RC25386 (WAT1305-086) May 29, 2013
Computer Science

IBM Research Report

On the Capacity of Memoryless Rewritable Storage Channels

Luis A. Lastras-Montaño[1], Michele M. Franceschini[1], Thomas Mittelholzer[2], Mayank Sharma[1]

[1]IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

[2]IBM Research Division - Zurich
Säumerstrasse 4
8803 Rüschlikon
Switzerland

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

On the Capacity of Memoryless Rewritable Storage Channels

Luis A. Lastras-Montaño, Michele M. Franceschini, Thomas Mittelholzer, Mayank Sharma

Abstract

A significant number of modern storage technologies, when written to, exhibit substantial variability in the outcomes of a write action. It is possible to mitigate the effect of the write uncertainty through the use of a feedback loop that rewrites the memory whenever judged necessary, in effect reshaping the write noise. This scheme highlights a trade-off between the storage capacity of the memory and the cost of writing to it, measured for example in the number of rewrites. The authors have developed the model of a *rewritable channel* to provide an explicit form for this trade-off and study other performance characteristics of such memories.

In this paper, we describe some initial results on the information-theoretic analysis of the rewritable channel. We first consider the problem of determining the capacity of this channel with input cost constraints, and obtain a variety of results from which we extract insights that we believe are of value to memory designers. Our results include an upper bound on capacity of the form $\log(\Gamma\kappa)$ where Γ is a constant that can be easily calculated from the channel's statistics and κ is an average cost parameter. We also provide a lower bound on capacity with a similar form. We analyze the particular case of uniform write noise in detail, obtaining a closed form expression for the capacity-cost trade-off for all possible cost parameters. We explore this formula from the *capacity per unit cost* perspective and establish that in order to achieve optimal energy and memory-wear per bit, it is sometimes strictly better to take advantage of the rewriting capability as opposed to writing only once; this observation has significant practical implications. We also include a discussion of the relevance of our work to three real emerging memory technologies.

I. Introduction

Memory technology developments have delivered to us media where uncertainty during the process of writing to the memory is a major consideration. To attack the problem of write uncertainty, memory designers have adopted the strategy of using feedback information gathered during the memory writing procedure. Obtaining the feedback information has an associated cost, and the memory designer's job is to carefully trade-off this cost to obtain desired memory capacity, performance, endurance and energy usage attributes.

Examples of memory technologies with a write process that has outcomes only statistically determined include phase change memory (PCM, [6]), spin-transfer torque magnetic RAM (STT-RAM, [7]) and, in general, various kinds of resistive RAM (RRAM, see for example [8], [9], [10]). The most prominent example of such a memory technology is flash, an enormously successful medium that relies heavily on feedback to attack write uncertainty even when only a single bit per memory cell is to be written.

In this article we aim to provide a few basic elements for the construction of a Shannon theory that is aimed at understanding write uncertainty in memories. Inspired by Shannon's own approach to reliable communication, we adopt a minimalist approach which focuses on a class of simple memory models and write controllers with the goal of establishing a foundation for later analysis of models of more relevance to real memory technologies. As a result of this methodology, our model and results will not be directly applicable to flash; nonetheless we will argue that despite our model's simplicity, we are able to partially capture some elements of other memory technologies such as PCM and STT-RAM.

Our work will illustrate a general trend in the class of rewritable memories and class of rewrite techniques that we consider. For analog media, this trend can be summarized by saying that storage capacity increases approximately logarithmically as a function of the effort spent writing to the memory. This result holds up with reasonable generality; as a matter of fact in some cases it holds exactly.

Our first result is an upper bound on storage capacity of the form

$$C(\kappa) \leq \log(\Gamma\kappa),$$

where κ measures the cost, in average number of write attempts, to write to the memory and where Γ is a parameter that depends on the shape of the noise distribution.

To complement this upper bound, we prove a lower bound based on a novel application of the principle of superposition coding, and the associated decoding by interference cancellation, to memories. This result states that for $\kappa_1 > \kappa_0$,

$$C(\kappa_1) \geq C(\kappa_0) + \log\left(\frac{\kappa_1}{\kappa_0}\right).$$

Some of the contents of this paper have appeared previously in conferences; see the references [1], [2], [3], [4], [5].

Luis A. Lastras-Montaño, Michele M. Franceschini and Mayank Sharma are with the IBM TJ Watson Research Center
Thomas Mittelholzer is with the IBM Zurich Research Laboratory

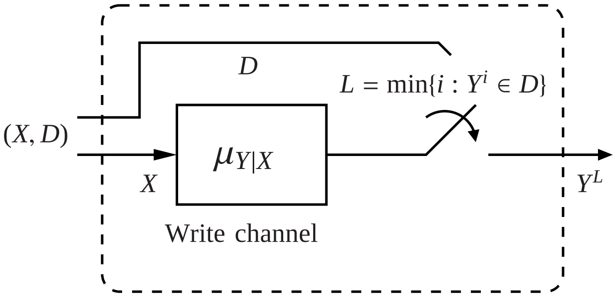

Fig. 1. Our model for rewriting in one memory cell. The statistical relation between the output Y^L and the input (X, D) is called the iterated channel.

In our work $\kappa_0 = 1$ corresponds to the classical notion of channel capacity with no rewrites (write only once) and therefore the result gives a lower bound on the rewrite channel capacity for any channel for which we know the classical storage capacity.

We also consider the specific case where a memory cell introduces noise that is uniformly distributed in some interval. In this case, we are able to obtain a closed form expression for channel capacity that shows the existence of a ”critical cost” beyond which capacity grows logarithmically but below which capacity grows slightly faster than a logarithm. For this model, we also show that sometimes it is possible to improve memory density and the cost per written bit simultaneously. In these settings, rewriting can be used in a memory not only as a tool to increase memory density, but also to reduce the cost of writing per user bit.

We close the introduction by giving a description of other works which are of relevance to our work, and a description of the organization of the article.

A. Related work

The information-theoretic study of the capacity/cost trade-off in memories created by write uncertainty appears to be unexplored prior to the article [1]. The present journal article in essence subsumes the results presented in [1], [2], [3], [4], [5] with the exception that the lower bound on capacity for the data-dependent model in [2] is not covered in detail in our results. The special case of two writes can be studied using the concept of channels with action dependent states which was introduced in [11]; in fact in this article the author explicitly gives results for a two write rewritable binary symmetric channel with possible corrupted feedback. Furthermore, the results in [11] could conceivably be used to study other two-stage versions of useful problems with physical relevance to rewritable memories. Other explorations of this subject treated the case of maximum number of iterations [12], [13], the discrete memoryless case and the case of noisy observations [13], the problem of computing rewritable channel capacity with maximum number of iterations using an efficient dynamic programming algorithm [14] and the problem of hidden states in rewritable channels for the uniform noise case [15].

In his article on zero error communication systems [16], Shannon studied what is now considered the classical notion of feedback is from a receiver back to the sender, proving the well known result that feedback does not increase channel capacity. Both physically and mathematically, the kind of feedback we consider is different from the classical notion of feedback, even if one were to assign a cost to the feedback. The essential difference is that in classical feedback, the receiver has access to all the data that is being sent by the transmitter, while in our model, overwritten data never reaches a receiver.

An early work treating the type of tradeoff that we explore in our article can be found in [17], where Cassutto et. al. introduced novel coding techniques for asymmetric errors that can be used for creating a trade-off between effective information density in a memory and the cost of writing to it. Our work is focused on the information-theoretic aspects of this type of trade-off, with an emphasis the simplest models of write uncertainty for which we can make non-trivial assertions.

B. Organization of the article

In Section II we lay the basic concepts for the study of i.i.d. rewritable channels with statistically independent write outcomes, including the definition of the capacity of the rewritable channel under an average number of write iterations constraint. The first goal of Section III is to give an information theoretic expression for such capacity, which is found in Theorem 1. This expression serves as a foundation for the rest of the article. Also in Section III we introduce the *uniform noise* rewritable channel model, which can be regarded as the simplest continuous input/output rewritable channel that one might consider. Some simple assertions about the capacity of this storage channel are made, with a full treatment of it deferred to Section VI.

The article then follows with a general upper bound on capacity (Section IV) and a general lower bound on capacity (Section V). In both of these sections as well as in Section VI, we have adopted the convention of stating the main theorem and then immediately holding a discussion about the implications of the theorem, deferring the actual proof to a later subsection. During such discussions, we use examples derived from the uniform noise rewritable channel or sometimes the Gaussian rewritable channel; in fact the reader will see that the insights gathered during the discussions in Section IV and Section V for the uniform noise case are useful in obtaining the full solution in Section VI.

To connect the results of this article to real memory technologies, we address two key issues. The first issue is to document whether write uncertainty is indeed a major consideration in memories. The second issue is whether the memory model that we are considering captures an important aspect of a real memory technology. We address these issues in Section VIII; a preview is that the former is easy to establish and that we have partial success in the latter.

Our conclusions can be found in Section IX.

II. Memory model

We propose a model of a memory that consists of discrete memory parts which we call cells. In response to a stimulus $x \in \mathcal{X}$, a cell takes on a value $y \in \mathcal{Y}$ according to a statistical relation $\mu_{Y|X}(y|x)$. The alphabets $\mathcal{X}$ and $\mathcal{Y}$ will be subsets of the real line in the particular examples examined in detail in this article, but of course can be modified to fit particular storage medium characteristics. We refer the reader to Section VIII for a short introduction on a selection of real memory technologies.

We shall be interested in storing information into a group of n cells. The role of the write controller is to accept a message $U \in \{1, \cdots, 2^{nR}\}$, where $R > 0$ is a rate parameter, to be encoded in the memory and to provide input signals to be applied to the memory so that a future requester can retrieve the intended message with very high probability.

The write controller plays a role similar to that of channel encoder in traditional transmission schemes. Nonetheless, an extra degree of freedom is allowed: the write controller may read after writing and decide to write again an arbitrary subset of the n cells. Physically this increases the time for writing and also increases the amount of energy spent for encoding. For memories that degrade after a large number of write/read cycles, the lifetime of the memory may also be affected. Nonetheless, a write controller may find it advantageous to exercise the rewrite ability in order to increase the storage capacity and/or reduce decoding error rates for the underlying memory system.

The cells are statistically independent from each other (spatial independence) and the outcome of the writing in one cell is statistically independent from the prior history of that cell - we call that temporal independence. Furthermore, all parameters of the statistical cell behavior are assumed to be known. As previewed at the beginning of this section, we will follow the convention that upper case letters denote random variables, and lower case letters denote realizations of random variables. We denote vectors with n entries with **bold** letters. Individual entries in a vector are denoted using the same letter without the bold font and with a subscript denoting the index within the vector. We will use an superscript (as in Y^i) to denote a time index. Suppose that $Q_{A|B}$ is some conditional probability law for random variables taking on values in an alphabet $\mathcal{A}$ where the conditioning random variable is in an alphabet $\mathcal{B}$. The notation $\mathbf{A} \overset{Q_{A|B}}{\longleftarrow} \mathbf{B}$ defines $\mathbf{A}$ to be a random vector with n entries, each entry obtained by passing the corresponding entry in $\mathbf{B}$ through the channel $Q_{A|B}$ in a stochastically independent manner.

A write process is associated with a cost, generally related to the number of iterations required to complete it. In this article, we will place a limit on the average number of iterations as this is the setting that is easiest to analyze, and also has physical significance; for example one expects the average number of iterations to be related to energy consumption during a write process. In practice, we will generally want to additionally place a limit on the maximum number of iterations since in some settings the time for encoding a message into the n cells is dominated by the worst time across all cells; this problem was addressed in [12].

Formally, a write controller is given by

1) A *stimuli generator* function
$$f : \mathcal{U} \to \mathcal{X}^n$$
where $\mathcal{U} = \{1, ..., 2^{nR}\}$ is the set of possible messages to be encoded in the memory, and
2) a stopping criterion for each cell $j \in \{1, \cdots, n\}$ represented by the function $D_j : \mathcal{U} \to 2^{\mathcal{Y}}$, where $2^{\mathcal{Y}}$ represents the power set of $\mathcal{Y}$.

We assume that every write is followed by a read. In what amounts to a significant specialization in our work, we assume in our article that the first action of a write controller is to write to all cells; one alternative is to allow a read before the first write which can affect what and which cells we subsequently write to. The latter is an interesting case that leads to a problem with additional richness; we will delay discussion of this to subsequent publications. We assume that all n cells are written at least once; if there is a prior agreement between encoder and decoder to not use a subset of the cells at all, then we regard that as a problem with a “smaller n”.

We assume that the message $U \in \mathcal{U}$ to be encoded in the memory is uniformly distributed over $\mathcal{U}$. The write process is defined by

$$\begin{array}{ccc} \mathbf{X} & = & f(U) \\ \mathbf{Y}^i & \overset{\mu_{Y|X}}{\longleftarrow} & \mathbf{X} \end{array}$$

for time indices $i = 1, 2, \cdots$, and where $\mathbf{Y}^i$ is statistically independent from $\mathbf{Y}^k$ conditional on $\mathbf{X}$, if $i \neq k$. The stopping criterion $D_j(U)$ $(j \in \{1, \cdots, n\})$ is used to determine when it is that a cell j has an acceptable content, *after* the input stimulus is applied and the resulting contents are observed through a read operation. Define the stopping time for the jth cell as

$$L_j = \min\{i \geq 1 : Y_j^i \in D_j(U)\}.$$

where Y_j^i denotes the jth entry of the vector $\mathbf{Y}^i$.

