Minimum MS. E. Gerber's Lemma

Or Ordentlich and Ofer Shayevitz, *Member, IEEE*

***Abstract*— Mrs. Gerber's Lemma lower bounds the entropy at the output of a binary symmetric channel in terms of the entropy of the input process. In this paper, we lower bound the output entropy via a different measure of input uncertainty, pertaining to the minimum mean squared error prediction cost of the input process. We show that in many cases our bound is tighter than the one obtained from Mrs. Gerber's Lemma. As an application, we evaluate the bound for binary hidden Markov processes, and obtain new estimates for the entropy rate.**

***Index Terms*— Binary symmetric channel, Mrs. Gerber's lemma, hidden Markov process.**

I. Introduction

MRS. GERBER'S Lemma [1] lower bounds the entropy of the output of a binary symmetric channel (BSC) in terms of the entropy of the input to the channel. More specifically, if $\mathbf{X} \in \{0,1\}^n$ is an n-dimensional binary random vector with entropy $H(\mathbf{X})$, $\mathbf{Z} \in \{0,1\}^n$ is an n-dimensional binary random vector with i.i.d. Bernoulli(α) components, statistically independent of $\mathbf{X}$, and $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$, Mrs. Gerber's Lemma states that

$$\frac{1}{n}H(\mathbf{Y}) \geq h\left(\alpha * h^{-1}\left(\frac{1}{n}H(\mathbf{X})\right)\right), \tag{1}$$

where $h(p) \triangleq -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function, $h^{-1}(\cdot)$ is its inverse function restricted to $[0, \frac{1}{2}]$ and $a * b \triangleq a(1-b) + b(1-a)$ denotes the binary convolution between two numbers $a, b \in [0,1]$. For $\mathbf{X}$ i.i.d., the inequality (1) is tight.

The inequality (1) is in fact a simple consequence of the conditional scalar Mrs. Gerber's Lemma, which states the following: If U is some random variable, $X|U = u \sim$ Bernoulli(P_u), and $Z \sim$ Bernoulli(α) is statistically independent of (X, U), we have that

$$H(X \oplus Z|U) \geq h\big(\alpha * h^{-1}(H(X|U))\big), \tag{2}$$

or alternatively,

$$\mathbb{E}h(\alpha * P_U) \geq h\big(\alpha * h^{-1}(\mathbb{E}h(P_U))\big). \tag{3}$$

Manuscript received May 31, 2015; accepted September 3, 2015. Date of publication September 17, 2015; date of current version October 16, 2015. O. Ordentlich was supported by the Admas Fellowship Program of the Israel Academy of Science and Humanities. The work of O. Shayevitz was supported in part by an ERC under Grant 639573, in part by CIG under Grant 631983, and in part by an ISF under Grant 1367/14.

O. Ordentlich is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: ordent@mit.edu).

O. Shayevitz is with the Department of Electrical Engineering Systems, Tel Aviv University, Tel Aviv 69102, Israel (e-mail: ofersha@eng.tau.ac.il).

Communicated by J. Chen, 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.2015.2479641

Since the publication of [1], many extensions, generalizations and results of a similar flavor have been found, see e.g., [2]–[5]. In this paper, we derive a lower bound on the entropy of the output $\mathbf{Y}$ in terms of the *minimum mean squared error predictability* of the input $\mathbf{X}$, as we define next.

Let π be some permutation of the coordinates $\{1, 2, \ldots, n\}$. We define the minimum mean squared error (MMSE) predictability of a binary vector $\mathbf{X}$ w.r.t. the permutation π as

$$\begin{aligned}
&\mathsf{MMSE}_\pi(\mathbf{X}) \\
&\quad\triangleq \sum_{i=1}^{n} \mathrm{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \ldots, X_{\pi(1)}\right) \\
&\quad\triangleq \sum_{i=1}^{n} \mathbb{E}\left(\mathrm{Var}\left(X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \ldots, X_{\pi(1)}\right)\right) \\
&\quad\triangleq \sum_{i=1}^{n} \mathbb{E}\left(P_i^\pi(1-P_i^\pi)\right),
\end{aligned} \tag{4}$$

where the random variable P_i^π is defined as

$$P_i^\pi \triangleq \Pr\left(X_{\pi(i)} = 1|X_{\pi(i-1)}, X_{\pi(i-2)}, \ldots, X_{\pi(1)}\right). \tag{5}$$

The *worst-case MMSE* predictability of a binary vector $\mathbf{X}$ is defined as

$$\overline{\mathsf{MMSE}}(\mathbf{X}) \triangleq \max_\pi \mathsf{MMSE}_\pi(\mathbf{X}). \tag{6}$$

Our main result is the following.

Theorem 1: Let $\mathbf{X}, \mathbf{Z}$ be two statistically independent n-dimensional random binary vectors, where $\mathbf{X}$ is arbitrary and $\mathbf{Z}$ is i.i.d. Bernoulli(α). Let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$. Then

$$\frac{1}{n}H(\mathbf{Y}) \geq h(\alpha) + (1-h(\alpha))\,4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}, \tag{7}$$

with equality if and only if $\mathbf{X}$ is memoryless with $\Pr(X_i = 1) \in \{0, \frac{1}{2}, 1\}$ for every $i = 1, \ldots, n$.

Remark 1: Mrs. Gerber's Lemma hinges on the convexity of the function $\phi(u) = h(\alpha * h^{-1}(u))$. Considering the function $g(x) = 4x(1-x)$ and its restricted inverse $g^{-1} : [0,1] \to [0, \frac{1}{2}]$, one could similarly define the function $g(\alpha * g^{-1}(u)) = 4\alpha(1-\alpha) + (1-2\alpha)^2 u$, which is convex (and in fact linear) in u. Using an argument analogous to the proof of Mrs. Gerber's Lemma, one can obtain

$$\begin{aligned}
\frac{1}{n}H(\mathbf{Y}) &\geq 4\frac{\overline{\mathsf{MMSE}}(\mathbf{Y})}{n} \\
&\geq g\left(\alpha * g^{-1}\left(4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}\right)\right) \\
&= 4\alpha(1-\alpha) + (1-2\alpha)^2 \cdot 4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n},
\end{aligned} \tag{8}$$

Alternatively, the inequality (8) can be easily established directly without invoking the convexity of $g(\alpha * g^{-1}(u))$. However, as a bound for the output entropy, (8) is strictly looser than our bound (7) for any $\alpha \notin \{0, \frac{1}{2}, 1\}$.

In Section II we prove an MMSE version of the conditional scalar Mrs. Gerber Lemma (2), which implies Theorem 1 as a simple corollary. In Section III we derive several MMSE-based extensions of Theorem 1, including a lower bound on $H(\mathbf{Y})$ for the setting where $\mathbf{Z}$ is not i.i.d. as well as an upper bound on $H(\mathbf{Y})$. Section IV compares our new bound to Mrs. Gerber's Lemma. As an application of Theorem 1, in Section V we develop a lower bound on the entropy rate of a binary hidden Markov process, which is shown to be considerably stronger than Mrs. Gerber's Lemma in certain scenarios. Furthermore, our MMSE-based scalar lower bound is combined with a bounding technique developed in [6] to obtain new estimates on the entropy rate of binary hidden Markov processes.

II. Proofs

Mrs. Gerber's Lemma is proved by first deriving the conditional scalar inequality (2) and then invoking the chain rule for entropy and convexity of the function $\phi(u) = h(\alpha * h^{-1}(u))$ to arrive at (1), see [1], [7]. Similarly, we begin by proving an MMSE version of (2) below, from which Theorem 1 will follow as a simple corollary.

