

Machine Learning

Lecture 2: Probabilistic Inference

Prof. Dr. Stephan Günnemann

Data Mining and Analytics
Technische Universität München

29.10.2018

Reading material

- Murphy [ch. 3.1 - 3.3]

Acknowledgements

- Slides are based on an older version by M. Sölch

We flip the same coin 10 times:



Probability that the next coin flip is **T**?

We flip the same coin 10 times:



Probability that the next coin flip is **T**?

~ 0 ~ 0.3 ~ 0.38 ~ 0.5 ~ 0.76 ~ 1

↳ 100%
Fail

30%?

This seems reasonable, but why?

30%?

This seems reasonable, but why?

Every flip is random. So every sequence of flips is random, i.e., it has some probability to be observed.

30%?

This seems reasonable, but why?

Every flip is random. So every sequence of flips is random, i.e., it has some probability to be observed.

For the i -th coin flip we write

$$p_i(F_i = \text{T}) = \theta_i$$

30%?

This seems reasonable, but why?

Every flip is random. So every sequence of flips is random, i.e., it has some probability to be observed.

For the i -th coin flip we write

$$p_i(F_i = \text{T}) = \theta_i$$

To denote that the probability distribution depends on θ_i , we write

$$p_i(F_i = \text{T} \mid \theta_i) = \text{Ber}(F_i = \text{T} \mid \theta_i) = \theta_i$$

i.e. $F_i \sim \text{Ber}(\theta_i)$

Note the i in the index! We are trying to reason about θ_{11} .

All the randomness of a sequence of flips is governed (*modeled*) by the parameters $\theta_1, \dots, \theta_{10}$:

$$p(\text{H T H H T H H H T H} \mid \theta_1, \theta_2, \dots, \theta_{10})$$

What do we know about $\theta_1, \dots, \theta_{10}$? Can we infer something about θ_{11} ?
At first sight, there is no connection.

Find θ_i 's such that that $p(\text{H T H H T H H H T H} \mid \theta_1, \theta_2, \dots, \theta_{10})$ is as high as possible. This is a very important principle:

Maximise the *likelihood* of our observation. (*Maximum Likelihood*)

$$???? \quad p(\text{H T H H T H H H T H} \mid \theta_1, \theta_2, \dots, \theta_{10}) \quad ????$$

We need to model this.

First assumption: The coin flips do not affect each other—independence.

$$\begin{aligned} & p(\text{H T H H T H H H T H} \mid \theta_1, \theta_2, \dots, \theta_{10}) \\ &= p_1(F_1 = \text{H} \mid \theta_1) \cdot p_2(F_2 = \text{T} \mid \theta_2) \cdot \dots \cdot p_{10}(F_{10} = \text{H} \mid \theta_{10}) \\ &= \prod_{i=1}^{10} p_i(F_i = f_i \mid \theta_i) \end{aligned}$$

Notice the i in p_i, θ_i ! This indicates: The coin flip at time 1 is different from the one at time 2, ...

But the coin does not change.

Second assumption: The flips are qualitatively the same—identical distribution.

$$\prod_{i=1}^{10} p_i(F_i = f_i \mid \theta_i) = \prod_{i=1}^{10} p(F_i = f_i \mid \theta)$$

In total: The 10 flips are *independent and identically distributed (i.i.d.)*.

Remember θ_{11} ? With the i.i.d. assumption we can link it to $\theta_1, \dots, \theta_{10}$.

Now we can write down the probability of our sequence with respect to θ :

$$\begin{aligned} \prod_{i=1}^{10} p(F_i = f_i \mid \theta) &= (1 - \theta)\theta(1 - \theta)(1 - \theta)\theta(1 - \theta)(1 - \theta)(1 - \theta)\theta(1 - \theta) \\ &= \theta^3(1 - \theta)^7 \end{aligned}$$

Under our model assumptions (i.i.d.):

$$p(\text{H T H H T H H H T H} \mid \theta) = \theta^3(1 - \theta)^7$$

Under our model assumptions (i.i.d.):

$$p(\text{H T H H T H H H T H} \mid \theta) = \theta^3(1 - \theta)^7$$

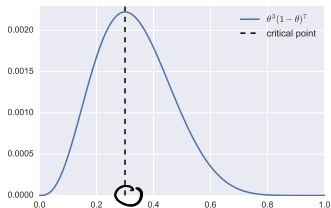
$\frac{3}{3+7} = 0.3 = \theta_{MLE} = \underset{\theta}{\text{ARGMAX}} \theta^3(1 - \theta)^7$

This can be interpreted as a function $f(\theta)$. We want to find the maxima (maximum likelihood) of this function.

