Probabilistic Graphical Models

Metropolis-Hastings

Learning Parameters

Summary: Gibbs Sampling

- Converts the hard problem of inference to a sequence of “easy” sampling steps
- Pros:
  - Probably the simplest Markov chain for PGMs
  - Computationally efficient to sample
- Cons:
  - Only applies if we can sample from product of factors
  - Often slow to mix, esp. when probabilities are very high
    - How can you move away from the current space?

Reversible Chains

Detailed Balance Equation:

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x)$$

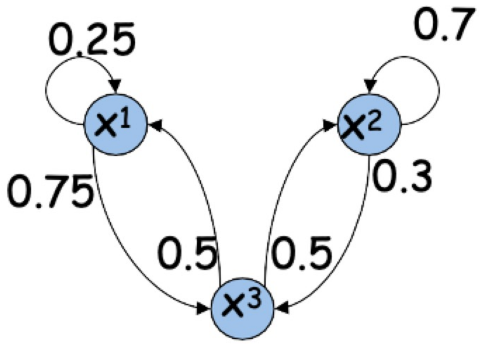
Definition: A Markov Chain is reversible if it satisfies the detailed balance equation for a unique distribution π

Reversible Chains

Detailed Balance Equation:

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x)$$

Definition: A Markov Chain is reversible if it satisfies the detailed balance equation for a unique distribution π



$$\begin{aligned} \pi(x^1) &= 0.2 \\ \pi(x^2) &= 0.5 \\ \pi(x^3) &= 0.3 \end{aligned}$$

$$\pi(x^1)T(x^1 \to x^3) = \pi(x^3)T(x^3 \to x^1)$$

Metropolis Hastings Chain

Proposal distribution $Q(x \to x')$

Acceptance probability: $A(x \to x')$

- At each state x, sample x' from $Q(x \to x')$
- Accept proposal with probability $\mathrm{A}(x \to x')$
  - If proposal accepted, move to x'
  - Otherwise stay at x

$$T(x \to x') = Q(x \to x')A(x \to x'), if\ x \neq x'$$

$$T(x \to x) = Q(x \to x) + \sum_{x \neq x'} Q(x \to x')[1 - A(x \to x')]$$

Acceptance Probability

Construct A such that detailed balance holds

$$\pi(\boldsymbol{x})T(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')T(\boldsymbol{x}' \to \boldsymbol{x})$$

$$\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')A(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})A(\boldsymbol{x}' \to \boldsymbol{x})$$

$$\frac{A(\boldsymbol{x} \to \boldsymbol{x}')}{A(\boldsymbol{x}' \to \boldsymbol{x})} = \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}$$

Acceptance Probability

Construct A such that detailed balance holds

$$\pi(\boldsymbol{x})T(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')T(\boldsymbol{x}' \to \boldsymbol{x})$$

$$\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')A(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})A(\boldsymbol{x}' \to \boldsymbol{x})$$

$A(x \to x') = \rho$
$A(x' \to x) = 1$

$$\frac{A(\boldsymbol{x} \to \boldsymbol{x}')}{A(\boldsymbol{x}' \to \boldsymbol{x})} = \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}$$

$$A(\boldsymbol{x} \to \boldsymbol{x}') = \min\left[1, \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}\right]$$

Proposal Distiribution

$$A(\boldsymbol{x} \rightarrow \boldsymbol{x}') = \min\left[1, \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \rightarrow \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')}\right]$$

- Q must be reversible:
  - $Q(x \rightarrow x') > 0 \Rightarrow Q(x' \rightarrow x) > 0$
- Opposing forces
  - Q should try to spread out, to improve mixing
  - But then acceptance probability often low

Theorem

Let Q be a proposal distribution, and *consider the Markov chain defined by equations (12.25)*

$$T(x \to x') = Q(x \to x')A(x \to x'), if\ x \neq x'$$

$$T(x \to x) = Q(x \to x) + \sum_{x \neq x'} Q(x \to x')[1 - A(x \to x')]$$

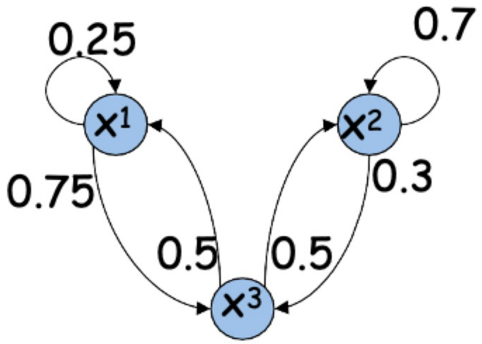
With $A(\boldsymbol{x} \to \boldsymbol{x}') = \min\left[1, \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}\right]$

If this Markov chain is regular, then it has the stationary distribution π

Example: Acceptance Probability

If $Q = T$, but you want to sample from a different stationary distribution $\pi'(x^1) = 0.6, \pi'(x^2) = 0.3, \pi'(x^3) = 0.1$

Find the Acceptance Probability



$$\begin{aligned} \pi(x^1) &= 0.2 \\ \pi(x^2) &= 0.5 \\ \pi(x^3) &= 0.3 \end{aligned}$$

$$\pi(x^1)T(x^1 \to x^2) = \pi(x^2)T(x^2 \to x^1)$$
$$\pi(x^2)T(x^2 \to x^3) = \pi(x^3)T(x^3 \to x^2)$$
$$\pi(x^3)T(x^1 \to x^3) = \pi(x^1)T(x^3 \to x^1)$$

Relationship to Gibbs Sampling

Gibbs Sampling is a special case of MH

- The GS proposal distribution is

$$Q(x_i', \mathbf{x}_{-i} \mid x_i, \mathbf{x}_{-i}) = P(x_i' \mid \mathbf{x}_{-i})$$

($\mathbf{x}_{-i}$ denotes all variables except $\mathbf{x_i}$)

- Applying Metropolis-Hastings with this proposal, we obtain:

$$A(x_i', \mathbf{x}_{-i} \mid x_i, \mathbf{x}_{-i}) = \min\left(1, \frac{P(x_i', \mathbf{x}_{-i})Q(x_i, \mathbf{x}_{-i} \mid x_i', \mathbf{x}_{-i})}{P(x_i, \mathbf{x}_{-i})Q\left(x_i', \mathbf{x}_{-i} \mid x_i, \mathbf{x}_{-i}\right)}\right)$$
$$= \min\left(1, \frac{P(x_i', \mathbf{x}_{-i})P(x_i \mid \mathbf{x}_{-i})}{P(x_i, \mathbf{x}_{-i})P\left(x_i' \mid \mathbf{x}_{-i}\right)}\right) = \min\left(1, \frac{P(x_i' \mid \mathbf{x}_{-i})P(\mathbf{x}_{-i})P(x_i \mid \mathbf{x}_{-i})}{P(x_i \mid \mathbf{x}_{-i})P(\mathbf{x}_{-i})P\left(x_i' \mid \mathbf{x}_{-i}\right)}\right)$$
$$= \min(1,1) = 1$$

GS is simply MH with a proposal that is always accepted!

