Penalized Maximum Likelihood Methods in Process Estimation

Zsolt Talata*

Abstract
Stationary ergodic processes with finite alphabets are estimated by finite memory processes from a sample, an n-length realization of the process, where the memory depth of the estimator process is also estimated from the sample using penalized maximum likelihood (PML). Under some assumptions on the continuity rate and the assumption of non-nullness, a rate of convergence in $\bar{d}$-distance is obtained, with explicit constants. The results show optimality of the PML Markov order estimator for not necessarily finite memory processes.

Key Words: finite memory estimator, Markov approximation, infinite memory, rate of convergence, penalized maximum likelihood, stationary ergodic process

1. Introduction

This paper is concerned with the problem of estimating stationary ergodic processes with finite alphabet from a sample, an observed length n realization of the process, with the $\bar{d}$-distance being considered between the process and the estimated one. The $\bar{d}$-distance was introduced by Ornstein [13] and became one of the most widely used metrics over stationary processes. Two stationary processes are close in $\bar{d}$-distance if there is a joint distribution whose marginals are the distributions of the processes such that the marginal processes are close with high probability (see Section 4 for the formal definition). The class of ergodic processes is $\bar{d}$-closed and entropy is $\bar{d}$-continuous, which properties do not hold for the weak topology [18].

Ornstein and Weiss [14] 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.

Csiszár and Talata [8] estimated the n-length part of a stationary ergodic process X by a Markov process of order k_n. The transition probabilities of this Markov estimator process are the empirical conditional probabilities, and the order $k_n \to +\infty$ does not depend on the sample. They obtained a rate of convergence of the Markov estimator to the process X in $\bar{d}$-distance, which consists of two terms. The first one 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 from a sample.

Model selection methods in various settings seek a tradeoff between the bias and the variation. There are classical results aiming at identifying the balance, see for instance the indices of resolvability in the work by Barron [2, 3, 4].

In this paper, the order k_n of the Markov estimator process is estimated from the sample. Some of the subsequent results were also presented at the IEEE International Symposium on Information Theory, Cambridge, Massachusetts, July 2012. The complete proofs of all of the results given in this paper are contained in [20]. The penalized maximum

*Department of Mathematics, University of Kansas, 1460 Jayhawk Boulevard, Lawrence, KS 66045

likelihood (PML) is a natural generalization of the Bayesian information criterion, that is often regarded as an approximation of the criteria derived from the minimum description length principle (see Section 3 for the formal definition). For the order estimation, PML with general penalty term is used. The resulted Markov estimator process finds a tradeoff between the bias and the variation as it uses shorter memory for faster memory decays of the process X. If the process X is a Markov chain, the PML order estimation recovers its order asymptotically with a wide range of penalty terms.

Not only an asymptotic rate of convergence result is obtained but also an explicit bound on the probability that the $\bar{d}$-distance of the above Markov estimator from the process X is greater than ε. It is assumed 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 and the restricted continuity rate is uniformly convergent. These conditions are usually assumed in this area [6, 9, 10, 12]. The summability of the continuity rate implies that the process is isomorphic to an i.i.d. process [5].

2. Infinite Memory Processes

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

$$P(x_i^j) = \Pr(X_i^j = x_i^j)$$

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

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

For $m = 0$, $P(a|x_{-m}^{-1}) = P(a)$.

The process X is called *non-null* if

$$p_{\text{inf}} = \min_{a \in A} \inf_{x_{-\infty}^{-1} \in A^\infty} P(a|x_{-\infty}^{-1}) > 0.$$

The *continuity rate* of the process X is

$$\gamma(k) = \sup_{x_{-\infty}^{-1} \in A^\infty} \sum_{a \in A} \left| P(a|x_{-k}^{-1}) - P(a|x_{-\infty}^{-1}) \right|.$$

If $\sum_{k=1}^{\infty} \gamma(k) < +\infty$, then the process X is said to have *summable continuity rate*.

