Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 3 & 4
INTRODUCTION TO INFORMATION THEORY

Chapter 2 of Elements of Information Theory, 2nd ed.

ENTROPY, RELATIVE ENTROPY, AND MUTUAL INFORMATION

OUTLINE

- Probability Review
- Entropy
- Joint entropy, conditional entropy
- Relative entropy, mutual information
- Chain rules
- Jensen’s inequality
- Data processing inequality
- Fano’s inequality

PROBABILITY REVIEWED

A discrete random variable X takes on values x from the discrete alphabet χ.
The probability mass function (pmf) is described by

$$p_X(x) = p(x) = \Pr\{X = x\}, for\ x\ \epsilon\ \chi$$

The joint probability mass function of two random variables X and Y taking on values in alphabets χ and ψ.

$$p_{X,Y}(x, y) = p(x, y) = \Pr\{X = x, Y = y\}, for\ x, y\ \epsilon\ \chi \times \psi$$

If $p_X(X = x) > 0$, the conditional probability that the outcome Y = y given that X = x is defined as:

$$p_{Y|X}(Y = y|X = x) = \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(x, y)}{p_{\boldsymbol{X}}(x)}$$

Product Rule

BASIC PROBABILITY RULES

Marginalization

$$p(y) = \sum_x p(x,y) = \sum_x p(y|x)p(x)$$

$$p(y) = \int_x p(x,y) = \int_x p(y|x)p(x)$$

Bayes' Rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Product Rule

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$$
$$= p_{X|Y}(x|y)p_Y(y)$$

Convention

- $0 \log 0 = 0$
- $a \log \frac{a}{0} = \infty$, if $a > 0$
- $0 \log \frac{0}{0} = 0$

INDEPENDENCE REVIEWED

The events $X = x$ and $Y = y$ are *statistically independent* if

$$\boxed{p(x, y) = p(x)p(y).}$$

The random variables X and Y defined over the alphabets χ and ψ , resp. are statistically independent if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \text{ for } \forall(x, y) \in \chi \times \psi$$

The variables $X_1, X_2, \dots, X_N$ are called independent if for all $(x_1, x_2, \dots, x_N) \in \chi_1 \times \chi_x \times \cdots \times \chi_N$

$$p(x_1, x_2, \dots, x_N) = \prod_{i=1}^{N} p_{X_i}(x_i)$$

They are furthermore called identically distributed if all variables X_i have the same distribution $p_X(x)$.

EXPECTED VALUE

1 Discrete random variable, finite case, taking $x_1, x_2, \ldots, x_N$ with prob. $p_1, p_2, \ldots, p_N$

$$\boxed{E[X] = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_k p_N}{p_1 + p_2 + \cdots + p_N}}$$

Sum to 1 if probability

2 Discrete random variable X, countable case, taking $x_1, x_2, \ldots$ with prob. $p_1, p_2, \ldots$

$$\boxed{E[X] = \sum_{i=1}^{\infty} x_i p_i}$$

3 Univariate continuous random variable:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x$$

General definition: random variable defined on a probability space (Ω, Σ, P), then the expected value of X, denoted by E[X], ⟨X⟩, $\overline{\mathrm{X}}$ or **E**[X], is defined as the Lebesgue integral

$$E[X] = \int_{\Omega} X\, dP = \int_{\Omega} X(\omega)\, P(\mathrm{d}\omega)$$

ENTROPY

Definition:

The **entropy** *H*(*X*) of a discrete random variable *X* with pmf $p_X(x)$ is given by

$$H(X) = -\sum_{x} p_X(x) \log p_X(x) = -E_{p_X(x)}[\log p_X(X)]$$

The **entropy** *H*(*X*) of a continuous random variable *X* with pdf $f_X(x)$ in support set S is given by

$$h(X) = -\int_S f_X(x) \log f_X(x) = -E_{f_X(x)}[\log f_X(X)]$$

Meaning:

- Measure of the uncertainty of the r.v.
- Measure of the amount of information required on the average to describe the r.v.

