Lecture 1

You are presented with a challenge: estimate the probability that a thumbtack lands on its head.

You have only 6 draws of this thumbtack:

i.e. H T H H T H

What is the probability of heads? It seems like $\frac{4}{6} = \frac{2}{3}$ is a good anwer. We will see here why.

Maximum Likelihood Estimation (MLE)

Let's restate the problem:

We assume the thumbtack flip follows a Bernoulli distribution with parameter θ

$$P(X) = \begin{cases} \theta & \text{if } x = H \\ 1-\theta & \text{if } x = T \end{cases}$$

Now we write the probabalty of the data D given θ, also called the likelihood of the data:

$$P(D|\theta) = P(X_1, X_2, X_3, X_4, X_5, X_6 | \theta)$$
$$= P(X_1|\theta)\, P(X_2|\theta) \dots P(X_6|\theta)$$

because the flips are IID

$$= \theta^4 (1-\theta)^2$$
$$= \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

where α_H = # of heads

α_T = # of tails

MLE consist in picking the value of θ that maximzes the likelihood.

$$\theta_{MLE} = \arg\max_{\theta} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

It is convenient to compute the log likelihood (LL)

$$LL = \ln \theta^{\alpha_H} (1-\theta)^{\alpha_T} = \ln \theta^{\alpha_H} + \ln (1-\theta)^{\alpha_T}$$
$$= \underline{\alpha^H \ln \theta + \alpha^T \ln (1-\theta)}$$

Because log is monotonic;

$$\arg\max_{\theta} \log P(D|\theta) = \arg\max P(D|\theta)$$

↳ Before you optimize you need to veify that the objective function is concave

There are many ways to show that a function is concave. Here we will show that the second derivative is always < 0. Reising convexity (concavity) properties will be useful late in the course.

$$\frac{\partial^2 LL}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left(\frac{\alpha^H}{\theta} - \frac{\alpha^T}{1-\theta}\right)$$
$$= \frac{-\alpha^H}{\theta^2} - \frac{\alpha^T}{(1-\theta)^2} < 0$$

Solve for θ_{MLE}:

$$\frac{\partial LL}{\partial \theta} = 0$$

$$\frac{\alpha^H}{\theta_{MLE}} - \frac{\alpha^T}{1-\theta_{MLE}} = 0$$

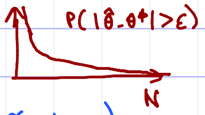
$$\theta_{MLE} = \frac{\alpha^H}{\alpha^H + \alpha^T}$$

↳ This corresponds to the intuitive answer $\frac{2}{3}$

How confident can we be of our answer? Are 6 flips enough? Would we feel more confident if we had 1000 flips? How much more confident?

We can use Hoeffding's inequality adapted for Bernoulli variables to compute a minimum sample size for an error of at most ε.

$$P(|\hat{\theta} - \theta^*| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}$$



$$N = \alpha_H + \alpha_T \ , \quad \hat{\theta} = \alpha_H / (\alpha_H + \alpha_T)$$

Maximum a posteriori estimation (MAP)

Now assume that instead of a thumbtack, for which it's hard to guess the probability of heads, you were tasked with finding the probability of a coin falling on heads.

Suddenly, 4 heads and 2 tails don't seem enough to say that the coin is biased (a biased coin has $\theta \neq 0.5$)

We have a strong prior on θ being 0.5 or very close to 0.5 for a normal coin.

However if we had observed 4000 Heads and 2000 Tails, we would think the coin is biased.

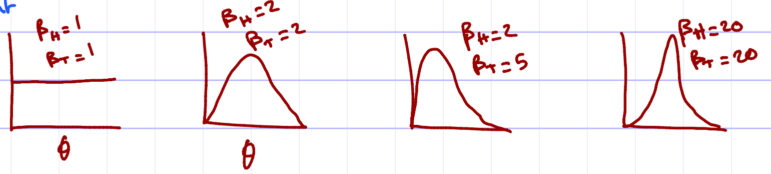
This can be modeled as a Bayesian estimation problem where our prior belief about the coin is expressed as a prior distribution $p(\theta)$.

For this problem, a convenient distribution is the Beta distribution. It allows us to express different beliefs about θ, the parameter of a Bernouilli distribution.

$$\theta \sim \text{Beta}(\beta_H, \beta_T) \leftarrow$$

Here we call the two parameters β_H & β_T because they effectively act as additional head & tail counts in the posterior distribution

$$p(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{\underbrace{B(\beta_H, \beta_T)}_{\text{normalizing constant}}}$$



$\leftarrow$ β_H, β_T affect the prior distribution

We compute the posterior distribution:

$$P(\theta | D) = \frac{P(D|\theta)P(\theta)}{P(D)} \quad \text{(bayes rule)}$$

$$\propto P(D|\theta)P(\theta) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\,\theta^{\beta_H-1}(1-\theta)^{\beta_T-1} = \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

We find $\theta_{MAP} = \arg\max_{\theta} \ \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$

To solve this, we repeat similar steps to above.

$$\theta_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

$\leftarrow$ as $N \to \infty$ the effect of the prior washes out

for small N, the prior can have a big effect.