Notes on Bayesian Inference

Alexander Hoyle

March 2026

1 Motivating Bayesian Inference

1.1 Likelihoods

Imagine I see the following data:

If I flip this coin again, what is the probability that this coin will give heads, i.e., what is $P(X = \texttt{H})$?

You might answer 1/2. Which is maybe reasonable, although I didn't indicate it was a fair coin. Here's another question. What is the most likely probability of heads that will have produced the above set of coin flips?

Answering this formally requires a little machinery. We're going to define a probability mass function: $P(X = \texttt{H}) = \theta$ with $\theta \in [0, 1]$, meaning $P(X = \texttt{T}) = (1 - \theta)$. We then say that a coin flip is a sample from this distribution, $X \sim P(X)$, with X a random variable that can take either the value `H` or `T`.

What is the joint probability of the flips we saw? Well, we know that the probability of getting one head (`H`) is θ. And the probability of getting two heads, `HH`, is $\theta \cdot \theta$.

So the probability of the exact sequence of flips we observed, `HHTTHHHHHT`, is $\theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta \cdot \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta)$

Or $\theta^7(1-\theta)^3$. That said, we usually are less interested in the probability of obtaining that exact sequence, and instead in the probability of getting a specific number of heads and tails. So we need to add up all possible ways of getting exactly 7 heads and 3 tails:

`HHHHHHHTTT`
`HHHHHHTHTT`
`HHHHHTHHTT`
...

Which, as you probably know, is just $\binom{10}{7}$. Indeed, the *Binomial* distribution is defined as n independent Bernoulli trials with k successes[1]:

$$P(X = k \mid \theta) = \text{Binom}(n, k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}. \tag{1}$$

So now we start to answer our question: what parameter θ would make the data we observed the most likely? This invites us to define the **likelihood function**

$$\mathcal{L}(\theta \mid \mathcal{D}) = P(\mathcal{D} \mid \theta) \tag{2}$$

[1] In our case, we are arbitrarily defining "success" as the coin landing on heads.

That is, the probability of seeing observed data $\mathcal{D}$ given a specific parameter value θ (note that θ can also represent a set of parameters, as in a neural network). Note the change in notation: $\mathcal{L}(\theta \mid \mathcal{D})$ and $P(\mathcal{D} \mid \theta)$ are the same thing, but we use the $\mathcal{L}$ to highlight that θ is the main variable (rather than the data $\mathcal{D}$).

We want to find the θ that maximizes this quantity—this is called **maximum likelihood estimation** (MLE):

$$\max_{\theta} \ \mathcal{L}(\theta \mid \mathcal{D}) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{3}$$

How do we maximize a function with respect to a parameter? We can take the partial derivative and set it to zero. However, taking derivatives of exponential functions is not very fun. Mercifully, we can apply the log, because it is a monotonic function (meaning the maximum/minimum of $f(x)$ is the same as $\log f(x)$). Hence an extremely common quantity in machine learning, which we will come back to time and time again, is the *log-likelihood*

$$\max_{\theta} \ \ell(\theta \mid \mathcal{D}) = \log \left[\binom{n}{k} \theta^k (1-\theta)^{n-k} \right] \tag{4}$$

$$= \log \binom{n}{k} + k \log \theta + (n-k) \log(1-\theta) \tag{5}$$

$\log \binom{n}{k}$ is a constant with respect to θ and disappears when differentiating. Taking the partial derivative and setting it to zero:

$$\frac{\partial \ \ell(\theta \mid \mathcal{D})}{\partial \theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \tag{6}$$

$$\frac{k}{\theta} = \frac{n-k}{1-\theta} \tag{7}$$

$$k(1-\theta) = (n-k)\theta \tag{8}$$

$$k = n\theta \tag{9}$$

$$\theta^*_{\text{MLE}} = \frac{k}{n} = \frac{7}{10} \tag{10}$$

Quite some work for a very intuitive and simple result! The probability that maximizes the data likelihood is just the empirical mean: that is, the number of successes over the number of trials. This should make sense: if θ were higher or lower than 0.7, say 0.9, then we'd expect more heads than 7 out of 10 (namely, we'd expect 9).

