

PROBABILISTIC MODELS – PART 5a: PARAMETER LEARNING IN BAYESIAN NETWORKS

Gerhard Widmer

Institute of Computational Perception
Johannes Kepler University
Linz, Austria

gerhard.widmer@jku.at
www.cp.jku.at/people/widmer



October 28, 2025

Presentation partly based on and inspired by [Koller & Friedman, 2009] and [Russell & Norvig, 2021], including the use of some figures from their books and/or lecture slides.

Many thanks to Daphne Koller, Nir Friedman, Stuart Russell, and Peter Norvig for making these available
(pgm.stanford.edu; aima.cs.berkeley.edu).

Do not distribute!

Goals of this Lecture

- ▶ Discuss why we may need to estimate parameters for given structures
- ▶ Recapitulate likelihood as an objective function
- ▶ Show how to compute the maximum likelihood (ML) parameters for a given model structure from training data
- ▶ Discuss problems involved in this
- ▶ Explain idea of Bayesian parameter estimation as an alternative to ML estimation
- ▶ Demonstrate Bayesian estimation for binary variables
- ▶ Briefly discuss the fully Bayesian approach: Bayesian model averaging

Outline

1 Motivation and Task

2 Maximum-Likelihood Estimation

ML Estimation for Unconditional Distributions

ML Estimation for Conditional Distributions

Likelihood Decomposition in Bayesian Networks

The Data Fragmentation Problem

Estimating Parameters of Gaussian Distributions

3 Bayesian Parameter Estimation

Motivation and Idea

Bayesian Estimation for a Binary Variable

The Full Approach: Bayesian Model Averaging

Motivation

Problem addressed in this chapter:

- ▶ Given a fixed dependency graph structure of a Bayesian network,
- ▶ learn the *parameters* (the complete CPDs) from a set of example events.

Motivation

- ▶ Sketching the structure of dependencies is relatively easy in many domains,
- ▶ but getting precise numerical parameters from experts may be much harder.
- ▶ May be easier to collect a large number of observations (example events).

Parameter learning is also a sub-problem in

- ▶ Structure learning (☞ Part 5.b)
- ▶ Learning from incomplete observations (☞ Part 6.a)

Task

Given:

- ▶ The graph structure $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ of a Bayesian network \mathcal{M} , specifying the dependencies (edges \mathcal{E}) over a given set of random variables \mathcal{X}
- ▶ A training set $\mathcal{D} = \{x_1, \dots, x_M\}$ of M samples from the world (i.e., from the true underlying distribution P^*)
- ▶ Assumption: Training instances x_i are *independent and identically distributed (i.i.d.)*

Find

- ▶ A complete set of parameter values $\hat{\theta}$ for the model (i.e., a complete set of conditional probabilities for the CPDs) such that resulting model $\mathcal{M}(\mathcal{G}, \hat{\theta})$ maximises some objective function $F(\mathcal{M}, \mathcal{D})$
- ▶ $\hat{\theta}$ can be thought of as one list/vector of numbers.

Two Families of Methods:

- ① Maximum-Likelihood (ML) Estimation
- ② Bayesian Parameter Estimation

Maximum Likelihood Parameter Estimation for Bayesian Networks

The Task:

Maximum Likelihood Parameter Estimation

Given:

- ▶ the structure \mathcal{G} of a Bayesian Network \mathcal{M}
- ▶ a data set $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$

Find:

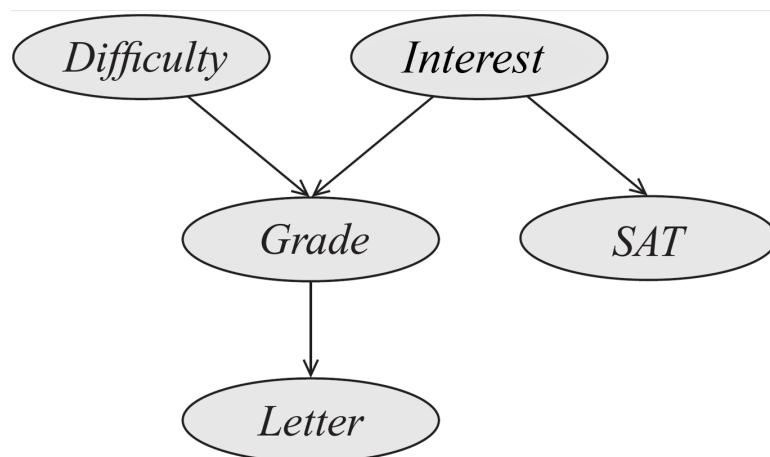
- ▶ parameters $\hat{\theta}$ such that

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\mathcal{M}_\theta : \mathcal{D}) = \arg \max_{\theta \in \Theta} P(\mathcal{D} | \mathcal{M}_\theta)$$

In words: Given a training set of observations and a fixed network structure, find numbers $\hat{\theta}$ for the CPDs (out of the infinite set Θ of possible parametrisations) such that the resulting model maximises the probability of the observed data.

