Capacity of Gaussian Channels With Duty Cycle and Power Constraints

Lei Zhang, Hui Li, and Dongning Guo, *Senior Member, IEEE*

***Abstract*— In many wireless communication systems, radios are subject to a duty cycle constraint, that is, a radio can only actively transmit signals over a fraction of the time. For example, it is desirable to have a small duty cycle in some low power systems; a half-duplex radio cannot keep transmitting if it wishes to receive useful signals; and a cognitive radio needs to listen and detect primary users frequently. This paper studies the capacity of point-to-point scalar discrete-time Gaussian channels subject to a duty cycle constraint as well as an average transmit power constraint. An idealized duty cycle constraint is first studied, which can be regarded as a requirement on the minimum fraction of nontransmissions or zero symbols in each codeword. Independent input with a unique discrete distribution is shown to achieve the channel capacity. In many situations, numerically optimized on-off signaling can achieve much higher rate than Gaussian signaling over a deterministic transmission schedule. This is in part because the positions of nontransmissions in a codeword can convey information. A more realistic duty cycle constraint is also studied, where the extra cost of transitions between transmissions and nontransmissions due to pulse shaping is accounted for. The capacity-achieving input is correlated over time and is hard to compute. A lower bound of the achievable rate as a function of the input distribution is shown to be maximized by a first-order Markov input process, whose stationary distribution is also discrete and can be computed efficiently. The results in this paper suggest that, under various duty cycle constraints, departing from the usual paradigm of intermittent packet transmissions may yield substantial gain.**

***Index Terms*— Capacity-achieving input, channel capacity, duty cycle, entropy rate, hidden Markov process (HMP), Markov process, Monte Carlo method, mutual information.**

Manuscript received May 10, 2012; revised August 14, 2013; accepted October 30, 2013. Date of publication December 20, 2013; date of current version February 12, 2014. This work was supported in part by the National Science Foundation under Grants CCF-0644344 and ECCS-1231828, and in part by DARPA under Grant W911NF-07-1-0028. D. Guo and L. Zhang were supported by the Institute of Network Coding at the Chinese University of Hong Kong. D. Guo was supported by the University Grants Committee of the Hong Kong Special Administrative Region under Project AoE/E-02/08. This paper was presented at the 2011 and 2012 IEEE International Symposium on Information Theory.

L. Zhang was with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA. He is now with Qualcomm Research, Bridgewater, NJ 08807 USA (e-mail: zhanglei.mike@gmail.com).

H. Li is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: mythlee@gmail.com).

D. Guo is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: dguo@northwestern.edu).

Communicated by S. Diggavi, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2295616

I. Introduction

In MANY wireless communication systems, a radio is designed to transmit actively only for a fraction of the time, which is known as its *duty cycle*. For example, low duty cycle signaling has been used to conserve power in some wideband systems [1], [2]. The physical half-duplex constraint also requires a radio to stop transmission over a frequency band from time to time if it wishes to receive useful signals over the same band. Thus wireless relays are subject to duty cycle constraint, and so are cognitive radios which have to listen to the channel frequently to avoid causing interference to primary users. The *de facto* standard solution under duty cycle constraint is to transmit sizable packets intermittently.

This work studies the fundamental question of what is the optimal signaling for a Gaussian channel with a duty cycle constraint as well as an average transmission power constraint. An important observation is that the signaling in nontransmission periods can be regarded as transmission of a special *zero* signal. We first make a simplistic and idealized assumption that the analog waveform corresponding to each transmitted symbol spans exactly one symbol interval. We restrict our attention to discrete-time scalar additive white Gaussian noise (AWGN) channels for simplicity, where the duty cycle constraint is equivalent to a requirement on the minimum fraction of zero symbols in each transmitted codeword, which is called the *idealized duty cycle constraint*. We then consider the case where a practical pulse shaping filter is used, e.g., for band-limited transmissions. As such, during a transition between a zero symbol and a nonzero symbol, the pulse waveform of the nonzero symbol leaks into the interval of the zero symbol. A *realistic duty cycle constraint* must include the extra cost incurred upon transitions between zero and nonzero symbols. The mathematical model of the preceding input-constrained channels is described in Section II.

Determining the capacity of a channel subject to various input constraints is a classical problem. It is well-known that Gaussian signaling achieves the capacity of a Gaussian channel with average input power constraint only. In addition, Zamir [3] shows that the information rate of an additive noise channel achievable using a white Gaussian input never incurs a loss of more than half a bit per sample with respect to the power constrained capacity. Furthermore, Smith [4] investigated the capacity of a scalar AWGN channel under both peak power constraint and average power constraint. The input distribution that achieves the capacity is shown to be discrete with a finite number of probability mass points. The discreteness of capacity-achieving distributions for various channels, including quadrature Gaussian channels, and Rayleigh-fading channels is also established in [5]–[10]. Chan *et al.* [11] studied the capacity-achieving input distribution for conditional Gaussian channels which form a general channel model for many practical communication systems.

The main results of this paper are summarized in Section III. In the case of the idealized duty cycle constraint, because all costs associated with the constraints can be decomposed into per-letter costs, the optimal input distribution is independent and identically distributed (i.i.d.) over time. Following [4] and [11], we use tools in complex analysis to show that the capacity-achieving input distribution for an AWGN channel with the idealized duty cycle constraint and the average power constraint is discrete. The absence of the peak power (i.e., bounded support) constraint and the additional duty cycle constraint make the derivation harder. Unlike in [4] and [11], the optimal distribution has an infinite number of probability mass points, whereas only a finite number of the points are found in every bounded interval. This allows efficient numerical optimization of the input distribution.

The case of the realistic duty cycle constraint is more challenging. Because the constraint concerns symbol transitions, the capacity-achieving input distribution is dependent over time, and becomes hard to compute. We develop a lower bound of the input-output mutual information as a function of the input distribution. It is proved that, under the realistic duty cycle constraint, a first-order Markov process maximizes the lower bound, the distribution of which is also discrete and can be computed efficiently. The main theorems for the cases of the idealized and the realistic duty cycle constraints are proved in Section IV and V, respectively.

We devote Section VI to the numerical methods and results. In order to compute the achievable rate when the input is a Markov Chain, a Monte Carlo method is introduced in Section VI-A to numerically compute the differential entropy rate of hidden Markov processes. Numerical results in Section VI-B demonstrate that, in the case of the idealized duty cycle constraint, using a numerically optimized discrete signaling achieves higher rates than using Gaussian signaling over a deterministic transmission schedule. For example, if the radio is allowed to transmit no more than half the time, i.e., the duty cycle is no greater than 50%, a near-optimal discrete input achieves 50% higher rate at 10 dB signal-to-noise ratio (SNR). In the case of the realistic duty cycle constraint, numerical results also show that the rate achieved by the Markov process is substantially higher than that achieved by any i.i.d. input. This suggests that, compared to intermittently transmitting packets using Gaussian or Gaussian-like signaling, it is more efficient to disperse nontransmission symbols within each packet to form codewords, which results in a form of *on-off* signaling.

One of the reasons for the superiority of on-off signaling is that the positions of nontransmission symbols can be used to convey information, the impact of which is particularly significant in the case of low SNR or low duty cycle. This has been observed in the past. For example, as shown in [12] (see also [13] and [14]), time sharing or time-division duplex (TDD) can fall considerably short of the theoretical limits in a relay network: The capacity of a cascade of two noiseless binary bit pipes through a half-duplex relay is 1.14 bits per channel use, which far exceeds the 0.5 bit achieved by TDD and even the 1 bit upper bound on the rate of binary signaling.

This work is restricted to a directed link from one point (say, T) to another point (say, R) where T is subject to the duty cycle constraint. There is no constraint on, or interference to, R's receiver, so that R can listen to the channel continuously. The reverse link from R to T, if it exists, does not interfere with the forward link. (For example, the reverse link may use a separate band.)

The situation where three or more nodes communicate using on-off signaling over the same frequency band has been studied in [15] and [16]. Under the half duplex constraint, the received signal of a node is erased during symbol intervals of its own active transmissions. Thus each node sees a multiaccess channel with erasures. With suitable error control coding, all nodes can communicate simultaneously over each frame, accomplishing *virtual* full-duplex communication. The reader is referred to [15] and [16] for an in-depth study of the scheme, called *rapid on-off-division duplex (RODD).*

II. System Model

Consider digital communication systems where coded data is mapped to waveforms for transmission. Usually there is a collection of pulse waveforms, where each pulse represents a symbol (or letter) from a discrete alphabet. We view nontransmission over a symbol interval as transmitting the all zero waveform. In other words, a symbol interval of nontransmission is simply regarded as transmitting a special symbol "0," which carries no energy.

As far as the capacity-achieving input is concerned, it suffices to consider the baseband discrete-time model for the AWGN channel. The received signal over a block of n symbols can be described by

$$Y_i = X_i + N_i \tag{1}$$

where $i = 1, \dots, n$, X_i denotes the transmitted symbol at time i and $N_1, \dots, N_n$ are independent standard Gaussian random variables. For simplicity, we assume that there is no intersymbol interference at the receiver. Each symbol modulates a continuous-time pulse waveform for transmission. If the width of all pulses were exactly one symbol interval, which is denoted by T, the duty cycle is equal to the fraction of nonzero symbols in a codeword. In practice, however, the pulse is usually wider than T, so that the support of the transmitted waveform is greater than the sum of the intervals corresponding to nonzero symbols due to leakage into intervals of adjacent zero symbols. To be specific, suppose the width of a pulse is $(1 + 2c)T$, then each transition between zero and nonzero symbols incurs an additional cost of up to $c\,T$ in terms of actual transmission time.

Let $1 - q$ denote the maximum duty cycle allowed. In this paper, we only consider the non-trivial cases where $0 < q < 1$, and require every codeword $(x_1, x_2, \cdots, x_n)$ to satisfy

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \neq 0\}} + \frac{1}{n}2c\left(1_{\{x_n=0, x_1\neq 0\}} + \sum_{i=1}^{n-1} 1_{\{x_i=0, x_{i+1}\neq 0\}}\right) \le 1-q \tag{2}$$

where $1_{\{\cdot\}}$ is the indicator function. Here the duty cycle is defined in a cyclic manner, where a transition between x_n and x_1 is also counted. This of course has vanishing impact as $n \to \infty$ and thus no impact on the capacity. Therefore, the transition cost is twice that of zero-to-nonzero transitions, because the number of nonzero-to-zero transitions and the number of zero-to-nonzero transitions are equal under the cyclic transition cost configuration. From now on, we refer to (2) as the *duty cycle constraint* (q, c). Note that the idealized duty cycle constraint is the special case $(q, 0)$. If $c \in [0, \frac{1}{2}]$, the left hand side of (2) is equal to the actual duty cycle. If $c > \frac{1}{2}$, the left hand side of (2) is an overestimate of the duty cycle. Nonetheless, we use constraint (2) for its simplicity. In addition, we consider the usual average input power constraint,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le \gamma. \tag{3}$$

In many wireless systems, the transmitter's activity is constrained in the frequency domain as well as in the time domain. In principle, the results in this paper also apply to the more general model where the duty cycle constraint is on the time-frequency plane.

III. Main Results

A. The Case of the Idealized Duty Cycle Constraint

Let μ denote the distribution of the channel input X. The set of distributions with duty cycle constraint $(q, 0)$ and power constraint γ is denoted by

$$\Lambda(\gamma, q) = \left\{\mu : \mu(\{0\}) \ge q,\ \mathsf{E}_\mu\left\{X^2\right\} \le \gamma\right\}. \tag{4}$$

Although implicit in (4), it should be understood that μ is a probability measure defined on the Borel algebra on the real number set, denoted by $\mathcal{B}(\mathbb{R})$.

Theorem 1: The capacity of the additive white Gaussian noise channel (1) with its idealized duty cycle no greater than $1-q$ and the average power no greater than γ is

$$C(\gamma, q) = \max_{\mu \in \Lambda(\gamma, q)} I(X; X+N) \tag{5}$$

where X follows distribution μ and N is standard Gaussian and independent of X. In particular, the following properties hold:

a) the maximum of (5) is achieved by a unique (capacity-achieving) distribution $\mu_0 \in \Lambda(\gamma, q)$;
b) μ_0 is symmetric about 0 and its second moment is exactly equal to γ; and
c) μ_0 is discrete with an infinite number of probability mass points, whereas the number of probability mass points in any bounded interval is finite.

The proof of Theorem 1 is relegated to Section IV. Property (b) suggests that the capacity-achieving input always exhausts the power budget. Property (c) indicates that the capacity-achieving input can be well approximated by some discrete inputs with finite alphabet, which can be computed using numerical methods. The achievable rate of numerically optimized input distribution is studied in Section VI.