Denote H(X) and H(p) as same when X is binary rv
Use log base 2

JOINT ENTROPY

Definition:

The **joint entropy** *H(X,Y)* on a pair of discrete r.v. *(X,Y)* with a joint distribution *p(x,y)* is defined as

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$
$$= -E_{p(x,y)} \log p(x,y)$$

CONDITIONAL ENTROPY

Definition:

The **conditional entropy** *H(Y|X)* on a pair of discrete r.v. *(X,Y)* with a joint distribution *p(x,y)* is defined as

$$H(Y|X) = -\sum_x p(x) H(Y|X=x)$$
$$= \sum_x p(x) \sum_y p(y|x) \log p(y|x)$$
$$= -\sum_x \sum_y p(x,y) \log p(y|x)$$
$$= -E_{p(x,y)} \log p(y|x)$$

CHAIN RULE

Theory (**Chain Rule**)

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

proof

Corollary

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

Remark

$$H(Y|X) \neq H(X|Y)$$
$$H(Y) - H(Y|X) = H(X) - H(X|Y)$$

RELATIVE ENTROPY

Definition:

The **relative entropy** (**Kullbuck-Leibler distance, K-L divergence**) between two probability mass function *p(x)* and *q(x)* is defined as

$$D(p||q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}$$

Meaning:

- **Distance** between two distributions
- A measure of the **inefficiency** of assuming that the distribution is *q* when the true distribution is *p*

Properties:

- Is non-negative
- $D(p||q) = 0$ if and only if *p*=*q*
- Is asymmetric : $D(p||q) \neq D(q||p)$
- Does not satisfy triangle inequality

Definition:

The **conditional relative entropy** between two probability mass function *p(x,y)* and *q(x,y)* is defined as

$$D(p(y|x)||q(y|x)) = \sum_{x \in \chi} p(y|x) \log \frac{p(y|x)}{q(y|x)} = E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}$$

MUTUAL INFORMATION

Definition:

Mutual information I(X;Y) is the relative entropy between the joint distribution p(x,y) and the product distribution p(x)p(y)

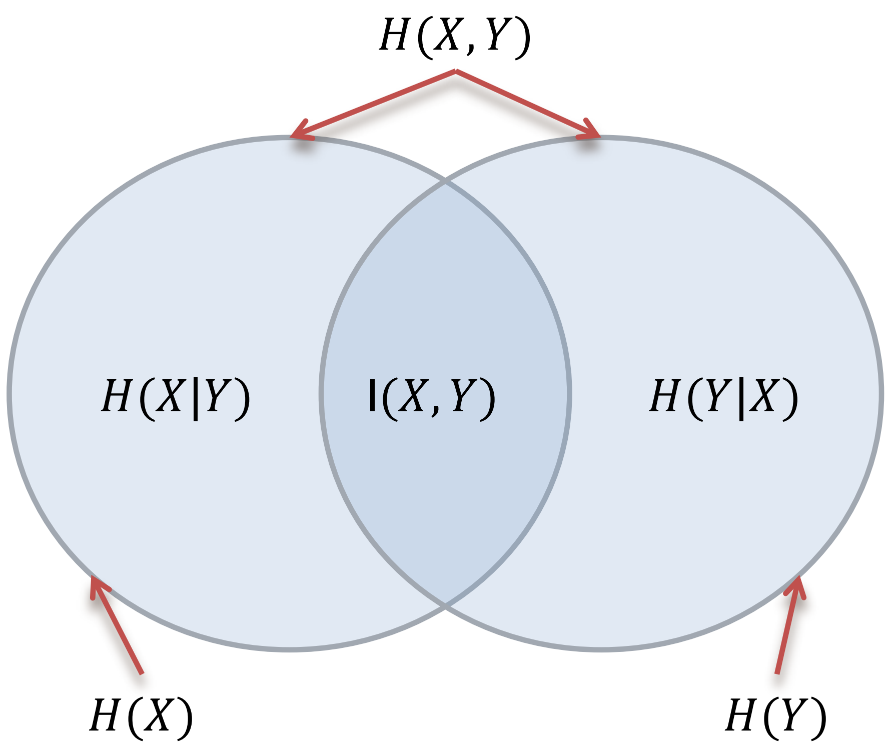
$$I(X;Y) = D(p(x,y)||p(x)p(y))$$
$$= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}$$

Definition:

Conditional mutual information I(X;Y|Z) is the reduction in the uncertainty of X due to knowledge of Y when Z is given

$$I(X;Y|Z) = D(p(x,y|z)||p(x|z)p(y|z))$$
$$= \sum_x \sum_y p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$
$$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$
$$= \mathrm{H(X|Z)} - \mathrm{H(X|Y,Z)}$$

RELATIONSHIP BETWEEN ENTROPY AND MUTUAL INFORMATION



Properties:

- I(X;Y) is the reduction of uncertainty of X due to the knowledge of Y (or *vise versa*)

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

- Is symmetric: X says about Y as much and Y says about X

- $I(X;Y) = H(Y) + H(X) - H(X,Y)$
  since $H(X,Y) = H(X) + H(Y|X)$
  by chain rule

- $I(X;X) = H(X)$ also called **self information**

VARIATIONS OF CHAIN RULES

Theorem **(chain rule for entropy)**
Let $X_1, X_2, \dots, X_n$ be drawn according to $p(x_1, x_2, \dots, x_n)$. Then,

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1)$$

Theorem **(chain rule for information)**
Let $X_1, X_2, \dots, X_n$ be drawn according to $p(x_1, x_2, \dots, x_n)$. Then,

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \dots, X_1)$$

Theorem **(chain rule for relative entropy)**
For joint pmf $p(x, y)$ and $q(x, y)$.

