CS 583: Probabilistic Graphical Models

Topic: Parameter Estimation
Chapter: 17

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Parameter estimation for BNs

- Assume the network structure is given
- The data $\mathcal{D}$ consists of fully observed instances of the network variables
  - $\mathcal{D} = \{x[1], x[2], \ldots, x[n]\}$
- Estimate the network parameters, i.e., learn the CPDs
- Two approaches
  1. Maximum likelihood estimation
  2. Bayesian estimation

SIMPLEST CASE – ONE VARIABLE

- Imagine we have a thumbtack
- Flip it, and it comes as heads or tails

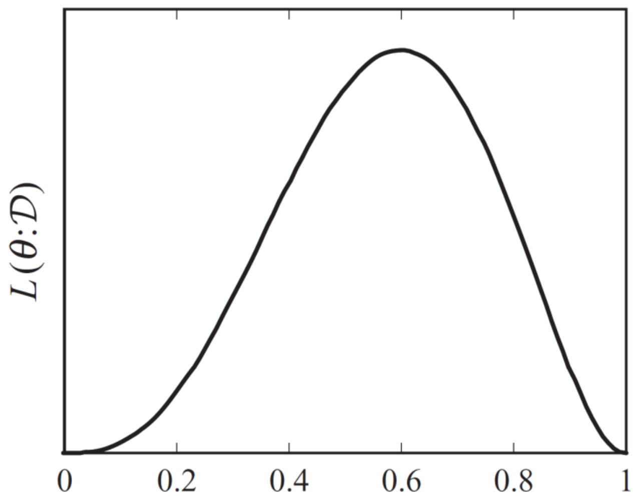
- Assume we flip it 100 times and it comes head 30 times
- What is θ?

THUMBTACK TOSSES

- Assume we have a set of thumbtack tosses
  - $\mathcal{D} = \{x[1], \ldots, x[n]\}$
- Also assume each toss, x[i], is IID
- We define a *hypothesis space* Θ
  - Θ is the set of all parameters $\theta \in [0, 1]$
- We formulate an *objective function*
  - The objective function tells us how good a given hypothesis (in this case θ) is

LIKELIHOOD

- What is the probability, or *likelihood*, of seeing the sequence H, T, T, H, H?
  - $\theta * (1-\theta) * (1-\theta) * \theta * \theta = \theta^3 (1-\theta)^2$



When is $L(\theta : \mathcal{D})$ maximum?

Likelihood/Log-Likelihood

- Number of heads = h, number of tails = t
- Likelihood: $L(\theta : \mathcal{D}) = \theta^h(1-\theta)^t$
- Log-likelihood: $l(\theta : \mathcal{D}) = h\ln\theta + t\ln(1-\theta)$
- Find θ that maximizes the log-likelihood
- Take derivate of $l(\theta : \mathcal{D})$ with respect to θ and set it to zero

Maximum likelihood for a multinomial

- Domain of X is {A, B, C}
- We see A a times, B b times, and C c times.
- P(X=A) is p, P(X=B) is q, and P(C) = $1 - p - q$
- What are p and q?
- Proof?

Constrained optimization

- Assume X can take k values
- $P(X=x_i) = \theta_i$
- Find $\boldsymbol{\theta}$ that maximizes the entropy
  - $H(X) = -\Sigma_i \theta_i \log_2 \theta_i$
- If we take the partial derive w.r.t. θ_i
  - …

Constrained optimization

$$
\begin{array}{lll}
\text{Find} & \boldsymbol{\theta} & \\
\text{maximizing} & f(\boldsymbol{\theta}) & \\
\text{subject to} & & \\
& c_1(\boldsymbol{\theta}) & = \quad 0 \\
& & \dots \\
& c_m(\boldsymbol{\theta}) & = \quad 0
\end{array}
$$

