ISIT'2013 1569717327

A Randomized Approach to the Capacity of Finite-State Channels

Guangyue Han
The University of Hong Kong
Email: ghan@hku.hk

***Abstract*—Inspired by the ideas from the field of stochastic approximation, we propose a randomized algorithm to compute the capacity of a finite-state channel with a Markovian input. When the mutual information rate of the channel is concave with respect to the chosen parameterization, we show that, at least for some practical channels, the proposed algorithm will converge to the capacity almost surely.**

I. Introduction

Discrete-time finite-state channels are a broad class of channels which have attracted plenty of interest in information theory; prominent examples of such channels include partial response channels [27], Gilbert-Elliott channels [21] and noisy input-restricted channels [34], which are widely used in a variety of real-life applications, including magnetic and optical recording [20], communications over band-limited channels with inter-symbol interference [7]. The computation of the capacity of a finite-state channel is notoriously difficult and has been open for decades. For a discrete memoryless channel with a discrete memoryless source at its input, the classical Blahut-Arimoto algorithm (BAA) can effectively compute the channel capacity, however, for almost all nontrivial finite-state channels, little is known about the channel capacity other than some numerically computed bounds; see, e.g., [34], [28], [2], [8] and references therein.

Recently, Vontobel *et al.* have proposed a generalized Blahut-Arimoto algorithm (GBAA) [32] to maximize the mutual information rate of a finite-state machine channel with a finite-state machine source at its input. This interesting algorithm has attracted a great deal of attention due to the observations that it fairly precisely approximates the channel capacity for a number of practical channels. For a finite-state channel, let X denote the input Markov process and Y its corresponding output process, which is, by definition, a *hidden Markov process* (see [5] and references therein). In contrast to the BAA, the proof of the convergence of the GBAA in [32] requires the extra assumption that $I(X;Y)$ and $H(X|Y)$ are both concave with respect to a chosen parameterization, which has been posed as Conjecture 74 in [32]. Example V.3, however, shows that the concavity conjecture is not true in general.

One of the hurdles encountered in computing the finite-state channel capacity is the problem of optimizing $H(Y)$, which naturally occurs in the formula of the capacity of a broad class of finite-state channels. More specifically, there has long been a lack of understanding on the following two issues:

(I) How to effectively compute the entropy rate of hidden Marov processes?
(II) How does the entropy rate of hidden Markov processes vary as the underlying Markov processes and the channels vary?

As elaborated below, recently, these two issues have been partially addressed by the information theory community.

Related work on (I). It is well known that $H(X)$ has a simple analytic formula; in stark contrast, there is no simple and explicit formula of $H(Y)$ for most non-degenerate channels ever since hidden Markov processes (or, more precisely, hidden Markov models) were formulated more than half a century ago.

The celebrated Shannon-McMillan-Breiman theorem states that the n-th order *sample entropy* $-\log p(Y_1^n)/n$ converges to $H(Y)$ almost surely. Based on this, efficient Monte Carlo methods for approximating $H(Y)$ were proposed independently by Arnold and Loeliger [1], Pfister, Soriaga and Siegel [25], Sharma and Singh [29]. One of the concerns that has not been addressed in these work is the "accuracy" of approximating $H(Y)$ using the sample entropy. In this regard, a central limit theorem (CLT) [26] for the sample entropy has been derived as a corollary of a CLT for the top Lyapunov exponent of a product of random matrices; a functional CLT has also been established in [15]. To some extent, these two CLTs suggested that the Monte Carlo methods are "accurate" in terms of approximating $H(Y)$. However, more quantitative description of the convergence behavior of the proposed methods, such as rate of convergence, are still lacking in these work.

Recently, we have obtained [9] a number of limit theorems for the sample entropy of Y. These limit theorems can be viewed as further refinements of the Shannon-McMillian-Breiman theorem, which is the backbone of information theory. More specifically, Theorem 1.2 in [9] is a CLT with an error-estimate, which can be used to characterize the rate of convergence of the Monte Carlo methods in [1], [25], [29], and Theorem 1.5 in [9] is a large deviation result, which gives a sub-exponential decaying upper bound on the probability of the sample entropy $-\log p(Y_1^n)/n$ deviating from $H(Y)$. Among many other applications, such as deriving non-asymptotic coding theorems [33], these theorems positively confirmed the effectiveness of using the Shannon-McMillan-Breiman theorem to approximate $H(Y)$.

