Multilevel Flash Memories: Channel Modeling, Capacities and Optimal Coding Rates

Xiujie Huang*, Aleksandar Kavcic*, Xiao Ma†, Guiqiang Dong‡, and Tong Zhang§
*Dept. of Elect. Eng., University of Hawaii, Honolulu, HI 96822 USA
†Dept. of Elect. and Commun. Eng., Sun Yat-sen University, Guangzhou, GD 510006 China
‡Skyera Inc., San Jose, CA 95131 USA
§Dept. of Elect., Comp. and Syst. Eng., Rensselaer Polytechnic Institute, Troy, NY 12180-3590 USA
Email: {xiujie,kavcic}@hawaii.edu, maxiao@mail.sysu.edu.cn, dongguiqiang@gmail.com, tong.zhang@ieee.org

***Abstract*—This paper is concerned with channel modeling and capacity evaluation of the multilevel flash memory with m levels. The m-level flash memory is modeled as an m-amplitude-modulation channel with input-dependent additive Gaussian noise whose standard deviation depends on the channel input. The capacity as well as the optimal coding rate of an m-level flash memory channel (m-LFMC) is given. If the channel output is observed after a (finite) quantizer, then the channel is further transformed into a discrete memoryless channel, which yields an approximation of the capacity for the m-LFMC. Actually, the determination of the capacity for the m-LFMC can be transformed into a two-step optimization problem, which can be numerically solved by an alternating iterative algorithm. This algorithm delivers not only the optimal input/level distribution but also the optimal values of levels. This algorithm also delivers the optimized number of levels at any given voltage-to-deviation ratio. Numerical results are presented to show the consistency with well-known Smith's results for the amplitude-limited AWGN channel and the applicability of the modeling method, and to reveal that a finite level quantization of the channel output for the m-LFMC suffers from a negligible loss of information rate compared to the capacity.**

***Keywords*—Amplitude-modulation channel; channel capacity; input-dependent additive Gaussian noise; multilevel flash memory; optimal coding rate.**

I. INTRODUCTION

As the demand for non-volatile data storage increases, flash memories are gaining attention. The original flash memory used only two levels to store one bit in one memory cell. However, a modern mainstream flash memory is a multilevel flash memory (MLFM), which stores more than one bit in one memory cell to improve the storage density and reduce the bit cost of flash memories. In our previous work [1], we investigated the general MLFM with m-levels, where the number m of levels can be any integer not less than two. In practice, designers have presented some MLFMs, where the number m of levels are powers of two, such as the first MLFM product presented by Bauer *et al.* in [2], the 4-level MLFM in [3] and the Intel StrataFlash$^{\text{TM}}$ memory in [4], all three of which had four levels and stored two bits in one cell, the 8-level MLFMs in [5], [6] storing three bits in one cell, and the 16-level MLFMs in [7], [8] storing four bits in one cell.

It is obvious that, as the number of levels increases, the capability of the MLFM could be enhanced. However, due to the complexity of the configuration (including the programming/reading techniques and inter-cell interferences), it is complicated to model precisely the MLFM channel. Hence, research on the information-theoretic channel capacity is sporadic, such as [9], [10]. In particular, in [9], the simple upper and lower bounds in single letter formulas on the capacity were presented and computed numerically when the probability distribution of the channel inputs are assumed to be known. In [10], the MLFM was quantized to different discrete memoryless channels (DMCs) by introducing different reading numbers of reference voltages. By optimizing the reference voltages, the mutual information of DMC could be maximized, and then the achievable rate of the MLFM could be obtained.

Although the capability is enhanced, the reliability of the MLFM could be decreased because the margins between adjacent levels (voltages) in a cell are reduced as the number of levels increases and various interferences (for example, inter-cell interference) could arise. To guarantee the reliability, two approaches are usually investigated and applied in MLFMs. One approach is the on-chip error correcting technique [11]. Up to date, various error correcting codes (ECCs) used in the MLMC have been presented, such as the BCH codes [12], Reed-Solomon codes [13], [14], LDPC codes [10], [15], trellis coded modulation [12], [16], [17] and rank modulation [18], [19]. Other approaches are signal processing methods, for example, the data postcompensation method [9], the data pre-distortion method [9], and the coupling canceller method [20], which could tolerate the inter-cell interference in MLFMs.

To address the information-theoretic issues of the MLFM, we first need to solve a key problem, i.e., channel modeling. The simplest model is a constrained communication system, namely, an amplitude-limited *input-independent* additive white Gaussian noise (AWGN) channel, whose channel capacity and properties were investigated in [21], [22]. In [21], [22], Smith proved that the capacity of the amplitude-limited AWGN channel is achieved by a unique discrete random variable taking values on a finite alphabet. Based on the current techniques and configuration, there exist two universal phenomena for the MLFM. One is that the device degrades with age and the degradation varies from cell to cell as mentioned in [4], [23]. The other is the inter-cell interference as mentioned in [24]. In this paper, building upon our previous work [1], we are interested in only the former, while the latter was discussed in [9], [25]. The contribution of this work is two-fold. First, in Section II, we model the MLFM with m levels, also called m-level flash memory channel (m-LFMC) as an m-amplitude-

modulation (m-AM) channel with input-dependent additive Gaussian noise (ID-AGN) whose standard deviation depends on the channel input (i.e., the voltage value of the level). The m-AM with ID-AGN channel can also be regarded as a constrained communication system [26], [27]. Second, we give the channel capacity and present a numerical method to evaluate it. Consequently, the optimal coding rate is obtained to guide the ECC code design.

Structure: The remainder of this paper is organized as follows. First, the m-LFMC is modeled as an m-AM channel with ID-AGN, as shown in Section II-A. Using channel output quantization, the channel can be considered as a discrete memoryless channel, as shown in Section II-B. Second, the channel capacities of the m-LFMC and the quantized channel are introduced in Sections III-A and III-B, respectively. The quantized capacity is an approximation of the capacity for the m-LFMC. Furthermore, the coding rate is defined in Section III-C. To evaluate the capacity, an alternating iterative algorithm is presented in Section IV, which delivers not only the optimal distribution of the channel input but also the optimal values of the channel input levels. Section V provides numerical results and discussions on the (quantized) capacities and optimal coding rates. We conclude this work in Section VI.