An Example

Graph Structure \mathcal{G} :



Training Set \mathcal{D} :

$(d^1, i^0, g^2, s^1, l^0)$
$(d^1, i^1, g^1, s^0, l^1)$
$(d^0, i^0, g^3, s^0, l^1)$
$(d^1, i^0, g^1, s^1, l^1)$
$(d^1, i^0, g^2, s^1, l^0)$
$(d^0, i^1, g^1, s^0, l^1)$
$(d^0, i^0, g^1, s^0, l^1)$
$(d^0, i^0, g^2, s^1, l^0)$
$(.. \quad .. \quad .. \quad .. \quad ..)$
$(.. \quad .. \quad .. \quad .. \quad ..)$

- ▶ **Find:** Parameters (conditional probabilities) $\hat{\theta}$ for tables $P(D), P(I), P(G|D, I), P(S|I), P(L|G)$ such that

$$L(\mathcal{M}_{\boldsymbol{\theta}} : \mathcal{D}) = \prod_{\mathbf{x}_i \in \mathcal{D}} P_{\mathcal{M}_{\boldsymbol{\theta}}}(\mathbf{x}_i)$$

becomes maximal.

Maximum Likelihood Parameter Estimation for Bayesian Networks

Overview of the following slides:

- ▶ First: A simple example: How to compute the ML estimate for the parameter of a binary (Bernoulli) distribution
- ▶ Computing the ML estimate for general categorical distributions
- ▶ Computing the ML estimate for *conditional* distributions
- ▶ Global and local decomposability of the likelihood function in Bayesian networks
- ▶ General algorithm for ML parameter estimation in Bayesian networks
(will turn out to be very simple ...)

MLE for an Unconditional Distribution: A Simple Example

The story:

Consider a thumbtack that is tossed N times, landing heads or tails certain numbers of times. Can you infer the thumbtack's general behaviour from this (i.e., the probability distribution that would best explain the actual observations)?



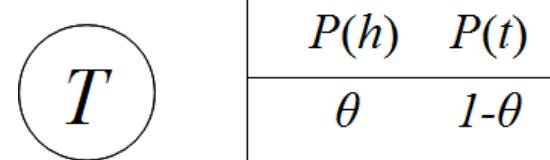
Note:

- ▶ Unlike a (fair) coin¹, a thumbtack is probably not going to display a uniform 50:50 distribution over heads and tails.

¹ Surprise: natural coins are *not* fair. A flipped coin is about 51% likely to land facing the same way up as it started, due to the physics of the problem. For a proof and experiments, see → Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical Bias in the Coin Toss. *SIAM Review* 49(2), 211-235.



The Story as a Graphical Model



Notes:

- ▶ The model consists of one binary variable only ($T = \text{'Toss'}$)
- ▶ The distribution over the two values h (heads) and t (tails) is determined by a single parameter θ :

$$\theta = P(\text{heads})$$

$P(\text{tails})$ is just $1 - \theta$

- ▶ This is known as a **Bernoulli Distribution**
- ▶ Our task: After N tosses with observed results (k heads), estimate θ in a maximum likelihood way.

A Simple Example

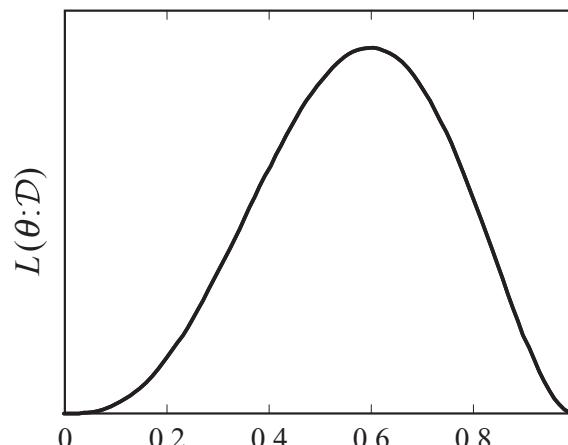
Assume the following sequence of observations:

- ▶ $N = 5$ tosses with result *heads, tails, tails, heads, heads*
- ▶ $\mathcal{D} = \{h, t, t, h, h\}$

The Likelihood Function for this data sequence:

$$\begin{aligned}
 \underline{L(\theta : \mathcal{D})} &= P(h | \theta) \cdot P(t | \theta) \cdot P(t | \theta) \cdot P(h | \theta) \cdot P(h | \theta) \\
 &= \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot \theta \\
 &= \underline{\theta^3(1 - \theta)^2} \quad [= \theta^k(1 - \theta)^{N-k}]
 \end{aligned}$$

