ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics:

- Bayes Nets: Parameter Learning
  - MLE, MAP, Bayesian Estimation

Readings: KF 16, 17.1-17.4; Barber 9.1-9.4

Dhruv Batra

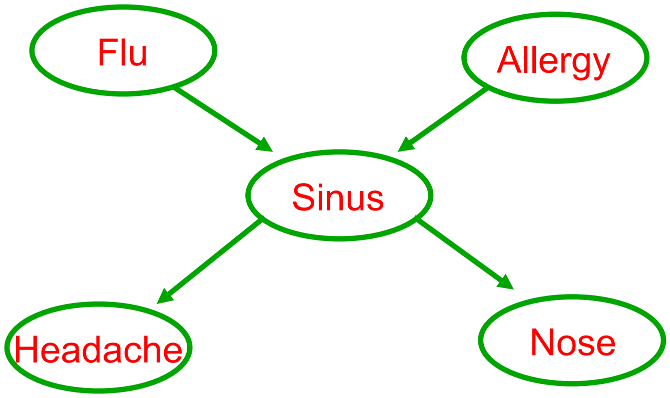
Virginia Tech

Administrativia

- HW1
  - Out soon
  - Due in 2 weeks: Feb 17, 11:59pm
  - Please please please please start early
  - Implementation: TAN, structure + parameter learning
  - Please post questions on Scholar Forum.

A general Bayes net

- Set of random variables



- Directed acyclic graph
  - Encodes independence assumptions
- CPTs
  - Conditional Probability Tables
- Joint distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Main Issues in PGMs

- Representation
  - How do we store $P(X_1, X_2, \ldots, X_n)$
  - What does my model mean/imply/assume? (Semantics)

- Learning
  - How do we learn parameters and structure of $P(X_1, X_2, \ldots, X_n)$ from data?
  - What model is the right for my data?

- Inference
  - How do I answer questions/queries with my model? such as
  - Marginal Estimation: $P(X_5 \mid X_1, X_4)$
  - Most Probable Explanation: argmax $P(X_1, X_2, \ldots, X_n)$

Learning Bayes Nets

True Distribution P*
(Maybe corresponds to a BN G*
maybe not)

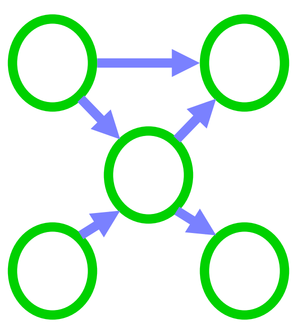
Domain Experts

Data

$\mathbf{x}^{(1)}$

…

$\mathbf{x}^{(m)}$

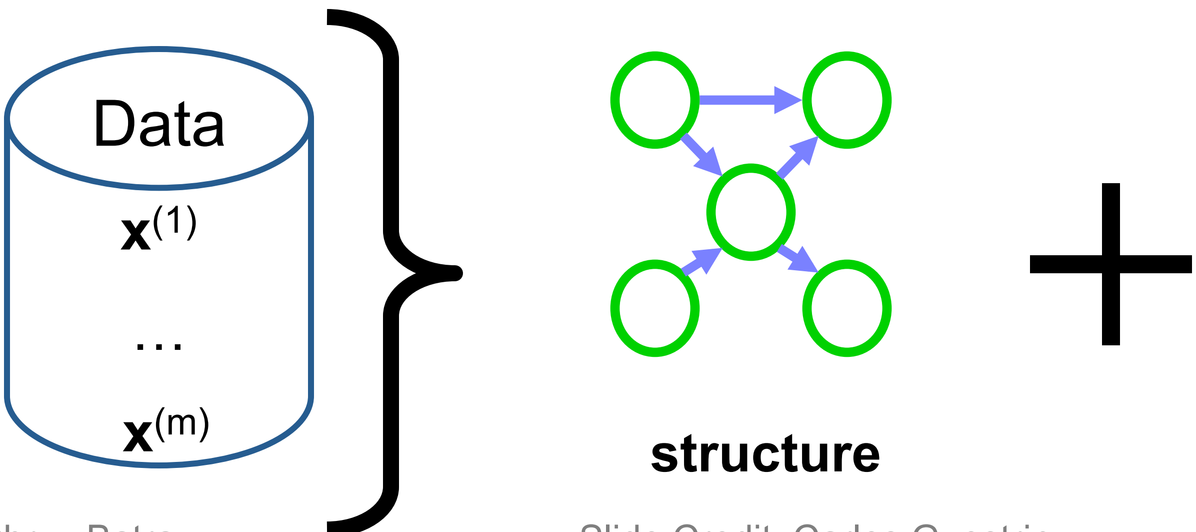


+

CPTs –
$P(X_i | \mathbf{Pa}_{X_i})$

Learning Bayes nets

	Known structure	Unknown structure
Fully observable data	Very easy	Hard
Missing data	Somewhat easy (EM)	Very very hard



CPTs – $P(X_i | \mathbf{Pa}_{X_i})$

parameters

Your first probabilistic learning algorithm

- After taking this ML class, you drop out of VT and join an illegal betting company.
- Your new boss asks you:
  - If Rafael Nadal & Stanislas Wawrinka play tomorrow, will Nadal win or lose W/L?
- You say: what happened in the past?
  - W, W, W, W, L
- You say: P(Nadal Wins) = …
- Why?

