Entropy Rate Estimation for Markov Chains with Large State Space

Yanjun Han
Jiantao Jiao
Chuan–Zheng Lee
Tsachy Weissman
Yihong Wu
Tiancheng Yu

Stanford EE
Berkeley EECS
Stanford EE
Stanford EE
Yale Stats
Tsinghua EE

NIPS 2018, Montréal, Canada
Entropy Rate Estimation

Entropy rate of a stationary process \( \{X_t\}_{t=1}^{\infty} \):

\[
\tilde{H} \triangleq \lim_{n \to \infty} \frac{H(X^n)}{n}, \quad H(X^n) = \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) \log \frac{1}{p_{X^n}(x^n)}.
\]
Entropy Rate Estimation

Entropy rate of a stationary process \( \{X_t\}_{t=1}^{\infty} \):

\[
\bar{H} \triangleq \lim_{n \to \infty} \frac{H(X^n)}{n}, \quad H(X^n) = \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) \log \frac{1}{p_{X^n}(x^n)}.
\]

- fundamental limit of the expected logarithmic loss when predicting the next symbol given all past symbols
Entropy Rate Estimation

Entropy rate of a stationary process $\{X_t\}_{t=1}^{\infty}$:

$$\tilde{H} \triangleq \lim_{n \to \infty} \frac{H(X^n)}{n}, \quad H(X^n) = \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) \log \frac{1}{p_{X^n}(x^n)}.$$ 

- fundamental limit of the expected logarithmic loss when predicting the next symbol given all past symbols
- fundamental limit of data compressing for stationary stochastic processes
Entropy Rate Estimation

Entropy rate of a stationary process $\{X_t\}_{t=1}^{\infty}$:

$$\tilde{H} \triangleq \lim_{n \to \infty} \frac{H(X^n)}{n}, \quad H(X^n) = \sum_{x^n \in \mathcal{X}^n} p_X(x^n) \log \frac{1}{p_X(x^n)}.$$ 

- fundamental limit of the expected logarithmic loss when predicting the next symbol given all past symbols
- fundamental limit of data compressing for stationary stochastic processes

Our Task
Given a length-$n$ trajectory $\{X_t\}_{t=1}^{n}$, estimate $\tilde{H}$. 

From Entropy to Entropy Rate

Theorem (Jiao–Venkat–Han–Weissman'15, Wu–Yang'16)

For discrete entropy estimation with support size $S$, consistent estimation is possible if and only if $n \gg \frac{S}{\log S}$. 
From Entropy to Entropy Rate

Theorem (Jiao–Venkat–Han–Weissman'15, Wu–Yang'16)

For discrete entropy estimation with support size $S$, consistent estimation is possible if and only if $n \gg \frac{S}{\log S}$.

Sample Complexity

\[ n \approx S \log S \]

\[ n \approx \infty \]

\[ n \approx 3/7 \]
From Entropy to Entropy Rate

Theorem (Jiao–Venkat–Han–Weissman'15, Wu–Yang'16)

For discrete entropy estimation with support size $S$, consistent estimation is possible if and only if $n \gg \frac{S}{\log S}$.

Sample Complexity

$$n \asymp \frac{S}{\log S}$$

\[\left\langle\begin{array}{c}
i.i.d. \text{ process} \\
\text{constant process}
\end{array}\right.\]
From Entropy to Entropy Rate

Theorem (Jiao–Venkat–Han–Weissman’15, Wu–Yang’16)

For discrete entropy estimation with support size $S$, consistent estimation is possible if and only if $n \gg \frac{S}{\log S}$.

Sample Complexity

\[ n \asymp \frac{S}{\log S} \quad \text{i.i.d. process} \]

\[ n \asymp \infty \quad \text{constant process} \]
From Entropy to Entropy Rate

Theorem (Jiao–Venkat–Han–Weissman’15, Wu–Yang’16)

For discrete entropy estimation with support size $S$, consistent estimation is possible if and only if $n \gg \frac{S}{\log S}$.

Sample Complexity

$$n \asymp \frac{S}{\log S}$$

\[\text{i.i.d. process} \quad n \asymp ? \quad \text{constant process} \quad n \asymp \infty\]
Assumption

The data-generating process \( \{X_t\}_{t=1}^n \) is a reversible first-order Markov chain with relaxation time \( \tau_{rel} \).
Assumption

The data-generating process $\{X_t\}_{t=1}^n$ is a reversible first-order Markov chain with relaxation time $\tau_{\text{rel}}$.

