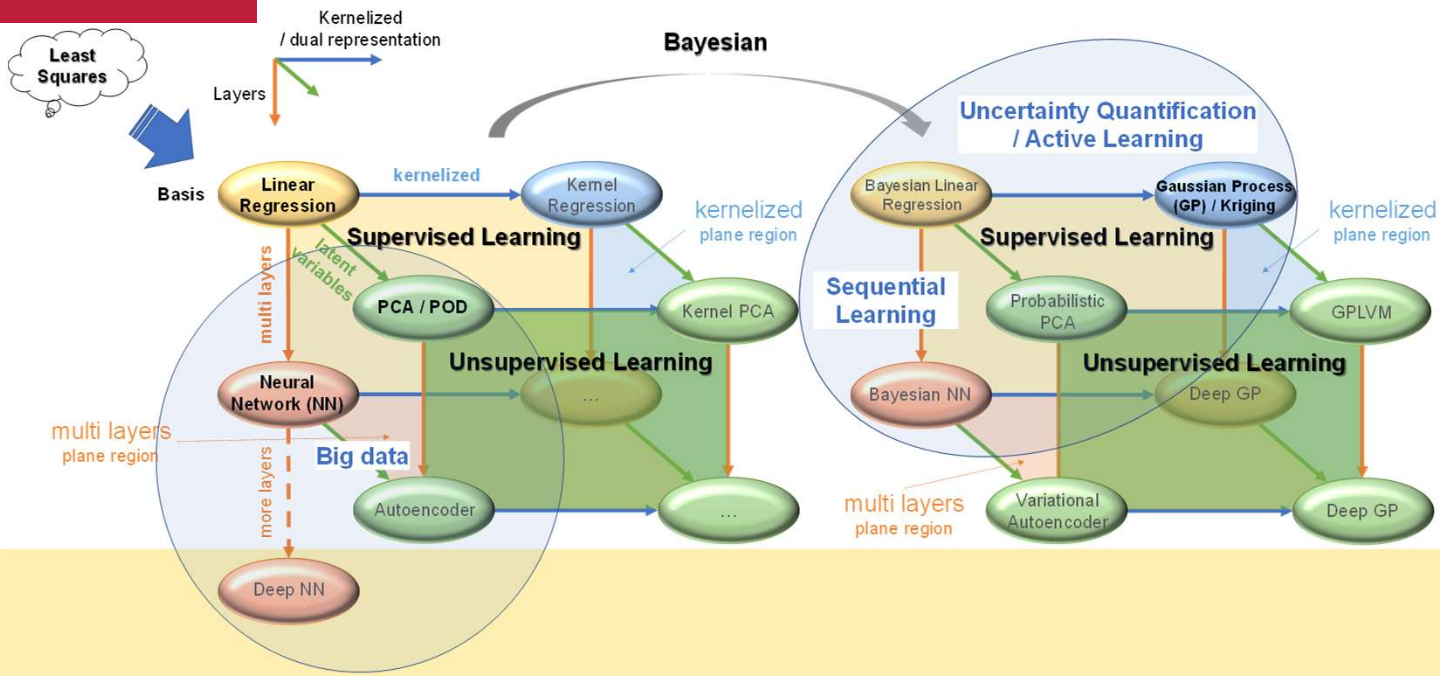
Technische Universität Braunschweig

IFL



Scientific Machine Learning

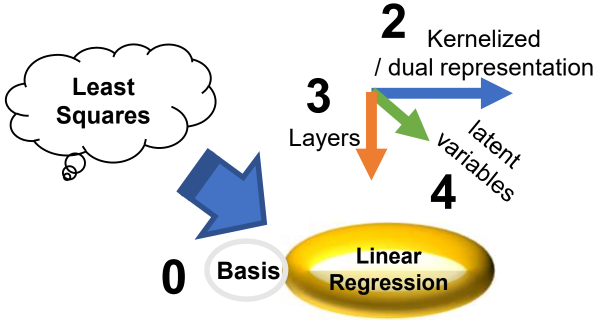
Lecture 5: Bayesian Statistics (1/2)

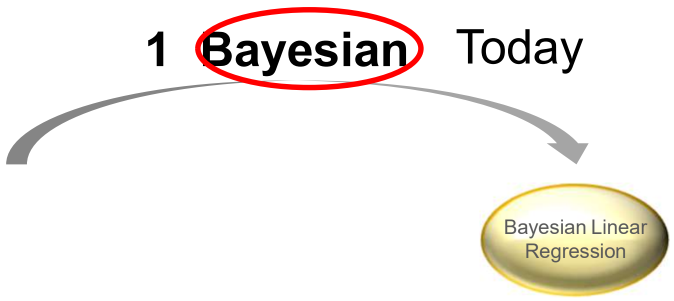
Dr. Daigo Maruyama

Prof. Dr. Ali Elham

Key Components

the three axes + Bayesian axis = four axes





Lecture content

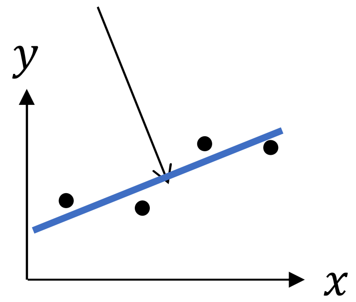
- Bayesian approach
- Summary (update of the current process)

There are no references about the context of this lecture.

Process in General (Review)

1. Define a probabilistic model $p(y|x, \boldsymbol{w})$ ⬅ Model definition
2. MLE $p(\mathbf{y}|\mathbf{X}, \boldsymbol{w}) \xrightarrow{\text{MLE}} \hat{\boldsymbol{w}}$ ⬅ Learning process
3. New prediction $p(y_{new}|x_{new}, \hat{\boldsymbol{w}})$ ⬅ Prediction process

$$y = w_1 x + w_0 = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

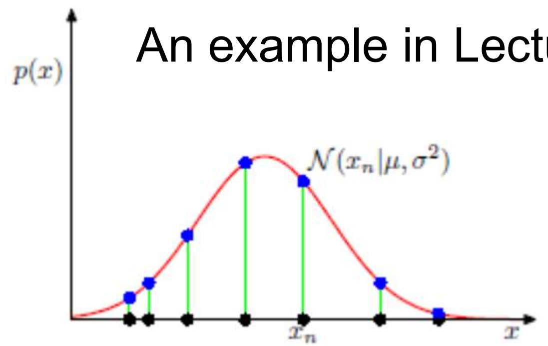


Data: $\mathcal{D}$

$$\mathbf{X} = \left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(N)}\right)^{\mathrm{T}}$$

$$\mathbf{y} = \left(y^{(1)}, y^{(2)}, \cdots, y^{(N)}\right)^{\mathrm{T}}$$

An example in Lecture 3



Bayes’ theorem (REVIEW of Lecture 2)

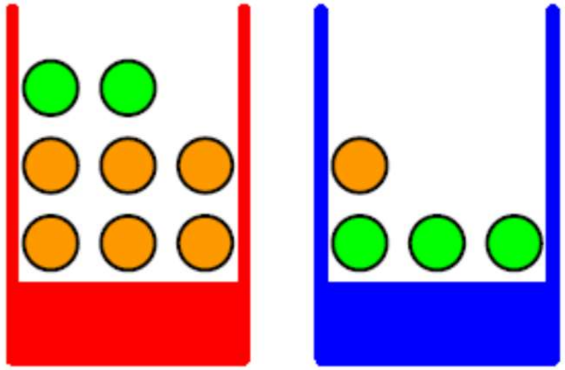
sum rule $P(Y) = \sum_X P(X, Y)$

product rule $P(X, Y) = P(X|Y)P(Y) \qquad P(X, Y) = P(Y|X)P(X)$

Bayes’ theorem $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Interpretation is important.

Let us consider
time flow / causality



Bayesian Approach

Relationship between parameters $\boldsymbol{w}$ and data $\mathcal{D}$

Bayes' theorem

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{\boxed{P(\mathcal{D}|\boldsymbol{w})}P(\boldsymbol{w})}{P(\mathcal{D})}$$

Likelihood function $P(\mathcal{D}|\boldsymbol{w})$

So far:

1. **Define a probabilistic model**
2. **Then, MLE (Maximum Likelihood Estimation)**

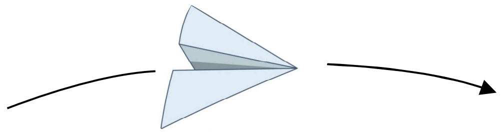
extended

In this process, the goal was to determine the likelihood function, then to maximize it.

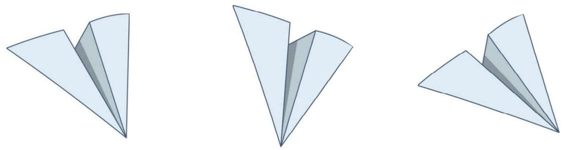
Bayesian Approach

Example:

You threw a paper plane three times.

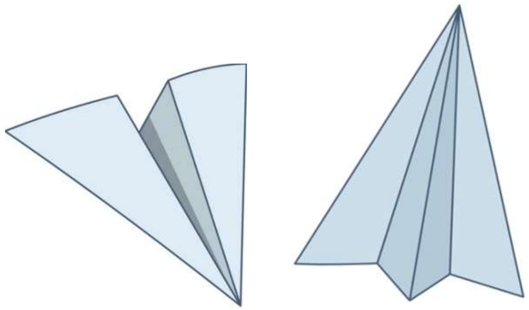


Observed Data: all the three times they were front.



front **front** **front**

Front **Back**



probability μ

probability $1-\mu$

$0 \leq \mu \leq 1$

What is the probability μ?

