# On the Entropy and Filtering of Hidden Markov Processes Observed Via Arbitrary Channels

Jun Luo and Dongning Guo
Department of Electrical Engineering & Computer Science
Northwestern University
Evanston, IL 60208, USA

***Abstract*— This paper studies the entropy and filtering of hidden Markov processes (HMPs) which are observations of discrete-time binary homogeneous Markov chains 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 this equation is guaranteed by martingale theory, its uniqueness follows from contraction mapping property. In order to compute this distribution, the fixed-point functional equation is firstly converted to a linear system through quantization and then solved numerically using quadratic programming. The entropy or differential entropy rate of the HMP can be computed in two ways: one by exploiting the average entropy of each input symbol conditioned on past observations, and the other by applying a relationship between the input-output mutual information and the stationary distribution obtained via filtering. Two examples, where the numerical method is applied to the binary symmetric channel (BSC) and additive white Gaussian noise (AWGN) channel, are presented. Unlike many other numerical methods, this numerical solution is not based on averaging over a long sample path of the HMP.**

## I. Introduction

Let $\{X_n\}$ be a binary Markov chain with symmetric crossover probability $\epsilon$. 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 $Y_{k+1}^{\infty}$, 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 of HMPs has been studied since as early as the 1950s. Blackwell obtained an expression of the entropy in terms of a 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 (filtering) viewpoint. In 1965, Wonham [2] used stochastic differential equation to study 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. The stationary distribution of this Markov process determines the quality of estimating $X_n$ using the past observations $Y_{-\infty}^{n}$. Furthermore, Nair *et al.* [4] used the techniques in [3] to study the behavior of filtering error probability and obtained tight bounds of the entropy rate in the rare-transition regime, i.e. when $\epsilon$ is small. A recent overview of statistical and information-theoretic aspects of HMPs is presented in [5].

In absence of an analytical solution, 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, random 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 the stationary distribution of this variable. 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 of HMPs using filtering techniques and numerical methods. A fixed-point functional equation whose solution characterizes the stationary *distribution* of the *conditional* distribution of $X_0$ given the past observations $Y_{-\infty}^{0}$ is derived in Section II. The existence of a solution to this equation is guaranteed by martingale theory, and Section III shows that the uniqueness is due to a contraction mapping property. Since no explicit analytical solution to this equation is known, a numerical method is developed in Section V using quadratic programming, which gives a good approximation. In addition, the entropy rate and the input-output mutual information are computed by two different methods. Like the numerical methods in [9] and [10], the scheme in this paper does not require a very long 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 scheme 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 crossover probability $\epsilon \in (0, 1/2)$. Let $\{Y_n\}$ be the observation of $\{X_n\}$ through a 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. For every $n = 1, 2, \ldots$, the entropy is related to the input-output mutual information of the channel

by

$$H(Y_1^n) = nH(Y|X) + I\left(X_1^n; Y_1^n\right) \tag{1}$$

where $H(Y|X)$ is the conditional entropy of the memoryless channel. Note that in case the alphabet $\mathcal{Y}$ is continuous and that $P_{Y|X}(\cdot|\pm 1)$ are densities, we shall replace the entropies in (1) by corresponding differential entropies. As far as the entropy or differential entropy of the output is concerned, it suffices to study the mutual information.

The input-output mutual information can also be written as

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} I(X_1 \dots X_n; Y_1 \dots Y_n) \\
&= H_2(\epsilon) - \lim_{n\to\infty} \frac{1}{n} H(X_1 \dots X_n | Y_1 \dots Y_n) \qquad (2)\\
&= H_2(\epsilon) - H(X_1 | X_0, Y_1^\infty) \qquad (3)
\end{aligned}$$

where $H_2(\cdot)$ is the binary entropy function.

One can treat $H(X_1|X_0, Y_1^\infty)$ in (3) as the expectation of the binary entropy of $X_1$ conditioned on $X_0, Y_1^\infty$, i.e.,

$$H(X_1|X_0, Y_1^\infty) = \mathsf{E}\left\{H_2(P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty))\right\} \tag{4}$$

where the conditional probability $P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)$ is a random number in $[0, 1]$ dependent of $X_0$ and $Y_1^\infty$. Consequently, in order to compute the entropy or input-output mutual information of the HMP, it suffices to obtain the distribution of $P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)$.

