Asymmetric Error Correction and Flash-Memory Rewriting using Polar Codes

Eyal En Gad, Yue Li, Joerg Kliewer, Michael Langberg,
Anxiao (Andrew) Jiang and Jehoshua Bruck

Abstract

We propose efficient coding schemes for two communication settings: 1. asymmetric channels, and 2. channels with an informed encoder. These settings are important in non-volatile memories, as well as optical and broadcast communication. The schemes are based on non-linear polar codes, and they build on and improve recent work on these settings. In asymmetric channels, we tackle the exponential storage requirement of previously known schemes, that resulted from the use of large Boolean functions. We propose an improved scheme, that achieves the capacity of asymmetric channels with polynomial computational complexity and storage requirement.

The proposed non-linear scheme is then generalized to the setting of channel coding with an informed encoder, using a multicoding technique. We consider specific instances of the scheme for flash memories, that incorporate error-correction capabilities together with rewriting. Since the considered codes are non-linear, they eliminate the requirement of previously known schemes (called polar write-once-memory codes) for shared randomness between the encoder and the decoder. Finally, we mention that the multicoding scheme is also useful for broadcast communication in Marton's region, improving upon previous schemes for this setting.

I. Introduction

In this paper we make several contributions to the design and analysis of error-correcting codes in two important communication settings: 1. asymmetric channel coding, and 2. channel coding with an informed encoder. Asymmetric channel coding is important for applications such as non-volatile memories, in which the electrical mechanisms are dominantly asymmetric [5]. Another important application is in optical communication, where photons may fail to be detected $(1 \rightarrow 0)$ but the creation of spurious photons $(0 \rightarrow 1)$ is impossible [15, Section IX]. Channel coding with an informed encoder is also important for non-volatile memories, since the memory state in these devices affects the fate of writing attempts. Channel coding with an informed encoder is also useful in broadcast communication, where it is used in Marton's coding scheme to achieves high communication rates (see [7, p. 210]).

The focus of this paper is on polar coding techniques, as they are both highly efficient in terms of communication rate and computational complexity, and are relatively easy to analyze and understand. Polar codes were introduced by Arikan in [1], achieving the symmetric capacity of binary-input memoryless channels. The first task that we consider in this paper is that of point-to-point communication over asymmetric channels. Several polar coding schemes for asymmetric channels were proposed recently, including a pre-mapping using Gallager's scheme [11, p. 208] and a concatenation of two polar codes [30]. A more direct approach was proposed in [17], which we consider in this paper. A similar approach is also considered in [22]. The scheme in [17] achieves the capacity of asymmetric channels using non-linear polar codes, but it uses large Boolean functions that require storage space that is exponential in the block length. We propose a modification for this scheme, that removes the requirement for the Boolean functions,

and thus reduces the storage requirement of the encoding and decoding tasks to a linear function of the block length.

The second contribution of this paper is a generalization of the non-linear polar-coding scheme to the availability of channel side information at the encoder. We call this scheme a polar multicoding scheme, and we prove that it achieves the capacity of channels with informed encoders. The capacity of such channels was characterized by Gelfand and Pinsker in [12]. This scheme is useful for non-volatile memories such as flash memories and phase change memories, and for broadcast channels. We focus mainly on the flash memory application.

A prominent characteristic of flash memories is that the response of the memory cells to a writing attempt is affected by the previous content of the memory. This complicates the design of error correcting schemes, and thus motivates flash systems to "erase" the content of the cells before writing, and by that to eliminate its effect. However, the erase operation in flash memories is expensive, and therefore a simple coding scheme that does not require erasures could improve the performance of solid-state drives significantly. We show two instances of the proposed polar multicoding scheme that aim to achieve this goal.

A. Relation to Previous Work

The study of channel coding with an informed encoder was initiated by Kusnetsov and Tsybakov [19], with the channel capacity derived by Gelfand and Pinsker [12]. The informed encoding technique of Gelfand and Pinsker was used earlier by Marton to establish an inner bound for the capacity region of broadcast channels [21]. Low-complexity capacity-achieving codes were first proposed for continuous channels, using lattice codes [32]. In discrete channels, the first low-complexity capacity-achieving scheme was proposed using polar codes, for the symmetric special case of information embedding [18, Section VIII.B]. A modification of this scheme for the application of flash memory rewriting was proposed in [4], considering a model called write-once memory. An additional scheme for the application of flash memory, based on randomness extractors, was also proposed recently [10].

Our work is concerned with a setup that is similar to those considered in [4], [10]. An important contribution of the current paper compared to [4], [10] is that our scheme achieves the capacity of a rewriting model that also includes noise, while the schemes in [4], [10] address only the noiseless case. Indeed, error correction is a crucial capability in flash memory systems. Our low-complexity achievability of the noisy capacity is done using a multicoding technique. Comparing with [10], the current paper allows an input cost constraint, which is important in rewriting models for maximizing the sum of the code rates over multiple rewriting rounds. Comparing with [4], the current paper also improves by removing the requirement for shared randomness between the encoder and the decoder, which limits the practical coding performance. The removal of the shared randomness is done by the use of non-linear polar codes. An additional coding scheme was proposed during the writing of this paper, which also does not require shared randomness [20]. However, the scheme in [20] considers only the noiseless case, and it is in fact a special case of the scheme in the current paper.

Polar coding for channels with informed encoders was implicitly studied recently in the context of broadcast channels, as the Marton coding scheme for broadcast communication contains an informed encoding instance as an ingredient. In fact, a multicoding technique similar to the one presented in this paper was recently presented for broadcast channels, in [13]. While we were unaware of the result of [13] and developed the scheme independently, this paper also has three contributions that were not shown in [13]. First, by using the modified scheme of non-linear polar codes, we reduce the storage requirement from an exponential function in the block length to a linear function. Secondly, we connect the scheme to the application of data storage and flash memory rewriting, that was not considered in the previous work. And thirdly, the analysis in [13] holds only for channels whose capacity-achieving distribution forms a certain degraded structure. In this paper we consider a specific noisy rewriting model, whose capacity-achieving distribution forms the required degraded structure, and by that we show that the scheme achieves the capacity of the considered flash-memory model.

Another recent paper on polar coding for broadcast channels was published recently by Mondelli et. al. [23]. That paper proposed a method, called "chaining", that allows to bypass the degraded structure requirement. In this paper we connect the chaining method to the flash-memory rewriting application and to our new non-linear polar coding scheme, and apply it to our proposed multicoding scheme. This allows for a linear storage requirement, together with the achievability of the informed encoder capacity and Marton's inner bound, eliminating the degraded structure requirement. Finally, we show an important instance of the chaining scheme for a specific flash-memory model, and explain the applicability of this instance in flash-memory systems.

The rest of the paper is organized as following. Section II proposes a new non-linear polar coding scheme for asymmetric channels, which does not require an exponential storage of Boolean functions. Section III proposes a new polar multicoding scheme for channels with informed encoders, including two special cases for the rewriting of flash memories. Finally, Section IV summarizes the paper.

II. ASYMMETRIC POINT-TO-POINT CHANNELS

Notation: For positive integers $m \leq n$, let $[m : n]$ denote the set $\{m, m+1, \ldots, n\}$, and let $[n]$ denote the set $[1 : n]$. Given a subset $\mathcal{A}$ of $[n]$, let $\mathcal{A}^c$ denote the complement of $\mathcal{A}$ with respect to $[n]$, where n is clear from the context. Let $x_{[n]}$ denote a vector of length n, and let $x_{\mathcal{A}}$ denote a vector of length $|\mathcal{A}|$ obtained from $x_{[n]}$ by deleting the elements with indices in $\mathcal{A}^c$.

Throughout this section we consider only channels with binary input alphabets, since the literature on polar codes with non-binary codeword symbols is relatively immature. However, the results of this section can be extended to non-binary alphabets without much difficulty using the methods described in [24]–[29]. The main idea of polar coding is to take advantage of the polarization effect of the Hadamard transform on the entropies of random vectors. Consider a binary-input memoryless channel model with an input random variable (RV) $X \in \{0, 1\}$, an output RV $Y \in \mathcal{Y}$ and a pair of conditional probability mass functions (pmfs) $p_{Y|X}(y|0), p_{Y|X}(y|1)$ on $\mathcal{Y}$. Let n be a power of 2 that denotes the number of channel uses, also referred to as the block length. The channel capacity is the tightest upper bound on the code rate in which the probability of decoding error can be made as small as desirable for large enough block length. The channel capacity is given by the mutual information of X and Y.

Theorem 1. *(Channel Coding Theorem) [6, Chapter 7] The capacity of the discrete memoryless channel $p(y|x)$ is given by*

$$C = \max_{p(x)} I(X; Y).$$

The Hadamard transform is a multiplication of the random vector $X_{[n]}$ over the field of cardinality 2 with the matrix $G_n = G^{\otimes \log_2 n}$, where $G = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ and $\otimes$ denotes the Kronecker power. In other words, G_n can be described recursively for $n \geq 4$ by the block matrix

$$G_n = \begin{pmatrix} G_{n/2} & 0 \\ G_{n/2} & G_{n/2} \end{pmatrix}.$$

The matrix G_n transforms $X_{[n]}$ into a random vector $U_{[n]} = X_{[n]} G_n$, such that the conditional entropy $H(U_i|U_{[i-1]}, Y_{[n]})$ is *polarized*. That means that for a fraction of close to $H(X|Y)$ of the indices $i \in [n]$, the conditional entropy $H(U_i|U_{[i-1]}, Y_{[n]})$ is close to 1, and for almost all the rest of the indices, $H(U_i|U_{[i-1]}, Y_{[n]})$ is close to 0. This result was shown by Arikan in [1], [2].

