

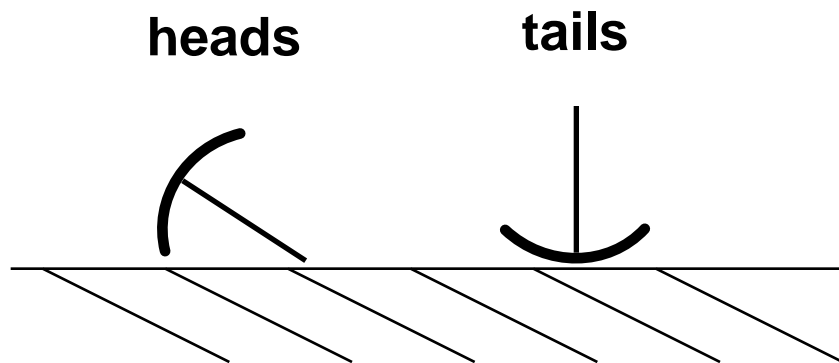
Basics of Statistical Estimation

Probabilistic Models for ML

- Joint Distribution can answer queries
 - $P(\text{Play Tennis}, \text{Weather})$ can be used to predict whether Aldo plays tennis based on the weather
- But:
 - **Where do the probabilities come from (learning)?**
 - How do we represent a joint compactly using conditional independencies? (representation – Bayes Nets)

Learning Probabilities: Classical Approach

Simplest case: Flipping a thumbtack



True probability θ is unknown

Given: flips generated independently with the same θ ,
(a.k.a. Independent and identically distributed data - iid),
Estimate: θ

Maximum Likelihood Principle

Choose the parameters that maximize the probability of the observed data

Maximum Likelihood Estimation

$$p(\text{heads} \mid \theta) = \theta$$

$$p(\text{tails} \mid \theta) = (1 - \theta)$$

$$p(\text{hhth...tth} \mid \theta) = \theta^{\#h} (1 - \theta)^{\#t}$$

(Number of heads is binomial distribution)

Computing the ML Estimate

- Use log-likelihood
- Differentiate with respect to parameter(s)
- Equate to zero and solve
- Solution:

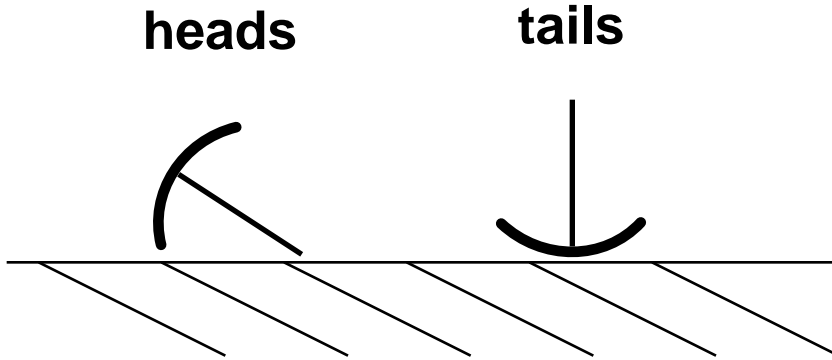
$$\theta = \frac{\#h}{\#h + \#t}$$

Sufficient Statistics

$$p(hhth\dots ttth | \theta) = \theta^{\#h} (1 - \theta)^{\#t}$$

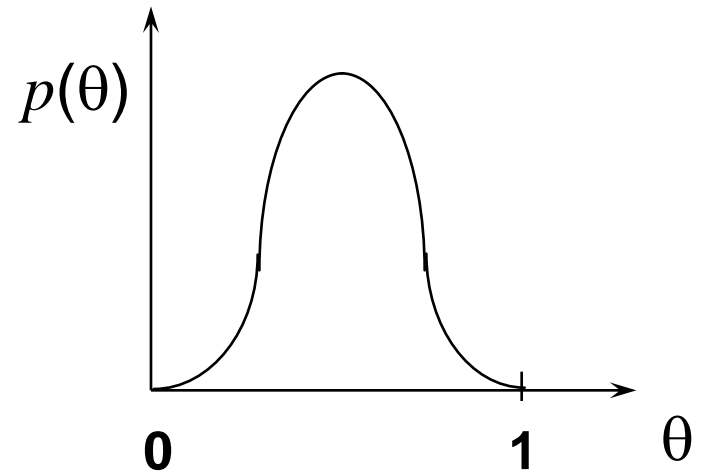
(#h,#t) are sufficient statistics

Bayesian Estimation



True probability θ is unknown

Bayesian probability density for θ



Use of Bayes' Theorem

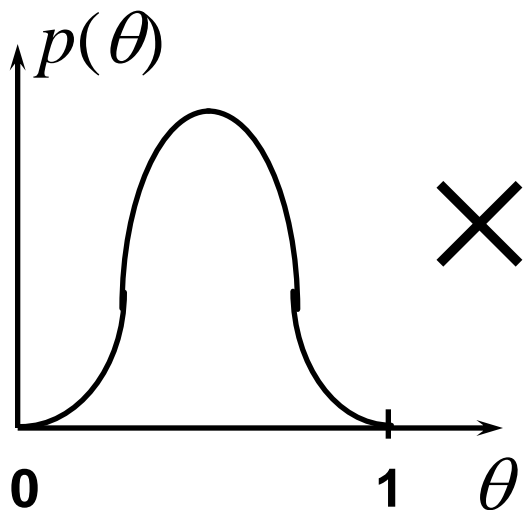
posterior

prior

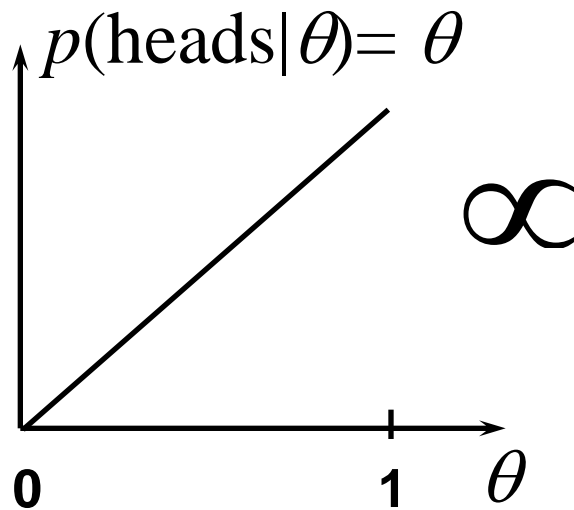
likelihood

$$p(\theta | \text{heads}) = \frac{p(\theta) p(\text{heads} | \theta)}{\int p(\theta) p(\text{heads} | \theta) d\theta}$$
$$\propto p(\theta) p(\text{heads} | \theta)$$

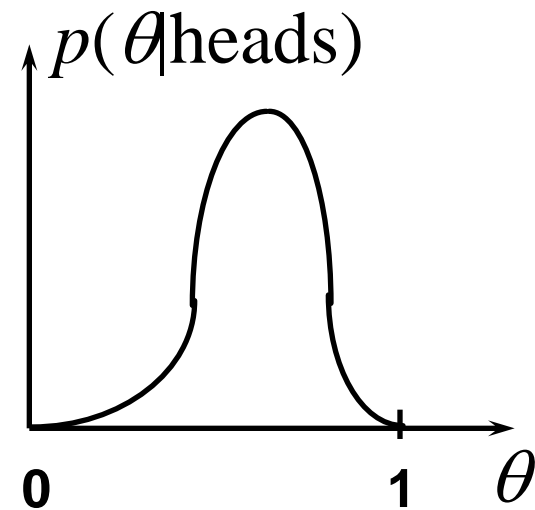
Example: Application to Observation of Single "Heads"



prior



likelihood



posterior

Probability of Heads on Next Toss

$$\begin{aligned} p(n + 1\text{th toss is } h \mid \mathbf{d}) &= \int p(X_{N+1} = h \mid \theta) p(\theta \mid \mathbf{d}) d\theta \\ &= \int \theta p(\theta \mid \mathbf{d}) d\theta \\ &= E_{p(\theta \mid \mathbf{d})}(\theta) \end{aligned}$$

MAP Estimation

- Approximation:
 - Instead of averaging over all parameter values
 - Consider only the **most probable value** (i.e., value with highest posterior probability)
- Usually a very good approximation, and much simpler
- MAP value \neq Expected value
- MAP \rightarrow ML for infinite data (as long as prior $\neq 0$ everywhere)

Prior Distributions for θ

- Direct assessment
- Parametric distributions
 - Conjugate distributions
(for convenience)
 - Mixtures of conjugate distributions

Conjugate Family of Distributions

Beta distribution:

$$p(\theta) = \text{Beta}(\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1} (1 - \theta)^{\alpha_t - 1} \quad \alpha_h, \alpha_t > 0$$

