On the Entropy Rate of Hidden Markov Processes Observed Through Arbitrary Memoryless Channels

Jun Luo, *Student Member, IEEE*, and Dongning Guo, *Member, IEEE*

***Abstract*—This paper studies the entropy rate of hidden Markov processes (HMPs) which are generated by observing a discrete-time binary homogeneous Markov chain through an arbitrary memoryless channel. A fixed-point functional equation is derived for the stationary distribution of an input symbol conditioned on all past observations. While the existence of a solution to the fixed-point functional equation is guaranteed by martingale theory, its uniqueness follows from the fact that the solution is the fixed point of a contraction mapping. The entropy or differential entropy rate of the HMP can then be obtained through computing the average entropy of each input symbol conditioned on past observations. In absence of an analytical solution to the fixed-point functional equation, a numerical method is proposed in which the fixed-point functional equation is first converted to a discrete linear system using uniform quantization and then solved efficiently. The accuracy of the computed entropy rate is shown to be proportional to the quantization interval. Unlike many other numerical methods, this numerical solution is not based on averaging over a sample path of the HMP.**

***Index Terms*—Blackwell's measure, contraction mapping, entropy rate, filtering, fixed-point functional equation, hidden Markov process.**

I. Introduction

Let $\{X_n\}$ be a binary homogeneous Markov chain with symmetric transition probability ϵ. Let $\{Y_n\}$ be the observation of $\{X_n\}$ through an arbitrary memoryless channel. Conditioned on X_k, the past, current and future observations, namely, $Y_{-\infty}^{k-1}$, Y_k and Y_{k+1}^{∞}, are independent. Without conditioning, however, the output $\{Y_n\}$ is not a Markov process. Such a process is called a hidden Markov process (HMP). The entropy (resp. differential entropy) rate of discrete (resp. continuous) HMPs is a classical open problem.

The entropy rate of HMPs has been studied since the 1950s. Blackwell expressed the entropy rate in terms of a complicated probability measure, which is the distribution of the conditional distribution of X_0 given the past observations $Y_{-\infty}^{0}$ [1]. Blackwell's work was followed by several authors who studied HMPs from the estimation-theoretic viewpoint. In 1965, Wonham [2] used stochastic differential equation to describe the evolution of the *posterior* probability distribution of the dynamical state given the output perturbed by Gaussian noise. Recently, Ordentlich and Weissman [3] presented a new approach for bounding the entropy rate of HMP by constructing an alternative Markov process corresponding to the log-likelihood ratio of estimating the positivity of current symbol X_0 based on the past observations $Y_{-\infty}^{0}$. Furthermore, Nair *et al.* [4] used the techniques in [3] to study the behavior of filtering error probability and obtained tight bounds for the entropy rate in the rare-transition regime, i.e., when ϵ is very small. An overview of statistical and information-theoretic aspects of HMPs is presented in [5].

In absence of an analytical solution due to the difficulty of Blackwell's measure, some other works use Monte Carlo simulation [6], sum-product method [7] and "prefixsets" method [8] to numerically compute the entropy rate by averaging over a long, randomly chosen sample path of the HMP. In addition, some deterministic computation methods based on quantized systems are suggested in [9] and [10] independently. In [9], density evolution is applied to a "forward variable" after quantization to obtain its stationary distribution. Reference [10] solves a linear system for the stationary distribution of the quantized Markov process to obtain a good approximation of the entropy rate.

This paper studies the entropy rate problem using filtering techniques and develops a new numerical method. A fixed-point functional equation is derived in Section II whose solution is the conditional cumulative distribution function (cdf) of the log-likelihood ratio of X_0 given the past observations $Y_{-\infty}^{-1}$ conditioned on $X_0 = +1$. Once the cdf is obtained, the entropy rate can be computed. While the existence of a solution to the fixed-point functional equation is guaranteed by martingale theory, Section III proves its uniqueness by showing that the solution corresponds to the fixed point of a contraction mapping. Since no explicit analytical solution to this equation (which corresponds to the cdf of Blackwell's measure) is known, numerical methods are developed in Section IV which give excellent approximations to the quantized cdf. In Section V, the accuracy of the numerically computed entropy rate is shown to be in the order of the quantization interval. Like the numerical methods in [9] and [10], the numerical method in this paper does not require a sample path of the HMP; rather, it is based on a direct computation of the filtering probability measure. While the numerical method provided in [10] quantizes the likelihood process, the numerical method in this paper quantizes the fixed-point functional equation.

II. Entropy Rate

Let $\{X_n\}$ be a stationary binary symmetric Markov chain with alphabet $\mathcal{X} = \{+1, -1\}$ and transition probability $\epsilon \in (0, 1/2)$. Let $\{Y_n\}$ be the observation of $\{X_n\}$ through a

Manuscript received ...; revised ... This work was presented in part at 42nd annual Conference on Information Science and Systems, Princeton, NJ, USA. This research was supported in part by NSF under Grant CCF-0644344 and by DARPA under Grant W911NF-07-1-0028.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA.

stationary memoryless channel characterized by two transition probability distributions $P_{Y|X}(\cdot|x)$, $x = \pm 1$, with alphabet $\mathcal{Y} \subset \mathbb{R}$.

Suppose $\mathcal{Y}$ is discrete. The entropy rate is related to the input-output mutual information of the memoryless channel by

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} H(Y_1^n) \\
&= \lim_{n\to\infty} \left(\frac{1}{n} H(Y_1^n|X_1^n) + \frac{1}{n} I\left(X_1^n; Y_1^n\right) \right) \qquad (1) \\
&= H(Y_1|X_1) + \lim_{n\to\infty} \left(\frac{1}{n} H(X_1^n) - \frac{1}{n} H(X_1^n|Y_1^n) \right) \\
&= H(Y_1|X_1) + H_2(\epsilon) - H(X_1|X_0, Y_1^\infty) \qquad (2)
\end{aligned}$$

where $H_2(\cdot)$ is the binary entropy function. Note that in case the alphabet $\mathcal{Y}$ is continuous and that $P_{Y|X}(\cdot| \pm 1)$ admit probability densities, we shall replace the entropies of Y by the corresponding differential entropies.