II. Channel Model

For an MLFM with m levels, each level has an intended *threshold voltage* [2]. By applying this voltage to the floating gate of a memory cell (transistor), the charge is maintained and then the data is stored in the cell. Affected by the configuration (including the programming/reading techniques) of the flash memory and device aging, the threshold voltage shift may vary from cell to cell. Hence, each level corresponds to a threshold voltage range [2]. In this paper, we focus on only the variation caused by device aging. For mathematical modeling, the variation of the threshold voltage is usually approximated by a Gaussian distribution and characterized by its probability density function (pdf). The following example illustrates the models of threshold voltage distributions for a 4-LFMC.

Example 1 (Threshold Voltage Distributions of a 4-LFMC): Consider a 4-LFMC. Let the intended voltages of the four levels be $x_0 = 0$, $x_1 = 3.25$, $x_2 = 4.55$ and $x_3 = 6.5$. Note that, throughout this paper, the voltage is measured in the unit of *volt*, which is omitted if no confusion arises. By default, the threshold voltage distribution model of the manufactured 4-LFMC is shown in Figure 1, where the noise at each level has the same variance and the pdf of the output for each level is depicted. As documented in [4], [23], the number of electrons of a cell decreases with time and some cells become defective as time elapses, which means that the cell has a long but finite lifetime and the degradation varies from cell to cell. Consequentially, the performance of the 4-LFMC gets gradually worse as the device ages. Suppose that, after three years, the threshold voltage distribution model of the 4-LFMC is shown in Figure 2, where every level experiences more noise than in Figure 1 and the first level x_0 is the most noisy level while the other three levels have almost the same noise. Again, suppose that, after five years, the threshold voltage distribution model of the 4-LFMC is shown in Figure 3, where every level has even more noise

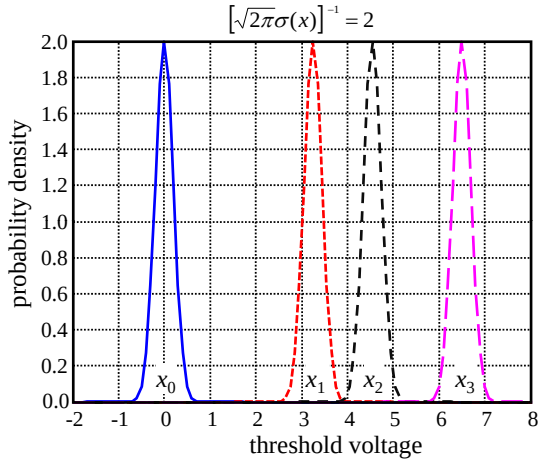

Figure 1. A threshold voltage distribution model for a 4-LFMC, in which the noise at each level has the same variance $\sigma(x) = \frac{1}{2\sqrt{2\pi}}$.

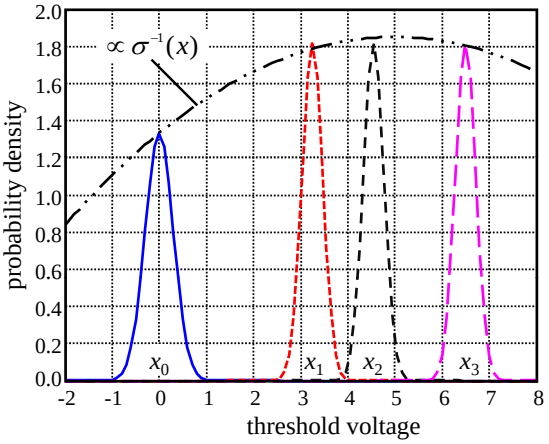

Figure 2. A threshold voltage distribution model for a 4-LFMC, in which the first level x_0 is the most noisy level while the other three levels have roughly the same noise.

than in Figure 2, while the first level x_0 and the last level x_3 are respectively the most noisy levels. This behavior can be easily modeled by a function $\sigma(x)$, which depends on the age of the device. As shown in Figures 2 and 3, the dash-dot-dot curve $\left[\sqrt{2\pi}\sigma(x)\right]^{-1}$ is (approximately) the envelope of the peaks of the level-output-pdfs. In Figure 1, the curve $\sigma(x)$ is assumed to be a constant, i.e., $\sigma(x) = \frac{1}{2\sqrt{2\pi}}$.

Models similar to Figures 2 and 3 for the 4-LFMC were introduced in [4], [12], [15]. In particular, in [4], [12], the model of the 2 bits/cell (i.e., 4-level) NOR flash memory showed that the first level x_0 had the highest noise variance and the last level x_3 had the second highest noise variance while the two middle levels had almost the same noise variances. In [15], the model of a 4-level NAND flash memory showed that, when no inter-cell interference occurred, the first level x_0 had the highest Gaussian noise and the other three levels had almost the same noises characterized by bounded Gaussian variables.

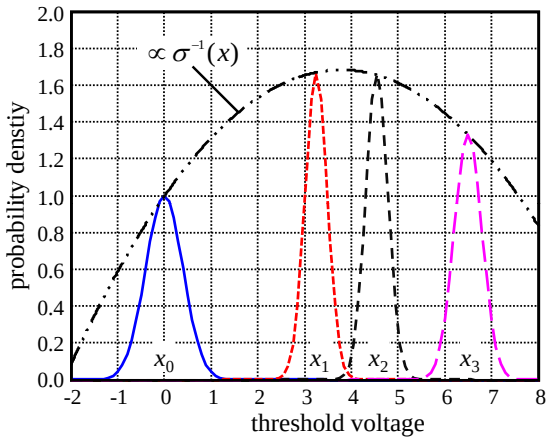

Figure 3. A threshold voltage distribution model for a 4-LFMC, in which the first level x_0 and the last level x_3 are respectively the most noisy levels while the two middle levels x_1 and x_2 have roughly the same noise.

A. The ID-AGN m-AM Channel Model

In this paper, an m-LFMC is modeled as an m-AM channel with ID-AGN. Specifically, it is characterized as follows.

