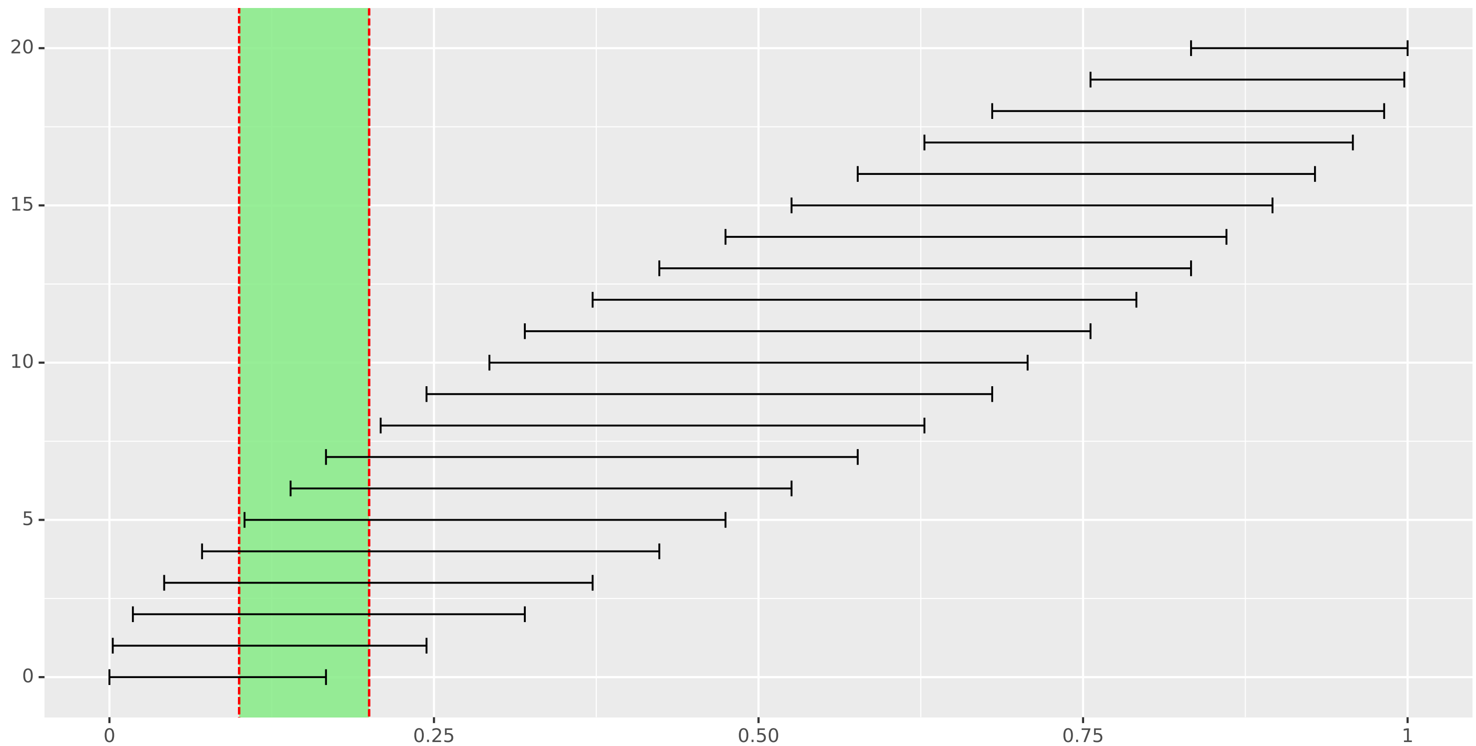
LTAT.02.004 Machine Learning II

Bayesian methods

Sven Laur
University of Tartu

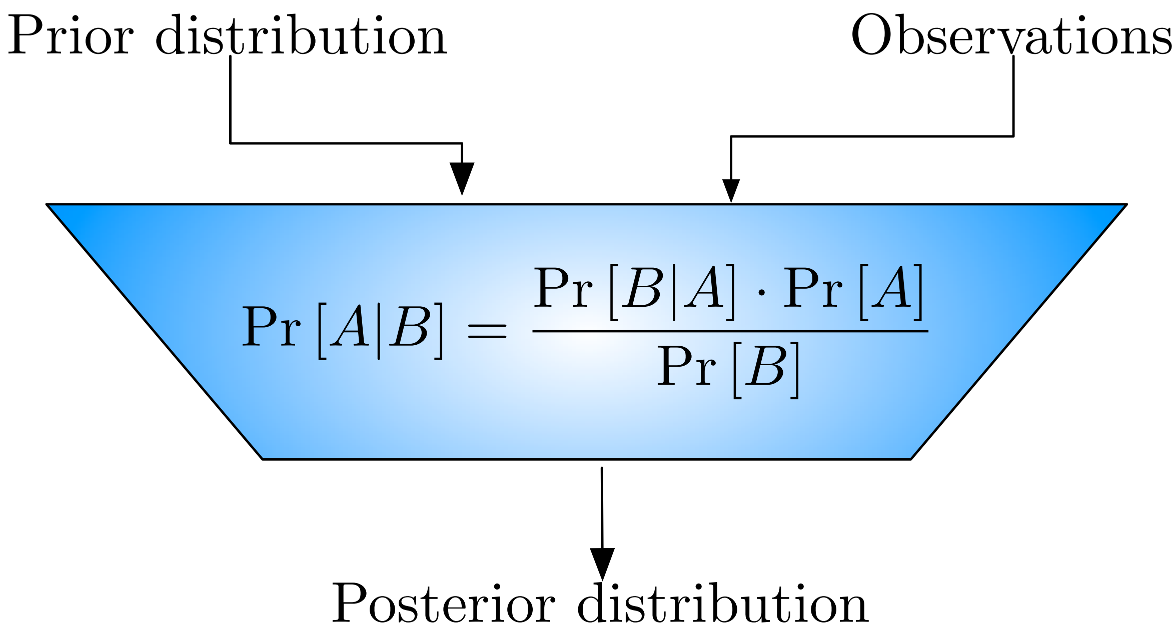
Bayesian methods

Confidence intervals vs background knowledge



- ▷ Confidence intervals do not capture background knowledge $p \in [0.1, 0.2]$.
- ▷ Thus we must accept absurd or suboptimal parameter estimations.

Bayesian inference procedure



- ▷ Prior distribution $\Pr[A]$ encodes the background knowledge