Form the Lagrangian:

$$F(\boldsymbol{\theta}, \boldsymbol{\lambda}) = f(\boldsymbol{\theta}) - \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{\theta})$$

Lagrange multipliers examples

1. Maximize x*y st. x+y = 10
2. Maximize x+y st. $x^2+y^2 = 1$
3. Entropy
4. Maximum likelihood estimate for a multinomial

ML for BNs

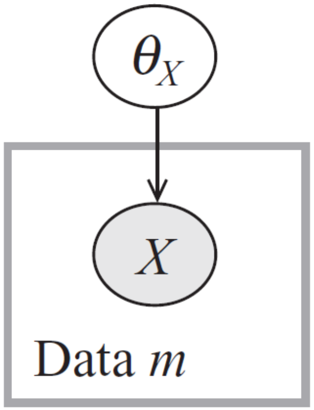
- Simple structure
  - $X \rightarrow Y$
- General structure
  - The key is that the parameters for each variable can be optimized independently
  - Examples

Bayesian estimation

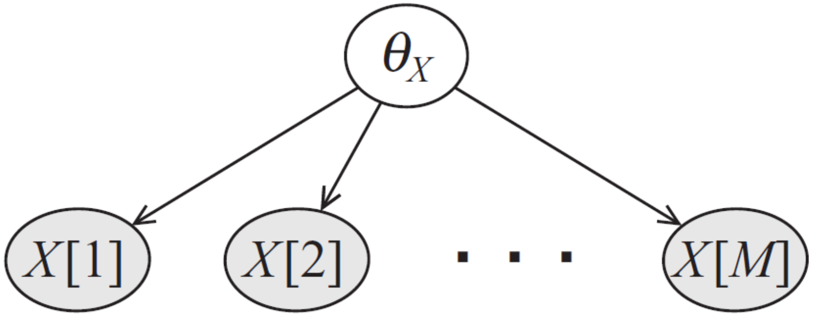
- Assume we flip a coin 10 times and we get 4 Heads, 6 Tails
  - What is P(C=H)?
- Assume we flip a thumbtack 10 times and we get 4 Heads, 6 Tails
  - What is P(T=H)?
- What if we repeat the flips 10M times and we get 4M Heads and 6M Tails?
- Bayesian estimation will let us encode our *prior knowledge*

INDEPENDENCE?

- Earlier, we assumed the tosses are independent
- This is true if we know θ
- If we don't know θ, then each toss tells us something about θ, thus the next toss



(a)

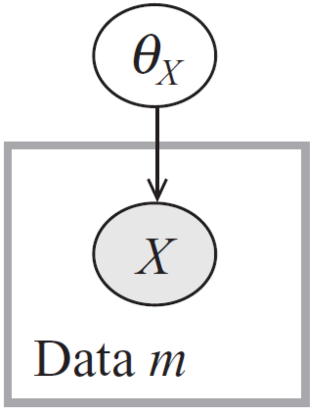


(b)

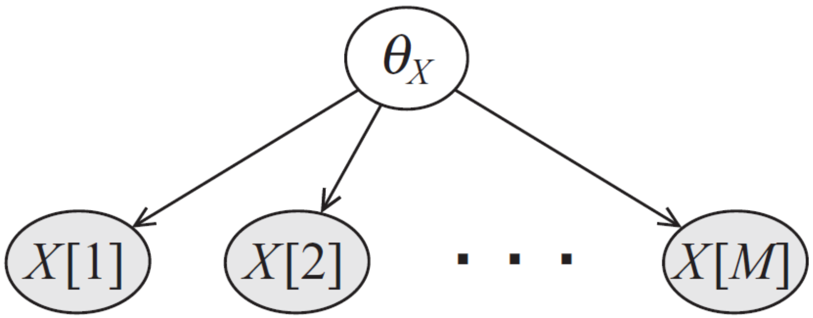
Bayesian Estimation

- Rather than a single θ, we will instead have a probability distribution, $P(\theta)$, over θ