Lemma 1: Let U be a random variable and let $X|U = u \sim \text{Bernoulli}(P_u)$. Denote the MMSE in estimating X from U by

$$\text{MMSE}(X|U) \triangleq \mathbb{E}\left(\text{Var}(X|U)\right) = \mathbb{E}\left(P_U(1-P_U)\right). \quad (9)$$

Let $Z \sim \text{Bernoulli}(\alpha)$ be statistically independent of (X, U). Then

$$H(X \oplus Z|U) \geq h(\alpha) + (1-h(\alpha))\, 4\text{MMSE}(X|U),$$

with equality if and only if $P_u \in \{0, \frac{1}{2}, 1\}$ for any value of u.

Proof: Since Z is statistically independent of (X, U) we have

$$H(X \oplus Z|U) = \mathbb{E}h(P_U * \alpha). \quad (10)$$

Let $V_U \triangleq P_U - \frac{1}{2}$ and note that

$$\begin{aligned} P_U * \alpha &= \left(\frac{1}{2} + V_U\right)(1-\alpha) + \alpha\left(\frac{1}{2} - V_U\right) \\ &= \frac{1}{2} + (1-2\alpha)V_U. \end{aligned} \quad (11)$$

Recall that the Taylor series expansion of the binary entropy function around $\frac{1}{2}$ is

$$h\left(\frac{1}{2} + \frac{p}{2}\right) = 1 - \sum_{k=1}^{\infty} \frac{\log(e)}{2k(2k-1)} p^{2k}, \quad (12)$$

and therefore, by (11) we have

$$\begin{aligned} h(P_U * \alpha) &= 1 - \sum_{k=1}^{\infty} \frac{\log(e)}{2k(2k-1)} (1-2\alpha)^{2k} (2V_U)^{2k} \\ &\geq 1 - 4V_U^2 \sum_{k=1}^{\infty} \frac{\log(e)}{2k(2k-1)} (1-2\alpha)^{2k} \quad (13) \\ &= 1 - 4V_U^2 + 4V_U^2 \left(1 - \sum_{k=1}^{\infty} \frac{\log(e)}{2k(2k-1)} (1-2\alpha)^{2k}\right) \\ &= 1 - 4V_U^2 + 4V_U^2 \cdot h\left(\frac{1}{2} + \frac{1-2\alpha}{2}\right) \\ &= 1 - 4V_U^2\,(1 - h(\alpha)), \quad (14) \end{aligned}$$

where (13) follows from the fact that $|2V_U| \leq 1$, and is satisfied with equality if and only if $V_U \in \{-\frac{1}{2}, 0, \frac{1}{2}\}$, which implies that $P_U \in \{0, \frac{1}{2}, 1\}$. Substituting (14) into (10) gives

$$\begin{aligned} H(X \oplus Z|U) &\geq 1 - (1-h(\alpha))\ 4\mathbb{E}(V_U^2) \\ &= 1 - (1-h(\alpha))\ 4\mathbb{E}\left(\frac{1}{2} - P_U\right)^2 \\ &= h(\alpha) + (1-h(\alpha))\ 4\mathbb{E}\left(P_U(1-P_U)\right), \end{aligned}$$

as desired. ∎

Remark 2: Note that the only property of the binary entropy function used in the proof above is that all coefficients of (nonzero) even order in its Taylor expansion around $\frac{1}{2}$ are negative, whereas all odd coefficients are zero. It follows that for any function $g : [0,1] \mapsto \mathbb{R}$ whose Taylor expansion around $\frac{1}{2}$ is of the form

$$g\left(\frac{1}{2} + \frac{p}{2}\right) = c_0 - \sum_{k=1}^{\infty} c_k (p)^{2k},$$

where $c_k \geq 0$ for all positive k we have

$$\mathbb{E}g\left(\alpha * P_U\right) \geq g(\alpha) + (c_0 - g(\alpha))\, 4\text{MMSE}(X|U).$$

Theorem 1 now follows as a straightforward corollary of Lemma 1.

Proof of Theorem 1: By the chain rule for entropy, for any permutation π we have

$$\begin{aligned} H(\mathbf{Y}) &= \sum_{i=1}^{n} H\left(Y_{\pi(i)}|Y_{\pi(i-1)}, \ldots, Y_{\pi(1)}\right) \\ &= \sum_{i=1}^{n} H\left(X_{\pi(i)} \oplus Z_{\pi(i)}|Y_{\pi(i-1)}, \ldots, Y_{\pi(1)}\right) \quad (15) \end{aligned}$$

$$\geq \sum_{i=1}^{n} h(\alpha) + (1-h(\alpha))\, 4\text{MMSE}\left(X_{\pi(i)}|Y_{\pi(i-1)}, \ldots, Y_{\pi(1)}\right). \quad (16)$$

Clearly

$$\begin{aligned} &\text{MMSE}\left(X_{\pi(i)}|Y_{\pi(i-1)}, \ldots, Y_{\pi(1)}\right) \\ &\quad \geq \text{MMSE}\left(X_{\pi(i)}|Y_{\pi(i-1)}, \ldots, Y_{\pi(1)}, Z_{\pi(i-1)}, \ldots, Z_{\pi(1)}\right) \\ &\quad = \text{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)}, \ldots, X_{\pi(1)}, Z_{\pi(i-1)}, \ldots, Z_{\pi(1)}\right) \\ &\quad = \text{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)}, \ldots, X_{\pi(1)}\right), \quad (17) \end{aligned}$$

where the last equality follows since the random variables $\{Z_i\}_{i=1}^n$ are statistically independent of $\{X_i\}_{i=1}^n$. Thus, for any permutation π we have

$$H(\mathbf{Y}) \geq nh(\alpha) + (1-h(\alpha))\,4\sum_{i=1}^{n}\mathrm{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)},\ldots,X_{\pi(1)}\right), \quad (18)$$

and (7) follows by maximizing (18) w.r.t. π. By Lemma 1, the inequality (16) is tight if and only if $\Pr\left(X_{\pi(i)}=1|Y_{\pi(i-1)},\ldots,Y_{\pi(1)}\right) \in \{0,\frac{1}{2},1\}$ for every i and every realization of the vector $\left(Y_{\pi(i-1)},\ldots,Y_{\pi(1)}\right)$, whereas for $0<\alpha<1$ the inequality (17) is tight if and only if $\mathbf{X}$ is memoryless. Thus, (7) holds with equality if and only if $\mathbf{X}$ is memoryless with $\Pr(X_i=1) \in \{0,\frac{1}{2},1\}$ for every $i=1,\ldots,n$. ∎

Remark 3: In the definition of $\overline{\mathsf{MMSE}}(\mathbf{X})$, the maximization is over all *fixed* permutations π. Alternatively, we could allow the prediction order to adaptively depend on the values of $\mathbf{X}$ seen thus far. Specifically, the first coordinate to be predicted is fixed in advance, but the following coordinate to be predicted can be dictated by the value in the previous coordinate, and so on. Clearly, the worst-case MMSE prediction cost for an adaptive prediction order, denoted by $\overline{\mathsf{MMSE}}^*(\mathbf{X})$, may be larger than $\overline{\mathsf{MMSE}}(\mathbf{X})$. By slightly modifying the proof of Theorem 1, it can be shown that it continues to hold even if we replace $\overline{\mathsf{MMSE}}(\mathbf{X})$ with $\overline{\mathsf{MMSE}}^*(\mathbf{X})$.

III. Extensions

In this section we derive several simple extensions of our main results. Since the proofs are quite similar to those of Lemma 1 and Theorem 1, we omit the full details and only sketch the differences instead.