Note that for given $x$ and $y$,

$$\begin{aligned}
&P_{X_1|X_0,Y_1,Y_2^\infty}(X_1|x, y, Y_2^\infty) \\
&= \frac{P_{X_1,X_0,Y_1|Y_2^\infty}(X_1, x, y|Y_2^\infty)}{P_{X_0,Y_1|Y_2^\infty}(x, y|Y_2^\infty)} \qquad (5)\\
&= \frac{P_{X_1|Y_2^\infty}(X_1|Y_2^\infty) P_{X_0|X_1}(x|X_1) P_{Y|X}(y|X_1)}{\sum_{x'=\pm 1} P_{X_1|Y_2^\infty}(x'|Y_2^\infty) P_{X_0|X_1}(x|x') P_{Y|X}(y|x')}. \qquad (6)
\end{aligned}$$

The distributions of $P_{X_i|Y_{i+1}^\infty}(+1|Y_{i+1}^\infty)$ is well-defined for every $i = 0, \pm 1, \dots$, on the $\sigma$-algebra generated by $Y_{i+1}^\infty$. In fact, the distributions for all $i$ are also identical, which is a direct consequence of the following proposition and the stationarity of the HMP.

*Proposition 1:* As $n \to \infty$,

$$P_{X_0|Y_1^n}(+1|Y_1^n) \to P_{X_0|Y_1^\infty}(+1|Y_1^\infty) \tag{7}$$

holds with probability 1.

*Proof:* For every $n = 1, 2, \dots$, define the conditional expectation $Z_n$ such that

$$Z_n = \mathsf{E}\left\{1_{\{X_0=+1\}} \,\middle|\, Y_1^n\right\} = P_{X_0|Y_1^n}(+1|Y_1^n). \tag{8}$$

Denote by $\mathcal{F}_n$ the $\sigma$-algebra generated by $Y_1^n$. Then $\{\mathcal{F}_n : n = 1, 2, \dots\}$ is a filtration and $\{(Z_n, \mathcal{F}_n) : n = 1, 2, \dots\}$ is a Doob martingale. Because $Z_n$ is uniformly bounded by 1, by Doob's martingale convergence theorem [15, Theorem 13.3.7],

$$\mathsf{E}\left\{1_{\{X_0=+1\}} \,\middle|\, Y_1^n\right\} \to \mathsf{E}\left\{1_{\{X_0=+1\}} \,\middle|\, Y_1^\infty\right\} \tag{9}$$

with probability 1. ∎

Note the above proof is also valid for convergence of 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 conditioning only changes the probability measure defined on the $\sigma$-algebra generated by $Y_1^n$.

Define the likelihood ratio of a binary random variable $B$ with any observation $U$ as

$$\Lambda_B(U) = \frac{P_{B|U}(+1|U)}{P_{B|U}(-1|U)} \ . \tag{10}$$

Then the computation of entropy rate boils down to obtaining the distribution of the log-likelihood ratio

$$L_{i+1} = \log \Lambda_{X_i}(Y_{i+1}^\infty) \ . \tag{11}$$

Note that $L_i$, $i = 0, \pm 1, \dots$, are identically distributed.

In the remainder of this section, we show that the distribution of $L_i$ satisfies a fixed-point functional equation using the fact that $L_i$ is a function of $L_{i+1}$ and $Y_i$.

### A. 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 cumulative distribution function (cdf) of $L_{i+1}$ conditioned on $X_i = +1$ be denoted by $F(\cdot)$, i.e.,

$$F(l) = \Pr\left\{L_{i+1} \le l | X_i = +1\right\}. \tag{12}$$

*Theorem 1:* The conditional cdf $F$ satisfies

$$\begin{aligned}
&F\left(\log \frac{\epsilon + (1-\epsilon)e^x}{\epsilon e^x + (1-\epsilon)}\right) \\
&= \epsilon + \mathsf{E}\left\{(1-\epsilon) F\left(x - r(W)\right) - \epsilon F\left(-x - r(W)\right)\right\} \qquad (13)
\end{aligned}$$

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

$$r(y) = \log \frac{P_{Y|X}(y|+1)}{P_{Y|X}(y|-1)} \ . \tag{14}$$

*Proof:* The log-likelihood ratio $L_i$ defined in (11) accepts a natural bound as follows

$$|L_i| \le \log \frac{1-\epsilon}{\epsilon}, \tag{15}$$

because in terms of estimating $X_{i-1}$, providing $Y_i^\infty$ is no better than providing $X_i$.

