

Information Theory

Mahdi Roozbahani
Georgia Tech

Outline

- Motivation 
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Uncertainty and Information

ML

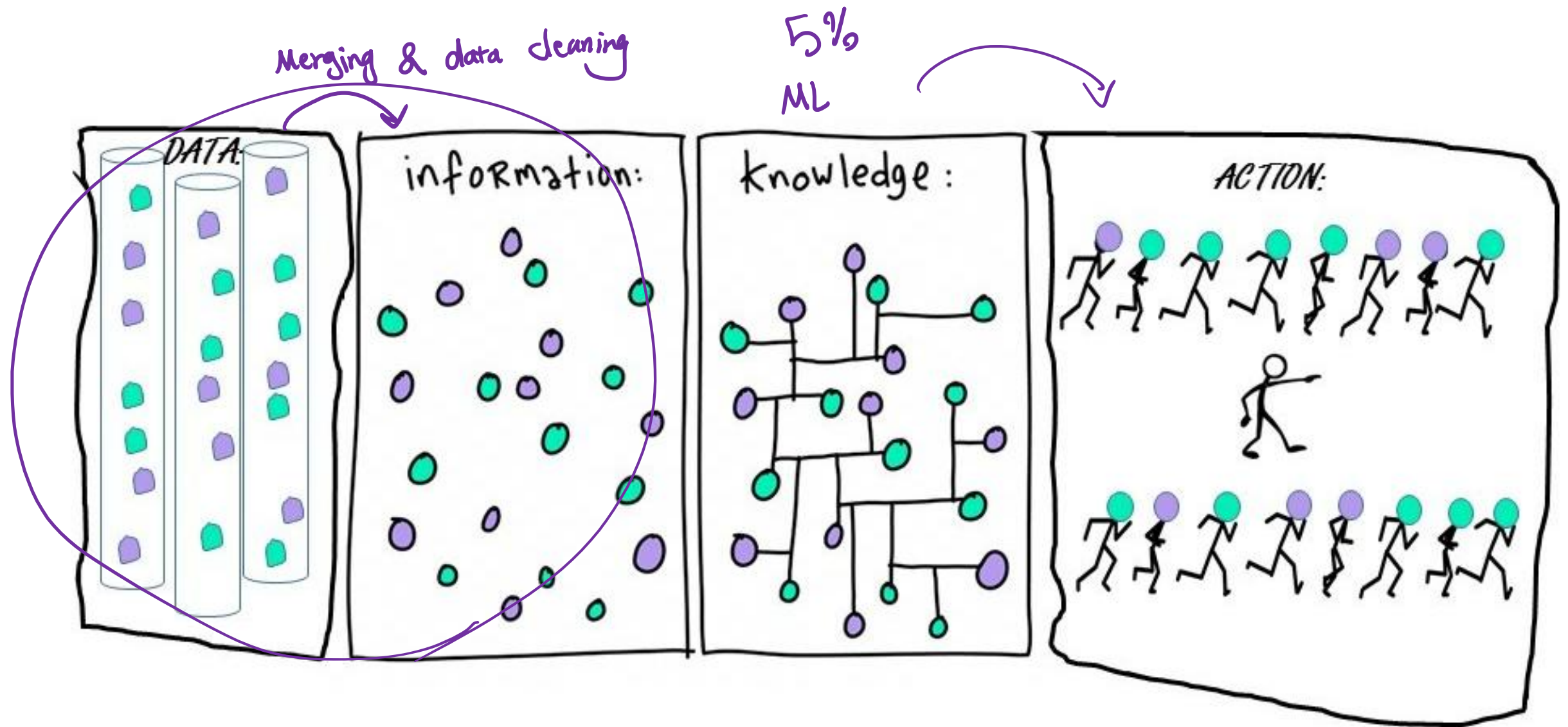
Information

Information is processed data whereas **knowledge** is **information** that is modeled to be useful.