Theorem 2. *(Polarization Theorem) [2, Theorem 1] Let $n, U_{[n]}, X_{[n]}, Y_{[n]}$ be defined as above. For any $\delta \in (0, 1)$, let*

$$H_{X|Y} \triangleq \left\{ i \in [n] : H(U_i|U_{[i-1]}, Y_{[n]}) \in (1-\delta, 1) \right\},$$

and

$$L_{X|Y} \triangleq \left\{ i \in [n] : H(U_i|U_{[i-1]}, Y_{[n]}) \in (0, \delta) \right\}.$$

Then

$$\lim_{n\to\infty} |H_{X|Y}|/n = H(X|Y) \quad \text{and} \quad \lim_{n\to\infty} |L_{X|Y}|/n = 1 - H(X|Y).$$

Note that $H(X|Y)$ denotes a conditional entropy, while $H_{X|Y}$ denotes a subset of $[n]$. It is also shown in [1] that the transformation G_n is invertible with $G_n^{-1} = G_n$, implying $X_{[n]} = U_{[n]}G_n$. This polarization effect can be used quite simply for the design of a coding scheme that achieves the capacity of symmetric channels with a running time that is polynomial in the block length. The capacity of symmetric channels is achieved by a uniform distribution on the input alphabet, i.e. $p(x) = 1/2$ [6, Theorem 7.2.1]. Since the input alphabet in this paper is binary, the capacity-achieving distribution gives $H(X) = 1$, and therefore we have

$$\lim_{n\to\infty} (1/n)|L_{X|Y}| = 1 - H(X|Y) = H(X) - H(X|Y) = I(X;Y) = C. \tag{1}$$

Furthermore, for each index in $L_{X|Y}$, the conditional probability $p(u_i|u_{[i-1]}, y_{[n]})$ must be close to either 0 or 1 (since the conditional entropy is small by the definition of the set $L_{X|Y}$). It follows that the RV U_i can be estimated reliably given $u_{[i-1]}$ and $y_{[n]}$. This fact motivates the capacity-achieving coding scheme that follows. The encoder creates a vector $u_{[n]}$ by assigning the subvector $U_{L_{X|Y}}$ with the source message, and the subvector $U_{L_{X|Y}^c}$ with uniformly distributed random bits that are shared with the decoder. The randomness sharing is useful for the analysis, but is in fact unnecessary for using the scheme (the proof of this fact is described in [1, Section VI]). The set $U_{L_{X|Y}^c}$ is called the *frozen set*. Equation (1) implies that this coding rate approaches the channel capacity. The decoding is performed iteratively, from index 1 up to n. In each iteration, the decoder estimates the bit u_i using the shared information or using a maximum likelihood estimation, according to the set membership of the iteration. The estimations of the bits u_i for which i is in $L_{X|Y}^c$ are always successful, since these bits were known to the decoder in advance. The rest of the bits are estimated correctly with high probability, leading to a successful decoding of the entire message with high probability.

However, this reasoning does not translate directly to asymmetric channels. Remember that the capacity-achieving input distribution of asymmetric channels is in general not uniform (see, for example, [14]), i.e. $p_X(1) \neq 1/2$. Since the Hadamard transform is bijective, it follows that the capacity-achieving distribution of the polarized vector $U_{[n]}$ is non uniform as well. The problem with this fact is that assigning uniform bits of message or shared randomness changes the distribution of $U_{[n]}$, and consequentially also changes the conditional entropies $H(U_i|U_{[i-1]}, Y_{[n]})$. To manage this situation, our approach is to make sure that the change in the distribution of $U_{[n]}$ is kept to be minor, and thus its effect on the probability of decoding error is also minor. To do this, we consider the conditional entropies $H(U_i|U_{[i-1]})$, for $i \in [n]$. Since the polarization happens regardless of the channel model, we can consider a channel for which the output Y is a deterministic variable, and conclude by Theorem 2 that the entropies $H(U_i|U_{[i-1]})$ also polarize. For this polarization, a fraction of $H(X)$ of the indices admit a high $H(U_i|U_{[i-1]})$. To ensure a minor change in the distribution of $U_{[n]}$, we restrict the assignments of uniform bits of message and shared randomness to the indices with high $H(U_i|U_{[i-1]})$.

The insight of the last paragraph motivates a modified coding scheme. The locations with high entropy $H(U_i|U_{[i-1]})$ are assigned with uniformly distributed bits, while the rest of the locations are assigned with the pmf $p(u_i|u_{[i-1]})$. Note that $p(u_{[n]}, x_{[n]})$ and $H(U_{[n]})$ refer to the capacity-achieving distribution of the channel, which does not equal to the distribution that the encoding process induces. Similar to the notation of Theorem 2, we denote the set of indices with high entropy $H(U_i|U_{[i-1]})$ by H_X. To achieve a reliable decoding, we place the message bits in the indices of H_X that can be decoded reliably, meaning that their entropies $H(U_i|U_{[i-1]}, Y_{[n]})$ are low. So we say that we place the message bits in the intersection $H_X \cap L_{X|Y}$. The locations whose indices are not in $L_{X|Y}$ must be known by the decoder in

advance for a reliable decoding. Previous work suggested to share random Boolean functions between the encoder and the decoder, drawn according to the pmf $p(u_i|u_{[i-1]})$, and to assign the indices in $(H_X \cap L_{X|Y})^c = H_X^c \cup L_{X|Y}^c$ according to these functions [13], [17]. However, we note that the storage required for those Boolean functions is exponential in n, and therefore we propose an efficient alternative.

To avoid the Boolean function, we divide the complement of $H_X \cap L_{X|Y}$ into three disjoint sets. First, the indices in the intersection $H_X \cap L_{X|Y}^c$ are assigned with uniformly distributed random bits that are shared between the encoder and the decoder. As in the symmetric case, this randomness sharing will in fact not be necessary, and a deterministic frozen vector could be shared instead. The rest of the bits of $U_{[n]}$ (those in the set H_X^c), are assigned *randomly* to a value u with probability $p_{U_i|U_{[i-1]}}(u|u_{[i-1]})$ (where $p_{U_i|U_{[i-1]}}$ is calculated according to the pmf $p_{U_{[n]},X_{[n]},Y_{[n]}}$, the capacity-achieving distribution of the channel). The indices in $H_X^c \cap L_{X|Y}$ could be decoded reliably, but not those in $H_X^c \cap L_{X|Y}^c$. Fortunately, the set $H_X^c \cap L_{X|Y}^c$ can be shown to be small (as we will show later), and thus we could transmit those locations separately with a vanishing effect on the code rate. The encoding of the vector $u_{[n]}$ is illustrated in Figure 1.

The see the intuition of why the code rate approaches the channel capacity, notice that the source message is placed in the indices in the intersection $H_X \cap L_{X|Y}$. The asymptotic fraction of this intersection can be derived as following.

$$|H_X \cap L_{X|Y}|/n = 1 - |H_X^c \cup L_{X|Y}^c|/n = 1 - |H_X^c|/n - |L_{X|Y}^c|/n + |H_X^c \cap L_{X|Y}^c|/n. \qquad (2)$$

The Polarization Theorem (Theorem 2) implies that $|H_X^c|/n \to 1 - H(X)$ and $|L_{X|Y}^c|/n \to H(X|Y)$. Since the fraction $|H_X^c \cap L_{X|Y}^c|$ vanishes for large n, we get that the asymptotic rate is $|H_X \cap L_{X|Y}|/n \to H(X) - H(X|Y) = I(X;Y)$, achieving the channel capacity.

For a more precise definition of the scheme, we use the so called Bhattacharyya parameter in the selection of subsets of $U_{[n]}$, instead of the conditional entropy. The Bhattacharyya parameters are polarized in a similar manner as the entropies, and are more useful for bounding the probability of decoding error. For a discrete RV Y and a Bernoulli RV X, the Bhattacharyya parameter is defined by

$$Z(X|Y) \triangleq 2 \sum_y \sqrt{p_{X,Y}(0,y) p_{X,Y}(1,y)}. \qquad (3)$$

Note that most of the polar coding literature is using a slightly different definition of the Bhattacharyya parameter, that coincides with Equation (3) when the RV X is distributed uniformly. We use the following relations between the Bhattacharyya parameter and the conditional entropy.

Proposition 3. *([2, Proposition 2])*

$$(Z(X|Y))^2 \le H(X|Y), \qquad (4)$$
$$H(X|Y) \le \log_2(1 + Z(X|Y)) \le Z(X|Y). \qquad (5)$$
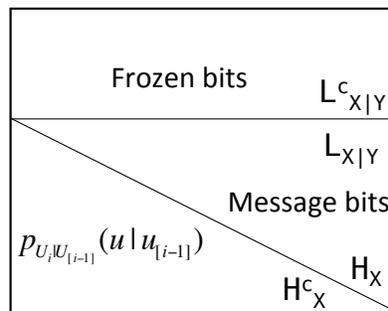

Fig. 1. Encoding the vector $u_{[n]}$.

We now define the set of high and low Bhattacharyya parameters, and work with them instead of the sets $H_{X|Y}$ and $L_{X|Y}$. For $\delta \in (0,1)$, define

$$\mathcal{H}_{X|Y} \triangleq \left\{ i \in [n] : Z(U_i|U_{[i-1]}, Y_{[n]}) \geq 1 - 2^{-n^{1/2-\delta}} \right\},$$

$$\mathcal{L}_{X|Y} \triangleq \left\{ i \in [n] : Z(U_i|U_{[i-1]}, Y_{[n]}) \leq 2^{-n^{1/2-\delta}} \right\}.$$

As before, we define the sets $\mathcal{H}_X$ and $\mathcal{L}_X$ for the parameter $Z(U_i|U_{[i-1]})$ by letting $Y_{[n]}$ be a deterministic vector. Using Proposition 3, it is shown in [17, combining Proposition 2 with Theorem 2] that Theorem 2 holds also if we replace the sets $H_{X|Y}$ and $L_{X|Y}$ with the sets $\mathcal{H}_{X|Y}$ and $\mathcal{L}_{X|Y}$. That is, we have

$$\lim_{n \to \infty} |\mathcal{H}_{X|Y}|/n = H(X|Y) \quad \text{and} \quad \lim_{n \to \infty} |\mathcal{L}_{X|Y}|/n = 1 - H(X|Y). \tag{6}$$

We now define our coding scheme formally. Let $m_{[|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|]} \in \{0,1\}^{|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|}$ be the realization of a uniformly distributed source message, and $f_{[|\mathcal{H}_X \cap \mathcal{L}_{X|Y}^c|]} \in \{0,1\}^{|\mathcal{H}_X \cap \mathcal{L}_{X|Y}^c|}$ be a deterministic frozen vector known to both the encoder and the decoder. We discuss how to find a good frozen vector in Appendix C-C. For a subset $\mathcal{A} \subseteq [n]$ and an index $i \in \mathcal{A}$, we use a function $r(i, \mathcal{A})$ to denote the *rank* of i in an ordered list of the elements of $\mathcal{A}$. The probabilities $p_{U_i|U_{[i-1]}}(u|u_{[i-1]})$ and $p_{U_i|U_{[i-1]}, Y^n}(u|u_{[i-1]}, y_{[n]})$ can be calculated efficiently by a recursive method described in [17, Section III.B].

