Consider the following model describing how a single neuron responds to houses and faces, which is given by a pair of conditional distributions:

<table>
<thead>
<tr>
<th># spikes</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P( # spikes</td>
<td>“house”)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>P( # spikes</td>
<td>“face”)</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Furthermore, suppose \( P(\text{house}) = 0.4, P(\text{face}) = 0.6 \)

1) What is the joint distribution \( P(\# \text{ spikes, stimulus}) \)?
2) What is the marginal distribution \( P(\# \text{ spikes}) \)?

**bonus warmup problems for today:**

- \( \log(ab) = ? \)
- \( \log \left( \frac{1}{a} \right) = ? \)
Information Theory

A mathematical theory of communication,
Claude Shannon 1948
Entropy

# yes/no questions needed, on average, to determine the value of a random variable
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

Strategy 1:

Q1: is it 1?
Q2: is it 2?

Can we do better???
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1  2  3  4  5  6  7  8

Strategy 2: Q1: is it >4?
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

Strategy 2: Q1: is it >4? N Y

1,2,3,4 5,6,7,8
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

Strategy 2:  
Q1: is it >4?  
N  Y  

1,2,3,4  5,6,7,8  

Q2: is it >2?  
N  Y  N  Y  is it >6?  

1,2  3,4  5,6  7,8
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

Strategy 2:

Q1: is it >4?  

1, 2, 3, 4  

5, 6, 7, 8

Q2: is it >2?  

1, 2  

3, 4  

5, 6  

7, 8

Q3: is it 2? is it 4? is it 6? is it 8?
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

Strategy 2:  Q1: is it >4?  N  Y
                   1,2,3,4  5,6,7,8

  Q2: is it >2?  N  Y  N  Y  is it >6?
                  1,2   3,4   5,6   7,8

  Q3: is it 2?  N  Y  N  Y  N  Y  is it 8?
                      1  2   3  4   5  6   7  8

• average: 3 questions
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

General result:

\[2^N \text{ options} \implies N \text{ binary questions}\]

Therefore: \[\# \text{ questions} = \log_2(K)\]

\# possibilities (unknowns)
motivating example #2: how many Y/N questions needed?

<table>
<thead>
<tr>
<th>P(X)</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

- 3 questions would still suffice
- But can we do better?
motivating example #2: how many Y/N questions needed?

P(X)  
0 0 0 1/8 1/8 1/8 1/8 1/2 1
2 3 4 5 6 7 8

Strategy: Q1: is it 8?  
N
Y

4,5,6,7  8 (done!)
motivating example #2: how many Y/N questions needed?

<table>
<thead>
<tr>
<th>P(X)</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Strategy:

- Q1: is it 8?
  - N
  - Y

- Q2: is it > 5?
  - N
  - Y

4,5,6,7

8 (done!)
motivating example #2: how many Y/N questions needed?

<table>
<thead>
<tr>
<th>P(X)</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Strategy:

Q1: is it 8?

N  Y

Q2: is it >5?

N  Y

Q3: is it 5?

N  Y  N  Y

is it 7?

N  Y  N  Y
motivating example #2: how many Y/N questions needed?

<table>
<thead>
<tr>
<th>P(X)</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Strategy:

Q1: is it 8?

Q2: is it >5?

Q3: is it 5?

is it 7?

- what is the average # of questions?
  \[
  \frac{1}{2} \times 1 \text{ question} + \frac{1}{2} \times 3 \text{ questions} = \frac{1}{2} + \frac{3}{2} = 2 \text{ questions on average!}
  \]
motivating example #2: how many Y/N questions needed?

<table>
<thead>
<tr>
<th>P(X)</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

General results:

• optimal strategy: divide probability in half w/ each question
• need N questions to identify options with probability $1/2^N$
• thus: $\log_2(K)$ questions for options with probability $1/K$
• or: $-\log_2(p)$ questions for options with probability $p$

(using: $\log(K) = -\log(1/K) = -\log(p)$)

code length
P(X) 0 0 0 1/8 1/8 1/8 1/8 1/2

1 2 3 4 5 6 7 8

Strategy:

Q1: is it 8?

N Y

Q2: is it >5?

N Y

Q3: is it 5?

N Y N Y

• or: -\log_2(p) questions for options with probability p

Y

1/8

8

4,5,6,7

4,5 6,7

4 5 6 7

Y

0

0

0

1/8

1/8

1/8

1/8

1/2

code:

8: Y
7: NYY
6: NYN
5: NNY
4: NNN

code length
• number of “yes/no” questions needed to identify x (on average)
exercises

Compute the entropy:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1(X)$</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$P_2(X)$</td>
<td>$\frac{1}{16}$</td>
<td>0</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>$P_1(X)$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Entropy

$$H(x) = \sum_{x} p(x) \log p(x)$$

• number of “yes/no” questions needed to identify x (on average)

for distribution on K bins,

• maximum entropy = $\log K$ (achieved by uniform dist)
• minimum entropy = 0 (achieved by all probability in 1 bin)
What about when the probabilities aren’t powers of 2?

X: A B
P₁(X) 1/3 2/3

formula still applies:

\[ H(X) = -P(A) \log P(A) - P(B) \log P(B) \]
\[ = -\frac{1}{3} \log \left( \frac{1}{3} \right) - \frac{2}{3} \log \left( \frac{2}{3} \right) \]
\[ \approx 0.91 \text{ questions “on average”} \]

But how could you achieve that?

ANSWER: consider longer blocks of symbols
What about when the probabilities aren’t powers of 2?

<table>
<thead>
<tr>
<th>X:</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁(X)</td>
<td>1/3</td>
<td>2/3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/9</td>
<td>2/9</td>
<td>2/9</td>
<td>4/9</td>
</tr>
</tbody>
</table>

Formula still applies:

\[
H(X) = -P(A) \log P(A) - P(B) \log P(B)
\]

\[
= -\frac{1}{3} \log(\frac{1}{3}) - \frac{2}{3} \log(\frac{2}{3})
\]

\[
\approx 0.91 \text{ questions “on average”}
\]

= 0.9444 questions / symbol

Shannon showed: converges to entropy as you make the blocks longer.
entropy: alternate derivation

Shannon: wanted a “surprise” function $h(\cdot)$ that had two properties:

- decreasing function $p(X)$
- the surprise of independent variables adds:

$$h(p(X, Y)) = h(p(X)) + h(p(Y))$$

if

$$p(X, Y) = p(X)p(Y)$$

Only function that has this property: $h(p) = -\log(p)$

- **entropy** = “average surprise” for values from $P(X)$
Conditional Entropy

\[ H(x|y) = -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \]

- averaged over \( p(y) \)
- entropy of \( x \) given some fixed value of \( y \)

Mutual Information

is a symmetric function two random variables \( x \) and \( y \) that quantifies how many bits of information \( x \) conveys about \( y \) and vice-versa. It can be written as the entropy of \( x \) given some fixed value of \( y \) averaged over \( p(y) \).
Conditional Entropy

\[
H(x|y) = - \sum_y p(y) \sum_x p(x|y) \log p(x|y)
\]

averaged entropy of \( x \) given some fixed value of \( y \)

\[
= - \sum_{x,y} p(x, y) \log p(x|y)
\]

\[
= H(x) \quad \text{if} \quad P(x, y) = P(x)P(y)
\]

“On average, how uncertain are you about \( x \) if you know \( y \)?”
exercise

Compute the conditional entropy:

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X</td>
<td>Y=0)</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
<td>1/2</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P(X</td>
<td>Y=1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(Y)</td>
<td>2/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>