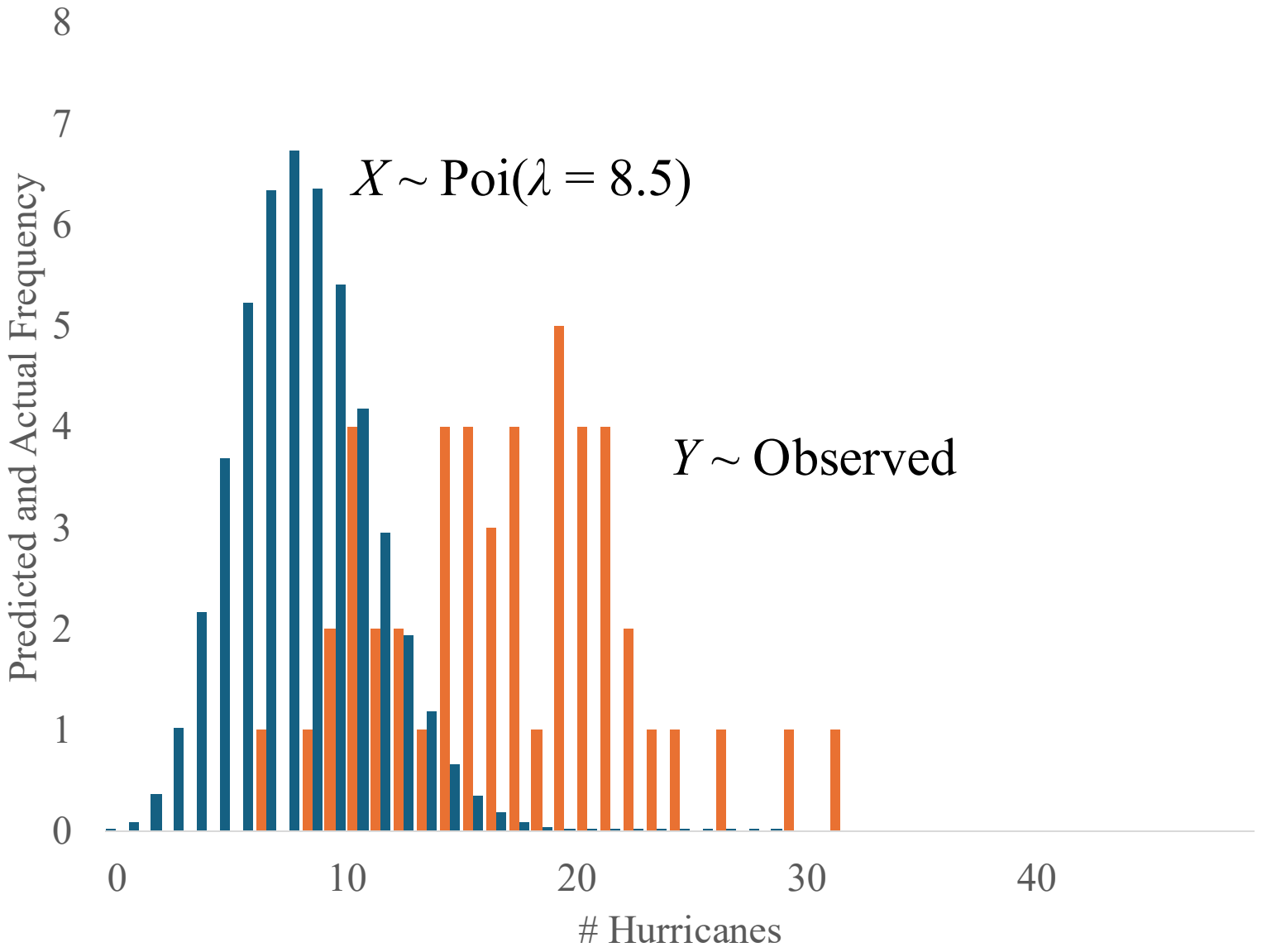
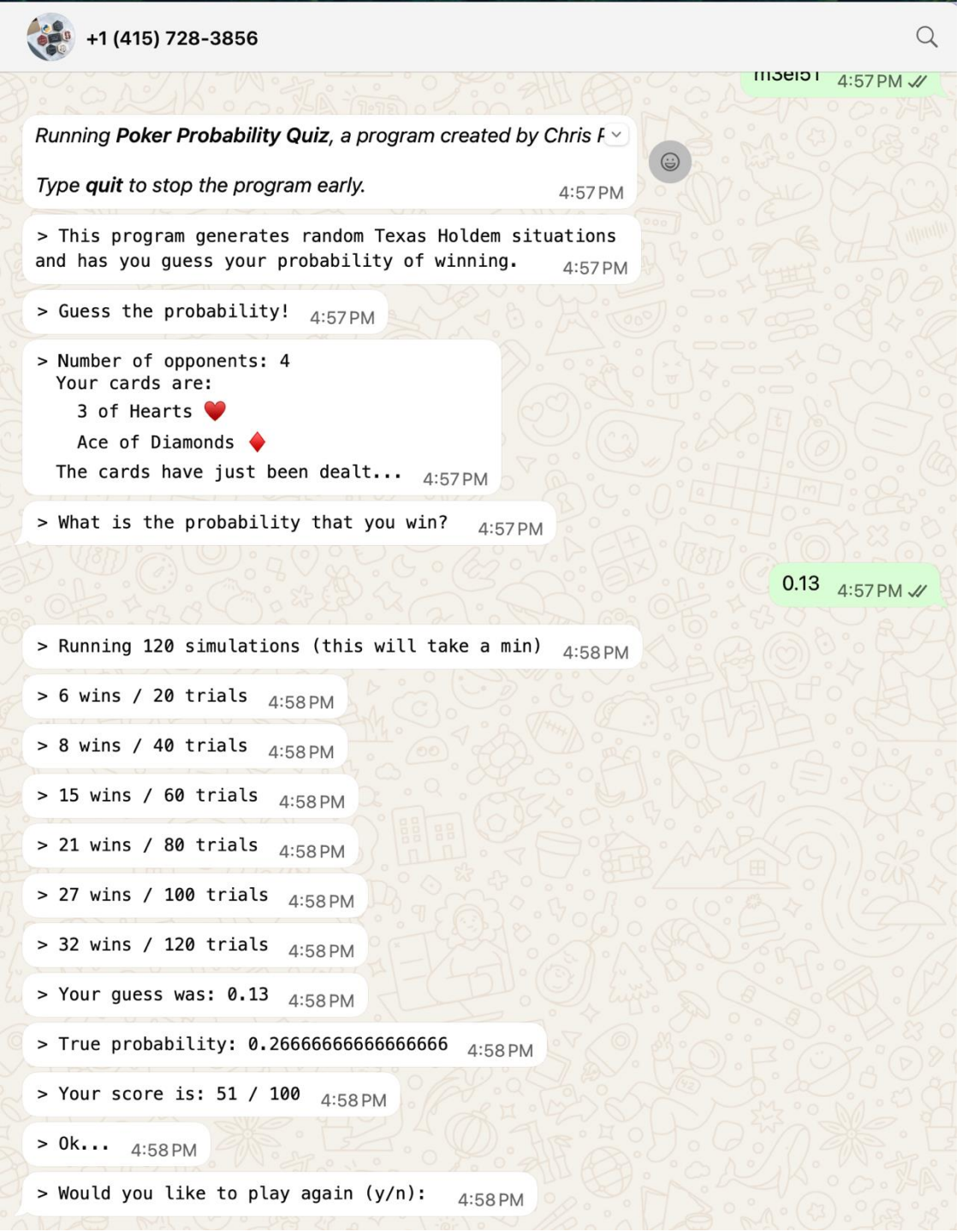
KL Divergence

