ENTROPY COMPUTATION IN PARTIALLY OBSERVED MARKOV CHAINS

François Desbouvries

Institut National des Télécommunications, Evry, France

Abstract. Let $X = \{X_n\}_{n\in\mathbb{N}}$ be a hidden process and $Y = \{Y_n\}_{n\in\mathbb{N}}$ be an observed process. We assume that (X,Y) is a (pairwise) Markov Chain (PMC). PMC are more general than Hidden Markov Chains (HMC) and yet enable the development of efficient parameter estimation and Bayesian restoration algorithms. In this paper we propose a fast (i.e., $O(N)$) algorithm for computing the entropy of $\{X_n\}_{n=0}^N$ given an observation sequence $\{y_n\}_{n=0}^N$.

Keywords: Entropy, Hidden Markov Models, Partially observed Markov Chains
PACS: 05.70.-a, 65.40.Gr, 02.50.-r, 02.50.Ga.

INTRODUCTION

Let $(X,Y) = \{X_n,Y_n\}_{n\in\mathbb{N}}$ be a joint process in which X is unobserved and Y is observed. We assume that X and Y are both discrete with $X_n \in \{1,\cdots,K\}$ and $Y_n \in \{1,\cdots,M\}$ for all $n \in \mathbb{N}$. Let $X_{i:j} = \{X_n\}_{i\leq n\leq j}$, $Y_{i:j} = \{Y_n\}_{i\leq n\leq j}$, $x_{i:j} = \{x_n\}_{i\leq n\leq j}$ and $y_{i:j} = \{y_n\}_{i\leq n\leq j}$ (upper case letters denote random variables (r.v.) and lower case letters their realizations). Let also $p(x_{i:j}|y_{i:j})$, say, denote the conditional probability that $X_{i:j} = x_{i:j}$ given $Y_{i:j} = y_{i:j}$. In some applications it is relevant to compute the entropy of $X_{0:N} = \{X_n\}_{n=0}^N$ given an observation $y_{0:N} = \{y_n\}_{n=0}^N$, i.e. we want to compute

$$H(X_{0:N}|y_{0:N}) = -\sum_{x_{0:N}} p(x_{0:N}|y_{0:N}) \log p(x_{0:N}|y_{0:N}). \tag{1}$$

The brute force computation of (1) requires $O(K^N)$ elementary operations. However, a fast (i.e., $O(K^2N)$) algorithm for computing (1) has been proposed recently [1] in the framework of Hidden Markov Chains (HMC) (see e. g. the recent tutorials [2] [3]), i.e. of processes (X,Y) satisfying

$$p(x_{n+1}|x_{0:n}) = p(x_{n+1}|x_n)\ ; \tag{2}$$

$$p(y_{0:N}|x_{0:N}) = \prod_{n=0}^{N} p(y_n|x_{0:N})\ ; \tag{3}$$

$$p(y_n|x_{0:N}) = p(y_n|x_n) \ \text{ for all } n,\ 0\leq n\leq N\ . \tag{4}$$

Now, HMC have been generalized recently to Pairwise Markov Chains (PMC) [4], i.e. to joint processes (X,Y) which satisfy

$$p(x_n,y_n|x_{0:n-1},y_{0:n-1}) = p(x_n,y_n|x_{n-1},y_{n-1}). \tag{5}$$

As we see from the definition, a PMC can be seen as a (vector) Markov chain in which one component is observed and the other one is hidden. Now, (2)-(4) imply (5), so any HMC is a PMC. The converse is not true, as can be seen at the local level, since in a PMC the transition probability reads

$$p(x_n,y_n|x_{n-1},y_{n-1}) = p(x_n|x_{n-1},y_{n-1})p(y_n|x_n,x_{n-1},y_{n-1}); \tag{6}$$

so an HMC is indeed a PMC in which $p(x_n|x_{n-1},y_{n-1})$ reduces to $p(x_n|x_{n-1})$ and $p(y_n|x_n,x_{n-1},y_{n-1})$ reduces to $p(y_n|x_n)$. In other words, making use of PMC enables to model rather complex physical situations, since at time n, conditionnally on the previous state x_{n-1}, the probability of the current state x_n may still depend on the previous observation y_{n-1}; and conditionnally on x_n, the probability of observation y_n may still depend on the previous state x_{n-1} and on the previous observation y_{n-1}.

It happens that it is possible to extend from HMC to PMC [4] the existing efficient Bayesian restoration or parameter estimation algorithms. As we shall see in this paper, it is also possible in the context of PMC to compute $H(X_{0:N}|y_{0:N})$ efficiently. More precisely, our aim here is to extend to PMC the algorithm of [1]; the algorithm we obtain remains $O(K^2N)$.

