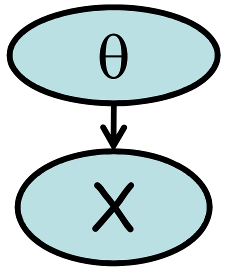
Probabilistic Graphical Models

Learning

Parameter Estimation

Bayesian Prediction

Bayesian Prediction



$\sim$ Dirichlet($\alpha_1,\ldots,\alpha_k$)

$$P(X) = \int_\theta P(X \mid \theta) P(\theta) d\theta$$

marginalizing over θ

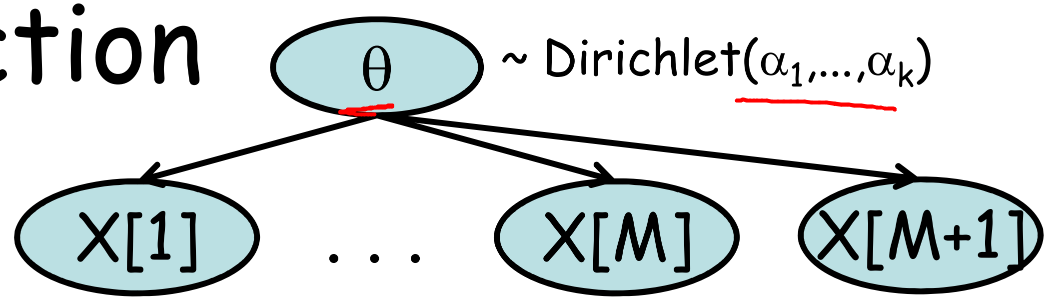
$$P(X = x^i \mid \theta) = \frac{1}{Z} \int_\theta \theta_i \cdot \underbrace{\prod_j \theta^{\alpha_j - 1}}_{\text{prior}} d\theta$$

$$= \frac{\alpha_i}{\sum_j \alpha_j = \alpha}$$

fraction of instances we've seen where x^i

- Dirichlet hyperparameters correspond to the number of samples we have seen

Bayesian Prediction



$\sim$ Dirichlet($\alpha_1,\ldots,\alpha_k$)

$$P(x[M+1] \mid x[1],\ldots,x[M])$$

$$= \int_\theta P(x[M+1] \mid x[1],\ldots,x[M],\theta)\, P(\theta \mid x[1],\ldots,x[M])\, d\theta$$

$$= \int_\theta P(x[M+1] \mid \theta)\, \boxed{P(\theta \mid x[1],\ldots,x[M])}\, d\theta$$

$\sim$ Dirichlet($\alpha_1+M_1,\ldots,\alpha_k+M_k$)

posterior over θ given D

$$P(X[M+1] = x^i \mid \theta, x[1],\ldots,x[M]) = \frac{\alpha_i + M_i}{\alpha + M}$$

$\alpha = \sum \alpha_i$

$M = \sum M_i$

- Equivalent sample size $\alpha = \alpha_1 + \ldots + \alpha_K$
  - Larger α ⇨ more confidence in our prior

Daphne Koller

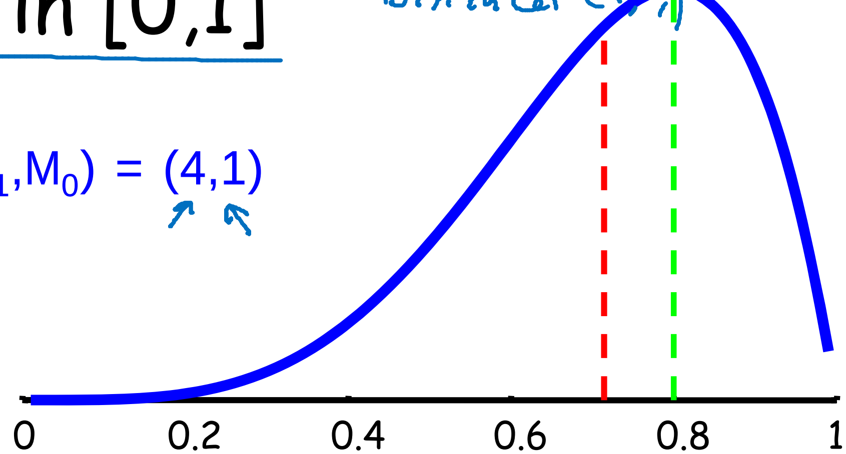
Example: Binomial Data

- Prior: uniform for θ in [0,1]