Chris Piech
CS109, Stanford University

How do you Compare Random Variables?





True probability: 0.27
Guess is: 0.13
Score is: 51/ 100

What's the score function?

Send a WhatsApp message to...

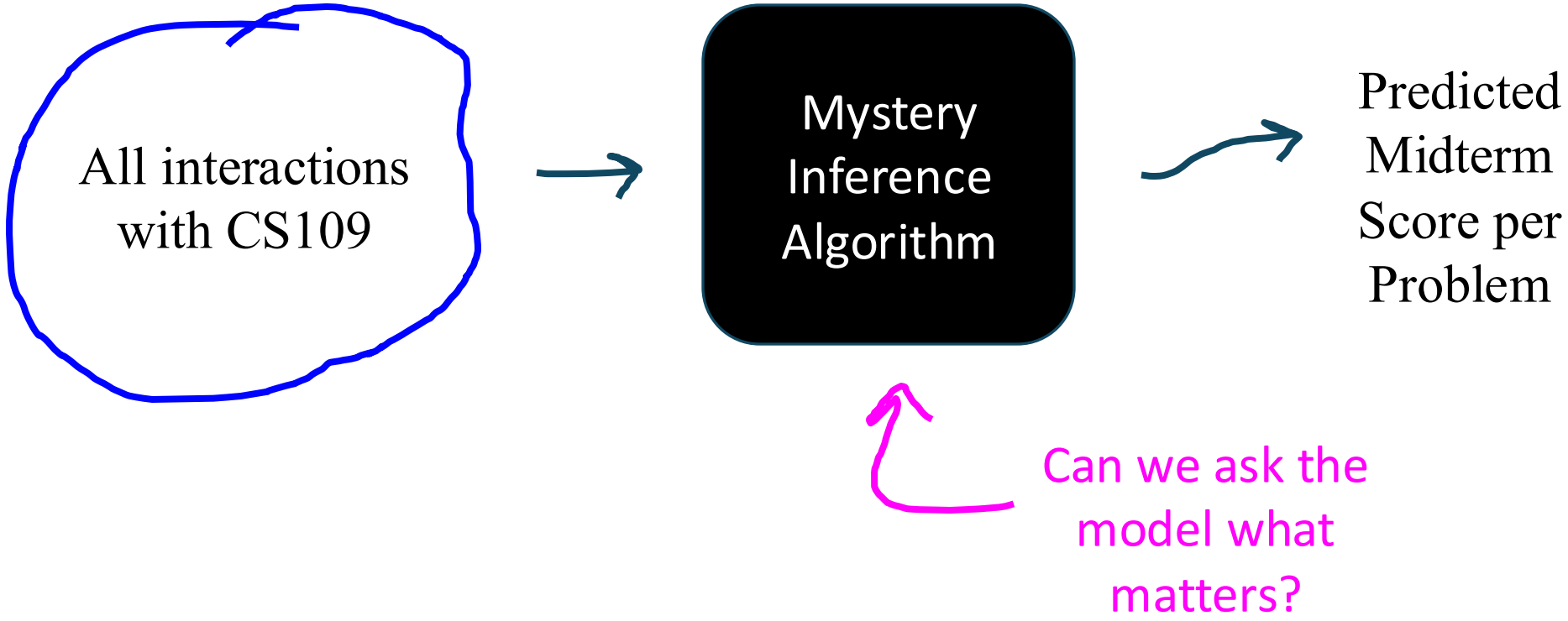
+1 (415) 728-3856

With the text message...

M3EI51

Or scan the QR code below.

What is the Best Way to Prepare for the Final?



I only look at this data in exceptional circumstances. You are adults.

Aside: Used to identify phone use in midterm. No false positives. Two true positives.

What is the best summary

summary_a = “Your mom wants you to call her”

summary_b = “Your pet elephant ate all our uranium”

summary_c = “…”

email = “Hi honey, I hope you are doing well at Stanford. Things are fine but your pet elephant…”

Review

Uncertainty of a Random Variable (Entropy)