Likelihood $L(\theta : \mathcal{D})$ over all possible values of θ :



A Simple Example

Our Goal: Find

$$\hat{\theta} = \arg \max_{\theta} L(\theta : \mathcal{D})$$

Note:

- ▶ This likelihood function $L(\theta : \mathcal{D})$ has exactly one maximum
 - ▶ Also true for the logarithm of L
 - ▶ Let's maximise the log-likelihood $\ell(\theta : \mathcal{D})$ instead:

$$\ell(\theta : \mathcal{D}) = \log [\theta^3 \cdot (1 - \theta)^2] = 3 \log \theta + 2 \log(1 - \theta)$$

Finding the arg max of $\ell(\theta : \mathcal{D})$: Calculate derivative

$$\frac{\partial}{\partial \theta} \ell(\theta : \mathcal{D}) = \frac{\partial}{\partial \theta} [3 \log \theta + 2 \log(1 - \theta)] = \frac{3}{\theta} - \frac{2}{1 - \theta}$$

... set to 0, and solve for θ :

$$\begin{aligned} \frac{3}{\theta} - \frac{2}{1-\theta} &= 0 \\ \Rightarrow \hat{\theta} &= 3/5 = 0.6 \end{aligned}$$

In words: Given the 5 observations above, the most likely value for θ is 0.6.

ML Estimates for Discrete Distributions

Given a set of observations $\mathcal{D} = \{x_1, \dots, x_M\}$ of a discrete variable X , the **Maximum Likelihood Estimate** for the parameter(s) of a

- ▶ **Bernoulli Distribution** over a binary variable $X = \{x^0, x^1\}$ is^a

$$\hat{\theta} = \hat{P}(x^0) = \frac{N[x^0]}{N[x^0] + N[x^1]} = \frac{N[x^0]}{|\mathcal{D}|}$$

where $N[x^i]$ is the number of occurrences of x^i in \mathcal{D} .

^aEasy to prove (computations analogous to above).

- ▶ **General Categorical Distribution** over a discrete variable

$X = \{x^0, \dots, x^k\}$ is^a

$$\hat{\theta}_i = \hat{P}(x^i) = \frac{N[x^i]}{\sum_j N[x^j]} = \frac{N[x^i]}{|\mathcal{D}|}$$

^aSlightly more difficult to prove, because $\sum_i \theta^i$ must be 1.0 (need *Lagrange multipliers* ...)



ML Estimates for Discrete Distributions

$$\hat{\theta}_i = \hat{P}(x^i) = \frac{N[x^i]}{|\mathcal{D}|}$$

In Words:

- ▶ The maximum likelihood estimate for a discrete probability is the proportion of occurrences of the relevant event in the observed data.
- ▶ Trivial?
- ▶ Have proven *in what sense* relative frequencies are optimal probability estimators: *Maximum Likelihood (ML)*
- ▶ but will see below that there are other possible (and meaningful) answers ...

ML Estimates for Conditional Distributions

Obvious extension to estimating **conditional probabilities**:

Given a set of observations $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ over a set of discrete variables \mathcal{X} , the Maximum Likelihood Estimate for the parameters of a **conditional distribution** $P(X | \mathbf{y})$ is

$$\hat{\theta}_{x^i | \mathbf{y}} = \hat{P}(x^i | \mathbf{y}) = \frac{N[x^i, \mathbf{y}]}{N[\mathbf{y}]}$$

where

- ▶ $N[x^i, \mathbf{y}]$ is the number of events in \mathcal{D} with $X = x^i$ and $\mathbf{Y} = \mathbf{y}$
- ▶ $N[\mathbf{y}] = \sum_j N[x^j, \mathbf{y}]$ is the number of events in \mathcal{D} with $\mathbf{Y} = \mathbf{y}$.

In Words:

- ▶ Estimate a conditional distribution $P(X | \mathbf{y})$ for specific parent values \mathbf{y} by collecting all the instances from \mathcal{D} that match \mathbf{y} , and counting the frequencies of the values x^i in this subset.
- ▶ This is what we have been doing all along in Ch.4b (Approximate Inference)

...

Full Problem: ML Parameter Estimation for a Complete Network

Given:

- ▶ Given a dataset $\mathcal{D} = \{x_1, \dots, x_M\}$
- ▶ and the graph structure \mathcal{G} of a Bayesian network

Estimate:

- ▶ Complete set of parameters $\hat{\theta}$ such that the resulting model $\mathcal{M} = (\mathcal{G}, \hat{\theta})$ has the highest likelihood w.r.t. \mathcal{D}

Central Question:

- ▶ How to estimate this large set of parameters such that they ***together*** give a Bayesian network (= full joint distribution over \mathcal{X}) with maximum likelihood?