A. Entropy and Posterior Probabilities

One can treat $H(X_1|X_0, Y_1^\infty)$ in (2) as the expectation of the binary entropy of X_1 conditioned on X_0, Y_1^∞, i.e.,

$$H(X_1|X_0, Y_1^\infty) = \mathsf{E}\left\{ H_2(P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)) \right\} \qquad (3)$$

where the conditional probability $P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)$ is a random variable taking its value in $[0, 1]$, which is a function of X_0 and Y_1^∞. Consequently, in order to compute the entropy rate of the HMP, it suffices to obtain the *distribution of the conditional probability* $P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)$. We further note that for given x and y,

$$\begin{aligned}
&P_{X_1|X_0,Y_1,Y_2^\infty}(X_1|x, y, Y_2^\infty) \\
&= \frac{P_{X_1|Y_2^\infty}(X_1|Y_2^\infty) P_{X_0|X_1}(x|X_1) P_{Y|X}(y|X_1)}{\sum_{x'\in\{\pm 1\}} P_{X_1|Y_2^\infty}(x'|Y_2^\infty) P_{X_0|X_1}(x|x') P_{Y|X}(y|x')}. \qquad (4)
\end{aligned}$$

Therefore, it is enough to find the distribution of $P_{X_1|Y_2^\infty}(+1|Y_2^\infty)$.

The random probability $P_{X_i|Y_{i+1}^\infty}(+1|Y_{i+1}^\infty)$ represents a limit. The distribution of $P_{X_i|Y_{i+1}^\infty}(+1|Y_{i+1}^\infty)$ for every $i = 0, \pm 1, \ldots$, is well-defined on the σ-algebra generated by Y_{i+1}^∞. In fact, these distributions are also identical, which is a direct consequence of the stationarity of the HMP and the fact that $P_{X_0|Y_1^n}(+1|Y_1^n) \to P_{X_0|Y_1^\infty}(+1|Y_1^\infty)$ with probability 1 as $n \to \infty$. The latter convergence can be established by considering a Doob martingale $\{(Z_n, \mathcal{F}_n) : n = 1, 2, \ldots\}$ where $\mathcal{F}_n$ is the σ-algebra generated by Y_1^n and $Z_n = P_{X_0|Y_1^n}(+1|Y_1^n) = \mathsf{E}\left\{1_{\{X_0=+1\}}|Y_1^n\right\}$ which converges to $\mathsf{E}\left\{1_{\{X_0=+1\}}|Y_1^\infty\right\} = P_{X_0|Y_1^\infty}(+1|Y_1^\infty)$ with probability 1 by Doob's martingale convergence theorem [11, Theorem 13.3.7]. Furthermore, the above convergence also applies to the probability distribution of $P_{X_0|Y_1^n}(+1|Y_1^n)$ conditioned on $X_0 = +1$ because $P_{X_0|Y_1^n}(+1|Y_1^n)$ is a function of Y_1^n and the conditioning only changes the probability measure defined on the σ-algebra generated by Y_1^n.

By (3) and (4), the computation boils down to obtaining the distribution of the log-likelihood ratio

$$L_i = \log \frac{P_{X_{i-1}|Y_i^\infty}(+1|Y_i^\infty)}{P_{X_{i-1}|Y_i^\infty}(-1|Y_i^\infty)} \ . \qquad (5)$$

Note that L_i, $i = 0, \pm 1, \ldots$, are identically distributed. In the remainder of this section, we show that the distribution of L_i satisfies a fixed-point functional equation using stationarity and the fact that L_i is a function of L_{i+1} and Y_i.

B. Symmetric Channels

Let $\{Y_n\}$ be the observation of $\{X_n\}$ through a symmetric memoryless channel characterized by $P_{Y|X}(y|x) = P_{Y|X}(-y|-x)$. Let the cdf of L_i conditioned on $X_{i-1} = +1$ be denoted by F, i.e.,

$$F(l) = \Pr\left\{ L_i \le l | X_{i-1} = +1 \right\}.$$

Theorem 1: The cdf F satisfies the following fixed-point functional equation:

$$\begin{aligned}
F(q_\epsilon(x)) = \mathsf{E}\big\{ &(1-\epsilon) F\left(x - r(W)\right) \\
&+ \epsilon(1 - F\left(-x - r(W)\right)) \big\} \qquad (6)
\end{aligned}$$

for all $x \in \mathbb{R}$, where $W \sim P_{Y|X}(\cdot| + 1)$,

$$r(y) = \log \frac{P_{Y|X}(y| + 1)}{P_{Y|X}(y| - 1)} \qquad (7)$$

and

$$q_\epsilon(x) = \log \frac{\epsilon + (1-\epsilon)e^x}{\epsilon e^x + (1-\epsilon)} \ . \qquad (8)$$

Proof: The key to the proof is the following evolution, which follows from the Bayes' rule and definition (5)

$$\begin{aligned}
L_i &= \log \frac{P_{Y_i^\infty|X_{i-1}}(Y_i^\infty| + 1)}{P_{Y_i^\infty|X_{i-1}}(Y_i^\infty| - 1)} \\
&= \log \frac{(1-\epsilon) P_{Y_i^\infty|X_i}(Y_i^\infty| + 1) + \epsilon P_{Y_i^\infty|X_i}(Y_i^\infty| - 1)}{\epsilon P_{Y_i^\infty|X_i}(Y_i^\infty| + 1) + (1-\epsilon) P_{Y_i^\infty|X_i}(Y_i^\infty| - 1)} \\
&= \log \frac{e^{\alpha + r(Y_i) + L_{i+1}} + 1}{e^{r(Y_i) + L_{i+1}} + e^\alpha} \qquad (9)
\end{aligned}$$

where $\alpha = \log[(1-\epsilon)/\epsilon]$. The log-likelihood ratio L_i defined in (5) accepts a natural bound, i.e.,

$$|L_i| \le \alpha, \qquad (10)$$

which is because in terms of estimating X_{i-1}, providing Y_i^∞ is no better than providing X_i. Define

$$h_\epsilon(l) = \log \frac{(1-\epsilon)e^l - \epsilon}{(1-\epsilon) - \epsilon e^l}$$

which is a monotonically increasing function of $l \in (-\alpha, \alpha)$. Then inverting relationship (9) gives

$$L_{i+1} = h_\epsilon(L_i) - r(Y_i) \ .$$

Let $F_{U|V}(u|v)$ denote the cdf of random variable U conditioned on $V = v$, i.e., $F_{U|V}(u|v) = \Pr\{U \le u | V = v\}$. Clearly, by change of variable,

$$\begin{aligned}
F_{L_i|Y_i,X_i}(l|y, x) &= F_{L_{i+1}|Y_i,X_i}(h_\epsilon(l) - r(y)|y, x) \\
&= F_{L_{i+1}|X_i}(h_\epsilon(l) - r(y)|x),
\end{aligned}$$

and thus

$$F_{L_i|X_i}(l|x) = \int_{\mathcal{Y}} F_{L_i|Y_i,X_i}(l|y,x)\,\mathrm{d}F_{Y|X}(y|x) = \int_{\mathcal{Y}} F_{L_{i+1}|X_i}(h_\epsilon(l)-r(y)|x)\mathrm{d}F_{Y|X}(y|x)\,. \quad (11)$$