Bayesian estimation



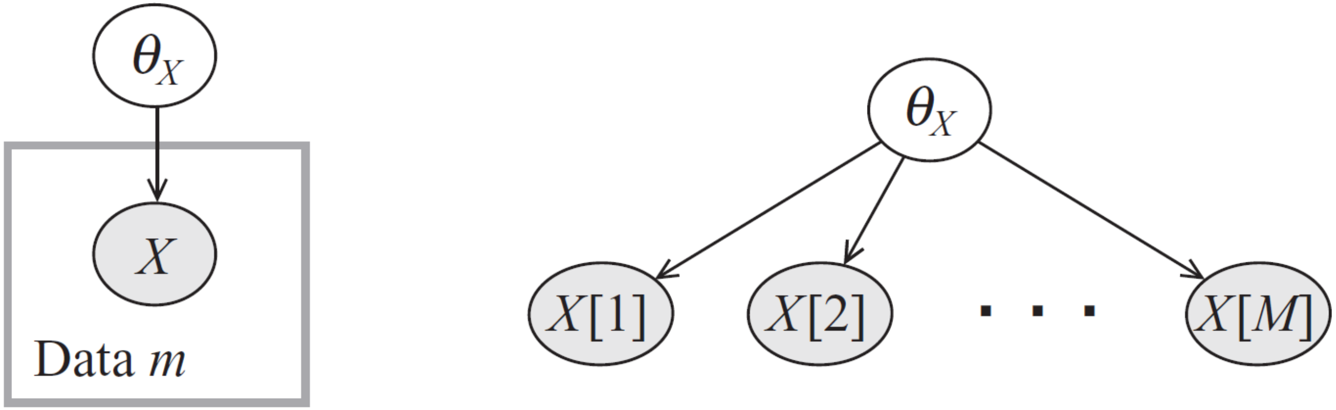
(a)



(b)

- We treat the parameter θ as a random variable
- We ascribe a prior probability to θ, P(θ), encoding our prior knowledge

PARAMETERS



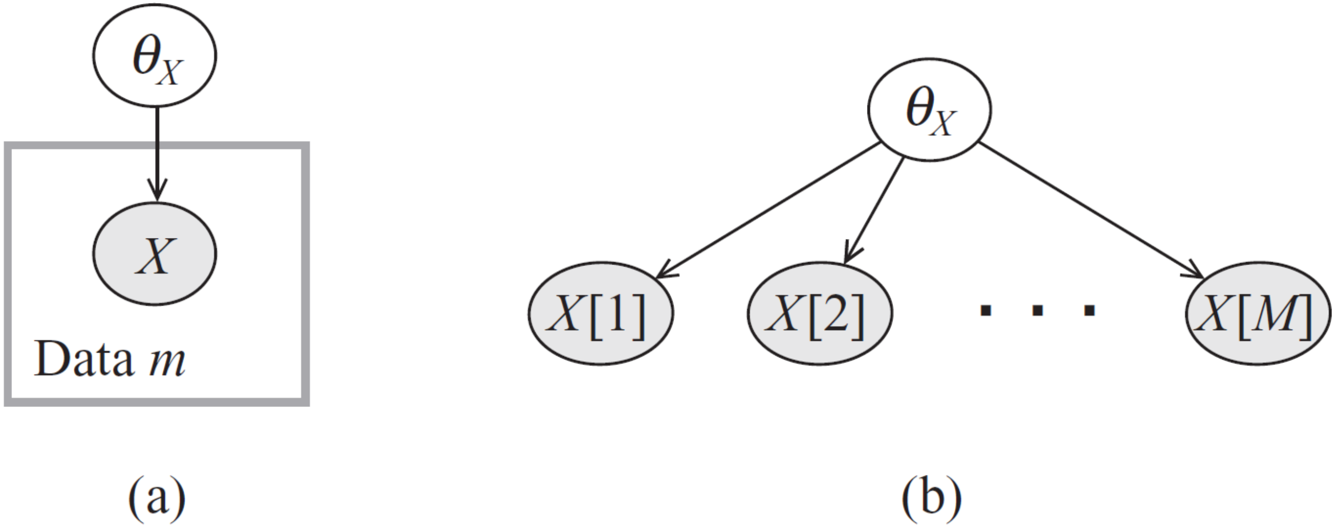
(a) (b)

