# A Randomized Algorithm for the Capacity of Finite-State Channels *

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

August 12, 2013

**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 the proposed algorithm will almost surely converge to the capacity of the channel and derive the rate of convergence. We also discuss the convergence behavior of the algorithm without the concavity assumption.

## 1 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 [45, 50], Gilbert-Elliott channels [37, 18] and noisy input-restricted channels [55], which are widely used in a variety of real-life applications, including magnetic and optical recording [36], communications over band-limited channels with inter-symbol interference [17]. 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) [2, 12] can effectively compute the channel capacity, however, for almost all nontrivial finite-state channels, little is known about the channel capacity other than some bounds (see, e.g., [55], [46], [5] and references therein), which are numerically computed using Monte Carlo approaches. The methods in these work are believed to produce fairly precise numerical approximations of the capacity of certain classes of finite-state channels, however there are no general proofs to justify such beliefs.

Recently, Vontobel *et al.* have proposed a generalized Blahut-Arimoto algorithm (GBAA) [53] 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

*A preliminary version of this paper has been presented in the 2013 ISIT.

attention due to the observations that it fairly precisely approximates the channel capacity for a number of practical channels. (Notably, some results that were derived in the context of the GBAA have proven to be useful for analyzing the Bethe entropy function of some graphical models that appear in the context of low-density parity-check codes [51] and for approximately computing the permanent of a non-negative matrix [52].) For a finite-state channel, let $X$ denote the input Markov process and $Y$ its corresponding output process, which, by definition, is a *hidden Markov process* [13]. In contrast to the BAA, the proof of the convergence of the GBAA depends on 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 [53]. Example 9.4, however, shows that the concavity conjecture is not true in general; for other examples showing $I(X;Y)$ and $H(X|Y)$ fail to be concave, see [32].

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 Markov 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. Here, we remark that Blackwell [11] showed that $H(Y)$ can be written as an integral of an explicit function on a simplex with respect to the Blackwell Measure. However, the Blackwell measure seems to be rather complicated for effective computation of $H(Y)$. Since 2000, there has been a rebirth of interest in computing and estimating $H(Y)$ in a variety of scenarios: the Blackwell measure has been used to bound $H(Y)$ [39], a variation on the classical Birch bounds [10] can be found in [16] and a new numerical approximation of $H(Y)$ has been proposed in [35]. Generalizing Blackwell's idea, an integral formula for the derivatives of $H(Y)$ has been derived in [44].

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 [4], Pfister, Soriaga and Siegel [42], Sharma and Singh [47]. However, more quantitative description of the convergence behavior of the proposed methods, such as rate of convergence, asymptotic normality and so on, are lacking in these work. Recently, a central limit theorem (CLT) [43] 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 [28]. To some extent, these two CLTs suggested that the Monte Carlo methods are "accurate" in terms of approximating $H(Y)$. There are also other related work in different contexts from outside the information theory community, such as [30, 27, 26].

Recently, we have obtained [19] 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 [19] is a CLT with an error-estimate, which can be used to characterize the rate of convergence of the Monte Carlo methods in [4, 42, 47], and Theorem 1.5 in [19] 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 [54], 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 function 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 [3, 29, 39, 40, 56, 57, 38, 41, 44]. Particularly in [56], for a special type of hidden Markov chain $Y$, the Taylor series expansion of $H(Y)$ is given under the assumption that $H(Y)$ is analytic. Under mild assumptions, analyticity of $H(Y)$ has been established in [20]; see also related work in [13, 56, 57, 1, 35, 44] and references therein. The framework in [20] has been generalized to continuous-state settings and further provides useful tools and techniques for our subsequent work, such as derivatives [21], asymptotics [22], concavity [23] 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 [19] and [20] are vital to this paper. Roughly speaking, [20] proves that the entropy rate of hidden Markov chains is a "nicely behaved" function; and [19] confirms that it can be "well-approximated" using Monte Carlo simulations. The simulator of the derivative of $I(X;Y)$ as specified in Section 4, which is crucial to this work, is an "offspring" of the two schools of thoughts in [20] and [19].

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 [31] 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 mixing 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. In particular, the proofs in Section 8 are largely inspired by [49], which has credited origins of some of its techniques to earlier work, such as [7, 31, 33]. However, neither the results nor the proofs in [49] or any of previous work imply our results; as a matter of fact, considerable amount of simplification and adaptation of the techniques in [49] have been incorporated into this work.

Although described in different languages, our settings are essentially the same as in [53]. 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 9.4), $I(X;Y)$ 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 2 and we then present our algorithm in Section 3. In Section 4, we propose a simulator for the derivative of $I(X;Y)$ and discuss its convergence behavior. The convergence of the algorithm is established in Section 5, while the rate of convergence of the algorithm with and without concavity conditions are derived in Sections 7 and 8, respectively. In Section 9, we discuss the capacity achieving distribution of a special class of finite-state channels.

# 2 Channel Model

In this section, we specify the channel model considered in this paper in greater detail, which is essentially the same as the one considered in [53].

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

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

Let $\Pi$ 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 $X' \in \Pi_{F,\epsilon}$ is 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 [34].

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 [36]. 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

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

(2.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 $s_{n-1}, s_n, x_n$.

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

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

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 (2.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 $\theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$, where $\Theta$ denote the entire parameter space. Then, naturally, $X = X(\theta)$ and $Y = Y(\theta)$ are also analytically parameterized by $\theta$. Under this parameterization, we would like to find $\theta^* \in \Theta$ such that $X(\theta^*)$ maximizes $I(X(\theta); Y(\theta))$.

**Remark 2.1.** By Condition (2.b), $X$ is necessarily a first-order Markov chain. This assumption is for notional convenience only: through a usual "reblocking" technique, the case for a Markov chain of any order can be boiled down to the first-order case.

**Remark 2.2.** One natural goal is to find $X \in \Pi_F$ to maximize $I(X;Y)$. However, in this paper, we will restrict our attention to $\Pi_{F,\epsilon}$ for a given $\epsilon > 0$; such restriction will be justified in Section 9.

# 3 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. Letting $A_n$ denote the event "$\theta_n + a_n g_{n^b}(\theta_n) \notin \Theta$", we propose to find $\theta^*$ through the following recursive procedure:

$$\theta_{n+1} = \begin{cases} \theta_n, \text{ if } A_n \text{ occurs,} \\ \theta_n + a_n g_{n^b}(\theta_n), \text{ otherwise;} \end{cases} \tag{1}$$

here $b > 0$, the initial $\theta_0$ is randomly selected from $\Theta$, and $g_{n^b}(\theta)$ is a to-be-specified simulator (see Section 4) for $I'(X(\theta); Y(\theta))$, where the derivative is taken with respect to $\theta$. Throughout the paper, we assume that

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

here, $\alpha, \beta$ are some "hidden" parameters involved in the definition of $g_{n^b}(\theta)$, which will be defined in Section 4.

# 4 A Simulator of $I'(X;Y)$

As stated in Section 1, albeit rather difficult to compute analytically, $I_n(X;Y)$ 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 [8], which is a well-established tool in proving limit theorems for mixing sequences; see, e.g., [14].

Now, consider a stationary stochastic process $Z = Z_{-\infty}^{\infty}$ satisfying the following assumptions:

(4.a) There exist $C', C'' > 0$ such that for all $z_{-n}^0$,

$$C' \le p(z_0|z_n^{-1}) \le C''.$$

(4.b) There exist $C > 0$, $0 < \lambda < 1$ such that for all $n$,

$$\psi_Z(n) \triangleq \sup_{U \in \mathcal{B}(Z_{-\infty}^{-n}), V \in \mathcal{B}(Z_0^{\infty}), P(U)>0, P(V)>0} |P(V|U) - P(V)|/P(V) \le C\lambda^n,$$

where $\mathcal{B}(Z_i^j)$ denotes the $\sigma$-field generated by $\{Z_k : k = i, i+1, \cdots, j\}$.

(4.c) There exist $C > 0$, $0 < \rho < 1$ such that for any two $z_{-m}^0, \hat{z}_{-\hat{m}}^0$ with $z_{-n}^0 = \hat{z}_{-n}^0$ (here $m, \hat{m} \ge n \ge 0$),

$$|p(z_0|z_{-m}^{-1}) - p(\hat{z}_0|\hat{z}_{-\hat{m}}^{-1})| \le C\rho^n.$$

**Remark 4.1.** Conditions (4.a)-(4.c) are the same ones used in Section 2 of [19], which are essential for establishing the main results in [19]. As observed in [19], Condition (2.a) implies that $Y$ and $(X, Y)$ both satisfy Conditions (4.a)-(4.c).

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 $iq + (i-1)p + 1 \leq j \leq iq + ip$, define

$$W_j = W_j(Z^j_{j-\lfloor q/2 \rfloor}) \triangleq -\frac{p'(Z^j_{j-\lfloor q/2 \rfloor})}{p(Z^j_{j-\lfloor q/2 \rfloor})} \log p(Z_j | Z^{j-1}_{j-\lfloor q/2 \rfloor}), \tag{3}$$

and furthermore

$$\zeta_i \triangleq W_{iq+(i-1)p+1} + \cdots + W_{iq+ip}, \quad S_n \triangleq \sum_{i=1}^{k(n)} \zeta_i. \tag{4}$$

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

**Definition 4.2.**

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

The following lemma, whose proof is somewhat similar to that of Lemma 3.3 in [19], gives an estimate of the variance of $S_n$. [19].

**Lemma 4.3.** *For $Z$ satisfying Conditions (4.a), (4.b) and (4.c),*

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

*Proof.* As in [19], using Condition (4.a), (4.b), we can deduce that for some $0 < \lambda < 1$,

$$E[(S_n - E[S_n])^2] = E[(\sum_{i=1}^{k} \zeta_i - \sum_{i=1}^{k} E[\zeta_i])^2] = kE[(\zeta_i - E[\zeta_i])^2] + O(k^2 \lambda^{q/2}).$$

So, to prove the lemma, it suffices to prove that for any $i \in \mathbb{N}$,

$$E[(\zeta_i - E[\zeta_i])^2] = O(pq^3).$$

Note that

$$E[(\zeta_i - E[\zeta_i])^2] = E[(\sum_{i=1}^{k} W_i - E[W_i])^2] = \sum_{i,j=1}^{k} E[(W_i - E[W_i])(W_j - E[W_j])]. \tag{5}$$

It is apparent that when $|j - i| \leq \lfloor q/2 \rfloor$,

$$E[(W_i - E[W_i])(W_j - E[W_j])] = O(q^2), \tag{6}$$

and one verifies, using Condition (4.a), (4.b), that when $|j-i| > \lfloor q/2 \rfloor$,

$$E[(W_i - E[W_i])(W_j - E[W_j])] = O(q^2 \lambda^{|j-i|-\lfloor q/2 \rfloor}). \tag{7}$$

Combining (5), (6) and (7), we then have

$$\begin{aligned} E[(\zeta_i - E[\zeta_i])^2] &= (\sum_{|j-i| \le \lfloor q/2 \rfloor} + \sum_{|j-i| > \lfloor q/2 \rfloor}) E[(W_i - E[W_i])(W_j - E[W_j])] \\ &= O(pq^3). \end{aligned}$$

The proof is then complete.