Use $F_{U|V}(u|v)$ to denote the cdf of random variable $U$ conditioned on $V = v$, i.e.

$$F_{U|V}(u|v) = \Pr\left\{U \le u | V = v\right\} \tag{16}$$

The key of the proof is the following evolution, which follows from the Bayes' rule and definition (11)

$$\begin{aligned}
L_i &= \log \frac{P_{X_{i-1}|Y_i^\infty}(+1|Y_i^\infty)}{P_{X_{i-1}|Y_i^\infty}(-1|Y_i^\infty)} \qquad (17)\\
&= \log \frac{P_{Y_i^\infty|X_{i-1}}(Y_i^\infty|+1)}{P_{Y_i^\infty|X_{i-1}}(Y_i^\infty|-1)} \qquad (18)\\
&= \log \frac{e^{\alpha + r(Y_i) + L_{i+1}} + 1}{e^{r(Y_i) + L_{i+1}} + e^\alpha} \ , \qquad (19)
\end{aligned}$$

where $\alpha = \log[(1-\epsilon)/\epsilon]$.

Define

$$h_\epsilon(l) = \log \frac{(1-\epsilon)e^l - \epsilon}{(1-\epsilon) - \epsilon e^l} \tag{20}$$

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

$$L_{i+1} = h_\epsilon(L_i) - r(Y_i)\ . \tag{21}$$

Clearly, by changing the variable,

$$\begin{aligned} F_{L_i|Y_i,X_i}(l|y,x) &= F_{L_{i+1}|Y_i,X_i}(h_\epsilon(l) - r(y)|y,x) &(22)\\ &= F_{L_{i+1}|X_i}(h_\epsilon(l) - r(y)|x), &(23)\end{aligned}$$

and thus

$$\begin{aligned} F_{L_i|X_i}(l|x) &= \int_{\mathcal{Y}} F_{L_i|Y_i,X_i}(l|y,x)\, \mathrm{d}P_{Y|X}(y|x) &(24)\\ &= \int_{\mathcal{Y}} F_{L_{i+1}|X_i}(h_\epsilon(l) - r(y))|x) \mathrm{d}P_{Y|X}(y|x)\ . &(25)\end{aligned}$$

Also, because $L_i$ is a function of $Y_i^\infty$, one can get

$$F_{L_i|X_{i-1}}(l|x) = (1-\epsilon)F_{L_i|X_i}(l|x) + \epsilon F_{L_i|X_i}(l|-x). \tag{26}$$

Since $L_i$ are identically distributed even conditioned on $X_{i-1} = +1$, one can define $F(l) \equiv F_{L_i|X_{i-1}}(l|+1)$, $l \in \mathbb{R}$. Furthermore, note the following fact by symmetry,

$$F_{L_i|X_{i-1}}(l|x) = 1 - F_{L_i|X_{i-1}}(-l|-x)\,. \tag{27}$$

Substituting from (25) into (26) and letting $x = +1$ yields

$$\begin{aligned} F(l) = \mathsf{E}\{&(1-\epsilon)F(h_\epsilon(l) - r(W)) \\ &+\epsilon(1 - F(-h_\epsilon(l) - r(W)))\}\ . \end{aligned} \tag{28}$$

The inverse of $h_\epsilon(l)$ is

$$q_\epsilon(x) = \log \frac{\epsilon + (1-\epsilon)e^x}{\epsilon e^x + (1-\epsilon)} \tag{29}$$

for $x \in \mathbb{R}$.

Equation (28) becomes (13) by letting $x = h_\epsilon(l)$ and hence $l = q_\epsilon(x)$. ∎

### B. Non-symmetric Channels

Consider a channel characterized by $P_{Y|X}(\cdot|+1)$ and $P_{Y|X}(\cdot|-1)$ which are in general not symmetric. Let $F_+(\cdot)$ (resp. $F_-(\cdot)$) 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

$$\begin{bmatrix} F_+(q_\epsilon(x)) \\ F_-(q_\epsilon(x)) \end{bmatrix} = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \begin{bmatrix} \mathsf{E}\{F_+(x - r(U))\} \\ \mathsf{E}\{F_-(x - r(V))\} \end{bmatrix} \tag{30}$$

for every $x \in \mathbb{R}$, where $q_\epsilon(x)$ is given in (29), $U \sim P_{Y|X}(\cdot|+1)$, $V \sim P_{Y|X}(\cdot|-1)$, and $U$, $V$ are independent.