- ▷ The model $\Pr[B|A]$ determines how the posterior $\Pr[A|B]$ is updated

Prior and likelihood

Likelihood $\mathcal{L}(\mathcal{D}|\mathcal{M})$ is a probability of observations $\mathcal{D}$ when the data generation model $\mathcal{M}$ is fixed. The model is fixed by the set of parameters.

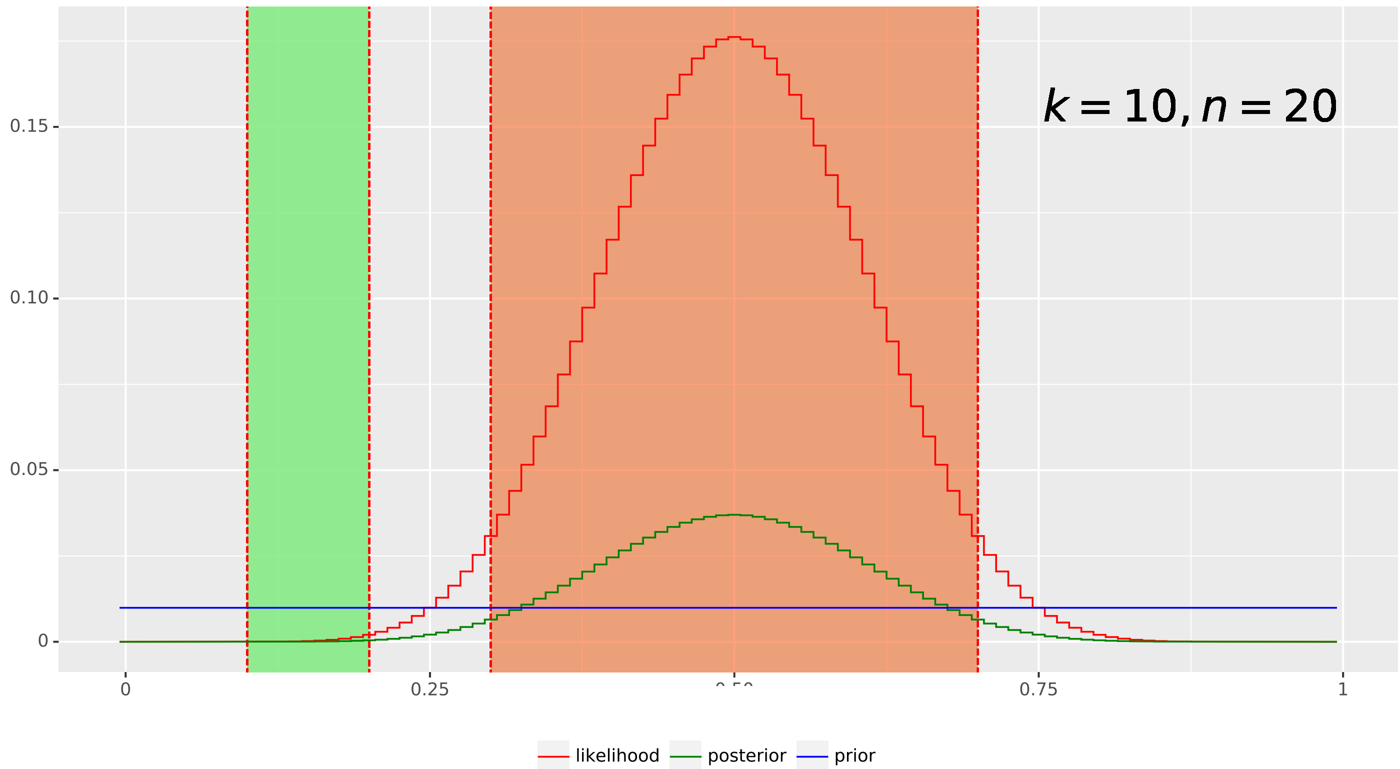
For coin flipping experiment the number of ones k is the observation and the coin bias p is the model paramater and thus

