Confidence Sets in Time–Series Filtering

Boris Ryabko
Institute of Computational Technology
of Siberian Branch of Russian Academy of Science
Siberian State University of Telecommunications and Informatics,
Novosibirsk, Russia
boris@ryabko.net

Daniil Ryabko
INRIA Lille,
daniil@ryabko.net

***Abstract*—The problem of filtering of finite–alphabet stationary ergodic time series is considered. A method for constructing a confidence set for the (unknown) signal is proposed, such that the resulting set has the following properties: First, it includes the unknown signal with probability γ, where γ is a parameter supplied to the filter. Second, the size of the confidence sets grows exponentially with the rate that is asymptotically equal to the conditional entropy of the signal given the data. Moreover, it is shown that this rate is optimal.**

I. Introduction

The problem of estimating a discrete signal $X_1, \dots, X_t$ from a noisy version $Z_1, \dots, Z_t$ has attracted attention of many researchers due to its great importance for statistics, computer science, image processing, astronomy, biology, cryptography, information theory and many other fields.

The main attention is usually focused on developing methods of estimation (denoising, or filtering) of the unknown signal, with the performance measured under a given fidelity criterion; see [7], [8] and references therein. Such an approach is close in spirit to the problem of point estimation in statistics.

An alternative approach, often considered in mathematical statistics, is that of constructing confidence sets. That is, one tries to use the data to construct a set that includes the unknown parameter (in our case, the signal) with a prescribed probability, while trying to keep the size of the set as small as possible (some classical examples of the use of this method in statistics in can be found in, e.g., [4]). Such a set is usually constructed as the set of most likely values of the parameter. The reason why this approach is of interest is as follows. In the presence of noise, the exact recovery of the signal is typically impossible, and thus, in such cases, any of its estimates is necessarily imperfect. The choice of a particular estimated signal from many likely estimates is largely arbitrary. Moreover, the optimal choice may depend on the specific application involved. The confidence–set approach effectively abstracts from the problem of choosing the "best" estimate, proposing, instead, a set of estimates. The performance of a method is then characterized by the size of the confidence set (depending on the confidence level). This is the approach and the problems considered in this work.

We consider a model in which the underlying noiseless signal and the resulting corrupted (noisy) signal (and thus the channel) are assumed to be stationary ergodic processes with finite alphabets. It is assumed that the probability distributions of the noiseless signal and the noisy channel are known. (Obviously, in such a case the distribution of the corrupted signal is known, too.) The results that we obtain establish the optimal rate of growth (with respect to time, or to the length of the signal) of the size of the confidence set, as well as a method for constructing such a set.

Let us consider an example that illustrates our approach and exposes the notation. Let the signal be binary (with the alphabet $\{0, 1\}$), and suppose that it is transmitted through a memoryless binary erasure channel (e.g. [1]). The binary erasure channel with erasure probability π is defined as a channel with binary input, ternary output (with the alphabet $\{0, 1, *\}$), and the probability of erasure π. The channel replaces each input symbol 0 or 1 with the (output) symbol $*$ with probability π (erasure), and places the input signal in the output otherwise (that is, with probability $1 - \pi$).

Suppose that the noiseless sequence is generated by an i.i.d. source P and $P\{X_i = 0\} = 0.9$, and let the erasure probability be any $\pi \in (0, 1)$. Suppose that the corrupted by noise sequence is as follows:

$$Z_1 ... Z_4 = 0 * 1 * \ .$$

Then we have the following probability distribution for the lossless signal:

$$P(\{X_1 ... X_4 = 0010\}) = 0.81,$$

$$P(\{X_1 ... X_4 = 0110\}) = 0.09,$$

$$P(\{X_1...X_4 = 0011\}) = 0.09,$$
$$P(\{X_1...X_4 = 0111\}) = 0.01.$$

If we take the confidence level $\gamma = 0.99$, the confidence set will contain three following sequences: $\{0010, 0110, 0011\}$.