□

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

Using similar techniques as in the proof of Theorem 1.1 in [20], the first theorem shows that on average, our simulator sub-exponentially converges to $I'(X;Y)$.

**Theorem 4.4.** *For some $0 < \rho_0 < 1$, we have*

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

*Proof.* Notice that for the Markov chain $X$, we have

$$H(X) = H(X_2|X_1).$$

So, by Remark 4.1, it suffices to prove that for any $Z$ satisfying Conditions (4.a)-(4.c), we have

$$\frac{E[S_n]}{kp} - H'(Z) = O(\rho_1^{\lfloor q/2 \rfloor}),$$

for some $0 < \rho_1 < 1$.

Note that for any $j$ with $iq + (i-1)p + 1 \le j \le iq + ip$, we have

$$\begin{aligned} E[W_j] &= -\sum_{z_{j-\lfloor q/2 \rfloor}^j} p(z_{j-\lfloor q/2 \rfloor}^j) \frac{p'(z_{j-\lfloor q/2 \rfloor}^j)}{p(z_{j-\lfloor q/2 \rfloor}^j)} \log p(z_j | z_{j-\lfloor q/2 \rfloor}^{j-1}) \\ &= -\sum_{z_{j-\lfloor q/2 \rfloor}^j} p'(z_{j-\lfloor q/2 \rfloor}^j) \log p(z_j | z_{j-\lfloor q/2 \rfloor}^{j-1}). \end{aligned}$$

Then, following [20], we can prove that for any small $\varepsilon$, we have

$$\sum_{z_1^n} |p'(z_n | z_1^{n-1})| = O((1+\varepsilon)^n).$$

This, together with Condition (4.c), implies that for some $0 < \rho_1 < 1$,

$$E[W_j] - H'(Z) = O(\rho_1^{\lfloor q/2 \rfloor}),$$

which further implies that for some $0 < \rho_1 < 1$

$$\frac{E[S_n]}{kp} - H'(Z) = \frac{E[S_n] - kpH'(Z)}{kp} = \frac{\sum_j (E[W_j] - H'(Z))}{kp} = O(\rho_1^{\lfloor q/2 \rfloor}).$$

□

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 4.5.** *For any $\varepsilon > 0$, there exist some $0 < \gamma, \delta < 1$ such that*

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

*Proof.* By Lemma 4.4 and Remark 4.1, it suffices to prove that for any $Z$ satisfying Conditions (4.a)-(4.c) and for any $\varepsilon > 0$, there exist $0 < \gamma, \delta < 1$ such that

$$P\left(\left|\frac{S_n - E[S_n]}{kp}\right| \geq \varepsilon\right) \leq \gamma^{n^\delta}. \tag{8}$$

By the Markov inequality, we have

$$P(S_n - E[S_n] \geq kp\varepsilon) = P\left(\frac{t(S_n - E[S_n])}{p^2} \geq \frac{tkp\varepsilon}{p^2}\right) \leq \frac{E[e^{t(S_n - E[S_n])/p^2}]}{e^{tk\varepsilon/p}}. \tag{9}$$

As in [19], applying Conditions (4.a) and (4.b), we then have

$$\begin{aligned} E[e^{t(S_n - E[S_n])/p^2}] &= E[e^{t\sum_{i=1}^{k-1}(\zeta_i - E[\zeta_i])/p^2} e^{t(\zeta_k - E[\zeta_k])/p^2}] \\ &= (1 + O(\lambda^{q(n)/2}))E[e^{t\sum_{i=1}^{k-1}(\zeta_i - E[\zeta_i])/p^2}]E[e^{t\zeta_k}], \end{aligned} \tag{10}$$

for some $0 < \lambda < 1$. An iterative application of (10) yields that for any $0 < t < 1$

$$\begin{aligned} E[e^{t(S_n - E[S_n])/p^2}] &= E[e^{t\sum_{i=1}^{k}(\zeta_i - E[\zeta_i])/p^2}] \\ &= (1 + O(\lambda^{q(n)/2}))^{k-1}(E[e^{t(\zeta_1 - E[\zeta_1])/p^2}])^k, \end{aligned} \tag{11}$$

as $n$ goes to infinity. By Condition (4.a), we have

$$\zeta_1 - E[\zeta_1] = O(pq), \text{ and thus, } O((\zeta_1 - E[\zeta_1])^2/p^4) = O(q^2/p^2) = o(1).$$

It then follows that for any $0 < t < 1$,

$$E[e^{t(\zeta_1 - E[\zeta_1])/p^2}] = 1 + o(1)t^2.$$

Choosing $t = n^{-(1-\alpha)/2}$, then, by (9) and (11), we deduce that

$$\begin{aligned} P\left(\frac{S_n - E[S_n]}{kp} \geq \varepsilon\right) &\leq \frac{E[e^{t(S_n - E[S_n])/p^2}]}{e^{tk\varepsilon/p}} \\ &\leq (1 + O(\lambda^{q(n)/2}))^k \frac{(1 + o(1)t^2)^{n^{1-\alpha}}}{(1 + t\varepsilon + O(1)t^2)^{n^{1-2\alpha}}} \\ &= O(e^{-n^{1/2-3\alpha/2}}). \end{aligned}$$

Noticing that $0 < \alpha < 1/3$ (and thus $1/2 - 3\alpha/2 < 0$), we conclude that for any $\varepsilon > 0$, there exists $0 < \gamma, \delta < 1$ such that

$$P\left(\frac{S_n - E[S_n]}{kp} \geq \varepsilon\right) \leq \gamma^{n^\delta}.$$

With a parallel argument, one verifies that for any $\varepsilon > 0$, there exists $0 < \gamma, \delta < 1$ such that

$$P\left(\frac{S_n - E[S_n]}{kp} \leq -\varepsilon\right) \leq \gamma^{n^\delta},$$

which immediately implies (8). The proof is then complete. □

The following theorem states that our simulator is asymptotically unbiased.

**Theorem 4.6.** *With probability* 1*,*

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

*as* $n$ *tends to* $\infty$*.*

*Proof.* It immediately follows from Theorem 4.5 and the Borel-Cantelli lemma. □

**Remark 4.7.** Other than confirming the effectiveness of the proposed simulator $g_n$ in Definition 4.2, Theorems 4.4, 4.5 and 4.6 are of interest in their own right: For any stationary process $Z$ satisfying Conditions (4.a), (4.b) and (4.c), the proofs of these theorems reveals that the sample path of $Z$ can be used to effectively simulate the derivative of $H(Z)$. This can be viewed as an extension (for a special class of processes) of the Shannon-Millan-Breiman theorem, which states that the sample path of a stationary and ergodic process can be used to simulate its entropy rate. For the proposed algorithm, an alternative simulator can be defined as follows: $g_n$ can be defined in the same way except that $\zeta_2, \zeta_3, \cdots, \zeta_n$ are taken as independent copies of $\zeta_1$. It can be easily checked that Theorems 4.4, 4.5 and 4.6 also holds for the redefined $g_n$. As a consequence, the convergence results later in this paper also hold for the redefined $g_n$.

**Remark 4.8.** In our notation, the following expression has been proposed in [53] 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 [53] 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.

**Remark 4.9.** For the purpose of effective computation to avoid accumulating round-off errors caused by multiplication, $W_j$ can be rewritten as

$$W_j = -\left(\frac{p'(Z_{j-\lfloor q/2 \rfloor})}{p(Z_{j-\lfloor q/2 \rfloor})} + \frac{p'(Z_{j-\lfloor q/2 \rfloor+1}|Z_{j-\lfloor q/2 \rfloor})}{p(Z_{j-\lfloor q/2 \rfloor+1}|Z_{j-\lfloor q/2 \rfloor})} + \cdots + \frac{p'(Z_j|Z_{j-\lfloor q/2 \rfloor}^{j-1})}{p(Z_j|Z_{j-\lfloor q/2 \rfloor}^{j-1})}\right) \log p(Z_j|Z_{j-\lfloor q/2 \rfloor}^{j-1}).$$

# 5 Convergence

Consider the following condition:

(5.a) $P(\cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n) = 0$, that is, $A_n, n \in \mathbb{N}$, only occurs finitely many times,

which will be assumed throughout the convergence analysis in the paper. Particularly, in this section, assuming (5.a), we will show that $\{I(X(\theta_n); Y(\theta_n))\}$ converges almost surely. Note that if $\Theta = \mathbb{R}^d$, then Assumption (5.a) will be trivially satisfied and the iteration in (1) can be simply written as

$$\theta_{n+1} = \theta_n + a_n g_{n^b}(\theta_n). \tag{12}$$

In fact, unless specified otherwise, we will simply assume that $\theta = \mathbb{R}$ in all the proofs in this paper to avoid obscuring the main idea. The proofs of the same results under Assumption (5.a) follow from parallel arguments only with an increasing level of notational complexity.

Henceforth, we will write

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

Note that under Assumption (2.a), Theorem 1.1 of [20] implies that

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

a key fact that will be implicitly used throughout the paper. Now, rewrite (12) as

$$\theta_{n+1} = \theta_n + a_n f'(\theta_n) + a_n R_n(\theta_n), \tag{13}$$

where

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

It can be easily verified that

$$\begin{aligned} f(\theta_{n+1}) - f(\theta_n) &= \int_0^1 f'(\theta_n + t(\theta_{n+1} - \theta_n))(\theta_{n+1} - \theta_n) dt \\ &= \int_0^1 f'(\theta_n)(\theta_{n+1} - \theta_n) dt + \int_0^1 (f'(\theta_n + t(\theta_{n+1} - \theta_n)) - f'(\theta_n))(\theta_{n+1} - \theta_n) dt \\ &= a_n f'(\theta_n)(f'(\theta_n) + R_n(\theta_n)) + \int_0^1 (f'(\theta_n + t(\theta_{n+1} - \theta_n)) - f'(\theta_n))(\theta_{n+1} - \theta_n) dt \\ &= a_n f'^2(\theta_n) + \hat{R}_n(\theta_n), \end{aligned} \tag{14}$$

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

*Proof.* Let

$$T_1 = \sum_{n=0}^{\infty} a_n f'(\theta_n) R_n(\theta_n), \quad T_2 = \sum_{n=0}^{\infty} \int_0^1 (f'(\theta_n + t(\theta_{n+1} - \theta_n)) - f'(\theta_n))(\theta_{n+1} - \theta_n) dt.$$

It suffices to prove that $T_1, T_2$ both converge almost surely.

