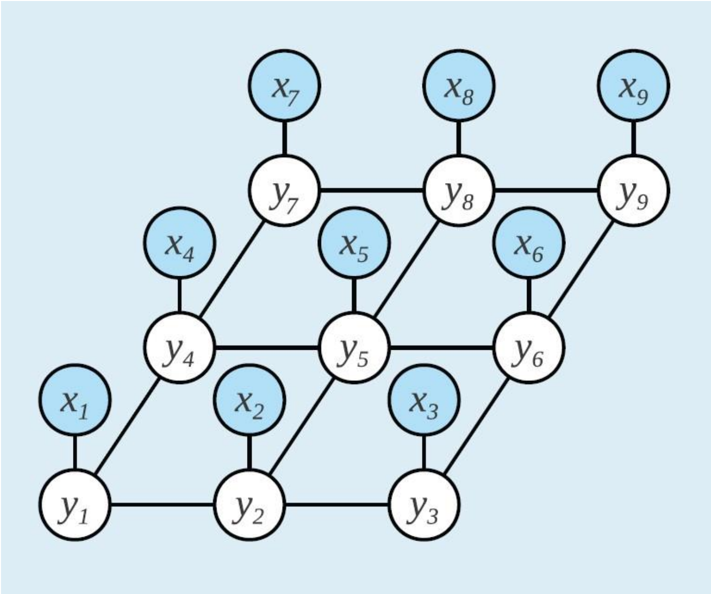
Probabilistic Graphical Models in Bioinformatics

Lecture 6: Learning Bayesian networks- Bayesian parameter estimation



Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Farvardin 21, 1398
(April 10, 2019)

Bayesian parameter estimation

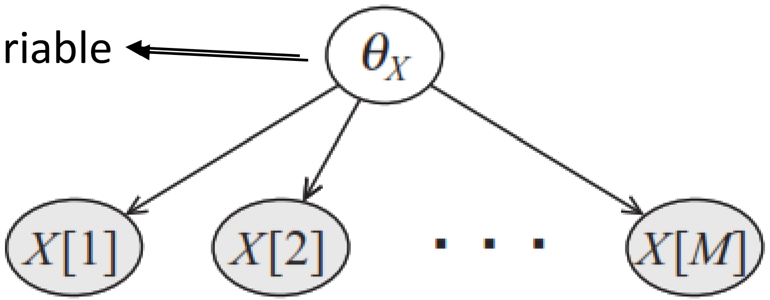
- Thumbtack example:
  - Get 3 heads out of 10 tosses
    - We conclude θ is 0.3.
- Coin example
  - Get 3 heads out of 10 tosses
    - It is less likely that we conclude the parameter of the coin is 0.3.
    - Why? Because we have a lot of prior knowledge about their behavior.
  - Now assume we get 300000 heads out of 1 million tosses.
    - We are willing to conclude this is a trick coin with the parameter 0.3.

- Maximum likelihood approach does not allow us to incorporate
  - Our prior knowledge that coin is fairer than thumbtack
  - Between 10 and one million tosses

Bayesian parameter estimation

- We encode our prior knowledge about θ with a probability distribution
- We create a joint distribution over the parameter θ and the data $X[1], \ldots, X[M]$

Random variable



Ground Bayesian network

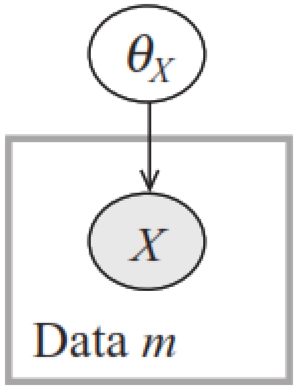


Plate model

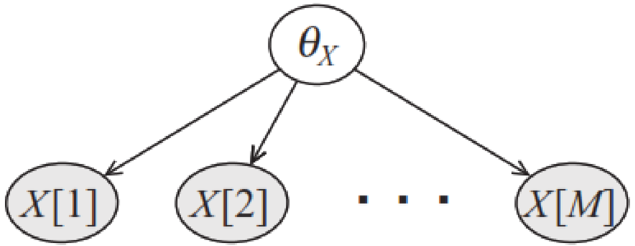
Tosses are conditionally independent given θ!