Construction 4.
Encoding
  Input: a message $m_{[|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|]} \in \{0,1\}^{|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|}$.
  Output: a codeword $x_{[n]} \in \{0,1\}^n$.
  1) *For i from 1 to n, successively, set*

$$u_i = \begin{cases} u \in \{0,1\} & \text{with probability } p_{U_i|U_{[i-1]}}(u|u_{[i-1]}) & \text{if } i \in \mathcal{H}_X^c \\ m_{r(i, \mathcal{H}_X \cap \mathcal{L}_{X|Y})} & & \text{if } i \in \mathcal{H}_X \cap \mathcal{L}_{X|Y} \\ f_{r(i, \mathcal{H}_X \cap \mathcal{L}_{X|Y}^c)} & & \text{if } i \in \mathcal{H}_X \cap \mathcal{L}_{X|Y}^c. \end{cases}$$

  2) *Transmit the codeword $x_{[n]} = u_{[n]} G_n$.*
  3) *Transmit the vector $u_{\mathcal{H}_X^c \cap \mathcal{L}_{X|Y}^c}$ separately using a linear, non-capacity-achieving polar code with a uniform input distribution (as in [1]). In practice, other error-correcting codes could be used for this vector as well.*
Decoding
  Input: a noisy vector $y_{[n]} \in \{0,1\}^n$.
  Output: a message estimation $\hat{m}_{[|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|]} \in \{0,1\}^{|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|}$.
  1) *Estimate the vector $u_{\mathcal{H}_X^c \cap \mathcal{L}_{X|Y}^c}$ by $\hat{u}_{\mathcal{H}_X^c \cap \mathcal{L}_{X|Y}^c}$.*
  2) *For i from 1 to n, set*

$$\hat{u}_i = \begin{cases} \arg\max_{u \in \{0,1\}} p_{U_i|U_{[i-1]}, Y_{[n]}}(u|u_{[i-1]}, y_{[n]}) & \text{if } i \in \mathcal{L}_{X|Y} \\ \hat{u}_{r(i, \mathcal{H}_X^c \cap \mathcal{L}_{X|Y}^c)} & \text{if } i \in \mathcal{H}_X^c \cap \mathcal{L}_{X|Y}^c \\ f_{r(i, \mathcal{H}_X \cap \mathcal{L}_{X|Y}^c)} & \text{if } i \in \mathcal{H}_X \cap \mathcal{L}_{X|Y}^c. \end{cases}$$

  3) *Return the estimated message $\hat{m}_{[|\mathcal{H}_X \cap \mathcal{L}_{X|Y}|]} = \hat{u}_{\mathcal{H}_X \cap \mathcal{L}_{X|Y}}$.*

We say that a sequence of coding schemes achieves the channel capacity if the probability of decoding error vanishes with the block length for any rate below the capacity.

Theorem 5. *Construction* 4 *achieves the channel capacity (Theorem* 1*) with a computational complexity of* $O(n \log n)$ *and a probability of decoding error of* $2^{-n^{1/2-\delta}}$ *for any* $\delta > 0$ *and large enough* n.

In the next section we show a generalized construction and prove its capacity-achieving property. Theorem 5 thus will follow as a corollary of the more general Theorem 15. We note here two differences between Construction 4 and the construction in [23, Section III.B]. First, in the encoding of Construction 4, the bits in the set $\mathcal{H}_X^c$ and set randomly, while in [23, Section III.B], those bits are set according to a maximum likelihood rule. And second, the vector $u_{\mathcal{H}_X^c \cap \mathcal{L}_{X|Y}^c}$ is being sent through a side channel in Construction 4, but not in [23, Section III.B]. These changes allow us to prove that the coding scheme achieves the channel capacity.

III. CHANNELS WITH NON-CAUSAL ENCODER STATE INFORMATION

In this section we generalize Construction 4 to the availability of channel state information at the encoder. We consider mainly the application of rewriting in flash memories, and present two special cases of the channel model for this application. In flash memory, information is stored in a set of n memory cells. We mainly focus on a flash memory type that is called Single-Level Cell (SLC), in which each cell stores a single information bit, and its value is denoted by either 0 or 1. We first note that the assumption of a memoryless channel is not exactly accurate in flash memories, due to a mechanism of cell-to-cell interference. However, we keep using this assumption, as it is nonetheless useful for the design of coding schemes with valuable practical performance. The main limitation of flash memories that we do tackle in this work is the high cost of changing a cell level from 1 to 0 (in SLC memories). To perform such a change, an expensive operation, called "block erasure", is required. To avoid this block erasure operation, information is rewritten over existing memory in the sense that no cell is changed from value 1 to 0. We thus consider the use of the information about the previous state of the cells in the encoding process. We model the memory cells as a channel with a discrete *state*, and we also assume that the state is memoryless, meaning that the states of different cells are distributed independently.

We assume that the state of the entire n cells is available to the writer prior to the beginning of the writing process. In communication terminology this kind of state availability is refereed to as "non causal". We note that this setting is also useful in the so called *Marton-coding* method for communication over broadcast channels. Therefore, the multicoding schemes that will follow serve as a contribution also in this important setting. One special case of the model which we consider is the noiseless write-once memory model. This model also serve as an ingredient for a type of codes called "rank-modulation rewriting codes" [9]. Therefore, the schemes proposed in this section can also be useful for the design of rank-modulation rewriting codes.

We represent the channel state as a Bernoulli random variable S with parameter β, which equals the probability $p(S = 1)$. A cell of state 1 can only be written with the value 1. Note that, intuitively, when β is high, the capacity of the memory is small, since only a few cells are available for modification in the writing process, and thus only a small amount of information could be stored. This also means that the choice of codebook has a crucial effect on the capacity of the memory in *future writes*. A codebook that contains many codewords of high Hamming weight (number of 1's in the codeword) would make the parameter β of future writes high, and thus the capacity of the future writes would be low. However, forcing the expected Hamming weight of the codebook to be low would reduce the capacity of the current write. To settle this trade-off, previous work suggested to optimize the sum of the code rates over multiple writes. It was shown that in many cases, constraints on the codebook Hamming weight (henceforth just weight) strictly increases the sum rate (see, for example, [16]). Therefore, we consider an input cost constraint in the model.

The most general model that we consider is a discrete memoryless channel (DMC) with a discrete memoryless (DM) state and an input cost constraint, where the state information is available non causally at the encoder. The channel input, state and output are denoted by x, s and y, respectively, and their respective finite alphabets are denoted by $\mathcal{X}, \mathcal{S}$ and $\mathcal{Y}$. The random variables are denoted by X, S and

Y, and the random vectors by $X_{[n]}, S_{[n]}$ and $Y_{[n]}$, where n is the block length. The state is distributed according to the pmf $p(s)$, and the conditional pmfs of the channel are denoted by $p(y|x,s)$. The input cost function is denoted by $b(x)$, and the input cost constraint is

$$\sum_{i=1}^{n} \mathsf{E}[b(X_i)] \leq nB,$$

where B is a real number representing the normalized constraint. The channel capacity with an informed encoder and an input cost constraint is given by an extension of the Gelfand-Pinsker Theorem[1].

Theorem 6. *(Gelfand-Pinsker Theorem with Cost Constraint) [7, Equation (7.7) on p. 186] The capacity of a DMC with a DM state $p(y|x,s)p(s)$ and an input cost constraint B when the state information is available non causally only at the encoder is*

$$C = \max_{p(v|s),x(v,s):\mathsf{E}(b(X))\leq B} (I(V;Y) - I(V;S)), \tag{7}$$

where V is an auxiliary random variable with a finite alphabet $\mathcal{V}$, $|\mathcal{V}| \leq \min\{|\mathcal{X} \cdot \mathcal{S}, \mathcal{Y} + \mathcal{S} - 1|\}$.

The main coding scheme that we present in this section achieves the capacity in Theorem 6. The proof of Theorem 6 considers a virtual channel model, in which the RV V is the channel input and Y is the channel output. Similar to the previous section, we limit the treatment to the case in which the RV V is binary. In flash memory, this case would correspond to a single-level cell (SLC) type of memory. As mentioned in Section II, an extension of the scheme to a non-binary case is not difficult. The non-binary case is useful for flash memories in which each cell stores 2 or more bits of information. Such memories are called Multi-Level Cell (MLC). We also mention that the limitation to binary random variables does not apply on the channel output Y. Therefore, the cell voltage in flash memory could be read more accurately at the decoder to increase the coding performance, similarly to the soft decoding method that is used in flash memories with LDPC codes. Another practical remark is that the binary-input model can be used in MLC memories by coding separately on the MSB and the LSB of the cells, as in fact is the coding method in current MLC flash systems.

The scheme that achieves the capacity of Theorem 6 is called Construction 16, and it will be described in Subsection III-C. The capacity achieving result is summarized in the following theorem, which will be proven in Subsection III-C.

Theorem 7. *Construction 16 achieves the capacity of the Gelfand-Pinsker Theorem with Cost Constraint (Theorem 6) with a computational complexity of $O(n \log n)$ and a probability of decoding error of $2^{-n^{1/2-\delta}}$ for any $\delta > 0$ and large enough n.*

Note that the setting of Theorem 7 is a generalization of the asymmetric channel-coding setting of Theorem 5, and therefore Construction 16 and Theorem 7 are in fact a generalization of Construction 4 and Theorem 5. Before we describe the code construction, we first show in Subsection III-A two special cases of the Gelfand-Pinsker model that are useful for the rewriting of flash memories. Afterwards, in subsections III-B and III-C, we will show two versions of the construction that correspond to generalizations of the two special cases.

A. Special Cases

We start with a special case that is quite a natural model for flash memory rewriting.

Example 8. *Let the sets $\mathcal{X}, \mathcal{S}$ and $\mathcal{Y}$ be all equal to $\{0,1\}$, and let the state pmf be $p_S(1) = \beta$. This model corresponds to a single level cell flash memory. We describe the cell behaviour after a bit x is attempted to*

---

[1]The cost constraint is defined slightly differently in this reference, but the capacity is not affected by this change.
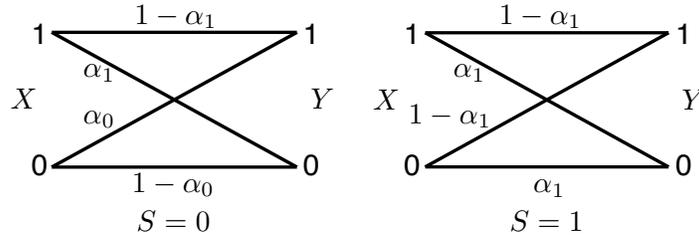
Fig. 2. Example 8: A binary noisy WOM model.

be written. When $s = 0$, the cell behaves as a binary asymmetric channel with input x, since the call state does not interfere with the writing attempt. When $s = 1$, the cell behaves as if a value of 1 was attempted to be written, regardless of the actual value x attempted. However, an error might still occur, during the writing process or anytime afterwards (for example, due to charge leakage). Thus, we can say that when $s = 1$, the cell behaves as a binary asymmetric channel with input 1. Formally, the channel pmfs are given by

$$p_{Y|XS}(1|x,s) = \begin{cases} \alpha_0 & \text{if } (x,s) = (0,0) \\ 1 - \alpha_1 & \text{if } (x,s) = (0,1) \\ 1 - \alpha_1 & \text{if } (x,s) = (1,0) \\ 1 - \alpha_1 & \text{if } (x,s) = (1,1) \end{cases} \tag{8}$$

The error model is also presented in Figure 2. The cost constraint is given by $b(x_i) = x_i$, since it is desirable to limit the amount of cells written to a value of 1.

Our coding-scheme construction for the setting of Theorem 6 is based on a more limited construction, which serves as a building block. We will start by describing the limited construction, and then show how to extend it for the model of Theorem 6. We will prove that the limited construction achieves the capacity of channels whose capacity-achieving distribution forms a certain stochastically degraded structure. We first recall the definition of stochastically degraded channels.