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

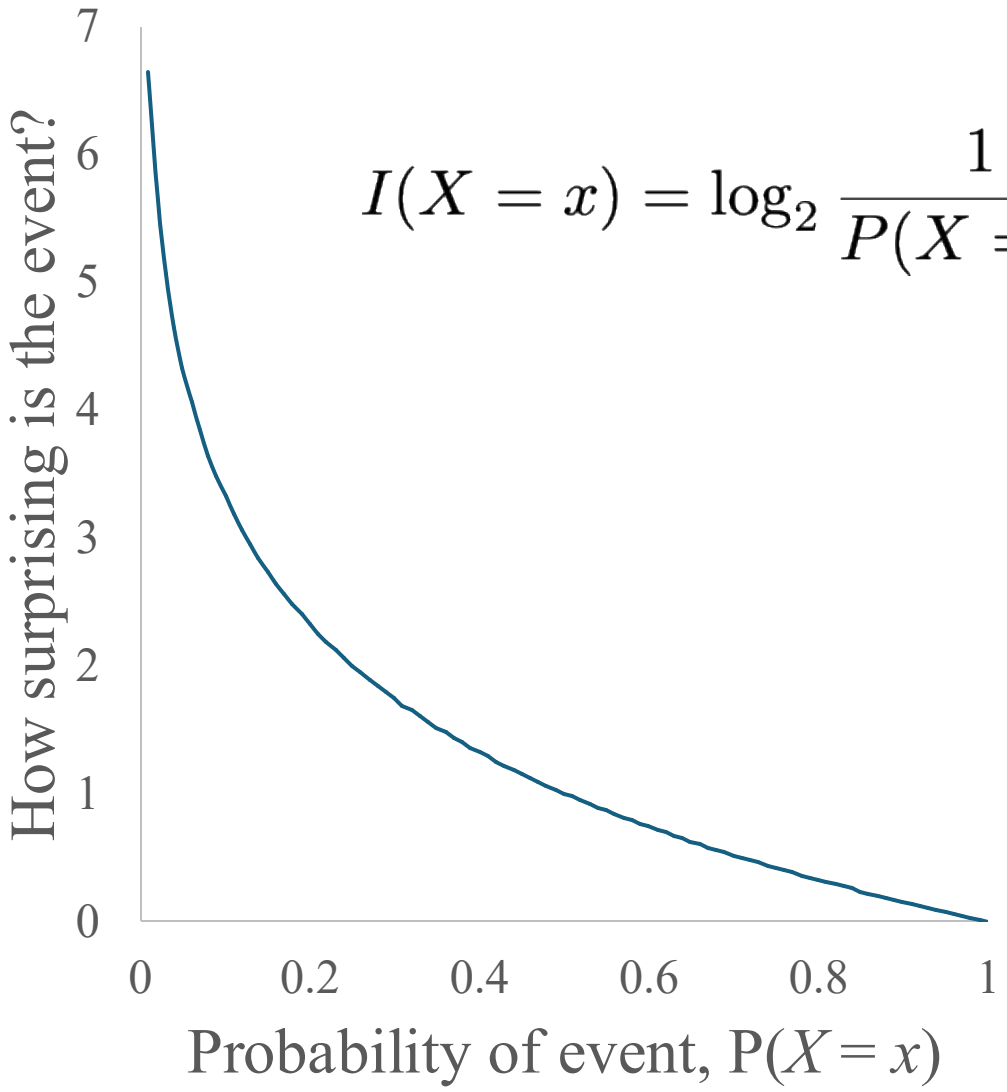
Calculates expected surprise

$$\text{Uncertainty}(X) = \sum_{x \in X} \log_2 \frac{1}{P(X = x)} \cdot P(X = x)$$

My preferred name for "entropy" aka H(X)

$$\text{Surprise}(X = x) = \log_2 \frac{1}{P(X = x)})$$

Surprise of an Event, $\mathrm{I}(X = x)$

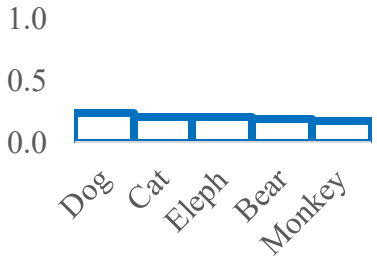


$$I(X = x) = \log_2 \frac{1}{P(X = x)}$$

Probability of Event $P(X = x)$	Surprise of Event $I(X = x)$
1	0
½	1
1/4	2
1/8	3
1/16	4
1/32	5
1/64	6

$I(X = x)$ stands for "Information Content" aka "Surprisal" aka "Self-Information"

Which Question is Better?

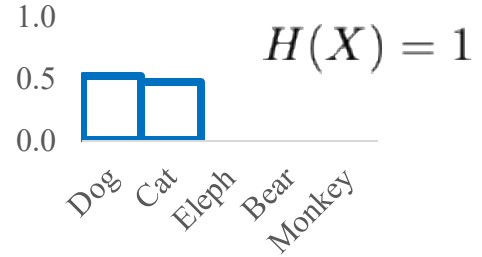


$H(X) = 2.3$

Is it a pet?

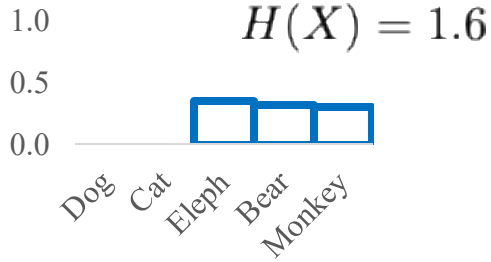
$E[H(X)] = 1.3$

yes
p = 0.44



$H(X) = 1$

no
p = 0.56

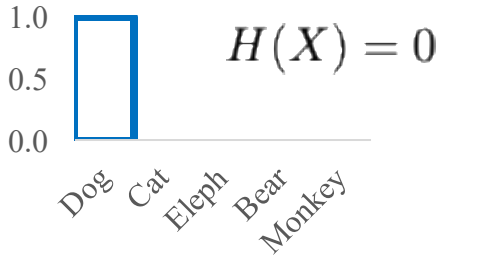


$H(X) = 1.6$

Is it a Dog?

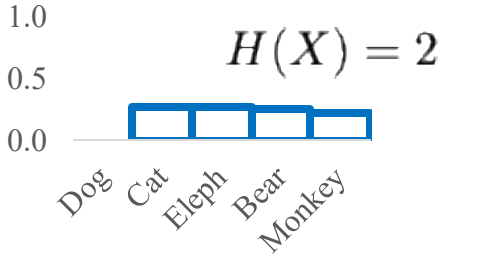
$E[H(X)] = 1.7$

yes
p = 0.23



$H(X) = 0$

no
p = 0.77



$H(X) = 2$

Information (Surprise) is additive

$$\text{Surprise}(E) = \frac{1}{\log_2 P(E)}$$

Let E and F be independent events. Show that I(E and F) = I(E) + I(F)

$$\begin{aligned} I(E \text{ and } F) &= \log_2 \frac{1}{P(E \text{ and } F)} \\ &= \log_2 \frac{1}{P(E) \cdot P(F)} \\ &= -\log_2 [P(E) \cdot P(F)] \\ &= -\log_2 P(E) - \log_2 P(F) \\ &= I(E) + I(F) \end{aligned}$$

Now your turn

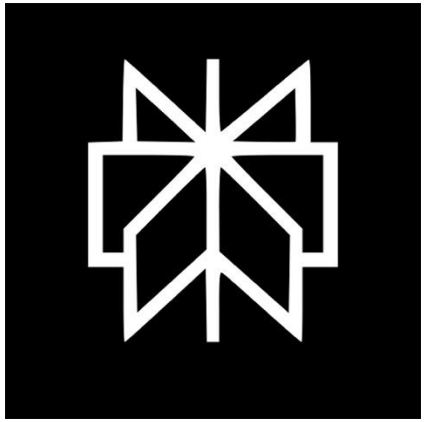
Entropy (Uncertainy) as # Decisions

$$H(X) = \sum_{x \in X} \log_2 \frac{1}{P(X = x)} \cdot P(X = x)$$

Consider a random variable X which can take on values $\{1, 2, \ldots n\}$ with equal probability. In other words $P(X = x) = 1/n$. What is the Entropy of X?

