Problems for Griffiths' Electrodynamics

Contents

1 Chapter 1: Probablities and Interference (Mackay Ch 2-3) 2

2 Chapter 2: Probabilities and Interference (Mackay Ch 2-3) 6

3 Chapter 3: Maximum Likelihood *Approximation* 10

4 Module 0: Quantitative Imaging 16

Homework 1 32

Homework 2 44

Homework 5 54

Homework 6 56

481 Lecture 1/16/24

1 Chapter 1: Probablities and Interference (Mackay Ch 2-3)

An ensemble: x random variable

$$A_x = (a_1, a_2, \ldots, a_n)$$
$$P_x = (p_1, p_2, \ldots, p_n)$$
$$p(x = a_i) = p_i$$

x takes value a_i with probability p_i

$$p \geq 0, \quad \sum_{a_i \in A_x} p(x = a_i) = 1$$

Short hand for $p(x = a_i)$ is $p(a_i), \quad p(x)$
Joint ensemble: X, Y ensembles

$$XY = \text{ordered pairs}(x, y) \quad x \in A_X, y \in A_Y$$
$$P(x, y) = \text{joint probability of x and y}$$

Marginal probability: $P(x, y) \to P(x), P(y)$

$$P(x) = \sum_{y \in A_y} P(x, y)$$
$$P_x(x = a_i) = \sum_{b \epsilon A_y} P_{XY}(x = a_i, y = b)$$

Conditional probability:

$$P(x = a_i | y = b_j) = \frac{P(x = a_i, y = b_j)}{P(y = b_j)}$$

"Probability of $x = a_i$ given that $y = b_j$ (is true)"

Example 1 $XY = 2$ successive letters in english alphabet. P_x and P_y are identical 'frequency of a letter in english'

$$A_{xy} = \{aa, ab, ac, \ldots, zz\}$$

$$P(y | x = \text{'}q\text{'})$$

Peak at $y = \text{'}u\text{'}$

$$\neq P_Y(y)$$

because x and y are not independent
X, Y "independent" if (and only if) $P(x, y) = P(x)P(y)$
Userful relations: $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$
For any assumption H

$$\forall H : \qquad P(x, y|H) = p(x, y|H)p(y|H)$$

'Sum rule':

$$P(x|H) = \sum_{y \in A_y} P(x, y|H) = \sum_{y \in A_y} P(x|y, H)P(y|H)$$

Lecture 1/18

Last time: Main point $P(y|x) \neq P(y)$
Useful relations: Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

where the joint relation is

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

this can be rewritten into *Baye's theorem*

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Example 2: Apply Baye's theorem Alex is test for a nast disease.

- Disease status: a (sick or healthy)
- Test outcome: b (positive or negative)

"Test is 95% reliable" or

$$P(+|\text{sick}) = 0.95, \quad P(-|\text{healthy}) = 0.95$$

Disease is nasty but rare $P(\text{sick}) = 0.01$; $P(\text{Healthy}) = 0.99$
Test is positive, what is the probability that Alex is sick? $P(\text{sick}|+) =?$

Solution Use Baye's theorem:

$$P(\text{sick}|+) = \frac{P(+|\text{sick})P(\text{sick})}{P(+)}$$

where $P(+)$ is the probability of a positive test result. This can be found using the sum rule

$$P(+) = P(+|\text{sick})P(\text{sick}) + P(+|\text{healthy})P(\text{healthy})$$

Thus

$$P(\text{sick}|+) = \frac{0.95 * 0.01}{0.95 * 0.01 + 0.05 * 0.99} = 0.161$$

It is useful to write the probabilities in a table

	$b = +$	$b = -$	$P(b)$
$a = \text{sick}$	$0.95 * 0.01$	$0.05 * 0.01$	0.01
$a = \text{healthy}$	$0.05 * 0.99$	$0.95 * 0.99$	0.99
$P(a)$	0.161	0.839	1

where columns represent the 95:5 reliable test.

Exclam!

$$P(S|+) \neq P(+|S)$$

A brief philosphical interlude... The 'Bayesian viewpoint':
Probability as degree of beliefs in propositions given assumptions & evidence, or Probability as 'freq of outcomes in repeat random experiments'

Forward and inverse problems

So far we have talked about Cond Prob, Baye's thrm, and and example.

Generative Model: Parameters $\Theta \to P(D|\Theta) \to (P)$ outcomes (data) AKA 'forward problem'
'a model' predicts an outcome given parameters. The model is a probablity distribution due to all the uncertainties and errors we have in the real world.

The Inverse Problem $P(\Theta|D)$
The inverse problem is the opposite of the forward problem (obviously). Also related to the issues regarding 'inference' and using Baye's theorem.

Example 3: A forward problem
An urn contains K balls, B balls are black, and $K - B$ balls are white. A ball is drawn at N times with replacement.

- $n_B = \#$ of times a black ball is drawn
- $P(n_B)$, average n_B?, STD?

With

$$f_B = \frac{B}{K}$$

The probability is given by the binomial distribution

$$P(n_B|N, f_B) = \binom{N}{n_B} f_B^{n_B}(1 - f_B)^{N-n_B}$$

The mean is $N * f_B$ and the STD is $\sqrt{N * f_B * (1 - f_B)}$

Example 4: An inverse problem
We have 11 urns, each with 10 balls. u is the number of black balls in each urn and the urns have $u = 0, 1, \ldots, 10$ black balls. Alex selects an urn at random and draws N balls at random with replacement. Bob wates Alex, but does not know which urn u was selected. For Bob, what is $P(u|N, n_B)$?
We have the data, but we are trying to infer the parameter u

Solution Use Baye's theorem

$$P(u|N, n_B) = \frac{P(n_B|u)P(u)}{P(n_B)}$$

where $P(n_B|u)$ is the 'forward' part from Ex 2, $P(u) = 1/11$, and $P(n_B)$ is the 'normalization' that makes it a valid prob. distribution:

$$P(n_B) = \sum_{u'} P(n_B|u')P(u')$$

Therefore

$$P(u|N, n_B) \propto \binom{N}{n_B} \left(\frac{u}{10}\right)^{n_B} \left(1 - \frac{u}{10}\right)^{N-n_B}$$

e.g. $n_B = 3, N = 10$
insert figure 1.2
The (0,0) point is impossible because we picked 3 black balls, and the urn $u = 0$ has no black balls. The same is true for the (10,10) point. The most likely point is $u = 3\ldots$

Exclam! This is known as 'Posterior Probabilty'

- Θ is the parameter
- D is the data
- $P(\Theta)$ is the prior
- $P(D|\Theta)$ is the likelihood: a function of D prob of data given param (sums to 1 over all options for D). As a function of $\Theta \rightarrow$ likelihood of Θ
- $P(\Theta|D)$ is the posterior
- $P(D)$ is the normalization

! **Probability of *data***

! **Likelihood of *parameters***

Role of Prior:

! You can't do inference without making assumptions

Lecture 1/23/24

Last time:

- Forward $p(data|param)$
- Inverse $p(param|data)$

Using Baye's theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{norm}}$$

Note: You can't do inference w/o working assumptions (prior) priors are subjective. From the inverse problem ex from last week: what is the probability that next ball Alex draws is black?

$$P(B) = \sum P(u)P(B|u)$$

Note: Infereince $\neq$ decision/choice of model. Inference is assigning probabilties to hypothesese.

Problem USB Cable frustrations "It takes 3 tries to plug in a USB cable"
During our first try to plug in the cable, we are collecting data. And if its wrong, we 'believe' that the orientation is wrong, thus we flip it believing that the 2nd try is the correct one. But in fact, this is wrong and the 3rd try is the correct one.

How to collect data?

Lecture 1/25/24

---

2 Chapter 2: Probabilities and Interference (Mackay Ch 2-3)

Example 5: Tossing a coin

- 3 times: H, H, H
- 10 times: H, H, ... H

what is the probability of the next toss being H?

Ex 5.1 Coin with freq of heads f_H is tossed N times and n_H heads. What is the probability of the next toss being H? (Ex 4 but with fixed unknown parameter)
Prior: subjective assumption (e.g. could be uniform) then do inference.

Ex 5.2 N tosses, n_H heads. What is the probability that the coin is biased? (Model Comparison)

Lecture 1/30/24

Last time: Simple inference (within a model) where we solve for $p(data|param)$ and now we move on to model comparison!

Ch 2: Model Comparaison Mackay Ch 3 & 28

A coint that is possibly bent has a frequency of heads f_H. For $N = 100$ tosses, $n_H = 90$ heads which is definitely a bent coin (biased).
For the case $N = 100$, $n_H = 55$, we are not sure if the coin is biased or not. The best fit to data is $f_H = 0.55$ we say that it is probably not bent from our intuition.
For the case $N = 10000$, $n_H = 5500$ we believe that the coin is more likely to be 'bent'

Which model? We know that the fair coin model fits the model less than the bent coin model, but we believe that the fair coin model fits the data better than the bent coin model. From "Occam's Razor" (simplicity): Accept the simplest explanation that fits the data. We would prefer the simpler fair coin model since it is simpler. This is mereley a ad hoc rule of thumb. But Bayesian Calculus naturally implements Occam's Razor.

Comparing hypthesis H_o (fair coin) and H_1 (bent coin) Warning! We should choose the hypothesis set before we see the data, otherwise it is cheating!

Big Picture Two levels of inference

- Level 1: Hypothesis set H_o with parameter f_H: Inferring $P(p_a) =?$
- Level 1: Hypothesis set H_o no params: no inference
- Level 2: Hypothesis set H_o, H_1: Inferring both $P(H_o)$ and $P(H_1)$

2.1 Coin tosses: 1-param model H_1 (L1 inference)
Outcomes: $X = \{a, b\}$ for heads and tails with probabilities p_a and $p_b = 1 - p_a$
Assumption: The prior on p_a is uniform
F Tosses: data = sequence, s = aaba... with $F_a = \#$ of a's and $F_b = \#$ of b's; $F_a + F_b = F$
The model:

$$P(s|p_a, F, H_1) = p_a^{F_a}(1 - p_a)^{F_b}$$

since the tosses are are specific sequence e.g. aaba... From the definition of H_1

$$p_a \in [0 \dots 1]$$

is equiprobable and the prior tells use that $p(p_a) = 1$

Questions Given a sequence s of F observations, with $\# \ a = F_a$ and $\# \ b = F_b$,

1. What is my posterior belief about p_a? or $P(p_a) =?$
2. What is the probability that next draw is a?

As this is a inverse problem, we use Baye's theorem

$$P(p_a|s, F, H_1) = \frac{P(s|p_a, F, H_1)P(p_a)|H_1}{P(s|F, H_1)}$$

the bottom takes the full probabilty of the data no matter the value of p_a and is the normalization

$$= \frac{p_a^{F_a}(1-p_a)^{F_b}(1)}{\int_0^1 p_a^{F_a}(1-p_a)^{F_b} dp_a}$$

where we use the sum rule for the denominator

$$\sum_{p_a} P(s|p_a, F, H_1)P(p_a|H_1)$$

but since it is a continuous variable, we use the integral instead of the sum. The math gives us the gamma function

$$\text{normalization factor} = \frac{F_a!F_b!}{(F_a+F_b+1)!}$$

Examples $s = aba$ vs $s = bbb$

$$P(p_a|s=aba) \propto p_a^2(1-p_a) \quad \text{vs} \quad P(p_a|s=bbb) \propto (1-p_a)^3$$

The first looks like a parabola and the second looks like a decaying cubic function. In each case, the most probable p_a is $2/3$ and 0 respectively which is shown by the data.

Probability of next toss is a We need to integrate over the prior to get the probability of the next toss being a.

$$P(\text{next}=a) = \int \mathrm{d}p_a\, P(\text{next}=a|p_a)P(p_a|s,F,H_1) = \int \mathrm{d}p_a\, P(p_a|s,F,H_1)p_a = \text{average of } p_a$$

the average of p_a for the first example is $3/5 = 0.6$ and for the second example is $1/5 = 0.2$

Conclusion: We found Probability of s given p_a and H_1 (Data given biased coin model) and the probability of p_a given s, F, H_1 (inference), or forward and inverse probabilities for the biased coind model H_1.

2.2 Zero-parameter model H_o (Fair coin) & model comparison where $p_a = 1/2$. The forward probability is

$$P(s|H_o) = \frac{1}{2^F}$$

Question: Given a string of F observations, what comparison can we make between the biased coin model and the fair coin model, H_o vs H_1?
The Hypothesis space is now $\{H_o, H_1\}$ where only models are under consideration. Using Baye's theorem again

$$P(H_o|s,F) = \frac{P(s|F,H_o)P(H_o)}{P(s|F)}$$

and

$$P(H_1|s,F) = \frac{P(s|F,H_1)P(H_1)}{P(s|F)}$$

where $P(s|F) = \sum_{H\in\{H_o,H_1\}} P(s|F,H)P(H)$. looking at the ratio of the two probabilities

$$\frac{P(H_1|s,F)}{P(H_0|s,F)} = \frac{P(s|F,H_o)}{P(s|F,H_1)}\frac{P(H_1)}{P(H_0)}$$

where the first fraction is what the data told us, and the second fraction is what we know before (prior).

Lecture 2/1/24

Last time: We discussed the zero-parameter model H_o (fair coin) and the one-parameter model H_1 (biased coin). We used Baye's theorem to compare the two models to find the ratio of the two probabilities

$$\mathcal{R} = \frac{P(H_1|s,F)}{P(H_o|s,F)} = \frac{P(H_1)}{P(H_o)} \frac{P(s|F,H_o)}{P(s|F,H_1)}$$

where we set no a priori model (prior) preference, so $P(H_1) = P(H_o) = 1/2$. So the ratio is

$$\mathcal{R} = \frac{P(s|F,H_1)}{P(s|F,H_o)} = \frac{\frac{F_a!F_b!}{(F_a+F_b+1)!}}{\frac{1}{2^F}} = \frac{2^F F_a!F_b!}{(F+1)!}$$

what does this plot look like? As the number of tosses goes to infinity, this ratio will go to the truth! Simulation is shown by Figure 2.1. where the the bent coin $p_a = 0.9$ probability goes to infinity as well

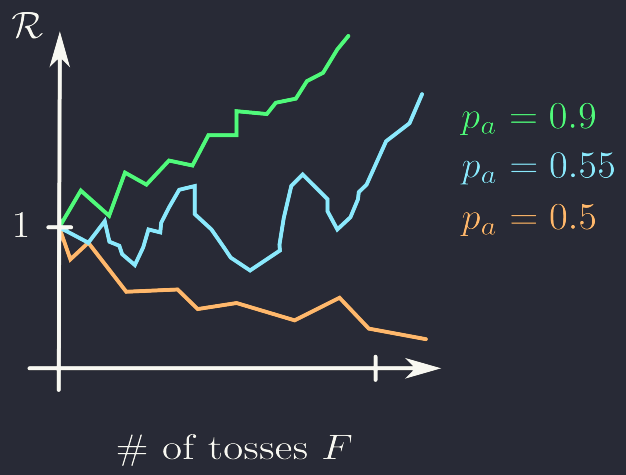


Figure 2.1: Ratio of the two probabilities as a function of the number of tosses

as the slightly biased coin (but at a slower pace) and the fair coin goes to zero. We know this from the probability

$$P(s|F,H_o) = \int_0^1 P(s|p_a,F,H_1)P(p_a|F,H_1)dp_a$$

NOTE: There exists a p_a that fits data better than H_o, but this evidence term includes averaging over p_a
Bayes theorem in the context of model comparison

$$\text{bayes} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

TAKEHOME: Bayesian model comparison naturally includes Occam's Razor!

2.4 P-values? Why not just use p-values? e.g.

$$F = 250 \qquad F_a = 141, F_b = 109$$

Do these data suggest that the coin is biased?

P-value: Probability to get data as extreme or move, assuming the null hypothesis is true.

