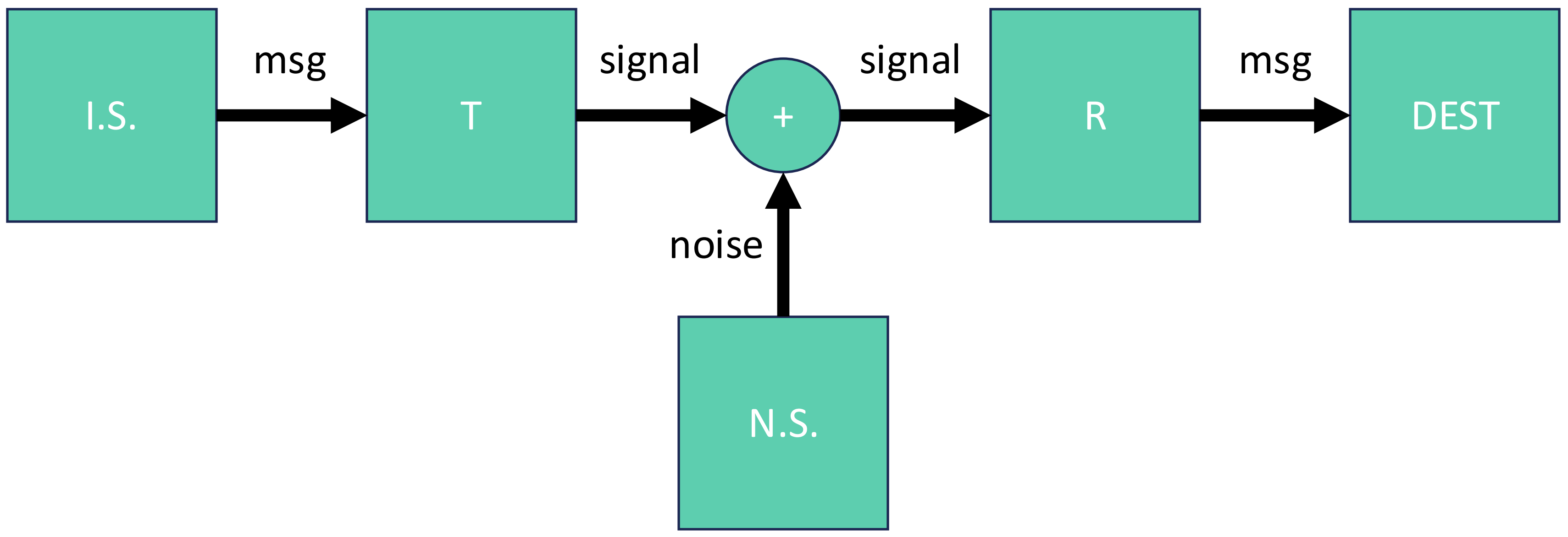
Information sources and measures

Daniel Hsu

COMS 6998-7 Spring 2025

Discrete information sources

Communication systems (Shannon, 1948)



Discrete information source

- Finite alphabet Σ
- A **discrete information source** is a stochastic process $(X_t)_{t \in \mathbb{N}}$ where each X_t has range Σ
  - Regard t as "position" or "time"
  - X_t is t-th symbol in message
- Simplifying assumption: stochastic process is stationary
  e.g., (X_1, X_3, X_5) has same distribution as (X_{11}, X_{13}, X_{15})
  - Still arbitrarily complicated; no finite description