Related work on (II). The behavior of $H(Y)$ (as a func-

tion of the underlying Markov chain and the channel) is of significance in a number of scientific disciplines; particularly in information theory, it is of great importance for computing/estimating the capacity of finite-state channels. However, some of the basic problems, such as smoothness (or even differentiability) of $H(Y)$, have long remained unknown. Recently, asymptotical behavior of $H(Y)$ has been studied in [16], [23], [35], [24]. Under mild assumptions, analyticity of $H(Y)$ has been established in [10]. The framework in [10] has been generalized to continuous-state settings and further provides useful tools and techniques for our subsequent work, such as derivatives [11], asymptotics [12], concavity [13] of $H(Y)$.

Equipped with ideas and techniques from the above-mentioned work on (I) and (II), we are more prepared to make further progress towards the computation of the channel capacity. In particular, the ideas and techniques in [9] and [10] are vital to this paper. Roughly speaking, [10] proves that the entropy rate of hidden Markov chains is a "nicely behaved" function; and [9] confirms that it can be "well-approximated" using Monte Carlo simulations. The simulator of the derivative of $I(X;Y)$ as specified in Section IV, which is crucial to this work, is an "offspring" of the two schools of thoughts in [10] and [9].

Stochastic approximation methods refer to a family of recursive stochastic algorithms, aiming to find zeroes or extrema of functions whose values can only be estimated via noisy observations. The extensive literature on stochastic approximation has grown up around two prototyipcal algorithms, the Robbins-Monro algorithm and the Kiefer-Wolfowitz algorithm, mainly concerning the convergence analysis on these two algorithms and their variants; we refer the reader to [17] for an exposition to the vast literature on stochastic approximation.

Inspired by the ideas in stochastic approximation, we propose a randomized algorithm to compute the capacity of a class of finite-state channels with input Markov processes supported on some mixing finite-type constraint. Bearing the same spirit as the Robbins-Monro algorithm and the Kiefer-Wolfowitz algorithm, the proposed algorithm, in many subtle respects, differs from both of them. The main task of this paper is to conduct a convergence analysis of the proposed algorithm, which employs some established ideas and techniques from the field of stochastic approximation [3], [17], [18], [31]. However, neither the results nor the proofs in any of previous work imply our results.

Although described in different languages, our settings are essentially the same as in [32]. On the other hand, as opposed to the GBAA, the concavity of $I(X;Y)$ alone is already sufficient to guarantee the convergence of our algorithm. Here, let us note that that for certain classes of channels (see Example V.3), $I(X;Y)$ is indeed concave with respect to certain parameterization, whereas $H(X|Y)$ fails to be concave with respect to the same parameterization.

Characterizing the maximal rate at which the information can be transmitted through a given channel, the capacity is the most fundamental notion in information theory. The capacity achieving distribution will further provide us insightful guidance towards designing coding schemes that actually achieve the promised capacity. Apparently, such an algorithm would be of fundamental significance to both information theoretic research and practical applications to tele-communications and data storage.

The organization of the paper is as follows. We first describe our channel model in greater detail in Section II and we then present our algorithm in Section III. In Section IV, we propose a simulator for the derivative of $I(X;Y)$ and discuss its convergence behavior. The convergence of the algorithm is established in V.

II. Channel Model

Let $\mathcal{X}$ be a finite alphabet and let

$$\mathcal{X}^2 = \{(i,j) : i,j \in \mathcal{X}\}.$$

Let Π denote the set of all stationary irreducible first-order Markov chain over the alphabet $\mathcal{X}$. For a given subset $F \subset \mathcal{X}^2$, define

$$\Pi_F = \{X \in \Pi : X_{i,j} = 0, \quad (i,j) \in F\},$$

where we have identified an irreducible first-order Markov chain with its transition probability matrix. Furthermore, for any $\epsilon > 0$, define

$$\Pi_{F,\epsilon} = \{X \in \Pi_F : X_{i,j} \geq \epsilon, \quad (i,j) \notin F\}.$$

Obviously, if some $X \in \Pi_{F,\epsilon}$ is primitive (namely, irreducible and aperiodic), then any other $X' \in \Pi_{F,\epsilon}$ is also primitive; in this case, we say F is a *mixing* finite-type constraint. Here, let us note that a mixing finite-type constraint can be defined in a much more general context; see [19].

