ENTROPY RATE FOR HIDDEN MARKOV CHAINS WITH RARE TRANSITIONS

YUVAL PERES AND ANTHONY QUAS

Abstract. We consider Hidden Markov Chains obtained by passing a Markov Chain with rare transitions through a noisy memoryless channel. We obtain asymptotic estimates for the entropy of the resulting Hidden Markov Chain as the transition rate is reduced to zero.

Let (X_n) be a Markov chain with finite state space S and transition matrix $P(p)$ and let (Y_n) be the Hidden Markov chain observed by passing (X_n) through a homogeneous noisy memoryless channel (i.e. Y takes values in a set T, and there exists a matrix Q such that $\mathbb{P}(Y_n = j|X_n = i, X_{-\infty}^{n-1}, X_{n+1}^{\infty}, Y_{-\infty}^{n-1}, Y_{n+1}^{\infty}) = Q_{ij}$). We make the additional assumption on the channel that the rows of Q are distinct. In this case we call the channel *statistically distinguishing*.

Finally we assume that $P(p)$ is of the form $I + pA$ where A is a matrix with negative entries on the diagonal, non-negative entries in the off-diagonal terms and zero row sums. We further assume that for small positive p, the Markov chain with transition matrix $P(p)$ is irreducible. Notice that for Markov chains of this form, the invariant distribution $(\pi_i)_{i\in S}$ does not depend on p. In this case, we say that for small positive values of p, the Markov chain is in a *rare transition regime*.

We will adopt the convention that H is used to denote the entropy of a finite partition, whereas h is used to denote the entropy of a process (the *entropy rate* in information theory terminology). Given an irreducible Markov chain with transition matrix P, we let $h(P)$ be the entropy of the Markov chain (i.e. $h(P) = -\sum_{i,j} \pi_i P_{ij} \log P_{ij}$ where π_i is the (unique) invariant distribution of the Markov chain and where as usual we adopt the convention that $0 \log 0 = 0$). We also let $H_{\text{chan}}(i)$ be the entropy of the output of the channel when the input symbol is i (i.e. $H_{\text{chan}}(i) = -\sum_{j\in T} Q_{ij} \log Q_{ij}$). Let $h(Y)$ denote the entropy of Y (i.e. $h(Y) = -\lim_{N\to\infty} \frac{1}{N} \sum_{w\in T^N} \mathbb{P}(Y_1^N = w) \log \mathbb{P}(Y_1^N = w))$

Theorem 1. *Consider the Hidden Markov Chain (Y_n) obtained by observing a Markov chain with irreducible transition matrix $P(p) = I+Ap$ through a statistically distinguishing channel with transition matrix Q. Then there exists a constant $C > 0$ such that for all small $p > 0$,*

$$h(P(p)) + \sum_i \pi_i H_{chan}(i) - Cp \le h(Y) \le h(P(p)) + \sum_i \pi_i H_{chan}(i),$$

where $(\pi_i)_{i\in S}$ is the invariant distribution of $P(p)$.

Date: March 23, 2009.

AQ's research was supported by NSERC; The authors thank BIRS where this research was conducted.

If in addition the channel has the property that there exist i, i' and j such that $P_{ii'} > 0$, $Q_{ij} > 0$ and $Q_{i'j} > 0$, then there exists a constant $c > 0$ such that

$$h(Y) \le h(P(p)) + \sum_i \pi_i H_{chan}(i) - cp.$$

The entropy rate in the rare transition regime was considered previously in the special case of a 0–1 valued Markov Chain with transition matrix $P(p) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$ and where the channel was the binary symmetric channel with crossover probability ϵ (i.e. $Q = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$). It is convenient to introduce the notation $g(p) = -p \log p - (1-p)\log(1-p)$. In [5], Nair, Ordentlich and Weissman proved that $g(\epsilon) - (1-2\epsilon)^2 p \log p/(1-\epsilon) \le h(Y) \le g(p) + g(\epsilon)$. For comparison, with our result, this is essentially of the form $g(\epsilon) + a(\epsilon)g(p) \le h(Y) \le g(p) + g(\epsilon)$ where $a(\epsilon) < 1$ but $a(\epsilon) \to 1$ as $\epsilon \to 0$ (i.e. $h(Y) = g(p) + g(\epsilon) - O(p \log p)$). A second paper due to Chigansky [1] shows that $g(\epsilon) + b(\epsilon)g(p) \le h(Y)$ for a function $b(\epsilon) < 1$ satisfying $b(\epsilon) \to 1$ as $\epsilon \to 1/2$ (again giving an $O(p \log p)$ error). Our result states in this case that there exist $C > c > 0$ such that $g(p)+g(\epsilon)-Cp \le h(Y) \le g(p)+g(\epsilon)-cp$ (i.e. $h(Y) = g(p) + g(\epsilon) - \Theta(p)$).