B. The Case of the Realistic Duty Cycle Constraint

Let X_k^n denote the subsequence $(X_k, X_{k+1}, \cdots, X_n)$, where $X_k^\infty = (X_k, X_{k+1}, \cdots)$. We also use shorthand $X^n = X_1^n$. Let μ denote the probability distribution of the process $X_1, X_2, \cdots$. We use μ_{X_i} to denote the marginal distribution of X_i, and μ_{X_i, X_j} to denote the joint probability distribution of (X_i, X_j). Denote the set of n-dimensional distributions which satisfy duty cycle constraint (q, c) and power constraint γ by

$$\Lambda^n(\gamma, q, c) = \left\{\mu :\ \mathsf{E}_\mu\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2\right\} \le \gamma,\ \frac{1}{n}\sum_{i=1}^{n}\left[\mu_{X_i}(\{0\}) - 2\,c\,\mu_{X_i, X_{(i \bmod n)+1}}(\{0\}\times(\mathbb{R}\backslash\{0\}))\right] \ge q\right\} \tag{6}$$

where

$$\mu_{X_i, X_j}(\{0\}\times(\mathbb{R}\backslash\{0\})) = P(X_i = 0, X_j \neq 0) \tag{7}$$

denotes the probability of a zero-to-nonzero transition and the modular operation

$$i \bmod n = \begin{cases} i, & \text{if } 1 \le i < n, \\ 0, & \text{if } i = n. \end{cases} \tag{8}$$

is used to express the cyclic transition cost configuration in (2).

The capacity of the AWGN channel (1) with duty cycle constraint (q, c) and power constraint γ is

$$C(\gamma, q, c) = \lim_{n\to\infty} \frac{1}{n} \max_{\mu \in \Lambda^n(\gamma, q, c)} I(X^n; Y^n). \tag{9}$$

The capacity is in fact achieved by a stationary input process. This is justified in Section V-A by showing that any nonstationary input process has a stationary counterpart with equal or greater input-output mutual information per symbol. Let us denote the set of stationary distributions which satisfy duty cycle constraint (q, c) and power constraint γ by

$$\Lambda(\gamma, q, c) = \left\{\mu :\ \mu \text{ is stationary}, \mathsf{E}_\mu\left\{X_1^2\right\} \le \gamma,\ \mu_{X_1}(\{0\}) - 2\,c\,\mu_{X_1, X_2}(\{0\}\times(\mathbb{R}\backslash\{0\})) \ge q\right\}. \tag{10}$$

Theorem 2: For any $\mu \in \Lambda(\gamma, q, c)$, let

$$L(\mu) = I(X_1; X_1 + N) - I(X_1; X_2^\infty) \tag{11}$$

where N is standard Gaussian and independent of X_1. The following properties hold:

a) $L(\mu)$ is a lower bound of the channel capacity;
b) The maximum of $L(\cdot)$ is achieved by a discrete first-order Markov process, denoted by μ^*;
c) μ^* satisfies the following property: Define $B_i = 1_{\{X_i \neq 0\}}$, $i = 1, 2, \ldots$. Then for every i, conditioned on B_i and B_{i+1}, the variables X_i and X_{i+1} are independent, and

$$L(\mu^*) = I(X_1; X_1 + N) - I(B_1; B_2). \tag{12}$$

The proof of Theorem 2 is relegated to Section V. Evidently, increasing the input power by scaling the input linearly not only maintains its duty cycle, but also increases the mutual information. Therefore, the optimal input distribution must exhaust the power budget γ.

In the special case of no transition cost, i.e., $c = 0$, the maximizer μ^* of the lower bound $L(\cdot)$ in Theorem 2 requires $I(B_1; B_2) = 0$, and is exactly the same as the capacity-achieving distribution μ_0 in Theorem 1.

IV. Proof of Theorem 1 (the Case of Idealized Duty Cycle Constraint)

This section is devoted to a proof of Theorem 1 for the case of the idealized duty cycle constraint $(q, 0)$. The conditional probability density function (pdf) of the output given the input of the AWGN channel (1) is

$$p_{Y|X}(y|x) = \phi(y - x) \tag{13}$$

where

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{14}$$

is the standard Gaussian pdf.

With the idealized duty cycle constraint, the capacity of the AWGN channel is achieved by an i.i.d. process and the duty cycle constraint reduces to a per symbol cost constraint. For given input distribution μ, the pdf of the output exists and is expressed as

$$p_Y(y;\mu) = \int p_{Y|X}(y|x)\,\mu(\mathrm{d}x) = \mathsf{E}_\mu\left\{\phi(y - X)\right\}. \tag{15}$$

Denote the relative entropy $D\left(p_{Y|X}(\cdot|x)\|p_Y(\cdot;\mu)\right)$ by $d(x;\mu)$, which is expressed as

$$d(x;\mu) = \int_{-\infty}^{\infty} p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{p_Y(y;\mu)} \mathrm{d}y\,. \tag{16}$$

The mutual information $I(\mu) = I(X; Y)$ is then

$$I(\mu) = \int d(x;\mu)\,\mu(\mathrm{d}x) = \mathsf{E}_\mu\left\{d(X;\mu)\right\}. \tag{17}$$

The capacity of the AWGN channel under per-letter duty cycle constraint and power constraint is evidently given by the supremum of the mutual information $I(\mu)$ where $\mu \in \Lambda(\gamma, q)$. The achievability and converse of this result can be established using standard techniques in information theory.

The proof of property (a) is presented in Section IV-A. Now suppose μ_0 is the unique capacity-achieving distribution, property (b) is established as follows. Since the mirror reflection of μ_0 about 0 is evidently also a maximizer of (5), the uniqueness requires μ_0 to be symmetric. Note that linear scaling of the input to increase its power maintains its duty cycle and cannot reduce the mutual information, as the receiver can add noise to maintain the same SNR. By the uniqueness of the maximizer μ_0, the power constraint must be binding, i.e., the second moment of μ_0 must be equal to γ. In order to prove property (c), we first establish a sufficient and necessary condition for μ_0 in Section IV-B and then apply it to show the discreteness of μ_0 in Section IV-C.

A. Existence and Uniqueness of μ_0

Let $\mathcal{P}$ denote the collection of all Borel probability measures defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which is a topological space with the topology of weak convergence [17]. We first establish the following lemma.

Lemma 1: $\Lambda(\gamma, q)$ is compact in the topological space $\mathcal{P}$.

Proof: According to [17], the topology of weak convergence on $\mathcal{P}$ is metrizable. Therefore, by Prokhorov's theorem [18], in order to prove that $\Lambda(\gamma, q)$ is compact in $\mathcal{P}$, it suffices to show that it is both tight and closed.

For any $\epsilon > 0$, there exits an $a_\epsilon > 0$, such that for all $\mu \in \Lambda(\gamma, q)$,

$$\mu(|X| > a_\epsilon) \le \frac{\mathsf{E}_\mu\left\{X^2\right\}}{a_\epsilon^2} \le \frac{\gamma}{a_\epsilon^2} < \epsilon \tag{18}$$

by Chebyshev's inequality. Choose $K_\epsilon = [-a_\epsilon, a_\epsilon]$, then K_ϵ is compact in $\mathbb{R}$ and $\mu(K_\epsilon) \ge 1 - \epsilon$ for all $\mu \in \Lambda(\gamma, q)$, thus $\Lambda(\gamma, q)$ is tight.

Let $B_m = \left[-\frac{1}{m}, \frac{1}{m}\right]$ for $m = 1, 2, \ldots$. Let $\{\mu_n\}_{n=1}^{\infty}$ be a convergent sequence in $\Lambda(\gamma, q)$ with limit μ_0. Since $\mu_n(B_m) \ge q$ for every m, n, we have [17, Section 3.1]

$$q \le \limsup_{n\to\infty} \mu_n(B_m) \le \mu_0(B_m), \tag{19}$$

and hence

$$\mu_0(\{0\}) = \mu_0\left(\bigcap_{m=1}^{\infty} B_m\right) = \lim_{m\to\infty} \mu_0(B_m) \ge q. \tag{20}$$

Moreover, let $f(x) = x^2$ which is continuous and bounded below. By weak convergence [17, Section 3.1], we have

$$\mathsf{E}_{\mu_0}\left\{X^2\right\} = \int f\,\mathrm{d}\mu_0 \le \liminf_{n\to\infty} \int f\,\mathrm{d}\mu_n \le \gamma. \tag{21}$$

Therefore, $\mu_0 \in \Lambda(\gamma, q)$, i.e., $\Lambda(\gamma, q)$ is closed, and the compactness of $\Lambda(\gamma, q)$ follows. ■

Since the mutual information $I(\mu)$ is continuous on $\mathcal{P}$ [19, Theorem 9], it must achieve its maximum on the compact set $\Lambda(\gamma, q)$. Hence the capacity-achieving distribution μ_0 exists.

According to [19, Corollary 2], the mutual information $I(\mu)$ is strictly concave. It is easy to see that $\Lambda(\gamma, q)$ is convex. Hence the capacity-achieving distribution μ_0 must be unique.

B. Sufficient and Necessary Conditions

Denote the finite-power set as

$$\Lambda(q) = \cup_{0\le\gamma<\infty}\Lambda(\gamma, q). \tag{22}$$

We first establish the following sufficient and necessary condition for the optimal input distribution.

Lemma 2: Let

$$f_\lambda(x;\mu) = d(x;\mu) - I(\mu) - \lambda(x^2 - \gamma). \tag{23}$$

Then $\mu_0 \in \Lambda(\gamma, q)$ achieves the capacity if and only if there exists $\lambda \ge 0$ such that $\lambda\mathsf{E}_{\mu_0}\left\{X^2 - \gamma\right\} = 0$ and $\mathsf{E}_\mu\left\{f_\lambda(X;\mu_0)\right\} \le 0$ for all $\mu \in \Lambda(q)$.

We call $x \in \mathbb{R}$ a point of increase of a measure μ if $\mu(O) > 0$ for every open subset O of $\mathbb{R}$ containing x. Let S_μ

be the set of points of increase of μ. Based on Lemma 2, we establish another sufficient and necessary condition for the optimal input distribution, which will be used to prove Property (c) of Theorem 1 in Section IV-C.

Lemma 3: Let

$$g_\lambda(x;\mu) = q f_\lambda(0;\mu) + (1-q) f_\lambda(x;\mu). \tag{24}$$

Then $\mu_0 \in \Lambda(\gamma, q)$ achieves the capacity if and only if there exists $\lambda \geq 0$ such that for every $x \in \mathbb{R}$,

$$g_\lambda(x;\mu_0) \leq 0\,. \tag{25}$$

Furthermore, $g_\lambda(x;\mu_0) = 0$ for every $x \in S_{\mu_0}\backslash\{0\}$.

In order to prove Lemma 2 and Lemma 3, we first derive some auxiliary technical results as follows. Let $\phi(\cdot)$ defined in (14) be extended to the complex plane $\mathbb{C}$. The relative entropy $d(x;\mu)$ defined in (16) can be extended to the complex plane $\mathbb{C}$ and has the following property:

Lemma 4: For any $\mu \in \Lambda(q)$ and $z \in \mathbb{C}$,

$$d(z;\mu) = \int_{-\infty}^{\infty} \phi(y-z) \log \frac{\phi(y-z)}{p_Y(y;\mu)} \mathrm{d}y \tag{26}$$

is a holomorphic function of z on $\mathbb{C}$. Consequently, $d(x;\mu)$ is a continuous function of x on $\mathbb{R}$.

Proof: It can be shown that $\int_{-\infty}^{\infty} \phi(y-z)\log\phi(y-z)\mathrm{d}y$ is a constant, thus a holomorphic function of z on $\mathbb{C}$. Therefore, it remains to prove that

$$\xi(z) = \int_{-\infty}^{\infty} \phi(y-z) \log p_Y(y;\mu)\mathrm{d}y \tag{27}$$

is a holomorphic function of z on $\mathbb{C}$.