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 (13) and (30) in the special case of discrete memoryless channels were also established.

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

An explicit solution to the fixed-point functional equations (13) and (30) is not available, despite the following result.

*Proposition 2:* The fixed-point functional equation (13) admits no more than one solution as a cdf.

*Proof:* First, rewrite (13) as (28) with the variable $l$ replaced by $u$. Let $\mathcal{S}$ denote the set of cdfs whose corresponding probability measure has support within in the region $[-\log(1-\epsilon)/\epsilon, \log(1-\epsilon)/\epsilon]$, which is denoted as $\Omega$. 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

$$d(F_1, F_2) = \int_{\mathbb{R}} |F_1(u) - F_2(u)|\, \mathrm{d}u. \tag{31}$$

Define the operator $\Psi$ over the set $\mathcal{S}$ as

$$(\Psi F)(u) = \begin{cases} 0, & u < -\log\frac{1-\epsilon}{\epsilon}; \\ G(u), & u \in \Omega; \\ 1, & u > \log\frac{1-\epsilon}{\epsilon}. \end{cases} \tag{32}$$

where

$$\begin{aligned} G(u) = \mathsf{E}\big\{&(1-\epsilon)F(h_\epsilon(u) - r(W)) \\ &+ \epsilon(1 - F(-h_\epsilon(u) - r(W)))\big\}. \end{aligned} \tag{33}$$

For simplicity, denote $f_1(x) - f_2(x)$ as $\{f_1 - f_2\}(x)$ where $f_1$ and $f_2$ are two functions. The key to the proof is the fact that $\Psi$ is a contraction mapping under the $L^1$ distance.

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

$$\begin{aligned} &d(\Psi F_1, \Psi F_2) \\ &= \int_\Omega \Big|\mathsf{E}\Big\{(1-\epsilon)\{F_1 - F_2\}(h_\epsilon(u) - r(W)) \\ &\qquad + \epsilon\{F_2 - F_1\}(-h_\epsilon(u) - r(W))\Big\}\Big|\mathrm{d}u &(34)\\ &\le \int_\Omega \mathsf{E}\Big\{\Big|(1-\epsilon)\{F_1 - F_2\}(h_\epsilon(u) - r(W)) \\ &\qquad + \epsilon\{F_2 - F_1\}(-h_\epsilon(u) - r(W))\Big|\Big\}\mathrm{d}u &(35)\\ &\le \int_\Omega \mathsf{E}\Big\{(1-\epsilon)\big|\{F_1 - F_2\}(h_\epsilon(u) - r(W))\big| \\ &\qquad + \epsilon\big|\{F_2 - F_1\}(-h_\epsilon(u) - r(W))\big|\Big\}\mathrm{d}u &(36)\\ &= \mathsf{E}\Big\{\int_\Omega \Big[(1-\epsilon)\big|\{F_1 - F_2\}(h_\epsilon(u) - r(W))\big| \\ &\qquad + \epsilon\big|\{F_1 - F_2\}(-h_\epsilon(u) - r(W))\big|\Big]\mathrm{d}u\Big\} &(37)\end{aligned}$$

where (35) and (36) follow from Jenson's inequality, and the order of integration and expectation is changed in (37) using Tonelli's theorem [11, p. 183].

Note that $q_\epsilon(\cdot)$ defined in (29) is the inverse of $h_\epsilon(\cdot)$. For the first term in the integrand of (37), one can obtain the following

$$\begin{aligned}\int_\Omega \big|F_1(h_\epsilon(u)-r(W))-F_2(h_\epsilon(u)-r(W))\big|\mathrm{d}u &= \int_{\mathbb{R}} \big|F_1(t)-F_2(t)\big|q_\epsilon'(t+r(W))\mathrm{d}t \quad (38)\\ &\le (1-2\epsilon)\int_{\mathbb{R}} \big|F_1(t)-F_2(t)\big|\mathrm{d}t \quad (39)\end{aligned}$$

with the inequality justified by the fact that $q_\epsilon'(t) \le 1-2\epsilon$ for all $t \in \mathbb{R}$.

