Consistency of the Plug-In Estimator of the Entropy Rate for Ergodic Processes

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

ISIT 2016, Barcelona

Entropy estimation

- Entropy estimation is well researched in the IID case:
  - Paninski (2004), *Estimating Entropy on m Bins Given Fewer Than m Samples*.
  - Valiant and Valiant (2011), *An $n/\log(n)$-Sample Estimator for Entropy and Support Size*.
  - Jiao, Venkat, Han, and Weissman (2015), *Minimax estimation of functionals of discrete distributions*.
- What about the general ergodic case?
  - Universal compression (some upper bound, researched).
  - Plug-in estimator (some lower bound, not researched yet).

Some notation

Entropy of a distribution: $H(p) = -\sum_{w:p(w)>0} p(w) \log p(w)$.

True distribution and block entropy:

$$p_k(w) = P(X_{i+1}^{i+k} = w),$$
$$H(k) = H(p_k).$$

Empirical distribution and plug-in estimator:

$$p_k(w, X_1^n) = \frac{1}{\lfloor n/k \rfloor} \sum_{i=1}^{\lfloor n/k \rfloor} 1\left\{ X_{i(k-1)+1}^{ik} = w \right\},$$
$$H(k, X_1^n) = H(p_k(\,\cdot\,, X_1^n)).$$

Some known facts

1. The plug-in estimator is biased and the bias is large:

$$\mathbb{E}\, H(k, X_1^n) \le H(k) \text{ since } \mathbb{E}\, p_k(w, X_1^n) = p_k(w).$$
$$H(k, X_1^n) \le \log \lfloor n/k \rfloor \text{ since } p_k(w, X_1^n) \ge \lfloor n/k \rfloor^{-1}.$$

2. For a fixed block length k and a stationary ergodic process, plug-in estimator is consistent and asymptotically unbiased:

$$\lim_{n\to\infty} H(k, X_1^n) = H(k) \text{ almost surely},$$
$$\lim_{n\to\infty} \mathbb{E}\, H(k, X_1^n) = H(k).$$

Can we estimate the entropy rate $h = \lim_{n\to\infty} H(k)/k$ if we let $k \to \infty$? What $n = n(k)$ should we choose?

A result by Marton and Shields (1994)

For the variational distance

$$|p - q| := \sum_{w} |p(w) - q(w)|,$$

we have

$$\lim_{k\to\infty} \left| p_k - p_k(\cdot, X_1^{n(k)}) \right| = 0,$$

if we put $n(k) \geq 2^{k(h+\epsilon)}$ for: IID processes, irreducible Markov chains, functions of irreducible Markov chains, ψ-mixing processes, and weak Bernoulli processes.

This result suggests that sample size $n(k) \approx 2^{k(h+\epsilon)}$ may be sufficient for estimation of block entropy $H(k)$.

Our result

Theorem

Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary ergodic process over a finite alphabet $\mathbb{X}$. For any $\epsilon > 0$ and $n(k) \geq 2^{k(h+\epsilon)}$, we have

$$\lim_{k\to\infty} \mathbb{E}\, H(k, X_1^{n(k)})/k = h,$$

$$\liminf_{k\to\infty} H(k, X_1^{n(k)})/k = h \text{ a.s.},$$

$$\forall_{\eta>0} \lim_{k\to\infty} P\left(H(k, X_1^{n(k)})/k - h > \eta\right) = 0.$$

This result is established using source coding
in a more general setting than Marton and Shields (1994).

The main idea of the proof

Let $D(k, X_1^n)$ be the number of distinct blocks of length k contained in the sample X_1^n. Formally,

$$D(k, X_1^n) = \left| \left\{ w \in \mathbb{X}^k : \exists_{i \in 1, \ldots, \lfloor n/k \rfloor} X_{(i-1)k+1}^{ik} = w \right\} \right|.$$

Quantity

$$K(k, X_1^n) = 2 \log k + \frac{n}{k} \left(H(k, X_1^n) + 2 \right) + \\ + 3k \log |\mathbb{X}| \left(D(k, X_1^n) + 1 \right)$$

is an upper bound for the length of a k-block code for X_1^n.

Observation: $K(k, X_1^n) \geq nh$ so $H(k, X_1^{n(k)})/k \to h$ if the number of distinct blocks $D(k, X_1^{n(k)})$ grows sufficiently slow.

A new upper bound for the number of distinct blocks

By the Markov inequality,

$$\begin{aligned}\mathbb{E}\, D(k, X_1^n) &\leq \sum_{w \in \mathbb{X}^k} \min\left[1, \mathbb{E}\left(\sum_{i=1}^{n/k} 1\left\{X_{(i-1)k+1}^{i+k} = w\right\}\right)\right] \\ &= \sum_{w \in \mathbb{X}^k} \min\left[1, \frac{n}{k} P(X_1^k = w)\right].\end{aligned}$$

Putting $\sigma(y) = \min[exp(y), 1]$,

$$\begin{aligned}\frac{k}{n}\mathbb{E}\, D(k, X_1^n) &\leq \mathbb{E}\,\sigma\left(-\log P(X_1^k) - \log\frac{n}{k}\right) \\ &\leq \frac{1}{m} + \left(1 - \frac{1}{m}\right)\sigma\left(mH(X_1^k) - \log\frac{n}{k}\right).\end{aligned}$$

Another application of the new bound

$\mathcal{I}$ — shift-invariant algebra.

Theorem

For a stationary process $(X_i)_{i=-\infty}^{\infty}$, *natural numbers* p *and* k, $n = pk$, *and a real number* $m \geq 1$,

$$\frac{H(X_1^n)}{n} - \frac{H(X_1^k|\mathcal{I})}{k} \leq \frac{2}{k} + \frac{2}{n}\log k + 3\log|\mathbb{X}| \times$$
$$\times \left(\frac{1}{m} + \left(1 - \frac{1}{m}\right)\sigma\left(mH(X_1^k|\mathcal{I}) - \log\frac{n}{k}\right) + \frac{k}{n}\right),$$

where $\sigma(y) = \min(\exp(y), 1)$.

The idea of the proof:

$$\frac{H(X_1^n)}{n} - \frac{H(X_1^k|\mathcal{I})}{k} \leq \mathbb{E}\left[\frac{K(k, X_1^n)}{n} - \frac{H(k, X_1^n)}{k}\right].$$

Some open problems

1. Does the equality

$$\lim_{k\to\infty} H(k, X_1^{n(k)})/k = h \text{ a.s.}$$

   hold true in some cases?

2. What happens for $\lim_{k\to\infty} k^{-1} \log n(k) = h$? Can we set $n(k)$ equal to some random stopping time, such as

$$n(k) = 2^{K(X_1^k)},$$

   where $K(X_1^k)$ is a length of a universal code for X_1^k?

3. The plug-in estimator is not optimal in the IID case. Can we propose a better estimator of the entropy rate for an arbitrary ergodic process?

www.ipipan.waw.pl/~ldebowsk