

# Lecture #4b: Maximum Likelihood Estimation and Fitting Probability Distributions

---

COMP3608: Intelligent Systems  
Inzamam Rahaman

# Uses of optimisation in achieving rational agents

---

- Recall our definition of rational agents
- Rational agents try to optimise some performance measure
- Optimisation is critical both directly and indirectly to achieving rational behaviour from agents
- So far, we looked at how optimisation helps with the problem directly
  - Now we look at indirect case that is monumentally important for when we tackle ML in a few lectures down the road

# Probability and Intelligence

---

- Many sources of uncertainty
- Hence, all models of the world carry a degree of uncertainty
- Hence, the world can be understood through the help of the language of uncertainty - probability!
  - We make judicious use of probability distributions to help model the world
- Will assume that you retained some basic intuition and knowledge of probability
  - Basic axioms, definition of a PMF, PDF, and independence, intuition of what a random variable is, and basic notions of a probability distribution
  - Will skip Bayesian stuff for now, but it will become important at the last topic of the course, so start looking over from now :)

# An important epistemic truism!

---

- “All models are wrong, but some models are useful”
- Much of what we will do here requires us to make simplifying assumptions
- World is too complicated to model with perfect accuracy
- Making these assumptions makes rationally navigating the world a tractable problem
- So please don't be too alarmed :)
  - But always stay mindful of this truth!



# The Relativity of Wrong

---

- My answer to him was, "John, when people thought the earth was flat, they were wrong. When people thought the earth was spherical, they were wrong. But if you think that thinking the earth is spherical is just as wrong as thinking the earth is flat, then your view is wronger than both of them put together."
- From Asimov's Essay, "The Relativity of Wrong"
  - Every science student should read at least once:  
<https://chem.tufts.edu/AnswersInScience/RelativityofWrong.htm>





# Frequentism vs Bayesianism

---

- Two competing definition on the “meaning” behind probability
- In the frequentism case, we assume probabilities describe the frequency of events
- In the bayesian case, we assume probabilities are meant to express degrees of belief
- Subtle differences that can lead to different methodological approaches
- We will look at a Frequentism approach today, bayesian approaches towards the end to the course



# Probability Distribution

---

- Have some random variable that can take on many different values
  - Discrete and continuous cases
- Probability of different values, can be described by probability mass functions (discrete) or proportionally described by a probability density function
- Several common patterns emerge in terms.
  - General Patterns that be refined in terms of using several parameters describe the shape of the distribution
  - Each pattern has common use cases
  - Usually we assume a particular pattern when dealing with some data







# Probability Distributions Notation

---

$$X \sim PD(\theta_1, \theta_2, \dots, \theta_n)$$

$X$  is distributed in accord with probability distribution  $PD$  with parameters  $\theta_1, \theta_2, \dots, \theta_n$

$$P(X = x \mid \theta_1, \theta_2, \dots, \theta_n) = PD(x; \theta_1, \theta_2, \dots, \theta_n)$$

# Probability Distributions Notation - Poission Distribution

---

$$X \sim \text{Pois}(\lambda = 4)$$

$X$  is distributed in accord with a Poission distribution with  
 $\lambda = 4$

$$P(X = x \mid \lambda = 4) = \frac{\lambda^x e^{-x}}{x!} = \frac{4^x e^{-x}}{x!}$$



# Fitting a distribution

---

- Sometimes, we can make a reasonable guess or assumption about what distribution our data follows
- ... but we don't know the parameters of the distribution!
- To meaningfully compute things or answer questions, we need to know the parameters as well!
- How to compute parameters

# Fitting a distribution

---

- Think of the distribution like an article of clothing
- Just like a tailor/seamstress can measure you to create the best suit or dress that we can fit you, we can do the same with data
- We need to find the **optimal** fit
- We need to measure the degree to which a distribution with a particular configuration of parameters fits the data
- This can be used as our objective function





# Likelihood function

---

- Suppose that we have a set of data points,  
 $D = \{x_1, x_2, \dots, x_n\}$
- And we are trying to measure the fit of a distribution  
 $PD(\theta_1, \theta_2, \dots, \theta_m)$  to  $D$
- How to do this?
- We still need to make a few assumptions
  - Main assumption i.i.d - independent and identically distributed

# I.I.D Assumption

---

- Two components:
- Identically - the same.
  - Assume that all data points are drawn from the same distribution with the same parameters
- Independent - the probability of two data points are independent,

$$P(x_1, x_2 | \theta_1, \theta_2, \dots, \theta_m) = P(x_1 | \theta_1, \theta_2, \dots, \theta_m)P(x_2 | \theta_1, \theta_2, \dots, \theta_m)$$




# I.I.D Assumption

---

- Two components:
- Identically - the same.
  - Assume that all data points are drawn from the same distribution with the same parameters
- Independent - the probability of two data points are independent,

$$P(x_1, x_2 | \theta_1, \theta_2, \dots, \theta_m) = P(x_1 | \theta_1, \theta_2, \dots, \theta_m)P(x_2 | \theta_1, \theta_2, \dots, \theta_m)$$