Physically, once a cell has met its stopping condition, the write controller will cease writing or reading to that cell. In a situation in which all cells which need to be written to in any given iteration are written at the same time, the write controller finishes the entire write operation at time $\max_{1\leq j\leq n} L_j$. The content of the cells after writing is then given by $Y_1^{L_1} \cdots Y_n^{L_n}$. Assuming that each write attempt has cost equal to one, the average cost associated with this write controller is given by

$$\frac{1}{n}\sum_{j=1}^{n} EL_j. \tag{1}$$

In this article we will wish to place an upper constraint on this average cost and we shall use the letter κ to denote the cost value of this constraint. As we discussed earlier, other cost metrics may be employed; we choose the one above on the basis of simplicity and relevance as the average is simultaneously suggestive of energy consumed and wear due to rewriting in memories. The reading mechanism in the memory is given by a decoding function $h : \mathcal{Y}^n \to \{1, \cdots, 2^{nR}\}$.

The probability of error of this write controller is given by

$$P(h(Y_1^{L_1} \cdots Y_n^{L_n}) \neq U).$$

We say that the rate R at a given cost κ is achievable if for every $\epsilon > 0$ there is a write controller with a stimuli generator function f and read mechanism g with rate R for a sufficiently large number of cells n, such that

$$P(h(Y_1^{L_1} \cdots Y_n^{L_n}) \neq U) \leq \epsilon, \qquad \frac{1}{n}\sum_{j=1}^{n} EL_j \leq \kappa.$$

The largest such R for a fixed cost κ is the capacity of this rewritable storage channel and we use the notation $C(\kappa)$ to denote it.

In real memories there are many other effects that can be taken into account. We list some of them below:

- Input stimuli affect an internal physical state, which in turn may only be observable indirectly through a statistical or deterministic mechanism (see Weissman [11] and Bunte, Lapidoth [13] for a treatment of noisy reads),
- The type of stimulus one can apply and the change in state due to an input stimulus can be a function of the current state of the cell.
- There could be parameters of a cell that are unknown to the writing mechanism at the beginning of the writing process but which can be learned during the writing process.
- Writing to a cell can disturb the contents of other cells. Similarly, write noise can be correlated.
- Reading before starting the writing can be taken advantage of.
- The memory is also subject to noise unrelated to writing.

Each memory technology will exhibit some combination of the issues above to various degrees of severity.

III. A first basic analysis

For the type of writing controllers considered in this article, obtaining a generic information-theoretic expression for storage capacity is elementary given the well established notion of channel capacity with input cost constraints. In this theory of channel capacity, given a channel $Q_{B|A}$ there is an input symbol cost function $c(a)$. A codeword $a_1, \cdots, a_n$ that is input into n independent copies of this channel has associated cost $n^{-1}\sum_{j=1}^{n} c(a_j)$. We can then ask the question: what is the maximum amount of information that can be transmitted over the channel $Q_{B|A}$ with unbounded block length and assuming we want the average cost of the codewords used in the code to be less than some value κ? The answer is given by

$$\max_{A:Ec(A)\leq\kappa} I(A; B)$$

where B is the result of passing the random variable A through $Q_{B|A}$ (see for example [18]).

In order to view our problem in the context above, we set the cost function to be

$$c(x, d) = \frac{1}{\mu_{Y|X}(d|x)}.$$

Note that

$$c(X_j, D_j(U)) = \frac{1}{\mu_{Y|X}(D_j(U)|X_j)} = E\left[L_j | X_j, D_j(U)\right]$$

and thus the average cost of the codewords is, as measured by the function $c(\cdot)$, equal to

$$E\left[\frac{1}{n}\sum_{j=1}^{n} c(X_j, D_j(U))\right] = \frac{1}{n}\sum_{j=1}^{n} EL_j$$

which is precisely the average cost of the write controller in (1).

Having identified a suitable candidate for the cost function, in order to complete the association to channel capacity with input cost constraints we focus on the notion of the *iterated channel* (see Figure 1), which is a single cell's view of the process of writing in a rewritable channel. The iterated channel has inputs X and D, and a single output Y^L. When X and D are input, the signal X is applied to a cell as many times as needed until the output of the cell falls inside D. The random variable L denotes the number of iterations until this happens, and Y^L is then the output of the cell. Note that $1/\mu_{Y|X}(d|x)$ is the expected number of iterations conditional on $X = x$ and $D = d$ being the inputs and thus $EL = E\left[1/\mu_{Y|X}(D|X)\right]$.

At this point it should be clear that the iterated channel can be thought of as a classical channel with inputs associated with a cost. In light of the discussion above, we have arrived to the following result:

Theorem 1: The capacity of the rewritable channel with statistically independent memory cells and with memoryless write outcomes governed by a law $\mu_{Y|X}$ is given by

$$C(\kappa) = \sup_{X,D:E\left[1/\mu_{Y|X}(D|X)\right]\leq\kappa} I(X, D; Y^L), \tag{2}$$

where X is a random variable taking values on the alphabet $\mathcal{X}$, D is a random subset of $\mathcal{Y}$ and Y^L is the output of the iterated channel implied by $\mu_{Y|X}(\cdot|\cdot)$. □

From now onwards, we will use the symbols X and D and Y^L to denote *single letter* random entities that will be used in entropy and mutual information expressions such as that one in (2), in contrast to the discussion in Section II where they were denoting, using appropriate subindices, specific quantities of a actual writing mechanism, as Theorem 1 ensures a connection between the analysis of single letter expressions and actual rewriting techniques for rewritable channels.

The remainder of the paper can be seen as a series of attempts to extract insights from Theorem 1 that we hope have value for practitioners or for further investigation of more elaborate information theoretic models. A first such attempt follows, where we investigate what may be regarded as the simplest continuous input/output alphabet rewritable channel.

A. The uniform noise model

Consider the i.i.d. uniform noise model with average cost constraint. We assume that $\mathcal{Y} = [-a/2, 1+a/2]$ and that $\mathcal{X} = [0, 1]$, where a is a positive real number (without any limitations on its magnitude) whose role will be evident shortly. We assume that $\mu_{Y|X}$ is such that during the cell operation $Y_i = x + W_i$ where W_i is a random variable uniformly distributed in the interval $[-a/2, a/2]$. Then we have

Theorem 2: For the i.i.d. uniform noise model with average cost constraint $\kappa \geq 1$,

$$\log\left\lfloor\frac{1+a}{a}\kappa\right\rfloor \leq C(\kappa) \leq \log\left(\frac{1+a}{a}\kappa\right). \tag{3}$$

Proof. For the upper bound, we start from Theorem 1. Let $\Delta = D \cap [X - a/2, X + a/2]$. The significance of Δ is that it is the *effective* stopping criterion, since given an input X, one may only reach values in the interval $[X - a/2, X + a/2]$. Write $I(X, D; Y^L) = h(Y^L) - h(Y^L|X, D)$. First note that $h(Y^L) \leq \log(1 + a)$. Next, using Jensen's inequality, we write

$$\begin{aligned} h(Y^L|X, D) &= E\log|\Delta| \\ &= -E\left[\log\frac{1}{|\Delta|}\right] \\ &\geq -\log E\left[\frac{1}{|\Delta|}\right] \\ &\geq -\log(\kappa/a), \end{aligned} \tag{4}$$

where $|\Delta|$ denotes the total length of the subset of the real line Δ; this immediately implies the upper bound.

The basic idea for the lower bound is to use rewrites for shaping the noise to be uniformly distributed with a smaller range. To this end, let $0 < b < a$. We construct from $\mathcal{Y}$ disjoint open intervals each of length b. Then one can obtain a cell storage capacity of $\log\lfloor\frac{1+a}{b}\rfloor$ bits by selecting as input the center of any of the $\lfloor\frac{1+a}{b}\rfloor$ intervals[1], and then attempting as many writes as necessary in order to fall within the desired interval of length b. The average number of iterations is then a/b; one then

[1] An adjustment to this description is needed at the boundaries.

simply sets b so that $a/b = \kappa$. □

In Section VI we explore the i.i.d. uniform noise model in more detail, giving an exact expression for the capacity/cost function. As we will see there, it turns out that there is a ”critical cost” $\kappa_0 \geq 1$ dependent on the value of a such that for all $\kappa > \kappa_0$, the upper bound in (3) is strictly tight. Furthermore, the same upper bound is not tight for $\kappa < \kappa_0$, assuming that $\kappa_0 > 1$.

IV. A general capacity upper bound

In this section, we give an upper bound on capacity that depends on a single statistic computed from the cell’s noise distribution. We will assume for this section that there exists a density $f_{Y|X}$ such that for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$\mu_{Y|X}((-\infty, y]|x) = \int_{-\infty}^{y} f_{Y|X}(\xi|x) d\xi.$$

Define the function $f_{\text{sup}} : \mathcal{Y} \to [0, \infty)$ as

$$f_{\text{sup}}(y) = \sup_{x \in \mathcal{X}} f_{Y|X}(y|x). \tag{5}$$

and let $D_{KL}(A||B)$ denote the Kullback-Leibler divergence between the distributions of the underlying random variables A and B. Our main result in here is given by

Theorem 3: For a given i.i.d. rewritable channel with write noise conditional density $f_{Y|X}$, assume that

$$\Gamma \triangleq \int_{\mathcal{Y}} f_{\text{sup}}(y) d\lambda(y) < +\infty$$

where $\lambda(\cdot)$ denotes the Lebesgue measure. Let Y_{sup} be a generic random variable distributed according to $\Gamma^{-1} f_{\text{sup}}(y)$. Then

$$C(\kappa) \leq \log(\Gamma\kappa) - \min_{X,D:E[1/\mu_{Y|X}(D|X)] \leq \kappa} D_{KL}\left(Y^L \,||\, Y_{\text{sup}}\right)$$

where X, D, L and Y^L are defined in the discussion preceding Equation (2). □

The proof of this result is deferred to Subsection IV-B. Since divergence is always non-negative, as a corollary of the above we have, for the same conditions as in Theorem 3:

Corollary 1: $C(\kappa) \leq \log(\Gamma\kappa)$ with equality if and only if $Y^L = Y_{\text{sup}}$ almost everywhere. □

This corollary is useful in that it does not require an optimization to give an interesting upper bound, and it is sometimes tight for the uniform noise case.

A. Discussion

As an example, note that for the uniform noise model, $f_{\text{sup}}(y) = 1/a$ and $|\mathcal{Y}| = 1 + a$. Therefore, $\Gamma = (1 + a)/a$ and we immediately obtain

$$C(\kappa) \leq \log\left(\frac{1 + a}{a}\kappa\right)$$

which matches the result in (3). In Section VI, we further analyze the uniform noise case and obtain an exact result for rewritable storage capacity. The key for the upper bound will be to obtain a sharp lower bound for the term $\min D_{KL}\left(Y^L \,||\, Y_{\text{sup}}\right)$, showing that Theorem 3 gives a tight upper bound for the uniform noise case.

One may consider instead a model where the width of the uniform noise depends on the input to the channel. This model, which we call the *data-dependent* uniform noise model, was explored in detail in [2]; as a matter of fact, Theorem 3 can be considered to be a generalization of the upper bound on the rewritable channel found in [2] to account for more general channel models and write controllers. For the data-dependent model, the uniform noise is dependent on the input stimulus x, i.e., $a = a(x)$ for the noise width parameter. The $a_{\min}(\cdot)$ function in [2], which plays a role similar to the function $f_{\text{sup}}(\cdot)$, was used to show the tightness of the upper bound for very special cases, where one can achieve $Y^L = Y_{\text{sup}}$ in Corollary 1.

For another application of Theorem 3 we consider the case of additive write noise:

Corollary 2: For an additive rewritable channel with noise density f_W, and a peak input stimulus constraint $X \in [\min(\mathcal{X}), \max(\mathcal{X})]$, we have

$$C(\kappa) \leq \log\left(1 + (\max(\mathcal{X}) - \min(\mathcal{X})) \sup_{\xi \in \mathcal{Y}} f_W(\xi)\right) + \log(\kappa).$$

□

The proof of this result is a straightforward calculation given Theorem 3.

