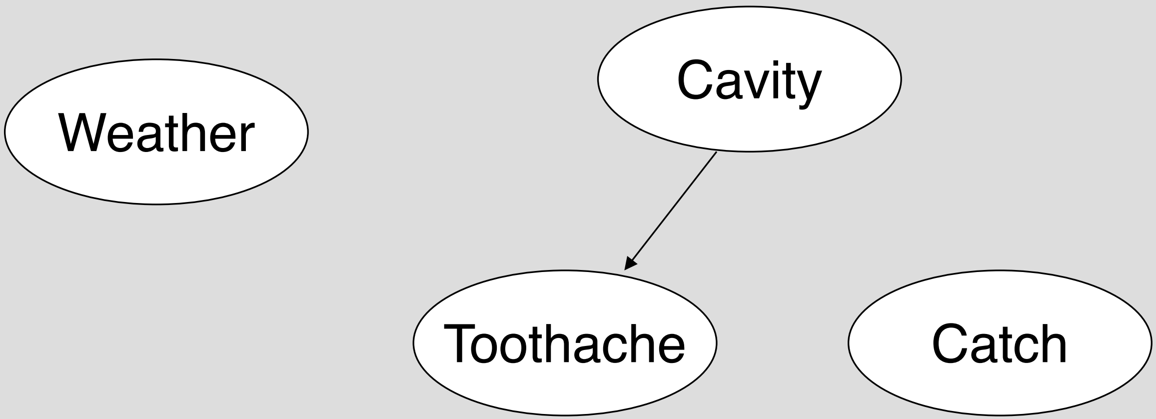
Example from book

- Now draw links from X to Y if X directly influences Y



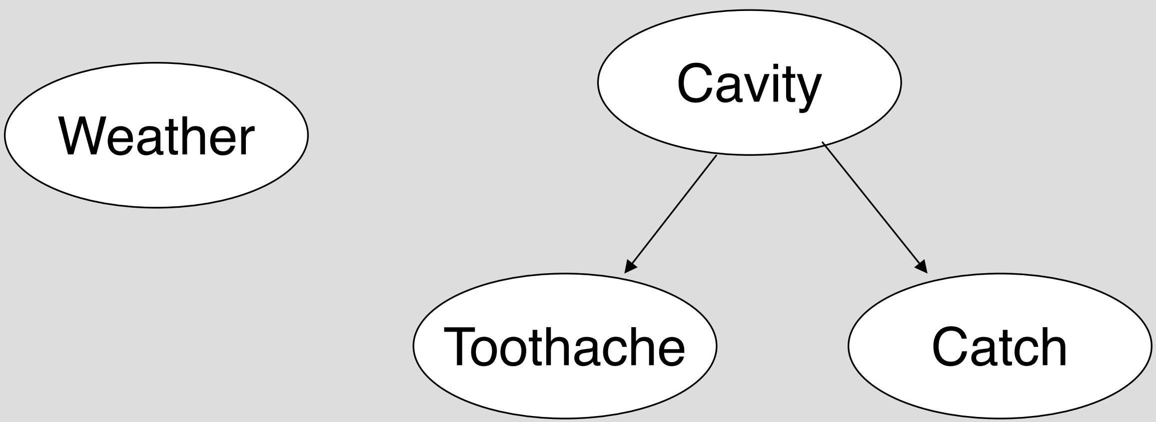
Example from book

- Cavity directly affects whether you have a Toothache



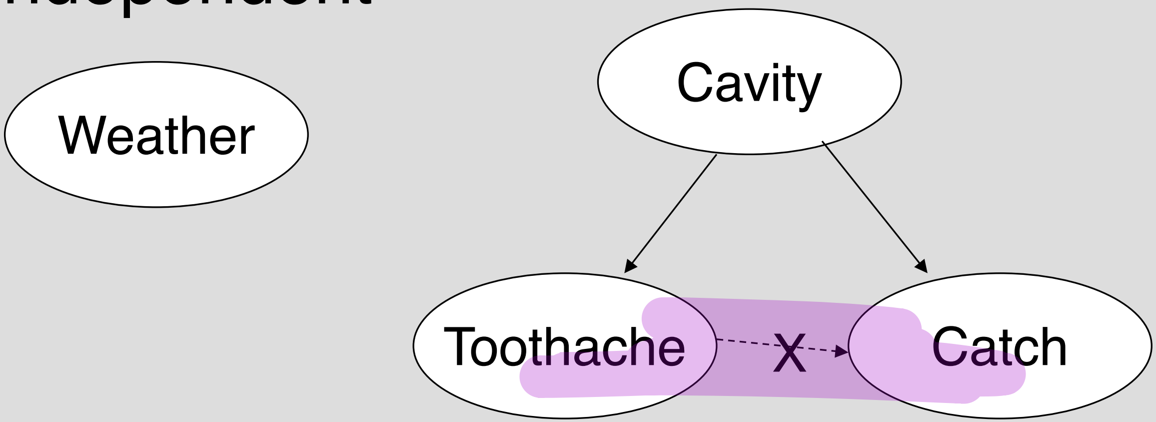
Example from book

- Cavity directly affects whether the probe Catches



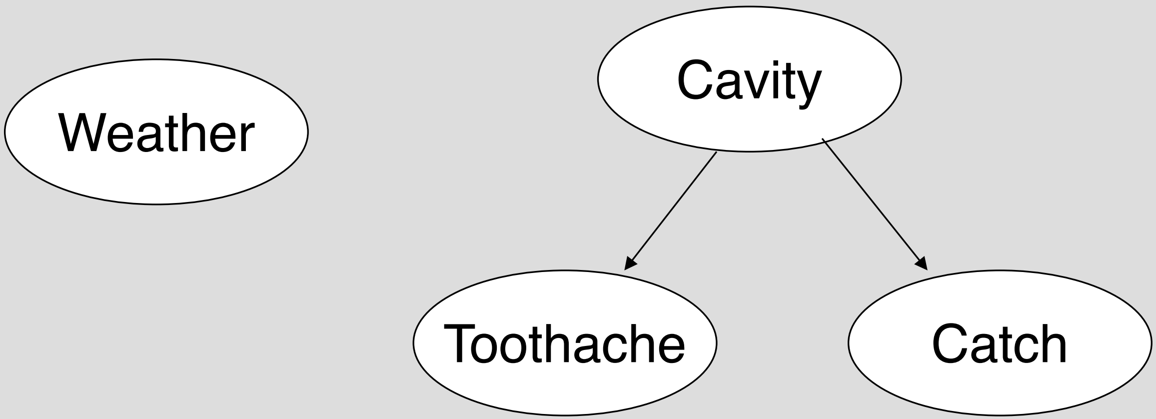
Example from book

- If you know if you have a cavity or not, then Toothache and Catch are independent



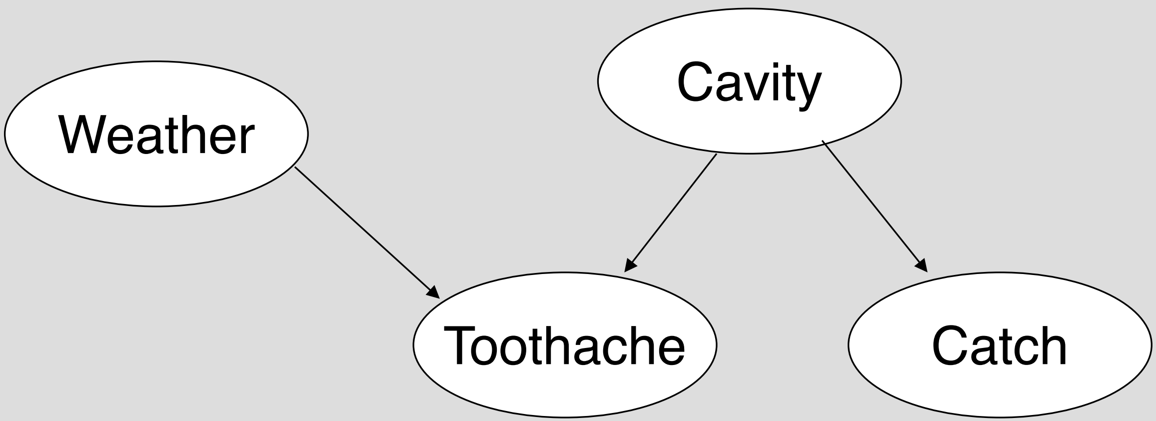
Example from book

- Weather is (in theory) independent of all other variables. (Is this true?)



Example from book

- If stormy weather makes toothaches more likely…



Directions of arrows

- Can have P(A,B) represented as
  - P(A|B), P(B) $A \leftarrow B$
  - P(B|A), P(A) $A \rightarrow B$
- How do you know which way to draw the arrows?
  - Usually best to go from causes to effects

Alarm example (Judea Pearl)

- Currently at work
- John (neighbor) calls to say alarm is ringing
- Mary (other neighbor) doesn’t call
- Sometimes alarm is set off by small earthquakes
- Is there a burglar?

Alarm example

- Variables:
  - Burglar (B): binary
  - Earthquake (E): binary
  - Alarm (A): binary
  - John calls (J): binary
  - Mary calls (M): binary
- [Example worked out on board, can be found R&N]

Global semantics

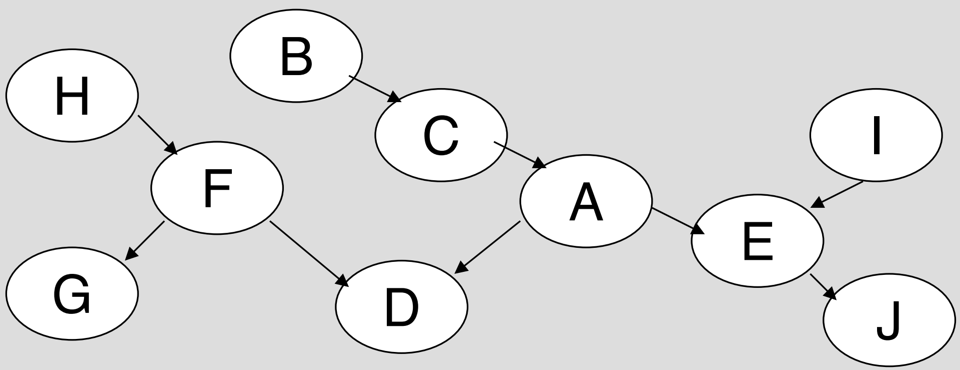
- Can get any joint distribution from Bayes net:
  - $P(X_1, X_2, \ldots X_n) = \Pi_{i=1:n} P(X_i | Parents(X_i))$
- What is probability of John calling, Mary calling, Alarm is on, but no burglary or earthquake?