Simplest BN

- One variable X
  - On board

Maximum Likelihood Estimation

- Goal: Find a good θ

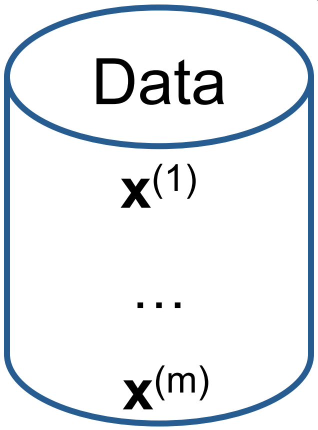
- What's a good θ?
  - One that makes it likely for us to have seen this data
  - Quality of θ = Likelihood(θ; D) = P(data | θ)

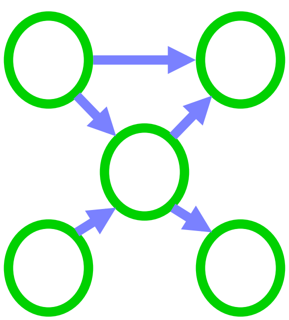
Why Max-Likelihood?

- Leads to “natural” estimators

- MLE is OPT if model-class is correct
  - Log-likelihood(θ) = entropy(P*) – KL(P*,P(D|θ))
  - Maximizing LL = minimizing KL

Learning the CPTs



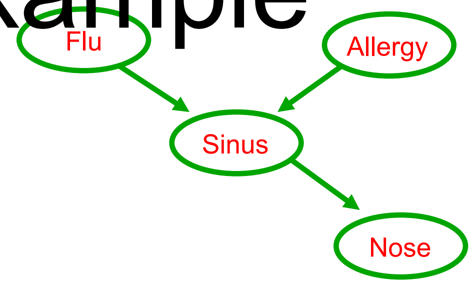


For each discrete variable X_i

$$\hat{P}_{MLE}(X_i = a \mid \mathrm{Pa}_{X_i} = b) = \frac{\mathrm{Count}(X_i = a, \mathrm{Pa}_{X_i} = b)}{\mathrm{Count}(\mathrm{Pa}_{X_i} = b)}$$

WHY??????????

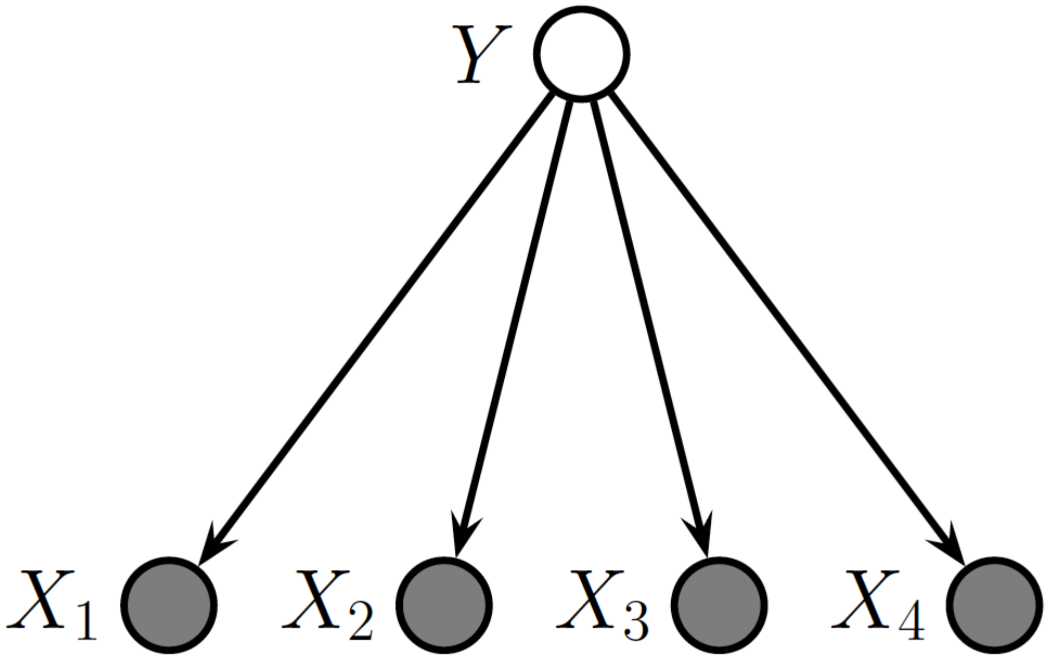
MLE of BN parameters – example



- Given structure, log likelihood of data:

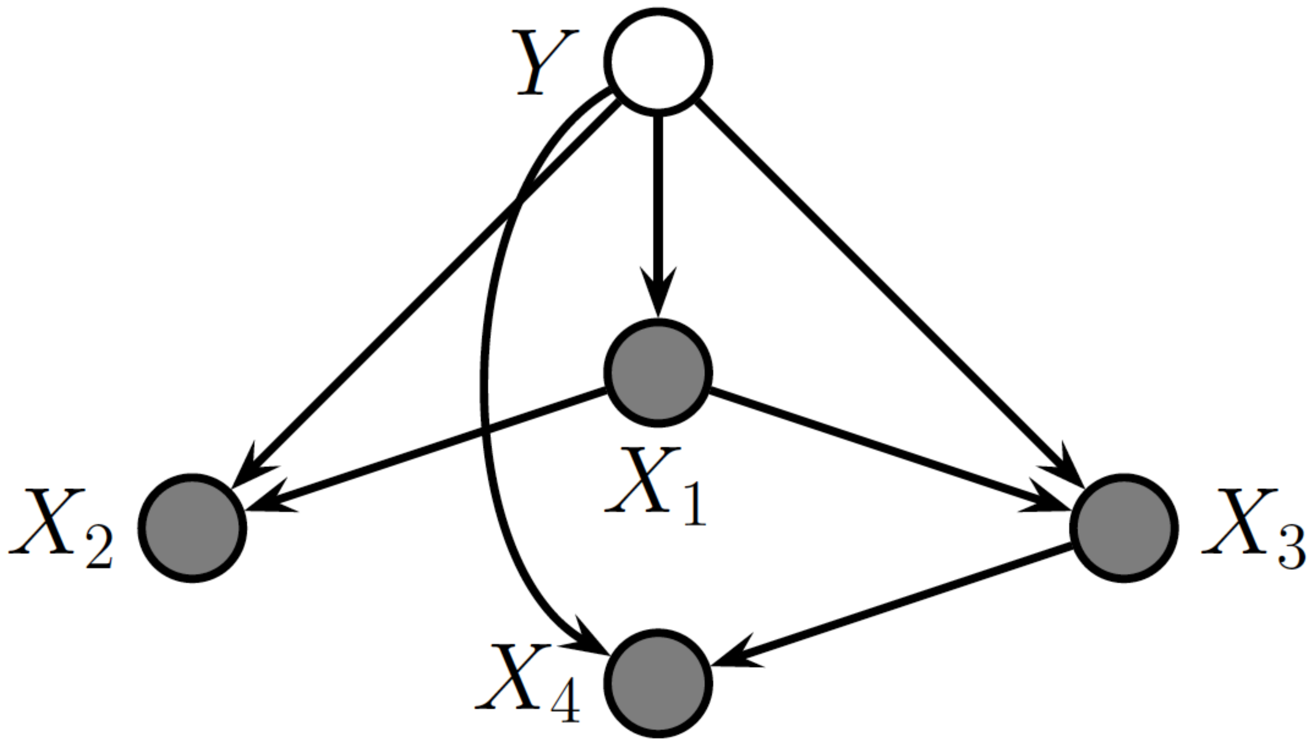
$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$