- Relaxation time $\tau_{\text{rel}} = (\text{spectral gap})^{-1} \geq 1$ characterizes the mixing time of the Markov chain
Assumption

The data-generating process $\{X_t\}_{t=1}^n$ is a reversible first-order Markov chain with relaxation time $\tau_{\text{rel}}$.

- Relaxation time $\tau_{\text{rel}} = (\text{spectral gap})^{-1} \geq 1$ characterizes the mixing time of the Markov chain.
- High-dimensional setting: state space $S = |\mathcal{X}|$ is large and may scale with $n$. 
Estimators

For first-order Markov chain:

\[
\bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \pi_i \bar{H}(X_1|X_0 = i)
\]

where \(\pi_i\) is the stationary distribution and \(\bar{H}(X_1|X_0 = i)\) is the conditional entropy.
Estimators

For first-order Markov chain:

\[ \tilde{H} = H(X_1|X_0) = \sum_{i=1}^{S} \pi_i H(X_1|X_0 = i) \]

- Estimate of \( \pi_i \): empirical frequency \( \hat{\pi}_i \) of state \( i \)
For first-order Markov chain:

\[ \bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \pi_i \sum_{j=1}^{i} H(X_1|X_0 = i) \]

- Estimate of \( \pi_i \): empirical frequency \( \hat{\pi}_i \) of state \( i \)
- Estimate of \( H(X_1|X_0 = i) \): estimate discrete entropy from samples \( X^{(i)} = \{X_j : X_{j-1} = i\} \)
Estimators

For first-order Markov chain:

\[ \bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \pi_i H(X_1|X_0 = i) \]

- Estimate of \( \pi_i \): empirical frequency \( \hat{\pi}_i \) of state \( i \)
- Estimate of \( H(X_1|X_0 = i) \): estimate discrete entropy from samples \( X^{(i)} = \{X_j : X_{j-1} = i\} \)

Estimators

- Empirical estimator: \( \bar{H}_{\text{emp}} = \sum_{i=1}^{S} \hat{\pi}_i \hat{H}_{\text{emp}}(X^{(i)}) \)
- Proposed estimator: \( \bar{H}_{\text{opt}} = \sum_{i=1}^{S} \hat{\pi}_i \hat{H}_{\text{opt}}(X^{(i)}) \)
Main Results

Empirical estimator $\tilde{H}_{\text{emp}}$

\[ \Theta\left(\frac{S}{\log^3 S}\right) \]

For a wide range of $\tau_{\text{rel}}$, sample complexity does not depend on $\tau_{\text{rel}}$. 
Main Results

Empirical estimator $\bar{H}_{\text{emp}}$

$n \asymp S^2$

$n \gtrsim S^2$

$\Theta\left(\frac{S}{\log^3 S}\right)$
Main Results

Empirical estimator $\bar{H}_{\text{emp}}$

$n \asymp S^2$

Proposed estimator $\bar{H}_{\text{opt}}$

$n \gtrsim S^2$

$1 \quad 1 + \Theta\left(\frac{\log^2 S}{\sqrt{S}}\right) \quad \Theta\left(\frac{S}{\log^3 S}\right)$
Main Results

Empirical estimator $\tilde{H}_{\text{emp}}$

\[
n \asymp S^2
\]

Proposed estimator $\tilde{H}_{\text{opt}}$

\[
n \asymp \frac{S}{\log S}
\]

For a wide range of $\tau_{\text{rel}}$, sample complexity does not depend on $\tau_{\text{rel}}$. 
Main Results

**Empirical estimator $\tilde{H}_{\text{emp}}$**

\[ n \asymp S^2 \]

**Proposed estimator $\tilde{H}_{\text{opt}}$**

\[ n \asymp \frac{S}{\log S} \]

For a wide range of $\tau_{\text{rel}}$, sample complexity does not depend on $\tau_{\text{rel}}$. 

\[ n \asymp \frac{S^2}{\log S} \]

\[ n \asymp \frac{S^2}{\log S} \]

\[ n \asymp \frac{S^2}{\log S} \]
Application: Fundamental Limits of Language Models

Figure: Estimated and achieved fundamental limits of language modeling

- Penn Treebank (PTB): 1.50 vs. 5.96 bits per word
- Googles One Billion Words (1BW): 3.46 vs. 4.55 bits per word