Definition 9. *[7, p. 112] A discrete memoryless channel (DMC) $W_1 : \{0,1\} \to \mathcal{Y}_1$ is stochastically degraded (or simply degraded) with respect to a DMC $W_2 : \{0,1\} \to \mathcal{Y}_2$, denoted as $W_1 \preceq W_2$, if there exists a DMC $W : \mathcal{Y}_2 \to \mathcal{Y}_1$ such that W satisfies the equation $W_1(y_1|x) = \sum_{y_2 \in \mathcal{Y}_2} W_2(y_2|x)W(y_1|y_2)$.*

Next, we bring the required property of channels whose capacity is achieved by the limited construction to be proposed.

Property 10. *There exist functions $p(v|s)$ and $x(v,s)$ that maximize the Gelfand-Pinsker capacity in Theorem 6 which satisfy the condition $p(y|v) \succeq p(s|v)$.*

It is an open problem whether the model of Example 8 satisfies the degradation condition of Property 10. However, we can modify the model such that it will satisfy Property 10. Specifically, we study the following model:

Example 11. *Let the sets $\mathcal{X}, \mathcal{S}$ and $\mathcal{Y}$ be all equal to $\{0,1\}$. The channel and state pmfs are given by $p_S(1) = \beta$ and*

$$p_{Y|XS}(1|x,s) = \begin{cases} \alpha & \text{if } (x,s) = (0,0) \\ 1 - \alpha & \text{if } (x,s) = (1,0) \\ 1 & \text{if } s = 1. \end{cases} \tag{9}$$

In words, if $s = 1$ the channel output is always 1, and if $s = 0$, the channel behave as a binary symmetric channel. The cost function is given by $b(x_i) = x_i$. The error model is also presented in Figure 3. This model can represent a writing noise, as a cell of state $s = 1$ is not written on and it never suffers errors.
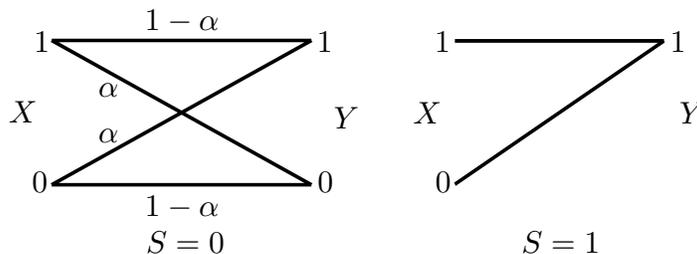
Fig. 3. Example 11: A binary WOM with writing noise.

We claim that the model of Example 11 satisfies the degradation condition of Property 10. To show this, we need first to find the functions $p(v|s)$ and $x(v,s)$ that maximize the Gelfand-Pinsker capacity in Theorem 6. Those functions are established in the following theorem.

Theorem 12. *The capacity of the channel in Example* 11 *is*

$$C = (1 - \beta)[h(\epsilon * \alpha) - h(\alpha)],$$

where $\epsilon = B/(1 - \beta)$ and $\epsilon * \alpha \equiv \epsilon(1 - \alpha) + (1 - \epsilon)\alpha$. The selections $\mathcal{V} = \{0,1\}$, $x(v,s) = v \wedge \neg s$ (where $\wedge$ is the logical AND operation, and $\neg$ is the logical negation), and

$$p_{V|S}(1|0) = \epsilon, \qquad p_{V|S}(1|1) = \frac{\epsilon(1 - \alpha)}{\epsilon * \alpha} \tag{10}$$

achieve this capacity.

Theorem 12 is similar to [16, Theorem 4], and its proof is described in Appendix A. Intuitively, the upper bound is obtained by assuming that the state information is available also at the decoder, and the lower bound is obtained by setting the functions $x(v,s)$ and $p(v|s)$ according to the statement of the theorem. The proof that the model in Example 11 satisfies the degradation condition of Property 10 is completed by the following lemma.

Lemma 13. *The capacity achieving functions of Theorem* 12 *for the model of Example* 11 *satisfy the degradation condition of Property* 10. *That is, the channel $p(s|v)$ is degraded with respect to the channel $p(y|v)$.*

Lemma 13 is proven in Appendix B, and consequently, the capacity of the model in Example 11 can be achieved by our limited construction. In the next subsection we describe the construction for channel models which satisfy Property 10, including the model in Example 11.

B. Multicoding Construction for Degraded Channels

Notice first that the capacity-achieving distribution of the asymmetric channel in Section II actually satisfies Property 10. In the asymmetric channel-coding case, the state can be thought of as a degenerate random variable (a RV which only takes a single value), and therefore we can choose W in Definition 9 to be degenerate as well, and by that satisfy Property 10. We will see that the construction that we present in this subsection is a generalization of Construction 4.

The construction has a similar structure as the achievability proof of the Gelfand-Pinsker Theorem (Theorem 6). The encoder first finds a vector $v_{[n]}$ in a similar manner to Construction 4, where the RV $X|Y$ in Construction 4 is replaced with $V|Y$, and the RV X is replaced with $V|S$. The vector $U_{[n]}$ is now the polarization of the vector $V_{[n]}$, meaning that $U_{[n]} = V_{[n]}G_n$. The RV V is taken according to the pmfs $p(v|s)$ that maximize the rate expression in Equation (7). The selection of the vector $u_{[n]}$ is illustrated in Figure 4. After the vector $u_{[n]}$ is chosen, each bit $i \in [n]$ in the codeword $x_{[n]}$ is calculated by the function

$x_i(v_i, s_i)$ that maximizes Equation (7). To use the model of Example 11, one should use the functions $p(v|s)$ and $x(v,s)$ according to Theorem 12. The key to showing that the scheme achieves the channel capacity is that the fraction $|\mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y}|/n$ can be shown to vanish for large n if the channel satisfies Property 10. Then, by the same intuition as in Equation (2) and using Equation (6), the replacements imply that the asymptotic rate of the codes is

$$\begin{aligned}
|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}|/n &= 1 - |\mathcal{H}^c_{V|S}|/n - |\mathcal{L}^c_{V|Y}|/n + |\mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y}|/n \\
&\to 1 - (1 - H(V|S)) - H(V|Y) + 0 \\
&= I(V;Y) - I(V;S),
\end{aligned}$$

achieving the Gelfand-Pinsker capacity of Theorem 6. We now describe the coding scheme formally.

Construction 14.

Encoding

   Input: *a message* $m_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}|]} \in \{0,1\}^{|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}|}$ *and a state* $s_{[n]} \in \{0,1\}^n$.
   Output: *a codeword* $x_{[n]} \in \{0,1\}^n$.

   1) *For each i from 1 to n, assign*

$$u_i = \begin{cases} u \in \{0,1\} & \text{with probability } p_{U_i|U_{[i-1]},S_{[n]}}(u|u_{[i-1]},s_{[n]}) & \text{if } i \in \mathcal{H}^c_{V|S} \\ m_{r(i,\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y})} & & \text{if } i \in \mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y} \\ f_{r(i,\mathcal{H}_{V|S} \cap \mathcal{L}^c_{V|Y})} & & \text{if } i \in \mathcal{H}_{V|S} \cap \mathcal{L}^c_{V|Y}. \end{cases} \quad (11)$$

   2) *Calculate $v_{[n]} = u_{[n]} G_n$ and for each $i \in [n]$, store the value $x_i(v_i, s_i)$.*
   3) *Store the vector $u_{\mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y}}$ separately using a point-to-point linear non-capacity-achieving polar code with a uniform input distribution. The encoder here does not use the state information in the encoding process, but rather treat it as an unknown part of the channel noise.*

Decoding

   Input: *a noisy vector* $y_{[n]} \in \{0,1\}^n$.

   Output: *a message estimation* $\hat{m}_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}|]} \in \{0,1\}^{|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}|}$.

   1) *Estimate the vector $u_{\mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y}}$ by $\hat{u}_{\mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y}}$.*
   2) *Estimate $u_{[n]}$ by $\hat{u}_{[n]}(y_{[n]}, f_{[|\mathcal{H}_{V|S} \cap \mathcal{L}^c_{V|Y}|]})$ as follows: For each i from 1 to n, assign*

$$\hat{u}_i = \begin{cases} \underset{u \in \{0,1\}}{\arg\max}\, p_{U_i|U_{[i-1]},Y_{[n]}}(u|u_{[i-1]},y_{[n]}) & \text{if } i \in \mathcal{L}_{V|Y} \\ \hat{u}_{r(i,\mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y})} & \text{if } i \in \mathcal{H}^c_{V|S} \cap \mathcal{L}^c_{V|Y} \\ f_{r(i,\mathcal{H}_{V|S} \cap \mathcal{L}^c_{V|Y})} & \text{if } i \in \mathcal{H}_{V|S} \cap \mathcal{L}^c_{V|Y}. \end{cases} \quad (12)$$
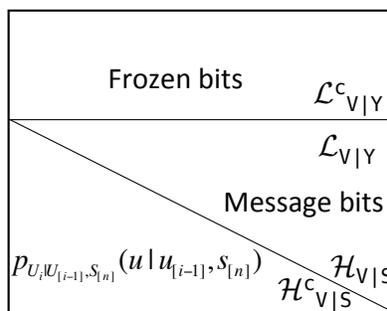

Fig. 4.   Encoding the vector $u_{[n]}$ in Construction 14.

3) Return the estimated message $\hat{m}_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}|]} = \hat{u}_{\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}}$.

The asymptotic performance of Construction 14 is stated in the following theorem.

Theorem 15. *If Property* 10 *holds, then Construction* 14 *achieves the capacity of Theorem* 6 *with a computational complexity of* $O(n \log n)$ *and a probability of decoding error of* $2^{-n^{1/2-\delta}}$ *for any* $\delta > 0$ *and large enough* n.

The proof of Theorem 15 is shown in Appendix C. The next subsection describes a method to remove the degradation requirement of Property 10. This allows to achieve also the capacity of the more realistic model of Example 8.

C. Multicoding Construction without Degradation

A technique called "chaining" was proposed in [23] that allows to achieve the capacity of models that do not exhibit the degradation condition of Property 10. The chaining idea was presented in the context of broadcast communication and point-to-point universal coding. We connect it here to the application of flash memory rewriting through Example 8. We note also that the chaining technique that follows comes with a price of a slower convergence to the channel capacity, and thus a lower non-asymptotic code rate.

