

# COMP2610 / COMP6261 Information Theory

## Lecture 5: Bernoulli, Binomial, Maximum Likelihood and MAP

**Thushara Abhayapala**

Audio & Acoustic Signal Processing Group  
School of Engineering,  
College of Engineering & Computer Science  
The Australian National University,  
Canberra, Australia.

# Announcements

## Assignment 1

- Will be released this week
- Worth 10% of Course total
- Due Friday 26 August 2022, 5:00 pm

# Last time

- Examples of application of Bayes' rule
  - ▶ Formalizing problems in language of probability
  - ▶ Eating hamburgers, detecting terrorists, ...
- Frequentist vs Bayesian probabilities

# The Bayesian Inference Framework

## Bayesian Inference

Bayesian inference provides us with a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\underbrace{p(Z|X)}_{\text{posterior}} = \frac{\overbrace{p(X|Z)}^{\text{likelihood}} \overbrace{p(Z)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}} = \frac{p(X|Z)p(Z)}{\sum_{Z'} p(X|Z')p(Z')}$$

**Prior:** Belief that someone is sick

**Likelihood:** Probability of testing positive given someone is sick

**Posterior:** Probability of being sick given someone tests positive

# This time

- The Bernoulli and binomial distribution (we will make much use of this henceforth in studying binary channels)
- Estimating probabilities from data
- Bayesian inference for parameter estimation

# Outline

- 1 The Bernoulli Distribution
- 2 The Binomial Distribution
- 3 Parameter Estimation
- 4 Bayesian Parameter Estimation
- 5 Wrapping up

- 1 The Bernoulli Distribution
- 2 The Binomial Distribution
- 3 Parameter Estimation
- 4 Bayesian Parameter Estimation
- 5 Wrapping up

# The Bernoulli Distribution

## Introduction

Consider a binary variable  $X \in \{0, 1\}$ . It could represent many things:

- Whether a coin lands heads or tails
- The presence/absence of a word in a document
- A transmitted bit in a message
- The success of a medical trial

Often, these outcomes (0 or 1) are not equally likely

What is a general way to model such an  $X$ ?



# The Bernoulli Distribution

## Definition

The variable  $X$  takes on the outcomes

$$X = \begin{cases} 1 & \text{probability } \theta \\ 0 & \text{probability } 1 - \theta \end{cases}$$

Here,  $0 \leq \theta \leq 1$  is a **parameter** representing the **probability of success**

For higher values of  $\theta$ , it is more likely to see 1 than 0

- e.g. a biased coin

# The Bernoulli Distribution

## Definition

By definition,

$$p(X = 1|\theta) = \theta$$

$$p(X = 0|\theta) = 1 - \theta$$

More succinctly,

$$p(X = x|\theta) = \theta^x(1 - \theta)^{1-x}$$

# The Bernoulli Distribution

0-1  $\frac{1}{2}$  flip

## Definition

By definition,

$$p(X = 1|\theta) = \theta$$

$$p(X = 0|\theta) = 1 - \theta$$

More succinctly,

$$p(X = x|\theta) = \theta^x(1 - \theta)^{1-x}$$

This is known as a **Bernoulli distribution** over binary outcomes:

$$p(X = x|\theta) = \text{Bern}(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

Note the use of the conditioning symbol for  $\theta$ ; will revisit later

# The Bernoulli Distribution

## Mean and Variance

The **expected value** (or mean) is given by:

$$\begin{aligned}\mathbb{E}[X|\theta] &= \sum_{x \in \{0,1\}} x \cdot p(x|\theta) \\ &= 1 \cdot p(X = 1|\theta) + 0 \cdot p(X = 0|\theta) \\ &= \theta.\end{aligned}$$

The **variance** (or squared standard deviation) is given by:

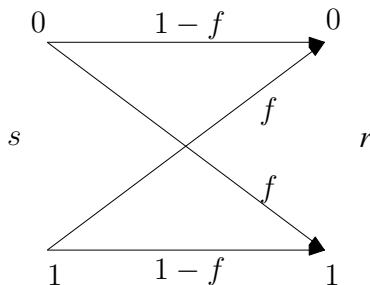
$$\begin{aligned}\mathbb{V}[X|\theta] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \theta)^2] \\ &= (0 - \theta)^2 \cdot p(X = 0|\theta) + (1 - \theta)^2 \cdot p(X = 1|\theta) \\ &= \theta(1 - \theta).\end{aligned}$$