Short Answer:²

- ▶ The set of parameters that maximise the likelihood of the complete model is identical to the parameters that individually maximise, in \mathcal{D} , the likelihood of each variable, given its parents (*Decomposition of the Likelihood Function*).
- 👉 We can solve the Maximum Likelihood problem for each variable separately!

²The proof is rather trivial and too boring for us here ...

Summary: The Complete Algorithm

ML Parameter Estimation for Discrete Bayesian Networks

Given:

- ▶ Graph structure \mathcal{G} of a Bayesian Network \mathcal{M} over discrete variables \mathcal{X}
- ▶ Data set $\mathcal{D} = \{x_1, \dots, x_M\}$

Find:

- ▶ Parameters $\hat{\theta} = \arg \max_{\theta} L(\mathcal{M}_{\theta} : \mathcal{D})$

Algorithm:

- ▶ For each variable X in \mathcal{G} , with its parents U :
 - for each possible assignment of values u to U :
 - estimate the parameters $\theta_{X|u}$ (i.e., one row in X 's CPD) as

$$\hat{\theta}_{X|u} = \hat{P}(X \mid u) = \frac{N[X, u]}{N[u]}$$

that is, by counting, for each value x of X , how often parent values u co-occur with $X = x$ in training set \mathcal{D} .

The Data Fragmentation Problem

$$\hat{\theta}_{x|u} = \hat{P}(x | u) = \frac{N[x, u]}{N[u]}$$

Problem:

- ▶ Estimate of $\hat{\theta}_{x|u}$ is based on $N[u]$ examples $\in \mathcal{D}$
- ▶ As number of parents U grows, number of different parent value combinations u grows exponentially
- ▶ No. of instances matching u in a fixed data set \mathcal{D} **shrinks exponentially**
- ▶ Will have very few (or no) instances for estimating parameters
- ▶ Will have large numbers of **unspecified distributions** (where $N[u] = 0$) or **zeroes** (where $N[x, u] = 0$) in the CPDs
 \Rightarrow very brittle behaviour and poor generalisation of the model.

Consequence:

- ▶ Important: keep number of parents per variable as small as possible!
- ▶ Avoid zeroes by **smoothing** the ML estimates.

Common Smoothing Method: Laplace Correction

Laplace Correction

Calculate smoothed estimate as

$$\hat{\theta}_{x|\mathbf{u}} = \hat{P}(x | \mathbf{u}) = \frac{N[x, \mathbf{u}] + \alpha}{N[\mathbf{u}] + \alpha k}$$

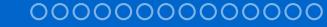
where $k = |\text{Val}(X)|$

and $\alpha > 0$ is a smoothing parameter.

Example: $\alpha = 1$

$$\hat{P}(x | \mathbf{u}) = \frac{N[x, \mathbf{u}] + 1}{N[\mathbf{u}] + k}$$

- ▶ Avoids zero estimates
- ▶ Slightly decreases high and increases low estimates
- ▶ Interesting interpretation as ‘pseudocounts’ (‘virtual observations’)
- ▶ Can be understood as the result of Bayesian estimation with a uniform prior and with posterior mean selection (see later)
- ▶ Look up ‘*Laplace’s Rule of Succession*’ and ‘*sunrise problem*’ on the Web ...



Estimating the Parameters of Gaussian Distributions

Remember from Part 3:

- ▶ In worlds with real-valued variables, CPDs become continuous conditional density functions
- ▶ Convenient (but restricted) model: the **Linear Gaussian Model**
- ▶ Will also become important in *Kalman Filters* (see  Part 6b).



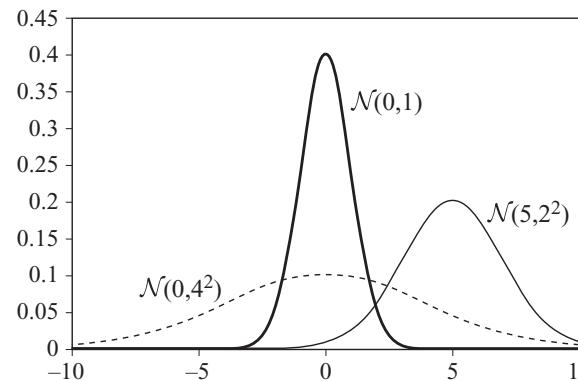
Must be able to estimate parameters for Gaussian Distributions!

Remember: The Normal (Gaussian) Distribution

Definition

A variable X has a **Normal Distribution** with mean μ and variance σ^2 , denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has the **Gaussian PDF**

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Task:

- ▶ Given a dataset $\mathcal{D} = \{x_1, \dots, x_M\}$ assumed to be normally distributed
- ▶ Compute the Maximum Likelihood Estimates for parameters $\theta = \{\mu, \sigma^2\}$.