The requirement of Construction 14 for degraded channels comes from the fact that the set $\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c$ needs to be communicated to the decoder in a side channel. If the fraction $(1/n)|\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c|$ vanishes with n, Construction 14 achieves the channel capacity. In this subsection we deal with the case that the fraction $(1/n)|\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c|$ does not vanish. In this case we have

$$
\begin{aligned}
|H_{V|S} \cap L_{V|Y}|/n =& 1 - |H_{V|S}^c \cup L_{V|Y}^c|/n \\
=& 1 - |H_{V|S}^c|/n - |L_{V|Y}^c|/n + |H_{V|S}^c \cap L_{V|Y}^c|/n \\
\rightarrow& I(V;Y) - I(V;S) + |H_{V|S}^c \cap L_{V|Y}^c|/n.
\end{aligned}
$$

The idea is then to store the subvector $u_{\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c}$ in a subset of the indices $\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}$ of an *additional code block* of n cells. The additional block is using the same coding technique as the original block. Therefore, it can use about $I(V;Y) - I(V;S)$ of the cells to store additional message bits, and by that to approach the channel capacity. We denote the by $\mathcal{R}$ the subset of $\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}$ in which we store the the subvector $u_{\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c}$ of the previous block. Note that the additional block also faces the same difficulty as the original block with the set $H_{V|S}^c \cap L_{V|Y}^c$. To solve this, we use the same solution, recursively, sending a total of k blocks, each of length n. Each block can store a source message of fraction that approaches the channel capacity. The "problematic" bits of block k (the last block) will then be stored using yet another block, but this block will be coded without taking the state information into account, and thus will not face the same difficulty. The last block is thus causing a rate loss, but this loss is of a fraction $1/k$, which vanishes for large k. The decoding is performed "backwards", starting from
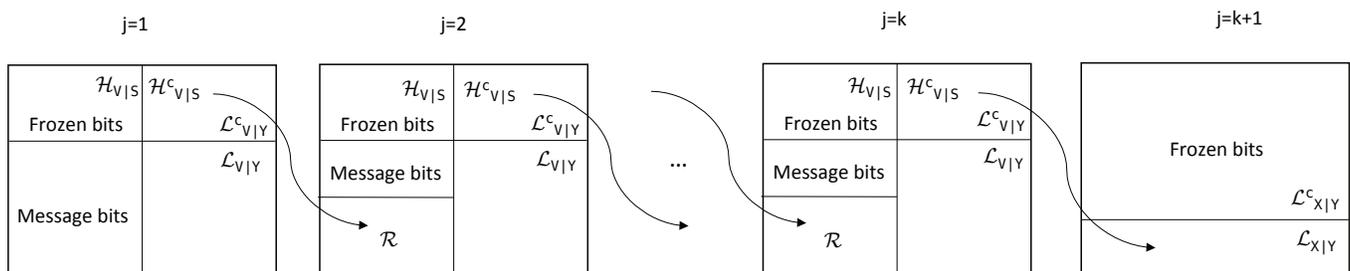

Fig. 5.   The chaining construction

the last block and ending with the first block. The chaining construction is illustrated in Figure 5. In the following formal description of the construction we denote the index i of the j-th block of the message by $m_{i,j}$, and similarly for other vectors. The vectors themselves are are also denoted in two dimensions, for example $x_{[n],[k]}$.

Construction 16.
Let $\mathcal{R}$ be an arbitrary subset of $\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}$ of size $|\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c|$.
Encoding
   Input: *a message* $m_{[(|\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}|-|\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c|],[k])} \in \{0,1\}^{k(|\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}|-|\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c|)}$ *and a state* $s_{[n],[k]} \in \{0,1\}^{kn}$.

   Output: *a codeword* $x_{[n],[k]} \in \{0,1\}^{kn}$.

1) *Let $u_{[n],0} \in \{0,1\}^n$ be an arbitrary vector. For each j from 1 to k, and for each i from 1 to n, assign*

$$
u_{i,j} = \begin{cases}
u \in \{0,1\} \quad \text{with probability } p_{U_i|U_{[i-1]},S_{[n]}}(u|u_{[i-1]},s_{[n]}) & \text{if } i \in \mathcal{H}_{V|S}^c \\
m_{r(i,\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}),j} & \text{if } i \in (\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}) \setminus \mathcal{R} \\
u_{r(i,\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c),j-1} & \text{if } i \in \mathcal{R} \\
f_{r(i,\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}^c),j} & \text{if } i \in \mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c.
\end{cases}
\tag{13}
$$

2) *For each j from 1 to k calculate $v_{[n],j} = u_{[n],j}G_n$, and for each $i \in [n]$, store the value $x_{i,j}(v_{i,j},s_{i,j})$.*
3) *Store the vector $u_{\mathcal{H}_V^c\cap\mathcal{L}_{V|Y}^c,k}$ separately using a point-to-point linear non-capacity-achieving polar code with a uniform input distribution. The encoder here does not use the state information in the encoding process, but rather treat it as an unknown part of the channel noise.*

Decoding
   Input: *a noisy vector* $y_{[n],[k]} \in \{0,1\}^k n$.

   Output: *a message estimation* $\hat{m}_{[|\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}|-|\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c|],[k]} \in \{0,1\}^{k|\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}|-|\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c|}$.

1) *Estimate the vector $u_{\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c,k}$ by $\hat{u}_{\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c,k}$, and let $\hat{u}_{\mathcal{R},k+1} = \hat{u}_{\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c,k}$.*
2) *Estimate $u_{[n],[k]}$ by $\hat{u}_{[n],[k]}(y_{[n],[k]}, f_{[|\mathcal{L}_{V|Y}^c\cap\mathcal{H}_{V|S}|],[k]})$ as follows: For each j down from k to 1, and for each i from 1 to n, assign*

$$
\hat{u}_i^j = \begin{cases}
\arg\max\limits_{u\in\{0,1\}} p_{U_i|U_{[i-1]},Y_{[n]}}(u|u_{[i-1],j},y_{[n],j}) & \text{if } i \in \mathcal{L}_{V|Y} \\
\hat{u}_{r(i,\mathcal{R}),j+1} & \text{if } i \in \mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c \\
f_{r(i,\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}^c),j} & \text{if } i \in \mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c.
\end{cases}
\tag{14}
$$

3) *Return the estimated message $\hat{m}_{[|\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y}|-|\mathcal{H}_{V|S}^c\cap\mathcal{L}_{V|Y}^c|],[k]} = \hat{u}_{(\mathcal{H}_{V|S}\cap\mathcal{L}_{V|Y})\setminus\mathcal{R},[k]}$.*

Constructions 14 and 16 can also be used for communication over broadcast channels in Marton's region, as described in [13], [23]. Constructions 14 and 16 improve on these previous results since they provably achieve the capacity with linear storage requirement.

Construction 16 achieves the capacity of Theorem 6 with low complexity, without the degradation requirement of Property 10. This result was stated in Theorem 7. The proof of Theorem 7 follows from Theorem 15 and the fact that the rate loss vanishes with large k. Construction 16 is useful for the realistic model of flash memory-rewriting of Example 8, using the appropriate capacity-achieving functions $p(v|s)$ and $x(v,s)$.

IV. CONCLUSION

In this paper we proposed three capacity-achieving polar coding schemes, for the settings of asymmetric channel coding and flash memory rewriting. The scheme for asymmetric channels improves on the scheme

of [17] by reducing the exponential storage requirement into a linear one. The idea for this reduction is to perform the encoding randomly instead of using Boolean functions, and to transmit a vanishing fraction of information on a side channel.

The second proposed scheme is used for the setting of flash memory rewriting. We propose a model of flash memory rewriting with writing noise, and show that the scheme achieves its capacity. We also describe a more general class of channels whose capacity can be achieved using the scheme. The second scheme is derived from the asymmetric-channel scheme by replacing the Shannon random variables X and $X|Y$ with the Gelfand-Pinsker random variables $V|S$ and $V|Y$.

The last proposed scheme achieves the capacity of any channel with non-causal state information at the encoder. We bring a model of noisy flash memory rewriting for which the scheme would be useful. The main idea in this scheme is called code chaining. Another potential application could be in information embedding (as in [3]), where the model is asymmetric

APPENDIX A

In this appendix we prove Theorem 12. A similar result was proven in [16, Theorem 4]. We show a different proof here, which we find more intuitive. Theorem 12 states that the capacity of the channel in Example 11 is

$$C = (1 - \beta)[h(\epsilon * \alpha) - h(\alpha)],$$

where $\epsilon = B/(1 - \beta)$ and $\epsilon * \alpha \equiv \epsilon(1 - \alpha) + (1 - \epsilon)\alpha$. An upper bound on the capacity can be obtained by assuming that the state information is available also to the decoder. In this case, the best coding scheme would ignore the cells with $s_i = 1$ (about a fraction β of the cells), and the rest of the cells would be coded according to a binary symmetric channel with an input cost constraint. It is optimal to assign a channel input $x_i = 0$ for the cells with state $s_i = 1$, such that those cells who do not convey information do not contribute to the cost. We now focus on the capacity of the binary symmetric channel with cost constraint. To comply with the expected input cost constraint B of the channel of Example 11, the expected cost of the input to the binary symmetric channel (BSC) must be at most $\epsilon = B/(1 - \beta)$. To complete the proof of the upper bound, we show next that the capacity of the BSC with cost constraint is equals to $h(\alpha * \epsilon) - h(\alpha)$. For this channel, we have

$$H(Y|X) = h(\alpha)p_X(0) + h(\alpha)p_X(1) = h(\alpha).$$

We are left now with maximizing the entropy $H(Y)$ over the input pmfs $p_X(1) \le \epsilon$. We have

$$\begin{aligned} p_Y(1) &= p_{Y|X}(1|0)p_X(0) + p_{Y|X}(1|1)p_X(1) \\ &= \alpha(1 - p_X(1)) + (1 - \alpha)p_X(1) \\ &= \alpha * p_X(1). \end{aligned}$$

Now since $p_X(1) \le \epsilon \le 1/2$ and $\alpha * p_X(1)$ is increasing in $p_X(1)$ below $1/2$, it follows that $p_Y(1) \le \alpha * \epsilon \le 1/2$ and therefore also that $H(Y) \le h(\alpha * \epsilon)$. So we have

$$\max_{p_X(1) \le \epsilon} I(X;Y) = \max_{p_X(1) \le \epsilon} (H(Y) - H(Y|X)) = h(\alpha * \epsilon) - h(\alpha).$$

This completes the proof of the upper bound.