1) Let X, Y and W denote the channel input, the channel output and the channel noise random variables, respectively. They have the relation:
$$Y = X + W. \tag{1}$$
2) The channel input X takes values from a finite alphabet $\mathcal{X}^{(m)} \triangleq \{x_0, x_1, \cdots, x_{m-1}\}$ under the constraint
$$a \le x_0 < x_1 < x_2 < \cdots < x_{m-2} < x_{m-1} \le b, \tag{2}$$
where a and b are the respective lowest and highest possible threshold voltages, and their difference is denoted by $V_m \triangleq b - a$. The finite alphabet $\mathcal{X}^{(m)}$ is called an m-AM signal set. Denote the collection of all such m-AM signal sets as $\mathscr{X}^{(m)}$, i.e., $\mathcal{X}^{(m)} \in \mathscr{X}^{(m)}$. In the following context, we also use the vector notation $\underline{x}$ to denote the m levels, i.e., $\underline{x} = (x_1, x_2, \cdots, x_{m-1})$.
3) The probability mass function (pmf) of X over $\mathcal{X}^{(m)}$ is denoted by $\underline{p} = (p_0, p_1, \cdots, p_{m-1})$ with $p_i = \Pr(X = x_i)$.
4) The noise W is an ID-AGN whose standard deviation depends on the realization of the channel input. That is, the noise W has mean zero and variance depending on the channel input $x \in \mathcal{X}^{(m)}$, i.e., $W \sim \mathcal{N}\big(0, \sigma^2(x)\big)$. In this paper, the function $\sigma(x)$ is assumed to be continuous and differentiable.

Therefore, the channel transition pdf, i.e., the channel law, is
$$f_{Y|X,\sigma(\cdot)}(y|x) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left\{-\frac{(y-x)^2}{2\sigma^2(x)}\right\}. \tag{3}$$
And the pdf of the channel output Y can be obtained as
$$f_{Y,\sigma(\cdot)}(y) = \sum_{i=0}^{m-1} p_i \, f_{Y|X,\sigma(\cdot)}(y|x_i). \tag{4}$$

Recall Example 1 of 4-AM channels with ID-AGN. At the time of manufacturing, the noise standard deviations for all levels are considered to be constant; see Figure 1. As the device ages, the noise standard deviations for different levels increase in different extents; see Figures 2 and 3. That is, the noise standard deviations for an aged device are level-dependent.

B. Quantized Channel Model

In practice, the channel output is often obtained by quantizing the real-valued channel output voltage Y. In this way, a discrete memoryless channel (DMC) of the m-LFMC is obtained when the channel inputs are known and fixed. Let $Q(\cdot)$ be a quantizer of real values and $\hat{Y}$ be the quantized channel output, i.e., $\hat{Y} = Q(Y)$. We assume that, using the quantizer $Q(\cdot)$, the set $\mathbb{R}$ is partitioned into a sequence of n intervals as
$$(r_0, r_1], (r_1, r_2], \cdots, (r_{n-2}, r_{n-1}], (r_{n-1}, r_n), \tag{5}$$
where r_0 and r_n may be finite or infinite. Each interval (r_j, r_{j+1}) is represented by a representation point y_j. Then the finite alphabet of quantized channel outputs is $\mathcal{Y} = \{y_0, y_1, \cdots, y_{n-1}\}$. The quantized DMC is characterized by the channel law (the channel transition probability) as
$$p_{\sigma(\cdot)}(y_j|x_i) = \int_{r_j}^{r_{j+1}} f_{Y|X,\sigma(\cdot)}(y|x_i)\, dy. \tag{6}$$
Consequently, the pmf of the quantized channel output $\hat{Y}$ can be obtained as
$$p_{\sigma(\cdot)}(y_j) = \sum_{i=0}^{m-1} p_i \, p_{\sigma(\cdot)}(y_j|x_i). \tag{7}$$

III. Channel Capacities and Optimum Coding Rates

From the previous section, i.e., Section II-A, we know that the m-LFMC is modeled as an m-AM channel with ID-AGN, parameterized by the m-AM signal set $\mathcal{X}^{(m)}$, the pmf $\underline{p} = (p_0, p_1, \cdots, p_{m-1})$ and the standard deviation function $\sigma(x)$. Therefore, to express the information-theoretic essentials of the m-LFMC, we introduce a new notation different slightly from the conventional one by inserting the subscript $\big(\mathcal{X}^{(m)}, \sigma(\cdot)\big)$ into the mutual information expression, i.e.,
$$I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y) \triangleq \sum_{i=0}^{m-1} \int_{-\infty}^{\infty} p_i \, f_{Y|X,\sigma(\cdot)}(y|x_i) \log\left(\frac{f_{Y|X,\sigma(\cdot)}(y|x_i)}{f_{Y,\sigma(\cdot)}(y)}\right) dy. \tag{8}$$
Similarly, the mutual information of the DMC is given as
$$I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;\hat{Y}) \triangleq \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} p_i \, p_{\sigma(\cdot)}(y_j|x_i) \log\left(\frac{p_{\sigma(\cdot)}(y_j|x_i)}{p_{\sigma(\cdot)}(y_j)}\right). \tag{9}$$

A. The Channel Capacity of the m-LFMC

Definition 1: The capacity of the m-LFMC with standard deviation function $\sigma(\cdot)$ is defined as
$$C_{m,\sigma(\cdot)} \triangleq \max_{\mathscr{X}^{(m)},\{\underline{p}\}} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y), \tag{10}$$

where the maximum is taken over all possible m-AM signal sets $\mathcal{X}^{(m)} = \{x_0, x_1, \cdots, x_{m-1}\} \in \mathscr{X}^{(m)}$ satisfying

$$a \le x_0 < x_1 < \cdots < x_{m-2} < x_{m-1} \le b \tag{11}$$

and all possible pmfs $\underline{p} = (p_0, p_1, \cdots, p_{m-1})$ satisfying

$$p_i \ge 0, \text{ and } \sum_{i=0}^{m-1} p_i = 1. \tag{12}$$

Remark 1. Recall Smith's result that the capacity of the amplitude-limited AWGN channel is achieved by a unique discrete random variable taking values on a finite alphabet [21], [22]. The main two differences between the m-AM channel with ID-AGN and the amplitude-limited AWGN channel are: the noise in the former is input-dependent, while in the latter it is independent of inputs; and the number of inputs is fixed to be m in the former, while in the latter the optimal (capacity-achieving) number of inputs is obtained by optimization.

Remark 2. Comparing with Ungerboeck's results of average energy limited AWGN channel with amplitude modulation [28], there are three main differences. First, the m-AM channel with ID-AGN for an m-LFMC is not average energy limited but amplitude limited (in the interval $[a, b]$). Second, the m-AM signal set is not fixed but can be optimized in the evaluation of its capacity. Third, the input distribution is not uniform but can be optimized too.

One of the main objectives in capacity research is numerical evaluation. To this end, a comprehensive understanding is necessary and can provide a methodology of evaluation. The following proposition gives an insight into the capacity $C_{m,\sigma(\cdot)}$ of the m-LFMC.