For $T_1$, note that

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

It follows from Theorem 4.4 that there exists $0 < \rho_0 < 1$ such that

$$\sum_{n=0}^{\infty} a_n |f'(\theta_n)||(f'_{n^b}(\theta_n) - f'(\theta_n))| \leq \sum_{n=0}^{\infty} a_n |f'(\theta_n)| \rho_0^{n^b} < \infty. \tag{15}$$

Then, using Lemma 4.3, one verifies that uniformly over all $\theta_n \in \Theta$,

$$\sum_{n=0}^{\infty} E[\{a_n^2 (f'(\theta_n))^2 R_n^2(\theta_n)\}] = \sum_{n=0}^{\infty} O\left(\frac{1}{n^{2a+b(1-3\beta)}}\right), \tag{16}$$

which converges since $2a+b-3b\beta > 1$. Noting that $\{a_n f'(\theta_n) R_n(\theta_n), \mathcal{B}(X_1^n)\}$ is a Martingale difference sequence and applying Doob's Martingale convergence theorem (see Theorem 2.8.7 of [48]), 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.

For $T_2$, it is easy to check that

$$\left| \int_0^1 (f'(\theta_n + t(\theta_{n+1} - \theta_n)) - f'(\theta_n))(\theta_{n+1} - \theta_n) dt \right| = O((\theta_{n+1} - \theta_n)^2) = O(a_n^2 (f'(\theta_n))^2) + O(a_n^2 R_n^2(\theta_n)).$$

Similarly as in deriving (15) and (16), we have

$$\sum_{n=0}^{\infty} a_n^2 (f'_{n^b}(\theta_n) - f'(\theta_n))^2 < \infty, \quad \sum_{n=0}^{\infty} E[a_n^2 (g_{n^b}(\theta_n) - f'_{n^b}(\theta_n))^2] < \infty,$$

and furthermore,

$$\sum_{n=0}^{\infty} a_n^2 (g_{n^b}(\theta_n) - f'_{n^b}(\theta_n))^2$$

converges almost surely. This, together with (15), further implies that

$$\sum_{n=0}^{\infty} a_n^2 |(g_{n^b}(\theta_n) - f'_{n^b}(\theta_n))(f'_{n^b}(\theta_n) - f'(\theta_n)|$$

converges almost surely. Recalling that

$$R_n(\theta_n) = g_{n^b}(\theta_n) - f'_{n^b}(\theta_n) + f'_{n^b}(\theta_n) - f'(\theta_n),$$

we conclude that

$$\sum_{n=0}^{\infty} a_n^2 R_n^2(\theta_n) < \infty,$$

which further implies that

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

converges almost surely. The proof is then complete. □

We are now ready for the following convergence theorem, whose proof closely follows that of Lemma 7 in [49], whose general ideas can be further traced back to the standard proof of the Martingale convergence theorem [48].

**Theorem 5.2.** *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 5.1, 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

$$|f'(\theta_{m_k})| \le \varepsilon, \quad |f'(\theta_{n_k})| \ge 2\varepsilon, \quad |f'(\theta_i)| \ge \varepsilon \tag{17}$$

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

$$\begin{aligned}
\varepsilon &\le |f'(\theta_{n_k}) - f'(\theta_{m_k})| \\
&= O(|\theta_{n_k} - \theta_{m_k}|) \\
&= O\left(\sum_{i=m_k}^{n_k-1} a_i |f'(\theta_i)|\right) + O\left(\left|\sum_{i=m_k}^{n_k-1} a_i R_i(\theta_i)\right|\right) \\
&= O\left(\sum_{i=m_k}^{n_k-1} a_i\right) + O\left(\left|\sum_{i=m_k}^{n_k-1} a_i R_i(\theta_i)\right|\right). \qquad (18)
\end{aligned}$$

As in the proof of Lemma 5.1, using the decomposition

$$R_n(\theta_n) = g_{n^b}(\theta_n) - f'(\theta_n) = g_{n^b}(\theta_n) - f'_{n^b}(\theta_n) + f'_{n^b}(\theta_n) - f'(\theta_n),$$

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 $\infty$. On the other hand, by (17), we have

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

This implies that as $k$ tends to $\infty$, $\sum_{i=m_k}^{n_k-1} a_i$ tends to zero, which, together with (18), further implies that

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

a contradiction. □

**Remark 5.3.** The fact that $\{f(\theta_n)\}$ converges almost surely does not necessarily imply that $\{\theta_n\}$ converges almost surely. In the following sections, we will prove that, under some assumptions, $\{\theta_n\}$ does converge almost surely.

**Numerical simulation.** Extensive numerical simulations of the algorithm have been conducted for a number of practical channels, (typically fast) convergence of $\{f(\theta_n)\}$ and $\{\theta_n\}$ have been observed. In particular, for a binary symmetric channel with $(1, \infty)$-RLL input constraint (see Example 9.4), our simulations yield the same capacity curve (as the crossover probability varies) as the GBAA; see Figure 15 in [53]. For the classical Gilbert-Elliott channel, it has been shown [37] that the capacity achieving distribution is the uniform distribution, fast convergence of our algorithm to the capacity and its achieving distribution has also been observed.

# 6 Some Estimations

In this section, assuming (5.a), we will derive some estimations that will be used in the later sections for convergence analysis.

For any $j \in \mathbb{N}$, let

$$A_j = a_1 + a_2 + \cdots + a_{j-1},$$

and for any $h > 0$ and any $n \in \mathbb{N}$, define

$$t(n,h) \triangleq \min\{k : a_n + a_{n+1} + \cdots + a_{k-1} \geq h\}.$$

Now, for any fixed $n_0 \in \mathbb{N}$, recursively define

$$n_{k+1} = t(n_k, h).$$

One then verifies that for $k$ sufficiently large,

$$A_{n_{k+1}} - A_{n_k} = \hat{O}(h), \quad n_k = \hat{O}(k^{1/(1-a)}), \tag{19}$$

where by $M = \hat{O}(N)$, we mean that there exist positive constants $C_1, C_2$ such that

$$C_1 N \leq M \leq C_2 N.$$

Now, an iterated application of

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

yields

$$\begin{aligned}
\theta_k &= \theta_n + \sum_{i=n}^{k-1} a_i f'(\theta_i) + \sum_{i=n}^{k-1} a_i R_i(\theta_i) \\
&= \theta_n + (A_k - A_n) f'(\theta_n) + \sum_{i=n}^{k-1} a_i R_i(\theta_i) + \sum_{i=n}^{k-1} a_i (f'(\theta_i) - f'(\theta_n)) \\
&= \theta_n + R_{n,k},
\end{aligned}$$

where

$$R_{n,k} = \sum_{i=n}^{k-1} a_i R_i(\theta_i) + \sum_{i=n}^{k-1} a_i (f'(\theta_i) - f'(\theta_n)). \tag{20}$$

Similarly, an iterated application of

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

yields

$$\begin{aligned}
f(\theta_k) - f(\theta_n) &= \int_0^1 f'(\theta_n + t(\theta_k - \theta_n))(\theta_k - \theta_n) dt \\
&= \int_0^1 f'(\theta_n)(\theta_k - \theta_n) dt + \int_0^1 (f'(\theta_n + t(\theta_k - \theta_n)) - f'(\theta_n))(\theta_k - \theta_n) dt \\
&= f'(\theta_n)((A_k - A_n) f'(\theta_n) + R_{n,k}) + \int_0^1 (f'(\theta_n + t(\theta_k - \theta_n)) - f'(\theta_n))(\theta_k - \theta_n) dt \\
&= (A_k - A_n) f'^2(\theta_n) + f'(\theta_n) R_{n,k} + \int_0^1 (f'(\theta_n + t(A_k - A_n)) - f'(\theta_n))(A_k - A_n) dt \\
&= (A_k - A_n) f'^2(\theta_n) + \hat{R}_{n,k}(\theta_n),
\end{aligned} \tag{21}$$

where

$$\hat{R}_{n,k}(\theta_n) = f'(\theta_n)R_{n,k} + \int_0^1 (f'(\theta_n + t(A_k - A_n)) - f'(\theta_n))(A_k - A_n)dt. \tag{22}$$

The following lemma introduces a positive random variable, $\tilde{C}_0$, and a constant, $\tau$, which will be referred to throughout the rest of the paper.

**Lemma 6.1.** *There exists a positive random variable $\tilde{C}_0$ such that for all $n$ and for any $\tau > 0$ with $2a + b - 3b\beta - 2\tau > 1$,*

$$\sup_{k \geq n} \left| \sum_{i=n}^{k} a_i R_i(\theta_i) \right| \leq \tilde{C}_0 n^{-\tau} \ \ a.s.$$

*Proof.* For any $\tau > 0$ with $2a + b - 3b\beta - 2\tau > 1$, as in the proof of Lemma 5.1, we deduce that $\sum_{i=1}^{\infty} i^{\tau} a_i R_i$ converges almost surely. Letting

$$T_n \triangleq \sum_{i=1}^{n} i^{\tau} a_i R_i(\theta_i),$$

we then have for any $k \geq n$,

$$\begin{aligned}
\sum_{i=n}^{k} a_i R_i(\theta_i) &= \sum_{i=n}^{k} (i^{\tau} a_i R_i(\theta_i)) i^{-\tau} \\
&= \sum_{i=n}^{k} (T_i - T_{i-1}) i^{-\tau} \\
&= \sum_{i=n}^{k} T_i i^{-\tau} - \sum_{i=n+1}^{k} T_{i-1} i^{-\tau} \\
&= \sum_{i=n}^{k} T_i i^{-\tau} - \sum_{i=n}^{k-1} T_i (i+1)^{-\tau} \\
&= T_k k^{-\tau} + \sum_{i=n}^{k-1} (i^{-\tau} - (i+1)^{-\tau}) T_i \\
&\leq (k^{-\tau} + \sum_{i=n}^{k-1} (i^{-\tau} - (i+1)^{-\tau})) \sup_i T_i \\
&= n^{-\tau} \sup_i T_i,
\end{aligned}$$

which immediately implies the lemma. □

In the following, to avoid notational cumbersomeness, we will use $C$ to denote a positive constant, which may not be the same on its each appearance.