X

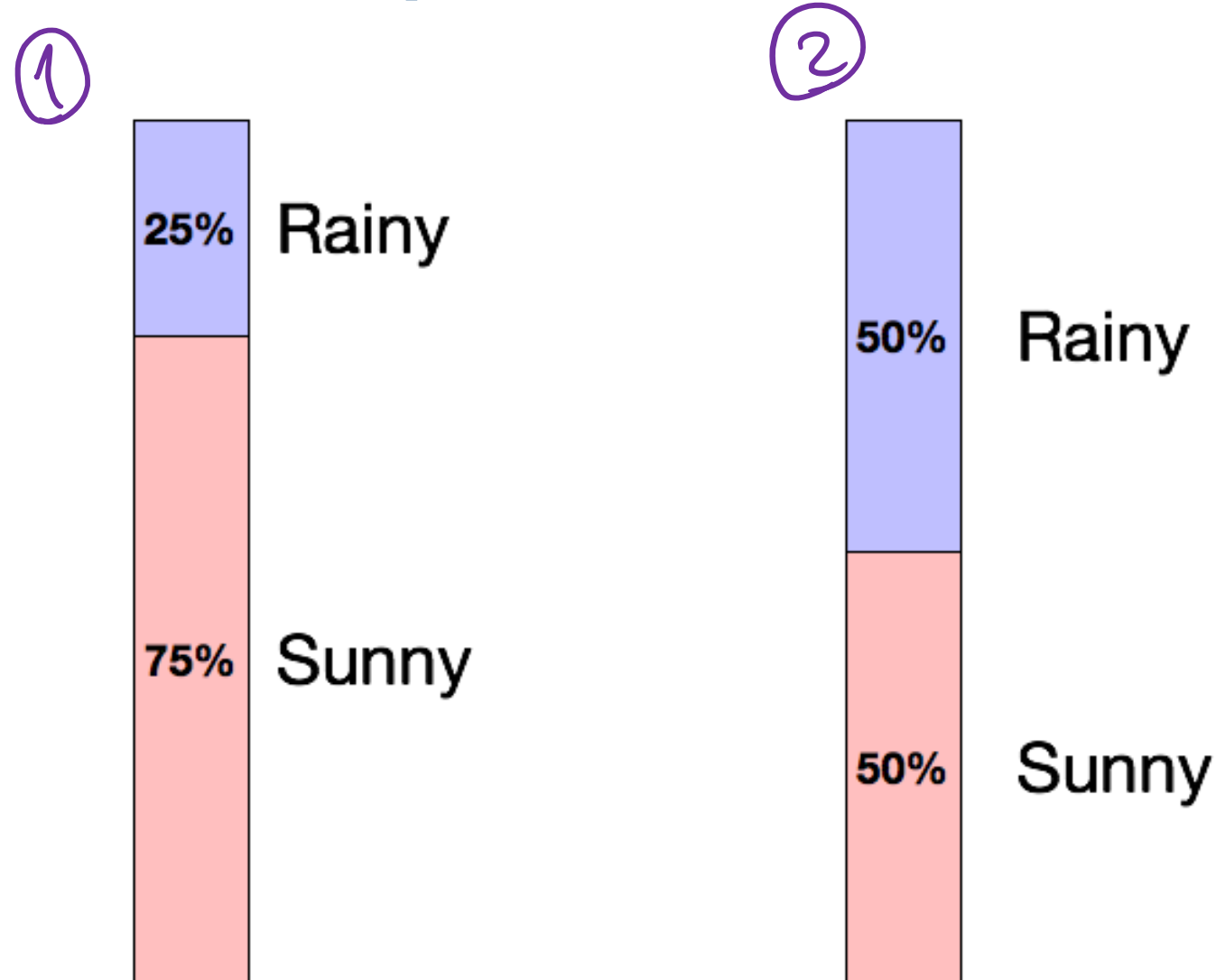
You need **information** to be able to get **knowledge**

- information $\neq$ knowledge
Concerned with abstract possibilities, not their meaning



Created by Bruce Campbell: "DIKA – ancient Chinese saying for get up and DO! Data-Information-Knowledge-Action."

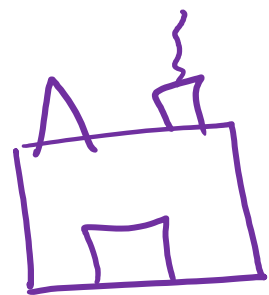
Uncertainty and Information



Which day is more uncertain?

How do we quantify uncertainty?

High entropy correlates to high information or the more uncertain

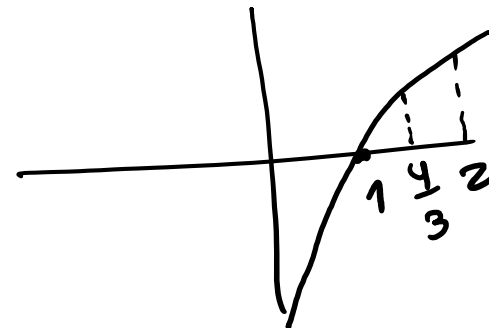


$\leadsto X = [\text{cat}, \text{cat}, \text{cat}, \text{dog}] \leadsto$

ML



frequency



$$P(X = \text{cat}) = 1 \quad I(x) = \log_2 \frac{1}{P(x)} \Rightarrow I(x = \text{cat}) = \log_2 1 = 0$$

Surprised
information

$$P(X = \text{cat}) = \frac{3}{4}$$

$$\textcircled{1} I(x = \text{cat}) = \log_2 \frac{1}{P(x = \text{cat})} = \log_2 \frac{4}{3}$$

$\textcircled{2}$

$$I(x = \text{dog}) = \log_2 \frac{1}{P(x = \text{dog})} = \log_2 2^2 = 2 \log_2 2 = 2$$

Entropy is for the system:

$$E[g(x)] = \sum g(x) P(x)$$

Entropy

$$\leftarrow H(x) = E[I(x)] = \sum_{\text{cat, dog}} I(x) P(x)$$

$$H(x) = P(X = \text{cat}) I(x = \text{cat}) + P(X = \text{dog}) I(x = \text{dog}) = \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4} \log_2 4$$