Remark 1. Since for any $x_{-k}^{-1} \in A^k$ and $z_{-m}^{-k-1} \in A^{m-k}$, $m \geq k$,

$$\inf_{x_{-\infty}^{-k-1}} P(a|x_{-\infty}^{-1}) \leq P(a|z_{-m}^{-k-1} x_{-k}^{-1}) \leq \sup_{x_{-\infty}^{-k-1}} P(a|x_{-\infty}^{-1}),$$

the above definition of continuity rate is equivalent to

$$\gamma(k) = \sup_{i > k} \max_{x_{-i}^{-1} \in A^i} \sum_{a \in A} \left| P(a|x_{-k}^{-1}) - P(a|x_{-i}^{-1}) \right|.$$

The *restricted continuity rate* of the process X is

$$\gamma(k|m) = \max_{x_{-m}^{-1} \in A^m} \sum_{a \in A} \left| P(a|x_{-k}^{-1}) - P(a|x_{-m}^{-1}) \right|, \quad k < m.$$

Similarly to Remark 1, note that the above definition is equivalent to

$$\gamma(k|m) = \max_{k<i\leq m} \max_{x_{-i}^{-1} \in A^i} \sum_{a \in A} \left| P(a|x_{-k}^{-1}) - P(a|x_{-i}^{-1}) \right|.$$

Hence, $\lim_{m\to+\infty} \gamma(k|m) = \gamma(k)$ for any fixed k. We say that the process X has *uniformly convergent restricted continuity rate* with parameters θ_1, θ_2, k_θ if

$$\gamma(k)^{\theta_1} \leq \gamma(k \,|\, \lceil \theta_2 k \rceil) \quad \text{if } k \geq k_\theta, \text{ for some } \theta_1 \geq 1, \theta_2 > 1.$$

The k-order *entropy* of the process X is

$$H_k = -\sum_{a_1^k \in A^k} P(a_1^k) \, \log P(a_1^k), \quad k \geq 1,$$

and the k-order *conditional entropy* is

$$h_k = -\sum_{a_1^{k+1} \in A^{k+1}} P(a_1^{k+1}) \, \log P(a_{k+1}|a_1^k), \quad k \geq 0.$$

Logarithms are to the base 2. It is well-known for stationary processes that the conditional entropy h_k is a non-negative decreasing function of k, therefore its limit exists as $k \to +\infty$. The *entropy rate* of the process is

$$\bar{H} = \lim_{k\to+\infty} h_k = \lim_{k\to+\infty} \frac{1}{k} H_k.$$

Note that $h_k - \bar{H} \geq 0$ for any $k \geq 0$.

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

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

where $P(x_1^k)$ is called initial distribution and $\left\{ P(a|a_1^k), a \in A, a_1^k \in A^k \right\}$ is called transition probability matrix. The case $k = 0$ corresponds to i.i.d. processes. The process X is of *infinite memory* if it is not a Markov chain for any order $k < +\infty$. For infinite memory processes, $h_k - \bar{H} > 0$ for any $k \geq 0$.

In this paper, we consider statistical estimates based on a sample X_1^n, an n-length part of the process. Let $N_n(a_1^k)$ denote the number of occurrences of the string a_1^k in the sample X_1^n

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

For $k \geq 1$, the empirical probability of the string a_1^k is

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

and the empirical conditional probability of $a \in A$ given a_1^k is

$$\hat{P}(a_{k+1}|\, a_1^k) = \frac{N_n(a_1^{k+1})}{N_{n-1}(a_1^k)}.$$

For $k = 0$, $\hat{P}(a_{k+1}|\, a_1^k) = \hat{P}(a_{k+1})$. The k-order *empirical entropy* is

$$\hat{H}_k(X_1^n) = -\sum_{a_1^k \in A^k} \hat{P}(a_1^k)\, \log \hat{P}(a_1^k), \quad 1 \le k \le n,$$

and the k-order *empirical conditional entropy* is

$$\hat{h}_k(X_1^n) = -\sum_{a_1^{k+1} \in A^{k+1}} \hat{P}(a_1^{k+1})\, \log \hat{P}(a_{k+1}|\, a_1^k), \quad 0 \le k \le n-1.$$

3. Penalized Maximum Likelihood

An information criterion assigns a score to each hypothetical model (here, Markov chain order) based on a sample, and the estimator will be that model whose score is minimal.

