Finite Memory Estimation of Infinite Memory Processes

Imre Csiszár and Zsolt Talata

***Abstract*— Stationary ergodic processes with finite alphabets are approximated by finite memory processes based on an n-length realization of the process. Under the assumptions of summable continuity rate and non-nullness, a rate of convergence in $\bar{d}$-distance is obtained, with explicit constants. Asymptotically, as $n \to \infty$, the result is near the optimum.**

I. Introduction

This paper deals with estimation of stationary ergodic processes based on a sample, an observed finite realization of the process, of length n. We consider the $\bar{d}$-distance between the process and the estimated one. This is one of the most widely used metrics over stationary processes, its properties include that entropy is $\bar{d}$-continuous and the class of ergodic processes is $\bar{d}$-closed [7], [8], [9].

Ornstein and Weiss [8] proved that for stationary processes isomorphic to i.i.d. processes, the empirical distribution of the $k(n)$-length blocks is a strongly consistent estimator of the $k(n)$-length parts of the process in $\bar{d}$-distance if and only if $k(n) \leq (\log n)/h$, where h denotes the entropy of the process.

In this paper, we estimate the n-length part of a process X by a Markov process of order $k(n)$. The transition probabilities of this Markov estimator process are the empirical conditional probabilities. We assume that the process X is non-null, that is, the conditional probabilities of the symbols given the pasts are separated from zero, and that the continuity rate of the process X is summable. These conditions are usually assumed in this area [3], [4], [5], [6]. The summability of the continuity rate implies that the process is isomorphic to an i.i.d. process [1].

We obtain not only an asymptotic rate of convergence result but also an explicit bound on the probability that the $\bar{d}$-distance of the Markov estimator from the process X is greater than ε. For i.i.d. processes, our rate asymptotically almost attains the best rate possible. Our results rely upon those in [4], [5], [6]. We are not aware of prior works addressing specifically the rate of convergence of statistical estimation of processes in $\bar{d}$-distance.

The work of I. Csiszár was supported in part by the Hungarian National Foundation for Scientific Research under Grant K76088. The work of Zs. Talata was supported in part by NSF under Grant DMS 0906929.

I. Csiszár is with the Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, POB 127, H-1364 Budapest, Hungary csiszar@renyi.hu

Zs. Talata is with the Department of Mathematics, University of Kansas, Lawrence, KS 66045-7523, USA talata@math.ku.edu

II. Finite Samples

Let $X = \{X_i, -\infty < i < +\infty\}$ be a stationary ergodic stochastic process with finite alphabet A. We write $X_i^j = X_i, \dots, X_j$ and $x_i^j = x_i, \dots, x_j \in A^{j-i+1}$ for $j \geq i$. For two strings $x_1^j \in A^j$ and $y_1^m \in A^m$, $x_1^j y_1^m$ denotes their concatenation $x_1, \dots, x_j, y_1, \dots, y_m \in A^{j+m}$. Write

$$P(x_j^m) = \Pr\{\, X_j^m = x_j^m \,\}$$

and, if $P(x_{-m}^{-1}) > 0$,

$$P(a|x_{-m}^{-1}) = \Pr\{\, X_0 = a \mid X_{-m}^{-1} = x_{-m}^{-1} \,\}.$$

The process X is called *non-null* if always $P(x_{-m}^{-1}) > 0$ and, in addition,

$$p_{\inf} = \inf_{m \geq 1} \; \min_{a \in A,\, x_{-m}^{-1} \in A^m} P(a|x_{-m}^{-1}) > 0.$$

The *continuity rate* of the process X is

$$\gamma(k) = \sup_{m \geq k} \; \max_{a \in A} \; \max_{x_{-m}^{-1},\, y_{-m}^{-1} \in A^m :\, x_{-k}^{-1} = y_{-k}^{-1}} \left| P(a|x_{-m}^{-1}) - P(a|y_{-m}^{-1}) \right|.$$

Let $\gamma = \sum_{k=1}^{\infty} \gamma(k)$ and

$$\alpha = \frac{1}{\prod_{j=1}^{+\infty} (1 - \gamma(j))}$$

and

$$\beta(k) = \frac{1 - (1 - |A|\gamma(k))^k}{k\gamma(k) \prod_{j=1}^{+\infty} (1 - |A|\gamma(j))^2}.$$

If $\gamma < +\infty$, the process X is said to have summable continuity rate. In this case, $\alpha < +\infty$ and $\beta(k) \leq \beta$ for some constant $0 < \beta < +\infty$.