**Lemma 6.2.** *Let $0 < h < 1$ and $\tilde{C}_0, \tau$ be as in Lemma 6.1, then we have*

*(1) there exists a constant $C > 0$ such that*

$$|f'(\theta_{t(n,h)})| \le C(\tilde{C}_0 n^{-\tau} + |f'(\theta_n)|).$$

*(2) there exists a constant $C > 0$ such that*

$$|\theta_{t(n,h)} - \theta_n| \le C(\tilde{C}_0 n^{-\tau} + h|f'(\theta_n)|).$$

*(3) there exists a constant $C > 0$ such that*

$$|R_{n,t(n,h)}| \le C(\tilde{C}_0 n^{-\tau} + h^2|f'(\theta_n)|).$$

*(4) there exists a constant $C > 0$ such that*

$$|\hat{R}_{n,t(n,h)}| \le C(\tilde{C}_0^2 n^{-2\tau} + \tilde{C}_0 n^{-\tau}|f'(\theta_n)| + h^2|f'(\theta_n)|^2).$$

*(5) there exists a constant $C > 0$ such that*

$$f(\theta_n) - f(\theta_{t(n,h)}) \le -(3/4 - 3Ch/2)h|f'(\theta_n)|^2 + C\tilde{C}_0^2 n^{-2\tau}(1 + 1/(2h^2)).$$

*(6) there exists $C > 0$ such that for sufficiently small $h$*

$$2(f(\theta_n) - f(\theta_{t(n,h)})) + |f'(\theta_n)||\theta_{t(n,h)} - \theta_n| \le (C + 1/(2h^2))\tilde{C}_0^2 n^{-2\tau}.$$

*(7) for any $\tau' < \tau$, there exists a positive constant $C$ such that for sufficiently small $h$, we have*

$$|\theta_{t(n,h)} - \theta_n| \le Cn^{\tau'}(f(\theta_{t(n,h)}) - f(\theta_n)) + C\tilde{C}_0^2 n^{-\tau'}.$$

*Proof.* In this proof, for notational simplicity, we will write $t(n,h)$ as $k$.

Note that there exists a positive constant $C$ such that

$$\begin{aligned}
|f'(\theta_k)| &\le |f'(\theta_n)| + |f'(\theta_k) - f'(\theta_n)| \\
&\le |f'(\theta_n)| + C|\theta_k - \theta_n| \\
&\le |f'(\theta_n)| + C\sum_{i=n}^{k-1} a_i|f'(\theta_i)| + C|\sum_{i=n}^{k-1} a_i R_i(\theta_i)|,
\end{aligned}$$

where we have applied (13). Applying Lemma 6.1, we then have

$$|f'(\theta_k)| \le C\tilde{C}_0 n^{-\tau} + |f'(\theta_n)| + C\sum_{i=n}^{k-1} a_i|f'(\theta_i)|.$$

Applying Gronwall's lemma, we then have for $n$ sufficiently large

$$|f'(\theta_k)| \le (C\tilde{C}_0 n^{-\tau} + |f'(\theta_n)|)\exp(C(a_n + a_{n+1} + \cdots + a_{k-1})) \le \exp(C)(C\tilde{C}_0 n^{-\tau} + |f'(\theta_n)|),$$

where we have used the fact that for $n$ large enough

$$a_n + a_{n+1} + \cdots + a_{k-1} \approx h < 1.$$

We have then established (1).

It then follows from (1) that for some $C$

$$\begin{aligned}|\theta_k - \theta_n| &\le \sum_{i=n}^{k-1} a_i |f'(\theta_i)| + |\sum_{i=n}^{k-1} a_i R_i(\theta_i)| \\ &\le (A_k - A_n)(C\tilde{C}_0 n^{-\tau} + C|f'(\theta_n)|) + \tilde{C}_0 n^{-\tau},\end{aligned}$$

which immediately implies (2).

Now, by (20) and (2), we have for some $C$

$$\begin{aligned}|R_{n,k}| &\le \tilde{C}_0 n^{-\tau} + C \sum_{i=n}^{k-1} a_i |\theta_i - \theta_n| \\ &\le \tilde{C}_0 n^{-\tau} + C^2 (A_k - A_n)(\tilde{C}_0 n^{-\tau} + (A_k - A_n)|f'(\theta_n)|),\end{aligned}$$

which establishes (3).

Furthermore, by (22), (2) and (3), we have

$$\begin{aligned}|\hat{R}_{n,k}| &\le |f'(\theta_n)||R_{n,k}| + C|\theta_k - \theta_n|^2 \\ &\le C\tilde{C}_0 n^{-\tau}|f'(\theta_n)| + C(A_k - A_n)^2 |f'(\theta_n)|^2 + 2C^3(\tilde{C}_0^2 n^{-2\tau} + (A_k - A_n)^2 |f'(\theta_n)|^2),\end{aligned}$$

which establishes (4).

It then follows from (21), (III) and (IV) and that for sufficiently large $n$

$$\begin{aligned}f(\theta_n) - f(\theta_k) &\le -(A_k - A_n)|f'(\theta_n)|^2 + |\hat{R}_{n,k}| \\ &\le -3h/4|f'(\theta_n)|^2 + C(\tilde{C}_0^2 n^{-2\tau} + \tilde{C}_0 n^{-\tau}|f'(\theta_n)| + h^2 |f'(\theta_n)|^2) \\ &\le -3h/4|f'(\theta_n)|^2 + C(\tilde{C}_0^2 n^{-2\tau} + \tilde{C}_0^2 n^{-2\tau}/(2h^2) + h^2|f'(\theta_n)|^2/2 + h^2|f'(\theta_n)|^2) \\ &\le -3h/4|f'(\theta_n)|^2 + C(\tilde{C}_0^2 n^{-2\tau}(1 + 1/(2h^2)) + 3h^2/2|f'(\theta_n)|^2) \\ &\le -(3/4 - 3Ch/2)h|f'(\theta_n)|^2 + C\tilde{C}_0^2 n^{-2c}(1 + 1/(2h^2)),\end{aligned}$$

which establishes (5).

It follows from (21), (3) and (4) that

$$\begin{aligned}f(\theta_k) - f(\theta_n) &= |f'(\theta_n)||(A_k - A_n)f'(\theta_n)| + \hat{R}_{n,k} \\ &= |f'(\theta_n)||\theta_k - \theta_n + R_{n,k}| + \hat{R}_{n,k} \\ &\ge |f'(\theta_n)|(|\theta_k - \theta_n| - |R_{n,k}|) - \hat{R}_{n,k} \\ &\ge |f'(\theta_n)||\theta_k - \theta_n| - C(\tilde{C}_0^2 n^{-2\tau} + \tilde{C}_0 n^{-\tau}|f'(\theta_n)| + (A_k - A_n)^2 |f'(\theta_n)|^2),\end{aligned}$$

which implies that

$$\begin{aligned}f(\theta_n) - f(\theta_k) + |f'(\theta_n)||\theta_k - \theta_n| &\le C(\tilde{C}_0^2 n^{-2\tau} + \tilde{C}_0 n^{-\tau}|f'(\theta_n)| + (A_k - A_n)^2 |f'(\theta_n)|^2) \\ &\le C(\tilde{C}_0^2 n^{-2\tau}(1 + 1/(2h^2)) + 3h^2/2|f'(\theta_n)|^2).\end{aligned}$$

Applying (V), we then have for sufficiently small $h$,

$$f(\theta_n) - f(\theta_k) + |f'(\theta_n)||\theta_k - \theta_n| \leq 2C(1 + 1/(2h^2))\tilde{C}_0^2 n^{-2\tau} + f(\theta_k - f(\theta_n)),$$

which can be rewritten as

$$2(f(\theta_n) - f(\theta_k)) + |f'(\theta_n)||\theta_k - \theta_n| \leq 2C(1 + 1/(2h^2))\tilde{C}_0^2 n^{-2\tau},$$

which establishes (6).

We next prove (7). If $|f'(\theta_n)| \leq n^{-\tau'}$, applying (II), we deduce that

$$|\theta_k - \theta_n| \leq C\tilde{C}_0 n^{-\tau'} + Ch^2 n^{-\tau'}. \tag{23}$$

It follows from (21) and (4) that

$$\begin{aligned}
|f(\theta_k) - f(\theta_n)| &\leq (A_k - A_n)f'^2(\theta_n) + |\hat{R}_{n,k}(\theta_n)| \\
&\leq (A_k - A_n)f'^2(\theta_n) + C(\tilde{C}_0^2 n^{-2\tau} + \tilde{C}_0 n^{-\tau}|f'(\theta_n)| + h^2|f'(\theta_n)|^2) \\
&\leq (A_k - A_n)f'^2(\theta_n) + C(\tilde{C}_0^2 n^{-2\tau'} + \tilde{C}_0 n^{-\tau'}|f'(\theta_n)| + h^2|f'(\theta_n)|^2) \\
&\leq C(\tilde{C}_0^2 n^{-2\tau'}(1 + 1/(2h^2))) + (h + 3Ch^2/2)|f'(\theta_n)|^2,
\end{aligned}$$

which, together with (23), immediately implies that for some $C$,

$$\begin{aligned}
|\theta_k - \theta_n| &\leq n^{\tau'}(f(\theta_k) - f(\theta_n)) + n^{\tau'}|f(\theta_k) - f(\theta_n)| + C\tilde{C}_0 n^{-\tau'} + Ch^2 n^{-\tau'} \\
&\leq n^{\tau'}(f(\theta_k) - f(\theta_n)) + C(\tilde{C}_0^2 n^{-2\tau'}(1 + 1/(2h^2))) \\
&+ (h + 3Ch^2/2)|f'(\theta_n)|^2 + C\tilde{C}_0 n^{-\tau'} + Ch^2 n^{-\tau'}.
\end{aligned} \tag{24}$$

On the other hand, if $|f'(\theta_n)| \geq n^{-\tau'}$, applying (6), we deduce that

$$\begin{aligned}
|\theta_k - \theta_n| &\leq 2|f'(\theta_n)|^{-1}(f(\theta_k - f(\theta_n))) + (C + 1/(2h^2))|f'(\theta_n)|^{-1}\tilde{C}_0^2 n^{-2\tau} \\
&\leq 2n^{\tau'}(f(\theta_k - f(\theta_n))) + (C + 1/(2h^2))\tilde{C}_0^2 n^{-\tau'}.
\end{aligned} \tag{25}$$

Combining (24) and (25), we then have established (7). □

# 7 Rate of Convergence with Concavity

In this section, we assume that

(7.a) $f(\theta)$ is strictly concave with respect to $\theta$. More precisely, there exists $\hat{\epsilon} > 0$ such that for any $\theta_1, \theta_2 \in \Theta$,

$$f'_t(t\theta_1 + (1-t)\theta_2) \geq \hat{\epsilon},$$

for all $0 \leq t \leq 1$.

(7.b) With probability 1, $\theta_n$ converges to the unique global maximum $\theta^*$ as $n$ tends to $\infty$.

Here, let us note that (7.a), together with Theorem 4.5, implies (5.a). With Assumptions (7.a) and (7.b), which, as argued in Section 9, can be satisfied for a class of finite-state channels, we will derive the convergence rate of $\{\theta_n\}$. Again, for notational convenience only, we assume that $\Theta = \mathbb{R}$ in the proofs.