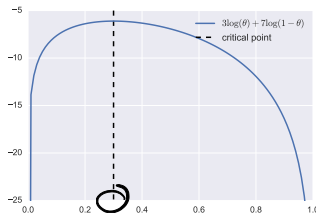
High-school math! Take the derivative $df/d\theta$, set it to 0, and solve for θ . Check these *critical points* by inserting them into the second derivative.

In principle, this is easy. But the second derivative of $f(\theta)$ is already ugly.

Luckily, monotonic functions preserve critical points.



$f(\theta)$

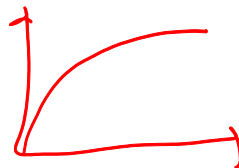


$\log(f(\theta))$

$f(\theta)$ has the same maxima as $\log f(\theta)$

$$\theta_{\text{MLE}} = 0.3$$



$$\begin{aligned} & \underset{x}{\text{ARGMAX}} f(x) \\ &= \underset{x}{\text{ARGMAX}} 5 \cdot f(x) \end{aligned}$$




Maximum Likelihood Estimation (MLE) for any coin sequence?

Maximum Likelihood Estimation (MLE) for any coin sequence?

$$\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$$

$|T|, |H|$ denote number of , , respectively.

Remember we wanted to find the probability the next coin flip is 

$$F_{11} \sim \text{Ber}(\theta_{\text{MLE}}) \quad p(F_{11} = \text{T}) = \text{Ber}(\theta_{\text{MLE}}) = \theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$$

This justifies 30% as a reasonable answer to our initial question.
Problem solved?!

Just for fun, a totally different sequence (*same coin!*):



$$\theta_{\text{MLE}} = 0.$$

But even a fair coin ($\theta = 0.5$) has 25% chance of showing this result!

We have *prior beliefs* that have nothing to do with math. Is there any chance to incorporate them?

Yes: Make the parameter θ itself a random variable.

$$\theta \sim p(\cdot)$$

LIKELIHOOD

We were asking the wrong question...

Instead of maximizing $p(\mathcal{D} \mid \theta)$ we are actually interested in $p(\theta \mid \mathcal{D})$

For *any* x , we want to be able to express $p(\theta = x \mid \mathcal{D})$, where θ is a *random variable* that models the observed sequence at hand ($\mathcal{D} = \text{Data}$).

$$\theta^3 (1-\theta)^2$$

We were asking the wrong question...

Instead of maximizing $p(\mathcal{D} \mid \theta)$ we are actually interested in $p(\theta \mid \mathcal{D})$

For *any* x , we want to be able to express $p(\theta = x \mid \mathcal{D})$, where θ is a *random variable* that models the observed sequence at hand ($\mathcal{D} = \text{Data}$).

Because we are talking about coin flips, we know that $p(\theta = x \mid \mathcal{D})$ only *makes sense* for $x \in [0, 1]$.

By Bayes' rule:

$$p(\theta = x \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta = x) \cdot p(\theta = x)}{p(\mathcal{D})}$$

The numerator consists of:

- $p(\mathcal{D} \mid \theta = x)$: We know this one from MLE, now with a fixed $\theta = x$! It is called *likelihood*.
- $p(\theta = x)$: Our prior belief in the value of θ before observing data. It is called *the prior*.

The denominator $p(\mathcal{D})$ is called the *evidence*

We call $p(\theta = x \mid \mathcal{D})$ the *posterior* (distribution) – i.e., our belief in the value of θ *after* observing data.

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}$$

How should we choose the prior $p(\theta = x)$?

There are no constraints except that:

(1) It **must** not depend on the data.

