02477 – Bayesian Machine Learning: Lecture 1

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

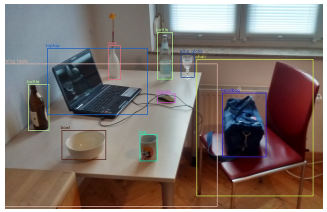
Outline for week 1

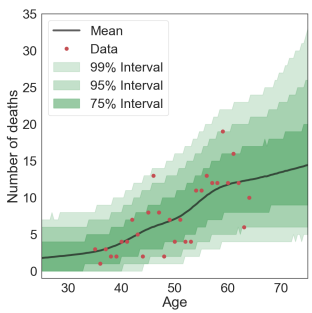
1. Introduction and course formalities
2. Bayesian machine learning
3. Brief recap on terminology from probability theory
4. The binomial model and maximum likelihood estimation
5. Bayesian inference and the Beta-binomial models
6. Introduction to the exercise sessions

Introduction and course formalities

Bayesian Machine Learning

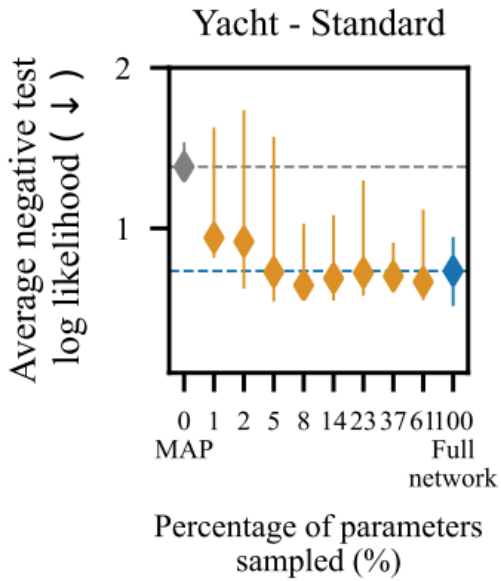
- Machine learning
  - Understanding and finding pattern in data
  - Making machines learn from data
  - Making predictions
- A multitude of applications
  - Object detection
  - Speech recognition and natural language processing
  - Self-driving cars
  - Spam & fraud detection
  - Recommender systems
  - Brain imaging
  - ...
- Bayesian statistics
  - Named after Thomas Bayes (1702 - 1761)
  - Mathematical framework for reasoning with uncertainty



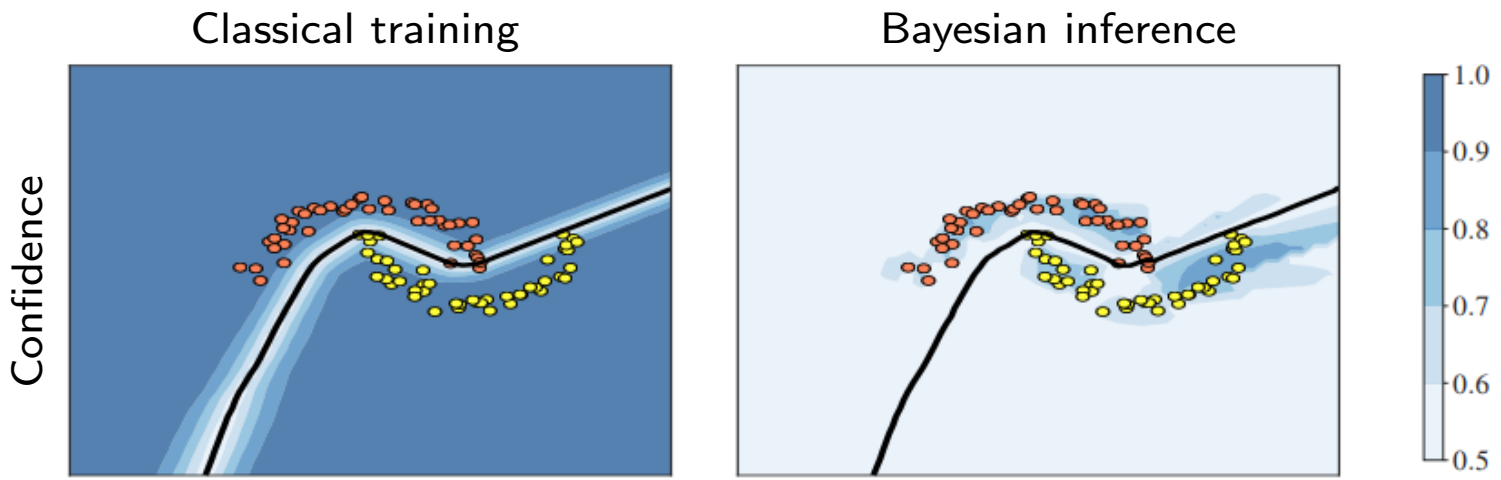


Motivation for studying Bayesian methods for machine learning

- Very flexible and intuitive modelling framework
- Can be less prone to overfitting
- Intuitive uncertainty quantification
- Can be computationally more intensive
- Deeper understanding of "traditional" methods
- Bayesian deep learning
  - Can improve generalization
  - May improve calibration and reduce overconfidence

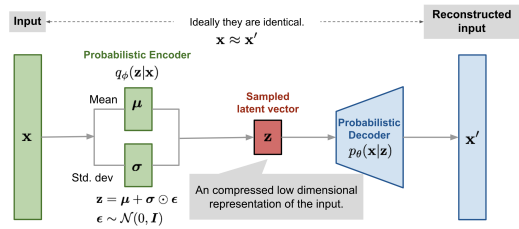


Kristiadi et al, 2019 and Sharma† et al, 2022

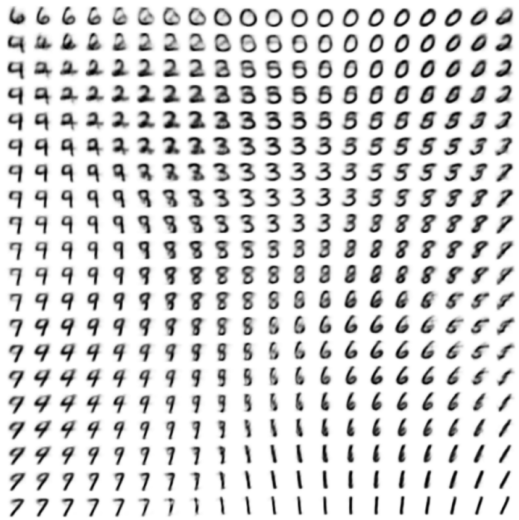


A probabilistic perspective on machine learning

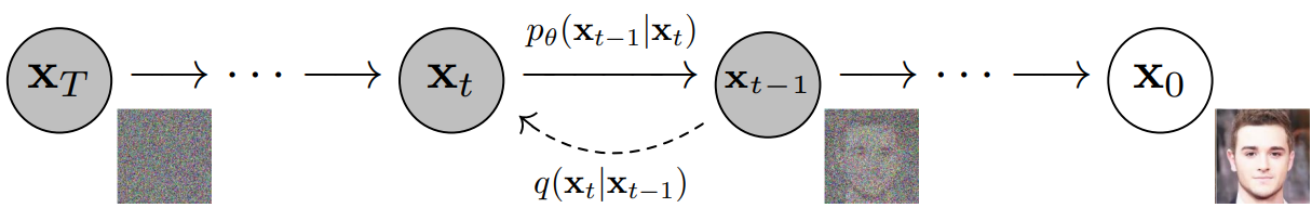
- We will study machine learning from a *probabilistic perspective*
- Interpreting classic methods in a probabilistic setting reveals a connection between loss functions (e.g. squared loss, cross entropy etc.) and probability distributions
- Gives *deeper insights into fundamentals* and enables us to *adapt methods to different types of data*
- The probabilistic framework is the *foundation for advanced machine learning* e.g. generative models such as variational autoencoders and diffusion models are typically explained using *variational methods*



https://lilianweng.github.io/posts/2018-08-12-vae/



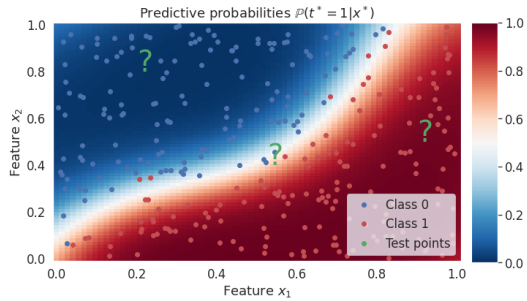
Kingma & Welling, 2014: Auto-encoding variational Bayes

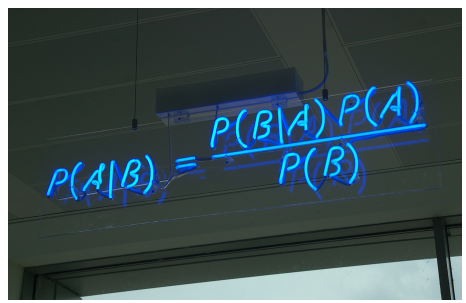


Ho et al, 2020: Denoising Diffusion Probabilistic Models

What to expect from this course?