## Example: Binary Symmetric Channel

Suppose a sender transmits messages  $s$  that are sequences of bits

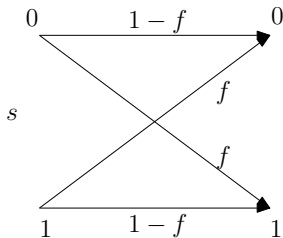
The receiver sees the bit sequence (message)  $r$

Due to noise in the channel, the message is flipped with probability  $0 \leq f \leq 1$



## Example: Binary Symmetric Channel

We can think of  $r$  as the outcome of a **random variable**, with conditional distribution given by:



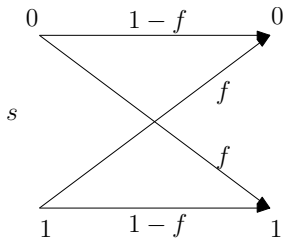
$$\begin{aligned} p(r=0|s=0) &= 1-f & p(r=0|s=1) &= f \\ p(r=1|s=0) &= f & p(r=1|s=1) &= 1-f \end{aligned}$$

If  $E$  denotes whether an error occurred, clearly

$$p(E=e) = \text{Bern}(e|f), \quad e \in \{0, 1\}.$$

## Example: Binary Symmetric Channel

We can think of  $r$  as the outcome of a **random variable**, with conditional distribution given by:



$$\begin{aligned} p(r=0|s=0) &= 1-f & p(r=0|s=1) &= f \\ p(r=1|s=0) &= f & p(r=1|s=1) &= 1-f \end{aligned}$$

If  $E$  denotes whether an error occurred, clearly

$$p(E=e) = \text{Bern}(e|f), \quad e \in \{0, 1\}.$$

Why?  $p(E=e) = p(r=1, s=0) + p(r=0, s=1)$  (mutually exclusive)

So  $p(E=e) = p(r=1|s=0)p(s=0) + p(r=0|s=1)p(s=1)$

This equals  $f$  regardless of the value of  $p(s=0)$ .

- 1 The Bernoulli Distribution
- 2 The Binomial Distribution
- 3 Parameter Estimation
- 4 Bayesian Parameter Estimation
- 5 Wrapping up



# The Binomial Distribution 二项分布.

## Introduction

$n$ 次伯努利分布.

Suppose we perform  $N$  independent Bernoulli trials

- e.g. we toss a coin  $N$  times
- e.g. we transmit a sequence of  $N$  bits across a noisy channel

Each trial has probability  $\theta$  of success

What is the distribution of the number of times ( $m$ ) that  $X = 1$ ?

- e.g. the number of times we obtained  $m$  heads
- e.g. the number of errors in the transmitted sequence

# The Binomial Distribution

## Definition

Let

$$Y = \sum_{i=1}^N X_i$$

where  $X_i \sim \text{Bern}(\theta)$ .

Then  $Y$  has a **binomial** distribution with parameters  $N, \theta$ :

$$p(Y = m) = \text{Bin}(m|N, \theta) = \binom{N}{m} \theta^m (1 - \theta)^{N-m}$$

for  $m \in \{0, 1, \dots, N\}$ . Here

$$\binom{N}{m} = \frac{N!}{(N-m)!m!}$$

is the # of ways we can we obtain  $m$  heads out of  $N$  coin flips

# The Binomial Distribution:

## Mean and Variance

It is easy to show that:

$$\mathbb{E}[Y] = \sum_{m=0}^N m \cdot \text{Bin}(m|N, \theta) = N\theta$$

$$\mathbb{V}[Y] = \sum_{m=0}^N (m - \mathbb{E}[m])^2 \cdot \text{Bin}(m|N, \theta) = N\theta(1 - \theta)$$

- Follows from linearity of mean and variance

$$\mathbb{E}[Y] = \mathbb{E} \left[ \sum_{i=1}^N X_i \right] = \sum_{i=1}^N \mathbb{E}[X_i] = N\theta$$

$$\mathbb{V}[Y] = \mathbb{V} \left[ \sum_{i=1}^N X_i \right] = \sum_{i=1}^N \mathbb{V}[X_i] = N\theta(1 - \theta)$$

# The Binomial Distribution:

## Example

Ashton is an excellent off spinner. The probability of him getting a wicket during a cricket match is  $\frac{1}{4}$ . (That is, on each attempt, there is a 1/4 chance he will get a wicket.)