From

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

trivially we have

$$\Delta_{n+1} - \Delta_n = -a_n f'(\theta_n) - a_n R_n(\theta_n),$$

where

$$\Delta_n \triangleq (\theta^* - \theta_n).$$

It immediately from the above two conditions that for $\theta$ sufficiently close to $\theta^*$

$$f(\theta) = \hat{O}(|\theta^* - \theta|^2), \qquad f'(\theta) = \hat{O}(|\theta^* - \theta|). \tag{26}$$

So, if $\theta_n$ is sufficiently close to $\theta^*$, we will have

$$f(\theta_n) = \hat{O}(\Delta_n^2), \qquad f'(\theta_n) = \hat{O}(|\Delta_n|).$$

Throughout the paper, by $M = \tilde{O}(N)$, we mean that there exists a positive random variable $\tilde{C}$ such that with probability 1,

$$|M| \leq \tilde{C} N.$$

In this section, we will prove that $\Delta_n$ is at most of order $\tilde{O}(n^{-\tau})$.

We first prove the following lemma.

**Lemma 7.1.** *There exists $l \in \mathbb{N}$ such that*

$$\liminf_{n\to\infty} n^\tau |\Delta_n| \leq l\tilde{C}_0.$$

*Proof.* Suppose, by way of contradiction, that for any $l$,

$$n^\tau |\Delta_n| \geq l\tilde{C}_0, \tag{27}$$

as long as $n$ is sufficiently large. First, pick $n_0$ sufficiently large such that (27) is satisfied and then recursively define

$$n_{k+1} = t(n_k, h).$$

for some $0 < h < 1$. We then have, for any feasible $k$,

$$\theta_{n_{k+1}} = \theta_{n_k} + (A_{n_{k+1}} - A_{n_k}) f'(\theta_{n_k}) + R_{n_k, n_{k+1}}.$$

It then follows from Lemma 6.2 (3) and (26) that $R_{n_k,n_{k+1}}$ is dominated by $|f'(\theta_{n_k})|$ as long as $l$ is chosen sufficiently large and $h$ is chosen sufficiently small. Noticing that due to the concavity of $f$, $\Delta_n$ always has the same sign as $f'(\theta_n)$, then we have

$$|\Delta_{n_{k+1}}| \leq |\Delta_{n_k}| - h/2|\Delta_{n_k}| \leq |\Delta_{n_k}| e^{-h/2},$$

an iterative application of which would yield

$$\Delta_{n_k} \le \Delta_{n_0} e^{-kh/2}.$$

It then follows that for any $k$

$$\Delta_{n_0} n_k^\tau e^{-kh/2} \ge n_k^\tau \Delta_{n_k} \ge l\tilde{C}_0.$$

This, together with the fact that (see (19))

$$n_k = \hat{O}(k^{1/(1-a)}),$$

as $k$ tends to infinity, implies that

$$\tilde{C}_0 \le 0,$$

which is a contradiction. □

**Theorem 7.2.**

$$|\Delta_n| = \tilde{O}(n^{-\tau}).$$

*Proof.* It is enough to prove that there exists an integer $l$ such that for all $n$ sufficiently large,

$$n^\tau |\Delta_n| \le l\tilde{C}_0.$$

By way of contradiction, suppose otherwise. Then, by Lemma 7.1, for any $l$ and arbitrarily large $N$, we can find $k_0 > m_0 > N$ such that

$$m_0^\tau \Delta_{m_0} \le 2l\tilde{C}_0, \quad k_0^\tau \Delta_{k_0} \ge 3l\tilde{C}_0,$$

$$\min_{m_0 < n \le k_0} n^\tau \Delta_n > 2l\tilde{C}_0, \quad \max_{m_0 \le n < k_0} n^\tau \Delta_n \le 3l\tilde{C}_0. \tag{28}$$

Now, for some $0 < h < 1$, let $m_1 = t(m_0, h)$. Then, for any $m_0 \le n \le m_1$, it follows from (28) and

$$\theta_n - \theta_{m_0} = (A_n - A_{m_0}) f'(\theta_{m_0}) + R_{m_0,n}, \quad |R_{m_0,n}| \le C(m_0^{-\tau}\tilde{C}_0 + (A_n - A_{m_0})^2 |f'(\theta_{m_0})|)$$

that

$$|\Delta_n - \Delta_{m_0}| = O(m_0^{-\tau})\tilde{C}_0.$$

Applying (28), we then deduce that for sufficiently small $h$

$$\begin{aligned} |n^\tau \Delta_n - m_0^\tau \Delta_{m_0}| &\le n^\tau |\Delta_n - \Delta_{m_0}| + (n^\tau - m_0^\tau)\Delta_{m_0} \\ &\le O(m_0^\tau) O(m_0^{-\tau})\tilde{C}_0 + o(m_0^\tau) 2l m_0^{-\tau} \tilde{C}_0, \end{aligned}$$

where we have used the fact that

$$n^\tau = O(m_0^\tau), \quad n^\tau - m_0^\tau = o(m_0^\tau).$$

It then follows that, with $l$ large enough and $h$ small enough, we have

$$|n^\tau \Delta_n - m_0^\tau \Delta_{m_0}| \le l\tilde{C}_0.$$

In particular, we have

$$|(m_0+1)^\tau \Delta_{m_0+1} - m_0^\tau \Delta_{m_0}| \le l\tilde{C}_0 \text{ and } |m_1^\tau \Delta_{m_1} - m_0^\tau \Delta_{m_0}| \le l\tilde{C}_0,$$

which further implies that

$$m_0^\tau \Delta_{m_0} \ge l\tilde{C}_0 \text{ and } m_1 < k_0,$$

respectively.

Now, for some $0 < h < 1$, we have

$$\theta_{m_1} = \theta_{m_0} + (A_{m_1} - A_{m_0}) f'(\theta_{m_0}) + R_{m_0,m_1},$$

and

$$|R_{m_0,m_1}| \le C(m_0^{-\tau}\tilde{C}_0 + (A_{m_1} - A_{m_0})^2 |f'(\theta_{m_0})|).$$

As in the proof of Lemma 7.1, if $l$ is chosen large enough, then $|f'(\theta_{m_0})|$ will dominate $|R_{m_0,m_1}|$. Again, due to the concavity of $f$, $\Delta_{m_0}$ always has the same sign as $f'(\theta_{m_0})$, then for sufficiently small $h > 0$, we have

$$|\Delta_{m_1}| \le |\Delta_{m_0}| - h/2|\Delta_{m_0}|.$$

Then, for $m_0$ sufficiently large such that

$$m_1^\tau < m_0^\tau/(1 - h/2),$$

we have

$$m_1^\tau |\Delta_{m_1}| \le m_1^\tau |\Delta_{m_0}|(1 - h/2) < m_0^\tau |\Delta_{m_0}| \le 2l\tilde{C}_0,$$

which is a contradiction to (28).

□

# 8 Rate of Convergence without Concavity

In this section, assuming (5.a) and

(8.a) with probability 1, $\theta_n \in Q$ for all $n$, where $Q$ is a compact subset of $\Theta$,

we derive the rate of convergence of our algorithm. Again, for notational convenience only, we assume that $\Theta = \mathbb{R}$.

As one of the main results in real algebraic geometry, the Lojasiewicz inequality [9], among many other applications, has been widely applied to the convergence analysis of a broad class of dynamical systems. In this section, we will first use the "function" version of the Lojasiewicz inequality (Lemma 8.1) to prove that $\{f(\theta_n)\}$ converges almost surely and derive the convergence rate, which can be further used to derive the convergence rate of $\{\theta_n\}$. Then, using the "variable" version of the Lojasiewicz inequality (Lemma 8.7), the rate of convergence can be refined. The above-mentioned framework is essentially due to Tadic [49], however, a comprehensive adaptation to our settings has been done in this section.

Following [49], we state the "function" version of the Lojasiewicz inequality as below.

**Lemma 8.1.** *For any compact set $Q \subset \Theta$ and real number $z \in f(Q)$, there exist real numbers $\delta_{Q,z} \in (0,1)$, $\mu_{Q,z} \in (1,2]$ and $M_{Q,z} \in [1,\infty)$ such that*

$$|f(\theta) - z| \leq M_{Q,z}|f'(\theta)|^{\mu_{Q,z}}$$

*for all $\theta \in Q$ satisfying $|f(\theta) - z| \leq \delta_{Q,z}$.*

From now on, we will set $\hat{f} = \lim_{n\to\infty} f(\theta_n)$ and write $\mu = \mu_{Q,\hat{f}}$. Define

$$\hat{\Delta}_n \triangleq \hat{f} - f(\theta_n).$$

We first prove the following lemma.

**Lemma 8.2.** *There exists a positive integer $l$ such that for all $n$ sufficiently large,*

$$n^{\mu\tau}\hat{\Delta}_n \geq -l\tilde{C}_0^{\mu}.$$

*Proof.* Suppose, by way of contradiction, that for any $l$, there exists some $n_0$,

$$n_0^{\mu\tau}\hat{\Delta}_{n_0} < -l\tilde{C}_0^{\mu}. \tag{29}$$

Then, by Lemma 6.2 (5), we have, for some $0 < h < 1$,

$$f(\theta_{n_0}) - f(\theta_{t(n_0,h)}) \leq -(3/4 - 3Ch/2)h|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau}(1 + 1/(2h^2)),$$

which implies for $h$ sufficiently small,

$$\begin{aligned}\hat{\Delta}_{\theta_{t(n_0,h)}}) - \hat{\Delta}_{n_0} &\leq -(3/4 - 3Ch/2)h|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau}(1 + 1/2h^2) \\ &\leq -h/2|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau}(1 + 1/(2h^2)).\end{aligned}$$

Choosing $l$ sufficiently large, then by Lemma 8.1 and (29), we deduce that for $n$ large enough,

$$-h/2|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau} \leq -h/4|f'(\theta_{n_0})|^2,$$

and therefore

$$\hat{\Delta}_{t(n_0,h)} - \hat{\Delta}_{n_0} \leq -h/4|f'(\theta_{n_0})|^2. \tag{30}$$

We then have

$$\hat{\Delta}_{t(n_0,h)} \leq \hat{\Delta}_{n_0} \leq -l\tilde{C}_0^{\mu} n_0^{-\mu\tau} \leq -l\tilde{C}_0^{\mu} t(n_0,h)^{-\mu\tau}.$$

Henceforth, recursively define

$$n_{k+1} = t(n_k, h).$$

It then follows that for any $k$,

$$\hat{\Delta}_{n_k} \leq \hat{\Delta}_{n_0} \leq -l\tilde{C}_0^{\mu} n_0^{-\mu\tau} < 0,$$

which is a contradiction to the fact that almost surely

$$\lim_{k\to\infty} \hat{\Delta}_{n_k} = 0.$$

□

In the remainder of this section, define

$$\hat{\tau} = \min(\mu\tau, \mu(1-a)/(2-\mu)).$$

**Lemma 8.3.** *There exists a positive integer $l$ such that*

$$\liminf_{n\to\infty} n^{\hat{\tau}}\hat{\Delta}_n \leq l\tilde{C}_0^{\mu}$$

*almost surely.*

*Proof.* Suppose, by way of contradiction, that for any $l$, we have

$$n^{\hat{\tau}}\hat{\Delta}_n \geq l\tilde{C}_0^{\mu}, \tag{31}$$

for all $n$ sufficiently large. By Lemma 6.2 (5), for any $0 < h < 1$, we have for $n_0$ large enough,

$$f(\theta_{n_0}) - f(\theta_{t(n_0,h)}) \leq -(3/4 - 3Ch/2)h|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau}(1 + 1/(2h^2)),$$

which implies that for $h$ sufficiently small

$$\begin{aligned}\hat{\Delta}_{t(n_0,h)} - \hat{\Delta}_{n_0} &\leq -(3/4 - 3Ch/2)h|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau}(1 + 1/(2h^2)) \\ &\leq -h/2|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau}(1 + 1/(2h^2)).\end{aligned}$$