First, by Jensen's inequality, we have

$$p_Y(y;\mu) = \mathsf{E}_\mu\left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-X)^2}{2}}\right\} \tag{28}$$

$$\geq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mathsf{E}_\mu\{(y-X)^2\}} \tag{29}$$

$$= e^{-\frac{1}{2}y^2 - ay - b} \tag{30}$$

where $a = -\mathsf{E}_\mu\{X\}$ and $b = \frac{1}{2}\left(\mathsf{E}_\mu\left\{X^2\right\} + \log(2\pi)\right)$ are real numbers due to the fact that $\mu \in \Lambda(q)$. Thus, $p_Y(y;\mu) \in [e^{-\frac{1}{2}y^2-ay-b}, 1]$, i.e.,

$$|\log P_Y(y;\mu)| \leq \frac{1}{2}y^2 + ay + b. \tag{31}$$

As a result, we have

$$|\phi(y-z)\log p_Y(y;\mu)| \leq \frac{1}{\sqrt{2\pi}}\left|e^{-\frac{(y-z)^2}{2}}\right|\left(\frac{1}{2}y^2 + ay + b\right) \tag{32}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mathrm{Re}(z))^2 - \mathrm{Im}^2(z)}{2}}\left(\frac{1}{2}y^2 + ay + b\right), \tag{33}$$

which is integrable. (Here $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ represent the real and imaginary parts of z, respectively.) It follows that $\xi(z)$ given by (27) exists for any $\mu \in \Lambda(q)$ and $z \in \mathbb{C}$.

Suppose U is an open and bounded subset of $\mathbb{C}$. There exists an $r > 0$ such that $|\mathrm{Re}(z)| \leq r$ and $|\mathrm{Im}(z)| \leq r$ for all $z \in U$. It is easy to check that

$$e^{-\frac{(y-\mathrm{Re}(z))^2}{2}} \leq e^{-\frac{y^2}{2}+|yr|} \tag{34}$$

$$\leq e^{-\frac{y^2}{2}+yr} + e^{-\frac{y^2}{2}-yr} \tag{35}$$

$$= e^{\frac{r^2}{2}}\left[e^{-\frac{1}{2}(y-r)^2} + e^{-\frac{1}{2}(y+r)^2}\right]. \tag{36}$$

Combining (33) and (36) yields that

$$|\phi(y-z)\log p_Y(y;\mu)| \leq \frac{e^{r^2}}{\sqrt{2\pi}}\left[e^{-\frac{1}{2}(y-r)^2} + e^{-\frac{1}{2}(y+r)^2}\right]\left(\frac{1}{2}y^2 + ay + b\right), \tag{37}$$

which is integrable. Therefore, we can say that the integral $\int_{-\infty}^{\infty}\phi(y-z)\log p_Y(y;\mu)\mathrm{d}y$ is uniformly convergent for all $z \in U$. Moreover, $\phi(y-z)\log p_Y(y;\mu)$ is a holomorphic function of z on U for each $y \in \mathbb{R}$. According to the differentiation lemma [20], $\xi(z)$ is a holomorphic function of z on U. It follows that $\xi(z)$ is holomorphic on the whole complex plane $\mathbb{C}$. Lemma 4 is thus established. ■

Let $F(\mu)$ be a real-valued function defined on the convex set $\Lambda(q)$ and $\mu_0 \in \Lambda(q)$. Define the weak derivative of $F(\mu)$ at μ_0 as

$$F'_{\mu_0}(\mu) = \lim_{\theta\to 0^+} \frac{F((1-\theta)\mu_0 + \theta\mu) - F(\mu_0)}{\theta} \tag{38}$$

whenever the limit exists. The following result, which finds its parallel in [7], [10] and [11] gives the weak derivative of the mutual information function $I(\mu)$.

Lemma 5: For $\mu \in \Lambda(q)$, the weak derivative of the mutual information function $I(\mu)$ at μ_0 is

$$I'_{\mu_0}(\mu) = \int d(x;\mu_0)\,\mu(\mathrm{d}x) - I(\mu_0). \tag{39}$$

Proof: Define $\mu_\theta = (1-\theta)\mu_0 + \theta\mu$ for all $\theta \in (0,1]$. It can be shown that

$$\frac{1}{\theta}\left(I(\mu_\theta) - I(\mu_0)\right) = \frac{1}{\theta}\int \left(d(x;\mu_\theta) - d(x;\mu_0)\right)\,\mu_\theta(\mathrm{d}x) + \frac{1}{\theta}\left(\int d(x;\mu_0)\,\mu_\theta(\mathrm{d}x) - I(\mu_0)\right) \tag{40}$$

$$= -\frac{1}{\theta}\int_{-\infty}^{\infty} p_Y(y;\mu_\theta)\log\frac{p_Y(y;\mu_\theta)}{p_Y(y;\mu_0)}\mathrm{d}y + \int d(x;\mu_0)\,\mu(\mathrm{d}x) - I(\mu_0). \tag{41}$$

Therefore, it suffices to show that

$$\lim_{\theta\to 0^+}\int_{-\infty}^{\infty}\frac{1}{\theta}p_Y(y;\mu_\theta)\log\frac{p_Y(y;\mu_\theta)}{p_Y(y;\mu_0)}\mathrm{d}y = 0. \tag{42}$$

In the remainder of this proof, we find a function independent of θ that dominates the integrand so that dominated convergence theorem can be used to establish (42) by exchanging the order of the limit and the integral therein.

Lemma 6: Let $\theta, a, b \in (0,1]$. Define

$$f(\theta) = \frac{(1-\theta)a + \theta b}{\theta}\log\frac{(1-\theta)a + \theta b}{a}\,, \tag{43}$$

then

$$|f(\theta)| \leq b + a - b\log b - b\log a\,. \tag{44}$$

Proof: It is easy to check that $f(1) = b\log\frac{b}{a}$, $f(0^+) = b - a$ and

$$f'(\theta) = \frac{b-a}{\theta} - \frac{a}{\theta^2}\log\left(1 - \theta + \frac{b}{a}\theta\right). \tag{45}$$

Define $g(\theta) = \theta(b-a) - a\log\left(1-\theta+\frac{b}{a}\theta\right)$ for $\theta \in (0,1]$, then we have

$$g'(\theta) = \frac{\theta(b-a)^2}{(1-\theta)a + \theta b} \geq 0. \tag{46}$$

Since $g(0^+) = 0$, $g(\theta) \geq 0$ for all $\theta \in (0,1]$. According to (45), we have $f'(\theta) = \frac{g(\theta)}{\theta^2} \geq 0$. It follows that for all $\theta \in (0,1]$,

$$b - a = f(0^+) \leq f(\theta) \leq f(1) = b\log\frac{b}{a}, \tag{47}$$

and hence

$$|f(\theta)| \leq \max\left\{|b-a|, \left|b\log\frac{b}{a}\right|\right\} \tag{48}$$

$$\leq b + a - b\log b - b\log a. \tag{49}$$

Lemma 6 is thus established. ■

Applying Lemma 6 with $a = p_Y(y;\mu_0)$ and $b = p_Y(y;\mu)$, we have

$$\left|\frac{1}{\theta}p_Y(y;\mu_\theta)\log\frac{p_Y(y;\mu_\theta)}{p_Y(y;\mu_0)}\right| \leq p_Y(y;\mu) + p_Y(y;\mu_0) - p_Y(y;\mu)\log p_Y(y;\mu) - p_Y(y;\mu)\log p_Y(y;\mu_0) \tag{50}$$

where the right hand side is an integrable function of y by the result that $-\int_{-\infty}^{\infty} p_Y(y;\mu_2)\log p_Y(y;\mu_1)\mathrm{d}y < \infty$ for any $\mu_1, \mu_2 \in \Lambda(q)$. In fact, as in the proof of Lemma 4 (see (31)), there exist $a, b \in \mathbb{R}$ such that $|\log p_Y(y;\mu_1)| \leq \frac{1}{2}y^2 + ay + b$. Therefore,

$$\int_{-\infty}^{\infty} |p_Y(y;\mu_2)\log p_Y(y;\mu_1)|\mathrm{d}y$$

$$\leq \int_{-\infty}^{\infty} p_Y(y;\mu_2)\left(\frac{1}{2}y^2 + ay + b\right)\mathrm{d}y \tag{51}$$

$$= \frac{1}{2}\mathsf{E}_{\mu_2}\left\{X^2\right\} + a\mathsf{E}_{\mu_2}\{X\} + b + \frac{1}{2} \tag{52}$$

$$< \infty \tag{53}$$

due to the assumption that $\mu_2 \in \Lambda(q)$.

Therefore, the dominated convergence theorem provides that

$$\lim_{\theta\to 0^+}\frac{1}{\theta}\int_{-\infty}^{\infty} p_Y(y;\mu_\theta)\log\frac{p_Y(y;\mu_\theta)}{p_Y(y;\mu_0)}\mathrm{d}y$$

$$= \int_{-\infty}^{\infty}\lim_{\theta\to 0^+}\frac{1}{\theta}p_Y(y;\mu_\theta)\log\frac{p_Y(y;\mu_\theta)}{p_Y(y;\mu_0)}\mathrm{d}y \tag{54}$$

$$= \int_{-\infty}^{\infty}\left(p_Y(y;\mu) - p_Y(y;\mu_0)\right)\mathrm{d}y \tag{55}$$

$$= 0. \tag{56}$$

Lemma 5 is thus proved. ■

We next prove Lemmas 2 and 3.

Proof of Lemma 2: Define the Lagrangian

$$J(\mu) = I(\mu) - \lambda\mathsf{E}_\mu\left\{X^2 - \gamma\right\} \tag{57}$$

where λ is the Lagrange multiplier. Since $\Lambda(q)$ defined in (22) is a convex set and $I(\mu) < \infty$ on $\Lambda(q)$, μ_0 is capacity-achieving if and only if there exists $\lambda \geq 0$ such that the following conditions hold [21]:

(i) $\lambda\mathsf{E}_{\mu_0}\left\{X^2 - \gamma\right\} = 0$;
(ii) for all $\mu \in \Lambda(q)$, $J(\mu_0) \geq J(\mu)$.

Due to concavity of $I(\mu)$, $J(\mu)$ is also concave. Condition (ii) is then equivalent to that the weak derivative $J'_{\mu_0}(\mu) \leq 0$ for all $\mu \in \Lambda(q)$.

By Lemma 5, the linearity of $\mathsf{E}_\mu\left\{X^2 - \gamma\right\}$ with respect to (w.r.t.) μ and Condition (i), $J'_{\mu_0}(\mu)$ can be easily calculated as

$$J'_{\mu_0}(\mu) = \mathsf{E}_\mu\left\{f_\lambda(X;\mu_0)\right\}. \tag{58}$$

Therefore, Condition (ii) is equivalent to $\mathsf{E}_\mu\{f_\lambda(X;\mu_0)\} \leq 0$ for all $\mu \in \Lambda(q)$. Thus Lemma 2 follows. ■

Up to this point, the results in Lemmas 2, 4 and 5 hold in general with $\Lambda(q)$ replaced by any set of distributions with finite power. The duty cycle constraint is only used in the proof of Lemma 3 in the following.

Proof of Lemma 3: The necessity part is shown as follows. Suppose μ_0 achieves the capacity, then by Lemma 2, there exists $\lambda \geq 0$ such that $\lambda\mathsf{E}_{\mu_0}\left\{X^2 - \gamma\right\} = 0$ and $\mathsf{E}_\mu\{f_\lambda(X;\mu_0)\} \leq 0$ for all $\mu \in \Lambda(q)$. For any $x \in \mathbb{R}\backslash\{0\}$, choose μ such that $\mu(\{0\}) = q$ and $\mu(\{x\}) = 1-q$, so by the fact that $\mu \in \Lambda(q)$, we have

$$0 \geq \mathsf{E}_\mu\{f_\lambda(X;\mu_0)\} = qf_\lambda(0;\mu_0) + (1-q)f_\lambda(x;\mu_0). \tag{59}$$

Due to the continuity of $d(x;\mu_0)$ by Lemma 4, $f_\lambda(x;\mu_0)$ is also continuous so that (59) holds for all $x \in \mathbb{R}$, i.e., $g_\lambda(x;\mu_0) \leq 0$ for every $x \in \mathbb{R}$.

To finish proving the necessity, it suffices to show that $g_\lambda(x;\mu_0) = 0$ for all $x \in S_{\mu_0}\backslash\{0\}$. Evidently, $g_\lambda(0;\mu_0) = f_\lambda(0;\mu_0)$ and by (17) and $\lambda\mathsf{E}_{\mu_0}\left\{X^2-\gamma\right\} = 0$,

$$\int f_\lambda(x;\mu_0)\,\mu_0(\mathrm{d}x) = 0\,. \tag{60}$$

Hence,

$$\int_{\mathbb{R}\backslash\{0\}} g_\lambda(x;\mu_0)\,\mu_0(\mathrm{d}x)$$

$$= \int g_\lambda(x;\mu_0)\,\mu_0(\mathrm{d}x) - g_\lambda(0;\mu_0)\mu_0(\{0\}) \tag{61}$$

$$\geq qf_\lambda(0;\mu_0) + (1-q)\int f_\lambda(x;\mu_0)\,\mu_0(\mathrm{d}x) - qf_\lambda(0;\mu_0) \tag{62}$$

$$= 0. \tag{63}$$

Since $g_\lambda(x;\mu_0) \leq 0$ for every $x \in \mathbb{R}$, (63) implies that on $\mathbb{R}\backslash\{0\}$, $g_\lambda(x;\mu_0) = 0$ μ_0-almost surely, so that $g_\lambda(x;\mu_0) = 0$ for all $x \in S_{\mu_0}\backslash\{0\}$ follows immediately.