- $P(X[i] = x^1 \mid \theta_x) = \theta$; $P(X[i] = x^0 \mid \theta_x) = (1 - \theta)$
- P(θ_X)?
  - A continuous distribution over the interval [0,1]

Posterior and prediction

- We are interested in
  - The probability of the next instance, given data
    - $P(x[M+1] \mid D)$
  - The posterior distribution of θ given data
    - $P(\theta \mid D)$

Factorization



$$
\begin{aligned}
P(x[1], x[2], \ldots, x[M], \theta_X) &= P(\theta_X) \prod_{m=1}^{M} P(x[m] \mid \theta_X) \\
&= P(\theta_X)\,\theta^{M[1]}(1-\theta)^{M[0]}
\end{aligned}
$$

Posterior and P(X[M+1] | D)

Posterior distribution

$$P(\theta_X \mid D) = \frac{P(x[1],\ldots,x[M] \mid \theta_X)P(\theta_X)}{P(x[1],\ldots,x[M])}$$

$$P(x[M+1] \mid D) = \int_0^1 P(x[M+1] \mid \theta_X, x[1],\ldots x[M])P(\theta_X \mid x[1],\ldots,x[M])d\theta$$

$$= \int_0^1 \underbrace{P(x[M+1] \mid \theta_X)}_{\theta \text{ or } 1-\theta \text{ (if binary)}} \underbrace{P(\theta_X \mid x[1],\ldots,x[M])}_{\text{Posterior}} d\theta$$

θ or 1-θ (if binary)

Posterior

Think of taking a weighted average

P(x[M+1] | D)

$$P(x[M+1] \mid x[1],\ldots,x[M]) \quad = \quad \int_0^1 P(x[M+1] \mid \theta_X) P(\theta_X \mid x[1],\ldots,x[M]) d\theta$$

$$= \quad \int_0^1 P(x[M+1] \mid \theta_X) \frac{P(\theta_X) P(x[1],\ldots,x[M] \mid \theta_X)}{P(x[1],\ldots,x[M])}$$

P(x[1], …, x[M]) is a constant

$$P(x[M+1] \mid x[1],\ldots,x[M]) \propto \int_0^1 P(x[M+1] \mid \theta_X) P(\theta_X) P(x[1],\ldots,x[M] \mid \theta_X) d\theta$$

Uniform prior

- We have a uniform prior over θ_X. That is, $p(\theta_X)=1$
- $P(X[M+1]=x^1 \mid x[1],\ldots,x[M])$?