B. Proof of Theorem 3

Our starting point is Theorem 1, which connects $C(\kappa)$ with the mutual information $I(X, D; Y^L)$, which we then upper bound:

$$\begin{aligned}
&I(X, D; Y^L) \\
\overset{(a)}{=}\quad & E_{Y^L,X,D}\left[\log \frac{f_{Y^L|X,D}(Y^L|X, D)}{f_{Y^L}(Y^L)}\right] \\
=\quad & E_{Y^L,X,D}\left[\log \frac{f_{Y^L|X,D}(Y^L|X, D)}{f_{Y^L}(Y^L)} \frac{f_{\text{sup}}(Y^L)}{\Gamma} \frac{\Gamma}{f_{\text{sup}}(Y^L)}\right] \\
=\quad & E_{Y^L,X,D}\left[\log \frac{f_{Y^L|X,D}(Y^L|X, D)}{f_{\text{sup}}(Y^L)}\Gamma\right] \\
& +E_{Y^L}\left[\log \frac{f_{\text{sup}}(Y^L)/\Gamma}{f_{Y^L}(Y^L)}\right\} \\
\overset{(b)}{=}\quad & E_{Y^L,X,D}\left[\log \frac{f_{Y^L|X,D}(Y^L|X, D)}{f_{\text{sup}}(Y^L)}\Gamma\right] \\
& -D_{KL}\left(Y^L \middle\| Y_{\text{sup}}\right) \\
\overset{(c)}{=}\quad & E_{Y^L,X,D}\left[\log \frac{f_{S|X}(Y^L|X)}{\mu_{Y|X}(D|X) f_{\text{sup}}(Y^L)}\Gamma\right] - D_{KL}\left(Y^L \middle\| Y_{\text{sup}}\right) \\
\overset{(d)}{\leq}\quad & E_{X,D}\left[\log \frac{\Gamma}{\mu_{Y|X}(D|X)}\right] - D_{KL}\left(Y^L \middle\| Y_{\text{sup}}\right) \\
\overset{(e)}{\leq}\quad & \log\left(\Gamma E\left[\frac{1}{\mu_{Y|X}(D|X)}\right]\right) - D_{KL}\left(Y^L \middle\| Y_{\text{sup}}\right) \\
\overset{(f)}{=}\quad & \log\left(\Gamma E L\right) - D_{KL}\left(Y^L \middle\| Y_{\text{sup}}\right) \\
\leq\quad & \log\left(\Gamma \kappa\right) - D_{KL}\left(Y^L \middle\| Y_{\text{sup}}\right)
\end{aligned}$$

where (a) follows from the definition of mutual information and (b) follows from the definition of the Kullback-Leibler (K-L) divergence. The equality (c) follows from the definition of the operation of the rewritable channel and the associated random variables Y^L, X, D from which it can be deduced that

$$f_{Y^L|X,D}(Y^L|X, D) = \begin{cases} \frac{f_{Y|X}(Y^L|X)}{\mu_{Y|X}(D|X)} & \text{if } Y^L \in D \\ 0 & \text{otherwise.} \end{cases}$$

The inequality (d) follows from the definition of the $f_{\text{sup}}(\cdot)$ function in (5) and the inequality (e) follows from Jensen's inequality. Finally, the equality (f) is due to the fact that the expectation of the time until success of a sequence of independent Bernoulli trials is equal to the inverse of the success probability.

Finally, recall that the distribution of the random variable Y^L is completely specified given the distributions of the inputs to the iterated channel X, D, which in turn are known to satisfy the constraint $E1/\mu_{Y|X}(D|X) \leq \kappa$. From this observation, we can readily lower bound the divergence term, obtaining the theorem statement.

V. Bounds based on superposition coding

The technique that we describe in this section for obtaining capacity lower bounds is based on the notion of *superposition coding*, in which a decoder employs *sequential decoding* in order to fully decode the message encoded in the memory. While these techniques are well known in information theory in the context of multiuser communications [19], they have not been applied, to our knowledge, to the problem of storage of information. In this context, we use them to create two virtual memories out of a physical memory. The first virtual memory appears to a decoder as a classical memory in which a single write (no rewrite iterations) was used to encode information. Upon decoding the message conveyed in the first virtual memory, the second virtual memory, which contains additional message bits, is decoded. While the amount of information that one may store in the first virtual memory by definition does not change with the average number of iterations, the same is not necessarily true for the second virtual memory, whose capacity, as we will show, grows at least as the logarithm of the average number of iterations. The total rewritable channel capacity is then the sum of these two capacities.

We previously argued that the storage capacity of the rewritable channel when the average number of allowed write attempts is at most κ is given by the formula

$$C(\kappa) = \sup_{X \in \mathcal{X}, D \subset \mathcal{Y}, E[1/\mu_{Y|X}(D|X)] \leq \kappa} I(X, D; Y^L). \tag{6}$$

In some problems we may be interested in further restricting the class of stimuli used in the write controller so as to incorporate a *stimulus cost constraint*, not to be confused with the constraint on the average number of iterations. Let $\rho : \mathcal{X} \to \mathcal{R}$ be a cost function, then referring back to Section II, we define the expected cost of the stimuli in the write controller to be

$$E\left[\frac{1}{n}\sum_{j=1}^{n}\rho(X_j)\right].$$

We would like to find the capacity/cost function when $E\left[n^{-1}\sum_{j=1}^{n}\rho(X_j)\right] \leq \rho^*$ for some ρ^*. As the reader may expect, this capacity/cost function is then given by

$$C(\kappa) = \sup_{X\in\mathcal{X},E[\rho(X)]\leq\rho^*,D\subset\mathcal{Y},E[1/\mu_{Y|X}(D|X)]\leq\kappa} I(X,D;Y^L). \tag{7}$$

The proof of this result can be easily obtained by following an argument similar to that in the beginning of Section III which led to Theorem 1. We will develop the results of this section in the more general context afforded us through (7).

The role that superposition coding and sequential decoding play in rewritable channels derives from a simple observation based on the chain rule for mutual information. Recall that the iterated channel (see Figure 1) has two inputs: a stimulus signal X that is used as a physical input to the cell, as well as a set D that determines when it is that the iterative write algorithm will finish. For a given marginal distribution on X, D, the corresponding storage rate is given by

$$I(X,D;Y^L) = I(X;Y^L) + I(D;Y^L|X).$$

The rewriting of the mutual information as a sum of two mutual information expressions emphasizes the fact that *sequential decoding* can be used when interpreting the contents of a memory. In particular, if we can build a decoder for recovering the first $I(X;Y^L)$ bits by, in a loose sense, retrieving X, then in principle we can recover an additional $I(D;Y^L|X)$ bits by building a decoder for the channel whose input is D, output is Y^L, under the assumption that both the encoder and decoder know X.

One can take this analogy further and construct a mechanism for storing information at a cost $\kappa_1 > \kappa_0$ by using an existing technique for storing information at cost κ_0 as a scaffold. These observations are the basis of the following result:

Theorem 4: Assume a rewritable channel with write uncertainty governed by a conditional probability law $\mu_{Y|X}$. We further assume that the cost of an input to this channel is measured using the cost function $\rho(\cdot)$, and that the expected cost of the stimuli of any write controller must be less than ρ^*. Further assume that for every $x \in \mathcal{X}$, $\mu_{Y|X}(\cdot|x)$ is absolutely continuous with respect to the Lebesgue measure. For any $1 \leq \kappa_0 < \kappa_1$, the capacity/cost function $C(\cdot)$ of the rewritable channel under a constraint on the average number of iterations satisfies

$$C(\kappa_1) \geq C(\kappa_0) + \log\left(\frac{\kappa_1}{\kappa_0}\right).$$

□

The proof of this result is deferred to Subsection V-C, with a teaser to the proof included in Subsection V-B.

A. Discussion

A particularly useful consequence of Theorem 4 is obtained by setting $\kappa_0 = 1$: for all $\kappa > 1$,

$$C(\kappa) \geq C(1) + \log(\kappa)$$

where $C(1)$ is simply the classical channel capacity of the channel $\mu_{Y|X}$. Thus we now have a lower bound on rewritable channel capacity for any conditional write noise channel for which we know classical capacity.

For another application, consider the i.i.d. uniform noise write channel with noise width a, and let

$$\kappa_0 = \frac{a}{1+a}\left\lceil\frac{1+a}{a}\right\rceil$$

Note that $\kappa_0 \geq 1$. Then in Theorem 2 the lower and upper bounds match when evaluated at κ_0 and therefore

$$C(\kappa_0) = \log\left(\frac{1+a}{a}\kappa_0\right),$$

that is, capacity is known exactly at κ_0. Furthermore, Theorem 4 implies that

$$\log\left(\frac{1+a}{a}\kappa\right) \tag{8}$$

is achievable for all $\kappa > \kappa_0$, and Theorem 2 assures us that (8) is also a capacity upper bound. We conclude that we have determined capacity exactly for all $\kappa \geq \kappa_0$. In Section VI we complete this analysis by examining the setting where $\kappa < \kappa_0$.

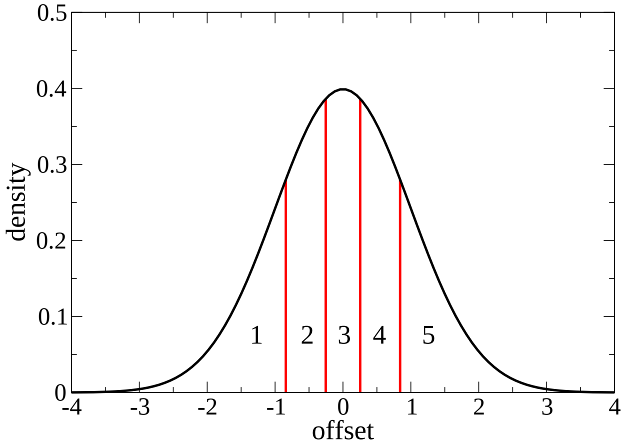

Fig. 2. An example of a partition for the unit variance, zero mean Gaussian density which has $M = 5$ bins, each with probability $1/5$. These bins are identified with a distinct integer.

As another example, suppose that $\mathcal{X} = \mathcal{R}$ and the law $\mu_{Y|X}(\cdot|x)$ describes an additive Gaussian noise channel with variance σ_W^2. If we further assume that the input X to the rewritable channel must satisfy an average stimulus cost constraint of the form

$$E\left[X^2\right] \le \sigma_X^2$$

for some $\sigma_X^2 > 0$, then Theorem 4 implies that

$$C(\kappa) \ge \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_W^2}\right) + \log(\kappa).$$

Let us now assume that $\mathcal{X} = [-\max(\mathcal{X}), \max(\mathcal{X})]$ with $\max(\mathcal{X}) < +\infty$ and that we do not have any average input stimulus constraint. As before, we assume that the medium has additive Gaussian write noise with variance σ_W^2. Then Theorem 4 combined with the results of Raginsky [20] on the channel capacity of Gaussian channels with small peak power constraints imply that as long as $\max(\mathcal{X}) \le 1.05\sigma_W$, then

$$C(\kappa) \ge \frac{1}{\mu^*}\frac{1}{2}\log\left(1 + \frac{\max(\mathcal{X})^2}{\sigma_W^2}\right) + \log(\kappa) \tag{9}$$

where μ^* is a constant satisfying $\mu^* \le 5/4$. On the other hand, note that for this setting,

$$f_{\text{sup}}(y) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_W^2}} & \text{for } y \in [-\max(\mathcal{X}), \max(\mathcal{X})] \\ \frac{\exp(-(y-\max(\mathcal{X}))^2/(2\sigma_W^2))}{\sqrt{2\pi\sigma_W^2}} & \text{for } y > \max(\mathcal{X}) \\ \frac{\exp(-(y+\max(\mathcal{X}))^2/(2\sigma_W^2))}{\sqrt{2\pi\sigma_W^2}} & \text{for } y < -\max(\mathcal{X}). \end{cases}$$

Therefore,

$$\int_{\mathcal{R}} f_{\text{sup}}(y)d\lambda(y) = 1 + \sqrt{\frac{2}{\pi}}\frac{\max(\mathcal{X})}{\sigma_W} < +\infty.$$

Thus we can use Theorem 3 to obtain, for every $\sigma_W > 0$,

$$C(\kappa) \le \log\left(1 + \sqrt{\frac{2}{\pi}}\frac{\max(\mathcal{X})}{\sigma_W}\right) + \log(\kappa)$$

as a counterpart to (9).

B. A preview of the proof: an example using the Gaussian rewritable channel

In order to motivate the arguments used to prove Theorem 4 we discuss informally the superposition coding concept using an example with Gaussian write noise. During this discussion we will temporarily stop using the single letter information theoretic characterization of rewritable channel capacity (Theorem 1) and instead will turn to arguments using actual block codes. Suppose that the memory cell write noise statistics $\mu_{Y|X}$ are such that the value of the cell after a write attempt is equal to the input stimulus plus an additive offset that is a unit variance Gaussian random variable. We shall further assume that we

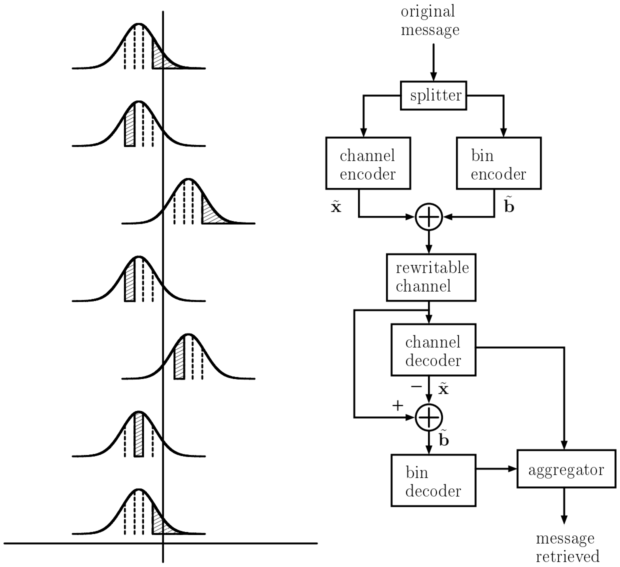

Fig. 3. In the left we give an example of a codeword of a superposition code for the Gaussian rewritable channel with $n = 7$ and $M = 5$. The center of each Gaussian is given by a vector $\tilde{\mathbf{x}} \subset C_{\mathcal{G}}$, while the bin shown as a filled area is selected by a vector $\tilde{\mathbf{b}}$; the codeword is given by $\tilde{\mathbf{x}} + \tilde{\mathbf{b}}$. In the right we show a general diagram of the encoding/decoding procedures for a superposition code.

will require an average input stimulus constraint of the form $E\{X^2\} \leq \rho^*$; this does not necessarily relate to any physically meaningful constraint on any memory and is only used to make it easier for the reader to grasp our point. Let $M > 1$ an integer, and partition the Gaussian density function in M bins each having the same probability. For the sake of simplicity, we will assume that these bins are open intervals, possibly stretching out to $-\infty$ or $+\infty$. Figure 2 shows a specific such partition for the case $M = 5$. Define these open intervals as

$$\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_M$$

where $\mathcal{B}_i \subset \mathcal{R}$.