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

JENSEN'S INEQUALITY

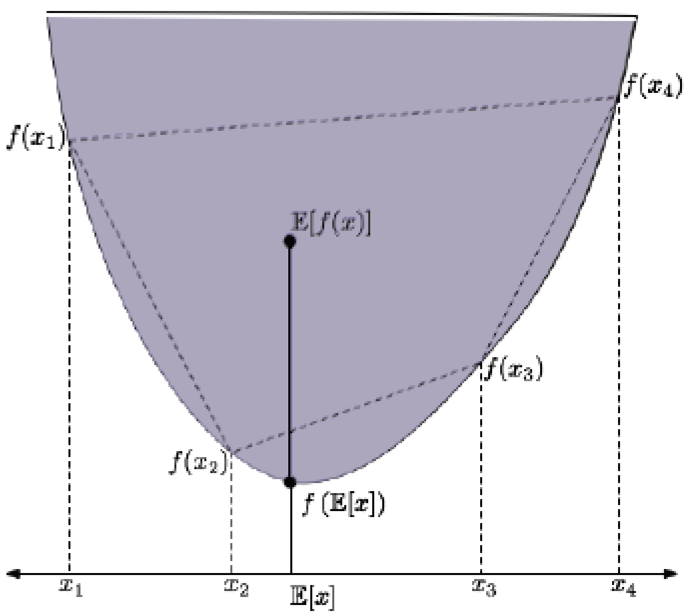
Theorem (Jensen's Inequality)

If f is a convex function and X is a random variable,

$$Ef(X) \geq f(EX)$$

Moreover, if f is strictly convex, the equality implies that X=EX with probability 1 (i.e. X is a constant)

proof



JENSEN'S INEQUALITY CONSEQUENCES

Theorem (Information Inequality)
Let $p(x), q(x), x \in \chi$, be two probability mass funcions. Then,

$$D(p||q) \geq 0$$

With equality if and only if *p(x)* = *q(x)* for all *x*.

proof

Corollary (No-negativity of mutual information)
For any two random variable X and Y. Then,

$$I(X;Y) \geq 0$$

With equality if and only if *X and Y are independent*.

proof

Corollary

$$D(p(y|x)||q(y|x)) \geq 0$$

With equality if and only if *p(y|x)* = *q(y|x)* for all y and *x* such that p(x) > 0.

Corollary

$$I(X;Y|Z) \geq 0$$

With equality if and only if *X and Y are independent given Z*.

JENSEN'S INEQUALITY CONSEQUENCES CONT.

Theorem [UPPER BOUND IN ENTROPY]

Let $H(X) \leq \log|\chi|$, where $|\chi|$ denotes the number of elements in the range of X, with equality if and only X has a uniform distribution over χ.

Proof Hint) show $D(p||u) = \log|\chi| - H(X)$, where $u(x) = \frac{1}{|\chi|}$

proof

Theorem **(Conditioning reduces entropy)**

$$H(X|Y) \leq H(X),$$

With equality if and only if X and Y are independent.

proof

NOTE>

The theorem says that knowing another r.v. Y can only reduce the uncertainty in X. Note that this in true only on the average. Specific H(X|Y=y) may be greater than or less than or euqal to H(X).

JENSEN'S INEQUALITY CONSEQUENCES CONT.

Theorem **(Independence Bound on Entropy)**

Let $X_1, X_2, \dots, X_n$ be drawn according to $p(x_1, x_2, \dots, x_n)$. Then

$$H(X_1, X_2, \dots, X_n) \le \sum_{i=1}^{n} H(X_i),$$

With equality if and only if X_i are independent.