Summary

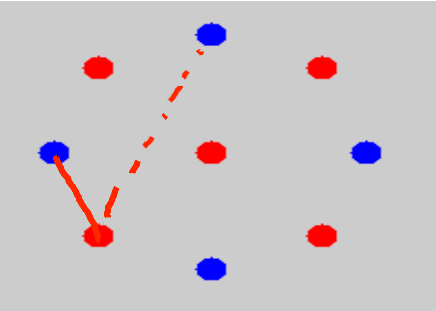
- MH is a general framework for building Markov chains with a particular stationary distribution
  - Requires a proposal distribution
  - Acceptance computed via detailed balance
- Tremendous flexibility in designing proposal distributions that explore the space quickly
  - But proposal distribution makes a big difference
  - and finding a good one is not always easy

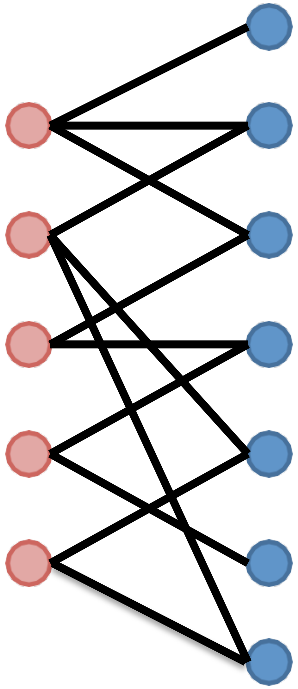
Gibbs Sampler is a special case of MH

MCMC for Matching

$X_i = j$ if i matched to j

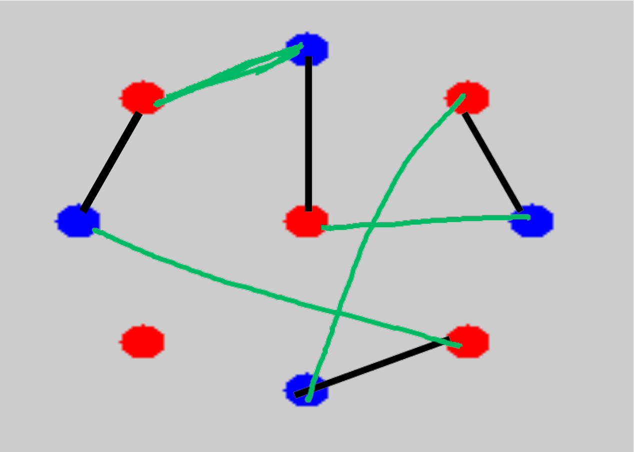
$$P(X_1 = v_1, \ldots, X_4 = v_4) \propto \begin{cases} \exp(-\sum_i \text{dist}(i, v_i)) & \text{if every } X_i \text{ has different value} \\ 0 & \text{otherwise} \end{cases}$$





Daphne Koller

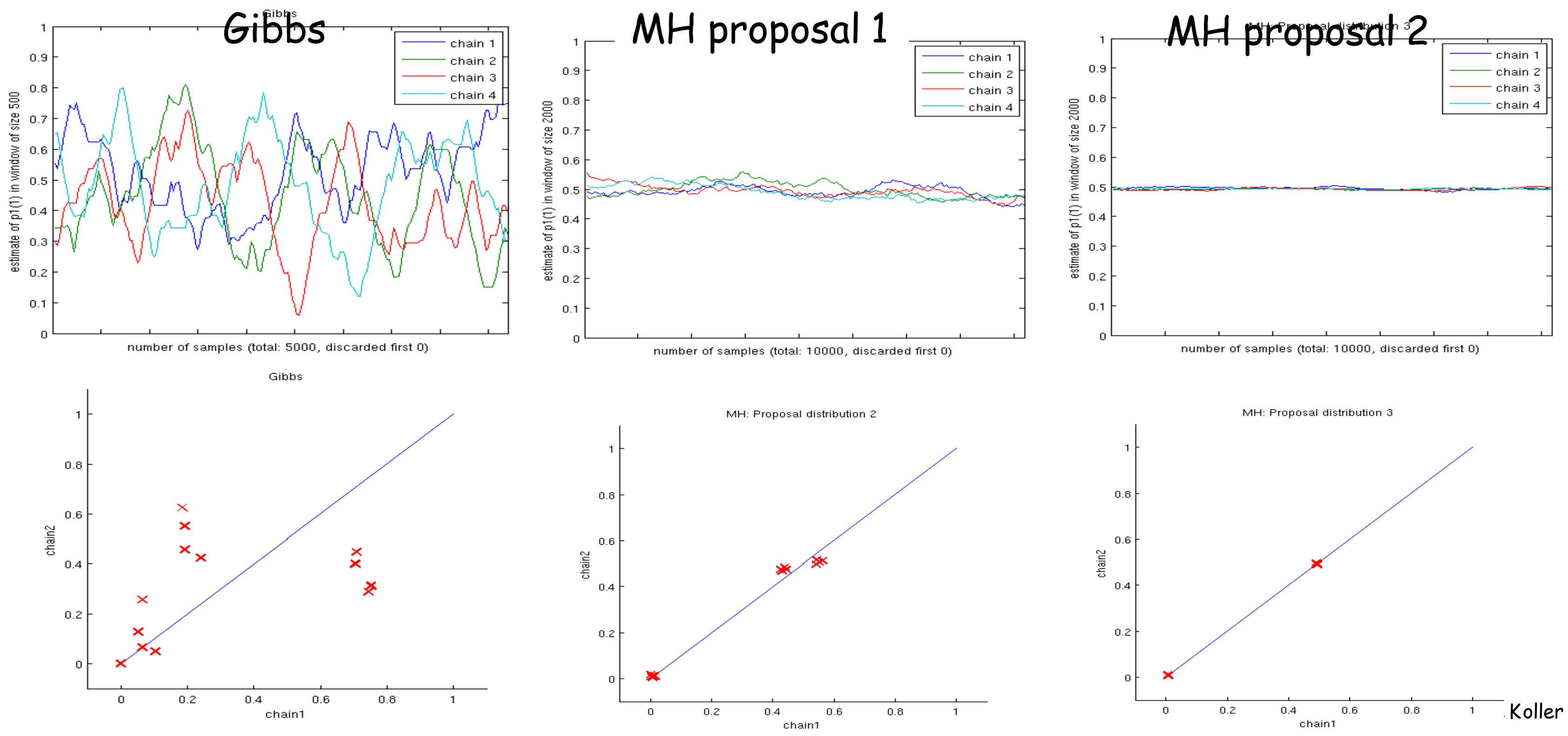
MH for Matching: AugmentingPath



1) randomly pick one variable X_i
2) sample X_i, pretending that all values are available
3) pick the variable whose assignment was taken (conflict), and return to step 2

- When step 2 creates no conflict, modify assignment to flip augmenting path

Example Results



Koller

Summary: Inference

Inference is about computing marginal and conditional distributions on a network

Exact Inference: Variable Elimination, Belief Propagation

Approximate Inference: Loopy Belief Propagation, Sampling-Based Inference (Forward Sampling, Importance Weighting, MCMC-Gibbs Sampling/MH sampling)

Approaches to learning parameters

Frequentist approach

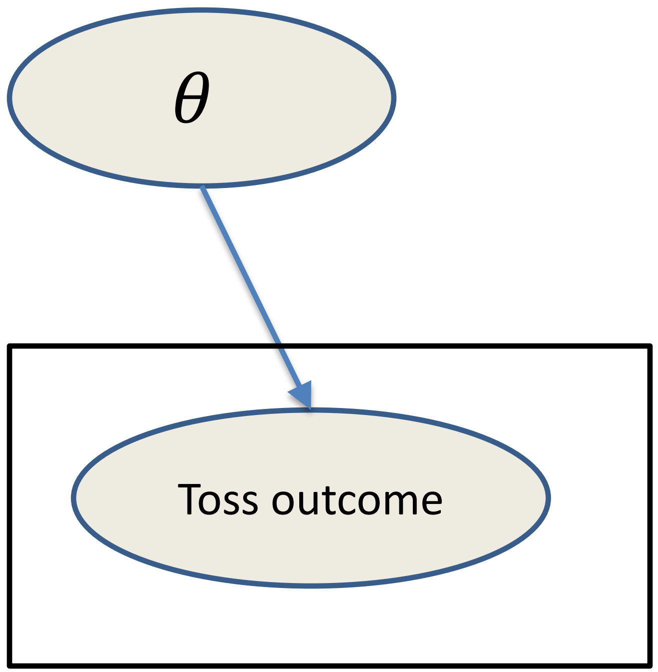
Parameters are numbers, I will try to identify the most likely number given my data.

Bayesian Approach

Parameters are numbers, but I have uncertainty about them, so I will treat them like random variables, that have distributions.