Proposition 1: When $\underline{x}$ is given, the mutual information $I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y)$ is concave with respect to (w.r.t.) $\underline{p}$; when $\underline{p}$ is given, the mutual information $I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y)$ is continuous and differentiable w.r.t. $\underline{x}$.

Proof sketch: The mutual information is expressed as

$$\begin{aligned} &I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y) \\ &\quad = h_{\mathcal{X}^{(m)},\sigma(\cdot)}(Y) - \sum_{i=0}^{m-1} p_i \log \sigma(x_i) - \frac{1}{2}\log(2\pi e) \end{aligned} \tag{13}$$

since the noise is input-dependent. When $\underline{x}$ is given, due to the linearity of $\sum p_i \log \sigma(x_i)$, we can prove that the mutual information is concave w.r.t. $\underline{p}$ by using the same method as in [29]. When $\underline{p}$ is given, the composition of elementary functions in (8) is continuous and differentiable w.r.t. $\underline{x}$ because $\sigma(x)$ is assumed to be continuous and differentiable. ■

B. Quantized Capacity

Denote by $\mathcal{Q}$ the collection of all possible quantizers, i.e., $\mathcal{Q} \triangleq \{Q(\cdot)\}$. Then an approximation of the channel capacity (10) is obtained as follows.

Definition 2: The *quantized capacity* of the m-LFMC with standard deviation function $\sigma(\cdot)$ is defined as

$$\hat{C}_{m,\sigma(\cdot)} \triangleq \max_{\mathscr{X}^{(m)},\{\underline{p}\},\mathcal{Q}} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;\hat{Y}), \tag{14}$$

where $I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;\hat{Y})$ is defined in (9), which is the mutual information between the channel input and the quantized channel output $\hat{Y}$ using the quantizer $Q(\cdot)$. The maximum in (14) is taken over all possible m-AM signal sets $\mathcal{X}^{(m)} = \{x_0, x_1, \cdots, x_{m-1}\} \in \mathscr{X}^{(m)}$ satisfying (11), all possible pmfs $\underline{p} = (p_0, p_1, \cdots, p_{m-1})$ satisfying (12) and all possible quantizers $Q(\cdot) \in \mathcal{Q}$.

If the channel input values $x_0, x_1, \ldots, x_{m-1}$ are known and the quantizer $Q(\cdot)$ is determined, then the quantized capacity in (14) is the capacity of a DMC (see the channel law in (6))

$$\hat{C}_{m,\sigma(\cdot)} \triangleq \max_{\{\underline{p}\}} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;\hat{Y}). \tag{15}$$

This capacity can be computed by the well-known Blahut-Arimoto algorithm [30], [31]. Through such capacities, the determination of quantization will be further discussed in Section V-C.

C. Equivalent Binary Code Rate of a Capacity-Achieving Code

As mentioned in the Introduction, ECCs are widely employed in MLFM products. So it is necessary to know the desired coding rate before we design a proper code. The capacity (10) provides an insight into how to pick the code rate of an equivalent binary code that achieves the capacity. Any binary code of a rate greater than the capacity achieving rate can not be used to guarantee the reliability of the MLFM channel; whereas, a binary code of the capacity-achieving rate can be constructed to achieve the capacity of the MLFM channel.

Definition 3: The *code rate* of an equivalent binary capacity-achieving code for the m-LFMC with standard deviation function $\sigma(\cdot)$ is defined as

$$R_m \triangleq \frac{C_{m,\sigma(\cdot)}}{\log_2 m}. \tag{16}$$

When the number of levels m is known (and fixed), then the code rate R_m serves as the upper limit of possible binary code rates that can guarantee reliable reception. If the number of levels m is undetermined, we have a chance to vary m, and thereby find a more appropriate code rate R_m that serves our design purposes. For instance, if $R_3 > R_4$ and $C_{3,\sigma(\cdot)} \approx C_{4,\sigma(\cdot)}$, we may want to use a binary code of the higher coding rate R_3, and thereby save on hardware complexity by using only $m = 3$ channel input levels as well as save on code complexity because binary codes of higher rates require fewer redundant bits. More discussion on this topic is provided in Section V-B.

IV. Evaluation of a Lower Bound on Capacity

To evaluate the capacity (10) of the m-LFMC, we turn to a two-step optimization problem

$$C_{m,\sigma(\cdot)} = \sup_{\underline{x}\in[a,b]^m} \sup_{\underline{p}\in[0,1]^m} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y)$$
$$\text{subject to } \begin{cases} a \le x_0 < x_1 < \cdots < x_{m-1} \le b \\ p_i \ge 0, \ i \in \{0, 1, \cdots, m-1\} \\ \sum_{i=0}^{m-1} p_i = 1 \end{cases}. \tag{17}$$

To solve the two-step optimization problem (17), we turn to two sub-problems.

Sub-problem I.

$$C(\underline{x}) = \max_{\underline{p}\in[0,1]^m} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y) \quad \text{subject to} \begin{cases} p_i \geq 0,\ i \in \{0,1,\cdots,m-1\} \\ \sum_{i=0}^{m-1} p_i = 1 \end{cases}. \tag{18}$$

When $\underline{x}$ is given, Sub-problem I is a conventional capacity problem for memoryless channel with finite inputs. Due to the concavity of the mutual information w.r.t. $\underline{p}$ shown in Proposition 1, the well-known algorithm, Blahut-Arimoto algorithm (BAA) [30]–[32] can be used to solve Sub-problem I.

Sub-problem II.

$$C(\underline{p}) = \max_{\underline{x}\in[a,b]^m} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y) \quad \text{subject to } a \leq x_0 < x_1 < \cdots < x_{m-1} \leq b. \tag{19}$$

The Karush-Kuhn-Tucker (KKT) conditions [33] of the sub-problem are that there exists $\mathbf{v}^* = (\underline{x}^*, \lambda^*, \mu^*)$ such that

$$\begin{cases} \left.\frac{\partial I}{\partial x_0}\right|_{\mathbf{v}^*} &= -\lambda^*, \\ \left.\frac{\partial I}{\partial x_{m-1}}\right|_{\mathbf{v}^*} &= \mu^*, \\ \left.\frac{\partial I}{\partial x_i}\right|_{\mathbf{v}^*} &= 0, \qquad i \in \{1,2,\cdots,m-2\} \\ x_0^* &\geq a, \\ x_{m-1}^* &\leq b, \\ x_{i-1}^* &< x_i^*, \qquad i \in \{1,2,\cdots,m-1\} \\ \lambda^* &\geq 0, \\ \mu^* &\geq 0, \\ \lambda^*(x_0^* - a) &= 0, \\ \mu^*(x_{m-1}^* - b) &= 0. \end{cases} \tag{20}$$

Note that the solution of (20) may be sub-optimal (a local solution) since the concavity of the mutual information w.r.t. $\underline{x}$ is unknown. However, a better $\underline{x}$ with a greater mutual information can be obtained by solving (20). The method to find such a better $\underline{x}$ is shown as below.