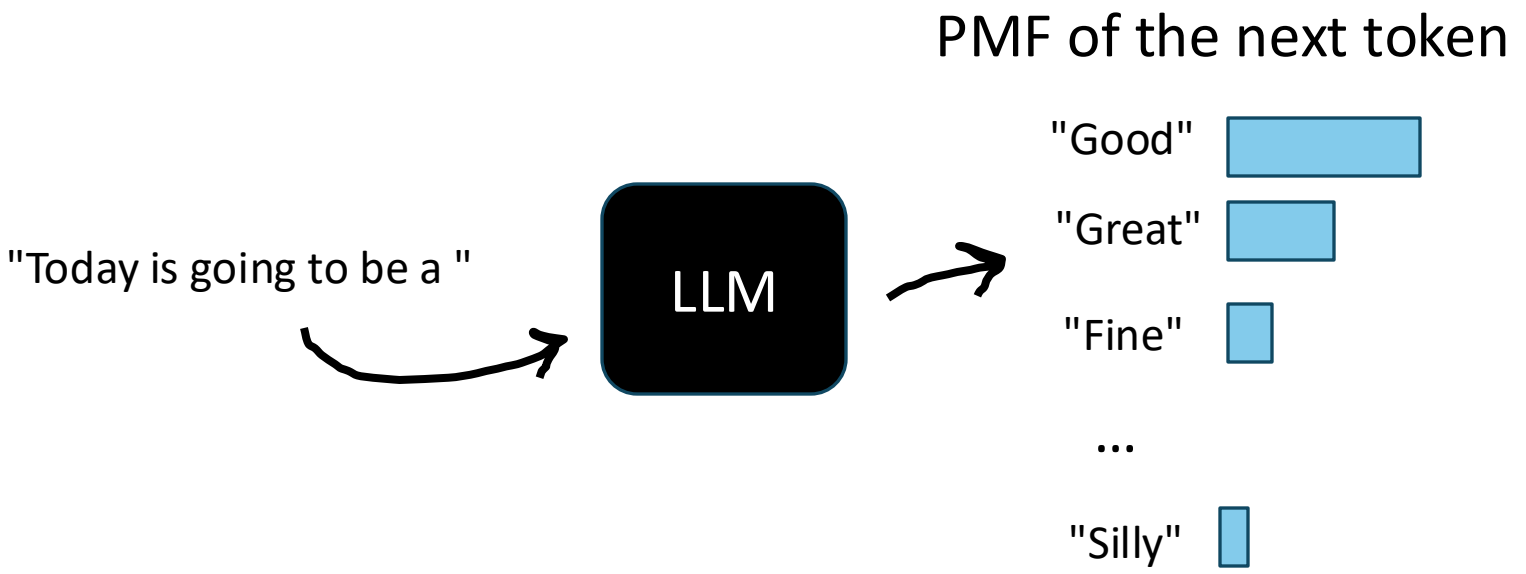
$$\begin{aligned} H(X) &= \sum_{i=1}^{n} \log_2 \frac{1}{P(X = x)} \cdot P(X = x) \\ &= \sum_{i=1}^{n} (\log_2 \frac{1}{\frac{1}{n}}) \cdot \frac{1}{n} \\ &= \sum_{i=1}^{n} (\log_2 n) \cdot \frac{1}{n} \\ &= n \cdot (\log_2 n) \cdot \frac{1}{n} = \log_2 n \end{aligned}$$

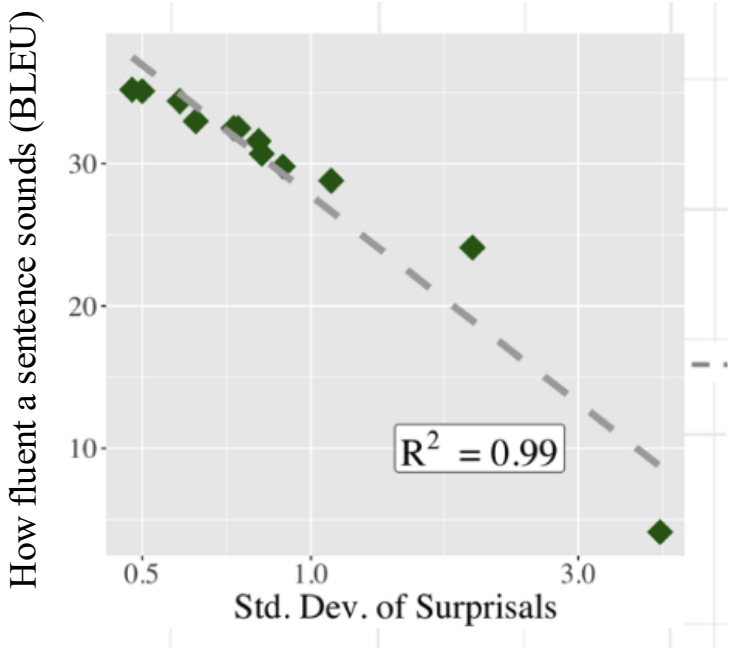
If I tell you entropy, can you calculate n?



Perplexity

Entropy of an LLM next token





Entropy of next token

3.5 1.5 2.1 1.1 1.2 1.3 2.2 1.2 5.9 0.3

How many binary questions?

"Today is going to be a **good** day. **We are going to learn about KL Divergence**"

Perplexity

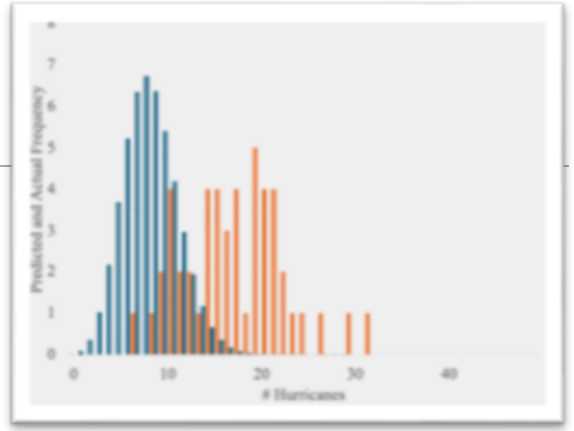
11 2.8 4.3 2.1 2.3 2.5 4.6 2.3 59.7 1.5

How many effective options left

End Review

Distance Between Two Distributions

Distance Between Two Distributions



Let poisson prediction be X

Let real data be Y

Three reasonable ideas

Total Variation (TV)

