

# **An introduction to Bayesian statistics**

Sensor fusion & nonlinear filtering

---

Lars Hammarstrand

# WHAT IS BAYESIAN STATISTICS?

---

- A statistical inference framework.
- Can be used for estimation, classification, detection, model selection, etc.
- **Key characteristic:** unknown quantities are described as random.

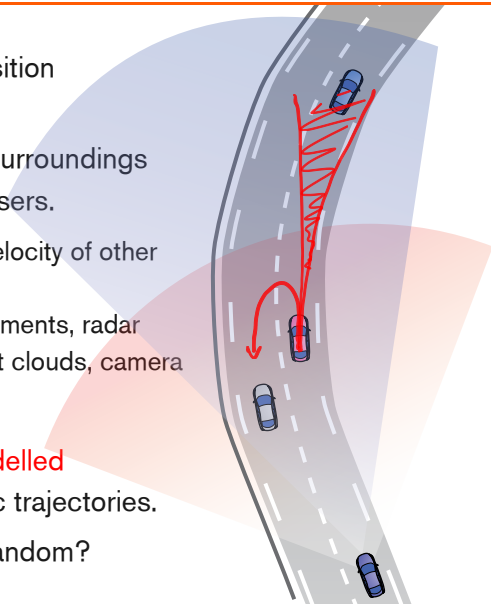
# APPLICATIONS OF BAYESIAN STATISTICS

- A medical application: analyze the disease of a patient.
  - **Quantity of interest**: the disease,  $\theta$ .
  - **Observations**: blood samples, temperature, comments by patient, etc.
- In Bayesian statistics  $\theta$  is described as random  
     $\rightsquigarrow$  we can make statements like: “based on our observations, patient has disease X with 97% probability”.
- **Possible concern**: is the disease random?



# APPLICATIONS OF BAYESIAN STATISTICS

- Self-driving vehicles rely on the ability to position surrounding vehicles.
- This enables the system safely navigate its surroundings without causing accidents with other road users.
  - **Quantity of interest:** relative position and velocity of other vehicles at the current time.
  - **Observations:** wheel speeds, INS measurements, radar detections (distance and angle), Lidar point clouds, camera images, etc.
- Bayesian statistics: **vehicle motions are modelled statistically**  $\leadsto$  helps us to rule out unrealistic trajectories.
- **Possible concern:** are the vehicle motions random?



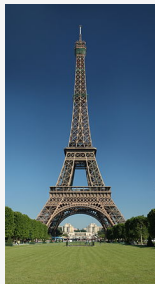
# COMPARISON: BAYES VS FREQUENTIST

- There are two main strategies to decision making:  
**Bayesian** and **frequentist statistics**.
- In **frequentist statistics**, the quantities of interest are described as **unknown and deterministic**.

## Bayes vs Frequentist

We wish to estimate the height of the Eiffel tower. Is the height random or not?

- **Frequentist perspective:** the tower has a certain height and is therefore not random.
- **Bayesian perspective:** we describe our uncertainties in the height stochastically  
⇒ height is described as random!



# OVERVIEW OF THE BAYESIAN STRATEGY

---

Suppose we wish to estimate  $\theta$  given measurements  $y$ .

Key steps in a Bayesian method:

1. **Modeling.** Model what we know about  $\theta$  (using a prior  $p(\theta)$ ) and the how the measurements  $y$  relate to  $\theta$  (using a density  $p(y|\theta)$ ).
2. **Measurement update.** Combine what we knew before (the prior) with our measurement (with  $p(y|\theta)$ , also called the likelihood) to summarize what we know about  $\theta$  ( $p(\theta|y)$ ).
3. **Decision making.** Given what we know about  $\theta$  (described by  $p(\theta|y)$ ) and a loss function, we compute *an optimal decision*.