$$\mathcal{L}[k|p] = \binom{n}{k} p^k (1-p)^{n-k}$$

Prior is a distribution over models that encodes our preferences of models before we observe any data.

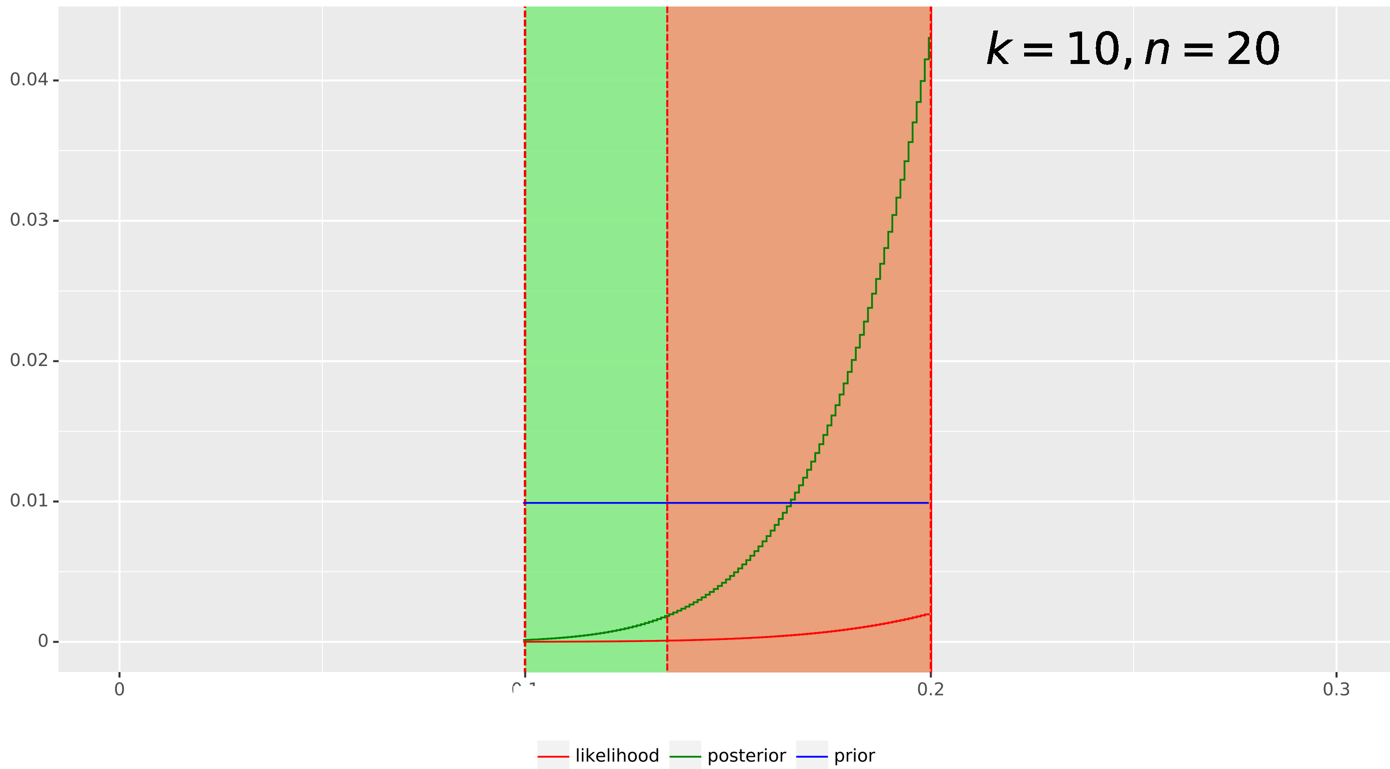
- ▷ Uninformative prior assigns uniform probability to all models.
- ▷ Uninformative prior is not well-defined for continuous parameters.