We begin with a straightforward extension of Theorem 1 to the conditional entropy $H(\mathbf{Y}|W)$ where $\mathbf{X}$ may depend on W, while $\mathbf{Z}$ and W are statistically independent. For a permutation π, we define

$$\mathsf{MMSE}_\pi(\mathbf{X}|W) \triangleq \sum_{i=1}^{n}\mathrm{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)},X_{\pi(i-2)},\ldots,X_{\pi(1)},W\right)$$

and

$$\overline{\mathsf{MMSE}}(\mathbf{X}|W) \triangleq \max_\pi \mathsf{MMSE}_\pi(\mathbf{X}|W).$$

Theorem 2: Let W be some random variable, and let $\mathbf{X},\mathbf{Z}$ be two n-dimensional random binary vectors, where $\mathbf{X}$ is arbitrary and $\mathbf{Z}$ is i.i.d. Bernoulli(α). Assume that $(\mathbf{X},W)$ is statistically independent of $\mathbf{Z}$, and let $\mathbf{Y}=\mathbf{X}\oplus\mathbf{Z}$. Then

$$\frac{1}{n}H(\mathbf{Y}|W) \geq h(\alpha) + (1-h(\alpha))\,4\frac{\overline{\mathsf{MMSE}}(\mathbf{X}|W)}{n},$$

with equality if and only if $\mathbf{X}|W=w$ is memoryless with $\Pr(X_i=1|W=w) \in \{0,\frac{1}{2},1\}$ for every $i=1,\ldots,n$ and every w.

Proof: The proof is omitted as it follows the exact same steps as in the proof of Theorem 1, where the conditioning on W is added where relevant. ∎

Next, we show that our lower bound can also be extended to the case of a binary noisy channel with memory. To that end, we first need to derive a simple generalization of Lemma 1.

Lemma 2: Let $U=(T,W)$, where T and W are statistically independent. Let X and Z be conditionally independent given U, such that $X|U=(t,w) \sim \mathrm{Bernoulli}(P_t)$ and $Z|U=(t,w) \sim \mathrm{Bernoulli}(\alpha_w)$. Let $\mathrm{MMSE}(X|U)=\mathrm{MMSE}(X|T)$ be as defined in (9). Then

$$H(X\oplus Z|U) \geq H(Z|W) + (1-H(Z|W))\,4\mathrm{MMSE}(X|T),$$

with equality if and only if $P_t \in \{0,\frac{1}{2},1\}$ for any value of t.

Sketch of Proof: The proof follows the same lines as the proof of Lemma 1. Since T and W are statistically independent, we have $H(X\oplus Z|U)=\mathbb{E}h(P_T*\alpha_W)$. By (14) we have that

$$h(P_T*\alpha_W) \geq 1-4\left(\frac{1}{2}-P_T\right)^2(1-h(\alpha_W)),$$

We therefore have

$$\mathbb{E}_U h(P_T*\alpha_W) \geq \mathbb{E}_U\left(1-4\left(\frac{1}{2}-P_T\right)^2(1-h(\alpha_W))\right) = 1-4\mathbb{E}_T\left(\frac{1}{2}-P_T\right)^2(1-\mathbb{E}_W h(\alpha_W)),$$

and the lemma follows by recalling that $4\mathbb{E}_T\left(\frac{1}{2}-P_T\right)^2 = 1-4\mathrm{MMSE}(X|T)$ and that $\mathbb{E}_W h(\alpha_W)=H(Z|W)$. ∎

As a simple corollary, we obtain the following.

Theorem 3: Let $\mathbf{X},\mathbf{Z}$ be two statistically independent n-dimensional random binary vectors, and let $\mathbf{Y}=\mathbf{X}\oplus\mathbf{Z}$. Then

$$H(\mathbf{Y}) \geq \max_\pi\Bigg\{H(\mathbf{Z}) + 4\mathsf{MMSE}_\pi(\mathbf{X}) - 4\sum_{i=1}^{n}H\left(Z_{\pi(i)}|Z_{\pi(i-1)},\ldots,Z_{\pi(1)}\right) \times \mathrm{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)},\ldots,X_{\pi(1)}\right)\Bigg\},$$

with equality if and only if $\mathbf{Z}$ is memoryless and $\mathbf{X}$ is memoryless with $\Pr(X_i=1) \in \{0,\frac{1}{2},1\}$ for every $i=1,\ldots,n$.

Proof: By the chain rule for entropy, for any permutation π we have

$$H(\mathbf{Y}) = \sum_{i=1}^{n}H\left(Y_{\pi(i)}|Y_{\pi(i-1)},\ldots,Y_{\pi(1)}\right) \geq \sum_{i=1}^{n}H\left(Y_{\pi(i)}|X_{\pi(i-1)},\ldots,X_{\pi(1)},Z_{\pi(i-1)},\ldots,Z_{\pi(1)}\right) \quad (19)$$

$$= \sum_{i=1}^{n}H\left(X_{\pi(i)}\oplus Z_{\pi(i)}|T_\pi^i,W_\pi^i\right) \quad (20)$$

where the random variables

$$T_\pi^i \triangleq \left(X_{\pi(i-1)},\ldots,X_{\pi(1)}\right)$$
$$W_\pi^i \triangleq \left(Z_{\pi(i-1)},\ldots,Z_{\pi(1)}\right)$$

are statistically independent, and $X_{\pi(i)}$ and $Z_{\pi(i)}$ are conditionally independent given (T_π^i, W_π^i), since $\mathbf{X}$ and $\mathbf{Z}$ are statistically independent. The inequality (19) is tight if and only if $\mathbf{X}$ and $\mathbf{Y}$ are both memoryless. Now, by Lemma 2 we have that

$$H\left(X_{\pi(i)} \oplus Z_{\pi(i)}|T_\pi^i, W_\pi^i\right) \geq H(Z_{\pi(i)}|W_\pi^i) + \left(1 - H(Z_{\pi(i)}|W_\pi^i)\right) 4\mathsf{MMSE}\left(X_{\pi(i)}|T_\pi^i\right).$$

Summing over i gives the desired result. ∎

A simple consequence of Theorem 3 is that if $\mathbf{X}$ and $\mathbf{Z}$ are statistically independent binary symmetric first-order Markov processes with transition probabilities q_1 and q_2, respectively, then $\frac{1}{n}H(\mathbf{Y}) \geq h(q_1) + 4q_2(1-q_2)(1-h(q_1))$. This bound uses the identity permutation $\pi = (1, \ldots, n)$. We note that a more clever choice of π, as used in Section V, can result in a better bound.

We end this section by deriving an upper bound on $H(\mathbf{Y})$ in terms of the best-case MMSE predictability of $\mathbf{X}$ from $\mathbf{Y}$

$$\underline{\mathsf{MMSE}}(\mathbf{X}|\mathbf{Y}) \triangleq \min_\pi \sum_{i=1}^n \mathrm{MMSE}\left(X_{\pi(i)}|Y_{\pi(i-1)}, \ldots, Y_{\pi(1)}\right).$$

To that end, we first upper bound $H(X \oplus Z|U)$ in terms of $\mathrm{MMSE}(X|U)$.

Lemma 3: Let U be some random variable and let $X|U = u \sim \text{Bernoulli}(P_u)$. Let $Z \sim \text{Bernoulli}(\alpha)$ be statistically independent of (X, U). Then

$$H(X \oplus Z|U) \leq h\left(\frac{1}{2} + \frac{1-2\alpha}{2}\sqrt{1 - 4\mathrm{MMSE}(X|U)}\right),$$

with equality if and only if $\left|P_u - \frac{1}{2}\right|$ does not depend on u.