## SELF-ASSESSMENT QUESTIONS

---

Which of the following statements are correct:

- Bayesian methods can be used to solve many types of decision making problems including estimation, detection and classification.
- We can model the height of the Eiffel tower as random only if we think that there are many similar towers with different heights.
- In Bayesian statistics we describe what we know about  $\theta$  (the quantity of interest) before observing any measurements.

Check all that apply.

# Bayes' rule – a first example

Sensor fusion & nonlinear filtering

---

Lars Hammarstrand



# BAYES' RULE: A FIRST EXAMPLE

## Selecting fruit from an urn

- An urn is selected at random (prob.  $1/2$ ,  $1/2$ ).  
From that urn we pick a fruit.



- If fruit is orange, what is probability that we chose the red urn?

# PROBABILITY THEORY

---

- Bayesian statistics is simple! We only need two rules:

## Conditional probability (product rule)

$$\Pr\{y, \theta\} = \Pr\{y|\theta\} \Pr\{\theta\}$$

## The law of total probability (sum rule)

$$\Pr\{y\} = \sum_{\theta} \Pr\{y, \theta\} \quad \text{discrete variables}$$

$$p(y) = \int_{\theta} p(y, \theta) d\theta \quad \text{continuous variables}$$

# PROBABILITY THEORY – BAYES' RULE

- Bayes' rule is a consequence of conditional probability,

$$p(y|\theta)p(\theta) = p(\theta|y)p(y).$$

## Bayes' rule

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

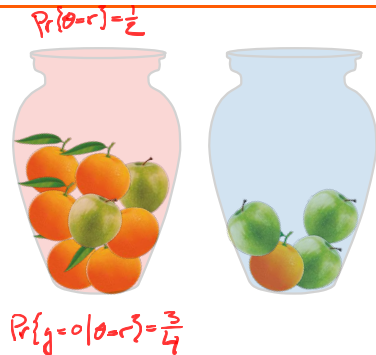
- **Usage of Bayes' rule:** express a relation of interest,  $p(\theta|y)$ , in terms of the relation that we know  $p(y|\theta)$ .
- Note that  $p(y) = \int_{\theta} p(y|\theta)p(\theta) d\theta$ .

# BAYES' RULE: A FIRST EXAMPLE

- Let  $\theta \in \{r, b\}$  be color of urn, and  $y \in \{o, a\}$  be the fruit.
- Question:** If fruit is orange, what is probability that we chose the red urn?

$$Pr\{\theta=r|y=o\} = \frac{Pr\{y=o|\theta=r\}Pr\{\theta=r\}}{Pr\{y=o\}} = \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{3}{4}$$

$$\begin{aligned} Pr\{y=o\} &= Pr\{y=o, \theta=r\} + Pr\{y=o, \theta=b\} = \\ &= \frac{3}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{2} \end{aligned}$$



# BAYES' RULE: A FIRST EXAMPLE

- Let  $\theta \in \{r, b\}$  be color of urn, and  $y \in \{o, a\}$  be the fruit.
- **Question:** If fruit is orange, what is probability that we chose the red urn?
- Bayes' rule gives

$$\Pr\{\theta = r | y = o\} = \frac{\Pr\{y = o | \theta = r\} \Pr\{\theta = r\}}{\Pr\{y = o\}}$$

where  $\Pr\{\theta = r\} = 1/2$ ,  $\Pr\{y = o | \theta = r\} = 3/4$  and

$$\begin{aligned}\Pr\{y = o\} &= \Pr\{y = o, \theta = r\} + \Pr\{y = o, \theta = b\} \\ &= \frac{3}{4} \frac{1}{2} + \frac{1}{4} \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

- Thus,  $\Pr\{\theta = r | y = o\} = \frac{3}{4}$ .



# **Building blocks of Bayesian models – Likelihoods, Priors and Posteriors**

Sensor fusion & nonlinear filtering

---

Lars Hammarstrand

# LIKELIHOODS, PRIORS AND POSTERIORIORS

## General problem formulation

- We are interested in an unknown parameter  $\theta \in \Theta$  for which we observe some related data  $y$ .
- Common problem types are **estimation** (e.g.,  $\Theta = \mathbb{R}^n$ ) and **detection** problems (e.g.,  $\Theta = \{-1, 1\}$ ).