For convenience, we denote the pdfs $f_{Y|X,\sigma(\cdot)}(y|x_i)$ in (3) and $f_{Y,\sigma(\cdot)}(y)$ in (4) and the mutual information $I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y)$ in (8) as $f(y|x_i)$, $f(y)$ and $I(X;Y)$, respectively.

We compute partial derivatives of the mutual information $I(X;Y)$. To this end, we first compute the partial derivatives of the transition pdf $f(y|x_i)$ w.r.t x_i for all $i \in \{0,1,\cdots,m-1\}$ as

$$\frac{\partial f(y|x_i)}{\partial x_i} = \begin{cases} f(y|x_i)\left[-\frac{\sigma'(x_i)}{\sigma(x_i)} + \frac{y-x_i}{\sigma^2(x_i)} + \frac{(y-x_i)^2\sigma'(x_i)}{\sigma^3(x_i)}\right], & \text{if } i=j \\ 0, & \text{if } i\neq j \end{cases}, \tag{21}$$

where $\sigma'(x_i) \triangleq \frac{d\sigma(x)}{d\,x_i}$ denotes the derivative of $\sigma(x_i)$ w.r.t. x_i. Then, using (8) and (13), the partial derivatives of the mutual information w.r.t. x_i for all $i \in \{0,1,\cdots,m-1\}$ are obtained

$$\begin{aligned} \frac{\partial}{\partial x_i} I(X;Y) &= -\int_{-\infty}^{\infty} \frac{\partial}{\partial x_i}\left(f(y)\ln f(y)\right) dy - \frac{p_i\sigma'(x_i)}{\sigma(x_i)} \\ &= \left[-\frac{p_i\sigma'(x_i)}{\sigma^3(x_i)}\int_{-\infty}^{\infty} f(y|x_i)\ln f(y)dy\right]\cdot x_i^2 \\ &\quad + \left[\frac{2p_i\sigma'(x_i)}{\sigma^3(x_i)}\int_{-\infty}^{\infty} yf(y|x_i)\ln f(y)dy \right. \\ &\quad \left. + \frac{p_i}{\sigma^2(x_i)}\int_{-\infty}^{\infty} f(y|x_i)\ln f(y)dy\right]\cdot x_i \\ &\quad + \left[-\frac{p_i\sigma'(x_i)}{\sigma^3(x_i)}\int_{-\infty}^{\infty} y^2 f(y|x_i)\ln f(y)dy \right. \\ &\quad - \frac{p_i}{\sigma^2(x_i)}\int_{-\infty}^{\infty} yf(y|x_i)\ln f(y)dy \\ &\quad \left. + \frac{p_i\sigma'(x_i)}{\sigma(x_i)}\int_{-\infty}^{\infty} f(y|x_i)\ln f(y)dy - \frac{p_i\sigma'(x_i)}{\sigma(x_i)}\right] \\ &\triangleq A_i x_i^2 + B_i x_i + C_i. \end{aligned} \tag{22}$$

Solving the KKT conditions (20) is equivalent to finding quantities $(\underline{x}, \lambda, \mu)$ that satisfy the equalities

$$\begin{cases} A_0x_0^2 + B_0x_0 + (C_0 + \lambda) &= 0 \\ \lambda(x_0 - a) &= 0 \end{cases}, \tag{23a}$$

$$\begin{cases} A_{m-1}x_{m-1}^2 + B_{m-1}x_{m-1} + (C_{m-1} - \mu) &= 0 \\ \mu(x_{m-1} - b) &= 0 \end{cases}, \tag{23b}$$

$$A_ix_i^2 + B_ix_i + C_i = 0,\ i \in \{1,2,\cdots,m-2\}, \tag{23c}$$

and the inequalities

$$\begin{cases} \lambda \geq 0 \\ \mu \geq 0 \\ a \leq x_0 < x_1 < \cdots < x_{m-2} < x_{m-1} \leq b \end{cases}. \tag{24}$$

Note that all quantities A_i, B_i and C_i depend on the input vector $\underline{x}$ and the standard deviation function $\sigma(\cdot)$ when the pmf $\underline{p}$ is given. To find the solution to the KKT conditions (23) by an iterative method, we assume that quantities A_i, B_i and C_i are independent of x_i. Then Eqns. (23) have at most $9 \times 2^{m-2}$ solutions. Moreover, under the full constraints in (24), the number of solutions may be much less than $9\times2^{m-2}$ (This happens in our numerical computations). Based on (23) and (24), we employ an iterative method to find a solution. Suppose that the input vector $\underline{x}^{(k)}$ is known at the beginning of the k-th iteration. Then solve Eqns. (23). Pick those solutions that satisfy all constraints in (24), and from them choose the one with the highest information rate as the improved input vector $\underline{x}^{(k+1)}$.

Based on the two sub-problems, an alternating iterative scheme is presented to solve problem (17). At each iteration, the two-stage alternating strategy shown below is employed.

Stage 1. Fix $\underline{x}$. Use the BAA to obtain the optimal $\underline{p}^*$

$$\underline{p}^* = \arg\max_{\underline{p}} I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y). \tag{25}$$

Stage 2. Fix $\underline{p}$. Solve (20) to obtain a better $\underline{x}^*$ such that

$$I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y)\big|_{\underline{x}^*} \geq I_{\mathcal{X}^{(m)},\sigma(\cdot)}(X;Y)\big|_{\underline{x}}. \tag{26}$$

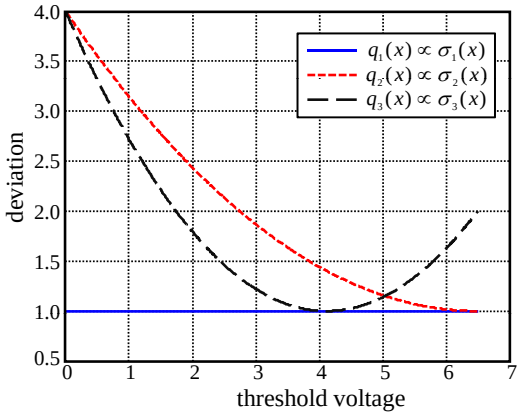

Figure 4. The standard deviation functions $\sigma_i(x)$ of the input-dependent Gaussian noise W: $\sigma_i(x) \propto q_i(x)$, where $i \in \{1, 2, 3\}$.

