

2011 Fall CS 598 RAR  
Probabilistic Graphical Models in AI

Lecture 12: Approximate Inference via Sampling  
(AKA "Particle" Methods)

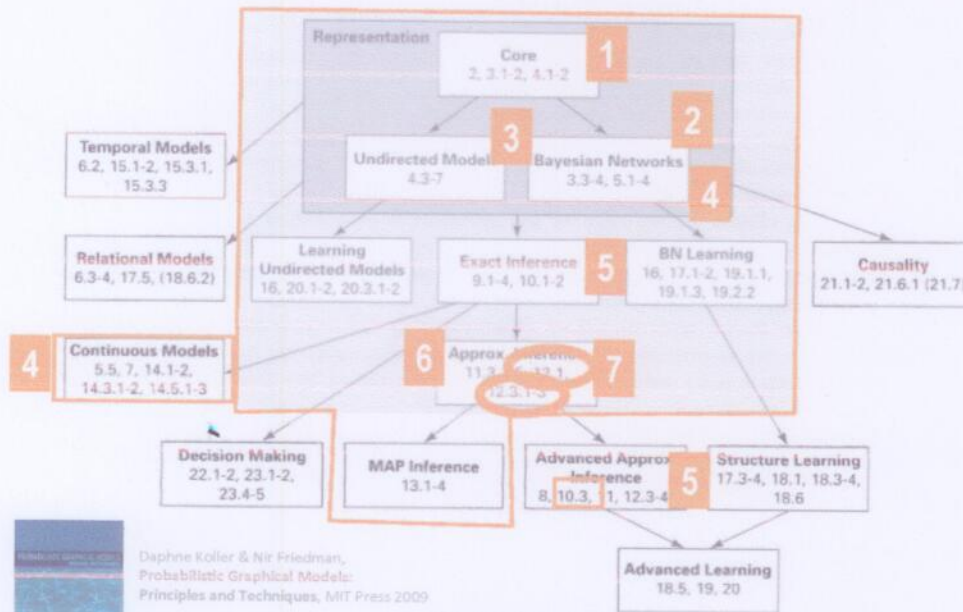
Rob A. Rutenbar  
Bliss Professor and Head



## PGM Class Overview: Where Are We?

- |           |              |              |
|-----------|--------------|--------------|
| ■ Week 1: | 8/23         | 8/25         |
| ■ Week 2: | 8/30         | 9/1          |
| ■ Week 3: | 9/6          | 9/8          |
| ■ Week 4: | <u>9/13</u>  | 9/15         |
| ■ Week 5: | 9/20         | 9/22         |
| ■ Week 6: | 9/27         | <u>9/29</u>  |
| ■ Week 7: | 10/4         | <u>10/6</u>  |
| ■ Week 8: | 10/11        | 10/13        |
| ■ Week 9: | 10/18        | <u>10/20</u> |
| ■ Week10: | 10/25        | <u>10/27</u> |
| ■ Week11: | <u>11/1</u>  | 11/3         |
| ■ Week12: | <u>11/8</u>  | 11/10        |
| ■ Week13: | 11/15        | 11/17        |
| ■ Week14: | Off, Thxgive |              |
| ■ Week15: | 11/29        | 12/1         |
| ■ Week16: | 12/6         | --           |
- 10/27, 11/1
    - Lec 12 Inference via Sampling
    - Read KF Chap 12.1, 12.2, 12.3
  - Acknowledgements/Sources
    - Koller/Friedman book, Chap 12
    - Andrew McCallum, Umass, CS691 Graphical Models, Lec 15 (Approx Inference by Sampling )
    - <http://www.cs.umass.edu/~mccallum/courses/gm2011/>
    - Ajit Singh, CMU, CS 10-708, Lec 11/10/2008, Approx Inference by Sampling
    - [http://www.cs.cmu.edu/~jtsingh/Class/10708\\_F08/index.html](http://www.cs.cmu.edu/~jtsingh/Class/10708_F08/index.html)

## Overall Gameplan: KF Chap 11 "Infer as Opt"



Slide 3

## You've Seen Some Elementary Sampling Ideas

- Suppose we have a (real valued) random variable  $X$ .

- $X$  takes values  $x \in \text{Val}(X) = \{x^1, x^2, \dots, x^V\}$ , with prob  $P(X=x^i)$
- What is the expected value,  $E[X]$ ?

$$E(X) = \sum_{x=x^i} P(X=x^i) \cdot x^i$$

- What if you observe a set of  $M$  samples of  $X$ ?

- Observe  $X = x[1], x[2], \dots, x[M]$ , all drawn from  $P(X)$  distrib
- How would you **approximate**  $E[X]$  from these observations?

$$E(X) \approx \frac{1}{M} \sum_{m=1}^M x[m]$$



## More Generally, for any Function $f(X)$

- Can get expected value  $E_p[f]$

$$E_p[f] = \sum_{x \in \text{val}(X)} f(x)P(X=x) \rightarrow E_p[f(X)] \approx \frac{1}{M} \sum_{\text{samples } m=1}^M f(x[m])$$

- Can also estimate **individual probabilities**,  $P(X=x)$

- KF notation: indicator function  $\mathbf{1}(x[m]=x) = \begin{cases} 1 & \text{if } x[m]=x \\ 0 & \text{else} \end{cases}$

$$P(X=x) = E[\mathbf{1}(X=x)] = \frac{1}{M} \sum_{m=1}^M \mathbf{1}(x[m]=x) = \frac{\# \text{ of } x \text{ in } M \text{ samples}}{M \text{ samples}}$$

- Key pt:

- Everything** interesting can be cast as finding  $E[\text{some func } f(X)]$

## Aside: Works in N Dimensions as Well...

- $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$  is a set of  $N$  random vars

- We also have a **joint** prob distribution  $P(\mathbf{X} = (x^1, x^2, \dots, x^n))$

- We observe a set of  $M$  values of  $\mathbf{X}$ , drawn from this distrib

$$\mathbf{X}[1] = (X_1[1], X_2[1], \dots, X_N[1])$$

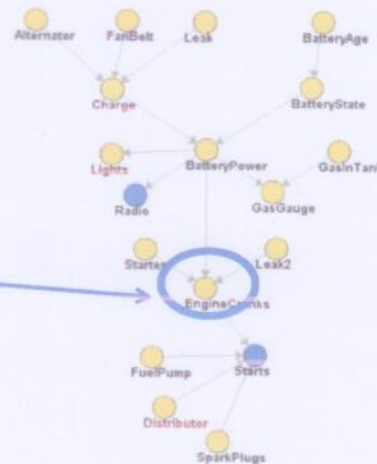
...

$$\mathbf{X}[M] = (X_1[M], X_2[M], \dots, X_N[M])$$

$$E[X_2] \approx \frac{1}{M} \sum_{m=1}^M X_2[m]$$

## Why We Care: Approx Marginals in PGMs

- If we can efficiently **sample** from the joint distribution defined by an arbitrary PGM, we can answer questions we care about -- approximately



- Like what?
  - Unconditional:  
 $P(Y=y)$
  - Conditional (evidence):  
 $P(Y=y \mid E=e)$

© Rob A. Rutenbar 2011

Slide 7

## About this Lecture

- All about doing approximate inference via **sampling**
  - Random sampling – samples from the “right” distribution
  - For a BN:  $\prod_i P(X_i \mid \mathbf{Pa}_{X_i})$  For a MN:  $(1/Z) \prod_i \phi_i$

- Lots of terminology flying by**

- Particles:** Another name for “samples”
- Monte Carlo:** Broad class of random sampling methods, good for doing things like  $E[X]$  and estimating  $P(X=x)$
- Markov Chain:** A particular class of probabilistic graphs – NOT PGMs – useful in connection with Markov Chains
- MCMC:** Markov Chain Monte Carlo .. topic at end of lec

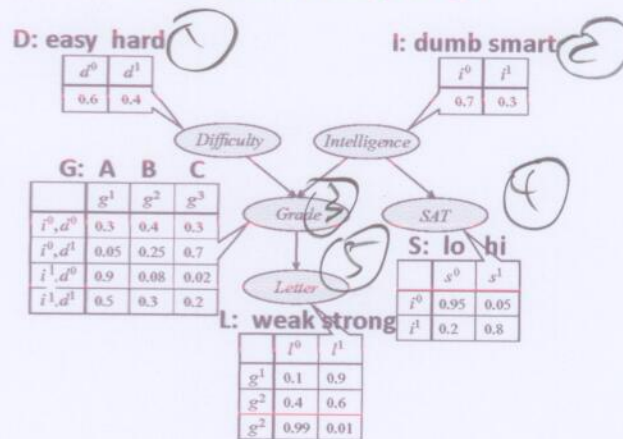
*Monte Carlo fix*

© Rob A. Rutenbar 2011

Slide 8

## Conceptually Easiest for BNs: Forward Sampling

- Sample nodes in topological order
  - ...ie, "forward" from roots to leaves, follow directed edges
- At each node
  - Draw a random sample from the local CPD at this node...
  - ...and which matches the values already selected for vars seen previously in the "forward" walk down BN