Cat, dog

$$p(x) = [0.5, 0.5]$$

$$H(x) = \frac{1}{2} \log_2 \frac{1}{\frac{1}{2}} + \frac{1}{2} \log_2 \frac{1}{\frac{1}{2}} = \frac{1}{2} \log_2^2 + \frac{1}{2} \log_2^2 = 1 = \log_2 k=2$$

$$X = [1, 2, 3, 4, 5, 6]$$

k number of sides

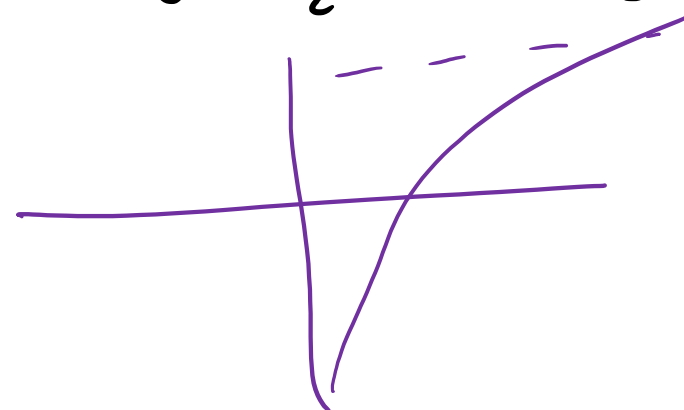
$$p(x) = [\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}]$$

$$H(x) = \frac{1}{6} \log_2 6 + \dots + \frac{1}{6} \log_2 6 = \log_2 6 = \log_2 k$$

$$X = [1, 2, 3, \dots, +\infty]$$

$$H(x) = \log_2 k \xrightarrow{+\infty}$$

$$H(x) \geq 0$$



Information

Let X be a random variable with distribution $p(x)$

$$I(X) = \log_{\{2\}} \left(\frac{1}{p(x)} \right)$$

MOTIVATION: COMPRESSION

- ▶ Suppose we observe a sequence of events:
 - ▶ Coin tosses
 - ▶ Words in a language
 - ▶ notes in a song
 - ▶ etc.
- ▶ We want to record the sequence of events in the smallest possible space.
- ▶ In other words we want the shortest representation which preserves all information.
- ▶ Another way to think about this: How much information does the sequence of events actually contain?

MOTIVATION: COMPRESSION

To be concrete, consider the problem of recording coin tosses in unary.

1 bit ← 0 00 → 2 bits
Approach 1:

T, T, T, T, H

H	T
0	00

$I(x=T)$

$I(x=H) \leadsto$ higher value

00, 00, 00, 00, 0

We used 9 characters

Which one has a higher probability: T or H?

Which one should carry more information: T or H?

MOTIVATION: COMPRESSION

To be concrete, consider the problem of recording coin tosses in unary.

T, T, T, T, H

Approach 2:

H	T
00	0

$0, 0, 0, 0, 00$

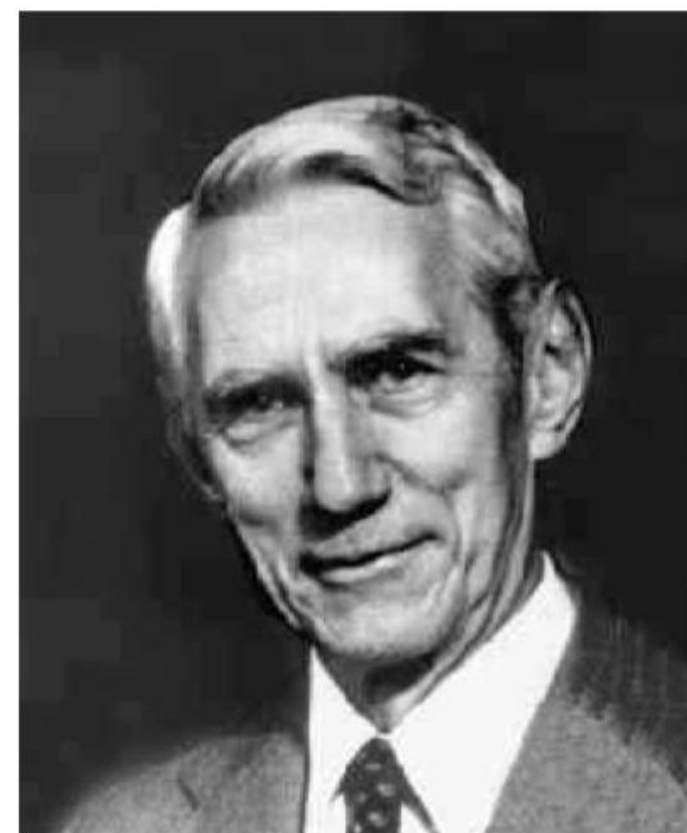
We used 6 characters

MOTIVATION: COMPRESSION

- ▶ Frequently occurring events should have short encodings
- ▶ We see this in english with words such as “a”, “the”, “and”, etc.
- ▶ We want to maximise the information-per-character
- ▶ seeing common events provides little information
- ▶ seeing uncommon events provides a lot of information

Information Theory

- Information theory is a mathematical framework which addresses questions like:
 - ▶ How much information does a random variable carry about?
 - ▶ How efficient is a hypothetical code, given the statistics of the random variable?
 - ▶ How much better or worse would another code do?
 - ▶ Is the information carried by different random variables complementary or redundant?