- Null hypothesis: Coin is fair (H_o)
- Our hypothesis: Coin is biased (H_1)
- mean $= F/2$
- $\sigma = \sqrt{F}/2$
- Our observation: $\frac{F_a - F/2}{\sqrt{F}/2} = 2.02\sigma$
- p-value $= 0.0497 < 0.05$!!!!

Google "a small p-value (< 0.05) indicates strong evidence against the null hypothesis so you reject it"

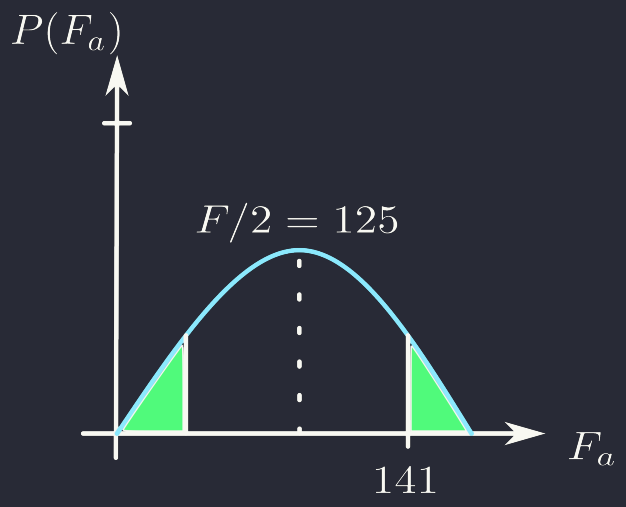


Figure 2.2: Finding p-value based on the Gaussian distribution

From sterling approximation

$$ln(k!) \approx kln(k) - k + \dots$$

With uniform prior on p_a

$$\mathcal{R} = \frac{2^2 50141!109!}{251!} = 0.61$$

if anything, there is weak evidence *against* coin being biased.

Non-uniform priors? For a reasonable family of priors, across the entire set of priors, strongest evidence for bias is $2.5:1$ (From Mackay) This differs from the p-value which is $20:1$.

3 Chapter 3: Maximum Likelihood *Approximation*

(Ch 22 Mackay)

GOAL: Connect to the stat you may have seen before. Going back to Example 4 (Urns and more urns)

- Unkown $u*$ selected at random
- 10 draws (with replacement): 3 black

- P(next draw = black) =?
- Most likeley $u : 3 \to$ predicts 0.3
- Correct answer: predicts 0.33

but the two numbers are kinda similar...
NOTE: Bayesian model comparison, not model selection, but complete enumeration of hypothesese (integration over hyp space) is computationally expensive (especially in high dimensions)
e.g. Comparing 2 models:

- 1 Gaussian: 2 parameters μ, σ
- 2 Gaussian ($a_1 G_1 + a_2 G_2$): 5 parameters $\mu_1, \sigma_1, \mu_2, \sigma_2, a_1/a_2$

This problem of an increasing number of parameters motivates *Max likelihood (ML) approximation*: instead of enumeration, focus on *1* hypothesis that maximized the likelihood function.

Max Likelihood Estimation (MLE)

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

instead of [assuming prior $\to$ compute posterior $\to$ integrate over hyp space] we just [compute the likelihoood unction $\to$ maximize it] (MLE).

3.1 A single Gaussian

- Data: $\{x_n\} \quad n = 1, \ldots, N$
- model: these observations were sampled from a gaussian with probability

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  where we have 2 parameters μ, σ to determine.

Log likelihood (multiplying likelihoods is hard, adding log likelihoods is easier)

$$\begin{aligned}\ln P(\{x_n\}|\mu,\sigma) &= \sum_{n=1}^{N} \left(-\ln\sqrt{2\pi\sigma^2} - \frac{(x_n-\mu)^2}{2\sigma^2} \right) \\ &= -N\ln\sqrt{2\pi\sigma^2} - \frac{N}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\end{aligned}$$

Sufficient statistics: Denote

$$\hat{x} \equiv \sum_n \frac{x_n}{N} \quad \text{empirical mean}$$

$$S = \sum_n (x_n - \hat{x})^2 \quad \text{sum of square deviations}$$

These two numbers refer to the sufficient statistics. From these we get the log likelihood

$$\ln P = -N\ln\sqrt{2\pi\sigma^2} - \frac{N(\mu-\hat{x})^2 + S}{2\sigma^2}$$

Thus the max likelihood estimate of μ, σ are

$$\mu_{ML} = \hat{x}$$

$$\sigma_{ML} = \sqrt{\frac{S}{N}} = \sqrt{\frac{\sum_n (x_n - \hat{x})^2}{N}}$$

If σ is known, then $P(\mu)$ is a Gaussian we know that $\sigma/\sqrt{N}$ is the width of the likelihood (error bars)

Lecture 2/6/24

Last time: We discussed familiar stats.

- Bayes Calculus in terms of $P(\theta)$ (params). Predictions of x

$$P(x) = P(x|\theta)P(\theta)d\theta \quad \text{is computationally hard}$$

- MLE: instead of full enumeration, focus on 1 hypothesis and its max likelihood

3.1 Fitting a single Gaussian

$$\theta = \{\sigma, \mu\} \quad P(D|\theta) = \prod_n^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right)$$

we get the sufficient stats

$$\mu_{ML} = \hat{x} = \frac{\sum_i x_i}{N}$$
$$\sigma_{ML} = \sqrt{\frac{\sum_i (x_i - \hat{x})^2}{N}}$$

Beyond the MLE: we can get the error bars on μ AKA "Standard error of the mean": $\sigma/\sqrt{N}$

HW 2 HINTS

- MAX LIKELIHOOD WORKS (WELL) FOR PREDICTIONS/ ESTIMATES WHEN MOST OF THE PROB WEIGH IS NEAR THE ML ESTIMATE

  THIS IS NOT ALWAYS THE CASE! (most of the prob weight can be located not near the ML, Most of the prob weight is around the center)

  e.g. For two gaussian with 2 clusters, fitting the model with 1 gaussian may have a super narrow but the MLE will tend to that narrow peak even though the data is not near that peak.

- MOST LIKELY $\neq$ TYPICAL / REPRESENTATIVE (Mackay 22)

3.2 Least square fitting: e.g. linear fit

- Dat: $\{y_n\}$ for each $\{x_n\}$
- Model: $y_n = ax_n + b+$ Gaussian noise of width σ
- Given x_n, σ, the params are a, b

Model (more formally):

$$P(y_n|x_n, a, b, \sigma) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - (ax_n + b))^2}{2\sigma^2}\right)$$

How do I infer a, b using the MLE: Log likelihood!

$$\ln P = C - \sum_{n=1}^N \frac{(y_n - (ax_n + b))^2}{2\sigma^2}$$

where C is a constant, and we must maximize over a, b. Maximing $\ln P$ over a, b is equivalent to minimizing sum of squares of residuals (deviation of y_n from the a, b).

! (a) Not magic or ad hoc

! (b) This is For Gaussian errors *only* (of same magnitude). LSQ $\leftrightarrow$ Gaussian

Takehome: MLE is widely use & often very sensible, but MLE $\neq$ not a silver bullet especially in high dimensions! (e.g. HW2)

Real world Example! How sensitive are our eyes?

- Participants look at dim flashes in a dark room overa time t with a height of the flash A (brightness)
- How low can A be for the flash to be detected?
- Experiemnt E_1: Flashes arrive randomly at some average rate. e.g. a flash but no response is a false negative while a false positive is a response but no flash (1 per 10 sec on average).
- Experiment E_2: First a bright pulse A_o (or beep of possible oncoming flash) that is easy to see, then 1 sec later, the there is either a flash of heigh A or no flash at all with prob p.

In both cases, both make A dimmer and measure for accuracy. We would expect that E_2 would allow us to detect dimmer flashes since we can expect.

Ground truth For E_2 when we know when to expect we let $f = 0$ as no flash and $f = 1$ a flash. For the perfect detector and noisy detector we have Figure 3.1. There also exista abackground noise b that is always present.

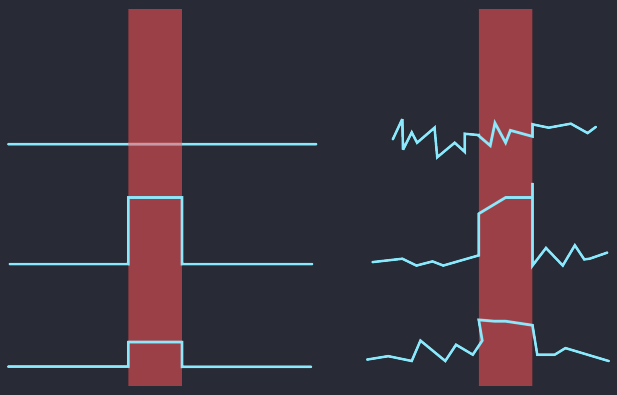


Figure 3.1: From top to bottom we have a no light $f = 0$, and two cases of $f = 1$ for a bright light and dim light. The Pefect detector (left) sees and appropriates with the correct response while the noisy (Gaussian) detector may have a incorrect response (especially for the dimmer signal).

Data For a noise time trace $I(t)$ over 5 seconds, we have a probability of a flash $P(f) \approx 0.5$.

$$P(D|A, f, \eta, E_2)$$

with parameters $A; f, \eta$ and the simplest version: A, η given an inference of f

E_2 The hypothesis space we have either 'Flash' or 'No Flash'. The expected model is a flash or no flash with Gaussian noise. We know the A and η. The parameters to infer are $f = 0, 1$ and the inference questions is $f =?$

E_1 The hypothesis: H_1 flash at t, H_o no flash. The model has known: A, η, b. Parameters: $f = 0, 1$ and t. Inference question: H_1 or H_o? Figure 3.2 shows the expectation of the model.

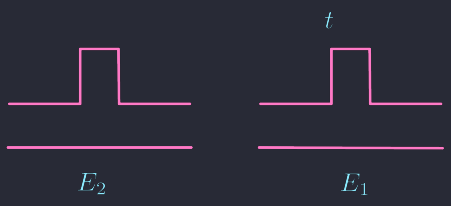


Figure 3.2: The expectation of the models given experiment E_1 and E_2. The top is for an expected model of a flash and no flash for bottom. NOTE that there also is Gaussian noise η added to both scenerios.

Approach we have $P(D|\text{param}) \to$ Bayes' Theorem

- E_2: Bayes' Theorem $\to P(f|D,\eta,A,b)$. If $f=1$ we are more likely to say we *saw it* with an error probability: (average of the probability of making a mistake over all data including False Positives and False Negatives)

$$\langle P(\text{wrong f}|D,\eta,A,b)\rangle \quad \text{data}$$

the error rate is a complicated integral (an average is a sum/trace/integral!):

$$\text{Error rate}(A,\eta,b) = \int \text{ddata}\, P(f=1|D)P(D|f=0,A,\eta)$$

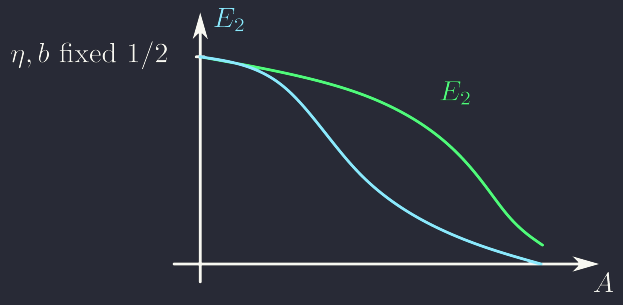


Figure 3.3: The error rate as a function of the brightness of the flash.

Simpler approach? We define I^* as a mean intensity over a window of interest. For E_2 we can easily find the window of interst, but for E_1 we could discriminate the window by finding the brightest flash and comparing it some threshold. Here lies two questions: how does a computer that computes whether or not there is a flash versus a human that is looking for a flash after 5 seconds.

If η is known, $P(D|f,A,b)$ depends only on I^* (sufficient statistics).

Version 2: Data: I^* is just *one* number. The probability given no flash or flash. Redefining noise η as expected noise of measurement over window length. As shown in Figure .
In E_2 we have a Gaussian distribution of the flash and no flash models, but in the E_1 the flash model is the same as we take the same window length of interest, but for the no flash model the model moves to the right as we have a likelihood of measuring a window length with MORE noise. The error probability for E_2 is: Looking at the midpoint of the two models, we can find the error as a sum of tail distribution (finding the weight of the outliers).

$$\text{error} = \int_{A/2}^{\infty} \frac{1}{\sqrt{2\pi\eta^2}\exp\left(-\frac{x^2}{2\eta^2}\right)}$$

the error is shown in Figure . If human interaction is close to Bayesian $\to$ specific *quant* prediction for perfomance, effect of having the cue, rate of P.

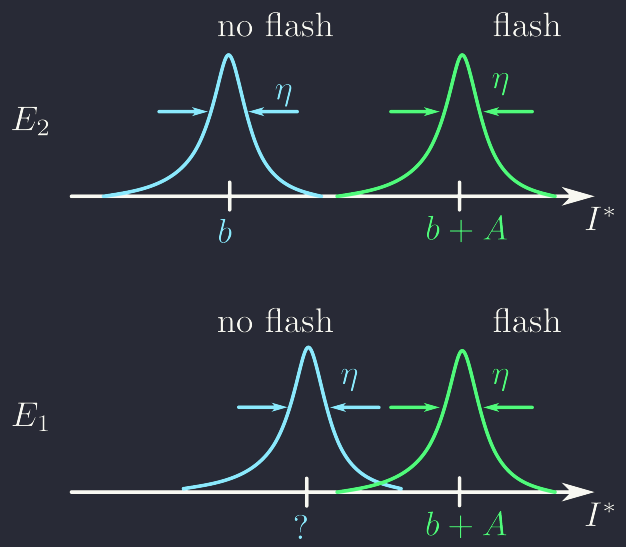


Figure 3.4: There is a shift in the no flash model in E_1

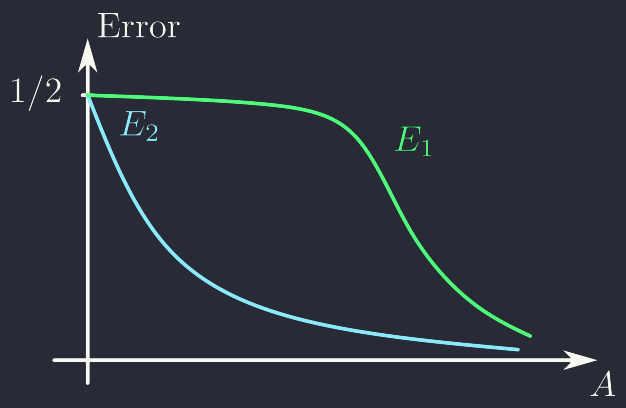


Figure 3.5: The error rate as a function of amplitude A.

Takehomes

- What is data? (non-trivial question)
- What is hyp? (not unique)
- Most straightforward method can be impossible
- Under the hood: Still Bayesian calculus.

4 Module 0: Quantitative Imaging

Lecture 2/8/24

Is Science Solved? The steps of science:

1. Gather Data + Build Model
2. Fit each model to data
3. Assign preferences to different models
4. Either Method 1: Choose which data to gather next, gather more data, and back to step (2),or Method 2: Decide whether to create new model, create new model, and back to step (2)

The not 'just math' part: Data → clever choice of features → model the features/ model the noise.

Takehomes:

- Choices in dataprocessing, Feature definitions, choice of data acquisition
- Depends on scieniic question: you need to know you subjective. Depends on the measurement: you need to know you experiment

Fly embyro patterning = perfec example; astonishingly precise, thus the precision of data anaylsis is the limiting factor

Role: Carries 'positional information'
If I konw Hb (hunchback protein concentration), I know somthing about where I am; Hb and x are not independent. Nature (funnel shaped) vs. Cell (narrow tube shaped) article arguments.

