Entropy Rate Estimation for Markov Chains with Large State Space

Yanjun Han*, Jiantao Jiao†, Chuan–Zheng Lee*, Tsachy Weissman*, Yihong Wu‡, Tiancheng Yu§

* Department of Electrical Engineering, Stanford University. Email: {yjhan, czlee, tsachy}@stanford.edu

† Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Email: jiantao@berkeley.edu

‡ Department of Statistics and Data Science, Yale University. Email: yihong.wu@yale.edu

§ Department of Electronic Engineering, Tsinghua University. Email: thueeyutc14@foxmail.com

Problem Formulation

Entropy of a random vector $X^n \in \mathcal{X}^n$:

$$H(X^n) \triangleq \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) \log \frac{1}{p_{X^n}(x^n)}.$$

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

$$\bar{H} \triangleq \lim_{n\to\infty} \frac{H(X^n)}{n}.$$

Entropy rate serves as the fundamental limit of:

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

Target: given a length-n trajectory $\{X_t\}_{t=1}^n$ from the stationary process, estimate $\bar{H}$.

Assumptions and Estimators

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

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

Estimators:

- Notation: $\hat{\pi}_i$ denotes the empirical frequency of state i, and $\mathbf{X}^{(i)} = \{X_j : X_{j-1} = i\}$ consists of sample states following state i.
- Empirical estimator: $\bar{H}_{\text{emp}} = \sum_{i=1}^S \hat{\pi}_i \hat{H}_{\text{emp}}(\mathbf{X}^{(i)})$, where $\hat{H}_{\text{emp}}(\cdot)$ is the empirical entropy estimator for i.i.d. data.
- Proposed estimator: $\bar{H}_{\text{opt}} = \sum_{i=1}^S \hat{\pi}_i \hat{H}_{\text{opt}}(\mathbf{X}^{(i)})$, where $\hat{H}_{\text{opt}}(\cdot)$ is any minimax rate-optimal entropy estimator for i.i.d. data [1, 2].

Performance of Empirical Estimator

Theorem:

- If $\tau_{\text{rel}} = O(\frac{S}{\log^3 S})$, the empirical entropy rate $\bar{H}_{\text{emp}}$ is consistent in estimating $\bar{H}$ if $n = \omega(S^2)$;
- For general $\tau_{\text{rel}} \geq 1$, the empirical entropy rate $\bar{H}_{\text{emp}}$ is not consistent in estimating $\bar{H}$ if $n = O(S^2)$.

Corollary: For a wide range of relaxation time, the sample complexity of the empirical estimator is $n \asymp S^2$.

Minimax Estimation

Theorem:

- If $\tau_{\text{rel}} = O(\frac{S}{\log^3 S})$, the proposed estimator $\bar{H}_{\text{opt}}$ is consistent in estimating $\bar{H}$ if $n = \omega(\frac{S^2}{\log S})$;
- If $\tau_{\text{rel}} \geq 1 + \Omega(\frac{\log^2 S}{\sqrt{S}})$, no estimator can be consistent in estimating $\bar{H}$ if $n = O(\frac{S^2}{\log S})$.

Corollary (dependence of optimal sample complexity on relaxation time):

- If $\tau_{\text{rel}} = 1$: sample complexity is $n \asymp \frac{S}{\log S}$;
- If $1 \leq \tau_{\text{rel}} \leq 1 + \Omega(\frac{\log^2 S}{\sqrt{S}})$: sample complexity is $O(\frac{S^2}{\log S})$ with unknown lower bound;
- If $1 + \Omega(\frac{\log^2 S}{\sqrt{S}}) \leq \tau_{\text{rel}} \lesssim \frac{S}{\log^3 S}$: sample complexity is $n \asymp \frac{S^2}{\log S}$;
- If $\tau_{\text{rel}} \gg \frac{S}{\log^3 S}$: sample complexity is $\Omega(\frac{S^2}{\log S})$ with unknown upper bound.

Application: Fundamental Limits of Language Modeling

Two aspects of language modeling:

- Achieving fundamental limit: train some language model which achieves a low cross-entropy rate (i.e., high efficacy)
- Estimating fundamental limit: provide an estimate of the entropy rate of the language (i.e., the optimal cross-entropy rate for any language model)

Dataset: Penn Treebank (PTB) and Googles One Billion Words (1BW) benchmarks

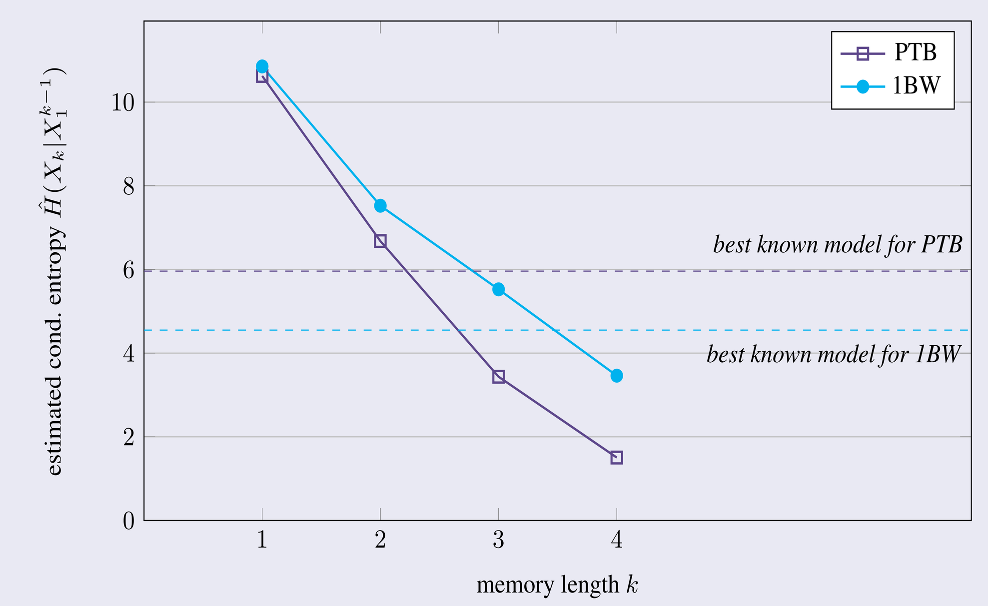

Figure: Estimated and achieved fundamental limits of language modeling

Contact

The authors are sorry to be absent due to visa reasons. If you have any questions and comments, feel free to email the authors.

References

[1] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. *Minimax estimation of functionals of discrete distributions.* IEEE Transactions on Information Theory, 61(5):2835–2885, 2015.

[2] Yihong Wu and Pengkun Yang. *Minimax rates of entropy estimation on large alphabets via best polynomial approximation.* IEEE Transactions on Information Theory, 62(6):3702–3720, 2016.