Entropy

- Entropy is the average surprise on finding out the outcome of some random variable X

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

- If $P(x)$ is the probability that the outcome is x, then $-\log P(x)$ is how surprised we would be if the outcome were x

- Since $P(x)$ ranges from 0 (for impossible events) to 1 (for certain events), the surprise ranges from infinity (for impossible events) to 0 (for certain events)

- Entropy is the weighted average of the surprise for all possible outcomes

Information theory

- A binary code for transmitting poker hands:

straight flush	0000
four of a kind	0001
full house	0010
flush	0100
straight	1000
three of a kind	0011
two pair	0101
pair	1001
high card	0111

Information theory

- An improved code, taking advantage of uneven probabilities:

straight flush	0.0000154	11111111
four of a kind	0.000240	11111110
full house	0.00144	1111110
flush	0.00196	111110
straight	0.00393	11110
three of a kind	0.0211	1110
two pair	0.0475	110
pair	0.422	10
high card	0.501	0

- Now the expected value of the message length $E[C]$ is 1.61 bits.

Relative entropy

- How many bits did we waste on average by using the wrong encoding?

$$
\begin{aligned}
E_P[(-\log_2 q(x)) - (-\log_2 p(x))] &= \sum_{x \in X} p(x)\,((-\log_2 q(x)) - (-\log_2 p(x))) \\
&= \sum_{x \in X} p(x)\,(\log_2 p(x) - \log_2 q(x)) \\
D(P||Q) &= \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}
\end{aligned}
$$

- This *relative entropy* (or *Kullback-Leibler divergence*) can be used as a measure of how closely two probability distributions agree.

- The entropy and divergence is usually measured in *bits* ($\log_2$) or *nats* ($\log_e$).

Log likelihood

- Often, we are interested in comparing how well two models q_1 and q_2 match an empirical distribution p.

- In this case, we can equivalently look at the *log likelihood*:

$$
\begin{aligned}
D(P||Q) &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) \log q(x)
\end{aligned}
$$

Minimizing $D(P||Q)$ is the same as maximizing:

$$
L_P(Q) = \sum_{x \in X} p(x) \log q(x)
$$

- Related to *cross entropy* $(-L_P(Q))$ and *perplexity* $(2^{-L_P(Q)})$.

Conditional entropy

- We can also define the joint and conditional entropy for combinations of random variables:

$$
\begin{aligned}
H(XY) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\
H(Y|X) &= \sum_{x\in X} p(x)H(Y|X=x) \\
&= \sum_{x\in X} p(x)\left(-\sum_{y\in Y} p(y|x)\log p(y|x)\right) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)
\end{aligned}
$$

- Like joint and conditional probability, joint and conditional entropy are related:

$$H(X,Y) = H(X) + H(Y|X)$$

Conditional entropy

- The *conditional entropy* $H(Y|X)$ measures the amount of uncertainty in Y after we know the value of X (on average)

$$H(Y|X) = H(XY) - H(X)$$

- Shannon's game gives us conditional entropy:

$$\begin{aligned} H(\text{letter}) &= 4.76 \\ H(\text{letter}|\text{previous letter}) &= 4.03 \\ H(\text{previous letter\&letter}) &= 8.79 \end{aligned}$$