Choosing $l$ sufficiently large, then by Lemma 8.1 and (29), we deduce that sufficiently large $n$,

$$-h/2|f'(\theta_{n_0})|^2 + C\tilde{C}_0^2 n_0^{-2\tau} \leq -h/4|f'(\theta_{n_0})|^2,$$

and therefore

$$\hat{\Delta}_{t(n_0,h)} - \hat{\Delta}_{n_0} \leq -h/4|f'(\theta_{n_0})|^2. \tag{32}$$

Now, recursively define

$$n_{k+1} = t(n_k, h).$$

An iterated application of (32) yields for some constant $C_1$,

$$\hat{\Delta}_{n_{k+1}} - \hat{\Delta}_{n_k} \leq -C_1 h\hat{\Delta}_{n_k}^{2/\mu}.$$

We then have two cases:

**Case $\mu = 2$:** For this case, we have,

$$\hat{\Delta}_{n_{k+1}} \leq (1 - C_1 h)\hat{\Delta}_{n_k}.$$

Recursively, we deduce that

$$\hat{\Delta}_{n_k} \leq \hat{\Delta}_{n_0}(1 - C_1 h)^k \leq \hat{\Delta}_{n_0} e^{-C_1 hk},$$

which implies that for any $k$,

$$\hat{\Delta}_{n_0} n_k^{\mu\tau} e^{-C_1 hk} \geq n_k^{\mu\tau}\hat{\Delta}_{n_k} \geq l\tilde{C}_0^{\mu}.$$

This, however, will yield $\tilde{C}_0 \leq 0$ when we take $k$ to $\infty$, which is a contradiction.

**Case $\mu < 2$:** For this case, it follows from

$$\hat{\Delta}_{n_k} - \hat{\Delta}_{n_{k+1}} \geq C_1 h \hat{\Delta}_{n_k}^{2/\mu}.$$

that

$$\int_{\hat{\Delta}_{n_{k+1}}}^{\hat{\Delta}_{n_k}} \frac{1}{u^{2/\mu}} du \geq \int_{\hat{\Delta}_{n_{k+1}}}^{\hat{\Delta}_{n_k}} \frac{1}{\hat{\Delta}_{n_k}^{2/\mu}} du = \frac{\hat{\Delta}_{n_k} - \hat{\Delta}_{n_{k+1}}}{\hat{\Delta}_{n_k}^{2/\mu}} \geq Ch,$$

which implies that for some positive constant $C_2$

$$\hat{\Delta}_{n_{k+1}}^{-2/\mu+1} - \hat{\Delta}_{n_k}^{-2/\mu+1} \geq C_2 h.$$

Recursively, we deduce that

$$\hat{\Delta}_{n_k}^{-2/\mu+1} \geq \hat{\Delta}_{n_0}^{-2/\mu+1} + C_2 hk,$$

and furthermore

$$n_k^{\hat{\tau}(-2+\mu)/\mu} \hat{\Delta}_{n_k}^{(-2+\mu)/\mu} \geq n_k^{\hat{\tau}(-2+\mu)/\mu} \hat{\Delta}_{n_0}^{-2/\mu+1} + C_2 n_k^{\hat{\tau}(-2+\mu)/\mu} kh.$$

It then follows from (31) and (19) that

$$\tilde{C}_0^{-2+\mu} l^{(-2+\mu)/\mu} \geq n_k^{\hat{\tau}(-2+\mu)/\mu} \hat{\Delta}_{n_k}^{(-2+\mu)/\mu} \geq O(k n_k^{\hat{\tau}(-2+\mu)/\mu}) \geq O(k^{\hat{\tau}(-2+\mu)/(-a+1)\mu+1}).$$

Now, one verifies that this gives us an contradiction if we take $k, l$ to $\infty$, as long as

$$\hat{\tau} \leq \mu(1-a)/(2-\mu), \text{ equivalently } \hat{\tau}(-2+\mu)/(-a+1)\mu + 1 \geq 0.$$

□

**Lemma 8.4.** *There exist an integer $l$ such that for all $n$ sufficiently large,*

$$n^{\hat{\tau}} \Delta_n \leq l \tilde{C}_0^2.$$

*Proof.* By way of contradiction, suppose otherwise. Then, by Lemma 8.3, for any $l$ and arbitrarily large $N$, we can find $k_0 > m_0 > N$ such that

$$m_0^{\hat{\tau}} \Delta_{m_0} \leq 2l\tilde{C}_0^2, \quad k_0^{\hat{\tau}} \Delta_{k_0} \geq 3l\tilde{C}_0^2,$$

$$\min_{m_0 < n \leq k_0} n^{\hat{\tau}} \Delta_n > 2l\tilde{C}_0^2, \quad \max_{m_0 \leq n < k_0} n^{\hat{\tau}} \Delta_n \leq 3l\tilde{C}_0^2. \tag{33}$$

For some $0 < h < 1$, let $m_1 = t(m_0, h)$. For any $m_0 \leq n \leq m_1$, as in the proof of Theorem 8.2, we derive

$$\hat{\Delta}_n - \hat{\Delta}_{m_0} \leq -h/4 |f'(\theta_{m_0})|^2, \tag{34}$$

which, together with Theorem 8.2 and (33), implies that

$$f'(\theta_{m_0})^2 \leq 4/h\tilde{C}_0^2 O(m_0^{-\hat{\tau}}) + \tilde{C}_0^2 O(m_0^{-\mu\tau}),$$

which, together with (21), further implies that for some $C > 0$,

$$|\hat{\Delta}_n - \hat{\Delta}_{m_0}| \le Ch|f'(\theta_{m_0})|^2 + C\tilde{C}_0^2 m_0^{-2\tau} \le \tilde{C}_0^2 O(m_0^{-\hat{\tau}}) + \tilde{C}_0^2 O(m_0^{-\mu\tau}) + C\tilde{C}_0^2 m_0^{-2\tau}.$$

It then follows that for sufficiently small $h$

$$\begin{aligned} |n^{\hat{\tau}}\hat{\Delta}_n - m_0^{\hat{\tau}}\hat{\Delta}_{m_0}| &\le n^{\hat{\tau}}|\hat{\Delta}_n - \hat{\Delta}_{m_0}| + (n^{\hat{\tau}} - m_0^{\hat{\tau}})\hat{\Delta}_{m_0} \\ &= O(m_0^{\hat{\tau}})(\hat{\Delta}_n - \hat{\Delta}_{m_0}) + o(m_0^{\hat{\tau}})\hat{\Delta}_{m_0} \\ &\le l\tilde{C}_0^2, \end{aligned}$$

where we have used the fact that

$$n^{\hat{\tau}} = O(m_0^{\hat{\tau}}), \quad n^{\hat{\tau}} - m_0^{\hat{\tau}} = o(m_0^{\hat{\tau}}).$$

In particular, we have

$$|(m_0+1)^{\hat{\tau}}\hat{\Delta}_{m_0+1} - m_0^{\hat{\tau}}\hat{\Delta}_{m_0}| \le l\tilde{C}_0 \text{ and } |m_1^{\hat{\tau}}\hat{\Delta}_{m_1} - m_0^{\hat{\tau}}\hat{\Delta}_{m_0}| \le l\tilde{C}_0^2,$$

which further implies that

$$m_0^{\hat{\tau}}\hat{\Delta}_{m_0} \ge l\tilde{C}_0^2 \text{ and } m_1 < k_0,$$

respectively.

Setting $n = m_1$ and rewriting (34), we have for some constant $C_1$,

$$\hat{\Delta}_{m_1} - \hat{\Delta}_{m_0} \le -C_1 h \hat{\Delta}_{m_0}^{2/\mu}.$$

We then consider two cases:

**Case $\mu = 2$:** For this case, we have for some positive constant $C_1$,

$$\hat{\Delta}_{m_1} \le (1 - C_1 h)\hat{\Delta}_{m_0}.$$

Then for $m_0$ large enough,

$$m_1^{\hat{\tau}}\hat{\Delta}_{m_1} \le (1 - C_1 h) m_1^{\hat{\tau}}\hat{\Delta}_{m_0} = (1 - C_1 h) m_0^{\hat{\tau}}(1 + o(1))\hat{\Delta}_{m_0} \le 2l\tilde{C}_0^2,$$

which yields a contradiction.

**Case $\mu < 2$:** For this case, as in the proof of Lemma 8.3, we have for some positive constant $C_2$,

$$\hat{\Delta}_{m_1}^{-2/\mu+1} \ge \hat{\Delta}_{m_0}^{-2/\mu+1} + C_2 h.$$

It then follows from (33) and (19) that for $l$ sufficiently large

$$\hat{\Delta}_{m_1}^{(-2+\mu)/\mu} \ge (2l\tilde{C}_0^2)^{(-2+\mu)/\mu} m_0^{-a+1} + C_2 h \ge (2l\tilde{C}_0^2)^{(-2+\mu)/\mu} m_1^{-a+1},$$

which implies that

$$m_1^{\hat{\tau}}\hat{\Delta}_{m_1} \le 2l\tilde{C}_0^2,$$

a contradiction.

□

The following theorem characterizes the rate of convergence of $\{f(\theta_n)\}$.

**Theorem 8.5.** *With probability* 1*, we have*

$$|\hat{\Delta}_n| = \tilde{O}(n^{-\hat{\tau}}).$$

*Proof.* It immediately follows from Lemmas 8.3 and 8.4. □

In the rest of this section, assuming

(8.b) $\mu\tau \geq (1-a)$,

we prove $\{\theta_n\}$ converges almost surely. Here, let us note that (8.b) can always be satisfied if $a, b, \beta$ are appropriately chosen such that $\tau$ is sufficiently large.

The following theorem characterizes the rate of convergence of $\{\theta_n\}$.