Loop over all possible values and calculate the ***absolute difference*** *in probability*

$$TV(X,Y) = \sum_i |P(X=i) - P(Y=i)|$$

Earth Movers (EMD)

Imagine one distribution is a ***lump of dirt****. How much work would it take to make it look just like the other?*

Solved using a Linear Program Solver

$O(n^3 \log n)$

Kullback Leibler (KL)

Expected ***excess surprise*** *from using Y as a model instead of X when the actual distribution is X.*

$$KL(X,Y) = \sum_x \log \frac{P(X=x)}{P(Y=x)} \cdot P(X=x)$$

KL Divergence Without Tears

$$\mathrm{KL}(X,Y) = \sum_{x \in X} \mathrm{ExcessSurprise}(x) \cdot P(X=x)$$

How much more surprising is x under Y than X?

$$= \sum_{x \in X} \Big[\mathrm{Surprise}(Y=x) - \mathrm{Surprise}(X=x)\Big] \cdot P(X=x)$$

Surprise!

$$= \sum_{x \in X} \left[\log_2 \frac{1}{P(Y=x)} - \log_2 \frac{1}{P(X=x)}\right] \cdot P(X=x)$$

Surprise!

$$= \sum_{x \in X} -\log_2 P(Y=x) + \log_2 P(X=x) \cdot P(X=x)$$

$1/x = x^{-1}$

$$= \sum_{x \in X} \log_2 \frac{P(X=x)}{P(Y=x)} \cdot P(X=x)$$

Log rules

People often use natural log

KL Divergence With Tears

Kullback–Leibler divergence

22 languages

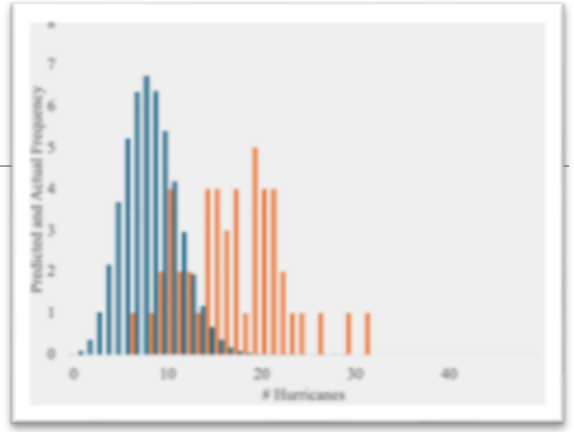
Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

In mathematical statistics, the **Kullback–Leibler** (**KL**) **divergence** (also called **relative entropy** and **I-divergence**[1]), denoted $D_{\mathrm{KL}}(P \parallel Q)$, is a type of statistical distance: a measure of how much an approximating probability distribution Q is different from a true probability distribution P.[2][3] Mathematically, it is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

KL Divergence in Code



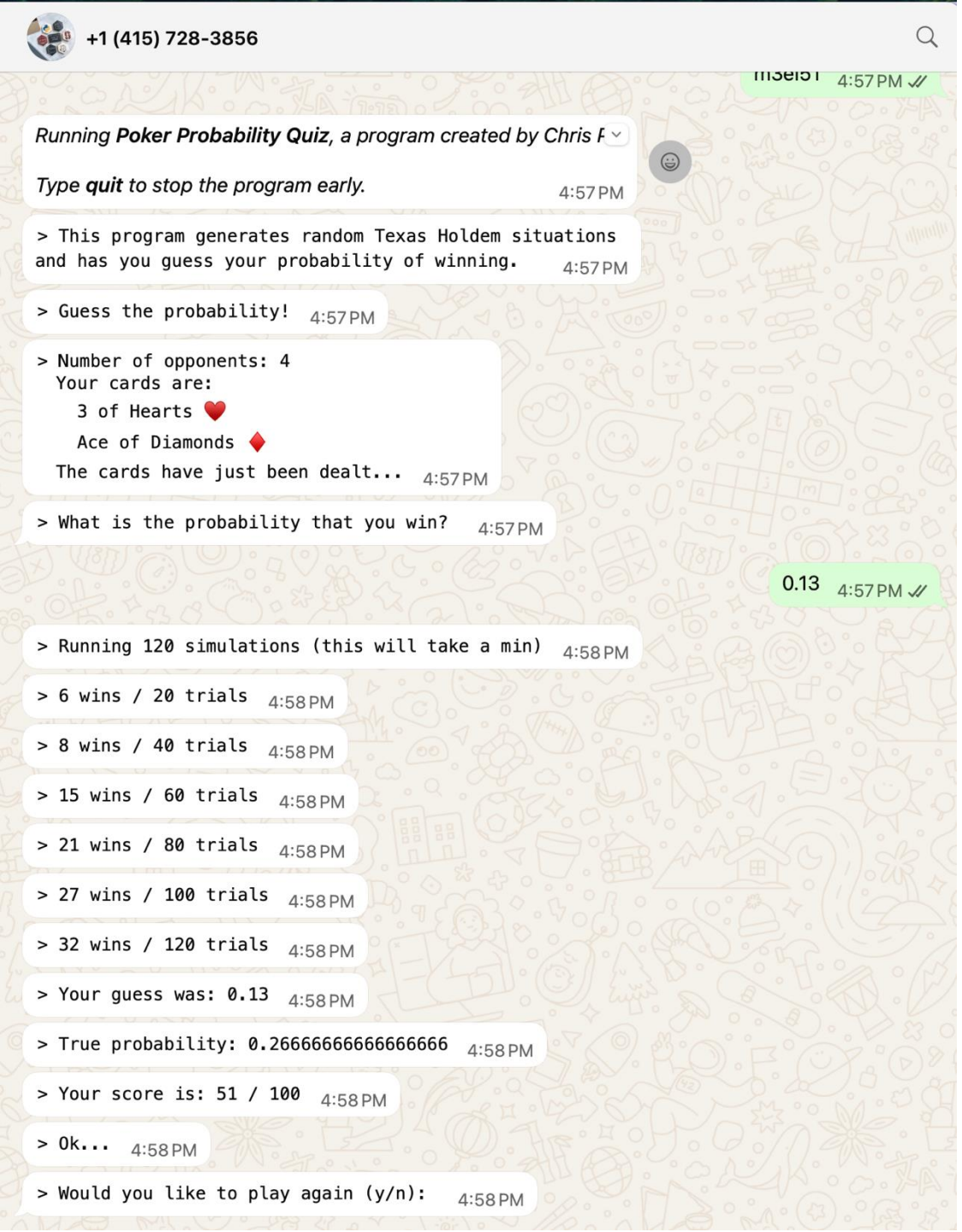
Let poisson prediction be X

Let real data be Y

```
from scipy import stats
import math

def kl_divergence(predicted_lambda, observed_pmf):
  """
  We predicted that the number of hurricanes would be
  X ~ Poisson(predicted_lambda) and observed a real world
  number of hurricanes Y ~ observed_pmf
  """
  X = stats.poisson(predicted_lambda)
  divergence = 0
  # loop over all the values of hurricanes
  for i in range(0, 40):
    pr_X_i = X.pmf(i)
    pr_Y_i = observed_pmf[i]
    excess_surprise_i = math.log(pr_X_i / pr_Y_i)
    divergence += excess_surprise_i * pr_X_i
  return divergence
```

$$KL(X, Y) \approx 0.376$$



True probability: 0.27
Guess is: 0.13
Score is: 51 / 100

What's the score function?