The goal of this paper is to describe a construction of confidence sets and to give an estimate of their size, for the case when the signal and noise are stationary ergodic processes with finite alphabets. It is shown that for any $\gamma \in (0,1)$ the size of the confidence set grows exponentially with the rate $h(X|Z)$, where $h(X|Z)$ is the limit (conditional) Shannon entropy. Moreover, we prove that this rate is minimal, which means that the suggested method of constructing confidence sets is asymptotically optimal.

It is worth noting that the information theory is deeply connected with mathematical statistics and signal processing; see, for example, [1], [2], [6], [9], [10], [11], [12], [13] and [7], [8], [5], correspondingly. In this paper a new connection of this kind is established: it is shown that the Shannon entropy determines the rate of growth of the size of the confidence set for the signal, given its version corrupted by stationary noise.

II. The confidence sets and their properties

We consider the case where the signal $X = X_1, X_2, \dots$ and its noisy version $Z = Z_1, Z_2, \dots$ are described by stationary ergodic processes with finite alphabets $\mathbf{X}$ and $\mathbf{Z}$ respectively. It is assumed that probability distributions of both processes are known, and, hence, the statistical structure of the noise corrupting the signal $X = X_1, X_2, \dots$ is known, too. Introduce the short-hand notation $X_{1..t}$ for $X_1, \dots, X_t$, and analogously for Z.

Informally, for any $\gamma \in (0,1)$ and any sequence $Z_1, \dots, Z_t$ we define the confidence set $\Psi^t_\gamma(Z_1, Z_2, \dots, Z_t)$ as follows: the set contains sequences $x_1, x_2, \dots, x_t$ whose probabilities $P(x_{1..t}|Z_{1..t})$ are maximal and sum to γ. This definition is not precise, since it is possible that the sum can not be made equal to γ exactly. That is why the formal definition of the confidence set will use randomization.

For this purpose, we order all sequences $X_{1..t}$ according their conditional probabilities, in the decreasing order. That is, enumerate all sequences $x_{1..t} \in \mathbf{X}^n$ in such a way that $(a_{1..t}) \in \mathbf{X}^t$ has a smaller index than $(b_{1..t}) \in \mathbf{X}^t$ if either $P(a_{1..t}|Z_{1..t}) > P(b_{1..t}|Z_{1..t})$, or $P(a_{1..t}|Z_{1..t}) = P(b_{1..t}|Z_{1..t})$ and $(a_{1..t})$ is lexicographically less than $(b_{1..t})$. Let j be the integer for which $\sum_{i=1}^{j-1} P(x^i_{1..t}|Z_{1..t}) \le \gamma$ and $\sum_{i=1}^{j} P(x^i_{1..t}|Z_{1..t}) > \gamma$. If $\sum_{i=1}^{j-1} P(x^i_{1..t}|Z_{1..t}) = \gamma$, then define $\Psi^t_\gamma(Z_{1..t})$ as the set $\{x^1_{1..t}, \dots, x^{j-1}_{1..t}\}$. Otherwise, $\Psi^t_\gamma(Z_{1..t})$ also contains $j-1$ first elements, and additionally the element $x^j_{1..t}$ with probability $(\gamma - \sum_{i=1}^{j-1} P(x^i_{1..t}|Z_{1..t}))/P(x^j_{1..t}|Z_{1..t})$. (Note that this procedure is commonly used in mathematical statistics for making the confidence level exactly γ.) When talking about the sizes of the confidence sets we refer to their expected (with respect to the randomization) size.

Next, we estimate the size of the described confidence set.