As before, we use n to denote the number of cells that where we will be encoding our message. Let $\epsilon > 0$ be a parameter, and let $C_{\mathcal{G}} \subset \mathcal{R}^n$ be a good code for the classical Gaussian channel with unit noise power and average input power constraint ρ^*. We assume that $C_{\mathcal{G}}$ satisfies

$$\frac{1}{n} \log |C_{\mathcal{G}}| \geq C(1) - \epsilon. \tag{10}$$

Thus, by good we mean a capacity-achieving code with vanishing probability of decoding error.

Recall that a write controller is specified by associating, for every message $u \in \{1, \cdots, 2^{nR}\}$, an input stimulus for every one of the n cells as well as a *stopping criterion*, which is represented as a subset of the output alphabet. As we shall shortly see, the codewords of $C_{\mathcal{G}}$ will be used as the input stimuli. We now discuss how to build the stopping criteria.

Let $C_{\mathcal{V}} \subset \{1, \cdots, M\}^n$, be a code over the n dimensional space of bin indices. We assume $C_{\mathcal{V}}$ to have cardinality

$$\frac{1}{n} \log |C_{\mathcal{V}}| \geq \log M - \epsilon \tag{11}$$

and to additionally possess the property that for every $\mathbf{v} \in C_{\mathcal{V}}$,

$$\sum_{i=1}^{M} \left| \frac{N(i|\mathbf{v})}{n} - \frac{1}{M} \right| < \epsilon \tag{12}$$

where $N(i|\mathbf{v})$ stands for the number of instances of i within the vector $\mathbf{v}$. We assume n to be large enough so that (10), (11) and (12) can all be accomplished. This can be done because for n large, the vast majority of vectors in $\{1, \cdots, M\}^n$ satisfy (12), a fact that follows from standard typical sequence arguments.

The vector v is regarded as a vector of bin indices. We define the jth entry of the vector $\tilde{\mathbf{b}}$ to be the bin $\mathcal{B}_{v_j}$:

$$\tilde{b}_j = \mathcal{B}_{v_j}.$$

We now refer the reader to Figure 3. Split a message to be written into the memory into two submessages, one with rate $C(1) - \epsilon$ bits/cell, and the other one with rate $\log(M) - \epsilon$ bits/cell. Select an input stimulus $\tilde{\mathbf{x}} \in C_{\mathcal{G}}$ according to the submessage, and select bin indices $\tilde{\mathbf{v}} \in C_{\mathcal{V}}$ according to the message section, associated with the vector of bins $\tilde{\mathbf{b}}$. If S is a subset of the real line and t is a real number, define the translated subset $S + t = \{\xi + t : \xi \in S\}$. Next define the stopping criterion for cell j to be

$$\tilde{d}_j = \tilde{b}_j + \tilde{x}_j.$$

This fully specifies the write controller. Figure 3 illustrates a sample input codeword, which in the case of the rewritable channel is given by an input stimulus and stopping criterion for each cell.

As before, now that we have an input stimulus and a stopping criteria for the n cells, each cell will be programmed independently. Recall that conditional on the input stimulus and the stopping criterion, the value in a cell after programming is distributed according to a ”clipped” version of the write noise distribution:

$$\mu_{Y^L|X,D}(y|x,d) = \mu_{Y|X}(y|x)/\mu_{Y|X}(d|x)$$

if $y \in d$ and $\mu_{Y^L|X,D}(y|x,d) = 0$ otherwise. Because of the property (12), and due to the construction of the partition of the Gaussian with M equal probability, the values of the cells, conditional on the stimuli, but *not* on the stopping criteria, have an approximately multivariate i.id. Gaussian distribution centered around the stimuli. Intuitively, when one does not condition on the stopping criteria, the effect of selecting a bin is to ”restore” the clipped conditional distribution mentioned above to a full non-clipped conditional distribution closely resembling the original Gaussian noise.

Given these observations, it is not very difficult to see that if ϵ is small enough, then a decoder can retrieve, with very high probability the first message section using a decoder for the code $\mathcal{C}_{\mathcal{G}}$. This implies that the decoder has also learned $\tilde{\mathbf{x}}$ and thus it may subtract it from the codeword read from the memory in order to deduce $\tilde{\mathbf{v}}$, from which the second message section can be retrieved using a decoder for $\mathcal{C}_V$.

By using the formula for the capacity of the Gaussian channel, we have then argued that this superposition code allows us to store approximately

$$\frac{1}{2}\log\left(1 + \frac{1}{\sigma_W^2}\right) + \log(M) \tag{13}$$

bits/cell at a cost of M average iterations (since each bin has probability $1/M$ on a given write attempt). Thus $C(M)$ is lower bounded by (13). In what follows, we prove the more general statement in Theorem 4 using a generalization of the above.

C. Proof of Theorem 4

The key structure in the proof rests on the idea that a capacity-achieving input distribution for cost κ_0 will be used to construct another (non necessarily capacity-achieving) input distribution for cost κ_1. Because of this, it will be necessary to introduce two related iterated channels; see the discussion in Section III and Figure 1 for a reminder on the notion of an iterated channel. Both iterated channels will have a common stimulus input random variable X. Let $\{Y^i\}_{i=1}^{+\infty}$ be a random process that is obtained by passing X through statistically independent copies of the channel $\mu_{Y|X}$. The random process $\{Y^i\}_{i=1}^{+\infty}$ will be also shared between the two iterated channels.

The iterated channel for cost κ_0 will use for the encoding set the input random variable $D_0 \subset \mathcal{Y}$, with the associated number of write iterations being

$$L_0 = \min\{i \geq 1 : Y^i \in D_0\}.$$

Similarly, the iterated channel for cost κ_1 will use for the encoding set the input random variable $D_1 \subset \mathcal{Y}$, and as before, the associated number of write iterations is defined as

$$L_1 = \min\{i \geq 1 : Y^i \in D_1\}.$$

We are now ready to develop the proof of the result. Let $\epsilon > 0$, and let X, D_0 be random variables such that for the rewritable channel with conditional probability law $\mu_{Y|X}$, output process $\{Y^i\}_{i=1}^{\infty}$ and random cost L_0, the average cost satisfies

$$E\{L_0\} \leq \kappa_0, \tag{14}$$

the stimulus cost satisfies $E\rho(X) \leq \rho^*$ and furthermore

$$I(X, D_0; Y^{L_0}) \geq C(\kappa_0) - \epsilon.$$

If $\kappa_0 = 1$, then we choose $D_0 = \mathcal{Y}$, and we choose X so as to achieve the classical capacity of the channel $\mu_{Y|X}$ within ϵ bits/cell. For any $d \subset \mathcal{Y}$ and $x \in \mathcal{X}$ with $\mu_{Y|X}(d|x) > 0$, let $\mu_{Y^{L_0}|X,D_0}(\cdot,|x,d)$ denote the conditional probability law of Y^{L_0} given values for X and D_0. Using the definition of the iterated channel, it can be easily shown that for $\psi \subset \mathcal{Y}$,

$$\mu_{Y^{L_0}|X,D_0}(\psi|x,d) = \frac{\mu_{Y|X}(\psi \cap d|x)}{\mu_{Y|X}(d|x)}. \tag{15}$$

Let $\psi \subset \mathcal{Y}$ be a set with Lebesgue measure zero, that is, $\lambda(\psi) = 0$. Because $\mu_{Y|X}(\cdot|x)$ is absolutely continuous, it follows that $\mu_{Y|X}(\psi|x) = 0$ and since $\mu_{Y|X}(\psi \cap d|x) \leq \mu_{Y|X}(\psi|x) = 0$, we thus conclude that whenever it is defined, the measure $\mu_{Y^{L_0}|X,D_0}(\cdot|x, d)$ is also absolutely continuous with respect to the Lebesgue measure.

It is a known fact from probability theory that the cumulative distribution function of a random variable whose probability law is absolutely continuous with respect to the Lebesgue measure is a continuous function in the standard real analysis sense. Applied to the setting at hand, we see that if $\mu_{Y|X}(d|x) > 0$, the cumulative distribution function

$$F_{Y^{L_0}|X,D_0}(\xi|x, d) = \frac{\mu_{Y|X}((-\infty, \xi] \cap d|x)}{\mu_{Y|X}(d|x)} \tag{16}$$

is a continuous function of ξ. Define the *gain* as

$$g \overset{\Delta}{=} \kappa_1/\kappa_0 \tag{17}$$

which necessarily satisfies $g > 1$. For any $\phi \in (0, 1)$, define the set $\gamma(\phi) \subset (0, 1)$ as follows:

$$\gamma(\phi) = \begin{cases} (\phi, \phi + 1/g) & \text{if } \phi + 1/g < 1 \\ (\phi, 1) \cup (0, -1 + \phi + 1/g) & \text{if } \phi + 1/g \geq 1 \end{cases}.$$

Next, for any $d \subset \mathcal{Y}$, $x \in \mathcal{X}$ and $\phi \in (0, 1)$, let

$$\pi_{d,x}(\phi) = \left\{\xi \in d : F_{Y^{L_0}|X,D_0}(\xi|x, d) \in \gamma(\phi)\right\}.$$

We now make use of the following basic result (see for example Billingsley [21], Section 14).

Lemma 1: Let A be a real valued random variable with a continuous cumulative distribution function $F_A(\xi) = \mu_A((-\infty, \xi])$. Then

$$\mu_A(\{a : F_A(a) \leq u\}) = u.$$

□

Using this lemma, we can then see that

$$\frac{\mu_{Y|X}(\pi_{d,x}(\phi)|x)}{\mu_{Y|X}(d|x)} = \frac{1}{g}. \tag{18}$$

Now let Φ be a random variable taking values in the alphabet $[0, 1]$ that additionally is statistically independent from X and D_0. The encoding set for cost κ_1 is now defined as

$$D_1 \quad = \quad \pi_{D_0,X}(\Phi). \tag{19}$$

We now consider X, D_1 to be the input distribution to the rewritable channel $\mu_{Y|X}$. The associated average cost can be evaluated with

$$\begin{aligned} E[L_1] &\overset{(a)}{=} E[E[L_1|X, D_1]] \\ &\overset{(b)}{=} E\left[\frac{1}{\mu_{Y|X}(D_1|X)}\right] \\ &\overset{(c)}{=} E\left[\frac{1}{\mu_{Y|X}(\pi_{D_0,X}(\Phi)|X)}\right] \\ &\overset{(d)}{=} E\left[\frac{g}{\mu_{Y|X}(D_0|X)}\right] \\ &\overset{(e)}{\leq} \kappa_1. \end{aligned} \tag{20}$$

In this development, (a) follows from the basic properties of conditional expectation, (b) follows from the fact that the mean of a geometric distribution with a trial success probability p is $1/p$, (c) follows from the definition of D_1 in (19), and (e) follows from the assumption (14). The step (d) follows from (18), which does not have any fundamental restriction on x, d, ϕ, and hence will also hold in the case the arguments are random variables.

We remark that the result $E[L_1] \leq \kappa_1$ holds regardless of our choice for the marginal distribution of Φ. Nonetheless, in what follows we will consider two explicit choices for the random variable Φ. The first choice works only when $g > 1$ is an integer. In this choice, Φ is a discrete random variable uniformly distributed on the set

$$\left\{0, \frac{1}{g}, \frac{2}{g}, \cdots, \frac{g-1}{g}\right\}.$$

In the second choice, which works for all $g > 1$, Φ will be a random variable uniformly distributed on the interval $(0, 1)$. In both cases, as stated previously, Φ will be statistically independent of X and D_0.

Since the second choice works for all $g > 1$, it suffices to prove the theorem. Nonetheless the first choice is associated with a far simpler decoding scheme and thus we believe there is value in including it in this proof. Note that the example in subsection V-B uses a scheme based on the first choice.

Using the chain rule for mutual information, write

$$\begin{aligned} & I(X, D_0, D_1; Y^{L_1}) \\ = \quad & I(X, D_0; Y^{L_1}) + I(D_1; Y^{L_1}|X, D_0) \\ = \quad & I(X, D_1; Y^{L_1}) + I(D_0; Y^{L_1}|X, D_1). \end{aligned} \tag{21}$$

Due to the construction of the iterated channels, it is not difficult to see that the following is a Markov chain:

$$Y^{L_1} \to (X, D_1) \to D_0. \tag{22}$$

Therefore from (21) we can deduce that

$$I(X, D_1; Y^{L_1}) \quad = \quad I(X, D_0; Y^{L_1}) + I(D_1; Y^{L_1}|X, D_0).$$

The following fundamental lemma characterizes the two quantities on the right. Its proof is included in the Appendix in order to improve the flow of this paper.