Similarly, one can upper bound the second term in (37), which, together with (39), leads to

$$d(\Psi F_1, \Psi F_2) \le (1-2\epsilon) d(F_1, F_2) \quad (40)$$

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

Note that the solution to (13) is a fixed point of the operator $\Psi$. Suppose there exist two cdfs $F_1^*$ and $F_2^*$ in $\mathcal{S}$ which satisfy (13), by the contraction mapping property of $\Psi$, one can get the following inequality

$$d(F_1^*, F_2^*) = d(\Psi F_1^*, \Psi F_2^*) \le (1-2\epsilon) d(F_1^*, F_2^*). \quad (41)$$

The inequality (41) implies that $d(F_1^*, F_2^*) = 0$. That is, $F_1^*$ and $F_2^*$ must be the same cdf. ∎

Note that the result in Proposition 2 also applies to (30) which can be shown using the same contraction mapping argument.

## IV. Computation of Entropy Rate and Mutual Information

In this section, we propose two methods for computing the input-output mutual information and hence the entropy rate of HMPs.

### A. Direct Method

Recall (1) and (3), the direct method only requires computing the conditional entropy $H(X_1|X_0, Y_1^\infty)$. Here, we re-derive the expression for the conditional probability $P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)$ as

$$\begin{aligned}&P_{X_1|X_0,Y_1^\infty}(+1|X_0, Y_1^\infty)\\ &= \frac{P_{X_1|Y_2^\infty}(+1|Y_2^\infty)P_{X_0|X_1}(X_0|+1)P_{Y|X}(Y_1|+1)}{\sum_{x'=\pm 1} P_{X_1|Y_2^\infty}(x'|Y_2^\infty)P_{X_0|X_1}(X_0|x')P_{Y|X}(Y_1|x')} \quad (42)\\ &= \big(1+\exp\left[-\alpha X_0 - r(Y_1) - L_2\right]\big)^{-1}. \quad (43)\end{aligned}$$

Therefore, in view of (4), one can write

$$\begin{aligned}&H(X_1|X_0, Y_1^\infty)\\ &= \mathsf{E}\left\{ H_2\left(\frac{1}{1+\exp\left[-\alpha X_0 - r(Y_1) - L_2\right]}\right) \,\middle|\, X_0, Y_1\right\}. \quad (44)\end{aligned}$$

Also note that for given $x$ and $y$

$$\begin{aligned}&P_{L_2|X_0,Y_1}(l|x,y)\\ &= \frac{\sum_{x'\in\{\pm1\}} P_{L_2,X_0,X_1,Y_1}(l,x,x',y)}{P_{X_0,Y_1}(x,y)} \quad (45)\\ &= \frac{P_{X_0,X_1,Y_1}(x,+1,y)\frac{\mathrm{d}F(l)}{\mathrm{d}l}}{P_{X_0,Y_1}(x,y)}\\ &\quad + \frac{P_{X_0,X_1,Y_1}(x,-1,y)\frac{-\mathrm{d}F(-l)}{\mathrm{d}l}}{P_{X_0,Y_1}(x,y)}. \quad (46)\end{aligned}$$

Therefore, in order to compute the entropy, it suffices to solve the fixed-point functional equation (13) (or (30)).

### B. Computation Via an Information-Estimation Formula

One can also compute the input-output mutual information of HMPs using a fundamental information–estimation relationship. In the following the computation is illustrated using the special case of binary symmetric channel (BSC). The following is a variant of a result due to Palomar and Verdú [12].

*Proposition 3 ( [12]):* Let $\{X_n\}$ and $\{Y_n\}$ be the respective input and output of a BSC with crossover probability $\delta \in (0,1)$. For every input distribution $P_{X^n}$,

$$\lim_{n\to\infty}\frac{1}{2n}\frac{\mathrm{d}}{\mathrm{d}\delta} I(X_{-n}^n; Y_{-n}^n) = \mathsf{E}\left\{\frac{\Lambda-1}{\Lambda+1}\log\frac{\Lambda+e^{-\beta}}{\Lambda+e^{\beta}}\right\} - \beta \quad (47)$$

where $\Lambda$ takes the limiting distribution of the likelihood ratio $\Lambda_{X_i}(Y^{n\setminus i})$ with $Y^{n\setminus i} = (Y_{-n}^{i-1}, Y_{i+1}^n)$ as $n > 2i \to \infty$.