Tractable approximation: $\boldsymbol{n}$-gram model

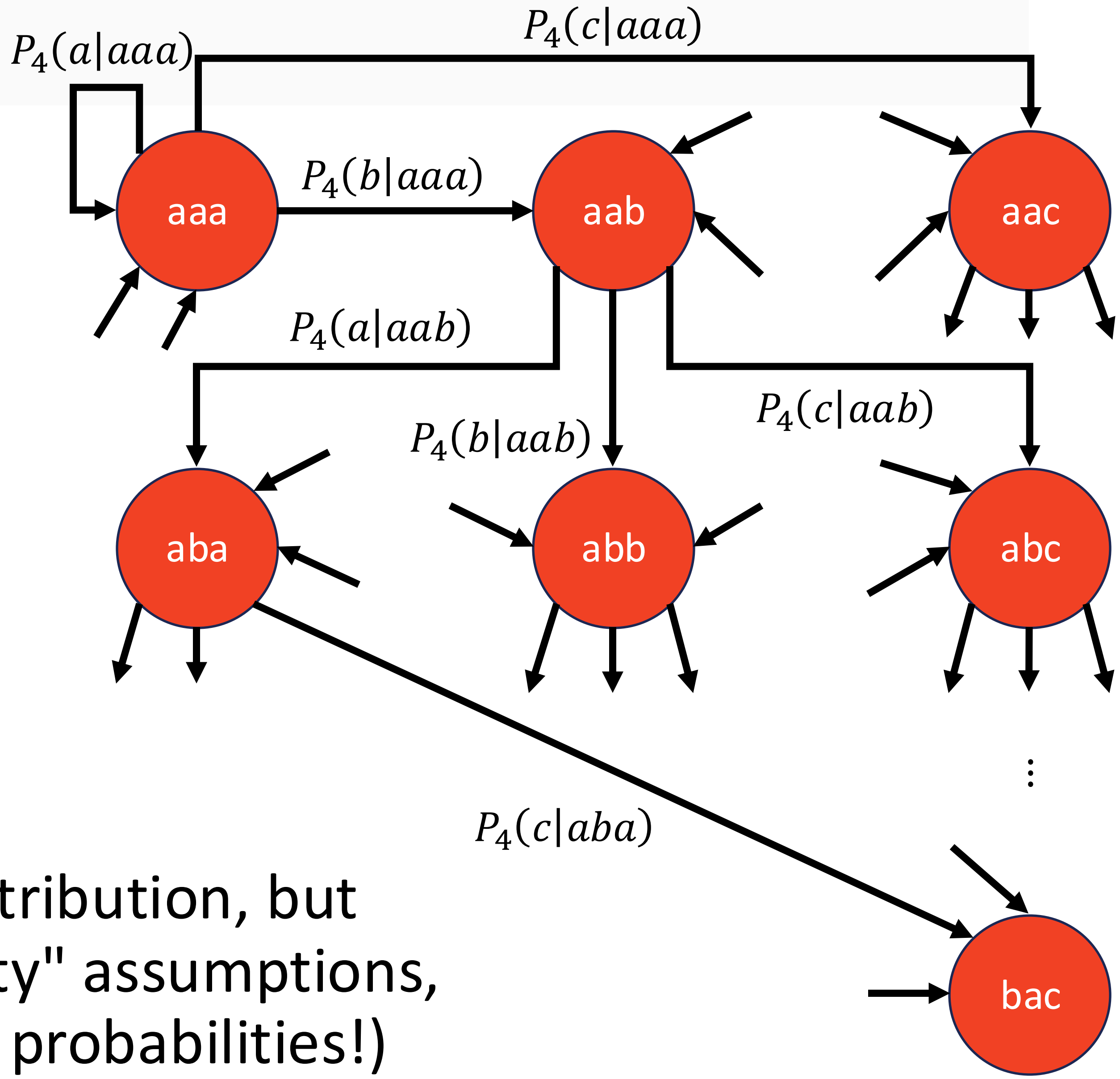
- Let P_n be law for symbols at n consecutive positions (for true I.S.)
  - $P_n(x_1, \dots, x_n) \geq 0$
  - $\sum_{x_1,\dots,x_n} P_n(x_1, \dots, x_n) = 1$
- For $n' < n$, can get $P_{n'}$ from P_n by marginalization
- Conditional law for n-th symbol given first $n-1$ symbols:

$$P_n(x_n | x_1, \dots, x_n) = \frac{P_n(x_1, \dots, x_n)}{P_{n-1}(x_1, \dots, x_{n-1})}$$

- The **$\boldsymbol{n}$-gram model** is the Markov chain over state space Σ^{n-1} defined in the "natural" way using this conditional law…

n-gram Markov chain

- Transition probability from state $(x_1, \dots, x_{n-1})$ to $(x_2, \dots, x_n)$ is $P_n(x_n | x_1, \dots, x_{n-1})$



(Should also define initial state distribution, but under "ergodicity" and "stationarity" assumptions, it is uniquely defined by transition probabilities!)

Specification of n-gram model

- For each state, specify transition probs. for $|\Sigma|$ possible next-states
- Meaning:
  - $n = 1$: remember nothing (no states)
  - $n = 2$: only remember last symbol ($|\Sigma|$ states)
  - $n = 3$: only remember last two symbols ($|\Sigma|^2$ states)

Generation based on n-gram model

Given $x_1, \dots, x_T$, generate next symbol according to n-gram model

- Only look at last $n-1$ symbols $x_{T-(n-2)}, \dots, x_T$
- Sample x_{T+1} according to conditional law
$$P_n\left(\cdot \,\middle|\, x_{T-(n-2)}, \dots, x_T\right)$$
- Same as starting in Markov chain state $x_{T-(n-2)}, \dots, x_T$ and taking one step

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

Prediction based on n-gram model

Given $x_1, \dots, x_T$, predict next symbol according to n-gram model

- Only look at last $n-1$ symbols $x_{T-(n-2)}, \dots, x_T$
- Predict x_{T+1} to be

$$\mathrm{argmax}_{x \in \Sigma} P_n\left(x \middle| x_{T-(n-2)}, \dots, x_T\right)$$

Measuring information

How much "information" is in a source?