Proof: Define the function $Q(t) \triangleq h\left(\frac{1}{2} + \sqrt{t}\right)$ and note that it is concave over $[0, \frac{1}{4}]$. By (10) and (11) we have

$$\begin{aligned} H(X \oplus Z|U) &= \mathbb{E}h\left(\frac{1}{2} + (1-2\alpha)\left(P_U - \frac{1}{2}\right)\right) \\ &= \mathbb{E}h\left(\frac{1}{2} + \sqrt{(1-2\alpha)^2\left(P_U - \frac{1}{2}\right)^2}\right) \\ &\leq h\left(\frac{1}{2} + \sqrt{\mathbb{E}\left((1-2\alpha)^2\left(P_U - \frac{1}{2}\right)^2\right)}\right) \\ &= h\left(\frac{1}{2} + (1-2\alpha)\sqrt{\mathbb{E}\left(\frac{1}{4} - P_U(1-P_U)\right)}\right) \\ &= h\left(\frac{1}{2} + \frac{1-2\alpha}{2}\sqrt{1 - 4\mathrm{MMSE}(X|U)}\right), \end{aligned}$$

as desired. ∎

Remark 4: In the special case where $\alpha = 0$, Lemma 3 reduces to the inequality

$$\mathbb{E}h(P_U) \leq h\left(\frac{1}{2} + \sqrt{\mathbb{E}\left(\frac{1}{2} - P_U)^2\right)}\right),$$

which was obtained by Wyner in [8, eq. (3.11)]

The function $F_\alpha(x) \triangleq h\left(\frac{1}{2} + \frac{1-2\alpha}{2}\sqrt{1-4x}\right)$ is concave and monotone non-decreasing for $x \in [0, \frac{1}{4}]$ and any value of $\alpha \in [0, \frac{1}{2}]$. Combining this with (15) and with Lemma 3 gives the following.

Theorem 4: Let $\mathbf{X}, \mathbf{Z}$ be two statistically independent n-dimensional random binary vectors, where $\mathbf{X}$ is arbitrary and $\mathbf{Z}$ is i.i.d. Bernoulli(α). Let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$. Then

$$\frac{1}{n}H(\mathbf{Y}) \leq h\left(\frac{1}{2} + \frac{1-2\alpha}{2}\sqrt{1 - 4\frac{\underline{\mathsf{MMSE}}(\mathbf{X}|\mathbf{Y})}{n}}\right),$$

with equality if and only if $\mathbf{X}$ is i.i.d.

IV. Comparison With Mrs. Gerber's Lemma

In this section we compare the performance of our MMSE-based bound to Mrs. Gerber's Lemma. First, we consider the family of random vectors with fixed $\overline{\mathsf{MMSE}}(\mathbf{X})$. Clearly, the bound from Theorem 1 is the same for all members of this family. However, the entropy $H(\mathbf{X})$ may vary within the family, and hence applying Mrs. Gerber's Lemma results in a range of bounds, which can be juxtaposed with the bound of Theorem 1. Similarly, we fix $H(\mathbf{X})$ and juxtapose Mrs. Gerber's Lemma with the range of bounds obtained by applying Theroem 1.

For the special case of $\alpha = 0$, Theorem 1 reads

$$H(\mathbf{X}) \geq 4\overline{\mathsf{MMSE}}(\mathbf{X}). \tag{21}$$

and Theorem 4 reads

$$\begin{aligned} H(\mathbf{X}) &\leq nh\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}}\right) \\ &\leq nh\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}}\right). \end{aligned} \tag{22}$$

Denote the RHS of (1) by

$$\mathsf{MGL}(\alpha, P_\mathbf{X}) \triangleq h\left(\alpha * h^{-1}\left(\frac{H(\mathbf{X})}{n}\right)\right),$$

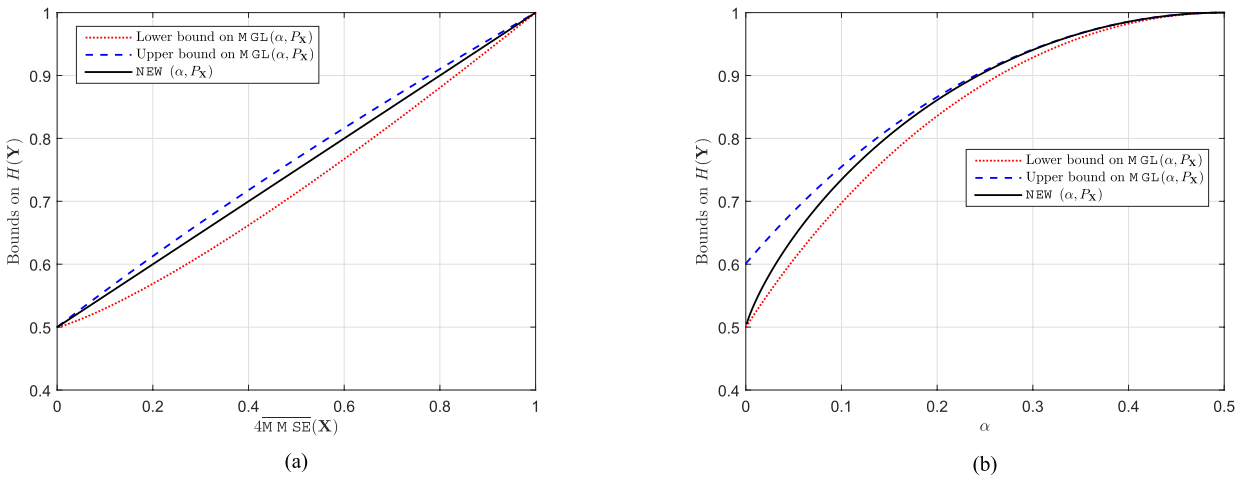
and the RHS of (7) by

$$\mathsf{NEW}(\alpha, P_\mathbf{X}) \triangleq h(\alpha) + (1 - h(\alpha))\, 4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}.$$

By (22) and (21) it follows that

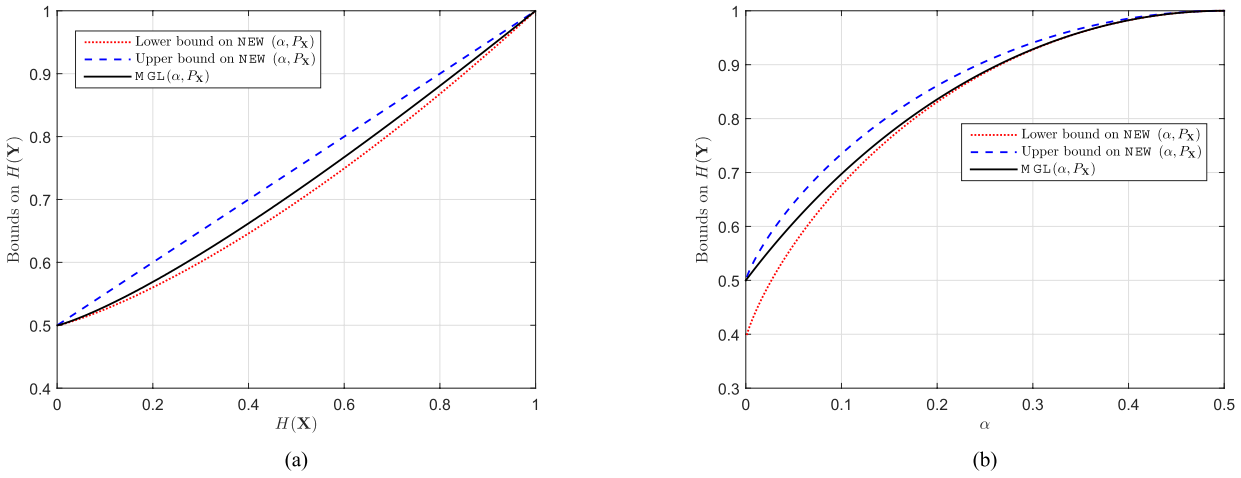
$$\begin{aligned} h\left(\alpha * h^{-1}\left(4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}\right)\right) &\leq \mathsf{MGL}(\alpha, P_\mathbf{X}) \\ &\leq h\left(\alpha * \left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\frac{\overline{\mathsf{MMSE}}(\mathbf{X})}{n}}\right)\right). \end{aligned} \tag{23}$$