Question: specify the local probability distributions?

Joint probabilistic model-2

- Local probability distributions
  - $P(\theta)$: our prior knowledge about θ. A continuous distribution over the interval [0, 1].
  - $P(\, X[m] \mid \theta\,)$: probability of data given parameter

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



  - **Notation:** since in Bayesian approach, we treat θ as a random variable, we write $P(\, X[m] \mid \theta\,)$ instead of $P(X[m] : \theta)$.

- Joint distribution

Likelihood function $\mathcal{L}(\theta : D)$

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

Joint probabilistic model-3

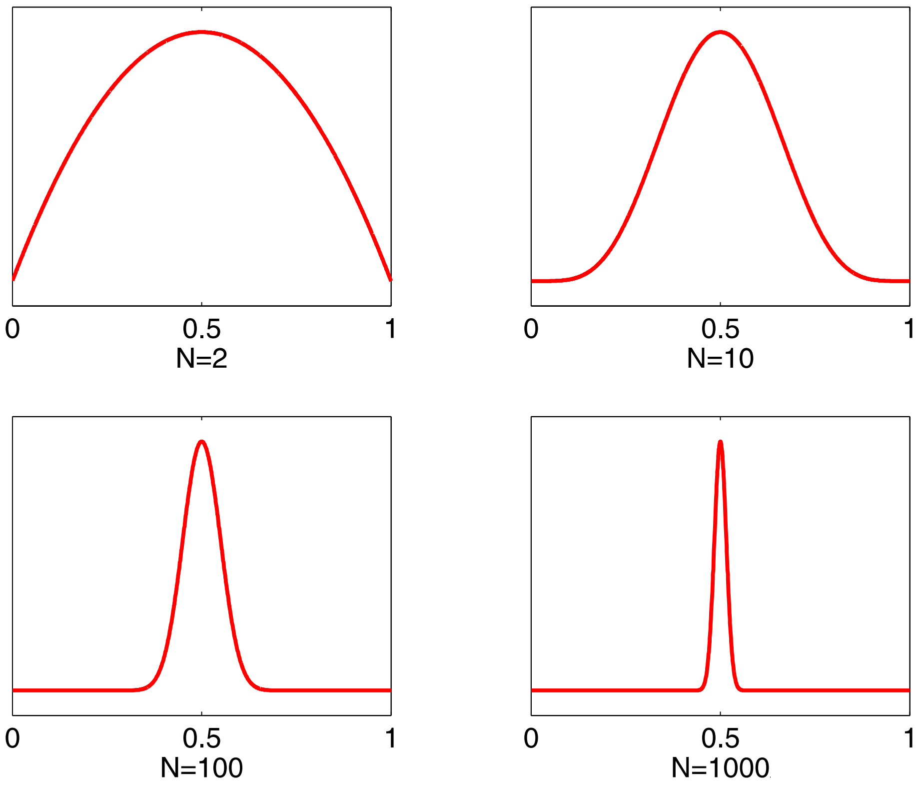
- Posterior distribution over θ:

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

- Hence, posterior is a proportional to the product of the likelihood and the prior

- If the prior is a uniform distribution (i.e., $P(\theta) = 1$ for all $\theta \in [0,1]$), then the posterior is the normalized likelihood function.

Example: posterior of θ for a uniform prior



Prediction

- **Question:** in the ML approach, how do we predict the probability over the next toss?

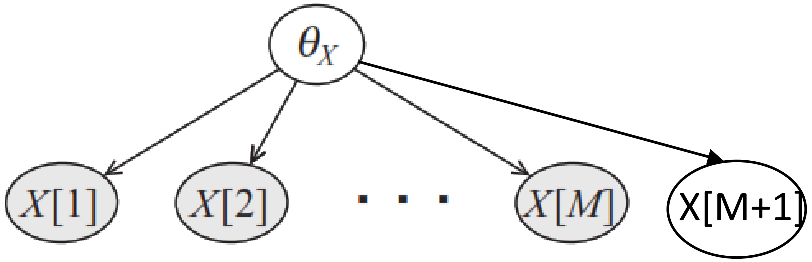
- In Bayesian framework, instead of selecting a single value from the posterior, we use entire posterior distribution $P(\theta \mid D)$ for prediction

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

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

Why?