Computing the ML Estimate for the Parameters

Maximisation of log-likelihood analogous to calculation in discrete case:

- ① Write down the (log-)likelihood function:

$$\begin{aligned} L(\boldsymbol{\theta} : \mathcal{D}) = p(\mathcal{D} | \boldsymbol{\theta}) &= \prod_{i=1}^M p(x_i | \boldsymbol{\theta}) = \prod_{i=1}^M \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ \ell(\boldsymbol{\theta} : \mathcal{D}) = \log p(\mathcal{D} | \boldsymbol{\theta}) &= \sum_{i=1}^M \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ &= M(-\log \sqrt{2\pi} - \log \sigma) - \sum_{i=1}^M \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

- ② Compute the derivative w.r.t. the parameters and set to 0:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^M (x_i - \mu) = 0 \quad \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma} = -\frac{M}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^M (x_i - \mu)^2 = 0$$

- ③ Solve the system of 2 equations:

$$\hat{\mu} = \frac{\sum_i x_i}{M} \quad \hat{\sigma} = \sqrt{\frac{\sum_i (x_i - \hat{\mu})^2}{M}}$$

Computing the ML Estimate for the Parameters

$$\hat{\mu} = \frac{\sum_i x_i}{M} \quad \hat{\sigma} = \sqrt{\frac{\sum_i (x_i - \hat{\mu})^2}{M}}$$

In Words:

- ▶ The ML estimate for the mean μ is the sample mean (average)
- ▶ The ML estimate for the standard deviation σ is the square root of the sample variance

Again:

- ▶ Not very surprising (?)
- ▶ But now we have proven *in what sense* these $\hat{\mu}$ and $\hat{\sigma}$ are optimal estimators
- ▶ This kind of calculation will be needed to estimate the parameters of linear Gaussian CPDs.

Remember: The Multivariate Gaussian

Definition

N random variables $\mathbf{X} = \{X_1, \dots, X_N\}$ have a **Joint (Multivariate) Normal Distribution** with mean vector μ and covariance Σ , denoted $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, if they have the **Multivariate Gaussian PDF**

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

where

- ▶ μ is the *mean vector*
- ▶ Σ is the $N \times N$ *covariance matrix*
- ▶ $|\Sigma|$ is the determinant of Σ

Task:

- ▶ Given a multidimensional dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ assumed to have a joint normal distribution
- ▶ Compute the Maximum Likelihood Estimates for parameters $\theta = \{\mu, \Sigma\}$.

Computing the ML Estimate for the Parameters

By calculations similar to the above:

- ① Write down the (log-)likelihood function:

$$L(\boldsymbol{\theta} : \mathcal{D}) = p(\mathcal{D} | \boldsymbol{\theta}) = \prod_{i=1}^M p(x_i) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

$$\ell(\boldsymbol{\theta} : \mathcal{D}) = \log p(\mathcal{D} | \boldsymbol{\theta}) = \dots$$

- ② Compute the derivative w.r.t. the parameters and set to 0:

... abracadabra ...

- ③ Search maximum: Solve the system of 2 equations:

$\hat{\boldsymbol{\mu}} = \frac{1}{M} \sum_i \mathbf{x}_i$	$\hat{\boldsymbol{\Sigma}} = \frac{1}{M} \sum_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$
--	---

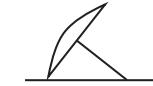
So now you know how to estimate Gaussians from data.

Will turn out useful later ...

Problems with the ML Estimate

Consider 3 different thumbtack experiments:

- ▶ **Experiment 1:** 10 random tosses. Result: 3 heads, 7 tails
- ▶ **Experiment 2:** 10 random tosses **with a coin** instead of the thumbtack. Result: 3 heads, 7 tails
- ▶ **Experiment 3:** 1, 000, 000 tosses **with a coin** instead of the thumbtack. Result: 300, 000 heads, 700, 000 tails.



Intuitively:

- ▶ **In case 1:** we find ML estimate $\hat{\theta} = 0.3$ ‘plausible’.
- ▶ **In case 2:** Given our prior knowledge about regular coins, and the small number of observations, we tend to believe that the 10 tosses gave an untypical result. We are inclined to believe that $\theta \approx 0.5$ is more likely.
- ▶ **In case 3:** Given this very strong empirical evidence, we may now be willing to believe that this is a trick coin, and that $\theta = 0.3$ after all ...



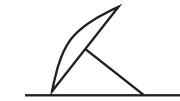
Problems with the ML Estimate

Problem with Maximum Likelihood Estimation:

- ▶ Does not distinguish between these situations
- ▶ Ignores the amount of evidence (and the resulting *uncertainty* of the estimate), and possible prior expectations
- ▶ Bases its decision on likelihood $P(\mathcal{D} | \theta)$
- ▶ and returns only a single estimate $\hat{\theta}$ (“point estimate”, “best guess”)

Here's a particularly problematic case:

- ▶ **Experiment 4:** 10 random tosses. Result: 0 heads, 10 tails



Consequence:

- ▶ ML estimate $\hat{\theta} = 0.0$
- ▶ In other words: “heads is impossible.”
- ▶ But: would you really want to bet all your money on “tails” for the next toss?
- ▶ ... or do you still have a certain ‘degree of belief’ that heads *might* be possible after all?