**Theorem 8.6.** *Assume that (8.b). Then, we have*

$$\sup_{k \geq n} |\theta_k - \theta_n| = \tilde{O}(n^{-(\hat{\tau}-(1-a)/2)}).$$

*Proof.* In this proof, we set

$$\tau' = (\hat{\tau} + (1-a))/2.$$

For some $0 < h < 1$, starting from a fixed $n_0$, recursively define

$$n_{k+1} = t(n_k, h).$$

Then, to prove the theorem, it suffices to prove that

$$\sup_{k \geq m} |\theta_{n_k} - \theta_{n_m}| = \tilde{O}(n_m^{-(\hat{\tau}+(1-a)/2)}). \tag{35}$$

Now, applying Lemma 6.2 (7), we deduce that for some $C > 0$

$$|\theta_{n_{i+1}} - \theta_{n_i}| \leq C n_i^{\tau'} (f(\theta_{n_{i+1}}) - f(\theta_{n_i})) + C\tilde{C}_0^2 n_i^{-\tau'}.$$

It then follows that for any $m \leq k$,

$$\begin{aligned}
|\theta_{n_k} - \theta_{n_m}| &\leq \sum_{i=m}^{k-1} |\theta_{n_{i+1}} - \theta_{n_i}| \\
&\leq C\tilde{C}_0^2 \sum_{i=m}^{k-1} n_i^{-\tau'} + C \sum_{i=m}^{k-1} (u(\theta_{n_i}) - u(\theta_{n_{i+1}})) n_i^{\tau'} \\
&\leq C\tilde{C}_0^2 \sum_{i=m}^{k-1} n_i^{-\tau'} + C \sum_{i=m+1}^{k} (n_i^{\tau'} - n_{i-1}^{\tau'}) |u(\theta_{n_i})| + C n_m^{\tau'} |u(\theta_{n_m})| + C n_k^{\tau'} |u(\theta_{n_k})|.
\end{aligned}$$

Applying (19), we deduce that

$$\sum_{i=m}^{k-1} n_i^{-\tau'} = \sum_{i=m}^{k-1} O(i^{-\tau'/(1-a)}) = O(m^{-\tau'/(1-a)+1}),$$

$$\sum_{i=m+1}^{k} (n_i^{\tau'} - n_{i-1}^{\tau'})|u(\theta_{n_i})| = \sum_{i=m}^{k} O((i-1)^{y/(1-a)-1} i^{-\hat{\tau}/(1-a)}) = O(m^{y/(1-a)-\hat{\tau}/(1-a)}),$$

$$n_m^{\tau'}|u(\theta_{n_m})| = O(m^{\tau'/(1-a)} m^{-\hat{\tau}/(1-a)}) = O(m^{(\tau'-\hat{\tau})/(1-a)}),$$

$$n_k^{\tau'}|u(\theta_{n_k})| = O(k^{\tau'/(1-a)} k^{-\hat{\tau}/(1-a)}) = O(k^{(\tau'-\hat{\tau})/(1-a)}).$$

We then immediately conclude that

$$|\theta_{n_k} - \theta_{n_m}| = O(n_m^{(\tau'-\hat{\tau})}),$$

which immediately implies (35).

□

The following "variable" version of the Lojasiewicz inequality will be used to refine the rates of convergence of $\{\theta_n\}$ and $\{f(\theta_n)\}$.

**Lemma 8.7.** *For each $\theta \in \Theta$, there exist real numbers $\delta_\theta \in (0,1)$, $\mu_\theta \in (1,2]$, $M_\theta \in [1,\infty)$ such that*

$$|f(\theta') - f(\theta)| \le M_\theta \|f'(\theta')\|^{\mu_\theta}$$

*for all $\theta' \in \Theta$ satisfying $\|\theta' - \theta\| \le \delta_\theta$.*

Theorem 8.6 implies that with probability 1, $\{\theta_n\}$ converges. From now on, let $\hat{\theta} = \lim_{n\to\infty} \theta_n$ and set $\mu = \mu_{\hat{\theta}}$. Then, with this redefined $\mu$, going through exactly the same arguments as in the proof of Theorems 8.5 and 8.6, we have the following two theorems.

**Theorem 8.8.** *For the above redefined $\mu$, Theorems 8.5 holds.*

**Theorem 8.9.** *For the above redefined $\mu$, assume (8.b). Then, we have*

$$|\theta_n - \hat{\theta}| = \tilde{O}(n^{-(\hat{\tau}-(1-a)/2)}).$$

# 9 Capacity Achieving Distribution of a Special Class of Channels

In this section, we restrict our attention to a special class of input-restricted finite-state channels with certain parameterization and we prove that for such channels operated at high SNR regime, the capacity will only be achieved at the interior of the parameter space and our algorithm converges almost surely.

More specifically, recalling that $X, Y$ denote the input, output processes of the channel over finite alphabets $\mathcal{X}$, $\mathcal{Y}$, respectively, we consider a class of parameterized memoryless channels such that

(9.a) the channel only has one state; in other words, at any time slot, the channel is characterized by the conditional probability $p(y|x)$.

(9.b) for some mixing finite-type constraint $F \subset \mathcal{X}^2$, $X \in \Pi_F$.

(9.c) the channel is parameterized by $\varepsilon \geq 0$ such that for each $x$ and $y$, $p(y|x)(\varepsilon)$ is an analytic function of $\varepsilon \geq 0$, which is not identically 0.

(9.d) there is a one-to-one (not necessarily onto) mapping $\Phi : \mathcal{X} \to \mathcal{Y}$, such that for any $x \in \mathcal{X}$, $p(\Phi(x)|x)(0) = 1$.

(9.e) $X$ is parameterized as in [53], that is,

$$\theta = (p(X_1 = w_1, X_2 = w_2) : (w_1, w_2) \notin F).$$

Under the above assumptions, $\varepsilon$ can be regarded as a parameter that quantifies noise, and $\Phi(x)$ is the noiseless output corresponding to input $x$. The regime of "small $\varepsilon$" corresponds to high SNR. Note that the output process $Y = Y(X, \varepsilon)$ depends on the input process $X$ and the parameter value $\varepsilon$; we will often suppress the notational dependence on $\varepsilon$ or $X$, when it is clear from the context. Prominent examples of such families include input-restricted versions of the binary symmetric channel with crossover probability $\varepsilon$, denoted by BSC($\varepsilon$),, and the binary erasure channel with erasure rate $\varepsilon$, denoted by BEC($\varepsilon$).

**General SNR regime.** By using an asymptotic formula of $I(X;Y)$, we show that for the above-mentioned channels, the capacity achieving $X$ must be primitive.

Assume that $X$ has period $e$ with period classes $D_1, D_2, \ldots, D_e$. Then, by the classical Perron-Frobenius theory, after necessary reindexing, its transition probability matrix $\Pi$ can be written as

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_{e-1} \\ D_e \end{array} \begin{array}{c} \begin{array}{ccccc} D_1 & D_2 & D_3 & \cdots & D_e \end{array} \\ \begin{pmatrix} 0 & B_1 & 0 & \cdots & 0 \\ 0 & 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & B_{e-1} \\ B_e & 0 & 0 & \cdots & 0 \end{pmatrix} \end{array}, \tag{36}$$

where we used the period classes to index the sub-blocks. In the following, let $\mathbf{B}$ denote the set of all entry indices of $\Pi$ corresponding to some $B_k$, that is,

$$\mathbf{B} = \{(i,j) : i \in D_k, j \in D_{k+1}, \text{ for } k = 1, \cdots, e-1\} \cup \{(i,j) : i \in D_e, j \in D_1\}.$$

Now, consider an analytic perturbation $\Pi(\delta)$ of $\Pi$, $\delta \geq 0$, where

(9.f) $\Pi(0) = \Pi$;

(9.g) for some $(i,j) \in \mathbf{B}$, $\Pi_{ij}(\delta)$ is not identically 0;

(9.h) for any $\delta \geq 0$, $\Pi(\delta)$ is still a stochastic matrix.

In other words, some non-$B$-entries in $\Pi$ are analytically perturbed; as a result, $Y$ is perturbed from $Y(0)$ to $Y(\delta)$. The following theorem describes the asymptotic behavior of $H(Y)$ under such a perturbation.

**Theorem 9.1.** *Under the aboved-mentioned perturbation as in (9.f)-(9.f), there exist* $C_1, C_2 > 0$ *such that*

$$C_1\delta \log 1/\delta \le H(Y(\delta)) - H(Y(0)) \le C_2\delta^{1/2}.$$

*Proof.* The proof is postponed to Appendix A. □

**Remark 9.2.** It follows from Condition (9.a) that $H(Y|X)$ is linear with respect to $\vec{p}$. Theorem 9.1, together with this fact, implies that there exist $C_1, C_2$ such that

$$C_1\delta \log 1/\delta \le I(X(\delta); Y(\delta)) - I(X(0); Y(0)) \le C_2\delta^{1/2},$$

which implies that, for any irreducible but not primitive $X$, any perturbation of $X$ as in (9.f)-(9.h) will strictly increase the mutual information. So, we conclude that the capacity achieving $X$ must be primitive, and thus Condition (2.a) holds.

**High SNR regime.** At the high SNR regime, that is, when $\varepsilon$ is close to 0, it has been established in [24] that there exists $\hat{\epsilon} > 0$ such that

(9.i) $I(X;Y)$, when restricted on $X \in \Pi_{F,\hat{\epsilon}}$, is strictly concave with respect to $\theta \in \Theta$.

(9.j) the capacity of the channel can be uniquely achieved within $\Pi_{F,\hat{\epsilon}}$.

As a consequence, we have the following theorem.

**Theorem 9.3.** *For the channel as in (9.a)-(9.d) operating at the high SNR regime and sufficiently small* $\hat{\epsilon}$*, under the iteration in (1),* $\{\theta_n\}$ *converges to the capacity achieving distribution with probability* 1*.*

*Proof.* Note that Condition (9.a) and Theorem 5.2 imply Conditions (7.a) and (7.b); and Condition(9.b) implies that the global maximum $\theta^*$ indeed corresponds to the capacity achieving distribution. The theorem then immediately follows. □

**Example 9.4.** Consider a binary symmetric channel with crossover probability $\varepsilon > 0$. Let $X$ be a binary input Markov chain with the transition probability matrix

$$\begin{bmatrix} 1-\pi & \pi \\ 1 & 0 \end{bmatrix}, \tag{37}$$

where $0 \le \pi \le 1$. Apparently, $X$ is supported on the so-called $(1,\infty)$-RLL constraint [34], which simply means that the string "11" is forbidden. Let $Y$ denote the corresponding output process. Assume that $X$ is parameterized by $\theta = (p(00), p(01), p(10))$, where $p(10) = 1$ is in fact a constant. It can be checked that Conditions (9.a)-(9.d) are all satisfied, so when $\varepsilon$ is sufficiently small, Conditions (9.i)-(9.j) are satisfied and thus Theorem 9.3 holds.

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

$$H(Y) = H(X) + \frac{\pi(2-\pi)}{1+\pi}\varepsilon \log(1/\varepsilon) + O(\varepsilon), \tag{38}$$

where the $O(\varepsilon)$-term is analytic with respect to $p$ (see Theorem 2.18 of [24]). It then follows that

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

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 $\theta$, which implies the strict convexity (rather than concavity) of $H(X|Y)$ when $\varepsilon$ is small enough. So, the concavity conjecture in [53] is not true in general, and thus the conditions guaranteeing the convergence of the GBAA are not satisfied.