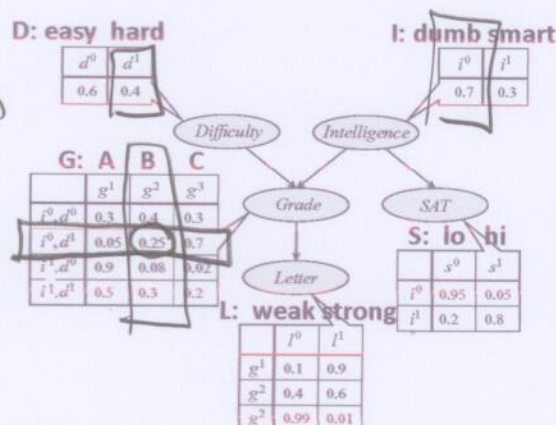


© Rob A. Rutenbar 2011

Slide 9

## BN Forward Sampling: Example

- Sample D  
*D = hard (0.4 prob)*
- Sample I  
*I = dumb (0.7 prob)*
- Sample G (depends on D, I)  
*G = B (prob 0.25, given hard/dumb)*



© Rob A. Rutenbar 2011

Slide 10



## Aside: Sampling from Multinomial Distrib

### How to sample G...?

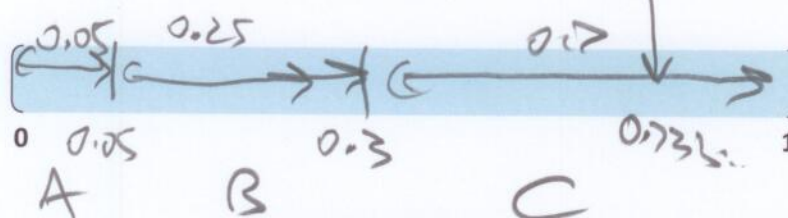
- Row of CPD adds up to 1
- Build a set of *contiguous buckets* across unit interval [0,1]
- Each bucket has **width**  $P(g^i)$
- Gen a *uniform* random  $r$  on [0,1], look at which bucket it lands in

| G: A B C   |       |       |       |
|------------|-------|-------|-------|
|            | $g^1$ | $g^2$ | $g^3$ |
| $i^0, d^0$ | 0.3   | 0.4   | 0.3   |
| $i^0, d^1$ | 0.05  | 0.25  | 0.7   |
| $i^1, d^0$ | 0.9   | 0.08  | 0.02  |
| $i^1, d^1$ | 0.5   | 0.3   | 0.2   |



Ex: Generate  $r = \text{rand}() = 0.73345321$

$r \Rightarrow \text{prob } G = C$



© Rob A. Rutenbar 2011

Slide 11

## BN: Forward Sampling

### 4. Sample S (depends on I)

$S = \text{lo}$  C prob 0.95, given dumb

### 5. Sample L (depends on G)

$L = \text{weak}$  C prob 0.1, given B

### Result: 1 sample from joint $P()$

- Now, repeat M times ( $M \sim \text{big}$ )
- Calculate the  $E_p[f()]$  as desired

(hard, dumb, B, lo, weak)

D: easy hard

|       |       |
|-------|-------|
| $d^0$ | $d^1$ |
| 0.6   | 0.4   |

I: dumb smart

|       |       |
|-------|-------|
| $i^0$ | $i^1$ |
| 0.7   | 0.3   |

| G: A B C   |       |       |       |
|------------|-------|-------|-------|
|            | $g^1$ | $g^2$ | $g^3$ |
| $i^0, d^0$ | 0.3   | 0.4   | 0.3   |
| $i^0, d^1$ | 0.05  | 0.25  | 0.7   |
| $i^1, d^0$ | 0.9   | 0.08  | 0.02  |
| $i^1, d^1$ | 0.5   | 0.3   | 0.2   |

Difficulty

Intelligence

Grade

SAT

Letter

S: lo hi

|       |       |
|-------|-------|
| $s^0$ | $s^1$ |
| 0.95  | 0.05  |
| $i^0$ | $i^1$ |
| 0.2   | 0.8   |

L: weak strong

|       |       |
|-------|-------|
| $l^0$ | $l^1$ |
| 0.1   | 0.9   |
| $g^1$ | $g^2$ |
| 0.4   | 0.6   |
| $g^2$ | $g^3$ |
| 0.99  | 0.01  |

## More Questions: How Big is M (#samples)?

- Can do theory in case that  $f(\cdot)$  is an indicator func, and we are trying to get marginals like  $\hat{P}(X=x) \approx (1/M) \sum_m 1(X=x)$ 
  - Technically, indicator function is a binary random var, each sample is an independent, identically distrib "Bernoulli trial".

- Chernoff bound (relative error of size  $\epsilon$ )

prob (answer is wrong by  $\epsilon$  relative)  $\rightarrow$

$$P[\hat{P}(X=x) \notin [(1-\epsilon)P(x), (1+\epsilon)P(x)]] \leq 2e^{-MP(x)\epsilon^2/3}$$

8

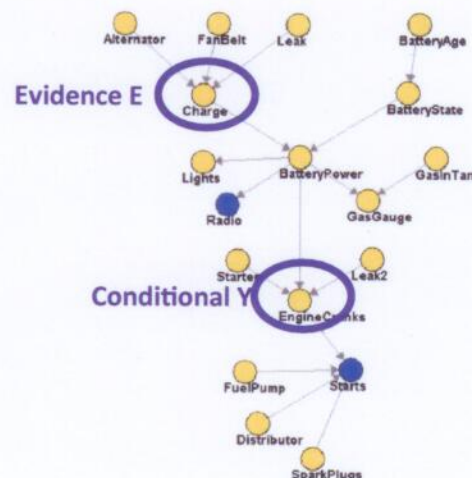
- KF Interpretation: to guarantee an accuracy of  $\epsilon$  with a probability of  $1-\delta$ , samples grows *logarithmically* with  $1/\delta$ , *quadratically* with  $1/\epsilon$ , and *linearly* with  $1/P(x)$
- In practice: **can't predict how many samples up front**

© Rob A. Rutenbar 2011

Slide 13

## Harder Sampling Task: $P(Y=y | E=e)$

- Why is this hard?
  - Because we need to generate samples that are...
    - (1) from the **correct** prob distribution...
    - (2) where evidence var  $E=e$  has the **correct** instance value
- The previous method won't work, we can't guarantee we get  $E=e$  in every random sample...



© Rob A. Rutenbar 2011

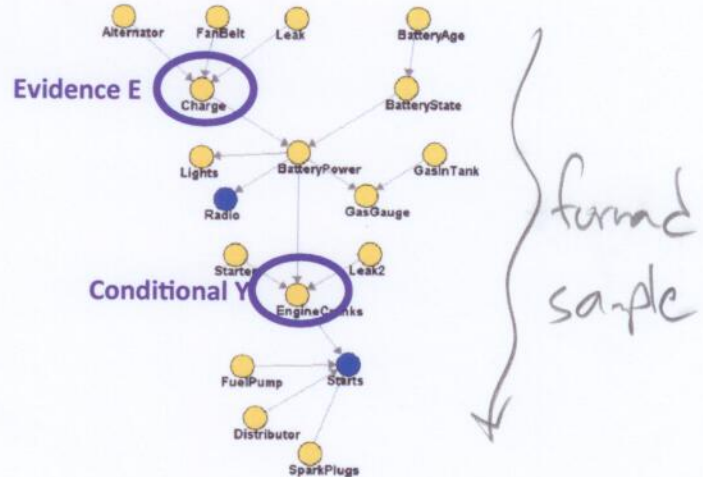
Slide 14



## Simple Solution for $P(Y=y | E=e)$ : Rejection Sampling

### Rejection method

- Set **NumSamples**=0
- Generate 1 random sample **S** using forward sampling method, as before
- If (evidence **E=e** in sample **S**) {  
Count this sample;  
**NumSamples**++;  
}
- else {reject this sample}
- Repeat till have **M** "correct" samples, each with **E=e**



© Rob A. Rutenbar 2011

Slide 15

## Problems with Rejection Sampling

### Yes, it works – but...

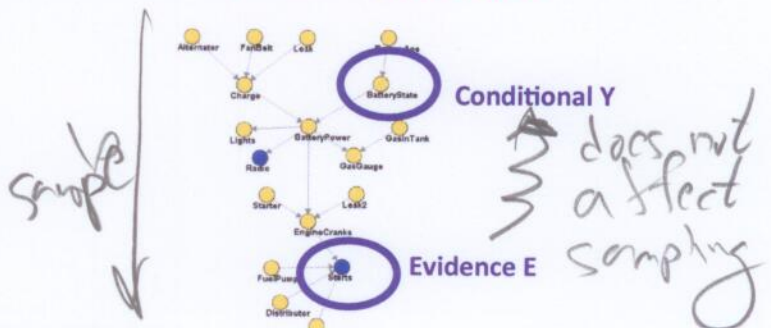
- Very **inefficient**
- What if  $P(E=e)$  is very **small**?  
Need **many** more samples now:
  - If needed **M** to get  $P(Y)$
  - Now need  $\sim M/P(E=e)$
  - This can be intractable

