Interactive Coding for Markovian Protocols

Assaf Ben-Yishai, Ofer Shayevitz and Young-Han Kim

Abstract—We address the problem of simulating an arbitrary Markovian interactive protocol over binary symmetric channels with crossover probability ε. We are interested in the achievable rates of reliable simulation, i.e., in characterizing the smallest possible blowup in communications such that a vanishing error probability (in the protocol length) can be attained. Whereas for general interactive protocols the output of each party may depend on *all* previous outputs of its counterpart, in a (first order) Markovian protocol this dependence is limited to the last observed output only. In the special case where there is no dependence on previous outputs (no interaction), the maximal achievable rate is given by the (one-way) Shannon capacity $1 - h(\varepsilon)$. For Markovian protocols, we first show that a rate of $\frac{2}{3}(1 - h(\varepsilon))$ can be trivially achieved. We then describe a more involved coding scheme and provide a closed-form lower bound for its rate at any noise level ε. Specifically, we show that this scheme outperforms the trivial one for any $\varepsilon < 0.044$, and achieves a rate higher than $\frac{1-h(\varepsilon)}{1+h(\varepsilon)+h(<\varepsilon(2-\varepsilon)>)} = 1 - \Theta(h(\varepsilon))$ as $\varepsilon \to 0$, which is order-wise the best possible. This should be juxtaposed with a result of Kol and Raz that shows the capacity for interactive protocols with alternating rounds is lower bounded by $1 - O(\sqrt{h(\varepsilon)})$.

I. INTRODUCTION

Suppose Alice and Bob would like to communicate using some interactive communication protocol, where at time point i Alice sends the bit X_i^A and Bob then replies with the bit X_i^B (after having observed Alice's transmission). The transcript associated with their protocol is therefore

$$X_1^A, X_1^B, X_2^A, X_2^B, \cdots, X_n^A, X_n^B.$$

where

$$X_i^A = f_i^A\left(\mathbf{X}_1^{i-1,B}\right); \quad X_i^B = f_i^B\left(\mathbf{X}_1^{i,A}\right). \quad (1)$$

The *transmission functions* $f_i^A(\cdot)$, $f_i^B(\cdot)$ depend on the time index i and the identity of the speaker (Alice or Bob) and are unknown to the other party. In general, these functions may depend on the entire set of past inputs observed by either Alice or Bob, i.e. $\mathbf{X}_1^{i-1,B}$ or $\mathbf{X}_1^{i,A}$ respectively. We refer to the transcript $X_1^A, X_1^B, X_2^A, X_2^B, \cdots, X_n^A, X_n^B$ as the *clean transcript*, where "clean" is used to indicate that Alice and Bob receive their counterpart's transmission without any noise.

Suppose now that Alice and Bob are connected through two independent binary symmetric channels (BSCs) with parameter ε. Namely, Alice receives Bob's transmission with additive noise: $Y_i^B = X_i^B + Z_i^B$, and Bob received Alice's transmission with additive noise $Y_i^A = X_i^A + Z_i^A$, where $\{Z_i^A, Z_i^B\}$ are mutually independent Bernoulli i.i.d. sequence with $\Pr(Z_i^A = 1) = \Pr(Z_i^B = 1) = \varepsilon$ and "+" is addition over $\mathbb{GF}(2)$. Alice and Bob would like to devise a coding scheme that would allow them to reliably simulate the clean transcript over the noisy BSCs. Reliable simulation in this context means that for any Markovian protocol, the probability of either Alice or Bob making an error in recovering the clean transcript goes to zero with the transcript length. To that end, they will need to exchange a larger number of bits; the communication rate of their coding scheme is hence defined to be the total number of bits in the clean transcript divided by the total number of channel uses consumed by their scheme. As usual, one is interested in characterizing the *capacity*, namely the maximal rate for which reliable simulation is possible.

The problem described above was originally introduced and studied by Schulman [1]. In this seminal work, he showed that reliable simulation with a positive rate (i.e., a positive capacity) can be achieved for any $\varepsilon \neq 1/2$. Kol and Raz [2] further studied the problem in the limit of $\varepsilon \to 0$ and introduced a scheme achieving a rate of $1 - O(\sqrt{h(\varepsilon)})$ (where $h(\cdot)$ denotes the binary entropy function). They also showed that for a larger class of protocols with non-alternating rounds the rate is upper bounded by $1 - \Omega(\sqrt{h(\varepsilon)})$.

This demonstrated a separation between one-way and interactive communications, as the one-way capacity is given by $1 - h(\varepsilon)$. In [3], Haeupler examined a more flexible channel model than ours, in which at every time slot Alice and Bob can independently decide if they want to use the channel as a transmitter or as a receiver. This flexibility can potentially lead to collisions, but was shown to eventually increase the achievable rate to $1 - O(\sqrt{\varepsilon})$. Haeupler also conjectured that this rate is order-wise tight under adaptive transmission order, i.e., that the rate of any such reliable scheme is upper bounded by $1 - \Omega(\sqrt{\varepsilon})$. We note that the general problem of exactly determining the capacity for any fixed ε in the interactive setup is still wide open.

In order to better understand the gap between the one-way and interactive setups for $\varepsilon \to 0$, Haeupler and Velingker [4] considered a more restrictive family of protocols that are "less interactive", where Alice and Bob have some limited average lookahead, i.e., can often speak for a while without requiring further input from their counterpart (hence, can use short error correcting codes). They showed (also for adversarial noise) that when this average lookahead is $\text{poly}(1/\varepsilon)$ then the capacity is $1 - O(h(\varepsilon))$, i.e., is order-wise the same as the one-way capacity.

In this work, rather than restricting the "interactiveness" of the protocol as above, we restrict the *memory* of the protocol. Specifically, we consider *Markovian protocols* for which the lookahead can be as short as 1 (highly interactive), but where Alice and Bob need only recall the last bit they have received. For these Markovian Protocols, we provide lower bounds for the capacity for all values of ε, and not only in the limit $\varepsilon \to 0$.

A. Markovian Protocols

A (first order) Markovian protocol is a protocol in which each party needs to know only the last transmission of its

counterpart in order to decide what to send next, and not the entire set of past transmissions. Namely,

$$X_i^A = f_i^A(X_{i-1}^B); \quad X_i^B = f_i^B(X_i^A).$$

where now, in contrast to (1), the transmission functions $f_i^A(\cdot)$, $f_i^B(\cdot)$ depend only on what was last received (X_{i-1}^B and X_i^A respectively). It is important to note that the non-interactive communication problem is a special case where $f_i^A(\cdot)$, $f_i^B(\cdot)$ are a sequence of constant valued functions that do not depend on the output of the second party.

The *rate* of any communication scheme that attempts to simulate the clean transcript is defined by

$$R = \frac{2n}{\tilde{n}}$$

where $2n$ is the length of the clean transcript, and $\tilde{n}$ is the number of channel uses required by the scheme.