The motivation for consideration of finite-type constraints mainly comes from magnetic recording, where input sequences are required to satisfy certain mixing finite-type constraints in order to eliminate the most damaging error events [20]. The most well known example is the so-called (d,k)-RLL constraint $\mathcal{S}(d,k)$ over the alphabet $\{0,1\}$, which forbids any sequence with fewer than d or more than k consecutive zeros in between two successive 1's.

In this paper, we are concerned with a discrete-time finite-state channel with some input constraint. Let X, Y, S denote the channel input, output and state processes over finite alphabets $\mathcal{X}, \mathcal{Y}$ and $\mathcal{S}$, respectively. Assume that

(II.a) For some mixing finite-type constraint $F \subset \mathcal{X}^2$ and some $\epsilon > 0$, $X \in \Pi_{F,\epsilon}$.

(II.b) (X,S) is a first-order stationary Markov chain whose transition probabilities satisfy

$$p(x_n, s_n | x_{n-1}, s_{n-1}) = p(x_n | x_{n-1}) p(s_n | x_n, s_{n-1}),$$

where $p(s_n | x_n, s_{n-1}) > 0$ for any s_n, x_n, s_{n-1}.

(II.c) the channel is stationary, and the channel transition probabilities satisfy

$$p(y_n, s_n | x_n, s_{n-1}) = p(s_n | x_n, s_{n-1}) p(y_n | x_n, s_n),$$

where $p(y_n|x_n, s_n) > 0$ for any y_n, x_n, s_n.

The capacity of the above channel is defined as

$$C_F = \sup I(X;Y) = \sup \lim_{n\to\infty} I_n(X;Y),$$

where the supremum is over all X satisfying (II.a) and

$$I_n(X;Y) \triangleq \frac{H(X_1^n) + H(Y_1^n) - H(X_1^n, Y_1^n)}{n}.$$

The fact that Y and (X, Y) are both hidden Markov processes makes it apparent that solutions to (I) and (II) are essential for computing C_F.

Assume that $\Pi_{F,\epsilon}$ is analytically parameterized by $\theta \in \Theta = \mathbb{R}^d$, $d \geq 1$. Then, naturally, $X = X(\theta)$ and $Y = Y(\theta)$ are also analytically parameterized by θ. Under this parameterization, we would like to find $\theta^* \in \Theta$ such that $X(\theta^*)$ maximizes $I(X(\theta); Y(\theta))$.

III. The Algorithm

For a given $1/2 < a < 1$, choose the so-called step sizes

$$a_n = \frac{1}{n^a}, \qquad n = 1, 2, \cdots;$$

apparently, $\{a_n\}$ satisfies

$$\sum_{n=0}^{\infty} a_n = \infty, \quad \sum_{n=0}^{\infty} a_n^2 < \infty,$$

which are the typical conditions imposed on step sizes in a generic stochastic approximation method. We propose to find θ^* through the following recursive procedure:

$$\theta_{n+1} = \theta_n + a_n g_{n^b}(\theta_n); \tag{1}$$

here $b > 0$, the initial θ_0 is randomly selected from Θ, and $g_{n^b}(\theta)$ is a to-be-specified simulator (see Section IV) for $I'(X(\theta); Y(\theta))$, where the derivative is taken with respect to θ. Throughout the paper, we assume that

$$0 < \beta < \alpha < 1/3, \quad 2a + b - 3b\beta > 1; \tag{2}$$

here, α, β are some "hidden" parameters involved in the definition of $g_{n^b}(\theta)$, which will be defined in Section IV.

IV. A Simulator of $I'(X;Y)$

As stated in Section I, albeit rather difficult to compute analytically, $I(X;Y)$ can be well-approximated via Monte Carlo simulations. In this section, we propose a simulator for $I'(X;Y)$. Needlessly to say, an effective simulator guaranteeing an "accurate" approximation to $I'(X;Y)$ is crucial to our algorithm. To some extent, our simulator is inspired by the Bernstein's blocking method [4], which is a well-established tool in proving limit theorems for mixing sequences; see, e.g., [6].