### Aside: just use ratios?

- Why not just calc both marginals  $P(Y,E)$ ,  $P(E)$  and do ratio  $P(Y,E=e)/P(E=e)$
- Answer: *still* hard to get low error, esp if  $P(E=e) = v$  small
- KF HW Prob 12.2

### More general problem with forward sampling:

- What if evidence is toward **leaves** of BN?
- Fixed  $E=e$  node only allows it to directly affects its **descendants**



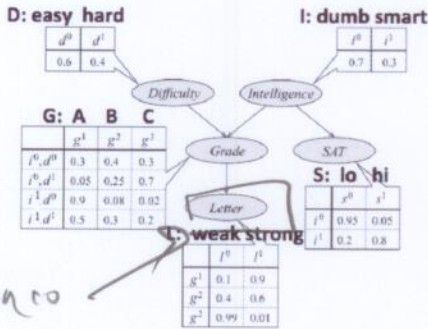
© Rob A. Rutenbar 2011

Slide 16



## Problems with Evidence and Forward Sampling...

- Good to be able to use the jargon properly, so let's analyze this statement from McCallum@Umass Lec15:
  - "If the evidence is in the leaves of the network, just sampling from the prior. Could be far from the posterior!"



$$\underbrace{P(X)}_{\text{PRIOR}} = \sum_e \underbrace{P(X|e)}_{\text{POSTERIOR}} \cdot \underbrace{P(e)}_{\text{EVIDENCE}}$$

you want  $P(X|E=e)$   
forward sampling doesn't get this  
right mainly sampling from  $P(X)$

© Rob A. Rutenbar 2011

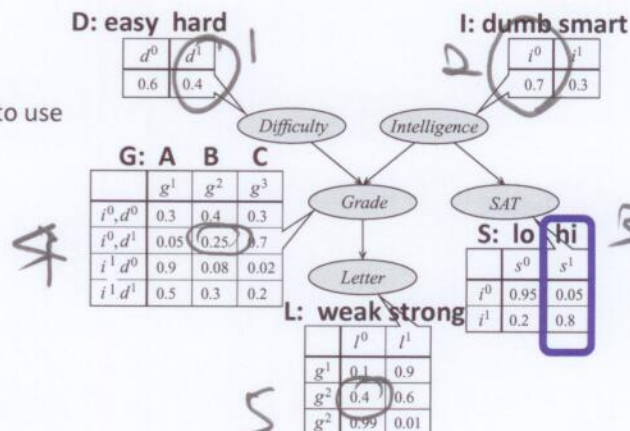
Slide 17

## Better Solution: Likelihood Weighting

- The intuition
  - Assume evidence is **SAT S=hi**
  - Lets try to force the sampling to use S=hi, like this...

- Sample D**
  - D=hard with Prob=0.4
- Sample I**
  - I=dumb with Prob=0.7
- Force S=SAT=hi**
  - Deterministic.
- Sample G (depends on D, I)**
  - G=B with Prob = 0.25
- Sample L (depends on G)**
  - L=weak with Prob 0.4

- Why is this **wrong**...?



© Rob A. Rutenbar 2011

Slide 18

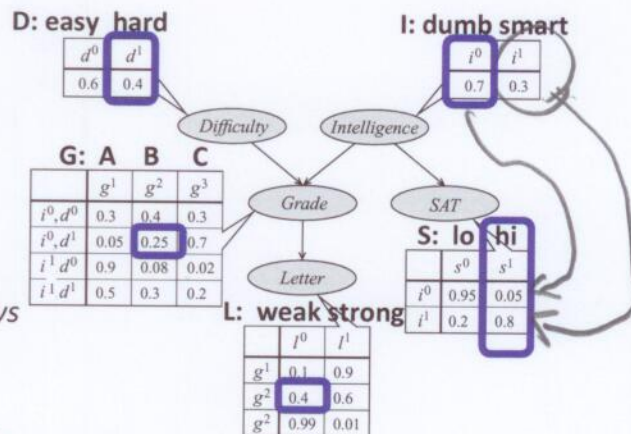
## Better Solution: Likelihood Weighting

### Wrong because...

- Evidence SAT=hi means Intell is *more likely* smart than dumb
- We will get  $P(I=Smart | SAT=hi)$  wrong, as a result of this
- In this naïve sampling we'll get  $P(I=smart | anything) = 0.3$  *always*

### To fix this: Weight samples

- Use CPD,  $P(S | I=i)$
- If sample  $I=smart$ , count this as **0.8** of a sample
- If sample  $I=dumb$ , count this as **0.05** of a sample



© Rob A. Rutenbar 2011

Slide 19

## Likelihood Weighting: Weights

### KF Algorithm 12.2

- Still a form of forward sampling, but we always get the evidence  $E=e$  right in each sample
- And, it returns not only a sample  $X[i]$  ("particle"), but also a weight  $w[i]$  for each sample
- Weight  $w[i] = \text{likelihood of evidence}$  in this particular sample
  - Product of probabilities for each  $E_i=e$  evidence variable, if we have more than one

### Lets look at algorithm

© Rob A. Rutenbar 2011



## Likelihood Weighting: KF Algorithm 12.2

### Algorithm 12.2 Likelihood-weighted particle generation

```

Procedure LW-Sample (
     $\mathcal{B}$ , // Bayesian network over  $\mathcal{X}$ 
     $Z = z$  // Event in the network
)
1  Let  $X_1, \dots, X_n$  be a topological ordering of  $\mathcal{X}$ 
2   $w \leftarrow 1$ 
3  for  $i = 1, \dots, n$ 
4     $u_i \leftarrow \mathcal{X}(\text{Pa}_{X_i})$  // Assignment to  $\text{Pa}_{X_i}$  in  $x_1, \dots, x_{i-1}$ 
5    if  $X_i \notin Z$  then
6      Sample  $x_i$  from  $P(X_i | u_i)$ 
7    else
8       $x_i \leftarrow z(X_i)$  // Assignment to  $X_i$  in  $z$ 
9       $w \leftarrow w \cdot P(x_i | u_i)$  // Multiply weight by probability of desired value
10 return  $(x_1, \dots, x_n), w$ 
    
```

same as forward sampling

get parent assignment  
if not evidence, no diff

if evidence, pick right

$E=e$  val  
but compute weight  
of sample

#### Returns:

- Samples, each with a likelihood weight
- $(\mathbf{X}[1], w[1]), (\mathbf{X}[2], w[2]), \dots (\mathbf{X}[M], w[M])$

© Rob A. Rutenbar 2011

Slide 21

## Likelihood Weighting: Use of Results

- (Sample, weight) =  $(\mathbf{X}[1], w[1]), (\mathbf{X}[2], w[2]), \dots (\mathbf{X}[M], w[M])$

$$\hat{P}(X = x | E = e) = \frac{\sum_{m=1}^M w[m] \cdot I[X[m] = x]}{\sum_{m=1}^M w[m]}$$

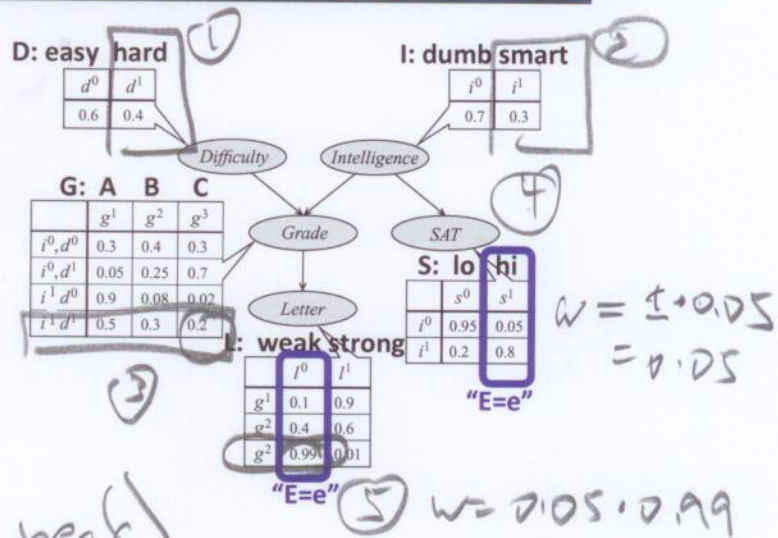
weights • A sample  
where  $X[m] = x$

add up weights

## Back to Our BN Ex

```

w=1
foreach (Xi in topo order)
  if (Xi is evidence var) {
    set sample xi = ei
    w = w * P(Xi | PaXi)
  }
  else
    sample xi val from CPD P(Xi | PaXi)
return sample x=(x1, ... xn), and weight w
    
```