Figure 1a depicts the lower and upper bounds on $\mathsf{MGL}(\alpha, P_\mathbf{X})$ from (23) as a function of $\overline{\mathsf{MMSE}}(\mathbf{X})$ along with $\mathsf{NEW}(\alpha, P_\mathbf{X})$, for $\alpha = 0.11$. It is seen that for all values of $\overline{\mathsf{MMSE}}(\mathbf{X})$ our bound is quite close to the upper bound on $\mathsf{MGL}(\alpha, P_\mathbf{X})$, and is often significantly stronger than the lower bound on $\mathsf{MGL}(\alpha, P_\mathbf{X})$. In general, for small values of α, $\mathsf{NEW}(\alpha, P_\mathbf{X})$ will be close to the lower bound on $\mathsf{MGL}(\alpha, P_\mathbf{X})$ and will approach the upper bound on $\mathsf{MGL}(\alpha, P_\mathbf{X})$ as α increases. Figure 1b demonstrates this phenomenon for $4\overline{\mathsf{MMSE}}(\mathbf{X}) = 0.5$.



(a) (b)

Fig. 1. Comparison between the lower and upper bounds on $\mathsf{MGL}(\alpha, P_{\mathbf{X}})$ from (23) and $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$. (a) $\alpha = 0.11$. (b) $\overline{\mathsf{MMSE}}(\mathbf{X}) = 0.5$.



(a) (b)

Fig. 2. Comparison between the lower and upper bounds on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$ from (25) and $\mathsf{MGL}(\alpha, P_{\mathbf{X}})$. (a) $\alpha = 0.11$. (b) $H(\mathbf{X}) = 0.5$.

Equivalently, by (21) and (22), we also have that

$$4nh^{-1}\left(\frac{H(\mathbf{X})}{n}\right)\left(1-h^{-1}\left(\frac{H(\mathbf{X})}{n}\right)\right) \le 4\overline{\mathsf{MMSE}}(\mathbf{X}) \le H(\mathbf{X}). \tag{24}$$

In fact, (24) holds for $4\mathsf{MMSE}_\pi(\mathbf{X})$ with any permutation π, and implies

$$h(\alpha)+(1-h(\alpha))4h^{-1}\left(\frac{H(\mathbf{X})}{n}\right)\left(1-h^{-1}\left(\frac{H(\mathbf{X})}{n}\right)\right) \le \mathsf{NEW}(\alpha, P_{\mathbf{X}}) \le h(\alpha)+(1-h(\alpha))H(\mathbf{X}) \tag{25}$$

Figure 2a depicts the lower and upper bounds on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$ from (25) as a function of $H(\mathbf{X})$ along with $\mathsf{MGL}(\alpha, P_{\mathbf{X}})$, for $\alpha = 0.11$. It is seen that for all values of $H(\mathbf{X})$, $\mathsf{MGL}(\alpha, P_{\mathbf{X}})$ is quite close to the lower bound on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$, and is often significantly weaker than the upper bound on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$. In general, for small values of α, $\mathsf{MGL}(\alpha, P_{\mathbf{X}})$ will be close to the upper bound on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$ and will approach the lower bound on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$ as α increases. Figure 2b demonstrates this phenomenon for $H(\mathbf{X}) = 0.5$.

V. Application: Lower Bound on the Entropy Rate of a Binary Hidden Markov Process

In this section we apply Theorem 1 to derive a simple lower bound on the entropy rate of a binary hidden Markov process. Let $X_1 \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ and for $m = 2, 3, \ldots$ let $X_m = X_{m-1} \oplus W_m$ where $\{W_m\}$ is an i.i.d. Bernoulli(q) process statistically independent of X_1. Clearly, the process $\{X_n\}$ is a symmetric first-order Markov Process. We define the hidden Markov process $Y_n = X_n \oplus Z_n$, where $\{Z_n\}$ is an i.i.d. Bernoulli(α) process statistically independent of the process $\{X_n\}$. Our goal in this section is to derive a lower bound on the entropy rate of $\{Y_n\}$ defined as

$$\overline{H}(Y) \triangleq \lim_{n\to\infty} \frac{H(Y_1, \ldots, Y_n)}{n}. \tag{26}$$

One very simple bound can be obtained by noting that $\overline{H}(X) = h(q)$ and applying Mrs. Gerber's Lemma (1) which gives

$$\overline{H}(Y) \ge h(\alpha * q). \tag{27}$$

We will see that in many cases our MMSE-based bound from Theorem 1 provides tighter bounds.

Note that for any π it holds that $\overline{\mathsf{MMSE}}(\mathbf{X}) \geq \mathsf{MMSE}_\pi(\mathbf{X})$ and therefore Theorem 1 implies that for any choice of π

$$\frac{1}{n}H(\mathbf{Y}) \geq h(\alpha) + (1-h(\alpha))\,4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n}. \tag{28}$$

Thus, in order to apply Theorem 1 we need to choose some π and evaluate $\mathsf{MMSE}_\pi(\mathbf{X})$. A trivial choice is the identity $\pi = \{1, 2, \ldots, n\}$, for which $\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n} = q(1-q)$ and our bound yields $\overline{H}(Y) \geq h(\alpha) + (1-h(\alpha))4q(1-q)$. It is easy to see that this choice of π yields the lower bound on $\mathsf{NEW}(\alpha, P_{\mathbf{X}})$ from (25), and hence, is strictly weaker than (27). We would therefore like to choose a permutation π that will incur a higher value of $\mathsf{MMSE}_\pi(\mathbf{X})$. Assume that $\log n$ is an integer. A natural candidate is the following

$$\pi = \left(n, \frac{n}{2}, \frac{n}{4}, \frac{3n}{4}, \frac{n}{8}, \frac{3n}{8}, \frac{5n}{8}, \frac{7n}{8}, \frac{n}{16}, \frac{3n}{16}, \ldots\right). \tag{29}$$

With this choice of π we have that if $\pi(i) = rn/2^k$, for some $r \in \{1, 3, \ldots, 2^{k-1}\}$, then

$$\begin{aligned}
&\mathrm{MMSE}\left(X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \ldots, X_{\pi(1)}\right)\\
&\quad\geq \mathrm{MMSE}\left(X_{\frac{rn}{2^k}}|X_{\frac{rn}{2^k}-\frac{n}{2^k}}, X_{\frac{rn}{2^k}+\frac{n}{2^k}}\right)\\
&\quad= \mathrm{MMSE}\left(X_m|X_{m-\frac{n}{2^k}}, X_{m+\frac{n}{2^k}}\right)\\
&\quad\triangleq \mathrm{MMSE}\left(\frac{n}{2^k}\right),
\end{aligned}$$

where the inequality follows from the Markovity of $\{X_n\}$ which implies that the conditional distribution of X_m given multiple samples from the past and the future of the process depends only on the nearest sample from the past and the nearest sample from the future. We therefore have

$$\begin{aligned}
\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n} &\geq \sum_{k=1}^{\log n}\frac{1}{2}\frac{2^k}{n}\mathrm{MMSE}\left(\frac{n}{2^k}\right)\\
&= \frac{1}{2}\sum_{k=1}^{\log n}2^{-(\log n-k)}\mathrm{MMSE}\left(2^{\log n-k}\right)\\
&= \frac{1}{2}\sum_{t=0}^{\log n-1}2^{-t}\mathrm{MMSE}\left(2^t\right).
\end{aligned} \tag{30}$$