## Assumption

- The observed data,  $y$ , is distributed as

$$y \sim p(y|\theta),$$

where  $p$  is a known distribution.

# LIKELIHOODS, PRIORS AND POSTERIORIORS

## Likelihood

- Since  $y$  is observed, we often view  $p(y|\theta)$  as a function of  $\theta$ ,

$$l(\theta|y) = p(y|\theta),$$

where  $l(\theta|y)$  is called the **likelihood** function.

- **Note:** the likelihood function is *not* a density w.r.t.  $\theta$ .

## Prior

- In **Bayesian statistics** we have a **prior** distribution on  $\theta$ ,  $p(\theta)$ .
- Prior means *earlier*, or before, and  $p(\theta)$  describes what we know *before* observing  $y$ .



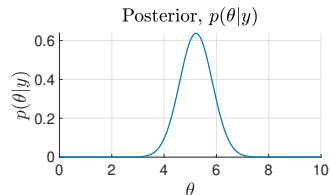
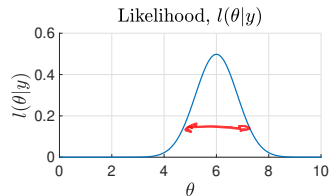
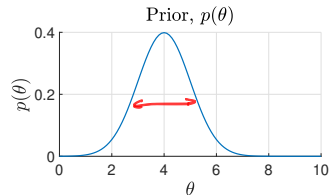
# LIKELIHOODS, PRIORS AND POSTERIOR

## Posterior

- One objective in Bayesian statistics is to compute the **posterior**

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

- Posterior means *after* and  $p(\theta|y)$  describes what we know *after observing*  $y$ .

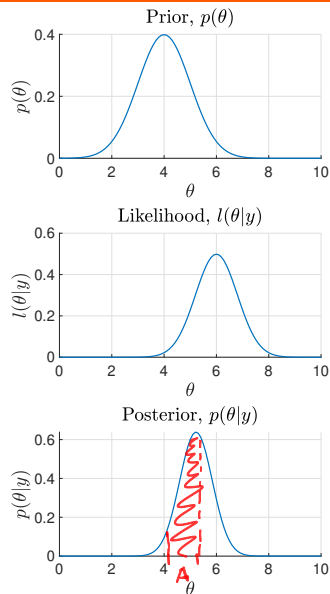


# LIKELIHOODS, PRIORS AND POSTERIORIORS

- We summarize this as

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

- Given the posterior,  $p(\theta|y)$  we can answer, e.g.,
  - What is the most probable  $\theta$ ?
  - What is the probability that  $\theta \in \mathcal{A}$ ?
  - What is the posterior mean of  $\theta$ ?
- We can also minimize expected costs in a decision theoretic manner.



## EXAMPLE: SCALAR IN GAUSSIAN NOISE

### Estimation of scalar in Gaussian noise

- Suppose we observe

$$y = \theta + v, \quad v \sim \mathcal{N}(0, \sigma^2)$$

such that  $p(y|\theta) = \mathcal{N}(y; \theta, \sigma^2) \propto \exp\{-(y - \theta)^2/(2\sigma^2)\}$ .

- A common non-informative prior on  $\theta$  is  $p(\theta) \propto 1$ .

- What is the posterior?

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta) \propto \exp(-(y - \theta)^2/(2\sigma^2)) \cdot 1 \propto \mathcal{N}(\theta; y, \sigma^2)$$

$$\Rightarrow p(\theta|y) = \mathcal{N}(\theta; y, \sigma^2)$$



# BAYESIAN APPROACH TO SENSOR FUSION

- Suppose we collect measurements from **two types of sensors**,  $y_1$  and  $y_2$ ,

## Bayesian fusion of independent observations

- We seek the posterior distribution:

$$p(\underbrace{\theta}_{\theta} | \underbrace{y_1, y_2}_{y}) \propto p(\theta) p(y_1, y_2 | \theta).$$

- It is often reasonable to assume that

$$p(y_1, y_2 | \theta) \approx p(y_1 | \theta) p(y_2 | \theta),$$

i.e., that measurements are **conditionally independent**.