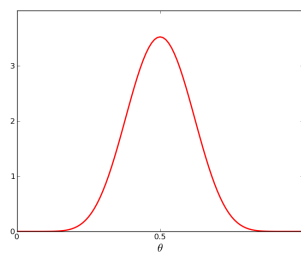
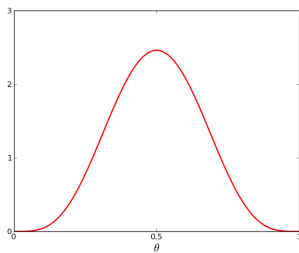
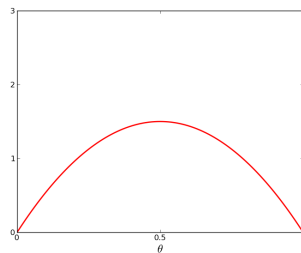
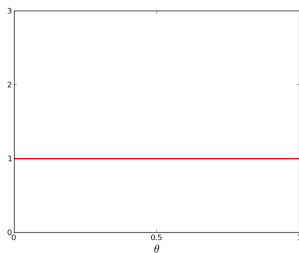
(2) $p(\theta = x) \geq 0 \quad \forall x$

(3) $\int p(\theta = x) dx = 1$

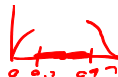
(2) and (3) have to hold on the support (i.e., feasible values) of θ .

This leaves room for (possibly subjective) model choices!

Some possible choices for the prior on θ :



MLE corresponds to having uniform prior



Question: What happens if $p(\theta = x) = 0$ for some particular x ?

Question: What happens if $p(\theta = x) = 0$ for some particular x ?

Recall:

$$\begin{array}{ccccccc} \text{posterior} & \propto & \text{likelihood} & \cdot & \text{prior} \\ p(\theta = x \mid \mathcal{D}) & \propto & p(\mathcal{D} \mid \theta = x) & \cdot & p(\theta = x) \end{array}$$

$$f(x) \cdot g(x)$$

$$f(x) = 2 \cdot x$$

$$g(x) = x + 3$$

Question: What happens if $p(\theta = x) = 0$ for some particular x ?

Recall:

posterior	\propto	likelihood	\cdot	prior
$p(\theta = x \mid \mathcal{D})$	\propto	$p(\mathcal{D} \mid \theta = x)$	\cdot	$p(\theta = x)$
BETA		Bernoulli	\cdot	BETA

Posterior will always be zero for that particular x regardless of the likelihood/data

Often, you choose the prior to make subsequent calculations easier.

$$\text{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}, \quad \theta \in [0, 1]$$

Likelihood
 $\theta^2 (1-\theta)^7$

Often, you choose the prior to make subsequent calculations easier.

$$\text{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}, \quad \theta \in [0, 1]$$

Note:

$\theta^3 \cdot (1-\theta)^7$

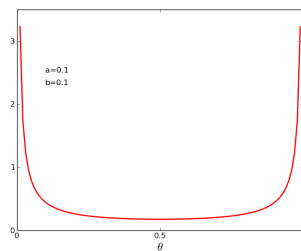
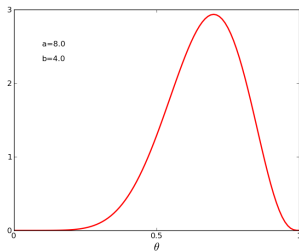
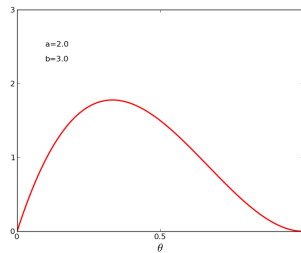
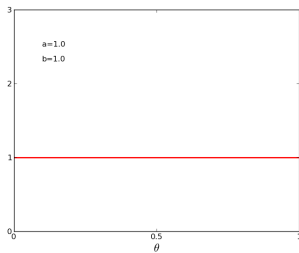
- We have seen this part before: $\theta^{a-1} (1-\theta)^{b-1}$
- $\Gamma(n) = (n-1)!$, if $n \in \mathbb{N}$

The fact that we have seen this before in parts is not a coincidence. We have chosen a *conjugate prior*, in this case the Beta distribution, that eases further computation.

In the MLE section, we had a similar functional form, but $(a-1)$ and $(b-1)$ were substituted by the number of **T** and **H**, respectively!