Also, because L_i is a function of Y_i^∞, one can get

$$F_{L_i|X_{i-1}}(l|x) = \sum_{x'=\pm x} F_{L_i|X_i,X_{i-1}}(l|x',x)P_{X_i|X_{i-1}}(x'|x) = (1-\epsilon)F_{L_i|X_i}(l|x) + \epsilon\, F_{L_i|X_i}(l|-x)\,. \quad (12)$$

Since the probability measure of L_i conditioned on X_{i-1} is stationary, one can define $F(l) \triangleq F_{L_i|X_{i-1}}(l|+1)$, $l \in \mathbb{R}$, which does not depend on i. Furthermore, note the following fact by symmetry,

$$F_{L_i|X_{i-1}}(l|x) = 1 - F_{L_i|X_{i-1}}(-l|-x)\,.$$

Substituting from (11) into (12) and letting $x=+1$ yields

$$F(l) = \mathsf{E}\big\{(1-\epsilon)F\left(h_\epsilon(l)-r(W)\right) + \epsilon\left(1-F\left(-h_\epsilon(l)-r(W)\right)\right)\big\}\,. \quad (13)$$

Note that the inverse of $h_\epsilon(\cdot)$ is $q_\epsilon(\cdot)$. Equation (13) becomes (6) by letting $x = h_\epsilon(l)$ and hence $l = q_\epsilon(x)$. □

C. Asymmetric Channels

Consider a channel characterized by $P_{Y|X}(\cdot|+1)$ and $P_{Y|X}(\cdot|-1)$ which are in general not symmetric. Let F_+ (resp. F_-) denote the cdf of L_i conditioned on $X_{i-1}=+1$ (resp. $X_{i-1}=-1$).

Theorem 2: The conditional cdfs F_+ and F_- satisfy

$$\begin{bmatrix} F_+(q_\epsilon(x)) \\ F_-(q_\epsilon(x)) \end{bmatrix} = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \begin{bmatrix} \mathsf{E}\left\{F_+(x-r(U))\right\} \\ \mathsf{E}\left\{F_-(x-r(V))\right\} \end{bmatrix} \quad (14)$$

for all $x \in \mathbb{R}$, where $q_\epsilon(x)$ is given in (8) and U, V are independent random variables with $U \sim P_{Y|X}(\cdot|+1)$ and $V \sim P_{Y|X}(\cdot|-1)$.

The proof is straightforward using the same technique developed in the proof of Theorem 1. Note that in [3] and [10], Ordentlich and Weissman studied the filtering process from a different perspective using an alternative Markov process. Formulas similar to (6) and (14) in the special case of discrete memoryless channels were also established.

D. Computation of Entropy Rate

Assuming that F, the conditional cdf of the log-likelihood ratio, is found, the entropy rate can be computed using (2), where the key is to compute the conditional entropy $H(X_1|X_0,Y_1^\infty)$. By (5) and (7), the conditional probability (4) can be rewritten using the log-likelihood ratio as

$$P_{X_1|X_0,Y_1^\infty}(+1|X_0,Y_1^\infty) = \left(1+\exp\left[-\alpha X_0 - r(Y_1) - L_2\right]\right)^{-1}\,.$$

Therefore, in view of (3), one can write

$$H(X_1|X_0,Y_1^\infty) = \mathsf{E}\left\{H_2\left(\left(1+\exp\left[-\alpha X_0 - r(Y_1) - L_2\right]\right)^{-1}\right)\right\}\,. \quad (15)$$

Also note that for given x and y and any subset A of $\mathbb{R}$,

$$\int_A P_{L_2|X_0,Y_1}(l|x,y)\mathrm{d}l = \int_A \frac{\sum_{x'\in\{\pm1\}} P_{L_2,X_0,X_1,Y_1}(l,x,x',y)}{P_{X_0,Y_1}(x,y)}\,\mathrm{d}l = \int_A \frac{\mathrm{d}F(l)}{1+\exp[-\alpha x - r(y)]} + \int_A \frac{-\mathrm{d}F(-l)}{1+\exp[\alpha x + r(y)]}\,. \quad (16)$$

Therefore, in order to compute the entropy, it suffices to solve the fixed-point functional equation (6) (or (14)).

An alternative method of computing the entropy rate is by first computing the input-output mutual information of HMPs using a fundamental information-estimation differential relationship due to Palomar and Verdú [12], and then using the decomposition (1). The key is still the computation of the same cdf, while this method has no particular advantage compared to the direct computation described in above.

III. Uniqueness of Solution to the Fixed-Point Functional Equation

Theorems 1 and 2 state that the (conditional) cdf of the log-likelihood ratio satisfies a fixed-point functional equation. An explicit solution to (6) or (14) is not available. An important question is that, is the solution to the fixed-point functional equation unique?

Proposition 1: Let $\mathcal{S}$ denote the set of cdfs whose corresponding probability measure has a support within the interval $\Omega = [-\alpha, \alpha]$. The fixed-point functional equation (6) admits no more than one solution in $\mathcal{S}$.