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



Prediction- thumbtack example

University of TEHRAN

- Assume the prior is uniform over θ in the interval [0, 1].
- We want to compute

head

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

- We use the following formula

$$P(X[M+1] = x^1 \mid x[1], \ldots, x[M]) = \int P(X[M+1] = x^1 \mid \theta) P(\theta \mid x[1], \ldots, x[M]) \mathrm{d}\theta$$

θ

Posterior ∝ likelihood x prior = $\theta^{M[1]}(1-\theta)^{M[0]}$

- Plugging this into integral

$$P(X[M+1] = x^1 \mid x[1], \ldots, x[M]) = \frac{1}{P(x[1], \ldots, x[M])} \int \theta \cdot \theta^{M[1]}(1-\theta)^{M[0]} d\theta.$$

- Doing all the math, we get

$$P(X[M+1] = x^1 \mid x[1], \ldots, x[M]) = \frac{M[1]+1}{M[1]+M[0]+2}.$$

Similar to the MLE: $\frac{M[1]}{M[1]+M[0]}$

Adds one *"imaginary"* sample to each count

Referred to as *Laplace's correction*

Non-uniform prior

- How do we pick a prior distribution?
  - Our choice of the prior should facilitate efficient update of the posterior as we get new data

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

- For reasons we will discuss later, an appropriate prior for coin example is the *Beta distribution*
  - Parameterized by two positive hyperparameters α_1 and α_0

$$E(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

$$\theta \sim \mathrm{Beta}(\alpha_1, \alpha_0); \qquad p(\theta) = \gamma \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_0 - 1}$$

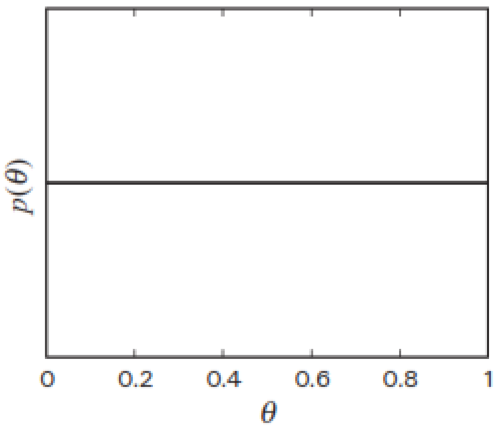
Normalizing constant

$$V(\theta) = \frac{\alpha_0 \alpha_1}{(\alpha_1 + \alpha_0)^2 (\alpha_1 + \alpha_0 + 1)}$$

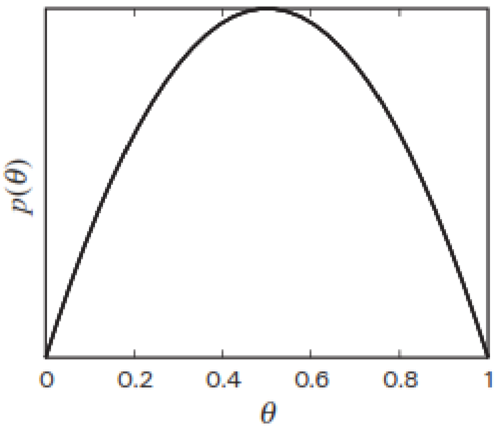
  - Intuitively, α_1 and α_0 correspond to the number of imaginary heads and tails that we have “seen” before starting the experiment.

Examples of Beta distributions

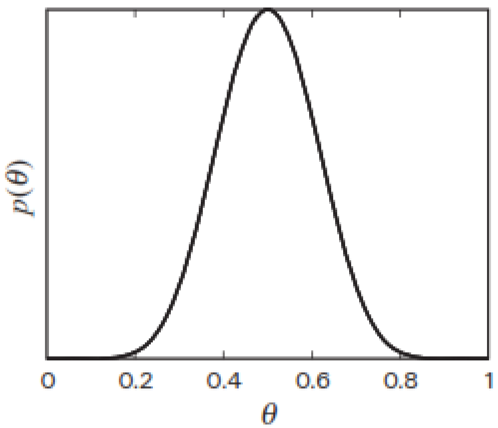
For different choices of hyperparameters