Name That Model



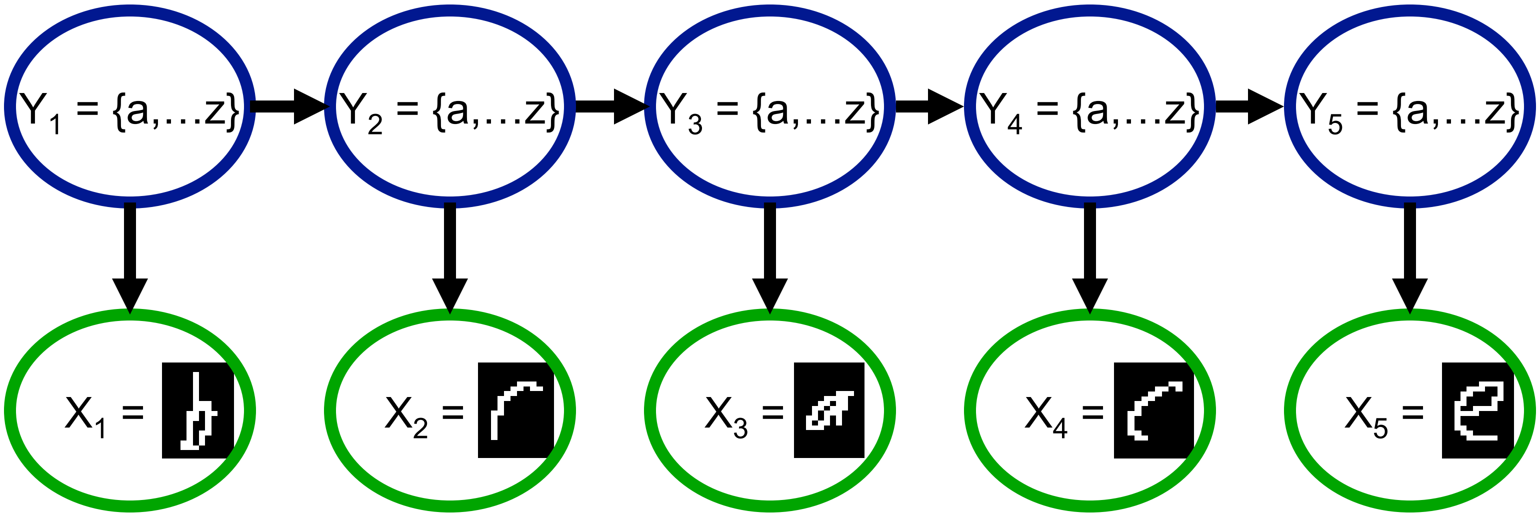
Naïve Bayes: $$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y)$$

Name That Model



Tree-Augmented Naïve Bayes (TAN)

Name That Model



Hidden Markov Model (HMM)

Figure Credit: Carlos Guestrin

How much data?

$$\hat{\theta}_{MLE} = \frac{m_H}{m_H + m_T}$$

- Last year:
  - 3 heads/wins; 2 tails/losses for Nadal.
    - You say: $\theta = 3/5$, I can prove it!
  - 30 heads/wins; 20 tails/losses for Nadal.
    - You say: Same answer, I can prove it!

Bayesian Estimation

- Boss says: What is I know Nadal is a better player on clay courts?
- You say: Bayesian it is then..

Priors

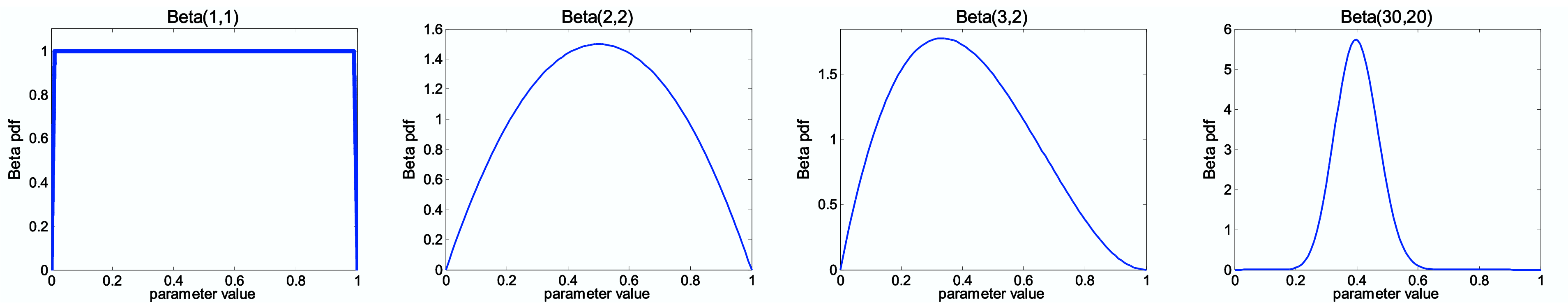
- What are priors?
  - Express beliefs before experiments are conducted
  - Computational ease: lead to “good” posteriors
  - Help deal with unseen data
  - Regularizers: bias us towards “simpler” models