Before giving the proof of the theorem, we explain briefly the underlying idea of the lower bound which is the main part. Since the transitions in the (X_n) sequence are rare, given a realization of (Y_n), the Y_n values allow one to guess (using the statistical-distinguishing property) the X_n values from which the Y_n values are obtained. This provides for an accurate reconstruction except that where there is a transition in the X_n's there is some uncertainty as to its location as estimated using the Y_n's. It turns out that by using maximum likelihood estimation, the transition locations may be pinpointed up to an error with exponentially small tail. Since the transitions occur with rate p, there is an $O(p)$ error in reconstructing (X_n) from (Y_n).

Proof of Theorem 1. Given a measurable partition $\mathcal{Q}$ of the space and an event A that is $\mathcal{F}$-measurable for a σ-algebra $\mathcal{F}$ then we will write $H(\mathcal{Q}|\mathcal{F}|A)$ for the entropy of $\mathcal{Q}$ relative to $\mathcal{F}$ with respect to the measure $\mathbb{P}_A(B) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. If $A_1, \ldots, A_k$ form an $\mathcal{F}$-measurable partition of the space, then we have the following equality:

$$H(\mathcal{Q}|\mathcal{F}) = \sum_{j=1}^{k} \mathbb{P}(A_j) H(\mathcal{Q}|\mathcal{F}|A_j). \tag{1}$$

Note that $((X_n, Y_n))_{n\in\mathbb{Z}}$ forms a Markov chain with transition matrix $\bar{P}$ given by $\bar{P}_{(i,j),(i',j')} = P_{ii'}Q_{i'j'}$ and invariant distribution $\bar{\pi}_{(i,j)} = \pi_i Q_{ij}$. The standard formula for the entropy of a Markov chain then gives $h(X,Y) = h(P(p)) + \sum_i \pi_i H_{\mathrm{chan}}(i)$. Since $h(X,Y) = h(Y) + h(X|Y)$, one obtains

$$h(Y) = h(X,Y) - h(X|Y) = h(P(p)) + \sum_i \pi_i H_{\mathrm{chan}}(i) - h(X|Y). \tag{2}$$

This establishes the basic upper bound in the theorem.

If the additional properties are satisfied (the existence of i, i' and j such that $P_{ii'} > 0$, $Q_{ij} > 0$ and $Q_{i'j} > 0$), then we need to show that $h(X|Y) \ge cp$ for some $c > 0$. To show this, we will demonstrate that $H(X_0|Y, X_{-\infty}^{-1}) \ge cp$. In fact, we show the stronger statement: $H(X_0|Y, (X_n)_{n\neq 0}) \ge cp$. Let A be the event that $X_{-1} = i$ and $X_1 = i'$ and $Y_0 = j$. Given that $X_{-1} = i$ and $X_1 = i'$, the probability that $X_0 = i$ is $P_{ii}P_{ii'}/\sum_k P_{ik}P_{ki'} = 1/2 + O(p)$ and similarly the probability that $X_0 = i'$ is $1/2 + O(p)$. Now, given that $X_{-1} = i$ and $X_1 = i'$ and $Y_0 = j$, one has from Bayes' theorem that the probability that $X_0 = i$ is $Q_{ij}/(Q_{ij} + Q_{i'j}) + O(p)$ and similarly the probability that $X_1 = i'$ is $Q_{i'j}/(Q_{ij} + Q_{i'j}) + O(p)$. It follows that $H(X_0|Y,(X_n)_{n\neq 0}|A) = k + O(p)$ for some positive constant k. Since A has probability $\Omega(p)$ we obtain the lower bound $h(X|Y) \geq cp$. From this we deduce the claimed upper bound for $h(Y)$:

$$h(Y) \leq h(X) + \sum_i \pi_i H_{\mathrm{chan}}(i) - cp.$$

We now establish the lower bound. We are aiming to show $h(X|Y) = O(p)$ (for which it suffices to show $H(X_0^{L-1}|Y) = O(Lp)$). Setting $L = |\log p|^4$ and letting $\mathcal{P}$ be a suitable partition, we estimate $H(X_0^{L-1}|Y,\mathcal{P})$ and use the inequality

$$H(X_0^{L-1}|Y) \leq H(X_0^{L-1}|Y,\mathcal{P}) + H(\mathcal{P}). \tag{3}$$

We define the partition $\mathcal{P}$ as follows: Set $K = |\log p|^2$ and let $\mathcal{P} = \{E_{\mathrm{m}}, E_{\mathrm{b}}, E_{\mathrm{g}}\}$. Here E_{m} (for many) is the event that there are at least two transitions in X_0^{L-1}, E_{b} (for boundary) is the event that there is exactly one transition and that it takes place within a distance K of the boundary of the block and finally E_{g} (for good) is the event that there is at most one transition and if it takes place, then it occurs at a distance at least K from the boundary of the block.

We have $\mathbb{P}(E_{\mathrm{m}}) = O(p^2L^2)$ and $H(X_0^{L-1}|E_{\mathrm{m}}) \leq L\log|S|$ from which we see that $\mathbb{P}(E_{\mathrm{m}})H(X_0^{L-1}|E_{\mathrm{m}}) = O(p^2L^3) = o(pL)$.

We have $\mathbb{P}(E_{\mathrm{b}}) = O(pK)$. Given that E_{b} takes place, there are $2K|S|(|S|-1) = O(K)$ possible values of X_0^{L-1} so that $\mathbb{P}(E_{\mathrm{b}})H(X_0^{L-1}|E_{\mathrm{b}}) = O(pK\log K) = o(pL)$.

Since the elements of the partition have probabilities $O(pK)$ or $1 - O(pK)$, we see that $H(\mathcal{P}) = o(pL)$.

It remains to show that $H(X_0^{L-1}|Y|E_{\mathrm{g}}) = O(pL)$. In fact we shall demonstrate $H(X_0^{L-1}|Y_0^{L-1}|E_{\mathrm{g}}) = O(pL)$. Given that the event E_{g} holds, the sequence X_0^{L-1} belongs to $B = \{a^L : a \in S\} \cup \{a^ib^{L-i} : a, b \in S, K \leq i \leq L-K\}$.

Given a sequence $u \in B$, we define the log-likelihood of u being the input sequence yielding the output Y_0^{L-1} by $L_u(Y_0^{L-1}) = \sum_{i=0}^{L-1} \log Q_{u_iY_i}$. Given that E_{g} holds, we define Z_0^{L-1} to be the sequence in B for which $L_Z(Y_0^{L-1})$ is maximized (breaking ties lexicographically if necessary). We will then show using large deviation methods that Z_0^{L-1} is a good reconstruction of X_0^{L-1} with small error.

We calculate for $u, v \in B$,

$$\begin{aligned}\mathbb{P}\left(L_v(Y_0^{L-1}) \geq L_u(Y_0^{L-1}) | X_0^{L-1} = u\right) &= \mathbb{P}\left(\sum_{i=0}^{L-1} \log(Q_{v_iY_i}/Q_{u_iY_i}) \geq 0\right) \\ &= \mathbb{P}\left(\sum_{i\in\Delta} \log(Q_{v_iY_i}/Q_{u_iY_i}) \geq 0\right),\end{aligned}$$

where $\Delta = \{i : u_i \neq v_i\}$. For each $i \in \Delta$, given that $X_0^{L-1} = u$, we have that $\log(Q_{v_iY_i}/Q_{u_iY_i})$ is an independent random variable taking the value $\log(Q_{v_ij}/Q_{u_ij})$ with probability Q_{u_ij}.

