02-680 Module 19

Essentials of Mathematics and Statistics

Aidan Jan

December 2, 2025

Maximum a Posteriori Estimation

Frequentist vs. Bayesian Schools

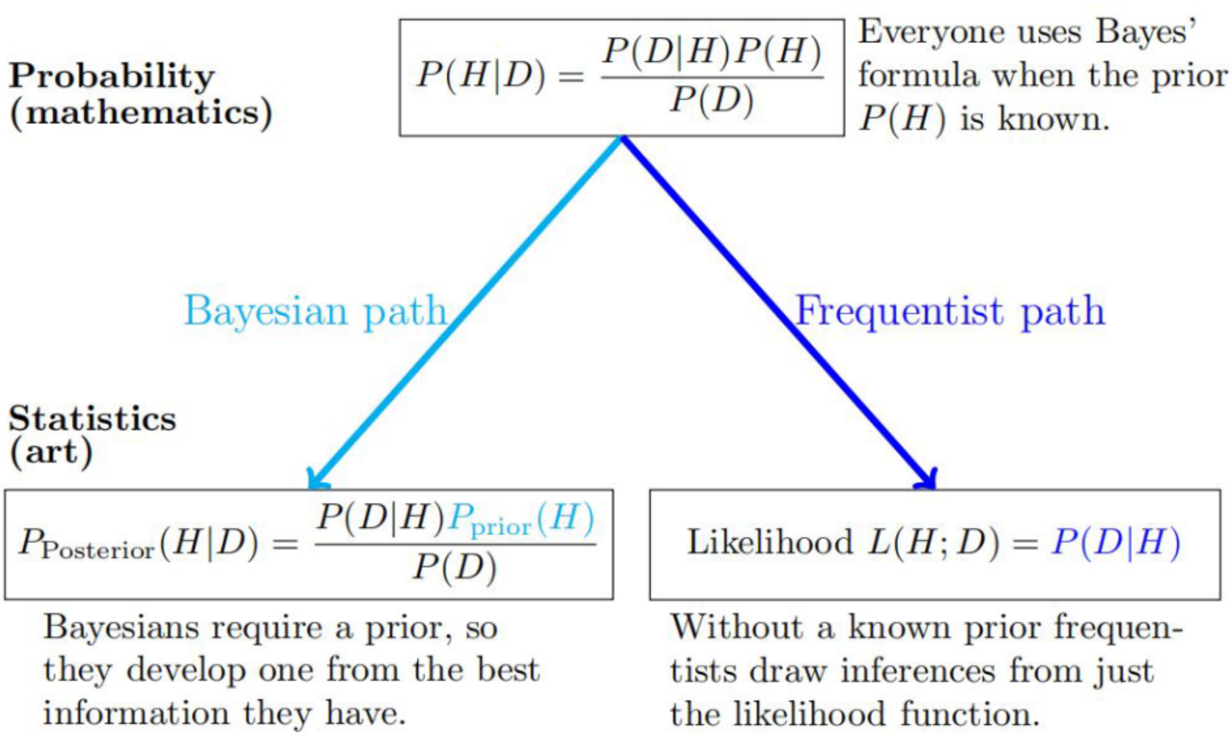
The **Frequentist School** assumes that H is fixed and not random. It uses likelihood only:

$$L(H|D) = p(D|H)$$

Probabilities reflect long-run frequencies. Relies only on observed data.

The **Bayesian School** takes H to be a hypothesis (parameter) and D some data. Different people will have different a priori beliefs, but we would still like to make useful inferences from the data.

When $p(H)$ is known, there is no disagreement, we will all just follow Bayes' Rule as written.



In practice, there is no universally-accepted prior. The main n philosophical difference concerns the **meaning of probability**.

- The Frequentist school represents the idea that *probabilities represent long-term frequencies of repeatable random experiments*

- **Objective interpretation**
- Example: 'A coin has 0.5 probability of tails' means that the relative frequency of tails goes to 0.5 as the number of flips goes to infinity.

- The Bayesian school represents the idea that *probability is an abstract concept that measures a state of knowledge or a degree of belief in a given population*
  - **Subjective interpretation**
  - Example: 'A coin has 0.5 probability of tails' means you "believe" that you will get tails 50% of the time.
  - That is, they consider a range of values each with its own probability of being true.