Send a WhatsApp message to...

+1 (415) 728-3856

With the text message...

M3EI51

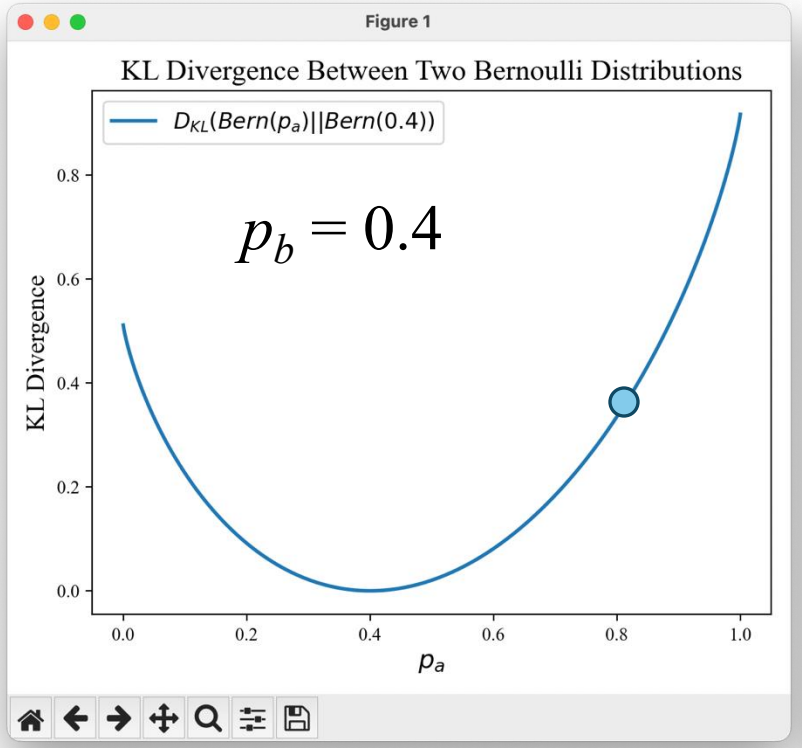
Or scan the QR code below.

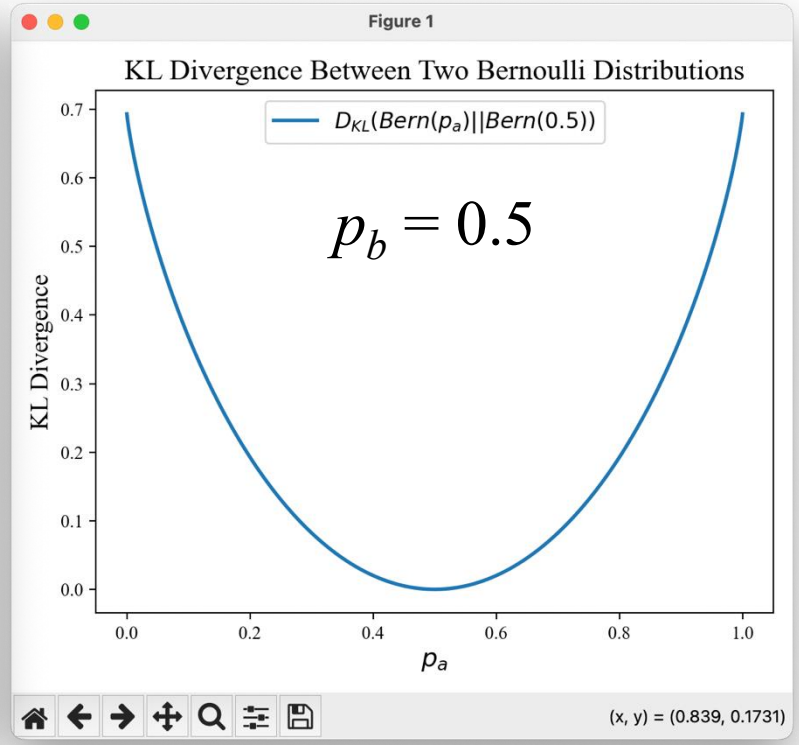
KL Divergence of Bernoulli

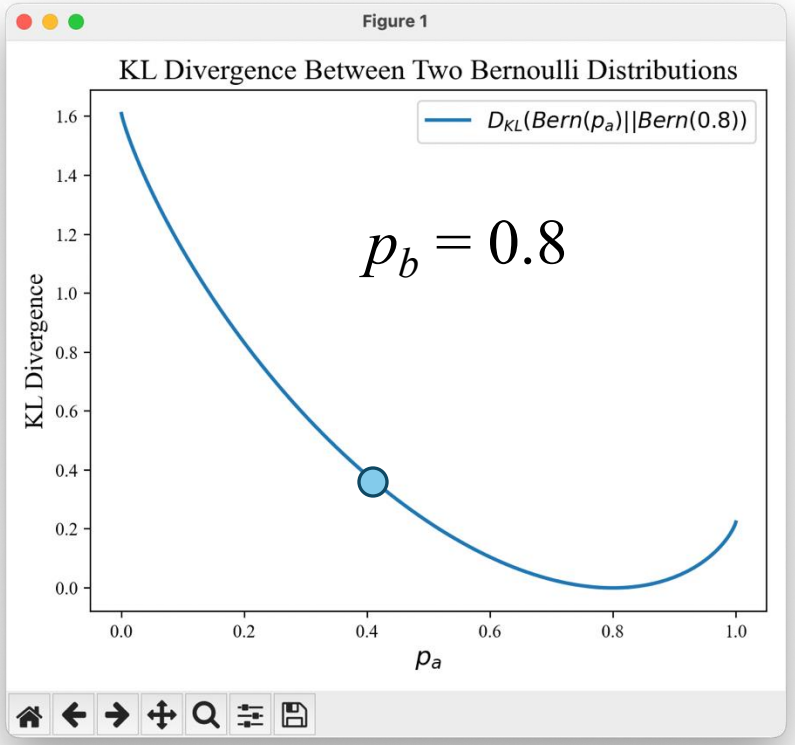
Let $X \sim \text{Bern}(p_a)$ and $Y \sim \text{Bern}(p_b)$. What is the KL Divergence between them?