EFFICIENT ENTROPY COMPUTATION IN PMC

From now on we assume that (X,Y) is a PMC, i.e. that (5) holds. Let us first recall [5] the following basic properties of entropy :

$$\begin{aligned} h(U,V/w) &= h(U/w)+h(V/U,w), &(7)\\ h(V/U,w) &= \sum_u h(V/u,w)p(u|w). &(8)\end{aligned}$$

Let us now address the computation of $H(X_{0:N}|y_{0:N})$. Let $0 \le n \le N$. From (7), (8) we get

$$\begin{aligned} H(X_{0:n}|y_{0:n}) &= H(X_n|y_{0:n})+H(X_{0:n-1}|X_n,y_{0:n})\\ &= H(X_n|y_{0:n})+\sum_{x_n} H(X_{0:n-1}|x_n,y_{0:n})p(x_n|y_{0:n}). &(9)\end{aligned}$$

On the other hand, from (5) we get

$$p(x_{0:n-2}|x_{n-1},x_n,y_{0:n}) = p(x_{0:n-2}|x_{n-1},y_{0:n-1}), \tag{10}$$

so $H(X_{0:n-1}|x_n,y_{0:n})$ in (9) can be computed recursively by

$$\begin{aligned} H(X_{0:n-1}|x_n,y_{0:n}) &= H(X_{n-1}|x_n,y_{0:n})+H(X_{0:n-2}|X_{n-1},x_n,y_{0:n})\\ &= H(X_{n-1}|x_n,y_{0:n})+\sum_{x_{n-1}} H(X_{0:n-2}|x_{n-1},x_n,y_{0:n})p(x_{n-1}|x_n,y_{0:n})\\ &\overset{(10)}{=} H(X_{n-1}|x_n,y_{0:n})+\sum_{x_{n-1}} H(X_{0:n-2}|x_{n-1},y_{0:n-1})p(x_{n-1}|x_n,y_{0:n}) &(11)\end{aligned}$$

It remains to compute $p(x_n|y_{0:n})$ and $p(x_{n-1}|x_n,y_{0:n})$ efficiently. This can be performed by an algorithm which extends to PMC [4] the (forward pass of) the Forward-Backward algorithm [6] [7] [8] [9], and which we now recall

$$\begin{aligned} p(x_{n-1},x_n|y_{0:n}) &= \frac{p(x_{n-1},x_n,y_{0:n})}{\sum_{x_{n-1},x_n} p(x_{n-1},x_n,y_{0:n})} \\ &\overset{(5)}{=} \frac{p(x_n,y_n|x_{n-1},y_{n-1})p(x_{n-1}|y_{0:n-1})p(y_{0:n-1})}{\sum_{x_{n-1},x_n} p(x_n,y_n|x_{n-1},y_{n-1})p(x_{n-1}|y_{0:n-1})p(y_{0:n-1})} \\ &= \frac{p(x_n,y_n|x_{n-1},y_{n-1})p(x_{n-1}|y_{0:n-1})}{\sum_{x_{n-1},x_n} p(x_n,y_n|x_{n-1},y_{n-1})p(x_{n-1}|y_{0:n-1})}, \qquad (12) \\ p(x_n|y_{0:n}) &= \sum_{x_{n-1}} p(x_{n-1},x_n|y_{0:n}), \qquad (13) \\ p(x_{n-1}|x_n,y_{0:n}) &= \frac{p(x_{n-1},x_n|y_{0:n})}{p(x_n|y_{0:n})}. \qquad (14) \end{aligned}$$

Let us summarize the discussion. We got the following algorithm :

Fast algorithm for computing $H(X_{0:N}|y_{0:N})$.

- At time $n-1$:
  - assume that we have $\{H(X_{0:n-2}|x_{n-1},y_{0:n-1})\}_{x_{n-1}=1}^K$, $\{p(x_{n-1}|y_{0:n-1})\}_{x_{n-1}=1}^K$.
- Iteration $n-1 \to n$:
  - compute $\{p(x_n|y_{0:n})\}_{x_n=1}^K$ and $\{p(x_{n-1}|x_n,y_{0:n})\}_{x_{n-1},x_n=1}^K$ via (12), (13) and (14);
  - compute $\{H(X_{n-1}|x_n,y_{0:n}) = -\sum_{x_{n-1}} p(x_{n-1}|x_n,y_{0:n}) \log p(x_{n-1}|x_n,y_{0:n})\}_{x_n=1}^K$;
  - compute $\{H(X_{0:n-1}|x_n,y_{0:n})\}_{x_n=1}^K$ via (11);
  - compute $H(X_n|y_{0:n}) = -\sum_{x_n} p(x_n|y_{0:n}) \log p(x_n|y_{0:n})$;
  - compute $H(X_{0:n}|y_{0:n})$ via (9).

Note that the algorithm is $O(K^2N)$, as was the original algorithm of [1]. Finally, we assumed that Y_n is a discrete r.v., but the extension to continuous emission probability densities is straightforward.

REFERENCES

1. D. Hernando, V. Crespi, and G. Cybenko, *IEEE Transactions on Information Theory* **51**, 2681–85 (2005).
2. Y. Ephraim, and N. Merhav, *IEEE Transactions on Information Theory* **48**, 1518–69 (2002).
3. O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, Springer-Verlag, 2005.
4. W. Pieczynski, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 634–39 (2003).
5. T. M. Cover, and J. A. Thomas, *Elements of Information Theory*, Wiley series in Telecommunications, Wiley Interscience, 1991.
6. L. E. Baum, and T. Petrie, *Ann. Math. Stat.* **37**, 1554–63 (1966).
7. L. E. Baum, and J. A. Eagon, *Bull. Amer. Meteorol. Soc.* **73**, 360–63 (1967).
8. L. R. Rabiner, *Proceedings of the IEEE* **77**, 257–286 (1989).
9. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, *IEEE Transactions on Information Theory* **20**, 284–87 (1974).