- We will work with several machine learning problems
  - Regression
  - Classification
  - Clustering
  - Optimal decision-making
  - ...
- A probabilistic modelling approach
  - Probability distributions as Lego-blocks
  - Linear models, Gaussian processes, Neural networks
  - Relation to common loss functions (cross-entropy, MSE etc.)
- Inference methods
  - Maximum likelihood
  - Bayesian inference
  - Laplace approximations
  - Variational inference
  - Markov chain Monte Carlo methods
- Bayesian perspective: understanding how and why
  - Generalizes "classical" training
  - Incorporating prior knowledge into models
  - Uncertainty quantification
  - Deeper insights into fundamentals
  - Emphasis on bridging gap between theory and intuition
- Prepare you for a M.Sc. thesis in advanced machine learning





Course formalities

- Flipped classroom
  - Exercise sessions every Monday 13-17
  - Pen & paper exercises and programming exercises
  - Engage with teachers/teaching assistant highly encouraged
- Location and other practical information will be announced on DTU Learn ASAP
- Assignments
  - 3 mandatory assignments
  - Groups of 3-5
  - Feedback for student and teachers
- Written exam
  - Curriculum: All course materials, e.g. slides, book chapter, exercises, assignments etc.
  - Details T.B.A.
- Teachers and teaching assistants
  - Course responsible: Michael Riis Andersen (Building 321, room 216, miri@dtu.dk)
  - TAs T.B.A.

Course plan

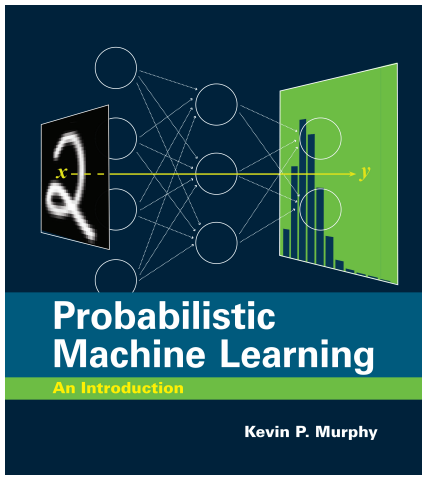
The plan is subject to change!

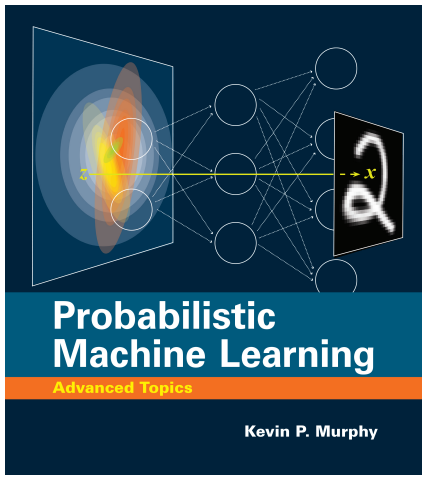
Week	Topic
1	Intro, basic concepts, Beta-Binomial model
2	Making predictions, grid approximations
3	Bayesian classification and Laplace approximations
4	Bayesian linear regression
5	Distributions on function spaces, Gaussian Processes
6	Gaussian process classification
7	Multi-class classification and decision theory
8	Monte Carlo & Markov Chain Monte Carlo methods
9	More on MCMC
10	Mixture models and variational inference
11	Black-box variational inference
12	More on variational inference
13	Bayesian neural networks

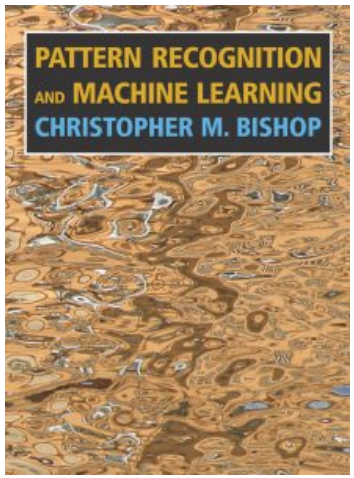
Course plan & list of reading material for each week can be on DTU Learn

Literature

- Main textbooks: Probabilistic Machine Learning by Kevin Murphy (vol. 1 & 2)
- Supplement: Pattern Recognition and machine learning by Christopher Bishop



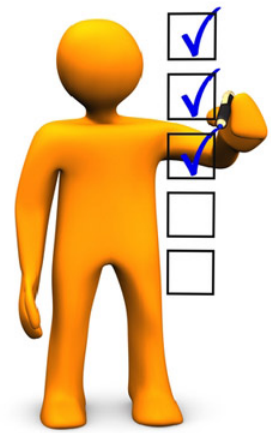




- All freely available as PDFs
  - https://probml.github.io/pml-book/book1.html
  - https://probml.github.io/pml-book/book2.html
  - https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/
- We will supplement with other resources in the second half of the course, e.g. book chapters, research papers etc

Course prerequisites

- 02450 Introduction to machine learning
- Math
  1. Linear algebra (Bishop appendix C)
  2. Calculus
  3. Probability theory (Bishop chap. 1, 2, appendix B)
  4. Statistics
- Programming
  - Python
  - Numpy, scipy, matplotlib etc
- It is indeed possible to complete the course without the all prerequisites, but you should *expect a significantly increased workload*



Continuous feedback

- I need your help to improve the course!
- Will announce feedback persons every week
- We meet on Monday at 16:45 (exact location TBA)
- Ideas for feedback
  - Were you able to follow the lecture?
  - How far did you make it in the exercise?
  - What was easy, what was difficult?
  - What did you like?
  - What can be improved?
  - etc etc.

Bayesian machine learning

Classical machine learning: supervised learning for regression

Common steps to fit a model to a given dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$

1. Choose a model

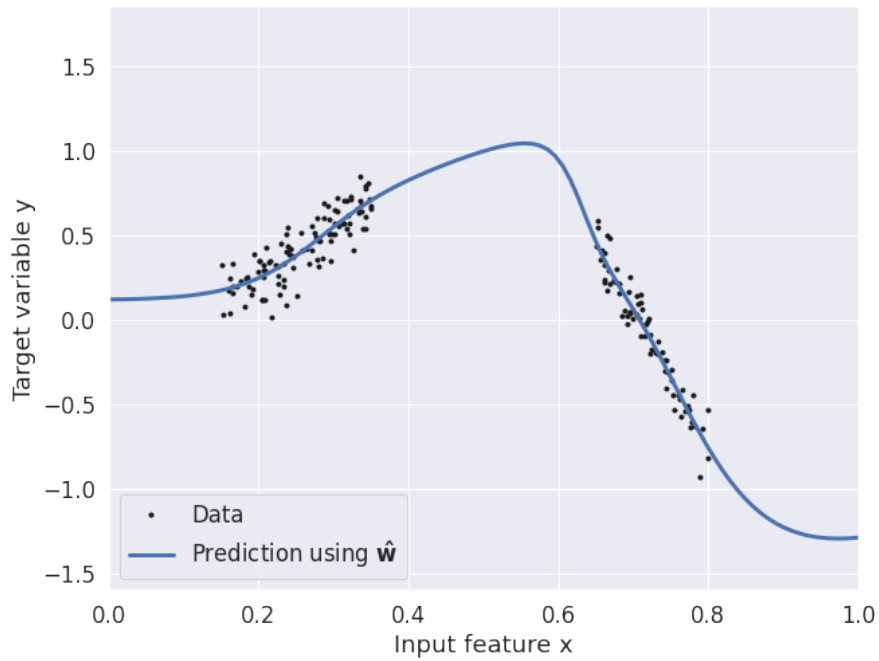
$$y_i = f(\boldsymbol{x}_i|\boldsymbol{w}) + e_i$$

2. Choose a loss function

$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

3. Find parameters $\boldsymbol{w}$ that minimizes the average loss $\mathcal{L}$ for the data set

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} [y - f(\boldsymbol{x}_i|\boldsymbol{w})]^2$$