Now, for $0 < \beta < \alpha < 1/3$, define

$$q = q(n) \triangleq n^\beta, \quad p = p(n) \triangleq n^\alpha, \quad k = k(n) \triangleq n/(n^\alpha + n^\beta).$$

For any j with $iq + (i-1)p + 1 \leq j \leq iq + ip$ and a stationary stochastic process Z, define

$$\begin{aligned} W_j &= W_j(Z_{j-\lfloor q/2 \rfloor}^j) \\ &\triangleq -\left(\sum_{i=j-\lfloor q/2 \rfloor}^{j} \frac{p'(Z_i | Z_{j-\lfloor j-\lfloor q/2 \rfloor}^{i-1})}{p(Z_i | Z_{j-\lfloor j-\lfloor q/2 \rfloor}^{i-1})} \right) \log p(Z_j | Z_{j-\lfloor q/2 \rfloor}^{j-1}), \end{aligned}$$

and furthermore

$$\zeta_i \triangleq W_{iq+(i-1)p+1} + \cdots + W_{iq+ip}, \quad S_n \triangleq \sum_{i=1}^{k(n)} \zeta_i.$$

Now, we are ready to define our simulator for $I'(X;Y)$.

Definition IV.1.

$$g_n = g_n(X_1^n, Y_1^n) \triangleq H'(X_2|X_1) + S_n(Y_1^n)/(kp) - S_n(X_1^n, Y_1^n)/(kp).$$

The following lemma, whose proof is somewhat similar to Lemma 3.3 in [9], gives an estimate of the variance of S_n when Z is replaced by Y or (X, Y).

Lemma IV.2. *Replacing Z by Y or (X, Y), we have*

$$E[(S_n - E[S_n])^2] = O(kpq^3).$$

The following three theorems characterize the performances of our simulator from different perspectives.

Using similar techniques as in the proof of Theorem 1.1 in [10], the first theorem shows that on average, our simulator sub-exponentially converges to $I'(X;Y)$.

Theorem IV.3. *For some $0 < \rho_0 < 1$, we have*

$$E[g_n(X_1^n, Y_1^n)] - I'(X;Y) = O(\rho_0^{\lfloor q/2 \rfloor}).$$

The following large deviation type lemma gives a sub-exponentially decaying upper bound on the tail probability of $g_n(X_1^n, Y_1^n)$ deviating from $I'(X;Y)$.

Theorem IV.4. *For any $\varepsilon > 0$, there exists some $0 < \gamma, \delta < 1$ such that ,*

$$P\left(|g_n(X_1^n, Y_1^n) - I'(X;Y)| \geq \varepsilon\right) \leq \gamma^{n^\delta}.$$

The following theorem states that our simulator is asymptotically unbiased.

Theorem IV.5. *With probability 1,*

$$g_n(X_1^n, Y_1^n) \to I'(X;Y),$$

as n tends to ∞.

Proof. It immediately follows from Theorem IV.4 and the Borel-Cantelli lemma. □

Remark IV.6. In our notation, the following expression has been proposed in [32] as a simulator of $I'(X;Y)$:

$$H(X_2|X_1) - \frac{p'(Y_1^n)}{p(Y_1^n)} \log p(Y_1^n)/n + \frac{p'(X_1^n, Y_1^n)}{p(X_1^n, Y_1^n)} \log p(X_1^n, Y_1^n)/n.$$

Extensive numerical experiments conducted in [32] suggest that this simulator converges to $I'(X;Y)$ almost surely as n tends to infinity, however, there is no rigorous proof for the convergence.

V. CONVERGENCE

In this section, we will show that under the iteration in (1), $\{f(\theta_n)\}$ converges almost surely. For notational simplicity only, we assume $\Theta = \mathbb{R}$.

Henceforth, we will write

$$f(\theta) = I(X(\theta); Y(\theta)), \quad f_n(\theta) = I_n(X(\theta); Y(\theta)).$$

Note that under the assumption (II.a), Theorem 1.1 of [10] implies that

> $f(\theta)$ is analytic and each of its derivatives is uniformly bounded over all $\theta \in \Theta$,

a key fact that may be used throughout the paper implicitly. Now, rewrite (1) as

$$\theta_{n+1} = \theta_n + a_n f'(\theta_n) + a_n R_n(\theta_n), \tag{3}$$

where

$$R_n(\theta_n) \triangleq g_{n^b}(\theta_n) - f'(\theta_n).$$

It can be easily verified that

$$f(\theta_{n+1}) - f(\theta_n) = a_n f'^2(\theta_n) + \hat{R}_n(\theta_n), \tag{4}$$

where

$$\hat{R}_n(\theta_n) \triangleq a_n f'(\theta_n) R_n(\theta_n)$$
$$+ \int_0^1 (f'(\theta_n + t(\theta_{n+1} - \theta_n)) - f'(\theta_n))(\theta_{n+1} - \theta_n) dt.$$