Posterior of an uninformed person



- ▷ With no preferences the posterior is concentrated around 0.5.
- ▷ Credibility interval $p \in [0.3, 0.7]$ contains 95% of posterior probability.

Posterior of an informed person



- ▷ With preferences the posterior is concentrated to the left of 0.2.
- ▷ Credibility interval $p \in [0.135, 0.2]$ contains 95% of posterior probability.

Beta distribution as a posterior

By increasing the number of grid points in the non-informative prior we reach a continuous distribution with a density function

$$p[p|k] = \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot p^k(1-p)^{n-k} \ .$$

This distribution is known as *beta distribution* Beta$(\alpha = k+1, \beta = n-k+1)$. The parameter value that maximises the posterior is

$$p_* = \frac{\alpha - 1}{\beta - \alpha} = \frac{k}{n} \ .$$

Maximum likelihood principle

If I have no background information to prefer one model to another then

$$\Pr[\mathcal{M}_i] = const$$

and thus

$$\Pr[\mathcal{M}_i|\mathcal{D}] = const \cdot \Pr[(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)|\mathcal{M}_i]$$

As a result I should choose a model that maximises *likelihood*

$$\Pr[(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)|\mathcal{M}_i]$$

The same principle is also applicable if the number of models is infinite.

Maximum a posteriori principle

Sometimes, we have extra background knowledge that makes some models more likely than the others:

$$\Pr[\mathcal{M}_i] \neq const$$

Then the model with largest likelihood is suboptimal choice and we should take a model with highest posterior probability

$$\Pr[\mathcal{M}_i|\mathcal{D}] \to \max \quad .$$

This method is known as *maximum a posteriori principle*.

In most cases, MAP estimates are defined so that they are *numerically and statistically more stable* than ML estimates.

Dice throwing vs coin flipping

A behaviour of a dice with faces $\{1, \ldots, m\}$ is determined by probabilities

