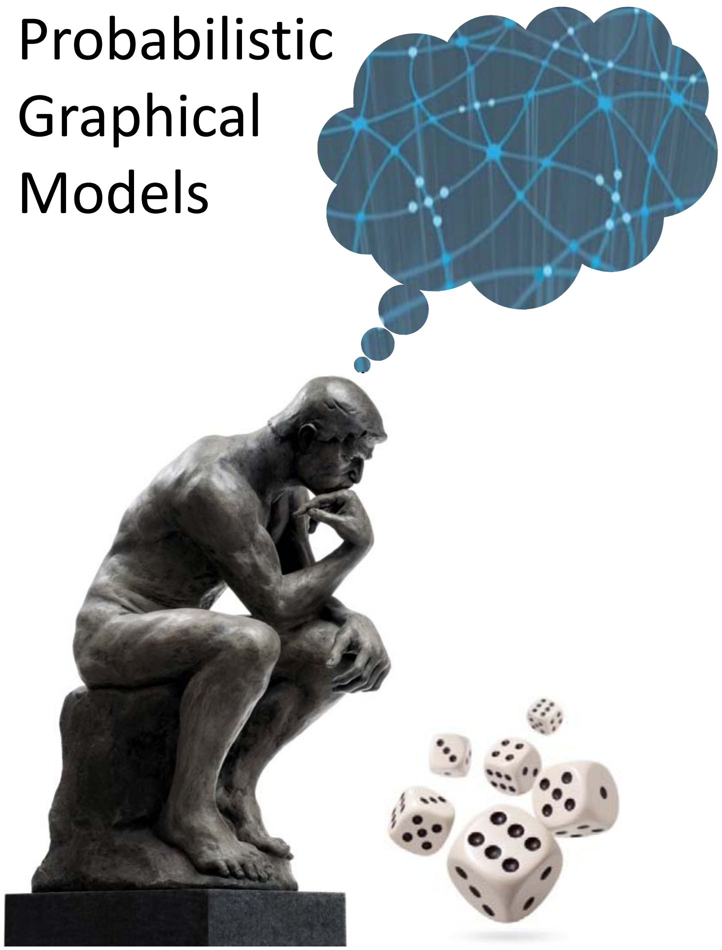
Probabilistic
Graphical
Models



Learning

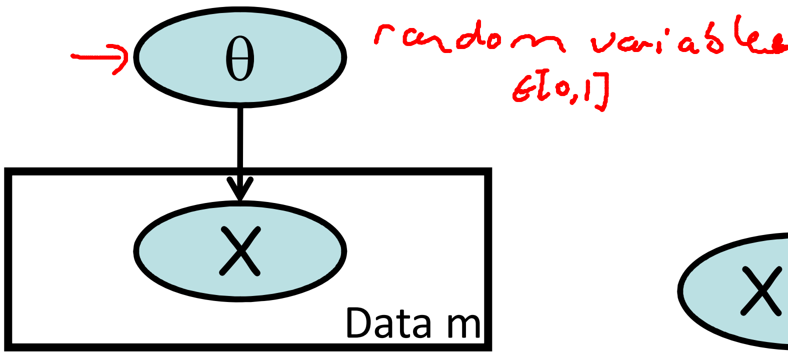
Parameter Estimation

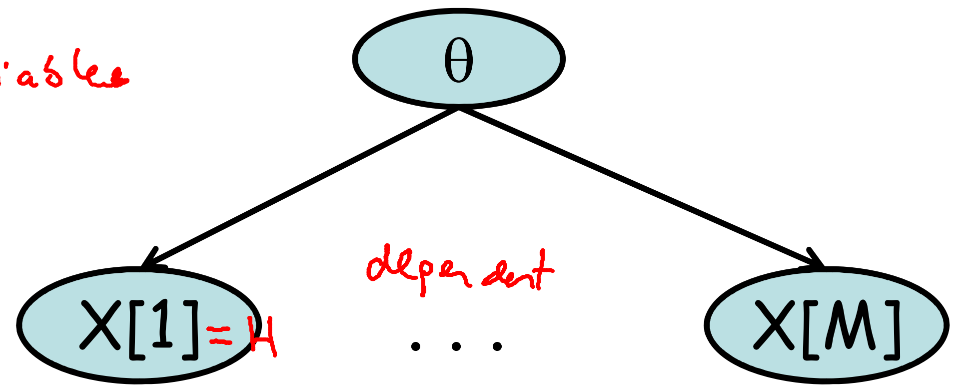
Bayesian Estimation

Limitations of MLE

- Two teams play 10 times, and the first wins 7 of the 10 matches
  - ⇨ Probability of first team winning = 0.7
- A coin is tossed 10 times, and comes out 'heads' 7 of the 10 tosses
  - ⇨ Probability of heads = 0.7
- A coin is tossed 10000 times, and comes out 'heads' 7000 of the 10000 tosses
  - ⇨ Probability of heads = 0.7

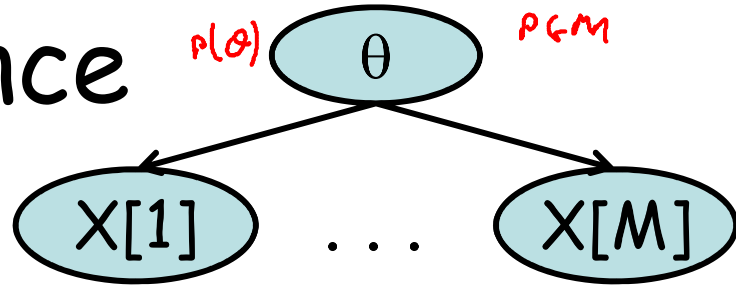
Parameter Estimation as a PGM





- 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



- Joint probabilistic model

$$
\begin{aligned}
P(x[1],\ldots, x[M], \theta) &= P(x[1],\ldots, x[M] \mid \theta)\, P(\theta) \\
&= P(\theta) \underbrace{\prod_{i=1}^{M} P(x[i] \mid \theta)}_{\text{likelihood function}} \\
&= P(\theta)\, \theta^{M_H} (1-\theta)^{M_T}
\end{aligned}
$$

$$
P(\theta \mid \underbrace{x[1],\ldots, x[M]}_{\text{data } D}) = \frac{\overbrace{P(x[1],\ldots, x[M] \mid \theta)}^{\text{likelihood}}\, \overbrace{P(\theta)}^{\text{prior}}}{P(x[1],\ldots, x[M])}
$$