The sufficiency part of Lemma 3 is established as follows. Suppose $g_\lambda(x;\mu_0) \leq 0$ for every $x \in \mathbb{R}$. By integrating

$g_\lambda(x;\mu_0)$ w.r.t. μ_0, we have

$$qg_\lambda(0;\mu_0) \geq \int g_\lambda(x;\mu_0)\,\mu_0(\mathrm{d}x) \quad (64)$$
$$= qg_\lambda(0;\mu_0) - (1-q)\lambda E_{\mu_0}\{X^2-\gamma\} \quad (65)$$
$$\geq qg_\lambda(0;\mu_0) \quad (66)$$

where (65) is due to (17) and $g_\lambda(0;\mu_0) = f_\lambda(0;\mu_0)$, and (66) follows from $\mathsf{E}_{\mu_0}\{X^2\} \leq \gamma$ since $\mu_0 \in \Lambda(\gamma, q)$. Hence, $\lambda\mathsf{E}_{\mu_0}\{X^2-\gamma\} = 0$ due to the fact that $q < 1$. Furthermore, for any $\mu \in \Lambda(q)$, by integrating $g_\lambda(x;\mu_0)$ w.r.t. μ, we have

$$qg_\lambda(0;\mu_0) \geq \int g_\lambda(x;\mu_0)\,\mu(\mathrm{d}x) \quad (67)$$
$$= qf_\lambda(0;\mu_0) + (1-q)\mathsf{E}_\mu\{f_\lambda(x;\mu_0)\}. \quad (68)$$

Because $g_\lambda(0;\mu_0) = f_\lambda(0;\mu_0)$ and $q < 1$, we have $\mathsf{E}_\mu\{f_\lambda(X;\mu_0)\} \leq 0$. Together with $\lambda\mathsf{E}_{\mu_0}\{X^2-\gamma\} = 0$ and Lemma 2, this implies that μ_0 must be capacity-achieving. ■

C. Discreteness of μ_0

With Lemma 3 established, we now prove Property (c) in Theorem 1.

Let $\lambda \geq 0$ satisfy condition (25) and $d(z;\mu)$ be defined in (26). We extend functions $f_\lambda(x;\mu)$ in Lemma 2 and $g_\lambda(x;\mu)$ in Lemma 3 to be defined on the whole complex plane $\mathbb{C}$ as (23) and (24), respectively, with x replaced by $z \in \mathbb{C}$. By Lemma 4, $d(z;\mu)$ is a holomorphic function of z on $\mathbb{C}$, hence so is $g_\lambda(z;\mu)$. According to Lemma 3, each element in the set $S_{\mu_0}\backslash\{0\}$ is a zero of the function $g_\lambda(z;\mu_0)$.

Next we show that for any bounded interval L of $\mathbb{R}$, $S_{\mu_0} \bigcap L$ is a finite set. Suppose, to the contrary, $S_{\mu_0} \bigcap L$ is infinite, then it has a limit point in $\mathbb{R}$ by the Bolzano-Weierstrass Theorem [20] and hence, $g_\lambda(z;\mu_0) = 0$ on the whole complex plane $\mathbb{C}$ by the Identity Theorem [22]. Then, by (16), (23) and (24), for every $x \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} \phi(y-x)r(y)\mathrm{d}y = 0 \quad (69)$$

where

$$r(y) = \log p_Y(y;\mu_0) + \lambda y^2 + c \quad (70)$$

and $c = \frac{1}{2}\log(2\pi e) + \frac{1}{1-q}\left(I(\mu_0) - qd(0;\mu_0) - \lambda\gamma\right) - \lambda$ is a constant.

As in the proof of Lemma 4, there exist $a, b \in \mathbb{R}$ such that $|\log p_Y(y;\mu_0)| \leq \frac{1}{2}y^2 + ay + b$. As a result, there exist some $\alpha, \beta > 0$ such that $|r(y)| \leq \alpha y^2 + \beta$. Since the convolution of $r(y)$ and the Gaussian density is equal to the zero function by (69), $r(y)$ must be the zero function according to [11, Corollary 9]. This requires the capacity-achieving output distribution $p_Y(y;\mu_0)$ be Gaussian, which cannot be true unless X is Gaussian, which contradicts the assumption that X has a probability mass at 0. Therefore, $S_{\mu_0} \bigcap L$ must be a finite set for any bounded interval L, which further implies that S_{μ_0} is at most countable.

Finally, we show that S_{μ_0} is countably infinite. Suppose, to the contrary, $S_{\mu_0} = \{x_i\}_{i=1}^M$ is a finite set with $\mu_0(\{x_i\}) = p_i$ and $|x_i| \leq B_1$ for all $i = 1, 2, \ldots, N$. For any $y > B_1$,

$$p_Y(y;\mu_0) = \sum_{i=1}^{M} p_i\phi(y-x_i) \leq e^{-\frac{(y-B_1)^2}{2}}. \quad (71)$$

For any $\epsilon > 0$, choose $B_2 > 0$ such that $\int_{-B_2}^{B_2}\phi(x)\mathrm{d}x > 1-\epsilon$. By (16), (23), (24) and (25), for any $x > B_1 + B_2$, we have

$$0 \geq -\int_{-\infty}^{\infty} \phi(y-x)\log p_Y(y;\mu_0)\mathrm{d}y - \lambda x^2 - (c+\lambda) \quad (72)$$
$$\geq \int_{x-B_2}^{x+B_2} \phi(y-x)\frac{1}{2}(y-B_1)^2\mathrm{d}y - \lambda x^2 - (c+\lambda) \quad (73)$$
$$= \int_{-B_2}^{B_2} \phi(t)\frac{1}{2}(x-B_1+t)^2\mathrm{d}t - \lambda x^2 - (c+\lambda) \quad (74)$$
$$\geq \frac{1}{2}(x-B_1)^2(1-\epsilon) - \lambda x^2 - (c+\lambda). \quad (75)$$

For (75) to hold for large x, λ must satisfy $\lambda \geq \frac{1}{2}$.

To finish the proof, it suffices to show that $\lambda < 1/2$ for any $\gamma > 0$, so that contradiction arises, which implies that S_{μ_0} must be countably infinite. For fixed $q \in (0,1)$, denote the Lagrange multiplier in (25) as $\lambda(\gamma)$. Denote $C_G(\gamma) = 1/2\log(1+\gamma)$, which is the channel capacity of a Gaussian channel with the average power constraint only. By the envelope theorem [21], $\lambda(\gamma)$ is the derivative of $C(\gamma, q)$ w.r.t. γ. Since $C(0,q) = C_G(0) = 0$ and the derivative of $C_G(\gamma)$ at $\gamma = 0$ is 1/2, we have $\lambda(0) \leq 1/2$, otherwise we could find a small enough γ such that $C(\gamma, q)$ would exceed $C_G(\gamma)$ which is obviously impossible. Next we show that $C(\gamma, q)$ is strictly concave for $\gamma \geq 0$. Suppose μ_1 and μ_2 are the capacity-achieving input distributions of (5) for different power constraints γ_1 and γ_2, respectively. Due to Property (b) in Theorem 1, μ_1 and μ_2 must be different. Define $\mu_\theta = \theta\mu_1 + (1-\theta)\mu_2$ for $\theta \in (0,1)$. It is easy to see that μ_θ satisfies that the duty cycle is no greater than $1-q$ and the average input power is no greater than $\theta\gamma_1 + (1-\theta)\gamma_2$. Now we have

$$C(\theta\gamma_1 + (1-\theta)\gamma_2, q) \geq I(\mu_\theta) \quad (76)$$
$$> \theta I(\mu_1) + (1-\theta)I(\mu_2) \quad (77)$$
$$= \theta C(\gamma_1, q) + (1-\theta)C(\gamma_2, q), \quad (78)$$

where (77) is due to the strict concavity of $I(\mu)$. Therefore, the strict concavity of $C(\gamma, q)$ for $\gamma \geq 0$ follows, which implies that $\lambda(\gamma) < \lambda(0) \leq 1/2$ for all $\gamma > 0$.

V. Proof of Theorem 2 (The Case of Realistic Duty Cycle Constraint)

A. Stationarity of the Capacity-Achieving Input Distribution

We first establish the fact that a stationary distribution achieves the capacity of the AWGN channel (1) with the realistic duty cycle constraint and power constraint.

Proposition 1: A stationary distribution[1] achieves

$$\max_{\mu \in \Lambda^n(\gamma,q,c)} I(X^n; Y^n). \quad (79)$$

[1]The stationarity of distribution ν on X^n satisfies

$$\nu_{X_s,\cdots,X_t} = \nu_{X_{s+k},\cdots,X_{t+k}}$$

for any index s, t, k satisfied

$$1 \leq s \leq t \leq n \quad 1 \leq s+k \leq t+k \leq n$$

Proof: Let $T_k(\cdot)$ as a k-cyclic-shift operator on $\mu \in \Lambda^n(\gamma, q, c)$, defined as

$$T_k(\mu) = \mu_{X_{k+1},\cdots,X_n,X_1,\cdots X_k} \tag{80}$$

where $k = 1, \cdots, n-1$, and specifically $T_0(\mu) = \mu$. For any distribution μ in $\Lambda^n(\gamma, q, c)$, a distribution ν on X^n can be defined as

$$\nu = \frac{1}{n}\sum_{k=0}^{n-1} T_k(\mu). \tag{81}$$

According the concavity of the mutual information $I(\cdot)$,

$$\begin{aligned} I(\nu) &= I\left(\frac{1}{n}\sum_{k=0}^{n-1} T_k(\mu)\right) && (82)\\ &\geq \frac{1}{n}\sum_{k=0}^{n-1} I(T_k(\mu)) && (83)\\ &= I(\mu) && (84)\end{aligned}$$

where $I(T_k(\mu)) = I(\mu)$ since the AWGN channel (1) is memoryless and time-invariant. Obviously ν is a stationary distribution and satisfies the duty cycle constraint and power constraint, i.e., $\nu \in \Lambda^n(\gamma, q, c)$, hence Proposition 1 is established. ■

According to Proposition 1, for any n, $I(X^n; Y^n)$ is maximized by a stationary distribution. Therefore as $n \to \infty$, the capacity in (9) is achieved by a stationary input distribution.

B. Input-output Mutual Information

Proposition 2: Let the input follow a stationary distribution $\mu \in \Lambda(\gamma, q, c)$. The limit of the input-output mutual information per symbol as a function of μ can be expressed as

$$I(\mu) = I(X; Y) - h(Y) + h(\mathscr{Y}) \tag{85}$$

where $I(X; Y)$ is the mutual information of the AWGN channel between the input X, which follows distribution μ_{X_1} and the corresponding output Y, $h(Y)$ is the differential entropy of Y and $h(\mathscr{Y})$ is the differential entropy rate of output process $\{Y_i\}$.

Proof: The mutual information between X^n and Y^n can be expressed using relative entropies

$$\begin{aligned} I(X^n; Y^n) &= D(P_{Y^n|X^n} \| P_{Y^n} | P_{X^n}) && (86)\\ &= D(P_{Y^n|X^n} \| P_{Y_1} \times \cdots \times P_{Y_n} | P_{X^n})\\ &\quad - D(P_{Y^n} \| P_{Y_1} \times \cdots \times P_{Y_n}) && (87)\\ &= \sum_{k=1}^{n} D(P_{Y_k|X_k} \| P_{Y_k} | P_{X^n})\\ &\quad - \mathsf{E}\left\{\log P_{Y^n}(Y^n) - \sum_{i=1}^{n} \log P_{Y_i}(Y_i)\right\} && (88)\\ &= nI(X; Y) - nh(Y) + h(Y^n). && (89)\end{aligned}$$

Then

$$\begin{aligned} I(\mu) &= \lim_{n\to\infty} \frac{1}{n} I(X^n; Y^n) && (90)\\ &= I(X; Y) - h(Y) + \lim_{n\to\infty} \frac{1}{n} h(Y^n) && (91)\\ &= I(X; Y) - h(Y) + h(\mathscr{Y}). && (92)\end{aligned}$$

Proposition 2 is established. ■

When the input is an i.i.d. random process, the output process is also i.i.d., $h(Y) = h(\mathscr{Y})$. This implies the following corollary.

