481 Lecture 1/16/24

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

An ensemble: x random variable

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

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'

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

$$P(y | x = `q')$$

Peak at $y = `u'$

$$\neq P_Y(y)$$

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':

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

2 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 $\Theta \to$ 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

---

3 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

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

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 3.1. where the the bent coin $p_a = 0.9$ probability goes to infinity as well

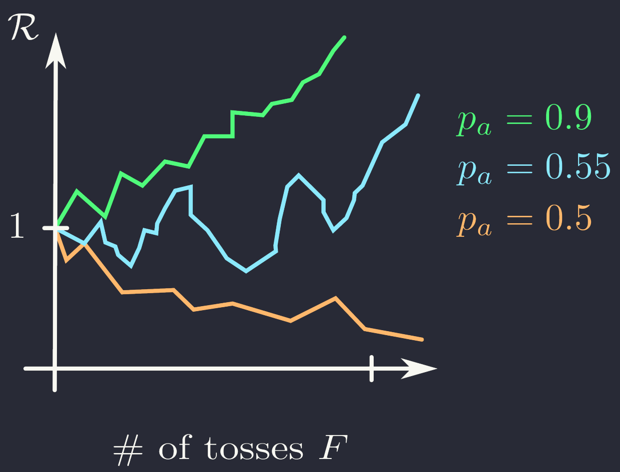

Figure 3.1: Ratio of the two probabilities as a function of the number of tosses

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"

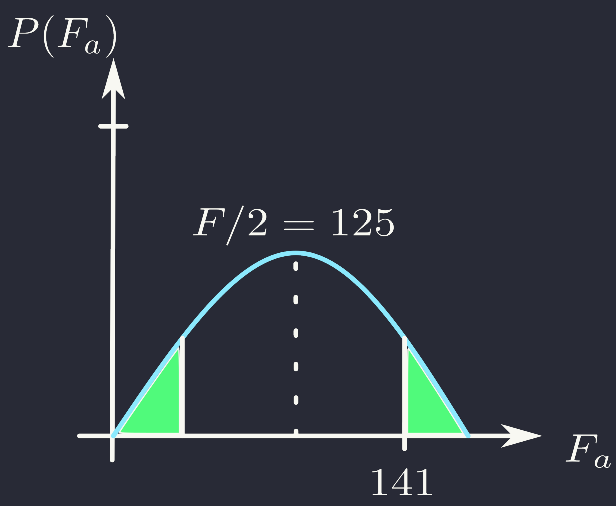

Figure 3.2: Finding p-value based on the Gaussian distribution

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 $20 : 1$.

4 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

- $P(\text{next draw} = \text{black}) =?$
- 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: $\{x_n\} \quad n = 1, \dots, N$
- 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

$$\begin{aligned}\hat{x} &\equiv \sum_n \frac{x_n}{N} \quad \text{empirical mean} \\ S &= \sum_n (x_n - \hat{x})^2 \quad \text{sum of square deviations}\end{aligned}$$

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

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

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 4.1. There also exista abackground noise b that is always present.

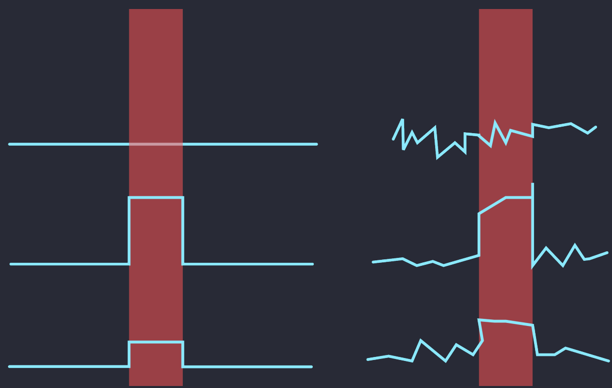

Figure 4.1: From top to bottom we have a no light $f = 0$, and two cases of $f = 1$ for a bright light and dim light. The Pefect detector (left) sees and appropriates with the correct response while the noisy (Gaussian) detector may have a incorrect response (especially for the dimmer signal).

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 4.2 shows the expectation of the model.

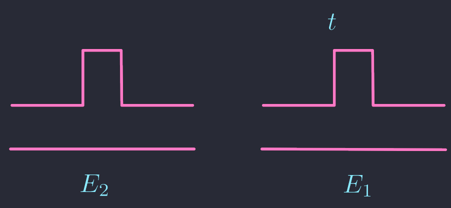

Figure 4.2: The expectation of the models given experiment E_1 and E_2. The top is for an expected model of a flash and no flash for bottom. NOTE that there also is Gaussian noise η added to both scenerios.

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!):

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

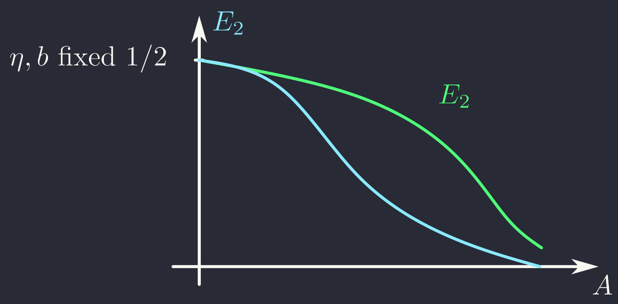

Figure 4.3: The error rate as a function of the brightness of the flash.