Interpretation: $a-1$ and $b-1$ are the numbers of **T** and **H** we think we would see, if we made $a+b-2$ many coin flips.

The Beta pdf for different choices of a and b :



Let us plug all we know into Bayes' Theorem:

$$p(\theta = x \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta = x) \cdot p(\theta = x)}{p(\mathcal{D})}$$

Let us plug all we know into Bayes' Theorem:

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

We know

$$p(\mathcal{D} | \theta = x) = x^{|\mathcal{T}|}(1-x)^{|\mathcal{H}|},$$
$$p(\theta = x) = p(\theta = x | a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}.$$

So we get:

$$p(\theta = x | \mathcal{D}) = \frac{1}{p(\mathcal{D})} \cdot x^{|\mathcal{T}|}(1-x)^{|\mathcal{H}|} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$
$$\propto x^{|\mathcal{T}|+a-1}(1-x)^{|\mathcal{H}|+b-1},$$

because $p(\mathcal{D})$ is constant w.r.t. x .

$$\text{Ans map } p(\theta | \mathcal{D}) = \frac{|\mathcal{T}|+a-1}{|\mathcal{T}|+|\mathcal{H}|+a+b-2} = \theta_{\text{MAP}}$$

We can find the maximum θ_{MAP} with the same algebra as in the MLE case:

$$\theta_{\text{MAP}} = \frac{|T| + a - 1}{|H| + |T| + a + b - 2}$$

Under our prior belief and after seeing the (i.i.d.) data, θ_{MAP} is the best guess for θ .

It is called the *maximum a posteriori estimate*. (MAP)

$$F_{11} \sim \text{Ber}(\theta_{\text{MAP}}) \quad p(F_{11} = \text{Ⓣ}) = \text{Ber}(\theta_{\text{MAP}}) = \theta_{\text{MAP}}$$

Nota bene: Remember that monotonic functions preserve critical points. Multiplication with a constant ($\neq 0$) is monotonic. For obtaining the minimum, it was not necessary to calculate the normalizing constant.

$$p(\theta = x \mid \mathcal{D}) \propto x^{|T|+a-1} (1-x)^{|H|+b-1} \cdot c$$

How do we find the multiplicative constant that turns “ \propto ” into “=”?

We have one more constraint: $\int p(\theta \mid \mathcal{D}) d\theta = 1$.

The right-hand side is proportional to a Beta pdf, which integrates to 1.

Consequently (by reverse-engineering), the posterior must also be Beta-distributed and the only constant that works is

$$c = \frac{\Gamma(|H| + |T| + a + b)}{\Gamma(|T| + a) \Gamma(|H| + b)} \quad \int_0^1 f(x) = \frac{1}{c}$$

Then we have

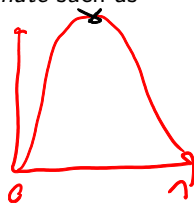
$$p(\theta = x \mid \mathcal{D}) = \text{Beta}(x \mid a + |T|, b + |H|)$$

Always remember this trick when you try to solve integrals that involve known pdfs (up to a constant factor)!

Now we have a full distribution over θ , rather than *point estimate* such as θ_{MLE} and θ_{MAP}

$$p(\theta = x \mid \mathcal{D}) = \text{Beta}(x \mid a + |T|, b + |H|),$$

i.e. we have a full Bayesian treatment of θ



Knowing that the mode of $\text{Beta}(\alpha, \beta)$ is $\frac{\alpha-1}{\alpha+\beta-2}$, for $\alpha, \beta > 1$

$$\frac{a+T-1}{a+T+b+H-2}$$

You can see that $\theta_{\text{MAP}} = \frac{|T|+a-1}{|H|+|T|+a+b-2}$ is just the mode of the posterior distribution

And using the uniform prior (i.e. $a = b = 1$) we obtain the MLE solution:

$$\theta_{\text{MLE}} = \frac{|T|+1-1}{|H|+|T|+1+1-2} = \frac{|T|}{|H|+|T|}$$

Remember we want to predict the next coin flip...