Global semantics

- Can get any joint distribution from Bayes net:
  - $P(X_1, X_2, \ldots X_n) = \Pi_{i=1:n} P(X_i | Parents(X_i))$
- What is probability of John calling, Mary calling, Alarm is on, but no burglary or earthquake?
  - P(j,m,a,~b,~e)=
    P(j|a)P(m|a)P(a|~b,~e) P(~b) P(~e) =
    .9 * .7 * .001 * .999 * .998 = 0.000628

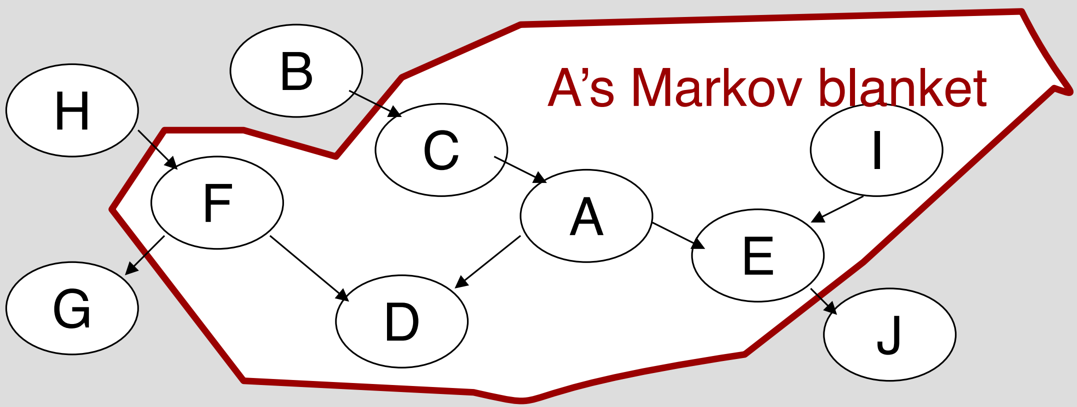
Local semantics

- There are two ways to express conditional independence relationships in Bayes nets:
  - 1) Each node is CI of non-descendants given its parents
  - 2) Each node is CI of others given its Markov blanket
    - Markov blanket: parents + children+ children's parents



Local semantics

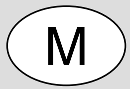
- There are two ways to express conditional independence relationships in Bayes nets:
  - 1) Each node is CI of non-descendants given its parents
  - 2) Each node is CI of others given its Markov blanket
    - Markov blanket: parents + children+ children's parents



Building Bayes Nets: Algorithm

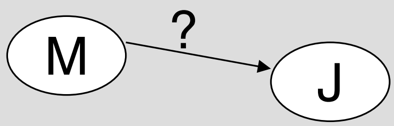
- Choose an ordering of variables $X_1 \ldots X_n$
- for i = 1..n
  - add X_i to the network
  - select Parents from $X_1..X_{i-1}$ s.t. $P(X_i|Parents(X_i))=P(X_i|X_1 \ldots X_{i-1})$

Alarm example, different order



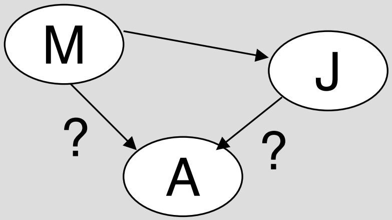
- Start with ordering M, J, A, B, E

Alarm example, different order



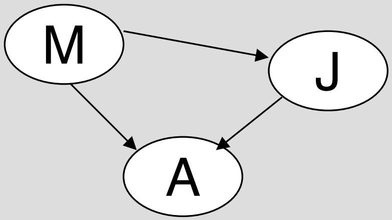
- Start with ordering M, J, A, B, E
- P(J|M)=P(J)?

Alarm example, different order



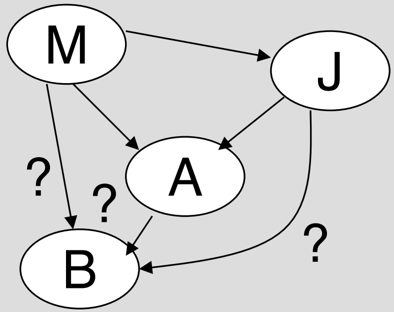
- $P(J|M)=P(J)$? NO
- $P(A|J,M) = P(A|J)$?
  $P(A|J,M) = P(A)$?

Alarm example, different order



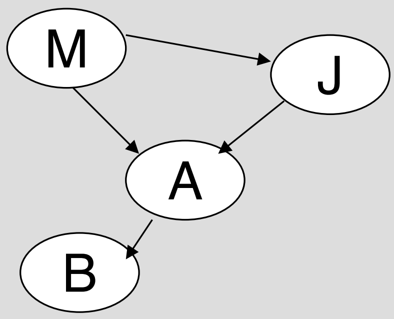
- $P(A|J,M) = P(A|J)$? NO
  $P(A|J,M) = P(A)$? NO

Alarm example, different order



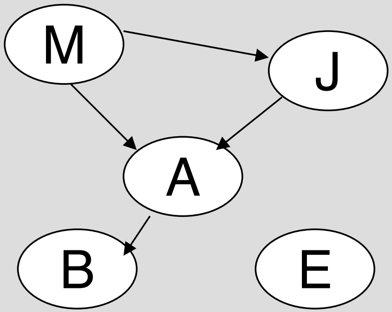
- P(B|A,J,M) = P(B|A)?
  P(B|A,J,M) = P(B)?

Alarm example, different order



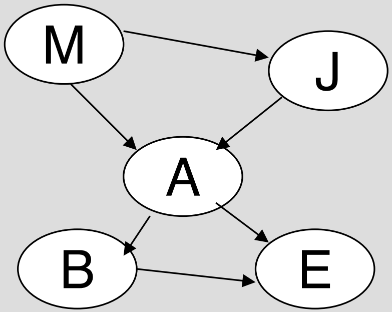
- P(B|A,J,M) = P(B|A)? YES
  P(B|A,J,M) = P(B)? NO

Alarm example, different order



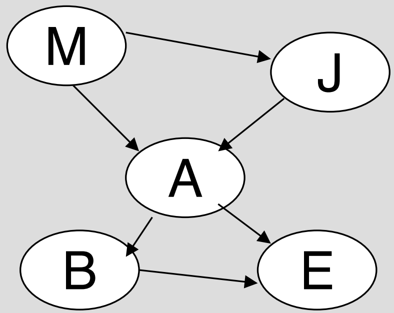
- P(E|B,A,J,M) = P(E|A)?
  P(E|B,A,J,M) = P(E|A,B)?

Alarm example, different order



- $P(E|B,A,J,M) = P(E|A)$? NO
  $P(E|B,A,J,M) = P(E|A,B)$? YES

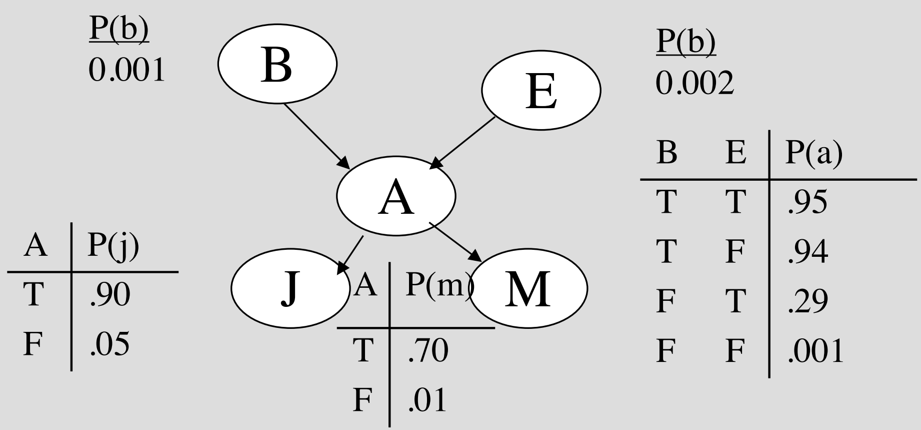
Alarm example, different order



- Conditional independence is hard in non-causal directions
  - How do you find P(J|M) in real life?
  - Difficult to know since the two are not causally related

Computing the joint probability

- Alarm example
- Compute: P(j,m,a,~b,~e)=
  P(j|a)P(m|a)P(a|~b,~e) P(~b) P(~e) =
  .9 * .7 * .001 * .999 * .998 = 0.000628



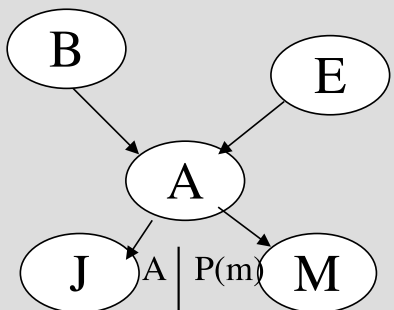
A	P(j)
T	.90
F	.05

A	P(m)
T	.70
F	.01

B	E	P(a)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

Missing information

- What's the probability distribution over Burglary given that John calls and Mary calls?



P(b)
0.001

P(b)
0.002

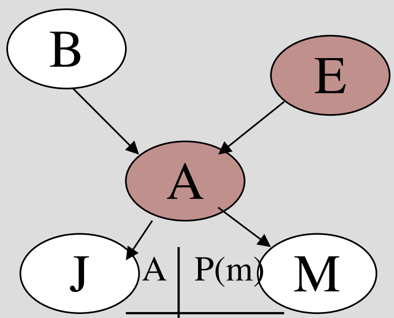
B	E	P(a)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(j)
T	.90
F	.05

A	P(m)
T	.70
F	.01

Missing information

- What's $P(B|j,m)$?
- The state of A, E are unknown.
  - So we must "sum them out" (marginalize).
  - These are called hidden variables.



P(b)
0.001

P(b)
0.002

B	E	P(a)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(j)
T	.90
F	.05

A	P(m)
T	.70
F	.01

Exact inference by enumeration

- First: redefine your quantity as the normalized sum over a joint distribution

- $P(B|j,m) = P(B,j,m) / P(j,m)$
  $= \alpha\, P(B,j,m)$
  $= \alpha\, \Sigma_e\, \Sigma_a\, P(B,\underline{e},\underline{a},j,m)$

"<u>e</u>" here means e or ~e

Exact inference by enumeration

- Next, replace joint distribution by CPT entries

$$\alpha \Sigma_e \Sigma_a P(B,\underline{e},\underline{a},j,m) =$$
$$\alpha \Sigma_e \Sigma_a P(B)\, P(\underline{e})\, P(\underline{a}|B,\underline{e})\, P(j|\underline{a})\, P(m|\underline{a})$$

- Where do these come from?

Exact inference by enumeration

- Now, push summations inwards

$\alpha \; \Sigma_e \, \Sigma_a \, P(B)\, P(\underline{e})\, P(\underline{a}|B,\underline{e})\, P(j|\underline{a})\, P(m|\underline{a}) =$

$\alpha \; P(B)\, \Sigma_e \, P(\underline{e})\, \Sigma_a \, P(\underline{a}|B,\underline{e})\, P(j|\underline{a})\, P(m|\underline{a})$

This reduces the number of products you need to do.

Exact inference by enumeration

- Now, start computing summations.
- Need two cases for inner summation
  - $\underline{e}$ is true, $\underline{e}$ is false
- Inner sum: Σ_a $P(\underline{a}|B,\underline{e})$ $P(j|\underline{a})$ $P(m|\underline{a})$
  - e is true:
    - (<.95,.29> * .9 * .7) + (<.05,.71> * .05 * .01)
      = <0.5985,0.1827> + <2.5e-05,3.55e-04>
      = <0.598525,0.183055>

    (vector is over <b, ~b>)

Exact inference by enumeration

- Now, start computing summations.
- Need two cases for inner summation
  - $\underline{e}$ is true, $\underline{e}$ is false
- Inner sum: $\Sigma_a\ P(\underline{a}|B,\underline{e})\ P(j|\underline{a})\ P(m|\underline{a})$
  - e is true:
    - <0.598525,0.183055>
  - e is false:
    - (<.94,.001> * .9 * .7)+(<.06,.999> * .05 * .01>)
      = <0.592236,0.0011295>

Exact inference by enumeration

- Bringing in the inner summation:

α P(B) Σ_e P($\underline{e}$) Σ_a P($\underline{a}$|B,$\underline{e}$) P(j|$\underline{a}$) P(m|$\underline{a}$) =

α P(B) Σ_e P($\underline{e}$) <<0.599,0.1839>$_{B|e}$, <0.593,0.001>$_{B|\sim e}$>

- Multiply each vector by P(e) or P(~e)

α P(B) (0.002*<0.599,0.1839> + 0.998*<0.593,0.001>) =

α P(B) <.541,.0014> =

α <.001,.999>*<.541,.0014> = α <.00054,.0014>

= <.278,.722> (diff from book because of rounding)

Hints for enumeration

- You will need to sum over hidden variables
- Basic math operations:
  - <a,b> * c = <a*c,b*c>
  - <a,b> + <c,d> = <a+c,b+d>
  - <a,b> * <c,d> = <a*c,b*d>
    - Not matrix multiplication!
- Keep track of what each part of the vector represents
- If you end up getting confused, break it down into individual cases.

