Language model: often given by conditional probability distributions:

\[
P(X^n | x_{1:i-1} x_{i+1:N}) \quad \text{for } n = 1, 2, \ldots, N
\]

W/ joint distribution

\[
P(X^n) = P(x_1) P(x_2 | x_1) P(x_3 | x_1, x_2) \ldots P(x_n | x_{n-1})
\]

To see last lecture notes + exercise class

Arithmetic coding:

Input: \( X^N \in \mathcal{A}^N \) to compress

Algorithm:

* \( Q \leftarrow 0, \ R \leftarrow \log_2(\frac{N}{Q}) + 1 \)
* For \( n = 1, 2, \ldots, N \):
  1. \( R \leftarrow Q + p \) \( R \leftarrow \log_2(\frac{N}{Q}) + 1 \)
  2. While \( R \leq \frac{1}{2} \) or \( Q \geq \frac{1}{2} \):
     - \( b \leftarrow \begin{cases} 1 & Q \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \)
     - Write \( b \)
     - \( R \leftarrow 2R - b \)
     - \( Q \leftarrow 2Q - b \)
  3. \( p \leftarrow R - Q \)

* Write \( \lceil \log_2 \frac{2^{-p}}{p} \rceil \) bits of binary expansion of \( \frac{Q + R}{2} \)

Average rate: \( \approx \frac{H(X^N)}{N} \) for large \( N \)
Joint Entropies

Joint distribution $P(x,y) \rightarrow H(xy)$

* marginal distributions: $P(x), P(y) \rightarrow H(x), H(y)$

\[ H(x) + H(y) \geq H(xy), \text{ iff } X,Y \text{ independent} \]

* conditional distributions: $P(y|x), P(x|y)$

\[ H(y|x) = \sum_x P(x) \cdot H(y|x=x) \]

\[ H(y|x) \geq 0, = 0 \text{ iff } Y = f(X) \text{ for some function } f \]

\[ \text{Pf: } = 0 \iff H(y|x=x) = 0 \forall x \iff \forall x, \exists y: P(y|x) = 1 \]

* conditional entropy:

\[ H(y|x) = H(xy) - H(x) \]

\[ \text{Pf: } H(y|x) = \sum_x P(x) P(y|x) \log \frac{1}{P(y|x)} \]

\[ = \sum_{x,y} P(x,y) \log \frac{P(x)}{P(x,y)} = H(x) - H(x). \]

* $H(y|x) \leq H(y), \text{ iff } X, Y \text{ independent}$

\[ \text{Pf: equivalent to } H(xy) \leq H(x) + H(y) \]

Ex: $X$ coin flips of biased coin until 1st heads

\[ H(N) = ? \]

\[ X = \begin{cases} 1 & \text{if 1st outcome is heads } \text{(N=1)} \\ 0 & \text{otherwise } \text{(N>1)} \end{cases} \]
\[ H(N) = H(p_1=p_3) / p \]

**Mutual Information:**

\[ I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \]

- \( I(X;Y) \geq 0 \), \( = 0 \) iff \( X, Y \) independent
- \( I(X;Y) \leq H(X), H(Y) \)
- \( I(X;Y) = D(P_{x,y} \parallel Q_{x,y}) \), where \( Q_{x,y} = P(x) P(y) \)

Recall: Relative Entropy

\[ D(P \parallel Q) = \sum P(x) \log \frac{P(x)}{Q(x)} \in [0, \infty) \]

- \( D(P \parallel Q) < \infty \) iff \( \forall x: Q(x) = 0 \Rightarrow P(x) = 0 \)
- Gibbs Inequality: \( D(P \parallel Q) \geq 0 \), \( = 0 \) iff \( P = Q \)

**Communicating over Noisy Channels**

Source --> Encoder --x--> Noisy Channel --> Decoder

where \( X \) is the input alphabet, \( Y \) the output alphabet

Discrete memoryless channel: \( Q(Y|X) \) and \( \text{probability dist.} \)

eg. 1) Binary symmetric channel: \[
\begin{array}{c|ccc}
& 0 & 1 & \\
0 & 1-f & f & \\
1 & f & 1-f & \\
\end{array}
\]

\( Q(0|0) = Q(1|1) = 1-f \)

\( Q(0|1) = Q(1|0) = f \)
How well can we communicate over each of them?

* If we allow no errors at all:  ① ② ③ ④
- Any \( y \) could come from either \( x \)
- \( y = 0 \) can come from any \( x \) (if sending 0 all the time is not informative)
- \( y = T \) can come from either \( x \)
- \( \) encode: \( 0 \rightarrow B \) \( \) decode: \( A, B, C \rightarrow 0 \) \( D, E, F \rightarrow 1 \)

* If we allow error: Can use Bayes' theorem to infer most likely \( x \):

\[
P(x|y) = \frac{Q(y|x)P(x)}{\sum Q(y|x')P(x')}
\]

assuming \( x \) come from some ensemble

For reliable communication, consider block encodings:

WANT: \( S = S' \) with high probability
Shannon's Noisy Coding Theorem (Informal): The "optimal" rate at which we can communicate reliably" is given by the capacity of the channel $C(y|x)$:

$$C(y|x) = \max_{P(x)} I(X; Y)$$

for $P(x|y) = P(x) \cdot Q(y|x)$

E.g., for the binary symmetric channel:

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

$$= H(Y) - H(\{f, 1-f\})$$

subject to $P$

$$\max_P H(f) = 1$$

since $P(Y=0) = f \cdot P + (1-f) \cdot P = \frac{1}{2}$ if $p = \frac{1}{2}$

$$\implies C(y|x) = \max_{P} I(X; Y) = 1 - H(\{f, 1-f\}) = \begin{cases} 0 & \text{if } f = \frac{1}{2} \\ 1 & \text{if } f = 0 \text{ or } f = 1 \end{cases}$$