Corollary 1: Among all i.i.d. distributions, the one that maximizes the mutual information under duty cycle constraint (q, c) and average power constraint γ can be solved from the following optimization:

$$\begin{aligned} &\underset{P_X}{\text{maximize}}\ I(X; Y) && (93)\\ &\text{subject to}\ P_X(0) - 2cP_X(0)(1 - P_X(0)) \geq q, && (94)\\ &\qquad \mathsf{E}\left\{X^2\right\} \leq \gamma. && (95)\end{aligned}$$

In the special case of no transition cost, i.e., $c = 0$, the result of (93) is equal to that of (5).

C. Proof of Theorem 2

The mutual information expressed by (85) is hard to optimize, even if the input is restricted to Markov processes. To simplify the matter, we introduce a lower bound of $I(\mu)$, which is given by $L(\mu)$ in (11). In order to finish the proof of Theorem 2, we separate Property (b) into three parts: 1) the achievability part, i.e., the maximum of $L(\cdot)$ is achieved by some input distribution in $\Lambda(\gamma, q, c)$, denoted by μ^*; 2) the Markov part, i.e., μ^* is a first-order Markov process; and 3) the discrete part, i.e., the stationary distribution of μ^* is discrete. In the following, the proof of Property (a) is first presented, then, assuming the existence of the maximizer μ^*, we show that it must be a discrete first-order Markov chain with Property (c), and finally the existence of μ^* is established to complete the proof of Theorem 2.

Property (a): Using the fact that processing reduces relative entropy and μ is specified as a stationary probability distribution, we have

$$\begin{aligned} &\frac{1}{n} D(P_{Y^n} \| P_{Y_1} \times P_{Y_2} \times \cdots \times P_{Y_n})\\ &\quad\leq \frac{1}{n} D(P_{X^n} \| P_{X_1} \times P_{X_2} \times \cdots \times P_{X_n}) && (96)\\ &\quad= \frac{1}{n}\sum_{k=1}^{n-1} D(P_{X_k|X_{k+1}^n} \| P_{X_k} | P_{X_{k+1}^n}) && (97)\\ &\quad= \frac{1}{n}\sum_{k=2}^{n} I(X_1; X_2^k). && (98)\end{aligned}$$

Therefore

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P_{Y_1} \times P_{Y_2} \times \cdots \times P_{Y_n}) \leq I(X_1; X_2^{\infty}) \tag{99}$$

using the fact that the Cesáro mean of sequence $I(X_1, X_2^k)$ is $I(X_1; X_2^{\infty})$. Applying (85), (87) and (99),

$$L(\mu) = I(X; Y) - I(X_1; X_2^{\infty}) \leq I(\mu) \leq C(\gamma, q, c). \tag{100}$$

Thus Property (a) is established.

In the following, we assume that the maximum of the lower bound $L(\cdot)$ is achieved, the proof of which is deferred to the end of this section.

The Markov part in Property (b): For any $\mu \in \Lambda(\gamma, q, c)$, which is not Markov in general, its first-order Markov counterpart ν is defined by

$$\nu_{X_1,\cdots,X_n} = \mu_{X_1}\mu_{X_2|X_1}\mu_{X_3|X_2}\cdots\mu_{X_n|X_{n-1}}. \tag{101}$$

Evidently, ν and μ have identical marginal distributions: $\nu_{X_i} = \mu_{X_i}$, and also identical joint distributions of all consecutive pairs: $\nu_{X_i,X_{i+1}} = \mu_{X_i,X_{i+1}}$. Therefore

$$\nu_{X_i}(\{0\}) = \mu_{X_i}(\{0\}) \tag{102}$$

and

$$\nu_{X_i,X_{i+1}}(\{x_i = 0, x_{i+1} \neq 0\}) = \mu_{X_i,X_{i+1}}(\{x_i = 0, x_{i+1} \neq 0\}). \tag{103}$$

Since $\mu \in \Lambda(\gamma, q, c)$, we have $\nu \in \Lambda(\gamma, q, c)$. Let $\{X_i\}$ follow distribution μ and $\{Z_i\}$ follow distribution ν. Then

$$\begin{aligned} I(Z_1; Z_2^\infty) &= I(Z_1; Z_2) + I(Z_1; Z_3^\infty|Z_2) && (104)\\ &= I(Z_1; Z_2) && (105)\\ &= I(X_1; X_2) && (106)\\ &\leq I(X_1; X_2^\infty) && (107) \end{aligned}$$

where equality holds if and only if $\{X_i\}$ is a first-order Markov process. By (11) and (107), $L(\nu) \geq L(\mu)$. So for any μ which maximizes $L(\mu)$, ν can be generated from μ by (101) with $L(\nu) \geq L(\mu)$. Therefore, the maximum of $L(\mu)$ can be achieved by a first-order Markov process.

Property (c): Suppose ν is a stationary first-order Markov process, and it can be sufficiently denoted as $\nu = \{\mathcal{X}, P_{X_2|X_1}\}$, where $\mathcal{X}$ is the state space of ν and $P_{X_2|X_1}$ is the transition probability distribution. Define a new first-order Markov process $\bar{\nu}$ from ν as follows.

Definition 1: Let $\bar{\nu}$, defined on the same state space $\mathcal{X}$ as ν, be a first-order Markov process denoted by $(\mathcal{X}, P_{Z_2|Z_1})$, where

$$P_{Z_2|Z_1}(z_2|z_1) = \begin{cases} \alpha & z_1 = 0\ z_2 = 0,\\ 1-\beta & z_1 \neq 0\ z_2 = 0,\\ \dfrac{1-\alpha}{\eta}P_X(z_2) & z_1 = 0\ z_2 \neq 0,\\ \dfrac{\beta}{\eta}P_X(z_2) & z_1 \neq 0\ z_2 \neq 0, \end{cases} \tag{108}$$

where

$$S_1 = \mathcal{X} \setminus \{0\} \tag{109}$$

and

$$\begin{aligned} \alpha &= P_{X_2|X_1}(0|0) && (110)\\ \beta &= P(X_2 \in S_1|X_1 \in S_1) && (111)\\ \eta &= P(X \in S_1). && (112) \end{aligned}$$

The process $\bar{\nu}$ is described by $(\mathcal{X}, \alpha, \beta, P_X)$, where P_X is the stationary distribution of ν. It is easy to prove that the stationary distribution P_Z of $\bar{\nu}$ is equal to P_X. Moreover, $\bar{\nu}$ satisfies the same power and duty cycle constraint ν satisfies, i.e., $\bar{\nu} \in \Lambda(\gamma, q, c)$. Furthermore, let $B_i = 1_{\{Z_i \neq 0\}}$, then

$$\begin{aligned} P_{B_2|B_1}(0|0) &= \alpha && (113)\\ P_{B_2|B_1}(1|1) &= \beta. && (114) \end{aligned}$$

From (108) to (114) we have,

$$P_{Z_2|Z_1}(z_2|z_1) = P_{B_2|B_1}(b_2|b_1)\frac{P_{Z_2}(z_2)}{P_{B_2}(b_2)}, \tag{115}$$

so Z_i and Z_{i+1} are independent given B_i and B_{i+1}.

Based on (115), it is easy to see that

$$\begin{aligned} I(Z_1; Z_2) &= \mathsf{E}\left\{\log\frac{P_{Z_2|Z_1}(z_2|z_1)}{P_{Z_2}(z_2)}\right\} && (116)\\ &= \mathsf{E}\left\{\log\frac{P_{B_2|B_1}(b_2|b_1)}{P_{B_2}(b_2)}\right\} && (117)\\ &= I(B_1; B_2) && (118)\\ &\leq I(X_1; X_2). && (119) \end{aligned}$$

The inequality in (119) follows since $X_1 \to X_2 \to B_2$ forms a Markov chain then $I(X_1; B_2) \leq I(X_1; X_2)$ [23] and $B_2 \to X_1 \to B_1$ also forms a Markov chain then $I(B_2; B_1) \leq I(B_2; X_1)$.

Therefore, by (11), we have $L(\bar{\nu}) \geq L(\nu)$. So for any ν which maximizes $L(\nu)$, $\bar{\nu}$ can be generated from ν by (108) with $L(\bar{\nu}) \geq L(\nu)$. Therefore, the maximum of $L(\mu)$ can be achieved by a first-order Markov process with Property (c).

The discreteness part in Property (b): The discreteness of the optimized input distribution which maximizes $L(\cdot)$ is proved in the following. According to the Markov part in Property (b) and Property (c), the lower bound $L(\cdot)$ is maximized by a first-order Markov process, the transition probability distribution of which $P_{X_2|X_1}$ can be expressed as

$$P_{X_2|X_1}(x_2|x_1) = P_{B_2|B_1}(b_2|b_1)\frac{P_X(x_2)}{P_{B_2}(b_2)} \tag{120}$$

where $B_i = 1_{\{X_i \neq 0\}}$. Then the maximum of $L(\mu)$ can be achieved by the following optimization

$$\begin{aligned} &\underset{q_0}{\text{maximize}}\ I_X(q_0) - I_B(q_0) && (121)\\ &\text{subject to}\ I_X(q_0) = \max_{P_X} I(X; Y) && (122)\\ &I_B(q_0) = \min_{P(B_2|B_1)} I(B_1; B_2) && (123)\\ &P_X(0) = P_{B_1}(0) = P_{B_2}(0) = q_0 && (124)\\ &q_0 - 2cq_0P_{B_2|B_1}(1|0) \geq q. && (125) \end{aligned}$$

Since given any $q_0 \geq q > 0$, $I_X(q_0) - I_B(q_0)$ can be maximized by the maximum of $I_X(q_0)$ and the minimum of $I_B(q_0)$, respectively. Suppose the maximum of (121) is achieved by $q_0 = q_0^*$. Obviously if $q_0^* = 1$, the input distribution has only one probability mass at zero which is discrete. In the following, we consider the case $q_0^* < 1$. The maximum of (121) must be achieved by P_X, which maximizes $I(X; Y)$ for $q_0 = q_0^*$. Therefore, the maximization in (122) is similar to the problem in Theorem 1. The difference to Theorem 1 is that the distribution P_X in (122) satisfies $P_X(0) = q_0^*$, whereas in Theorem 1 the distribution P_X satisfies $P_X(0) \geq q$. Define

$$\Lambda_0(\gamma, q_0^*) = \left\{\mu : \mu(\{0\}) = q_0^*,\ \mathsf{E}_\mu\left\{X^2\right\} \leq \gamma\right\} \tag{126}$$

where μ is the marginal input distribution of the first-order Markov process. We can establish the following lemma.

Lemma 7: $\Lambda_0(\gamma, q_0^*)$ is compact in the topological space $\mathcal{P}$.

Proof: As mentioned in Lemma 1, the topology of weak convergence on $\mathcal{P}$ is metrizable with the Lévy-Prohorov metric [17] and defined as

$$\pi(\mu, \nu) = \inf\left\{\delta : \mu(A) \le \nu(A^{(\delta)}) + \delta \text{ and } \nu(A) \le \mu(A^{(\delta)}) + \delta \text{ for all } A \subseteq \mathcal{B}\right\} \quad (127)$$

for any $\mu, \nu \in \mathcal{P}$, where

$$A^{(\delta)} = \begin{cases} \emptyset, & A = \emptyset; \\ \{x : \inf_{a \in A} |x - a| < \delta\}, & \text{otherwise.} \end{cases} \quad (128)$$

Similarly as in the proof of Lemma 1, it suffices to show that $\Lambda_0(\gamma, q_0^*)$ is both tight and closed in $\mathcal{P}$. The tightness can be shown by the same arguments as in Lemma 1. In the following, we prove that $\Lambda_0(\gamma, q_0^*)$ is closed in $\mathcal{P}$.