Note that one can decompose $\Lambda_{X_i}(Y^{n\setminus i})$ as follows,

$$\log\Lambda_{X_i}(Y^{n\setminus i}) = \log\Lambda_{X_i}(Y_{-n}^{i-1}) + \log\Lambda_{X_i}(Y_{i+1}^n). \quad (48)$$

Thus, the limit distribution of $\Lambda_{X_i}(Y^{n\setminus i})$ as $n\to\infty$ is easy to compute because the respective distribution of $\Lambda_{X_i}(Y_{-\infty}^{i-1})$ and $\Lambda_{X_i}(Y_{i+1}^\infty)$ can be solved from the fixed-point functional equations by Theorem 1 or 2.

Since the right hand side of (47) can be evaluated for every $0 < \delta < 1/2$, the mutual information for any given $\delta \in (0,1)$ can be obtained as an integral, also using the fact that the mutual information is equal to 0 with $\delta = 1/2$.

## V. Numerical Methods

Since no explicit analytic solution to the fixed-point functional equations is known, we develop numerical method to compute it in this section.

Note that $L_i$ accepts a natural bound (15), one can sample $F(\cdot)$ arbitrarily finely to obtain a good approximation. Here we give some examples to illustrate how this quantization-based method can be applied to solving fixed-point functional equations arising in this paper. One may also find that reference [10], as well, 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] discretizing the transition matrix of the alternative Markov process.

### A. Example 1: Binary Symmetric Channel

For a BSC with crossover probability $\delta \in (0, 1/2)$, the fixed-point equation can be written as:

$$\begin{aligned} F(q_\epsilon(x)) = &(1-\epsilon)(1-\delta)F(x-\beta) + (1-\epsilon)\delta F(x+\beta) \\ &+ \epsilon(1-\delta)(1-F(-x-\beta)) \\ &+ \epsilon\delta(1-F(-x+\beta)) \end{aligned} \tag{49}$$

for every $x \in \mathbb{R}$, where $\beta = \log(1/\delta - 1)$.

In fact, the log-likelihood ratio $L_i$ accepts a bound which is tighter than (15). Let the supremum of $L_i$ be $x^*$, which is given in [3], and derived here for completeness. Since $q_\epsilon(x)$ is an increasing contraction mapping of $x$ for every $\epsilon$, $x^*$ satisfies the following boundary condition

$$q_\epsilon(x^* + \beta) = x^* \tag{50}$$

which gives

$$x^* = \log\left[\frac{e^\alpha(1-e^{-\beta})}{2} + \sqrt{\frac{e^{2\alpha}(1-e^{-\beta})^2}{4} + e^{-\beta}}\right]. \tag{51}$$

Thus, one can apply a quantizer with $M$ levels to $[-x^* - \beta, x^* + \beta]$, and denote the resulted sample sequence by $\hat{x}$. Denote the $M$-sample sequence of $F(\cdot)$ evaluated on $\hat{x}$ by an $M \times 1$ vector $\boldsymbol{F}$. Since the right hand side of (49) is a linear combination of shifted versions of $F(\cdot)$, one can multiply $\boldsymbol{F}$ by a matrix $\mathbf{K}$ together with the help of an auxiliary constant vector $\mathbf{d}$ to obtain the discretized expression. The discretization of left hand side of (49) involves quantizing the logarithm, which is a contraction mapping from $\mathbb{R}$ to $(-\alpha, \alpha)$. One can simply quantize the values of $F(\cdot)$ evaluated at $q_\epsilon(\hat{x}_i), i = 1, \cdots, M$, where $\hat{x}_i$ is the $i$th element of $\hat{x}$, to the nearest sample point in $\boldsymbol{F}$. This can be done by pre-multiplying $\boldsymbol{F}$ by a scrambling matrix $\boldsymbol{P}$. Thus, one can convert the non-linear fixed-point functional equation (49) to the following linear system

$$\mathbf{P}\boldsymbol{F} = \mathbf{d} + \mathbf{K}\boldsymbol{F}\ . \tag{52}$$

When a uniform quantizer is used, the matrices $\mathbf{P}$ and $\mathbf{K}$ have the structure depicted in Fig. 1 and Fig. 2 respectively. In Fig. 1, the elements on the curve are all 1's and the rest of the elements in the matrix are 0's. In Fig. 2, all elements on line (1) take the value $(1-\epsilon)(1-\delta)$, all elements on line (2) take the value $(1-\epsilon)\delta$, on line (3) take the value $-\epsilon(1-\delta)$ and on line (4) take the value $-\epsilon\delta$. The rest of the elements in the matrix are 0's.