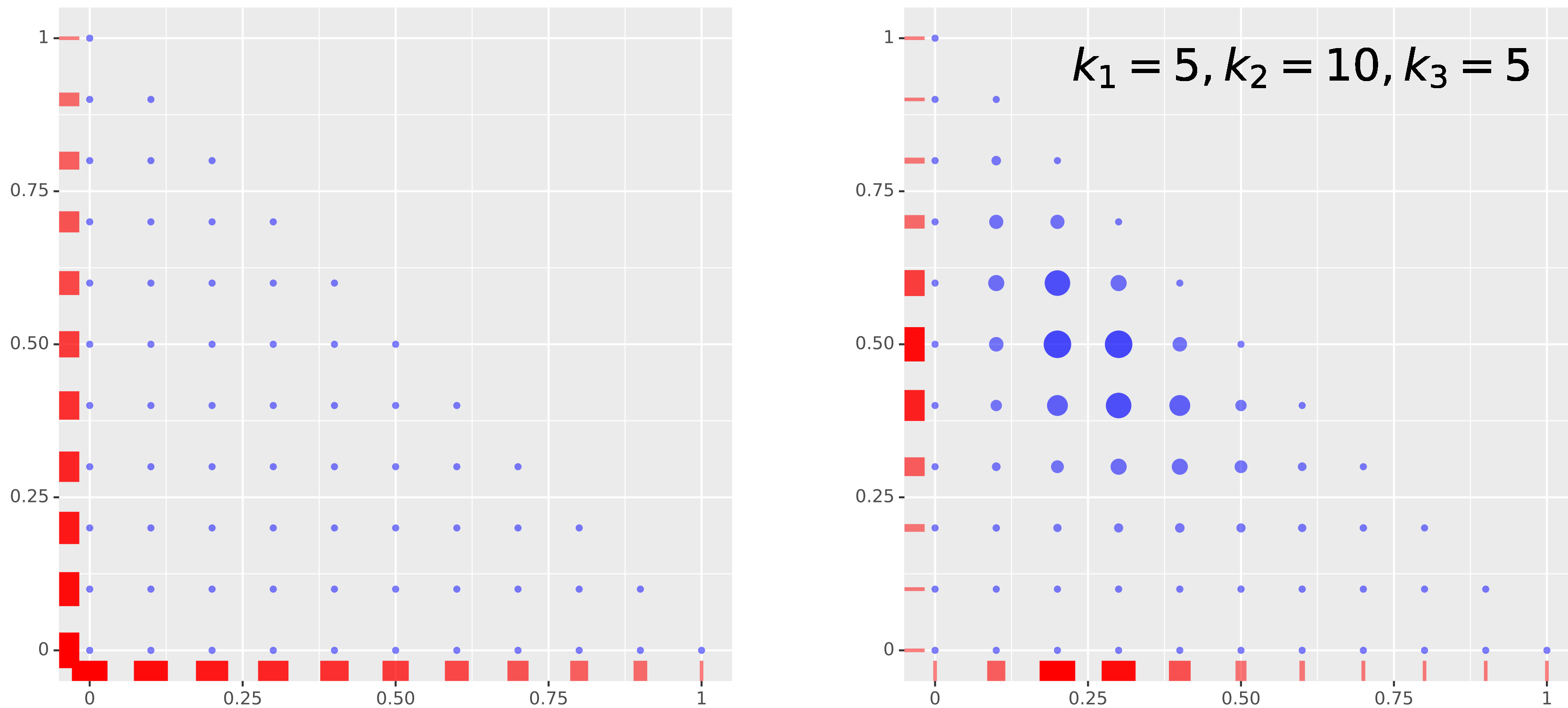
$$p_1 = \Pr[D_i = 1], \quad \ldots, \quad p_m = \Pr[D_i = m]$$

Reduction to coin flipping

- Let B_i denote the event that $D_i = j$.
- Then $B_1, \ldots, B_n$ is a coinflipping sequence with bias $\Pr[B_i = 1] = p_j$.
- ~~Non-informative prior for dice throwing goes to the non-informative prior.~~
- Informative priors can be marginalised to the right format.
- The same reduction can be done for all faces of the dice.

Caution: Marginal posteriors do not determine the full posterior in general.

Illustration



- ▷ Uniform prior over parameter pairs yields non-uniform marginal priors.
- ▷ The joint MAP estimate coincides with the marginal MAP estimates.

Dirichlet distribution as a posterior

By increasing the number of grid points in the non-informative prior over simplex we reach a continuous distribution with a density function

$$p[p_1, \ldots, p_m | k_1, \ldots, k_m] = \frac{\Gamma(n+m)}{\Gamma(k_1+1) \cdots \Gamma(k_m+1)} \cdot p_1^{k_1} \cdots p_m^{k_m} \ .$$

This distribution is known as *Dirichlet distribution*

$$\text{Dirichlet}(\alpha_1 = k_1 + 1, \ldots, \alpha_m = k_m + 1) \ .$$

The parameter value that maximises the posterior is

$$p_i^* = \frac{\alpha_i - 1}{\alpha_1 + \ldots \alpha_m - m} = \frac{k_i}{n} \ .$$

Laplace smoothing

Assume that we throw a dice with m faces and B_i encodes the event that the dice lands on a specific face. Then it is natural to assign the maximum prior probability to the parameter value $p_* = \frac{1}{m}$.

Such prior can be defined through a following though experiment:

- ▷ We start with non-informative prior.
- ▷ We observe all possible outcomes of the dice α times.
- ▷ We use the resulting posterior as a prior for real observations.

Thus the posterior can be obtained by starting with non-informative prior and observing $k + \alpha$ ones among $n + m\alpha$ throws.

- ▷ The ratio $p = \frac{k+\alpha}{n+m\alpha}$ is the maximal aposteriori estimate for p.