# SELF-ASSESSMENT

---

The posterior distribution is  $p(\theta|y) \propto p(y|\theta)p(\theta)$ .

It is also true that:

- The normalization factor is **not always unique**?
- The posterior  $p(\theta|y)$  can **always be uniquely determined** from the fact that  $\int p(\theta|y) d\theta = 1$ ?
- The posterior distribution can **only be uniquely determined if** it is proportional to a well known distribution, e.g., a Gaussian.

Only one statement is correct.

# Bayesian Decision Theory

Sensor fusion & nonlinear filtering

---

Lars Hammarstrand

# BAYESIAN DECISION PRINCIPLE

---

- How can we use  $p(\theta|y)$  to make decisions?
- **Examples** of decision problems
  - How to control a self-driving vehicle.
  - How to invest money.
  - Select medicine to give to a patient
  - Estimate a parameter vector (may represent temperature, distance, etc).

## Basic principle of Bayesian decision theory

- Minimize expected loss  
or, equivalently,
- Maximize expected utility.

# DECISION THEORY – A TOY EXAMPLE

## Choosing a course

- A student wants to decide whether to take a course or not.
- Suppose  $\theta \in \{\text{good course, fair course, bad course}\}$  and

	good course	fair course	bad course
$\Pr\{\theta y\}$	0.3	0.3	0.4

- If the loss function is

	good course	fair course	bad course
Take	0	5	30
Not take	20	5	0

should he/she then take the course?



# MINIMUM POSTERIOR EXPECTED LOSS

- We often study loss functions  $C(\theta, a)$  instead of utility.  
(Typically,  $C \geq 0$ .)
- Let  $\hat{\theta}$  denote an estimate of  $\theta$ .

## Optimal Bayesian decisions

Minimize the posterior expected loss

$$\hat{\theta} = \arg \min_a \mathbb{E} \{ C(\theta, a) | y \}$$

where  $\mathbb{E} \{ C(\theta, a) | y \} = \int_{\Theta} C(\theta, a) p(\theta | y) d\theta$

- **Note:**  $y$  is given (fixed) and  $\theta$  is random.

# SELF-ASSESSMENT

---

To make an optimal Bayesian decision it is sufficient to know:

- The prior,  $p(\theta)$ , the likelihood,  $p(y|\theta)$ , and a loss function  $C(\theta, a)$ .
- The likelihood,  $p(y|\theta)$ , and a loss function  $C(\theta, a)$ .
- The posterior distribution,  $p(\theta|y)$ , and a loss function,  $C(\theta, a)$ .

Check all statements that apply.

# COMPARISON: BAYES VS FREQUENTIST

Frequentist	Bayes
$\theta$ is fixed and unknown  $\Rightarrow \theta$ is deterministic	Uncertainties in $\theta$ are described stochastically  $\Rightarrow \theta$ is random
Maximum likelihood (ML) most famous estimator $\hat{\theta}_{ML} = \arg \max_{\theta} l(\theta y)$	Minimum mean square error and maximum a posteriori estimators, e.g., $\hat{\theta}_{MAP} = \arg \max_{\theta} p(\theta)l(\theta y)$
Study performance by averaging over $y$ for fixed $\theta$	Make decisions conditioned on the observation $y$ .

- **Note 1:** most Bayesians also study frequentist performance.
- **Note 2:** many frequentists agree that parameters may be random in some situations.