Lemma 2: If Φ is chosen as a uniform random variable on $(0, 1)$, we have

$$\begin{aligned} I(D_1; Y^{L_1}|X, D_0) \quad &\geq \quad \log(g) \\ I(X, D_0; Y^{L_1}) \quad &= \quad I(X, D_0; Y^{L_0}). \end{aligned}$$

The same result holds if $g > 1$ is an integer and Φ is chosen to be uniformly distributed on $\{0, 1/g, \cdots, (g-1)/g\}$.

□

In light of this result, we then find that

$$\begin{aligned} I(X, D_1; Y^{L_1}) \quad &\geq \quad I(X, D_0; Y^{L_0}) + \log(g) \qquad (23) \\ &\geq \quad C(\kappa_0) - \epsilon + \log(g). \qquad (24) \end{aligned}$$

Using (20), the fact that $E\rho(X) \leq \rho^*$ and the characterization of rewritable channel capacity in (7), as appropriate, we have $C(\kappa_1) \geq I(X, D_1; Y^{L_1})$. Finally,

$$C(\kappa_1) \geq C(\kappa_0) + \log\left(\frac{\kappa_1}{\kappa_0}\right) - \epsilon$$

where we have used the definition of g in (17). Since this holds for every $\epsilon > 0$, we have proved the theorem.

VI. A sharp result for the i.i.d. uniform noise model

Our goal for this section is to obtain a sharp characterization of the i.i.d. uniform noise model with an average cost constraint. This will involve improving both the upper and lower bounds given in the earlier sections.

The following result fully characterizes $C(\kappa)$:

Theorem 5: Let a be a given noise parameter, and let $N = \lceil \frac{1+a}{a} \rceil$. Then the capacity of the uniform noise rewritable channel with noise width a is given by

$$C(\kappa) = \log\left(\frac{1+a}{a}\kappa\right) - D_{KL}\left(\pi_{\mathcal{S}} \middle\| \frac{\ell_{\mathcal{S}}}{1+a}\right) \tag{25}$$

for $\kappa < \kappa_0$ and by

$$C(\kappa) = \log\left(\frac{1+a}{a}\kappa\right)$$

for $\kappa \geq \kappa_0$, where $D_{KL}(p\|q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$ and

$$\begin{aligned} \ell_{\mathcal{S}} \quad &= \quad (N-1)((N-1)a - 1), \quad \pi_{\mathcal{S}} \ = \ 1 - \frac{1-(N-2)a}{a}\kappa \\ \kappa_0 \quad &= \quad N\frac{a}{1+a}. \end{aligned}$$

□

The proof of this result is deferred to Subsection VI-B.

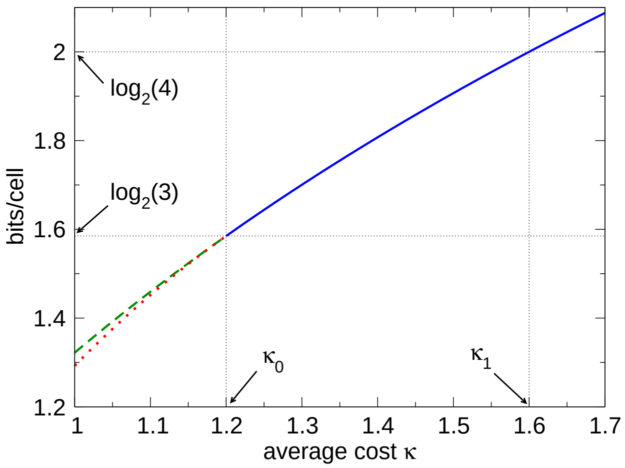

Fig. 4. The capacity for the uniform noise rewritable channel with $a = 2/3$. For $\kappa < \kappa_0$ the capacity is the dotted curve that is below the dashed curve. For $\kappa \geq \kappa_0$, the capacity is given by the solid curve. This solid curve, which is a logarithm with an offset, is extended to the left of κ_0 (as the dashed curve) to illustrate the gap between the upper bound $\log(\kappa(1 + a)/a)$ and the actual capacity.

A. *Discussion*

In Figure 4 we illustrate the capacity for a rewritable channel with $a = 2/3$. For $\kappa < \kappa_0$, capacity is given by the dotted curve and for $\kappa \geq \kappa_0$ it is given by the solid curve, and where we defined, for $i \in \{0, 1, 2\}$, $\kappa_i = \frac{a}{1+a}(N + i)$. The value of κ_0 is the same critical cost as defined in Theorem 5. The gap between the dotted and dashed curves for the regime $\kappa < \kappa_0$ corresponds to the divergence term in (25).

The form of this capacity result is similar to the upper bound of Theorem 3 in that there is a logarithmically growing term with a divergence subtracting from it. Indeed we will see that Theorem 3 is precisely tight in the uniform noise case. This was already made evident for $\kappa > \kappa_0$ in the discussion leading to Equation (8) where we combined Theorem 3 (in the particular weakened form of Corollary 1) and Theorem 4, and will be shown by lower bounding the divergence term in Theorem 3. The key is to have a good notion of what the optimal output distribution's properties are.

For $\kappa < \kappa_0$, the optimal output distribution f_{Y^L} for a capacity-achieving input distribution is piecewise constant and assumes two values whereas for $\kappa > \kappa_0$, the optimal output distribution is uniform on the interval $[-a/2, 1 + a/2]$. Since for $\kappa < \kappa_0$ there are only two possible values for f_{Y^L}, it is possible to segment the $\mathcal{Y}$ in two regions, each associated with one of the values. In our development, the region associated with the higher probability value is labeled $\mathcal{S}$, which stands for "center region" for reasons that will be evident in the proof. It turns out that the divergence term is measuring the discrepancies in the probability of $\mathcal{S}$ under the optimal output distribution of Y^L and the distribution of Y_{sup}.

The input distribution attaining capacity is by no means unique, with two capacity-achieving distributions potentially having little resemblance to each other. This will be made evident by the proofs of the lower bounds for the $\kappa < \kappa_0$ and $\kappa \geq \kappa_0$ cases, which are quite distinct yet both constructions may be used to achieve capacity at κ_0. It is important to note though that for $\kappa = \kappa_0$, either construction leads to a uniform output distribution.

Finally, we observe that the case $\kappa < \kappa_0$ includes the case $\kappa = 1$, that is, the classical channel capacity problem with no rewrites. The particular case $\kappa = 1$ is proposed as an exercise in the textbook of Cover & Thomas [19, Chapter 9 "Uniform distributed noise"]. As we will show, an optimal assignment to X, D in the rewritable channel case can be obtained by having the statistics of X be identical throughout the range $\kappa \in [1, \kappa_0]$. The additional bits that can be stored in the rewritable channel by allowing a cost larger than 1 are thus obtained by an appropriate manipulation of the stopping criterion D.

B. *Proof*

The proof consists of 4 different arguments, addressing upper and lower bounds when $\kappa \leq \kappa_0$ and when $\kappa > \kappa_0$.

1) The upper bound when $\kappa > \kappa_0$: In here, we simply invoke either Theorem 2 or Theorem 3.

2) The lower bound when $\kappa > \kappa_0$: In here, we could simply invoke the argument in the discussion after Theorem 4, but we will give a slightly different (and self contained) argument which is better matched to the lower bound in the case $\kappa < \kappa_0$. After specifying the random variables X, D, we will employ the relation $I(X, D; Y^L) = I(X; Y^L) + I(D; Y^L|X)$ to evaluate the lower bound. Define $\mathcal{Z}_i = [-a/2+i\frac{1+a}{N}, -a/2+(i+1)\frac{1+a}{N}]$. Define J to be a random variable uniformly distributed over $\{0, 1, \cdots, N-1\}$. Let $\{x_0, \cdots, x_{N-1}\}$ be any collection of distinct values such that $x_i \in [0, 1]$ and $\mathcal{Z}_i \subset [x_i - a/2, x_i + a/2]$. Define $X = x_J$. Now refer to Figure 5. Choose the random variable V to be uniformly distributed on $\mathcal{Z}_J$. Suppose that $V + a/\kappa \leq \max \mathcal{Z}_J$, then we choose $D = [V, V + a/\kappa]$. Otherwise, we choose

$$D = [V, \max \mathcal{Z}_J] \bigcup \left[\min \mathcal{Z}_J, a/\kappa - \frac{1+a}{N} + V\right].$$

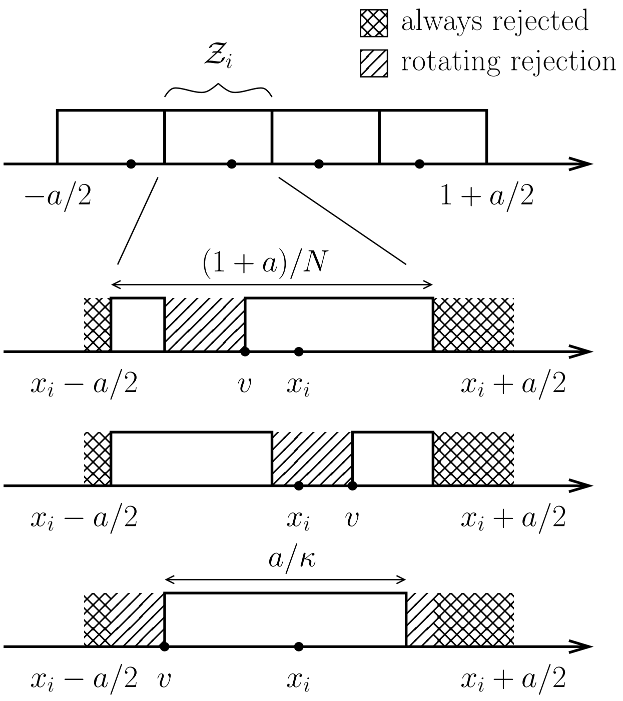

Fig. 5. Illustration of general technique for output distribution shaping used in the lower bound for $\kappa > \kappa_0$.

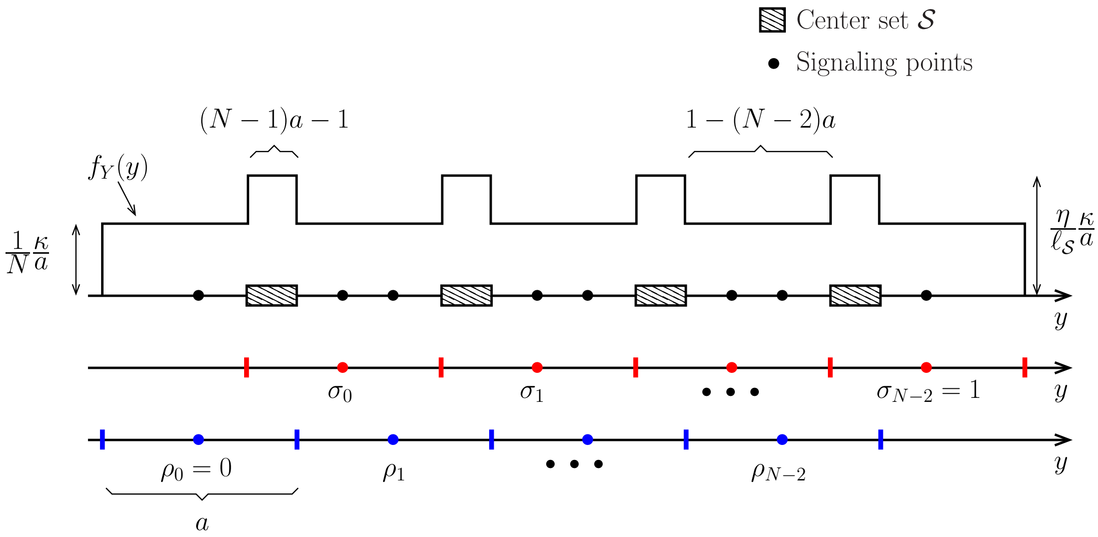

Fig. 6. Optimal output distribution and construction of the center set $\mathcal{S}$ for the case $\kappa < \kappa_0$. There are two groups of intervals, each with $N-1$ adjacent intervals of length a.

Note that $Ea/|D| = \kappa$; note that $D \subset \mathcal{Z}_J$. By our assumptions, $I(X;Y^L) = I(J;Y^L)$ and $I(J;Y^L) = \log N - H(J|Y^L)$. But since $Y^L \in D \subset \mathcal{Z}_J$ we have $H(J|Y^L) = 0$. From the identity $I(D;Y^L|X) = h(Y^L|X) - h(Y^L|X,D)$, it can be seen that given X, Y^L is uniformly distributed over the interval $\mathcal{Z}_J$. Similarly, we have that $Y^L \to D \to X$ and given D, Y^L is uniformly distributed over D, which has length a/κ. Thus $I(D;Y^L|X) = \log((1+a)/N) - \log a/\kappa$. Taken together and after basic algebra, we have the result of the theorem.