Plate Models

Plate models can represent repetition



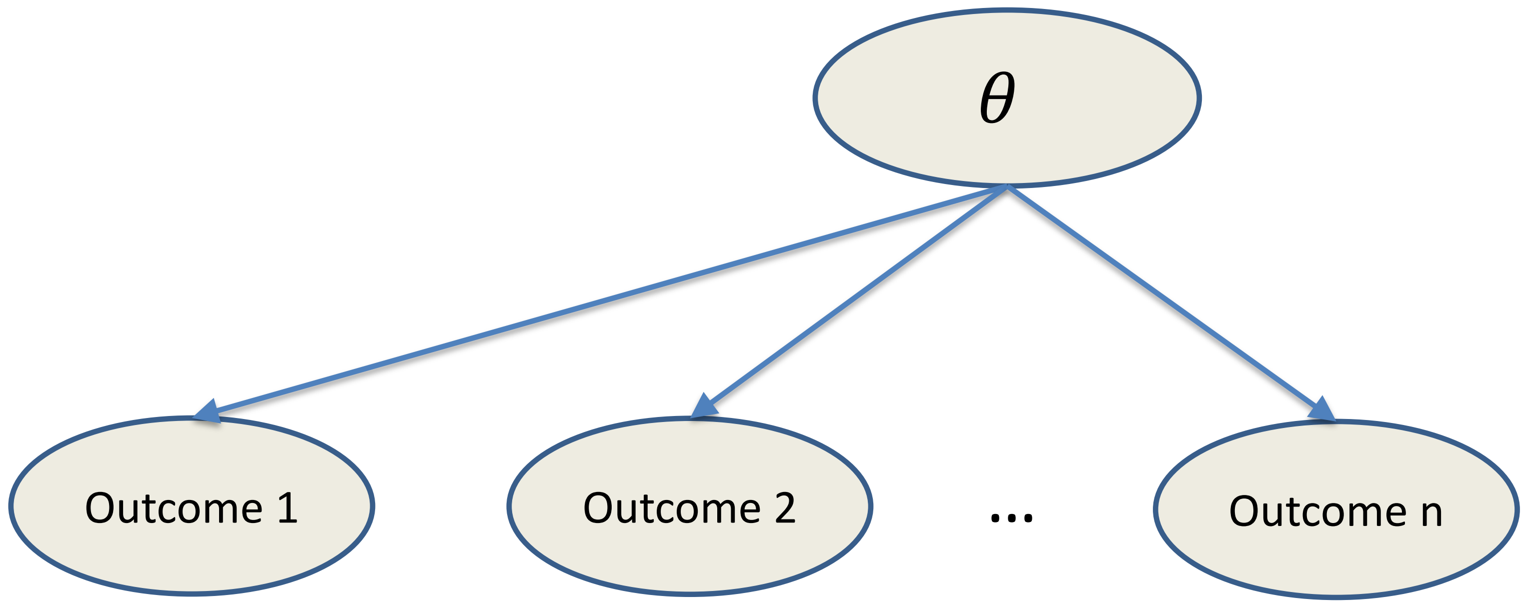
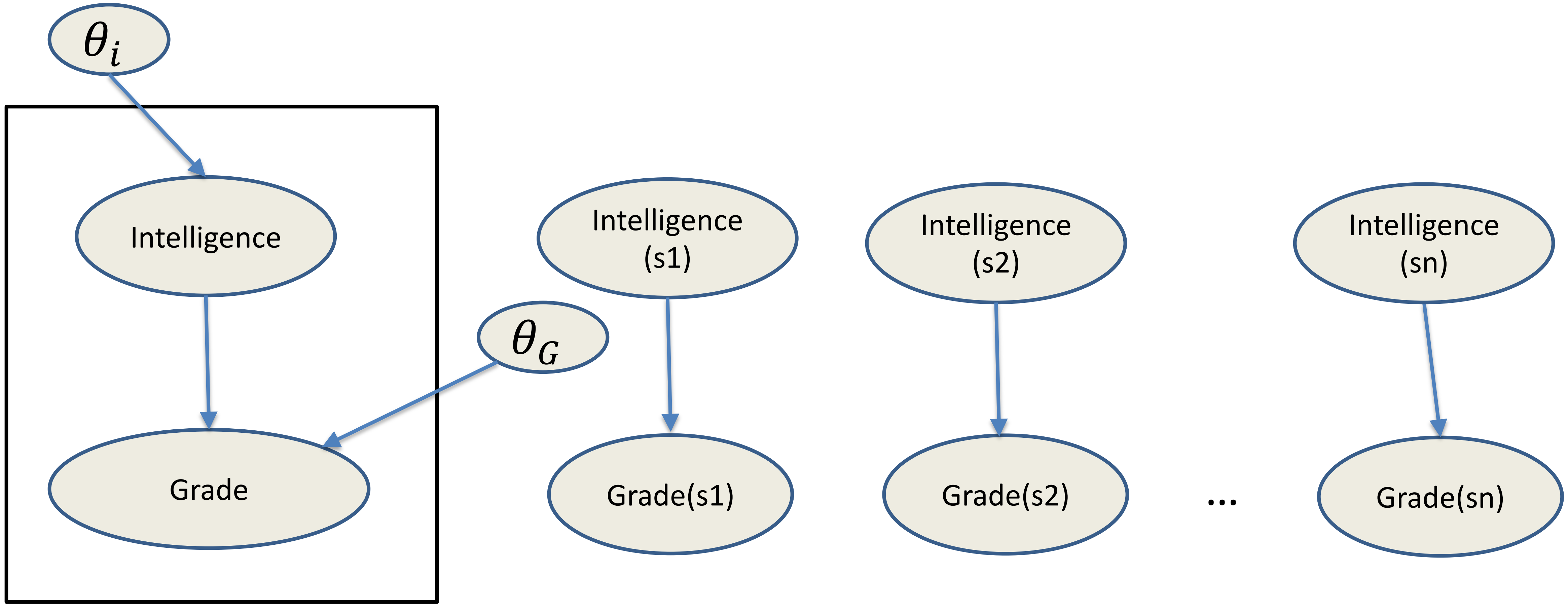


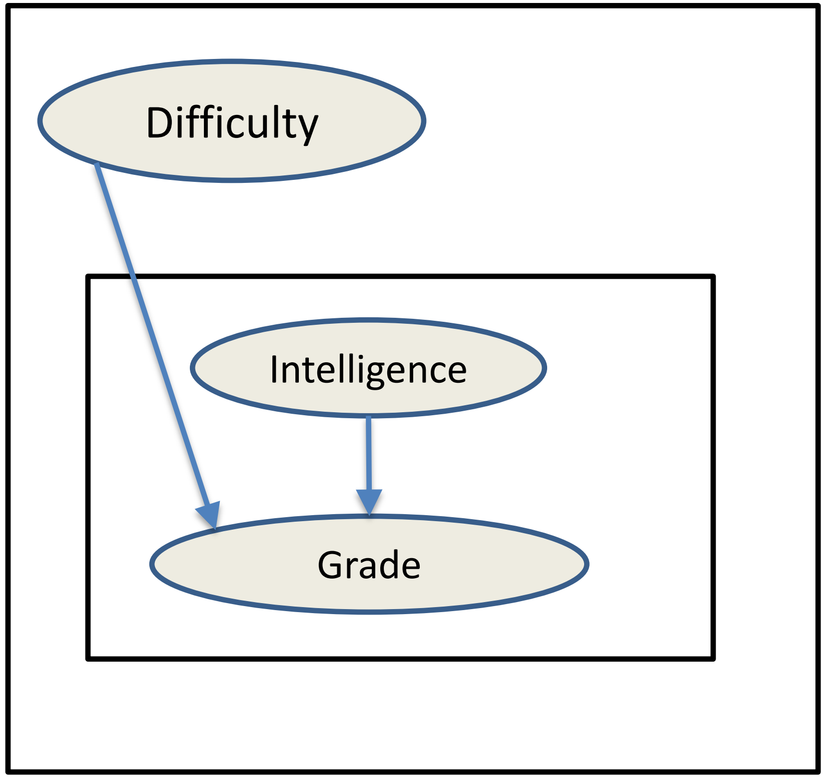
Plate Models

Plate models can represent repetition



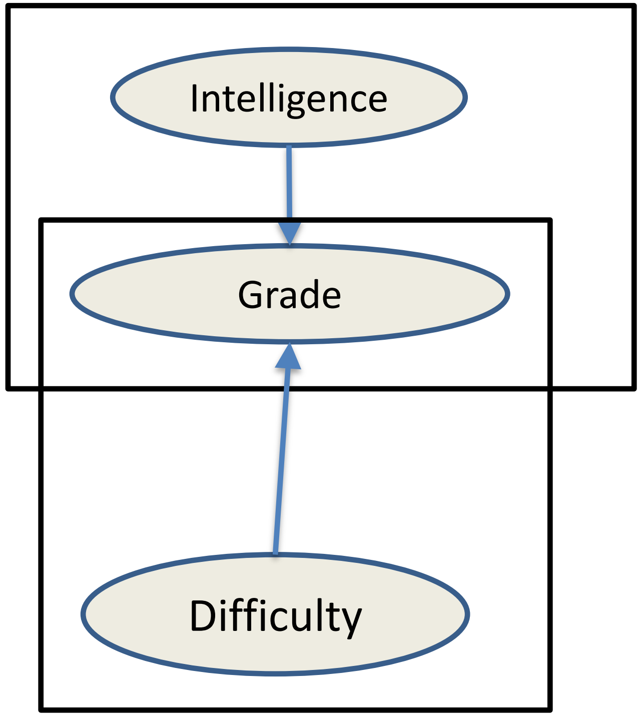
Nested Plate Models

Difficulty is a property of the course
Intelligence is a property of the course and the student