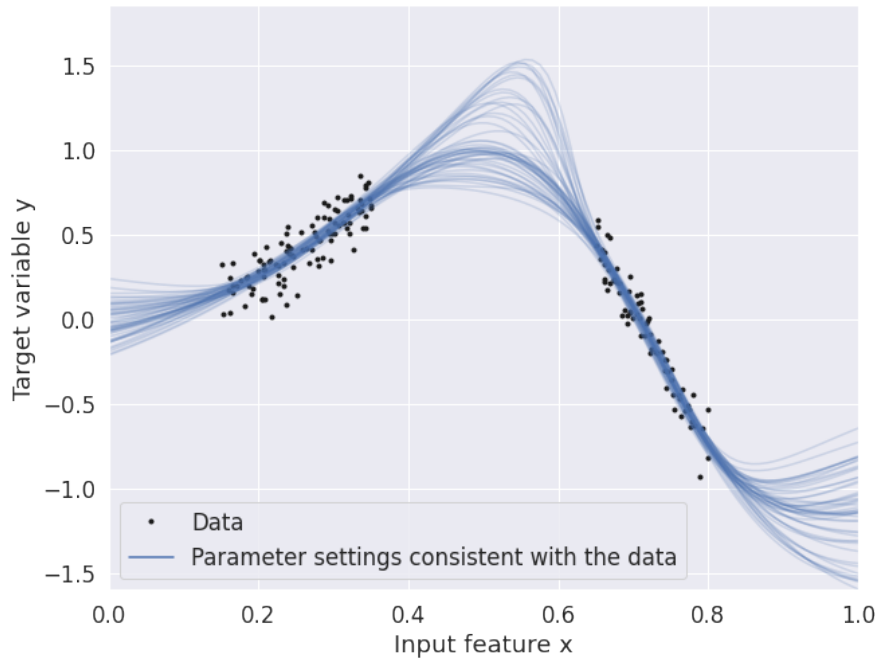


4. Make predictions for $\boldsymbol{x}^*$ using estimated parameters $\boldsymbol{w}$

$$y^* = f(\boldsymbol{x}^*|\hat{\boldsymbol{w}})$$

Model ambiguity due to finite data

- For a given model, there may be several sets of model parameters consistent with the data
- Model parameters $\hat{\boldsymbol{w}}_1$, $\hat{\boldsymbol{w}}_2$, and $\hat{\boldsymbol{w}}_3$ are all consistent with the data (as measured by training loss)
- Often many parameter settings consistent with data, but each can lead to very different predictions
- Classical machine learning: we choose *one* of these sets of parameters
- Bayesian machine learning: take *all* sets of model parameters consistent with the data into account



Bayesian inference and marginalization

- The *posterior distribution* $p(\boldsymbol{w}|\boldsymbol{y})$ measures how much weight to assign to each parameter set

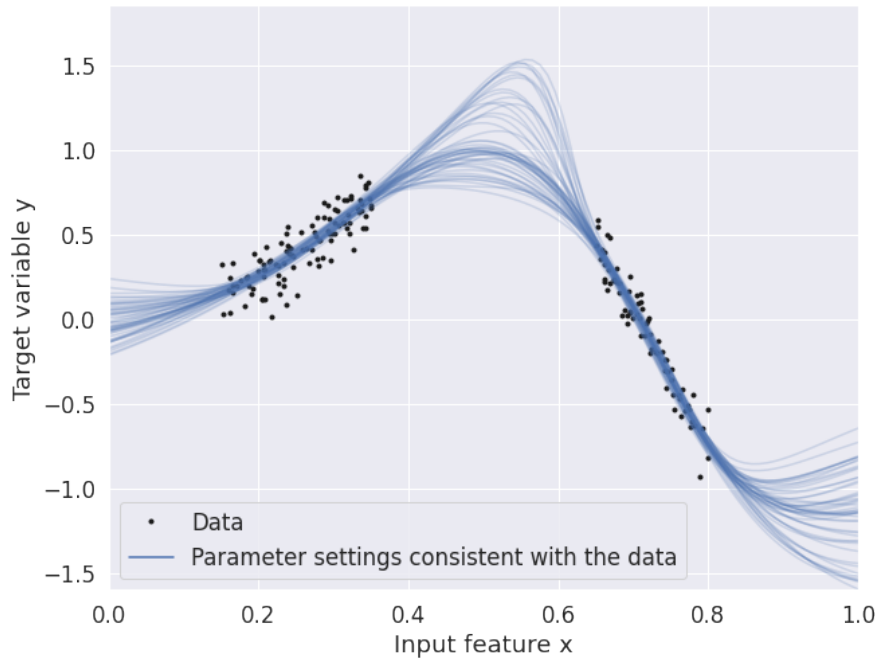
- Making predictions using a weighted average of all possible parameter sets

$$y^* = \sum_{i=1}^{M} f(\boldsymbol{x}^*|\boldsymbol{w}_i)p(\boldsymbol{w}_i|\boldsymbol{y}),$$

- Often, we have infinitely many parameter settings

$$y^* = \int f(\boldsymbol{x}^*|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{y})\mathrm{d}\boldsymbol{w}$$

- The process takes the uncertainty about the parameters into account and is called *marginalization*

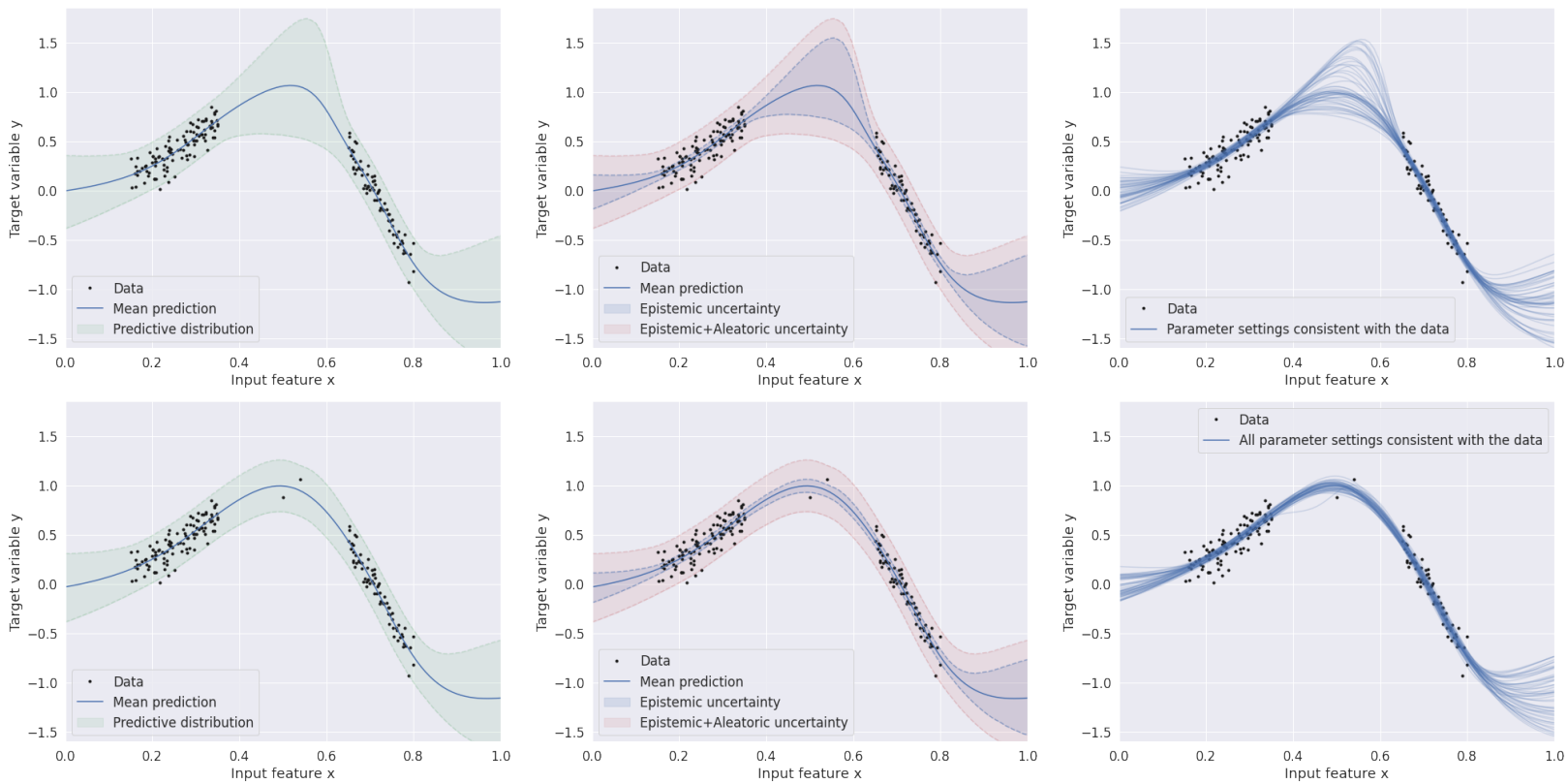


Uncertainty quantification

Two sources of uncertainty

1. *Epistemic uncertainty* is due to lack of knowledge (e.g. often due to a limited data set). Also sometimes called the *reducible* uncertainty.

2. *Aleatoric uncertainty* refers to the inherent randomness (e.g. measurement noise). Also sometimes called the *irreducible* uncertainty.



Bayesian machine learning

- In Bayesian methods, *all variables* (e.g. both parameters and data) are represented using *probability distributions*
- *Bayes' rule* provides a systematic way to combine data with prior knowledge

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- *Likelihood* $p(\boldsymbol{y}|\boldsymbol{w})$ - distribution of data $\boldsymbol{y}$ given a specific set of parameters $\boldsymbol{w}$
- *Prior* $p(\boldsymbol{w})$ - prior belief about parameters $\boldsymbol{w}$ before seeing any data
- *Posterior* $p(\boldsymbol{w}|\boldsymbol{y})$ - contains all knowledge about parameters after seeing data $\boldsymbol{y}$
- Why use priors?
  1. can encode domain knowledge
  2. can help prevent overfitting
  3. can generate artificial datasets from the model
- Why distributions rather than point estimates?
  1. Easy uncertainty quantification
  2. Avoid making predictions when uncertain is too large
  3. Better decision-making

Example: how can uncertainty improve decision making?