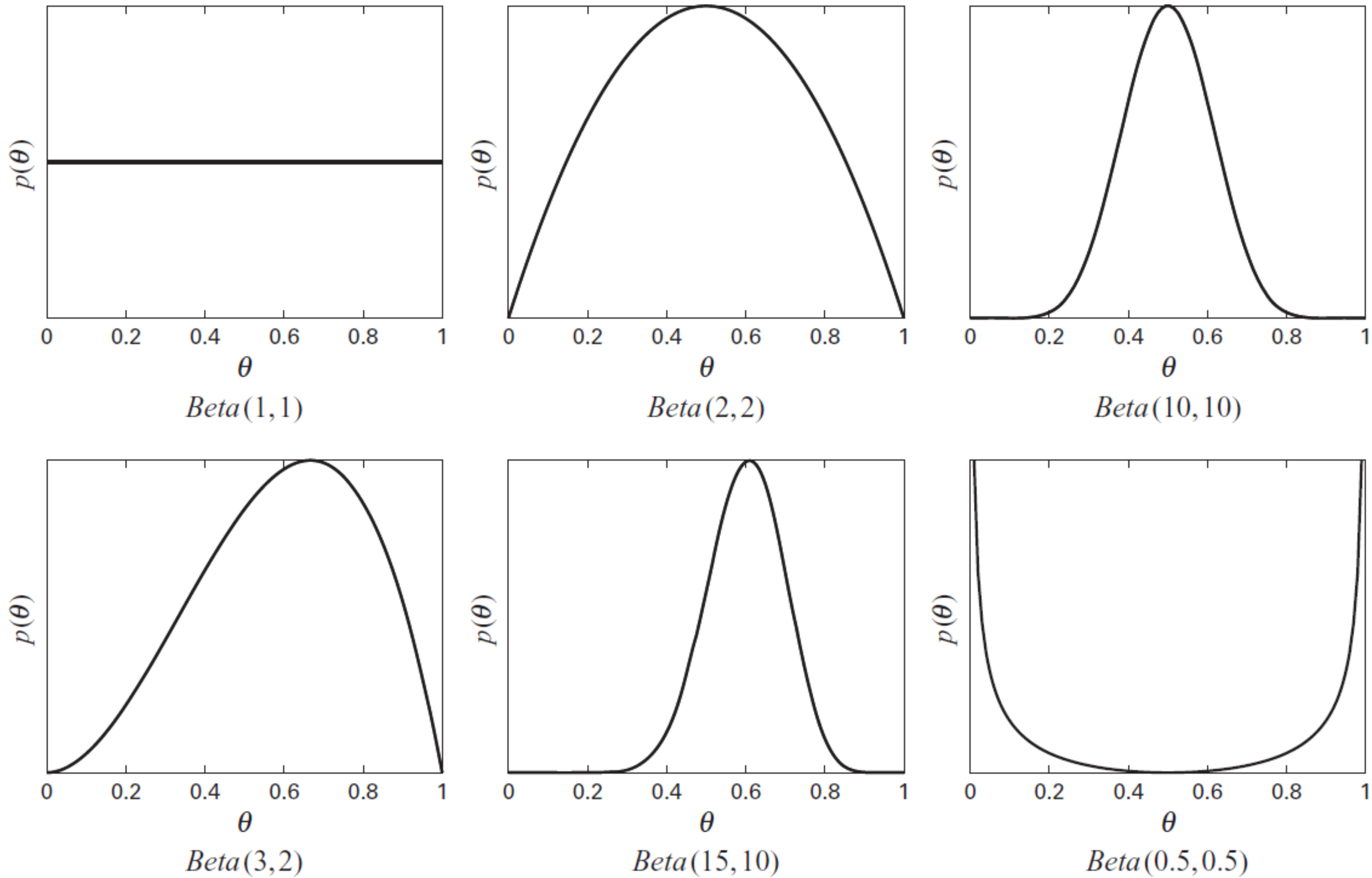
Uniform prior

- We have a uniform prior over θ_X. That is, $p(\theta_X)=1$
- $P(X[M+1]=x^1 \mid x[1],\ldots,x[M])$? That is, $P(X[M+1]=x^1 \mid D)$?
- For the binary case, $P(X[M+1]=x^1 \mid D) = (t + 1) / (t + f + 2)$, where t is the number of True cases and f is the number of False cases in D
- This is also called *Laplace smoothing*
- What about the posterior, $P(\theta \mid D)$, if the prior $P(\theta)$ is uniform?