$$
\begin{aligned}
KL(X,Y) &= \sum_{x \in X} \left[\log \frac{1}{P(Y=x)} - \log \frac{1}{P(X=x)} \right] \cdot P(X=x) \\
&= \left[\log \frac{1}{P(Y=1)} - \log \frac{1}{P(X=1)} \right] \cdot P(X=1) + \left[\log \frac{1}{P(Y=0)} - \log \frac{1}{P(X=0)} \right] \cdot P(X=0) \\
&= \left[\log \frac{1}{p_b} - \log \frac{1}{p_a} \right] \cdot p_a + \left[\log \frac{1}{1-p_b} - \log \frac{1}{1-p_a} \right] \cdot (1-p_a) \\
&= p_a \log \frac{p_a}{p_b} + (1-p_a) \log \frac{1-p_a}{1-p_b}.
\end{aligned}
$$

KL Divergence of Bernoulli





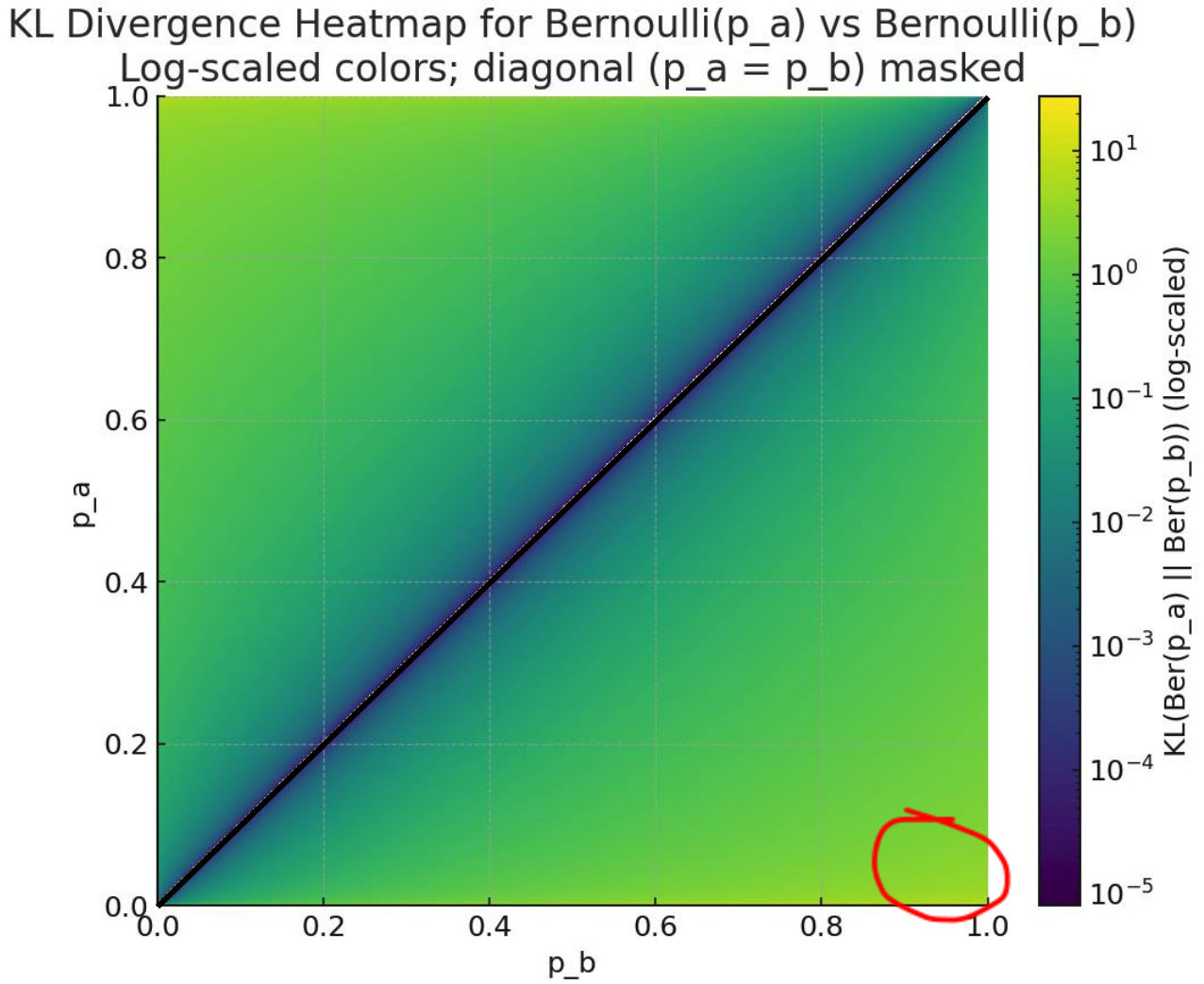


Note: not symmetric

$$D_{\mathrm{KL}}(0.8\|0.4) = 0.8\log\frac{0.8}{0.4} + 0.2\log\frac{0.2}{0.6}$$
$$= 0.8\log(2) + 0.2\log(1/3)$$
$$= 0.8(0.6931) + 0.2(-1.0986)$$
$$= 0.5545 - 0.2197 = 0.3348.$$

$$D_{\mathrm{KL}}(0.4\|0.8) = 0.4\log\frac{0.4}{0.8} + 0.6\log\frac{0.6}{0.2}$$
$$= 0.4\log(0.5) + 0.6\log(3)$$
$$= 0.4(-0.6931) + 0.6(1.0986)$$
$$= -0.2772 + 0.6592 = 0.3820.$$

KL Divergence of Bernoulli



```python
def score_guess(guess_p, true_p):
    # Avoid issues with log(0) by bounding probabilities away from 0 and 1
    epsilon = 1e-12
    guess_p = min(max(guess_p, epsilon), 1 - epsilon)
    true_p = min(max(true_p, epsilon), 1 - epsilon)

    kl_divergence = (
        true_p * math.log(true_p / guess_p)
        + (1 - true_p) * math.log((1 - true_p) / (1 - guess_p))
    )

    # normalize it to be 0 to 100
    max_score = 100
    c = 10
    score = max_score * math.exp(-c * kl_divergence)
    return int(score)
```

Indus Valley Script. A language?