Proof: First, rewrite (6) as (13) with the variable l replaced by u. For any two cdfs F_1 and F_2 in $\mathcal{S}$, the L^1 distance $d(F_1,F_2)$ is given by the following

$$d(F_1,F_2) = \int_{\mathbb{R}} |F_1(u) - F_2(u)|\,\mathrm{d}u\,.$$

Define the operator Ψ on the set $\mathcal{S}$ as

$$(\Psi F)(u) = \begin{cases} 0, & \text{if } u < -\alpha; \\ \mathsf{E}\left\{(1-\epsilon)F\left(h_\epsilon(u)-r(W)\right) + \epsilon\left(1-F\left(-h_\epsilon(u)-r(W)\right)\right)\right\}, & \text{if } u \in \Omega; \\ 1, & \text{if } u > \alpha. \end{cases} \quad (17)$$

Note that the image ΨF is increasing, equal to 0 at $u < -\alpha$ and equal to 1 at $u > \alpha$. Thus Ψ is an injection on $\mathcal{S}$. For simplicity, let us denote $f_1(x) - f_2(x)$ as $f_{1,2}(x)$ where f_1 and f_2 are two functions. The key to the proof is the fact that Ψ is a contraction mapping under the L^1 distance. For any

two cdfs $F_1,\, F_2 \in \mathcal{S}$,

$$\begin{aligned} d(\Psi F_1, \Psi F_2) &= \int_\Omega \Big| \mathsf{E}\Big\{ (1-\epsilon) F_{1,2}(h_\epsilon(u) - r(W)) \\ &\qquad + \epsilon F_{2,1}(-h_\epsilon(u) - r(W)) \Big\} \Big| \mathrm{d}u \\ &\le \int_\Omega \mathsf{E}\Big\{ (1-\epsilon) \big| F_{1,2}(h_\epsilon(u) - r(W)) \big| \\ &\qquad + \epsilon \big| F_{2,1}(-h_\epsilon(u) - r(W)) \big| \Big\} \mathrm{d}u \qquad (18) \\ &= \mathsf{E}\Big\{ \int_\Omega \Big[(1-\epsilon) \big| F_{1,2}(h_\epsilon(u) - r(W)) \big| \\ &\qquad + \epsilon \big| F_{1,2}(-h_\epsilon(u) - r(W)) \big| \Big] \mathrm{d}u \Big\}, \qquad (19) \end{aligned}$$

where (18) follows from Jenson's inequality, and the order of integration and expectation is exchanged in (19) using Tonelli's theorem [13, p. 183]. Note that $q_\epsilon(\cdot)$ defined in (8) is the inverse of $h_\epsilon(\cdot)$. For the first term in the integrand of (19), one can obtain the following by change of variable,

$$\begin{aligned} \int_\Omega \big| F_{1,2}(h_\epsilon(u) - r(W)) \big| \mathrm{d}u &= \int_{\mathbb{R}} \big| F_{1,2}(t) \big| q_\epsilon'(t + r(W)) \mathrm{d}t \\ &\le (1-2\epsilon) \int_{\mathbb{R}} \big| F_{1,2}(t) \big| \mathrm{d}t \qquad (20) \end{aligned}$$

where the inequality is because $q_\epsilon'(t) \le 1 - 2\epsilon$ for all $t \in \mathbb{R}$. Similarly, one can upper bound the second term in (19), which, together with (20), leads to

$$d(\Psi F_1, \Psi F_2) \le (1-2\epsilon) d(F_1, F_2) \qquad (21)$$

with $0 < 1 - 2\epsilon < 1$. Therefore, Ψ is a contraction mapping.

Note that the solution to (6) is a fixed point of the operator Ψ. Suppose there exist two cdfs F_1^* and F_2^* in $\mathcal{S}$ which satisfy (6), by the contraction mapping property of Ψ, one can get the following inequality

$$d(F_1^*, F_2^*) = d(\Psi F_1^*, \Psi F_2^*) \le (1-2\epsilon) d(F_1^*, F_2^*),$$

which implies that $d(F_1^*, F_2^*) = 0$. Thus, F_1^* and F_2^* must be the same cdf. □

Note that the assertion in Proposition 1 also applies to the asymmetric case (14), which can be shown using the same contraction mapping argument.

IV. Numerical Methods

Since no explicit analytical solution to the fixed-point functional equation is known, a numerical method is developed in this section to compute it. Although this method is based on the result for symmetric channels, it can be extended to asymmetric channels without much modification. The numerical method represents a significant improvement over a previous method proposed in [14], the convergence of which was not established.

Noting that L_i accepts a natural bound (10), one can sample F arbitrarily finely to obtain a good approximation. After sampling, the fixed-point functional equation can be converted to a linear system which has at least one feasible solution. One may find that reference [10] also utilizes a linearized system method to approximate the stationary distribution of an alternative Markov process. The method in this paper differs from the one in [10] by discretizing the fixed-point functional equation while the one in [10] discretizes the transition kernel of the alternative Markov process.

Let M be the number of samples from the support Ω, which is defined in Section III. Thus, the step size Δ_M is given by

$$\Delta_M = \frac{2}{M} \log \frac{1-\epsilon}{\epsilon}.$$

We take two extra samples outside Ω and get the vector of sample points $\hat{x}$. Formally,

$$\hat{x}_i = -\log \frac{1-\epsilon}{\epsilon} - \frac{\Delta_M}{2} + i\Delta_M,\ i = 0, \ldots, M+1. \qquad (22)$$

Let $\hat{F}$ be the vector that consists of values of F evaluated at the points of $\hat{x}$, i.e., the ith element of $\hat{F}$ is obtained by $\hat{F}(i) = F(\hat{x}_i)$. One can define the following discrete representation $\hat{\Psi}$ based on Ψ in (17).

$$(\hat{\Psi}\hat{F})(i) \triangleq \begin{cases} 0, & \text{if } i = 0; \\ \mathsf{E}\Big\{ (1-\epsilon) \hat{F}\left(Q\left(h_\epsilon(\hat{x}_i) - r(W)\right)\right) & \\ \quad + \epsilon \left(1 - \hat{F}\left(Q\left(-h_\epsilon(\hat{x}_i) - r(W)\right)\right)\right) \Big\}, & \text{if } 1 \le i \le M; \\ 1, & \text{if } i = M+1, \end{cases} \qquad (23)$$

where $Q(\cdot)$ is a uniform quantizer given by the following,

$$Q(x) = \begin{cases} 0, & \text{if } x < -\log \frac{1-\epsilon}{\epsilon}; \\ i, & \text{if } \hat{x}_i - \frac{\Delta_M}{2} \le x < \hat{x}_i + \frac{\Delta_M}{2},\ 1 \le i \le M; \\ M+1, & \text{if } x \ge \log \frac{1-\epsilon}{\epsilon}. \end{cases}$$

Clearly, $\hat{\Psi}$ maps an $(M+2)$ vector $\hat{F}$ to another $(M+2)$ vector $\hat{\Psi}\hat{F}$. If fact, equation (23) is equivalent to a linear system to be given in the following context.

