ECE 587 / STA 563: Lecture 4 – Entropy Rates & Markov Chains

Information Theory
Duke University, Fall 2024

History: Written by Galen Reeves (-2023). Updated by Henry Pfister (2024-).
Last Modified: September 18, 2024

Outline of lecture:

4.1 Stochastic Processes

- A discrete-time **stochastic process** $\{X_i : i \in \mathbb{N}\}$ is an indexed collection of random variables $X_i \colon \Omega \to \mathcal{X}$ defined on a common probability space. The variables need not be independent nor identically distributed. Assuming $|\mathcal{X}| < \infty$, the distribution on $X^n = (X_1, X_2, \dots, X_n)$ for each n is defined by the joint pmf

$$p_{X^n}(x^n) = p(x_1, x_2, \dots, x_n).$$

- A stochastic process provides a useful probabilistic model. Examples include:
  - temperature each day
  - sequence of letters in a book
  - the closing price of the stock market each day.
  - the state of a deck of cards after each shuffle
  - location of a molecule at each millisecond.

- A stochastic process is **stationary** if the joint distribution of all subsets is invariant to shifts in the time index, i.e.

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] = \mathbb{P}[X_{1+k} = x_1, X_{2+k} = x_2, \cdots, X_{n+k} = x_n]$$

  for all $n, k \in \mathbb{N}$ and all $x^n \in \mathcal{X}^n$.

  Examples of Stationary Processes:
  - Deck of cards when shuffled starting from a uniformly random permutation? Yes
  - Deck of cards when shuffled starting from a sorted deck? No
  - The sequence of random variables $\{X_n\}$ given by $X_n = Y Z_n$ where Y is not constant and $\{Z_n\}$ is IID and independent of Y? Yes

- A stationary stochastic process is **ergodic** if, for all $m \in \mathbb{N}$ and all functions $f\colon \mathcal{X}^m \to \mathbb{R}$ satisfying $\mathbb{E}[|f(X_1, X_2, \ldots, X_m)|] < \infty$, the time average converges to the ensemble average

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_{1+i}, X_{2+i}, \ldots, X_{m+i}) = \mathbb{E}[f(X_1, X_2, \ldots, X_m)].$$

  Examples of Stationary Ergodic Processes:

  - Deck of cards when shuffled starting from a uniformly random permutation? Yes
  - Deck of cards when shuffled starting from a sorted deck? No
  - The sequence of random variables $\{X_n\}$ given by $X_n = YZ_n$ where Y is not constant and $\{Z_n\}$ is IID and independent of Y? No

4.2 Markov Chains

- A discrete-time stochastic process $\{X_1, X_2, \ldots\}$ is said to be a *Markov chain* or *Markov process* if for all $n = 1, 2, \ldots,$

$$\mathbb{P}[X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_1 = x_1] = \mathbb{P}[X_{n+1} = x_{n+1}|X_n = x_n]$$

  for all $x^{n+1} \in \mathcal{X}^{n+1}$.

- A Markov chain is **time-invariant** if the conditional probability $p(x_{n+1} \mid x_n)$ does not depend on n, i.e.

$$\mathbb{P}[X_{n+1} = b|X_n = a] = \mathbb{P}[X_2 = b|X_1 = a] \qquad \text{for all } a, b \in \mathcal{X}$$

- A time-invariant Markov Chain with finite alphabet $\mathcal{X} = \{1, 2, \ldots, M\}$ is characterized by an $M \times M$ probability transition matrix P where

$$P_{i,j} = \mathbb{P}[X_{n+1} = j|X_n = i], \qquad i, j \in \{1, 2, \ldots, M\}$$

  The vector of probabilities at time n is denoted by

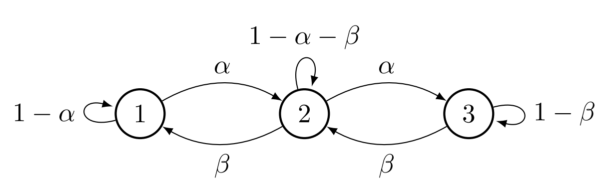
$$\mu_n = (\mu_{n,1}, \ldots, \mu_{n,M}) = (\mathbb{P}[X_n = 1], \mathbb{P}[X_n = 2], \ldots, \mathbb{P}[X_n = M])$$

  and is updated according to

$$\mu_{n+1,j} = \sum_{i=1}^{M} \mathbb{P}[X_{n+1} = j, X_n = i] = \sum_{i=1}^{M} \mu_{n,i} P_{i,j} = [\mu_n P]_j$$

- **Example:** Consider the following time invariant Markov chain with $\mathcal{X} = \{1, 2, 3\}$ and probability transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ \beta & 1-\alpha-\beta & \alpha \\ 0 & \beta & 1-\beta \end{bmatrix}$$