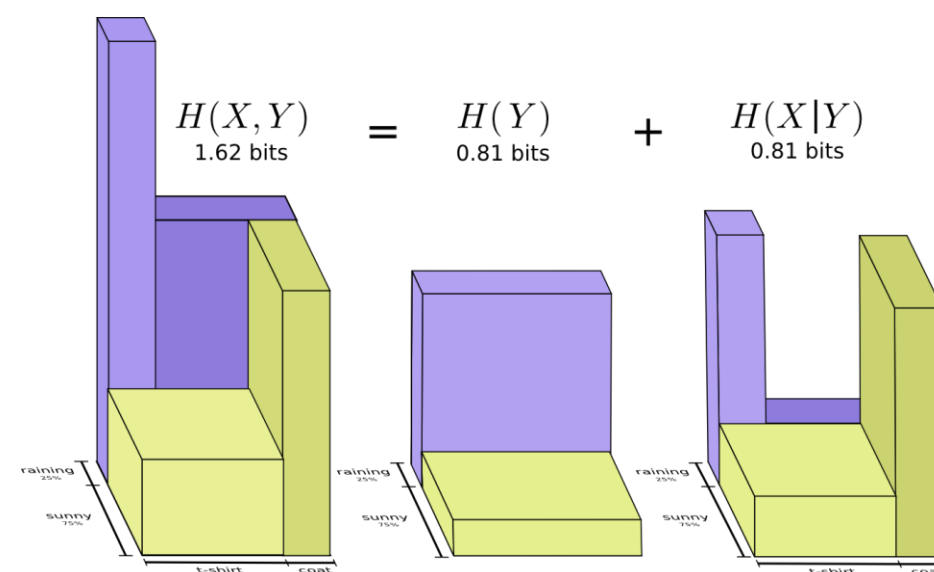
Claude Shannon

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

$$\underline{H(x, y) = H(x|y) + H(y)}$$

$$H(x, y) = H(x) + H(y)$$

$X \perp Y$



Outline

- Motivation
- Entropy 
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Entropy

- Entropy $H(Y)$ of a random variable Y

$$H(Y) = - \sum_{k=1}^K P(y = k) \log_2 P(y = k)$$

$$H(Y) = \sum P(y) \log_2 \frac{1}{P(y)} = \sum P(y) \log_2 P(y)^{-1} = - \sum P(y) \log_2 P(y)$$

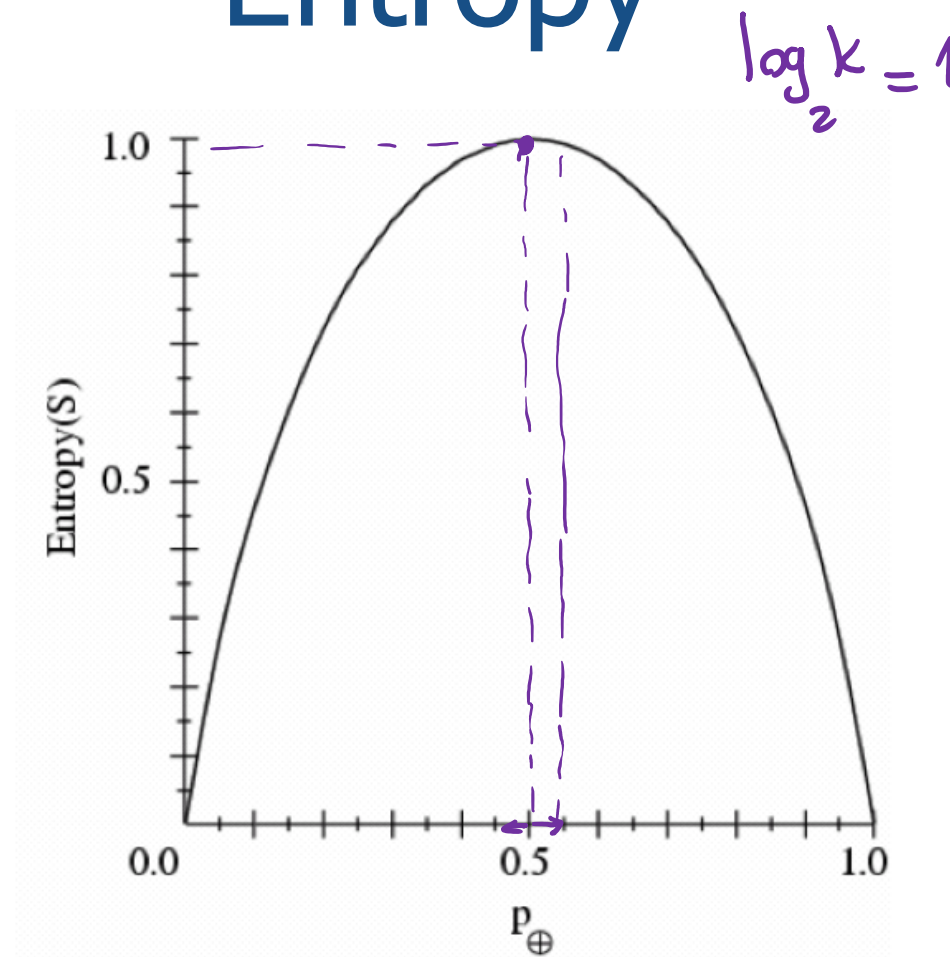
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- Information theory:

Most efficient code assigns $-\log_2 P(Y = k)$ bits to encode the message $Y = k$, So, expected number of bits to code one random Y is:

$$- \sum_{k=1}^K P(y = k) \log_2 P(y = k)$$

Entropy

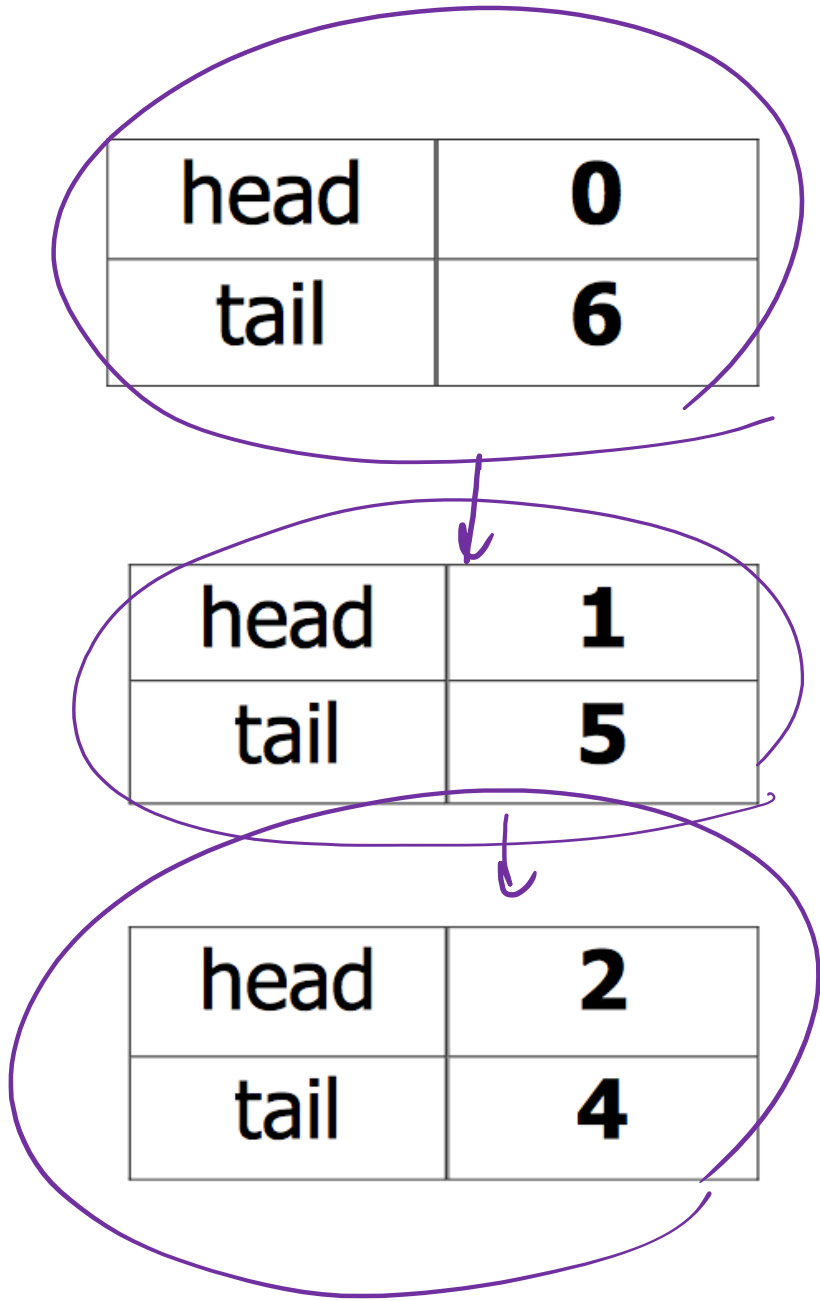


- S is a sample of coin flips
- p_+ is the proportion of heads in S
- p_- is the proportion of tails in S
- Entropy measure the uncertainty of S

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Entropy Computation: An Example

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$



head	0
tail	6

$$P(h) = 0/6 = 0. \quad P(t) = 6/6 = 1$$

$$\text{Entropy} = -0 \log 0 - 1 \log 1 = -0 - 0 = 0$$

head	1
tail	5

$$P(h) = 1/6 \quad P(t) = 5/6$$

$$\text{Entropy} = - (1/6) \log_2 (1/6) - (5/6) \log_2 (5/6) = 0.65$$

head	2
tail	4

$$P(h) = 2/6 \quad P(t) = 4/6$$

$$\text{Entropy} = - (2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$$

$$\begin{array}{cc} \text{cat} & \text{dog} \\ P = [0.8 & 0.2] \\ \uparrow & \\ Q = [0.6 & 0.4] \end{array}$$

Properties of Entropy

$$\sum p_i \log p_i^{-1} \leq \sum p_i \log q_i^{-1}$$

$$-\sum p_i \log p_i \leq -\sum p_i \log q_i \Rightarrow \sum p_i \log p_i \geq \sum p_i \log q_i$$

$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i}$$

$$\Rightarrow \sum p_i \log \frac{p_i}{q_i} \geq 0$$

$$\sum p_i \log p_i - \sum p_i \log q_i \geq 0$$

$$\sum p_i (\log p_i - \log q_i) \geq 0$$

$$\sum p_i \log \frac{p_i}{q_i} \geq 0$$

1. Non-negative: $H(P) \geq 0$

2. Invariant wrt permutation of its inputs:

$$H(p_1, p_2, \dots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \dots, p_{\tau(k)})$$