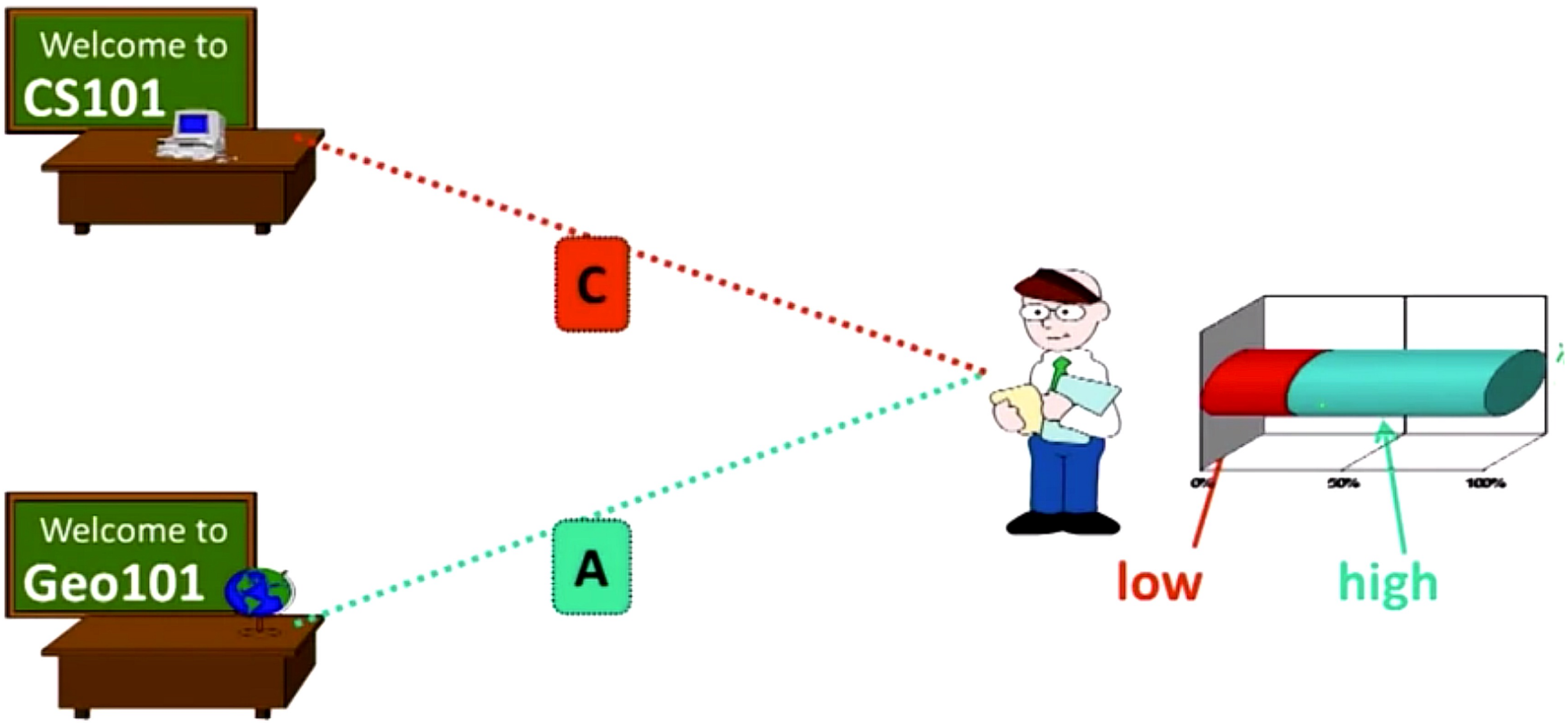


Overlapping Plate Models

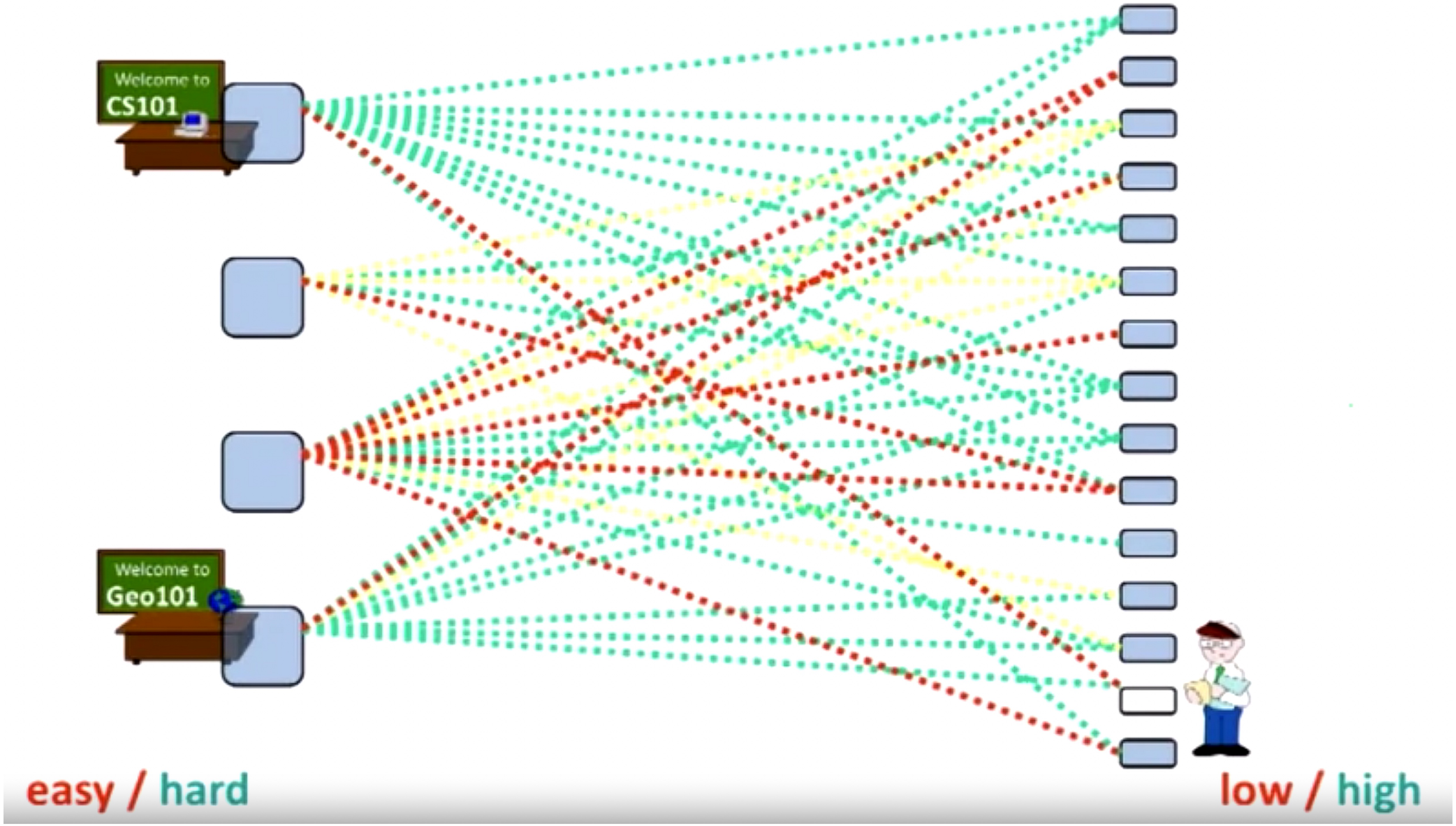
Difficulty is a property of the course
Intelligence is a property of the course and the student



Why?