Example from Section 4.6.7.3 in Murphy1

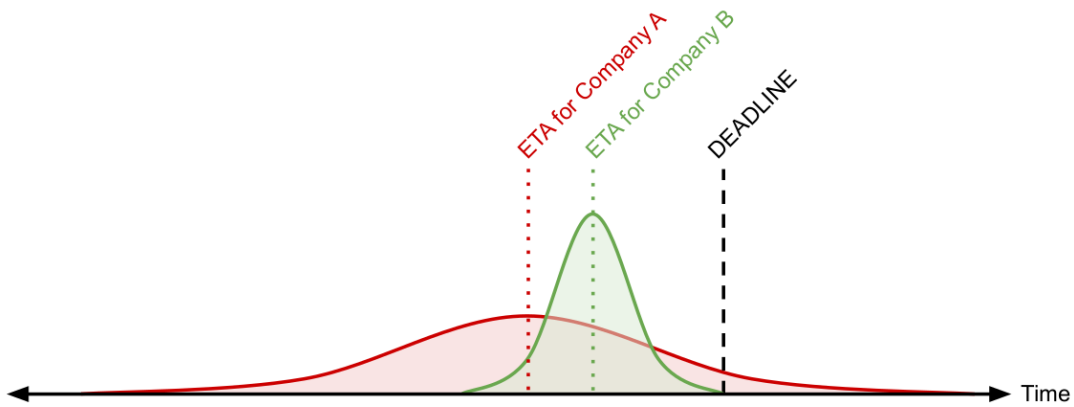


Figure 4.21: Distribution of arrival times for two different shipping companies. ETA is the expected time of arrival. A's distribution has greater uncertainty, and may be too risky. From https://bit.ly/39bc4XL. Used with kind permission of Brendan Hasz.

- Suppose you run a company that have promised to delivery a package to a customer before some deadline
- Goal: predict the *delivery time* for the package if shipped with company A rather than company B
- Based on previous shipments, we have estimated the mean and variance for the delivery times of the two company
- On *average company A is faster*, but due to the large variance we might not meet the deadline. Hence, *despite company B being slower on average, it might be the best option*.

Brief recap on terminology from probability theory

Probability notation and terminology

	Discrete distributions	**Continuous distributions**
Sample space	$\{0, 1\}$, $\{1, 2, 3, 4\}$, $\{\text{cat}, \text{dog}\}$	$\mathbb{R}$, $\mathbb{R}_+$, $[0, 1]$
Representation	Probability mass function (PMF) $0 \leq p(x) \leq 1$ $\sum_x p(x) = 1$	Probability density functions (PDF) $p(x) \geq 0$ $\int p(x)\mathrm{d}x = 1$ $p(x \in [a, b]) = \int_a^b p(x)\mathrm{d}x$
Mean	$\mathbb{E}[x] = \sum_x x\, p(x)$	$\mathbb{E}[x] = \int x\, p(x)\, \mathrm{d}x$
Variance	$\mathbb{V}[x] = \sum_x (x - \mathbb{E}[x])^2\ p(x)$	$\mathbb{V}[x] = \int (x - \mathbb{E}[x])^2\ p(x)\, \mathrm{d}x$
General expectations	$\mathbb{E}[f(x)] = \sum_x f(x)\, p(x)$	$\mathbb{E}[f(x)] = \int f(x)\, p(x)\, \mathrm{d}x$
Joint distribution	$p(x, y)$	$p(x, y)$
Conditional distribution	$p(x\|y) = \frac{p(x,y)}{p(y)}$	$p(x\|y) = \frac{p(x,y)}{p(y)}$
Sum rule (marginalization)	$p(x) = \sum_y p(x, y)$	$p(x) = \int p(x, y)\, \mathrm{d}y$
Product rule	$p(x, y) = p(x\|y)p(y)$	$p(x, y) = p(x\|y)p(y)$
Independence	$p(x, y) = p(x)p(y)$	$p(x, y) = p(x)p(y)$

Common operations in probabilistic machine learning

- Evaluation: evaluating the probability density of a random variable X taking value x

$$p(X = x) = p(x) \qquad \text{(Lazy notation)}$$

- Conditioning: Deriving the distribution of X conditioned on observed data $\mathcal{D}$

$$p(x|\mathcal{D}) = \frac{p(x, \mathcal{D})}{p(\mathcal{D})}$$

- Maximization: What is the most likely value for a given distribution?

$$\hat{x} = \arg\max_x p(x|\mathcal{D})$$

- Marginalizing: E.g. accounting for unobserved quantities

$$p(x) = \int p(x, y)dy$$

- Sampling: Generate samples from a given some probability distribution

$$x^{(i)} \sim p(x|\mathcal{D})$$

- Compute expectations and moments

$$\mu = \mathbb{E}[x] = \int xp(x)dx$$

- Constructing intervals: E.g. find values ℓ and u such that

$$p(\ell \leq x \leq u) = 0.95$$

The basic building blocks of probabilistic machine learning

Distribution	**Outcome space**	**Examples**
Bernoulli	$0, 1$	Binary data: Cancer/not cancer, click/no click
Binomial	$0, 1, ..., N$	k of out N successes
Poisson	$0, 1, 2, ...$	Count data: number of cancer cells in image
Categorical	$1, 2, 3, ..., K$	Multi-class: cat/dog/bird
Beta	$[0, 1]$	Reasoning about probabilities
Dirichlet	Probability simplex	Reasoning about probability vectors
Gaussian	$\mathbb{R}$	Real numbers
Student's t	$\mathbb{R}$	Real numbers
Gamma	$\mathbb{R}_+$	Non-negative real times, e.g. waiting times

The binomial model and maximum likelihood estimation

Motivating example: A/B testing

Your company's website has two ads. Ad A has been shown $N_A = 123$ times and generated $y_A = 12$ clicks, and Ad B has been shown $N_B = 145$ times and generated $y_B = 20$ clicks.

- What can we say about the click-rates for the two ads? Which one is best?
- We can calculate the sample click-rates

$$\hat{\theta}_A = \frac{y_A}{N_A} = \frac{12}{123} \approx 0.098, \qquad \hat{\theta}_B = \frac{y_B}{N_B} = \frac{20}{145} \approx 0.138$$

- Should we trust these estimates or ask for more data?
- What is the probability that the population click-rate for ad B is below 10%?
- What is the probability that ad B generates mores clicks than ad A?

We can answer such questions using the *beta-binomial model*

Goal for the rest of the lecture

- The *beta-binomial model* is a Bayesian model for estimating proportions, e.g.
  1. What proportion of users clicked the banner?
  2. What proportion of subject recovered after a certain medical treatment?
  3. What proportion of test images is correct classified?
  4. ...

- We will first look at the probabilistic building blocks
  1. Bernoulli distribution
  2. Binomial distribution
  3. Beta distribution

- Maximum likelihood inference

- Bayesian inference

- How does all this apply to A/B testing?

Recap: The Bernoulli distribution

Our first building block

- For a binary random variable $x \in \{0, 1\}$, where 1 represents "success"

$$P(x = 1|\theta) = \theta \qquad P(x = 0|\theta) = 1 - \theta,$$

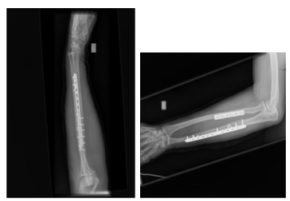
  where $0 \leq \theta \leq 1$

- *Bernoulli distribution* is a distribution over binary variables

$$\text{Ber}(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

Elbow

Forearm



- The mean of Bernoulli variable

$$\mathbb{E}[x] = \theta$$

- The variance is

$$\mathbb{V}[x] = \theta(1 - \theta)$$

Click me!