Key Differences

- Bayesian: Prior + Likelihood →Posterior. (Subjective probability)
- Frequentist: Likelihood only. (Objective probability)

Bayesians' Approach to Parameter Estimation

Let's look at the coin flip example from before:

$$D = X_1, X_2, \cdots, X_n, \quad \text{where } X_i \sim \text{Bernouli}(a)$$

We can further summarize D into c_H and c_T representing the counts of heads and tails, respectively.

We saw last time that

$$\hat{\alpha}_{MLE} = \frac{c_H}{c_H + c_T}$$

But this is assuming we know nothing about a ahead of time. What if we believe that it is 50/50, so we can add what are called pseudocounts to the input c_{H_0} and c_{T_0}, and thus compute

$$\hat{\alpha}_{MLE-PC} = \frac{c_H + c_{H_0}}{c_H + c_T + c_{H_0} + c_{T_0}}$$

Let's assume we have some experiment where we throw a coin 100 times, and we want to know α, $c_H = 0$ and $c_T = 100$. Vanilla MLE would say that the probability is zero.

$$\hat{\alpha}_{MLE} = \frac{c_H}{c_H + c_T} = \frac{0}{0 + 100} = 0$$

But, we have a small belief that this is a fair coin, so let's assume we add the pseudocounts $c_{H_0} = c_{T_0} = 1$, in that case

$$\hat{\alpha}_{MLE-PC} = \frac{c_H + c_{H_0}}{c_H + c_T + c_{H_0} + c_{T_0}} = \frac{0 + 1}{0 + 100 + 1 + 1} = \frac{1}{102}$$

If we're more confident in our prior and set $c_{H_0} = c_{T_0} = 100$, then

$$\hat{\alpha}_{MLE-PC} = \frac{c_H + c_{H_0}}{c_H + c_T + c_{H_0} + c_{T_0}} = \frac{0 + 100}{0 + 100 + 100 + 100} = \frac{1}{3}$$

Pseudocounts

Pseudocounts are a way of **exerting your belief**.

- Larger pseudocounts represent a strong prior belief. (Will have a greater effect on the posterior estimate).
- Small pseudocounts represent a weak prior belief. (Will have a smaller effect on the posterior estimate).

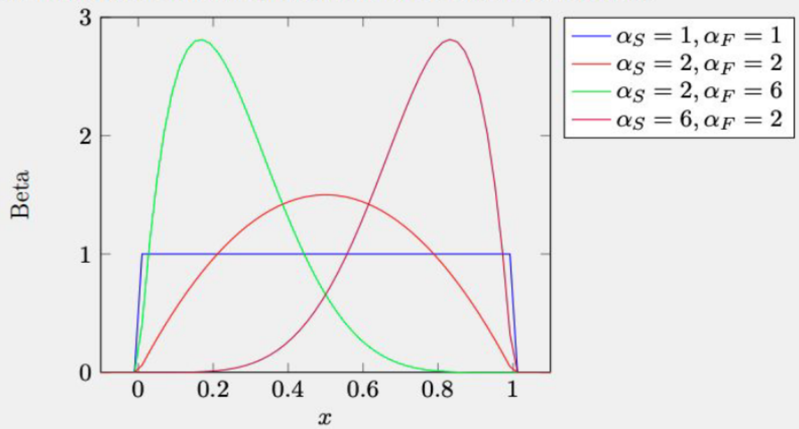
As the sample size goes to infinity, data will dominate the estimate.

An Additional Distribution: Beta

If we model the prior as a Beta distribution on c_{H_0} and c_{T_0} (that is $p(\theta) \sim \text{Beta}(c_{H_0}, c_{T_0})$).

An additional distribution: Beta

The Beta distribution visually looks like an un-balanced normal.



$$f(x; \alpha_S, \alpha_F) = \frac{x^{(\alpha_S - 1)}(1-x)^{(\alpha_F)}}{B(\alpha_S, \alpha_F)}$$

where

$$B(\alpha_S, \alpha_F) = \frac{\Gamma(\alpha_S, \alpha_F)}{\Gamma(\alpha_S)\Gamma(\alpha_F)} \text{ and } \Gamma(a) = (a-1)!.$$

When $X \sim Beta(\alpha_S, \alpha_F)$, $\mathbb{E}[X] = \frac{\alpha_S}{\alpha_S + \alpha_F}$.