Bayesian Parameter Estimation

Bayesian Parameter Estimation:

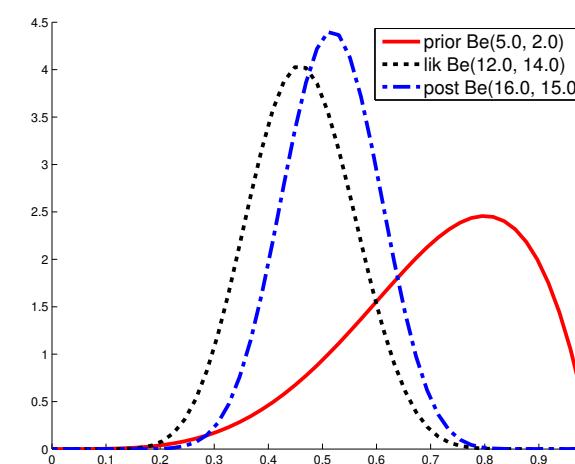
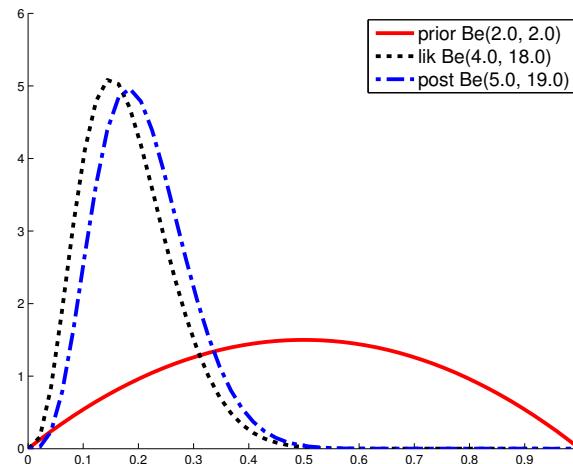
- ▶ Does not rely on a single parameter estimate $\hat{\theta}$, but learns a **probability distribution** over all possible parameter values
- ▶ In words: all parameter values are possible, but with different probabilities
- ▶ Computes **Posterior Distribution** $p(\theta | \mathcal{D})$ over parameter θ , given evidence \mathcal{D} , via Bayesian reasoning:

$$p(\theta | \mathcal{D}) = \frac{P(\mathcal{D} | \theta)p(\theta)}{P(\mathcal{D})}$$

The posterior distribution combines

- ▶ **strength of evidence** (via *likelihood* $P(\mathcal{D} | \theta)$, which can be calculated for any given \mathcal{D} and θ), and
- ▶ **subjective expectations** about plausible parameter settings (via *prior distribution* $p(\theta)$, which you can define).
- ▶ $P(\mathcal{D})$ is hard to calculate (integral!), but not needed if we are only interested in the shape (and the max or mean) of the posterior.

Bayesian Parameter Estimation



Posterior distribution $p(\theta|\mathcal{D})$ as a compromise between **prior expectations** $p(\theta)$ and information provided by observations (**likelihood** $P(\mathcal{D}|\theta)$). Left: strong evidence (peaked likelihood) and non-specific expectations (broad prior) result in posterior that is close to the likelihood – the data wins. Right: Stronger opinion regarding plausible parameter values leads to posterior that is a compromise between prior expectations and what the data would consider most likely. (Figures from [Murphy, 2012])

Different options, if we need to choose one particular estimate $\hat{\theta}$:

- ▶ $\hat{\theta} = \arg \max_{\theta} P(\mathcal{D} | \theta)$ – the **Maximum Likelihood (ML)** estimate
- ▶ $\hat{\theta} = \arg \max_{\theta} p(\theta | \mathcal{D})$ – the **Maximum A Posteriori (MAP)** estimate
- ▶ $\hat{\theta} = \int_{\theta} \theta \cdot p(\theta | \mathcal{D}) d\theta$ – the **Posterior Mean** (*expected value of the posterior*)

Bayesian Parameter Estimation for a Binary Variable

Consider again the thumbtack:

- ▶ Two possible outcomes
- ▶ Best modeled with a Bernoulli distribution, one parameter $\theta = P(h)$
- ▶ Assume a sequence \mathcal{D} of N observations, k of which produced heads (h)



Likelihood Function:³

$$P(\mathcal{D} \mid \theta) = \theta^k (1 - \theta)^{N-k}$$

³See previous thumbtack example above.