From the discussion of Sub-problem II, $\underline{x}^*$ may be a local solution. This sub-optimality also implies that a lower bound on the capacity $C_{m,\sigma(\cdot)}$ of the m-LFMC is evaluated.

V. Numerical Results and Discussions

In this section, we first numerically evaluate the capacities of different m-LFMCs using the alternating iterative scheme given in Section IV. We also interpret the results and put them in context with respect to prior works [21], [22]. Second, we estimate the optimal coding rate for the MLFM, which reveals a relationship between the capacity and the optimal number of levels. Third, the quantized capacities of the obtained DMCs using finite-level quantizations of the channel output are also numerically computed.

A. Lower Bounds on the Capacity

Let the lowest and highest threshold voltages be $a = 0$ and $b = 6.5$, respectively. Then the difference is $V_m = b - a = 6.5$. We introduce a new parameter $\sigma > 0$ that severs as the varying noise parameter in our computations. Let $q_i(x)$ where $i \in \{1, 2, 3\}$ be continuous and differentiable functions as shown in Figure 4. We consider three different standard deviation functions $\sigma(x)$, denoted as

$$\sigma_i(x) = q_i(x) \cdot \sigma, \text{ where } i \in \{1, 2, 3\}. \tag{27}$$

We allow the parameter σ to vary such that the *voltage-to-deviation ratio* (VDR) V_m/σ acts as an effective signal-to-noise ratio. We assume that the intended threshold voltage level x_0 (usually corresponding to the erased state) is 0.

We present results for $m \leq 5$, i.e., we consider multilevel flash memory channels with at most 5 levels. We consider three different m-LFMCs whose standard deviation functions are $\sigma_1(x)$, $\sigma_2(x)$ and $\sigma_3(x)$. The lower bounds on capacities of m-LFMCs with deviation functions $\sigma_1(x)$, $\sigma_2(x)$ and $\sigma_3(x)$ are shown in Figures 5, 6 and 7, respectively.

From Figure 5, we make the following observations.

1) When the VDR is less than 10.5 dB, i.e., $20 \log_{10}(V_m/\sigma) \leq 10.5$ dB, 2-LFMC, 3-LFMC, 4-LFMC and 5-LFMC have the same rates.
2) When the VDR is less than 15 dB, 3-LFMC, 4-LFMC and 5-LFMC have the same rates.
3) When the VDR is less than 18 dB, 4-LFMC and 5-LFMC have the same rates.

Furthermore, we observe (not explicitly shown in the figure) that in the VDR regime between 10.5 dB and 15 dB, the optimized lower bound is achieved with $m^* = 3$ levels, even if, say, the constraint allows up to $m = 5$ levels. This implies that, for a fixed VDR, there is an optimal (minimal) number of levels m^* for a given MLFM channel. Increasing the number of levels m beyond m^* does not further increase the capacity (nor the computed lower bound).

Suppose that the number of levels is unknown. Then the MLFM channel is an amplitude-limited channel with ID-AGN, whose capacity is defined as

$$C_{\sigma(\cdot)} \triangleq \max_m C_{m,\sigma(\cdot)}, \tag{28}$$

where $\sigma(x)$ is the standard deviation function of the ID-AGN. The previous observations from Figure 5 imply that 2-LFMC, 3-LFMC and 4-LFMC can achieve the capacity $C_{\sigma_1(\cdot)}$ as defined in (28) in the cases of VDR $\leq$ 10.5 dB, 10.5 dB $<$ VDR $\leq$ 15 dB and 15 dB $<$ VDR $\leq$ 18 dB, respectively. In other words, at a given VDR less than 10.5 dB, a 2-LFMC is "optimal"; at a given VDR less than 15 dB, a 3-LFMC is "optimal"; at a given VDR less than 18 dB, a 4-LFMC is "optimal". Naturally, as the VDR increases, the optimal number of levels does not decrease. This is consistent with prior work [21], [22], which showed that for the amplitude-limited AWGN channel, the capacity is achieved by a discrete channel input distribution over a *finite* alphabet.

Similar conclusions hold for the other two channels with noise standard deviation functions $\sigma_2(x)$ and $\sigma_3(x)$. Namely, even if the constraint is set to be, say, $m = 5$, at low VDRs the optimal number of threshold levels m^* is less than 5. For example, as shown in Figure 7, the optimal number of levels is $m^* = 4$ in the VDR regime between 12 dB and 14.5 dB even when a 5-LFMC with noise standard deviation function $\sigma_3(x)$ is considered. In the case that VDR is equal to 14 dB, using the lower bound optimizing algorithm presented in Section IV, we obtain that the optimal number of levels is $m^* = 4$ with assignment $x_0^* = 0$, $x_1^* \approx 2.718$, $x_2^* \approx 4.212$ and $x_3^* = 6.5$ and pdf $p_0^* \approx 0.274$, $p_1^* \approx 0.171$, $p_2^* \approx 0.271$ and $p_3^* \approx 0.284$, shown in Figure 8. Again, this is consistent with the literature [21], [22] for the amplitude-limited AWGN channel, even though in m-LFMC the noise standard deviation $\sigma(x)$ is input-dependent.

B. Optimal Binary Code Rate R^*

From the previous subsection, we know that for a given VDR, there is an optimal number of levels m^* that achieves that capacity $C_{\sigma(\cdot)}$ in (28). In other words, m^* is the minimal number of channel input levels required to achieve the capacity. Hence, the corresponding *optimal binary code rate* R^* is given by

$$R^* \triangleq \frac{C_{\sigma(\cdot)}}{\log_2 m^*}. \tag{29}$$

R^* is the rate of the equivalent binary capacity-achieving code that achieves the capacity using the smallest possible number

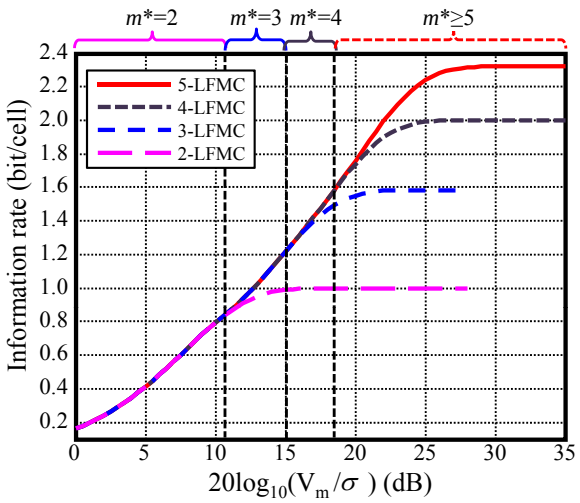

Figure 5. The information rates of m-LFMCs with $m \in \{2, 3, 4, 5\}$ when the standard deviation function is $\sigma_1(x)$. The numbers m^* on the top of the figure indicate that 2-LFMC, 3-LFMC and 4-LFMC can achieve the (computed) maximum rates in the cases of VDR $\leq$ 10.5 dB, 10.5 dB $<$ VDR $\leq$ 15 dB and 15 dB $<$ VDR $\leq$ 18 dB, respectively.