His coach commands him to make 10 attempts of wickets in a particular game.

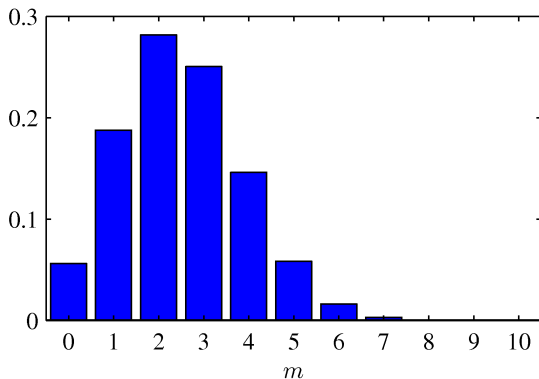
- 1 What is the probability that he will get exactly three wickets?  
 $\text{Bin}(3|10, 0.25)$

- 2 What is the expected number of wickets he will get?  
 $\mathbb{E}[Y]$ , where  $Y \sim \text{Bin}(\cdot|10, 0.25)$ .  $10 \times 0.25 = 2.5$

- 3 What is the probability that he will get at least one wicket?  
 $\sum_{m=1}^{10} \text{Bin}(m|N=10, \theta=0.25) = 1 - \text{Bin}(m=0|N=10, \theta=0.25)$

# The Binomial Distribution:

Example: Distribution of the Number of Wickets



Histogram of the binomial distribution with  $N = 10$  and  $\theta = 0.25$ . From Bishop (PRML, 2006)

(A plot of the function  $m \mapsto \text{Bin}(m|N = 10, \theta = 0.25)$ , for  $m \in \{0, \dots, 10\}$ )

- 1 The Bernoulli Distribution
- 2 The Binomial Distribution
- 3 Parameter Estimation**
- 4 Bayesian Parameter Estimation
- 5 Wrapping up

# The Bernoulli Distribution: Parameter Estimation

Consider the set of observations  $\mathcal{D} = \{x_1, \dots, x_N\}$  with  $x_i \in \{0, 1\}$ :

- The outcomes of a sequence of coin flips
- Whether or not there are errors in a transmitted bit string

Each observation is the outcome of a random variable  $X$ , with distribution

$$p(X = x) = \text{Bern}(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

for some parameter  $\theta$

# The Bernoulli Distribution: Parameter Estimation

We know that

$$X \sim \text{Bern}(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

But often, we **don't know** what the value of  $\theta$  is

- The probability of a coin toss resulting in heads
- The probability of the word *defence* appearing in a document about sports

What would be a reasonable **estimate for  $\theta$  from  $\mathcal{D}$ ?**



# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

Say that we observe

$$\mathcal{D} = \{0, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

Intuitively, which seems more plausible:  $\theta = \frac{1}{2}$ ?  $\theta = \frac{1}{5}$ ?

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

Say that we observe

$$\mathcal{D} = \{0, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

**If** it were true that  $\theta = \frac{1}{2}$ , **then** the probability of this sequence would be

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{10} p(x_i|\theta) \\ &= \prod_{i=1}^{10} \frac{1}{2} \\ &= \frac{1}{2^{10}} \\ &\approx 0.001. \end{aligned}$$

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

Say that we observe

$$\mathcal{D} = \{0, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

**If** it were true that  $\theta = \frac{1}{5}$ , **then** the probability of this sequence would be

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{10} p(x_i|\theta) \\ &= \left(\frac{1}{5}\right)^2 \cdot \left(\frac{4}{5}\right)^8 \\ &\approx 0.007. \end{aligned}$$

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

We can write down how likely  $\mathcal{D}$  is under the Bernoulli model. Assuming independent observations:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^N p(x_i|\theta) = \prod_{i=1}^N \theta^{x_i} (1 - \theta)^{1-x_i}$$

We call  $L(\theta) = p(\mathcal{D}|\theta)$  the likelihood function

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

We can write down how likely  $\mathcal{D}$  is under the Bernoulli model. Assuming independent observations:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^N p(x_i|\theta) = \prod_{i=1}^N \theta^{x_i} (1 - \theta)^{1-x_i}$$