There are many different ways for numerically solving the linear system (52), such as Gaussian elimination and QR factorization. Some are quite efficient if we utilize the sparsity of the matrices $\mathbf{P}$ and $\mathbf{K}$. For ease of imposing the monotonicity of the cdf $F$, we choose to solve the following quadratic programming with linear convex constraints.

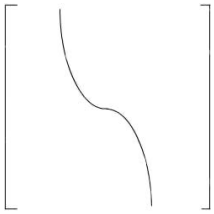

Fig. 1. Structure of Matrix $\mathbf{P}$

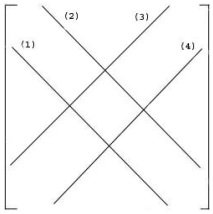


Fig. 2. Structure of Matrix $\mathbf{K}$

$$\begin{aligned} \min_{\boldsymbol{F}} \quad & \|(\mathbf{K}-\mathbf{P})\boldsymbol{F} + \mathbf{d}\|^2 \\ s.t. \quad & F_i = 0, \qquad 1 \le i \le M_1; \\ & F_i - F_{i-1} \ge 0, \qquad M_1 + 1 \le i \le M_1 + M_2 + 1; \\ & F_i = 1, \qquad M_1 + M_2 + 1 \le i \le M_1 + M_2 + M_3\ , \end{aligned} \tag{53}$$

where $F_i$ denotes the $i$th element of $\boldsymbol{F}$, and $M_1$, $M_2$ and $M_3$ denote the number of samples in the intervals $[-x^* - \beta, -x^*)$, $[-x^*, x^*]$ and $(x^*, x^* + \beta]$ respectively.

Since the feasible region of the quadratic programming is a compact convex set, the optimal solution exists. The uniqueness of solution requires a detailed study on the matrix $(\mathbf{K} - \mathbf{P})$, which in general is difficult. Fortunately, since the cdf of $L_i$ conditioned on $X_{i-1} = +1$ is unique, this quadratic programming does give a unique solution when the maximum sampling interval length is small.

There are many methods for solving the quadratic programming problem (53), among which we use the active set method [13]. Although the number of iterations depends on the initial test value, this method can give a quadratic convergence rate [14], which makes the entire computation fast.

Once the cdf $F(\cdot)$ is obtained, the entropy and input-output mutual information can be computed using either the direct method (see Section IV-A) or an information-estimation relationship (see Section IV-B). If the latter method is employed, one can utilize FFT and IFFT to compute the distribution of the left hand side of (48) conditioned on either $X_i = -1$ or $X_i = +1$, because the pdf of the sum of two independent random variables is the convolution of the pdf of each of the two random variables. As long as the distributions of $\log\Lambda$ conditioned respectively on $X_i = -1$ and $X_i = +1$ are obtained, one can average them with equal probability to obtain the distribution of $\log\Lambda$.

Fig. 3 gives the numerical results for BSC. The entropy rate of the output process is plotted as a function of $\epsilon$ and $\delta$. A uniform quantizer is used for computation.

### B. Example 2: AWGN Channel

For AWGN channel described by the following model

$$Y = \sqrt{\gamma}X + N \tag{54}$$

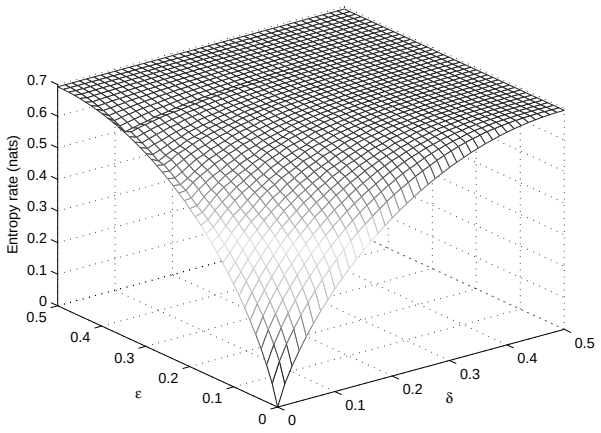


Fig. 3. Entropy Rate of BSC with respect to the transition probability of the Markov chain $\epsilon$ and that of the BSC $\delta$.