If we then want to find the posterior, $p(\theta|\mathcal{D})$,

$$\begin{aligned}
p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\
&\propto p(\mathcal{D}|\theta)p(\theta) \\
&= \binom{\alpha_H + \alpha_T}{\alpha_H} \cdot p^{\alpha_H} \cdot (1-p)^{\alpha_T} \cdot \beta(\alpha_{H_0}, \alpha_{T_0}) \\
&= \binom{\alpha_H + \alpha_T}{\alpha_H} \cdot p^{\alpha_H} \cdot (1-p)^{\alpha_T} \cdot \frac{p^{(\alpha_{H_0}-1)}(1-p)^{\alpha_{T_0}}}{\beta(\alpha_{H_0}, \alpha_{T_0})} \\
&\sim \text{Beta}(\alpha_H + \alpha_{H_0}, \alpha_T + \alpha_{T_0})
\end{aligned}$$

A beta distribution is the conjugate distribution of the binomial distribution.

Maximum a Posteriori (MAP) Estimation

As a reminder

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} p(\mathcal{D}|\theta)$$

On the other hand, if we want to include information of our prior knowledge, then we have **MAP**:

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} p(\theta|\mathcal{D}) = \arg\max_{\theta \in \Theta} p(\mathcal{D}|\theta) \cdot p(\theta)$$

Note that in both cases we're making a **point estimate** of θ. We still don't have the whole picture, we're usign data to estimate our model, but MAP is **partially Bayesian**.

Example: Known Prior

There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair and have probability 0.5 of heads.
- Type B coins are bent and have probability 0.6 of heads.
- Type C coins are bent and have probability 0.9 of heads.

Suppose you have a drawer containing 10 coins: 5 of type A, 3 of type B, and 2 of type C. You reach into the drawer and pick a coin at random. The coin is flipped once and you get tails. What is the probability it is type A? Type B? Type C?

We can create the following table:

Hypothesis θ	Prior $p(\theta)$	Likelihood $p(\mathcal{D}\|\theta)$	Bayes Numerator $p(\mathcal{D}\|\theta) \cdot p(\theta)$	Posterior $p(\theta\|\mathcal{D})$
A	0.5	0.5	0.25	$\propto 0.510$
B	0.3	0.4	0.12	$\propto 0.490$
C	0.2	0.1	0.02	$\propto 0.041$

Thus, $\hat{\theta}_{MAP} = A$.

What if you then flip the same coin another 9 times, so including the first coin we have $a_H = 6$ and $a_T = 4$? We get:

Hypothesis θ	Prior $p(\theta)$	Likelihood $p(\mathcal{D}\|\theta)$	Bayes Numerator $p(\mathcal{D}\|\theta) \cdot p(\theta)$	Posterior $p(\theta\|\mathcal{D})$
A	0.5	0.97×10^{-3}	4.88×10^{-4}	$\propto 0.570$
B	0.3	1.19×10^{-3}	3.58×10^{-4}	$\propto 0.418$
C	0.2	0.05×10^{-3}	0.11×10^{-4}	$\propto 0.012$

Notice in the table above, the prior does not change. In this case, it is still true that $\hat{\theta}_{MAP} = A$, but $\hat{\theta}_{MLE} = B$.

Example: Unknown Prior

Assume we have a similar scenario but this time we don't know the prior. We follow a similar procedure to that for MLE: take the zero point of the **log probability**.

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln p(\mathcal{D}|\theta)p(\theta) = 0$$

Suppose we want to know the probability of head p of a new coin. (1 if heads, 0 if tails). Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Find and estimate the parameter $\theta = \{p\}$. α_H, α_T represents the number of heads and tails separately.

MAP estimation: To estimate parameter p, find p that maximizes the likelihood $\times$ prior.

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} P(\theta|\mathcal{D}) = \arg\max_{\theta \in \Theta} P(\mathcal{D}|\theta)P(\theta)$$

We can simplify:

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{p\in\Theta} \log P(p|\mathcal{D}) \\
&= \arg\max_{p\in\Theta} \log P(\mathcal{D}|p) + \log P(p) \\
&= \arg\max_{p\in\Theta} \log p^{\alpha_H} \cdot (1-p)^{\alpha_T} + \log p^{\alpha_{H_0}-1}(1-p)^{\alpha_{T_0}-1} \\
&= \arg\max_{p\in\Theta} \log p^{\alpha_H} + \log(1-p)^{\alpha_T} + \log p^{\alpha_{H_0}-1} + \log(1-p)^{\alpha_{T_0}-1} \\
&= \arg\max_{p\in\Theta} \alpha_H \log p + (\alpha_T)\log(1-p) + (\alpha_{H_0}-1)\log p + (\alpha_{T_0}-1)\log(1-p) \\
&= \arg\max_{p\in\Theta} (\alpha_H + \alpha_{H_0} - 1)\log p + (\alpha_T + \alpha_{T_0} - 1)\log(1-p)
\end{aligned}
$$

Now for the MAP estimation, we differentiate with respect to p.