- Conjugate Priors
  - Prior is conjugate to likelihood if it leads to itself as posterior
  - Closed form representation of posterior

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\alpha_H - 1}(1-\theta)^{\alpha_T - 1}}{B(\alpha_H, \alpha_T)} \sim Beta(\alpha_H, \alpha_T)$$

- Demo:
  - http://demonstrations.wolfram.com/BetaDistribution/



Slide Credit: Carlos Guestrin

Posterior

- Benefits of conjugate priors

$$P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$$

$$P(\theta) = \frac{\theta^{\alpha_H - 1}(1-\theta)^{\alpha_T - 1}}{B(\alpha_H, \alpha_T)} \sim Beta(\alpha_H, \alpha_T)$$

$$P(\theta \mid \mathcal{D}) \;\; \propto \;\; P(\mathcal{D} \mid \theta) P(\theta)$$

$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

MAP for Beta distribution

$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

- MAP: use most likely parameter:

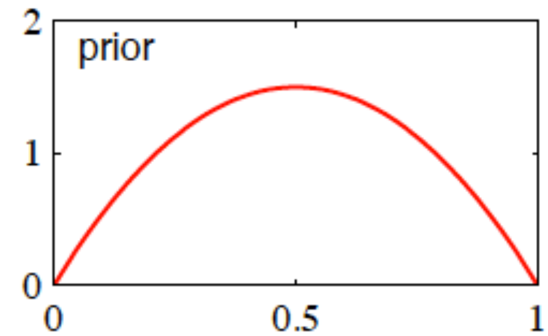
$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) =$$

- Beta prior equivalent to extra W/L matches
- As *m* → inf, prior is “forgotten”
- **But, for small sample size, prior is important!**

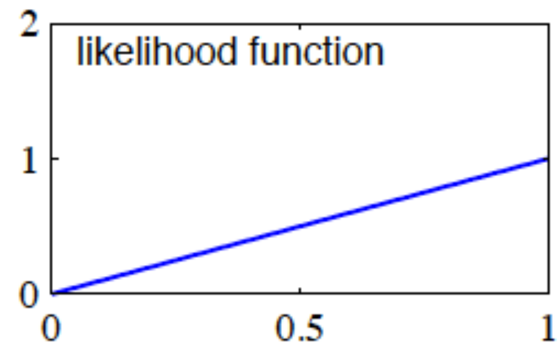
Slide Credit: Carlos Guestrin

Effect of Prior

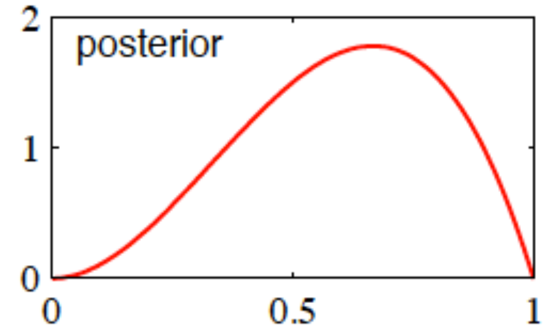
- Prior = Beta(2,2)
  - $\theta_{prior} = 0.5$