The probability of error attained by a scheme is defined to be the maximal probability that either Alice or Bob fail to exactly simulate the clean transcript, where the maximum is taken over all possible Markovian protocols. A sequence of schemes with rate at least R and error probability approaching zero is said to achieve the rate R. The capacity for Markovian protocols over BSCs is the supremum over all such achievable rates, and is denoted by $C_{\text{Markov}}(\varepsilon)$. Note that $C_{\text{Markov}}(\varepsilon)$ cannot exceed the one-way Shannon capacity of the BSC, i.e.,

$$C_{\text{Markov}}(\varepsilon) \leq 1 - h(\varepsilon), \tag{2}$$

as this is the maximal achievable rate for the special case of non-interactive protocols. Below we derive lower bounds on the Markovian capacity.

B. Main Result

Our main result is the following.

Theorem 1. *The capacity for Markovian protocols over BSCs with crossover probability ε is lower bounded by*

$$C_{\text{Markov}}(\varepsilon) \geq \max \left\{ R_0(\varepsilon), \sup_{K, M \in \mathbb{N}} R_1(\varepsilon, K, M) \right\}$$

where $R_0(\varepsilon) \stackrel{\text{def}}{=} \frac{2}{3}(1 - h(\varepsilon))$,

$$R_1(\varepsilon, K, M) \stackrel{\text{def}}{=} \frac{1 - h(\varepsilon)}{1 + h(\varepsilon) + \ell(\varepsilon, K, M)},$$

and $\ell(\varepsilon, K, M)$ is defined in (29). The following upper bounds for $\ell(\varepsilon, K, M)$ are easily computable and can be used to lower bound $R_1(\varepsilon, K, M)$. The first bound is

$$\ell(\varepsilon, K, M) \leq h\left(<\varepsilon(2 - \varepsilon)>\right)$$

(where $<x> \stackrel{\text{def}}{=} \min(x, \frac{1}{2})$) and the second, tighter upper bound is $\ell(\varepsilon, K, M) \leq \check{\ell}(\varepsilon)$ where

$$\check{\ell}(\varepsilon) = \sum_{k=1}^{\infty} (\varepsilon(2 - \varepsilon))^2 (1 - \varepsilon(2 - \varepsilon))^{k-1} \log(k + 1).$$

The rates $R_0(\varepsilon)$ and $R_1(\varepsilon, K, M)$, (with $K = 100, M = 400$) normalized by the BSC capacity $1 - h(\varepsilon)$, are plotted in Fig. 1. It can be seen that $R_1(\varepsilon, K, M)$ is superior for $\varepsilon < 0.044$, and $R_0(\varepsilon)$ is superior otherwise. Moreover, analyzing $R_1(\varepsilon, K, M)$ for small ε, the following can be shown:

Corollary 1. *For $\varepsilon \to 0$*

$$C_{\text{Markov}}(\varepsilon) = 1 - \Theta(h(\varepsilon)).$$

In light of the trivial upper bound (2), this rate is order-wise the best possible. Moreover, it is order-wise higher than the lower bound of $1 - O(\sqrt{h(\varepsilon)})$ obtained by Kol and Raz [2] for interactive protocols with alternating rounds a non-adaptive transmission schedule.

The remainder of the paper is dedicated to the proof of Theorem 1 and its corollary, and is organized follows: In Subsection II-A, we present Scheme #1, which is a very simple scheme that achieves $R_0(\varepsilon)$. In Subsection II-B, we present Scheme #2 which is more involved and achieves $\sup_{K,M} R_1(\varepsilon, K, M)$ which is larger than $R_0(\varepsilon)$ for any $\varepsilon < 0.044$. The analysis of Scheme #2, including the description of a designated compression protocol, its behavior for large n, and numeric evaluation of $R_1(\varepsilon, K, M)$ are given in Section III.

II. CODING SCHEMES

A. Scheme #1

We observe that the transmission functions $f_i^A(\cdot)$, $f_i^B(\cdot)$, are binary functions that map a single input bit to a single output bit. We note that there are only four such functions, μ_1, μ_2, μ_3, μ_4 as in the following table:

Y	μ_1: $X = Y + 0$	μ_2: $X = Y + 1$	μ_3: $X = 0$	μ_4: $X = 1$
0	0	1	0	1
1	1	0	0	1

We observe that μ_1 and μ_2 are *linear*, i.e. $X = Y + \xi$ and ξ is either 0 or 1. μ_3 and μ_4 are constant functions, namely, the output is 0 or 1 regardless the input. In the sequel we refer to the locations where μ_3 and μ_4 are used as "stuck positions".

Having this simple notion stated, we note both and Alice and Bob can compress their own transmission functions using $2n$ bits. We also note that every party, having the transmission functions of its counterpart, can simulate the entire clean transcript. So, we can state the following reliable interaction protocol:

1) Alice compresses all her transmission function using $2n$ bits
2) Alice sends them to Bob using a capacity achieving channel code with rate $1 - h(\varepsilon)$. The number of required transmissions from Alice to Bob at this step is $2n/(1 - h(\varepsilon) + o(1))$ with error probability $O(1/\text{poly}(n))$.
3) Bob, having all Alice's transmission functions, can simulate the clean transcript.
4) Bob can feed his side of the transcript to Alice, requiring n information bits over a channel with capacity $1 - h(\varepsilon)$. So overall $n/(1 - h(\varepsilon) + o(1))$ channel uses are needed, with error probability $O(1/\text{poly}(n))$.

So, overall $\tilde{n} = 3n/(1 - h(\varepsilon) + o(1))$ channel uses are required (with error probability $O(1/\text{poly}(n))$) hence the rate is

$$R_0(\varepsilon) = \frac{2n}{3n/(1 - h(\varepsilon))} = \frac{2}{3}(1 - h(\varepsilon)).$$

B. Scheme #2

The improved achievable rate introduced here is based on running the protocol disregarding the channel errors (as if the channels were clean), followed by several rounds designated to correct the errors. This scheme is found to be better that

to the trivial scheme when the channel noise is low. We start by running the "clean" protocol, namely Alice and Bob use the Markovian transmission functions on their noisy inputs, $X_i^A = f_i^A(Y_{i-1}^B)$ and $X_i^B = f_i^B(Y_i^A)$ requiring $2n$ channel uses. Then, Alice can describe to Bob the errors of the first round using Slepian-Wolf [5] coding protected by a channel code. After this step, the stuck positions are transmitted from side to side using a designated compression algorithm. Finally, the protocol is corrected, using the linearity of the transmission functions in places where they are linear, and reseting at stuck position (as will be elaborated in the sequel).

Let us summarize these steps and give the rate calculation:

1) Both parties perform interaction disregarding the channel errors. Overall $2n$ channel uses.

2) Alice describes Bob the errors that occurred on the channel connecting them (i.e. the channel from Alice to Bob) using Slepian-Wolf coding over a noisy channel. This step requires $n(h(\varepsilon) + o(1))/(1 - h(\varepsilon) + o(1))$ channel uses (with error probability $O(1/\text{poly}(n))$). Then Bob feeds the errors back to Alice using simple typical set coding (not Slepian-Wolf). These steps are repeated replacing the roles of Alice and Bob. All in all the channel are used $4n(h(\varepsilon) + o(1))/(1 - h(\varepsilon) + o(1))$ times (with error probability $O(1/\text{poly}(n))$). At the end of this step both parties are aware of all channel errors on both sides.