Lemma V.1. $\sum_{n=0}^{\infty} \hat{R}_n(\theta_n)$ *converges almost surely.*

Proof. Let

$$T_1 = \sum_{n=0}^{\infty} a_n f'(\theta_n) R_n(\theta_n),$$

and

$$T_2 = \sum_{n=0}^{\infty} \int_0^1 (f'(\theta_n + t(\theta_{n+1} - \theta_n)) - f'(\theta_n))(\theta_{n+1} - \theta_n) dt.$$

It suffices to prove that T_1, T_2 both converge almost surely. We will only prove the convergence of T_1, since the proof for T_2 is similar.

Note that

$$T_1 = \sum_{n=0}^{\infty} a_n f'(\theta_n)(g_{n^b}(\theta_n) - f'_{n^b}(\theta_n)) + \sum_{n=0}^{\infty} a_n f'(\theta_n)(f'_{n^b}(\theta_n) - f'(\theta_n)).$$

It follows from Theorem IV.3 that there exists $0 < \rho_0 < 1$ such that

$$\sum_{n=0}^{\infty} a_n |f'(\theta_n)| |(f'_{n^b}(\theta_n) - f'(\theta_n))| \leq \sum_{n=0}^{\infty} a_n |f'(\theta_n)| \rho_0^{n^b} < \infty. \tag{5}$$

Then, using Lemma IV.2, one verifies that uniformly over all $\theta_n \in \Theta$,

$$\sum_{n=0}^{\infty} E[\{a_n^2 (f'(\theta_n))^2 R_n^2(\theta_n)\}] = \sum_{n=0}^{\infty} O\left(\frac{1}{n^{2a+b(1-3\beta)}}\right), \tag{6}$$

which converges since $2a + b - 3b\beta > 1$. Noting that $\{a_n f'(\theta_n) R_n(\theta_n)\}$ is a Martingale difference sequence (with respect to the σ-field generated by $\{X_1^n\}$) and applying Doob's Martingale convergence theorem (see Theorem 2.8.7 of [30]), we deduce that

$$\sum_{n=0}^{\infty} a_n f'(\theta_n)(g_{n^b}(\theta_n) - f'_{n^b}(\theta_n))$$

converges with probability 1. The almost sure convergence of T_1 then follows. □

We are now ready for the following convergence theorem, whose proof closely follows that of Lemma 7 in [31], which can be further can be traced back to the standard proof of the Martingale convergence theorem [30].

Theorem V.2. *With probability* 1*, we have*

$$\lim_{n\to\infty} f'(\theta_n) = 0 \text{ and } \lim_{n\to\infty} f(\theta_n) \text{ exists}.$$

Proof. Recall that

$$f(\theta_{n+1}) - f(\theta_n) = a_n f'^2(\theta_n) + \hat{R}_n(\theta_n),$$

an iterative application of which implies

$$f(\theta_n) = f(\theta_0) + \sum_{i=0}^{n-1} a_i (f'(\theta_i))^2 + \sum_{i=0}^{n-1} \hat{R}_i(\theta_i).$$

Applying Lemma V.1, we deduce that with probability 1,

$$\sum_{i=0}^{\infty} a_i (f'(\theta_i))^2 < \infty,$$

which, in return, implies that $\lim_{n\to\infty} f(\theta_n)$ exists and furthermore there is a subsequence $\{\theta_{n_j}\}$ such that $f'(\theta_{n_j})$ converges to 0 as j tends to infinity.

We now prove that

$$\lim_{n\to\infty} f'(\theta_n) = 0.$$

By way of contradiction, suppose otherwise. Then, there exists $\varepsilon > 0$ such that there exist infinite sequences $m_k, n_k, k = 1, 2, \cdots$, such that

$$|f'(\theta_{m_k})| \leq \varepsilon, \quad |f'(\theta_{n_k})| \geq 2\varepsilon, \quad |f'(\theta_i)| \geq \varepsilon \tag{7}$$

for all $m_k + 1 \leq i \leq n_k$. It then follows that

$$\begin{aligned} \varepsilon &\leq |f'(\theta_{n_k}) - f'(\theta_{m_k})| \\ &= O\left(\sum_{i=m_k}^{n_k-1} a_i\right) + O\left(\left|\sum_{i=m_k}^{n_k-1} a_i R_i(\theta_i)\right|\right). \end{aligned} \tag{8}$$