It is well known (and easy to verify using elementary calculus) that for a given probability distribution π on a set T, the probability distribution σ maximizing

$\sum_{j\in T}\pi_j\log(\sigma_j/\pi_j)$ is $\sigma=\pi$ (for which the maximum is 0). Accordingly we see that given that $X_0^{L-1}=u$, $L_v(Y_0^{L-1})-L_u(Y_0^{L-1})$ is the sum of $|\Delta|$ random variables, each having one of $|S|(|S|-1)$ distributions, each with negative expectation. It follows from Hoeffding's Inequality [3] that there exist C and η such that $\mathbb{P}(L_v(Y_0^{L-1})\geq L_u(Y_0^{L-1})|X_0^{L-1}=u)\leq C\eta^{|\Delta|}$ where η is independent of p. Since there at most $2|S|$ sequences in B differing from u in exactly k places, we obtain

$$\mathbb{P}(Z_0^{L-1}\text{ differs from }X_0^{L-1}\text{ in }k\text{ places}|E_{\mathrm{g}})\leq 2|S|C\eta^k. \tag{4}$$

In particular, conditioned on E_{g}, $H(X_0^{L-1}|Z_0^{L-1}|Z_0\neq Z_{L-1})=O(1)$. If $Z_0=Z_{L-1}$ then X_0^{L-1} can differ in no places or at least K places. It follows that $H(X_0^{L-1}|Z_0^{L-1}|Z_0=Z_{L-1})=O(K\eta^K)=o(Lp)$.

We have $\mathbb{P}(Z_0\neq Z_{L-1})=\mathbb{P}(X_0\neq X_{L-1})+O(\eta^K)=O(Lp)$. This gives

$$\begin{aligned}H(X_0^{L-1}|Y_0^{L-1}|E_{\mathrm{g}})&\leq H(X_0^{L-1}|Z_0^{L-1}|E_{\mathrm{g}}\cap\{Z_0=Z_{L-1}\})\mathbb{P}(Z_0=Z_{L-1}|E_{\mathrm{g}})\\&\quad+H(X_0^{L-1}|Z_0^{L-1}|E_{\mathrm{g}}\cap\{Z_0\neq Z_{L-1}\})\mathbb{P}(Z_0\neq Z_{L-1}|E_{\mathrm{g}})\\&=O(pL).\end{aligned}$$

This completes the proof that $H(X_0^{L-1}|Y)=O(pL)$ so that $h(X|Y)=O(p)$ and $h(Y)=h(X)+\sum_i\pi_iH_{\mathrm{chan}}(i)+\Theta(p)$ as required.

□

We note that as part of the proof we attempt a reconstruction of (X_n) from the observed data (Y_n). In our case, the reconstruction of the nth symbol of X_n depended on past and future values of Y_m. A related but harder problem of filtering is to try to reconstruct X_n given only Y_1^n. This problem was addressed in essentially the same scenario by Khasminskii and Zeitouni [4], where they gave a lower bound for the asymptotic reconstruction error of the form $Cp|\log p|$ for an explicit constant C (i.e. for an arbitrary reconstruction scheme, the probability of wrongly guessing X_n is bounded below in the limit as $n\to\infty$ by $Cp|\log p|$). By the above, if one is allowed to use future as well as past observations then the asymptotic reconstruction error is $O(p)$.

References

[1] P. Chigansky. The entropy rate of a binary channel with slowly switching input. Available on arXiv: cs/0602074v1, 2006.

[2] R. S. Ellis. *Entropy, large deviations and statistical mechanics.* Springer, 1985.

[3] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.*, 58:13–30.

[4] R. Z. Khasminskii and O. Zeitouni. Asymptotic filtering for finite state Markov chains. *Stochastic Process. Appl.*, 63:1–10, 1996.

[5] C. Nair, E. Ordentlich, and T. Weissman. Asymptotic filtering and entropy rate of a hidden Markov process in the rare transitions regime. In *International Symposium on Information Theory*, pages 1838–1842, 2005.

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
E-mail address: peres(a)microsoft.com

Department of Mathematics and Statistics, University of Victoria, Victoria, BC V8W 3R4, CANADA
E-mail address: aquas(a)uvic.ca