You can imagine **a coin** as well.

→ Then, $\mu = 0.5$?

But why?

Bayesian Approach

Let's do the MLE.

1. Define a probabilistic model
2. Then, MLE

For the multiple trials for the binary result, the probability distribution is

Binomial distribution $$\boxed{p(m|\mu) = \binom{3}{m}\mu^m(1-\mu)^{3-m}}$$ $0 \le \mu \le 1$

Probabilistic model

strictly, $p(m|N=3,\mu)$ ($N=3$ is fixed and omitted.)

Now we could obtain data $\mathcal{D}$:

$\mathcal{D}$: $m = 3$

$$\boxed{p(m=3|\mu) = \binom{3}{3}\mu^3(1-\mu)^0 = \mu^3}$$

Likelihood function

MLE:
maximize $p(m=3|\mu)$

and find $\hat{\mu}$ where it maximize p

$$\hat{\mu} = \max_{\mu} p(m=3|\mu)$$

Bayesian Approach

$$\hat{\mu} = \max_{\mu} p(m = 3|\mu)$$
$$= \max_{\mu} \mu^3$$

$$0 \leq \mu \leq 1$$

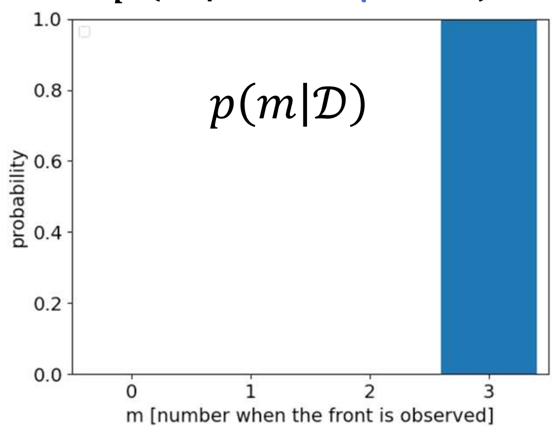
$\boxed{\hat{\mu} = 1}$ The probability where the front is observed is 1.

Probabilistic model

$$p(m|\mu) = \binom{3}{m} \mu^m (1-\mu)^{3-m}$$

What is the probability of m?

$p(m|N=3, \hat{\mu}=1)$



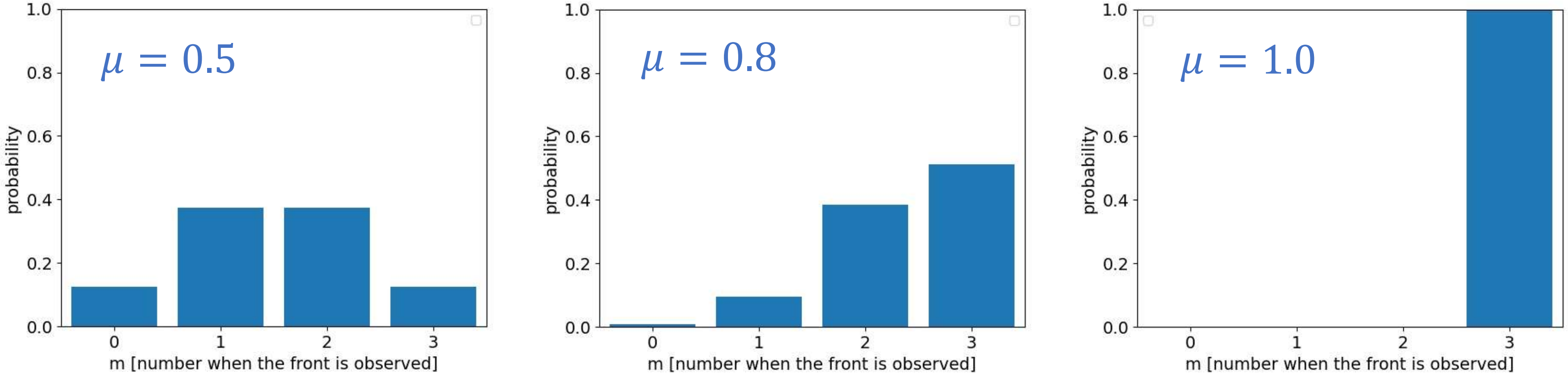
GOAL: Let's use this for prediction (for arbitrary m).

$$p(m|\hat{\mu}=1) = \binom{3}{m} 1^m (1-1)^{3-m}$$
$$= \binom{3}{m}$$

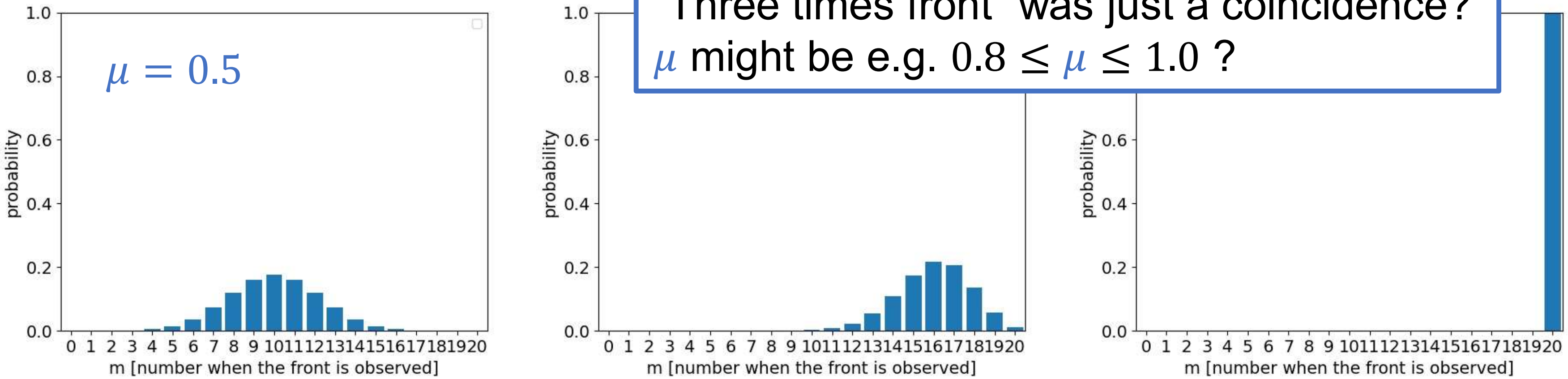
Bayesian Approach

$0 \leq \mu \leq 1$

$N = 3$



$\mu = 0.8$ seems to be more realistic?
“Three times front” was just a coincidence?
μ might be e.g. $0.8 \leq \mu \leq 1.0$?



Bayesian Approach

What was the problem?

- Maybe too few trials

➡ If we try $N = 20$, the front might observed less than that at least not 20.

- Difficult to uniquely determine the parameter μ

➡ μ might be between 0.8 and 1.0?

- Contrary to your belief? or your past experiences?

➡ The true μ might be around 0.5 in my belief / past experiences.

Bayesian Approach

STOP before doing MLE!

We can manage to use all the information of the likelihood function.

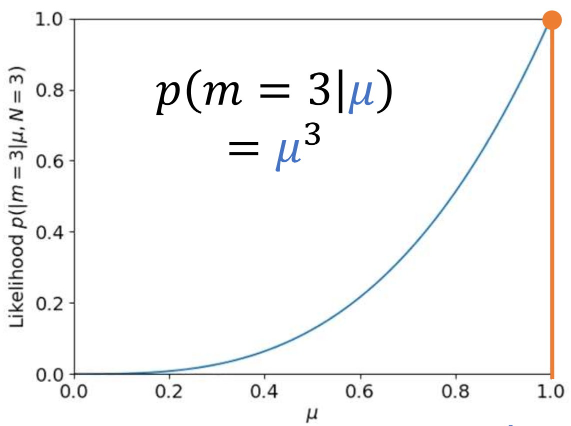
Data $\mathcal{D}$: $m = 3$

$$p(m = 3|\mu) = \binom{3}{3}\mu^3(1-\mu)^0 = \mu^3$$

$N = 3$

$0 \le \mu \le 1$

Likelihood function



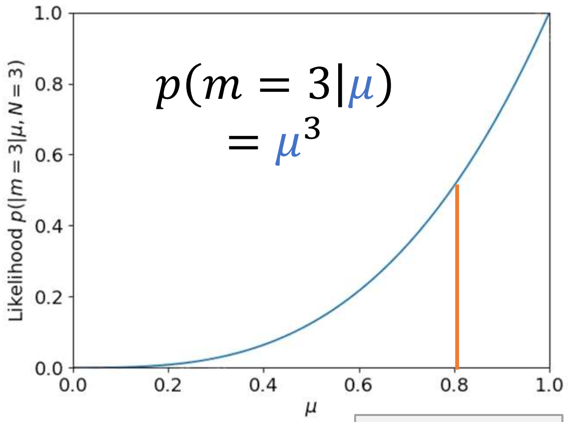
$$\min_{\mu} p(m = 3|\mu)$$

MLE

$\hat{\mu} = 1$

Bayesian Approach

Likelihood function



mean $\hat{\mu} = 0.8$

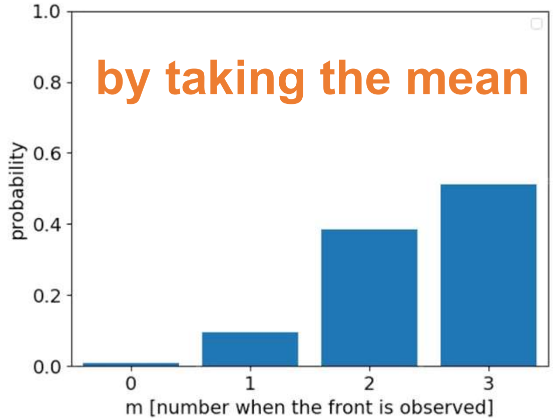
How about taking **the expectation**?