As in the proof of Lemma V.1, we deduce that $\sum_{n=0}^{\infty} a_n R_n(\theta_n)$ converges almost surely, and hence $\left|\sum_{i=m_k}^{n_k-1} a_i R_i(\theta_i)\right|$ tends to 0 as k goes to ∞. On the other hand, by (7), we have

$$\varepsilon^2 \sum_{i=m_k}^{n_k-1} a_i \leq \sum_{i=m_k}^{\infty} a_i (f'(\theta_i))^2.$$

This implies that as k tends to ∞, $\sum_{i=m_k}^{n_k-1} a_i$ tends to zero, which, together with (8), further implies that

$$\varepsilon \leq \lim_{k\to\infty} |f'(\theta_{n_k}) - f'(\theta_{m_k})| = 0,$$

a contradiction. □

Through the following example, we show that at least for some practical channels, our algorithm converges to the capacity almost surely.

Example V.3. Consider a binary symmetric channel with crossover probability $\varepsilon > 0$. Let X be a binary input Markov chain with the transition probability matrix

$$\left[\begin{array}{cc} 1-\pi & \pi \\ 1 & 0 \end{array} \right], \tag{9}$$

where $0 \leq \pi \leq 1$. Apparently, X is supported on the so-called $(1, \infty)$-RLL constraint [19], which simply means that the string "11" is forbidden. Let Y denote the corresponding output process. Assume that X is parameterized as in [32], that is, $\theta = (p_{ij} : i, j = 0, 1)$, where $p_{ij} = P(X_1 = i, X_2 = j)$. If ε is sufficiently small, it has been established in [13] that, roughly speaking, the capacity will be achieved in the "interior" of the parameter space, where $I(X;Y)$ is strictly concave (with respect to θ). This, together with Theorem V.2, implies that in high SNR regime, our algorithm will converge to the capacity achieving distribution almost surely.

On the other hand, it has been shown that for the output process Y, as $\varepsilon \to 0$,

$$H(Y) = H(X) + \frac{\pi(2-\pi)}{1+\pi}\varepsilon \log(1/\varepsilon) + O(\varepsilon), \tag{10}$$

where the $O(\varepsilon)$-term is analytic with respect to p (see Theorem 2.18 of [14]). It then follows that

$$\begin{aligned} H(X|Y) &= H(X) + H(Y|X) - H(Y) \\ &= H(\varepsilon) - \frac{p(2-\pi)}{1+\pi}\varepsilon \log(1/\varepsilon) + O(\varepsilon), \end{aligned}$$

where $H(\varepsilon) = \varepsilon \log 1/\varepsilon + (1-\varepsilon)\log 1/(1-\varepsilon)$. One can readily verify that $-\pi(2-\pi)/(1+\pi)$ is strictly convex with respect to θ, which implies the strict convexity (rather than concavity) of $H(X|Y)$ when ε is small enough. So, the concavity conjecture in [32] is not true in general, and thus the conditions guaranteeing the convergence of the GBAA are not satisfied.

REFERENCES

[1] D. M. Arnold and H.-A. Loeliger. The information rate of binary-input channels with memory. *IEEE ICC*, pp. 2692–2695, 2001.

[2] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavcic, W. Zeng, Simulation-based computation of information rates for channels with memory. *IEEE Trans. Info. Theory*, vol. 52, no. 8, pp. 3498–3508, 2006.

[3] A. Benveniste, M. Metivier and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, 1990.

[4] S. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, vol. 97, pp. 1–59, 1927.

[5] M. Boyle and K. Petersen. Hidden Markov processes in the context of symbolic dynamics. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems, London Mathematical Society Lecture Note Series*, vol. 385, pp. 5–71, 2011.

[6] R. Bradley. *Introduction to Strong Mixing Conditions*. Volumes 1,2 and 3. Kendrick Press, 2007.

[7] G. D. Forney, Jr. Maximum likelihood sequence estimation of digital sequences in the presence of inter-symbol interference. *IEEE Trans. Info. Theory*, vol. 18, no. 3, pp. 363–378, 1972.

[8] A. Goldsmith and P. Varaiya. Capacity, mutual information, and coding for finite-state Markov channels. *IEEE Trans. Info. Theory*, vol. 42, no. 3, pp. 868–886, 1996.