The lower bound is obtained by considering the selections $\mathcal{V} = \{0, 1\}$, $x(v, 0) = v$, $x(v, 1) = 0$ and

$$p_{V|S}(1|0) = \epsilon, \qquad p_{V|S}(1|1) = \frac{\epsilon(1 - \alpha)}{\epsilon * \alpha}, \tag{15}$$

and calculating the rate expression directly. Notice first that the cost constraint is met since

$$p_X(1) = p_{X|S}(1|0)p_S(0) = p_{V|S}(1|0)p_S(0) = \epsilon(1 - \beta) = B.$$

We need to show that $H(V|S) - H(V|Y) = (1-\beta)[H(\alpha * \epsilon) - H(\alpha)]$. Given the distributions p_S and $p_{V|S}$, the conditional entropy $H(V|S)$ is

$$H(V|S) = \sum_{s\in\{0,1\}} p_S(s)H(V|S=s)$$
$$= p_S(0)H(V|S=0) + p_S(1)H(V|S=1)$$
$$= (1-\beta)H(\epsilon) + \beta H\left(\frac{\epsilon(1-\alpha)}{\epsilon * \alpha}\right)$$

To compute the conditional entropy $H(V|Y)$, we first compute the probability distribution of the memory output Y as follows:

$$p_Y(0) = \sum_{v\in\{0,1\}} p_{Y|VS}(0|v,0)p_{V|S}(v|0)p_S(0)$$
$$= (1-\beta)((1-\alpha)(1-\epsilon) + \alpha\epsilon)$$
$$= (1-\beta)(\alpha * (1-\epsilon)),$$
$$p_Y(1) = 1 - p_Y(0)$$
$$= (1-\beta)(\alpha * \epsilon) + \beta.$$

The conditional distribution $p_{V|Y}$ is given by

$$p_{V|Y}(1|0) = \sum_{s\in\{0,1\}} p_{VS|Y}(1,s|0)$$
$$= \sum_{s\in\{0,1\}} \frac{p_{Y|VS}(0|1,s)p_{VS}(1,s)}{p_Y(0)}$$
$$= \sum_{s\in\{0,1\}} \frac{p_{Y|VS}(0|1,s)p_{V|S}(1|s)p_S(s)}{p_Y(0)}$$
$$= \frac{\alpha\epsilon}{\alpha * (1-\epsilon)},$$
$$p_{V|Y}(1|1) = \sum_{s\in\{0,1\}} p_{VS|Y}(1,s|1)$$
$$= \sum_{s\in\{0,1\}} \frac{p_{Y|VS}(1|1,s)p_{VS}(1,s)}{p_Y(1)}$$
$$= \sum_{s\in\{0,1\}} \frac{p_{Y|VS}(1|1,s)p_{V|S}(1|s)p_S(s)}{p_Y(1)}$$
$$= \frac{(1-\alpha)\epsilon(1-\beta) + \frac{\epsilon(1-\alpha)}{\epsilon*\alpha}\beta}{(1-\beta)(\alpha * \epsilon) + \beta}$$
$$= \frac{\epsilon(1-\alpha)}{\epsilon * \alpha}.$$

Therefore we have

$$H(V|Y) = \sum_{y\in\{0,1\}} p_Y(y)H(V|Y=y)$$

$$=(1-\beta)(\alpha*(1-\epsilon))H\left(\frac{\alpha\epsilon}{\alpha*(1-\epsilon)}\right)+(\beta+(1-\beta)(\alpha*\epsilon))H\left(\frac{\epsilon(1-\alpha)}{\epsilon*\alpha}\right),$$

and then

$$H(V|S)-H(V|Y)=(1-\beta)\left[H(\epsilon)-(\alpha*(1-\epsilon))H\left(\frac{\alpha\epsilon}{\alpha*(1-\epsilon)}\right)-(\alpha*\epsilon)H\left(\frac{\epsilon(1-\alpha)}{\epsilon*\alpha}\right)\right]$$

$$=(1-\beta)\left[H(\epsilon)+\alpha\epsilon\log_2\frac{\alpha\epsilon}{\alpha*(1-\epsilon)}+(1-\alpha)(1-\epsilon)\log_2\frac{(1-\alpha)(1-\epsilon)}{\alpha*(1-\epsilon)}\right.$$

$$\left.+\alpha(1-\epsilon)\log_2\frac{\alpha(1-\epsilon)}{\alpha*\epsilon}+\epsilon(1-\alpha)\log_2\frac{\epsilon(1-\alpha)}{\alpha*\epsilon}\right]$$

$$=(1-\beta)[H(\alpha*\epsilon)+H(\epsilon)+\alpha\epsilon\log_2(\alpha\epsilon)+(1-\alpha)(1-\epsilon)\log_2(1-\alpha)(1-\epsilon)$$

$$+\alpha(1-\epsilon)\log_2\alpha(1-\epsilon)+\epsilon(1-\alpha)\log_2\epsilon(1-\alpha)]$$

$$=(1-\beta)[H(\alpha*\epsilon)+H(\epsilon)-H(\alpha)-H(\epsilon)]$$

$$=(1-\beta)[H(\alpha*\epsilon)-H(\alpha)].$$

APPENDIX B

In this appendix we prove Lemma 13. We need to show that, using the functions of Therem 12, there exists a DMC $W:\{0,1\}^2\to\{0,1\}^2$ such that

$$p_{S|V}(s|v)=\sum_{y\in\{0,1\}^2}p_{Y|V}(y|v)W(s|y). \tag{16}$$

To define such channel W, we first claim that

$$p_{Y|V,S}(1|v,0)p_{V|S}(x|0)=(\epsilon*\alpha)p_{V|S}(x|1). \tag{17}$$

Equation (17) follows directly from Equation (10) since

$$\frac{p_{Y|V,S}(1|0,0)p_{V|S}(0|0)}{p_{V|S}(0|1)}=\frac{\alpha(1-\epsilon)}{\frac{\alpha(1-\epsilon)}{\epsilon*\alpha}}=\epsilon*\alpha,$$

$$\frac{p_{Y|V,S}(1|1,0)p_{V|S}(1|0)}{p_{V|S}(1|1)}=\frac{(1-\alpha)\epsilon}{\frac{(1-\alpha)\epsilon}{\epsilon*\alpha}}=\epsilon*\alpha.$$

Next, we claim that $\frac{p_{V,S}(v,1)}{p_{V,Y}(v,1)}=\frac{\beta}{(\epsilon*\alpha)(1-\beta)+\beta}$ for any $v\in\{0,1\}$, and therefore that $\frac{p_{V,S}(v,1)}{p_{V,Y}(v,1)}\in[0,1]$. This follows from

$$\frac{p_{V,S}(v,1)}{p_{V,Y}(v,1)}\overset{(a)}{=}\frac{p_{V|S}(v|1)p_S(1)}{p_{Y,V|S}(1,v|0)p_S(0)+p_{Y,V|S}(1,v|1)p_S(1)}$$

$$\overset{(b)}{=}\frac{p_{V|S}(v|1)\beta}{p_{Y|V,S}(1|v,0)p_{V|S}(v|0)(1-\beta)+p_{Y|V,S}(1|v,1)p_{V|S}(x|1)\beta}$$

$$\overset{(c)}{=}\frac{p_{V|S}(v|1)\beta}{(\epsilon*\alpha)p_{V|S}(v|1)(1-\beta)+p_{V|S}(v|1)\beta}$$

$$=\frac{\beta}{(\epsilon*\alpha)(1-\beta)+\beta},$$

where (a) follows from the law of total probability, (b) follows from the definition of conditional probability, and (c) follows from Equations (9) and (17).

Since $\frac{p_{V,S}(v,1)}{p_{V,Y}(v,1)}$ is not a function of x and is in $[0,1]$, we can define W as following:

$$W(s|y) \triangleq \begin{cases} 1 & \text{if } (s,y) = (0,0) \\ 1 - \frac{p_{S|V}(1|v)}{p_{Y|V}(1|v)} & \text{if } (s,y) = (0,1) \\ \frac{p_{S|V}(1|v)}{p_{Y|V}(1|v)} & \text{if } (s,y) = (1,1) \\ 0 & \text{if } (s,y) = (1,0). \end{cases}$$

We show next that Equation (16) holds for W defined above:

$$\sum_{y \in \{0,1\}} p_{Y|V}(y|v) W(s|y) = p_{Y|V}(0|v) W(s|0) + p_{Y|V}(1|v) W(s|1)$$

$$= \left[p_{Y|V}(0|v) + p_{Y|V}(1|v) \left(1 - \frac{p_{S|V}(1|v)}{p_{Y|V}(1|v)} \right) \right] \mathbb{1}[s = 0] + p_{S|V}(1|v) \mathbb{1}[s = 1]$$

$$= \left[1 - p_{S|V}(1|v) \right] \mathbb{1}[s = 0] + p_{S|V}(1|v) \mathbb{1}[s = 1]$$

$$= p_{S|V}(0|v) \mathbb{1}[s = 0] + p_{S|V}(1|v) \mathbb{1}[s = 1]$$

$$= p_{S|V}(s|v).$$

So the channel W satisfies Equation (16) and thus the lemma holds.

APPENDIX C

In this appendix we prove Theorem 15. The complexity claim of Theorem 15 is explained in [17, Section III.B]. We start with the asymptotic rate of Construction 14. We want to show that $\lim_{n \to \infty} (1/n) |\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c| = 0$. Since $p_{S|V}$ is degraded with respect to $p_{Y|V}$, it follows from [13, Lemma 4] that $\mathcal{L}_{V|S} \subseteq \mathcal{L}_{V|Y}$, and therefore that $\mathcal{L}_{V|S}^c \supseteq \mathcal{L}_{V|Y}^c$. So we have

$$\lim_{n \to \infty} (1/n) |\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|Y}^c| \le \lim_{n \to \infty} (1/n) |\mathcal{H}_{V|S}^c \cap \mathcal{L}_{V|S}^c| = 0,$$

where the equality follows from the definition of the sets. To complete the proof of the theorem, we need to show that for some frozen vector, the input cost meets the design constraint, while the decoding error probability vanishes sub-exponentially fast in the block length. We show this result using the probabilistic method. That is, we first assume that the frozen vector is random, drawn uniformly, and analyze the input cost and error probability in this case, and then we show the existence of a good vector using Markov's inequalities and the union bound. Denote the uniformly distributed random vector that represents the frozen vector by $F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}$. We show next that the expected input cost exceeds the design constraint B by a vanishing amount.

A. Expected Input Cost

Define

$$b^n(X_{[n]}) = \sum_{i=1}^{n} b(X_i).$$

For a state vector $s_{[n]}$ and the encoding rule (11) with uniform frozen vector $F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}$, each vector $u_{[n]}$ appears with probability

$$\left(\prod_{i \in \mathcal{H}_{V|S}^c} p_{U_i|U_{[i-1]},S_{[n]}}(u_i|u_{[i-1]}, s_{[n]}) \right) 2^{-|\mathcal{H}_{V|S}|}.$$

Remember that the joint pmf $p_{S_{[n]},U_{[n]}}$ refers to the capacity-achieving distribution of the channel. The expected cost is expressed as

$$\mathsf{E}_{F_{[|\mathcal{H}_{V|S}\cap\mathcal{L}^c_{V|Y}|]}}(b^n(X_{[n]})) = \sum_{u_{[n]},s_{[n]}} p_{S_{[n]}}(s_{[n]}) \left(\prod_{i\in\mathcal{H}^c_{V|S}} p_{U_i|U_{[i-1]},S_{[n]}}(u_i|u_{[i-1]},s_{[n]}) \right) 2^{-|\mathcal{H}_{V|S}|} b^n(u_{[n]}G_n).$$

Define the joint pmf

$$q_{S_{[n]},U_{[n]}} \equiv p_{S_{[n]}}(s_{[n]}) \left(\prod_{i\in\mathcal{H}^c_{V|S}} p_{U_i|U_{[i-1]},S_{[n]}}(u_i|u_{[i-1]},s_{[n]}) \right) 2^{-|\mathcal{H}_{V|S}|}. \tag{18}$$