Let $B_m = [-1/m, 1/m]$ for $m = 1, 2, \ldots$. Let $\{\mu_n\}_{n=1}^{\infty}$ be a convergent sequence in $\Lambda_0(\gamma, q_0^*)$ with limit μ_0. For any $m \in \mathbb{N}$, there exists an $n(m)$ such that $\pi(\mu_n, \mu_0) < 1/m$ for all $n > n(m)$. By the definition of π in (127), we have for any $m \in \mathbb{N}$ and $n > n(m)$,

$$\mu_0(\{0\}) \le \mu_n(B_m) + \frac{1}{m}, \quad (129)$$

and

$$\mu_n(\{0\}) \le \mu_0(B_m) + \frac{1}{m}. \quad (130)$$

For any $n \in \mathbb{N} \bigcup \{0\}$, we have

$$\mu_n(\{0\}) = \mu_n\left(\bigcap_{m=1}^{\infty} B_m\right) = \lim_{m \to \infty} \mu_n(B_m), \quad (131)$$

so for any $m \in \mathbb{N}$, there exists an $n'(m)$ such that $\mu_n(B_m) \le \mu_n(\{0\}) + \frac{1}{m}$ for all $n > n'(m)$. Therefore, according to (129), (130) and the fact that $\mu_n(\{0\}) = q_0^*$ for $n \in \mathbb{N}$, for all $m \in \mathbb{N}$ and $n > \max\{n(m), n'(m)\}$,

$$q_0^* - \frac{2}{m} \le \mu_0(\{0\}) \le q_0^* + \frac{2}{m}. \quad (132)$$

Thus we have $\mu_0(\{0\}) = q_0^*$ by letting $m \to \infty$.

Moreover, let $f(x) = x^2$ which is continuous and bounded below. By weak convergence [17, Section 3.1], we have

$$\mathsf{E}_{\mu_0}\left\{X^2\right\} = \int f \mathrm{d}\mu_0 \le \liminf_{n \to \infty} \int f \mathrm{d}\mu_n \le \gamma. \quad (133)$$

Together with $\mu_0(\{0\}) = q_0^*$, we have $\mu_0 \in \Lambda_0(\gamma, q_0^*)$, i.e., $\Lambda_0(\gamma, q_0^*)$ is closed, and the compactness of $\Lambda_0(\gamma, q_0^*)$ follows. ∎

Now P_X can be proved to be discrete by following the same development as in the proof of Theorem 1 with Lemma 1 substituted by Lemma 7. Because P_X is the stationary distribution of the Markov process, the maximum of the lower bound $L(\cdot)$ is achieved by a discrete first-order Markov process.

The achievability part of Property (b): As shown above, the maximum of $L(\mu)$ is achieved by the optimization problem in (121) to (125). In the following, we prove that $I_X(q_0)$ in (122) and $I_B(q_0)$ in (123) are continuous functions of q_0 on closed interval $[q, 1]$.

First we show that $I_X(q_0)$ is continuous on $[q, 1]$. As in Section IV, define the mutual information $I(\mu) = I(X; Y)$ for all $\mu \in \mathcal{P}$. For fixed γ, we rewrite $\Lambda_0(\gamma, q_0)$ in (126) as $\Gamma(q_0)$, which is a compact-valued correspondence from $[q, 1]$ to $\mathcal{P}$ according to Lemma 7. Since $I(\mu)$ is continuous on $\mathcal{P}$ [19, Theorem 9], $I(\mu)$ is also a continuous function defined on $[q, 1] \times \mathcal{P}$. In order to apply the maximum theorem [24, Theorem 9.14] to establish the continuity of $I_X(q_0)$, it suffices to show that the correspondence Γ is continuous (i.e., both upper and lower hemicontinuous) on $[q, 1]$.

We first prove that Γ is upper hemicontinuous. To this end, assume that $q_0^n \to q_0^0$ holds in $[q, 1]$ and $\mu_n \in \Gamma(q_0^n)$ for each $n = 1, 2, \ldots$. It suffices to show that there exists a subsequence $\{\mu_{n_k}\}_{k=1}^{\infty}$ of $\{\mu_n\}_{n=1}^{\infty}$ that converges to some point $\mu_0 \in \Gamma(q_0^0)$ [24, Proposition 9.8]. Since $\mu_n \in \left\{\mu : \mathsf{E}_{\mu}\left\{X^2\right\} \le \gamma\right\}$ which is compact in $\mathcal{P}$ by similar proof as in Lemma 1, it has a convergent subsequence $\{\mu_{n_k}\}_{k=1}^{\infty}$ with limit μ_0 satisfying $\mathsf{E}_{\mu_0}\left\{X^2\right\} \le \gamma$. Therefore, it remains to show that $\mu_0(\{0\}) = q_0^0$ to complete the proof of the upper hemicontinuity of Γ. Similarly as in Lemma 7, let $B_m = \left[-\frac{1}{m}, \frac{1}{m}\right]$ for $m = 1, 2, \ldots$. For any $m \in \mathbb{N}$, there exists an $n(m)$ such that $\pi(\mu_{n_k}, \mu_0) < \frac{1}{m}$ for all $n_k > n(m)$, where π is defined in (127). Therefore, we have for any $m \in \mathbb{N}$ and $n_k > n(m)$,

$$\mu_0(\{0\}) \le \mu_{n_k}(B_m) + \frac{1}{m}, \quad (134)$$

and

$$\mu_{n_k}(\{0\}) \le \mu_0(B_m) + \frac{1}{m}. \quad (135)$$

For any $n \in \{n_k\}_{k=1}^{\infty} \bigcup \{0\}$, we have

$$\mu_n(\{0\}) = \mu_n\left(\bigcap_{m=1}^{\infty} B_m\right) = \lim_{m \to \infty} \mu_n(B_m), \quad (136)$$

so for any $m \in \mathbb{N}$, there exists an $n'(m)$ such that $\mu_n(B_m) \le \mu_n(\{0\}) + \frac{1}{m}$ for all $n > n'(m)$. Therefore, according to (134) and (135), for all $m \in \mathbb{N}$ and $n_k > \max\{n(m), n'(m)\}$,

$$q_0^{n_k} - \frac{2}{m} \le \mu_0(\{0\}) \le q_0^{n_k} + \frac{2}{m}. \quad (137)$$

Thus, due to the assumption that $q_0^{n_k} \to q_0^0$, we have $\mu_0(\{0\}) = q_0^0$ by letting $m \to \infty$.

Next we establish the lower hemicontinuity of Γ. To this end, assume that $q_0^n \to q_0^0$ holds in $[q, 1]$ and $\mu_0 \in \Gamma(q_0^0)$. It suffices to show that there exists a subsequence $\{q_0^{n_k}\}_{k=1}^{\infty}$ of $\{q_0^n\}_{n=1}^{\infty}$ and $\mu_{n_k} \in \Gamma(q_0^{n_k})$ for each $k = 1, 2, \ldots$ such that μ_{n_k} converges to μ_0 in $\mathcal{P}$ [24, Proposition 9.9]. Let $\nu_1, \nu_2 \in \mathcal{P}$ satisfy $\nu_1(\{0\}) = 1$ and $\nu_2(\{0\}) = q, \nu_2(\{\sqrt{\gamma}\}) = 1 - q$, respectively. For all $n = 1, 2, \ldots$, define

$$\mu_n = \begin{cases} \frac{1-q_0^n}{1-q_0^0}\mu_0 + \frac{q_0^n - q_0^0}{1-q_0^0}\nu_1, & q_0^n \ge q_0^0, q_0^0 < 1; \\ \frac{q_0^n - q}{q_0^0 - q}\mu_0 + \frac{q_0^0 - q_0^n}{q_0^0 - q}\nu_2, & q_0^n \le q_0^0, q_0^0 > q. \end{cases} \quad (138)$$

It is easy to check that $\mathsf{E}_{\mu_n}\left\{X^2\right\} \le \gamma$ and $\mu_n(\{0\}) = q_0^n$, thus $\mu_n \in \Gamma(q_0^n)$. It is also easy to check that $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu_0$ for any bounded and continuous function f. Thus μ_n weakly converges to μ_0 [17, Section 3.1], and the lower hemicontinuity of Γ follows.

Next we show that $I_B(q_0)$ in (123) is continuous on $[q, 1]$. Define $\tau = P_{B_2|B_1}(1|0)$. By the stationary property and (125), τ must satisfy

$$0 \le \tau \le \min\left\{1, \frac{q_0 - q}{2cq_0}, \frac{1-q_0}{q_0}\right\}. \tag{139}$$

Denote a compact-valued correspondence Π from $[q, 1]$ to $\mathbb{R}$ such that $\Pi(q_0) = \left[0, \min\left\{1, \frac{q_0-q}{2cq_0}, \frac{1-q_0}{q_0}\right\}\right]$. Π is a continuous correspondence on $[q, 1]$ according to [25, Theorem A.1]. Define a continuous function $\bar{H}_2(\cdot)$ on $\mathbb{R}$ which extends the binary entropy function as follows:

$$\bar{H}_2(x) = \begin{cases} -x\log(x) - (1-x)\log(1-x), & 0 < x < 1; \\ 0, & \text{otherwise.} \end{cases} \tag{140}$$

Define function $f(q_0, \tau)$ on $[q, 1] \times \mathbb{R}$ as

$$f(q_0, \tau) = \begin{cases} \bar{H}_2(q_0) - q_0\bar{H}_2(\tau) - (1-q_0)\bar{H}_2\left(\frac{\tau q_0}{1-q_0}\right), & q_0 < 1; \\ 0, & q_0 = 1. \end{cases} \tag{141}$$

It is easy to see that $f(q_0, \tau)$ is continuous on $[q, 1] \times \mathbb{R}$ and $I(B_1; B_2)$ in (123) equals to $f(q_0, \tau)$ whenever $q_0 \in [q, 1]$ and τ satisfies (139). The continuity of $I_B(q_0)$ on $[q, 1]$ then follows by the maximum theorem [24, Theorem 9.14].

Because $I_X(q_0)$ and $I_B(q_0)$ are continuous on $[q, 1]$, the maximum of (121) can be achieved by some $q_0^* \in [q, 1]$. Also from the derivation above, $I_X(q_0^*)$ is achieved by some distribution $\mu^* \in \Gamma(q_0^*)$ since $I(X; Y) = I(\mu)$ is continuous on the compact set $\Gamma(q_0^*)$. Similarly, $I_B(q_0^*)$ is achieved by some $\tau^* \in \Pi(q_0^*)$ since $I(B_1; B_2) = f(q_0^*, \tau)$ is continuous on the compact set $\Pi(q_0^*)$. Thus the achievability part of Property (b) follows.

Based on Theorem 2, in order to find the lower bound of the capacity, we can maximize $L(\mu)$ and obtain an optimized discrete first-order Markov input $\mu^* = \{\mathcal{X}, \alpha, \beta, P_X\}$ in $\Lambda(\gamma, q, c)$. Let μ_0 denote the capacity-achieving distribution, then

$$I(\mu_0) \ge I(\mu^*) \ge L(\mu^*). \tag{142}$$

In Section VI-A, we develop a computationally efficient scheme to determine μ^*, the rate achieved by which is an approximation of the capacity.

VI. Numerical Methods and Results

A. Computation of the Entropy of Hidden Markov Processes

In order to numerically calculate the mutual information (85), it is important to compute the differential entropy rate of a HMP generated by Markov input through the AWGN channel. Computing the (differential) entropy rate of HMPs is a hard problem. Most works in this area focus on the entropy rate of the binary Markov input through various channels. Reference [26] solves a linear system for the stationary distribution of the quantized Markov process to obtain a good approximation of the entropy rate for the HMP output generated by binary Markov input through a binary symmetric channel. In [27], the entropy rate of HMP generated by binary-symmetric Markov input through arbitrary memoryless channels is studied and a numerical method is presented based on quantizing a fixed-point functional equation. Based on these existing studies, a Monte Carlo algorithm is provided in this paper to compute the differential entropy rate of HMPs generated from a m-state Markov chain ($m \ge 3$) through the AWGN channel. We sketch the main ideas in our algorithm for computing the differential entropy rate in this subsection.

Based on Blackwell's work [28], the entropy of HMPs can be expressed as an expectation on the distribution of the conditional distribution of X_0 given the past observations $Y_{-\infty}^0$. In order to estimate $P_{X_0|Y_{-\infty}^0}$, first define the log-likelihood ratio:

$$L_n^{(i)} = \log\frac{P_{X_n|Y^n}(x^{(i)}|Y^n)}{P_{X_n|Y^n}(x^{(0)}|Y^n)}, \quad i = 0, 1, \cdots, m-1 \tag{143}$$

where m is the number of the states of Markov Chain, $x^{(i)} \in \mathcal{X}$ is the ith state and $\mathcal{X}$ is the state space of Markov Chain. It is obviously that $L_n^{(0)} = 0$. Then given $\boldsymbol{L}_n = \left(L_n^{(0)}, L_n^{(1)}, \cdots, L_n^{(m-1)}\right)$, $P_{X_n|Y^n}$ can be calculated as

$$P_{X_n|Y^n}(x^{(i)}|Y^n) = \frac{e^{L_n^{(i)}}}{\sum_{i=0}^{m-1} e^{L_n^{(i)}}} \tag{144}$$

and when $n \to \infty$, (144) converges to $P_{X_0|Y_{-\infty}^0}$.