3. For any other probability distribution $\{q_1, q_2, \dots, q_k\}$:

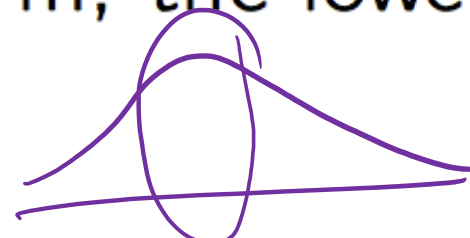
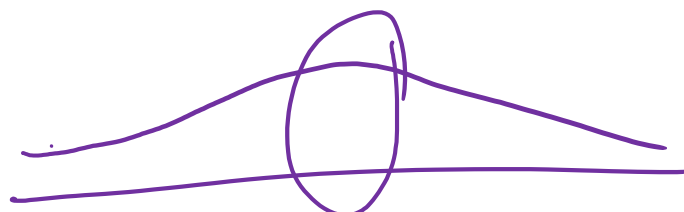
$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i} \leq \sum_i p_i \cdot \log \frac{1}{q_i}$$

KL divergence

predicted pdf

4. $H(P) \leq \log k$, with equality iff $p_i = 1/k \ \forall i$

5. The further P is from uniform, the lower the entropy.



Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information 
- Cross-Entropy and KL-Divergence

$$P(T = \text{cold}) = 0.3$$

$$P(M = \text{low}) = 0.6$$

$$P(M = \text{low}, T = \text{cold}) = 0.1$$

$$P(M = \text{low} | T = \text{mild}) = \frac{0.4}{0.5}$$

$$P(T = \text{mild} | M = \text{low}) = \frac{0.4}{0.6}$$

Joint Entropy

humidity

Temperature

	cold	mild	hot	
low	0.1	0.4	0.1	0.6
high	0.2	0.1	0.1	0.4
	0.3	0.5	0.2	1.0

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548 = -P_c \log P_c - P_m \log P_m - P_h \log P_h$
- $H(M) = H(0.6, 0.4) = 0.970951 = -0.3 \log 0.3 - 0.5 \log 0.5 - 0.2 \log 0.2$
- $H(T) + H(M) = 2.456431 = -0.6 \log 0.6 - 0.4 \log 0.4$
- **Joint Entropy:** consider the space of (t, m) events $H(T, M) = \sum_{t,m} P(T=t, M=m) \cdot \log \frac{1}{P(T=t, M=m)}$
 $H(\underline{0.1}, \underline{0.4}, \underline{0.1}, \underline{0.2}, \underline{0.1}, \underline{0.1}) = 2.32193 = -0.1 \log_2 0.1 - \dots - 0.1 \log_2 0.1$

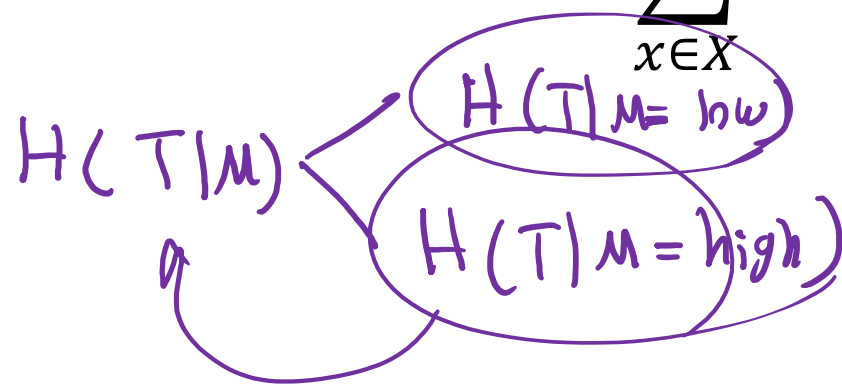
$$H(T, M) = H(T|M) + H(M) \quad \text{if } T \perp M \quad H(T) + H(M) = H(T, M)$$

Notice that $H(T, M) \leq H(T) + H(M) !!!$

$$H(T, M) = H(T|M) + H(M) = H(M|T) + H(T)$$

Conditional Entropy

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x)}{p(x, y)}$$



$$P(M = low) H(T|M = low) +$$

$$P(M = high) H(T|M = high)$$

Table showing the joint probability distribution of T and M :

	cold	mild	hot	
low	1/6	4/6	1/6	1.0
high	2/4	1/4	1/4	1.0

$$\sum_m P(M = m) H(T|M = m)$$

Conditional Entropy:

- $H(T|M = low) = H(1/6, 4/6, 1/6) = 1.25163$ $-\frac{1}{6} \log_2 \frac{1}{6} - \frac{4}{6} \log_2 \frac{4}{6} - \frac{1}{6} \log_2 \frac{1}{6}$
- $H(T|M = high) = H(2/4, 1/4, 1/4) = 1.5$

- **Average Conditional Entropy** (aka equivocation):

$$H(T/M) = \sum_m P(M = m) \cdot H(T|M = m) =$$

$$0.6 \cdot H(T|M = low) + 0.4 \cdot H(T|M = high) = 1.350978$$

Conditional Entropy

$$P(M = m|T = t)$$

	cold	mild	hot
low	1/3	4/5	1/2
high	2/3	1/5	1/2
	1.0	1.0	1.0

Conditional Entropy:

- $H(M|T = cold) = H(1/3, 2/3) = 0.918296$
- $H(M|T = mild) = H(4/5, 1/5) = 0.721928$
- $H(M|T = hot) = H(1/2, 1/2) = 1.0$
- Average Conditional Entropy (aka Equivocation):
 $H(M/T) = \sum_t P(T = t) \cdot H(M|T = t) =$
 $0.3 \cdot H(M|T = cold) + 0.5 \cdot H(M|T = mild) + 0.2 \cdot H(M|T = hot) = 0.8364528$

Conditional Entropy

- Conditional entropy $H(Y|X)$ of a random variable Y given X_i

Discrete random variables:

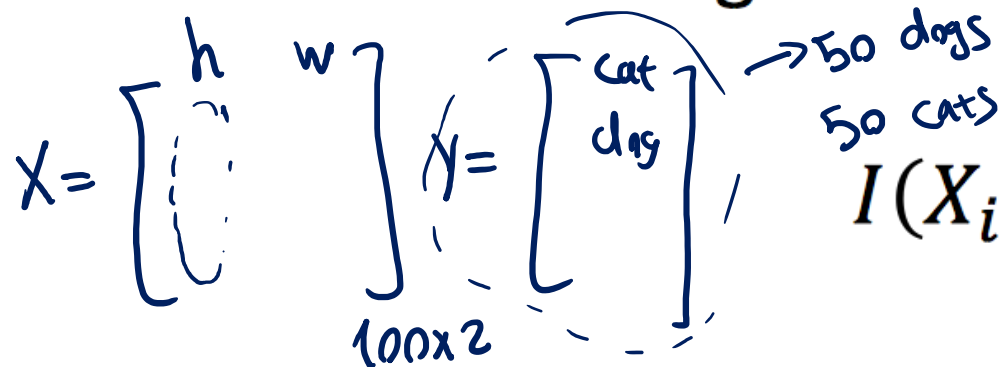
$$H(Y|X) = \sum_{x \in X} p(x_i) H(Y|X = x_i) = \sum_{x \in X, y \in Y} p(x_i, y_i) \log \frac{p(x_i)}{p(x_i, y_i)}$$

Mixed setting: Continuous (over x) and discrete (over y):

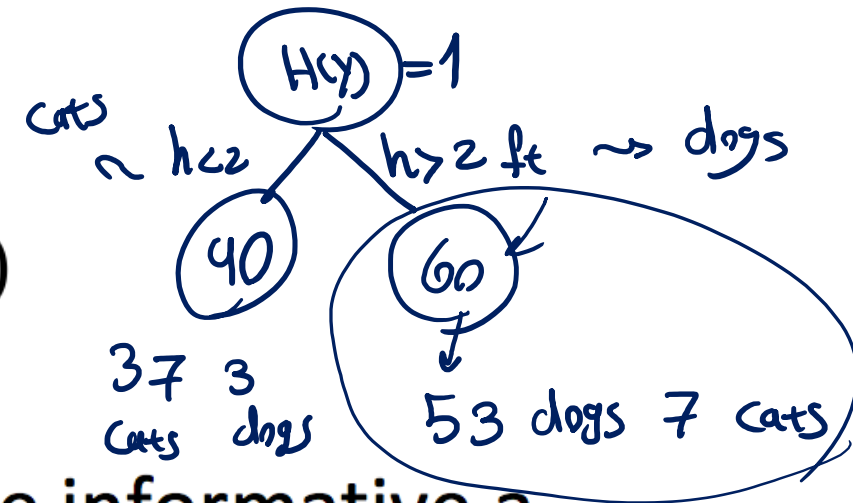
$$H(Y|X) = - \int \left(\sum_{k=1}^K p(y = k|x_i) \log_2(y = k|x_i) \right) p(x_i) dx_i$$

Mutual Information

- Mutual information: quantify the reduction in uncertainty in Y after seeing feature X_i



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



- The more the reduction in entropy, the more informative a feature.

- Mutual information is symmetric

- $I(X_i, Y) = I(Y, X_i) = H(X_i) - H(X_i|Y)$
- $I(Y|X) = \int \sum_k^K p(x_i, y = k) \log_2 \frac{p(x_i, y = k)}{p(x_i)p(y = k)} dx_i$
- $= \int \sum_k^K p(x_i|y = k)p(y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} dx_i$