Start here

Can you pass CSE 5522?

- Break into small groups (5-6 people)
- Think of 8 or so variables that affect whether you will pass 5522
  - Be creative, go deeper into root causes
- Build the corresponding Bayes' net
- When time is up, we'll put them on the board to share

Compact CPT representations

- CPTs get big depending on number, arity of parents
- Even with many parents, there are several ways to get smaller CPTs

Parameters and continuous distributions

- What about continuous ranges?
  - Infinite number of probabilities
  - Can’t explicitly enumerate parameters
- Probability density function: express a continuous distribution in a succinct manner
  - **Uniform distribution**: even probability over range
    - 2 parameters
  - **Gaussian (normal) distribution**: “bell curve”
    - 2 parameters in univariate case
  - Other distributions, e.g. binomial*, Beta, Dirichlet, Poisson* (*discrete)

Uniform distribution

- “The temperature is evenly distributed between 18 and 26 degrees C”
- P(Temperature=x)= U[18,26](x) = 0.125/C
  - NOTE: the probability that the temperature is 20 degrees is NOT 0.125
  - For any continuous distribution, meaning of P(X=c) is the prob. that X falls into an interval around c divided by the width of the interval, as the interval size goes to zero:

$$P(X = c) = \lim_{dx \to 0} P(c \leq X \leq c + dx)/dx$$

Uniform distribution

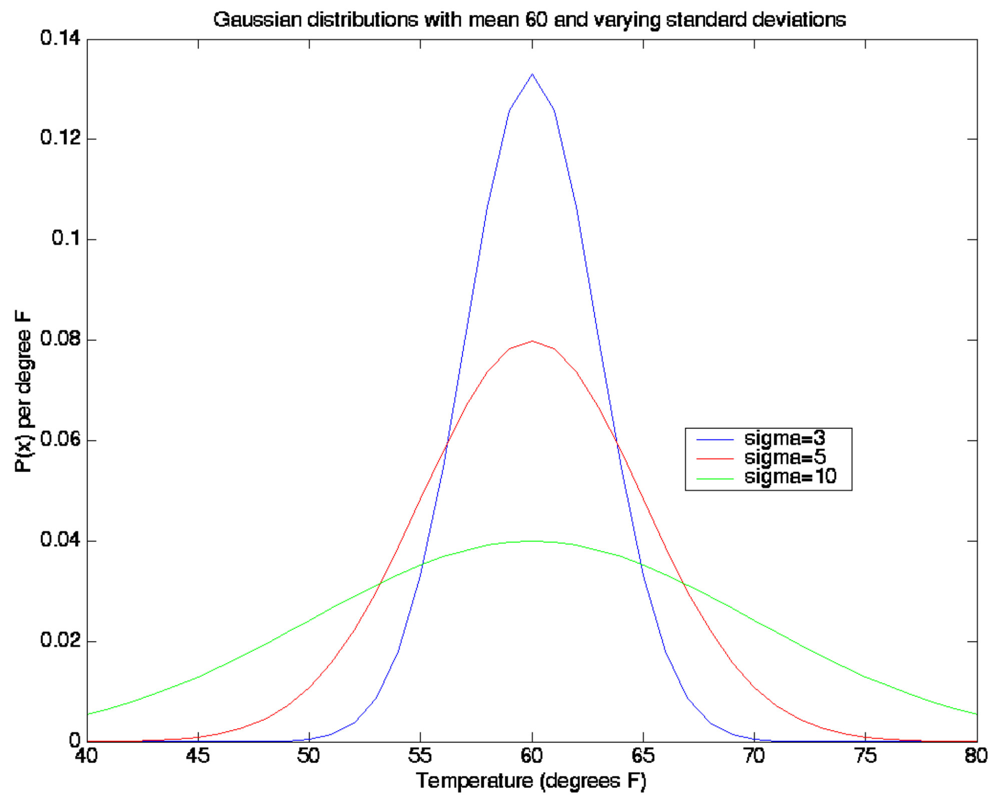
- P(Temperature=x)= U[18,26](x) = 0.125/C
  - P(20<Temp<22) = 2C * 0.125/C = 0.25
  - P(19<Temp<19.5) = 0.5C * 0.125/C = 0.0625
  - P(17<Temp<19)=
    1C*0/C + 1C*0.125/C = 0.125
- Distribution characterized by 2 numbers:
  - upper bound, lower bound
  - U[lb,ub](x) = 1/(ub-lb) /unit

Normal Distribution

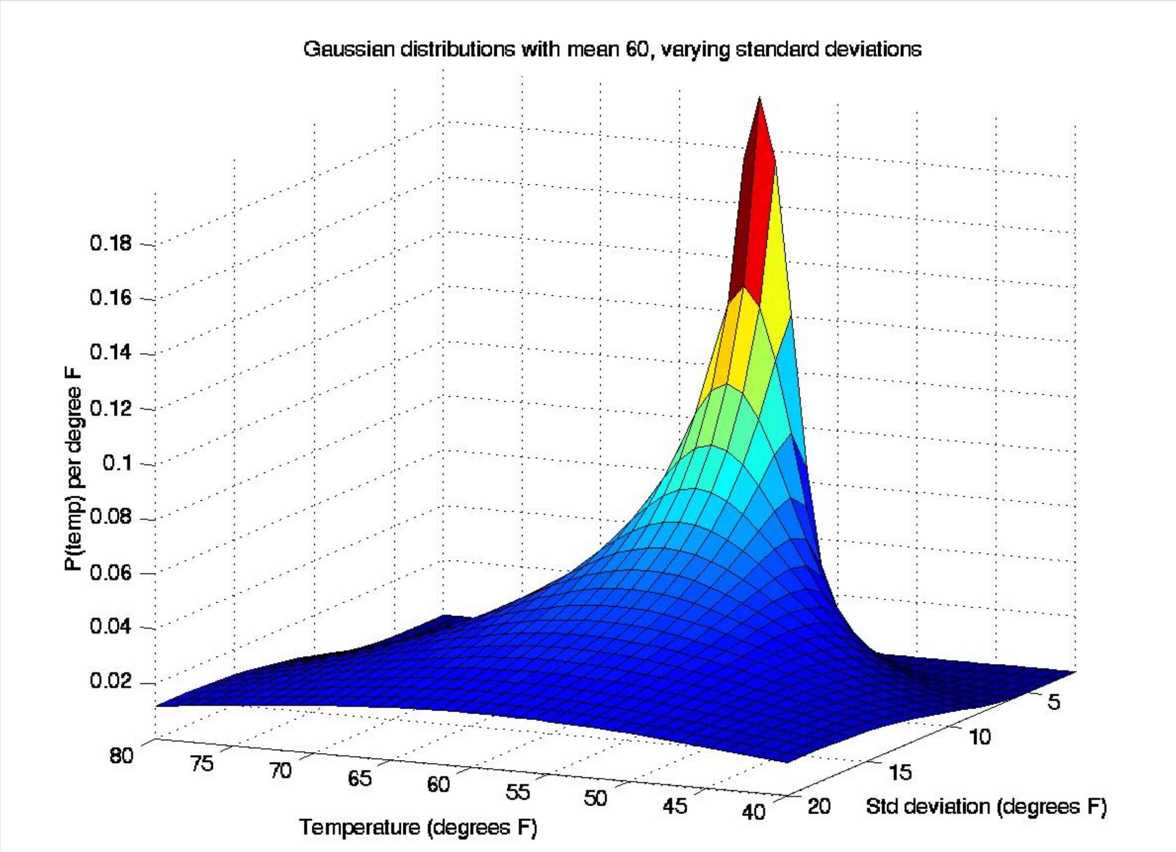
- Also characterized by two parameters (for one-dimensional case)
  - Mean: μ
  - Standard deviation: σ
- One-dimensional formula:

$$P(x) = N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

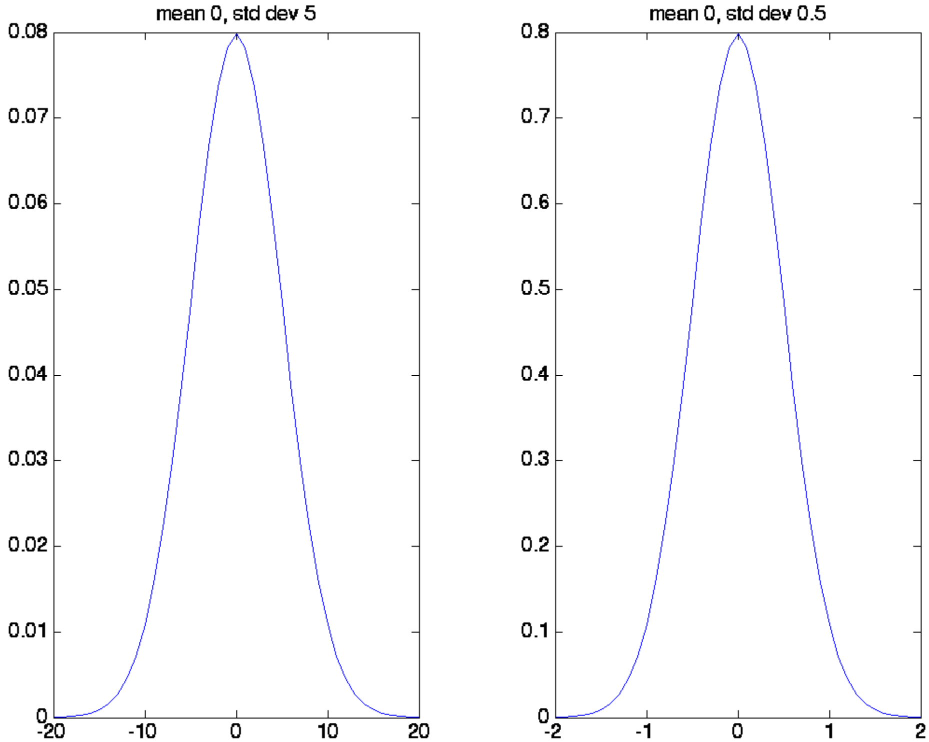
Gaussian (normal) distributions



Another view of the same data



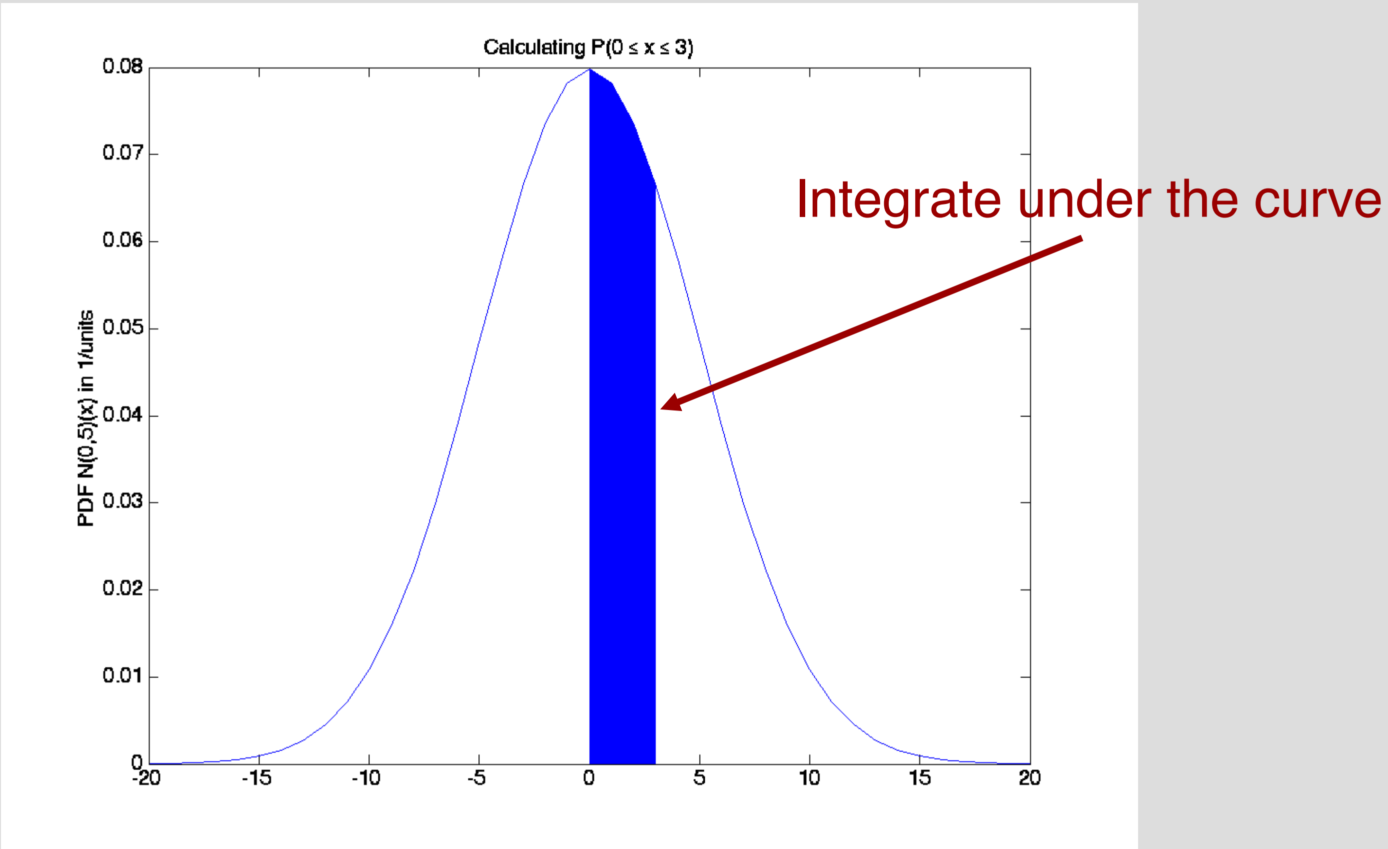
It’s all a matter of scaling



Scale a

Scale b=a/10

Finding probabilities from PDFs



Two dimensional Gaussian

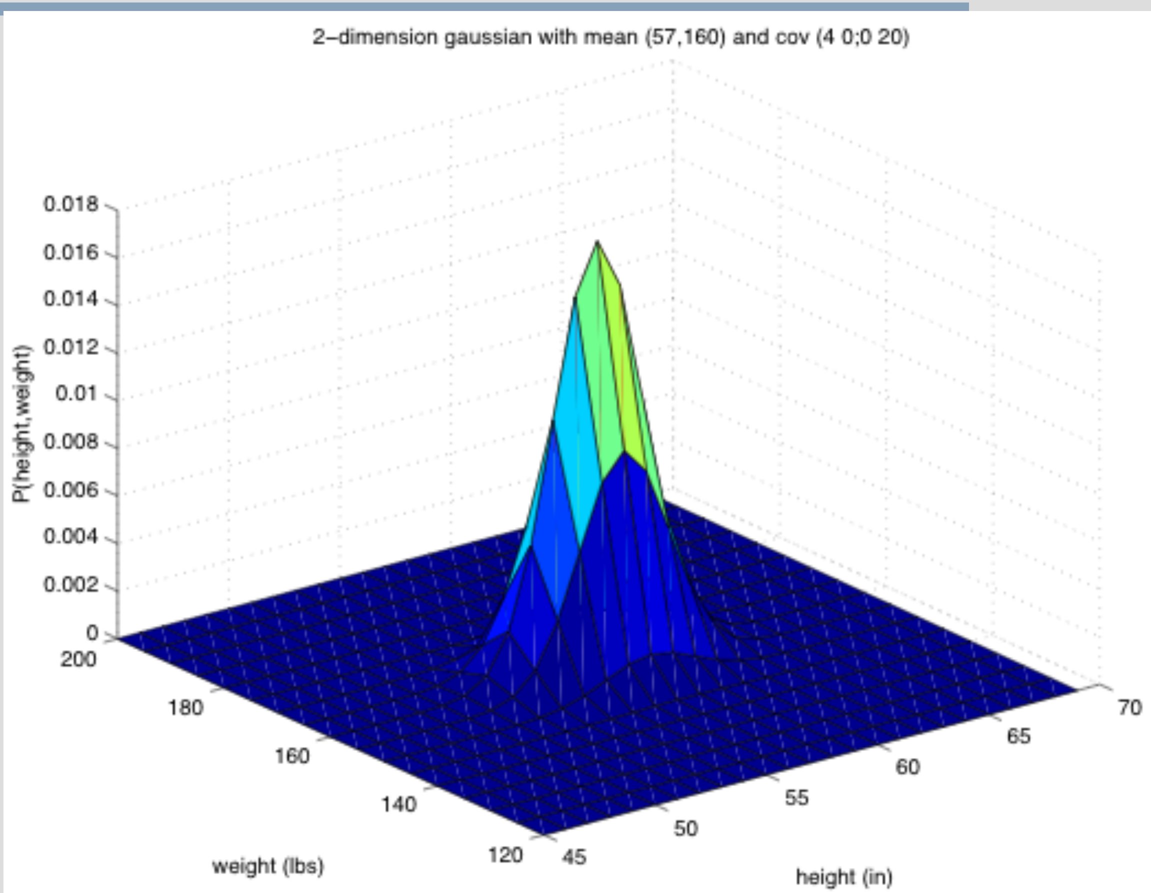
- We can have data points with more than one dimension
  - e.g. height and weight
- One dimensional gaussian:

$$P(x) = N(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Multidimensional gaussian

$$N(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)}$$

Two dimensional Gaussian



Two dimensional Gaussian

- Mean is just a vector
  - Height/weight: (57,160)
- Covariance is a matrix
  - If variables are independent, then covariance is diagonal
  - Diagonal:

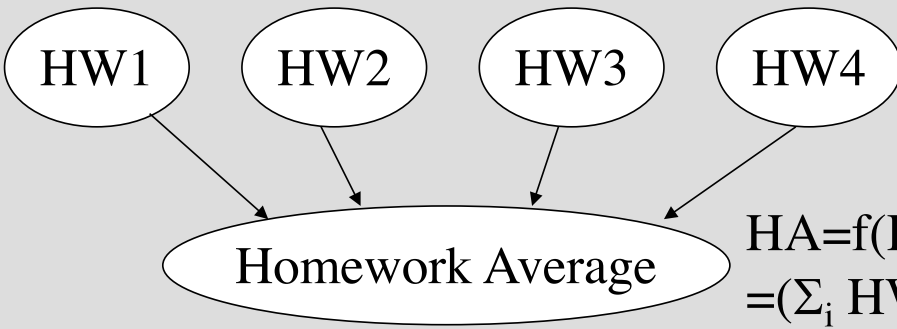
$$\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

Non-diagonal:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

Deterministic nodes

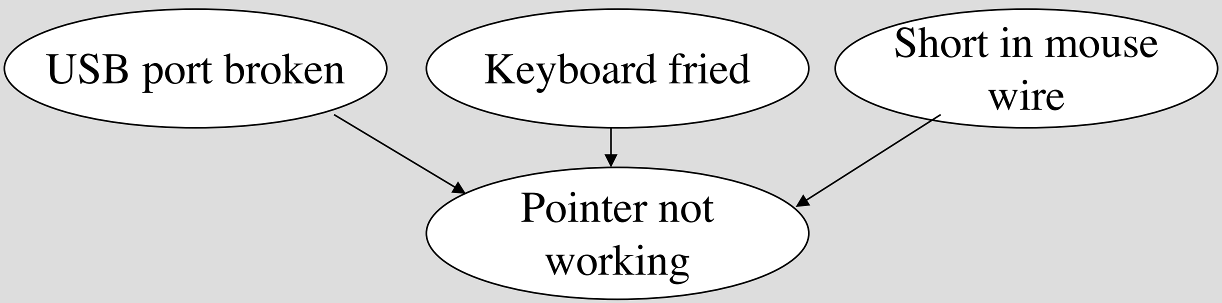
- If one node can be expressed as deterministic function of others, then more compact



$HA=f(HW_1 \ldots HW_4)$
$=(\Sigma_i HW_i - \min_i(HW_i))/3$

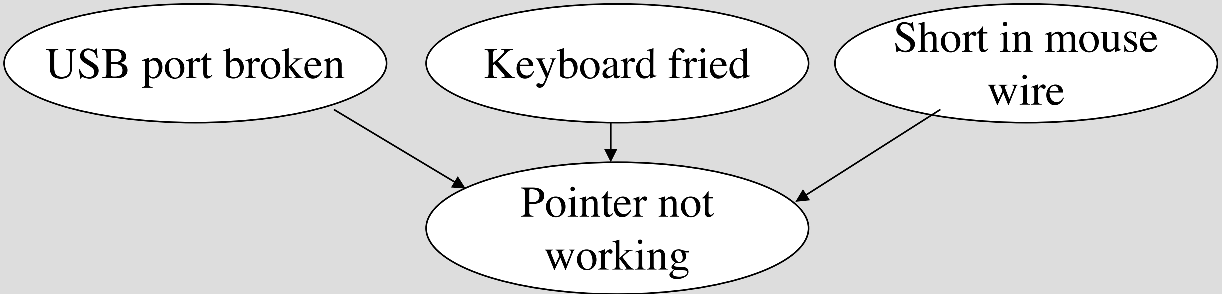
Noisy-or

- Like "or" in logic, but with uncertainty
  - x or y -> z in logic: P(z|x,y)=1, P(~z|x,~y)=0
  - With uncertainty: P(~z|x,~y) >0
    - "Inhibition" probability
- Assume we know all causes of an event
- We assume that inhibition of parents is independent



Noisy-or

- Specify inhibition probabilities
  - P(~p | u, ~k, ~s) = .05
    P(~p | ~u, k, ~s) = .10
    P(~p | ~u, ~k, s) = .5
    P(~p | ~u, ~k, ~s) = 1 (assumption)
- Can calculate "noisy-or" from these
  - P(p | u, ~k, s) = 1-P(~p|u,~k,s) = 1-(.05*.5)=.975
- Linear in number of parents

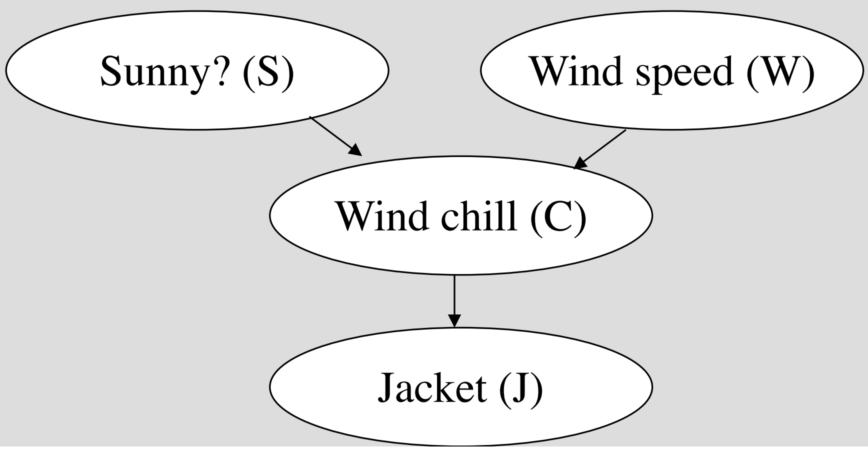


Bayes nets with continuous variables

- Problem: the perceived temperature (wind chill) depends on whether its sunny and the wind speed; if the wind chill is low then I’ll wear a jacket