In addition, $L_{n+1}^{(i)}$ can be calculated from $\boldsymbol{L}_n$ iteratively as

$$\begin{aligned} L_{n+1}^{(i)} &= \log\frac{P_{X_{n+1}|Y_1^{n+1}}(x^{(i)}|Y_1^{n+1})}{P_{X_{n+1}|Y_1^{n+1}}(x^{(0)}|Y_1^{n+1})} \qquad (145)\\ &= \log\frac{P_{Y_{n+1}|X_{n+1}}(Y_{n+1}|x^{(i)})}{P_{Y_{n+1}|X_{n+1}}(Y_{n+1}|x^{(0)})} \\ &\quad + \log\frac{\sum_{k=0}^{m-1} P_{X_2|X_1}(x^{(i)}|x^{(k)}) P_{X_n|Y_1^n}(x^{(k)}|Y_1^n)}{\sum_{k=0}^{m-1} P_{X_2|X_1}(x^{(0)}|x^{(k)}) P_{X_n|Y_1^n}(x^{(k)}|Y_1^n)} \qquad (146)\\ &= \log\frac{P_{Y_{n+1}|X_{n+1}}(Y_{n+1}|x^{(i)})}{P_{Y_{n+1}|X_{n+1}}(Y_{n+1}|x^{(0)})} + F^{(i)}(\boldsymbol{L}_n) \qquad (147)\\ &= R^{(i)}(Y_{n+1}) + F^{(i)}(\boldsymbol{L}_n) \qquad (148) \end{aligned}$$

where

$$R^{(i)}(Y_{n+1}) = (x^{(i)} - x^{(0)})\, Y_{n+1} - \frac{1}{2}\left((x^{(i)})^2 - (x^{(0)})^2\right), \tag{149}$$

$$F^{(i)}(\boldsymbol{L}_n) = \log\frac{\sum_{k=0}^{m-1} P_{X_2|X_1}(x^{(i)}|x^{(k)})\, e^{L_n^{(k)}}}{\sum_{k=0}^{m-1} P_{X_2|X_1}(x^{(0)}|x^{(k)})\, e^{L_n^{(k)}}}. \tag{150}$$

For the HMPs observed through the AWGN channel (1), the differential entropy can be computed as [28]:

$$h(\mathscr{Y}) = \lim_{n\to\infty} -\iint r(y, \boldsymbol{l}_n)\log r(y, \boldsymbol{l}_n)\, dy\, dP_{\boldsymbol{L}_n}(\boldsymbol{l}_n) \tag{151}$$

where

$$r(y, \boldsymbol{L}_n) = \sum_{i=0}^{m-1} \frac{\phi(y - x^{(i)})}{\sum_{j=0}^{m-1} e^{L_n^{(j)}}} \sum_{k=0}^{m-1} e^{L_n^{(k)}} P_{X_2|X_1}(x^{(i)}|x^{(k)}). \quad (152)$$

In order to compute the entropy rate of HMPs based on (151), the key is to estimate the distribution of $\boldsymbol{L}_n$, $P_{\boldsymbol{L}_n}$. In [26], for binary Markov input and the binary symmetric channel, $\boldsymbol{L}_n$ is considered as a one-dimensional M-state Markov chain by quantizing the dynamic system expressed in (148). Then the distribution of $\boldsymbol{L}_\infty$ is the stationary distribution of the quantized Markov process and can be computed easily through eigenvector solving method. In this paper, because the number of states of the Markov input m is larger than 2 and the HMPs is observed through the AWGN channel, directly quantizing the dynamic system (148) will generate a quantized Markov chain with M^{m-1} states, which is very difficult to deal with when large M is selected for good estimation precision.

According to (148), since $\boldsymbol{L}_{n+1}$ is only dependent on $\boldsymbol{L}_n$ and Y_{n+1}, $\{\boldsymbol{L}_n\}$ can be considered as a Markov process. In order to compute the stationary probability distribution $P_{\boldsymbol{L}_\infty}$, we can evolve the distribution of $\boldsymbol{L}_n$ based on (148) from any initial distribution $P_{\boldsymbol{L}_0}$. When n is large enough, the distribution $P_{\boldsymbol{L}_n}$ converges to $P_{\boldsymbol{L}_\infty}$. A Monte Carlo algorithm for approximating $h(\mathscr{Y})$ is introduced as follows:

1) Initialize M particles $\{\boldsymbol{L}_{0,1}, \cdots, \boldsymbol{L}_{0,M}\}$, $\boldsymbol{L}_{0,k} = \left(0, L_{0,k}^{(1)}, \cdots, L_{0,k}^{(m-1)}\right)$ can be simply sampled from the $(m-1)$-dimensional uniform distribution with each dimension on $\left[-\max(x^{(i)}), \max(x^{(i)})\right]$, $i = 1, \ldots, m-1$;
2) For $n = 0, 1, 2, \cdots, N$, iteratively evolve the particles $\{\boldsymbol{L}_{n,1}, \cdots, \boldsymbol{L}_{n,M}\}$ based on (148), where each $y_{n+1,k}$ is sampled according to $r(y, \boldsymbol{L}_{n,k})$;
3) When N is large enough, $\{\boldsymbol{L}_{N,k}\}$ can be used to estimate $h(\mathscr{Y})$ as

$$h(\mathscr{Y}) \approx -\frac{1}{M} \sum_{k=1}^{M} \int r(y, \boldsymbol{L}_{N,k}) \log r(y, \boldsymbol{L}_{N,k})\, dy. \quad (153)$$

When M is very large, the histogram method can be used to describe $\boldsymbol{L}_{N,k}$ in order to reduce the computational load.

B. Numerical Results

1) Idealized Duty Cycle Constraint $(q, 0)$*:* One implication of Theorem 1 is that directly computing the capacity-achieving input distribution requires solving an optimization problem with infinite number of variables which is prohibitive. Assuming any upper bound on the number of probability mass points, however, a numerical optimization over the mutual information can yield a suboptimal input distribution and a lower bound on the channel capacity. As we increase the number of mass points, the lower bound can be further refined. We take this approach to numerically compute a good approximation of the channel capacity by optimizing over a sufficient number of probability mass points.

Given the duty cycle and power constraints, we first numerically optimize the mutual information by a 3-point input distribution (including a mass at 0), then increase the number of probability mass points by 2 at a time to improve the mutual information, until the improvement is less than 10^{-3}.

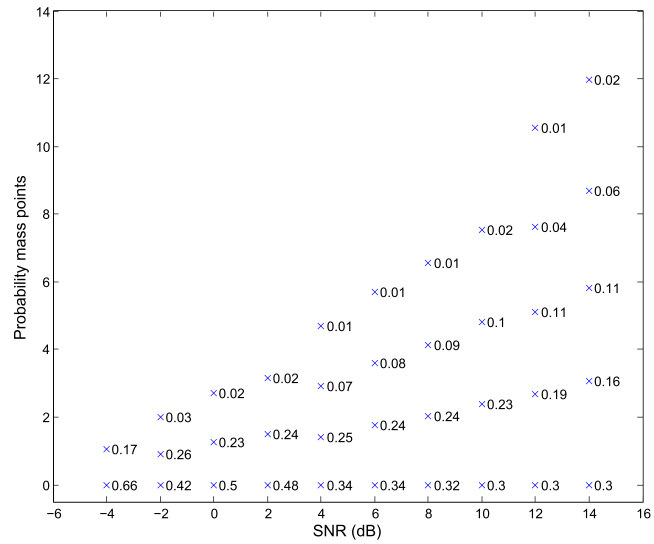

Fig. 1. Suboptimal input distribution for $P_X(0) \geq q = 0.3$.

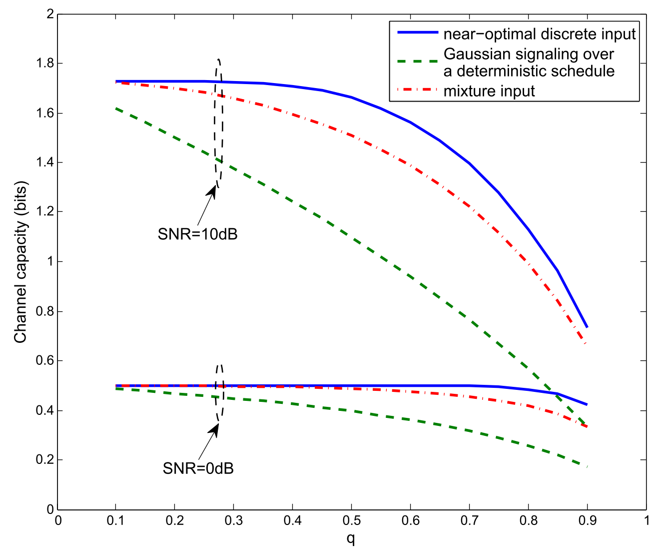

Fig. 2. Achievable rates under duty cycle constraint for 0 dB and 10 dB SNRs.

First consider the case that the duty cycle is no greater than 70%, i.e., $P_X(0) \geq q = 0.3$. For different SNRs, the mass points of the near-optimal input distribution with finite support along with the corresponding probability masses are shown in Fig. 1. Due to symmetry, only the positive half of the input distribution is plotted. We can see that as the SNR increases, more masses are put on higher-amplitude points, whereas the probability mass at zero achieves its lower bound 0.3 eventually.

In Fig. 2, we compare the rate achieved by the near-optimal input distribution and the rate achieved by a conventional scheme using Gaussian signaling over a deterministic schedule, which is $(1-q)$ times the Gaussian channel capacity without duty cycle constraint. It is shown in the figure that there is substantial gain for both 0 dB and 10 dB SNRs by using discrete input over Gaussian signaling with a deterministic schedule. For example, when the SNR is 10 dB, given the duty cycle is no more than 50%, the discrete input

TABLE I

$P_{X_2|X_1}$ AND P_X FOR $q = 0.5$, $c = 1.0$, SNR = 8dB

		X_2				
		0.0000	3.9281	-3.9281	7.1398	-7.1398
	0.0000	0.8342	0.0605	0.0605	0.0224	0.0224
	3.9281	0.4923	0.1852	0.1852	0.0687	0.0687
X_1	-3.9281	0.4923	0.1852	0.1852	0.0687	0.0687
	7.1398	0.4923	0.1852	0.1852	0.0687	0.0687
	-7.1398	0.4923	0.1852	0.1852	0.0687	0.0687
	P_X	0.7481	0.0919	0.0919	0.0341	0.0341

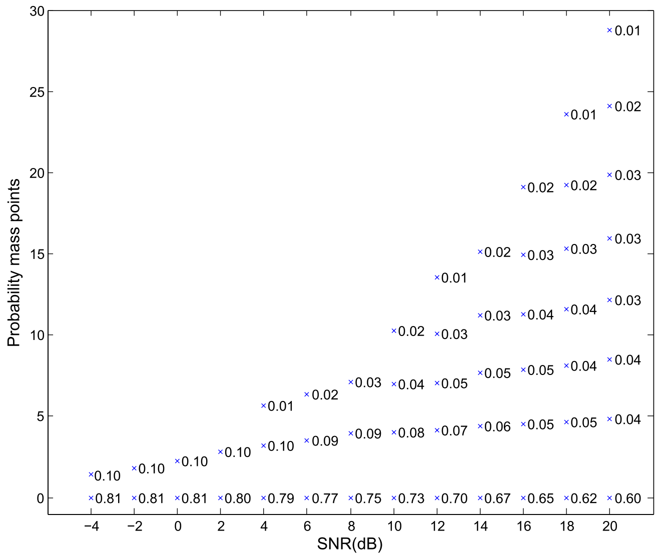

Fig. 3. The marginal distribution of the stationary Markov input. Duty cycle ≤ 0.5, transition cost $c = 1.0$.

distribution achieves 50% higher rate. Hence departing from the usual paradigm of intermittent packet transmissions may yield significant gains.

We also plot in Fig 2 the achievable rate by a superposition coding, where the input distribution is a mixture of Gaussian and a point mass at 0. We first decode the support of the input to find out the positions of nonzero symbols, and then the Gaussian codeword conditioned on the support. It is shown in the figure that the near-optimal discrete input achieves higher rate compared with the mixture input.

2) Realistic Duty Cycle Constraint (q, c)*:* In this subsection, numerical results of the lower bound of channel capacity and the suboptimal input distribution are provided based on the results in Section V and VI-A.

We first seek a discrete Markov chain with finite alphabet that maximizes the objective $L(\mu)$ defined in (11). Once the optimal Markov distribution μ^* is determined, we compute the achievable rate $I(\mu^*)$ according to (85).

