Learning about a coin

Parameter inference: maximum likelihood, maximum a posteriori, and fully Bayesian analysis

Maximilian Soelch

Technische Universität München

We flip the same coin 10 times:

H T H H T H H H T H

Probability that the next coin flip is T?

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

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

$$p_i\big(F_i = \textcircled{T}\big) = \theta_i.$$

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

$$p_i\big(F_i = \textcircled{T} \mid \theta_i\big).$$

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 $\theta_1, \ldots, \theta_{10}$:

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

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

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

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

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

We need to model this
First assumption: The coin flips do not affect each other—independence.

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

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 significantly.

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.):

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

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 $\mathrm{d}f/\mathrm{d}\theta$, 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*.

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

Maximum Likelihood Estimation (MLE) for *any* coin sequence?

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

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

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

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

H H

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

But even a fair coin 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 itself a random variable.

For *any* x, we want to be able to express $p(\theta = x \mid \mathcal{D})$, where $\mathcal{D}$ is a *random variable* that models the observed sequence at hand ($\mathcal{D} =$ 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 $p(\mathcal{D})$, called *evidence*, is not so important in our case.)

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

Prior, likelihood and posterior.

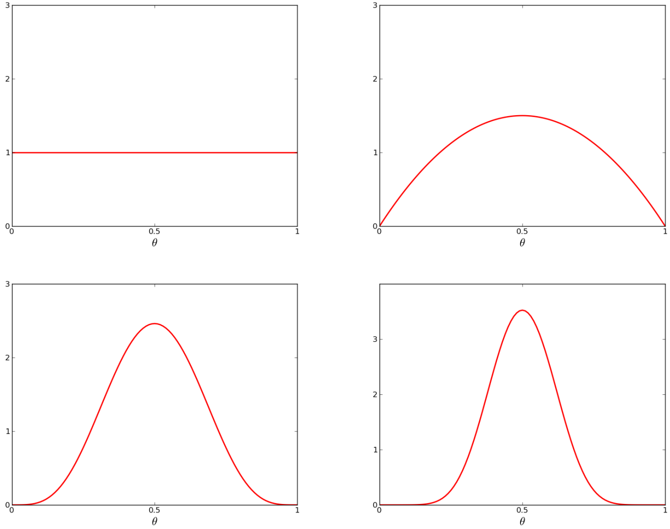
How should we choose the prior $p(\theta = x)$? There are no observations, so there are no assumptions/constraints/…, except that

$$\int p(\theta = x)\,\mathrm{d}x = 1 \qquad \left(\text{or shorter: } \int p(\theta)\,\mathrm{d}\theta = 1\right)$$

has 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:



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

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

Decode!

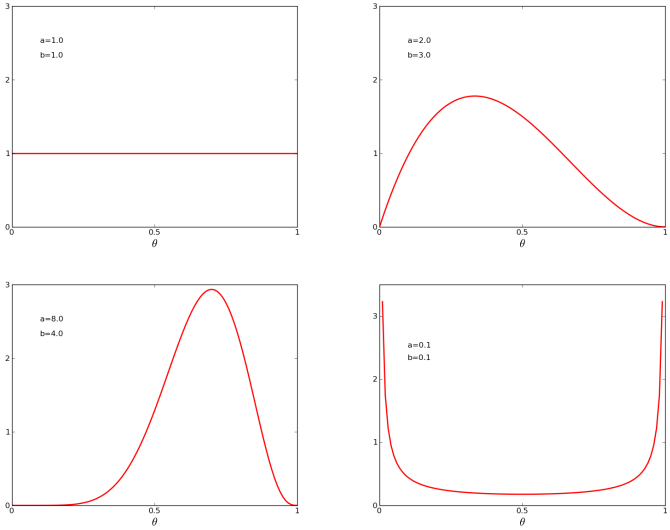
- We have seen this part before: $\theta^{a-1}(1-\theta)^{b-1}$
- $\Gamma(n) = (n-1)!$, if $n \in \mathbb{N}$

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:

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

We know

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

So we get:

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

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

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

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

$$\frac{\Gamma(|H|+|T|+a+b)}{\Gamma(|T|+a)\Gamma(|H|+b)}.$$

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

We now know that

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

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

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

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)

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.

Is that it? No... We can go even deeper!

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

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

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!

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

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

We call this a *(fully) Bayesian* analysis, because rather than optimising out parameters, we integrate them out, i.e., we incorporate the full Bayesian structure of our model.

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.)

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

Product Rule + conditional independence assumption!

Continue by plugging in all formulae we get

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

Is that it? No... We can go even deeper!
We could put a *hyperprior* on the parameters a and b. But not now...

How many flips?

For MLE we had

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

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

Hoeffding's Inequality for a *sampling complexity bound*:

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

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

Application: German Tank Problem

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`

What we learned

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

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