Collective Inference



Formal definition

A plate model $\mathcal{M}_{\text{plate}}$ defines, for each template attribute $A \in \aleph$ with argument signature $U_1, \ldots, U_k$:

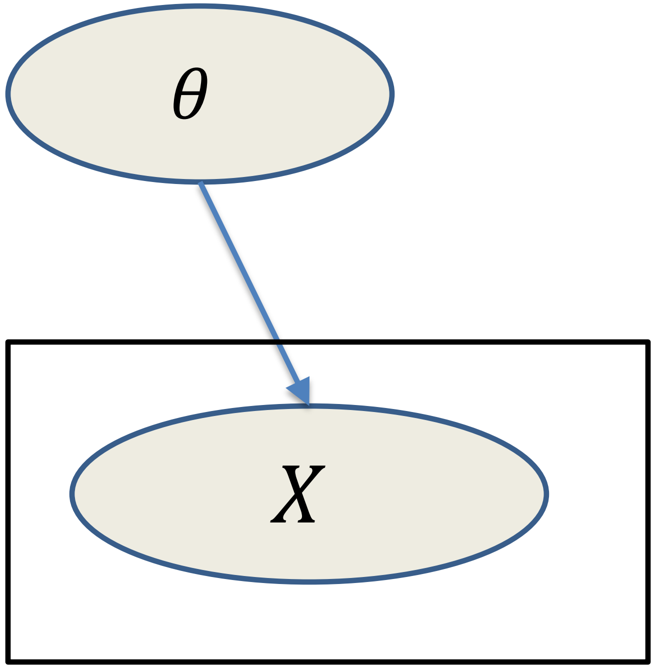
- a set of template parents

$$\mathrm{Pa}_A = \{B_1(\boldsymbol{U}_1), \ldots, B_l(\boldsymbol{U}_l)\}$$

such that for each $B_i(\boldsymbol{U}_i)$, we have that $\boldsymbol{U}_i \subseteq \{U_1, \ldots, U_k\}$. The variables $\boldsymbol{U}_i$ are the argument signature of the parent B_i.

- a template CPD $P(A \mid \mathrm{Pa}_A)$.

Back to Learning: iid as plate models



$$P(x[m] \mid \theta) = \begin{cases} \theta & x[m] = x^1 \\ 1-\theta & x[m] = x^0 \end{cases}$$

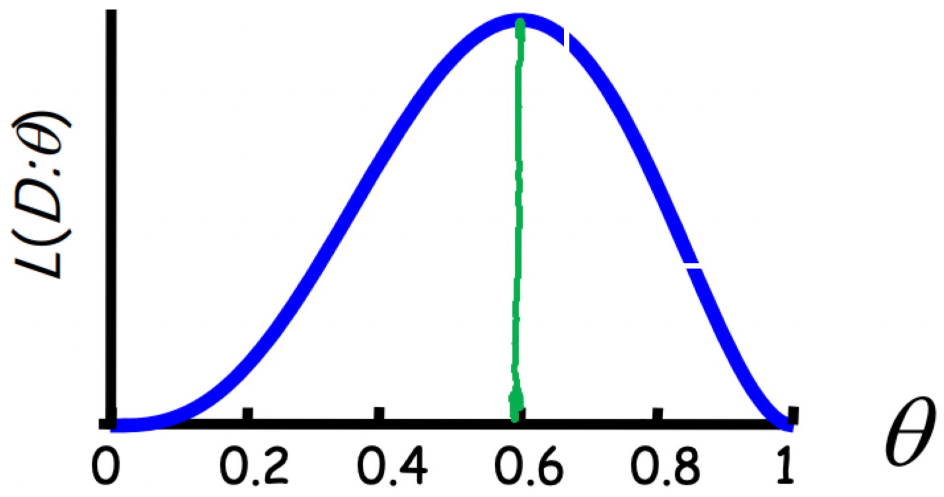
Maximum Likelihood Estimator

Find θ that maximizes the likelihood of the data

$$\sum x_i \ heads$$

$$n - \sum x_i \ tails$$

$$L(x_1, \dots, x_n; \theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$



Maximum Likelihood Estimator

- Observations: M_H heads and M_T tails
- Find θ maximizing likelihood
- Equivalent to maximizing log-likelihood

- $LL(\theta: M_H, M_T) = M_H \log\theta + M_T \log(1-\theta)$
- Differentiating the log-likelihood and solving for θ:

$$\hat{\theta} = \frac{M_H}{M_H + M_\tau}$$

Sufficient Statistics

For computing θ in the coin toss example, we only needed M_H and M_T since

$$L(\theta : D) = \theta^{M_H}(1-\theta)^{M_T}$$

M_H and M_T are sufficient statistics

A statistic $t = T(X)$ is sufficient for underlying parameter θ precisely if the conditional probability distribution of the data X, given the statistic $t = T(X)$, does not depend on the parameter θ.

$$T(D) = T(D') \Rightarrow L(\theta; D) = L(\theta; D')$$

Factorization Theorem:

T is sufficient for θ if and only if nonnegative functions g and h can be found such that

$$f_\theta(x) = h(x) g_\theta(T(x))$$

Sufficient Statistics

Multinomial distribution

For a dataset D over variable X with k values, the sufficient statistics are counts $\langle M_1, \ldots, M_k \rangle$ where M_i is the # of times that $X[m] = x^i$ in D

$$L(\theta : D) = \prod_{i=1}^{k} \theta^{M_i}$$

Gaussian distribution: $f(X) \sim N(\mu, \sigma^2)$ if $f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Rewrite as

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-x^2 \frac{1}{2\sigma^2} + x \frac{\mu}{-\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$$

Sufficient statistics for Gaussian: $\sum x^2, \sum x$, n

Maximum Likelihood Estimation

Maximum Likelihood Estimation

- MLE Principle: Choose θ to maximize $L(D:\Theta)$

- Multinomial MLE: $\widehat{\theta}_i = \frac{M_i}{\sum_j M_j}$

- Gaussian MLE:

$$\hat{\mu} = \frac{1}{M}\sum_m x[m]$$

$$\hat{\sigma} = \sqrt{\frac{1}{M}\sum_m (x[m] - \hat{\mu})^2}$$

Maximum Likelihood Estimation: Summary

- Maximum likelihood estimation is a simple principle for parameter selection given *D*

- Likelihood function uniquely determined by sufficient statistics that summarize *D*

- MLE has closed form solution for many parametric distributions

MLE for Bayes Nets

Parameters

$$\theta_{x^0}, \theta_{x^1},$$

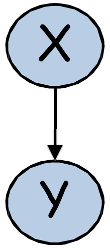
$$\theta_{y^0|x^0}, \theta_{y^1|x^0}, \theta_{y^0|x^1}, \theta_{y^0|x^1}$$