Let $\| \cdot \|_2$ denote the L^2 norm on the space $\mathbb{R}^{M+2}$. Let $\hat{\mathcal{S}}$ denote the set of all feasible vectors of $\hat{F}$, i.e., the set of vectors with value 0 for the first element and 1 for the last element and intermediate element values increasing with the index. It is easy to check that $\hat{\mathcal{S}}$ is a compact convex subset of $\mathbb{R}^{M+2}$. The key to the numerical method is that $\hat{\Psi}$ is a nonexpansive mapping on $\hat{\mathcal{S}}$, which also implies the continuity of $\hat{\Psi}$.

For any two points $\hat{F}_1$ and $\hat{F}_2$ in $\hat{\mathcal{S}}$, we adopt $\hat{F}_{1,2}(i)$ as the abbreviation of $\hat{F}_1(i) - \hat{F}_2(i)$, it is true that

$$\begin{aligned} \|\hat{\Psi}\hat{F}_1 - \hat{\Psi}\hat{F}_2\|_2^2 &= \sum_{i=1}^{M} \Big| \mathsf{E}\Big\{ (1-\epsilon) \hat{F}_{1,2}\left(Q\left(h_\epsilon(\hat{x}_i) - r(W)\right)\right) \\ &\qquad + \epsilon \hat{F}_{2,1}\left(Q\left(-h_\epsilon(\hat{x}_i) - r(W)\right)\right) \Big\} \Big|^2 \\ &\le \sum_{i=1}^{M} \mathsf{E}\Big\{ (1-\epsilon) \Big| \hat{F}_{1,2}\left(Q\left(h_\epsilon(\hat{x}_i) - r(W)\right)\right) \Big|^2 \\ &\qquad + \epsilon \Big| \hat{F}_{2,1}\left(Q\left(-h_\epsilon(\hat{x}_i) - r(W)\right)\right) \Big|^2 \Big\} \qquad (24) \\ &= \mathsf{E}\Big\{ \sum_{i=1}^{M} \Big[(1-\epsilon) \Big| \hat{F}_{1,2}\left(Q\left(h_\epsilon(\hat{x}_i) - r(W)\right)\right) \Big|^2 \\ &\qquad + \epsilon \Big| \hat{F}_{2,1}\left(Q\left(-h_\epsilon(\hat{x}_i) - r(W)\right)\right) \Big|^2 \Big] \Big\}, \end{aligned}$$

where (24) follows from Jensen's inequality. Because $[h_\epsilon(\hat{x}_0), \ldots, h_\epsilon(\hat{x}_{M+1})]$ is an expansion of sequence $[\hat{x}_0, \ldots, \hat{x}_{M+1}]$, namely $|h_\epsilon(\hat{x}_i) - h_\epsilon(\hat{x}_j)| > |\hat{x}_i - \hat{x}_j|$ when $i \neq j$, given $W = w$ for any w, one must have

$$\sum_{i=1}^{M} \left| \hat{F}_{1,2} \left(Q \left(h_\epsilon(\hat{x}_i) - r(w) \right) \right) \right|^2 \leq \|\hat{F}_1 - \hat{F}_2\|_2^2$$

and

$$\sum_{i=1}^{M} \left| \hat{F}_{2,1} \left(Q \left(-h_\epsilon(\hat{x}_i) - r(w) \right) \right) \right|^2 \leq \|\hat{F}_1 - \hat{F}_2\|_2^2.$$

Therefore,

$$\|\hat{\Psi}\hat{F}_1 - \hat{\Psi}\hat{F}_2\|_2 \leq \|\hat{F}_1 - \hat{F}_2\|_2,$$

which implies that $\hat{\Psi}$ is nonexpansive and hence continuous. Therefore, by Brouwer's fixed point theorem [15], the continuity of $\hat{\Psi}$ implies that there exists at least one $\hat{F}^*$ such that

$$\hat{\Psi}\hat{F}^* = \hat{F}^*. \tag{25}$$

A. A Linear System Method

To compute a solution $\hat{F}^*$, note that equation (25) is equivalent to the following linear system

$$(\boldsymbol{I} - \boldsymbol{A})\hat{F}^* = \boldsymbol{d}, \tag{26}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{d} = [0, \epsilon, \ldots, \epsilon, 1]^T$ is an $(M+2) \times 1$ vector, and $\boldsymbol{A}$ is the probability weight matrix with the element $a_{i,j}$ given by

$$a_{i,j} = \begin{cases} 0, & \text{if } i = 1; \\ (1-\epsilon)p_{i,j}^+ - \epsilon p_{i,j}^-, & \text{if } 2 \leq i \leq M+1, \text{where} \\ & p_{i,j}^+ = \Pr\{Q\left(h_\epsilon(\hat{x}_i) - r(W)\right) = j\} \\ & p_{i,j}^- = \Pr\{Q\left(-h_\epsilon(\hat{x}_i) - r(W)\right) = j\}; \\ 1, & \text{if } i = M+2. \end{cases}$$

Note that the weights $p_{i,j}^+$ and $p_{i,j}^-$ can be easily computed because $p_{i,j}^+$ is the probability of $r(W)$ taking values in $(h_\epsilon(\hat{x}_i) - \hat{x}_j - \frac{\Delta_M}{2}, h_\epsilon(\hat{x}_i) - \hat{x}_j + \frac{\Delta_M}{2}]$ and $p_{i,j}^-$ is the probability of $r(W)$ taking values in $(-h_\epsilon(\hat{x}_i) - \hat{x}_j - \frac{\Delta_M}{2}, -h_\epsilon(\hat{x}_i) - \hat{x}_j + \frac{\Delta_M}{2}]$. There may exist more than one solution in $\hat{\mathcal{S}}$ to the linear system (26). However, the accuracy analysis in Section V shows that no matter which solution is used to compute the entropy rate, the resulting error is in the order of the length of the quantization interval.