Lecture 2/15/24

Presentation: Entropy & Mutual Info

takehome: Information content in a random variable $X \rightleftharpoons$ Entropy $H(X)$

! Not arbitrayry, but natural & unique: Info content $H(X)$ has a *discrete distribution*

i $H(\{p_i\}) \geq 0$

ii $H(\{p_i\}) = 0$ iff $p_i = 1$ and others are 0

iii If X, Y are independent $Z = (X, Y)$ and $H(Z) = H(X) + H(Y)$

iv $p_i = \{1/N, 1/N, \ldots, 1/N\}$, $H(X)$ should be monotonic in N

Strengths of MI: Not arbitrary: X, Y are not independent, so $P(X) \neq P(X|Y)$

i $I(X;Y) = I(Y;X)$

ii $I(X;Y) \leq H(X)$

iii $I(X;Y) = 0$ iff X, Y are independent

iv If X, Y are related deterministically, then $I(X;Y) = H(X) = H(Y)$

If we know that they are not indepenent $H(X,Y) = H(X|Y) + I(X,Y) + H(Y|X)$ and $H(X,Y) \neq H(X) + H(Y)$ (there is overlap for dependent variables)

MOST IMPORTANT THING: Data processing inequality: "Data processing can only destroy information". X only knows about Y and Z only knows about X,

$$X \to Y \to Z; \quad P(X,Y,Z) = P_x(X)P_y(Y|X)P_z(Z|Y)$$

thus

$$I(X;Y) \geq I(X;Z)$$

Weaknesses:

! Estimating from data can require a lot of data.

! Information $\neq$ Useful Information. i.e. pure noise can have a *lot* of information

Lecture 2/20/24

Personal Thoughts: When we make a decision we have to consider what happens to the data explicitly. Normalizing does not make the data directly comparable to other data. Its easy to identify noise, so we should think twice when we compare it to other things

- Houchmandzadeh et al.: Limitations of data → Data analysis (had a subtle flaw) → one Conclusion
- Gregor et al. Better data (very careful exp) → extrene careful data analysis → opposite conclusion.

Takehomes:

- Smart people make mistakes
- If it's too good to be true, it might be???
- Data is never what you think it is
- Details matter

Lecture 2/22/24

Methods: How did we collect the data?

	Nature	Cell
Microscope	Scanning Confocal	Scanning Two Photon
Embryo	Fixed (dead)	Live
Labeling	Immunostaining	"Bcd-GFP" fusion protein

How to attach fluorophores 101:

Immunostaining: Washing off too much could wash off the pertinant proteins. It is washed multiple times to get rid of the background fluorescence. When we wash with the neutral buffer saline. First the Primary antibody (Ab) is washed off, then the second Ab, anti-rabbit, is washed off. TLDR; Fix, label, wash

- Strengths: Multiple washes leads to more(amplify) signal ; Not genetic engineering (easier); more colors!; dead sometimes an advantage for storage
- Weaknesses: Multiple washes leads to more background fluorescence; The wash may dilute?; amplifies unevenly; both random and systematic (place in cell); dead; fixation leads to deformation, shrinking, etc.

Protein fusion Genetically modify the fly to have GFP (Green Fluorescent Protein) fused to the protein of interest. TLDR; Engineer, add to protein of interest + GFP.

- Strengths: we're not adding wash; it's alive!; direct readout
- Weaknesses: GFP not bright? Limited Fluorophores (The fly has to make the bright stuff), ,so less bright less photostable; does the fusion protein still work the same?

Fluorescescing too much can lead to bleaching (death) of the protein. What kind of fluorophore is being used?

Errors & Noise: Fixed: Variable age at collection, mechanical deform, labeling efficiency (targets or not targets are labeled). Live: GFP can alter function, Impact by details of cell environment (maturation time of fluoresce).

Microscope & Imaging: Laser shoots stuff to scanning (moving mirrors) and a fluorescence detector collects data.

Confocal: Only stuff in the focal plane is collected (closer and further stuff is out of focus). Everything in the laser is fluorescing and fluorophores have a limited lifetime(can bleach quicker).

Two Photon: Infrared laser only excites the fluorophores in the place we want. In the image, the outside part has an exact concentration, so we can compare this to the inside part.

Takehomes: Expression of Protein vs Position in Embryo (the canonical example): Is 15 embryos enought? Is 1000 embryos enough? Looking at this picture: here is a plot, but is this the position in the embryo? no, it's the position in the image. Is it a picture of the image? no, its a picutre of 1D projection of the image. Is this expression? no, its tagged ptoteins... no its fluorescences... no its pixel values! More steps, more noise, more errors, there is complexity! The subject expert has the role to give us the answers to these questions. I don't know what the microscope is doing is bad!

Lecture 2/27/24

New Experiment! mRNA expression 101: We design a probe that take a mRNA sequence and attaches *several* fluorophores per mRNA strand. Some image info: The bright spots are transcription sites i.e. lots of mRNA in one place. Advantage of spot counting over intensity:

- If indeed single molecules*, robust to variations in intensity.
- Robust to background uncertainty
- Spatial Positioning
- Absolute units

Disadvantages:

- Undercounting if close together
- Detection threshold (if too dim, false positive: if too bright, false negative)
- single molecules?

The undercounting problem. Beyond what density? Resolution of your microscope i.e. point spread function PSF. ~ 1 spot per μm^3.

Recorded count vs. Tot fluorescence: Some questions to think about:

- Intercepts? Y-axis due to background fluorescence (non-zero intensity at zero mRNA)
- Scatter? Variance increases as number of independent variables increases
- Shape? Goes up,

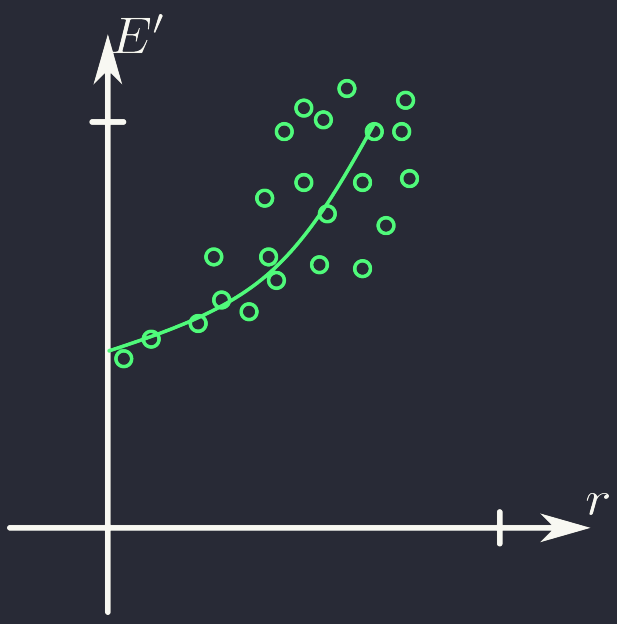


Figure 4.1: Recorded count vs. Tot fluorescence

Fitting a line gives us the slope and the intercept. The slope is the fluorescence per count. The intercept is the background fluorescence. At higher counts the slope is less linear, and gets a litter steeper due to the saturation of the fluorophores. How to pick a threshold? Bayesian Inference! If this model deviates up, we are increasingly undercounting.

False Postives & False Negatives... Some unanswered questions within Experimental detail and data analysis.

- What percent of mRNA are detected? How do we get this?
  - Count the total mRNA in the embryo: Double stain; stain the same thing twice e.g. what we want to count is red and what we dont count is green.
  - Compare with sample where we know exact mRNA count
  - Maybe there are multiple ways to label mRNA, compare those
  - Bayesian: method & number of mRNA detected?
- Are they single molecules?
  - Double stain again
- Fluctuations of the Hb and Kr are anticorrelated!

each nucleus has a δHb and δKr above or below. Claim: The anticorrelation thus Fancy theory of repression etc.

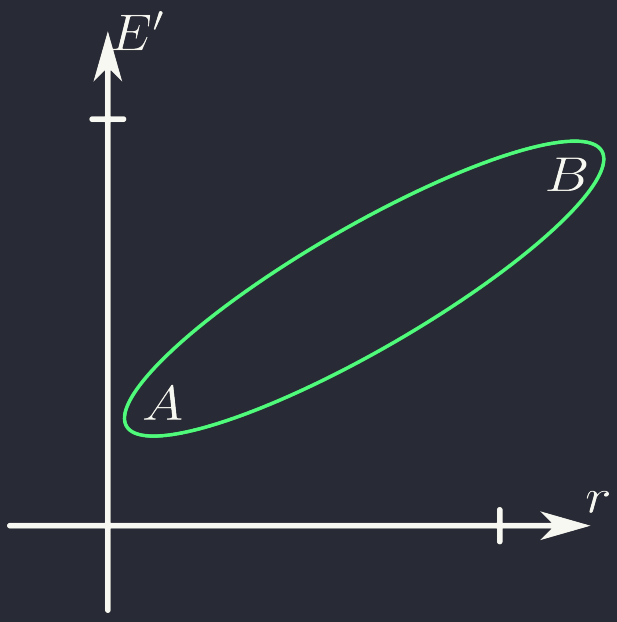


Figure 4.2: Fluctuations of the Hb and Kr are anticorrelated

Lecture 2/29/24

Last time: Single-molecule imaging:

- Spot counting vs total fluorescence
- Advantage & Issues of each

Quantitative assesment of method performance:

- If two methods, plot against each other
  - discussed expectatinons
  - info encoded in different features (and what we learned)
- leveraging a sample as its own control:
  - 2-color labeling (double stain)
  - inject extra known signal into an image:

Takehomes: Alot of similar issues within completely different fields(From imaging to sequencing coming soon).

- My thoughts: Know what physical tools you are working with(we can't just blindly trust a machine does exactly what we want it to do), without knowing how a microscope works within the two methods— confocal vs two photon—we wouldn't have known about which data is possibly better or accurate. The basic knowledge of physical tools is important! But the more you know, the better assesment you can make. Try to connect the data to the actual thing we are studying.
- You don't need to be a subject expert to ask good questions(we are novices in biology); Research papers are not textbook
- If you are the one doing this(you can't know everything within the data pipeline) i.e. in your *own* area, you *must* be an expert on details ("Oh, I'm not sure exxactly how X was done..." red flag)
- False positives and negatives: a tradeoff
- Assign weight of evidence to data points thoughtfully (least square fitting assume every data point is equal in weight and error bars) i.e. Bayesian!
- Cannot understand signal without first understanding noise i.e. your instrument, analysis, etc. what can we trust more than others? The physicist spends their first year measuring zero.
- Put error bars on error bars!
- Careful experiment goes hand in hand with careful data analysis: From the picture of the spots on a black background, the signal to noise ratio is quite high, but we must still be careful. Be even more precise!
- This is all usually in the supplementary material, so read it!

Mutual information

- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Lecture 3/19/24

16S Sequencing! Paper 1: What is 16S Gene?

- the blue regions are conserved (its apart of the RNA) and adds to the function of the ribosome itself
- You can't circularize the DNA, so you have to cut up parts of it and sequence it.

What and Why *E. coli*? And What is the purpose of the first paper?

- *E. coli* is a very well understood bacteria, so it is a good way to look at how well the method works.
- Purpose: Find surreptitious OTUs that add error/ inflated biodiversity

Errors in pyrosequencing 101:

- In long homopolymer(big problem) steteches, the peaks are hard to distinguish. If we have a lot of Cs, one error is magnified. It's easier to detect which color flashes versus how much it flashes
- Why do we care? We thought that the Pyrosequencing told us there is a LOT of biodiversity, but it was just fake news. We can't understand signal if we don't understand noise.
- We must understand the sequencing method and its errors.
- There is also PCR errors!
- We have errors in the methods i.e. contamination
- We have to cut out reads that are low quality
- The machine anaylsis also has errors. . . errors arise at different levels.

What is a quality score?

- Why is the theoretical number not 1? The DNA of *E. coli* have several copies, but they don't have to be exactly identical.
- Percentage Identity threshold: How identical are the sequences? 97 is the standard and the OTU number is around 1 because the errors make the 5 theoretical OTUs indiscernible.
- What is a quality score? How sure is the machine at the sequencing? The machine doesn't know about the databases and the biology, but it can accurately measure the amount of noise i.e. the peaks we measure are.
- The machine is doing Bayesian inference! The machine may have a peak that is not 100% sure what base it is, so the machine makes a guess while accounting for the noisy peak.

Truth vs Sequncer:

- A cartoon sequence space would have 1 point, but the sequencer has a cloud of points due to the pcr & pyrosequencing errors.
- In the E. coli we have 5 theoretical sequences, but the sequencer gave us 643 sequences!

Clustering:

- Even though there are 5 theoretical OTUs a clustering threshold may give us 2 OTUs since some of the sequences are more similar than the other. The radius of the clustering roughly the error rate of the sequencer. Lax identity theshold will lose this information.
- The sequencer has 300(or # of bp) dimensions with error in each dimension

Reads:

- We are looking at why and where the erros are, so we use unfiltered set?

- The binomial distribution is used to estimate errors because we expect them to be independent.

Chimera:

- We have to manually seperate chimeras
- chimeras come from the PCR step

Summary:

- Don't trust signal without understanding noise

Example: Say we have 300 bp sequence, and we have a 2% per base error for 10,000 copies of the same sequence. The error free sequence chance would be

$$(0.98)^{300}$$

and for 10,000 copies we would have

$$10,000(0.98)^{300} \approx 20$$

error free sequences within the 10,000 copies. So the average is equal to far from typical.

- If we understand errors well, we should be able to back out the truth
- Where is this happening in the 2nd paper? In the image context we can infer 2 points from a noisy image, but in a high dimensional space, its much harder...
- The 2nd paper responds to papers that think that the pyrosequencing led to inflated biodiversity.

Lecture 3/21/24

Paper 2: Looking at Figure 1

- How do we get more than the expected OTUs? Noise makes more OTUs.
- In the Template Samples, we have a definite understanding of the OTUs in a sample, but we have a more expected OTUs theoretically and experimentally in each clustering & linkage method.

Problems in clustering:

- What is table 2 asessing? Why MS vs PW? How do we define the distance between points (before clustering)? This is difficult because not only is there substitutions errors, but there are insertions and deletions i.e. one insertsion will make everything after it different

$$ATGCGCGC$$
$$AGCGCGC$$

Takehomes:

- Errors in both data collection and analysis is a big problem especially in each method.
- Using a baseline or something that we understand very well can be good to study a new method.
- Establish noise before analyzing data/signal

Bigger Picture from Imaging to Sequencing: My thoughts

- Quanitifying Fluorescence as a function of mRNA count vs. Quantifying OTUs as a function of sequence biodiversity.
- Data collection should be well understood

Our thoughts: Conceptual

- Choices of data analysis require understanding of exp. details i.e. the device, preprocessing of data, etc.
- Data is never what you think it is i.e. OTU is not as well defined as we think it is.
- Can't understand signal without first understanding noise (spend the first months measuring zero)
- Cool technology, but fully realizing its potential requires careful approach to analysis

Analogy:

- Imagaing

From Figure 4.3, in the sequeincing we have raw truth and the pcr amp adds pcr error, then the sequencerg adds sequencing error and the experiment leads to more OTUs than expected. Questions:

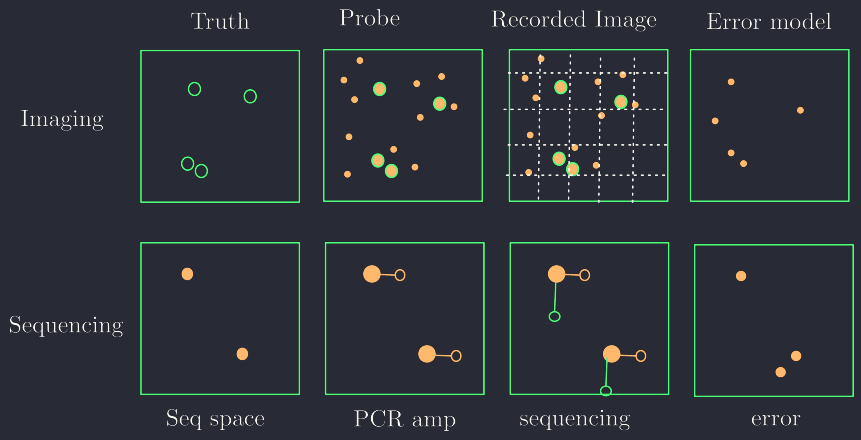


Figure 4.3: The Analogy of Imaging and Sequencing

- Error model? The size of the ball with errors in the sequencer step roughly relates to the amount of error/error rate. We can't just use the 3% biological difference as a threshold.
- Which two steps correspond to the clustering? A single point is not an OTU so we have to cluster similar points.