3) Bob, knowing all channel errors on both channels divides his interaction functions ,$f_i^A(\cdot)$, into segments that start and end with a channel error (on either channel direction). Then, the first "stuck position" (i.e. μ_3 of μ_4) is conveyed to Alice using the protocol elaborated in Subsection III-A. The maximal (i.e. worst case) number of bits used for the description is denoted by $n\ell(\varepsilon)$ and should be conveyed using a capacity achieving channel code requiring $n\ell(\varepsilon)/(1 - h(\varepsilon) + o(1))$ channel uses in total.

4) Having all this data, Alice can simulate Bob's clean transcript. Assume that from $1 \leq i \leq j$ both Alice and Bob have only linear transmission functions. Then, due to the linearity of the transmission at both parties, Bob's clean transcript $\hat{X}_i^B$ can be simulated by canceling the error at both sides:

$$\hat{X}_i^B = Y_i^B + \sum_{l=1}^{i} Z_l^A + \sum_{l=1}^{i} Z_i^B.$$

5) Whenever there is a "stuck position" for either party, the processing of previous errors is reset. For example, if Alice receives $Y_i^B = 0$, and knows the value of Z_i^B and the fact that f_i^B is either μ_3 or μ_4, then $X_i^B = Y_i^B + Z_i^B$, disregarding previous noise values. Note that in non-stuck positions X_i^B is not necessarily equal to $Y_i^B + Z_i^B$. This is because X_i^B is defined as Bob's transmission in the hypothetical noiseless interaction, and not as his transmission in step 1.

6) Steps 3,4 and 5 are repeated by appropriately exchanging the roles of Alice and Bob.

The rate attained by this scheme is therefore

$$R_1(\varepsilon, K, M) = \frac{2n}{2n + n(4h(\varepsilon) + 2\ell(\varepsilon, K, M))/(1 - h(\varepsilon))}$$

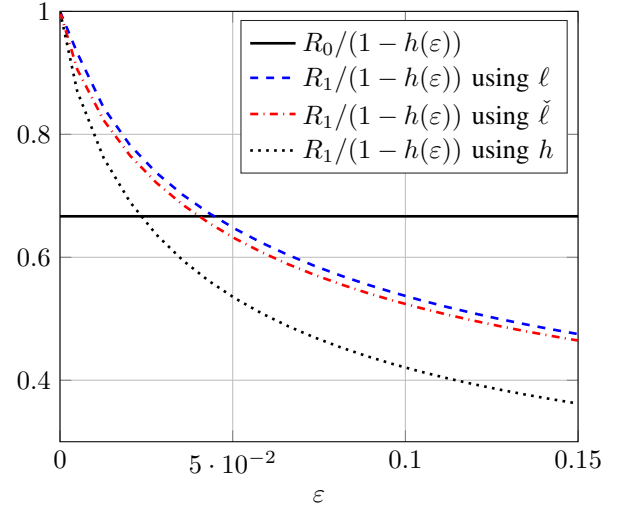$$= \frac{1 - h(\varepsilon)}{1 + h(\varepsilon) + \ell(\varepsilon, K, M)}$$

Fig. 1. Achievable rates normalized by $1 - h(\varepsilon)$. R_1 is computed with $K = 100, M = 400$.

In the sequel, we will be mostly concerned with the computation of the achievable rate R_1. We will also provide a simple lower bound on R_1 which is easier to compute, by upper bounding $\check{\ell}(\varepsilon) \geq \ell(\varepsilon)$ (see (30)):

$$\sup_{K,M} R_1(\varepsilon, K, M) \geq \frac{1 - h(\varepsilon)}{1 + h(\varepsilon) + \check{\ell}(\varepsilon)}.$$

The achievable rates are depicted in Fig. 1. R_1 is computed using $\ell(\varepsilon, K, M)$ and two corresponding upper bounds $\check{\ell}(\varepsilon)$ and a trivial entropy bound that is elaborated in the next section.

It is important to note that for $\varepsilon > 0.044$ the description of the errors and stuck positions in scheme #2 causes it to be less efficients than scheme #1 as seen in the figure. On the other hand, $\sup_{K,M} R_1(\varepsilon, K, M)$ is better that $R_0(\varepsilon)$ when the channel noise is low and approach 1 as the ε go to zero.

It is of interest to compare this results to [2]. Taking the trivial upper bound $\ell(\varepsilon) \leq h(<\varepsilon(2 - \varepsilon)>)$ given in Subsection III-A we can assess the behavior for small ε by:

$$\sup_{K,M} R_1(\varepsilon, K, M) \geq \frac{1 - h(\varepsilon)}{1 + h(\varepsilon) + h(\varepsilon(2 - \varepsilon))} = 1 - \Theta(h(\varepsilon)),$$

hence, the capacity for Markovian protocols scales like the Shannon capacity in this limit. This should be juxtaposed with the upper bound (for general protocols) of $1 - \Omega(\sqrt{h(\varepsilon)})$ given in [2]. This shows a gap between the capacities of general protocols and Markovian protocols. We note that [2] assumes non-adaptive transmission order, which is satisfied by our scheme.

It was shown in [3] a higher rate of $1 - O(\sqrt{\varepsilon})$ can be achieved for general protocols under adaptive transmission order. This rate is still outperformed by our scheme (for Markovian protocols).

III. ANALYSIS OF SCHEME #2

In this section we analyze the performance of the scheme introduced in II-B. In particular, we define and analyze a novel compression algorithm designated for the compression of the stuck position.

A. Compression of the Stuck Positions

We consider the fixed binary sequence $\phi^n = (\phi_1, \ldots, \phi_n)$, $\phi_i \in \{0, 1\}$ which describes the "stuck positions" in the original problem. Namely, ϕ^n describes Bob's "stuck positions", and is equal to 1 if $f_i^B(\cdot) = \mu_3 = 0$ or $f_i^B(\cdot) = \mu_4 = 1$. We also consider the i.i.d random sequence $\mathbf{z}^n = (z_1, \ldots, z_n)$, $z_i \in \{0, 1\}$ with marginal probability $\Pr(z_i = 1) = p$, where p is the probability that there is at least one error on the channel from Alice to Bob or vice versa, i.e. $p = 1 - (1 - \varepsilon)^2 = \varepsilon(2 - \varepsilon)$.

It is useful to think of the interlaced picture:

$$z_1 \qquad z_2 \qquad z_3 \qquad \cdots$$
$$\phi_1 \qquad \phi_2 \qquad \phi_3 \quad \cdots$$

The sequence $\mathbf{z}^n$ is parsed into segments of the form $(1, \mathbf{0}^{k-1})$, $k > 0$, where $\mathbf{0}^{k-1}$ denotes a row vector of zeros with $k - 1$ elements.