We just have started to be conscious of **the distribution**!

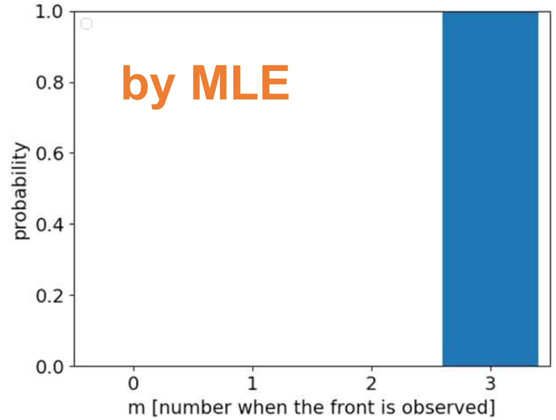
Bayesian perspectives

MLE is not always the best option?

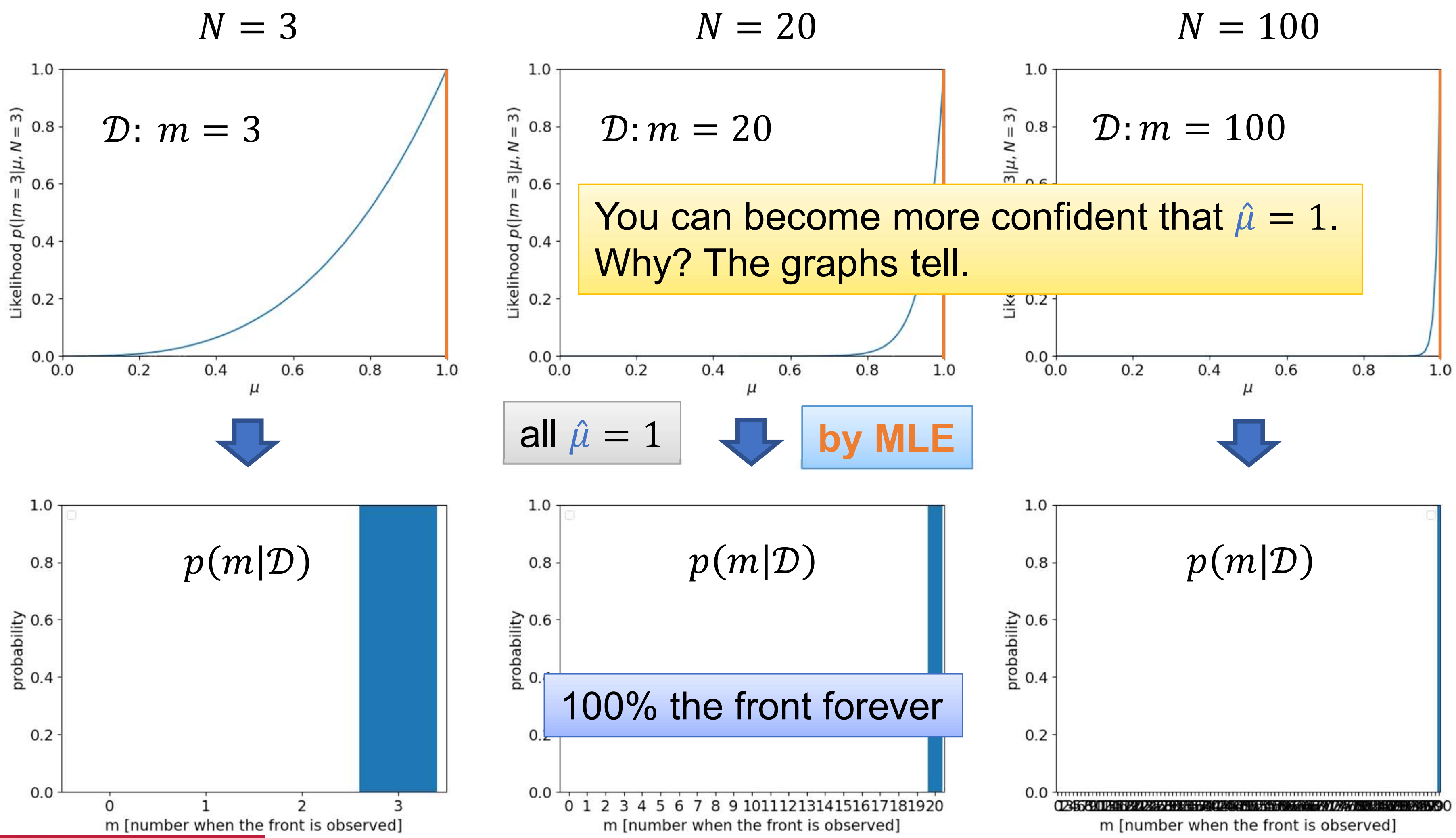
$\mu = 0.8$



$\mu = 1.0$



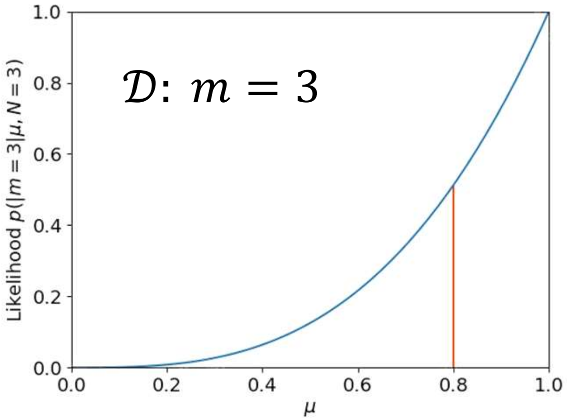
Bayesian Approach

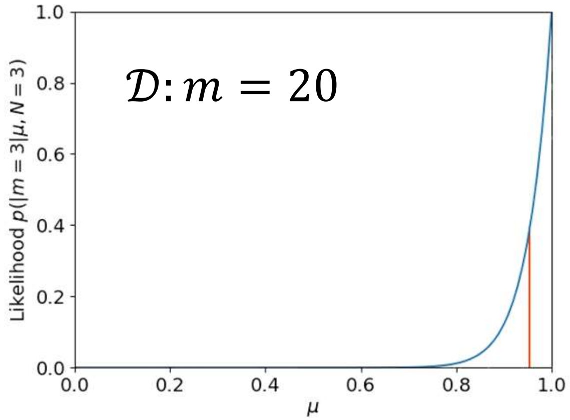


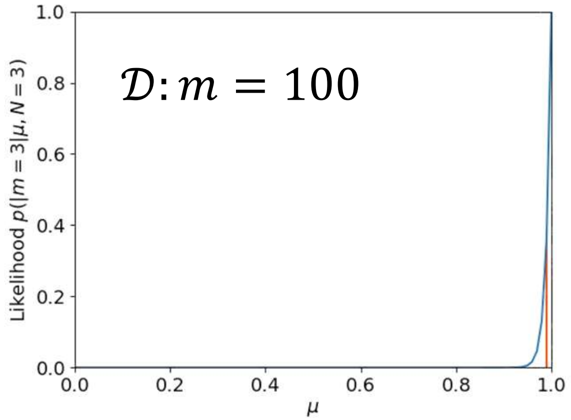
Bayesian Approach

by taking the mean

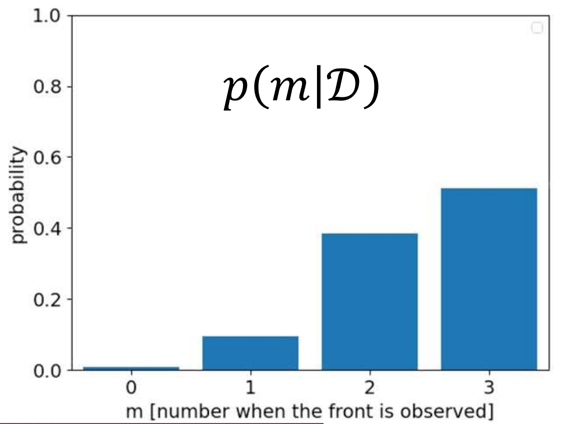
$N = 3$	$N = 20$	$N = 100$

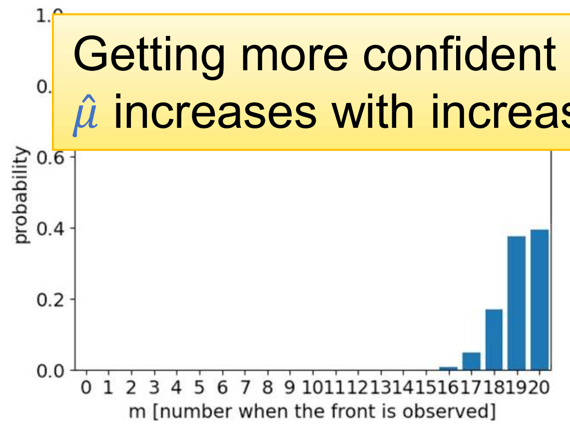


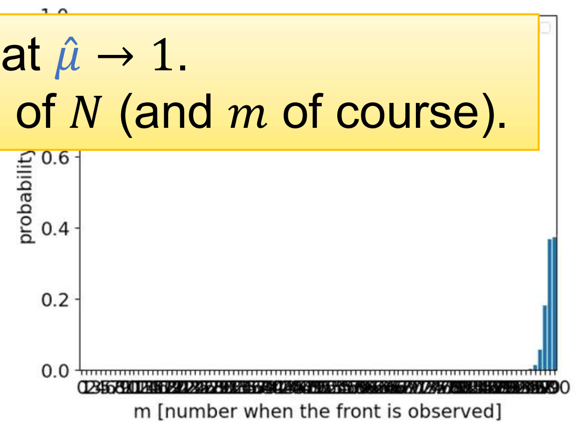




$\hat{\mu} = 0.8$ | $\hat{\mu} = 0.954$ | $\hat{\mu} = 0.990$





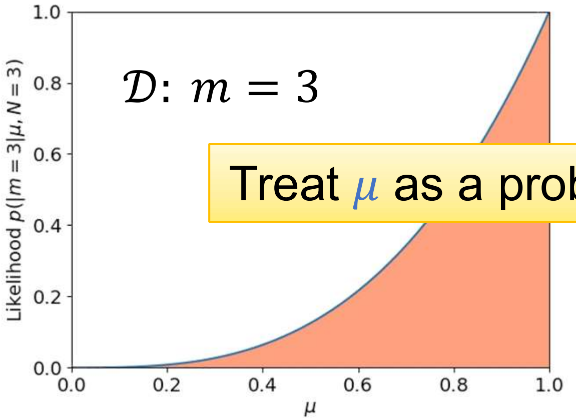


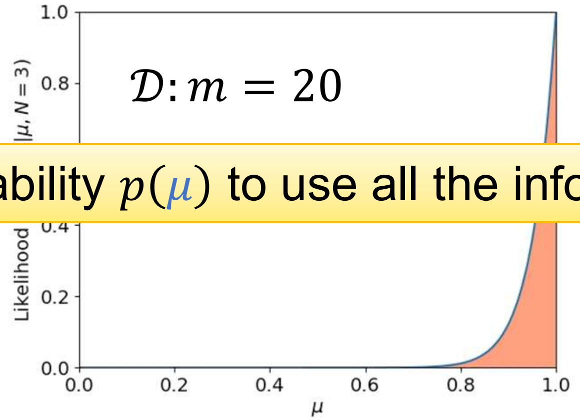