Lecture 3/26/24

Presentation Part 3: Question and more questions, and some answers

- Why use RNA? It expressive functional genes so we can amplify this unique part of the genome rather than the DNA itself.
- All our cells are same DNA, but different tissues will express different genes via RNA which is what we are measuring.
- RNA circularizes and is reverse transcibed, so DNA can be tagged differently and thus you can sequence them in parallel(machine will two different RNA sequences).
- UMAP: Take high dimensional data and project on to 2D space. It's an algorithm. And Default parameters are bad
- Expression not equal function:

Pros and Cons

- No one reads the documentation, so all the tools can be used with default values etc.

Lecture 4/2/24

Presentation: Neuron map of the retina

- Hidden Truth in retina to electrod arrays: Goal is to identify presence of neurons; when and where they spike.
- which of the sea of soup of stuff is truly the neuron firing. The spiking of neuron spike is representative of the signal we see in F. But in the image we have more than just one electrode firing. We see that there is a cloud of point spread functions.
- What is on F is a point spread function of neuron F. Another neuron may fire differently as shown in G.
- First extract the error of the machine, which is a challenge.

Lecture 4/4/24

- Why point-wise median and not mean? The mean is sensitive to outliers, so the median is more robust to outliers. If you have a lot of noise in the array with mean zero, the mean will not account for the peak at neuron 47. If we have neuron 47 fire and a millisecond later neuron 57 fires, the mean will take the middle value, and the median will stick to the mean.
- a_ij has a combination of the neurons that fire together.
- Fitting problem for amplitude, we can have an infinite number of points in time
- Refractory period violation: either the electrode is not dense enough to catch the neuron firing or the neuron is not firing (multiple neuron is caught by the same electrode)
- Artifical Template: Compare with a signal with fake spikes to see if it distinguished the artificial waveforms. We don't have a ground truth yet (where we know which exact neuron fires at what time).
- To templates that are spatially close but slightly different. The correlation vs. time interval.
- Two neurons close together may have a similar waveform. But a slight difference in the electrode body sometimes the algorithm can't distinguish, but it only happens 0.2% of the time (weird trough at 0ms).
- Time derivative of the template lines up with PCA 1 with lines up with the discrete smapling.
- Completeness: The distribution of the template is very similar to the distribution of the epifluorescence imaging. A very heterogenous system will give a better comparison: a homogenous system will not give much information.

Lecture 4/9/24

Data (High Dimensional) → Chosen features (simplified cluster)

- Equivalent to choice of similarity metric in data space (some similarity metrics are utterly useless, e.g. euclidean distance does not say anything about retinal map of movie)
- Non-trivial, non-unique, question specific → "Data = signal + noise", but also we must figure out what counts as signal and noise

Most informative representation appropriate for your question
Precise timing of spikes: Does it matter?

- No: "rate coding" (how active and frequent something spikes can't distinguished)
- Yes: "time coding"

Data anaylsis choies ↔ Scientific choices.

Every Operation that discards data ↔ Implicitly, multiple assumptions about your data

- about "what matters"
- about nature of data

Have you checked those assumptions? before discarding info, even if sure, always look at what you discarded.

- Looks: as you expected: good data
- Not as expected: even better! We know that this data we were about to discard changes things.

HW9 160 neurons, 297 iterations of same movie. operations:

- For each neuron compute fraction of of times (e.g. 270 0f 297) is was active in a given time bin. This discards information where we Implicitly assume that the time bin is exactly lined up with the movie frame.

Lecture 4/11/24

Floating Point Two concrete ways in which you COULD say it is connected to the subject of the class:

- The error in floating point is a form of "noise" in the data.
- The computer automatically "preprocesses" the data in a way that is not always what you want.

160 Cats Repeat a game w/owner 297 times. Record if cat moved a paw.

- Different cat, different game(stimuli)
- If given a different stimuli, we don't know the behavior of all the cats.
- HW 9: Different neurons, different stimuli.
- ! Dot product of spike trains is meaningless
- Main source of info: repeatability across 297 trials
- Neuron 72: As we go further in each iteration 1 to 257 we see that the retina has "learned": To extract the effect: Ideas
  - Average over horizontal window and loook at d/dy. Global Activity
  - First 50 vs last 50 replicates: Compare # of spikes is region of interest
  -

Lecture 4/16/24

Designing a Metric Desired Criteria: When

- Range from 0 to 1
- Test examples: should score high or low
- Look at ambiguous examples yourself

Its not guaranteed the metric will satisfy all things.

Extent of non-independence between X and Y: Mutual Information!

$$I(X, Y) = I(f(x), g(x))$$

for any deterministic $f(x)$ adn $g(x)$.

Firing Streaks? How do you find them?

- Look at the average length of the streaks?
- The number of streaks in a given time window(frequency)?

But first we need to look at the data more, i.e., the individual Neuron data shows some streaks... and the original data with averages is not actually a streak, but bands of activity that are close together.

! "Metric is supposed to telle me if phenomenon is real"... NO! The metric is to convince a reasonable reviewer that your effect is real.

- The metric is always less biased than you are, so you must know (or be reasonably convinced) that the observation is a real effect.

Bayesian Optimal are we?

Lecture 4/18/24

Takehomes: What's interesting? Climate change denial and its effect on the scientific community, Lewandowsky et al.

- Maybe even despite the uncertainty of scientific discussion, we should warrant

Scientists are supposed

- Science is not occuring outised societal context
- Ask good questions includes recognizing framing
- Rather than being universally skeptical

Class response:

- Pressure to succeed & overvalued dissenting opinions
- How do we assess attention given to a scientific topic/issue?
- Even when we are working in good faith (data is real and worked on), we can still be wrong in the analysis.

Paper 1: Possible Artifacts of data biases in the recent global surface warming hiatus, Karl et al.

- Define an estimae of global temperature
- Determine the trend of global temp over the past 100 years

Challenges:

- Instrument observation at a specifc location changes over time: They must correct data collected: Differneces in technology over the decades e.g. buckets vs engine intake
  - Number of data points per time
  - Change in coverage: much denser now vs 1940s
- The taks itself is challenging/ ill-defined. What is "global temp"

Approaching Data Noise before signal and Opening black boxes: How do we ask good questions? Details vs Big Picture: $\rightarrow$ which detials matter?

- Presentations: Communicating your science
  - Define the usecases and limitations: Try to draw a line at the amount of truth your presentation/ research has revealed, e.g. this explains that but not another thing.
  - Understand the background knowledge of the scientific audience.
  - Features of a bad presentation:
  - Anticipating questions and answering them when they arise and not much later or before.
    - too detailed $\leftrightarrow$ not detailed enough

Homework 1

Due 1/30 12pm

---

1. (a) For case 1.1 the marginal probabilty is

$$P(x=0)=0.2, \quad P(x=1)=0.8$$
$$P(y=0)=0.6, \quad P(y=1)=0.4$$

For 1.2

$$P(x=0)=0.4, \quad P(x=1)=0.6$$
$$P(y=0)=0.6, \quad P(y=1)=0.4$$

(b) 1.1

$$P(x=0|y=0)=1/5, \quad P(x=1|y=0)=4/5$$
$$P(x=0|y=1)=1/5, \quad P(x=1|y=1)=4/5$$

1.2

$$P(x=0|y=0)=1/3, \quad P(x=1|y=0)=2/3$$
$$P(x=0|y=1)=3/7, \quad P(x=1|y=1)=4/7$$

(c) Variables x and y are independent iff $P(x,y)=P(x)P(y)$. For 1.1

$$P(x=0,y=0)=0.12 \quad \text{and} \quad P(x=0)P(y=0)=0.2(0.6)=0.12$$
$$P(x=0,y=1)=P(x=0)P(y=1)=0.2(0.4)=0.08\ldots$$
$$P(x,y)=P(x)P(y)$$

So x and y are independent for 1.1. You can also see that the condition of y does not change the marginal probability of x. For 1.2 there is a simple counterexample

$$P(x=0,y=0)=0.1 \quad \text{and} \quad P(x=0)P(y=0)=0.4(0.3)=0.12$$
$$P(x=0,y=0)\neq P(x)P(y)$$

So x and y are not independent (dependent) for 1.2. You can also see that the conditional probability is not the same as the marginal probability for both cases.

2. For two random variables x and y to be independent, it must be true that

$$P(x,y)=P(x)P(y) \tag{1}$$

and from the definition of conditional probability

$$P(x|y=y_o)=\frac{P(x,y=y_o)}{P(y=y_o)}$$

substituting (1) into the joint probability

$$P(x|y=y_o)=\frac{P(x)P(y=y_o)}{P(y=y_o)}$$
$$P(x|y=y_o)=P(x)$$

3. (a) Since the two thrown dice are independent, the fair dice has 36 possible outcomes $A_{xy} = \{(1,1),(1,2),\ldots,(6,6)\}$ with equal probability

$$P(x,y) = P(x)P(y) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$$

The probability distribution of the sum of the two dice $P(S)$ is

$$P(S) = \begin{cases} 1/36 & S = 2,12 \\ 2/36 & S = 3,11 \\ 3/36 & S = 4,10 \\ 4/36 & S = 5,9 \\ 5/36 & S = 6,8 \\ 6/36 & S = 7 \end{cases}$$

where $S = x + y$. For the absolute difference of the two dice $D = |x - y|$

$$P(D) = \begin{cases} 2/36 & D = 5 \\ 4/36 & D = 4 \\ 6/36 & D = 3 \\ 8/36 & D = 2 \\ 10/36 & D = 1 \\ 6/36 & D = 0 \end{cases}$$

for the difference $D = 0$ there are 6 possible outcomes $(1,1),(2,2),\ldots,(6,6)$, for $D = 1$ there are 10 possible outcomes $(1,2),(2,1),(2,3),(3,2),\ldots,(5,6),(6,5)$, and so on.
(b) For 100 dice, the probability distribution of the sum of the dice $P(S)$ would be roughly

$$P(S) = \begin{cases} 1/6^{100} & S = 100,600 \\ 100/6^{100} & S = 101,599 \\ 5050/6^{100} & S = 102,598 \\ \vdots & \vdots \\ 1.52\times 10^{76}/6^{100} & S = 350 \end{cases}$$

first we find the mean of 1 independent dice roll (μ_1)':

$$\mu_1 = \sum_x P(x)x = \frac{1}{6}\sum_{x=1}^{6} x = \frac{21}{6} = 3.5$$

because the mean of N independent dice rolls is the sum of the means of each dice roll

$$\mu_N = \sum_{i=1}^{N} \mu_i$$

thus the mean of 100 dice rolls is

$$\mu_{100} = 100\cdot\mu_1 = \boxed{350}$$

To find the Standard Deviation we first find the variance of 1 independent dice roll (σ_1^2):

$$\begin{aligned} \operatorname{Var}[x] = \operatorname{E}[(x - \operatorname{E}[x])^2] &= \operatorname{E}[x^2 - 2x\operatorname{E}[x] + \operatorname{E}[x]^2] \\ &= \operatorname{E}[x^2] - 2\operatorname{E}[x]^2 + \operatorname{E}[x]^2\operatorname{E}[1] = \operatorname{E}[x^2] - \operatorname{E}[x]^2 \end{aligned}$$

or in summation notation

$$\sigma_1^2 = \sum_x P(x)(x-\mu_1)^2 = \frac{1}{6}\sum_{x=1}^{6}(x-3.5)^2 = \frac{17.5}{6} = 2.9167$$

for 2 independent variables x and y

$$\begin{aligned}\mathrm{Var}[x+y] &= \mathrm{E}[((x+y)-\mathrm{E}[x+y])^2] \\ &= \mathrm{E}[(x-\mathrm{E}[x])+(y-\mathrm{E}[y])^2] \\ &= \mathrm{E}[(x-\mathrm{E}[x])^2+(y-\mathrm{E}[y])^2+2(x-\mathrm{E}[x])(y-\mathrm{E}[y])] \\ &= \mathrm{E}[(x-\mathrm{E}[x])^2]+\mathrm{E}[(y-\mathrm{E}[y])^2]+2\,\mathrm{E}[(x-\mathrm{E}[x])(y-\mathrm{E}[y])] \\ &= \mathrm{Var}[x]+\mathrm{Var}[y]+2\,\mathrm{E}[(x-\mathrm{E}[x])(y-\mathrm{E}[y])]\end{aligned}$$

where the third term is

$$\begin{aligned}\mathrm{E}[(x-\mathrm{E}[x])(y-\mathrm{E}[y])] &= \mathrm{E}[xy-x\,\mathrm{E}[y]-y\,\mathrm{E}[x]+\mathrm{E}[x]\,\mathrm{E}[y]] \\ &= \mathrm{E}[xy]-\mathrm{E}[x]\,\mathrm{E}[y]\end{aligned}$$

and for independent variables x and y the third term is zero. Thus the variance of the sum of N independent dice rolls is

$$\mathrm{Var}[N] = N\sigma_1^2 = 100 \cdot 2.9167 = 291.67$$

and the standard deviation is

$$\sigma_N = \sqrt{\mathrm{Var}[N]} = \sqrt{291.67} = \boxed{17.08}$$

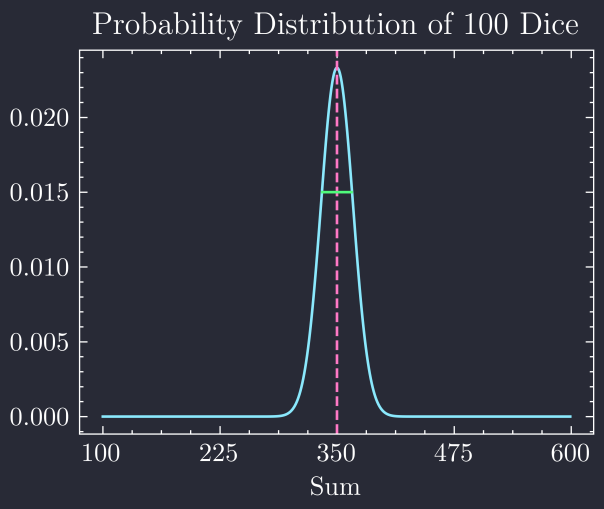


Figure 1.4: Probability distribution of the sum of 100 dice rolls. The mean is 350 and the standard deviation is ≈ 17.

The sketch of probability distribution of the sum of 100 dice rolls is shown in Figure 1.4.

4. (a) Assuming that there is an equal likelihood of the order of the age of the three brothers being any of the 6 possible permutations:

$$\mathrm{Age}_{A,B,F} = \{ABF, AFB, BAF, BFA, FAB, FBA\}$$

where we denote the first element of a permutation as the oldest brother and the last element as the youngest brother.
The probability that Fred (F) is older than Bob (B) is $\boxed{1/2}$ from both the 3 possible permutations or by realizing that there are only two equal outcomes when looking at only the age of Fred vs Bob.
(b) Given that Fred is older than Alex (A), we can eliminate the 3 permutations where Alex is older. Thus the probability that Fred is older than Bob is $\boxed{2/3}$.