Data

$$(x^1, y^1), \ldots (x^m, y^m)$$

X	
x^0	x^1
0.7	0.3

$\theta_{x^0}, \theta_{x^1}$



$\theta_{y^0|x^0}, \theta_{y^1|x^0}, \theta_{y^0|x^1}, \theta_{y^0|x^1}$

	Y	
X	y^0	y^1
x^0	0.95	0.05
x^1	0.2	0.8

MLE for Bayesian Networks

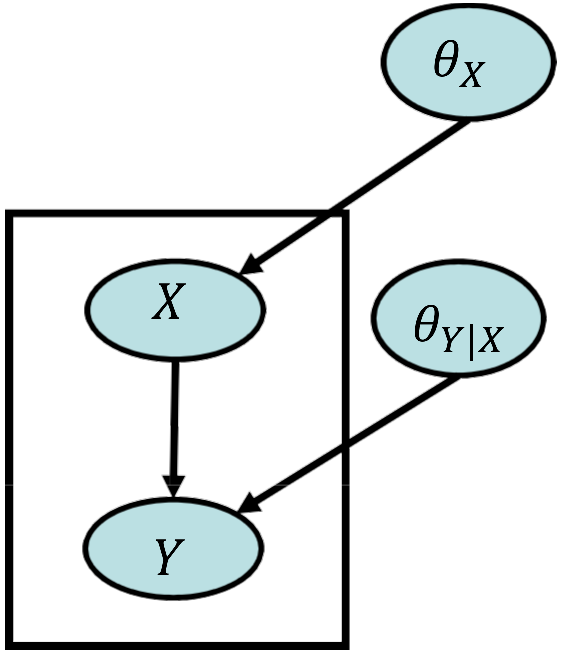
$\{\theta_x : x \in \mathrm{Val}(X)\}$

$\{\theta_{y|x} : x \in \mathrm{Val}(X), y \in \mathrm{Val}(Y)\}$

$$L(\Theta; D) = \prod_{m=1}^{M} P(x[m], y[m] : \theta)$$

$$= \prod_{m=1}^{M} P(x[m] : \theta) P(y[m] \mid x[m] : \theta)$$

$$\prod_{m=1}^{M} P(x[m] : \theta) \prod_{m=1}^{M} P(y[m] \mid x[m] : \theta)$$



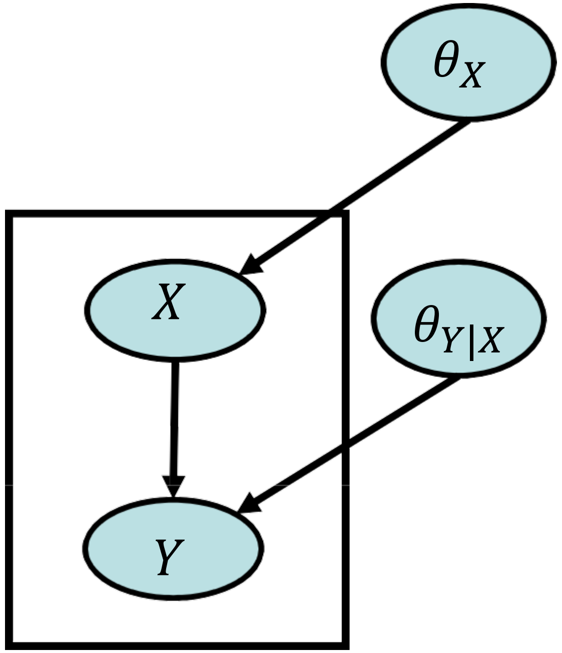
MLE for Bayesian Networks

$$L(\Theta : D) = \prod_{m} P(x[m] : \Theta)$$

$$= \prod_{m}^{m} P(x_i[m] \mid \boldsymbol{U}_i[m] : \Theta_i)$$

$$= \prod_{i} \prod_{m} P(x_i[m] \mid \boldsymbol{U}_i[m] : \Theta_i)$$

$$= \prod_{i} L_i(D : \Theta_i)$$



if $\theta_{X_i | U_i}$ are disjoint, then MLE can be computed by maximizing each local likelihood separately

For table CPDs, further decomposition

MLE limitations

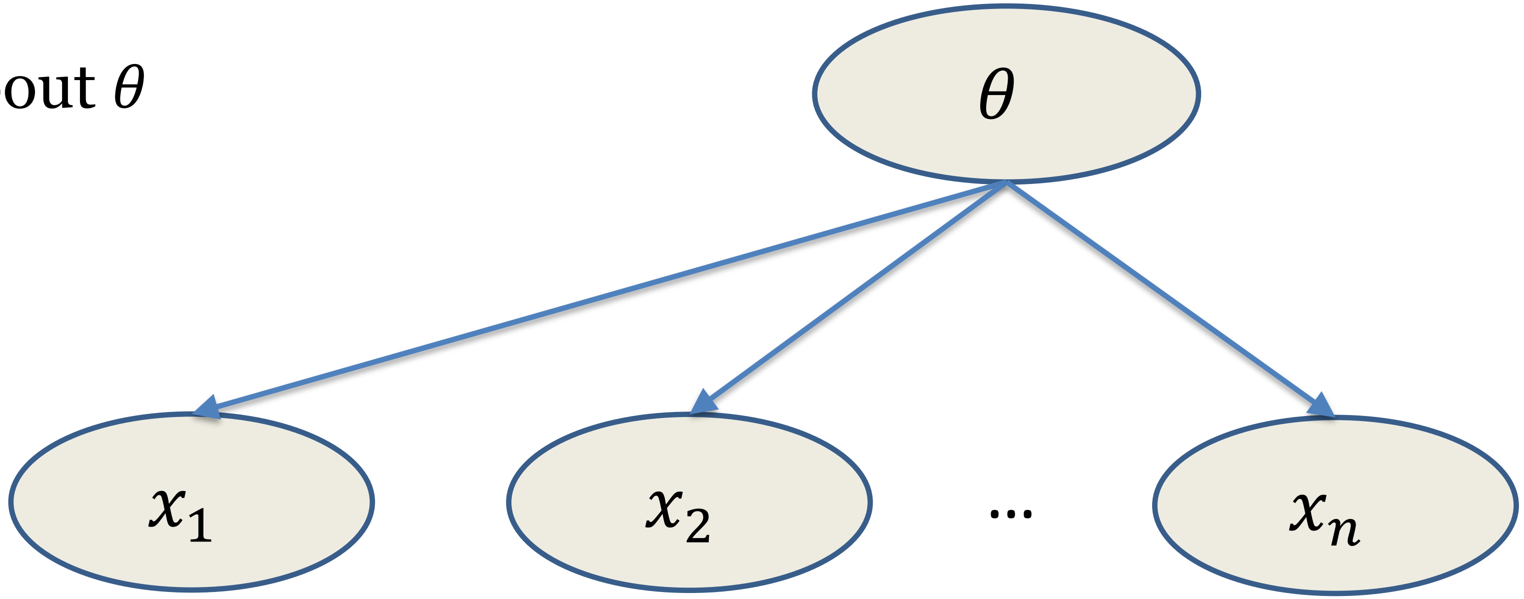
- Two teams play 10 times, and the first wins 7 of the 10 matches
  $\Rightarrow$ Probability of first team winning = 0.7
- A coin is tossed 10 times, and comes out 'heads' 7 of the 10 tosses
  $\Rightarrow$ Probability of heads = 0.7
- A coin is tossed 10000 times, and comes out 'heads' 7000 of the 10000 tosses
  $\Rightarrow$ Probability of heads = 0.7
- Before the first game, you cannot have an opinion on which team will win