- A time-invariant finite-state Markov process is completely specified by three items:
  (1) the finite set of all states $\mathcal{X}$
  (2) the $|\mathcal{X}| \times |\mathcal{X}|$ transition probability matrix P
  (3) the starting distribution μ_1.
- A distribution μ is a **stationary distribution** of a time-invariant Markov process if

$$\mu = \mu P$$

- **Example:** For the Markov process described in the previous example, the stationary distribution is

$$\mu = \left(\frac{1}{1+r+r^2}, \frac{r}{1+r+r^2}, \frac{r^2}{1+r+r^2} \right)$$

  where $r = \alpha/\beta$. Note that if $\alpha = \beta$, then this is the uniform distribution since each state has probability 1/3.
- A finite-state Markov Process is *irreducible* if it is possible to go from any state to any other state in a finite number of steps (i.e. it can't get stuck in a subset of the states). Otherwise, it is *reducible*.
- A finite-state Markov Process is *periodic* if the states can be grouped into disjoint subsets so that all transitions from one subset lead to the next subset. Otherwise, it is *aperiodic*.
- **Theorem:** Consider a time-invariant finite-state Markov process. Then,
  (1) If *irreducible*, then the stationary distribution μ is unique and the process is ergodic.
  (2) If *irreducible* and *aperiodic*, then for any initial distribution μ_1, the distribution μ_n will converge to the stationary distribution μ as $n \to \infty$.
  (3) If it is *irreducible*, then it is stationary if, and only if, the starting distribution μ_1 equals the unique stationary distribution.

4.3 Entropy Rate

- With dependent sequences, how does the entropy $H(X^n)$ grow with n?

  (1) Definition 1: Average entropy per symbol

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}$$

  (2) Definition 2: Rate of information innovation

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_1, X_2, \ldots, X_{n-1})$$

- If $\{X_i\}$ are iid $\sim p(x)$, then both limits exists and are equal to $H(X)$.

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{nH(X)}{n} = H(X)$$

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_1, X_2, \ldots, X_{n-1}) = H(X)$$

- **Theorem:** For a stationary process, the limits in $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and are equal.

- **Proof:** $H'(\mathcal{X})$ exists
  - Observe that

$$\begin{aligned} H(X_n \mid X_1, \dots, X_{n-1}) &\le H(X_n \mid X_2, \dots, X_{n-1}) && \text{conditioning can't inc. entropy} \\ &= H(X_{n-1} \mid X_1, \dots, X_{n-2}) && \text{since stationary} \end{aligned}$$

  - Therefore, $H(X_n \mid X_1, \dots, X_{n-1})$ is non-increasing in n
  - Also $H(X) \ge 0$
  - The monotone convergence theorem for real sequences shows that the limit must exist.

- **Proof:** $H(\mathcal{X}) = H'(\mathcal{X})$
  - By the chain rule for entropy;

$$\frac{1}{n} H(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1})$$

  - From the arguments above, we know that

$$H(X_i \mid X_1, \dots, X_{i-1}) \to H'(\mathcal{X})$$

  - To conclude, we use the fact that convergence of a sequence implies convergence of its Cesàro mean:

$$a_n \to a \quad \Longrightarrow \quad \underbrace{\frac{1}{n} \sum_{i=1}^{n} a_i}_{\text{Cesàro mean}} \to a$$

  Proof: For all $\epsilon > 0$, there is an $N(\epsilon) < \infty$ such that $|a - a_n| < \epsilon$ for all $n > N(\epsilon)$. Let $B(\epsilon) = \sum_{i=1}^{N(\epsilon)} (a - a_i)$. Then, for all $m \ge B(\epsilon)/\epsilon$, we have

$$\left| a - \frac{1}{m} \sum_{i=1}^{m} a_i \right| = \frac{1}{m} \left| \sum_{i=1}^{m} (a - a_i) \right| \le \frac{1}{m} \left| \sum_{i=1}^{N(\epsilon)} (a - a_i) \right| + \frac{1}{m} \left| \sum_{i=N(\epsilon)+1}^{m} (a - a_i) \right| < \epsilon + \epsilon.$$

- Recall that the AEP for iid sources says

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \to H(X)$$

- **Theorem:** (AEP for stationary sources) For a stationary ergodic process

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \to H(\mathcal{X})$$

  - The proof is given in Section 16.8 of [CT].
  - You will see an example of a process that is stationary but not ergodic in your homework.
  - Informally, the term ergodic means that "time averages converge to ensemble averages"

4.4 Card Shuffling and MCMC

- **Theorem:** For a stationary time-invariant Markov process. the entropy rate is given by

$$H(\mathcal{X}) = H(X_2 \mid X_1)$$

where the conditional entropy is calculated using the *stationary* distribution.