Mutual information

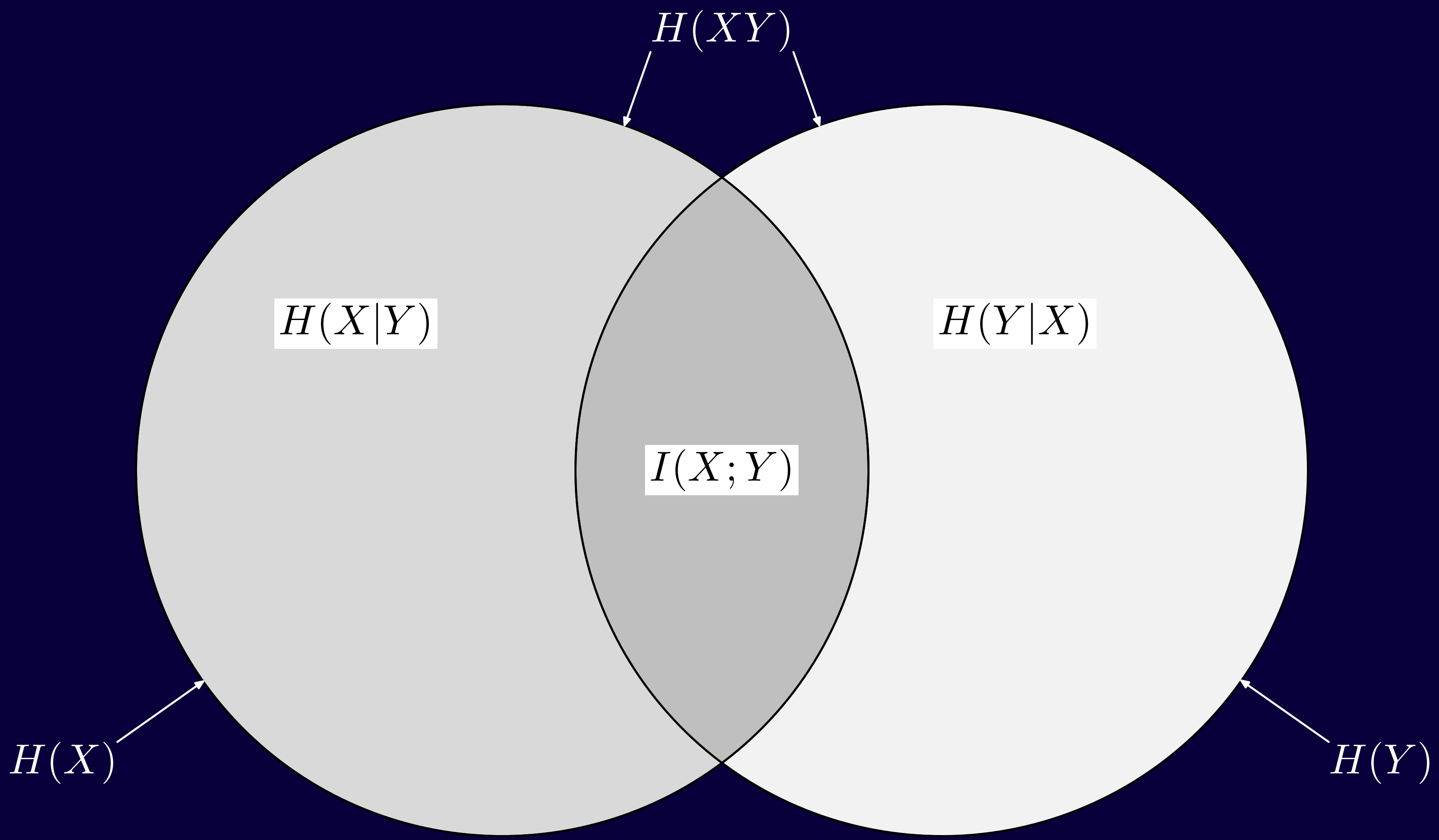
- The *average mutual information* $I(X;Y)$ measures how much knowing the value of one random variable reduces the uncertainty about another:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= I(Y;X) \\ &\geq 0 \end{aligned}$$

- $I(X;Y)$ is the expected value of:

$$I(x,y) = \log\frac{p(x,y)}{p(x)p(y)}$$

Mutual information



Mutual information

- If X and Y are independent, then $I(X;Y) = 0$.
- The average MI of dependent variables depenends on their entropy:

$$\begin{aligned} I(X;X) &= H(X) - H(X|X) \\ &= H(X) - H(X) + H(X) \\ &= H(X) \end{aligned}$$

- Average MI is also related to relative entropy:

$$I(X;Y) = D(P(X,Y)||P(X)P(Y))$$

Recall that if X and Y are independent, then $P(X,Y) = P(X)P(Y)$.

Mutual information

- Corpus linguistics, lexicography
  - $I(w_1, w_2) \gg 0$ means w_1 occurs together with w_2 *more* often than you would expect by chance
  - $I(w_1, w_2) \ll 0$ means w_1 occurs together with w_2 *less* often than you would expect by chance
  - $I(w_1, w_2) \approx 0$ means there is no evidence for a relationship between w_1 and w_2

- For word counts:

$$I(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} = \log \frac{f(w_1, w_2) \times N}{f(w_1) \times f(w_2)}$$

Mutual information

$f(ik)$	4094
$f(mijn)$	951
$f(ik, mijn)$	528
$I(ik, mijn)$	2.61

$f(ik)$	4094
$f(voor)$	4491
$f(ik, voor)$	449
$I(ik, voor)$	0.137

$f(ik)$	4094
$f(zij)$	1321
$f(ik, zij)$	60
$I(ik, zij)$	-1.00

Mutual information

-

$f(regering)$	150
$f(partij)$	133
$f(regering, partij)$	4
$I(regering, partij)$	3.17

-

$f(jongen)$	137
$f(meisje)$	112
$f(jongen, meisje)$	10
$I(jongen, meisje)$	4.88

-

$f(jong)$	95
$f(oud)$	118
$f(jong, oud)$	5
$I(jong, oud)$	4.32

Mutual information

$f(tweede)$	296
$f(kamer)$	194
$f(tweede, kamer)$	31
$I(tweede, kamer)$	4.60

$f(verenigde)$	66
$f(naties)$	16
$f(verenigde, naties)$	13
$I(verenigde, naties)$	9.11

Mutual information

Main verb	Object noun	MI	Joint freq
drink	martinis	12.6	3
drink	cup of water	11.6	3
drink	champagne	10.9	3
drink	beverage	10.8	8
drink	cup of coffee	10.6	2
drink	cognac	10.6	2
drink	beer	9.9	29
drink	cup	9.7	6
drink	coffee	9.7	12
drink	toast	9.6	4
drink	alcohol	9.4	20
drink	wine	9.3	10
drink	fluid	9.0	5
drink	liquor	8.9	4
drink	tea	8.9	5
drink	milk	8.7	8
drink	juice	8.3	4
drink	water	7.2	43
drink	quantity	7.1	4