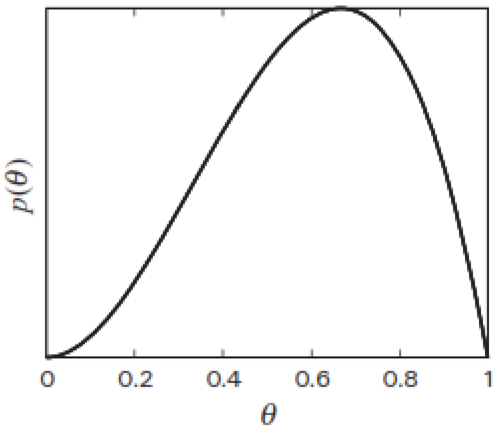
$Beta(1,1)$



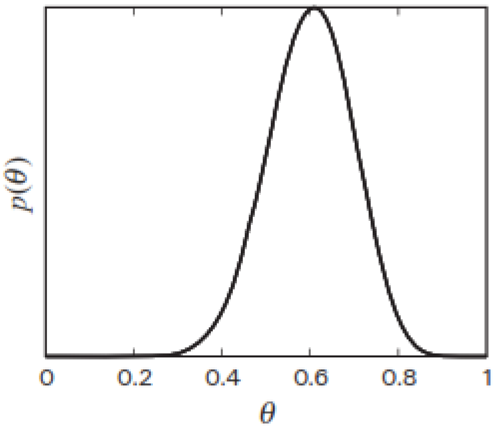
$Beta(2,2)$



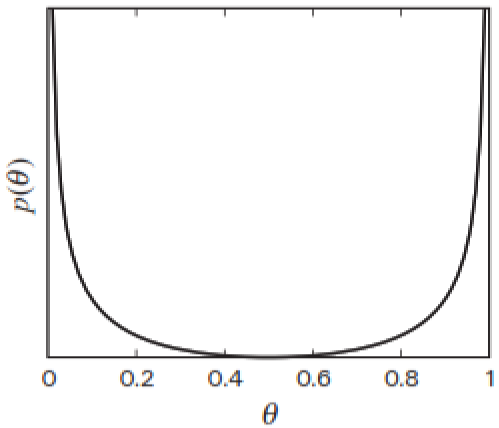
$Beta(10,10)$



$Beta(3,2)$



$Beta(15,10)$



$Beta(0.5,0.5)$

Posterior for Beta prior

- **Prior:** $\theta \sim \mathrm{Beta}(\alpha_1, \alpha_0)$ with density $p(\theta) = \gamma\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$

- **Data:** M[1] heads and M[0] tails

- **Likelihood:** $P(\, D \mid \theta\,) = \theta^{M[1]}(1-\theta)^{M[0]}$

- **Posterior:**

$$\begin{aligned} P(\theta \mid x[1], \ldots, x[M]) &\propto P(x[1], \ldots, x[M] \mid \theta) P(\theta) \\ &\propto \theta^{M[1]}(1-\theta)^{M[0]} \cdot \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \\ &= \theta^{\alpha_1 + M[1] - 1}(1-\theta)^{\alpha_0 + M[0] - 1}, \end{aligned}$$

which is precisely $Beta(\alpha_1 + M[1], \alpha_0 + M[0])$

- We say *Beta* distribution is conjugate to the *Bernoulli* likelihood function

Bayesian prediction

- **Question:** Assume $P(\theta) = Beta(\alpha_1, \alpha_0)$, and consider a single coin toss X. Compute the marginal probability $P(X[1] = x^1)$, based on $P(\theta)$.

$$
\begin{aligned}
P(X[1] = x^1) &= \int_0^1 P(X[1] = x^1 \mid \theta) \cdot P(\theta) d\theta \\
&= \int_0^1 \theta \cdot P(\theta) d\theta = \frac{\alpha_1}{\alpha_1 + \alpha_0}.
\end{aligned}
$$

- **Question:** compute the probability over the next toss given observations so far

$$
P(X[M+1] = x^1 \mid x[1], \ldots, x[M]) = \frac{\alpha_1 + M[1]}{\alpha + M}
$$

Posterior distribution tells us that we have seen $\alpha_1 + M[1]$ heads (imaginary and real) + $\alpha_0 + M[0]$ tails.