1.2 Priors

Let's get back to our coin. Say this time I flip a one-franc coin five times and get four heads—following our previous logic, the most likely value for θ is $\frac{4}{5}$. Let's extend this a little further. I flip a one-franc coin ten times and get nine heads—now my estimate of θ is $\frac{9}{10}$. Does this feel right to us? That the probability of a real franc coin landing heads is 0.9?

Instinctively, I would say no. In fact, if 100 people all flipped 10 fair coins, roughly 1 would get 9 heads. Could we maybe encode this idea that, even if we observe 9 heads, maybe we're not entirely convinced the coin is unfair/biased?

This question gets at the "main" distinction in statistics: *frequentism* and *Bayesian*. In the frequentist case, we assume there is some "true" parameter that we don't have access to, and we have to observe data until our estimators converge to that value—it acknowledges that the estimates are noisy the less data we've seen, but only samples can inform those estimates. In

the *Bayesian* case, the focus of this lecture, we can incorporate prior knowledge or beliefs before we've seen data, and the data lets us update those beliefs.

To encode this idea, we will bring out a little more machinery: the **prior density**:

$$p(\theta) \tag{11}$$

On notation: we use uppercase $P(\cdot)$ for probability mass functions (discrete distributions, like the Binomial) and lowercase $p(\cdot)$ for probability density functions (continuous distributions). You will see both going forward.[2]

The prior is a continuous distribution over the possible values of our parameters. I want to avoid discussing the functional form for a moment and just visualize it

Let's think about what this plot means—this is a probability density function. There is more probability mass here at the center, around $\theta = 0.5$, meaning that these values are a bit more likely. If I play around with the shape, I can increase that probability, decrease it, or make it sort of even across the board.

What I am doing here is encoding my *belief* about θ. This is a choice I am making—as in me, a person with subjective ideas about the world. I'm going to say that I'm pretty confident in the Swiss mint and feel that the probability density should be centered pretty strongly around 0.5.

Alright, let's now look at the functional form:

$$p(\theta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{12}$$

Don't worry too much about the function $\mathrm{B}(\cdot)$: it is a normalization constant to make sure that the total probability is one.

$$\mathrm{B}(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \equiv \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta \tag{13}$$

1.3 Posteriors

Ok. So this is something I've called a "prior". But now let's say I've observed my data—my ten coin flips—maybe I want to change my belief about θ? Sure, nine out of ten could be good luck, but if I got *29* out of 30 heads, I would probably wonder about that coin.

So, it makes sense to think of the distribution over θ conditioned on some observed data $\mathcal{D}$, plus my original beliefs about theta (which are encoded in α and β):

$$p(\theta \mid \mathcal{D}, \alpha, \beta) \tag{14}$$

Which I will just write as $p(\theta \mid \mathcal{D})$ because we're going to treat α and β as fixed. This term is called the **posterior distribution**.

How should we think about this quantity? This is where Bayes rule comes in. As a reminder

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{15}$$

[2] Text in blue is AI-generated.

Let's put this into use here (recall that the likelihood $\mathcal{L}(\theta \mid \mathcal{D})$ is the same quantity as $p(\mathcal{D} \mid \theta)$, just viewed as a function of θ):

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \tag{16}$$

$$= \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\int p(\mathcal{D} \mid \theta')p(\theta')\,\mathrm{d}\theta'} \tag{17}$$

Now, this denominator—called the marginal likelihood—is basically intractable. We can't easily integrate over all possible values of θ (and it is effectively impossible when θ may be multiple parameters). However, $p(\mathcal{D})$ is fixed and does not depend on θ, so we say that the posterior is *proportional to* the numerator

$$p(\theta \mid \mathcal{D}) = \frac{1}{Z}p(\mathcal{D} \mid \theta)p(\theta) \tag{18}$$

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta) \tag{19}$$

You can recognize the quantity $p(\mathcal{D} \mid \theta)$ as the likelihood from before. The above formulation is one of the fundamental mechanisms in Bayesian inference:

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{20}$$

Let's continue with our example to see how this actually gets us to what we want: the posterior.