5. (a) Given that the probability of choosing a black ball from an urn is $f_B = \frac{B}{K}$, the probability distribution of choosing n_B black balls from N draws is

$$\boxed{P(n_B|N, f_B) = \binom{N}{n_B} f_B^{n_B}(1 - f_B)^{N-n_B}}$$

(b) Finding the mean and standard deviation of n_B is quite similar to Problem 3. Since each draw is independent, the mean of n_B is the sum of the means of each draw! In the case of drawing one ball has a binary outcome

$$n_B(N = 1) = \begin{cases} 1 & \text{black ball with probability } f_b \\ 0 & \text{white ball } (1 - f_B) \end{cases}$$

thus the mean of n_B for one draw is

$$\mu_1 = 1(f_B) + 0(1 - f_B) = f_B$$

and the variance is

$$\sigma_1^2 = (1 - f_B)^2 f_B + (0 - f_B)^2(1 - f_B) = f_B(1 - f_B)$$

for N draws the means add up to

$$\boxed{\mu = N f_B}$$

and the variances add only due to the independent nature of the draws

$$\sigma^2 = N f_B(1 - f_B)$$

thus the standard deviation is

$$\boxed{\sigma = \sqrt{N f_B(1 - f_B)}}$$

For $K = 20$, and $B = K$; $f_B = 5/20 = 0.25$. And for $N = 5$ we have the ratio

$$\frac{\sigma}{\mu} = \frac{\sqrt{5 \cdot 0.25 \cdot 0.75}}{5 \cdot 0.25} = \frac{\sqrt{15}}{5} \approx \boxed{0.77}$$

and for $N = 20$

$$\frac{\sigma}{\mu} = \frac{\sqrt{1000 \cdot 0.25 \cdot 0.75}}{1000 \cdot 0.25} = \frac{\sqrt{30}}{100} \approx \boxed{0.05}$$

6. (a) Dividing the time period T in to M intervals where each interval has a probability $r\,\mathrm{d}t$ or $\mathrm{d}t = T/M$. The probability of no events occuring in time T is

$$\lim_{M\to\infty} (1 - r\,\mathrm{d}t)^M = \lim_{M\to\infty} \left(1 - \frac{rT}{M}\right)^M = \boxed{e^{-rT}}$$

(b) For $n_T = x$ events occuring in M this is similar to the binomial distribution as $M \to \infty$.

$$\lim_{M\to\infty} \frac{M!}{x!(M - x)!} \left(\frac{rT}{M}\right)^x \left(1 - \frac{rT}{M}\right)^{M-x}$$

canceling out some terms...

$$\frac{M!}{(M - x)!} = \frac{M(M - 1)\cdots(M - x + 1)(M - x)!}{(M - x)!} = M(M - 1)\cdots(M - x + 1)$$

and now the facorial has x terms, so we can write it as

$$\begin{aligned}\frac{M(M-1)\cdots(M-x+1)}{M^x} &= \frac{M}{M}\frac{M-1}{M}\cdots\frac{M-x+1}{M} \\ &= 1\cdot\left(1-\frac{1}{M}\right)\cdots\left(1-\frac{x-1}{M}\right)\end{aligned}$$

and as $M\to\infty$ the terms in the product go to 1, so the product goes to 1. Thus we are left with

$$\lim_{M\to\infty}\frac{(rT)^x}{x!}\left(1-\frac{rT}{M}\right)^{M-x} = \lim_{M\to\infty}\frac{(rT)^x}{x!}\left(1-\frac{rT}{M}\right)^{M}\left(1-\frac{rT}{M}\right)^{-x}$$

the second term is the limit of the exponential function

$$\lim_{M\to\infty}\left(1-\frac{rT}{M}\right)^{M} = e^{-rT}$$

and the third term tends to 1 as $M\to\infty$. Thus the probability of x events occuring in time T is

$$\boxed{P(x)=\frac{(rT)^x}{x!}e^{-rT}}$$

where $x=n_T$ for the sake of brevity in notation.
(c) The mean of n_T is

$$\begin{aligned}\mu &= \sum_{x=0}^{\infty} xP(x) \\ &= \sum_{x=0}^{\infty} x\frac{(rT)^x}{x!}e^{-rT} \\ &= e^{-rT}\sum_{x=1}^{\infty} x\frac{(rT)(rT)^{x-1}}{x(x-1)!} \\ &= e^{-rT}(rT)\sum_{x=1}^{\infty}\frac{(rT)^{x-1}}{(x-1)!}\end{aligned}$$

the the first term of the sum is zero which is why the sum starts at $x=1$. The sum is also the Taylor series expansion of e^{rT} if we let $n=x-1$ so

$$\sum_{x=1}\frac{(rT)^{x-1}}{(x-1)!} = \sum_{n=0}\frac{(rT)^n}{n!} = e^{rT}$$

Therefore the mean of n_T is

$$\mu = (rT)e^{-rT}e^{rT} = rT$$

The variance of n_T is

$$\sigma^2 = \mathrm{E}[x^2]-\mathrm{E}[x]^2$$

the first term is solved similarly to the mean

$$\begin{aligned}
\mathrm{E}[x^2] &= \sum_{x=0}^{\infty} x^2 P(x) \\
&= \sum_{x=0}^{\infty} x^2 \frac{(rT)^x}{x!} e^{-rT} \\
&= e^{-rT} \sum_{x=1}^{\infty} x^2 \frac{(rT)(rT)^{x-1}}{x(x-1)!} \\
&= (rT) e^{-rT} \sum_{x=1}^{\infty} x \frac{(rT)^{x-1}}{(x-1)!} \\
&= (rT) e^{-rT} \left[\sum_{x=1}^{\infty} (x-1) \frac{(rT)^{x-1}}{(x-1)!} + \sum_{x=1}^{\infty} \frac{(rT)^{x-1}}{(x-1)!} \right] \qquad x = [(x-1)+1] \\
&= (rT) e^{-rT} \left[(rT) \sum_{x=2}^{\infty} \frac{(rT)^{x-2}}{(x-2)!} + \sum_{n=0}^{\infty} \frac{(rT)^n}{n!} \right] \qquad n = x-1 \\
&= (rT) e^{-rT} \left[(rT) \sum_{l=0}^{\infty} \frac{(rT)^l}{l!} + \sum_{n=0}^{\infty} \frac{(rT)^n}{n!} \right] \qquad l = x-2 \\
&= (rT) e^{-rT} \left[(rT) e^{rT} + e^{rT} \right] \\
&= (rT)^2 e^{-rT} e^{rT} + (rT) e^{-rT} e^{rT} \\
&= (rT)^2 + rT
\end{aligned}$$

and the variance is

$$\sigma^2 = (rT)^2 + rT - (rT)^2 = rT$$

Therefore the mean and standard deviation of n_T are

$$\boxed{\mu = rT \quad \text{and} \quad \sigma = \sqrt{rT}}$$

7. Using Bayes' theorem for the outcome $X = \{7, 3, 4, 2, 5, 3\}$ is

$$P(A|7,3,4,2,5,3) = P(A|X) = \frac{P(X|A)P(A)}{P(X)}$$

where the probability of choosing dice A is 1 in 3—$P(A) = 1/3$. The conditional probabilty $P(X|A)$ is the probability of rolling the outcome X given that dice A is chosen:

$$P(X|A) = \frac{1 \times 4 \times 2 \times 4 \times 2 \times 4}{20^6} = \frac{256}{20^6}$$

and the probability of rolling the outcome X is given by the sum rule

$$P(X) = P(X|A)P(A) + P(X|B)P(B) + P(X|C)P(C)$$

where $P(B) = P(C) = 1/3$ and the conditional probabilities $P(X|B)$ and $P(X|C)$ are

$$P(X|B) = \frac{2 \times 3 \times 2 \times 2 \times 2 \times 3}{20^6} = \frac{144}{20^6}$$

$$P(X|C) = \frac{2^6}{20^6} = \frac{64}{20^6}$$

thus the probability of choosing dice A given the outcome X is

$$P(A|X) = \frac{\frac{256}{20^6} \cdot \frac{1}{3}}{\frac{256}{20^6} \cdot \frac{1}{3} + \frac{144}{20^6} \cdot \frac{1}{3} + \frac{64}{20^6} \cdot \frac{1}{3}} = \frac{256}{464} \approx \boxed{0.55}$$

with the knowledge that terms cancel out, the probability of the die being B is

$$P(B|X) = \frac{144}{464} \approx \boxed{0.31}$$

and the probability of the die being C is

$$P(C|X) = \frac{64}{464} \approx \boxed{0.14}$$

8. (a) Given that the bus arrives on average every 5 minutes, the average wait time is 5 minutes. And the bus that just left Sally would have left an average of 5 minutes ago. From the code, taking the mean value of the wait times is also ≈ 5 minutes.
(b) Therefore the average time between two buses is the sum in the time Sally is waiting for the bus and how long the missed bus has been gone: 10 minutes.
(c) The paradox is that 'we' think that after waiting for 5 minutes the bus will arrive, but the average time between buses is 10 minutes, so we are waiting longer than we expect to intuitively. This is because the conditional probability of Sally getting to the bus stop where the interval between buses is less than 5 minutes given that she has waited for a time t is less as time goes on. And the probability that Sally arrived at the bus stop where the interval between buses is more than 5 minutes given that she has waited for a time t is more as time goes on.
(d)

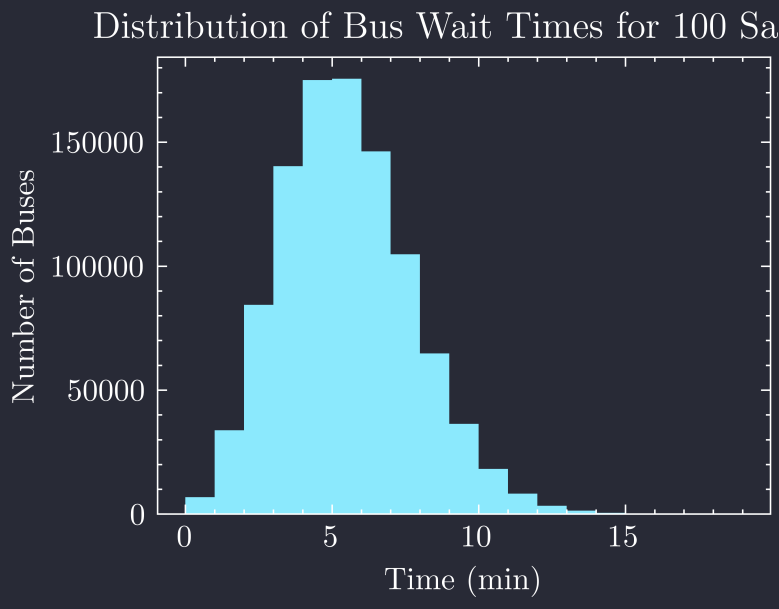


Figure 1.5: Mean of 5.00 min and Time between buses of 10.00 min.

PYTHON CODE BELOW

hw1_python

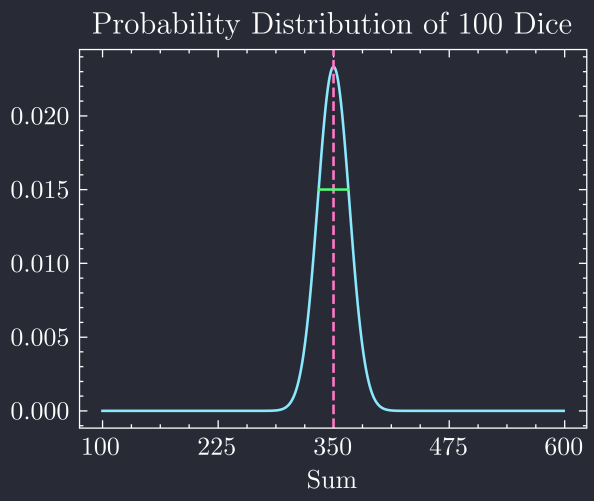
January 30, 2024

[ ]:
```
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt
import matplotlib as mpl
import scienceplots

# Science plot package + Dracula theme
plt.style.use(['science', 'dark_background'])
plt.rcParams['axes.facecolor'] = '#282a36'
plt.rcParams['figure.facecolor'] = '#282a36'
colorcycle = ['#8be9fd', '#ff79c6', '#50fa7b', '#bd93f9', '#ffb86c', '#ff5555',␣
 ↪'#f1fa8c',
'#6272a4']
plt.rcParams['axes.prop_cycle'] = mpl.cycler(color=colorcycle)
white = '#f8f8f2' # foreground

# change dpi
plt.rcParams['figure.dpi'] = 1024
```

[ ]:
```
# highly optimized function to count the number of ways to get a sum of s with␣
 ↪d dice
def count_ways_to_sum(sum_target, num_dice):
    # Initialize a 2D array to store results of subproblems
    dp = [[0] * (sum_target + 1) for _ in range(num_dice + 1)]

    # Base case: there is one way to get a sum of 0 (no dice)
    dp[0][0] = 1

    # Fill the dp table using the convolution formula
    for d in range(1, num_dice + 1):
        for s in range(1, sum_target + 1):
            for k in range(1, 7):
                if s - k >= 0:
                    dp[d][s] += dp[d-1][s-k]

    return dp[num_dice][sum_target]
```

```python
# Example usage:
sum_100_ways = count_ways_to_sum(100, 100)
sum_101_ways = count_ways_to_sum(101, 100)
sum_102_ways = count_ways_to_sum(102, 100)
sum_350_ways = count_ways_to_sum(350, 100)

print(f"Number of ways to get a sum of 100 with 100 dice: {sum_100_ways}")
print(f"Number of ways to get a sum of 101 with 100 dice: {sum_101_ways}")
print(f"Number of ways to get a sum of 102 with 100 dice: {sum_102_ways}")
print(f"Number of ways to get a sum of 350 with 100 dice: {sum_350_ways:.2e}")

# for loop to get a function of sum to combinations
sums = np.arange(100, 600, 1)
ways = []
for i in sums:
    ways.append(count_ways_to_sum(i, 100) / 6**100)

# plotting
plt.figure()
plt.plot(sums, ways)
plt.xlabel('Sum')
plt.title('Probability Distribution of 100 Dice')

# Add x-axis label at 350 as "mean"
plt.axvline(x=350, color=colorcycle[1], linestyle='--')

# Add tick label at 500/4 intervals
plt.xticks(np.arange(100, 601, 125))

# Add standard deviation at x= 350 +/- 17 as horizontal line
plt.hlines(y=0.015, xmin=350-17, xmax=350+17, color=colorcycle[2])

plt.show()

print (f"Probability of getting a sum of 350 with 100 dice: {sum_350_ways /␣
↪6**100:.2e}")
# sum up ways
print (f"Probability distribution adds to 1: {sum(ways):.2e}")
```

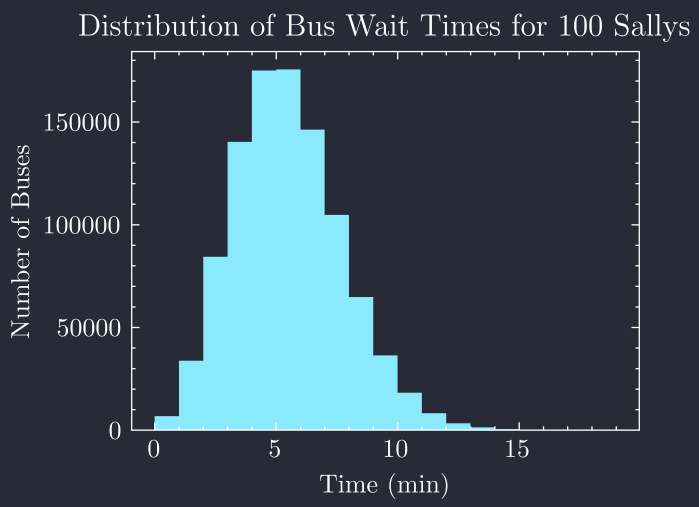
```
Number of ways to get a sum of 100 with 100 dice: 1
Number of ways to get a sum of 101 with 100 dice: 100
Number of ways to get a sum of 102 with 100 dice: 5050
Number of ways to get a sum of 350 with 100 dice: 1.52e+76
```




```
Probability of getting a sum of 350 with 100 dice: 2.33e-02
Probability distribution adds to 1: 1.00e+00
```