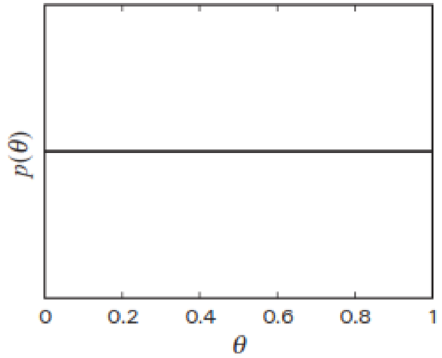
Conjugacy

The *posterior distribution* is in the *same parametric family* as the prior but with **new parameters values**

Bayesian approach vs MLE

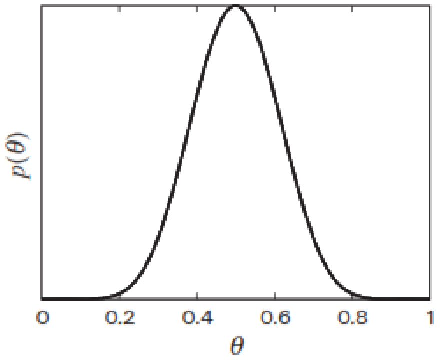
- In Bayesian approach, we can incorporate our prior knowledge.
- The distinction between coin and thumbtack can be captured by the strength of the prior

Thumbtack



$Beta(1,1)$

Coin



$Beta(10,10)$

Question: compare MLE & Bayesian prediction

Data 1: 3 heads in 10 tosses

Data 2: 300 heads in 1000 tosses

- The distinction between a few samples and many samples is captured by peakedness of our posterior.

More about Bayesian approach

- The MLE approach attempts to find the parameter $\hat{\theta}$ that are "best" given the data.

- Bayesian approach does not attempt to find such a *point estimate.* Instead, it tries to update our beliefs about θ*'s* values according to the evidence.

- In the Bayesian approach, we treat parameters, θ, as random variables. Then we describe a joint distribution $P(D, \theta)$ over the data and the parameters.

$$P(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(\mathcal{D})}.$$

- $P(D)$ is called marginal likelihood of the data (i.e., integration of the likelihood over all possible parameter assignments).

$$P(\mathcal{D}) = \int_{\Theta} P(\mathcal{D} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

Conjugate prior for multinomial distribution

- Let X has K values $x^1, \ldots, x^K$.
- Multinomial likelihood

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_k \theta_k^{M[k]}.$$

- Conjugate prior: Dirichlet distribution

$$\boldsymbol{\theta} \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$$

$$P(\boldsymbol{\theta}) \propto \prod_k \theta_k^{\alpha_k - 1}$$

- Observed data : $M[1], \ldots, M[K]$. Posterior distribution $P(\theta \mid D)$:

$$Dirichlet(\alpha_1 + M[1], \ldots, \alpha_K + M[K])$$

Dirichlet distribution

- Generalized the Beta distribution
- Specified by a set of hyperparameters $\alpha_1, \ldots, \alpha_K$
  - denoted as $\boldsymbol{\theta} \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$
  - with the density function $P(\boldsymbol{\theta}) \propto \prod_k \theta_k^{\alpha_k - 1}$
  - With $E[\theta_k] = \frac{\alpha_k}{\alpha}$ where $\alpha = \sum_j \alpha_j$
- Dirichlet hyperparameters are often called *pseudo-counts.*
- It represents the number of times we have seen the different outcomes in our prior experience
- Total α is often called the *equivalent sample size.*