Bayesian Parameter Estimation for a Binary Variable

Prior:

- ▶ Can choose a class of distributions that fits what we want to express
- ▶ Must be defined over interval $[0, 1]$ (range of θ)
- ▶ Ideally: should have a functional form that is ‘compatible’ with the form of the likelihood, so as to again give a posterior of the same form⁴
(This would also naturally permit *incremental belief updates*)
- ▶ Consequence: prior should be of the form

$$p(\theta) \propto \theta^{\gamma_1} (1 - \theta)^{\gamma_2}$$

for some parameters γ_1 and γ_2

- ▶ Then the posterior could be easily calculated by adding up the exponents:

$$\begin{aligned} p(\theta | \mathcal{D}) &\propto P(\mathcal{D} | \theta)p(\theta) = \overbrace{\theta^k (1 - \theta)^{N-k}}^{\text{likelihood}} \overbrace{\theta^{\gamma_1} (1 - \theta)^{\gamma_2}}^{\text{prior}} \\ &= \theta^{k+\gamma_1} (1 - \theta)^{N-k+\gamma_2} \end{aligned}$$

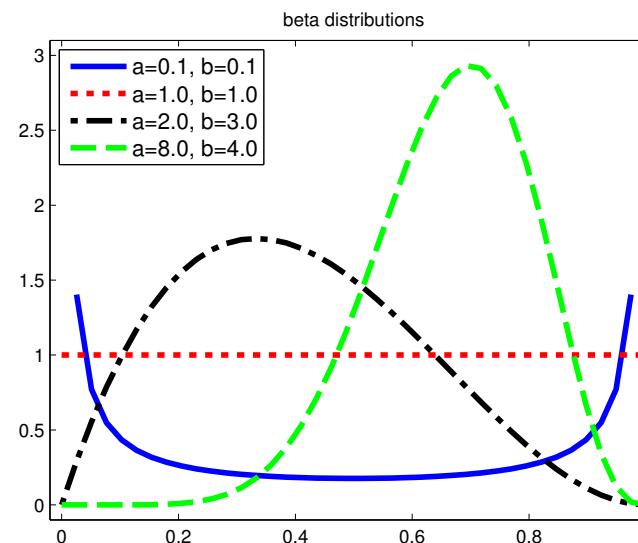
⁴When prior and posterior have the same form, we say that the prior is a **conjugate prior** for the given likelihood.

Bayesian Parameter Estimation for a Binary Variable

Conjugate Prior for the Bernoulli likelihood: the *Beta distribution*

$$\text{Beta}(\theta \mid a, b) \propto \theta^{a-1} (1 - \theta)^{b-1}$$

- ▶ Defined over $\theta \in [0, 1]$
- ▶ Parameters a, b determine the shape of the distribution
- ▶ Can use these ‘hyper-parameters’ to encode our prior belief about θ
- ▶ Special case: $a = b = 1$ gives a *uniform distribution* over $[0, 1]$





Bayesian Parameter Estimation for a Binary Variable

Posterior:

$$\begin{aligned}
 p(\theta | \mathcal{D}) &\propto \theta^k (1-\theta)^{N-k} \cdot \text{Beta}(\theta | a, b) \propto \theta^{k+a-1} (1-\theta)^{N-k+b-1} \\
 &= \text{Beta}(\theta | k+a, N-k+b)
 \end{aligned}$$