→ (hard, smart, C, hi, weak)  
 weight = (0.05 \* 0.99)

## This is a Special Case of: Importance Sampling

### Very big idea in random sampling methods

- Worth talking about in general
- Know this:

$$E_P[f(X)] = \sum_{X=x} f(x)P(x) \approx \frac{1}{M} \sum_{m=1}^M f(x[m])$$

Samples from P(X)

### New assumptions

- It's really *hard* to sample  $x[m]$  from  $P(X)$
- ...but we can find a  $Q(X)$  prob dist "similar" to  $P(X)$ , from which it is very *easy* to sample  $x[m]$  values
- ...and, for any sampled  $x[m]$ , *easy* to calculate  $P(X=x[m])$



## Important Sampling: Basic Derivation

### Want:

$$E_P[f(X)] = E_Q[\text{some new function of } f(x)] \approx \frac{1}{M} \sum_{m=1}^M \text{same new function of } f(x[i])$$

Going to sample these from "easy"  $Q(X)$

$$E_P(f(X)) = \sum_{x=x} f(x) \cdot P(x) = \sum_{x=x} f(x) P(x) \frac{Q(x)}{Q(x)}$$

$$= \sum_{x=x} \left[ \underbrace{f(x) \frac{P(x)}{Q(x)}}_{\substack{\text{a new func} \\ \downarrow x}} \right] \cdot \underbrace{Q(x)}_{\substack{\text{in distrib} \\ Q}} \approx \frac{1}{M} \sum_{m=1}^M f(x) \frac{P(x)}{Q(x)}$$

★ sample from  $Q$ !

© Rob A. Rutenbar 2011

Slide 25

## Importance Sampling: Basic Result

$$E_P[f(X)] = E_Q \left[ f(X) \frac{P(X)}{Q(X)} \right] \approx \frac{1}{M} \sum_{m=1}^M f(x[m]) \cdot \frac{P(x[m])}{Q(x[m])}$$

★ draw samples from  $Q(x)$  distrib, not  $P(x)$

### In English

- Too hard to sample  $x[m]$  from  $P(X)$ . So **don't**.
- Sample (randomly)  $x[m]$  from  $Q(X)$  instead.
- Evaluate the sum above, and "correct" each summand  $f(x[m])$  with the  $P()/Q()$  term.
- And you get the Expectation you were looking for!

### Technical restrictions

- Need "dominance":  $Q(X) > 0$  whenever  $P(X) > 0$
- Helps a lot to have  $D(P || Q)$  **small**, ie, form of  $Q$  matters a **lot**

© Rob A. Rutenbar 2011

Slide 26

## Importance Sampling: Simple Example

Source: Course at UC Berkeley

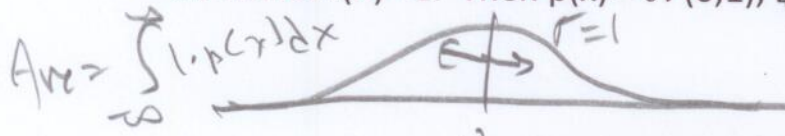
Lecture Notes for Stat 578C  
Statistical Genetics

20 October 1999

©ERIC C. ANDERSON  
(subbin' for E.A. THOMPSON)

### Monte Carlo Methods and Importance Sampling

- Suppose  $X$  is a normal RV, with distrib  $\mathcal{N}(0,1)$ 
  - Suppose we want to use random sampling to approx the area under the normal  $\mathcal{N}(0,1)$  bell curve.
  - So we let  $f(X) = 1$ . Then  $p(x) = \mathcal{N}(0,1)$ ,  $E_p[f(X)]$  is this area:



Eric C. Anderson, UC Berkeley 1999, Stat 578C

$$E_p[f(X)] = \int_{-\infty}^{\infty} 1 \cdot p(x) dx = \int_{-\infty}^{\infty} p(x) dx = 1$$

sample from  $\mathcal{N}(0,1)$

- Lets do this via importance sampling...

© Rob A. Rutenbar 2011

Slide 27

## Importance Sampling: Simple Example

- Mechanics are the same, even tho  $X$  is continuous

$$E_p[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx = \int_{-\infty}^{\infty} 1 \cdot p(x)dx \approx \int_{-50}^{50} p(x)dx \approx \frac{1}{M} \sum_{m=1}^M \frac{p(x[m])}{q(x[m])}$$

$f(x) = 1$  so this is  $\pm 50$  Samples drawn from  $Q(X)$

- Lets pick a few proposal  $Q(\cdot)$  distrib, see what happens
  - Note: this is random sampling. Every time we run this experiment, we get a different answer
  - So, we run the sampling experiment many times, and we look at the distribution of those results
  - Criterion for a "good"  $Q$ : low variance (spread) on this distr

© Rob A. Rutenbar 2011

Slide 28



## Importance Sampling: Simple Example

- Mechanics are the same, even tho X is continuous

$$E_p[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx = \int_{-\infty}^{\infty} 1 \cdot p(x)dx \approx \int_{-50}^{50} p(x)dx \approx \frac{1}{M} \sum_{m=1}^M \frac{p(x[m])}{q(x[m])}$$

Samples drawn from Q(X)

duplicate

- Lets pick a few proposal Q( ) distrib, see what happens

- Note: this is random sampling. Every time we run this experiment, we get a different answer
- So, we run the sampling experiment many times, and we look at the distribution of those results
- Criterion for a "good" Q: low variance (spread) on this distr

© Rob A. Rutenbar 2011

Slide 29

## Importance Sampling: Simple Example

easy to sample from!!

- First Q(X) is uniform

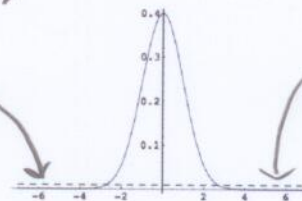
- Uniform on [-50,50]

- M=5000 samples

$$\frac{1}{M} \sum_{m=1}^M \frac{p(x[m])}{q(x[m])}$$

uniform Q(X)

Gaussian (X<sub>g</sub>)

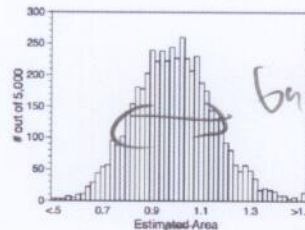


weight = 1/q

-50 0 +50

- Don't get confused: this is a lousy result

- We don't want a bell curve!
- We want integral == 1!
- Mean is right, but spread is bad
- Q uniform is a **lousy** proposal dist!



(a) Uniform

bad spread

right answer = 1!

Eric C. Anderson, UC Berkeley 1999, Stat 578C

© Rob A. Rutenbar 2011

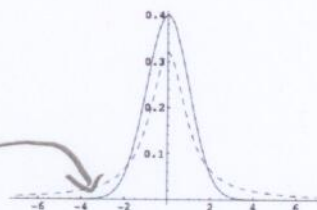
Slide 30

## Importance Sampling: Simple Example

- First  $Q(X)$  is *Standard Cauchy*

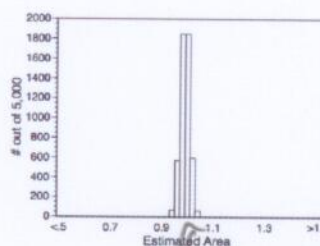
- $M=5000$  samples

$$\frac{1}{M} \sum_{m=1}^M \frac{p(x[m])}{q(x[m])} \leftarrow \underbrace{q(x) = \frac{1}{\pi(1+x^2)}}_{\text{Cauchy } Q(X)}$$



- This a much better result!

- Mean is right, but spread is now very tight, ie, if we run this experiment sampling many times, we "mostly" get  $\approx 1$ , right answer
- $Q$  Cauchy is a **good** proposal dist!



(b)  $t_1$  Dist.

Eric C. Anderson, UC Berkeley 1999, Stat 578C

© Rob A. Rutenbar 2011

Slide 31

## Importance Sampling: Problem...

$$E_p[f(X)] = E_Q \left[ f(X) \frac{P(X)}{Q(X)} \right] \approx \frac{1}{M} \sum_{m=1}^M f(x[m]) \cdot \frac{P(x[m])}{Q(x[m])} \quad \text{Oops!}$$

- ...but, what if I *can't* really calculate  $P(x[m])$ ?