Once an approximation of the cdf F is obtained, the entropy rate can be computed as is discussed in Section II-D.

B. An Iterative Method

As an alternative to solving the linear system (26), we propose an iterative method which is simpler to implement. Since $\hat{\mathcal{S}}$ is a compact convex subset of the strictly convex Banach space $\mathbb{R}^{M+2}$ and $\hat{\Psi}$ is a continuous nonexpansive mapping, according to Theorem 2 in [16], one can conclude that for any initial point $\hat{F}_0 \in \hat{\mathcal{S}}$ and $t \in (0, 1)$, the sequence $\{\hat{\Psi}_t^n(\hat{F}_0)\}$ converges to a fixed point of $\hat{\Psi}$, where $\hat{\Psi}_t$ is an alternative mapping which is given by $\hat{\Psi}_t(\hat{F}) = (1-t)\hat{F} + t\hat{\Psi}(\hat{F})$.

The above result suggests an algorithm which approximates a fixed point of the discrete mapping $\hat{\Psi}$. Starting from any initial point $\hat{F}_0$, we repeatedly apply $\hat{\Psi}_t$ to $\hat{F}_0$ and terminate when the distance between two successive resulting points is smaller than some pre-specified threshold.

Two figures are provided in the following in order to illustrate the effectiveness of the iterative method. Fig. 1 shows the numerically computed entropy rate in the case of observing the Markov process through a memoryless binary symmetric channel (BSC). The entropy rate of the HMP is plotted as a function of the transition probability ϵ of the Markov chain and the crossover probability δ of the BSC. Fig. 2 plots one numerical approximation of the cdf F using the iterative method in the BSC case with $\epsilon = 0.05$ and $\delta = 0.2$. It appears that the cdf F is rather complicated with infinite amount of details. (In certain cases the support of the probability measure of the log-likelihood ratio is a Cantor set [17, Section V].)

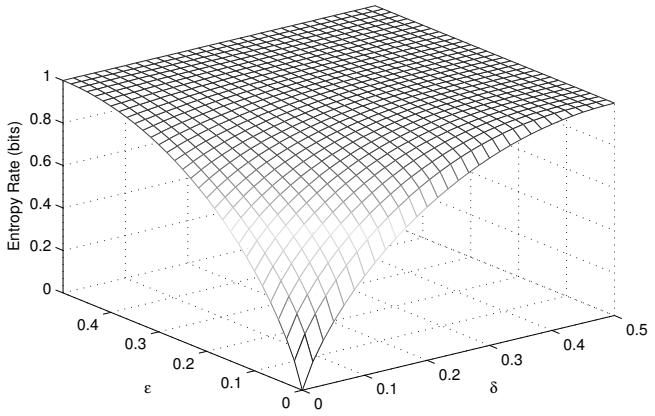

Fig. 1. The entropy rate as a function of the transition probability of the Markov chain (ϵ) and the crossover probability of the BSC (δ).

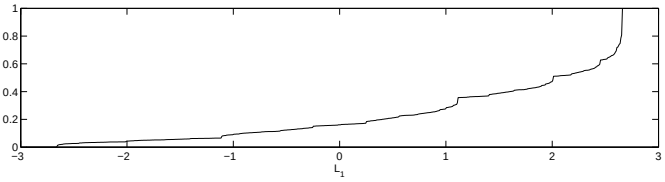

Fig. 2. The numerically computed cdf F of a BSC case with $\epsilon = 0.05$, $\delta = 0.2$ and 800 samples used. One may find complicated and seemingly infinite amount of details when zooming in the figure.

V. Accuracy of the Numerical Method

The purpose of this section is to analyze the accuracy of the numerical method for computing the entropy rate.

Theorem 3: Let H denote the conditional entropy (15) and let $\hat{H}$ denote the computed entropy rate using the numerical method described in Section IV. The accuracy of the the numerical method, i.e., $|H - \hat{H}|$, is upper bounded as

$$|H - \hat{H}| \leq \frac{\mathcal{K}}{2\epsilon}\Delta_M, \tag{27}$$

where M is the number of samples used in the method, Δ_M is the length of the quantization interval given by $\Delta_M = \frac{2}{M}\log\frac{1-\epsilon}{\epsilon}$ and

$$\mathcal{K} = \log\frac{1-\epsilon}{\epsilon} + \frac{1}{2}\mathsf{E}\left\{|r(Y_1)|\right\}.$$

Remark: One may check that $\mathsf{E}\{|r(Y_1)|\} < \infty$ for most of channels of interest in communications, such as BSC, additive white Gaussian noise channel, and channels with additive noise following Rayleigh, Rician or Laplacian distribution.

Proof: Computing the entropy rate of the output process $\{Y_n\}$ can be reduced to computing the conditional entropy H. For simplicity, we introduce the following notations

$$\begin{aligned} S(l) &= \exp(-\alpha X_0 - r(Y_1) - l), \\ B_1 &= 1 + \exp(-\alpha X_0 - r(Y_1)), \\ B_2 &= 1 + \exp(\alpha X_0 + r(Y_1)). \end{aligned}$$

Recall (15) and (16), one may have the following,

$$\begin{aligned} H &= \mathsf{E}\left\{H_2\left((1+\exp\left[-\alpha X_0 - r(Y_1) - L_2\right])^{-1}\right)\right\} \\ &= \mathsf{E}\left\{\int_{-\alpha}^{\alpha} H_2\left(\frac{1}{1+S(l)}\right)\left(\frac{\mathrm{d}F^*(l)}{B_1} + \frac{-\mathrm{d}F^*(-l)}{B_2}\right)\right\}, \end{aligned}$$

where F^* is the solution to the fixed-point functional equation (system) $\Psi F = F$. Define

$$K(l) \triangleq \mathsf{E}\left\{H_2\left(\frac{1}{1+S(l)}\right)\frac{1}{B_1} + H_2\left(\frac{1}{1+S(-l)}\right)\frac{1}{B_2}\right\}.$$