Dirichlet (1,1)

$$P(\theta) = \frac{1}{Z}\prod_k \theta_k^{\alpha_k - 1}$$

$(M_1, M_0) = (4,1)$

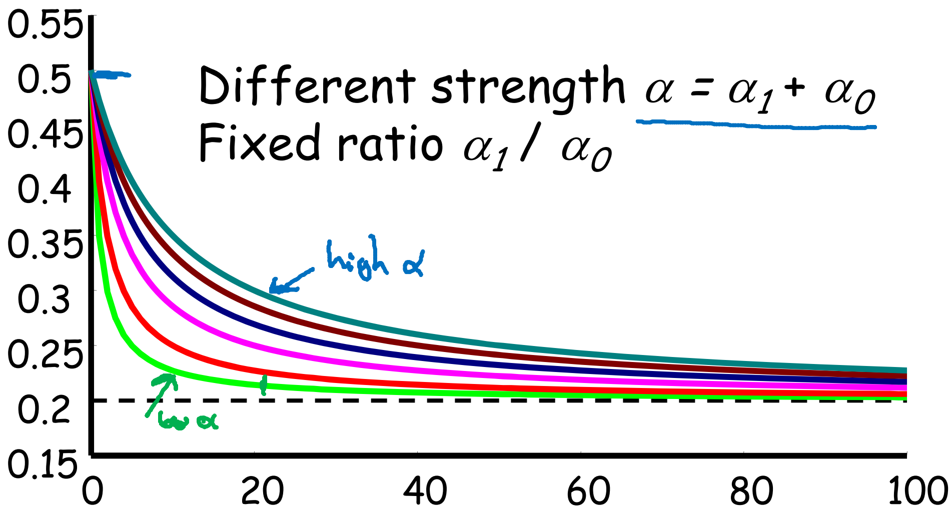


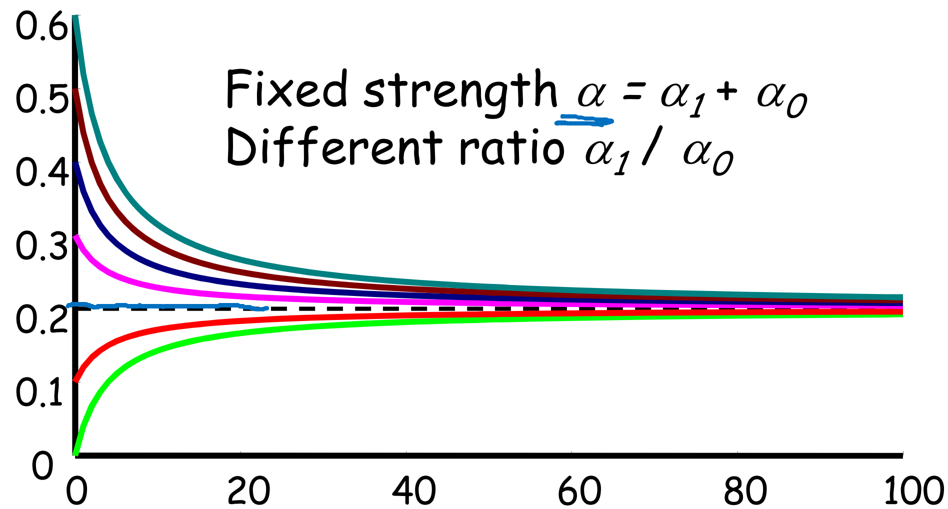
$$\frac{\alpha_1 + M_1}{\alpha + M} = \frac{1+4}{2+5}$$