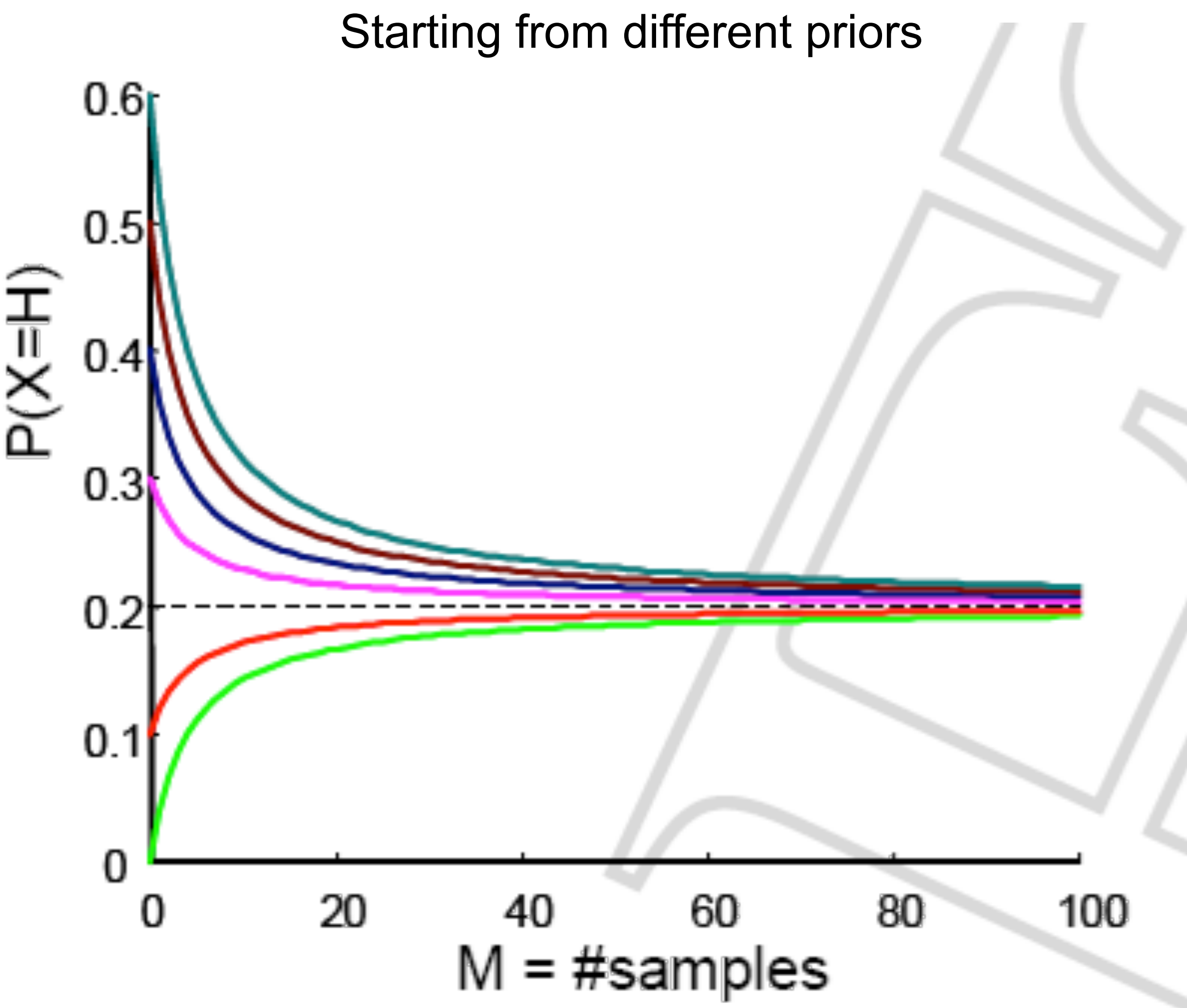
- Dataset = {H}
  - $L(\theta) = \theta$
  - $\theta_{MLE} = 1$



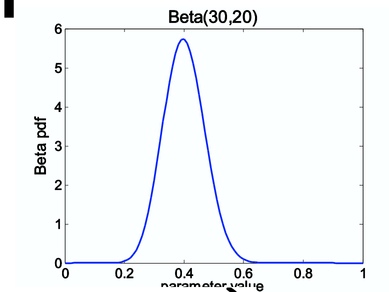
- Posterior = Beta(3,2)
  - $\theta_{MAP} = (3-1)/(3+2-2) = 2/3$