Therefore, the conditional entropy can be computed by

$$H = \int_{-\alpha}^{\alpha} K(l)\mathrm{d}F^*(l).$$

The derivative of $K(l)$ is bounded as

$$\begin{aligned} |K'(l)| &= \left|\mathsf{E}\left\{\frac{S(l)\log S(l)}{[1+S(l)]^2 B_1} - \frac{S(-l)\log S(-l)}{[1+S(-l)]^2 B_2}\right\}\right| \\ &\le \mathsf{E}\left\{\frac{1}{2+1/S(l)+S(l)}|-\alpha X_0 - r(Y_1) - l| \right. \\ &\qquad \left. + \frac{1}{2+1/S(-l)+S(-l)}|-\alpha X_0 - r(Y_1) + l|\right\} \\ &\le \mathsf{E}\left\{\frac{1}{4}(2\alpha + |r(Y_1)|) + \frac{1}{4}(2\alpha + |r(Y_1)|)\right\} \\ &\le \alpha + \frac{1}{2}\mathsf{E}\left\{|r(Y_1)|\right\} = \mathcal{K}. \end{aligned}$$

Using the numerical method in Section IV, the computed entropy rate $\hat{H}$ can be represented by

$$\hat{H} = \sum_{i=0}^{M} K\left(\hat{x}_i + \frac{1}{2}\Delta_M\right)\hat{P}_i,$$

where $\hat{x}_i$ is given in (22), and $\hat{P}_i = \hat{F}^*(i+1) - \hat{F}^*(i)$ for $i = 0, \ldots, M$. In order to calculate the deviation $|H - \hat{H}|$, we introduce the cdf $\tilde{F}$ which is a piecewise-constant extension of $\hat{F}^*$. Formally, $\tilde{F}$ is given by

$$\tilde{F}(x) = \hat{F}^*(Q(x)), \quad x \in \mathbb{R}.$$

Note that

$$\hat{H} = \int_{\mathbb{R}} K(l)\mathrm{d}\tilde{F}(l).$$

Therefore, the deviation

$$\begin{aligned} |H - \hat{H}| &= \left|\int_{\mathbb{R}} K(l)\mathrm{d}(F^*(l) - \tilde{F}(l))\right| \\ &= \left|\int_{\mathbb{R}} (F^*(l) - \tilde{F}(l))K'(l)\mathrm{d}l\right| \qquad (28) \\ &\le \mathcal{K}\int_{\mathbb{R}} \left|F^*(l) - \tilde{F}(l)\right|\mathrm{d}l = \mathcal{K}d(F^*, \tilde{F}) \end{aligned}$$

where (28) follows from integration by parts and the fact that $|F^*(l) - \tilde{F}(l)|$ is 0 outside $[-\alpha, \alpha]$. Thus, the accuracy of the numerical computation reduces to characterizing the L^1 distance between F^* and $\tilde{F}$.

We want to show that $\tilde{F}$ satisfies the following equality at the values $\{\hat{x}_i\}$, i.e.,

$$(\Psi\tilde{F})(\hat{x}_i) = \tilde{F}(\hat{x}_i), \quad 0 \le i \le M+1. \qquad (29)$$

First, it is easy to see that for $i = 0$ and $i = M+1$, the equality holds naturally by the definition of Ψ. Secondly, for the remaining values of i, the following is true,

$$\begin{aligned} (\Psi\tilde{F})(\hat{x}_i) &= \mathsf{E}\Big\{(1-\epsilon)\tilde{F}(h_\epsilon(\hat{x}_i) - r(W)) \\ &\qquad + \epsilon(1 - \tilde{F}(-h_\epsilon(\hat{x}_i) - r(W)))\Big\} \\ &= \mathsf{E}\Big\{(1-\epsilon)\hat{F}^*(Q(h_\epsilon(\hat{x}_i) - r(W))) \\ &\qquad + \epsilon(1 - \hat{F}^*(Q(-h_\epsilon(\hat{x}_i) - r(W))))\Big\} \\ &= \hat{F}^*(i) = \tilde{F}(\hat{x}_i). \end{aligned}$$

Note that equation (21) suggests $d(\Psi^i\tilde{F}, \Psi^{i-1}\tilde{F}) \le (1-2\epsilon)^{i-1}d(\Psi\tilde{F}, \tilde{F})$. In addition, according to the contraction mapping property,

$$\lim_{n\to\infty} \Psi^n\tilde{F} = F^*.$$

Therefore,

$$\begin{aligned} d(F^*, \tilde{F}) &= \lim_{n\to\infty} d(\Psi^n\tilde{F}, \tilde{F}) \\ &\le \lim_{n\to\infty} \sum_{i=1}^{n} d(\Psi^i\tilde{F}, \Psi^{i-1}\tilde{F}) \\ &\le \lim_{n\to\infty} \sum_{i=1}^{n} (1-2\epsilon)^{i-1} d(\Psi\tilde{F}, \tilde{F}) \\ &= \frac{1}{2\epsilon} d(\Psi\tilde{F}, \tilde{F}). \end{aligned}$$

Thus, obtaining the bound of $d(\Psi\tilde{F}, \tilde{F})$ suffices to characterize $d(F^*, \tilde{F})$.

Claim: $d(\Psi\tilde{F}, \tilde{F}) \le \Delta_M$.

Proof: For the piecewise constant cdf $\tilde{F}$, and for each interval $[\hat{x}_i - \Delta_M/2, \hat{x}_i + \Delta_M/2), i = 1, \ldots, M$, the value of $\Psi\tilde{F}$ at $\hat{x}_i$ is equal to $\tilde{F}$, while the value at positions left to $\hat{x}_i$ is smaller than $\tilde{F}$ and the value at positions right to $\hat{x}_i$ is larger than $\tilde{F}$. In addition, $\Psi\tilde{F}$ is increasing. Thus, $\tilde{F}(\hat{x}_{i-1}) \le \Psi\tilde{F}(x) \le \tilde{F}(\hat{x}_i)$ for $\hat{x}_{i-1} \le x \le \hat{x}_i, 1 \le i \le M+1$. Because the maximum value for a cdf is 1, one can upper