Definition 2. *For an information criterion*

$$\mathrm{IC}_{X_1^n}(\,\cdot\,) : \mathbb{N} \to \mathbb{R}^+,$$

the Markov order estimator bounded by $r_n < n$, $r_n \in \mathbb{N}$, *is*

$$\hat{k}_{IC}(X_1^n|\, r_n) = \arg\min_{0 \le k \le r_n} \mathrm{IC}_{X_1^n}(k).$$

Remark 3. Here, the number of candidate Markov chain orders based on a sample is finite, therefore the minimum is attained. If the minimizer is not unique, the smallest one will be taken as $\arg\min$.

A popular approach to choosing information criteria is the minimum description length (MDL) principle [15, 4]. In particular, the normalized maximum likelihood (NML) [19] and the Krichevsky–Trofimov (KT) [11] code lengths are natural information criteria because the former minimizes the worst case maximum redundancy for the model class of k-order Markov chains, while the latter does so, up to an additive constant, with the average redundancy. The Bayesian information criterion (BIC) [16] can be regarded as an approximation of the NML and KT code lengths. The family of penalized maximum likelihood (PML) is a generalization of BIC.

The likelihood of the sample X_1^n with respect to a k-order Markov chain model of the process X with some transition probability matrix $\left\{ Q(a_{k+1}|a_1^k), a_{k+1} \in A, a_1^k \in A^k \right\}$, by (1), is

$$P'(X_1^n) = P'(X_1^k) \prod_{a_1^{k+1} \in A^{k+1}} Q(a_{k+1}|\, a_1^k)^{N_n(a_1^{k+1})}.$$

For $0 \le k < n$, the *maximum likelihood* is the maximum in $Q(a_{k+1}|a_1^k)$ of the second factor above, which equals

$$\mathrm{ML}_k(X_1^n) = \prod_{a_1^{k+1} \in A^{k+1}} \hat{P}(a_{k+1}|\, a_1^k)^{N_n(a_1^{k+1})}.$$

Note that $\log \mathrm{ML}_k(X_1^n) = -(n-k)\hat{h}_k(X_1^n)$.

Definition 4. *Given a penalty function pen(n), a non-decreasing function of the sample size n, for a candidate order $0 \le k < n$ the PML criterion is*

$$\begin{aligned}\mathrm{PML}_{X_1^n}(k) &= -\log \mathrm{ML}_k(X_1^n) + (|A|-1)|A|^k \mathit{pen}(n) \\ &= (n-k)\,\hat{h}_k(X_1^n) + (|A|-1)|A|^k \mathit{pen}(n).\end{aligned}$$

The k-order Markov chain model of the process X is described by the conditional probabilities $\left\{ Q(a_{k+1}|a_1^k), a_{k+1} \in A, a_1^k \in A^k \right\}$, and $(|A|-1)|A|^k$ of these are free parameters.

The second term of the PML criterion, which is proportional to the number of free parameters of the k-order Markov chain model, is increasing in k. The first term, for a given sample, is known to be decreasing in k. Hence, minimizing the criterion yields a tradeoff between the goodness of fit of the sample to the model and the complexity of the model.

Remark 5. If $\mathrm{pen}(n) = \frac{1}{2}\log n$, the PML criterion is called *Bayesian information criterion* (BIC), and if $\mathrm{pen}(n) = 1$, *Akaike information criterion* (AIC) [1].

4. Statistical Estimation of Processes

The problem of statistical estimation of stationary ergodic processes by finite memory processes is considered, and the following distance is used. The per-letter Hamming distance between two strings x_1^n and y_1^n is

$$d_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \ne y_i), \quad \text{where } \mathbb{I}(a \ne b) = \begin{cases} 1 & \text{if } a \ne b \\ 0 & \text{if } a = b \end{cases},$$

and the $\bar{d}$*-distance* between two random sequences X_1^n and Y_1^n 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 the distributions of X_1^n and Y_1^n.

The process X is estimated by a Markov chain of order $k = k_n$ from the sample in the following way.