Getting more confident that $\hat{\mu} \rightarrow 1$.
$\hat{\mu}$ increases with increase of N (and m of course).

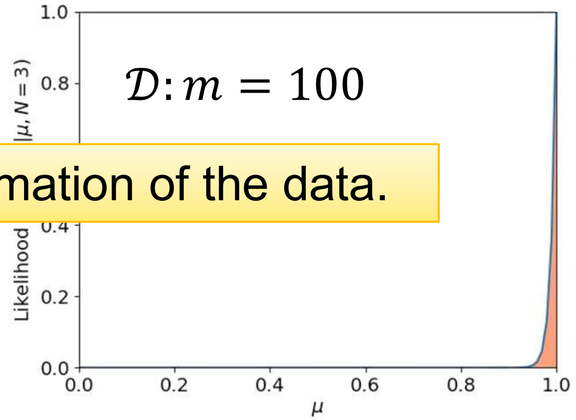
Bayesian Approach

by using all the information

$N = 3$ | $N = 20$ | $N = 100$





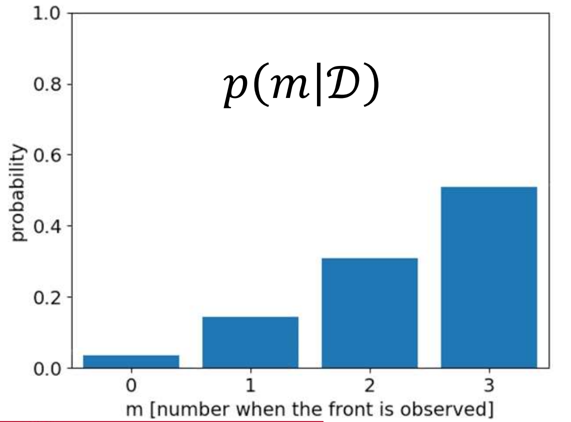


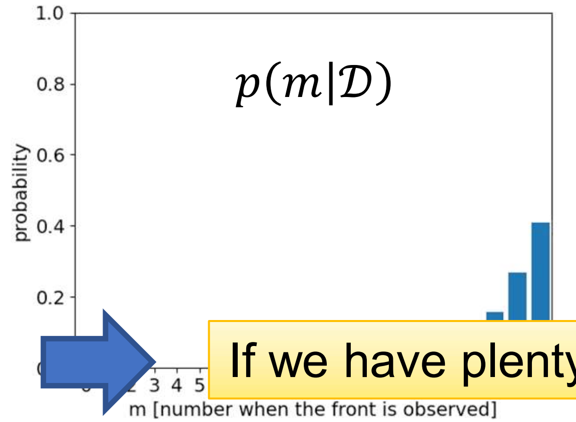
Treat μ as a probability $p(\mu)$ to use all the information of the data.

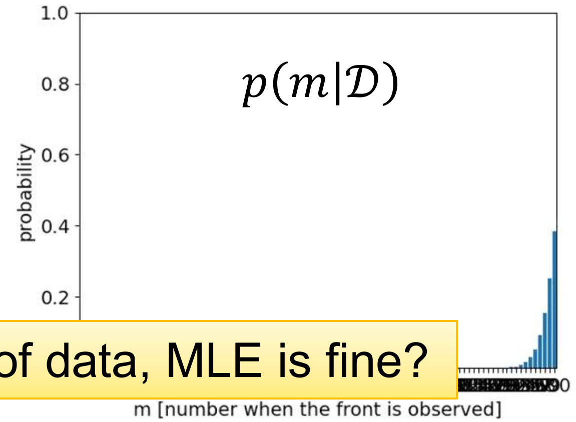
$p(\mu|m = 3)$

$p(\mu|m = 20)$

$p(\mu|m = 100)$



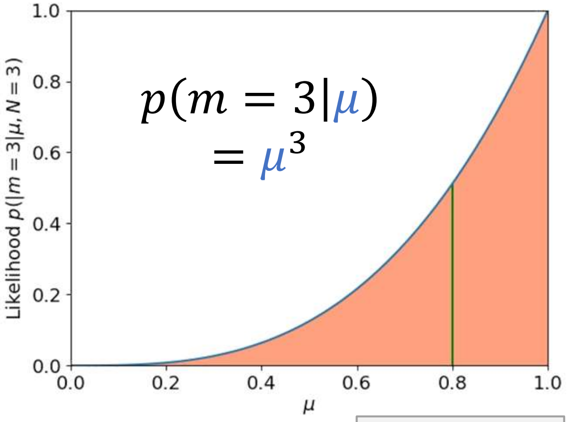




If we have plenty of data, MLE is fine?

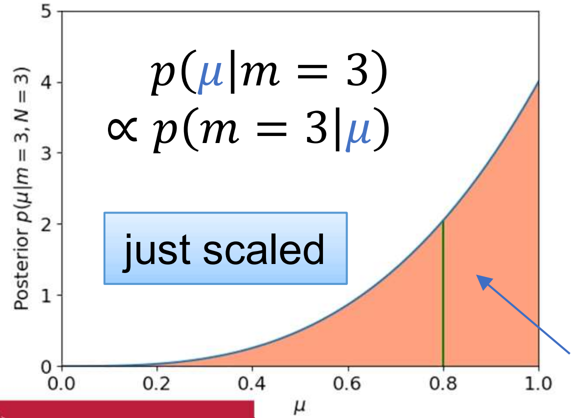
Bayesian Approach

Likelihood function



$\hat{\mu}=0.8$

Posterior function



area = 1
as probability

We have to treat $p(\mu|m=3)$, not $p(m=3|\mu)$.