Recap: The Binomial distribution

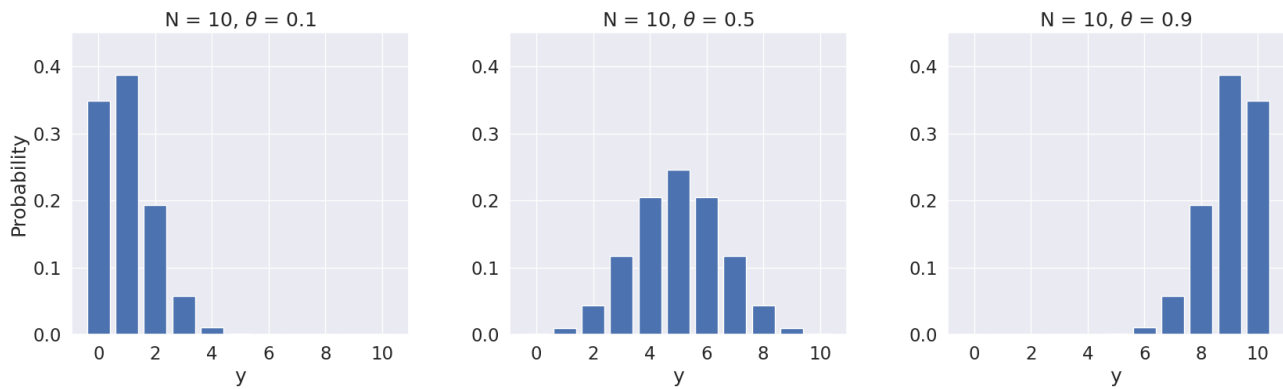
Modelling sequences of independent Bernoulli trials

- For a sequence of N independent Bernoulli trials $x_i \sim \text{Ber}(\theta)$ for $i = 1, \ldots, N$, the number of succeses $y = \sum x_i$ is said to follow a *Binomial distribution*

$$\text{Bin}(y|N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- Example: Suppose we flip a fair coin $N = 10$ times, the probability of getting $y = 4$ heads is given by

$$\text{Bin}(y = 4|N = 10, \theta = 0.5) = \binom{10}{4} 0.5^4 (1-0.5)^{10-4} \approx 0.21$$



- Calculating the mean of y using *linearity* of expectations

$$\mathbb{E}[y] = \mathbb{E}\left[\sum_{i=1}^{N} x_i\right] = \sum_{i=1}^{N} \mathbb{E}[x_i] = \sum_{i=1}^{N} \theta = N\theta$$

The Binomial distribution and maximum likelihood I

- Assume we collected a data set $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ of N independent Bernoulli trials with probability θ. How to estimate θ?

- Let $y = \sum x_i$ denote number of successes, then

$$P(y|\theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- The *likelihood function* is defined as $\mathcal{L}(\theta) \equiv P(y|\theta)$

- The likelihood function measures: what is the probability of the observed data given the parameter value θ

- *Maximum likelihood:* We can estimate the parameters by maximizing the likelihood function $\mathcal{L}$ wrt. θ

$$\hat{\theta}_{\text{MLE}} \equiv \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta \log \mathcal{L}(\theta) = y/N$$

- We *maximize* the log likelihood function by *differentiating, equating to zero and solving* for θ.

$$\log \mathcal{L}(\theta) = \log \left[\binom{N}{y} \theta^y (1-\theta)^{N-y} \right] = \log \left[\binom{N}{y} \right] + y \log(\theta) + (N-y) \log(1-\theta)$$

The Binomial distribution and maximum likelihood II

$$\log \mathcal{L}_{\mathcal{D}}(\theta) = \log \left[\binom{N}{y} \right] + y \log(\theta) + (N - y) \log(1 - \theta)$$

- We *maximize* the log likelihood function by *differentiating, equating to zero and solving* for θ.

- Computing the derivative

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta} \log \mathcal{L}_{\mathcal{D}}(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta} \log \left[\binom{N}{y} \right] + \frac{\mathrm{d}}{\mathrm{d}\theta} \log(\theta) y + \frac{\mathrm{d}}{\mathrm{d}\theta} \log(1 - \theta) (N - y) \\ &= \frac{1}{\theta} y - \frac{1}{1 - \theta} (N - y) = 0 \end{aligned}$$

- Solving for θ yields the *maximum likelihood estimator*

$$\hat{\theta}_{\mathrm{MLE}} = \frac{y}{N} = \frac{1}{N} \sum_{n=1}^{N} x_i$$

- The solution is equal to the empirical mean of the dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$

The Binomial distribution and maximum likelihood: Quiz

- The *likelihood function* is defined as $\mathcal{L}(\theta) \equiv P(y|\theta)$ and the *maximum likelihood estimator* is

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta) = \frac{y}{N}$$

- Spend 4 minutes DTU Learn quiz: "Lecture 1: Likelihoods"

The Binomial distribution and maximum likelihood III

Small data and overfitting

- Suppose an ad is shown $N = 3$ times and observe $y = 0$ clicks, then

$$\hat{\theta}_{\text{MLE}} = 0/3 = 0$$

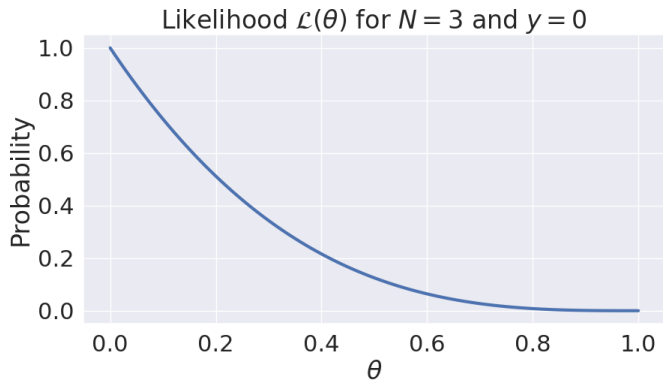
- What is the probability of exactly 0 clicks in the next 50 views? We *plug-in* the maximum likelihood estimator

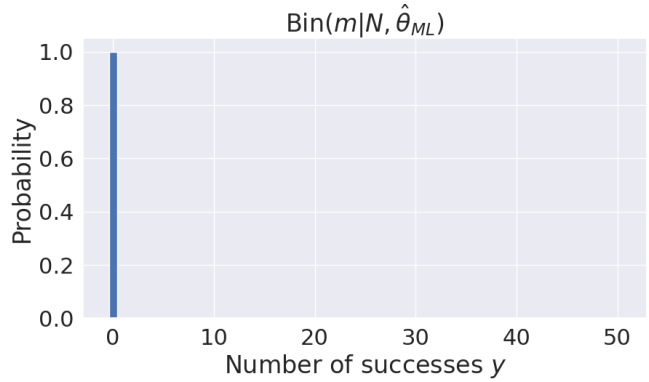
$$\begin{aligned} P[y = 0|\theta_{\text{MLE}}] &= \text{Bin}(0|N = 50, \theta_{\text{MLE}}) \\ &= \binom{50}{0} \theta_{\text{MLE}}^{0} (1 - \theta_{\text{MLE}})^{50-0} \\ &= 1 \end{aligned}$$

- According to this model, we are *absolutely sure* that there will be exactly 0 heads in the next 50 views based on information from *only 3 observations*.

- Does this seem like a reasonable conclusion?

$$\text{Bin}(m|N, \theta) = \binom{N}{m} \theta^m (1 - \theta)^{N-m}$$





Bayesian inference and the Beta-binomial models

Bayesian inference for θ

Example continued

- We observed $N = 3$ views and $y = 0$ clicks, then $\theta_{\text{MLE}} = 0/3 = 0$. This is *overfitting* and this is a common problem when using maximum likelihood for small data sets.

- We can reduce this effect using Bayesian inference

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- *The prior* $p(\theta)$ represents our prior belief about θ **before** seeing the data

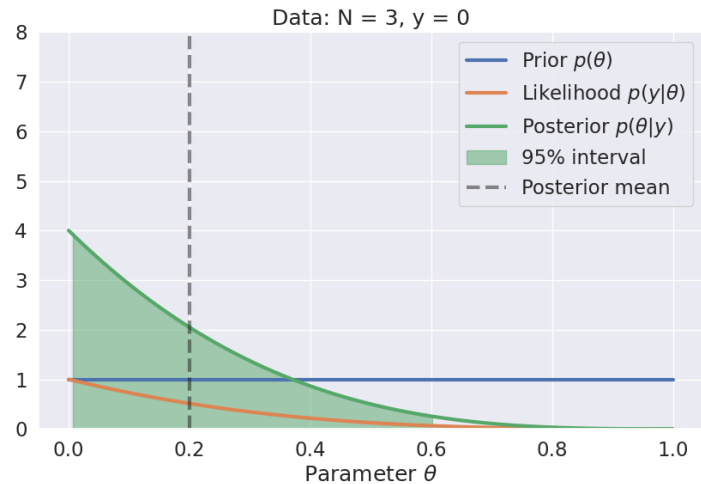
- The *likelihood* $p(y|\theta)$ represents our information from data

- After observing y, *the posterior distribution* $p(\theta|y)$ summarizes all our available information about θ

Bayesian inference in images

- Bayes' rule gives for a probability distribution for θ conditioned on the observed value for y

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$



- We can estimate θ using the mean of the posterior distribution

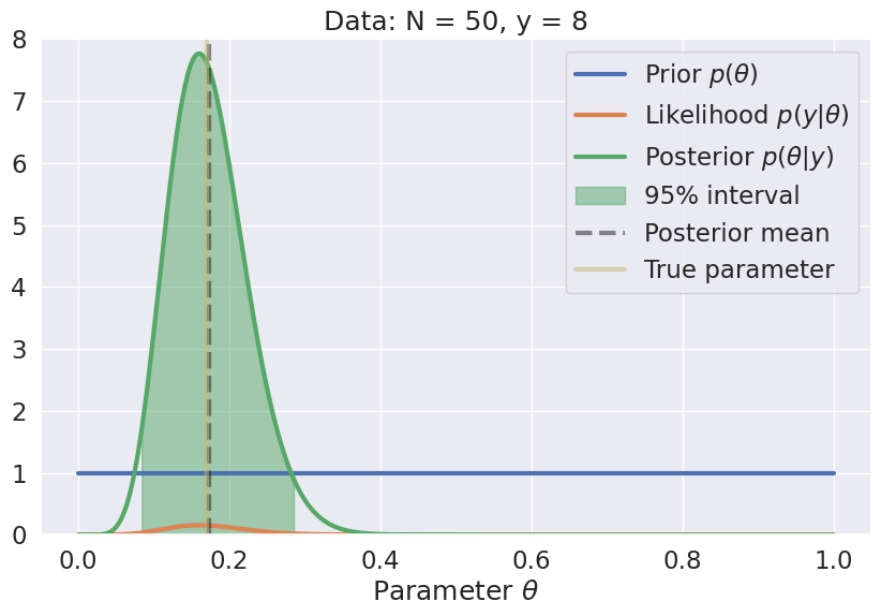
$$\theta_{\text{Bayes}} = \mathbb{E}\left[\theta|y\right] \equiv \int \theta\, p(\theta|y) \mathrm{d}\theta = 0.2$$

- and use *credibility intervals* of the posterior to quantify the uncertainty

$$P(\theta \in [0.01, 0.60]\,|y) = 0.95$$

What happens as we collect more data?

