Models

- When we talk about coin tosses, we are assuming a model
- For example: that there is a fixed probability of heads/tails
- There is one parameter we need: $f = \text{the probability of the coin landing on heads for any one toss}$

Heads for a Single Toss

- What’s the probability for a single coin toss given our model of coin tosses?
  - $P(H | f=0.5)$
- What other assumptions have we made?
  - The outcome is a discrete binary outcome
  - The probabilities are fixed and don’t change over time
  - The outcomes are independent from each other

The Binomial Distribution

- What’s the probability of the number of heads $r$ if the coin is tossed $N$ times?
  - $P(r | f, N) = \frac{N!}{(N-r)!r!} f^r (1-f)^{N-r}$
  - The latter part of the equation is for calculating the joint probability
  - The previous part is for looking at the fact that there’s more than one way to get 4 heads and 1 tail in 5 tosses

Likelihood

- If probability of event $x$ dependent on model parameters $p$ is: $P(x | p)$
- We can talk about the likelihood of the model given the parameters: $L(p | x)$
- The goal then comes to finding the parameters that make the observed data most likely...
- What is this?

Search

- Yes… maximizing the likelihood is another form of search.

The Difference Between Them

- Probability:
  - We know the parameters and want to predict outcome
- Likelihood:
  - We have data and want to estimate a model that fits it
An Example

- Let's try this for coin tossing…
- Say we know that when we tossed a coin 100 times, we obtained 56 heads.
  \[ P(r \mid f, N) = \frac{N!}{(N-r)!r!} f^r (1-f)^{N-r} \]
- Plug in the numbers!
- Is our coin biased? Maybe bent?

Complexity

- The more complex the data, the more parameters, and the harder things get to estimate
- Iterative MLE (maximum likelihood estimate):
  - Typically not feasible to evaluate the likelihood at all points in the problem’s parameter space
  - This is where optimization or minimization techniques come in handy

Practical Considerations

- In the coin example, the parameters only depend on the latter half of the formula
- Thus, when we maximize the likelihood, we can drop the rest of the formula to make a very hard problem simpler
  \[ P(r \mid f, N) = \frac{N!}{(N-r)!r!} f^r (1-f)^{N-r} \]

Optimization Techniques

- Play hotter-colder across the parameter space
- Are normally concerned with how likelihood changes when going across the space

Log-likelihood

- It’s often easier to work with the natural log of the likelihood, since if you multiply lots of small numbers, you’ll quickly end up with a precision problem
  \[ a = bc \]
- Log-likelihoods are simply added together and just get more negative and larger
  \[ \log(a) = \log(b) + \log(c) \]
Optimizing the Likelihood

- Is the same as minimizing the negative log-likelihood

\[-\log P(d \mid h) - \log P(h)\]

For the Coin Toss

Local Minima

Important to Note

- The parameters may not uniquely explain the data!
- Just think about those clouds that circle the top of a mountain…

Murder

- "Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type ‘O’ blood. The blood groups of the two traces are found to be of type ‘O’ (a common type in the local population, having frequency 60%) and of type ‘AB’ (a rare type, with frequency 1%). Do these data (type ‘O’ and ‘AB’ blood were found at scene) give evidence in favour of the proposition that Oliver was one of the two people present at the crime?"

What Happens…

- When you don’t have all the data?
- How can we handle this?
Expectation Maximization

- An iterative technique for when we have some visible data, but not all and we want to estimate the distribution of the data
- Many learning algorithms are an extension or may be explained in terms of the EM algorithm
- The book explains specifics