  
 $x_1$  and  $x_2$

# Deriving the likelihood function

---

- The likelihood function measures how probable our data is under our probability distribution
- Hence, our Likelihood function is the product of the probability of each data point under the model (i.i.d) assumption



$$\mathcal{L}(D, \theta_1, \theta_2, \dots, \theta_m) = \prod_{i=1}^n \text{PD}(x_i; \theta_1, \theta_2, \dots, \theta_m)$$

i.i.d assumption

# Example

---

- Suppose that I have a coin (that may not be fair), and I let the flip coming up heads as a success
- When we have two outcomes, a “success” and a “failure”, we model using a Bernoulli distribution
- Assume that we flip the coin 5 times and get the following sequence of results: HHTTH.
- What is the probability of success (i.e. the parameter of the Bernoulli distribution)?

$$\mathcal{L}(HHTTH, \theta) = P(HHTTH | \theta)$$

$$P(HHTTH | \theta) = P(H | \theta)P(H | \theta)P(T | \theta)P(T | \theta)P(H | \theta)$$

$$P(HHTTH | \theta) = \theta\theta(1 - \theta)(1 - \theta)\theta$$

$$P(HHTTH | \theta) = \theta^3(1 - \theta)^2$$



$$\mathcal{L}(HHTTH, \theta) = P(HHTTH | \theta)$$

$$P(HHTTH | \theta) = P(H | \theta)P(H | \theta)P(T | \theta)P(T | \theta)P(H | \theta)$$

$$P(HHTTH | \theta) = \theta\theta(1 - \theta)(1 - \theta)\theta$$

$$P(HHTTH | \theta) = \theta^3(1 - \theta)^2$$

$$\max_{\theta \in \mathbb{R}} \theta^3(1 - \theta^2)$$

Products are hard to work with. :(

# Transforming a problem

---

- Suppose that we want to maximise  $f$ 
  - i.e.  $x^* = \operatorname{argmax}_{x \in \mathcal{X}} f(x)$
- If we have another function,  $g$ , is monotonically increasing, then
  - i.e.  $x^* = \operatorname{argmax}_{x \in \mathcal{X}} f(x) = \operatorname{argmax}_{x \in \mathcal{X}} g(f(x))$



# Log-likelihood

---

- Log is monotonically increasing
- Hence,  $x^* = \operatorname{argmax}_{x \in \mathcal{X}} f(x) = \operatorname{argmax}_{x \in \mathcal{X}} \log(f(x))$
- Hence, by taking log, we can transform the likelihood  
replace the product with a sum!
- We call this the log-likelihood

# Taking natural log

---

- Recall that

$$\log a \cdot b = (\log a) + (\log b)$$

- Hence,

$$\log \prod_{i=1}^n x_i = (\log x_1) + (\log x_2) + \dots + (\log x_n)$$

$$\log \prod_{i=1}^n x_i = \sum_{i=1}^n \log x_i$$

Hence

$$\operatorname{argmax}_{\theta \in \mathbb{R}} \theta^3(1 - \theta^2) = \operatorname{argmax}_{\theta \in \mathbb{R}} \log(\theta^3(1 - \theta^2))$$

$$\operatorname{argmax}_{\theta \in \mathbb{R}} \theta^3(1 - \theta^2) = \operatorname{argmax}_{\theta \in \mathbb{R}} \log((\theta^3) + \log(1 - \theta)^2)$$

$$\operatorname{argmax}_{\theta \in \mathbb{R}} \theta^3(1 - \theta^2) = \operatorname{argmax}_{\theta \in \mathbb{R}} 3 \log(\theta) + 2 \log(1 - \theta)$$



$$\frac{d \left( 3 \log(\theta) + 2 \log(1 - \theta) \right)}{d\theta} = \frac{3}{\theta} - \frac{2}{1 - \theta}$$

Can solve analytically

$$\frac{3}{\theta} - \frac{2}{1-\theta} = 0$$

$$\frac{3(1-\theta) - 2\theta}{\theta(1-\theta)} = 0$$

$$\frac{3 - 3\theta - 2\theta}{\theta(1-\theta)} = 0$$

$$\frac{3 - 5\theta}{\theta(1-\theta)} = 0$$

$$3 - 5\theta = 0$$

$$\theta = \frac{3}{5}$$

# MLE using Gradient Descent

---

- Some distributions don't have closed form solutions for Log-likelihood
- Or solving closed forms are difficult
- We can use gradient descent to solve such cases!
- Need to transform maximisation problem to minimisation problem



# Negative log-likelihood

---

- Recall, by multiplying by  $-1$ , we can transform a maximisation problem to a minimisation problem
- Hence, by negating the log-likelihood, we get a loss function we can use with gradient descent

# General NLLs

---

- The loss function need not be tied to a specific dataset
- Can derive NLLs to use as loss functions that are expressed in terms of a generalised dataset