We wish to describe the position of the first $\phi_j = 1$ in every segment. For example, consider the following interlaced sequence

$$\begin{array}{ccccccc} \mathbf{z} = & 1 & 0 & 0, & 1 & 0, & 1, & 1 \\ \phi = & 0 & ① & 1 & ① & 1 & ① \end{array}$$

The parsed segments are separated by commas, and the appearances of the first $\phi = 1$ are circled.

First, we note that the total number of the first stuck positions is trivially upper bounded by the number of segments, which is the total number of errors. So, the total number of the first stuck positions is with high probability smaller than $n(p + o(1))$ and can be described via universal compression using less than $n(h(<p>) + o(1))$ bits. We note that this naive compression method does not use the fact that both sides know the error positions and can take advantage of them in order to improve the compression rate.

An improved compression algorithm can use the knowledge of the vector $\mathbf{z}$ as follows: segments of length k are grouped and the empirical distribution of the appearance of the first 1 is calculated. Then, universal compression is applied for every k based on these distributions. We denote the vector of empirical distribution related to segments of length k by $\boldsymbol{\pi}_k = \{\pi_{k,l}\}_{l=0}^k$. The first k elements of this vector comprise the fraction of these segments that start at some z_i, and whose first appearance of $\phi_j = 1$ thereafter is at $j = i + l$. $\pi_{k,k}$ is the fraction of the segments that contain no $\phi_j = 1$.

Let L denote the overall length of the stuck positions description (with high probability), normalized by n. In the sequel we shall prove that L converges to an asymptotic value $\bar{L}$, which is more easily computable.

First, we define the empirical distribution $\pi_{k,l}$ (for $0 \leq l \leq k$) as the ratio between the counters $N_{k,l}$ and N_k:

$$\pi_{k,l} = \frac{N_{k,l}}{N_k}. \tag{3}$$

The counters $N_{k,l}$ are defined as:

$$N_{k,l} \stackrel{\text{def}}{=} \sum_{i=1}^n \mathbb{1}_{k,l}(i),$$

where indicator $\mathbb{1}_{k,l}(i)$ for $0 \leq l < k$ is one only if and only if $\mathbf{z}_i^{i+k} = (1, \mathbf{0}^{k-1}, 1)$ and

$$\phi_{i+j} = \begin{cases} 1 & \text{for } j = l \\ 0 & \text{for } 0 \leq j < l. \end{cases}$$

The indicator $\mathbb{1}_{k,k}(i)$ is one only if $\mathbf{z}_i^{i+k} = (1, \mathbf{0}^{k-1}, 1)$ and $\phi_{i+j} = 0$ for $0 \leq j < k$. The denominator of (3) is defined as

$$N_k \stackrel{\text{def}}{=} \sum_{l=0}^k N_{k,l} = \sum_{i=1}^n \mathbb{1}\left(\mathbf{z}_i^{i+k} = (1, \mathbf{0}^{k-1}, 1)\right)$$

where the second equality is by construction.

Having the counters and the resulting empirical distribution vectors $\boldsymbol{\pi}_k$, we can calculate the average description length L. It is useful to use two schemes, one for $k \leq K_n$ and one for $k > K_n$, with K_n defined in the sequel. For $k < K_n$ we use universal compression which requires for every k: $N_k H(\boldsymbol{\pi}_k)$ bits for the compression where $H(\cdot)$ is the entropy function of a probability vector. Additional bits are also required for the lossless description of the probability vectors $\boldsymbol{\pi}_k$ for $k \leq K_n$. We denote this number of bits by W.

For $k > K_n$ we describe the location of the first stuck position using the simplifying assumption that $\pi_{k,l} = \frac{1}{k+1}$ (for all $0 \leq l \leq k$), shared by both the receiver and transmitter. So, the number of bits for every value of k is $\lceil \log(k+1) \rceil$. All in all, the average description length L is

$$L = \frac{1}{n} \left[\sum_{k=1}^{K_n} N_k H(\boldsymbol{\pi}_k) + W + \sum_{k=K_n+1}^n N_k \lceil \log(k+1) \rceil \right]$$

It is useful write L as

$$L = S_1 + \frac{W}{n} + S_2 \tag{4}$$

where

$$S_1 \stackrel{\text{def}}{=} \sum_{k=1}^{K_n} \frac{N_k}{n} H(\boldsymbol{\pi}_k), \tag{5}$$

$$S_2 \stackrel{\text{def}}{=} \sum_{k=K_n+1}^n \frac{N_k}{n} \lceil \log(k+1) \rceil. \tag{6}$$

In the sequel we prove that L converges to its asymptotic value by proving that the counters $N_{k,l}$ and N_k converge to their expected values. It is now useful to introduce "spectrum vector" $\{a_m\}$, and write $\mathbb{E}N_{k,l}$ and $\mathbb{E}N_k$ as functions of this vector. Let

$$a_m = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\left(\phi_{i-m} = 1, \phi_{i-m+1}^{i-1} = \mathbf{0}^{m-1}, \phi_i = 1\right) \tag{7}$$

for $m = 1, \ldots, n$. Namely, a_m is the fraction of elements in ϕ^n which are equal to 1, and their nearest preceding 1 in $\mathbf{z}^n$ is exactly m time instances earlier. In order to take care of the edge effects we set $\phi_0 = 1$ and $\phi_i = 0$ for $i < 0$.

Let us now calculate the related expectations:

$$\mathbb{E}N_k = \sum_{i=1}^n \mathbb{E}\mathbb{1}\left(\mathbf{z}_i^{i+k} = (1, \mathbf{0}^{k-1}, 1)\right) = np^2(1-p)^{k-1} \tag{8}$$

And for $0 \leq l < k$

$$\mathbb{E}N_{k,l} = \sum_{i=1}^n \mathbb{E}\left[\mathbb{1}_{k,l}(i)\right]$$

$$= \sum_{i=1}^n p^2(1-p)^{k-1}\mathbb{1}\left(\phi_i^{i+l-1} = \mathbf{0}^{l-1} \text{ AND } \phi_{i+l} = 1\right)$$

$$\stackrel{(a)}{=} p^2(1-p)^{k-1}n \sum_{m=l+1}^n a_m$$

where (a) follows by counting the number of elements in ϕ^n that are one, and whose distance to their preceding one is more than $l+1$ (note that l starts at zero).