where $N \sim \mathcal{N}(0,1)$, the conditional cdf $F$ satisfies

$$F(q_\epsilon(x)) = \epsilon + \mathsf{E}\left\{(1-\epsilon)F\left(x-2W\right) - \epsilon F\left(-x-2W\right)\right\} \tag{55}$$

for every $x \in \mathbb{R}$, $W \sim \mathcal{N}(\sqrt{\gamma}, 1)$.

To linearize (55), one needs to also quantize the support of distribution of $W$. Because it is a standard Gaussian distribution, one can take samples on a finite interval, e.g. $[-5, 5]$, and get a good approximation. In this case, the right hand side in linearized (55) will be expressed as the superposition of shifted versions of $F(\cdot)$ due to different quantization levels of $W$. The matrix $K$ in (52) is dense in this case.

Numerical results for AWGN channel with a uniform quantizer is illustrated in Fig. 4. The differential entropy rate is plotted as a function of the crossover probability $\epsilon$ and the signal-to-noise ratio $\gamma$.

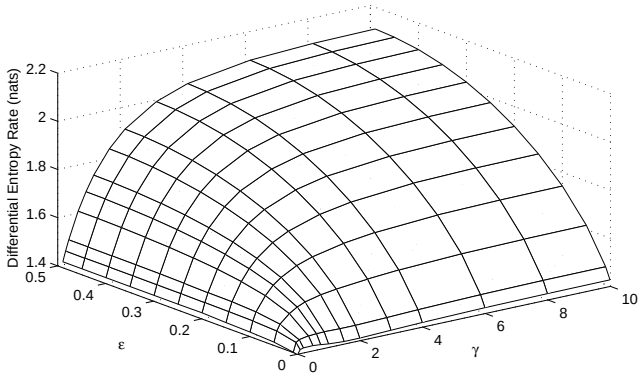


Fig. 4. Differential Entropy Rate of AWGN Channel w.r.t. $\delta$ and $\gamma$.

## VI. Conclusion

This paper derives fixed-point functional equations to characterize the stationary distribution of an input symbol from a binary symmetric Markov chain conditioned on the past observations under general channel models. The existence and uniqueness of the solution to such a fixed-point functional equation are justified using the martingale theory and a contraction mapping property respectively. Although in general these equations cannot be solved analytically, numerical methods have been developed to give an effective approximation. The resulting distribution allows straightforward computation of the entropy rate of the hidden Markov process.

## 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," (San Antonio, Texas), pp. 117–122, Inf. Th. Workshop, October 24-29 2004.

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

[5] Y. Ephraim and N. Merhav, "Hidden Markov Processes," in *IEEE Trans. Information Theory*, vol. 48, no. 6, June 2002.

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

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

[8] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen, and H. Hollmann, "On the entropy rate of a hidden Markov model," (Chicago, IL), p. 12, Int. Symp. Inf. Th., June-July 2004.

[9] H. Pfister, "On the Capacity of Finite State Channels and the Analysis of Convolutional Accumulate-m Codes", Ph.D thesis, pp. 148–151, 2003.

[10] E. Ordentlich, and T. Weissman, "Approximations for the Entropy Rate of a Hidden Markov Process," (Adelaide, Australia), Int. Symp. Inf. Th., Sep. 4-9 2005.

[11] F. Jones, "Lebesgue Integration on Euclidean Space (Revised Edition)," Jones and Bartlett Publishers, 2001.

[12] D. P. Palomar and S. Verdú, "Representation of mutual information via input estimates," in *IEEE Trans. Information Theory*, vol. 53, no. 2, Feb. 2007.

[13] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. Academic Press, 1981.

[14] T. Coleman and Y. Li, "A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM J. Optim.*, vol. 6, no. 4, pp. 1040 – 1058, 1996.

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