[ ]:
```python
# 481 Problem 8d
# Simulation of 10000 buses and 100 Sally's

# constants
N = 10000 # number of buses
t_avg = 5 # average time between buses

# using poisson distribution old way
t_poisson = np.random.poisson(t_avg, N)

# simulating 100 Sally's
for i in range(99):
    t_poisson = np.concatenate((t_poisson, np.random.poisson(t_avg, N)))

# checking if it is 100 Sally's
print(len(t_poisson) / 10000)

# plotting
plt.figure()
plt.hist(t_poisson, bins=np.max(t_poisson))
plt.xlabel('Time (min)')
plt.ylabel('Number of Buses')
plt.title('Distribution of Bus Wait Times for 100 Sallys')
plt.show()
```

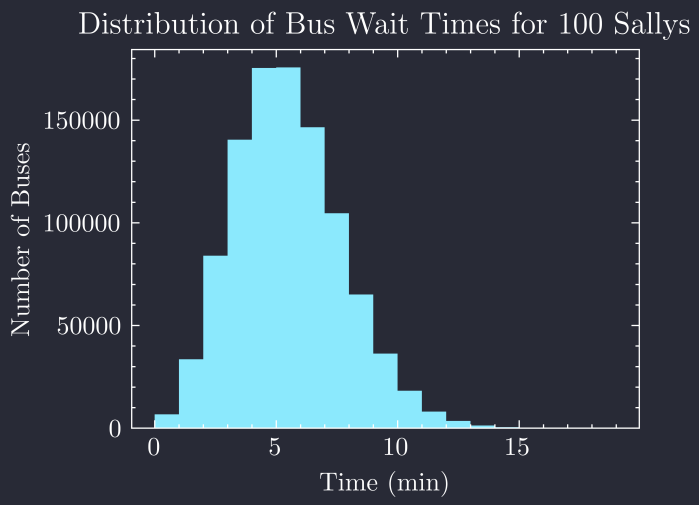
```
100.0
```



[ ]:
```
# implementing hard code
# random number generator
rand = np.random.default_rng(seed=42)

# Probability of 1 bus given 5 min avg
p = np.exp(-5)

# wait time for 1 bus
def wait_time():
    time = 0
    prod = 1.0
    while True:
        U = rand.random()
        prod *= U
        if prod > p:
            time += 1
        else:
            return time

def wait_time_100():
    times = []
    for i in range(100):
        times.append(wait_time())
    return times

# simulating 10000 buses
times = []
for i in range(10000):
    times += wait_time_100()
```

```
print(np.mean(times))

# plotting
plt.figure()
plt.hist(times, bins=np.max(times))
plt.xlabel('Time (min)')
plt.ylabel('Number of Buses')
plt.title('Distribution of Bus Wait Times for 100 Sallys')
plt.show()
```

```
5.001477
```



[ ]:
```
# time of bus that just passed
last_bus_times = []
for i in range(1, len(times)):
    if times[i] == 0:
        last_bus_times.append(times[i] - times[i - 1])

print(np.mean(last_bus_times))

# time in between buses
between_bus_times = []
for i in range(1, len(times)):
    between_bus_times.append(times[i] + times[i-1])

print(np.mean(between_bus_times))
```

```
-5.012881255552266
10.002945002945003
```

Homework 2

Due 2/6 12pm

1.

$$P(r|\lambda) = \exp(-\lambda)\frac{\lambda^r}{r!}$$

(a) Taking the log of the likelihood function:

$$L(\lambda) = \ln P(r|\lambda) = -\lambda + r\ln\lambda - \ln r!$$

finding the maximum by taking the derivative with respect to λ and setting it to zero:

$$\frac{dL}{d\lambda} = -1 + \frac{r}{\lambda} = 0 \implies \hat{\lambda} = r$$

so the maximum likelihood estimate for λ is $\hat{\lambda} = r$.
(b) Given the derivative with respect to the function $\ln\lambda$:

$$\frac{\mathrm{d}}{\mathrm{d}(\ln\lambda)}u^n = nu^n, \qquad \frac{\mathrm{d}}{\mathrm{d}(\ln\lambda)}\ln\lambda = 1$$

we can find the curvature of the log likelihood function:

$$\frac{\mathrm{d}}{\mathrm{d}(\ln\lambda)}L(\lambda) = -\lambda + r = 0 \implies \hat{\lambda} = r$$
$$\frac{\mathrm{d}^2}{\mathrm{d}(\ln\lambda)^2}L(\lambda) = -\lambda = k$$

For a normal distribution with width σ, the curvature is $k = -1/\sigma^2$. So the width is approximately

$$\sigma \propto \frac{1}{\sqrt{-k}} = \frac{1}{\sqrt{\lambda}}$$

and the 95% confidence interval at the MLE is approximately

$$\hat{\lambda} \pm 2\sigma = r \pm \frac{2}{\sqrt{\hat{\lambda}}}$$

(c) Given the new Poisson distribution

$$P(r|\lambda) = \exp(-(\lambda + b))\frac{(\lambda + b)^r}{r!}$$

the log likelihood function is

$$L(\lambda) = -(\lambda + b) + r\ln(\lambda + b) - \ln r!$$

and the maximum likelihood estimate for λ is

$$\frac{dL}{d\lambda} = -1 + \frac{r}{\lambda + b} = 0 \implies \hat{\lambda} = r - b$$

the value $\hat{\lambda} = 9 - 13 = -4$ is not physically meaningful, so the MLE will be the lowest possible value for λ which is $\hat{\lambda} = 0$. From this we can infer that the remote star is very dim. The Bayesian posterior distribution for λ is

$$P(\lambda|r) = \frac{P(r|\lambda)P(\lambda)}{P(r)}$$

and sketched in the figure below

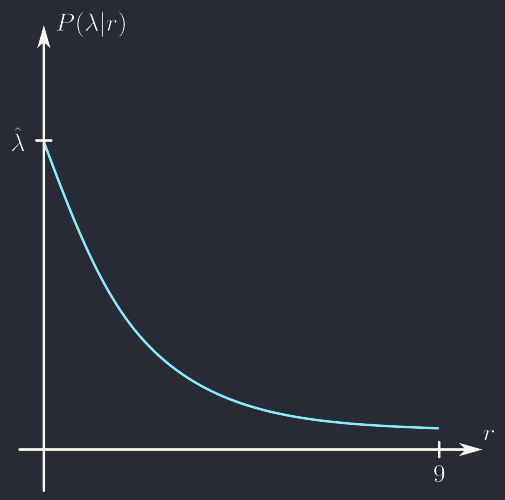


Figure 1.6: The posterior distribution for λ given r

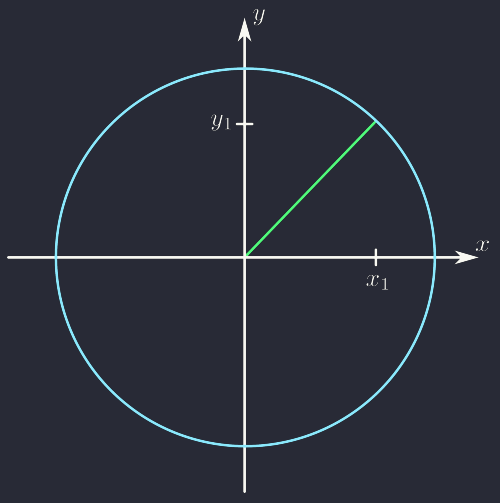


Figure 1.7: Segment of Gaussian distribution at $\{x_1, y_1\}$

2. (a) From the geometric picture as shown in Figure 1.7, the segment of the Gaussian is a circle with radius $\rho = \sqrt{x_1^2 + y_1^2}$ and the circumference is $2\pi\rho$ which directly relates to the extra factor of ρ and canceling the 2π in the denominator. This is also related to the Jacobian when transforming from Cartesian to polar coordinates when computing the integral. Using the integral of a 1D Gaussian:

$$\int_{-\infty}^{\infty} \exp(-ax^2)\,\mathrm{d}x = \sqrt{\frac{\pi}{a}}$$

Verifying that the integral is normalized:

$$\frac{1}{2\pi\sigma^2} \iint_{-\infty}^{\infty} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \mathrm{d}x\,\mathrm{d}y = \frac{1}{\sigma^2} \int_0^{\infty} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \rho\,\mathrm{d}\rho$$

$$\frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y = \frac{1}{\sigma^2} \int_0^{\infty} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \rho\,\mathrm{d}\rho$$

$$\frac{1}{2\pi\sigma^2} \sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma^2} = \frac{1}{\sigma^2} \int_0^{\infty} \sigma^2 \exp(-u)\,\mathrm{d}u$$

$$\frac{2\pi\sigma^2}{2\pi\sigma^2} = \left[-e^{-u}\right]\Big|_0^{\infty} = 1$$

so both integrals are normalized.

(b)

$$P(\rho) = \frac{\rho}{\sigma_w^k} \exp\left(-\frac{\rho^2}{2\sigma_w^2}\right)$$

The sketch for the distribution of $P(\rho)$ is shown in Figure 1.8.

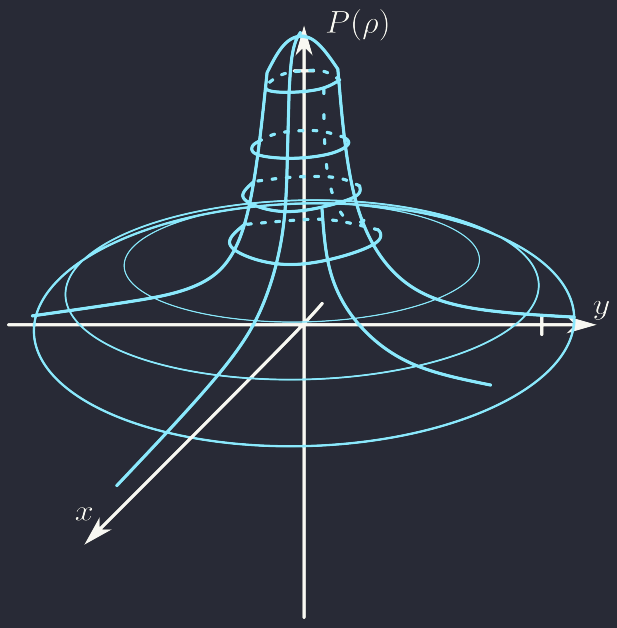


Figure 1.8: The distribution of ρ for $k = 1000$

(c) The standard deviation is

$$\sigma_w = \sqrt{\frac{\sum(w-\mu)^2}{k}}$$

and because the distribution is centered at the origin, the mean is $\mu = 0$, so

$$\sigma_w = \sqrt{\frac{\rho^2}{k}}$$

Since most of the probability mass lies around ρ or equivalently a radius of the thin shell where $\rho = r = \sigma_w\sqrt{k}$, and the thickness of the shell is equivalent to the standard deviation of the gaussian $= r/\sqrt{k}$.
(d) Taking the ratio of the probability density from the origin to a point $\rho = \omega_w\sqrt{k}$ away:

$$\frac{P(0)}{P(\sigma_w\sqrt{k})} = \exp\left\{\frac{\sigma_w^2 k}{2\sigma_w^2}\right\} = \exp\left\{\frac{k}{2}\right\}$$

(e) For a shell to contain 95% of the probability mass, $\sigma_w = 2$ and thus the radius and thickness of the shell are

$$r = \sigma_w\sqrt{k} = 2\sqrt{1000} = 63.25, \qquad \frac{r}{\sqrt{1000}} = 2$$

and the probability density is $\exp\{500\} \approx 10^{217}$ larger at the origin than at the edge of the shell.
(f) For a 1% difference in σ_w at the origin, the radius term is at zero so the exponents are 1, so the ratio of the probability densities is

$$\frac{(1.01\sigma_w)^k}{\sigma_w^k} = 1.01^k \approx 20959.$$

(g) Because of the large amount of parameters, the MLE would have an expected value at where the probability density is the largest, which is at the origin, but as we have seen, the probability mass is almost all at the edge of the thin shell of radius $\sigma_w\sqrt{k}$. The narrow peak at the origin will overwhelm the MLE and thus we don't see the full picture of the distribution.

3. (a) From class

$$\mathcal{R} = \frac{\frac{F_a!F_b!}{(F+1)!}}{1/2^F} = \frac{2^F F_a!F_b!}{(F+1)!}$$

(b) Taking the log of the ratio and using Stirling's approximation as shown in the code we get the plot of the three simulations in Figure 1.9

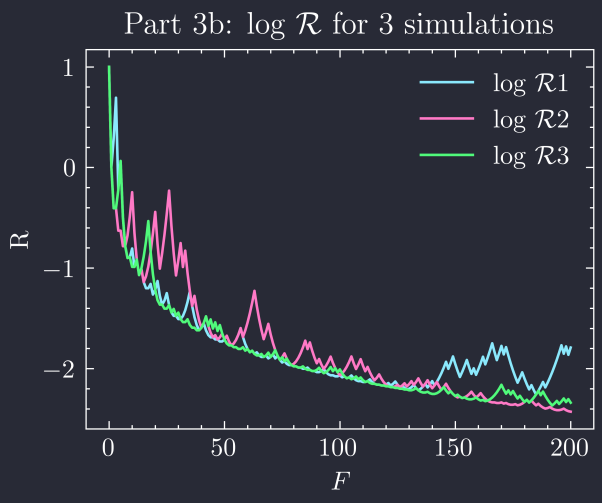


Figure 1.9: The log of the ratio of the likelihoods for the three simulations

(c) The trajectory of the two biased coin models are shown in Figure 1.10 and 1.11.

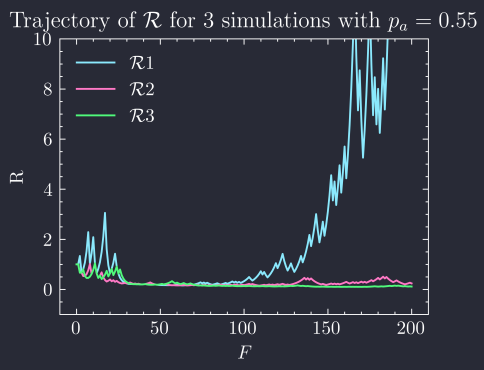


Figure 1.10: There isn't clear evidence for a bias in the first 200 flips

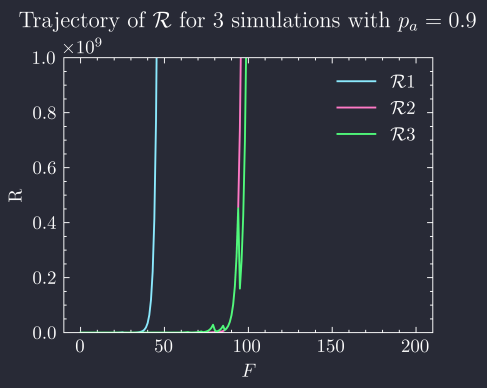


Figure 1.11: A clear bias is shown in the first 200 flips here

and the posterior distribution for the two biased coin models are shown in Figure 1.12 and 1.13.