- How much information is in length T sequences from the I.S.?
- Hartley (1928): it's related to the number of possible messages, and the logarithm of that number is "natural"

$$\log(|\Sigma|^T)$$

- But this doesn't take into account relative frequencies of messages

Shannon's entropy

- Let N be total number of possible messages (e.g., $N = |\Sigma|^T$)
- Let $p = (p_1, \ldots, p_N)$ where p_i is probability of i-th possible message
- **Entropy** of p (or **entropy** of $X \sim p$):

$$H(p) = H(X) = \sum_{i=1}^{N} p_i \log \frac{1}{p_i}$$

- Shannon derived this formula by:
  - Writing down some axioms that any reasonable measure of information should satisfy
  - Deriving the formula as the only possible formula that satisfies the axioms

Example

- $H\big((0.5,0.5)\big) = 0.5\log 2 + 0.5\log 2 = \log 2$
- $H\big((0.25,0.75)\big) = 0.25\log 4 + 0.75\log 1.333\ldots < \log 2$
- $H\big((\epsilon, 1-\epsilon)\big) = \epsilon\log\frac{1}{\epsilon} + (1-\epsilon)\log\frac{1}{1-\epsilon} = \frac{\log 1/\epsilon}{1/\epsilon} + (1-\epsilon)\log\frac{1}{1-\epsilon}$
- Uniform distribution on N possible messages:

$$H(p) = \sum_{i=1}^{N} p_i \log\frac{1}{p_i} = \log N$$

Axioms from Aczel, Forte, Ng (1974)

- Expansible: $H\big((p_1, \ldots, p_N)\big) = H\big((p_1, \ldots, p_N, 0)\big)$
- Symmetric: unaffected by relabeling the messages
- Additive: If X and Y are independent, then $H(X, Y) = H(X) + H(Y)$
- Subadditive: $H(X, Y) \leq H(X) + H(Y)$
- Small for small probabilities: $\lim_{\epsilon \to 0^+} H\big((\epsilon, 1 - \epsilon)\big) = 0$
- Normalized: $H\big((0.5, 0.5)\big) = \log 2$

Only Shannon's entropy satisfies these axioms!

Additivity

$$H(X,Y) = \sum_{x,y} p_{x,y} \log \frac{1}{p_{x,y}}$$

$$= \sum_{x,y} p_{x,y} \log \frac{1}{p_x p_y}$$

$$= \sum_{x,y} p_{x,y} \left(\log \frac{1}{p_x} + \log \frac{1}{p_y} \right)$$

$$= \sum_{x} p_x \log \frac{1}{p_x} + \sum_{y} p_y \log \frac{1}{p_y} = H(X) + H(Y)$$

Properties of entropy

- $H(p) \geq 0$
  - $H(p) = 0$ iff $X \sim p$ is a constant
- $H(p) \leq \log N$ for all $X \sim p$ on N possible values
  - $H(p) = \log N$ iff p is uniform distribution
- $H(p)$ is (strictly) concave function of p
- If A is doubly-stochastic $N \times N$ matrix, then $H(p) \leq H(Ap)$
  - Here we regard p as an N-vector
  - $H(p) = H(Ap)$ iff A is a permutation matrix

Conditional entropy

Define **conditional entropy** $H(Y|X)$ to be average of conditional distribution of Y given $X = x$ for each x but weighted by p_x:

$$H(Y|X) = \sum_x p_x \sum_y p_{y|x} \log \frac{1}{p_{y|x}}$$
$$= \sum_{x,y} p_{x,y} \log \frac{1}{p_{y|x}}$$

- On average (over X), how much information is in Y when X is known
- If X and Y are independent, then $p_{y|x} = p_y$ and $p_{x,y} = p_x p_y$, so
$$H(Y|X) = H(Y)$$