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}p} \log P(\mathcal{D}|p) \cdot \log P(p) = 0 \\
\frac{\alpha_H + \alpha_{H_0} - 1}{p} + \frac{\alpha_T + \alpha_{T_0} - 1}{(1-p)}(-1) = 0 \\
\hat{\theta}_{MAP} = p = \frac{\alpha_H + \alpha_{H_0} - 1}{\alpha_H = \alpha_T + \alpha_{H_0} - \alpha_{T_0} - 2}
\end{aligned}
$$

Example: MAP Estimation for Poisson Distribution

Recall the Poisson distribution is

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \forall k \in \Omega$$

Poisson distribution's conjugate prior $P(\Theta)$ is the Gamma distribution.

First, write down the log of $P(\mathcal{D}|\lambda)P(\lambda)$.

$$
\begin{aligned}
p(\mathcal{D}|\lambda)p(\lambda) &= e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod x_i!} \cdot \frac{\lambda^{k-1}e^{-\lambda/\theta}}{\theta^k \Gamma(k)} \\
\ln\left(p(\mathcal{D}|\lambda)p(\lambda)\right) &= \ln\lambda\left(k - 1 + \sum_{i=1}^n x_i\right) - \lambda\left(n + \frac{1}{\theta}\right) - \sum_{i=1}^n \ln(x_i!) - k\ln\theta - \ln\Gamma(k)
\end{aligned}
$$

Now, maximize this log likelihood $\times$ prior. (Take the derivative in terms of λ, and set to zero. Then, solve for λ)

$$
\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\lambda} \ln\lambda\left(k - 1 + \sum_{i=1}^n x_i\right) - \lambda\left(n + \frac{1}{\theta}\right) - \sum_{i=1}^n \ln(x_i!) - k\ln\theta - \ln\Gamma(k) \\
0 &= \frac{\mathrm{d}}{\mathrm{d}\lambda} \ln\lambda\left(k - 1 + \sum_{i=1}^n x_i\right) - \lambda\left(n + \frac{1}{\theta}\right) \\
0 &= -\left(n + \frac{1}{\epsilon}\right) + \frac{1}{\lambda}\left(k - 1 + \sum_{i=1}^n x_i\right) \\
k - 1 + \sum_{i=1}^n x_i &= \lambda\left(n + \frac{1}{\epsilon}\right) \\
\lambda_{MAP} &= \frac{k - 1 + \sum_{i=1}^n x_i}{n + \frac{1}{\epsilon}}
\end{aligned}
$$

Example: MAP Estimation for Normal Distribution

Same process as before. The prior and likelihood are:

$$P(\theta) = \mu \sim \mathcal{N}(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(\frac{1}{2\sigma_0^2}(\mu - \mu - 0)^2\right)$$
$$P(x|\theta = \{\mu, \sigma^2\}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

The log of the likelihood $\times$ prior is

$$\ln(P(x|\theta = \{\mu, \sigma\})P(\theta)) = -\ln 2\pi - \ln\sigma - \ln\sigma_0 - \frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]$$

Now, we maximize by setting the derivative in terms of μ to zero.

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\mu} - \ln 2\pi - \ln\sigma - \ln\sigma_0 - \frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right] \\
0 &= \frac{\mathrm{d}}{\mathrm{d}\mu} - \frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right] \\
0 &= -\frac{1}{2}\left[2 \cdot \frac{(x-\mu)}{\sigma^2} \cdot (-1) + 2 \cdot \frac{(\mu-\mu_0)}{\sigma_0^2} \cdot (1)\right] \\
0 &= \frac{(x-\mu)}{\sigma^2} - \frac{(\mu-\mu_0)}{\sigma_0^2}
\end{aligned}$$

Therefore, we get

$$\mu_{MAP} = \frac{x\sigma_0^2 + \mu_0\sigma^2}{\sigma_0^2 + \sigma^2}$$

We can repeat the process for σ by taking the derivative in terms of σ instead. In that case, we will get

$$\sigma_{MAP}^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$