3) The upper bound when $\kappa < \kappa_0$: Let $\mathcal{S} \subset [-a/2, 1+a/2]$ be (for now) an arbitrary subset. Define $D_{KL}(p\|q)$ to be the divergence between two Bernoulli random variables with parameters p and q, respectively, and recall that from Theorem 2, we have that

$$C(\kappa) \quad \leq \quad \log\left(\frac{1+a}{a}\kappa\right) - \min_{X,D:EL\leq\kappa} D_{KL}(Y^L\|Y_{\text{sup}})$$

$$\begin{aligned}&\leq \log\left(\frac{1+a}{a}\kappa\right)\\ &\quad - \min_{X,D:EL\leq\kappa} D_{KL}(P(Y^L \in \mathcal{S})||P(Y_{\text{sup}} \in \mathcal{S})) \qquad (26)\end{aligned}$$

where the last inequality follows from the fact that divergence becomes smaller when the space of events is discretized, a fact that can be easily derived using the log-sum inequality. The upper bound we will derive will be obtained by suitably choosing the set $\mathcal{S}$ as illustrated in Figure 6, and then obtaining a good lower bound for $P(Y^L \in \mathcal{S})$. The choice for $\mathcal{S}$ in this case is motivated by a "guess" of what is the optimal output distribution; of course the upper bound will hold for all possible input distributions. Because $1/a$ is not an integer, it is not possible to fit an integer number of intervals of length a in the output alphabet $[-a/2, 1+a/2]$. What will be done instead is to create two groups of intervals, each of which has $N-1$ contiguous intervals of length a. The first group will have its first interval centered on 0, and the second group will have its last interval centered on 1. Mathematically, let

$$\rho_i = ai \qquad (27)$$

$$\sigma_i = ai + 1 - (N-2)a \qquad (28)$$

for $i \in \{0, \cdots, N-2\}$ for the first and second groups, respectively. Next define

$$\mathcal{S} = \bigcup_{i=0}^{N-2}\left[\sigma_i - \frac{a}{2}, \rho_i + \frac{a}{2}\right]. \qquad (29)$$

It is immediate that

$$|\mathcal{S}| = (N-1)((N-1)a-1) \overset{\Delta}{=} \ell_{\mathcal{S}}. \qquad (30)$$

and therefore, after recalling that for the uniform noise model Y_{sup} is also uniformly distributed,

$$P(Y_{\text{sup}} \in \mathcal{S}) = \frac{\ell_{\mathcal{S}}}{1+a} \qquad (31)$$

which is a constant as κ is varied.

We next derive a lower bound on $P(Y^L \in \mathcal{S})$ under the assumption that $E\{a/|D|\} \leq \kappa$ (as per Equation (26)). Write

$$P(Y^L \in \mathcal{S}) = \int P(Y^L \in \mathcal{S}|D=d)d\mu_D(d) \qquad (32)$$

where μ_D is the probability law governing the input distribution, which takes on all possible encoding sets that are reachable fully by the noise width a. When d is selected as an encoding interval, the writing mechanism chooses a stimulus x such that $d \subset [-a/2+x, a/2+x]$ and repeats the stimulus x until the cell contains a value in d.

It can be checked that the construction of $\mathcal{S}$ is such that no matter what the choice of $x \in [0,1]$ is, we have that

$$|\mathcal{S}^c \cap [-a/2+x, a/2+x]| = 1-(N-2)a$$

where $\mathcal{S}^c = [-a/2, 1+a/2] \setminus \mathcal{S}$. Therefore

$$\begin{aligned}|\mathcal{S} \cap d| &= |d| - |\mathcal{S}^c \cap d|\\ &\geq |d| - |\mathcal{S}^c \cap [-a/2+x, a/2+x]|\\ &= |d| - 1 + (N-2)a.\end{aligned}$$

Next, in reference to (32) and using the assumption of uniform rewritable channel noise, we have that

$$\begin{aligned}P(Y^L \in \mathcal{S}) &= \int \frac{|\mathcal{S}\cap d|}{|d|}d\mu(d)\\ &\geq \int \frac{|d| - 1 + (N-2)a}{|d|}d\mu(d)\\ &\overset{(a)}{\geq} 1 - \frac{1-(N-2)a}{a}\kappa\\ &= \pi_{\mathcal{S}} \qquad (33)\end{aligned}$$

where in the step (a) we made use of the assumption that $E\{a/|D|\} \leq \kappa$. Since we are only considering $\kappa < \kappa_0$, we have that

$$P(Y^L \in \mathcal{S}) > 1 - N\frac{1-(N-2)a}{1+a} = \frac{\ell_{\mathcal{S}}}{1+a} = P(Y_{\text{sup}} \in \mathcal{S}). \qquad (34)$$

Since the divergence on binary random variables $D_{KL}(p||q)$ is monotonically increasing for fixed q and for increasing $p > q$, it follows that

$$\min_{X,D:EL\leq\kappa} D_{KL}(P(Y^L\in\mathcal{S})||P(Y_{\text{sup}}\in\mathcal{S}))\geq D_{KL}\left(\pi_{\mathcal{S}}||\frac{\ell_{\mathcal{S}}}{1+a}\right)$$

with equality if and only if $P(Y\in\mathcal{S})=\pi_{\mathcal{S}}$, thereby proving the result.

4) The lower bound when $\kappa<\kappa_0$: We refer the reader to Figures 6 and 7 to support this discussion. The lower bound will be obtained by defining, for a given cost κ, random variables X,D appropriately and then showing that $I(X,D;Y^L)$ meets the upper bound on capacity evaluated at κ. The random variable X will take values on the discrete set $\{\rho_0,\rho_1,\cdots,\rho_{N-2},\sigma_0,\sigma_1,\cdots,\sigma_{N-2}\}$ according to the following prescription:

$$P(X=\rho_i) \quad = \quad \frac{N-1-i}{N(N-1)} \tag{35}$$

$$P(X=\sigma_i) \quad = \quad \frac{i+1}{N(N-1)} \tag{36}$$

for every $i\in\{0,\cdots,N-2\}$. We now specify how the random variable $D\subset[-a/2+X,a/2+X]$ is chosen. Suppose that either $X=\rho_i$ or $X=\sigma_i$ for some i. Choose V to be a random variable uniformly distributed over the interval

$$\left[\sigma_i-\frac{a}{2},\rho_i+\frac{a}{2}\right]$$

which has length $(N-1)a-1$. Let η be such that

$$\frac{a}{1-(N-2)a+\eta}=\kappa \tag{37}$$

and let $D^*=[V,V+\eta]$ if $V+\eta<\rho_i+\frac{a}{2}$ and

$$D^*=[\sigma_i-\frac{a}{2},V+\eta-a+\sigma_i-\rho_i)\cup(V,\rho_i+\frac{a}{2}]$$

otherwise; note that in either case, $|D^*|=\eta$. Finally, if $X=\rho_i$, define

$$D=D^*\cup\left[\rho_i-\frac{a}{2},\sigma_i-\frac{a}{2}\right]$$

and if $X=\sigma_i$, define

$$D=D^*\cup\left[\rho_i+\frac{a}{2},\sigma_i+\frac{a}{2}\right].$$

We pause to note that $|D|=1-(N-2)a+\eta$. Thus because of (37), we have $Ea/|D|\leq\kappa$ as required. Recall the relation between X,D and Y^L: the input X is sent to the cell as many times as required until $Y^L\in D$. Thus conditioned on X and D, Y^L is uniformly distributed on D. Then we have $h(Y^L|X,D)=\log\frac{a}{\kappa}$. We now analyze the marginal output distribution $f_{Y^L}(y)$. We do this by splitting the analysis in sub-cases.

Suppose that there exists some $i\in\{0,\cdots,N-2\}$ such that

$$y\in\left[\sigma_i-\frac{a}{2},\rho_i+\frac{a}{2}\right].$$

Then $f_{Y^L}(y)=P(X=\rho_i)f_{Y^L|X}(y|\rho_i)+P(X=\sigma_i)f_{Y^L|X}(y|\sigma_i)$. Note that for $v\in[\sigma_i-a/2,\rho_i+a/2]$ we have $f_{V|X}(v|\sigma_i)=f_{V|X}(v|\rho_i)=1/((N-1)a-1)$ and that for $y\in D^*$, $f_{Y^L|XV}(y|\rho_i,v)=f_{Y^L|XV}(y|\rho_i,v)=1/(1-(N-2)a)=\kappa/a$ and zero for $y\in(D^*)^c$. Then

$$\begin{aligned}
f_{Y^L|X}(y|\rho_i) &= \int f_{Y^L|XV}(y|\rho_i,v)f_{V|X}(v|\rho_i)dv\\
&= \int_{D^*} f_{Y^L|XV}(y|\rho_i,v)\frac{1}{(N-1)a-1}dv\\
&= \frac{\kappa}{a}\frac{\eta}{(N-1)a-1}=f_{Y^L|X}(y|\sigma_i)
\end{aligned}$$

and therefore

$$f_{Y^L}(y) \quad = \quad \frac{\kappa}{a}\frac{\eta}{(N-1)a-1}(P(X=\rho_i)+P(X=\sigma_i))$$

$$\quad = \quad \frac{1}{N-1}\left(\frac{\kappa}{a}\frac{\eta}{(N-1)a-1}\right) \tag{38}$$

$$\quad = \quad \frac{\eta}{\ell_{\mathcal{S}}}\frac{\kappa}{a} \tag{39}$$

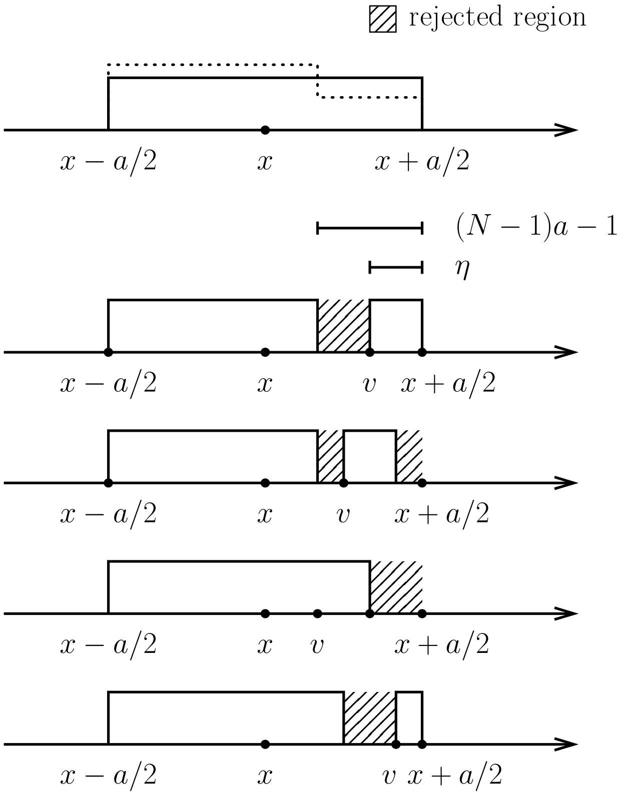

Fig. 7. Illustration of general technique for output distribution shaping used in the lower bound for $\kappa < \kappa_0$. In the proof of this lower bound, the random variable V selects the starting point for a region of length η which is included in the encoding set.

where $\ell_{\mathcal{S}}$ is given by the assumptions of the theorem. The key remark is that the above expression depends neither on i nor on y. Now let $y \in [-a/2, \sigma_0 - a/2)$. Then we have that

$$f_{Y^L}(y) = P(X = \rho_0)\frac{1}{1-(N-2)a+\eta} = \frac{1}{N}\frac{\kappa}{a} \tag{40}$$

where to arrive to (40) we have additionally used (37). Through an identical analysis, we can show that for $y \in (\rho_{N-2} + a/2, 1 + a/2]$, $f_{Y^L}(y)$ has exactly the same value as in (40).

Finally, for $y \in (\rho_{i+1} - a/2, \sigma_i + a/2)$ for $i \in \{0, \cdots, N-3\}$,

$$f_{Y^L}(y) = \frac{P(X = \rho_{i+1}) + P(X = \sigma_i)}{1-(N-2)a+\eta} = \frac{1}{N}\frac{\kappa}{a} \tag{41}$$

where in the last step we have used (37) again, in addition to (35,36). Note that (41) is identical to (40). Compare (40,41) with (38). It can be easily checked that

$$\frac{\eta}{\ell_{\mathcal{S}}} \geq \frac{1}{N}$$

with equality being met when $\kappa = \kappa_0$, which justifies the manner in which we have depicted $f_{Y^L}(y)$ in Figure 6.

We now summarize the results obtained in (38, 40, 41). The function $f_{Y^L}(y)$ takes on at most two values. If $y \in \mathcal{S}$ (refer to (29)), then (38) holds whereas if $y \in [-a/2, 1+a/2] \setminus \mathcal{S}$, $f_{Y^L}(y)$ is given by the value in (40,41). This implies that

$$\begin{aligned} h(Y^L) &= H_b(P(Y^L \in \mathcal{S})) + P(Y^L \in \mathcal{S}) \log|\mathcal{S}| \\ &\quad + P(Y^L \notin \mathcal{S}) \log(1 + a - |\mathcal{S}|), \end{aligned}$$

where $H_b(\cdot)$ denotes the binary entropy function. Finally recall from (30) that $|\mathcal{S}| = \ell_{\mathcal{S}}$. Using this and (38), we get

$$\begin{aligned} P(Y^L \in \mathcal{S}) &= \eta\frac{\kappa}{a} = \left(\frac{a}{\kappa} + (N-2)a - 1\right)\frac{\kappa}{a} \\ &= 1 - \frac{1-(N-2)a}{a}\kappa = \pi_{\mathcal{S}}. \end{aligned}$$

Combined with the previous finding that $h(Y^L|DX) = \log(a/\kappa)$, we obtain the result stated in the theorem.