Intuitively, $q_{S_{[n]},U_{[n]}}$ refers to the distribution imposed by the encoding procedure of Construction 14. Then we have

$$\begin{aligned}
\mathsf{E}_{F_{[|\mathcal{H}_{V|S}\cap\mathcal{L}^c_{V|Y}|]}}(b^n(X_{[n]})) &= \mathsf{E}_{q_{S_{[n]},U_{[n]}}}[b^n(U_{[n]}G_n)] \\
&\leq \mathsf{E}_{p_{S_{[n]},U_{[n]}}}[b^n(U_{[n]}G_n)] + \max_x b(x) \|p_{S_{[n]},U_{[n]}} - q_{S_{[n]},U_{[n]}}\| \\
&= nB + \max_x b(x) \|p_{S_{[n]},U_{[n]}} - q_{S_{[n]},U_{[n]}}\|,
\end{aligned}$$

where $\|\cdot\|$ is the L_1 distance and the inequality follows from the triangle inequality. We will now prove that

$$\mathsf{E}_{F_{[|\mathcal{H}_{V|S}\cap\mathcal{L}^c_{V|Y}|]}}(b^n(X_{[n]})) \leq nB + 2^{-n^{1/2-\delta}},$$

by showing that $\|p_{U_{[n]},S_{[n]}} - q_{U_{[n]},S_{[n]}}\| \leq 2^{-n^{1/2-\delta}}$. To do this, we will prove a slightly stronger relation that will be used also for the proof of the probability of decoding error. We first define the joint pmf

$$q_{S_{[n]},U_{[n]},Y_{[n]}} \equiv q_{U_{[n]},S_{[n]}}(u_{[n]},s_{[n]})p_{Y_{[n]}|U_{[n]},S_{[n]}}(y_{[n]}|u_{[n]},s_{[n]}).$$

Then we notice that

$$\begin{aligned}
\|p_{U_{[n]},S_{[n]}} - q_{U_{[n]},S_{[n]}}\| &= \sum_{u_{[n]},s_{[n]}} |p(u_{[n]},s_{[n]}) - q(u_{[n]},s_{[n]})| \\
&= \sum_{u_{[n]},s_{[n]}} \left| \sum_{y_{[n]}} [p(s_{[n]},u_{[n]},y_{[n]}) - q(s_{[n]},u_{[n]},y_{[n]})] \right| \\
&\leq \sum_{s_{[n]},u_{[n]},y_{[n]}} |p(s_{[n]},u_{[n]},y_{[n]}) - q(s_{[n]},u_{[n]},y_{[n]})| \\
&= \|p_{S_{[n]},U_{[n]},Y_{[n]}} - q_{S_{[n]},U_{[n]},Y_{[n]}}\|,
\end{aligned}$$

where the inequality follows from the triangle inequality. The proof of the expected cost is completed with the following lemma, which will be used also for analyzing the probability of decoding error.

Lemma 17.

$$\|p_{S_{[n]},U_{[n]},Y_{[n]}} - q_{S_{[n]},U_{[n]},Y_{[n]}}\| \leq 2^{-n^{1/2-\delta}}. \tag{19}$$

Proof: Let $D(\cdot\|\cdot)$ denote the relative entropy. Then

$$\begin{aligned}
&\|p_{S_{[n]},U_{[n]},Y_{[n]}} - q_{S_{[n]},U_{[n]},Y_{[n]}}\| \\
&\quad = \sum_{s_{[n]},u_{[n]},y_{[n]}} |p(s_{[n]},u_{[n]},y_{[n]}) - q(s_{[n]},u_{[n]},y_{[n]})|
\end{aligned}$$

$$\stackrel{(a)}{=} \sum_{s_{[n]},u_{[n]},y_{[n]}} |p(u_{[n]}|s_{[n]}) - q(u_{[n]}|s_{[n]})| p(s_{[n]}) p(y_{[n]}|u_{[n]},s_{[n]})$$

$$\stackrel{(b)}{=} \sum_{s_{[n]},u_{[n]},y_{[n]}} \left| \prod_{i=1}^{n} p(u_i|u_{[i-1]},s_{[n]}) - \prod_{i=1}^{n} q(u_i|u_{[i-1]},s_{[n]}) \right| p(s_{[n]}) p(y_{[n]}|u_{[n]},s_{[n]})$$

$$\stackrel{(c)}{=} \sum_{s_{[n]},u_{[n]},y_{[n]}} \left| \sum_i [p(u_i|u_{[i-1]},s_{[n]}) - q(u_i|u_{[i-1]},s_{[n]})] \right| p(s_{[n]}) p(y_{[n]}|u_{[n]},s_{[n]})$$

$$\cdot \prod_{j=1}^{i-1} p(u_j|u_{[j-1]},s_{[n]}) \prod_{j=i+1}^{n} q(u_j|u_{[j-1]},s_{[n]})$$

$$\stackrel{(d)}{\leq} \sum_{i \in \mathcal{H}_{V|S}} \sum_{s_{[n]},u_{[n]},y_{[n]}} \left| p(u_i|u_{[i-1]},s_{[n]}) - q(u_i|u_{[i-1]},s_{[n]}) \right| p(s_{[n]}) \prod_{j=1}^{i-1} p(u_j|u_{[j-1]},s_{[n]})$$

$$\cdot \prod_{j=i+1}^{n} q(u_j|u_{[j-1]},s_{[n]}) p(y_{[n]}|u_{[n]},s_{[n]})$$

$$= \sum_{i \in \mathcal{H}_{V|S}} \sum_{s_{[n]},u_{[i]}} \left| p(u_i|u_{[i-1]},s_{[n]}) - q(u_i|u_{[i-1]},s_{[n]}) \right| \prod_{j=1}^{i-1} p(u_j|u_{[j-1]},s_{[n]}) p(s_{[n]})$$

$$\stackrel{(e)}{=} \sum_{i \in \mathcal{H}_{V|S}} \sum_{s_{[n]},u_{[i-1]}} p(u_{[i-1]},s_{[n]}) \| p_{U_i|U_{[i-1]}=u_{[i-1]},S_{[n]}=s_{[n]}} - q_{U_i|U_{[i-1]}=u_{[i-1]},S_{[n]}=s_{[n]}} \|$$

$$\stackrel{(f)}{\leq} \sum_{i \in \mathcal{H}_{V|S}} \sum_{s_{[n]},u_{[i-1]}} p(u_{[i-1]},s_{[n]}) \sqrt{2\ln 2} \sqrt{D(p_{U_i|U_{[i-1]}=u_{[i-1]},S_{[n]}=s_{[n]}} \| q_{U_i|U_{[i-1]}=u_{[i-1]},S_{[n]}=s_{[n]}})}$$

$$\stackrel{(g)}{\leq} \sum_{i \in \mathcal{H}_{V|S}} \sqrt{(2\ln 2) \sum_{s_{[n]},u_{[i-1]}} p(u_{[i-1]},s_{[n]}) D(p_{U_i|U_{[i-1]}=u_{[i-1]},S_{[n]}=s_{[n]}} \| q_{U_i|U_{[i-1]}=u_{[i-1]},S_{[n]}=s_{[n]}})}$$

$$= \sum_{i \in \mathcal{H}_{V|S}} \sqrt{(2\ln 2) D(p_{U_i} \| q_{U_i}|U_{[i-1]},S_{[n]})}$$

$$\stackrel{(h)}{=} \sum_{i \in \mathcal{H}_{V|S}} \sqrt{(2\ln 2)[1 - H(U_i|U_{[i-1]},S_{[n]})]},$$

where

(a) follows from the fact that $p(s_{[n]}) = q(s_{[n]})$ and $p(y_{[n]}|u_{[n]},s_{[n]}) = q(y_{[n]}|u_{[n]},s_{[n]})$,

(b) follows from the chain rule,

(c) follows from the telescoping expansion

$$B_{[n]} - A_{[n]} = \sum_{i=1}^{n} A_{[i-1]} B_{[i:n]} - \sum_{i=1}^{n} A_{[i]} B_{[i+1:n]}$$

$$= \sum_{i=1}^{n} (B_i - A_i) A_{[i-1]} B_{[i+1:n]},$$

where $A_{[j:k]}$ and $B_{[j:k]}$ denote the products $\prod_{i=j}^{k} A_i$ and $\prod_{i=j}^{k} B_i$, respectively,

(d) follows from the triangular inequality and the fact that $p(u_i|u_{[i-1]},s_{[n]}) = q(u_i|u_{[i-1]},s_{[n]})$ for all $i \in \mathcal{H}_{V|S}^c$ (according to Equation (18)),

(e) follows from the chain rule again,

(f) follows from Pinsker's inequality (see, e.g., [6, Lemma 11.6.1]),

(g) follows from Jensen's inequality and

(h) follows from the facts that $q(u_i|u_{[i-1]}, s_{[n]}) = 1/2$ for $i \in \mathcal{H}_{V|S}$ and from [13, Lemma 10].

Now if $i \in \mathcal{H}_{V|S}$, we have

$$
\begin{aligned}
1 - H(U_i|U_{[i-1]}, S_{[n]}) &\leq 1 - [Z(U_i|U_{[i-1]}, S_{[n]})]^2 \\
&\leq 2^{-2n^{1/2-\delta}},
\end{aligned}
\tag{20}
$$

where the first inequality follows from Proposition 3, and the second inequality follows from the fact that i is in $\mathcal{H}_{V|S}$. This completes the proof of the lemma. ∎

B. Probability of Decoding Error

Let $\mathcal{E}_i$ be the set of pairs of vectors $(u_{[n]}, y_{[n]})$ such that $\hat{u}_{[n]}$ is a result of decoding $y_{[n]}$, and $\hat{u}_{[i]}$ satisfies both $\hat{u}_{[i-1]} = u_{[i-1]}$ and $\hat{u}_i \neq u_i$. The block decoding error event is given by $\mathcal{E} \equiv \cup_{i \in \mathcal{L}_{V|Y}} \mathcal{E}_i$. Under decoding given in (12) with an arbitrary tie-breaking rule, every pair $(u_{[n]}, y_{[n]}) \in \mathcal{E}_i$ satisfies

$$
p_{U_i|U_{[i-1]}, Y_{[n]}}(u_i|u_{[i-1]}, y_{[n]}) \leq p_{U_i|U_{[i-1]}, Y_{[n]}}(u_i \oplus 1|u_{[i-1]}, y_{[n]}).
\tag{21}
$$

Consider the block decoding error probability $p_e(F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]})$ for the random frozen vector $F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}$. For a state vector $s_{[n]}$ and the encoding rule (11), each vector $u_{[n]}$ appears with probability

$$
\left(\prod_{i \in \mathcal{H}_{V|S}^c} p_{U_i|U_{[i-1]}, S_{[n]}}(u_i|u_{[i-1]}, s_{[n]}) \right) 2^{-|\mathcal{H}_{V|S}|}.
$$