It now only remains to calculate

$$\begin{aligned}
\mathrm{MMSE}(\ell) &= \mathrm{MMSE}(X_n|X_{n-\ell}X_{n+\ell})\\
&= \mathbb{E}\left(P_1^\ell(X_{n+\ell}, X_{n-\ell})P_0^\ell(X_{n+\ell}, X_{n-\ell}))\right)
\end{aligned} \tag{31}$$

where the random variable $P_i^\ell(X_{n+\ell}, X_{n-\ell})$ is defined as

$$\begin{aligned}
&P_i^\ell(x_{n+\ell}, x_{n-\ell})\\
&\quad\triangleq \Pr(X_n = i|X_{n-\ell} = x_{n-\ell}, X_{n+\ell} = x_{n+\ell})\\
&\quad= \frac{P(X_{n+\ell} = x_{n+\ell}, X_n = i, X_{n-\ell} = x_{n-\ell})}{P(X_{n-\ell} = x_{n-\ell}, X_{n+\ell} = x_{n+\ell})}\\
&\quad= \frac{P(X_{n+\ell} = x_{n+\ell}|X_n = i)P(X_n = i|X_{n-\ell} = x_{n-\ell})}{P(X_{n+\ell} = x_{n+\ell}|X_{n-\ell} = x_{n-\ell})},
\end{aligned}$$

for $i = 0, 1$. Let $P_k \triangleq \Pr(X_{n+k} \neq X_n)$. With this notation we have that if $x_{n+\ell} \neq x_{n-\ell}$ then

$$P_1^\ell(x_{n+\ell}, x_{n-\ell}) = P_0^\ell(x_{n+\ell}, x_{n-\ell}) = \tfrac{P_\ell(1-P_\ell)}{P_{2\ell}}. \tag{32}$$

On the other hand, if $x_{n+\ell} = x_{n-\ell}$ we have

$$P_1^\ell(x_{n+\ell}, x_{n-\ell})P_0^\ell(x_{n+\ell}, x_{n-\ell}) = \tfrac{P_\ell^2}{1-P_{2\ell}}\tfrac{(1-P_\ell)^2}{1-P_{2\ell}}. \tag{33}$$

It therefore follows that

$$\begin{aligned}
\mathrm{MMSE}(\ell) &= \Pr(X_{n+\ell} \neq X_{n-\ell})\left(\frac{P_\ell(1-P_\ell)}{P_{2\ell}}\right)^2\\
&\quad+\Pr(X_{n+\ell} = X_{n-\ell})\left(\frac{P_\ell(1-P_\ell)}{1-P_{2\ell}}\right)^2\\
&= (P_\ell(1-P_\ell))^2\left(\frac{1}{P_{2\ell}}+\frac{1}{1-P_{2\ell}}\right)\\
&= \frac{(P_\ell(1-P_\ell))^2}{P_{2\ell}(1-P_{2\ell})}.
\end{aligned} \tag{34}$$

Note that

$$\begin{aligned}
P_k &= \Pr(X_{n+k} \neq X_n)\\
&= \Pr\left(\left(\prod_{i=n+1}^{n+k}(-1)^{W_i}\right) = -1\right)\\
&= \frac{1-\mathbb{E}\left(\prod_{i=n+1}^{n+k}(-1)^{W_i}\right)}{2}\\
&= \frac{1-(1-2q)^k}{2}
\end{aligned} \tag{35}$$

Substituting (35) into (34) gives

$$\begin{aligned}
\mathrm{MMSE}(\ell) &= \frac{\left(\frac{1}{4}\left(1-(1-2q)^{2\ell}\right)\right)^2}{\frac{1}{4}\left(1-(1-2q)^{2\ell}\right)\left(1+(1-2q)^{2\ell}\right)}\\
&= \frac{1}{4}\cdot\frac{1-(1-2q)^{2\ell}}{1+(1-2q)^{2\ell}}.
\end{aligned} \tag{36}$$

Substituting (36) into (30) gives

$$\begin{aligned}
\lim_{n\to\infty}4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n} &\geq \sum_{t=0}^{\infty}2^{-(t+1)}\frac{1-(1-2q)^{2^{t+1}}}{1+(1-2q)^{2^{t+1}}}\\
&\geq \sum_{t=1}^{\infty}2^{-t}\frac{1-(1-2q)^{2^t}}{1+(1-2q)^{2^t}},
\end{aligned} \tag{37}$$

and consequently we get the following theorem.

Theorem 5: Let $\{X_n\}$ be a first-order Markov process with parameter q, $\{Z_n\}$ be an i.i.d. Bernoulli(α) process statistically independent of $\{X_n\}$ and $Y_n = X_n \oplus Z_n$. Then

$$\overline{H}(Y) \geq h(\alpha) + (1-h(\alpha))\sum_{t=1}^{\infty}2^{-t}\frac{1-(1-2q)^{2^t}}{1+(1-2q)^{2^t}}.$$

Remark 5: For every $\alpha \in (0, 1/2)$ there exist a $q_\alpha > 0$ such that the bound from Theorem 5 outperforms Mrs. Gerber's Lemma for all $q \in (0, q_\alpha)$. For example, $q_{0.11} \approx 0.212$. As discussed in the previous section, q_α increases with α and approaches $1/2$ as $\alpha \to 1/2$.

It will be instructive to study the behavior of the RHS of (37) in the limit of $q \to 0$. To this end we write,

for some $0 < \gamma < 1$ such that $-\gamma \log q$ is an integer

$$\begin{aligned}
&\lim_{n\to\infty} 4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n} \\
&\quad \geq \sum_{t=1}^{\infty} 2^{-t}\frac{1-(1-2q)^{2^t}}{1+(1-2q)^{2^t}} \\
&\quad \geq \sum_{t=1}^{-\gamma\log q} 2^{-t}\frac{1-(1-2q)^{2^t}}{2} \\
&\quad = \sum_{t=1}^{-\gamma\log q} 2^{-(t+1)}\left(2^{t+1}q - \sum_{k=2}^{2^t}(-1)^k\binom{2^t}{k}(2q)^k\right) \\
&\quad \geq \sum_{t=1}^{-\gamma\log q} 2^{-(t+1)}\left(2^{t+1}q - \sum_{k=2}^{2^t}(2^t)^k(2q)^k\right) \\
&\quad \geq \sum_{t=1}^{-\gamma\log q} q - 2^{-(t+1)}\sum_{k=2}^{2^t}\left(2^{t+1}q\right)^k. \qquad (38)
\end{aligned}$$

Using the fact that $\sum_{k=2}^{m} r^k = \frac{r^2-r^{m+1}}{1-r} \leq \frac{r^2}{1-r}$ for $0 < r < 1$, we further bound (38) as

$$\begin{aligned}
\lim_{n\to\infty} 4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n} &\geq \sum_{t=1}^{-\gamma\log q} q - 2^{-(t+1)}\frac{\left(2^{t+1}q\right)^2}{1-2^{t+1}q} \\
&= \sum_{t=1}^{-\gamma\log q} q - \frac{2^{t+1}q^2}{1-2^{t+1}q} \\
&\geq -\gamma\, q\log q\left(1-\frac{2q^{1-\gamma}}{1-2q^{1-\gamma}}\right). \qquad (39)
\end{aligned}$$

For $q \to 0$ we can take $\gamma = 1 - 1/\sqrt{-\log q}$ such that

$$\begin{aligned}
\lim_{n\to\infty} 4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{n} &\geq -q\log(q)\left(1-\varepsilon_q'\right) \\
&= h(q)\left(1-\varepsilon_q\right) \qquad (40)
\end{aligned}$$

where $\varepsilon_q', \varepsilon_q \to 0$ as $q \to 0$. We have therefore obtained that

$$\lim_{q\to 0}\lim_{n\to\infty} 4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{nh(q)} = \lim_{q\to 0}\lim_{n\to\infty} 4\frac{\mathsf{MMSE}_\pi(\mathbf{X})}{H(\mathbf{X})} \geq 1.$$

Thus, we have seen that while the trivial choice $\pi' = \{1, 2, \ldots, n\}$ yields $\mathsf{MMSE}_{\pi'}(\mathbf{X})$ that meets the lower bound from (24), the more clever choice of π given in (29) yields $\mathsf{MMSE}_\pi(\mathbf{X})$ that meets the upper bound from (24) in the limit.