Bayesian prediction with Dirichlet prior

University of TEHRAN

- Posterior distribution $P(\theta \mid D)$

$$Dirichlet(\alpha_1 + M[1], \ldots, \alpha_K + M[K])$$

- Prediction

$$P(x[M+1] = x^k \mid \mathcal{D}) = \frac{M[k] + \alpha_k}{M + \alpha}.$$

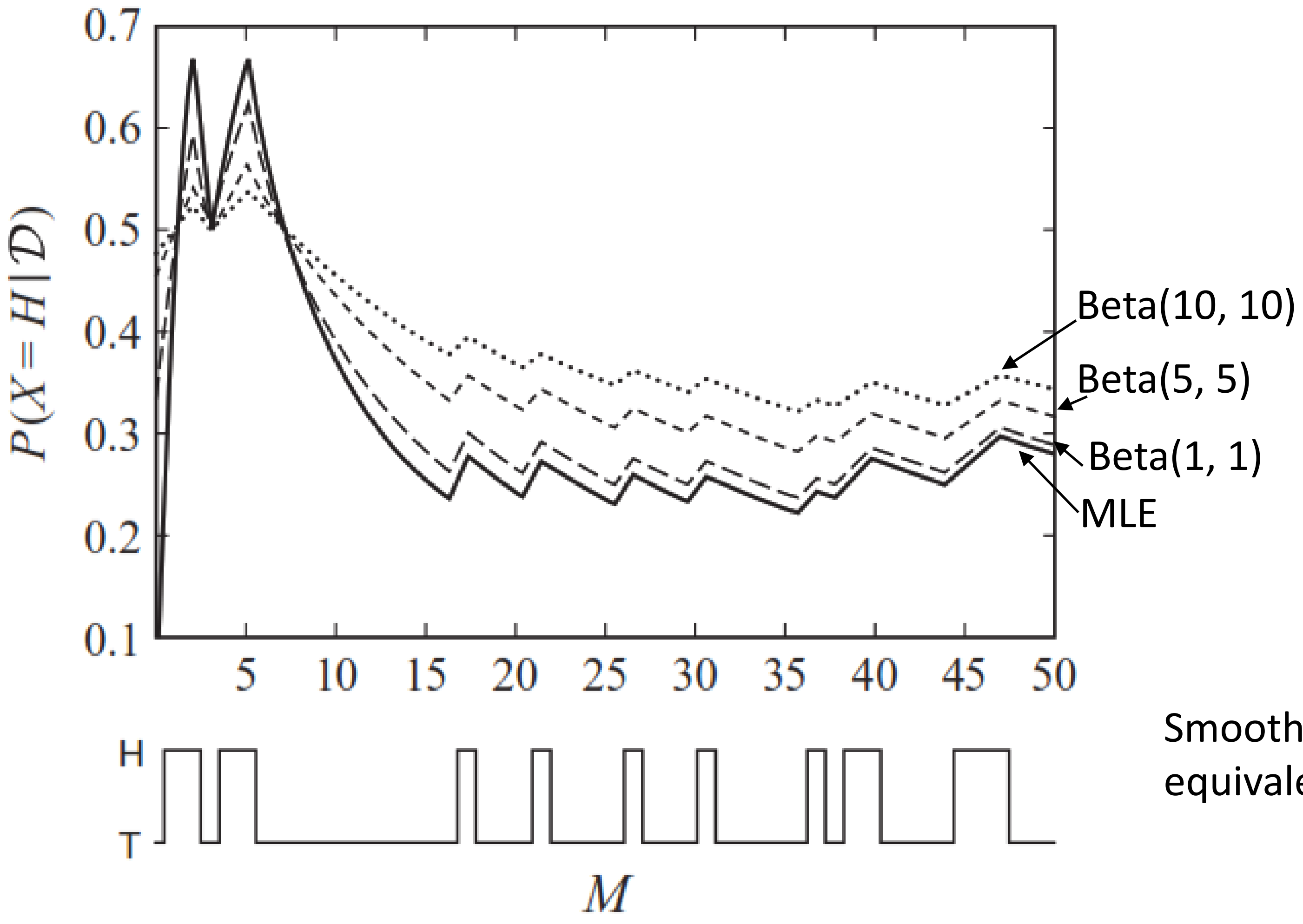
- We can rewrite the prediction as

$$P(x[M+1] = x^k \mid \mathcal{D}) = \underbrace{\frac{\alpha}{M+\alpha}\theta'_k}_{\text{Prior mean}} + \frac{M}{M+\alpha} \cdot \underbrace{\frac{M[k]}{M}}_{\text{ML estimate}}$$

where $\theta'_k = \frac{\alpha_k}{\alpha}$

- The prediction is a weighted average of the prior mean and the MLE.
- The weights are determined by α – the confidence of the prior- and M – the number of observed samples.

The effect of different priors



Sequence of tosses:

Smoother estimate with higher equivalent sample size