Beta Distribution

- $\theta \sim \text{Beta}(\alpha,\beta)$ if $P(\theta) = \gamma\theta^{\alpha-1}(1-\theta)^{\beta-1}$ where γ is a normalizing constant
- Mean: $\alpha/(\alpha+\beta)$
- Mode: $(\alpha-1)/(\alpha+\beta-2)$
- Note that the mode is closer to the mean when α and β are large
- Read more at
  - https://en.wikipedia.org/wiki/Beta_distribution

Beta Distribution



Beta distribution

- What is $P(X[M+1]=x^1 \mid D)$ if the prior is Beta(α,β)?
  - $P(X[M+1]=x^1 \mid D) = (p + \alpha) / (p + \mathrm{n} + \alpha + \beta)$
- What is the posterior, $P(\theta \mid D)$, if the prior is Beta(α,β)?
  - $P(\theta \mid D) = \mathrm{Beta}(p + \alpha, n + \beta)$
- α and β work like pseudo-counts for the positive and negative cases respectively
- What values to choose for α and β?
  - It depends on our belief and the strength of our belief

Dirichlet priors

- Generalizes the Beta distribution for multinomials

$$\theta \sim Dirichlet(\alpha_1, \ldots, \alpha_K) \text{ if } P(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- What is $P(X[M+1]=x^i \mid D)$ if the prior is Dirichlet?
  - $P(X[M+1]=x^i \mid D) = (n_i+\alpha_i) / (|D| + \alpha)$ where n_i is the number of times the i^{th} case appears in D and $\alpha = \alpha_1 + \alpha_2 + \ldots + \alpha_K$
- What is the posterior, $P(\theta \mid D)$, if the prior is Dirichlet?
  - $P(\theta \mid D) = \text{Dirichlet}(n_1 + \alpha_1, n_2 + \alpha_2, \ldots, n_K + \alpha_K)$