- Markov Chain Monte Carlo (MCMC) is a widely used technique for approximating integrals and expectations. The basic idea is as follows:
  - Suppose you want to compute the expectation $\mathbb{E}[f(X)]$ where $X \sim Q$
  - If you could draw samples $\{X_i\}$ iid $\sim Q$, then you could approximate the expectation using the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}[f(X)]$$

  - However, drawing independent samples can be tricky. One method is to setup an irreducible and aperiodic Markov chain $\{Y_n\}$ whose stationary distribution is equal to Q. Then, if one starts from an arbitrary point Y_1 and waits *long enough*, the distribution of Y_n will converge to Q and Y_n will be independent of the starting point.
  - They key question is to determine how long it takes before the distribution of Y_n is sufficiently close to Q and sufficiently independent of Y_1.

- Let X be the state of a deck of card. There are $|\mathcal{X}| = 52!$ possibilities. Thus, if X is distributed uniformly, the the entropy is

$$H(X) = \log 52! = \sum_{i=1}^{52} \log(i) \approx 226 \quad \text{(bits)}$$

- Markov chain of one-at-a-time shuffling.
  - Let $X_1, X_2, \ldots, X_n$ be a sequence of card shuffles where X_{n+1} is the state of the deck after a card is selected uniformly at random from one of the 52 locations, and placed on the top.
  - The entropy satisfies

$$H(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_1) = H(X_{n+1} \mid X_n) = \log 52 \approx 5.7 \quad \text{(bits)}$$

  - By symmetry, the stationary distribution of this Markov chain is the uniform distribution. The Markov chain is irreducible and aperiodic, and thus

$$H(X_n) \to \log(52!) \quad \text{as } n \to \infty.$$

- How many one-at-a-time shuffles needed to produce a state that is uniformly distributed and independent of the starting point X_0?
  - Independence requires that $I(X_0; X_n) = 0$. Thus the necessary and sufficient condition for independent and uniformly distributed is

$$H(X_n|X_0) = \log 52!$$

- Let $N_1, N_2, N_3, \ldots$ be the sequence where N_n denote the location of the card that is moved to the top in the n-th shuffle.
- The joint entropy can be expressed as

$$H(X_0, X_1, X_2, \cdots, X_n) = H(X_0, N_1, N_2, \cdots N_n) = H(X_0) + nH(N_1)$$

and also as

$$\begin{aligned} H(X_0, X_1, \cdots, X_n) &= H(X_0) + H(X_1, \cdots, X_n \mid X_0) \\ &= H(X_0) + H(X_n \mid X_0) + H(X_1, \cdots, X_{n-1} \mid X_n, X_0) \end{aligned}$$

- Thus, the conditional entropy after n shuffles is upper bounded by

$$H(X_n \mid X_0) \leq nH(N_1) = n \log(52)$$

This provides a lower bound on the minimum number of shuffles that are needed:

$$\text{number of shuffles} \leq \frac{\log 52!}{\log 52} \approx 39.6 \qquad \Longrightarrow \qquad H(X_n \mid X_0) < \log(52!)$$

- The bound above is necessary, but is it sufficient?
No, In fact, it can be shown that

$$H(X_n|X_0) < \log 52! \qquad \text{for all } n$$

In other words, the there is always some memory from the initial state X_0

- Modified one-at-a-time shuffle: Lets instead try a modified one-at-a-time shuffle: on the n-th shuffle, we draw a number M uniformly on $\{n, n+1, n+2, \ldots, 52\}$ and place the Mth card on the top. After $n = 52$ no more changes are made to the deck.
  - The joint entropy can be decomposed as

$$H(X_0, X_1, \ldots, X_n) = H(X_0, M_1, M_2, \ldots M_n) = H(X_0) + \sum_{i=1}^{n} H(M_i)$$

  - Given the first state X_0 and last state X_n, one can reconstruct the sequence $\{M_i\}_{i=1}^n$ by looking at the top cards in the deck. Thus,

$$H(X_1, \cdots, X_{n-1} \mid X_n, X_0) = 0$$

  - Combining the above displays yields:

$$H(X_{51} \mid X_0) = \sum_{i=1}^{51} H(M_i) = \sum_{i=1}^{51} \log(52 - i + 1) = \log 52!$$

So, after exactly 51 shuffles we have a truly random permutation, regardless of our starting point.

4.5 Entropy of the English Language [Optional]

- Is the english language a stationary process? Probably not. Nevertheless, lest try to estimate its entropy rate.
- Assume 27 characters in alphabet (26 letters plus space).
- Models (estimate probabilities from text):

  (a) 0th-order Markov chain (iid):

  $$H(\mathcal{X}) \approx 4.76 \qquad \text{bits per letter}$$

  (b) 1st order Markov chain:

  $$H(\mathcal{X}) \approx 4.03 \qquad \text{bits per letter}$$

  (c) 4th order Markov chain:

  $$H(\mathcal{X}) \approx 2.8 \qquad \text{bits per letter}$$

- Estimate by asking people to guess the next letter until they get it correct. The *order* of their guesses reflects their estimate of the *order* of their conditional probabilities for the next letter. (Shannon 1952).

  $$H(\mathcal{X}) \approx 1.3 \qquad \text{bits per letter}$$