Definition 6. *The* empirical k-order Markov estimator *of a process X based on the sample X_1^n is the stationary Markov chain, denoted by $\hat{X}[k]$, of order k with transition probability matrix $\left\{ \hat{P}(a_{k+1}|\, a_1^k), a_{k+1} \in A, a_1^k \in A^k \right\}$. 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.*

The order k of the empirical Markov estimator $\hat{X}[k]$ is estimated from the sample, using the PML criterion. The estimated order needs to be bounded to guarantee an accurate assessment of the memory decay of the process.

The optimal order can be smaller than the upper bound if the memory decay of the process is sufficiently fast. Define

$$K_n(r_n, \gamma, f(n)) = \min\left\{ \lfloor r_n \rfloor,\ k \ge 0 : \gamma(k) < f(n) \right\},$$

where $f(n) \searrow 0$ and $r_n \nearrow \infty$. Since γ is a decreasing function, K_n increases in n but does not exceed r_n. It is less than r_n if γ vanishes sufficiently fast, and then the faster γ vanishes, the slower K_n increases.

The process estimation result of the paper is the following.

Theorem 7. *For any non-null stationary ergodic process with summable continuity rate and uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ, and for any $\mu_n > 0$, the empirical Markov estimator of the process with the order estimated by the bounded PML Markov order estimator $\hat{k}_n = \hat{k}_{PML}(X_1^n | \, \eta \log n)$, $\eta > 0$, with $\frac{1}{2}\log n \leq$ pen$(n) \leq \mathcal{O}(\sqrt{n})$ satisfies*

$$\Pr\left(\bar{d}\left(X_1^n, \hat{X}[\hat{k}_n]_1^n\right) > \frac{\beta_2}{p_{inf}^2} \max\left\{\bar{\gamma}\left(\left\lfloor \frac{\eta}{\theta_2}\log n\right\rfloor\right), n^{-\frac{1}{4\theta_1}\left(1-4\eta\log\frac{|A|^4}{p_{inf}}\right)}\right\} + \frac{1}{n^{1/2-\mu_n}}\right)$$
$$\leq \exp\left(-c_4 \, 4^{\mu_n \log n - |\log p_{inf}| \left(K_n\left(\eta\log n, \bar{\gamma}, \frac{c}{n} pen(n)\right) + \frac{\log\log n}{\log|A|}\right)}\right)$$
$$+ \exp\left(-\frac{c_5\eta^3}{\log n} n^{\eta \, 2\log|A|}\right) + 2^{-s_n pen(n)}$$

if $n \geq n_0$, where $c > 0$ is an arbitrary constant, $s_n \to \infty$ and $\beta_2, c_4, c_5, n_0 > 0$ are constants depending only on the distribution of the process.

Remark 8. If the process X is a Markov chain of order k, then the restricted continuity rate is uniformly convergent with parameters $\theta_1 = 1$, $\theta_2 > 1$ arbitrary (arbitrarily close to 1), $k_\theta = k+1$, and if n is sufficiently large, $K_n = k$ and

$$\max\left\{\bar{\gamma}\left(\left\lfloor \frac{\eta}{\theta_2}\log n\right\rfloor\right), n^{-\frac{1}{4\theta_1}\left(1-4\eta\log\frac{|A|^4}{p_{\text{inf}}}\right)}\right\} = n^{-\frac{1}{4\theta_1}\left(1-4\eta\log\frac{|A|^4}{p_{\text{inf}}}\right)}.$$

An application of the Borel–Cantelli lemma in Theorem 7 yields the following asymptotic result.