The probability that the next coin flip is **T**, given observations \mathcal{D} and prior belief a, b :

$$p(F = \textbf{T} \mid \mathcal{D}, a, b)$$

This is called the *posterior predictive* distribution

Remember we want to predict the next coin flip...

The probability that the next coin flip is **T**, given observations \mathcal{D} and prior belief a, b :

$$p(F = \textbf{T} \mid \mathcal{D}, a, b)$$

This is called the *posterior predictive* distribution

A flip depends on θ , but we don't know what specific value for θ we should use. So integrate over *all* possible values of θ !

$$p(F = \textbf{T} \mid \mathcal{D}, a, b) = \int_0^1 p(F = \textbf{T}, \theta \mid \mathcal{D}, a, b) d\theta$$

How is this different from before? We use *all possibilities*, weighted by the posterior at the same time, rather than just the ML or MAP estimate.

We call this a (*fully*) *Bayesian* analysis, because rather than optimizing parameters, we integrate them out.

To calculate the integral from the previous slide, we rewrite the coin flip (Bernoulli) density:

$$p(F = f \mid \theta) = p(f \mid \theta) = \theta^f (1 - \theta)^{1-f}$$

(Here, f is boolean with values 1 for tails and 0 for heads.)

$$p(x) = \int_{\mathcal{Y}} p(x, y) dy$$

BERNOULLI:

$\mathcal{D} \leadsto \theta_{MAP}$

$p(f \mid \theta_{MAP})$

$$\begin{aligned} p(f \mid \mathcal{D}, a, b) &= \int_0^1 p(f, \theta \mid \mathcal{D}, a, b) d\theta \\ &= \int_0^1 p(f \mid \theta, \mathcal{D}, a, b) p(\theta \mid \mathcal{D}, a, b) d\theta \\ &= \int_0^1 p(f \mid \theta) p(\theta \mid \mathcal{D}, a, b) d\theta \end{aligned}$$

Product Rule + conditional independence assumption!

POSTERIOR
DISTRIBUTION
[OVER θ]

Continue by plugging in all formulas we get

$$\begin{aligned}
 p(f | \mathcal{D}, a, b) &= \int_0^1 p(f | \theta) p(\theta | \mathcal{D}, a, b) d\theta \\
 &= \int_0^1 \theta^f (1 - \theta)^{1-f} \frac{\Gamma(|T| + a + |H| + b)}{\Gamma(|T| + a) \Gamma(|H| + b)} \theta^{|T|+a-1} (1 - \theta)^{|H|+b-1} d\theta \\
 &= \frac{\Gamma(|T| + a + |H| + b)}{\Gamma(|T| + a) \Gamma(|H| + b)} \int_0^1 \theta^{f+|T|+a-1} (1 - \theta)^{|H|+b-f-1} d\theta \\
 &= \frac{\Gamma(|T| + a + |H| + b)}{\Gamma(|T| + a) \Gamma(|H| + b)} \frac{\Gamma(f + |T| + a) \Gamma(|H| + b - f + 1)}{\Gamma(|T| + a + |H| + b + 1)} \\
 &= \frac{(|T| + a)^f (|H| + b)^{(1-f)}}{|T| + a + |H| + b}
 \end{aligned}$$

Handwritten notes:

- $\int_0^1 f(x)$ (red)
- $f(x) \propto \text{BETA}$ (red)
- $\int_0^1 f(x) = 1/c$ (red)
- $p(f=1) = \frac{T+a}{T+a+H+b}$ (red)

We could also put a *hyperprior* on the parameters a and b . But not now...

$$\int_0^1 c \cdot f(x) = 1 \Leftrightarrow \int_0^1 f(x) = \frac{1}{c}$$

We can find the solution faster if we are clever...

Lets consider only $p(F = \textcircled{T} \mid \mathcal{D}, a, b)$

BETA($\theta \mid T+a, H+b$)

$$\begin{aligned} & \int_0^1 p(F = \textcircled{T} \mid \theta) p(\theta \mid \mathcal{D}, a, b) d\theta \\ &= \int_0^1 \theta \cdot p(\theta \mid \mathcal{D}, a, b) d\theta \\ &= \mathbb{E}_p[\theta] \quad // \text{ Expectation w.r.t. the posterior} \\ &= \frac{|T| + a}{|T| + a + |H| + b} \end{aligned}$$

BETA
 $E[x] = \frac{\alpha}{\alpha + \beta}$

Which is the same as $\frac{(|T|+a)^f (|H|+b)^{(1-f)}}{|T|+a+|H|+b}$ with $f = 1$

Caution! Only true for this particular case.