The per-letter Hamming distance between two strings x_1^n and y_1^n is

$$d_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \neq y_i),$$

where

$$\mathbb{I}(a \neq b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{if } a = b. \end{cases}$$

The $\bar{d}$*-distance* between two random sequences X_1^n and Y_1^n with distributions P_X and P_Y, respectively, is defined by

$$\bar{d}(X_1^n, Y_1^n) = \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}}\, d_n(\tilde{X}_1^n, \tilde{Y}_1^n),$$

where the minimum is taken over all the joint distributions $\mathbb{P}$ of $\tilde{X}_1^n$ and $\tilde{Y}_1^n$ whose marginals are equal to P_X and P_Y. The *Kullback–Leibler distance* of these sequences is

$$D(X_1^n \| Y_1^n) = \sum_{a_1^n \in A^n} P_X(a_1^n) \frac{P_X(a_1^n)}{P_Y(a_1^n)}.$$

The process X is a *Markov chain* of order k if for each $n > k$ and $x_1^n \in A^n$

$$P(x_1^n) = P(x_1^k) \prod_{i=k+1}^{n} P(x_i | x_{i-k}^{i-1}),$$

where $P(x_1^k)$ is called initial distribution and $P(\cdot|\cdot)$ is called transition probability matrix. The case $k = 0$ corresponds to an i.i.d. process. The k-order Markov approximation of a process X is the Markov chain, denoted by $X[k]$, of order k whose transition probabilities are $P(a|a_1^k)$, $a \in A$, $a_1^k \in A^k$.

Let $N_n(a_1^k)$ denote the number of occurrences of the string a_1^k in the sample X_1^n

$$N_n(a_1^k) = \left| \left\{ i : X_{i+1}^{i+k} = a_1^k, 0 \le i \le n-k \right\} \right|.$$

The empirical probability of the string a_1^k is

$$\hat{P}_n(a_1^k) = \frac{1}{n-k+1} N_n(a_1^k).$$

The *empirical k-order Markov approximation* of a process X based on the sample X_1^n is the stationary Markov chain, denoted by $\hat{X}[k]$, of order k whose transition probabilities are the empirical conditional probabilities

$$\hat{P}_n(a|a_1^k) = \frac{N_n(a_1^k a)}{N_{n-1}(a_1^k)}, \quad a \in A,\ a_1^k \in A^k.$$

If the initial distribution of a stationary Markov chain with these transition probabilities is not unique, then any of these initial distributions can be taken.

Theorem 1: Let X be a non-null stationary ergodic process with summable continuity rate. Then, for any $\varepsilon > 0$, the empirical k-order Markov approximation of the process satisfies

$$\Pr\left\{ \bar{d}\left(X_1^n, \hat{X}[k]_1^n\right) > \varepsilon \right\} \le 2e^{1/e} |A|^{k+2} \exp\left\{ -\frac{(n-k)\, p_{\inf}^{2k+2}}{16e|A|^3(|A|\gamma + p_{\inf})(k+1)} \left[\left(\frac{\varepsilon - \beta(k) p_{\inf}^{-2}\, \gamma(k)}{\alpha+1} \right)^2 - \frac{k\, |\log p_{\inf}|}{2n} \right] \right\}.$$

Proof: The proof relies upon the results in [4], [5], [6]. See [2] for the details. ■

III. Asymptotic Results

Based on the results for finite samples of size n, we will derive asymptotic bounds as $n \to \infty$.

Theorem 2: Let X be a non-null stationary ergodic process with summable continuity rate. Then, for any $\mu > 0$, the empirical $(\nu \log n)$-order Markov approximation of the process satisfies

$$\bar{d}\left(X_1^n, \hat{X}[\nu \log n]_1^n\right) \le \frac{\beta(\nu \log n)}{p_{\inf}^2} \gamma(\nu \log n) + \frac{1}{n^{1/2-\mu}}$$

eventually almost surely as $n \to \infty$, if

$$\nu < \frac{\mu}{|\log p_{\inf}|}.$$

Proof: For $\varepsilon = \beta(\nu \log n) p_{\inf}^{-2}\, \gamma(\nu \log n) + 1/(n^{1/2-\mu})$ and $k = \nu \log n$, Theorem 1 yields