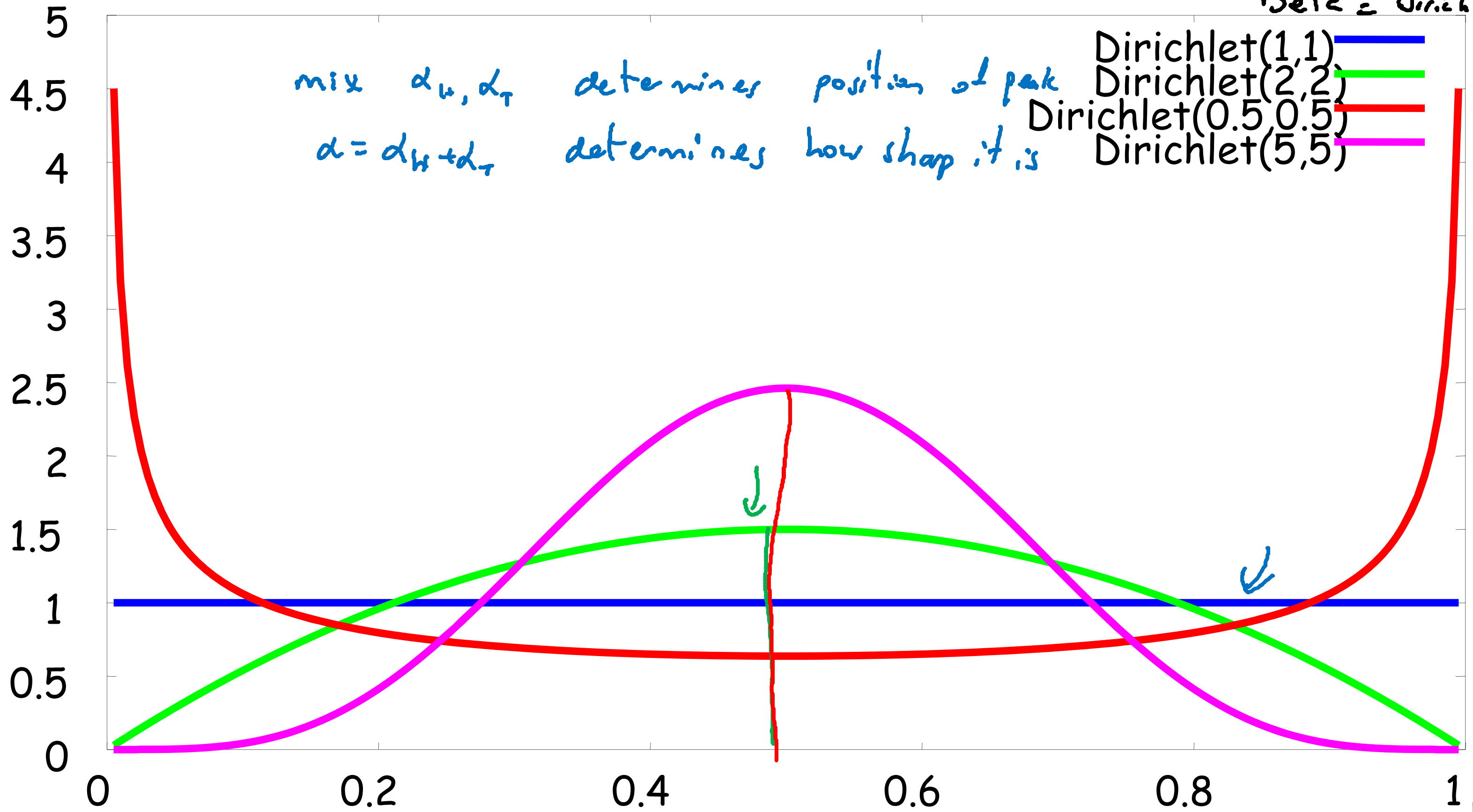
constant relative to θ

Dirichlet Distribution

- θ is a multinomial distribution over k values
- Dirichlet distribution $\theta \sim \text{Dirichlet}(\underbrace{\alpha_1,\ldots,\alpha_k}_{\text{hyperparameters}})$
  - where $P(\theta) = \frac{1}{Z}\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$ and $Z = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}$ $\quad \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$
- Intuitively, hyperparameters α_i correspond to the number of samples we have seen

Dirichlet Distributions

Beta = Dirichlet(α_1, α_2)



Dirichlet Priors & Posteriors

$$\overbrace{P(\theta \mid D)}^{\text{posterior}} \propto \overbrace{P(D \mid \theta)}^{\text{likelihood}}\,\overbrace{P(\theta)}^{\text{prior}}$$

$\theta_i^{M_i+\alpha_i-1}$

multinomial θ

M_i = # instances with x^i

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

- If P(θ) is Dirichlet and the likelihood is multinomial, then the posterior is also Dirichlet
  - Prior is Dir($\alpha_1,\ldots,\alpha_k$)
  - Data counts are $M_1,\ldots,M_k$
  - Posterior is Dir($\alpha_1+M_1,\ldots\alpha_k+M_k$)
- Dirichlet is a conjugate prior for the multinomial

prior, posterior have the same form

Summary

- Bayesian learning treats parameters as random variables
  - Learning is then a special case of inference
- Dirichlet distribution is conjugate to multinomial
  - Posterior has same form as prior
  - Can be updated in closed form using sufficient statistics from data