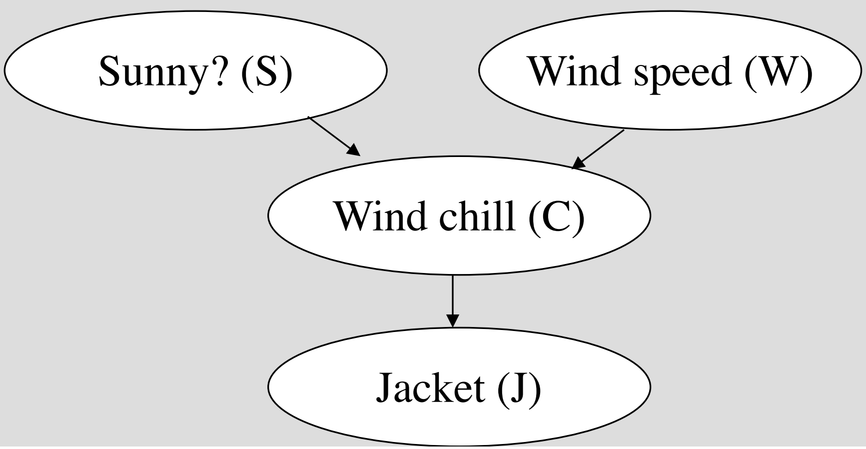
Bayes nets with continuous variables

- Problem: the perceived temperature (wind chill) depends on whether its sunny and the wind speed; if the wind chill is low then I’ll wear a jacket



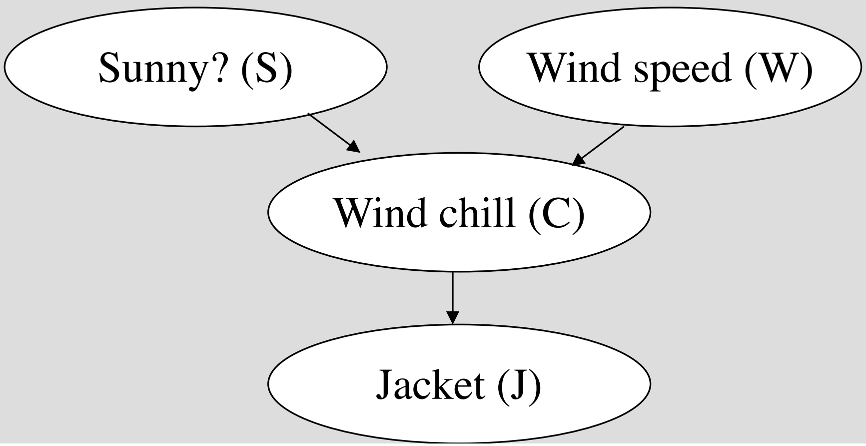
Bayes nets with continuous variables

- P(S) = <a,1-a>
- P(W) = N(μ,σ)
- P(C|W,s)=??? (continuous)



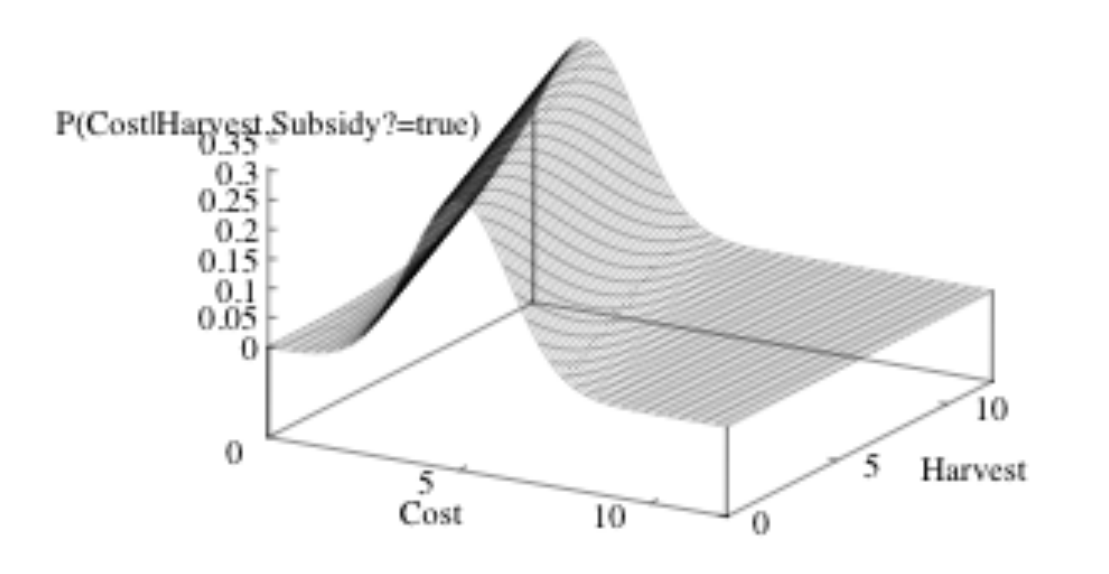
Bayes nets with continuous variables

- $P(S) = \langle a, 1-a \rangle$
- $P(W) = N(\mu, \sigma)$
- $P(C|W,s) = N(a_s w + b_s, \sigma_s)$
- $P(C|W,\sim s) = N(a_{\sim s} w + b_{\sim s}, \sigma_{\sim s})$



Linear Gaussian

- Example from book: cost depends on harvest (continuous), subsidy (binary)



What about P(jacket | Chill)?

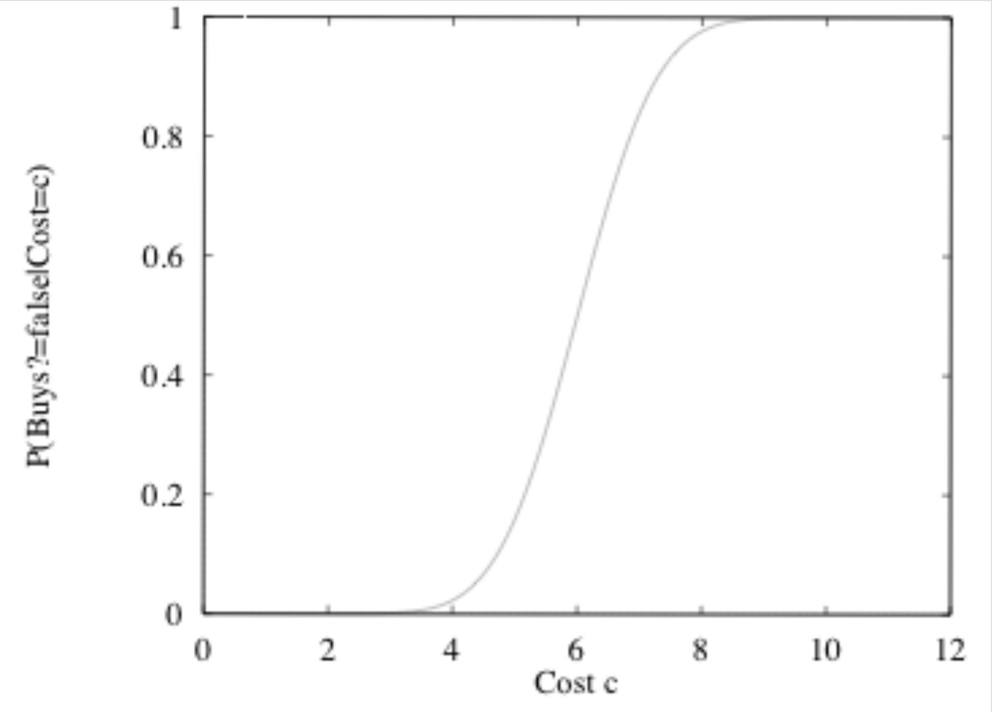
- Need to threshold -- at some temperature, I will get my jacket
  - P(j | C<θ)=1, P(j | C>=θ)=0
- In real life, don't know where boundary is
- Probit distribution: assume there is noise around decision boundary

$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)\,dx \qquad P(j \mid c) = \Phi(-(c - \theta)/\sigma)$$

threshold/inflection point

noise deviation

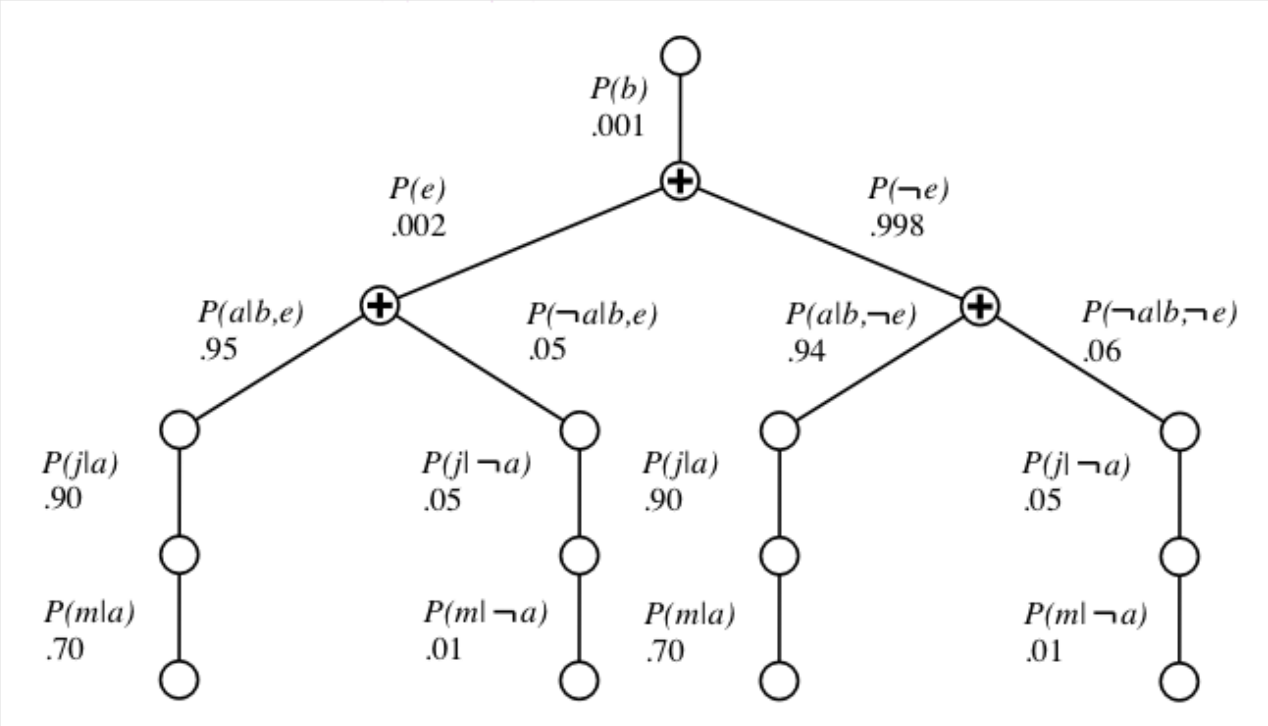
Probit distribution



Bayesian Inference Algorithms

- Already showed exact inference
  - Can always use this
  - However, exact inference is NP-hard in multiply-connected networks
    - Singly-connected network (polytree) is linear in #CPT entries
      - (at most one undirected path between nodes)
  - Enumeration is also inefficient