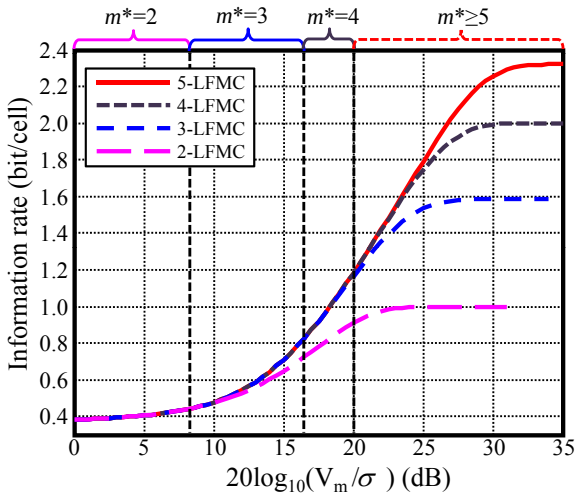

Figure 6. The information rates of m-LFMCs with $m \in \{2, 3, 4, 5\}$ when the standard deviation function is $\sigma_2(x)$. The numbers m^* on the top of the figure indicate that 2-LFMC, 3-LFMC and 4-LFMC can achieve the (computed) maximum rates in the cases of VDR $\leq$ 8 dB, 8 dB $<$ VDR $\leq$ 16.5 dB and 16.5 dB $<$ VDR $\leq$ 20 dB, respectively.

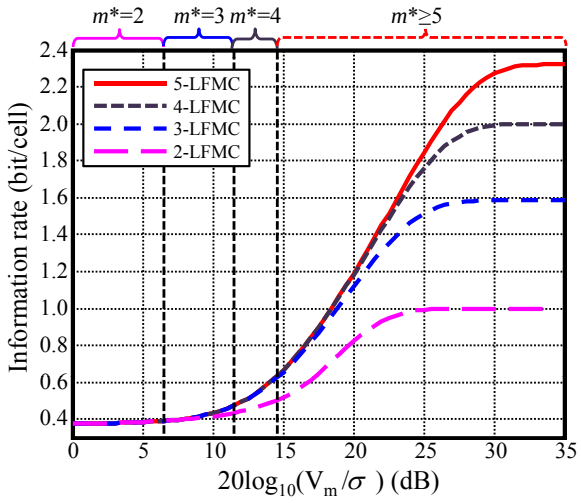

Figure 7. The achievable rates of m-LFMCs with $m \in \{2, 3, 4, 5\}$ when the standard deviation function is $\sigma_3(x)$. The numbers m^* on the top of the figure indicate that 2-LFMC, 3-LFMC and 4-LFMC can achieve the (computed) maximum rates in the cases of VDR $\leq$ 6.5 dB, 6.5 dB $<$ VDR $\leq$ 11.5 dB and 11.5 dB $<$ VDR $\leq$ 14.5 dB, respectively.

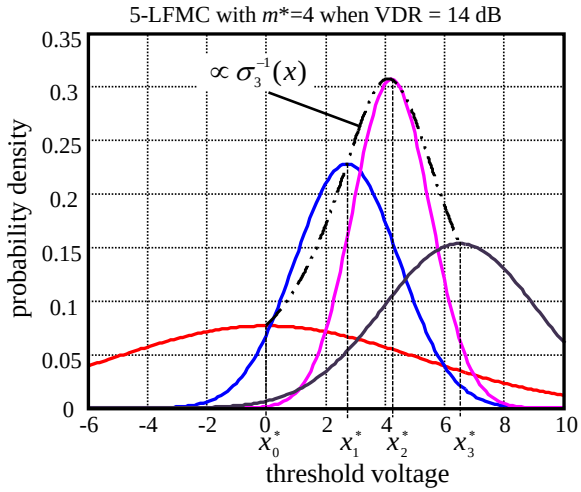

Figure 8. The pdfs of channel output distributions around the optimal threshold voltage levels when the m-LFMC with the standard deviation function $\sigma_3(x)$ and $m = 5$ is used at VDR $= 14$ dB. The optimal number of levels m^* is 4 with assignment $x_0^* = 0$, $x_1^* \approx 2.718$, $x_2^* \approx 4.212$ and $x_3^* = 6.5$ and pdf $p_0^* \approx 0.274$, $p_1^* \approx 0.171$, $p_2^* \approx 0.271$ and $p_3^* \approx 0.284$.

of levels m^*. Consequently, since m^* is the smallest number of levels that still guarantees the achievability of capacity, it follows that R^* is the highest possible rate of an equivalent binary code that can achieve the capacity $C_{\sigma(\cdot)}$. Hence, we refer to R^* as the *optimal rate*.

Figure 9 shows the optimal coding rate R^* for different optimized number of levels m^*, where $m^* \leq 8$, when the standard deviation function of the channel noise is $\sigma_1(x)$. Figure 10 shows the optimal coding rate R^* for different optimized number of levels m^*, where $m^* \leq 15$, when the standard deviation function of the channel noise is $\sigma_2(x)$. Figure 11 shows the optimal coding rate R^* for different optimized number of levels m^*, where $m^* \leq 15$, when the standard deviation function of the channel noise is $\sigma_3(x)$.