Remark 6: The permutation π from (29) can be found by a greedy algorithm that constructs the permutation vector sequentially by choosing in the ith step

$$\pi(i) = \operatorname*{argmax}_{j\in[n]\setminus\{\pi(1),\ldots,\pi(i-1)\}} \mathrm{MMSE}\left(X_j|X_{\pi(1)},\ldots,X_{\pi(i-1)}\right),$$

where $[n] \triangleq \{1,\ldots,n\}$. The asymptotic optimality of π from (29) for symmetric Markov chains may suggest that such a greedy algorithm will always yield the permutation vector that maximizes $\mathsf{MMSE}_\pi(\mathbf{X})$. This is, unfortunately, not true in general. As a counterexample consider the vector $\mathbf{X} = (X_1, X_2)$ with

$$\Pr(X_1=0, X_2=0) = \frac{1}{2}; \quad \Pr(X_1=0, X_2=1) = 0$$
$$\Pr(X_1=1, X_2=0) = \varepsilon; \quad \Pr(X_1=1, X_2=1) = \frac{1}{2}-\varepsilon$$

for which $\mathrm{Var}(X_1) > \mathrm{Var}(X_2)$ but

$$\mathrm{Var}(X_2) + \mathrm{MMSE}(X_1|X_2) > \mathrm{Var}(X_1) + \mathrm{MMSE}(X_2|X_1)$$

for ϵ small enough.

Substituting (40) into Theorem 5 gives that for small q

$$\overline{H}(\mathbf{Y}) \geq h(\alpha) + (1-h(\alpha))h(q)(1-\varepsilon_q). \qquad (41)$$

Note that this bound has an infinite slope at $q = 0$. This is always better than the Cover-Thomas type of bounds $\overline{H}(Y) \geq H(Y_m|Y_{m-1},\ldots,Y_1,X_0)$ derived in [9, Th. 4.5.1] which are always smaller than $h(q^{*m} * \alpha)$, where q^{*m} denotes convolving q with itself m times. Both bounds evaluate to $h(\alpha)$ at $q = 0$, but the derivative of the latter is finite for any finite m. Thus, for small q our bound is better than the Cover-Thomas bound of any order.

The bound (41) is weaker than the best known lower bounds on $\overline{H}(Y)$ in the rare transition regime. For example, in [10] it is shown that $\overline{H}(Y) \geq h(\alpha) - \frac{(1-2\alpha)^2}{1-\alpha} q\log q$, whereas in [11] this was improved to $\overline{H}(Y) \geq h(\alpha) + h(q) - Cq$ for some $C > 0$. However, the two bounds mentioned above are "tailor-made" to hidden Markov models, whereas (41) follows from applying our generic bound from Theorem 1 to the special case of a hidden Markov model. In the next subsection we will show that the scalar version of our MMSE-based bound, stated in Lemma 1 can be used to enhance such a "tailor-made" bound for Markov chains.

A. Bound based on the Ordentlich-Weissman Method

In [6], E. Ordentlich and T. Weissman cleverly observed that the entropy rate of a binary symmetric first-order hidden Markov process can be expressed as

$$\overline{H}(\mathbf{Y}) = \mathbb{E}\left(\frac{e^{W_i}}{1+e^{W_i}} * q * \alpha\right), \qquad (42)$$

where the auto-regressive process W_i is defined as

$$W_i = R_i \ln\frac{1-\alpha}{\alpha} + S_i f(W_{i-1}) \qquad (43)$$

for

$$f(t) = \ln\frac{e^t(1-q)+q}{qe^t+(1-q)} \qquad (44)$$

and i.i.d. processes $\{R_i\}$ and $\{S_i\}$ statistically independent of W_0, with distributions

$$R_i = \begin{cases} 1 & w.p.\ 1-\alpha \\ -1 & w.p.\ \alpha \end{cases}; \quad S_i = \begin{cases} 1 & w.p.\ 1-q \\ -1 & w.p.\ q. \end{cases} \qquad (45)$$

The expectation in (42) is taken under the assumption that W_0 is distributed according to the (unique) stationary distribution of the process $\{W_i\}$, and is therefore well-defined. In [6], upper and lower bounds on $\overline{H}(\mathbf{Y})$ were derived by analyzing the support of the process $\{W_i\}$. Here, we apply Lemma 1 in

order to derive a lower bounds on $\overline{H}(\mathbf{Y})$. To this end, we set $X|W_i \sim \text{Bernoulli}\left(\frac{e^{W_i}}{1+e^{W_i}}\right)$ and find a lower bounds on

$$\text{MMSE}(X|W_i) = \mathbb{E}\left(\frac{e^{W_i}}{(1+e^{W_i})^2}\right).$$

Let $F \triangleq e^{f(W_{i-1})}$ and $\eta = \frac{1-\alpha}{\alpha}$, such that $e^{W_i} = \eta^{R_i} F^{S_i}$. We have

$$\begin{aligned}
&\mathbb{E}\left(\frac{e^{W_i}}{(1+e^{W_i})^2}|F\right) \\
&= (1-\alpha)(1-q)\frac{\eta F}{(1+\eta F)^2} + (1-\alpha)q\frac{\eta/F}{(1+\eta/F)^2} \\
&\quad + \alpha(1-q)\frac{F/\eta}{(1+F/\eta)^2} + \alpha q\frac{1/(\eta F)}{(1+1/(\eta F))^2} \\
&= ((1-\alpha)(1-q)+\alpha q)\frac{\eta F}{(1+\eta F)^2} \\
&\quad + ((1-\alpha)q+\alpha(1-q))\frac{F/\eta}{(1+F/\eta)^2} \qquad (46)\\
&= (1-\alpha*q)\frac{\eta F}{(1+\eta F)^2} + (\alpha*q)\frac{F/\eta}{(1+F/\eta)^2} \\
&\triangleq g(F), \qquad (47)
\end{aligned}$$

where we have used the fact that $e^{W_i}/(1+e^{W_i})^2 = e^{-W_i}/(1+e^{-W_i})^2$ in (46). Let $\mathcal{S}$ be the support of the random variable F. Clearly,

$$\text{MMSE}(X|W_i) = \mathbb{E}g(F) \geq \min_{s\in S} g(s) \qquad (48)$$

In [6, eq. (44) and (45)] it is shown that $\mathcal{S} \subseteq [1/F_{\max}, F_{\max}]$, where

$$F_{\max} \triangleq \frac{(\eta-1)(1-q)+\sqrt{4\eta q^2+(\eta-1)^2(1-q)^2}}{2\eta q}. \qquad (49)$$

Let $g_1(F) \triangleq \frac{\eta F}{(1+\eta F)^2}$ and $g_2(F) \triangleq \frac{F/\eta}{(1+F/\eta)^2}$, and note that $g_2(1/F) = g_1(F)$ and that $g(F) = (1-\alpha*q)g_1(F) + (\alpha*q)g_2(F)$. For $F \geq 1$ we have that $g_2(F) \geq g_1(F)$, whereas for $F < 1$ we have that $g_1(F) > g_2(F)$. Since $(1-\alpha*q) \geq (\alpha*q)$ (recall that we assume $\alpha, q \leq 1/2$), we must have that

$$\min_{s\in[1/F_{\max}, F_{\max}]} g(s) = \min_{s\in[1,F_{\max}]} g(s). \qquad (50)$$