Bayesian Inference

- Given a fixed θ, tosses are independent

- If θ is unknown, tosses are not marginally independent

each toss tells us something about θ



Bayesian Inference

- Given a fixed θ, tosses are independent

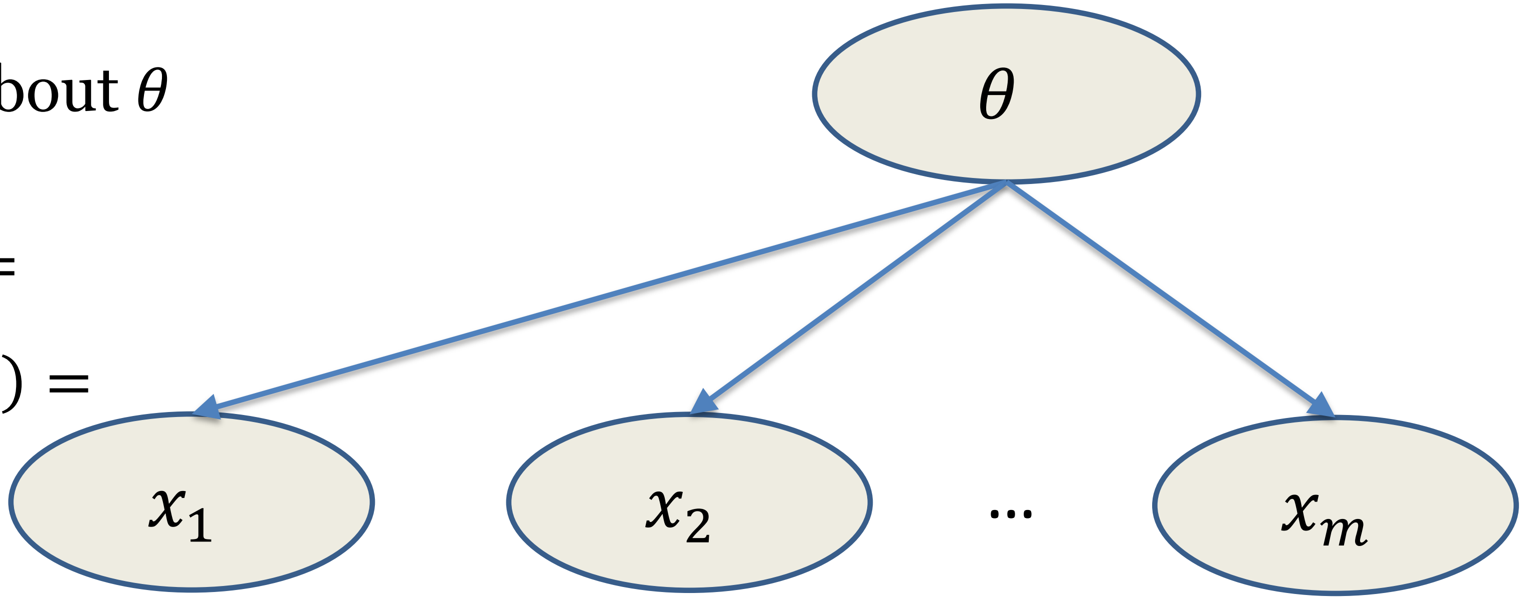
- If θ is unknown, tosses are not marginally independent

each toss tells us something about θ

$$P(\mathrm{x}[1], \dots, x[m], \theta) =$$

$$\mathrm{P}(\mathrm{x}[1], \dots, x[m], |\theta) P(\theta) =$$

$$P(\theta) \prod_{i}^{m} P(x[i]|\theta)$$

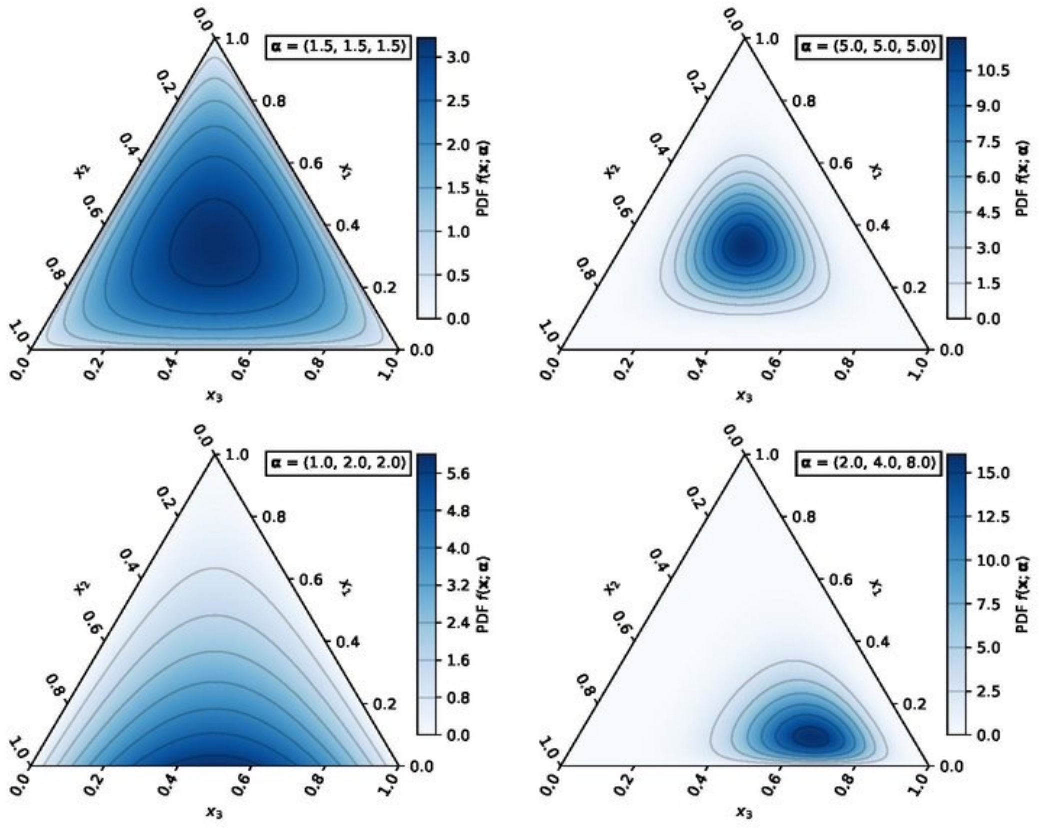


Bayesian Inference for Multinomial

Dirichlet distribution

$$f(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, a_k) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{a_i - 1}, \theta_i \in [0,1] \\ 0, \quad otherwise \end{cases}$$

where $B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}, \alpha_0 = \sum_{i=1}^{K} \alpha_i$



Bayesian Inference for Multinomial

$$P(D \mid \theta) = \prod_{i=1}^{k} \theta_i^{M_i}$$

$$P(\theta) \propto \prod_{i=1}^{k} \theta_i^{a_i}$$

$$P(D|\theta)P(\theta) \propto \prod_{i=1}^{k} \theta_i^{a_i + M_i}$$

Update only uses sufficient statistics

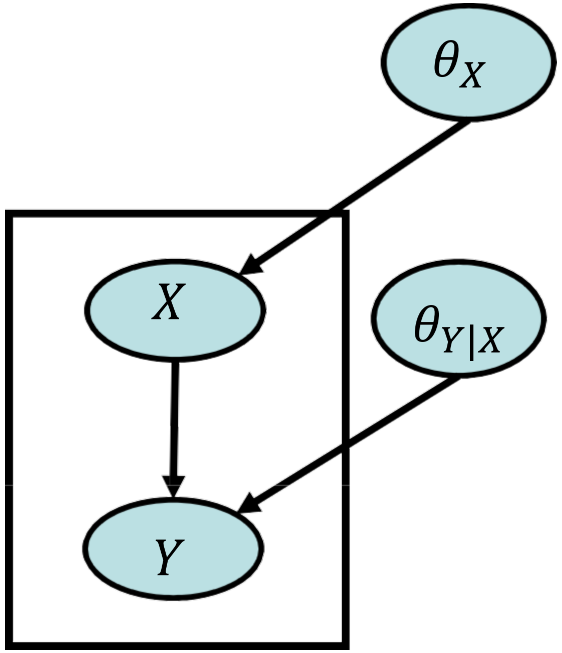
Bayesian Estimation for BNs

- Instances are independent given the parameters - (X[m'],Y[m']) are d-separated from (X [m], Y[m]) given θ

- Parameters for individual variables are independent a priori $P(\theta) = \prod P\left(\theta_{X_i} | P_a(X_i)\right)$

- Posteriors for θ are also independent given the data:
- $P\left(\theta_x, \theta_{Y|X} \middle| D\right) = P\left(\theta_x | D\right) P(\theta_{Y|X} \middle| D)$

  As in MLE, we can solve each estimation problem separately



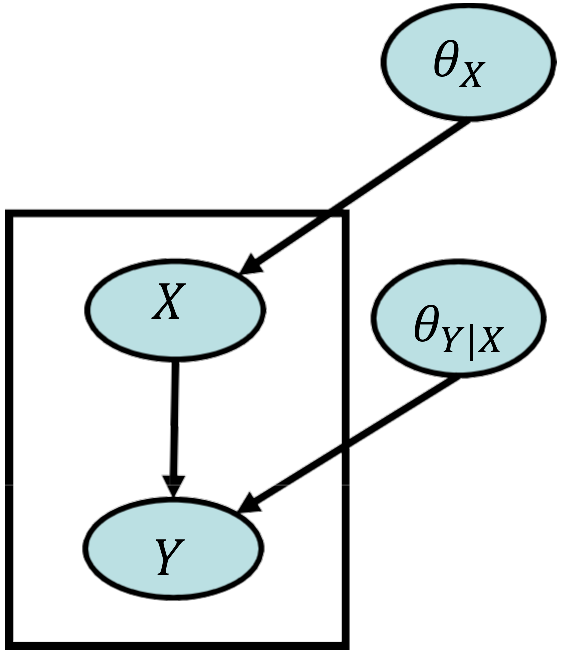
Bayesian Estimation for BNs

- Instances are independent given the parameters - (X[m'],Y[m']) are d-separated from (X [m], Y[m]) given θ