In this paper $\mu^* = (\mathcal{X}, \alpha, \beta, P_X)$ is used to approximate the optimum distribution μ_0 through maximizing $L(\cdot)$. It is obvious that the optimized μ^* is symmetric about 0. Table I is the transition probability matrix $P_{X_2|X_1}$ and stationary probability P_X for $q = 0.5$, $c = 1.0$ and SNR = 8 dB. The symmetry of the transition probability matrix is evident, as conditioned on indications whether the two consecutive symbols are zero or nonzero, they are independent.

Fig. 3 shows the stationary (marginal) distribution for suboptimal Markov input. In order to compensate the transition cost, additional fraction of zero symbol should be transmitted, i.e., $P_X(0) > q$. As the SNR increases, more and more weights are put on distant constellation points, where less and less weights are put on the zero letter.

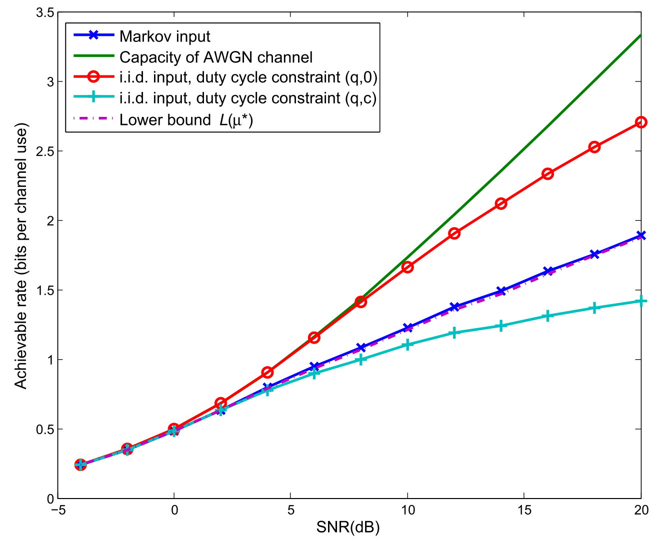

Fig. 4. The achievable rate vs. the SNR. Duty cycle ≤ 0.5, transition cost $c = 1.0$.

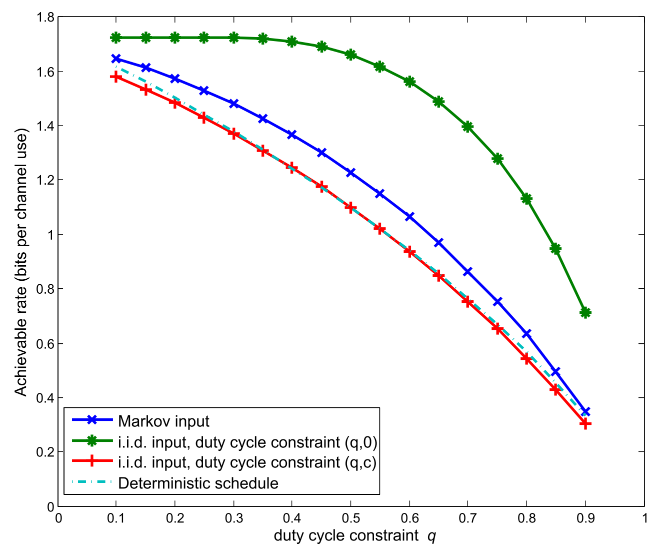

Fig. 5. The achievable rate vs. the duty cycle, SNR = 10 dB and transition cost $c = 1.0$.

In Fig. 4, the rates achieved by various optimized input distributions are plotted against the SNR. The rate achieved by the optimized Markov input μ^* is compared with the capacity of AWGN channel with only power constraint, the rate achieved by i.i.d. input with idealized duty cycle constraint $(q, 0)$, i.e., μ_0 in Theorem 1, the rate achieved by i.i.d. input with duty cycle constraint (q, c) calculated from (93) which is the optimal i.i.d. input distribution in terms of achievable rate, and the lower bound $L(\mu^*)$. It is observed that the optimized Markov input achieves higher rate than the i.i.d. input distribution under duty cycle constraint (q, c).

Figs. 5 and 6 demonstrate the sensitivity of the achievable rates to the duty cycle parameter q and the transition cost c,

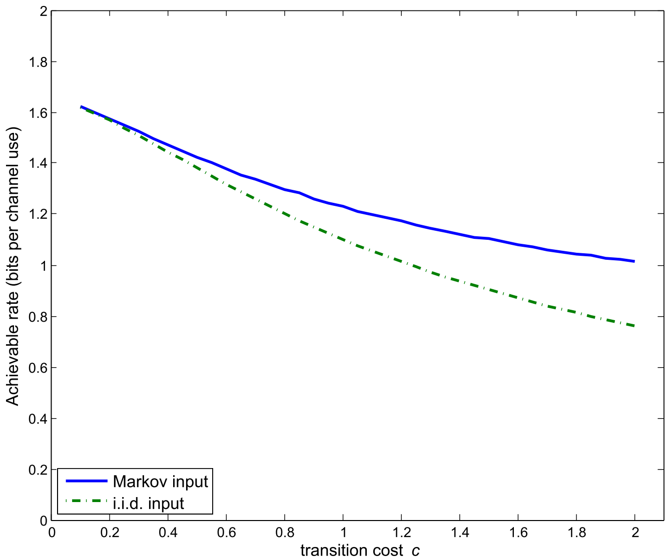

Fig. 6. The achievable rate vs. transition cost, with SNR = 10 dB and $q = 0.5$.

respectively. The performance of Markov inputs is superior to i.i.d. inputs as well as Gaussian signaling with a deterministic schedule. Fig. 5 shows that the performance of i.i.d. input is similar to the deterministic schedule, which implies that different from the case under the idealized duty cycle constraint, i.i.d. input is not a good choice under the realistic duty cycle constraint.

VII. Concluding Remarks

In this paper we have studied the impact of the duty cycle constraint on the capacity of AWGN channels. Under the idealized constraint, the optimal distribution has a finite number of probability mass points in a bounded interval. This allows efficient numerical optimization of the input distribution. Under the realistic duty cycle constraint, the capacity-achieving input is harder to compute. We develop techniques for computing a near-optimal input distribution. This input takes the form of a discrete first-order Markov process, which matches the "Markov" nature of the duty cycle constraint. The numerical results show that under the duty cycle constraint, departing from the usual paradigm of intermittent packet transmissions may yield substantial gain.

Acknowledgment

The authors would like to thank the anonymous reviewers for useful comments. D. Guo and L. Zhang would like to thank Dr. Raymond Yeung and Dr. Shuo-Yen Robert Li for hosting them in the Institute of Network Coding at the Chinese University of Hong Kong during the production of the first draft of this paper. The authors would also like to thank Terence Chan for sharing the code for the numerical results in [11], Yihong Wu and Sergio Verdú for their helpful comments, and Ryan Keating for proofreading the final manuscript.

References

[1] J. M. Rabaey, M. J. Ammer, J. L. da Silva, Jr., D. Patel, and S. Roundy, "PicoRadio supports ad hoc ultra-low power wireless networking," *Computer*, vol. 33, no. 7, pp. 42–48, Jul. 2000.

[2] F. Baccelli, N. Khude, R. Laroia, J. Li, T. Richardson, S. Shakkottai, *et al.*, "On the design of device-to-device autonomous discovery," in *Proc. 4th Int. Conf. Commun. Syst. Netw.*, Jan. 2012, pp. 1–9.

[3] R. Zamir, "A Gaussian input is not too bad," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1362–1367, Jun. 2004.

[4] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Inf. Control*, vol. 18, no. 3, pp. 203–219, 1971.

[5] S. Shamai, "Capacity of a pulse amplitude modulated direct detection photon channel," *IEE Proc. Commun., Speech Vis.*, vol. 137, no. 6, pp. 424–430, Dec. 1990.

[6] S. Shamai and I. Bar-David, "The capacity of average and peak-power-limited quadrature Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1060–1071, Jul. 1995.

[7] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.

[8] M. Katz and S. Shamai, "On the capacity-achieving distribution of the discrete-time noncoherent and partially coherent AWGN channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2257–2270, Oct. 2004.

[9] M. C. Gursoy, H. V. Poor, and S. Verdú, "The noncoherent Rician fading channel—Part I: Structure of the capacity-achieving input," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2193–2206, Sep. 2005.

[10] J. Huang and S. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2336–2351, Jul. 2005.

[11] T. H. Chan, S. Hranilovic, and F. R. Kschischang, "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2073–2088, Jun. 2005.

[12] T. Lutz, C. Hausl, and R. Kötter, "Bits through deterministic relay cascades with half-duplex constraint," in *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 369–381, Jan. 2012.

[13] G. Kramer, "Communication strategies and coding for relaying," in *Wireless Networks* (The IMA Volumes in Mathematics and its Applications), vol. 143. New York, NY, USA: Springer-Verlag, 2007, pp. 163–175.

[14] T. Lutz, G. Kramer, and C. Hausl, "Capacity for half-duplex line networks with two sources," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 2010, pp. 2393–2397.

[15] D. Guo and L. Zhang, "Virtual full-duplex wireless communication via rapid on-off-division duplex," in *Proc. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, 2010, pp. 412–419.

[16] L. Zhang and D. Guo, "Virtual full duplex wireless broadcasting via compressed sensing," *IEEE/ACM Trans. Netw.*, to appear in 2014.

[17] D. W. Stroock, *Probability Theory: An Analytic View*. New York, NY, USA: Cambridge Univ. Press, 1993.

[18] Y. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Theory Probab. Appl.*, vol. 1, no. 2, pp. 157–214, 1956.

[19] Y. Wu and S. Verdú, "Functional properties of MMSE," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1453–1457.

[20] S. Lang, *Complex Analysis*. New York, NY, USA: Springer-Verlag, 1999.

[21] D. G. Luenberger, *Optimization by Vector Sapce Methods*. New York, NY, USA: Wiley, 1969.

[22] W. Rudin, *Real and Complex Analysis*. New York, NY, USA: McGraw-Hill, 1986.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.

[24] R. K. Sundaram, *A First Course in Optimization Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[25] C. Aliprantis and G. Camera, "A correspondence-theoretic approach to dynamic optimization," *Macroecon. Dyn.*, vol. 13, no. 1, pp. 97–117, 2009.

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

[27] J. Luo and D. Guo, "On the entropy rate of hidden Markov processes observed through arbitrary memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1460–1467, Apr. 2009.

[28] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Proc. 1st Prague Conf. Inf. Theory, Statist. Decision Funct., Random Process.*, Prague, Czechoslovakia, 1957, pp. 13–20.

Lei Zhang received the B.E. and M.S. degrees in electronics engineering from Tsinghua University, Beijing, China, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering and computer science from Northwestern University, Evanston, IL, USA, in 2012.

He is currently with Qualcomm Research, Bridgewater, NJ. His research interests include wireless ad hoc networks, compressed sensing, and information theory.

Dr. Zhang is a recipient of the Student Paper Award of the IEEE International Symposium on Information Theory in 2011. He also received the Walter P. Murphy Fellowship and Terminal Year Fellowship from Northwestern University in 2007 and 2011, respectively.

Hui Li joined the faculty of the University of Science and Techology of China, Hefei, China, in 2004, where he is currently an Associate Professor in the Department of Electronic Engineering and Information Science. He received the B.Eng. degree and Ph.D. from the University of Science and Technology of China. He has held a visiting position at Northwestern University, Evanston, IL, USA, in 2011–2012. His research interests include wireless communications and information theory.

Dongning Guo (S'98–M'05–SM'11) joined the faculty of Northwestern University, Evanston, IL, USA, in 2004, where he is currently an Associate Professor in the Department of Electrical Engineering and Computer Science. He received the B.Eng. degree from the University of Science & Technology of China, the M.Eng. degree from the National University of Singapore, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, USA. He was an R&D Engineer in the Center for Wireless Communications (now the Institute for Infocom Research), Singapore, from 1998 to 1999. He has held visiting positions at the Norwegian University of Science and Technology in summer 2006 and in the Institute of Network Coding at the Chinese University of Hong Kong in 2010–2011. He has served as an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY, and is currently an Editor of *Foundations and Trends in Communications and Information Theory*, and a Guest Editor of the IEEE JSAC SPECIAL ISSUE ON FULL-DUPLEX WIRELESS COMMUNICATIONS AND NETWORKS.

Dongning Guo received the Huber and Suhner Best Student Paper Award in the International Zurich Seminar on Broadband Communications in 2000 and is a co-recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2010 (with Y. Zhu and M. L. Honig). He is also a recipient of the National Science Foundation Faculty Early Career Development (CAREER) Award in 2007. His research interests are in information theory, communications, and networking.