Another intuitive way to think about conditional entropy

$$H(Y|X) = \sum_{x,y} p_{x,y} \log \frac{1}{p_{y|x}}$$

$$= \sum_{x,y} p_{x,y} \log \frac{p_x}{p_{x,y}}$$

$$= \sum_{x,y} p_{x,y} \log \frac{1}{p_{x,y}} - \sum_{x,y} p_{x,y} \log \frac{1}{p_x}$$

$$= H(X,Y) - H(X)$$

- How much information of (X,Y) is left after taking out that from X

Effect of conditioning

- Using subadditivity of entropy,

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

- Therefore,

$$H(Y) \geq H(Y|X)$$

- Conditioning can only reduce entropy

Regarding (conditional) entropy as expected values

- If $X \sim p$, then

$$H(X) = \mathbb{E}\left[\log \frac{1}{p_X}\right]$$

- If $(X, Y) \sim p$, then

$$H(Y|X) = \mathbb{E}\left[\log \frac{1}{p_{Y|X}}\right]$$

Entropy rate of a source

Entropy rate

- Consider information source $(X_t)_{t\in\mathbb{N}}$
  - Shorthand: $X_{1:T} = (X_1, \dots, X_T)$
  - If symbols are IID, then
$$H((X_{1:T})) = T \cdot H(X_1)$$
    - Grows linearly with sequence length
- **Entropy rate** (i.e., per-symbol entropy):
$$\frac{1}{T} H\big((X_{1:T})\big)$$
  - Interested in this for large T (or limit $T \to \infty$)
    - Larger T means more "structure" about I.S. is captured
  - Easy upper bound for all sources: $\log|\Sigma|$
    - E.g., for $\Sigma = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \dots, \mathrm{z}\}$, upper bound is ≈ 4.7

Limiting entropy rate

- Conditional entropy of a symbol given the preceding symbols:
$$F_k \coloneqq H(X_k|X_{1:k-1}) = H(X_{1:k}) - H(X_{1:k-1})$$
  - Can write $H(X_{1:T})$ as telescoping sum:
  $$H(X_{1:T}) = F_1 + F_2 + \cdots + F_T$$
  so **entropy rate is average of these conditional entropies**
- By stationarity and "conditioning can only reduce entropy":
$$H(X_1) = H(X_2) \geq H(X_2|X_1)$$
  - Therefore
  $$F_1 \geq F_2 \geq \cdots \geq F_T$$
- By monotone convergence, there's a limit: $F_\infty = \lim_{T\to\infty} F_T$
- So $H(X_{1:T})/T \geq F_\infty$ as well

Using the n-gram approximation

- Shannon (1948, 1951) wanted to know entropy rate of printed English
  - But this information source is too unwieldy
  - What if we use an n-gram approximation?
- Let $(X_t)_{t\in\mathbb{N}}$ be stochastic process that describes printed English
- Let $(Y_t)_{t\in\mathbb{N}}$ be stochastic process governed by n-gram approximation
- Question:

$$\frac{1}{T}H(X_{1:T}) \approx \frac{1}{T}H(Y_{1:T})?$$

- For $k \leq n$, law of k consecutive Y_t's is same as that for X_t's

$$H(Y_{1:k}) - H(Y_{1:k-1}) = H(X_{1:k}) - H(X_{1:k-1}) = F_k$$

- What about $k > n$?

Where does the n-gram approximation go wrong?

- For $k > n$: Y_k only depends on previous $n - 1$ symbols

$$H(Y_k | Y_{1:k-1}) = H\left(Y_k \middle| Y_{k-(n-1):k-1}\right) = F_n$$

- If we write entropy rate of n-gram approximation as average of conditional entropies (just like before), we get (for $T \geq n$)

$$\frac{1}{T} H(Y_{1:T}) = \frac{1}{T}(F_1 + F_2 + \cdots + F_n + \underbrace{F_n + \cdots + F_n}_{T - n \text{ times}})$$

- Since $F_1 \geq F_2 \geq \cdots \geq F_T$, we also have

$$\frac{1}{T} H(Y_{1:T}) \geq \frac{1}{T}(F_1 + F_2 + \cdots + F_n + F_{n+1} + \cdots + F_T) = \frac{1}{T} H(X_{1:T})$$

- So n-gram approximation's entropy rate is an upper-bound

Does the n-gram approximation work in the "limit"?

$$\frac{1}{T}H(Y_{1:T}) = \frac{1}{T}(F_1 + F_2 + \cdots + F_n + F_n + \cdots + F_n)$$
$$\leq F_n + \frac{n(F_1 - F_n)}{T}$$

- If we take $T \to \infty$ and then $n \to \infty$, we have
$$\frac{1}{T}H(Y_{1:T}) \to F_\infty$$
- By sandwiching, F_∞ is (true) limiting entropy rate
- Shannon opts to simply use (estimate of) F_n as an upper-bound on F_∞