- MLE for P(X[6]=1)=4/5
- Bayesian prediction is 5/7

Effect of Priors

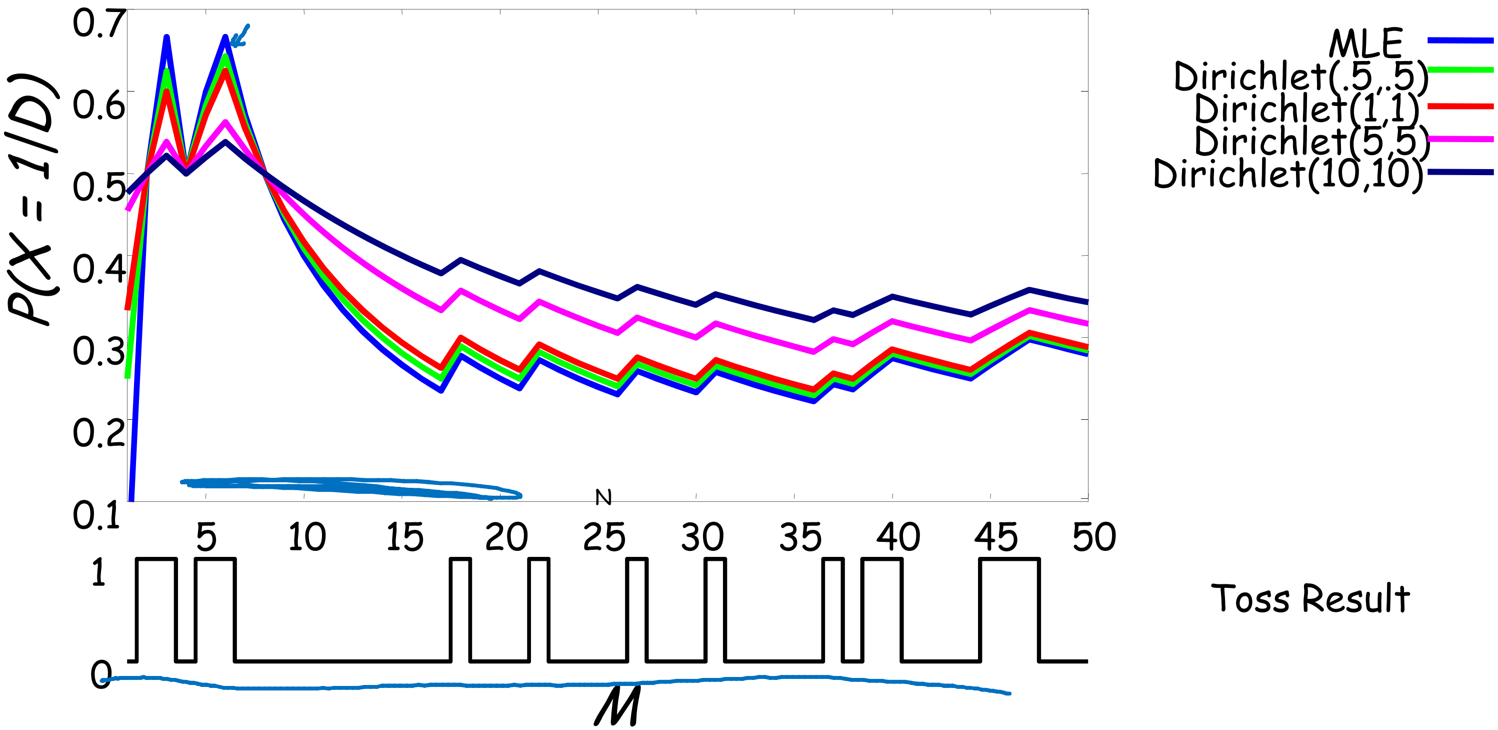
- Prediction of $P(X=1)$ after seeing data with $M_1 = \frac{1}{4} M_0$ as a function of sample size M





Effect of Priors

- In real data, Bayesian estimates are less sensitive to noise in the data



Summary

- Bayesian prediction combines sufficient statistics from imaginary Dirichlet samples and real data samples
- Asymptotically the same as MLE
- But Dirichlet hyperparameters determine both the prior beliefs and their strength