Bayes' theorem

$$p(\mu|m)=\frac{p(m|\mu)p(\mu)}{p(m)}=\frac{likelihood\times prior}{p(m)}$$

$$=p(m|\mu)\times\boxed{\frac{p(\mu)}{p(m)}}$$

no info in advance

no relation with μ

constant (scaling)

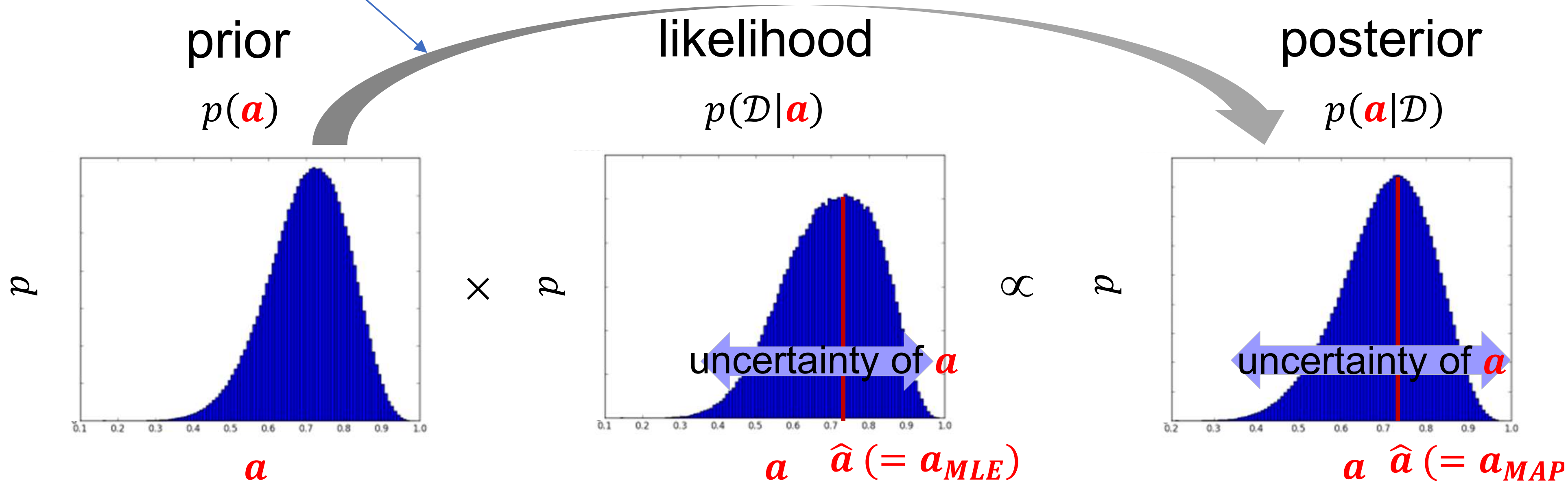
$$posterior\propto likelihood$$

more generalized

$$posterior\propto likelihood\times prior$$

Bayesian Approach

The prior (of $\boldsymbol{a}$) is **updated** to the posterior when data given.



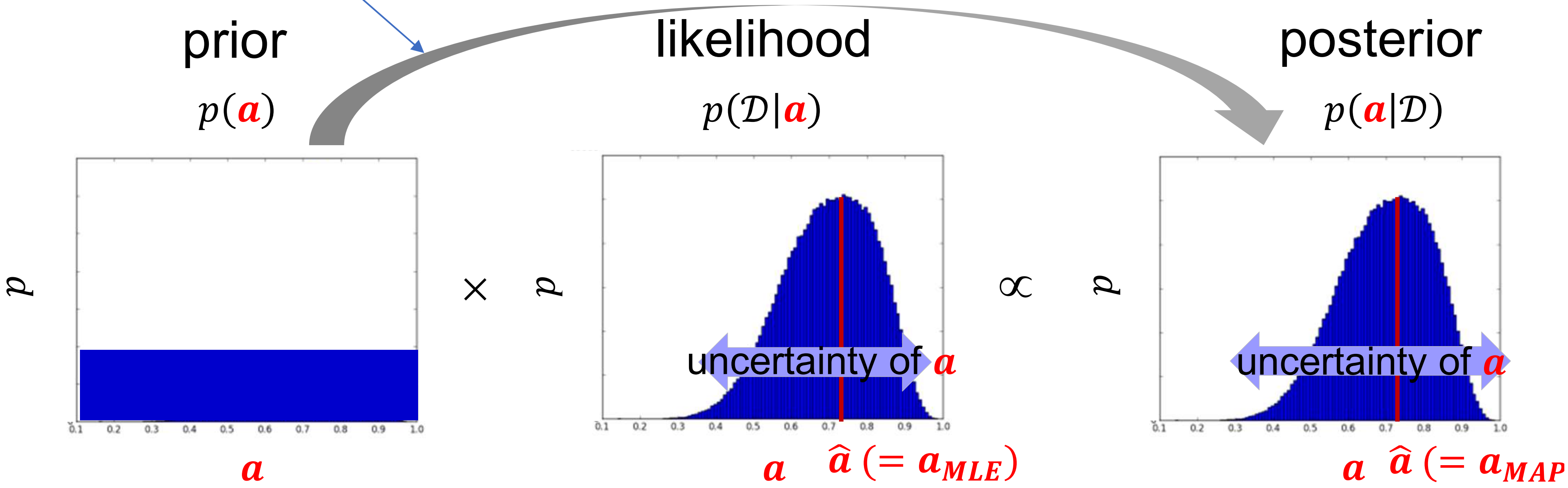
$$prior \times likelihood \propto posterior$$

Bayes' theorem: update of probability

- Everything can be treated as a probability.
- The probability is updated by information.

Bayesian Approach

The prior (of $\boldsymbol{a}$) is **updated** to the posterior when data given.



$$likelihood \propto posterior$$

Bayes' theorem: update of probability

- Everything can be treated as a probability.
- The probability is updated by information.

Bayesian Approach

Another example:

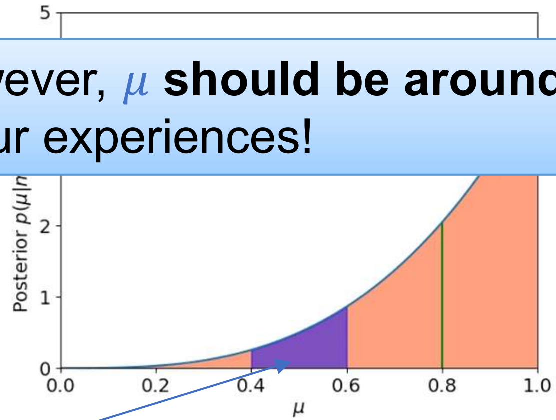
You threw a coin three times.

Observed Data: all the three times they were heads.

heads **heads** **heads**

Heads	**Tails**	
		
probability μ	probability $1-\mu$	$0 \leq \mu \leq 1$

We know several concepts to evaluate μ.

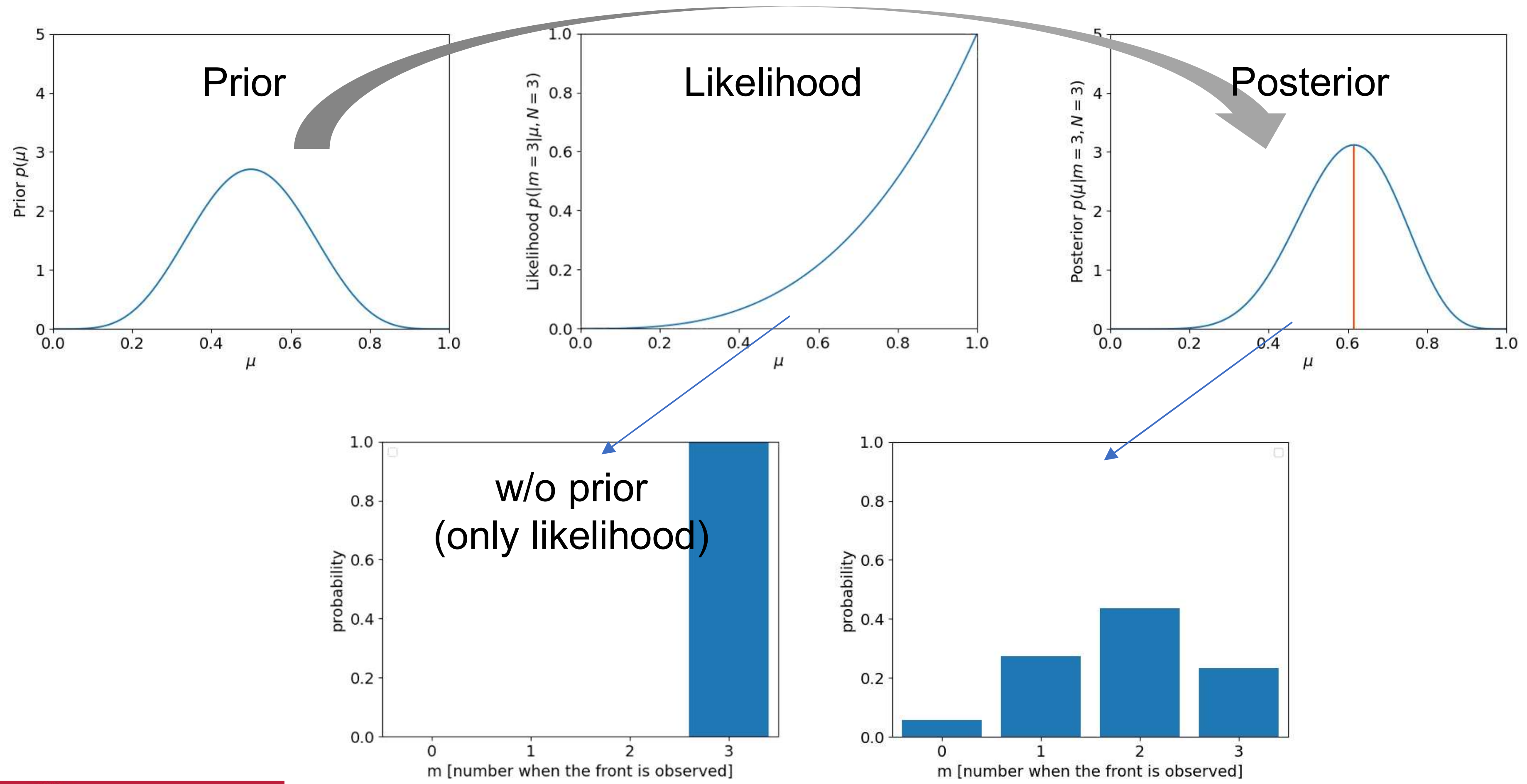


However, μ **should be around 0.5** in our experiences!