- Example:

- In a BN, you have evidence. So although you can calculate joint factored  $P(X)$ , you can't easily get  $P(X|E=e)$ , which takes the place of the " $P(X)$ " in above want-to-sample-from dist
- In a MN, you don't have  $Z$ . So, although you can calculate unnormalized  $\tilde{P} = \prod \phi$ , you can't really get probability  $P(X)$ , which is again what you really need in above formula

- OK – *now* what?

© Rob A. Rutenbar 2011

Slide 32



## Importance Sampling w/o Knowing P(X)

- Assume we know an *unnormalized* form for P :  $P(X) = \frac{1}{Z} \tilde{P}(X)$ 
  - In particular, we *can* get  $\tilde{P}(X)$
- Scenarios where/why this makes sense
  - BN:
    - You can get factored form  $P() = \prod P(X_i | \mathbf{Pa}_{X_i})$
    - But you want  $P(Y|E=e) = P(Y,e)/P(e)$ . So:  $Z = P(e)$  here
  - MN
    - You can get unnormalized  $\tilde{P} = \prod \phi$
    - But you can't get real prob =  $(1/Z)\tilde{P}$ . So:  $Z$  is just partition function as usual, here

© Rob A. Rutenbar 2011

Slide 33

## Importance Sampling w/o Knowing P(X)

- As before:
  - Assume we can find distrib Q "close" to P
  - And, we can easily sample from it
- Derivation defines a new 'weight' term:
  - Weight  $w(x) = \tilde{P}(x)/Q(x)$
- New problem:  $Z == ?$

$$E_P[f(X)] = \sum_x f(x)P(x)$$

$$= \sum_x Q(x)f(x) \frac{P(x)}{Q(x)}$$

$$= \frac{1}{Z} \sum_x Q(x)f(x) \frac{\tilde{P}(x)}{Q(x)}$$

$$= \frac{1}{Z} E_Q \left[ f(x) \cdot \frac{\tilde{P}(x)}{Q(x)} \right]$$

$$= \frac{1}{Z} E_Q [f(x) \cdot w(x)]$$

$= \frac{1}{Z} \tilde{P}(x)$

define  $w(x)$

© Rob A. Rutenbar 2011

Slide 34

## Importance Sampling w/o Knowing P(X)

- Observe:  $w(x)$  is *itself* a random variable

- Recall:  $w(x) = P(x)/Q(x)$
- Just *another* function of  $X$

- What is  $E_Q[w(X)]...$ ?

$$E_{Q(X)}[w(X)] = \sum_x Q(x) \frac{\tilde{P}(x)}{Q(x)} = \sum_x \tilde{P}(x) = Z.$$

Write this...

$$\begin{aligned} E_P[f(X)] &= \sum_x f(x)P(x) \\ &= \sum_x Q(x)f(x) \frac{P(x)}{Q(x)} \\ &= \frac{1}{Z} \sum_x Q(x)f(x) \frac{\tilde{P}(x)}{Q(x)} \\ &= \frac{1}{Z} E_Q \left[ f(x) \cdot \frac{\tilde{P}(x)}{Q(x)} \right] \\ &= \frac{1}{Z} E_Q [f(x) \cdot w(x)] \end{aligned}$$

## Importance Sampling w/o Knowing P(X)

- Put it all together...

$$\begin{aligned} E_P[f(X)] &= \frac{1}{Z} E_Q[f(x) \cdot w(x)] \\ &= \frac{E_Q[f(x) \cdot w(x)]}{E_Q[w(x)]} \\ &= \frac{\sum_{m=1}^M f(x[m]) \cdot w(x[m])}{\sum_{m=1}^M w(x[m])} \end{aligned} \quad \left. \vphantom{\sum_{m=1}^M} \right\} \text{sampled from } Q(X) \text{ distrib}$$

uniform prob dist

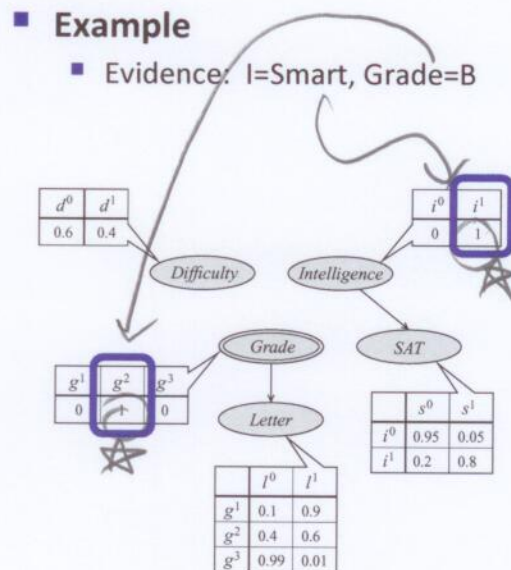
$$w(x[m]) = \frac{\tilde{P}(x[m])}{Q(x[m])}$$

easy to sample from distrib



## So, How Do We Get "Easy" Q Distrib?

- Turns out to be a 'vivid' solution for BNs
- Mutilated network**
  - Suppose we want  $P(Y|E=e)$
  - For every evidence var...
  - Remove** edges to its parents
  - Set CPD on node to **deterministically** set  $E=e$
  - Forward sample from this new "mutilated" network
  - This is the proposal  $Q()$  dist



© Rob A. Rutenbar 2011

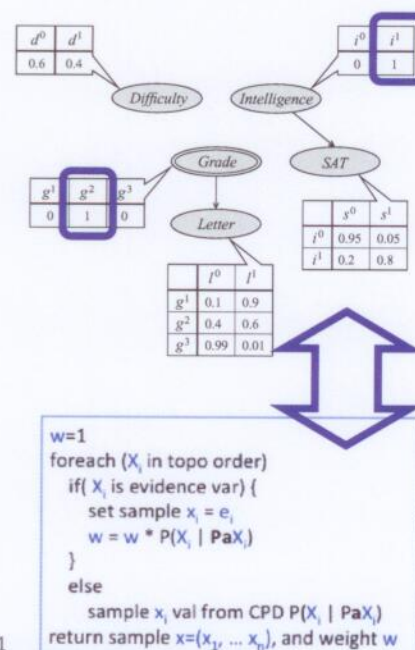
Slide 37

## Result: Mutilated Prop Q == Likelihood Weighting

- Nice result
  - Mutilate network with evidence
  - Sample from new network ( $=Q$ )
  - Use formula to get  $P(Y=y|E=e)$  ie,  $f(X) = 1(Y=y)$

$$E_P[f(X)] = \frac{\sum_{m=1}^M f(x[m]) \cdot w(x[m])}{\sum_{m=1}^M w(x[m])}$$

- This is **exactly same** as likelihood weighting!



© Rob A. Rutenbar 2011

Slide 38

## ...Unfortunate, Confusing Names in KF, tho...

### ■ Unnormalized importance sampling

- We know  $P(X)$  (ie, normalized)
- We sample using  $Q(\cdot)$  distrib

$$E_P[f(X)] = \sum_{m=1}^M f(x[m]) \cdot \frac{P(x[m])}{Q(x[m])}$$

### ■ Normalized importance sampling

- We don't know  $P(X)$
- We have to use unnorm  $P(X)$
- We sample using  $Q(\cdot)$  distrib

$$E_P[f(X)] = \frac{\sum_{m=1}^M f(x[m]) \cdot w(x[m])}{\sum_{m=1}^M w(x[m])}$$

$\hat{P}(x)$   
 $w \propto \frac{\hat{P}}{Q}$

- For several reasons (KF 12.2.3.5) books says *this* is most used in practice...

## Next: Markov Chain Monte Carlo MCMC

### ■ Problems with forward sampling methods

- Work best in BNs since they have a "direction"
- Don't really work in MNs, esp graphs with loops
- Problems with evidence
  - Evidence is toward root: we see effect in descendants
  - Evidence toward leaves: have to rely on weights (importance sampling, likelihood weights) to connect effects of evidence to nondescendants. Not always easy.