Corollary 9. *For any non-null stationary ergodic process with summable continuity rate and uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ, the empirical Markov estimator of the process with the order estimated by the bounded PML Markov order estimator $\hat{k}_n = \hat{k}_{PML}(X_1^n | \, r_n)$ with $\frac{1}{2}\log n \leq$ pen$(n) \leq \mathcal{O}(\sqrt{n})$ and*

$$\frac{5\log\log n}{2\log|A|} \leq r_n \leq o(\log n)$$

satisfies

$$\bar{d}\left(X_1^n, \hat{X}[\hat{k}_n]_1^n\right) \leq \frac{\beta_2}{p_{inf}^2} \max\left\{\bar{\gamma}\left(\left\lfloor \frac{r_n}{\theta_2}\right\rfloor\right), n^{-\frac{1}{4\theta_1}}\right\} + \frac{(\log n)^{c_6}}{\sqrt{n}} 2^{|\log p_{inf}| K_n\left(r_n, \bar{\gamma}, \frac{c}{n} pen(n)\right)}$$

eventually almost surely as $n \to +\infty$, where $c > 0$ is an arbitrary constant, and $\beta_2, c_6 > 0$ are constants depending only on the distribution of the process.

Remark 10. In Corollary 9, 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 order and the parameters of the Markov chain based on a sample. If the memory decay of the process is slow, the bias is essentially $\gamma(\lfloor r_n/\theta_2 \rfloor)$, and the variance is maximal. If the memory decay is sufficiently fast, then the rate of the estimated order $\hat{k}_n$ and the rate of K_n are smaller, therefore the variance term is smaller while the bias term is smaller as well. The result, however, shows the optimality of the PML Markov order estimator in the sense that it selects an order which is small enough to allow the variance to decrease but large enough to keep the bias below a polynomial threshold.

5. Discussion

In this paper, stationary ergodic processes have been estimated by finite memory processes from a sample, where the memory depth of the estimator process is also estimated from the sample using PML. Under some assumptions on the process, a rate of convergence in $\bar{d}$-distance has been obtained. The results show an optimality of the PML Markov order estimator for not necessarily finite memory processes. In [20], the PML Markov order estimator has been shown to be consistent with the oracle-type order estimate under some assumptions on the process. The consistency result requires larger penalty terms for PML than the process estimation result. This reflects the expectation that the estimation of the structure parameter needs larger penalty terms than the estimation of the sampling distribution; see, for example, [17] and [16].

Acknowledgment

In this research, Talata is supported by the NSF grant DMS 0906929.

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *2nd International Symposium on Information Theory,* (Tsahkadsor, 1971), pp. 267–281, Akadémia Kiadó, Budapest, 1972.

[2] A. R. Barron, L. Birgé and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields,* vol. 113, no. 3, pp. 301–413, 1999.

[3] A. R. Barron, T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory,* vol. 37, pp. 1034–1054, Jul. 1991.

[4] A. R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory,* vol. 44, pp. 2743–2760, Oct. 1998.

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

[6] X. Bressaud, R. Fernandez, and A. Galves, "Speed of $\bar{d}$-convergence for Markov approximations of chains with complete connections. A coupling approach," *Stochastic Process. Appl.,* vol. 83, pp. 127–138, 1999.

[7] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators," *IEEE Trans. Inform. Theory,* vol. 48, pp. 1616–1628, Jun. 2002.

[8] I. Csiszár and Zs. Talata, "On Rate of Convergence of Statistical Estimation of Stationary Ergodic Processes," *IEEE Trans. Inform. Theory,* vol. 56, pp. 3637–3641, 2010.

[9] 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.

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

[11] R. E. Krichevsky and V. K. Trofimov, “The performance of universal encoding,” *IEEE Trans. Inform. Theory,* vol. 27, pp. 199–207, Mar. 1981.

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

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

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

[15] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* Singapore: World Scientific, 1989.

[16] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.,* vol. 6, pp. 461–464, 1978.

[17] R. Shibata, “Asymptotically efficient selection of the order of the model for estimating parameters of a linear process,” *Ann. Statist.,* vol. 8, pp. 147–164, 1980.

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

[19] J. Shtarkov, “Coding of discrete sources with unknown statistics,” in *Topics in information theory (Second Colloq., Keszthely, 1975),* pp. 559–574. Colloq. Math. Soc. János Bólyai, vol. 16, North-Holland, Amsterdam, 1977.

[20] Zs. Talata, “Divergence Rates of Markov Order Estimators and Their Application to Statistical Estimation of Stationary Ergodic Processes,” manuscript.