The original question was:

Given our 10 flips (\mathcal{D}), what is the probability of the next flip showing tails?

$$\rightsquigarrow p(F_{11} = \text{T} \mid \mathcal{D})$$

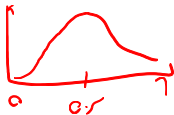
No θ !

We simplified our problem to point estimates of parameters rather than probabilities. MAP used a distribution on parameters, but eventually only maximized it.

Only Fully Bayesian analysis answered the original question:

$$p(F = \text{T} \mid \mathcal{D}, a, b) = \int_0^1 p(F = \text{T}, \theta \mid \mathcal{D}, a, b) d\theta$$

Example...



HERE FOR HEADS

Given the prior $a = b = 5$, and the counts $|H| = 6$, $|T| = 8$

$$\theta_{\text{MLE}} = \frac{6}{14} \approx 0.43 \quad \theta_{\text{MAP}} = \frac{10}{22} \approx 0.45 \quad \theta_{\text{FB}} = \frac{11}{24} \approx 0.46$$

How about if we have $|H| = 304$, $|T| = 306$?

$$\theta_{\text{MLE}} = \frac{304}{610} \approx 0.50 \quad \theta_{\text{MAP}} = \frac{308}{618} \approx 0.50 \quad \theta_{\text{FB}} = \frac{309}{620} \approx 0.50$$

Remember, θ_{MLE} and θ_{MAP} are point estimates, while θ_{FB} is derived via $p(F = \text{Ⓣ} \mid \mathcal{D})$

For MLE we had $\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$

Clearly, we get the same result for $|T| = 1, |H| = 4$ and $|T| = 10, |H| = 40$. Which one is *better*? Why?

For MLE we had $\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$

Clearly, we get the same result for $|T| = 1, |H| = 4$ and $|T| = 10, |H| = 40$. Which one is *better*? Why?

How many flips? Hoeffding's Inequality for a *sampling complexity bound*:

$$p(|\theta_{\text{MLE}} - \theta| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2} \leq \delta,$$

where $N = |T| + |H|$

For example, I want to know θ , within $\epsilon = 0.1$ error, with probability at least $1 - \delta = 0.99$

We have:

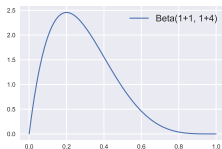
$$N \geq \frac{\ln(2/\delta)}{2\epsilon^2} \rightarrow N \approx 265$$

How many flips?

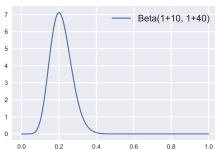
We had $p(\theta = x \mid \mathcal{D}) = \text{Beta}(x \mid a + |T|, b + |H|)$

Visualize the posterior (for given prior, e.g. $a = b = 1$):

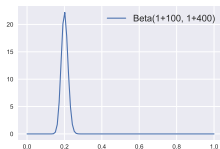
$$|T| = 1, |H| = 4$$



$$|T| = 10, |H| = 40$$



$$|T| = 100, |H| = 400$$



With more data the posterior becomes more peaky – we are more certain about our estimate of θ

Application: German Tank Problem

1, 2, 3, 4, 5, 6, ...
↑ ↑
π π

During WWII, the allies needed to estimate the number of tanks being produced by Germany.

Data? Serial numbers from parts of destroyed tanks.

Month	Bayesian est.	Intelligence est.	German records
June 40	169	1000	122
June 41	244	1550	271
August 42	327	1550	342

Source (and elaborate analysis):

https://en.wikipedia.org/wiki/German_tank_problem

- Maximum likelihood
- Maximum a posteriori
- Fully Bayesian analysis
- Prior, Posterior, Likelihood
- The i.i.d. assumption
- Conjugate prior

- Monotonic transforms for optimization.
- Solving integrals by reverse-engineering densities.