Proof Hint> use chain rule of entropy

LOG-SUM INEQUALITY

Theorem (Log sum inequality)
For nonnegative numbers $a_1, a_2, \dots, a_n$ and $b_1, b_2, \dots, b_n$. Then,

$$\sum_{i=1}^{n} a_1 \log(\frac{a_i}{b_i}) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left((\sum\nolimits_{i=1}^{n} a_i)/(\sum\nolimits_{i=1}^{n} b_i)\right)$$

with equality if and only if $\frac{a_i}{b_i} = \text{constant}$.

proof

Convention

- $0 \log 0 = 0$
- $\mathrm{a} \log \frac{a}{0} = \infty$, if $a > 0$
- $0 \log \frac{0}{0} = 0$

LOG-SUM INEQUALITY CONSEQUENCES

Theorem **(Convexity of relative entropy)**
$D(p||q)$ is convex in the pair (p,q), so that for pmf's (p_1, q_1) and (p_2, q_2), we have for all $0 \leq \lambda \leq 1$:

$$D(\lambda p_1 + (1-\lambda)p_2)||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2, q_2)$$

Theorem **(Concavity of entropy)**
For $X \sim p(x)$, we have that

$H(p) \coloneqq H_p(X)$is concave function of p(x).

LOG-SUM INEQUALITY CONSEQUENCES CONT.

Theorem (*Concavity of the mutual information in $p(x)$*)
Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Then, $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$.

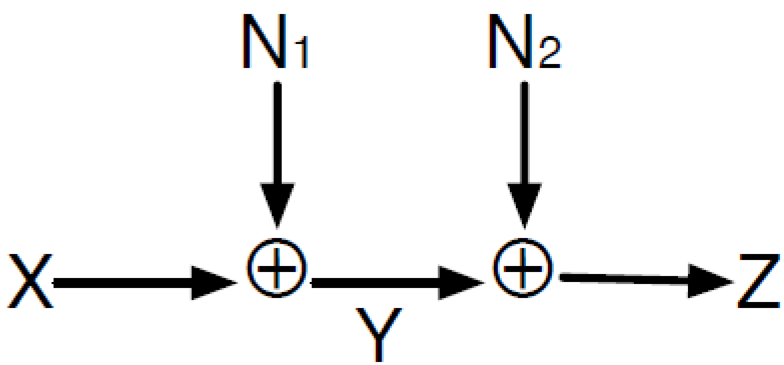
Theorem (*Convexity of the mutual information in $p(y|x)$*)
Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Then, $I(X; Y)$ is a convex function of $p(y|x)$ for fixed $p(x)$

MARKOV CHAINS

Definition:
X, Y, Z form a Markov chain in that order ($X \rightarrow Y \rightarrow Z$) iff

$$p(x, y, z) = p(x)p(y|x)p(z|y) \equiv p(z|y, x) = p(z|y)$$

With equality if and only if X *and* Y *are independent given* Z.



$X \rightarrow Y \rightarrow Z$ iff X and Z are conditionally independent given Y

$X \rightarrow Y \rightarrow Z \Rightarrow Z \rightarrow Y \rightarrow X$. Thus, we can write $X \leftrightarrow Y \leftrightarrow Z$.

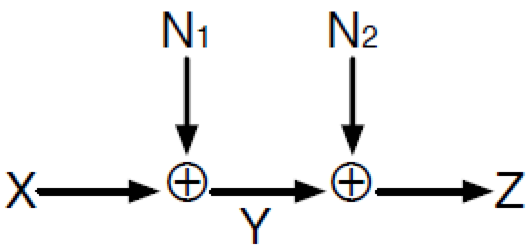
DATA-PROCESSING INEQUALITY

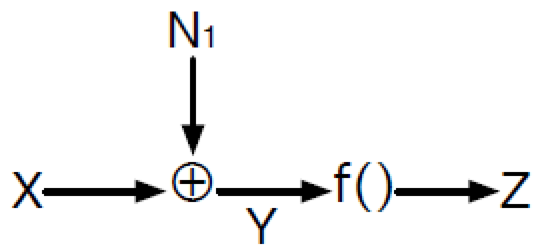
proof

Theorem (Data-processing inequality)
If $X \rightarrow Y \rightarrow Z$, *then*

$$I(X;Y) \geq I(X;Z)$$

with equality iff $I(X;Y|Z) = 0$.





Corollary

$$\text{If } \mathrm{Z} = \mathrm{f(Y)}, \text{ then } I(X;Y) \geq I(X;f(Y)).$$

Corollary
If $X \rightarrow Y \rightarrow Z$, then

$$I(X;Y) \geq I(X;Y|Z)$$

SUFFICIENT STATISTIC

Definition:
A function *T*(*X*) is said to be a ***sufficient statistic*** relative to the family $\{f_\theta(x)\}$if the conditional distribution of *X*, given *T*(*X*) = *t*, is independent of θ for any distribution on θ (***Fisher-Neyman***):

$$f_\theta(x) = f(x|t)f_\theta(t) \Rightarrow \theta \rightarrow T(X) \Rightarrow I\big(\theta; T(X)\big) \geq I(\theta; X)$$

Hence, $I(\theta; X)$= $I\big(\theta; T(X)\big)$ for a sufficient statistics (suf stat. preserves mutual info.)