### ■ **MCMC methods** are a different class of samplers

- Esp good for these problems, esp for inference on MNs



## MCMC Methods

- 2 ideas
- We will generate a (long) **sequence** of samples
  - First samples won't be very good – eg, maybe they look like the prior, not like the posterior that we seek
  - But longer we run the sequence, the more the series converges to be samples of the distribution we want
- Each new sample  $X[i+1]$  depends on prev sample  $X[i]$ 
  - This is the classical “Markov” constration, that the history on which we depend only goes back “one step”

## MCMC: About the Name...

- **Markov Chain**
  - The basic mechanism is that we seek to create a Markov Chain, which, when we “run it” will visit states – samples – with the probability distribution we seek
  - Hope is its easier to build/run the chain and have it eventually converge to  $P(\cdot)$ , than to try to get this in some direct manner from our PGM
- **Monte Carlo**
  - Broad name for a huge class of random sampling methods, that seek to do things like approx expectations, integrals, etc, using random sampling
  - Named after gambling hub in Monte Carlo, Monaco

## Aside: About Markov Chains

### Easy analogy: Probabilistic finite state machine

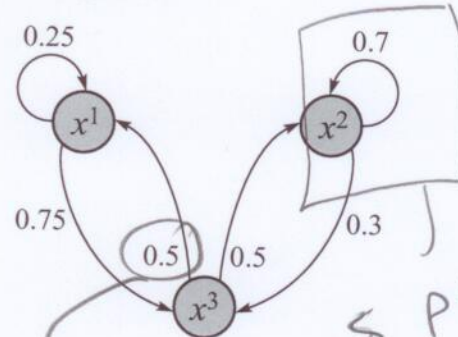
- States (like an FSM)
- Transitions (like an FSM)
- No inputs (unlike an FSM)
- Each edge has a probability

### Behavior

- At time  $t$ , chain in state  $X^{(t)}=x$
- At time  $t+1$ , chain transitions to one of its connected neighbors,  $X^{(t+1)}=x'$

- Prob of transition  $x \rightarrow x' = P[X^{(t+1)}=x' | X^{(t)}=x] = \mathcal{T}(x \rightarrow x')$

### KF Fig 12.4



$\sum_{x'} P_{rob} = 1$

$P(X^{t+1}=x' | X^t=x^3) = 0.5$

© Rob A. Rutenbar 2011

Slide 43

## Aside: About Markov Chains

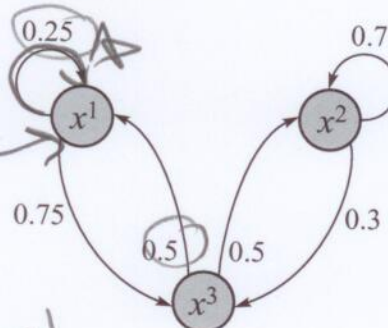
### Easy to answer some basic questions about chain

### KF Fig 12.4

### Ex: what is $P[\text{next state is } x]$ ?

- Easy to answer from diagram

$x=x'$



$P(X^{t+1}=x')$

$= P(X^{t+1}=x' | X^t=x^1) \cdot P(X^t=x^1)$

$+ P(X^{t+1}=x' | X^t=x^3) \cdot P(X^t=x^3)$

to get  $P(X^{t+1})$ , need  $P(X^t)$

© Rob A. Rutenbar 2011

Slide 44



## Aside: About Markov Chains

- Suppose we know, at current time step  $t$ , probability that we are in each of states

- Vector:

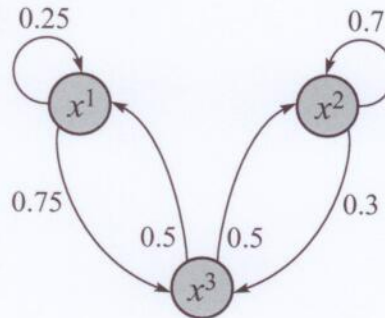
$$(\pi_1(t), \pi_2(t), \pi_3(t))$$

- Ex: suppose we start chain in state  $x^1$  at  $t=0$ . Then vector is:

$$(1, 0, 0)$$

or, Prob = 1 that at time  $t=0$ , in  $x^1$

KF Fig 12.4



## Aside: About Markov Chains

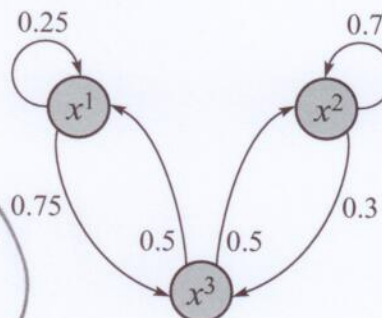
- We can write 1-step transition probability update in a nice matrix form...

$$(\pi_1^{t+1}, \pi_2^{t+1}, \pi_3^{t+1})$$

$$= (\pi_1^t, \pi_2^t, \pi_3^t) \begin{pmatrix} 0.25 & 0 & 0.5 \\ 0 & 0.7 & 0.3 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

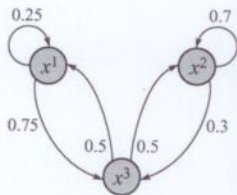
$T$

KF Fig 12.4



## Aside: About Markov Chains

- Suppose we run chain *forever*, ie,  $t \rightarrow \infty$ 
  - Does this vector of per-state probabilities **converge** to a constant distribution, ie, the "statistics" of the chain don't change anymore?



$$\pi^{t+1} = \pi^t T$$

use counts to  
estimate  $(\pi_1, \pi_2, \pi_3)$   
do these stop changing?

© Rob A. Rutenbar 2011

Slide 47

## Aside: About Markov Chains

- Surprisingly easy to solve for this: **Stationary distrib  $\pi(X)$** 
  - Just use the one-step update matrix form....

$$\pi^{t+1} = \pi^t T$$

if converges expect  $t \rightarrow \infty$

$$\pi = \pi T$$

$$\pi I = \pi T$$

$$0 = \pi (T - I)$$

aha!

© Rob A. Rutenbar 2011

Slide 48



## Aside: About Markov Chains

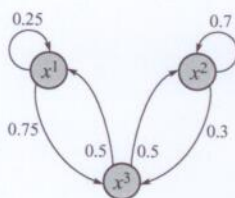
- Does **every** Markov Chain have a stationary distrib  $\pi(X)$ ?
  - Nope. (Sorry)
  - Can't even guarantee a single stationare dist; in some chains, the stationary dist you arrive at depends on the *starting* dist, ie, these are called Periodic Markov Chains
- We want chains that have one startionary distribution, arrived at from any starting distribution.
  - Such chains are said to be **Ergodic**
  - Condition: There is a non-zero probability of getting from state  $x$  to state  $x'$  in a finite number ( $k$ ) of steps, for all pairs of states in the chain

© Rob A. Rutenbar 2011

Slide 49

## From Markov Chains to MCMC

- Why do we care about this?
  - Because it turns out we can **design** Markov Chains that have stationary distributions that converge asymptotically to the complex  $P(\cdot)$  distribis inherent in our prob graphical models
  - Often the easiest way to gen these complex samples



$t=0$   $t=1$   $t=2$   $t=3$      \*\*\*      $t=10e6$   $t=10e6+1$   $t \rightarrow \infty$   
 ●   ●   ●   ●     ●   ●     \*\*\*

$PGM, P() = \frac{1}{Z} \pi \phi \approx$ 
 sampled!

© Rob A. Rutenbar 2011

Slide 50

## Famous MCMC Method: Gibbs Sampler

- Basic idea, illustrated with 4 vars

- Assume its **hard** to sample from full joint prob  $P(W, X, Y, Z)$
- Assume its **easy** to sample from conditional distributions, eg
  - Sample  $W$  from  $P(W \mid X=x, Y=y, Z=z)$
  - Sample  $X$  from  $P(X \mid W=w, Y=y, Z=z)$
  - Sample  $Y$  from  $P(Y \mid W=w, X=x, Z=z)$
  - Sample  $Z$  from  $P(Z \mid W=w, X=x, Y=y)$

vars condition on  
other values = const

- Gibbs sampler mechanics

- From a starting sample value  $(w_0, x_0, y_0, z_0)$ , repeated draw random samples using the 'script' above
- Each draw updates just 1 var, based on value of previous samples
- ie, its Markov, **next** sample depends on most **recent** sample

© Rob A. Rutenbar 2011

Slide 51

## Gibbs Sampler, More Formally

- Same example: trying to sample from  $P(W, X, Y, Z)$
- Lets write  $w_i \sim P(W \mid x, y, z)$  to mean...
  - We **sample**  $X=x_i$  from conditional distrib  $P(W \mid X=x, Y=y, Z=z)$
  - Then Gibbs sampler for this  $P()$  runs like this

1st: pick  $(w_0, x_0, y_0, z_0)$  vals

$$\left. \begin{aligned} x_1 &\sim P(X \mid w_0, y_0, z_0) \\ y_1 &\sim P(Y \mid w_0, x_1, z_0) \\ z_1 &\sim P(Z \mid w_0, x_1, y_1) \\ w_1 &\sim P(W \mid x_1, y_1, z_1) \end{aligned} \right\}$$

© Rob A. Rutenbar 2011

Slide 52

$$x_2 \sim P(X \mid w_1, y_1, z_1) \quad \text{etc}$$



## Gibbs Sampler: What It Does

- **Lovely result:**

- Gibbs sampler will converge (in the Markov Chain sense) to a stationary distribution that is  $P(W,X,Y,Z)$
    - ...that is, if you wait long enough, samples from Gibbs process, which ~~up~~ just 1 var at a time, using conditionals, will be same as sampling from full joint  $P()$ , for arbitrary  $P$
- sample*

- **Nice properties**

- Easy to do **evidence**: just restrict conditionals to set the evidence vars to the "right" values
  - Every new sampled var is (more or less) immediately a function of every of var in problem. **Unlike** forward sampling
  - Relatively **easy** to do, given factor-graph representation

