Entropy Rate Estimation for Markov Chains with Large State Space

Yanjun Han	Stanford EE
Jiantao Jiao	Berkeley EECS
Chuan–Zheng Lee	Stanford EE
Tsachy Weissman	Stanford EE
Yihong Wu	Yale Stats
Tiancheng Yu	Tsinghua EE

NIPS 2018, Montréal, Canada

Entropy Rate Estimation

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

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

Entropy Rate Estimation

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

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

- 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 $\{X_t\}_{t=1}^{\infty}$:

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

- 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 $\{X_t\}_{t=1}^{\infty}$:

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

- 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

Our Task

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

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

$\longleftrightarrow$

i.i.d. process — 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

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

$\longleftrightarrow$

i.i.d. process — 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

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

i.i.d. process — 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

$n \asymp \frac{S}{\log S}$ $\quad$ $n \asymp ?$ $\quad$ $n \asymp \infty$

i.i.d. process $\longleftrightarrow$ constant process

Assumption

Assumption

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

Assumption

Assumption

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

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

Assumption

Assumption

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

- Relaxation time $\tau_{\mathsf{rel}} = (\text{spectral gap})^{-1} \geq 1$ 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:

$$\bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \underbrace{\pi_i}_{\text{stationary distribution}} \overbrace{H(X_1|X_0 = i)}^{\text{conditional entropy}}$$

Estimators

For first-order Markov chain:

$$\bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \underbrace{\pi_i}_{\text{stationary distribution}} \overbrace{H(X_1|X_0 = i)}^{\text{conditional entropy}}$$

- Estimate of π_i: empirical frequency $\hat{\pi}_i$ of state i

Estimators

For first-order Markov chain:

$$\bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \underbrace{\pi_i}_{\text{stationary distribution}} \overbrace{H(X_1|X_0 = i)}^{\text{conditional entropy}}$$

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

Estimators

For first-order Markov chain:

$$\bar{H} = H(X_1|X_0) = \sum_{i=1}^{S} \underbrace{\pi_i}_{\text{stationary distribution}} \overbrace{H(X_1|X_0 = i)}^{\text{conditional entropy}}$$

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

Estimators

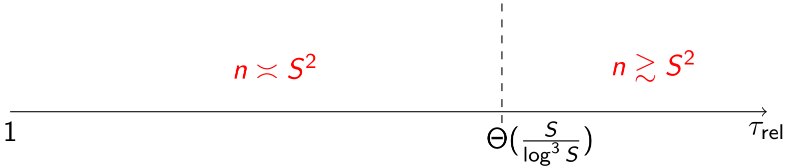
- Empirical estimator: $\bar{H}_{\text{emp}} = \sum_{i=1}^{S} \hat{\pi}_i \hat{H}_{\text{emp}}(\mathbf{X}^{(i)})$
- Proposed estimator: $\bar{H}_{\text{opt}} = \sum_{i=1}^{S} \hat{\pi}_i \hat{H}_{\text{opt}}(\mathbf{X}^{(i)})$

Main Results

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

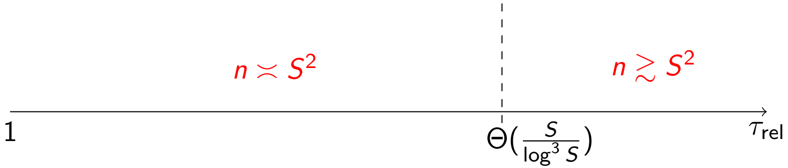
Main Results

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



Main Results

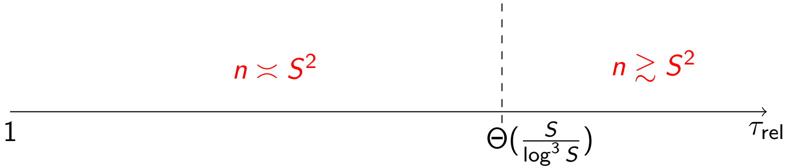
Empirical estimator $\bar{H}_{\mathrm{emp}}$



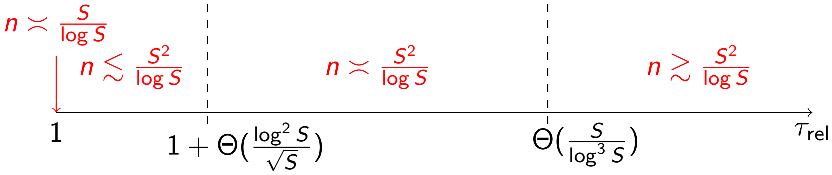
Proposed estimator $\bar{H}_{\mathrm{opt}}$

Main Results

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

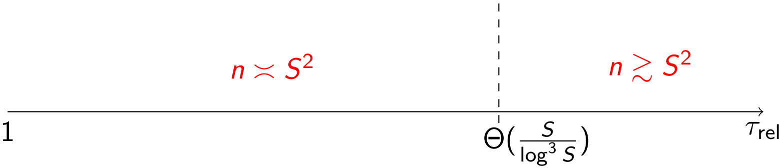


Proposed estimator $\bar{H}_{\mathrm{opt}}$

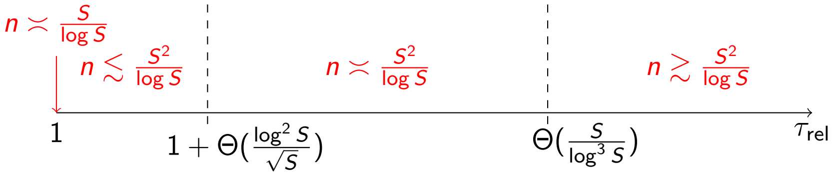


Main Results

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



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



For a wide range of τ_{rel}, sample complexity does not depend on τ_{rel}.

Application: Fundamental Limits of Language Models

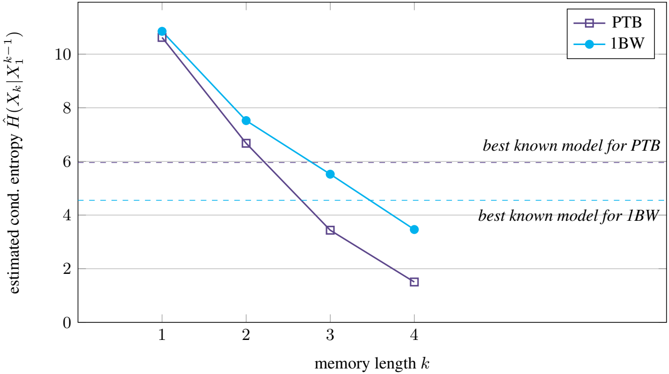


Figure: Estimated and achieved fundamental limits of language modeling

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