Theorem 1. *Let an (unknown) signal $X = X_1X_2, \dots$ and its noisy version $Z = Z_1Z_2, \dots$ be stationary ergodic processes with finite alphabets. Then, for every $\gamma \in (0,1)$, all $t \in \mathbb{N}$ and almost every $Z_1, \dots, Z_t$ the confidence set $\Psi^t_\gamma(Z_1, \dots, Z_t)$ contains the unknown $(X_1, \dots, X_t)$ with probability γ:*

$$P\{X_{1..t} \in \Psi^t_\gamma(Z_{1..t})\} = \gamma, \tag{1}$$

while, with probability 1, the size of the set $\Psi^t_\gamma(Z_1, \dots, Z_t)$ grows exponentially with the exponent rate that is equal to the conditional entropy:

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbf{E}|\Psi^t_\gamma(Z_1, \dots, Z_t)| = h(X|Z) \ a.s., \tag{2}$$

where the expectation is with respect to the randomization used in constructing the confidence sets.

Proof: The proof of (1) immediately follows from the construction of the set $\Psi^t_\gamma(Z_1Z_2...Z_t)$.

The proof of (2) will be based on the Shannon-McMillan-Breiman theorem [1], [3], which for the conditional entropy implies the following:

Lemma 1 (Shannon-McMillan-Breiman)**.** *$\forall \varepsilon > 0, \forall \delta > 0$, for almost all $Z_1, Z_2, \dots$ there exists n' such that if $n > n'$ then*

$$P\left\{\left|-\frac{1}{n}\log P(X_{1..n}|Z_{1..n}) - h(X|Z)\right| < \varepsilon\right\} \ge 1-\delta. \tag{3}$$

Take any $\varepsilon > 0$ and some constant $\delta > 0$ to be specified later, $n > n'$, and rewrite (3) as follows:

$$P\left(2^{-n(h(X|Z)+\varepsilon)} \le P(X_{1..n}|Z_{1..n}) \le 2^{-n(h(X|Z)-\varepsilon)}\right) \ge 1-\delta. \tag{4}$$

¿From this inequality it follows that there are at least $(1-\delta)2^{n(h(X|Z)-\varepsilon)}$ strings $x_1, \dots, x_n$ such that for each of them we have $P(x_{1..n}|Z_{1..n}) \ge 2^{-n(h(X|Z)+\varepsilon)}$. Therefore, if we fix any δ that satisfies

$$(1-\delta)2^{-\frac{h(X|Z)-\varepsilon}{h(X|Z)+\varepsilon}} \ge \gamma,$$

then we have

$$|\Psi^t_\gamma(Z_{1..n})| \le \gamma 2^{-n(h(X|Z)+\varepsilon)},$$

so that

$$\frac{1}{n}\log|\Psi^t_\gamma(Z_{1..n})| \le h(X|Z)+\varepsilon+O(1/n) \quad (5)$$

for $n > n'$. Having taken into account that (5) holds for every $\varepsilon > 0$ we obtain (2). ∎

III. Optimality of the confidence set

Theorem 2. *Let an (unknown) signal $X = X_1X_2,\dots$ and its noisy version $Z = Z_1Z_2,\dots$ be stationary ergodic processes with finite alphabets $\mathbf{X}$ and $\mathbf{Z}$. Let $\Phi^t_\gamma(Z_{1..t})$, be confidence sets, such that for some $\gamma \in (0,1)$ we have $P\left(X_{1..t} \in \Phi^t_\gamma(Z_{1..t})\right) \ge \gamma$ for all $t \in \mathbb{N}$ and almost all $Z_{1...t} \in \mathbf{Z}^t$. Then, with probability 1,*

$$\liminf_{t\to\infty} \frac{1}{t}\log|\Phi^t_\gamma(Z_1,\dots,Z_t)| \ge h(X|Z). \quad (6)$$

Proof: The proof will use the Shannon-McMillan-Breiman theorem (4). As before, we take any $\varepsilon > 0$ and fix $\delta := \gamma/2$. Then from some n on we have (4). Let Υ be a confidence set for this n and a certain γ. Define

$$\Phi = \Big\{x_{1..n} : 2^{-n(h(X|Z)+\varepsilon)} \le P(x_{1..n}|Z_{1..n}) \le 2^{-n(h(X|Z)-\varepsilon)}\Big\}. \quad (7)$$