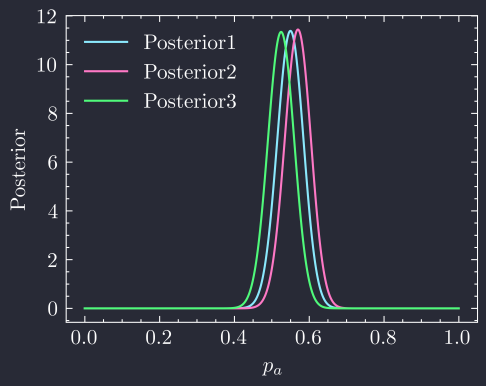


Figure 1.12: We can see that the distribution is centered around roughly $p_a = 0.55$ as expected

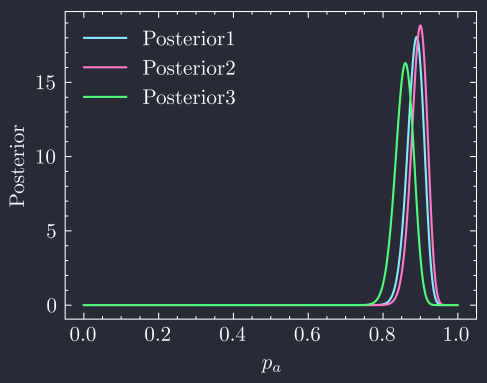


Figure 1.13: There is definitely a bias towards $p_a = 0.9$ in this distribution

(d) From the Gaussian, we can approximate the error bars as $p_a - \frac{1}{2}$, so

$$p_a - 0.5 = \frac{\sigma}{\sqrt{F}} \to F = \frac{\sigma^2}{(p_a - 0.5)^2}$$

so we can get a rough estimate of the number of flips needed to distinguish between the two models for within 2-3 standard deviations as shown in 1.14 and 1.15.

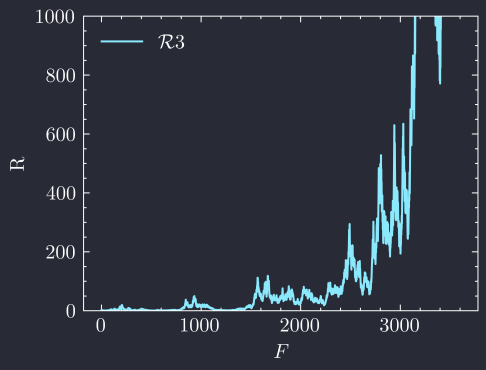


Figure 1.14: For $\sigma = 3$ the model is clearly distinguishable after ≈ 3600 flips

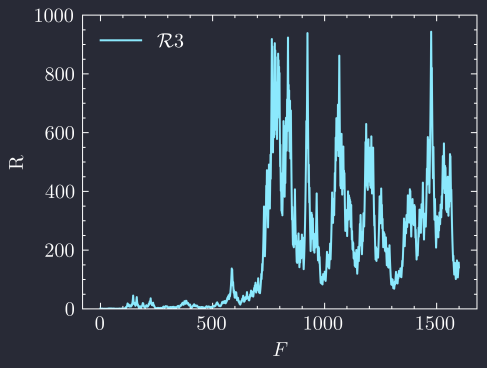


Figure 1.15: For $sigma = 2$ the model does seem to to be biased, but it is not as confident as the previous case

(e) It is suprising to find out that finding evidence against $\mathcal{H}_1$ would be slower than finding evidence for it, but in hindsight it makes sense for cases where the biased probabilities are close to the unbiased coin model since we would need an exponential number of flips as the bent coin model probabilities is closer to $p_o = 1/2$.
From the log of the ratio

$$\begin{aligned}\ln \mathcal{R} &= F \ln 2 + \ln F_a! + \ln F_b! - \ln(F+1)! \\ &= F \ln 2 + \ln(F - F_b)! + \ln(F - F_a)! - \ln(F+1)! \\ &= F \ln 2 + (F - F_b)\ln(F - F_b) - (F - F_b) + \frac{1}{2}\ln 2\pi(F - F_b) \dots\end{aligned}$$

taking the derivative

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}F}\ln \mathcal{R} &= \ln 2 + 1 + \ln(F - F_b) - 1 + \frac{1}{2}\frac{1}{F - F_b} \dots \\ &= \ln 2 + \ln(F_a) + \frac{1}{2F_a} + \ln(F_b) + \frac{1}{2F_b} - \ln(F+1) - \frac{1}{2(F+1)}\end{aligned}$$

The fastest $\log \mathcal{R}$ can grow is when F is very large, and the fractional terms at very large F will be negligible and the log terms grow much slower (e.g. $\ln(10^{20}) \approx 46$), so the derivative will be be approximately a constant which relates to a linear growth. The fastest $\log \mathcal{R}$ can fall is when F is small and where the $-\ln(F+1)$ term will dominate thus the growth will be approximately logarithmic.

PYTHON CODE BELOW

```python
# %%
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt
import matplotlib as mpl
import scienceplots
import gmpy2 as gm

# Science plot package + Dracula theme
plt.style.use(['science', 'dark_background'])
plt.rcParams['axes.facecolor']    = '#282a36'
plt.rcParams['figure.facecolor']  = '#282a36'
colorcycle = ['#8be9fd', '#ff79c6', '#50fa7b', '#bd93f9', '#ffb86c', '#ff5555', '#f1fa8c
    ',
'#6272a4']
plt.rcParams['axes.prop_cycle'] = mpl.cycler(color=colorcycle)
white    = '#f8f8f2' # foreground

# change dpi
plt.rcParams['figure.dpi'] = 1024

# %%
# 3b
# Calculating the ratio R
# Stirling's approximation
def power(n, exp):
    result = 1
    for   in range(exp):
        result = gm.mul(n, result)
    return result

def fact(n):
    if n < 10:
        return np.math.factorial(n)
    a = power(n, n)
    b = gm.exp(-n)
    c = gm.sqrt(2 * np.pi * n)
    return a * b * c

def logratio(F_a,F_b):
    return (F_a + F_b) * gm.log(2) + gm.log(fact(F_a)) + gm.log(fact(F_b)) - gm.log(fact
    (F_a + F_b + 1))

def flip():
    if np.random.rand() < 0.5:
        return 1
    else
        return 0

def simulate(n):
    F_a = 0
    F_b = 0
    R = []
    for   in range n):
        if flip():
            F_a += 1
        else
            F_b += 1
        R.append(logratio(F_a,F_b))
    return R
print(fact(100))
print(2 * gm.log(fact(100)))
print(gm.log(fact(201)))
print(logratio(100, 100))
# numpy array with 1 to 200 on the x-axis
n = np.arange(0, 201)
# calculate the ratio for each n
# R = simulate(200)
```

```
# simulating 3 times and plotting them side by side
plt.figure()
for   in range(3):
    R = simulate(200)
    # Plot on the first subplot
    plt.plot(n, R, label='log $\mathcal{R}$' + str(i+1))
    plt.xlabel('$F$')
    plt.ylabel('R')
plt.title('Part 3b: log $\mathcal{R}$ for 3 simulations')
plt.legend()
plt.show()

# plot the ratio
# plt.plot(n, R, label='R')
# plt.xlabel('$F$')
# plt.ylabel('R')
# plt.legend()
# plt.show()

# %%
# for p_a = 0.55
def bflip():
    if np.random.rand() < 0.55:
        return 1
    else
        return 0

def ybflip():
    if np.random.rand() < 0.9:
        return 1
    else
        return 0

Fa_toss_count_55 = []
Fa_toss_count_90 = []

def ratio(F_a,F_b):
    return fact(F_a) * fact(F_b) / fact(F_a + F_b + 1) * 2 ** (F_a + F_b)

def simulate_b(n):
    F_a = 0
    F_b = 0
    R = [1]
    for   in range(n):
        if bflip():
            F_a += 1
        else
            F_b += 1
        R = np.append(R, ratio(F_a,F_b))
    return R, F_a

print(2 ** 100 * gm.factorial(99) * gm.factorial(100) / gm.factorial(201))

print(2 ** 500 * gm.factorial(275) * gm.factorial(225) / gm.factorial(501))
def simulate_yb(n):
    F_a = 0
    F_b = 0
    R = [1]
    for   in range(n):
        if ybflip():
            F_a += 1
        else
            F_b += 1
        R = np.append(R, ratio(F_a,F_b))
    return R, F_a

# simulating 3 times and plotting them
plt.figure()
for   in range(3):
    R, F_a = simulate_b(200)
```

```
    Fa_toss_count_55.append(F_a)
    plt.plot(n, R, label='$\mathcal{R}$' + str(i+1))
plt.xlabel('$F$')
plt.ylabel('R')
plt.title('Trajectory of $\mathcal{R}$ for 3 simulations with $p_a = 0.55$')
plt.ylim(-1, 10)
plt.legend()

# for p_a = 0.9
plt.figure()
for i in range(3):
    R, F_a = simulate_vb(200)
    Fa_toss_count_90.append(F_a)
    plt.plot(n, R, label='$\mathcal{R}$' + str(i+1))
plt.xlabel('$F$')
plt.ylabel('R')
plt.title('Trajectory of $\mathcal{R}$ for 3 simulations with $p_a = 0.9$')
plt.legend()
plt.ylim(0, 1e9)
plt.show()


# %%
# posterior distribution for p_a = 0.55
def normalconst(F_a):
    return gm.factorial(F_a) * gm.factorial(200 - F_a) / gm.factorial(201)

def posterior(F_a, p_a):
    return p_a ** (F_a) * (1 - p_a) ** (200 - F_a) / normalconst(F_a)

for i, F_a in enumerate(Fa_toss_count_55):
    p_a = np.linspace(0, 1, 1000)
    plt.plot(p_a, posterior(F_a, p_a), label='Posterior' + str(i+1))
plt.xlabel('$p_a$')
plt.ylabel('Posterior')
plt.legend()
plt.show()

# posterior distribution for p_a = 0.9
for i, F_a in enumerate(Fa_toss_count_90):
    p_a = np.linspace(0, 1, 1000)
    plt.plot(p_a, posterior(F_a, p_a), label='Posterior' + str(i+1))
plt.xlabel('$p_a$')
plt.ylabel('Posterior')
plt.legend()
plt.show()

# %%
# back-envelope calculation for p_a close to 0.5
# for a Gaussian the width of the error bars is sigma / sqrt(n)
# the width should be less than (p_a - 0.5)
# p_a - 0.5 > sigma / sqrt(n)
# n > sigma ** 2 / (p_a - 0.5) ** 2
def backenvelope(sigma):
    return sigma ** 2 / (0.55 - 0.5) ** 2
print(backenvelope(3))
print(backenvelope(2))

n = np.arange(0, 3601)
plt.figure()
R, F_a = simulate_b(3600)
plt.plot(n, R, label='$\mathcal{R}$' + str(i+1))
plt.xlabel('$F$')
plt.ylabel('R')
plt.ylim(0, 1e3)
plt.legend()


plt.figure()
R, F_a = simulate_b(1600)
plt.plot(np.arange(0,1601), R, label='$\mathcal{R}$' + str(i+1))
plt.xlabel('$F$')
```

```
plt.ylabel('R')
plt.ylim(0, 1e3)
plt.legend()

# %%
# 2 e
print(np.e ** 500)
```

Homework 5

Due 2/27 4pm

---

1. 3 Advantages of Counting mRNA Molecules to Measure Gene Expression:

- Discrete Numbers: We can quantify the 'exact' number of mRNA moleucules that contribute to protein expression. We can also compare this to the total number of mRNA molecules to get a better picture of how much stuff is expressing the protein.
- More stuff: We can account for a parts of the cell that are low in intensity but still contribute to the overall expression of the protein.
- We are not counting empty space where there are no mRNA molecules. The intensity recording method would count regions of the cell(that do not have any mRNA molecules) as a data point which could skew the results.

2. 3-5 Potential Sources of Error

- Is the field of view (FOV) representative of the entire cell? The location of where this image is taken with respect to the cell could be a major source of error; we could make a mistake by taking a picture of the same small region across multiple cells, and this would not accurately represent the entire cell.
- Is the labeling process accurate? Does it account for all the mRNA molcules that relate to protein expression, and can we be sure that it doens't also label other unrelated molecules which have nothing to do with protein expression? Also
- How do we know what counts as a single molecule in relation to a fluorescing spot? There are spots of different intensities and sizes, so it could be possible that a large bright spot is actually multiple mRNA molecules.
- How do we know that the mRNA molecules are not degrading over time? That is, there is a finite time for the mRNA to fluoresce before it degrades, and we could be missing data.

3. Most problematic: The labeling procees is the most problematic because there is so much biological complexity that we can't account for. Unless we have a perfect labeling process, there can be errors from not accounting for all the mRNA molecules and also accidentally labeling other molecules that are not related to protein expression.

4. Lets say we have N total number of particles in a volume V. The probability of finding n_o particles in a volume v_o can be given by the binomial distribution:

$$P(n_o) = \frac{N!}{n_o!(N-n_o)!} f^{n_o}(1-f)^{N-n_o}$$

where the frequency

$$f = \frac{\lambda}{N} = \frac{rT}{N}$$

is the ratio of the average number of particles λ to the total number N, and r is the rate of particles entering this volume over a time T. After some mathy stuff:

$$\begin{aligned} P(n_o) &= \frac{N(N-1)\dots(N-n_o+1)\cancel{(N-n_o)!}}{n_o!\cancel{(N-n_o)!}} \left(\frac{\lambda}{N}\right)^{n_o} \frac{\left(1-\frac{\lambda}{N}\right)^N}{\left(1-\frac{\lambda}{N}\right)^{n_o}} \\ &= \frac{N(N-1)\dots(N-n_o+1)}{N^{n_o}} \left(1-\frac{\lambda}{N}\right)^{-n_o} \frac{\lambda^{n_o}}{n_o!} \left(1-\frac{\lambda}{N}\right)^N \end{aligned}$$

and for large number of total particles $N \to \infty$, we have two terms that go to 1:

$$\begin{aligned}\frac{N(N-1)\dots(N-n_o+1)}{N^{n_o}} &= \frac{N}{N}\frac{N-1}{N}\cdots\frac{N-n_o+1}{N} \\ &= 1\left(1-\frac{1}{N}\right)\dots\left(1-\frac{n_o-1}{N}\right) \approx 1\end{aligned}$$

and

$$\left(1-\frac{\lambda}{N}\right)^{-n_o} \to 1$$

And from the limit definition of the exponential function:

$$\lim_{N\to\infty}\left(1+\frac{-\lambda}{N}\right)^{N} \to e^{-\lambda}$$

So we finally get

$$P(n_o) = \frac{\lambda^{n_o}}{n_o!}e^{-\lambda}$$

thus obeying Poisson statistics.

5. This 'anomaly' is perhaps due to the fact that the molecule counts are centered around the detected nucleus center. In Problem 4, we assumed that this test volume was a randomly chosen volume, and the researchers at Fancy University have chosen a test volume that is dependent on focusing around nucleus centers for each molecule count. This method would disregard volumes that are not related to a nucleus, so we have added a bias in our method of data analysis. If we were to exclude this automatic nucleus centering i.e. we define this $100\,\mu m^3$ cylinder randomly in our FOV, we may get something closer to Poissonian noise.