By the definition of conditional probability and the law of total probability, the probability of error $p_e(F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]})$ is given by

$$
\begin{aligned}
\mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}[p_e] = \sum_{u_{[n]}, s_{[n]}, y_{[n]}} p_{S_{[n]}}(s_{[n]}) &\left(\prod_{i \in \mathcal{H}_{V|S}^c} p_{U_i|U_{[i-1]}, S_{[n]}}(u_i|u_{[i-1]}, s_{[n]}) \right) 2^{-|\mathcal{H}_{V|S}|} \\
&\cdot p_{Y_{[n]}|U_{[n]}, S_{[n]}}(y_{[n]}|u_{[n]}, s_{[n]}) \mathbb{1}[(u_{[n]}, y_{[n]}) \in \mathcal{E}].
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}[p_e] &= q_{U_{[n]}, Y_{[n]}}(\mathcal{E}) \\
&\leq \|q_{U_{[n]}, Y_{[n]}} - p_{U_{[n]}, Y_{[n]}}\| + p_{U_{[n]}, Y_{[n]}}(\mathcal{E}) \\
&\leq \|q_{U_{[n]}, Y_{[n]}} - p_{U_{[n]}, Y_{[n]}}\| + \sum_{i \in \mathcal{L}_{V|Y}} p_{U_{[n]}, Y_{[n]}}(\mathcal{E}_i),
\end{aligned}
$$

where the first inequality follows from the triangle inequality. Each term in the summation is bounded by

$$
\begin{aligned}
p_{U_{[n]}, Y_{[n]}}(\mathcal{E}_i) &\leq \sum_{u_{[i]}, y_{[n]}} p(u_{[i]}, y_{[n]}) \mathbb{1}\left[p(u_i|u_{[i-1]}, y_{[n]}) \leq p(u_i \oplus 1|u_{[i-1]}, y_{[n]}) \right] \\
&\leq \sum_{u_{[i]}, y_{[n]}} p(u_{[i-1]}, y_{[n]}) p(u_i|u_{[i-1]}, y_{[n]}) \sqrt{\frac{p(u_i \oplus 1|u_{[i-1]}, y_{[n]})}{p(u_i|u_{[i-1]}, y_{[n]})}} \\
&= Z(U_i|U_{[i-1]}, Y_{[n]}) \\
&\leq 2^{-n^{1/2-\delta}},
\end{aligned}
$$

where the last inequality follows from the fact that i belongs to the set $\mathcal{L}_{V|Y}$.

To prove that $\mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}[p_e] \leq 2^{-n^{1/2-\delta'}}$ for some $\delta' > \delta$, we are left with showing that $\|p_{U_{[n]},Y_{[n]}} - q_{U_{[n]},Y_{[n]}}\| \leq 2^{-n^{1/2-\delta}}$. Notice that

$$
\begin{aligned}
2\|p_{U_{[n]},Y_{[n]}} - q_{U_{[n]},Y_{[n]}}\| &= \sum_{u_{[n]},y_{[n]}} |p(u_{[n]},y_{[n]}) - q(u_{[n]},y_{[n]})| \\
&= \sum_{u_{[n]},y_{[n]}} \left| \sum_{s_{[n]}} [p(s_{[n]},u_{[n]},y_{[n]}) - q(s_{[n]},u_{[n]},y_{[n]})] \right| \\
&\leq \sum_{s_{[n]},u_{[n]},y_{[n]}} |p(s_{[n]},u_{[n]},y_{[n]}) - q(s_{[n]},u_{[n]},y_{[n]})| \\
&= 2\|p_{S_{[n]},U_{[n]},Y_{[n]}} - q_{S_{[n]},U_{[n]},Y_{[n]}}\|,
\end{aligned}
$$

where the inequality follows from the triangle inequality. Lemma 17 now completes the proof that $\mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}[p_e] = 2^{-n^{1/2-\delta}}$.

C. Existence of a Good Frozen Vector

We showed that

$$
\mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}(b^n(X_{[n]})) \leq nB + 2^{-n^{1/2-\delta}}.
$$

In this subsection we will be interested to find a frozen vector which satisfies a slightly higher expected cost, namely

$$
\mathsf{E}(b^n(X_{[n]})) \leq nB + 1/n^2.
$$

Remember that B is a constant. We take a uniformly distributed frozen vector, and bound the probability that the expected cost exceeds $nB + 1/n^2$, using Markov's inequality. The following inequalities hold for large enough n, and we are not concerned here with the exact values required.

$$
\begin{aligned}
P(\mathsf{E}(b^n(X_{[n]})) \geq nB + 1/n^2) &\leq \mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}(b^n(X))/(nB + 1/n^2) \\
&\leq (nB + 2^{-n^{1/2-\delta}})/(nB + 1/n^2) \\
&\leq (nB + n^{-3})/(nB + n^{-2}) \\
&= 1 - (n^{-2} - n^{-3})/(nB + n^{-2}) \\
&\leq 1 - (n^{-2.1})/(nB + n^{-2}) \\
&\leq 1 - 1/(Bn^{3.1} + n^{0.1}) \\
&\leq 1 - 1/(Bn^{3.2}) \\
&\leq 1 - 1/n^4.
\end{aligned}
$$

We now apply Markov's inequality on the probability of decoding error:

$$
P(p_e \geq n^5 2^{-n^{1/2-\delta}}) \leq \frac{\mathsf{E}_{F_{[|\mathcal{H}_{V|S} \cap \mathcal{L}_{V|Y}^c|]}}(p_e)}{n^5 2^{-n^{1/2-\delta}}} \leq \frac{2^{-n^{1/2-\delta}}}{n^5 2^{-n^{1/2-\delta}}} = 1/n^5.
$$

By the union bound, the probability that either $\mathsf{E}(b^n(X_{[n]})) \geq nB + 1/n^2$ or $p_e \geq n^5 2^{-n^{1/2-\delta}}$ is at most

$$
1 - 1/n^4 + 1/n^5 \leq 1 - 1/n^5.
$$

This implies that the probability that both $\mathsf{E}(b^n(X_{[n]})) \leq nB + 1/n^2$ and $p_e \leq n^5 2^{-n^{1/2-\delta}}$ is at least $1/n^5$. So there exists at least one frozen vector for which the desired properties for both the expected

cost and the probability of error hold. Furthermore, such a vector can be found by repeatedly selecting a frozen vector uniformly at random, until the required properties hold. The properties can be verified efficiently with close approximations by the upgradation and degradation method proposed in [31]. The expected number of times until a good vector is found is polynomial in the block length (at most n^5). This completes the proof of Theorem 15.

REFERENCES

[1] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theor.*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[2] ——, "Source polarization," in *Proc. IEEE Int. Symp. on Inf. Theor. (ISIT)*, June 2010, pp. 899–903.

[3] R. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Tran. on Inf. Theor.*, vol. 49, no. 5, pp. 1159–1180, May 2003.

[4] D. Burshtein and A. Strugatski, "Polar write once memory codes," *IEEE Trans. Inf. Theor.*, vol. 59, no. 8, pp. 5088–5101, Aug. 2013.

[5] Y. Cassuto, M. Schwartz, V. Bohossian, and J. Bruck, "Codes for asymmetric limited-magnitude errors with application to multilevel flash memories," *IEEE Trans. on Inf. Theor.*, vol. 56, no. 4, pp. 1582–1595, April 2010.

[6] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006.

[7] A. El Gamal and Y. Kim, *Network information theory*. Cambridge University Press, 2012.

[8] E. En Gad, Y. Li, J. Kliewer, M. Langberg, A. Jiang, and J. Bruck, "Polar coding for noisy write-once memories," in *Proc. IEEE Int. Symp. on Inf. Theor. (ISIT)*, June 2014, pp. 1638–1642.

[9] E. En Gad, E. Yaakobi, A. Jiang, and J. Bruck, "Rank-modulation rewrite coding for flash memories," *Available at http://arxiv.org/abs/1312.0972*, Dec 2013.

[10] A. Gabizon and R. Shaltiel, "Invertible zero-error dispersers and defective memory with stuck-at errors," in *APPROX-RANDOM*, 2012, pp. 553–564.

[11] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.

[12] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Problems of Control Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[13] N. Goela, E. Abbe, and M. Gastpar, "Polar codes for broadcast channels," *Available at http://arxiv.org/abs/1301.6150*, Jan 2013.

[14] S. Golomb, "The limiting behavior of the z-channel (corresp.)," *IEEE Trans. on Inf. Theor.*, vol. 26, no. 3, pp. 372–372, May 1980.

[15] J. Gordon, "Quantum effects in communications systems," *Proceedings of the IRE*, vol. 50, no. 9, pp. 1898–1908, Sept 1962.

[16] C. Heegard, "On the capacity of permanent memory," *IEEE Trans. Inf. Theor.*, vol. 31, no. 1, pp. 34–42, January 1985.

[17] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Trans. Inf. Theor.*, vol. 59, no. 12, pp. 7829–7838, 2013.

[18] S. Korada and R. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Inf. Theor.*, vol. 56, no. 4, pp. 1751–1768, April 2010.

[19] A. Kusnetsov and B. S. Tsybakov, "Coding in a memory with defective cells,," *translated from Problemy Peredachi Informatsii*, vol. 10, no. 2, pp. 52–60, 1974.

[20] X. Ma, "Write-once-memory codes by source polarization," *Available at http://arxiv.org/abs/1405.6262*, May 2014.

[21] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. on Inf. Theor.*, vol. 25, no. 3, pp. 306–311, May 1979.

[22] M. Mondelli, I. Hassani, and R. Urbanke, "How to achieve the capacity of asymmetric channels," *Available at http://arxiv.org/abs/1406.7373*, June 2014.

[23] M. Mondelli, S. H. Hassani, I. Sason, and R. Urbanke, "Achieving Marton's region for broadcast channels using polar codes," *Available at http://arxiv.org/abs/1401.6060*, January 2014.

[24] R. Mori and T. Tanaka, "Channel polarization on q-ary discrete memoryless channels by arbitrary kernels," in *IEEE Int. Symp. on Inf. Theor. (ISIT)*, June 2010, pp. 894–898.

[25] W. Park and A. Barg, "Polar codes for q-ary channels, $q = 2^r$," *IEEE Trans. Inf. Theor.*, vol. 59, no. 2, pp. 955–969, Feb 2013.

[26] A. Sahebi and S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, Sept 2011, pp. 1718–1725.

[27] E. Sasoglu, "Polar codes for discrete alphabets," in *Proc. IEEE Int. Symp. on Inf. Theor. (ISIT)*, July 2012, pp. 2137–2141.

[28] E. Sasoglu, I. Telatar, and E. Arikan, "Polarization for arbitrary discrete memoryless channels," in *IEEE Information Theory Workshop (ITW)*, Oct 2009, pp. 144–148.

[29] E. Sasoglu, "Polarization and polar codes," *Foundations and Trends in Communications and Information Theory*, vol. 8, no. 4, pp. 259–381, 2012. [Online]. Available: http://dx.doi.org/10.1561/0100000041

[30] D. Sutter, J. Renes, F. Dupuis, and R. Renner, "Achieving the capacity of any dmc using only polar codes," in *IEEE Information Theory Workshop (ITW)*, Sept 2012, pp. 114–118.

[31] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Trans. on Inf. Theor.*, vol. 59, no. 10, pp. 6562–6582, Oct 2013.

[32] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theor.*, vol. 48, no. 6, pp. 1250–1276, Jun 2002.