But the probability when $0.4 \leq \mu \leq 0.6$ is **only 10.4%**.

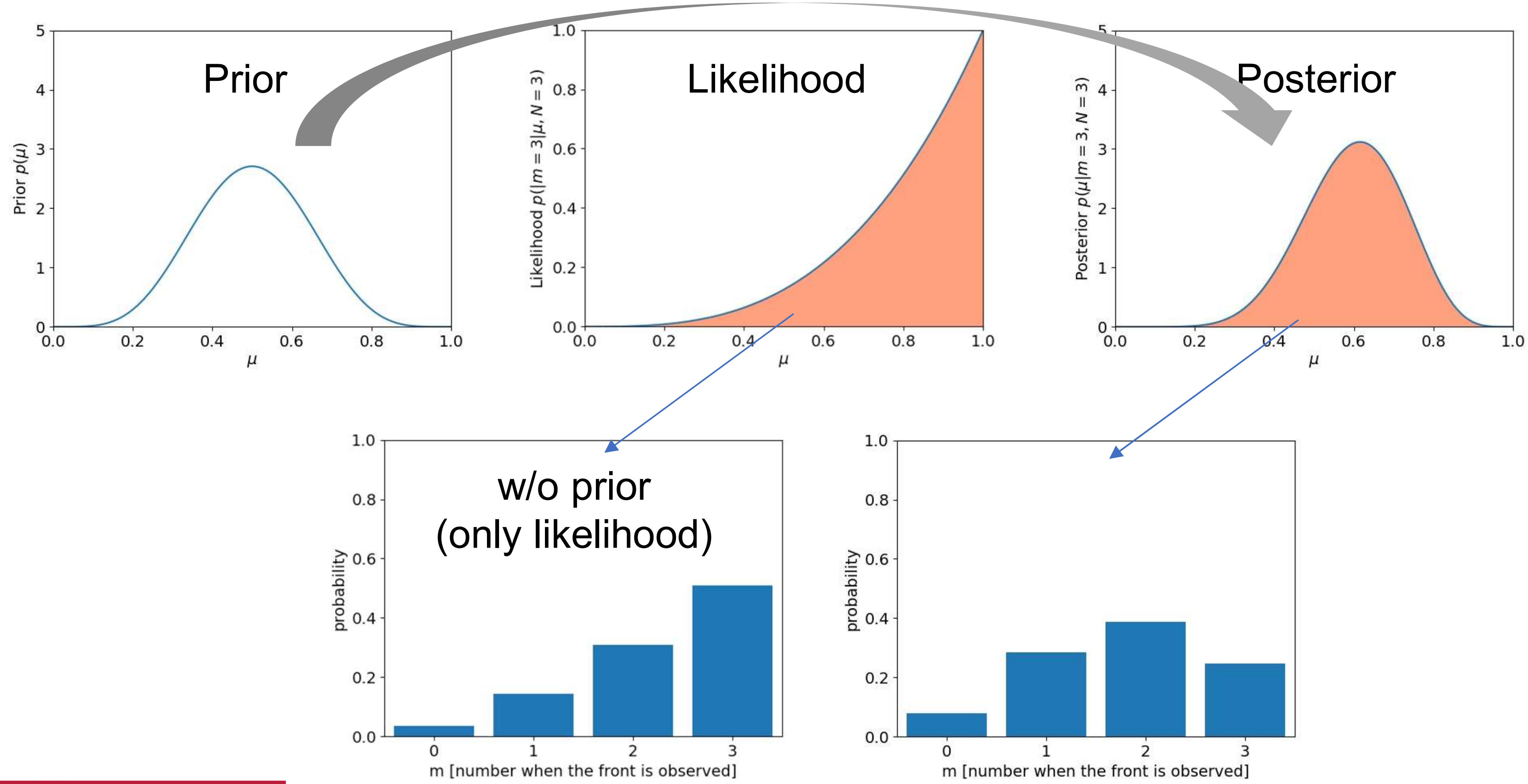
Bayesian Approach

If we have strong belief that μ is around 0.5.



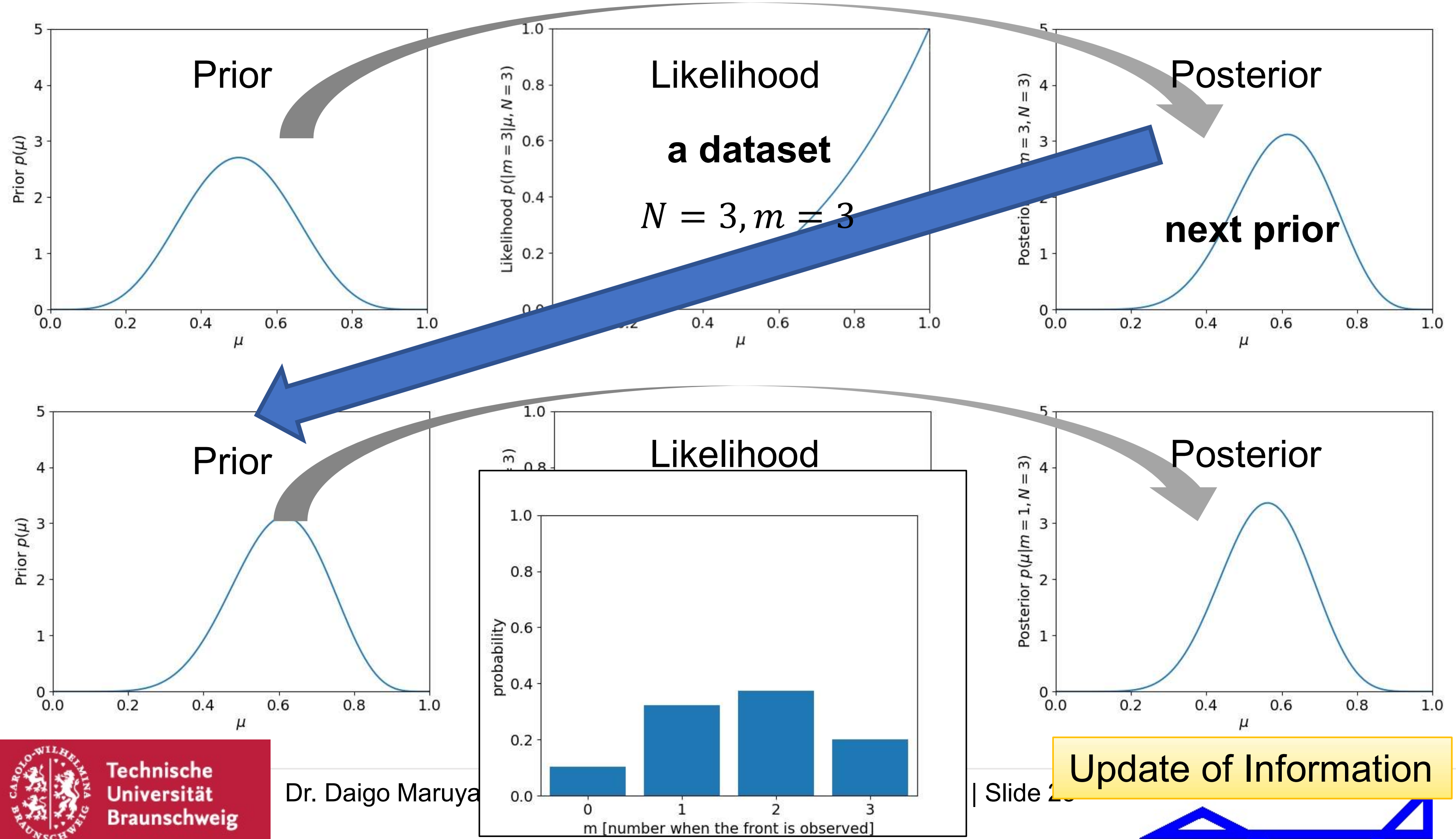
Bayesian Approach

If we have strong belief that μ is around 0.5.