- Simulated data: $x_i \sim \text{Ber}(\theta_0)$ for $i = 1, \ldots, N$, where $\theta_0 = 0.17$



- The posterior concentrates as we collect more data

Bayesian analysis and probabilistic reasoning

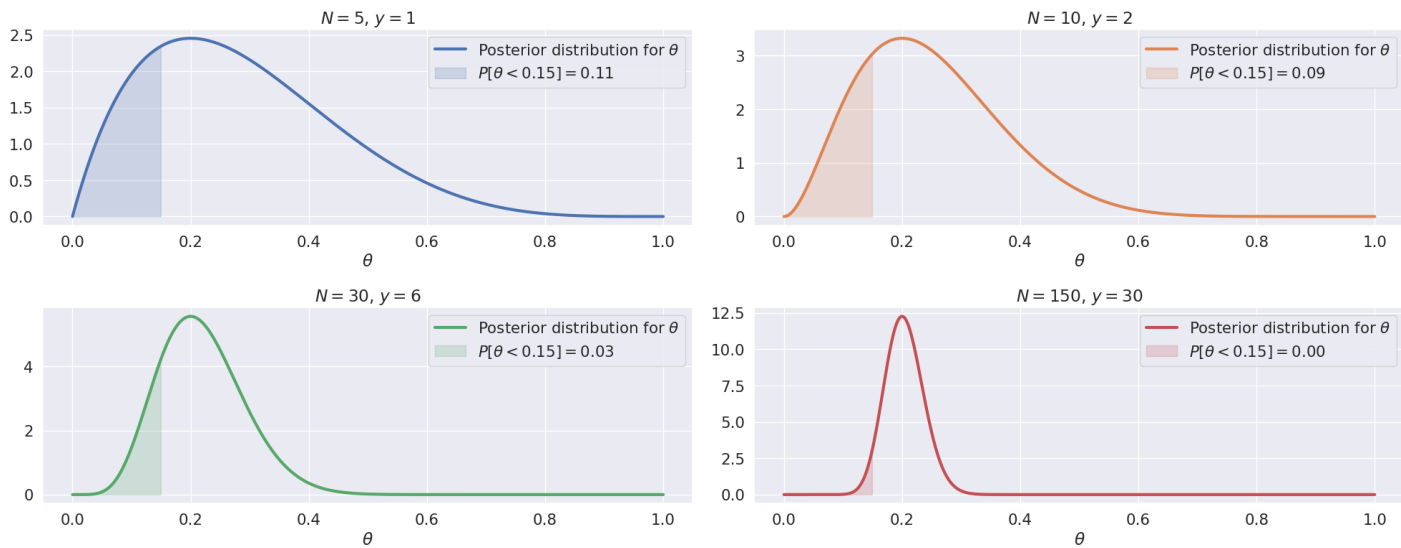
From point estimates to probability distributions

- Bayesian analysis provides a *probability distribution* summarizing our knowledge of θ rather than a *point estimate* like $\hat{\theta}_{\text{MLE}}$

- Many common questions can be answered using *posterior summaries*, e.g. mode, mean, standard deviation, intervals, tail probabilities etc.

$$p(\theta < 0.15|y) = \int_0^{0.15} p(\theta|y)\,\mathrm{d}\theta$$

Examples



The prior distribution: how to choose?

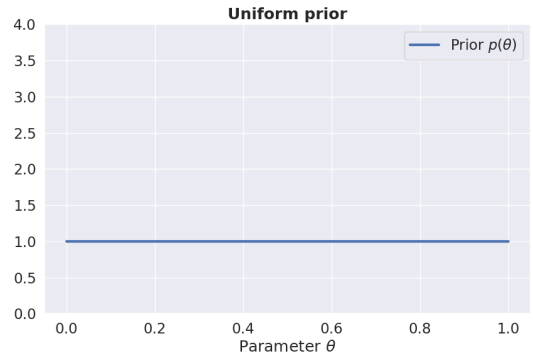
- The prior distribution $p(\theta)$ should reflect our prior assumptions before seeing the data

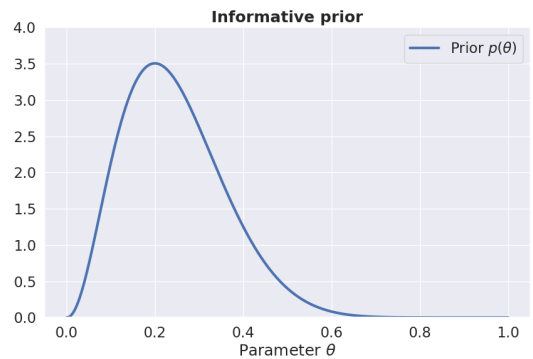
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$$

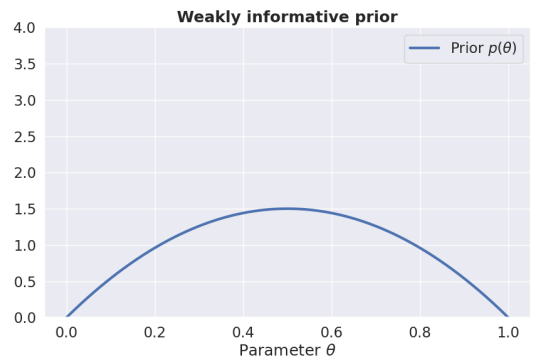
- Different types of priors
  - Uniform priors
  - Informative priors
  - Weakly informative priors
  - Priors for mathematical convinience

- Where does prior knowledge come from?
  - Previous experiments
  - Domain experts
  - Regularization

- Specifying a prior forces us to be explicit about our assumptions



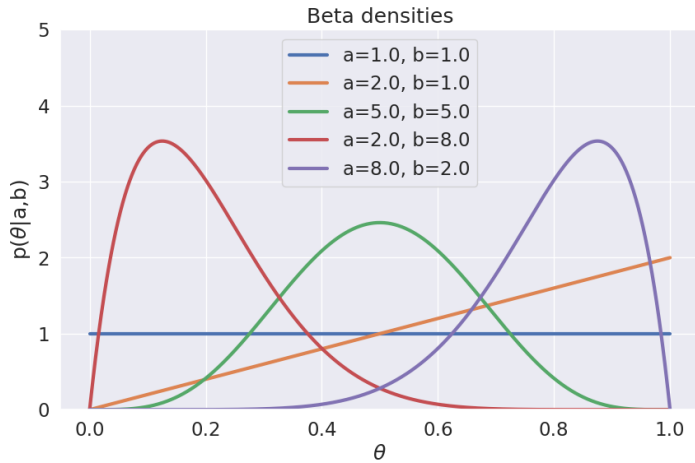




The Beta distribution

A mathematically convenient prior distribution for θ

- *Beta distribution* is a family of distributions for a random variable $\theta \in [0, 1]$ in the unit interval



- The density of the Beta distribution is given by

$$p(\theta|a, b) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

where $B(a, b)$ is a normalization constant given by

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\mathrm{d}x$$

- Mean and variance

$$\mathbb{E}[\theta] = \frac{a}{a+b}$$

$$\mathbb{V}[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$$

The functional form of a Beta distribution

- The density function of Beta distribution is given

$$p(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

- We know density functions integrate to one $\int p(\theta|a,b) = 1$, and hence

$$\int p(\theta|a,b)\mathrm{d}\theta = \int \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = \frac{1}{B(a,b)}\int \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = 1$$

- This implies

$$B(a,b) = \int \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta$$

- Example

$$p(\theta) \propto \theta^5(1-\theta)^4 = \theta^{6-1}(1-\theta)^{5-1} \quad \Rightarrow \quad \int \theta^{6-1}(1-\theta)^{5-1} = B(6,5)$$

- Therefore we know that

$$p(\theta) = \frac{1}{B(6,5)}\theta^{6-1}(1-\theta)^{5-1} = \mathrm{Beta}(\theta|6,5)$$

- We say that the *functional form* of a Beta density is $f(\theta) = \theta^{a-1}(1-\theta)^{b-1}$.