Effect of Prior



Using Bayesian posterior



- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

  - Integral is often hard to compute

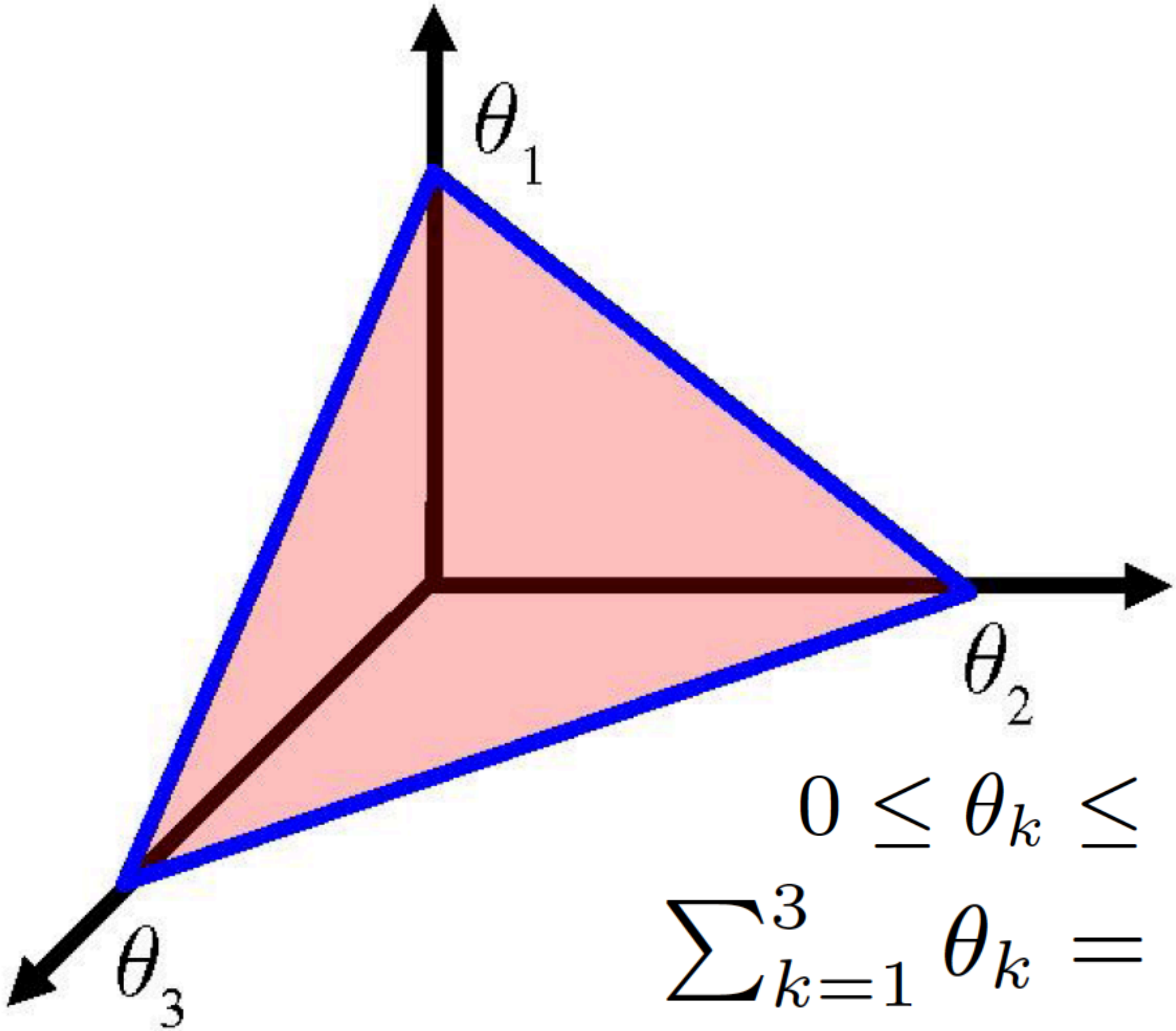
Bayesian learning for multinomial

- What if you have a k sided coin???
- Likelihood function if **categorical**:
- **Conjugate** prior for multinomial is **Dirichlet**:

$$\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$$

- **Observe** *m* data points, m_i from assignment i, **posterior**:
- **Prediction**:

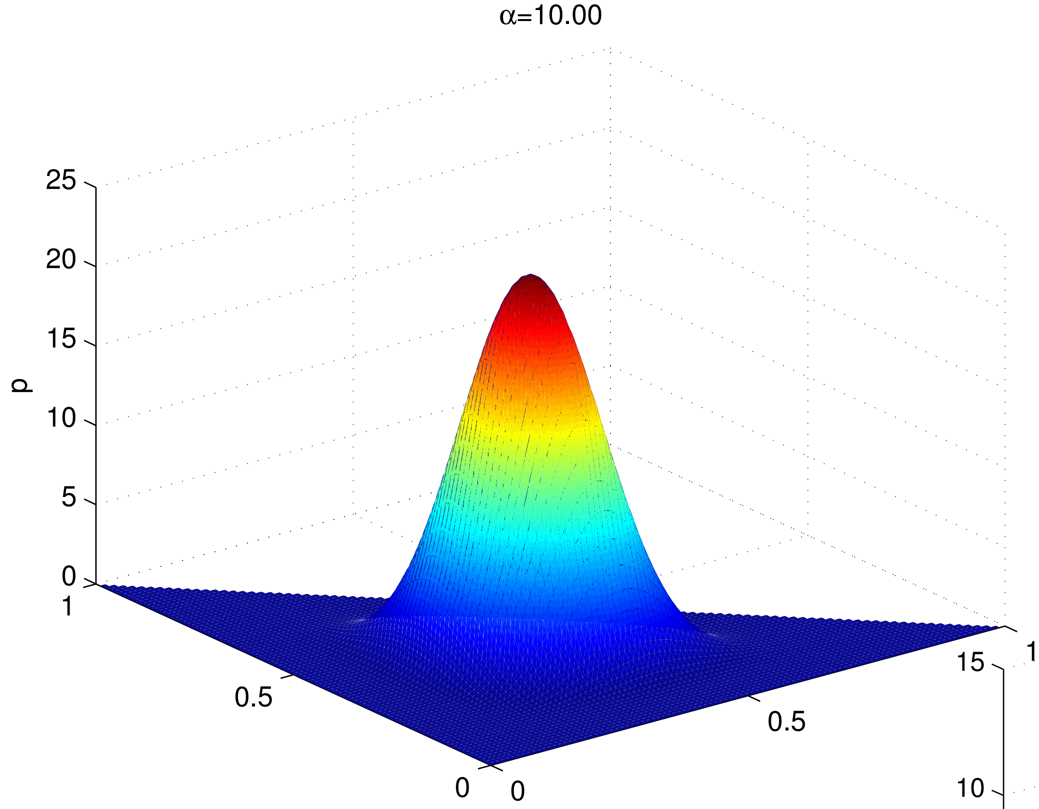
Simplex

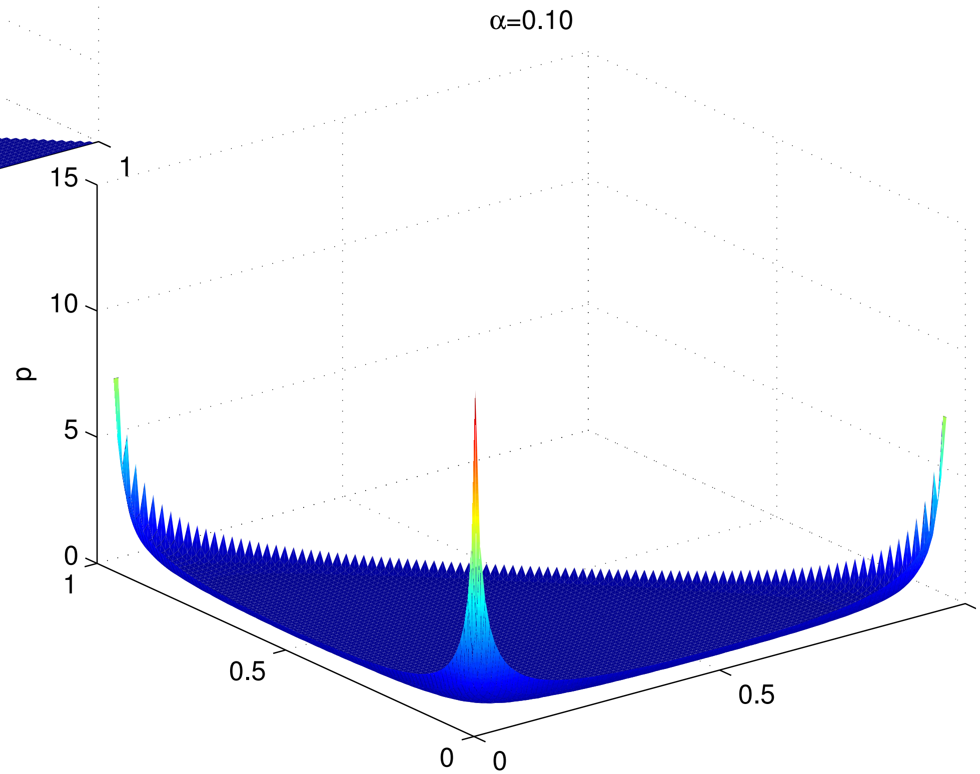


$$0 \leq \theta_k \leq 1$$

$$\sum_{k=1}^{3} \theta_k = 1$$

Dirichlet Probability Densities





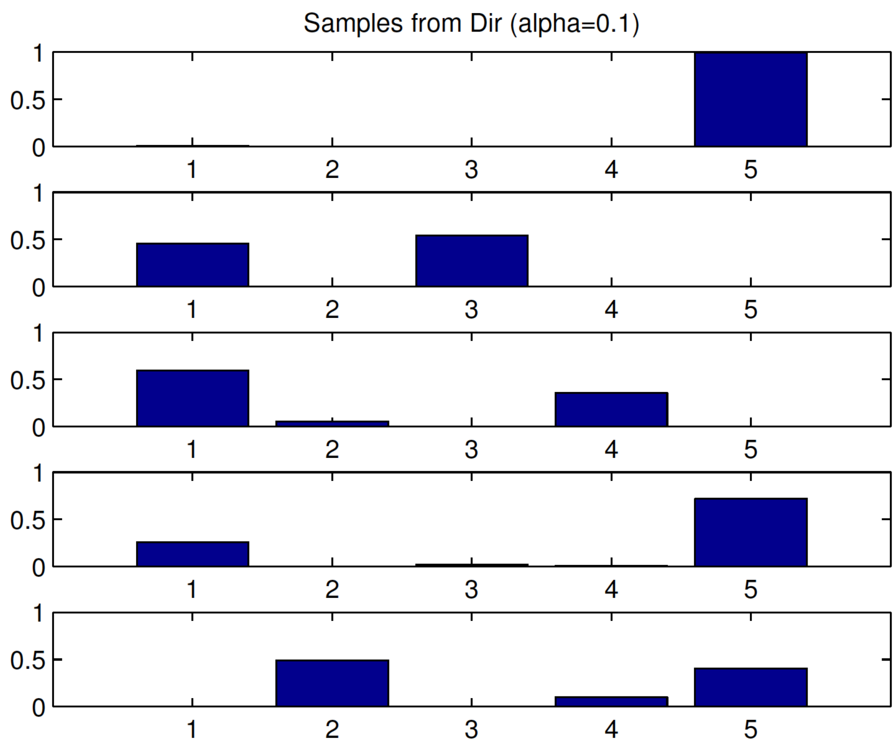
Mean: $$\mathbb{E}(\theta_i) = \frac{\alpha_i}{\sum_j \alpha_j}$$