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.

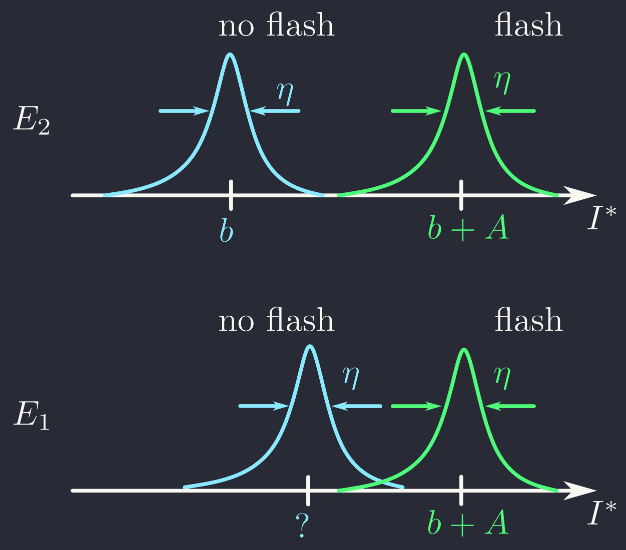

Figure 4.4: There is a shift in the no flash model in E_1

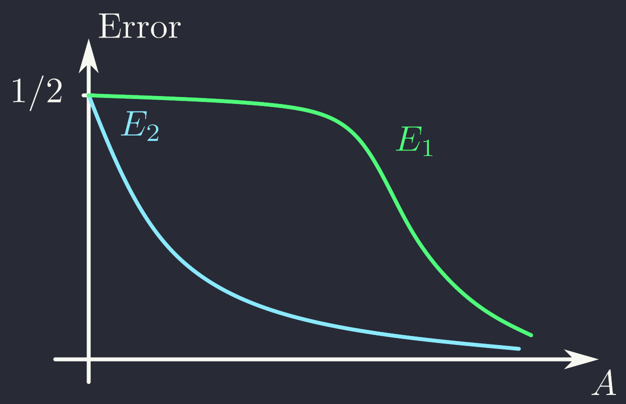

Figure 4.5: The error rate as a function of amplitude A.

Takehomes

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

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 $\rightarrow$ Data analysis (had a subtle flaw) $\rightarrow$ one Conclusion
- Gregor et al. Better data (very careful exp) $\rightarrow$ extrene careful data analysis $\rightarrow$ 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. $\sim$ 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,

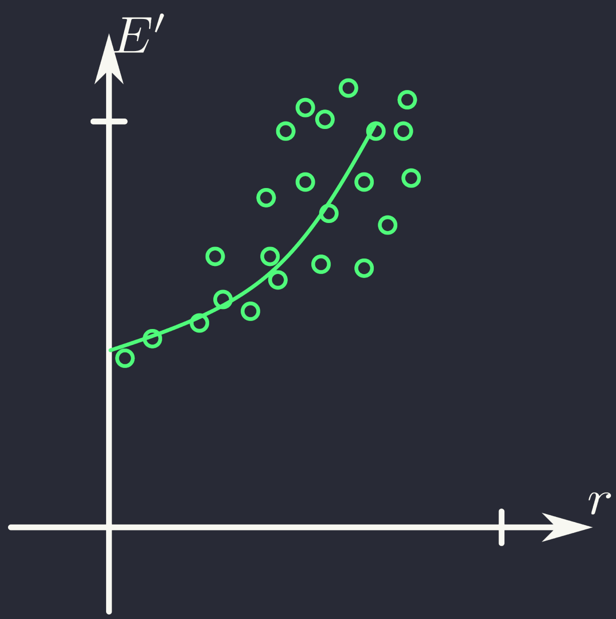

Figure 4.6: Recorded count vs. Tot fluorescence

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.

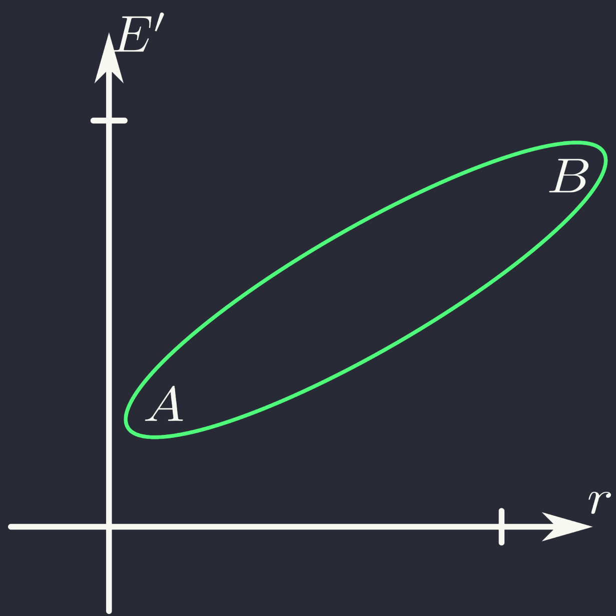

Figure 4.7: Fluctuations of the Hb and Kr are anticorrelated

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

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