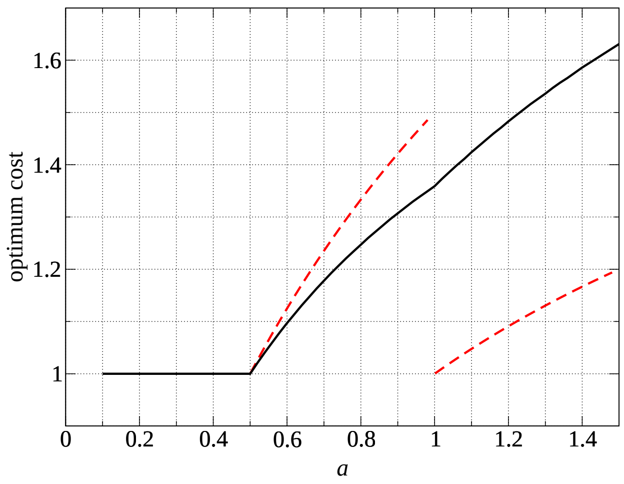

Fig. 8. In bold, plot of the optimum average cost attaining capacity per unit cost (κ_{up}) as function of the channel noise parameter a. The other plot (shown only for $a > 1/2$) is κ_0 as a function of a (see Theorem 5 for a definition of κ_0).

VII. On energy and wear optimized storage in rewritable channels

Suppose one is not interested in operating a memory with a given number of bits/cell, but rather, what one wants is to obtain an optimum ratio of information bits to the expected cost of writing. Our choice of write cost metric throughout this paper, namely, average number of iterations, can be regarded as a rough proxy for the amount of energy and/or wear (when the latter is relevant to the memory technology) that the memory is undergoing during the write process. This observation motivates us to explore the notion of *capacity per unit cost* [22] in the context of rewritable channels.

The capacity per unit cost for the uniform noise model can be obtained from

$$C_{uc}(a) = \sup_{\kappa>1} \frac{C(\kappa, a)}{\kappa} \tag{42}$$

where we have added a second parameter in the expression $C(\kappa, a)$ to explicitly denote the noise width of the associated channel. Let

$$\kappa_{uc}(a)$$

be such that

$$\lim_{\kappa \to \kappa_{uc}(a)} \frac{C(\kappa, a)}{\kappa} = C_{uc}(a)$$

be the limiting cost attaining the supremum. A particularly interesting question is whether $\kappa_{uc}(a)$ is strictly larger than 1 or not. Whenever it is, it follows that one can improve the cost of writing to memory and memory density simultaneously by allowing additional iterations beyond the first write to the memory. The fact that this can happen is not immediately obvious, but it is a phenomenon that does arise in the context of rewritable storage and in particular, the uniform noise model.

In order to illustrate this, we refer the reader to Figures 8 and 9. In Figure 8 we plot $\kappa_{uc}(a)$ against a using a bold line. As we can see, $\kappa_{uc}(a) = 1$ in the range $a \in (0, 0.5]$; for $a > 1/2$, we have $\kappa_{uc}(a) > 1$ and monotonically increasing. Also in Figure 8 we can find a superimposed plot of κ_0 as a function[2] of a for $a > 1/2$, that is, the *critical cost* such that for costs larger than this critical cost, capacity is strictly given by the expression $\log_2(\kappa(1 + a)/a)$. Note that for values $a \in [0.5, 1)$, the optimum cost happens to be in the sub-critical cost region $\kappa < \kappa_0$, while for $a > 1$, then optimum cost happens strictly in the regime $\kappa > \kappa_0$. One particular significance of the $\kappa > \kappa_0$ region, as seen earlier during the detailed discussion of the uniform noise model, is that coding strategies are far simpler than in the sub-critical cost regime.

In Figure 9 we have a plot of $C(\kappa_{uc}(a))$ (not $C_{uc}(a)$) as a function of a. The highlights of this plot are that in the regime $a > 1/2$, capacity per unit cost decreases much more slowly than in the regime $a < 1/2$. As a matter of fact, when $a > 1$ capacity remains strictly flat:

$$C(\kappa_{uc}(a)) = \log_2 e \quad \text{if} \quad a > 1$$

This of course does not mean that $C_{uc}(a)$ is flat in this regime; as a matter of fact the optimum cost keeps increasing according to

$$\kappa_{uc}(a) = \frac{a}{1+a} e \quad \text{if} \quad a > 1.$$

[2] We do not plot critical cost for $a < 1/2$ to maintain legibility of the plot

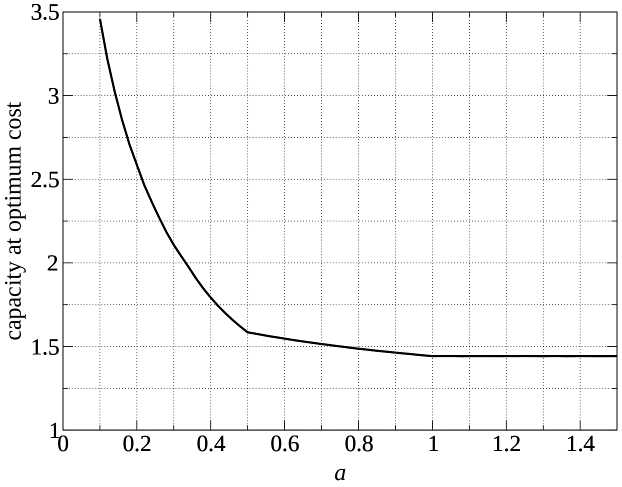

Fig. 9. Plot of capacity evaluated at the optimum cost κ_{up}, as function of the channel noise parameter a.

Still, it is interesting that the capacity per memory cell, when operated at the optimum information bit/cost point flattens out for very noisy channels.

We finalize this discussion by pointing out that this picture can change dramatically whenever we allow a first read before a write. In this case, one can model a memory in which there is a symbol that costs zero to write - the symbol that already exists in the memory at that position. In this case, the interesting average cost regime starts from $\kappa = 0$ instead of $\kappa = 1$. We expect the results in [22] to be of further relevance in the optimum capacity per unit cost analysis in this case.

VIII. Write uncertainty in memory technologies

In the following, we will describe the basic mechanisms leading to write uncertainty in phase-change memory (PCM) [23], spin-transfer torque magnetic RAM (STT-MRAM) [7] and, in general, various kinds of resistive RAM (RRAM) [8], [9], [10]. We also discuss, for PCM and STT-RAM, how a rewritable channel model is relevant for such memories.

A. Phase-Change Memory

PCM [23] is a memory technology that stores information in a cell by controlling the phase (either crystalline or amorphous) of a material. Electrical pulses flowing through the memory cell are used to melt or anneal (crystallize) phase change material. A change in phase corresponds to a change in the perceived electrical resistance of the memory cell. This allows one to sense the contents of the cell by estimating its resistance. With proper control, PCM can be programmed into a continuous span of resistance values, thus making it an analog memory suitable for multiple bits per cell operation.

Write uncertainties in PCM are associated with a number of phenomena. A first reason is due to the steep relationship between the energy of the programming signal and the final resistance [23]. Another important reason for write uncertainty in PCM is the so-called resistance drift. This phenomenon manifests as a noisy increase in resistance [24], [25] which has been empirically described as a logarithmic Wiener process [26]. The fast random evolution of PCM cell resistance immediately after programming can be held in part responsible for the uncertain outcome of a write operation; in particular, it can be characterized as a noisy read.

A programming pulse in PCM can be classified in two rough categories: pulses that cause melting of the cell, and pulses that do not cause melting of the cell. A RESET pulse, which is intended to make a cell attain a high resistance state, or the family of partial-RESET pulses (see [27]) which are able to attain intermediate levels of resistance, belong to the first category. A pulse that partially anneals a PCM cell starting from its current state belongs to the second category.

Programming pulses that cause melting of the cell will have an outcome that is not highly correlated to the prior state of the cell, because the organization of the atoms prior to the pulse is in a sense ”forgotten” during the melting phase. Thus if one is employing a programming technique in which every pulse includes a melting phase, we will be in a position to model such programming with some variant of a rewritable channel with independent write outcomes. It turns out that such programming techniques have been proposed in the past. For example in [28], a multilevel programming technique based on a melting pulse with an ”annealing” tail was used to demonstrate the potential of attaining 4 bit/cell densities in PCM. On the other hand, more recent work (see again [27]) uses a combination of pulses that melt and pulses that do not melt in order to attain its programming goals. To understand programming techniques such as the latter, the theory of rewritable channels needs to be extended appropriately.

B. Magnetic RAM

Magnetic RAMs are memory technologies that store data in the magnetization state of a medium enclosed in a memory cell and exploit the magnetoresistive effect to sense the content of the cell by estimating its resistance which in turn is used to predict the magnetization state. Among magnetic RAM technologies, the most promising one is today the spin-transfer torque MRAM [7] (STT-MRAM), which promises high density of integration and low power operation. STT-MRAM is naturally suited to store a digital content, in that it is designed to store two distinct states of magnetization. Nonetheless STT-MRAM is a valid example of memoryless rewritable memory in that it is affected by write uncertainty [7], [29], and, under the condition that the cell is not in the desired state (i.e., convergence has not yet been attained) the outcome of a programming operation does not depend on the previous programming operations. This suggests that an i.i.d. binary symmetric rewritable channel with statistically independent write outcomes is an appropriate model for some methods for programming MRAM; for one such Write-and-verify technique see [30].

C. Resistive RAM

Even though the broad category of resistive RAM includes all memories which ultimately store information by controlling and estimating the resistance of memory cells, the term resistive RAM (RRAM) is often used to denote a more specific category of memories that store information in an insulating medium based on metal oxides and that control the resistance of the memory cell by changing the chemical composition of the medium, usually in a localized volume within the memory cell. This RRAM category includes several technologies, such as the HP memristor [8], based on a titanium oxide medium, as well as other similar approaches based on different oxides, such as, for example hafnium oxide.

Most RRAM memory technologies rely on the formation of an initial conductive filament within the insulator medium. Programming is then accomplished by controlling the thickness or the continuity of said filament, by applying a current that is large enough to induce atom migration. The direction of the current (in the case of bipolar devices) or the shape of the current signal (in the case of bipolar devices) controls whether the cell resistance increases or decreases.

RRAM, like PCM, is an intrinsically analog memory and fine control of the programming operation can be used to obtain a storage density larger than one bit per cell [31].

As in all other nonvolatile memory technologies, the outcome of a write operation in RRAM is affected by uncertainty. This uncertainty manifests itself as a relatively broad distribution of the stored resistance value, upon the application of a given programming pulse [32]. In general, RRAM, similarly to PCM, exhibits a dependency between the initial cell state and the final cell state after the application of a given programming pulse. However, arguably techniques akin to those used in PCM can be used to break such dependency and obtain a memoryless rewritable channel behavior from RRAM. To the best of our knowledge, an accurate RRAM statistical model, suitable for defining a proper rewritable channel model, has not yet been proposed.

IX. Concluding remarks

In this article, we have taken initial steps towards the formulation of an information theory for rewritable memories that sheds light on the fundamental trade-off between storage capacity and the cost for writing to a memory. Our focus has been, for simplicity reasons, concentrated in an i.i.d. memory model. For this model, we have given capacity upper and lower bounds that show that in some cases capacity is sandwiched between two curves of the form $C + \log \kappa$ for some C. The lower bound is a particularly interesting and unusual application of the concept of superposition coding in storage channels that can be used to link rewritable channel capacity and classical channel capacity. These bounds, together with some additional refinements, were used to give an exact formula of capacity for the uniform noise channel model. An exploration of this result shows that rewritable channels exhibit a dual behavior when one wants to write information in an optimal storage rate per unit cost manner. For "cleaner" channels, one should not do any additional iterations beyond the first write; that is, we should treat this as a regular storage channel. For noisier channels, one should always do some number of iterations to attain this optimum operating point. This example and the capacity results above are examples of the types of insight with practical value that information theory can shed in this important technological area.

Much work remains to be done. From an information theory perspective, we believe that there is still room to improve significantly the models being studied while keeping the information theory tractable. Furthermore, we foresee that some of these extensions will require the introduction of ideas from control and learning theory resulting in a very rich and interesting field with significant practical impact potential.

X. Appendix

A. Proof of Lemma 2

1) Proof of $I(D_1; Y^{L_1} | X, D_0) \geq \log(g)$: By using the definition of D_1 in (19), we have

$$I(D_1; Y^{L_1} | X, D_0) \quad = \quad I(\pi_{D_0,X}(\Phi); Y^{L_1} | X, D_0). \tag{43}$$

We next argue that

$$I(\pi_{D_0,X}(\Phi); Y^{L_1}|X, D_0) \quad = \quad I(\Phi; Y^{L_1}|X, D_0) \tag{44}$$

which can be shown if we demonstrate that the function $\pi_{d,x}(\cdot)$ is invertible for any d, x with $\mu_{Y|X}(d|x) > 0$. Suppose that there exist $\phi_1, \phi_2 \in (0, 1)$ such that $\phi_1 \neq \phi_2$ and

$$\pi_{d,x}(\phi_1) = \pi_{d,x}(\phi_2)$$

Let ν be any open interval satisfying $\nu \subset \gamma(\phi_2) \setminus \gamma(\phi_1)$. Such ν must exist since $\phi_1 \neq \phi_2$. Then the set

$$\left\{\xi \in d : F_{Y^{L_0}|X,D_0}(\xi|x, d) \in \nu\right\} \subset \pi_{d,x}(\phi_2) = \pi_{d,x}(\phi_1) \tag{45}$$

must be empty as otherwise we arrive to a contradiction. Choose two points $u_1, u_2 \in \nu$ with $u_1 < u_2$. Since the function $F_{Y^{L_0}|X,D_0}(\cdot|x, d)$ is continuous, there exists $\xi_1 < \xi_2$ such that

$$F_{Y^{L_0}|X,D_0}(\xi_i|x, d) = u_i \qquad i \in \{1, 2\}$$

The set $(\xi_1, \xi_2) \cap d$ must be nonempty, since

$$\begin{aligned} &\mu_{Y^{L_0}|X,D_0}((\xi_1, \xi_2) \cap d|x, d) \\ &\quad = \quad F_{Y^{L_0}|X,D_0}(\xi_2|x, d) - F_{Y^{L_0}|X,D_0}(\xi_1|x, d) \\ &\quad = \quad u_2 - u_1 > 0 \end{aligned}$$

but $(\xi_1, \xi_2) \cap d$ is a subset of the set in the left of (45) which in turn is empty. Since this is a contradiction, it establishes the invertibility of $\pi_{d,x}(\phi)$.