# Cost functions in Bayesian decision theory

Sensor fusion & nonlinear filtering

---

Lars Hammarstrand

# BAYESIAN DECISION THEORY – SUMMARY

- Bayesian decision theory relies on
  1. Likelihood:  $p(y|\theta)$
  2. Prior distribution:  $p(\theta)$
  3. Loss function:  $C(\theta, a)$
- Combining likelihood and prior gives posterior

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

## Posterior and loss gives decisions

$$\hat{\theta} = \arg \min_a \int_{\Theta} C(\theta, a) p(\theta|y) d\theta.$$

# THE QUADRATIC LOSS FUNCTION

## Minimum mean squared error estimator, MMSE

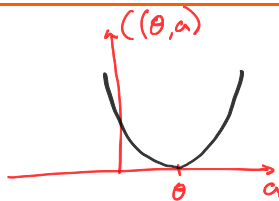
- Parameter estimation,  $\theta \in \Theta = \mathbb{R}^n$
- Most common loss function is the **quadratic loss**

$$C(\theta, a) = \|\theta - a\|_2^2 = (\theta - a)^T (\theta - a)$$

- Let:  $\bar{\theta} = \mathbb{E}\{\theta|y\}$ ,  $\mathbf{P} = \text{Cov}\{\theta|y\} = \mathbb{E}\{(\theta - \bar{\theta})(\theta - \bar{\theta})^T | y\}$

- Optimal estimator:  $E\{C(\theta, a)|y\} = E\{(\theta - a)^T(\theta - a)|y\} = E\{(\underbrace{\theta - \bar{\theta}}_{\text{zero mean}} + \underbrace{\bar{\theta} - a}_{\text{Determ}})^T(\theta - \bar{\theta} + \bar{\theta} - a)|y\}$   
 $= \underbrace{E\{(\theta - \bar{\theta})^T(\theta - \bar{\theta})|y\}}_{\text{Tr}\{\mathbf{P}\}} + \underbrace{E\{(\theta - \bar{\theta})^T|y\}}_{=0}(\bar{\theta} - a) + 0 + (\bar{\theta} - a)^T(\bar{\theta} - a)$

$$\hat{\theta}_{\text{MMSE}} = \min_a \arg E\{C(\theta, a)|y\} = \bar{\theta}$$



# THE 0 – 1 LOSS FUNCTION

## Maximum a-posteriori estimator, MAP

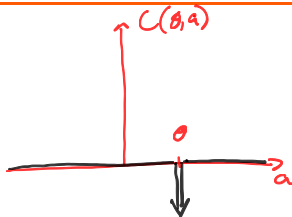
- Parameter estimation,  $\theta \in \Theta = \mathbb{R}^n$ .
- Another common choice is the **0 – 1 loss** function

$$C(\theta, a) = -\delta(\theta - a),$$

$\delta(\cdot)$  is the Dirac's delta function.

- Optimal estimator:  $\mathbb{E}\{C(\theta, a)|y\} = \mathbb{E}\{-\delta(\theta - a)|y\}$   
 $= -\int p(\theta|y)\delta(\theta - a) d\theta = -p(\theta|y)\Big|_{\theta=a}$

$$\begin{aligned}\Rightarrow \hat{\theta}_{\text{MAP}} &= \arg \min_a -p(\theta|y)\Big|_{\theta=a} \\ &= \arg \max_{\theta} p(\theta|y)\end{aligned}$$



# SELF-ASSESSMENT

---

Suppose  $p(\theta|y) = \mathcal{N}(\theta; \bar{\theta}, \mathbf{P})$ . The MMSE and MAP estimators are, respectively,

- $\bar{\theta} + \text{tr}\{\mathbf{P}\}$  and  $\bar{\theta}$ .
- $\bar{\theta}$  and  $\mathcal{N}(\theta; \bar{\theta}, \mathbf{P})$ .
- $\bar{\theta}$  and  $\bar{\theta}$ .
- $\bar{\theta} + \text{tr}\{\mathbf{P}\}$  and  $\mathcal{N}(\theta; \bar{\theta}, \mathbf{P})$ .

Only one statement is correct.