$$p(\theta \mid \mathcal{D}) \propto \binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{21}$$

$$\propto \theta^k(1-\theta)^{n-k}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{22}$$

$$\propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \tag{23}$$

Well...this looks sort of familiar now, doesn't it? In fact, it has exactly the functional form of the Beta distribution, without the normalizing constant. But we can work out that constant easily now, because we know that this posterior has to integrate to 1:

$$1 = \int_0^1 p(\theta \mid \mathcal{D})\,\mathrm{d}\theta \tag{24}$$

$$1 = \int_0^1 \frac{1}{Z}\cdot\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}\,\mathrm{d}\theta \tag{25}$$

$$Z = \int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}\,\mathrm{d}\theta \equiv \mathrm{B}(k+\alpha, n-k+\beta) \tag{26}$$

So this process returned a *new* Beta distribution, just with different parameters: $k+\alpha$ and $n-k+\beta$

$$p(\theta \mid \mathcal{D}) = \frac{1}{\mathrm{B}(k+\alpha, n-k+\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \tag{27}$$

When a posterior follows the same functional form as the prior (given a likelihood), we call this a *conjugate prior*. We can also maximize this quantity to get the most likely value of the posterior under the prior:

$$\theta^*_{\mathrm{MAP}} = \frac{k+\alpha-1}{n+\alpha+\beta-2} \tag{28}$$

This is called **maximum a posteriori** (MAP) estimation. There are several important things to note here: the first is that this quantity is the *mode* (highest density point) of the posterior, not its mean. The mean of a Beta(α, β) distribution is $\frac{\alpha}{\alpha+\beta}$, so the **posterior mean** is:

$$\mathbb{E}[\theta \mid \mathcal{D}] = \frac{k+\alpha}{n+\alpha+\beta} \tag{29}$$

When the posterior is symmetric, the mode and the mean coincide, but in general they differ. Note also that, as n grows large, both the MAP and the posterior mean converge to the MLE $\frac{k}{n}$—the data eventually overwhelms the prior.

The second thing to note is that you actually don't need to worry about the normalizing constant to get the MAP estimate: you can just maximize the unnormalized (proportional) quantity!

Last, this process also sheds light on an interpretation for the α and β parameters. They now have been updated with the number of successes, k and the number of failures, $n-k$:

$$\theta^*_{\text{MAP}} = \frac{\#\text{ observed heads} + \#\text{ prior heads}}{\#\text{ observed flips} + \#\text{ prior flips}} \tag{30}$$

Instead of coins, let's think of this prior as your beliefs about someone before going into a relationship. You're young so you're not that committed, but they're cool. But now you start going on dates, and each date, they're kind of shitty—"did you mean to wear that ugly sweater"—this keeps going and going and pretty soon you realize: oh, they're not a good person, and I shouldn't have given them the benefit of the doubt.

Then later, they come back and have said, "oh I've changed"—you might start with this posterior as a *new* prior.

1.4 The Posterior Predictive

Let's return to the question we started with: if I flip this coin again after some number of flips, what is the probability it lands heads? The MLE approach would simply plug in the point estimate: $P(X = \texttt{H}) = \theta^*_{\text{MLE}} = \frac{k}{n}$. But this ignores our uncertainty about θ. If we've only seen 3 flips, we probably shouldn't bet the house on $\theta = \frac{2}{3}$.

The Bayesian answer is to average over all possible values of θ, weighted by how plausible each value is given our data—that is, weighted by the posterior:

$$P(X = \texttt{H} \mid \mathcal{D}) = \int_0^1 P(X = \texttt{H} \mid \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta \tag{31}$$

$$= \int_0^1 \theta\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta \tag{32}$$

$$= \mathbb{E}[\theta \mid \mathcal{D}] = \frac{k+\alpha}{n+\alpha+\beta} \tag{33}$$

This is the **posterior predictive distribution** (for the next observation). It turns out to equal the posterior mean, which makes intuitive sense: your best prediction for the next flip is your current average belief about θ.

Why not just use the MAP or MLE estimate? Consider a case where you've seen 1 head in 1 flip. The MLE says $\theta = 1$—the coin always lands heads. The posterior predictive with a Beta$(2, 2)$ prior gives $\frac{1+2}{1+2+2} = \frac{3}{5} = 0.6$: shifted toward heads, but not fanatically so. More generally, whenever we have limited data and real uncertainty about θ, averaging over that uncertainty gives better-calibrated predictions than committing to a single value.