$$\Pr\left\{ \bar{d}\left(X_1^n, \hat{X}[\nu \log n]_1^n\right) > \frac{\beta(\nu \log n)}{p_{\inf}^2} \gamma(\nu \log n) + \frac{1}{n^{1/2-\mu}} \right\}$$
$$\le 2e^{1/e} |A|^{2+\nu \log n} \exp\left\{ -\frac{p_{\inf}^2}{16e|A|^3(|A|\gamma + p_{\inf})(\alpha+1)^2} \frac{(n - \nu \log n)\, n^{-2\nu |\log p_{\inf}|}}{(1+\nu \log n) n} \left[n^{2\mu} - \frac{\nu\, |\log p_{\inf}| (\alpha+1)^2 \log n}{2} \right] \right\},$$

which is summable in n and allows application of the Borel–Cantelli lemma. ■

Remark 1: While the parameter ν in Theorem 2 depends on the actual process X, it holds for all X satisfying the hypothesis that any $k_n \to +\infty$ with $k_n = o(\log n)$ guarantees $\bar{d}\left(X_1^n, \hat{X}[k_n]_1^n\right) \to 0$, almost surely.

Remark 2: In Theorem 2, in the upper bound the first term is the bias due to the error of the approximation of the process by a Markov chain. The second term is the variation due to the error of the estimation of the parameters of the Markov chain based on a sample.

Remark 3: If the process is i.i.d., then $\bar{d}\left(X_1^n, \hat{X}[0]_1^n\right) = \bar{d}\left(X[0]_1^n, \hat{X}[0]_1^n\right)$ equals the variation distance [9], whose order is $n^{-1/2}$. Since in this case $\gamma(k) = 0$, $k = 0, 1, \ldots$, the order of the upper bound in Theorem 2 cannot be improved significantly.

Corollary 1: Let X be a non-null stationary ergodic process with continuity rate $\gamma(k) = c' \exp(-ck)$, $k = 1, 2, \ldots$, where $c', c > 0$ are constants. Then, for any $\mu > 0$, the empirical $(\nu \log n)$-order Markov approximation of the process satisfies

$$\bar{d}\left(X_1^n, \hat{X}[\nu \log n]_1^n\right) \le \frac{2}{n^{1/2-\mu}}$$

eventually almost surely as $n \to \infty$, if

$$\nu < \frac{\mu}{|\log p_{\text{inf}}|}$$

and

$$c' \leq \frac{p_{\text{inf}}^2}{\beta} \quad \text{and} \quad c \geq \frac{1}{\nu}\left(\frac{1}{2} - \mu\right).$$

REFERENCES

[1] H. Berbee, "Chains with infinite connections: uniqueness and Markov representation," *Probab. Theory Related Fields,* vol. 76, no. 2, pp. 243–253, 1987.

[2] I. Csiszár and Zs. Talata, "On Rate of Convergence of Statistical Estimation of Stationary Ergodic Processes," *IEEE Trans. Inform. Theory,* accepted.

[3] D. Duarte, A. Galves, and N. Garcia, "Markov approximation and consistent estimation of unbounded probabilistic suffix trees," *Bull. Braz. Math. Soc, New Series,* vol. 37, no. 4, pp. 581–592, 2006.

[4] R. Fernández and A. Galves, "Markov Approximations of Chains of Infinite Order," *Bull. Braz. Math. Soc, New Series,* vol. 33, pp. 295–306, 2002.

[5] A. Galves and F. Leonardi, "Exponential inequalities for empirical unbounded context trees," *In and Out of Equilibrium 2,* (Sidoravicius, V.; Vares, M.E., Eds.), Progress in Probability, vol. 60, pp. 257-270, 2008.

[6] K. Marton, "Measure Concentration for a Class of Random Processes," *Probab. Theory Relat. Fields,* vol. 110, pp. 427–439, 1998.

[7] D. S. Ornstein, "An Application of Ergodic Theory to Probability Theory," *Ann. Probab.* vol. 1, no. 1, pp. 43–58, 1973.

[8] D. S. Ornstein and B. Weiss, "How sampling reveals a process," *Ann. Probab.* vol. 18, no. 3, pp. 905–930, 1990.

[9] P. Shields, *The ergodic theory of discrete sample paths.* Providence, RI: American Mathematical Society, 1996.