- Parameters for individual variables are independent a priori $P(\theta) = \prod P\left(\theta_{X_i} | P_a(X_i)\right)$

- Posteriors for θ are also independent given the data:
- $P\left(\theta_x, \theta_{Y|X} \middle| D\right) = P\left(\theta_x \middle| D\right) P(\theta_{Y|X} \middle| D)$

  As in MLE, we can solve each estimation problem separately



- Posteriors of θ can be computed independently

– For multinomial $\theta_{X|\boldsymbol{u}}$ if prior is Dirichlet($a_{x^1|\boldsymbol{u}}, \ldots, a_{x^k|\boldsymbol{u}}$)

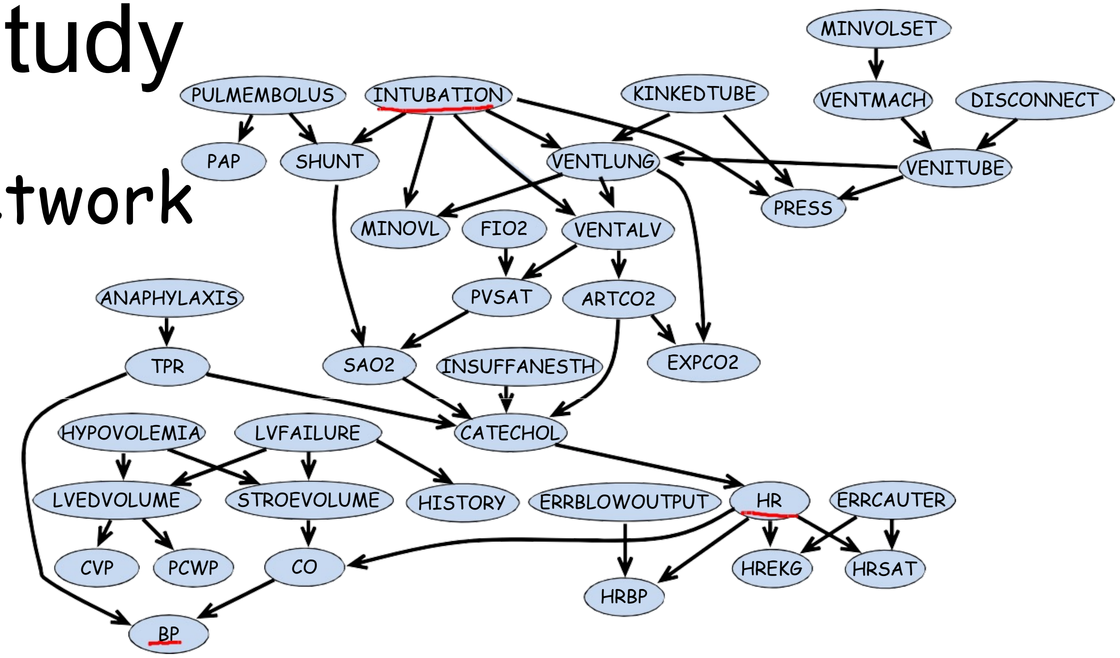
– posterior is Dirichlet($a_{x^1|\boldsymbol{u}} + M[x^1, \boldsymbol{u}], \ldots, a_{x^k|\boldsymbol{u}} + M[x^k, \boldsymbol{u}]$)

Equivalent Sample size

- We need hyperparameter $\alpha_{x|\boldsymbol{u}}$ for each node X, value x, and parent assignment $\boldsymbol{u}$
  - Prior network with parameters Θ_o
  - Equivalent sample size parameter α
  - $\alpha_{x|\boldsymbol{u}} = \alpha P(x, \boldsymbol{u}|\Theta_0)$

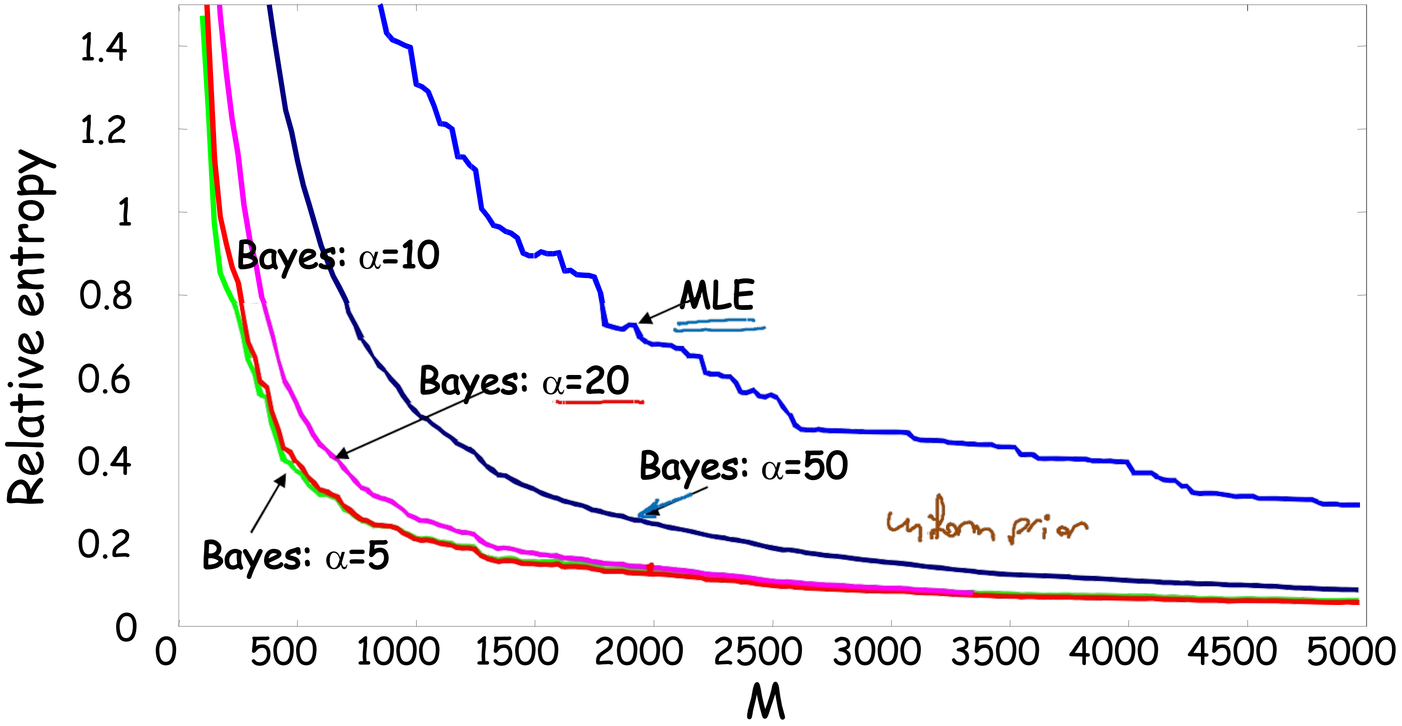
Case Study

- ICU-Alarm network
  - 37 variables
  - 504 params
- Experiment
  - Sample instances from network
  - Relearn parameters



Daphne Koller

Case Study: ICU Alarm Network



Daphne Koller

Summary

- In Bayesian networks, if parameters are independent a priori, then also independent in the posterior

- For multinomial BNs, estimation uses sufficient statistics $M[x, \boldsymbol{u}]$

$$\hat{\theta}_{x|u} = \frac{M[x, \boldsymbol{u}]}{M[\boldsymbol{u}]}$$

MLE

$$E(x|\boldsymbol{u}, D) = \frac{\alpha_{x,u} + M[x, \boldsymbol{u}]}{\alpha_u + M[\boldsymbol{u}]}$$

Bayesian (Dirichlet)

- Bayesian methods require choice of prior
  - can be elicited as prior network and equivalent sample size