Mode: $$\hat{\theta}_i = \frac{\alpha_i - 1}{\sum_j \alpha_j - k}$$

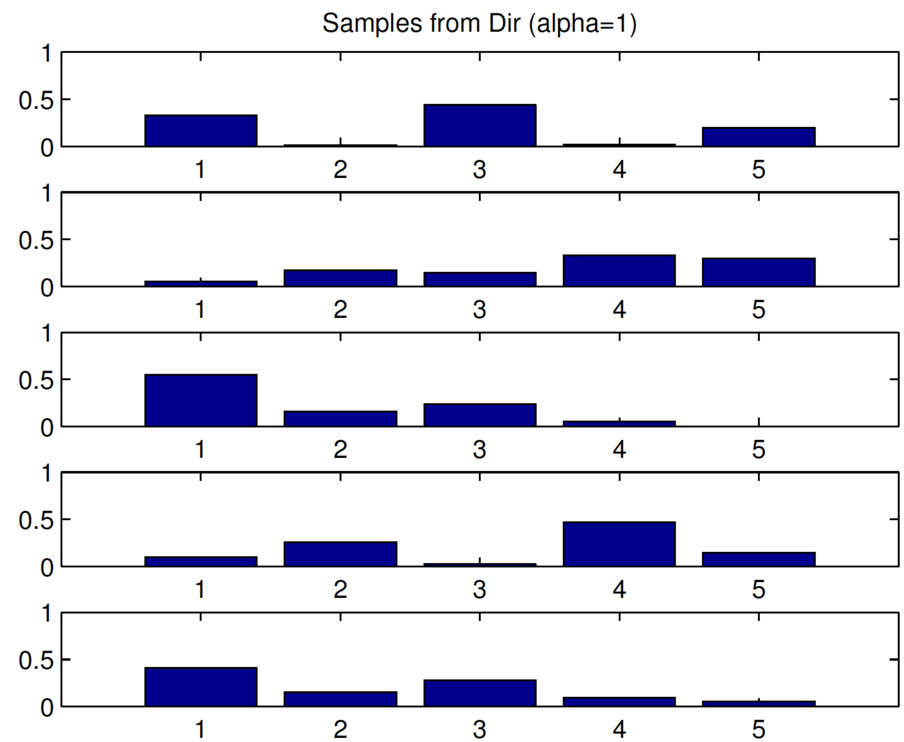
Dirichlet Probability Densities

- Matlab Demo
  - Written by Iyad Obeid

Dirichlet Samples



$$\mathrm{Dir}(\theta \mid 0.1, 0.1, 0.1, 0.1, 0.1)$$



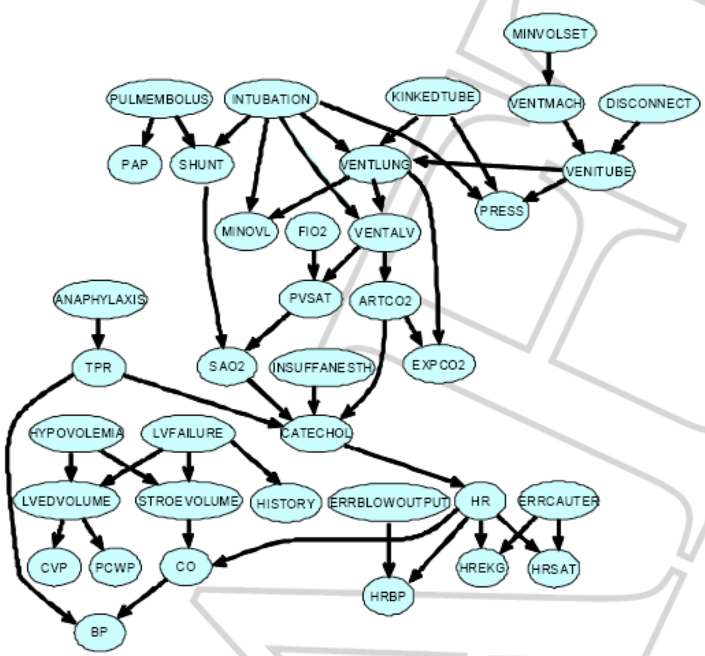
$$\mathrm{Dir}(\theta \mid 1.0, 1.0, 1.0, 1.0, 1.0)$$

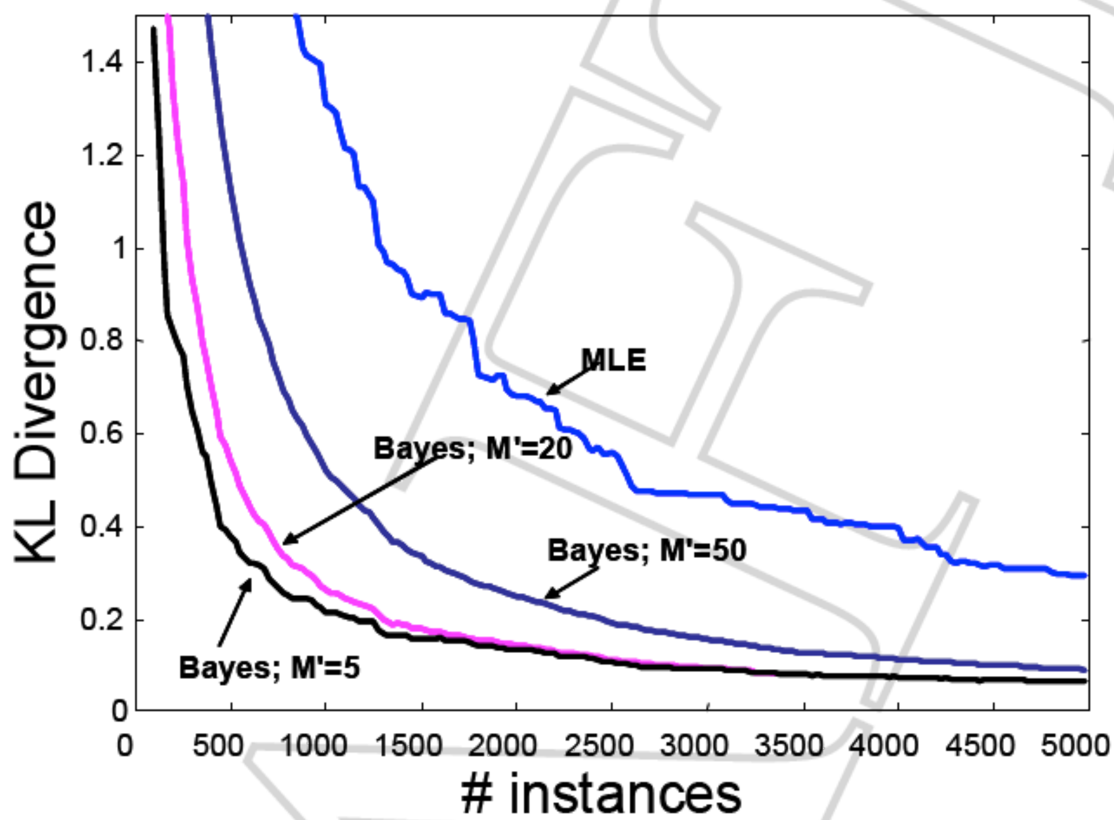
Slide Credit: Erik Sudderth

Priors for BN CPTs

- Consider each CPT: $P(X_i|\mathbf{Pa(X_i)=b})$
- Conjugate prior:
  - Dirichlet($\alpha_{Xi=1|\mathbf{Pa(Xi)=b}},\ldots,\alpha_{Xi=k|\mathbf{Pa(Xi)=b}}$)
- More intuitive:
  - prior counts

An example


MINVOLSET
PULMEMBOLUS
INTUBATION
KINKEDTUBE
VENTMACH
DISCONNECT
PAP
SHUNT
VENTLUNG
VENITUBE
PRESS
MINOVL
FIO2
VENTALV
ANAPHYLAXIS
PVSAT
ARTCO2
TPR
SAO2
INSUFFANESTH
EXPCO2
HYPOVOLEMIA
LVFAILURE
CATECHOL
LVEDVOLUME
STROEVOLUME
HISTORY
ERRBLOWOUTPUT
HR
ERRCAUTER
CVP
PCWP
CO
HREKG
HRSAT
HRBP
BP


KL Divergence
instances
MLE
Bayes; M'=20
Bayes; M'=50
Bayes; M'=5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000