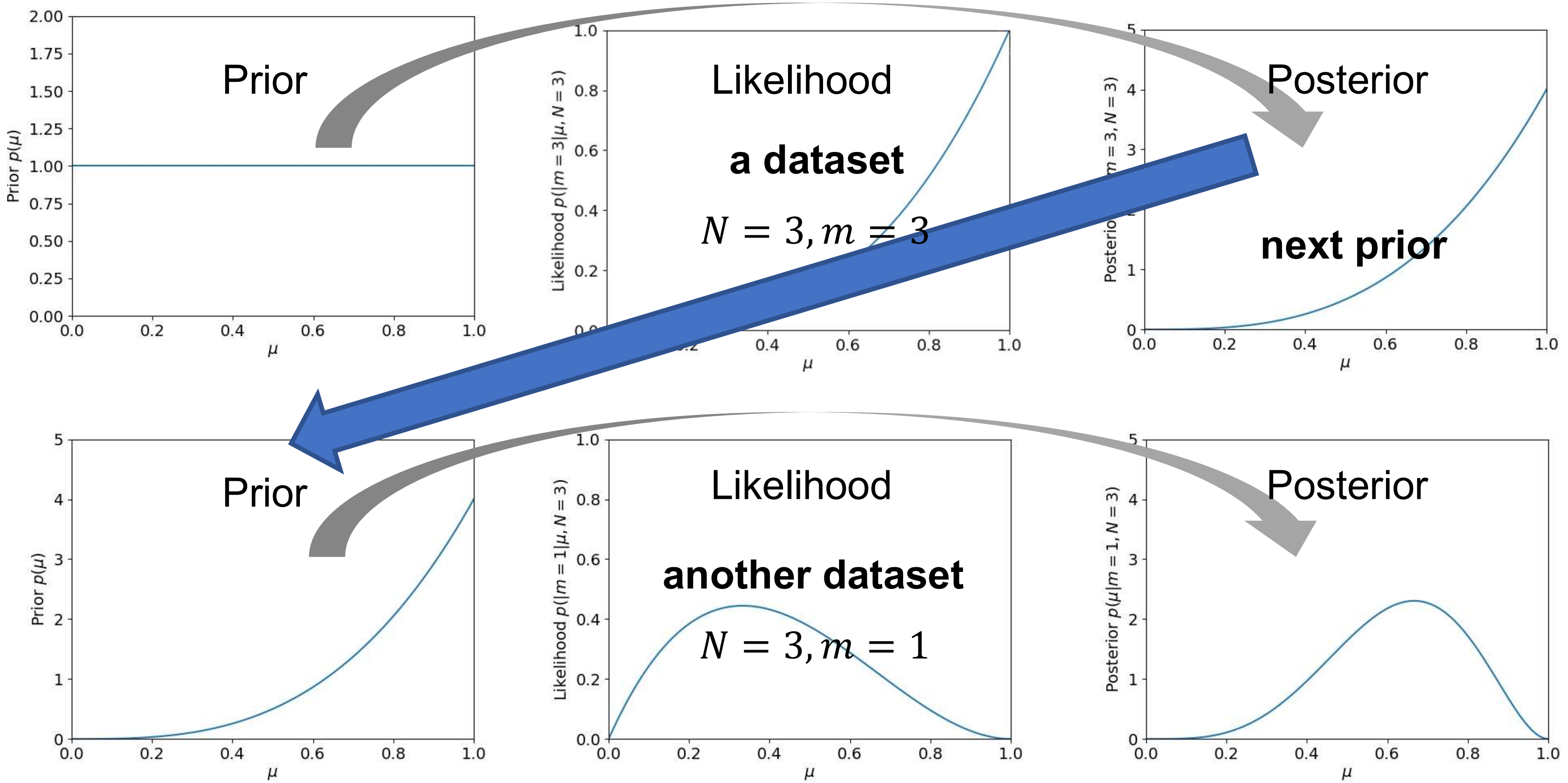
Bayesian Approach

If we have strong belief that μ is around 0.5.



Bayesian Approach

Even if we start from “no information”.



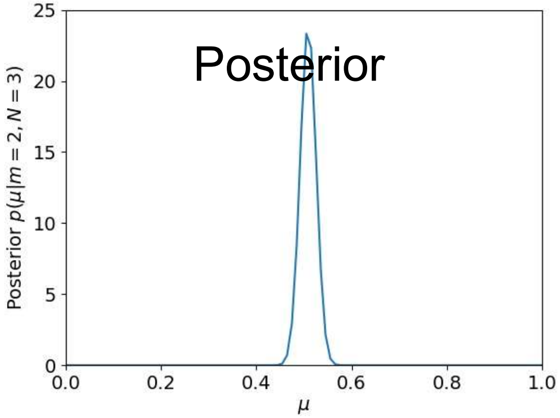
Update of Information

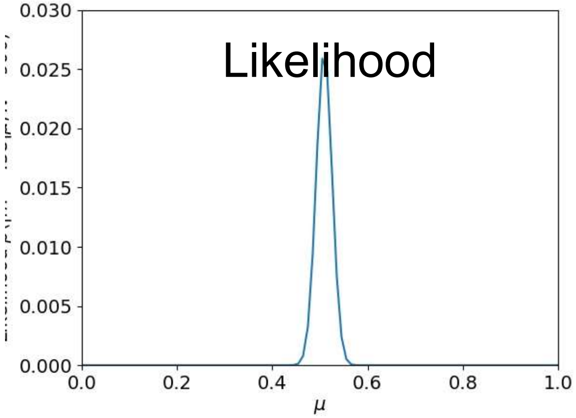
Bayesian Approach

… after 300 datasets
(300 times 3 (=900) trials) …

due to the law of large numbers

$N = 900, m = 458$

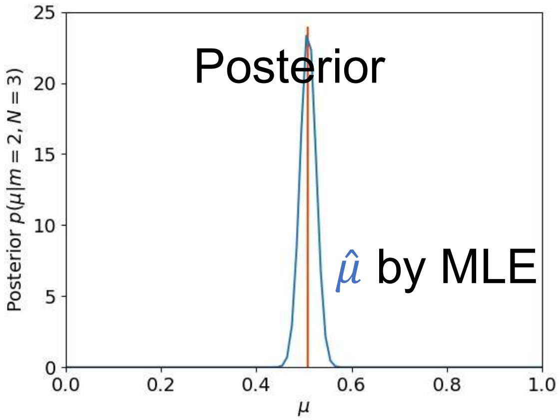


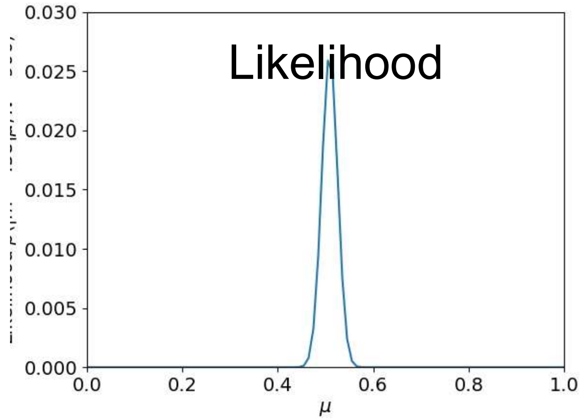


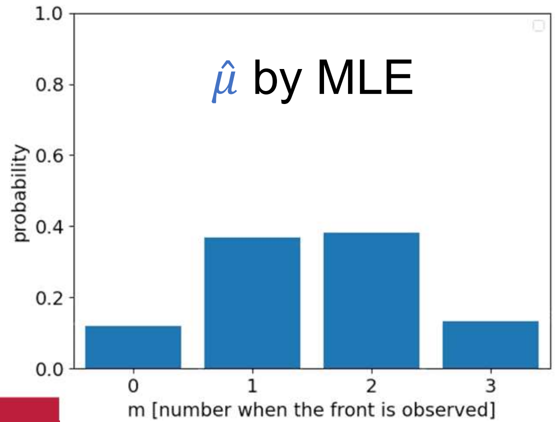
Bayesian Approach

… after 300 datasets
(300 times 3 (=900) trials) …

$N = 900, m = 458$







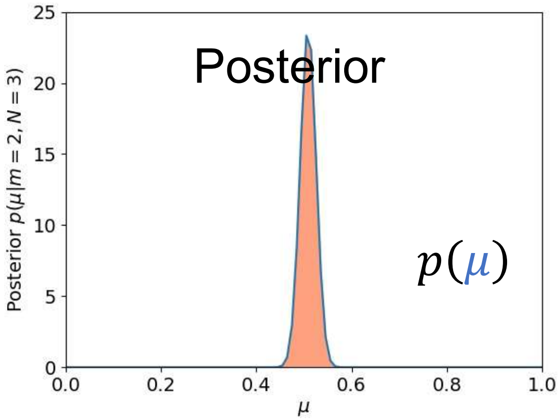
It does not influence to the result if we use,