Straightforward algebra gives

$$\begin{aligned}
&\text{sign}\left(g'(s)\right) \\
&= \text{sign}\left((\eta-s)(1+\eta s)^3 - \frac{1-\alpha*q}{\alpha*q}(\eta s-1)(\eta+s)^3\right).
\end{aligned}$$

Note that $\text{sign}(g'(1)) = -1$, and therefore if the equation $\text{sign}\left(g'(s)\right) = 0$ does not have any real solution in $[1, F_{\max})$ then we must have

$$\min_{s\in[1/F_{\max}, F_{\max}]} g(s) = g(F_{\max}). \qquad (51)$$

Otherwise, $\min_{s\in[1/F_{\max}, F_{\max}]} g(s)$ is obtained either in one of the solutions of $\text{sign}\left(g'(s)\right) = 0$ in the interval $[1, F_{\max})$, or in $F_{\max}$. The equation $\text{sign}\left(g'(s)\right) = 0$ is equivalent to

$$\begin{aligned}
&\eta\left(\frac{1-\alpha*q}{\alpha*q}+\eta^2\right)s^4 + \left(3\eta^2\frac{1}{\alpha*q} - \eta^4 - \eta\right)s^3 \\
&+ 3\eta\frac{1-2(\alpha*q)}{\alpha*q}(\eta^2-1)s^2 + \left(\frac{1-\alpha*q}{\alpha*q}\eta^4 + 1 - 3\frac{\eta^2}{\alpha*q}\right)s \\
&- \eta\left(1+\frac{1-\alpha*q}{\alpha*q}\eta^2\right) = 0, \qquad (52)
\end{aligned}$$

Let $\mathcal{S}^*$ be the set of solutions to the equation (52) in $[1, F_{\max})$. We conclude that $\text{MMSE}(X|W_i) \geq g(F^*)$ where

$$F^* = \operatorname*{argmin}_{s\in(\mathcal{S}^*\cup F_{\max})} g(s). \qquad (53)$$

and this combined with (42) and Lemma 1 yields the following.

Theorem 6: Let $\{X_n\}$ be a first-order Markov process with parameter q, $\{Z_n\}$ be an i.i.d. Bernoulli(α) process statistically independent of $\{X_n\}$ and $Y_n = X_n \oplus Z_n$. Then

$$\overline{H}(Y) \geq h(\alpha*q) + (1-h(\alpha*q))\,g(F^*),$$

where F^* is defined by (49), (52) and (53), $g(\cdot)$ is defined in (47), and $\eta = \frac{1-\alpha}{\alpha}$.

In Figure 3 we plot the bound from Theorem 6 for $\alpha = 0.11$ and $q \in [0, 0.5]$. For comparison, we also plot the lower bound from [6, Corollary 4.8 and Lemma 4.10], and it is seen that for small values of q our new bound improves upon that of [6].

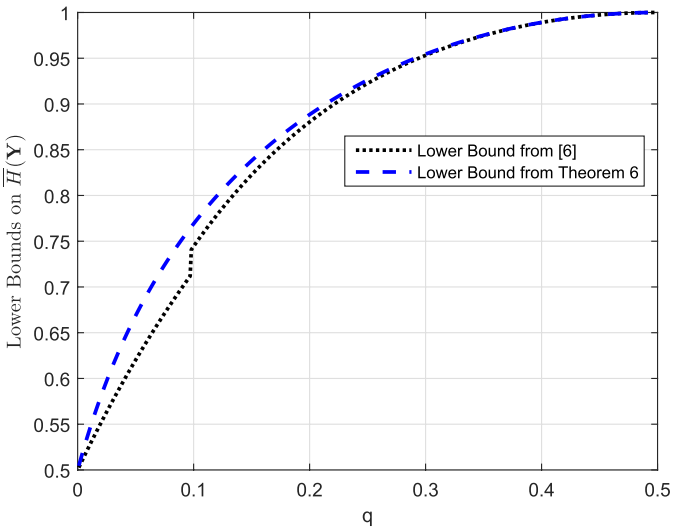


Fig. 3. Comparison between the lower bound from Theorem 6 and the lower bound from [6, Corollary 4.8 and Lemma 4.10] for $\alpha = 0.11$ and q ranging between 0 and $\frac{1}{2}$.

ACKNOWLEDGMENT

The authors thank Tsachy Weissman for his valuable comments, and for pointing out the observation stated in Remark 3.

REFERENCES

[1] A. D. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications–I," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov. 1973.

[2] H. S. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974.

[3] N. Chayat and S. Shamai, "Extension of an entropy property for binary input memoryless symmetric channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 1077–1079, Sep. 1989.

[4] S. Shamai and A. D. Wyner, "A binary analog to the entropy-power inequality," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1428–1430, Nov. 1990.

[5] V. Jog and V. Anantharam, "The entropy power inequality and Mrs. Gerber's lemma for groups of order 2^n," in *Proc. ISIT*, Jul. 2013, pp. 594–598.

[6] E. Ordentlich and T. Weissman, "Bounds on the entropy rate of binary hidden Markov processes," in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems* (London Mathematical Society Lecture note Series), vol. 385. Cambridge, U.K.: Cambridge Univ. Press, Jun. 2011.

[7] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K., Cambridge Univ. Press, 2011.

[8] A. D. Wyner, "The common information of two dependent random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

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

[11] Y. Peres and A. Quas, "Entropy rate for hidden Markov chains with rare transitions," in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems* (London Mathematical Society Lecture note Series), vol. 385. Cambridge, U.K.: Cambridge Univ. Press, Jun. 2011.

Or Ordentlich received the B.Sc. degree (cum laude) in 2010, M.Sc. degree (summa cum laude) in 2011, and completed his Ph.D. studies in 2015, all in electrical engineering at Tel Aviv University, Israel. He is currently a postdoctoral fellow in the Laboratory of Information and Decision Systems at the Massachusetts Institute of Technology (MIT), Cambridge.

Or is the recipient of the MIT - Technion Postdoctoral Fellowship, the Adams Fellowship awarded by the Israel Academy of Sciences and Humanities, the Thalheimer Scholarship for graduate students, the Advanced Communication Center (ACC) Feder Family Award for outstanding research work in the field of communication technologies (2011,2014), and the Weinstein Prize for research in signal processing (2011,2013,2014).

Ofer Shayevitz (M'08) received the B.Sc. degree (summa cum laude) from the Technion Institute of Technology, Haifa, Israel, in 1997 and the M.Sc. and Ph.D. degrees from the Tel-Aviv University, Tel Aviv, Israel, in 2004 and 2009, respectively, all in electrical engineering. He is currently a Senior Lecturer in the Department of EE - Systems at the Tel Aviv University, and also serves as the head of the Advanced Communication Center (ACC). Before joining the department, he was a postdoctoral fellow in the Information Theory and Applications (ITA) Center at the University of California, San Diego, from 2008 to 2011, and worked as a quantitative analyst with the D. E. Shaw group in New York from 2011 to 2013. Prior to his graduate studies, he served as an engineer and team leader in the Israeli Defense Forces from 1997 to 2003, and as an algorithms engineer at CellGuide from 2003 to 2004. Dr. Shayevitz is the recipient of the ITA postdoctoral fellowship (2009 - 2011), the Adams fellowship awarded by the Israel Academy of Sciences and Humanities (2006 - 2008), the Advanced Communication Center (ACC) Feder Family award (2009), and the Weinstein prize (2006 - 2009).