FANO'S INEQUALITY

Problem: using the observation of r.v. Y. we want to guess the value of X that is correlated to r.v. Y.
-> Fano's inequality relates the probability of error in guessing the r.v. X to its conditional entropy H(X|Y).
* We can estimate X for Y with 0 prob. Of error if and only if H(X|Y) = 0;

Theorem (Fano's inequality)
For any estimator $\hat{X}$ such that $X \rightarrow Y \rightarrow \hat{X}$, with probability of error $P_e = \Pr(X \neq \hat{X})$, we have

proof

$$H(P_e) + P_e \log|\chi| \geq H(X|\hat{X}) \geq H(X|Y)$$

$$g(Y) = \hat{X}$$

This inequality can be weekend to

$$1 + P_e \log|\chi| \geq H(X|Y)$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log|\chi|}$$

NOTE: Fano's bound is a loose bound, but sufficient for many cases of interest.

FANO'S INEQUALITY CONSEQUENCES

Corollary

Let $p = Pr(X \neq Y)$. Then

$$H(p) + p\log|\chi| \geq H(X|Y) \,.$$

Corollary

Let $\mathrm{P_e} = Pr\big(X \neq \hat{X}\big)$, and $\hat{X}\colon \psi \to \chi$; Then

$$H(P_e) + P_e \log(|\chi| - 1) \geq H(X|Y) \,.$$

* Range of possible outcome changed to $|\chi| - 1$.

Remark:

Suppose that ther is no knowledge of Y. Thus, X must be guessed. Without any information. Let $X \in \{1,2, \dots, m\}$ and $p_1 \geq p_2 \geq \cdots \geq p_m$. Then the best guess of X is $\hat{X} = 1$ and the resulting probability of error is $P_e = 1 - p_1$. Fano's inequality becomes

$$H(P_e) + P_e \log|m - 1| \geq H(X)$$

The pmf

$$(p_1, p_2, \dots, p_m) = (1 - P_e, \frac{P_e}{m-1}, \dots, \frac{P_e}{m-1})$$

achieves this bound with equality.

FANO'S INEQUALITY CONSEQUENCES

Lemma

If X and X' are i.i.d. with entropy H(X), assume the probability at X=X' is given by

$$P(X = X') = \sum_x p^2(x).$$

Then

$$Pr(X = X') \geq 2^{-H(X)}$$

with equality if and only if X has a uniform distribution.

Corollary

Let X, X' be independent with X ~ p(x), X'~r(x), x, x' ∈ χ , then

$$Pr(X = X') \geq 2^{-H(p)-D(p||r)}$$
$$Pr(X = X') \geq 2^{-H(r)-D(r||p)}$$

with equality if and only if X has a uniform distribution.

Chapter 4 of Elements of Information Theory, 2nd ed.

ENTROPY RATES OF A STOCHASTIC PROCESS

STOCHASTIC PROCESSES

- What about the notion of entropy of a general random process?

Definition: A *stochastic process* $\{X_i\}$ is an indexed sequence of random variables.

Definition: A *discrete-time stochastic process* $\{X_i\}_{i \in \mathcal{I}}$ is one for which we associate the discrete index set $\mathcal{I} = \{1, 2, \ldots\}$ with time.

Entropy: $H(\{X_i\}) = H(X_1) + H(X_2|X_1) + \cdots = \infty$ (often)

MOTIVATION: Should probably normalize by n somehow.

ENTROPY RATE

- *Entropy Rate:* The *entropy rate* of a stochastic process $\{X_i\}$ is defined by

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

when the limit exists. We can also define an alternative notion:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1).$$

- Entropy rate estimates the additional entropy per new sample.
- Gives lower bound on number of code bits per sample.
- If the X_i are not i.i.d the entropy rate limit may not exist.
- X_i i.i.d. random variables: $H(\mathcal{X}) = H(X_i)$

STATIONARY PROCESSES

Definition: A discrete-time stochastic process is said to be *stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index; that is,

$$\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\}$$

for every n and every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Lemma: For a stationary stochastic process, $H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.

Lemma: Cesáro mean If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.

Theorem: For a stationary stochastic process, $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and are equal:

$$H(\mathcal{X}) = H'(\mathcal{X}).$$