Entropy rate of printed English

- Plug-in existing frequency tables for 1-grams, 2-grams, 3-grams:

$$F_n = H(P_n) - H(P_{n-1})$$

F_0	F_1	F_2	F_3
4.7	4.14	3.56	3.3

- No n-gram tables for $n > 3$, but have word frequency tables
  - Estimate: k-th most frequent word in English has frequency $0.1/k$ (Zipf's law)
  - To have this normalize properly, only consider 12366 words
  - Plug-in to formula to get entropy of English word distribution:

$$9.72$$

  - Average English word has 4.5 letters, so get per-letter entropy estimate:

$$\frac{9.72}{4.5} = 2.16$$

Shannon's prediction game

- Thesis: English speakers implicitly know/use distribution of English
- Game (simple version):
  - Choose passage of English text $x_{1:T}$
  - For $t = 1, \ldots, T$:
    - Speaker guesses x_t
    - If correct, tell the speaker to record a null symbol █
    - Else, reveal x_t to speaker
- Example run of the game:

THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG READING LAMP ON THE DESK SHED

████ROO██████NOT█V█████I██████SM████OBL████REA███████
███O██████D████SHED█

Redacted sequence is a perfect encoding

- Theorem: For any sequence $x_{1:T}$, resulting "redacted" sequence produced by in game has same information as original sequence
- Proof: Can perfectly recover any redacted symbol x_t using the Speaker and redacted prefix
- (Shannon also has another version of the game based on ranks)

ML version of Shannon's game

- Construct $\mathrm{NN}\colon \Sigma^* \to \Delta(\Sigma)$ (say, using deep learning)
  - Write $\mathrm{NN}(x|x_{1:t-1})$ for probability assigned to x by evaluating NN on $x_{1:t-1}$
- Choose passage of English text $x_{1:T}$ (not used in construction of NN)
- For $t = 1, \ldots, T$:
  - Record **log-loss** of NN on t-th symbol x_t

$$\log \frac{1}{\mathrm{NN}(x_t|x_{1:t-1})}$$

- Average log-loss on entire passage $x_{1:T}$:

$$\frac{1}{T}\sum_{t=1}^{T} \log \frac{1}{\mathrm{NN}(x_t|x_{1:t-1})}$$

Why log-loss?

- Let p be probability distribution, and $X \sim p$
- Let q be another probability distribution
- Then

$$\mathbb{E}\left[\log \frac{1}{q_X}\right] = \sum_x p_x \left(\log \frac{p_x}{q_x} + \log \frac{1}{p_x}\right) = \underbrace{\sum_x p_x \log \frac{p_x}{q_x}}_{\mathrm{RE}(p,q)} + H(p)$$

  - … where $\mathrm{RE}(p,q)$ is **relative entropy** (a.k.a. Kullback-Leibler (KL) divergence) from p to q
  - **Gibb's inequality**: $\mathrm{RE}(p,q) \geq 0$ with equality iff $p = q$
- So, as function of q, expected log-loss is minimized by $q = p$

Why average los-loss over sequence?

- Log loss on t-th symbol X_t, in expectation:

$$\mathbb{E}\left[\log\frac{1}{\mathrm{NN}(X_t|X_{1:t-1})}\right] = \mathbb{E}\left[\mathbb{E}\left[\log\frac{1}{\mathrm{NN}(X_t|X_{1:t-1})}\middle|X_{1:t-1}\right]\right]$$
$$\geq \mathbb{E}\left[\mathbb{E}\left[\log\frac{1}{P_t(X_t|X_{1:t-1})}\middle|X_{1:t-1}\right]\right]$$
$$= H(X_t|X_{1:t-1}) = F_t$$

- So average log-loss, in expectation, is

$$\geq \frac{1}{T}\sum_{t=1}^{T} F_t = \frac{1}{T}H(P_T)$$

- **Average log-loss, in expectation, gives upper-bound on entropy rate**

Recap

- Entropy rate of source – per symbol entropy over long sequences
- Can upper-bound using:
  - Entropy rate of n-gram approximations
  - Average of log-losses in sequential prediction (in expectation)

Proof of Gibbs' inequality

- Let $X \sim p$, and by (strict) convexity of negative logarithm:

$$\mathrm{RE}(p,q) = \mathbb{E}\left[\log \frac{p_X}{q_X}\right] = \mathbb{E}\left[-\log \frac{q_X}{p_X}\right]$$

$$\geq -\log\left(\mathbb{E}\left[\frac{q_X}{p_X}\right]\right)$$

$$= -\log\left(\sum_x q_x\right)$$

$$= -\log(1) = 0$$

- Equality holds iff q_X/p_X is constant function (i.e., $p = q$)