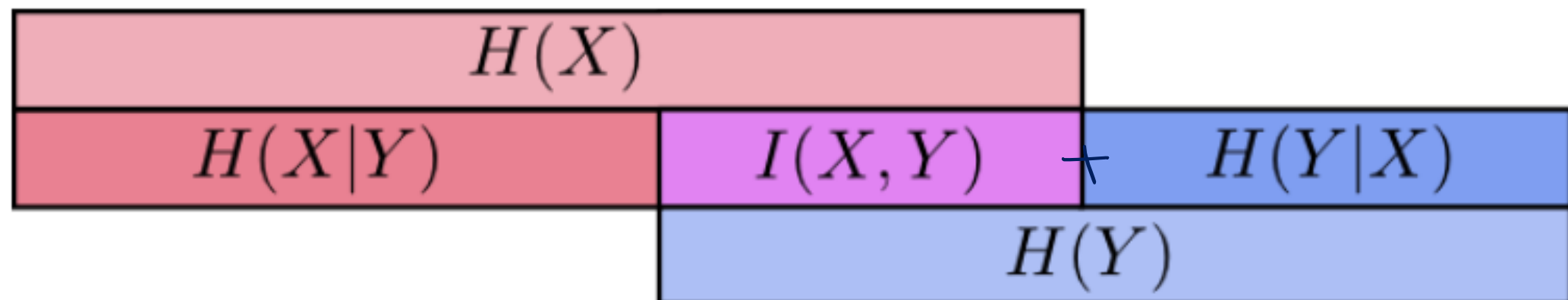
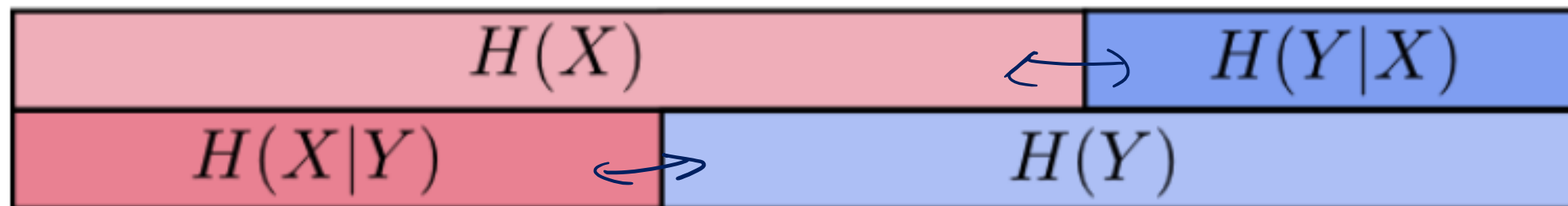
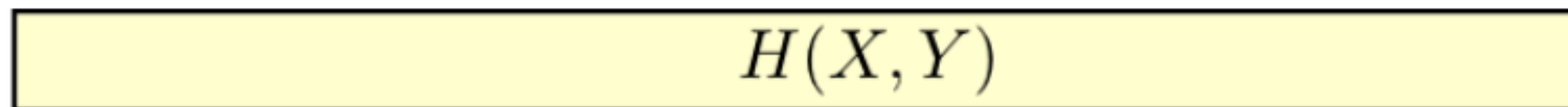
Properties of Mutual Information

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) \\ &= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x, y) \cdot \log \frac{1}{P(x|y)} \\ &= \sum_{x,y} P(x, y) \cdot \log \frac{P(x|y)}{P(x)} \\ &= \sum_{x,y} P(x, y) \cdot \log \frac{P(x, y)}{P(x)P(y)} \end{aligned}$$

Properties of Average Mutual Information:

- Symmetric
- Non-negative
- Zero iff X, Y independent

CE and MI: Visual Illustration



Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence ←



Let's work on this subject in our Optimization lecture

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

loss function $\leftarrow C_E = H(p, q) = - \sum_{x \in \mathcal{X}} p(x) \log q(x) = \underbrace{H(P)}_{\text{actual pdf}} + \underbrace{KL[P][Q]}_{\text{predicted pdf}}$

This is because:

$$\sum p(x) \log \frac{1}{q(x)} = \sum p(x) g(x) = E[g(x)] = E\left[\log \frac{1}{q(x)}\right]$$

$$H(p, q) = E_p[l_i] = E_p\left[\log \frac{1}{q(x_i)}\right]$$

$$H(p, q) = \sum_{x_i} p(x_i) \log \frac{1}{q(x_i)}$$

$$H(p, q) = - \sum_x p(x) \log q(x).$$

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\begin{aligned} \mathbf{KL}[P(S) \parallel Q(S)] &= \sum_s P(s) \log \frac{P(s)}{Q(s)} \geq 0 \\ &= \underbrace{\sum_s P(s) \log \frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P] = \overset{\text{CE}}{H(P, Q)} - H(P) \end{aligned}$$

Excess cost in bits paid by encoding according to Q instead of P .

KL Divergence is
a **KIND OF**
distance
measurement

log function is concave or convex?

$$\begin{aligned} \text{lower bound} \leftarrow \mathbf{KL}[P \parallel Q] &= \sum_s P(s) \log \frac{Q(s)}{P(s)} \xrightarrow{\text{ub}} \text{upper bound} \\ \sum_s P(s) \log \frac{Q(s)}{P(s)} &\leq \log \sum_s P(s) \frac{Q(s)}{P(s)} \quad \text{By Jensen Inequality} \\ &= \log \sum_s Q(s) = \log 1 = 0 \end{aligned}$$

So $\mathbf{KL}[P \parallel Q] \geq 0$. Equality iff $P = Q$

When $P = Q$, $KL[P \parallel Q] = 0$

Take-Home Messages

- Entropy
 - A measure for uncertainty
 - Why it is defined in this way (optimal coding)
 - Its properties
- Joint Entropy, Conditional Entropy, Mutual Information
 - The physical intuitions behind their definitions
 - The relationships between them
- Cross Entropy, KL Divergence
 - The physical intuitions behind them
 - The relationships between entropy, cross-entropy, and KL divergence