The calculation of $\mathbb{E}N_{k,k}$ is different:

$$
\begin{aligned}
\mathbb{E}N_{k,k} &= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{k,k}(i)\right] \\
&= p^2(1-p)^{k-1} \sum_{i=1}^{n} \mathbb{1}\left(\phi_i^{i+k-1} = \mathbf{0}^k\right) \\
&\overset{(a)}{=} p^2(1-p)^{k-1} n \sum_{m=k+1}^{n} a_m(m-k)
\end{aligned}
$$

The equality (a) follows by observing that for every ϕ segment $(1, \mathbf{0}^{m-1}, 1)$ there exist $m-k$ placements of a z sequence $(1, \mathbf{0}^{k-1}, 1)$ that contain no $\phi = 1$. This notion is illustrated below:

$$
\begin{aligned}
\mathbf{z} &= \quad 1\overset{k-1}{\overbrace{0\cdots0}}1 \\
\phi &= \quad 1\underset{m-1}{\underbrace{0\cdots\cdots\cdots0}}1
\end{aligned}
$$

It is also easy to verify that $\mathbb{E}N_k = \sum_{l=0}^{k} \mathbb{E}N_{k,l}$. Let us define the probabilities

$$
\begin{aligned}
\bar{\pi}_{k,l} &\overset{\text{def}}{=} \frac{\mathbb{E}N_{k,l}}{\mathbb{E}N_k} \\
&= \begin{cases} \sum_{m=l+1}^{n} a_m, & \text{for } 0 \le l < k \\ \sum_{m=k+1}^{n} a_m(m-k), & \text{for } 0 \le l = k \end{cases}
\end{aligned} \quad (9)
$$

and define $\bar{L}$ based on the definition of S_1 in (5), replacing N_k with $\mathbb{E}N_k$, $\boldsymbol{\pi}_k$ with $\bar{\boldsymbol{\pi}}_k$:

$$
\bar{L} = \frac{1}{n} \sum_{k=1}^{K_n} np^2(1-p)^{k-1} H\left(\bar{\boldsymbol{\pi}}_k\right)
$$

We are now ready to state Theorem 2.

B. Convergence of the Compression Rate

Theorem 2. *For $K_n = \frac{\beta \ln n}{-\ln(1-p)}$ for every $\varepsilon > 0$*

$$
\lim_{n\to\infty} \Pr\left(L > \bar{L} + \varepsilon\right) = 0.
$$

Recalling (4), L is composed of three elements : S_1, W/n and S_2. Proving that W/n and S_2 converge to zero is simple and is deferred to the end of this subsection. Analyzing the convergence of S_1 is more involved and is now handled. The proof is based on two elements: the convergence of the counters $N_{k,l}$ to their expected value (Lemma 1), and the smoothness of the entropy function (Lemma 2). Let us start by giving the lemmas and then use them to prove the theorem.

Lemma 1. *The following inequalities hold any $t \ge 0$:*

$$
\Pr\left(|N_{k,l} - \mathbb{E}N_{k,l}| \ge t\right) \le 2\exp\left(-\frac{t^2}{16n}\right) \quad (10)
$$

$$
\Pr\left(|N_k - \mathbb{E}N_k| \ge t\right) \le 2\exp\left(-\frac{t^2}{16n}\right) \quad (11)
$$

$$
\Pr\left(\sum_{k=K+1}^{n} N_k - \mathbb{E}\sum_{k=K+1}^{n} N_k \ge t\right) \le \exp\left(-\frac{t^2}{4n}\right) \quad (12)
$$

Proof. The proof is based on a straightforward application of the bounded difference inequality. We start by citing the inequality:

Theorem 3 (*Bounded difference inequality* [6, Theorem 3.18]). *Let $\mathbf{x}^n$ be a random independent series, and $f(\mathbf{x}^n)$ a scalar function, then for any $t \ge 0$ the following hold:*

$$
\Pr(f(\mathbf{x}^n) - \mathbb{E}f(\mathbf{x}^n) \ge t)
$$
$$
\le \exp\left(-\frac{t^2}{4\left\|\sum_{i=0}^{n} |D_i^- f|^2\right\|_\infty}\right) \quad (13)
$$
$$
\Pr(f(\mathbf{x}^n) - \mathbb{E}f(\mathbf{x}^n) \le -t)
$$
$$
\le \exp\left(-\frac{t^2}{4\left\|\sum_{i=0}^{n} |D_i^+ f|^2\right\|_\infty}\right). \quad (14)
$$

where

$$
D_i^- f \overset{\text{def}}{=} f(\mathbf{x}^n) - \inf_x f(\mathbf{x}^{i-1}, x, \mathbf{x}_{i+1}^n)
$$

and

$$
D_i^+ f \overset{\text{def}}{=} \sup_x f(\mathbf{x}^{i-1}, x, \mathbf{x}_{i+1}^n) - f(\mathbf{x}^n).
$$

We shall use the theorem by setting $f(\mathbf{x}^n) = N_{k,l}$ where $\mathbf{x}^n$ is the noise series $\mathbf{z}^n$ (i.i.d $\text{Ber}(p)$). Using this, the elements of $\mathbf{z}^n$ determine the error segments in which the counters $N_{k,l}$ are calculated. We observe that changing a single element of $\mathbf{z}^n$ can leave the number of segments unchanged or change them by at most two. The maximal change is achieved in the following situation:

$$
\mathbf{z}^n = (\ldots, 1, \mathbf{0}^{k-1}, x_i, \mathbf{0}^{k-1}, 1 \ldots)
$$

in which changing x_i from zero to one (respectively from one to zero) will increase (respectively decrease) the number of segments by two. Since changing the number of segments by two will change the counter $N_{k,l}$ by at most two we can conclude that $D_i^+ f \le 2$ and $D_i^- f \le 2$, and

$$
\left\|\sum_{i=0}^{n} |D_i^- f|^2\right\|_\infty \le n (2)^2 = 4n
$$

and similarly $\left\|\sum_{i=0}^{n} |D_i^+ f|^2\right\|_\infty$ can be bounded by the same value. Using this and (13) and (14) we obtain (10). Note that the same bound and the same argument also holds for the total number of segments N_k expressed in (11).

Finally, (12) follows from the fact that changing x_i from one to zero will create at most one new segment with $k \ge K+1$. So $D_i^+ \le 1$ and $\left\|\sum_{i=0}^{n} |D_i^+ f|^2\right\|_\infty \le n$. $\qquad\square$

Lemma 2 ([7, Lemma 2.7]). *If $d(P, Q) = \Theta \le \frac{1}{2}$ then*

$$
|H(P) - H(Q)| \le -\Theta \log \frac{\Theta}{|\mathcal{X}|}
$$

where $d_{\text{TV}}(P, Q)$ is the total variation distance between the distributions P and Q on $\mathcal{X}$:

$$
d_{\text{TV}}(P, Q) \overset{\text{def}}{=} \sum_{x\in\mathcal{X}} |P(x) - Q(x)|.
$$

Having these two lemmas at hand, we are ready to prove the Theorem 2.