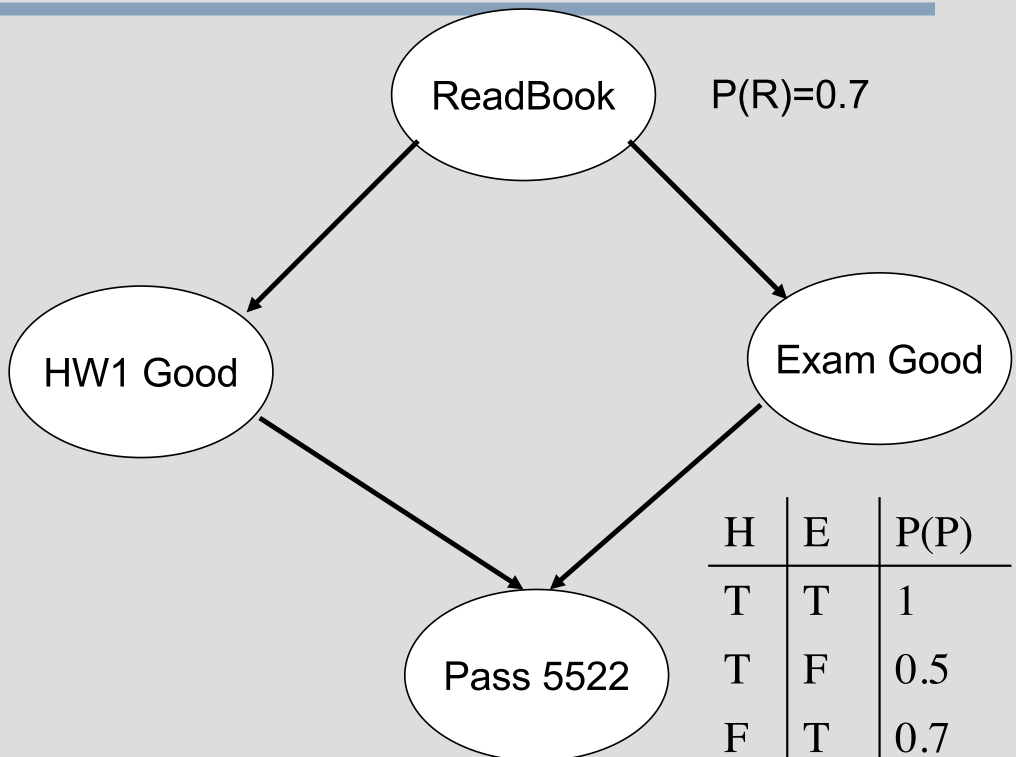
Enumeration Evaluation Tree



Inference options

- Variable elimination
  - Exact inference, still NP-hard
- Approximate inference
  - Randomly generate instances according to CPTs
  - Compute probabilities by counting instances

Sampling from an empty network



P(R)=0.7

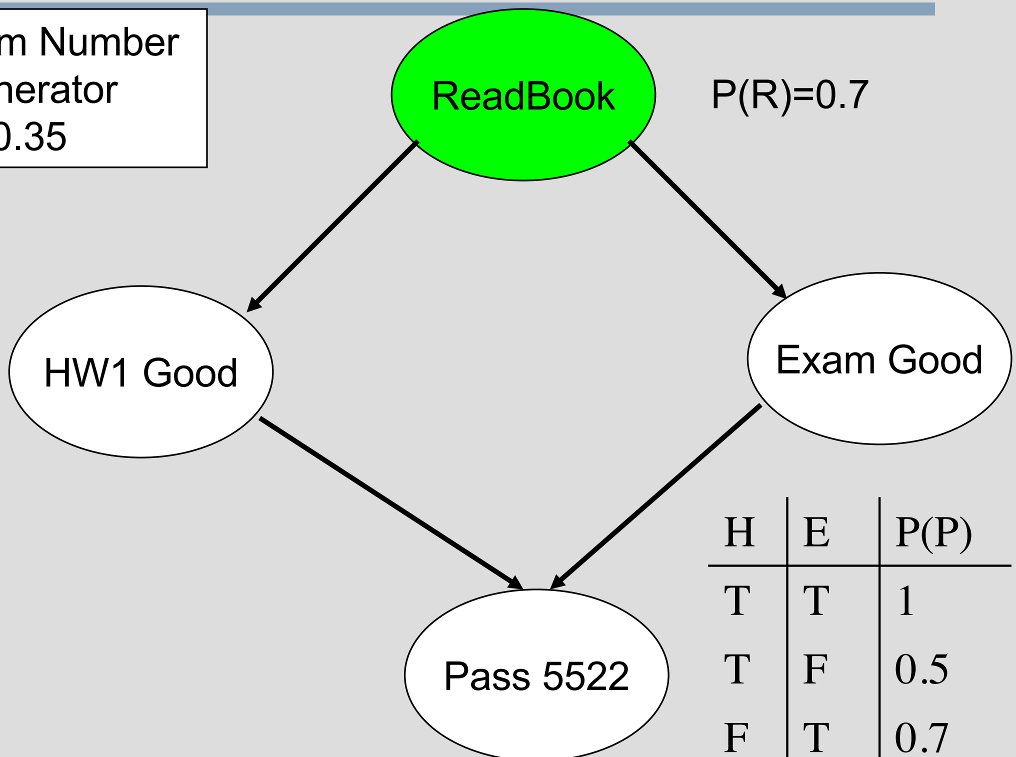
R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Sampling from an empty network

Random Number Generator
0.35



P(R)=0.7

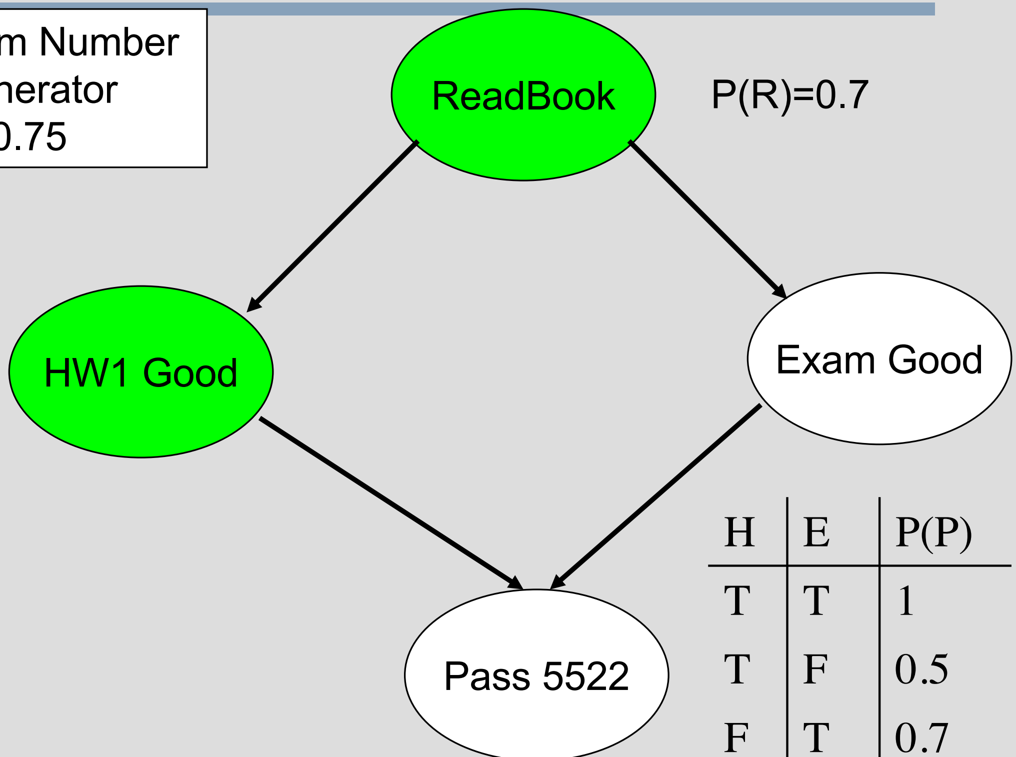
R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Sampling from an empty network

Random Number Generator
0.75



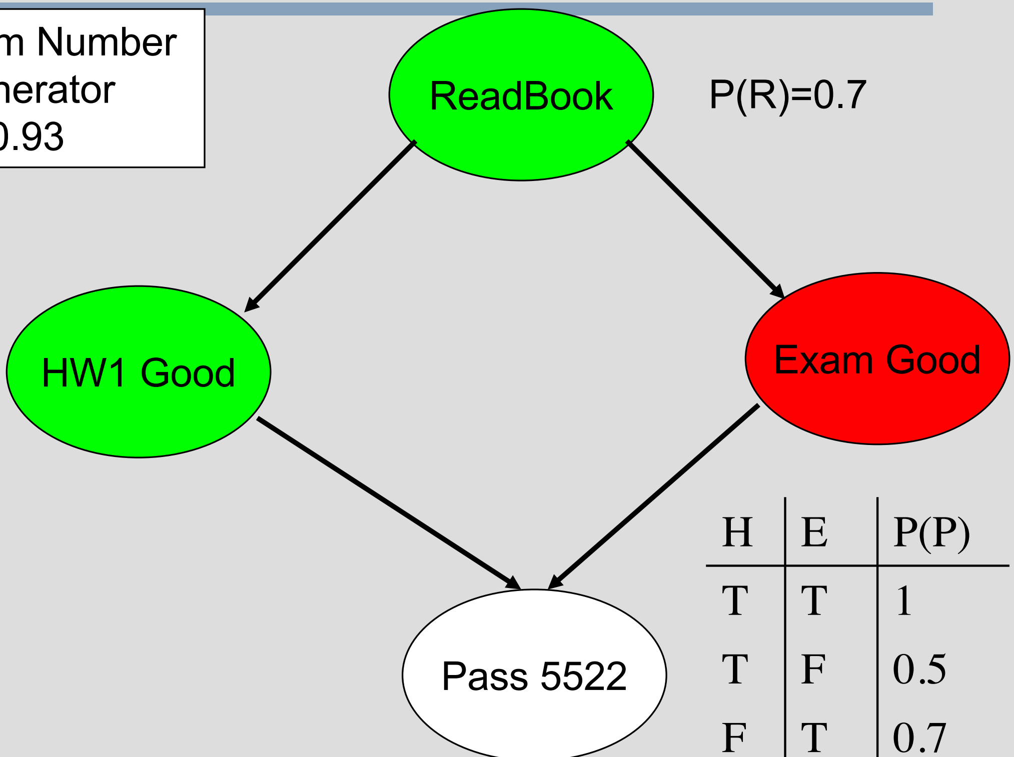
R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Sampling from an empty network