- $\hat{\mu}$
- $p(\mu)$

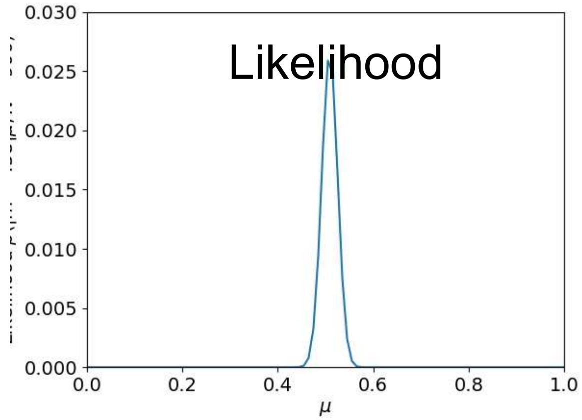
- Uncertainty
- Big data

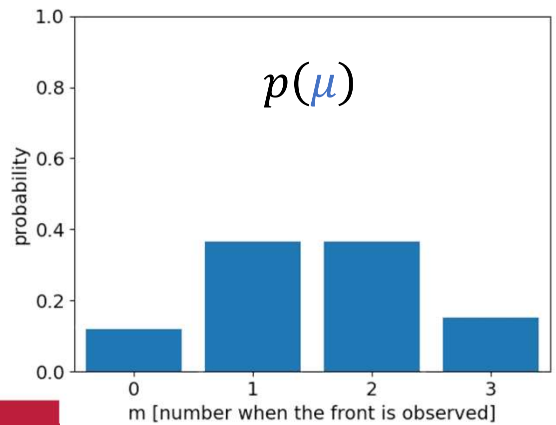
Bayesian Approach

… after 300 datasets
(300 times 3 (=900) trials) …



$N = 900, m = 458$





It does not influence to the result if we use,

- $\hat{\mu}$
- $p(\mu)$

- Uncertainty
- Big data

Bayesian Approach

Data treatment / Parameter treatment	**Batch**
Point estimate	Frequentist approach

Bayesian Approach

Data treatment / Parameter treatment	Batch	Sequential
Point estimate	Frequentist approach	Bayesian approach
Probability Distribution	Bayesian approach	Bayesian approach

Bayesian Approach

1. Define a probabilistic model
2. Then, **MLE**

A probability of m when data $\mathcal{D}$ is given

Probabilistic model

$$p(m|\mu) = \binom{3}{m} \mu^m (1-\mu)^{3-m} \quad \Rightarrow \quad p(m|\mathcal{D}) = p(m|\hat{\mu}) \quad \text{goal}$$

Likelihood function

$$\mathcal{D}: m_{Data} = 3$$

$$p(\mathcal{D}|\mu) = \mu^3 \quad \Rightarrow$$

- **Deterministic** $\hat{\mu}$ by MLE

Bayesian Approach

1. Define a probabilistic model
2. Then, **compute the posterior** (several choices)

Probabilistic model

$$p(m|\mu) = \binom{3}{m} \mu^m (1-\mu)^{3-m}$$

$$p(m|\hat{\mu})$$

$$p(m|\mathcal{D}) = \boxed{\int p(m|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu}$$

Why this formulation?

Likelihood function

$\mathcal{D}: m_{Data} = 3$

$p(\mathcal{D}|\mu) = \mu^3$

Prior distribution

$p(\mu)$: arbitrary

Posterior distribution $p(\mu|\mathcal{D})$

- **Deterministic** $\hat{\mu}$ — by MLE / by mean
  Point estimate
- **Stochastic** $p(\mu)$
  Probability distribution

Probability Theory (Rules of Probability) – Review Lecture 2

marginal distribution

sum rule $p(y) = \int p(x,y)\mathrm{d}x$

product rule $p(x,y) = p(x|y)p(y)$

$$p(y) = \int p(x,y)\mathrm{d}x$$

$$= \int p(y|x)p(x)\mathrm{d}x$$

Bayes' theorem $p(y|x) = \dfrac{p(x|y)p(y)}{p(x)}$

We need to get familiar with this transformation process.

Bayesian Approach

$$p(y) = \int p(y|x)p(x)\mathrm{d}x$$

by the sum and product rules

$y \rightarrow m$
$x \rightarrow \mu$

$$p(m) = \int \underbrace{p(m|\mu)}_{\text{Probabilistic model}} \underbrace{p(\mu)}_{\text{A probability of } \mu} \mathrm{d}\mu$$

Probabilistic model

A probability of μ
prior
posterior

$p(\mu) \rightarrow p(\mu|\mathcal{D})$

Posterior

$$p(m|\mathcal{D}) = \int p(m|\mu) \underbrace{p(\mu|\mathcal{D})}_{\text{A probability of } \mu \text{ when data } \mathcal{D} \text{ is given}} \mathrm{d}\mu$$

A probability of μ when data $\mathcal{D}$ is given

A probability of m
when data $\mathcal{D}$ is given

Bayesian Approach

1. Define a probabilistic model
2. Then, **compute the posterior** (several choices)

parametric probability distributions (see Lecture 3)

Your model for the output,
which is parametrized by parameters

1. Define a probabilistic model

$$p(output) = \int (Probabilistic\ Model) \times (Posterior)\, \mathrm{d}(parameter)$$

Information of the parameters - a probability distribution
(obtained by data)

Predictive distribution (goal)

2. Then, **compute the posterior**

Bayesian Approach

Another Meaning / Aspect / Interpretation

$$p(m|\mathcal{D}) = \int p(m|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu$$

$$= \int p(m,\mu|\mathcal{D})\mathrm{d}\mu$$

REVIEW of Lecture 2:
Meaning of
- Joint distribution
- Marginal distribution

1. Clarify all the stochastic variables ➡ $\underset{\text{output}}{m}$ and $\underset{\text{parameter}}{\mu}$

2. Try to fine **the joint probability** $p(m,\mu)$, which include all the information.

3. Our objective is only **the marginal distribution** $p(m)$ $\quad p(m) = \int p(m,\mu)\mathrm{d}\mu$

We are not interested in the parameter μ anymore.

Bayesian Approach

$$p(m|\mathcal{D}) = \int p(m|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu$$

We know the **probabilistic model** $p(m|\mu)$ (because “we” defined it by ourselves).
But how can we obtain the **posterior** $p(\mu|\mathcal{D})$?

We know the **likelihood function** $p(\mathcal{D}|\mu)$.

We have been computing this to maximize it.

Posterior

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

Likelihood $p(\mathcal{D}|\mu) = \mu^3$

Prior $p(\mu)$: arbitrary. $\int p(\mu)\mathrm{d}\mu = 1$

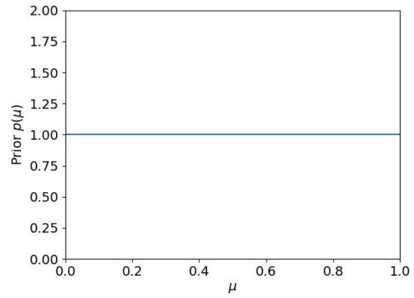
Bayesian Approach

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}$$
$$= \frac{\mu^3 \times 1}{p(\mathcal{D})}$$
$$= 4\mu^3$$

$$\int p(\mu|\mathcal{D})\mathrm{d}\mu = 1$$

$p(\mu)$ can be arbitrary. e.g. $p(\mu) = 1$

$$\int p(\mu)\mathrm{d}\mu = 1$$
$$0 \leq \mu \leq 1$$



The posterior $p(\mu|\mathcal{D})$ can be obtained without computing $p(\mathcal{D})$.

Actually $p(\mathcal{D})$ can be computed.

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu$$
$$= \int(\mu^3 \times 1)\mathrm{d}\mu = \frac{1}{4}$$

Bayes' theorem
(another expression)

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}$$

by the sum and product rules

Bayesian Approach

$$p(m|\mathcal{D}) = \int p(m|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu$$

$$= \int \binom{3}{m} \mu$$

The output distributions are non closed-form in real applications.
Exceptions: Gaussian Processes (Lectures 7,8)

$$= 4\binom{3}{m} \underbrace{\int \mu^{3+m}(1-\mu)^{3-m}\mathrm{d}\mu}_{\text{any closed-form expressions?}}$$

$$= \cdots$$

$$p(m|\mathcal{D}) = \iiint \cdots \mathrm{d}\mu_1 \mathrm{d}\mu_2 \mathrm{d}\mu_3 \cdots$$

In practical applications, the number of the parameters is large and $p(m|\mathcal{D})$ requires multiple integral.

In practice, numerical approaches are used.
(e.g. MCMC presented in Lecture 2 and Lecture 4
– explained in Lecture 12)

The results in the slide 16 (and the above equation) was obtained by using the MCMC.

Bayesian Approach

Let's try to apply this concept to **the curve fitting problem**.

Probabilistic model

$$p(t|x,\boldsymbol{w},\sigma) = \mathcal{N}(t|y(x,\boldsymbol{w}),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\{t-y(x,\boldsymbol{w})\}^2}{2\sigma^2}\right\}$$

Likelihood function

$$p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma) = \prod_{i=1}^{N}\mathcal{N}(t_i|y(x_i,\boldsymbol{w}),\sigma^2) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{\{t_i-y(x_i,\boldsymbol{w})\}^2}{2\sigma^2}\right]$$

Posterior distribution $p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t})$

Predictive distribution

$$p(t|x,\mathbf{x},\mathbf{t}) = \underbrace{\iint p(t|x,\boldsymbol{w},\sigma)p(\boldsymbol{w},\sigma|\mathbf{x},\mathbf{t})\mathrm{d}\boldsymbol{w}\mathrm{d}\sigma}_{\text{probabilistic model} \times \text{posterior}} = ?$$

Summary

Process
1. Define a probabilistic model
2. Then, **compute the posterior** (several choices)
   - Point estimate
   - Probability distribution ← computations hard

Next lecture
- Connections with the regularization techniques
- Uncertainty
- Large numbers of dataset
- etc.

Using the curve fitting problem

A generalized perspective by the Bayes approach