Deriving the analytical posterior for the Beta prior

- The beta distribution is a particular convenient choice for Binomial likelihoods

$$\underbrace{p(\theta|a_0, b_0)}_{\text{prior distribution}} = \frac{1}{B(a_0, b_0)}\theta^{a_0-1}(1-\theta)^{b_0-1} \qquad \underbrace{p(y|\theta)}_{\text{likelihood}} = \binom{N}{y}\theta^y(1-\theta)^{N-y}$$

- Bayes rule states

$$\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta) \\
&= \underbrace{\binom{N}{y}\theta^y(1-\theta)^{N-y}}_{\text{Binomial PMF}} \underbrace{\frac{1}{B(a_0, b_0)}\theta^{a_0-1}(1-\theta)^{b_0-1}}_{\text{Beta density}} \\
&\propto \theta^y(1-\theta)^{N-y}\theta^{a_0-1}(1-\theta)^{b_0-1} \\
&= \theta^{y+a_0-1}(1-\theta)^{N-y+b_0-1} \\
&\propto \text{Beta}(\theta|y + a_0, N - y + b_0)
\end{aligned}$$

- Key take-away: The posterior distribution is a another Beta distribution with parameters

$$\begin{aligned}
a &= a_0 + y \\
b &= b_0 + N - y
\end{aligned}$$

- The Beta distribution is said to be *conjugate* to the binomial distribution because the posterior is of the same *functional form* as the prior

The posterior mean

- The *key equations* for the beta-binomial model

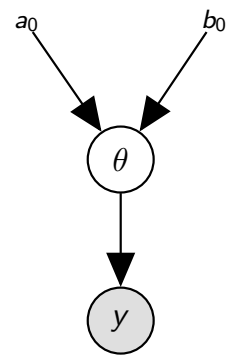
$$p(\theta) = \text{Beta}(\theta|a_0, b_0) \quad \text{(Prior)}$$

$$p(y|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y} \quad \text{(Likelihood)}$$

$$p(\theta|y) = \text{Beta}(\theta|a_0 + y, b_0 + N - y) \quad \text{(Posterior)}$$

- The posterior mean is a compromise between the prior mean and the maximum likelihood solution

$$\mathbb{E}[\theta|y] = \frac{a}{a+b} = \frac{a_0 + y}{a_0 + b_0 + N}$$



- We can interpret a_0 and b_0 as *pseudo observations* of prior successes and failures, respectively

Quick exercise

- The *key equations* for the beta-binomial model

$$p(\theta) = \text{Beta}(\theta|a_0, b_0) \qquad \text{(Prior)}$$

$$p(y|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y} \qquad \text{(Likelihood)}$$

$$p(\theta|y) = \text{Beta}(\theta|a_0 + y, b_0 + N - y) \qquad \text{(Posterior)}$$

$$\mathbb{E}[\theta|y] = \frac{a_0 + y}{a_0 + b_0 + N} \qquad \text{(Posterior mean)}$$

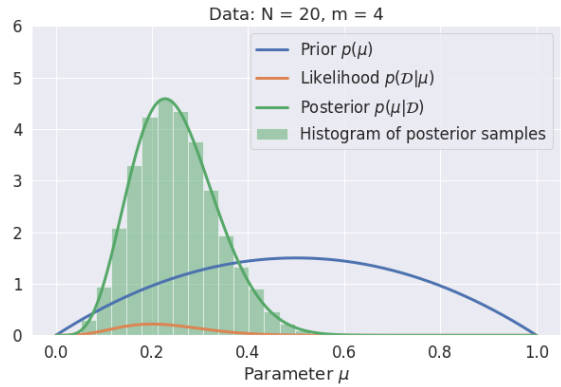
Exercise
Assuming we have observed the following data $N = 20$ views and $y = 4$ click-rates and assume the prior is a Beta distribution with $a_0 = 2$ and $b_0 = 2$, compute ...

1. the prior mean
2. the parameters a and b for the posterior distribution
3. the posterior mean

Exercise - follow up I

Calculating posterior summaries in practice

- Assume we have obtained our posterior of interest and that our goal is to estimate the *posterior mean* $\mathbb{E}_{p(\theta|y)}[\theta]$ and the *probability* $P(\theta < 0.15|y)$



- If we can generate *samples* from the posterior, then we can estimate the posterior mean using *Monte Carlo estimation*:

$$\mathbb{E}_{p(\theta|y)}[\theta] = \int \theta\, p(\theta|y)\, \mathrm{d}\theta \approx \frac{1}{S}\sum_{i=1}^{S} \theta^{(i)}, \qquad \theta^{(i)} \sim p(\theta|y)$$

  - This works for any function of θ:

$$\mathbb{E}_{p(\theta|y)}[f(\theta)] = \int f(\theta)\, p(\theta|y)\, \mathrm{d}\theta \approx \frac{1}{S}\sum_{i=1}^{S} f(\theta^{(i)}), \qquad \theta^{(i)} \sim p(\theta|y)$$

  - We can estimate the probability by counting the fraction of samples below 0.15

$$P(\theta < 0.15|y) = \int_0^{0.15} p(\theta|y)\mathrm{d}\theta = \int \mathbb{I}\left[\theta < 01.5\right] p(\theta|y)\mathrm{d}\theta \approx \frac{1}{S}\sum_{i=1}^{S} \mathbb{I}\left[\theta^{(i)} < 0.15\right]$$

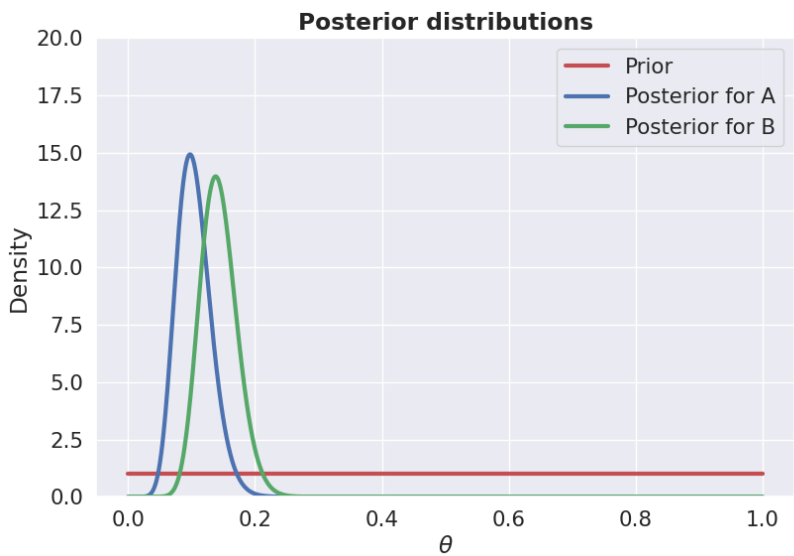
Posterior summaries for A/B testing example

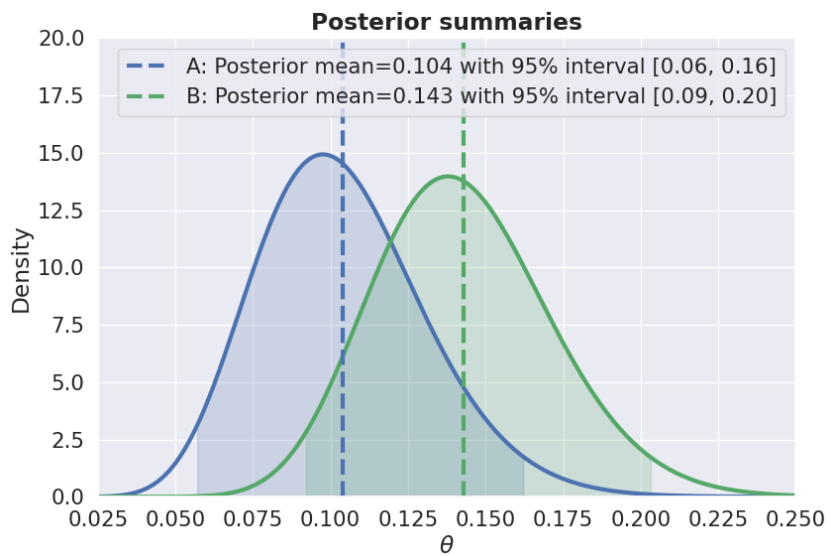
- Ad A has been shown $N_A = 123$ times and generated $y_A = 12$ clicks, and Ad B has been shown $N_B = 145$ times and generated $y_B = 20$ clicks.

- We will use uniform Beta-priors with $a_0 = b_0 = 1$ for both

- We compute posterior distribution for each ad (using $\mathcal{D}$ to denote observed data)

$$p(\theta_A|\mathcal{D}_A) = \text{Beta}(\theta_A|13, 112) \qquad p(\theta_B|\mathcal{D}_B) = \text{Beta}(\theta_B|21, 126)$$





Yes, but which one is better? A or B?