The Indus civilization - one of the world's earliest urban societies - emerged 5,300 years ago. The script they used is one of the last remaining undeciphered alphabets.

We are using entropy analysis to show that the script represents a true linguistic system. Let **`all_examples`** be the list of all recorded examples of the script, where each item in the list is one example string:

```
all_examples = [' 𓀀 𓃰 𓃰 𓃱 ',
                ' 𓃱 𓀀 𓄿 𓄿 𓃰 ',
                ' 𓀀 𓄿 ',
                ...]
```

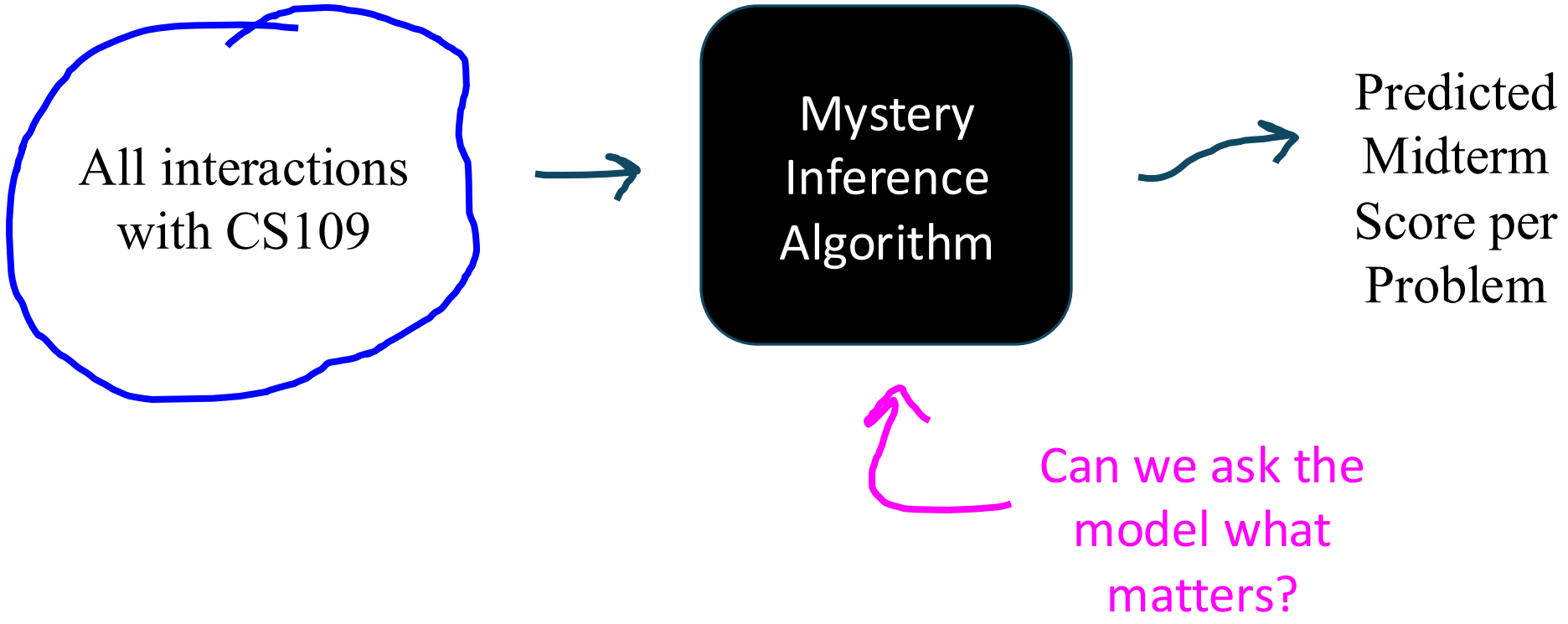
Let X be the first char
Let Y be the second char

a) What is $P(Y = y \mid X = x)$?
b) Expected reduction in surprise about Y when you learn the value of X?

Contrasting Related Concepts

Concept	Definition	Mathematical Expression
KL Divergence	Expected **excess surprise** from using Y as a model instead of X when the actual distribution is X.	$\mathrm{KL}(X,Y) = \sum_{x \in X} \Big[\text{Surprise}(Y=x) - \text{Surprise}(X=x)\Big] \cdot P(X=x)$ $= \sum_{x \in X} \log \frac{P(X=x)}{P(Y=x)} \cdot P(X=x)$
Mutual Information	**Expected reduction in surprise** about X when you learn the value of Y. Equivalently, the **KL Divergence** if you assumed X and Y were independent when in fact they are not.	$MI(X,Y) = \sum_{x,y} \big[\text{Surprise}(X=x) - \text{Surprise}(X{=}x \mid Y=y)\big]\, P(X{=}x, Y{=}y)$ $= \sum_{x,y} P(X=x, Y=y) \log \frac{P(X=x, Y=y)}{P(X=x)\,P(Y=y)}.$
Cross-Entropy	**Expected surprise** when the true outcomes come from X, but you measure surprise using Y.	$H(X,Y) = \sum_{x \in X} \text{Surprise}(Y=x) \cdot P(X{=}x)$ $= \sum_{x \in X} \frac{1}{\log P(Y=x)} \cdot P(X{=}x)$

What is the Best Way to Prepare for the Final?



Aside: No false positives. Two true positives

According to my model

feature	mutual_information_gain
PEP Inference Question	1.72
PSet App Big	1.72
Video Post PEP	0.92

RMSE of 10 points

Why mutual information and not correlation (covariance)?

Mutual Information between Strings

string_a = “this is a spam email”

string_b = “I need help storing ~~by~~ my bit coin”

What is the best summary

summary_a = “Your mom wants you to call her”

summary_b = “Your pet elephant ate all our uranium”

summary_c = “…”

email = “Hi honey, I hope you are doing well at Stanford. Things are fine but your pet elephant…”

Amazing Result

Which email does this summary describe?

Summary:
Your mom wants you to call her

Candidates:
[1] Hi honey, I hope you are doing well at Stanford. Things are fine but your pet elephant…
[2] I need duck tape and inference code. No time to explain.
[3] Meeting moved to 3pm

Return your result as a probability mass function:
{”pmf": {1:..., 2:..., 3:...}}

In her research Juliette uses this to trace student misconceptions (in English) over time as they solve open ended tasks

Representation Learning with Contrastive Predictive Coding

Neat Question

Does Information (Surprise) Integrate to 1?

$$I(E) = \log_2\left(\frac{1}{P(E)}\right)$$