We call  $L(\theta) = p(\mathcal{D}|\theta)$  the **likelihood function**

**Maximum likelihood principle:** We want to **maximize** this function wrt  $\theta$

The parameter for which the observed sequence has the highest probability

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

Maximising  $p(\mathcal{D}|\theta)$  is equivalent to maximising  $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$

转积为和

$$\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^N \log p(x_i|\theta) = \sum_{i=1}^N [x_i \log \theta + (1 - x_i) \log(1 - \theta)]$$

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

Maximising  $p(\mathcal{D}|\theta)$  is equivalent to maximising  $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^N \log p(x_i|\theta) = \sum_{i=1}^N [x_i \log \theta + (1 - x_i) \log(1 - \theta)]$$

Setting  $\frac{d\mathcal{L}}{d\theta} = 0$  we obtain:

$$\theta_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i$$

# The Bernoulli Distribution: Parameter Estimation:

## Maximum Likelihood

Maximising  $p(\mathcal{D}|\theta)$  is equivalent to maximising  $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^N \log p(x_i|\theta) = \sum_{i=1}^N [x_i \log \theta + (1 - x_i) \log(1 - \theta)]$$

Setting  $\frac{d\mathcal{L}}{d\theta} = 0$  we obtain:

$$\theta_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i$$

The proportion of times  $x = 1$  in the dataset  $\mathcal{D}$ !



# The Bernoulli Distribution:

## Parameter Estimation — Issues with Maximum Likelihood

Consider the following scenarios:

- After  $N = 3$  coin flips we obtained 3 ‘tails’
  - ▶ What is the estimate of the probability of a coin flip resulting in ‘heads’?
- In a small set of documents about sports, the words *defence* never appeared.
  - ▶ What are the consequences when predicting whether a document is about sports (using Bayes’ rule)?

# The Bernoulli Distribution:

## Parameter Estimation — Issues with Maximum Likelihood

Consider the following scenarios:

- After  $N = 3$  coin flips we obtained 3 ‘tails’
  - ▶ What is the estimate of the probability of a coin flip resulting in ‘heads’?
- In a small set of documents about sports, the words *defence* never appeared.
  - ▶ What are the consequences when predicting whether a document is about sports (using Bayes’ rule)?

These issues are usually referred to as **overfitting**

- Need to “smooth” out our parameter estimates
- Alternatively, we can do Bayesian inference by considering **priors** over the parameters

- 1 The Bernoulli Distribution
- 2 The Binomial Distribution
- 3 Parameter Estimation
- 4 Bayesian Parameter Estimation**
- 5 Wrapping up

# The Bernoulli Distribution:

## Parameter Estimation: Bayesian Inference

Recall:

$$\underbrace{p(\theta|X)}_{\text{posterior}} = \frac{\overbrace{p(X|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

If we treat  $\theta$  as a random variable, we may have some prior belief  $p(\theta)$  about its value

- e.g. we believe  $\theta$  is probably close to 0.5

Our prior on  $\theta$  quantifies what we believe  $\theta$  is likely to be, before looking at the data

Our posterior on  $\theta$  quantifies what we believe  $\theta$  is likely to be, after looking at the data

# The Bernoulli Distribution:

Parameter Estimation: Bayesian Inference

The **likelihood** of  $X$  given  $\theta$  is

$$\text{Bern}(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

For the **prior**, it is mathematically convenient to express it as a **Beta distribution**:

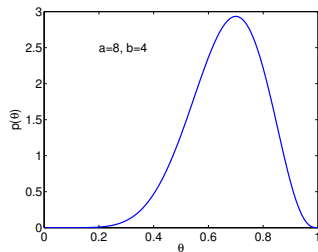
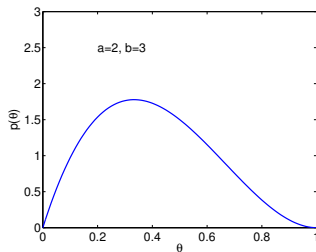
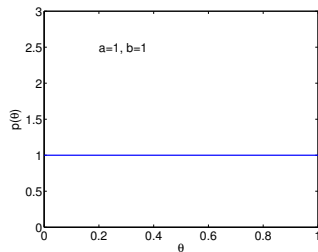
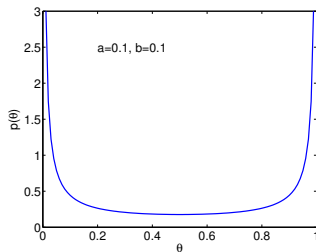
$$\text{Beta}(\theta|a, b) = \frac{1}{Z(a, b)} \theta^{a-1} (1 - \theta)^{b-1},$$

where  $Z(a, b)$  is a suitable normaliser

We can tune  $a, b$  to reflect our belief in the range of likely values of  $\theta$