# Appendices

# A Proof of Theorem 9.1

First of all, we define

$$Z(\delta) = Z(X_1^n(\delta)) = \begin{cases} 0 & (X_i(\delta), X_{i+1}(\delta)) \in \mathbf{B} \text{ for all } i \in \{1, \cdots, n-1\} \\ 1 & (X_i(\delta), X_{i+1}(\delta)) \notin \mathbf{B} \text{ for exactly one } i \in \{1, \cdots, n-1\} \\ 2 & (X_i(\delta), X_{i+1}(\delta)) \notin \mathbf{B} \text{ for more than one } i \in \{1, \cdots, n-1\} \end{cases}.$$

Next, applying the Birch bound [10], we derive the following key inequality for this proof:

$$\frac{H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta))}{n-m} \le H(Y) \le \frac{H(Y_1^n|X_0(\delta), Z(\delta))}{n} + \frac{H(Z(\delta))}{n} + \frac{H(X_0(\delta))}{n}, \quad (39)$$

for any $m \le n$.

**The lower bound part.** We first prove that there exists $C_1 > 0$ such that

$$H(Y(\delta)) \ge H(Y(0)) + C_1 \delta \log 1/\delta,$$

which immediately implies the lower bound part of the theorem. In this part, we set

$$n = \sqrt{\log \delta} \text{ and } m = n/2. \quad (40)$$

By definition, we have

$$\begin{aligned} H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta))/(n-m) &= \sum_{x_0} p^\delta(x_0, Z=0) H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta)=0)/(n-m) \\ &+ \sum_{x_0} p^\delta(x_0, Z=1) H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta)=1)/(n-m) \\ &+ \sum_{x_0} p^\delta(x_0, Z=2) H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta)=2)/(n-m). \\ &\triangleq T_1 + T_2 + T_3 \end{aligned}$$

where $p^\delta(x_0, Z=0)$ means $P(X_0(\delta) = x_0, Z(\delta) = 0)$.

We next give estimates for the each of three terms defined as above.

For $T_3$, notice that $n\delta < 1$ for sufficiently small $\delta$ and then

$$\sum_{x_0} p^\delta(x_0, Z=2) \le n^2 (C_0\delta)^2 + n^3 (C_0\delta)^3 + \cdots \le \frac{C_0^2}{1 - nC_0\delta} n^2 \delta^2,$$

for some $C_0 > 0$. It then follows that

$$\begin{aligned} T_3 &= \sum_{x_0} p^\delta(x_0, Z=2) H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta)=2)/(n-m) \\ &\le \sum_{x_0} p^\delta(x_0, Z=2) H(Y_{m+1}^n(\delta))/(n-m) \\ &\le \sum_{x_0} p^\delta(x_0, Z=2) \log|\mathcal{Y}| \\ &= O(n^2\delta^2). \end{aligned} \tag{41}$$

For $T_2$, one verifies that for any $x_0$, there exist constants $C_1, C_2 > 0$, $0 < \lambda_1 < \lambda_2 < 1$ such that

$$C_1 n\delta\lambda_1^n \le p^\delta(y_1^n|x_0, Z=1) \le C_2 n\delta\lambda_2^n.$$

Similarly, for any $x_0$, there exist $C_3, C_4 > 0$, and the same $0 < \lambda_1 < \lambda_2 < 1$ as above such that

$$C_3 m\lambda_1^m \le p^\delta(y_1^m|x_0, Z=1) \le C_4 m\lambda_2^m.$$

It then follows that for any $x_0$,

$$C_5\delta\lambda_2^n/\lambda_1^m \le p^\delta(y_{m+1}^n|y_1^m, X_0, Z=1) \le C_6\delta\lambda_2^n/\lambda_1^m,$$

which, together with (40), implies that

$$H(Y_{m+1}^n(\delta)|Y_1^m(\delta), X_0(\delta), Z(\delta)=1) = \hat{O}(\log 1/\delta) + O(n\log\lambda_2) + O(m\log\lambda_1).$$

This, together with the fact

$$p(x_0, Z=1) = \hat{O}(n\delta),$$

implies that

$$T_2 = \hat{O}(\delta\log 1/\delta) + O(n\delta\log\lambda_2) + O(m\delta\log\lambda_1). \tag{42}$$

For $T_1$, notice that it can be rewritten as

$$T_1 = \sum p^\delta(y_1^n, x_0, Z=0) \log p^\delta(y_{m+1}^n|y_1^m, x_0, Z=0)/(n-m).$$

One then verifies that

$$\left|p^\delta(y_1^n, x_0, Z=0) - p^0(y_1^n, x_0, Z=0)|_{\delta=0}\right| = O(n\delta)p^0(y_1^n, x_0, Z=0),$$

which implies that

$$\left|\frac{\sum p^0(y_1^n, x_0, Z=0)\log p^\delta(y_{m+1}^n|y_1^m, x_0, Z=0)}{n-m} - \frac{\sum p^\delta(y_1^n, x_0, Z=0)\log p^\delta(y_{m+1}^n|y_1^m, x_0, Z=0)}{n-m}\right| = O(n\delta).$$

When fixing $x_0$ and assuming $Z=0$, the analyticity argument in [20] can be used to prove that

$$\sum p^0(y_1^n, x_0, Z=0)\log p^\delta(y_{m+1}^n|y_1^m, x_0, Z=0)/(n-m)$$

exponentially converges to an analytic function of $\delta$. It then follows that for some $0 < \rho < 1$

$$T_1 = H(Y(0)) + O(\rho^m) + O(\delta). \tag{43}$$

Combining (41), (42) and (43), we then have

$$H(Y_{m+1}^n(\delta)|Y_1^m, X_0(\delta), Z(\delta))/(n-m) = H(Y(0)) + \hat{O}(\delta \log 1/\delta).$$

**The upper bound part.** We then prove that there exists $C_2 > 0$,

$$H(Y(\delta)) \leq H(Y(0)) + C_2\delta^{1/2},$$

which immediately implies the upper bound part of the theorem. For this part, setting

$$n = \delta^{-1/2} \text{ and } m = 0. \tag{44}$$

Using a parallel argument as in the lower bound part, we can still derive (41), (42) and (43) and then

$$H(Y_1^n|X_0(\delta), Z(\delta))/n = H(Y)_{\delta=0,0} + O(\delta^{1/2}).$$

It can verified that

$$p(Z(\delta) = 1) = O(n\delta), \quad p(Z(\delta) = 2) = O(n^2\delta^2),$$

which, together with the straightforward observation that $H(X_0(\delta))/n = O\left(\frac{1}{n}\right)$, implies that

$$H(Z(\delta)) = -\sum_{i=0}^{2} p(Z(\delta) = i) \log p(Z(\delta) = i) = O(n\delta \log \delta) + O(n\delta \log n),$$

and consequently

$$\frac{H(Z(\delta))}{n} = O(\delta \log \delta) + O(\delta \log n).$$

The upper bound part then follows from all the above estimates and (39).

# References


[1] A. Allahverdyan. Entropy of hidden Markov processes via cycle expansion. *J. Stat. Phys.*, vol. 133, pp. 535–564, 2008.

[2] S. Arimoto. An algorithm for computing the capacity of arbitrary memoryless channels. *IEEE Trans. Info. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[3] L. Arnold, V. M. Gundlach and L. Demetrius. Evolutionary formalism for products of positive random matrices. *Annals of Applied Probability*, vol. 4, pp. 859–901, 1994.

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

[5] 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.

[6] L. R. Bahl, J. Cocke, F. Jelinek and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Info. Theory*, vol. 20, no. 2, pp. 284-287, 1974.

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

[8] 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.

[9] E. Bierstone and P. Milman. *Semianalytic and Subanalytic Sets, Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, vol. 56, no, 1, pp. 5-42, 1988.

[10] J. J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Statist.*, vol. 33, pp. 930–938, 1962.

[11] D. Blackwell. The entropy of functions of finite-state Markov chains. *Trans. First Prague Conf. Information Thoery, Statistical Decision Functions, Random Processes*, pp. 13–20, 1957.

[12] R. E. Blahut. Computation of channel capacity and rate distortion functions. *IEEE Trans. Info. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[13] 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.

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

[15] H. Chen. *Stochastic approximation and its applications*. Kluwer Academic Publishers, 2002.

[16] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen and H. Hollmann. On the entropy rate of a hidden Markov model. *IEEE ISIT*, pp. 12, 2004.

[17] 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.

[18] 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.

[19] G. Han. Limit theorems in hidden Markov models. *IEEE Trans. Info. Theory*, vol. 59, no. 3, pp. 1311-1328, 2013.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] G. Han and B. Marcus. Analyticity of entropy rate of continuous-state hidden Markov chains. Submitted to *Stochastic Processes and Their Applications*.

[26] N. Haydn. The central limit theorem for uniformly strong mixing measures. arXiv:0903.1325, 2009.

[27] N. Haydn and S. Vaienti. Fluctuations of the metric entropy for mixing measures. *Stochastics and Dynamics*, vol. 4, pp. 595–627, 2004.

[28] 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.

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

[30] I. Kontoyiannis. Asymptotic recurrence and waiting times for stationary processes. *J. Theor. Prob.*, vol. 11, pp. 795–811, 1998.

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

[32] Y. Li and G. Han. On the concavity conjecture for the generalized Blahut-Arimoto algorithm. Preprint.

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

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

[35] J. Luo and D. Guo. On the entropy rate of hidden Markov processes observed through arbitrary memoryless channels. *IEEE Trans. Info. Theory*, vol. 55, pp. 1460–1467, 2009.

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

[37] 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.

[38] 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.

[39] E. Ordentlich and T. Weissman. On the optimality of symbol by symbol filtering and denoising. *IEEE Trans. Info. Theory*, vol. 52, no. 1, pp. 19–40, 2006.

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

[41] 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.

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

[43] 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.

[44] H. D. Pfister. The capacity of finite-state channels in the high-noise regime. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems, London Mathematical Society Lecture Note Series*, vol. 385, pp. 179–222, 2011.

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

[46] 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.

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

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

[49] 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.

[50] H. Thapar and A. Patel. A class of partial response systems for increasing storage density in magnetic recording. *IEEE Trans. Magn.*, vol. 23, no. 5, pp. 3666–3668, 1987.

[51] P. O. Vontobel. Connecting the Bethe entropy and the edge zeta function of a cycle code. *IEEE ISIT*, 2010.

[52] P. O. Vontobel. The Bethe Permanent of a Non-Negative Matrix. To appear in *IEEE Trans. Info. Theory*.

[53] 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.

[54] 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.

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

[56] 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.

[57] O. Zuk, E. Domany, I. Kanter and M. Aizenman. Taylor series expansions for the entropy rate of hidden Markov Processes. *IEEE ICC*, 2006.