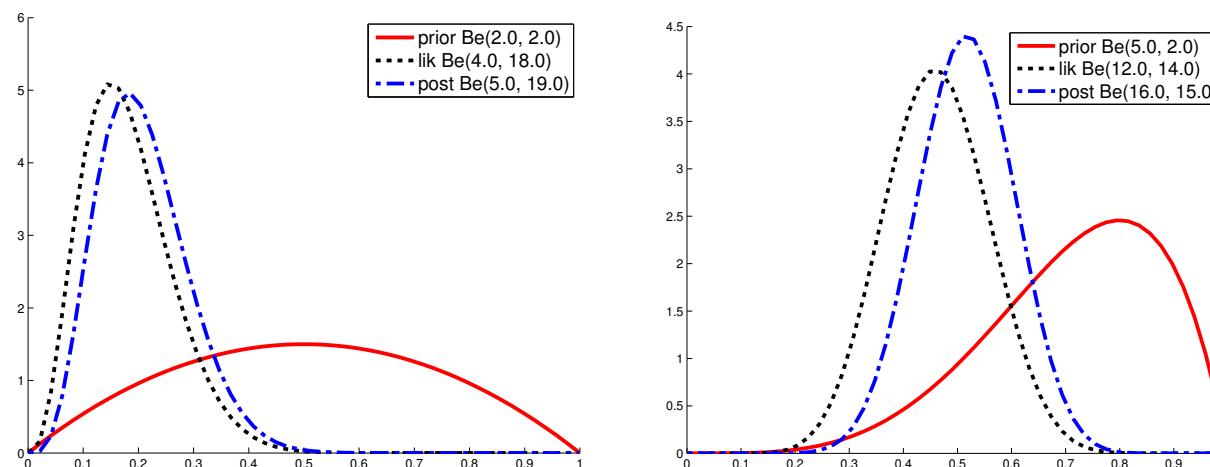


Figure : Left: Updating a $\text{Beta}(2, 2)$ prior with a Binomial likelihood with counts $k = 3, N - k = 17$ to yield a $\text{Beta}(5, 19)$ posterior. Right: Updating a $\text{Beta}(5, 2)$ prior with a Binomial likelihood with counts $k = 11, N - k = 13$ to yield a $\text{Beta}(16, 15)$ posterior.
(From [Murphy, 2012])

How to Choose a Final Single Estimate $\hat{\theta}$?

Obvious Choices:

- ▶ The **Maximum Likelihood (ML)** estimate = the **mode** of the **likelihood**.

In the thumbtack case:

$$\hat{\theta}_{ML} = \arg \max_{\theta} P(\mathcal{D} \mid \theta) = \frac{k}{N}$$

- ▶ The **Maximum A Posteriori (MAP)** estimate = the **mode** of the **posterior**.

In the thumbtack case:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\theta \mid \mathcal{D}) = \frac{k + a - 1}{N + a + b - 2}$$

- ▶ The **Posterior Mean** = the **expected value (mean)** of the **posterior**.

In the thumbtack case:

$$\hat{\theta}_{pmean} = \int_{\theta} \theta \cdot p(\theta \mid \mathcal{D}) d\theta = \frac{k + a}{N + a + b}$$

Special Case: Uniform Prior

Assume we know nothing about how thumbtacks behave:

- ▶ Consider all values of θ equally likely *a priori*
- ▶ Model this with a uniform prior: $p(\theta) = \text{Beta}(\theta | 1, 1)$

Resulting estimates for θ :

- ▶ *Maximum Likelihood (ML):*

$$\hat{\theta}_{ML} = \frac{k}{N}$$

- ▶ *Maximum A Posteriori (MAP):*

$$\hat{\theta}_{MAP} = \frac{k + a - 1}{N + a + b - 2} = \frac{k}{N}$$

With a uniform prior, this is equal to the ML estimate (logical ...)

- ▶ *Posterior Mean:*

$$\hat{\theta}_{pmean} = \frac{k + a}{N + a + b} = \frac{k + 1}{N + 2}$$

☞ Naturally leads to **Laplace Smoothing** (see earlier)!

To Summarise ...

In Bayesian parameter estimation, the probability estimate $\hat{\theta}$ we get depends on ...

- ▶ our prior assumptions (the prior $p(\theta)$, which may be uniform)
- ▶ the strength of the evidence (the likelihood $P(\mathcal{D} | \theta)$)
- ▶ and the way we choose our final estimate $\hat{\theta}$ from the posterior (*mode* ($\arg \max$) or *mean*)

But why force ourselves to rely on one single point estimate $\hat{\theta}$ at all?

- ▶ Let's consider all values of θ possible, with different probabilities!
- ▶ Let's use all of them for doing inference!
- ▶ That's what a true Bayesian will do ...
- ▶ Leads us to full *Bayesian Model Averaging*

Bayesian Model Averaging

The Basic Idea of Bayesian Model Averaging:

- ▶ Result of parameter estimation is not a single parameter hypothesis $\hat{\theta}$, but a probability distribution $p(\theta | \mathcal{D})$ over all possible θ (i.e., over possible models with a given graph structure \mathcal{G})
- ▶ Instead of selecting one specific parameter setting $\hat{\theta}$ for doing inference with our model (i.e., answer queries like $P(x | \mathcal{M}_{\hat{\theta}})$), we could use **all possible parameter settings** (weighted with their respective probabilities) to perform inference!

$$P(x | \mathcal{D}) = \int_{\theta} P(x | \mathcal{M}_{\theta}) p(\theta | \mathcal{D}) d\theta$$

- ▶ This is the “true Bayesian approach”
- ▶ ... but too complex to pursue further here.



What you should remember of this section

- ▶ Definition of the likelihood function
- ▶ How to compute the ML estimates for discrete distributions, and for Gaussians
- ▶ The complete procedure for learning the parameters of a BN
- ▶ The data fragmentation problem and the idea of smoothing
- ▶ The basic idea of Bayesian parameter estimation, and why it is important
- ▶ Concepts such as conjugate prior, MAP estimate, posterior mean.
(No need to remember the detailed formulas for the exam)

Literature

Koller, Daphne and Friedman, Nir (2009).

Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT Press.

Murphy, Kevin P. (2012).

Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press.