In summary, if we know D_0, X and $\pi_{D_0,X}(\Phi)$ we can retrieve Φ; clearly also if we know D_0, X and Φ we can construct $\pi_{D_0,X}(\Phi)$. This establishes (44). Combining (43) and (44), we get

$$I(D_1; Y^{L_1}|X, D_0) = I(\Phi; Y^{L_1}|X, D_0). \tag{46}$$

We now need to specialize the result according to the choice of Φ. If Φ is chosen as a discrete random variable, then it is readily seen that

$$\begin{aligned} I(\Phi; Y^{L_1}|X, D_0) \quad &= \quad H(\Phi|X, D_0) - H(\Phi|X, D_0, Y^{L_1}) \\ &= \quad H(\Phi) - H(\Phi|X, D_0, Y^{L_1}) \\ &\overset{(d)}{=} \quad H(\Phi) \\ &= \quad \log(g) \end{aligned} \tag{47}$$

where (d) follows from the fact that knowledge of X, D_0 reveals a partition of D_0 in the form of

$$\left\{\pi_{D_0,X}\left(\frac{j}{g}\right)\right\}_{j=0}^{g-1}$$

An element of this partition was chosen by Φ as the encoding set for the iterated channel for cost κ_1. But further knowledge Y^{L_1} selects a unique element of this partition, thus revealing Φ.

On the other hand, if Φ is chosen as uniformly distributed on $[0, 1]$, then we employ differential entropy instead to obtain

$$\begin{aligned} I(\Phi; Y^{L_1}|X, D_0) \quad &= \quad h(\Phi|X, D_0) - h(\Phi|X, D_0, Y^{L_1}) \\ &\overset{(e)}{=} \quad -h(\Phi|X, D_0, Y^{L_1}) \\ &\overset{(f)}{\geq} \quad -\log\left(\frac{1}{g}\right) = \log(g) \end{aligned} \tag{48}$$

where (e) follows from the assumption that Φ is statistically independent from X and D_0 and from the fact that the differential entropy of a random variable uniformly distributed on an interval of unit length is zero. The step (f) can be deduced as follows: knowledge of X, D_0 and Y^{L_1} reveals that

$$\Phi \in \{\phi : Y^{L_1} \in \pi_{D_0,X}(\phi)\}.$$

On the other hand,

$$|\{\phi : Y^{L_1} \in \pi_{D_0,X}(\phi)\}| = 1/g.$$

Finally, recall that the differential entropy of a random variable with a bounded support is upper bounded by the logarithm of the length of the support. This establishes (f) and hence the the first part of the Lemma.

2) Proof of $I(X, D_0; Y^{L_1}) = I(X, D_0; Y^{L_0})$: We claim that for either of the two choices for Φ (the first choice being valid only when g is an integer), the joint distribution of (X, D_0, Y^{L_0}) is identical to the joint distribution of (X, D_0, Y^{L_1}), hence implying the result of the lemma. An examination of the probability law of Y^{L_0} and Y^{L_1} conditioned on specific values of X, D_0 will suffice for this purpose.

We do not discuss the case when g is an integer and that Φ is chosen to be uniformly distributed over the discrete alphabet $\{0, 1/g, \cdots, (g-1)/g\}$, since it is quite easy to see particularly after reading the following proof for the case in which Φ is uniformly distributed in the real interval $(0, 1)$. Let $\mathcal{I}(\cdot)$ be the indicator function of a boolean event, that is, it is equal to one if the event in the argument is true and equal to zero otherwise. The following result will be convenient in establishing our desired result.

Lemma 3: Let $F_A(\xi)$ be the cumulative distribution function of a random variable A whose probability law μ_A is absolutely continuous with respect to the Lebesgue measure. Further define

$$\pi(\phi) = \{a : F_A(a) \in \gamma(\phi)\}$$

Then for any μ_A-measurable set δ

$$\int_0^1 \mu_A(\delta \cap \pi(\phi)) d\lambda(\phi) = \mu_A(\delta)/g.$$

where the integral above is with respect to the Lebesgue measure.
Proof: Let $F_A(\delta)$ denote the image of the set δ through the function F, that is,

$$F_A(\delta) = \{F_A(\xi) : \xi \in \delta\}$$

Next write

$$\begin{aligned} \int_0^1 \mu_A(\delta \cap \pi(\phi)) d\lambda(\phi) &= \int_0^1 \int_{F(\delta)} \mathcal{I}(\xi \in \gamma(\phi)) d\lambda(\xi) d\lambda(\phi) \\ &= \int_{F(\delta)} \int_0^1 \mathcal{I}(\xi \in \gamma(\phi)) d\lambda(\phi) d\lambda(\xi) \\ &= \int_{F(\delta)} \frac{1}{g} d\lambda(\xi) \\ &= \mu_A(\delta)/g. \end{aligned}$$

□

We now proceed with the main proof. Let $\delta \subset d_0$. Then

$$\begin{aligned} &\mu_{Y^{L_1}|X,D_0}(\delta|x, d_0) \\ &\quad= \int_0^1 \mu_{Y^{L_1}|X,D_0,\Phi}(\delta|x, d_0, \phi) d\lambda(\phi) \\ &\quad= \int_0^1 \frac{\mu_{Y^{L_0}|X,D_0}(\delta \cap \pi_{d_0,x}(\phi)|x, d_0)}{\mu_{Y^{L_0}|X,D_0}(\pi_{d_0,x}(\phi)|x, d_0)} d\lambda(\phi) \\ &\quad= g \int_0^1 \mu_{Y^{L_0}|X,D_0}(\delta \cap \pi_{d_0,x}(\phi)|x, d_0) d\lambda(\phi) \end{aligned}$$

where the last equality follows from (15) and (18). At this moment, we invoke Lemma 3 to deduce that

$$\begin{aligned} &\int_0^1 \mu_{Y^{L_0}|X,D_0}(\delta \cap \pi_{d_0,x}(\phi)|x, d_0) d\lambda(\phi) \\ &\quad= (1/g)\mu_{Y^{L_0}|X,D_0}(\delta|x, d_0) \end{aligned}$$

This proves the lemma. □

References

[1] L.A. Lastras-Montaño, M. M. Franceschini, T. Mittelzholer, M. Sharma. Rewritable Storage Channels. In *International Symposium on Information Theory and Its Applications, 2008.*, pages 1–6, Dec. 2008.

[2] T. Mittelholzer, M. Franceschini, L.A. Lastras-Montaño, I. Elfadel, M. Sharma. Rewritable channels with data-dependent noise. In *2009 International Conference on Communications*, pages 1–6, June 2009.

[3] M. Franceschini, L.A. Lastras-Montaño, T. Mittelholzer, and M. Sharma. The role of feedback in rewritable storage channels [lecture notes]. *IEEE Signal Processing Magazine*, pages 190–194,222, November 2009.

[4] L.A. Lastras-Montaño, T. Mittelholzer, M.M. Franceschini. Superposition coding in rewritable channels. In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–9, Jan. 2010.

[5] L.A. Lastras-Montaño, M. M. Franceschini, T. Mittelholzer. The capacity of the uniform noise rewritable channel with average cost. In *Proceedings of ISIT 2010*, July 2010.

[6] G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, B. Rajendran, S. Raoux, and R. Shenoy. Phase change memory technology. *Journal of Vacuum Science & Technology B*, 28(2), 2010.

[7] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 459 –462, December 2005.

[8] D. B Strukov, G. S Snider, D. R Stewart, and R. S Williams. The missing memristor found. *Nature*, 453(7191):8083, 2008.

[9] I.G. Baek, M.S. Lee, S. Seo, M.J. Lee, D.H. Seo, D.-S. Suh, J.C. Park, S.O. Park, H.S. Kim, I.K. Yoo, U.-In. Chung, and J.T. Moon. Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses. In *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, pages 587 – 590, December 2004.

[10] An Chen, S. Haddad, Yi-Ching Wu, Tzu-Ning Fang, Zhida Lan, S. Avanzino, S. Pangrle, M. Buynoski, M. Rathor, Wei Cai, N. Tripsas, C. Bill, M. VanBuskirk, and M. Taguchi. Non-volatile resistive switching for advanced memory applications. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 746 –749, December 2005.

[11] T. Weissman. Capacity of channels with action dependent states. *IEEE Transactions on Information Theory*, 56(11), 2010.

[12] T. Mittelholzer; L.A. Lastras-Montaño; M. Sharma; M. Franceschini. Rewritable storage channels with limited number of rewrite iterations. In *Proceedings of ISIT 2010*, July 2010.

[13] C. Bunte, A. Lapidoth. On the storage capacity of rewritable memories. In *2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, pages 402–405, Nov. 2010.

[14] C. Bunte, A. Lapidoth. Computing the capacity of rewritable memories. In *2011 IEEE International Symposium on Information Theory Proceedings*, Aug 2011.

[15] R. Venkataramanan, S. Tatikonda, L.A. Lastras-Montaño and M. Franceschini. Coding strategies for rewritable channels with hidden state. In *Submitted to IEEE Int. Symp. on Information Theory (ISIT), 2012.*, 2012.

[16] C.E. Shannon. The zero-error capacity of a noisy channel. *IRE Transactions on Information Theory*, IT-2:8–19, 1956.

[17] Y. Cassuto, M. Schwartz, V. Bohossian, and J. Bruck. Codes for asymmetric limited-magnitude errors with application to multilevel flash memories. *Information Theory, IEEE Transactions on*, 56(4):1582–1595, April 2010.

[18] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.

[19] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[20] M. Raginsky. On the information capacity of gaussian channels under small peak power constraints. In *Proc. Forty-Sixth Annual Allerton Conference*, pages 286–293, September 2008.

[21] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Los Alamitos, CA, 1995.

[22] S. Verdú. On channel capacity per unit cost. *IEEE Transactions on Information Theory*, 36(5), September 1990.

[23] G. W Burr, M. J Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A Lastras, A. Padilla, et al. Phase change memory technology. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, 28:223, 2010.

[24] A. Pirovano, A. L Lacaita, F. Pellizzer, S. A Kostylev, A. Benvenuti, and R. Bez. Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials. *Electron Devices, IEEE Transactions on*, 51(5):714719, 2004.

[25] D. Ielmini, A. L Lacaita, and D. Mantegazza. Recovery and drift dynamics of resistance and threshold voltages in phase-change memories. *Electron Devices, IEEE Transactions on*, 54(2):308315, 2007.

[26] M. Franceschini, L. A. Lastras-Montano, A. Jagmohan, M. Sharma, R. Cheek, and M.-H. Lee. A Communication-Theoretic approach to phase change storage. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1 –6, May 2010.

[27] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam, and E. Eleftheriou. Programming algorithms for multilevel phase-change memory. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pages 329 –332, may 2011.

[28] T. Nirschl, J.B. Phipp, T.D. Happ, G.W. Burr, B. Rajendran, M.-H. Lee, A. Schrott, M. Yang, M. Breitwisch, C.-F. Chen, E. Joseph, M. Lamorey, R. Cheek, S.-H. Chen, S. Zaidi, S. Raoux, Y.C. Chen, Y. Zhu, R. Bergmann, H.-L. Lung, and C. Lam. Write strategies for 2 and 4-bit multi-level phase-change memory. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pages 461 –464, dec. 2007.

[29] J. Li, C. Augustine, S. Salahuddin, and K. Roy. Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement. In *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, page 278283, 2008.

[30] Hongbin Sun, Chuanyin Liu, Nanning Zheng, Tai Min, and Tong Zhang. Design techniques to improve the device write margin for mram-based cache memory. In *Proceedings of the 21st edition of the great lakes symposium on Great lakes symposium on VLSI*, GLSVLSI '11, pages 97–102, New York, NY, USA, 2011. ACM.

[31] M. Terai, Y. Sakotsubo, S. Kotsuji, and H. Hada. Resistance Controllability of Ta_2 O_5 TiO_2 Stack ReRAM for Low-Voltage and Multilevel Operation. *Electron Device Letters, IEEE*, 31(3):204206, 2010.

[32] H.Y. Lee, Y.S. Chen, P.S. Chen, P.Y. Gu, Y.Y. Hsu, S.M. Wang, W.H. Liu, C.H. Tsai, S.S. Sheu, P.C. Chiang, W.P. Lin, C.H. Lin, W.S. Chen, F.T. Chen, C.H. Lien, and M. Tsai. Evidence and solution of over-RESET problem for HfO_X based resistive memory with sub-ns switching speed and high endurance. In *Electron Devices Meeting (IEDM), 2010 IEEE International*, pages 19.7.1 –19.7.4, December 2010.