By definition, $\sum_{x_{1..n}\in\Upsilon} P(x_{1..n}|Z_{1..n}) \ge \gamma$. ¿From this and (4) we obtain

$$\sum_{x_{1..n}\in\Upsilon\cap\Phi} P(x_{1..n}|Z_{1..n}) \ge \gamma-\delta.$$

¿From this and (7) we get

$$|\Upsilon| \ge |\Upsilon\cap\Phi| \ge (\gamma-\delta)2^{n(h(X|Z)-\varepsilon)}.$$

Hence,

$$\liminf_{t\to\infty}\frac{1}{n}\log|\Upsilon| \ge h(X|Z)-\varepsilon.$$

Since this inequality is true for any confidence set Υ and any $\varepsilon > 0$, we obtain (6). ∎

IV. Discussion

To the best of our knowledge, the problem of constructing a confidence set for the unknown signal was not considered before, that is why there are many quite natural and obvious extensions and generalizations of the present work. First, it is interesting to consider this problem for certain specific classes of distributions of the signal and noise, such as i.i.d. and Markov sources. For these classes of sources it should be possible to obtain rates of convergence in those statements that in this work are only asymptotic, for example in (2).

Second, a natural question is to find a construction of the confidence set for the cases where the signal is multi-dimensional. This is particularly important for applications, many of which are concerned with denoising such objects as photographs or video fragments. Another interesting generalization is the case where the alphabets are (subsets of), for example, the Euclidean space. This generalization can be also interesting from the practical point of view. Finally, the case where statistics of the noise and/or signal are unknown is obviously of great theoretical and practical interest.

Acknowledgments

Boris Ryabko was partially supported by the Russian Foundation for Basic Research (grant no. 09-07-00005). Daniil Ryabko was partially supported by the French Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through CPER 2007-2013, ANR projects EXPLO-RA (ANR-08-COSI-004) and Lampada (ANR-09-EMER-007), and by Pascal-2.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 2006.

[2] I. Csiszar and P.C. Shields. Notes on information theory and statistics. In *Foundations and Trends in Communications and Information Theory*, 2004.

[3] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, NY, USA, 1968.

[4] M.G. Kendall and A. Stuart. *The advanced theory of statistics; Vol.2: Inference and relationship*. London, 1961.

[5] E. Ordentlich, G. Seroussi, S. Verdu, and K. Viswanathan. Universal algorithms for channel decoding of uncompressed sources. *Information Theory, IEEE Transactions on*, 54(5):2243 –2262, May 2008.

[6] J. Rissanen. Universal coding, information, prediction, and estimation. *Information Theory, IEEE Transactions on*, 30(4):629 – 636, July 1984.

[7] J. Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, 46(7):2537 – 2543, November 2000.

[8] T. Roos, P. Myllymaki, and J. J. Rissanen. MDL denoising revisited. *IEEE Transactions on Signal Processing*, 57(9):3347 – 3360, 2009.

[9] B. Ryabko. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Transactions on Information Theory*, 55:4309–4315, 2009.

[10] B. Ryabko and J. Astola. Universal codes as a basis for time series testing. *Statistical Methodology*, 3:375–397, 2006.

[11] D. Ryabko. Testing composite hypotheses about discrete-valued stationary processes. In *Proc. IEEE Information Theory Workshop (ITW'10)*, pages 291–295, Cairo, Egypt, 2010. IEEE.

[12] D. Ryabko. Testing composite hypotheses about discrete ergodic processes. *Test*, page (to appear), 2011.

[13] D. Ryabko and B. Ryabko. Nonparametric statistical inference for ergodic processes. *IEEE Transactions on Information Theory*, 56(3):1430–1435, 2010.