[9] G. Han. Limit theorems in hidden Markov models. To appear in *IEEE Trans. Info. Theory*.

[10] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5251–5266, 2006.

[11] G. Han and B. Marcus. Derivatives of entropy rate in special families of hidden Markov chains. *IEEE Trans. Info. Theory*, vol. 53, no. 7, pp. 2642–2652, 2007.

[12] G. Han and B. Marcus. Asymptotics of input-constrained binary symmetric channel capacity. *Annals of Applied Probability*, vol. 19, no. 3, pp. 1063–1091, 2009.

[13] G. Han and B. Marcus. Asymptotics of entropy rate in special families of hidden Markov chains. *IEEE Trans. Info. Theory*, vol. 56, no. 3, pp. 1287–1295, 2010.

[14] G. Han and B. Marcus. Concavity of the mutual information rate for input-restricted memoryless channels at high SNR. *IEEE Trans. Info. Theory*, vol. 58, no. 3, pp. 1534–1548, 2012.

[15] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of finite state channels based on Lyapunov exponents of random matrices. *IEEE Trans. Info. Theory*, vol. 52, no. 8, pp. 3509–3532, 2006.

[16] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Theoretical Computer Science*, vol. 395, pp. 203–219, 2008.

[17] H. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.

[18] L. Ljung. *System Identification: Theory for the User*, 2nd edition, Prentice Hall, 1999.

[19] D. Lind and B. Marcus. *An introduction to symbolic dynamics and coding*, Cambridge University Press, 1995.

[20] B. Marcus, R. Roth and P. H. Siegel. Constrained systems and coding for recording channels. *Handbook of Coding Theory*, Elsevier Science, 1998.

[21] M. Mushkin and I. Bar-David. Capacity and coding for the Gilbert-Elliott channel. *IEEE Trans. Info. Theory*, vol. 5, no. 6, pp. 1277–1290, 1989.

[22] C. Nair, E. Ordentlich and T. Weissman. Asymptotic filtering and entropy rate of a hidden Markov process in the rare transitions regime. *IEEE ISIT*, pp. 1838–1842, 2005.

[23] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov processes. *IEEE ITW*, pp. 117–122, 2004.

[24] Y. Peres and A. Quas. Entropy rate for hidden Markov chains with rare transitions. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems, London Mathematical Society Lecture Note Series*, vol. 385, pp. 172–178, 2011.

[25] H. D. Pfister, J. Soriaga and P. H. Siegel. The achievable information rates of finite-state ISI channels. *IEEE GLOBECOM*, pp. 2992–2996, 2001.

[26] H. D. Pfister. On the capacity of finite state channels and the analysis of convolutional accumulate-m codes. Ph.D. thesis, University of California at San Diego, USA, 2003.

[27] J. Proakis. *Digital Communications*, 4th ed. McGraw-Hill, New York, 2000.

[28] S. Shamai (Shitz) and Y. Kofman. On the capacity of binary and Gaussian channels with run-length limited inputs. *IEEE Trans. Commun.*, vol. 38, pp. 584–594, 1990.

[29] V. Sharma and S. Singh. Entropy and channel capacity in the regenerative setup with applications to Markov channels. *IEEE ISIT*, pp. 283, 2001.

[30] W. Stout. *Almost sure convergence*, New York, Academic Press, 1974.

[31] V. Tadic. Analyticity, Convergence, and Convergence Rate of Recursive Maximum-Likelihood Estimation in Hidden Markov Models. *IEEE Trans. Info. Theory*, vol. 56, no, 12, pp. 6406–6432, 2010.

[32] P. O. Vontobel, A. Kavcic, D. Arnold and H.-A. Loeliger. A generalization of the Blahut-Arimoto algorithm to finite-state channels. *IEEE Trans. Info. Theory*, vol. 54, no. 5, pp. 1887–1918, 2008.

[33] E. Yang and J. Meng. Non-asymptotic equipartition properties for independent and identically distributed sources. Preprint, available at http://ita.ucsd.edu/workshop/12/files/paper/paper_306.pdf.

[34] E. Zehavi and J. Wolf. On runlength codes. *IEEE Trans. Info. Theory*, vol. 34, no. 1, pp. 45–54, 1988.

[35] O. Zuk, I. Kanter and E. Domany. The entropy of a binary hidden Markov process. *J. Stat. Phys.*, vol. 121, no. 3-4, pp. 343–360, 2005.