Resulting posterior distribution:

$$p(\theta \mid h \text{ heads}, t \text{ tails}) \propto \theta^{\#h + \alpha_h - 1} (1 - \theta)^{\#t + \alpha_t - 1}$$

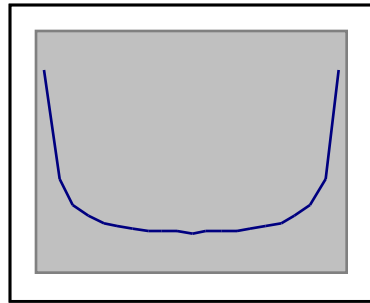
Estimates Compared

- Prior prediction: $E(\theta) = \frac{\alpha_h}{\alpha_h + \alpha_t}$
- Bayesian posterior prediction $E(\theta) = \frac{\#h + \alpha_h}{\#h + \alpha_h + \#t + \alpha_t}$
- MAP estimate: $\theta = \frac{\#h + \alpha_h - 1}{\#h + \alpha_h - 1 + \#t + \alpha_t - 1}$
- ML estimate: $\theta = \frac{\#h}{\#h + \#t}$

Intuition

- The hyperparameters α_h and α_t can be thought of as imaginary counts from our prior experience, starting from "pure ignorance"
- Equivalent sample size = $\alpha_h + \alpha_t$
- The larger the equivalent sample size, the more confident we are about the true probability

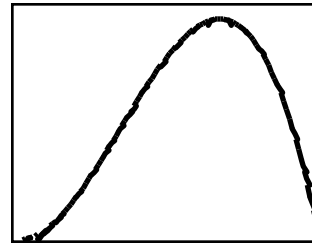
Beta Distributions



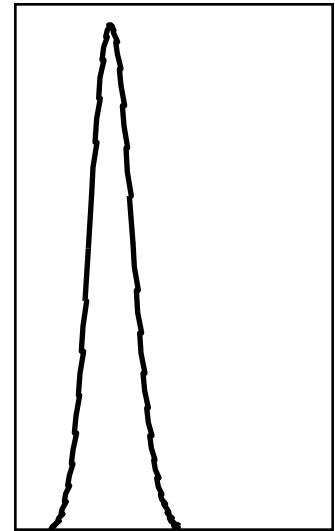
Beta(0.5, 0.5)



Beta(1, 1)



Beta(3, 2)



Beta(19, 39)

Assessment of a Beta Distribution

Method 1: Equivalent sample

- assess α_h and α_t
- assess $\alpha_h + \alpha_t$ and $\alpha_h / (\alpha_h + \alpha_t)$

Method 2: Imagined future samples

$$p(\text{heads}) = 0.2 \text{ and } p(\text{heads} \mid 3 \text{ heads}) = 0.5 \Rightarrow \alpha_h = 1, \alpha_t = 4$$

$$\text{check: } 0.2 = \frac{1}{1+4}, \quad 0.5 = \frac{1+3}{1+3+4}$$

Generalization to m Outcomes (Multinomial Distribution)

Dirichlet distribution:

$$p(\theta_1, \dots, \theta_m) = \text{Dirichlet}(\alpha_1, \dots, \alpha_m) \propto \prod_{i=1}^m \theta_i^{\alpha_i - 1}$$

$$\sum_{i=1}^m \theta_i = 1 \quad \alpha_i > 0$$

Properties:

$$E(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^m \alpha_i}$$

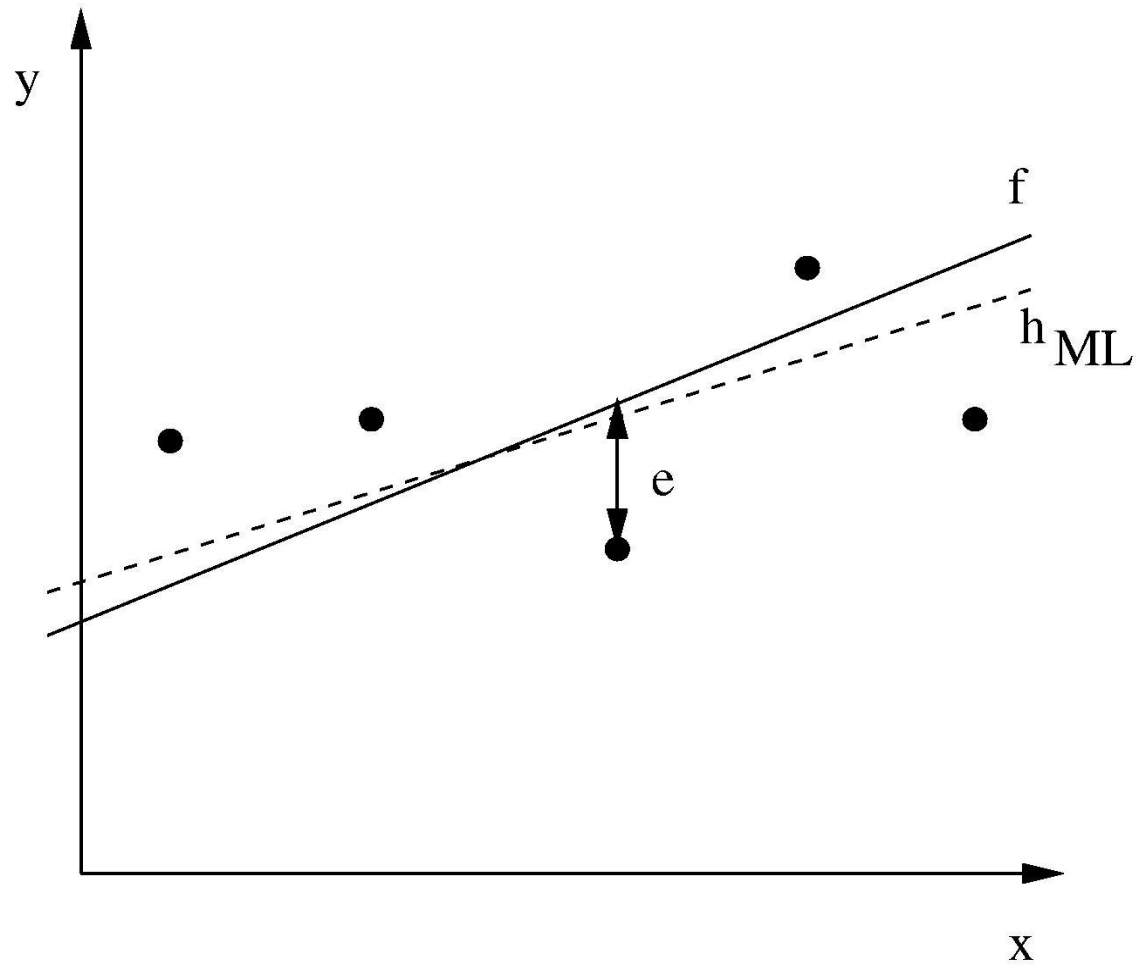
$$p(\theta | N_1, \dots, N_m) \propto \prod_{i=1}^m \theta_i^{\alpha_i + N_i - 1}$$

Other Distributions

Likelihoods from the exponential family

- Binomial
- Multinomial
- Poisson
- Gamma
- Normal

Learning a Real-Valued Function



Consider any real-valued target function f

Training examples $\langle x_i, d_i \rangle$, where d_i is noisy training value

- $d_i = f(x_i) + e_i$
- e_i is random variable (noise) drawn independently for each x_i according to some Gaussian distribution with mean=0

Then the maximum likelihood hypothesis h_{ML} is the one that minimizes the sum of squared errors:

$$h_{ML} = \arg \min_{h \in H} \sum_{i=1}^m (d_i - h(x_i))^2$$

Maximum likelihood hypothesis:

$$\begin{aligned} h_{ML} &= \operatorname{argmax}_{h \in H} p(D|h) = \operatorname{argmax}_{h \in H} \prod_{i=1}^m p(d_i|h) \\ &= \operatorname{argmax}_{h \in H} \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{d_i - h(x_i)}{\sigma} \right)^2} \end{aligned}$$

Maximize natural log of this instead ...

$$\begin{aligned}h_{ML} &= \operatorname{argmax}_{h \in H} \sum_{i=1}^m \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \left(\frac{d_i - h(x_i)}{\sigma} \right)^2 \\&= \operatorname{argmax}_{h \in H} \sum_{i=1}^m -\frac{1}{2} \left(\frac{d_i - h(x_i)}{\sigma} \right)^2 \\&= \operatorname{argmax}_{h \in H} \sum_{i=1}^m -(d_i - h(x_i))^2 \\&= \operatorname{argmin}_{h \in H} \sum_{i=1}^m (d_i - h(x_i))^2\end{aligned}$$