Random Number Generator
0.93



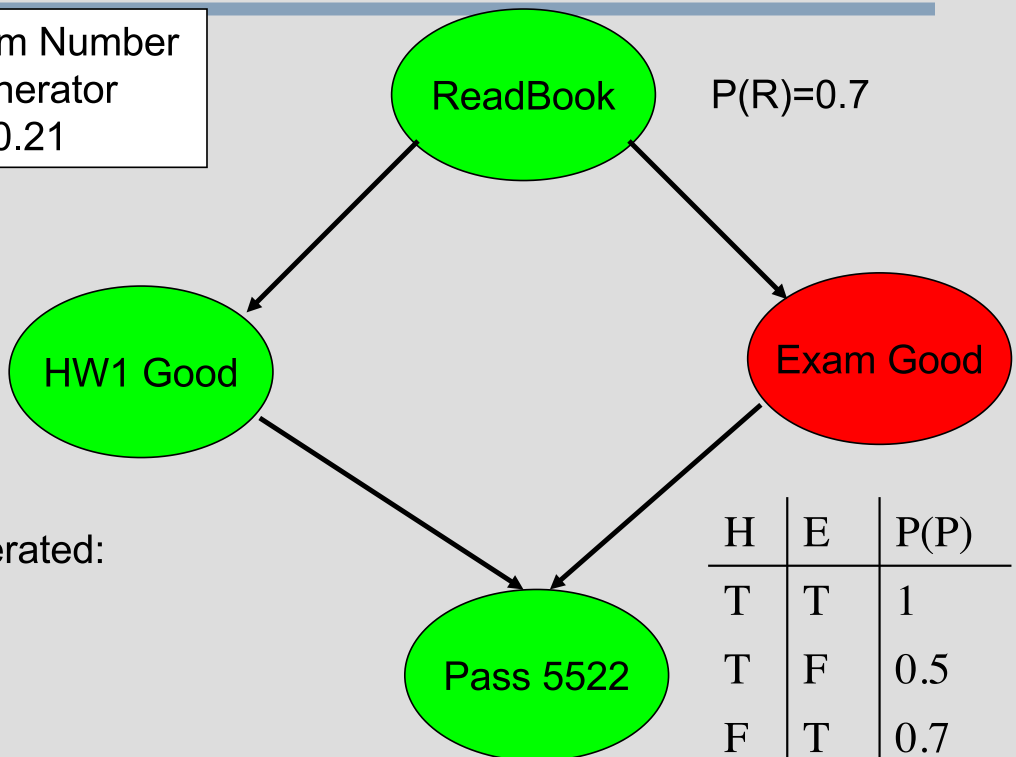
R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Sampling from an empty network

Random Number Generator
0.21



P(R)=0.7

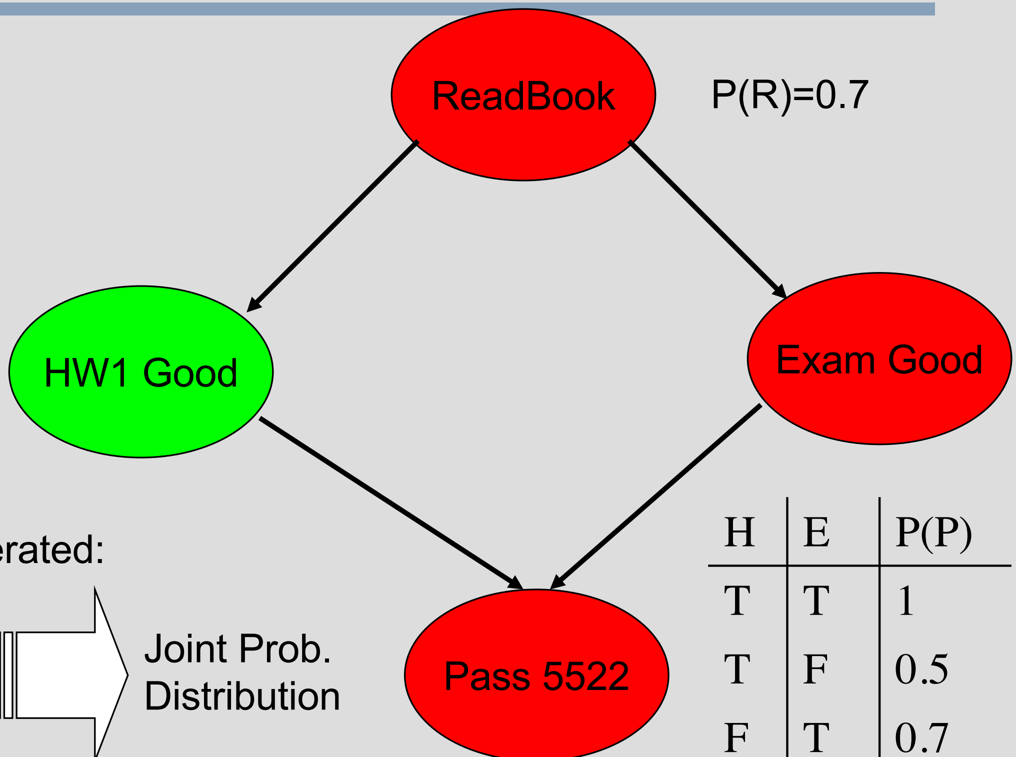
R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Events generated:
r,h,¬e,p

Sampling from an empty network



P(R)=0.7

R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

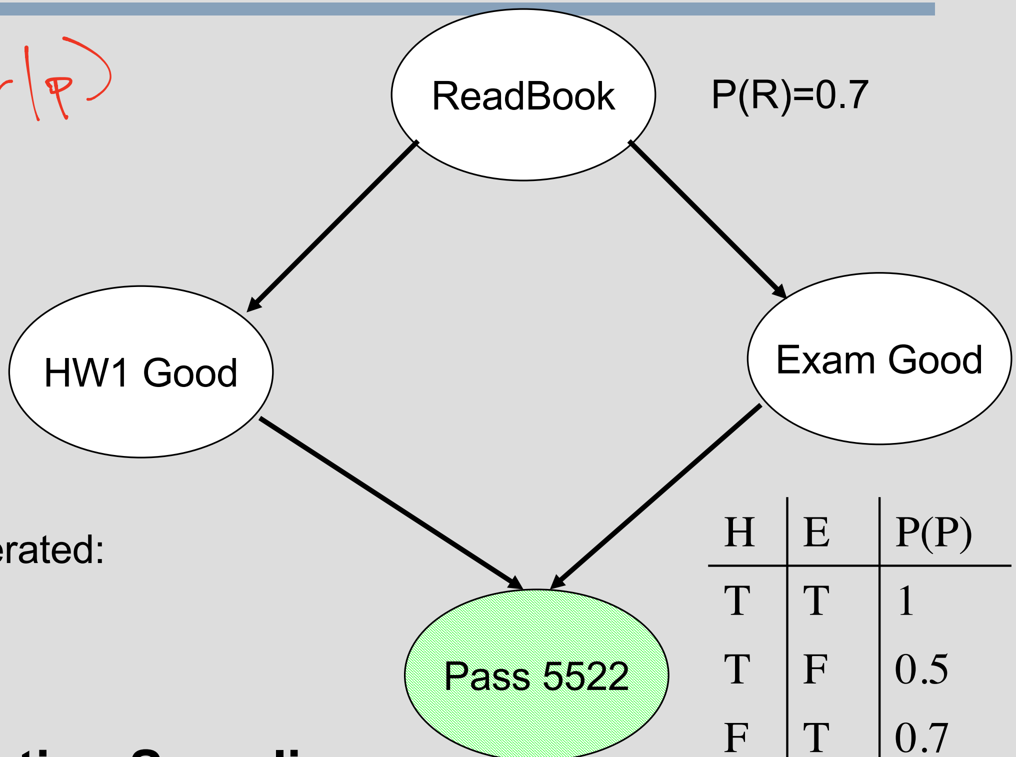
H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Events generated:
r,h,¬e,p
¬r,h,¬e,¬p
…

Joint Prob.
Distribution

What happens when you have evidence?

P(r|p)



P(R)=0.7

R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

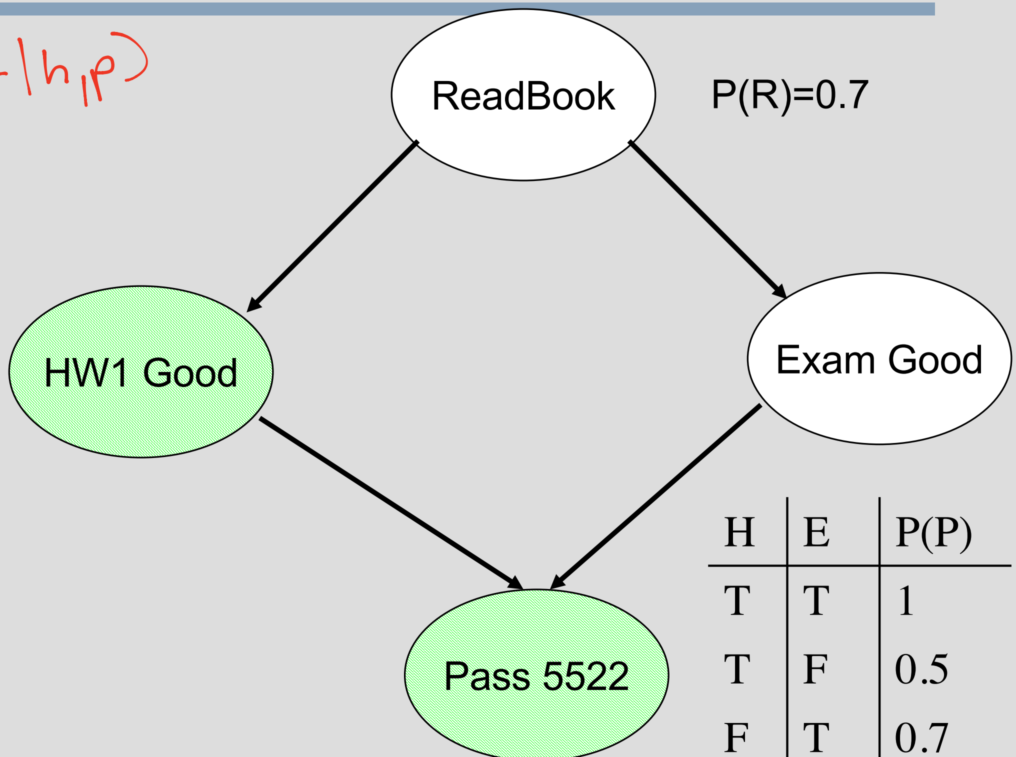
H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

Events generated:
r,h,¬e,p
~~¬r,h,¬e,¬p~~
…

Rejection Sampling

What happens when you have evidence?

$P(r|h,p)$



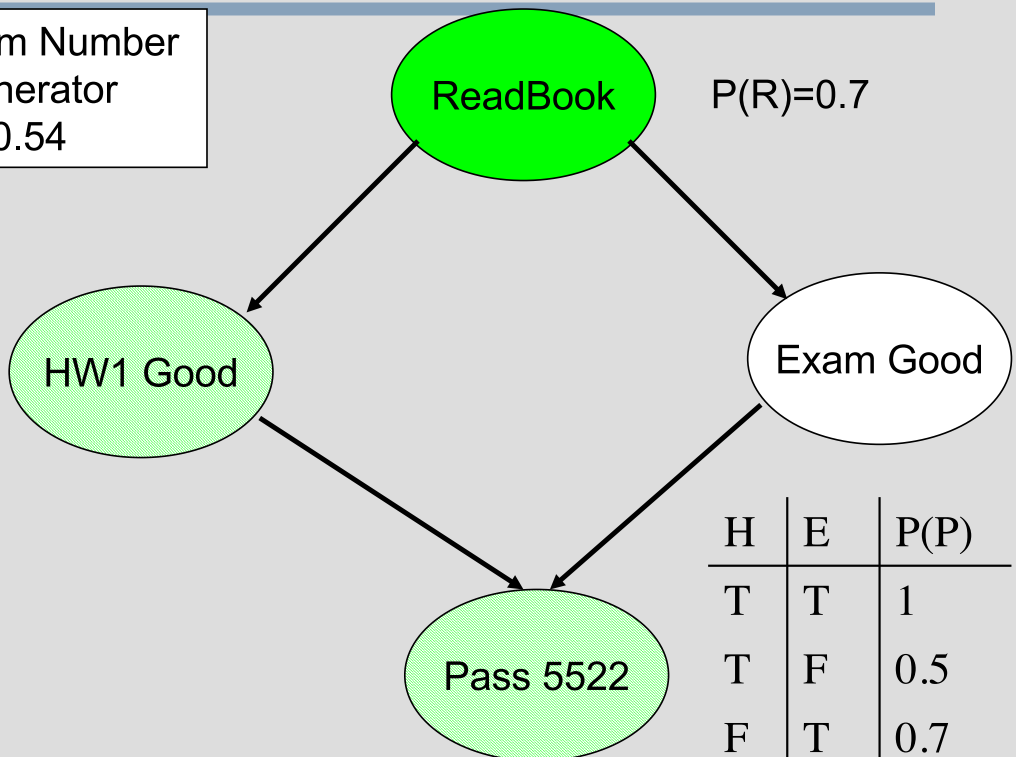
R	P(H)
T	.9
F	.3

R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

What happens when you have evidence?