Proof of Theorem 2. We start by proving that S_1 is asymptotically upper bounded by $\bar{L}$. We first upper bound $|H(\boldsymbol{\pi}_k) - H(\bar{\boldsymbol{\pi}}_k)|$ by upper bounding the variation distance

$\Theta = |\pi_k - \bar{\pi}_k|$. Using (10) and (11) and the union bound it follows that

$$\Pr\left(\frac{N_{k,l}}{N_k} \geq \frac{\mathbb{E}N_{k,l} + t}{\mathbb{E}N_k - t}\right) \leq 4\exp\left(-\frac{t^2}{16n}\right).$$

where $t < \mathbb{E}N_k$. Using (3), (9) and the fact that $\mathbb{E}N_k = np^2(1-p)^{k-1}$ we can also write

$$\Pr\left(\frac{N_{k,l}}{N_k} \geq \frac{\mathbb{E}N_{k,l} + t}{\mathbb{E}N_k - t}\right) = \Pr\left(\pi_{k,l} \geq \frac{np^2(1-p)^{k-1}\bar{\pi}_{k,l} + t}{np^2(1-p)^{k-1} - t}\right)$$

$$= \Pr\left(\pi_{k,l} - \bar{\pi}_{k,l} \geq \frac{t(\bar{\pi}_{k,l} + 1)}{np^2(1-p)^{k-1} - t}\right)$$

and hence

$$\Pr\left(\pi_{k,l} - \bar{\pi}_{k,l} \geq \frac{t(\bar{\pi}_{k,l}+1)}{np^2(1-p)^{k-1} - t}\right) \leq 4\exp\left(-\frac{t^2}{16n}\right). \quad (15)$$

Similarly

$$\Pr\left(\frac{N_{k,l}}{N_k} \leq \frac{\mathbb{E}N_{k,l} - t}{\mathbb{E}N_k + t}\right)$$

$$= \Pr\left(\pi_{k,l} - \bar{\pi}_{k,l} \leq -\frac{t(\bar{\pi}_{k,l}+1)}{np^2(1-p)^{k-1} + t}\right)$$

$$\leq 4\exp\left(-\frac{t^2}{16n}\right). \quad (16)$$

Combining (15) and (16) taking into account that (15) is tighter, we obtain

$$\Pr\left(|\pi_{k,l} - \bar{\pi}_{k,l}| \geq \frac{t(\bar{\pi}_{k,l}+1)}{np^2(1-p)^{k-1} - t}\right) \leq 8\exp\left(-\frac{t^2}{16n}\right).$$

Summing up for $l = 1, \ldots, k$ and using the fact that $\sum_{i=0}^{k} \bar{\pi}_{k,l} = 1$ we get the following inequality for the variation distance $\Theta = d_{TV}(\pi_k, \bar{\pi}_k)$:

$$\Pr\left(\Theta \geq \frac{t(k+2)}{np^2(1-p)^{k-1} - t}\right) \leq 8(k+1)\exp\left(-\frac{t^2}{16n}\right).$$

Now, we can set $t = n^\alpha$ with $\alpha \in (\frac{1}{2}, 1)$ and obtain

$$\Pr\left(\Theta \geq \frac{k+2}{n^{1-\alpha}p^2(1-p)^{k-1} + 1}\right) \leq 8(k+1)\exp\left(-\frac{n^{2\alpha-1}}{16}\right).$$

Finally, Lemma 2 implies that

$$\Pr\left(|H(\pi_k) - H(\bar{\pi}_k)| \geq \varepsilon_k\right) \quad (17)$$

$$\leq 8(k+1)\exp\left(-\frac{n^{2\alpha-1}}{16}\right).$$

where

$$\varepsilon_k \stackrel{\text{def}}{=} -\frac{k+2}{n^{1-\alpha}p^2(1-p)^{k-1} - 1}\log\left(\frac{(k+2)/(k+1)}{n^{1-\alpha}p^2(1-p)^{k-1} - 1}\right).$$

Clearly, for any fixed k we have that $\varepsilon_k \stackrel{n\to\infty}{\longrightarrow} 0$. Trivially, the upper side of the bound in (17) also holds:

$$\Pr(H(\pi_k) - H(\bar{\pi}_k) \geq \varepsilon_k) \leq 8(k+1)\exp(-\frac{n^{2\alpha-1}}{16})(18)$$

Let us now bound the summands of S_1. Using (18), (10) and (8) and the union bound implies that

$$\Pr\left(\frac{N_k}{n}H(\pi_k) \geq (p^2(1-p)^{k-1} + n^{\alpha-1})(H(\bar{\pi}_k) + \varepsilon_k)\right)$$

$$\leq (8k+10)\exp\left(-\frac{n^{2\alpha-1}}{16}\right). \quad (19)$$

Rearranging (19) we obtain

$$\Pr\left(\frac{N_k}{n}H(\pi_k) - p^2(1-p)^{k-1}H(\bar{\pi}_k)\right.$$

$$\geq H(\bar{\pi}_k)n^{\alpha-1} + (p^2(1-p)^{k-1} + n^{\alpha-1})\varepsilon_k\bigg)$$

$$\leq (8k+10)\exp\left(-\frac{n^{2\alpha-1}}{16}\right).$$

Summing up for $k = 1, \ldots, K_n$ and noticing that the following bounds hold for $1 \leq k \leq K_n$:

$$p^2(1-p)^{k-1} \leq p^2$$
$$H(\bar{\pi}_k) \leq \log(K_n + 1)$$
$$\varepsilon_k \leq \varepsilon_{K_n}$$

we obtain

$$\Pr\left(S_1 - \bar{L} \geq \epsilon_1\right) \leq \delta_1. \quad (20)$$

where

$$\epsilon_1 \stackrel{\text{def}}{=} K_n \log(K_n + 1)n^{\alpha-1} + (p^2 + n^{\alpha-1})K_n\varepsilon_{K_n}$$

and

$$\delta_1 \stackrel{\text{def}}{=} K_n(8K_n + 10)\exp\left(-\frac{n^{2\alpha-1}}{16}\right).$$

Setting

$$K_n = \frac{\beta \ln n}{-\ln(1-p)} \quad (21)$$

with $\beta \in (0, 1-\alpha)$ yields

$$(1-p)^{K_n} = n^{-\beta} \quad (22)$$

which assures that $\varepsilon_{K_n} \stackrel{n\to\infty}{\longrightarrow} 0$ and also $\epsilon_1 \stackrel{n\to\infty}{\longrightarrow} = 0$ and $\delta_1 \stackrel{n\to\infty}{\longrightarrow} 0$.

Let us now prove that S_2 converges to zero in probability. For $k > K_n$ we describe the location of the first stuck position using $\lceil \log(k+1) \rceil$ bits for every value of k. Therefore

$$S_2 = \sum_{k=K_n+1}^{n} \frac{N_k}{n} \lceil \log(k+1) \rceil$$

$$\leq \sum_{k=K_n+1}^{n} \frac{N_k}{n} (\log(k+1) + 1)$$

$$\leq \left(\sum_{k=K_n+1}^{n} \frac{N_k}{n}\right)(\log(n+1) + 1) \quad (23)$$

Recalling (12) and using only the upper side of the bound

$$\Pr\left(\sum_{k=K_n+1}^{n} \frac{N_k}{n} \geq \sum_{k=K+1}^{n} p^2(1-p)^{k-1} + \frac{t}{n}\right)$$

$$\leq 2\exp\left(-\frac{t^2}{4n}\right).$$

and noticing that

$$\sum_{k=K+1}^{n} p^2(1-p)^{k-1} \leq \sum_{K+1}^{\infty} p^2(1-p)^{k-1} = p(1-p)^{K_n}$$

we have

$$\Pr\left(\sum_{k=K_n+1}^{n}\frac{N_k}{n}\geq p(1-p)^{K_n}+\frac{t}{n}\right)\leq 2\exp\left(-\frac{t^2}{4n}\right)$$

setting as before $t=n^\alpha$ with $\alpha\in(\frac{1}{2},1)$ and recalling (22) we obtain

$$\Pr\left(\sum_{k=K_n+1}^{n}\frac{N_k}{n}\geq pn^{-\beta}+n^{\alpha-1}\right)\leq 2\exp\left(-\frac{n^{2\alpha-1}}{4}\right)$$