## Gibbs Sampler

- **2 things to discuss**

- **Exactly how is this procedure a Markov Chain...?**

- Not entirely obvious
  - Worth going through a tiny example to show connection

- **Why is this called a "Gibbs" sampler?**

- Because, if your joint prob distrib  $P()$  is in factored, Gibbs form, the mechanics of getting the conditionals is easy(ish)

## Gibbs Sampler as a Markov Chain

### Small example:

- 2 vars, X, Y

| X | Y | P(X,Y) |
|---|---|--------|
| 0 | 0 | 0.1    |
| 0 | 1 | 0.2    |
| 1 | 0 | 0.3    |
| 1 | 1 | 0.4    |

| X | P(X) |
|---|------|
| 0 | 0.3  |
| 1 | 0.7  |

| Y | P(Y) |
|---|------|
| 0 | 0.4  |
| 1 | 0.6  |

#### Explaining the Gibbs Sampler

George Casella; Edward I. George

*The American Statistician*, Vol. 46, No. 3 (Aug., 1992), 167-174.

| X | P(X   Y=0)       | P(X   Y=1)       |
|---|------------------|------------------|
| 0 | $0.1/0.4 = 0.25$ | $0.3/0.6 = 0.5$  |
| 1 | $0.2/0.4 = 0.5$  | $0.4/0.6 = 0.67$ |

| Y | P(Y   X=0)       | P(Y   X=1)       |
|---|------------------|------------------|
| 0 | $0.1/0.3 = 0.33$ | $0.2/0.3 = 0.67$ |
| 1 | $0.2/0.3 = 0.67$ | $0.4/0.7 = 0.57$ |

conditions we use for Gibbs sampler  
 $x_{i+1} \sim P(X | y_i)$   
 $y_{i+1} \sim P(Y | x_{i+1}) \rightarrow \text{repeat}$

reality  
this!

30-40/40-60

© Rob A. Rutenbar 2011

Slide 55

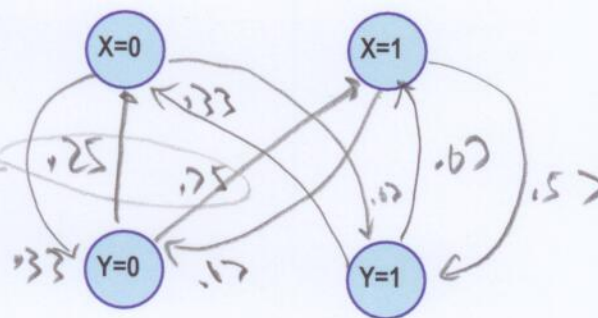
## Gibbs Sampler as a Markov Chain

### Use conditionals to draw the implicit Markov Chain

- 4 states: X=0, X=1, Y=0, Y=1

| X | P(X   Y=0) | P(X   Y=1) |
|---|------------|------------|
| 0 | 0.25       | 0.33       |
| 1 | 0.33       | 0.67       |

| Y | P(Y   X=0) | P(Y   X=1) |
|---|------------|------------|
| 0 | 0.33       | 0.43       |
| 1 | 0.67       | 0.57       |



$x_{i+1} \sim P(X | y_i)$   
 $y_{i+1} \sim P(Y | x_{i+1})$

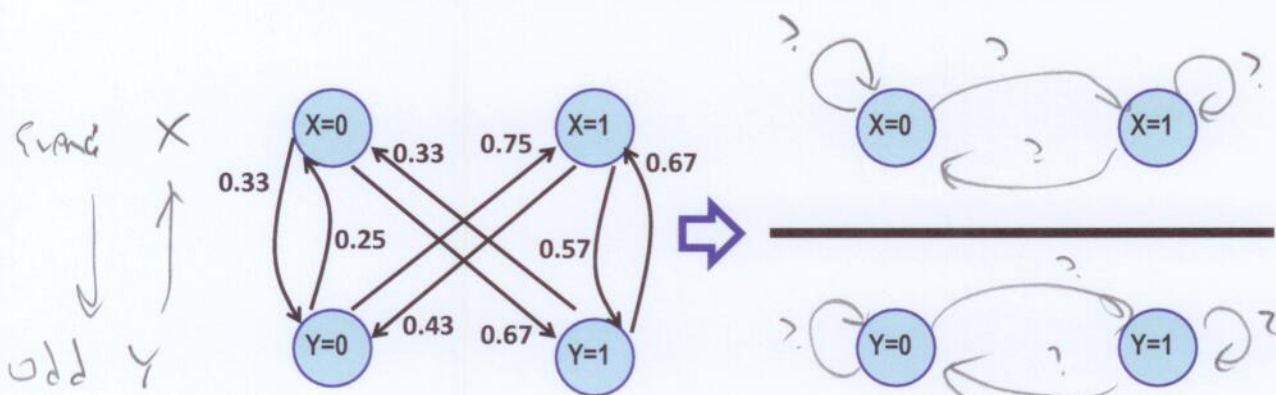
© Rob A. Rutenbar 2011

Slide 56



## Gibbs Sampler as a Markov Chain

- Interestingly, this MC does not have a stationary distrib!
- It's technically "periodic": Even cycles are only Xs, odd are Ys
- So, can we transform into a pair of 'equiv' X-only, Y-only MCs?

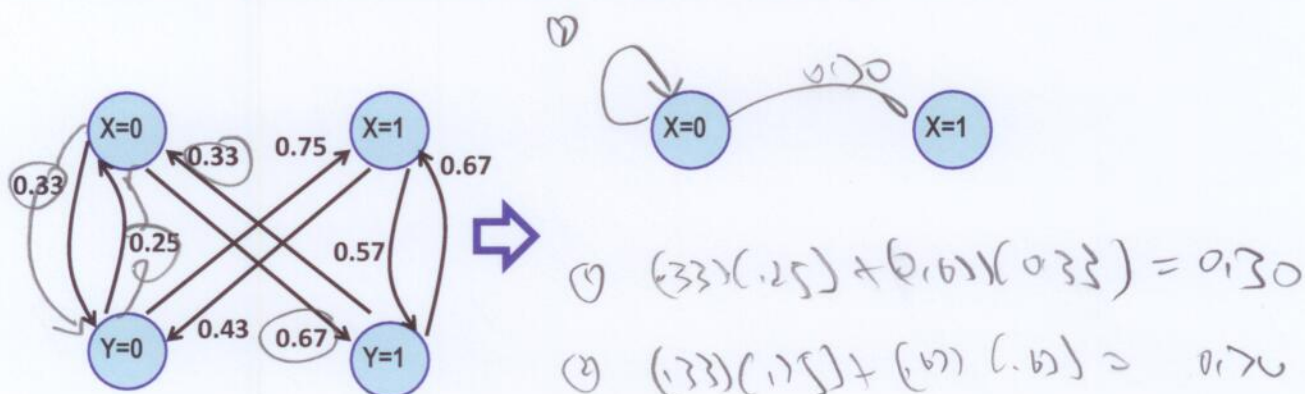


© Rob A. Rutenbar 2011

Slide 57

## Gibbs Sampler as a Markov Chain

- Look at X chain.
- For each possible edge, ask "how can we get here?"
- Look at paths  $X_i \rightarrow Y \rightarrow X_j$  Add up probs appropriately

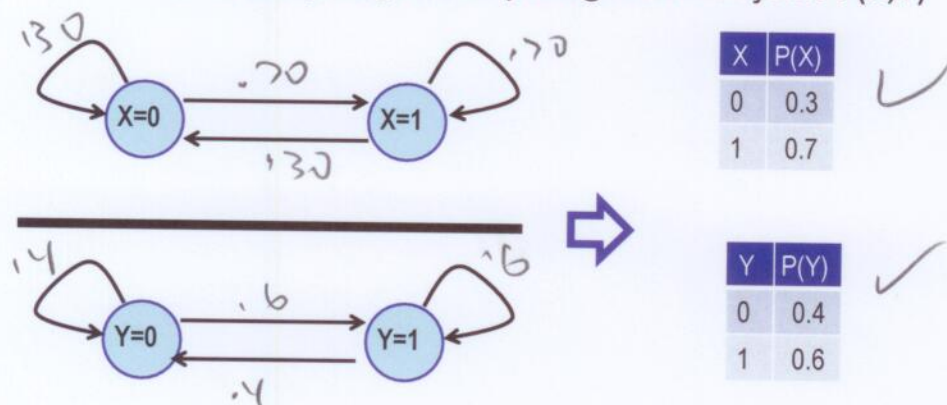


© Rob A. Rutenbar 2011

Slide 58

## Gibbs Sampler as a Markov Chain

- So, if you compute all edges, what do you get?
  - You get chains that (obviously) get **right marginals** for X and Y!
  - Turns out you get everything from full joint  $P(X,Y)$