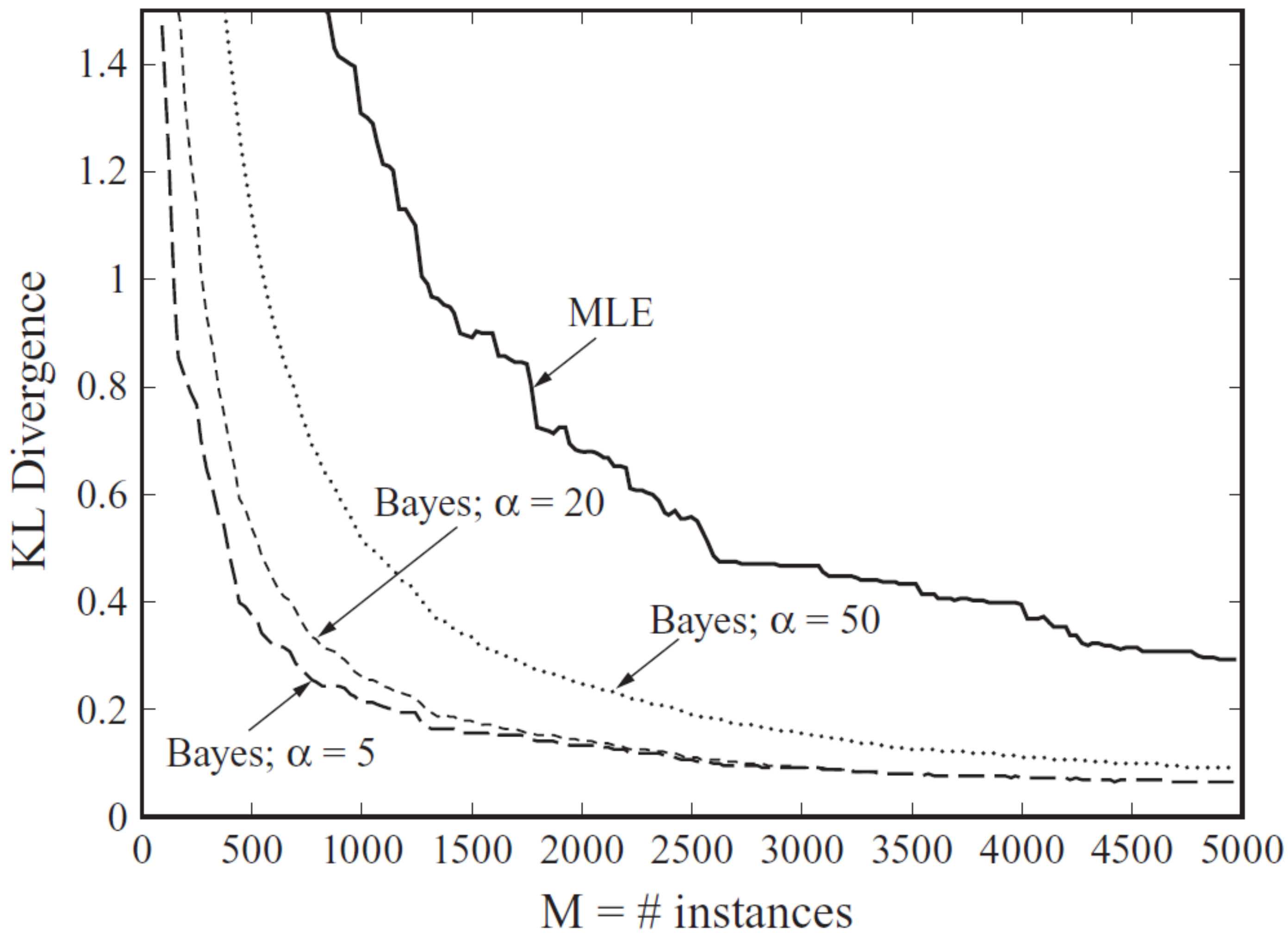
Bayesian estimation

- In MLE for BNs, we optimized each parameter independently
- Can we do the same for Bayesian estimation for BNs?
  - Only if the prior also factorizes wrt the BN
- What about the priors? How do we choose them?
  1. Ask the prior for each variable to an expert
  2. Use the same prior for all variables
     - This is called the *K2 prior*
  3. Imagine a dataset D' of imaginary instances
     - The number of imaginary instances for x is $|D'| * P'(x, pa(x))$
     - This is called the *BDe prior*
     - What is P'?
       - Could be anything; e.g., a marginally independent distribution

Bayesian estimation examples

- Try a dataset using
  - MLE
  - Bayesian
    - K2
    - BDe

ICU Alarm network – fig 17.c.1



Missing Data

- Note: the following information is from Chapter 19

Missing Data

1. Accidental
   - E.g., sensor failure
2. Intentional
   - E.g., not all tests are ordered for medical diagnosis
3. Hidden variable
   - E.g., cluster assignment; hidden cause

APPROACHES

1. Ignore the data points with missing values
   - Not economical (throwing data away), not necessarily accurate (if missing intentionally), and might not be possible (hidden variables)
2. Gradient optimization
3. Expectation Maximization

Expectation Maximization (EM)

- Initialize θ
- Iterate
  - Expectation
    - $M[x, pa(x)] = \sum P(x, pa(x) \mid observed)$ for each X
  - Maximization
    - $\theta_x = M[x, pa(x)] \, / \, M[pa(x)]$ for each X

EM Examples

- Simple network
  - $X \rightarrow Y$
  - Let's see a few cases:
    - X is missing in a small percentage; sometimes X and sometimes Y is missing; X is a hidden variable
- A more complicated network:
  - A disease (D) variable and three test variables (T_1, T_2 and T_3)
  - A naïve Bayes structure: D is the parent of T_1, T_2 and T_3
  - T_1 is the fastest but least accurate test. T_2 requires more time but is more accurate. T_3 requires the most time but also is the most accurate test.
  - The doctors order T_1 for everyone first. Depending on the results, they might order T_2, and depending on its results, they might order T_3
  - D might be totally hidden or might be observed for only a small subset of the data