Now, we use (23) and further loosen the bound, obtaining

$$\Pr\left(S_2\geq(pn^{-\beta}+n^{\alpha-1})\left(\log(n+1)+1\right)\right)$$
$$\leq 2\exp(-\tfrac{n^{2\alpha-1}}{4}).$$

Defining

$$\epsilon_2\overset{\text{def}}{=}\left(pn^{-\beta}+n^{\alpha-1}\right)\log(n+1)$$

and

$$\delta_2\overset{\text{def}}{=}2\exp\left(-\frac{n^{2\alpha-1}}{4}\right).$$

we obtain

$$\Pr\left(S_2\geq\epsilon_2\right)\leq\delta_2 \tag{24}$$

where clearly $\epsilon_2\overset{n\to\infty}{\longrightarrow}=0$ and $\delta_2\overset{n\to\infty}{\longrightarrow}0$.

Lastly, we can show that W/n converges to zero in probability by representing the values in $\boldsymbol{\pi}_k$ for $k=1,\ldots,K_n$ using $\log(K_n)$ bits each. There are overall $\sum_{k=1}^{K_n}(k+1)=K_n(K_n+3)/2$ such elements so, the total number of required bits is $W=K_n(K_n+3)\log(K_n+1)$. Setting K_n as in (21) clearly yields $W/n\overset{n\to\infty}{\longrightarrow}=0$. Combining this, (20), (24) and applying the union concludes the proof. $\qquad\square$

C. Numerical Evaluations of the Compression Rate

In the previous subsection, we proved that L is asymptotically upper bounded by $\bar{L}$. However, $\bar{L}$ is a function of the spectrum vector $\mathbf{a}$. Therefore, an upper bound for $\bar{L}$ should be related to the maximization of $\bar{L}$ w.r.t $\mathbf{a}$. In this subsection we explicitly write this (convex) optimization problem, and provides some numeric evaluations. We note that $\mathbf{a}$ is a vector of length $n\to\infty$. Since our optimization tools are limited to vectors with finite dimension, we limit the size of $\mathbf{a}$, and bound the residue inflicted by this process.

We start by recalling (4): $L=S_1+\frac{W}{n}+S_2$, however, in contrast to the definitions in (5) and (6), we define S_1 and S_2 with K that is a fixed number, and not an increasing function in n. Namely

$$S_1=\sum_{k=1}^{K}\frac{N_k}{n}H\left(\boldsymbol{\pi}_k\right),$$

$$S_2=\sum_{k=K+1}^{n}\frac{N_k}{n}\left\lceil\log(k+1)\right\rceil.$$

Having a fixed K, the number of bits required for the description of the universal codebooks, W can be trivially upper bounded by $K^2\log(K+1)$ thus clearly $\frac{W}{n}\overset{n\to\infty}{\longrightarrow}0$. In the previous subsection, we showed that L is asymptotically upper bounded by $\bar{L}$, for K_n defined in (21). It is possible to show by steps similar to the ones used in the previous

subsection that for a fixed K, S_1 and S_2 are asymptotically upper bounded by the following terms respectively

$$\bar{S}_1(p,K)\overset{\text{def}}{=}\sum_{k=1}^{K}p^2(1-p)^{k-1}H(\bar{\boldsymbol{\pi}}_k),$$

$$\bar{S}_2(p,K)\overset{\text{def}}{=}\sum_{k=K+1}^{n}p^2(1-p)^{k-1}\left\lceil\log(k+1)\right\rceil.$$

Thus the total description length can be written as

$$L(p,K)=\bar{S}_1(p,K)+\bar{S}_2(p,K)$$

We first note that we can trivially upper bound all $H(\bar{\boldsymbol{\pi}}_k)$ by $\log(k+1)$ yielding the following bound

$$L(p,K)\leq\check{L}(p)\overset{\text{def}}{=}\sum_{k=1}^{\infty}p^2(1-p)^{k-1}\log(k+1). \tag{25}$$

Let us now write $\bar{S}_1(p,L)$ as a convex optimization problem in $\mathbf{a}$, and numerically evaluate its optimum. We recall that $\bar{\boldsymbol{\pi}}_k$ can be written in terms of $\mathbf{a}$ as in (9). This relation can be stated in using matrix/vector notation by introducing the set of matrices B_k with sizes $(k+1)\times n$ with the following element. For $1\leq i\leq k$

$$[B_k]_{i,j}=\begin{cases}1 & \text{for } j\geq i\\0 & \text{otherwise}\end{cases}$$

and for $i=k+1$

$$[B_k]_{k+1,j}=\begin{cases}j-k & \text{for } j\geq k\\0 & \text{otherwise}\end{cases}$$

The matrix B_k can also be written as follows:

$$B_k=\begin{array}{c}\\1\\2\\3\\\vdots\\k\\k+1\end{array}\begin{array}{c}\begin{array}{ccccccc}1 & 2 & 3 & \cdots & k & k+1 & k+2\end{array}\\\left[\begin{array}{ccccccc}1 & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots\\0 & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots\\0 & 0 & 1 & \cdots & 1 & 1 & 1 & \cdots\\ & & & \ddots & & & \\0 & 0 & 0 & \cdots & 1 & 1 & 1 & \cdots\\0 & 0 & 0 & \cdots & 0 & 1 & 2 & \cdots\end{array}\right]\end{array}$$

Recalling the definition of $\{a_m\}$ in (7), and taking into account that all the sequences of all lengths $m=1,\ldots,n$ construct the original sequence ϕ^n (whose length is n) gives $\sum_{m=1}^{n}mna_m=n$ hence $\sum_{m=1}^{n}ma_m=1$. Also note that $a_m\geq 0$ for all $m=1,\ldots,n$.

So, an upper bound for $\bar{S}_1(p,K)$ denoted by $\tilde{S}_1(p,K)$ can be computed as follows:

$$\tilde{S}_1(p,K)=\max_{\mathbf{a}}\sum_{k=1}^{K}p^2(1-p)^{k-1}H\left(B_k\mathbf{a}\right)$$
$$\text{s.t. } a_i\geq 0\quad\forall i,\quad\sum ia_i=1$$

We note that the constraints are convex and that the function to be maximized is the sum of the composition of convex function ($H(\cdot)$) with linear functions, hence is also convex.