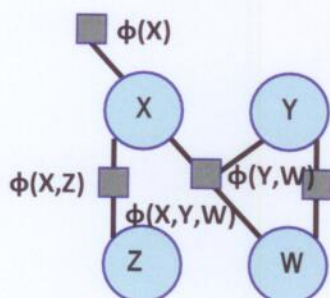


© Rob A. Rutenbar 2011

Slide 59

## So, Why Called a "Gibbs Sampler"

- Because, if you have a factored Gibbs form for joint  $P$ , all these sample-from-conditional-mechanics are easy(ish)



Factors:  $\phi(X)$ ,  $\phi(X,Z)$ ,  $\phi(X,Y,W)$ ,  $\phi(Y,W)$

How to get:  $P(Y \mid x,z,w)$ ?

$$\begin{aligned}
 & \frac{1}{Z} \phi(X) \phi(X,Z) \phi(X,Y,W) \phi(Y,W) \\
 & \sum_{Y=y} \frac{1}{Z} \phi(X) \phi(X,Z) \phi(X,Y,W) \phi(Y,W) \\
 & = \frac{\phi(X) \phi(X,Z) \phi(X,Y,W) \phi(Y,W)}{\phi(X) \phi(X,Z) \sum_{Y=y} \phi(X,Y,W) \phi(Y,W)} = \frac{\pi_{Y=y}}{\sum_{Y=y} \pi_{Y=y}}
 \end{aligned}$$

© Rob A. Rutenbar 2011

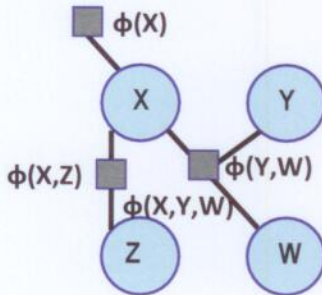
Slide 60



## So, Why Called a "Gibbs Sampler"

- NOTE: You have to calculate the distribution  $P(Y|x,z,w)$

- This is NOT a number, it's a prob distribution; you have to calculate all the value since you have to sample from this



Suppose  $\text{val}(Y) = \{a, b, c, d\}$

val

| $Y=y$ | $P(Y x,z,w)$ |
|-------|--------------|
| a     | val          |
| b     | val          |
| c     | val          |
| d     | val          |

use formula a prev slide & sum

## So, Why Called a "Gibbs Sampler"

- Gibbs sampling in general case:  $X_i \sim P(X_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$

- Works for either BN (directed) or MN (undirected) models
- Conditional probability *always* simplifies as per prev example
- For BNs: only need CPDs of  $X_i$ , and its children
- For MNs: only need factors in Markov Blanket of  $X_i$

In English: most stuff in directed for  
cancels most of time

## Beyond Gibbs Sampling

- Huge universe of advanced methods
  - Lots of problems to deal with, for example
  - Classical Gibbs sampling may be slow to converge: ie, the “mixing time” or “burn in” of the chain is long, gets to right distribution only after a very, very long time
  - Classical Gibbs sampling also doesn’t do well when distribution is very “peaky” or deterministic, or when variables are highly correlated
- Lots of tricks and methods to attack these problems
  - See KF book for some examples

© Rob A. Rutenbar 2011

Slide 63

## One Recent Example: IEEE 2008 ICIP Conf

MAP problem!

### BLIND RESTORATION OF BLURRED PHOTOGRAPHS VIA AR MODELLING AND MCMC

Tom E. Bishop<sup>a</sup>, Rafael Molina<sup>b</sup>, James R. Hopgood<sup>a</sup>

a) IDCOM, Joint Research Institute for Signal & Image Processing,  
School of Engineering & Electronics, The University of Edinburgh, Edinburgh, EH9 3JL, UK  
b) Dept. Ciencias de la Computación e I. A., Univ. de Granada, 18071 Granada, Spain.  
t.e.bishop@ed.ac.uk, rms@decsai.ugr.es, james.hopgood@ed.ac.uk

#### ABSTRACT

We propose a new image and blur prior model, based on non-stationary autoregressive (AR) models, and use these to blindly deconvolve blurred photographic images, using the Gibbs sampler. As far as we are aware, this is the first attempt to tackle a real-world blind image deconvolution (BID) problem using Markov chain Monte Carlo (MCMC) methods. We give examples with simulated and real out-of-focus images, which show the state-of-the-art results that the proposed approach provides.

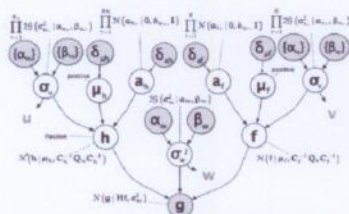


Fig. 1. Graphical model showing relationships between variables



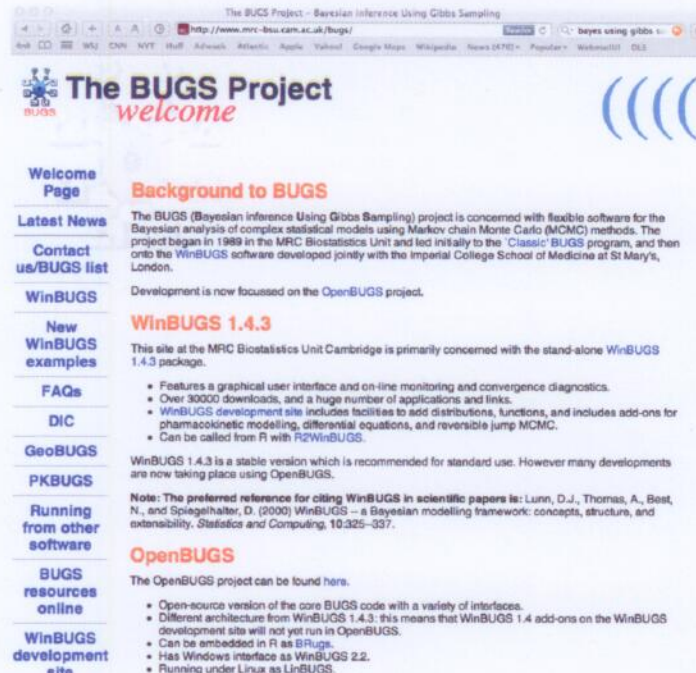
Fig. 2. Experimental results: (a) – (c) Exp. 1; (d) – (e) Exp. 2

© Rob A. Rutenbar 2011

Slide 64



## Lots of Code Around to do Gibbs Sampling



The BUGS Project - Bayesian Inference Using Gibbs Sampling  
http://www.mrc-bsu.cam.ac.uk/bugs/

**The BUGS Project**  
welcome

**Welcome Page**  
Latest News  
Contact us/BUGS list  
WinBUGS  
New WinBUGS examples  
FAQs  
DIC  
GeoBUGS  
PKBUGS  
Running from other software  
BUGS resources online  
WinBUGS development site

**Background to BUGS**  
The BUGS (Bayesian Inference Using Gibbs Sampling) project is concerned with flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. The project began in 1989 in the MRC Biostatistics Unit and led initially to the 'Classic' BUGS program, and then onto the WinBUGS software developed jointly with the Imperial College School of Medicine at St Mary's, London.  
Development is now focussed on the OpenBUGS project.

**WinBUGS 1.4.3**  
This site at the MRC Biostatistics Unit Cambridge is primarily concerned with the stand-alone WinBUGS 1.4.3 package.  
• Features a graphical user interface and on-line monitoring and convergence diagnostics.  
• Over 30000 downloads, and a huge number of applications and links.  
• WinBUGS development site includes facilities to add distributions, functions, and includes add-ons for pharmacokinetic modelling, differential equations, and reversible jump MCMC.  
• Can be called from R with R2WinBUGS.  
WinBUGS 1.4.3 is a stable version which is recommended for standard use. However many developments are now taking place using OpenBUGS.  
**Note:** The preferred reference for citing WinBUGS in scientific papers is: Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2009) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 19:325-337.

**OpenBUGS**  
The OpenBUGS project can be found here.  
• Open-source version of the core BUGS code with a variety of interfaces.  
• Different architecture from WinBUGS 1.4.3: this means that WinBUGS 1.4 add-ons on the WinBUGS development site will not yet run in OpenBUGS.  
• Can be embedded in R as BRugs.  
• Has Windows interface as WinBUGS 2.2.  
• Running under Linux as LinBUGS.

Slide 65

## Summary

- Can use **random sampling** to do approximate inference
- Two big approaches covered
  - Forward sampling (mostly for directed models like BNs)
    - Has some issues with E=e evidence
    - Can address via likelihood weighting, importance sampling
  - Gibbs sampling (works for either BNs or MNs)
    - Form of MCMC, uses sequence of samples
    - Works when you can compute  $P(x_i \mid \text{all other vars})$
    - Has some issues with convergence rate of MC sequences, highly correlated sets of variables