Homework 6

Due 3/7

---

1. (a) The joint entropy is (where $\log = \log_2$)

$$\begin{aligned} H(V,T) &= \sum_{V,T} P(V,T) \log\left(\frac{1}{P(V,T)}\right) \\ &= \left[\frac{6}{16}\log(16) + \frac{4}{32}\log(32) + \frac{2}{8}\log(8) + \frac{1}{4}\log(4)\right] = \frac{54}{16} \\ H(V,T) &= \boxed{3.38 \text{ bits}} \end{aligned}$$

Given the marginal probability

$$\begin{aligned} P(V = \text{Sunny}) &= \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{4}{16} = \frac{1}{4} \\ P(V = \text{Cloudy \& dry}) &= \frac{1}{16} + \frac{1}{8} + \frac{1}{32} + \frac{1}{32} = \frac{8}{32} = \frac{1}{4} \\ P(V = \text{Cloudy \& rain}) &= \frac{1}{4} \\ P(V = \text{Cloudy \& snow}) &= \frac{1}{4} \end{aligned}$$

marginal entropy of V is

$$\begin{aligned} H(V) &= \sum_{V} P(V) \log\left(\frac{1}{P(V)}\right) \\ &= \frac{1}{4}\log(4) + \frac{1}{4}\log(4) + \frac{1}{4}\log(4) + \frac{1}{4}\log(4) \\ H(V) &= \boxed{2 \text{ bits}} \end{aligned}$$

And given the marginal probability

$$\begin{aligned} P(T = \text{Miserably Cold}) &= \frac{1}{16} + \frac{1}{16} + \frac{1}{8} + \frac{1}{4} = \frac{1}{2} \\ P(T = \text{Very Cold}) &= \frac{1}{4} \\ P(T = \text{Cold}) &= \frac{1}{8} \\ P(T = \text{Chilly}) &= \frac{1}{8} \end{aligned}$$

marginal entropy of T is

$$\begin{aligned} H(T) &= \sum_{T} P(T) \log\left(\frac{1}{P(T)}\right) \\ &= \frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \frac{1}{8}\log(8) + \frac{1}{8}\log(8) \\ H(T) &= \boxed{1.75 \text{ bits}} \end{aligned}$$

(b) The conditional entropy of T given $V = v$ is

$$H(T|V=v) = \sum_{T} P(T|V=v) \log\left(\frac{1}{P(T|V=v)}\right)$$

and from Bayes' theorem

$$P(T|V=v) = \frac{P(V=v,T)}{P(V=v)}$$

So for $v =$ Sunny:

$$H(T|V=\text{Sunny}) = \frac{1}{4}\log(4) + \frac{1}{4}\log(4) + \frac{1}{4}\log(4) + \frac{1}{4}\log(4) = \boxed{2 \text{ bits}}$$

For $v =$ Cloudy & dry:

$$H(T|V=\text{Cloudy \& dry}) = \frac{1}{4}\log(4) + \frac{1}{2}\log(2) + \frac{1}{8}\log(8) + \frac{1}{8}\log(8) = \boxed{1.75 \text{ bits}}$$

For $v =$ Cloudy & rain:

$$H(T|V=\text{Cloudy \& rain}) = \frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \frac{1}{8}\log(8) + \frac{1}{8}\log(8) = \boxed{1.75 \text{ bits}}$$

For $v =$ Cloudy & snow:

$$H(T|V=\text{Cloudy \& snow}) = \log(1) = \boxed{0 \text{ bits}}$$

this makes sense since its *always* miserably cold given it is Cloudy & snowing.
(c) The conditional entropy as an average

$$\begin{aligned} H(T|V) &= \sum_V P(V)[H(T|V=v)] \\ &= \frac{1}{4}H(T|V=v) \\ H(T|V) &= \frac{1}{4}(2 + 1.75 + 1.75 + 0) = \boxed{1.38 \text{ bits}} \end{aligned}$$

(d) Using product rule on the joint entropy:

$$\begin{aligned} H(V,T) &= \sum_{V,T} P(V,T)\log\left(\frac{1}{P(T|V)P(T)}\right) \\ &= \sum_{V,T} P(V,T)\log\left(\frac{1}{P(T|V)}\right) + \sum_{V,T} P(V,T)\log\left(\frac{1}{P(T)}\right) \end{aligned}$$

and from sum the sum rule:

$$\begin{aligned} P(T) &= \sum_V P(V,T) \\ &= \sum_V P(T|V)P(V) \end{aligned}$$

so

$$H(V,T) = H(T|V) + H(T) \implies H(T|V) = H(V,T) - H(T) = 3.38 - 2 = 1.38 \text{ bits}$$

which confirms the result from part (c), and we can also see that

$$\begin{aligned} & H(V,T) = H(T) + H(V|T) \\ \implies & H(V|T) = H(V,T) - H(T) = 3.38 - 1.75 = \boxed{1.63 \text{ bits}} \end{aligned}$$

(e) The mutual information is

$$\begin{aligned} I(V;T) &= H(V) - H(V|T) \quad \text{or} \quad H(T) - H(T|V) \\ &= 2 - 1.63 = \boxed{0.37 \text{ bits}} \end{aligned}$$

2. (a) For a Gaussian defined by the PDF(Probability Density Function)

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

the entropy is (here log is the natural logarithm $\log_e = \ln$ i.e. unit of nats)

$$\begin{aligned} H(P) &= -\int_{-\infty}^{\infty} P(x)\log(P(x))\,\mathrm{d}x \\ &= -\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\left(\frac{-x^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)\right)\mathrm{d}x \\ &= \frac{1}{\sqrt{8\pi\sigma^6}}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x + \frac{\log\left(\sqrt{2\pi\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x \end{aligned}$$

and using some useful Gaussian integrals:

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-ax^2}\,\mathrm{d}x &= \sqrt{\frac{\pi}{a}} \\ \int_{-\infty}^{\infty} x^2 e^{-ax^2}\,\mathrm{d}x &= \frac{1}{2}\sqrt{\frac{\pi}{a^3}} \end{aligned}$$

where $a = \frac{1}{2\sigma^2}$, so

$$\begin{aligned} H(P) &= \frac{1}{\sqrt{8\pi\sigma^6}}\left[\frac{1}{2}\sqrt{8\pi\sigma^6}\right] + \frac{\log\left(\sqrt{2\pi\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}\sqrt{2\pi\sigma^2} \\ &= \frac{1}{2} + \frac{1}{2}\log\left(2\pi\sigma^2\right) \\ &= \frac{1}{2}\left[1 + \log\left(2\pi\sigma^2\right)\right] \end{aligned}$$

and $H(P)$ can be negative when

$$\begin{aligned} & 1 + \log\left(2\pi\sigma^2\right) < 0 \\ \implies \sigma^2 < \frac{1}{2\pi e} \quad &\text{or} \quad \sigma < \frac{1}{\sqrt{2\pi e}} \end{aligned}$$

(b) Since ξ and X are independent, the variance of $Y = \xi + X$ is the sum of the variances

$$\mathrm{Var}(Y) = \mathrm{Var}(\xi) + \mathrm{Var}(X) = \sigma_\xi^2 + \sigma_X^2$$

(c) From the Sum rule

$$P_Y(y) = \sum_{Z=z} P_\xi(\xi = z) P_X(X)$$

we can change the discrete case to a continuous one by integrating over all possible values of $\xi = z$ to find the probability density function $P_Y(y)$:

$$\begin{aligned} P_Y(y) &= \int_{-\infty}^{\infty} P_\xi(\xi = z) P_X(X|\xi = z)\,\mathrm{d}z \\ &= \int_{-\infty}^{\infty} P_{X,\xi}(X, \xi = z)\,\mathrm{d}z \end{aligned}$$

Since ξ and X are independent, $P_{X,\xi}(X,\xi) = P_X(X)P_\xi(\xi)$, and $X = Y - \xi$:

$$\begin{aligned} P_Y(y) &= \int_{-\infty}^{\infty} P_X(X = y - z) P_\xi(\xi = z)\,\mathrm{d}z \\ &= \int_{-\infty}^{\infty} P_X(y - z) P_\xi(z)\,\mathrm{d}z \end{aligned}$$

(d) We can just plug in the Gaussian PDFs for P_X and P_ξ where we assume the means are zero:

$$\begin{aligned}
P_Y(y) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(y-z)^2}{2\sigma_X^2}} \frac{1}{\sqrt{2\pi\sigma_\xi^2}} e^{-\frac{z^2}{2\sigma_\xi^2}} \,\mathrm{d}z \\
&= \frac{1}{2\pi\sigma_X\sigma_\xi} \int_{-\infty}^{\infty} e^{-\frac{y^2-2yz+z^2}{2\sigma_X^2} - \frac{z^2}{2\sigma_\xi^2}} \,\mathrm{d}z \\
&= \frac{1}{2\pi\sigma_X\sigma_\xi} e^{-\frac{y^2}{2\sigma_X^2}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma_X^2} - \frac{z^2}{2\sigma_\xi^2} + \frac{yz}{\sigma_X^2}} \,\mathrm{d}z \\
&= \frac{1}{2\pi\sigma_X\sigma_\xi} \left(\frac{2\pi\sigma_X^2\sigma_\xi^2}{\sigma_X^2+\sigma_\xi^2}\right)^{1/2} e^{-\frac{y}{2(\sigma_X^2+\sigma_\xi^2)}} \\
&= \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}} e^{-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)}}
\end{aligned}$$

which is also a Gaussian with our expected variance (add variances)!
(e) The mutual information is

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

so

$$\begin{aligned}
H(Y) &= -\int_{-\infty}^{\infty} P_Y(y)\log(P(y))\,\mathrm{d}y \\
&= -\frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)}} \log\left(\frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}} e^{-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)}}\right) \mathrm{d}y \\
&= -\frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)}} \left[-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)} - \log\Big(\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}\Big)\right] \mathrm{d}y \\
&= \frac{1}{\sqrt{8\pi(\sigma_X^2+\sigma_\xi^2)^3}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)}} \,\mathrm{d}y + \frac{\log\Big(\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}\Big)}{\sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2(\sigma_X^2+\sigma_\xi^2)}}
\end{aligned}$$

using

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-ay^2} \,\mathrm{d}x &= \sqrt{\frac{\pi}{a}} = \sqrt{2\pi(\sigma_X^2+\sigma_\xi^2)} \\
\int_{-\infty}^{\infty} y^2 e^{-ay^2} \,\mathrm{d}x &= \frac{1}{2}\sqrt{\frac{\pi}{a^3}} = \frac{1}{2}\sqrt{8\pi(\sigma_X^2+\sigma_\xi^2)^3}
\end{aligned}$$

where $a = \frac{1}{2(\sigma_X^2+\sigma_\xi^2)}$, so

$$H(Y) = \frac{1}{2} + \frac{1}{2}\log\big(2\pi(\sigma_X^2+\sigma_\xi^2)\big)$$

and

$$\begin{aligned}
-H(Y|X) &= -\int_{-\infty}^{\infty} P_X(x) H(Y|X=x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} P_X(x) \left[\int_{-\infty}^{\infty} P_Y(y|X=x)\log(P_Y(y|X=x))\,\mathrm{d}y\right] \mathrm{d}x
\end{aligned}$$

in the second integral we can use $y = x + \xi$ so

$$P_Y(y|X = x) = P_\xi(\xi = y - x|X = x)$$

and since ξ and X are independent

$$P_\xi(\xi = y - x|X = x) = P_\xi(\xi)$$

so

$$\begin{aligned} -H(Y|X) &= \int_{-\infty}^{\infty} P_X(x) \left[\int_{-\infty}^{\infty} P_\xi(\xi) \log(P_\xi(\xi)) \, \mathrm{d}\xi \right] \mathrm{d}x \\ &= -\int_{-\infty}^{\infty} P_X(x) H(\xi) \, \mathrm{d}x = -H(\xi) \end{aligned}$$

and from part (a) we know that $H(\xi) = \frac{1}{2}\left[1 + \log\left(2\pi\sigma_\xi^2\right)\right]$, so

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= \frac{1}{2} \log \left(\frac{2\pi(\sigma_X^2 + \sigma_\xi^2)}{2\pi\sigma_\xi^2} \right) \\ &= \frac{1}{2} \log \left(\frac{\sigma_X^2 + \sigma_\xi^2}{\sigma_\xi^2} \right) \\ &= \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_\xi^2} \right) \end{aligned}$$

- If σ_X is large and σ_ξ is small, then $I(X,Y)$ is large.
- If σ_X is small and σ_ξ is large, then $I(X,Y)$ is small or zero

4 (a) Given

$$m(n) = \frac{m(n+1)}{Nb_i} \implies m(n-1) = \frac{m(n)}{Nb_i}$$

the expected value of $x = \log(m(n))$ is

$$\mathrm{E}[x] = \sum_i x_i p_i$$

where the probability of horse i wins is p_i, so

$$\begin{aligned} \mathrm{E}[x] &= \sum_i (\log(m(n-1)) + \log(Nb_i)) p_i \\ &= \sum_i \log(m(n-1)) p_i + \sum_i \log Np_i + \sum_i \log(b_i) p_i \\ &= \sum_i \log(m(n-2) Nb_i) p_i + \sum_i \log Np_i + \sum_i \log(b_i) p_i \end{aligned}$$

which is a recursive structure so we get

$$\begin{aligned} &= \log(m(0)) + n(\mathrm{E}[\log N] + \mathrm{E}[\log b_i]) \\ &= \log(m(0)) + n \log N + n \, \mathrm{E}[\log b_i] \end{aligned}$$

(b) Finding where the derivative is zero and since we only care about the case of maximizing

$$\mathrm{E}[\log b_i] = \sum_i p_i \log(b_i)$$

and since b_i is a normalized vector:

$$\sum_i b_i = 1$$

we can use the method of Lagrange multipliers to find the maximum of $\mathrm{E}[\log b_i]$ (from Cover & Thomas Chapter 6):

$$\mathcal{L} = \sum_i p_i \log(b_i) + \lambda \sum_i b_i$$

and differentiating with respect to b_i:

$$\frac{\partial \mathcal{L}}{\partial b_i} = \frac{p_i}{b_i} + \lambda = 0 \implies b_i = -\frac{p_i}{\lambda}$$

and from the constraint:

$$\sum_i b_i = -\frac{\sum_i p_i}{\lambda} \implies \lambda = -1 \implies b_i = p_i$$

and to find the growth rate we can use

$$\mathrm{E}[x] = \log(m(0)) + \lambda_{\max}$$

where $n = 1$ is the expected growth rate after one race:

$$\begin{aligned}
\lambda_{\max} &= \log N + \sum_i p_i \log(p_i) \\
&= \log N - \sum_i p_i \log\left(\frac{1}{p_i}\right) \\
&= \log N - H(p)
\end{aligned}$$

(c) The max capital minus our capital based on our bet $b_i = q_i$ is

$$\begin{aligned}
\log(m(n)) &= \log(m(0)) + n \log N + n \sum_i p_i \log\left(p_i \frac{q_i}{p_i}\right) \\
&= \log(m(0)) + n \log N + n \sum_i p_i \log(p_i) + n \sum_i p_i \log\left(\frac{q_i}{p_i}\right) \\
&= \log(m(0)) + n \log N - nH(p) - nD_{KL}(p||q)
\end{aligned}$$

and since the maximimum capital is

$$m_{\max}(n) = m(0) e^{n \log N - nH(p)}$$

we can see the capital falls off exponentially

$$\begin{aligned}
m(n) &= m(0) e^{n \log N - nH(p) - nD_{KL}(p||q)} \\
&= m_{\max}(n) e^{-nD_{KL}(p||q)}
\end{aligned}$$

(d) Given

$$m(n+1) = \frac{1}{p_i} b_i m(n) \implies m(n) = \frac{1}{p_i} b_i m(n-1)$$

so the expected value of $\log(m(n))$:

$$\mathrm{E}[\log(m(n))] = \sum_i p_i \left(\log(m(n-1)) + \log\left(\frac{b_i}{p_i}\right) \right)$$

and since $b_i = p_i$ for the optimal strategy:

$$\begin{aligned}
\mathrm{E}[\log(m(n))] &= \mathrm{E}[\log(m(n-1))] + \sum_i p_i \log(1) \\
&= \mathrm{E}[\log(m(n-2))] + \sum_i p_i \log(1) + \sum_i p_i \log(1) \\
&\dots \\
&= \log(m(0)) + n \log(1) \\
&= \log(m(0))
\end{aligned}$$

so we end up with the same money we started with! Therefore the long term capital growth is 0.
(e) (f) :(