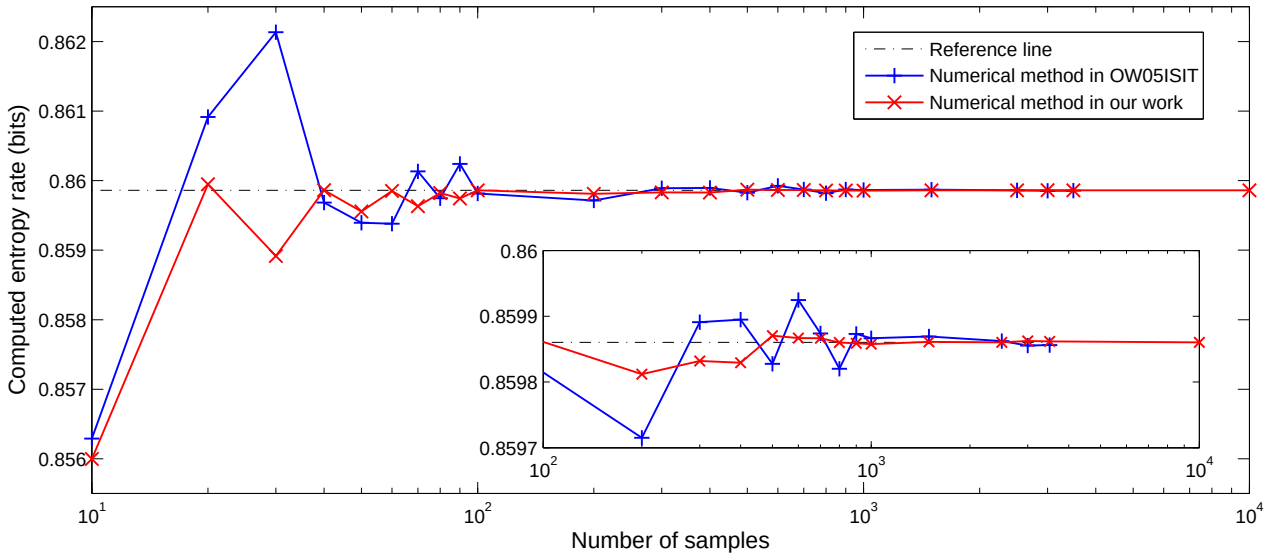

Fig. 3. Computed entropy rates of a BSC case with $\epsilon = 0.05$ and $\delta = 0.2$ using two numerical methods. The curve marked by '$\times$' is produced using the iterative method in Section IV-B and the curve marked by '+' is produced using the numerical method in [10]. The black dot-dash line stands for a reference of the real entropy rate. The inset shows a magnified portion of the original plot with number of samples ranging from 100 to 10000.

bound $d(\Psi\tilde{F}, \tilde{F})$ by Δ_M. □

This suffices to conclude the accuracy (27), as opposed to $O(\Delta_M \log \frac{1}{\Delta_M})$ in [10]. □

Numerical results suggest that the entropy rate computed using the numerical method provided in this paper converges rapidly as the number of samples grows. In addition, the iterative method given in Section IV-B is of low complexity due to the iterative structure. Fig. 3 plots the computed entropy rates using the iterative method in Section IV-B together with those using the numerical method in [10]. A BSC case with $\epsilon = 0.05$ and $\delta = 0.2$ is used in this numerical result. And the entropy rates are plotted against the number of samples used in the computation.

Although both accuracy bounds in [10] and in this paper can be rigorously proved, they are not tight enough for a direct comparison. Depending on ϵ and δ, the numerical method in [10] may perform better or worse than the one in this paper.

ACKNOWLEDGMENT

We thank the reviewers for their constructive comments. We would also like to thank Erik Ordentlich and Tsachy Weissman for helpful discussions. Dongning Guo would also like to thank Sergio Verdú for introducing the entropy rate problem to him in 2002.

REFERENCES

[1] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, pp. 13–20, Prague, Czechoslovakia, House Chechoslovak Acad. Sci., 1957.

[2] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *SIAM J. Control Optim.*, vol. 2, pp. 347–368, 1965.

[3] E. Ordentlich and T. Weissman, "New bounds on the entropy rate of hidden Markov processes," in *Proc. IEEE Inform. Theory Workshop*, pp. 117–122, San Antonio, TX, USA, 2004.

[4] C. Nair, E. Ordentlich, and T. Weissman, "Asymptotic filtering and entropy rate of a hidden Markov process in the rare transitions regime," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 1838–1842, Adelaide, Australia, Sep. 4-9 2005.

[5] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518 – 1569, 2002.

[6] H. Pfister, J. Soriaga, and P. Siegel, "On the achievable information rates of finite state ISI channels," in *Proc. IEEE GLOBECOM*, pp. 2992–2996, San Antonio, TX, USA, Nov. 2001.

[7] D. Arnold and H. Leoliger, "The information rate of binary-input channels with memory," in *Proc. IEEE Int. Conf. Communications*, pp. 2692–2695, Helsinki, Finland, Jun. 2001.

[8] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen, and H. Hollmann, "On the entropy rate of a hidden Markov model," in *Proc. IEEE Int. Symp. Inform. Theory*, p. 12, Chicago, IL, USA, 2004.

[9] H. Pfister, *On the Capacity of Finite State Channels and The Analysis of Convolutional Accumulate-m Codes*. PhD thesis, University of California, San Diego, 2003.

[10] E. Ordentlich and T. Weissman, "Approximations for the entropy rate of a hidden Markov process," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 2198–2202, Adelaide, Australia, 2005.

[11] K. B. Athreya and S. N. Lahiri, *Measure Theory and Probability Theory*. New York: Springer, 2006.

[12] D. P. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inform. Theory*, vol. 53, pp. 453–470, Feb. 2007.

[13] F. Jones, *Lebesgue Integration on Euclidean Space (revised edition)*. Jones and Bartlett Publishers, 2001.

[14] J. Luo and D. Guo, "On the entropy rate of hidden Markov processes observed through arbitrary memoryless channels," in *Proc. Conference on Information Science and Systems*, pp. 1025–1030, Princeton, NJ, USA, March 19–21 2008.

[15] R. V. Vohra, *Advanced Mathematical Economics*. London and New York: Routlege, 2005.

[16] C. S. Wong, "Approximation to fixed points of generalized nonexpansive mappings," *Proceedings of the American Mathematical Society*, vol. 54, pp. 93–97, Jan. 1976.

[17] G. Han and B. Marcus, "Analyticity of entropy rate in families of hidden Markov chains," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 2193–2197, Adelaide, Australia, Sept. 4–9 2005.