- Let's introduce the difference of the click rates

$$\theta_D = \theta_B - \theta_A$$

- We compute θ_D for each pair of samples

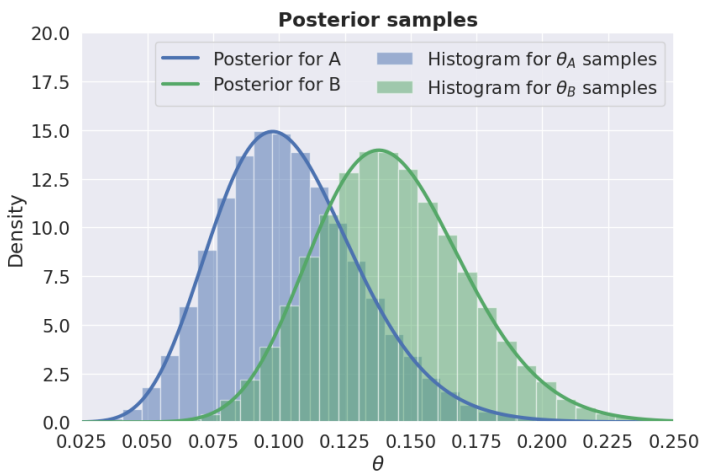
$$\theta_D^{(i)} = \theta_B^{(i)} - \theta_A^{(i)}$$

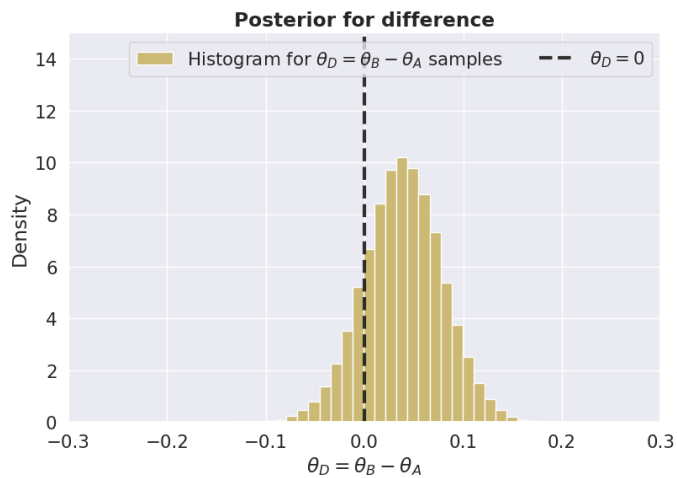
- Posterior probability that B is A than A

$$P\left(\theta_B > \theta_A | \mathcal{D}\right) = P\left(\theta_D > 0 | \mathcal{D}\right) \approx 0.85$$

- Calculating posterior mean and credibility interval from posterior samples $\theta_D^{(i)}$

$$\mathbb{E}\left[\theta_D | \mathcal{D}\right] = 0.039 \qquad P\left(\theta_D \in [-0.038, 0.116] \,|\mathcal{D}\right) \approx 0.95$$





Main takeaways for today

1. Bayesian inference represents all variables using *probability distributions*

2. We update *from prior to posterior* belief using Bayes' rule

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

3. The *posterior summarizes all information* about θ after we observed data

4. The Beta-binomial model is Bayesian approach for *estimating proportions* that uses the binomial distribution as likelihood and the beta distribution as prior

5. As we collect more and more data, the posterior distribution concentrates and becomes more and more independent of the prior

6. Distributions can be summarized using *modes, means, intervals, probabilities* etc. and these can be easily computed via *sampling*

Introduction to the exercise sessions

Intro to exercise

- On DTU Learn you will find an exercise for each week in notebook format
  - We will spent all 4 hours from 13-17 working with the exercises
  - Mix of pen&paper, programming and discussion questions
- The purpose of the exercise for this week is to get familiar with
  - Basic Bayesian terminology
  - The Beta-binomial model
  - Application to A/B testing
- The programming part is based on Python (which is a prerequisite)
  - Numpy, scipy, matplotlib, seaborn packages
  - We will use JaX for numerical computations
  - JaX is a state-of-the-art framework for machine learning with a numpy interface
  - See https://jax.readthedocs.io/en/latest/quickstart.html
- Feel free to collaborate with your peers
- Ask for help!
  - Ask for help when stuck
  - Use teachers/TAs to check your understanding
  - Engage in discussion to practice
- Feedbacks persons: Meet at 16:45

Binary variables and uncertainty quantification: example (bonus slide)

- We model the user click behavior as Bernoulli distributed with probability μ, where $x = 1$ means click and $x = 0$ means no-click such that

Click me!

$$P(\text{click}|\mu) = P(x = 1|\mu) = \mu$$

- We know $0 \leq P(\text{click}|\mu) \leq 1$, but when are we most uncertain about the outcome?
  a) $\mu = 1$?
  b) $\mu = 0.5$?
  c) $\mu = 0$?

- When $\mu = 1$, we are absolute sure that the user is going to click the ad, i.e. no uncertainty

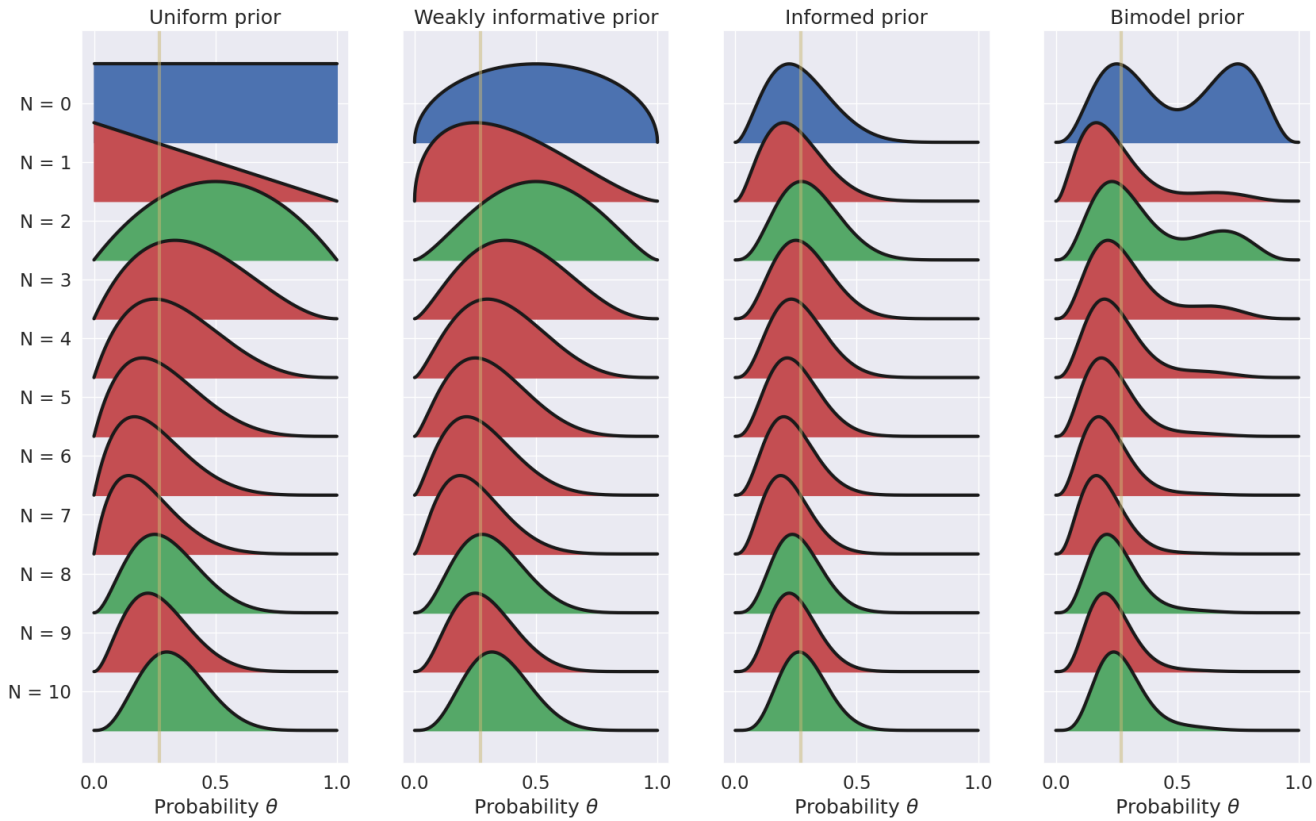
- When $\mu = 0$, we are absolute sure that the user is *not* going to click the ad, i.e. no uncertainty

- When $\mu = 0.5$, we have maximum uncertainty because

$$P(\text{click}|\mu) = P(\text{not click}|\mu) = 0.5$$

The effect of the prior distribution (bonus slide)

- A Bayesian analysis starts with a *prior distribution* $p(\mu)$ representing our knowledge of μ *before* we observe any data



- The prior has a strong influence when the sample size is small, but its effect disappears when the sample size grows

Plots inspired by Rasmus Baath