# Beta Prior

## Examples



# Beta Prior and Binomial Likelihood:

## Beta Posterior Distribution

Recall that for  $\mathcal{D} = \{x_1, \dots, x_N\}$ , the likelihood under a Bernoulli model is:

$$p(\mathcal{D}|\theta) = \theta^m (1 - \theta)^\ell,$$

where  $m = \#(x = 1)$  and  $\ell \stackrel{\text{def}}{=} N - m = \#(x = 0)$ .

# Beta Prior and Binomial Likelihood:

## Beta Posterior Distribution

Recall that for  $\mathcal{D} = \{x_1, \dots, x_N\}$ , the likelihood under a Bernoulli model is:

$$p(\mathcal{D}|\theta) = \theta^m(1 - \theta)^\ell,$$

where  $m = \#(x = 1)$  and  $\ell \stackrel{\text{def}}{=} N - m = \#(x = 0)$ .

For the prior  $p(\theta|a, b) = \text{Beta}(\theta|a, b)$  we can obtain the posterior:

$$\begin{aligned} p(\theta|\mathcal{D}, a, b) &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{p(\mathcal{D}|a, b)} \\ &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{\int_0^1 p(\mathcal{D}|\theta)p(\theta|a, b)d\theta} \\ &= \text{Beta}(\theta|m + a, \ell + b). \end{aligned}$$



# Beta Prior and Binomial Likelihood:

## Beta Posterior Distribution

Recall that for  $\mathcal{D} = \{x_1, \dots, x_N\}$ , the likelihood under a Bernoulli model is:

$$p(\mathcal{D}|\theta) = \theta^m(1 - \theta)^\ell,$$

where  $m = \#(x = 1)$  and  $\ell \stackrel{\text{def}}{=} N - m = \#(x = 0)$ .

For the prior  $p(\theta|a, b) = \text{Beta}(\theta|a, b)$  we can obtain the posterior:

$$\begin{aligned} p(\theta|\mathcal{D}, a, b) &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{p(\mathcal{D}|a, b)} \\ &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{\int_0^1 p(\mathcal{D}|\theta)p(\theta|a, b)d\theta} \\ &= \text{Beta}(\theta|m + a, \ell + b). \end{aligned}$$

Can use this as our new prior if we see more data!

# Beta Prior and Binomial Likelihood:

## Beta Posterior Distribution

Now suppose we choose  $\theta_{\text{MAP}}$  to maximise  $p(\theta|\mathcal{D})$   
(MAP= Maximum *A Posteriori*)

One can show that

$$\theta_{\text{MAP}} = \frac{m + a - 1}{N + a + b - 2}$$

cf. the estimate that did not use any prior,

$$\theta_{\text{ML}} = \frac{m}{N}$$

The prior parameters  $a$  and  $b$  can be seen as adding some “fake” trials!

What values of  $a$  and  $b$  ensure  $\theta_{\text{MAP}} = \theta_{\text{ML}}$ ?  $a = b = 1$ . Make sense?  
(Note that the choice of the beta distribution was not accidental here — it is the “conjugate prior” for the binomial distribution. )

- 1 The Bernoulli Distribution
- 2 The Binomial Distribution
- 3 Parameter Estimation
- 4 Bayesian Parameter Estimation
- 5 Wrapping up

# Summary

- Distributions involving binary random variables
  - ▶ Bernoulli distribution
  - ▶ Binomial distribution
- Bayesian inference: Full posterior on the parameters
  - ▶ Beta prior and binomial likelihood  $\rightarrow$  Beta posterior
- **Reading:** Mackay §23.1 and §23.5; Bishop §2.1 and §2.2

## Next time

- Entropy

# Acknowledgement

These slides were originally developed by Professor Robert C. Williamson.