Random Number Generator
0.54



P(R)=0.7

R	P(H)
T	.9
F	.3

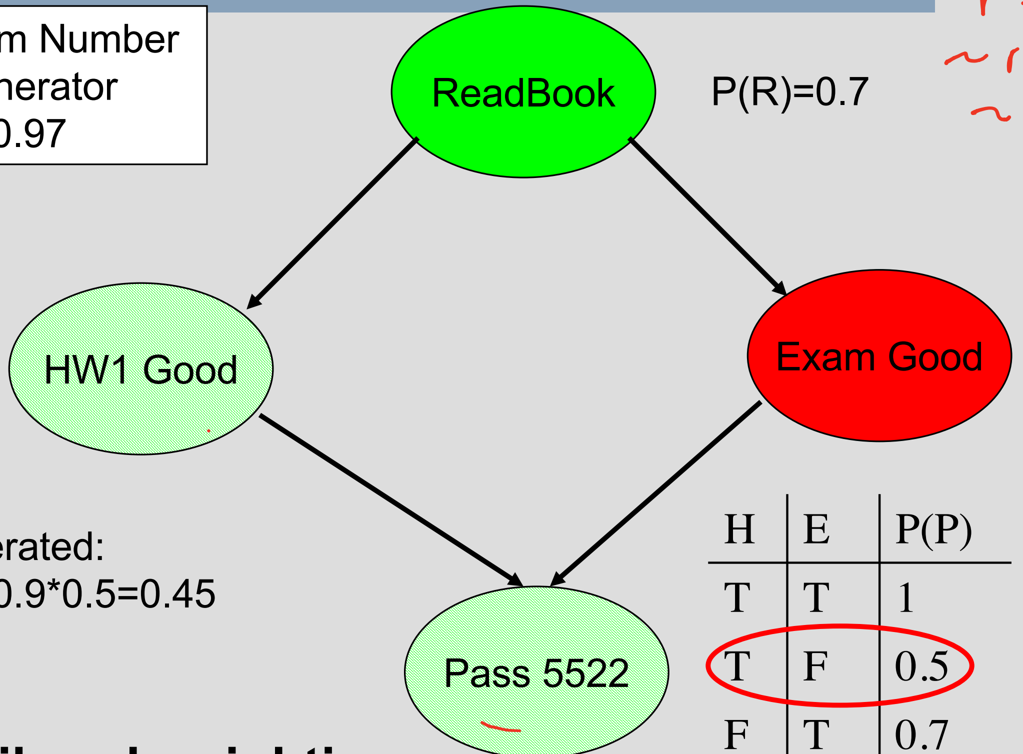
R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

What happens when you have evidence?

rehp 0.9
r~ehp 0.45
~rehp 0.3
~r~ehp 0.15

Random Number Generator
0.97



P(R)=0.7

R	P(H)
T	.9
F	.3

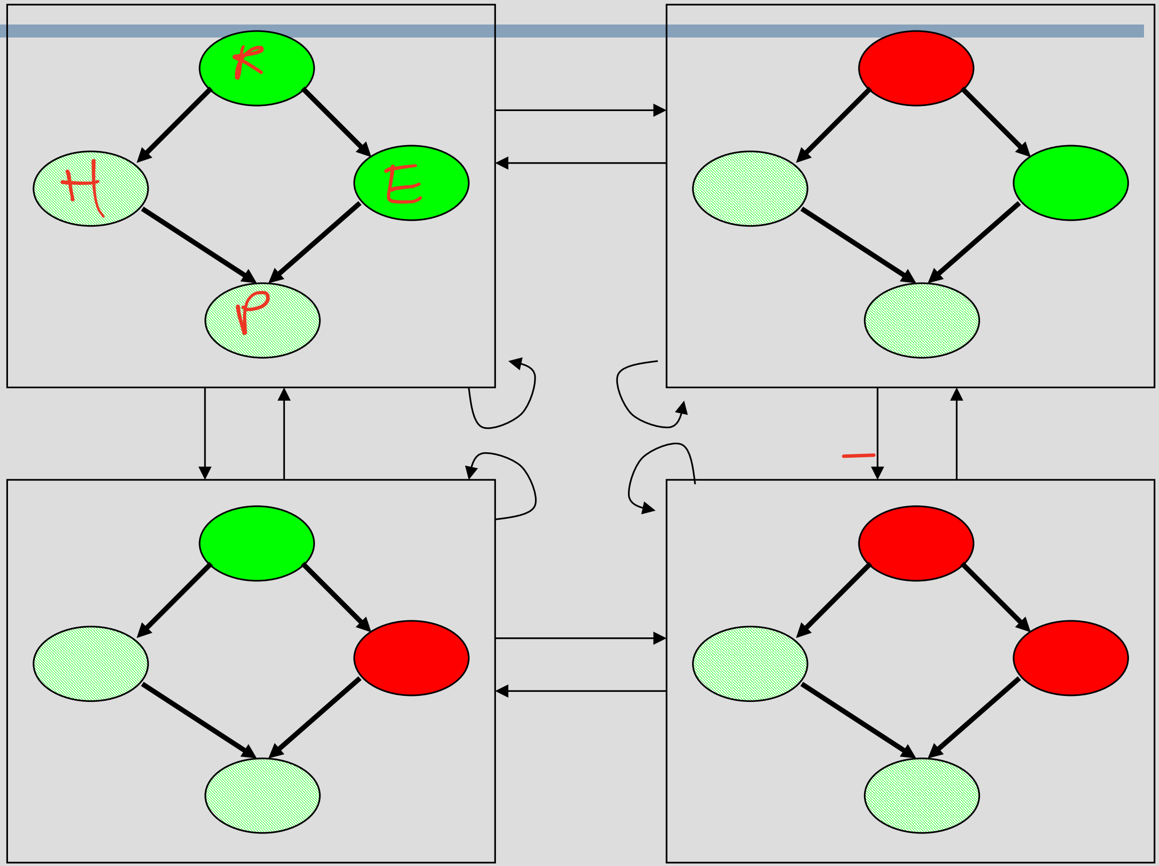
R	P(E)
T	.9
F	.1

H	E	P(P)
T	T	1
T	F	0.5
F	T	0.7
F	F	0.1

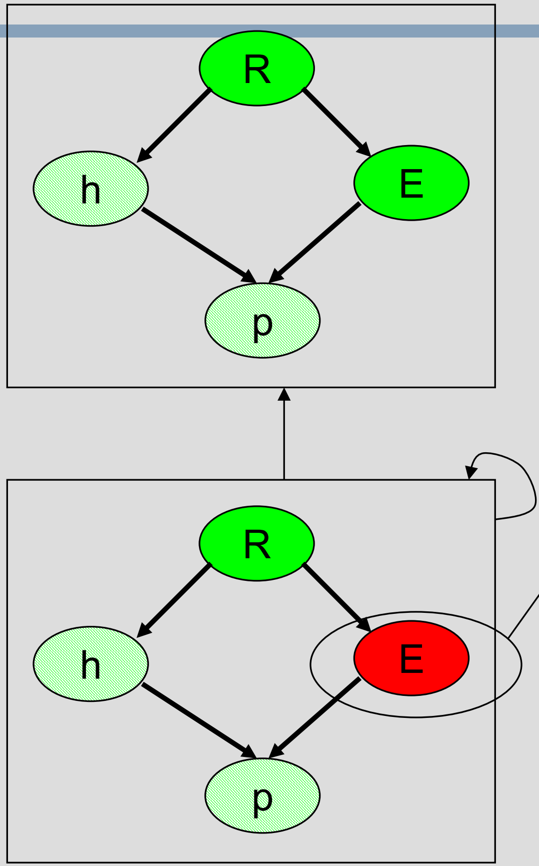
Events generated:
r,h,¬e,p w=0.9*0.5=0.45
…

Likelihood weighting

Markov Chain Monte Carlo



Markov Chain Monte Carlo



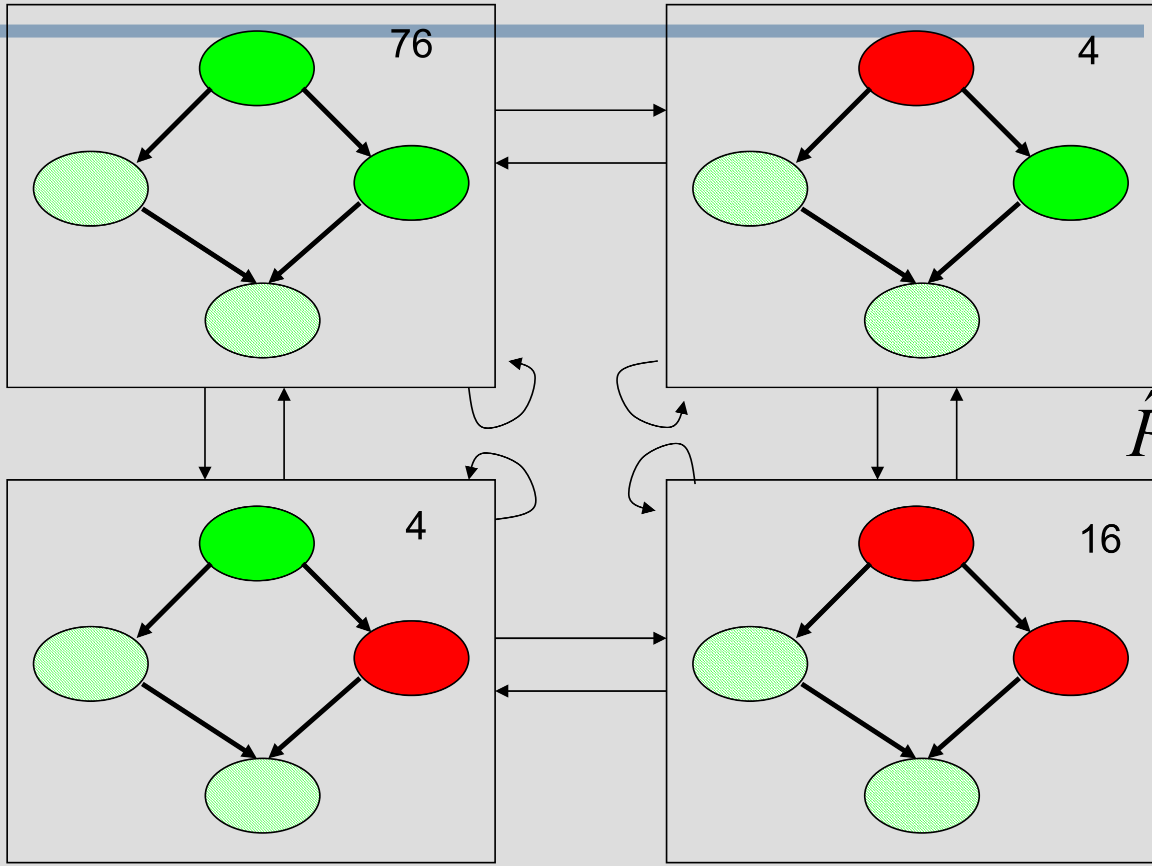
$$P(e' \mid MB(E)) =$$

$$\alpha P(E \mid Par(E)) \prod_j P(Z_j \mid Par(Z_j)) =$$

$$\alpha P(E \mid r) P(p \mid E, h)$$

Gibbs Sampling

Markov Chain Monte Carlo



$$\hat{P}(R \mid h, p) = ?$$