A more convenient parameterization is obtained by normalizing $\mathbf{a}$ to be a probability vector. To that end, B_k should be replaced with C_k as follows

$$
C_k =
\begin{array}{c}
\\ 1 \\ 2 \\ 3 \\ \vdots \\ k \\ k+1
\end{array}
\begin{array}{cccccccc}
1 & 2 & 3 & \cdots & k & k+1 & k+2 & \\
\left[\begin{array}{ccccccc}
1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{k} & \frac{1}{k+1} & \frac{1}{k+2} & \cdots \\
0 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{k} & \frac{1}{k+1} & \frac{1}{k+2} & \cdots \\
0 & 0 & \frac{1}{3} & \cdots & \frac{1}{k} & \frac{1}{k+1} & \frac{1}{k+2} & \cdots \\
 & & & \ddots & & & & \\
0 & 0 & 0 & \cdots & \frac{1}{k} & \frac{1}{k+1} & \frac{1}{k+2} & \cdots \\
0 & 0 & 0 & \cdots & 0 & \frac{1}{k+1} & \frac{2}{k+2} & \cdots
\end{array}\right]
\end{array}
\tag{26}
$$

This yields the following optimization problem

$$
\tilde{S}_1(p,K) = \max_{\mathbf{a}} \sum_{k=1}^{n} p^2 (1-p)^{k-1} H(C_k \mathbf{a}) \tag{27}
$$

$$
\text{s.t. } a_i \geq 0 \quad \forall i, \quad \sum a_i = 1
$$

We would like evaluate $\tilde{S}_1(p,K)$ by numerically optimizing (27). It is clear that the length of the vector $\mathbf{a}$ should be limited to some fixed value (denoted by M). We denote the reduced size vector by $\tilde{\mathbf{a}}$ and derive it from $\mathbf{a}$ by:

$$
\tilde{a}_i = \begin{cases} a_i, & \text{for } i = 1, \ldots, M-1 \\ \sum_{j=M}^{n} a_j & \text{for } i = M \end{cases}
$$

We also define $\tilde{C}_k$ by cutting only the first M columns of C_k. Noticing the definition of C_k in (26) and the fact that both $\mathbf{a}$ and $\tilde{\mathbf{a}}$ are probability vector gives the following bounds

$$
\left[\tilde{C}_k \tilde{\mathbf{a}} - C_k \mathbf{a}\right]_i \leq \frac{\tilde{a}_M}{M}
$$

for $i \in [1, k]$ and

$$
\left|\left[\tilde{C}_k \tilde{\mathbf{a}} - C_k \mathbf{a}\right]_{k+1}\right| = \tilde{a}_M \left|\frac{M-k}{M} - 1\right| = \frac{k \tilde{a}_M}{M}
$$

Therefore, the variation distance is bounded by

$$
d_{\text{TV}}(\tilde{C}_k \tilde{\mathbf{a}}, C_k \mathbf{a}) \leq \frac{2k \tilde{a}_M}{M}
$$

Lemma 2 requires that $d_{\text{TV}}(\tilde{C}_k \tilde{\mathbf{a}}, C_k \mathbf{a}) \leq \frac{1}{2}$, so in order to comply we set $M = 4K$ and obtain the bound:

$$
H(C_k \mathbf{a}) < H(\tilde{C}_k \tilde{\mathbf{a}}) - \frac{2k \tilde{a}_M}{M} \log \frac{2k \tilde{a}_M}{(k+1)M}.
$$

Finally, the following finite-dimensional convex optimization problem provides a computable upper bound for $\check{S}_1(p,K,M) \geq \tilde{S}_1(p,K)$ that holds for any n large enough:

$$
\check{S}_1(p,K,M) \overset{\text{def}}{=} \tag{28}
$$

$$
\max_{\tilde{\mathbf{a}} \in \mathbb{R}^M} \sum_{k=1}^{K} \left[p^2(1-p)^{k-1} H\left(\tilde{C}_k \tilde{\mathbf{a}}\right) - \frac{2k \tilde{a}_M}{M} \log \frac{2k \tilde{a}_M}{(k+1)M} \right]
$$

$$
\text{s.t.} \quad \tilde{a}_i \geq 0, \quad \sum_{i=1}^{M} \tilde{a}_i = 1
$$

and lastly

$$
L(p,K,M) \leq \check{S}_1(p,K,M) + \sum_{k=K+1}^{n} p^2(1-p)^{k-1} \log(k+1).
$$
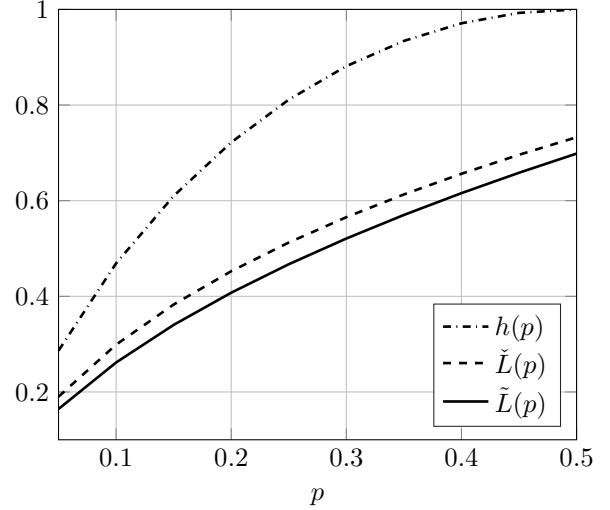
Fig. 2. $h(p)$, $\tilde{L}(p)$ and $\check{L}(p)$ as a function of p.

We evaluated $\check{L}(p)$ and $L(p,K,M)$ for $K = 100$ and $M = 400$. The results are depicted in Fig. 2 including the trivial bound $h(p)$.

We are only left with relating $L(p,K,M)$ to $\ell(\varepsilon,K,M)$. We note that p corresponds to the event of one of more errors on the channel between Alice and Bob and vice versa. So, $p = 1 - (1-\varepsilon)^2 = \varepsilon(2-\varepsilon)$ and

$$
\ell(\varepsilon,K,M) = L(\varepsilon(2-\varepsilon),K,M), \tag{29}
$$

where $\tilde{L}(\cdot)$ is given in (28). A simpler upper bound can be obtained using (25),

$$
\sup_{K,M} \ell(\varepsilon,K,M) \leq \check{\ell}(\varepsilon) \overset{\text{def}}{=} \check{L}(\varepsilon(2-\varepsilon)). \tag{30}
$$

REFERENCES

[1] L. J. Schulman, "Coding for interactive communication," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1745–1756, 1996.
[2] G. Kol and R. Raz, "Interactive channel capacity," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 715–724.
[3] B. Haeupler, "Interactive channel capacity revisited," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 226–235.
[4] B. Haeupler and A. Velingker, "Bridging the capacity gap between interactive and one-way communication," in *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2017, pp. 2123–2142.
[5] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on information Theory*, vol. 19, no. 4, pp. 471–480, 1973.
[6] R. Van Handel, *Probability in High Dimension, ORF 570, Lecture notes, Princeton University*, 2014.
[7] I. Csiszár and J. Körner, *Information Theory*, Cambridge University Press, second edition, 2011.