C. Quantized Capacities

In this subsection, we present the capacities of the quantized m-LFMC, where the values of the channel inputs are known and fixed and the standard deviation function of the noise is $\sigma_1(x)$. We explore an $(m * 2^k)$-level quantizer, in which the output is quantized into $(m * 2^k)$ intervals in a (non-uniform) way such that, around each level, there are 2^k equi-spaced intervals (which are indexed by k bits). An example of the quantization for the 4-LFMC with $x_0 = 0$, $x_1 = 3.25$, $x_2 = 4.55$ and $x_3 = 6.5$ is shown in Figure 12. Around each level, it is uniformly quantized into four intervals. Actually, it is a non-uniform quantization over $\mathbb{R}$. Then the channel becomes an DMC with four inputs and sixteen outputs. As mentioned in Section III-B, the quantized channel is a

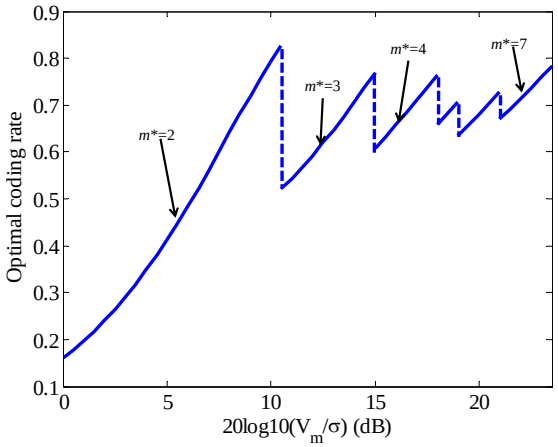

Figure 9. The optimal coding rates for m-LFMCs when the standard deviation function is $\sigma_1(x)$.

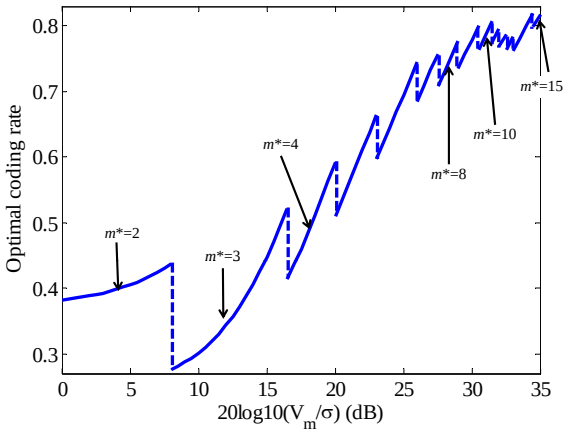

Figure 10. The optimal coding rates for m-LFMCs when the standard deviation function is $\sigma_2(x)$.

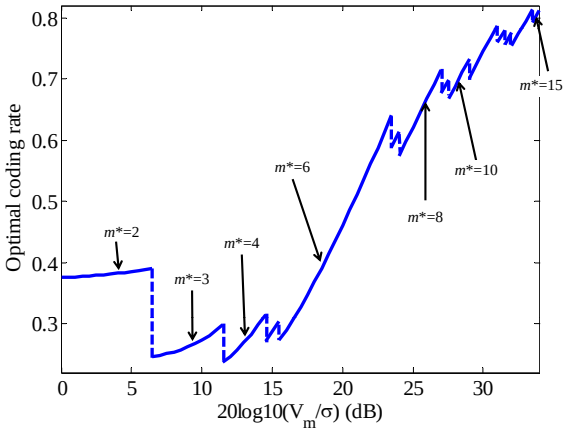

Figure 11. The optimal coding rates for m-LFMCs when the standard deviation function is $\sigma_3(x)$.

DMC whose capacity can be computed by the Blahut-Arimoto algorithm.

Suppose that six different $(m*2^k)$-level quantizers, where $k=0$, $k=1$, $k=2$, $k=3$, $k=4$ and $k=5$ are used.

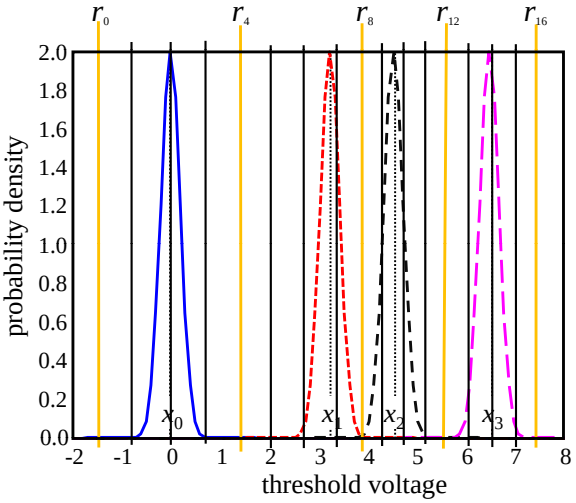

Figure 12. A (non-unform) 16-level quantization of the channel output for the 4-LFMC, where $x_0=0$, $x_1=3.25$, $x_2=4.55$ and $x_3=6.5$.

These quantizers are denoted by Q0, Q1, Q2, Q3, Q4 and Q5, respectively. Figures 13 and 14 show the capacities of different quantized DMCs for the 2-LFMC, where the channel inputs are fixed $\mathcal{X}^{(2)}=\{0,6.5\}$. Figures 15 and 16 show the capacities of different quantized DMCs for the 4-LFMC, where the channel inputs are fixed $\mathcal{X}^{(4)}=\{0,3.25,4.55,6.5\}$. Figures 17 and 18 show the capacities of different quantized DMCs for the 8-LFMC, where the channel inputs are fixed $\mathcal{X}^{(4)}=\{0,1,2,3,4,5,6,6.5\}$. Also shown in these figures are the exact capacities without quantization. From these figures, we can see that,

1) as the number of quantization bits around each level (i.e., k) increases, the quantized capacity does not decrease;
2) in lower VDR regimes, a finite quantization induces some loss of capacity, as shown in Figures 13, 15, and 17;
3) in high VDR regimes, the quantized capacity of the 3-bit quantization around each level almost matches the exact capacity without quantization, as shown in Figures 14, 16 and 18.

Hence, for an m-LFMC with given channel inputs, the $m*2^k$-level quantization (where k could be very small - at most 3) is good enough to practically approach the channel capacity.

VI. Conclusion

In this paper, the m-level flash memory was modeled as an m-AM channel with ID-AGN, in which the standard deviation of noise depends on the channel input. The capacity and the optimal coding rate of the m-LFMC were given. A simpler DMC was also derived by channel output quantization, which drove an approximation of the capacity for the m-LFMC. The determination of the capacity of the m-LFMC is an optimization problem, which can be transformed into two optimization sub-problems. One can be solved by the Blahut-Arimoto algorithm. The other can be solved by finding the solution to KKT conditions. Based on these, an alternating iterative algorithm was presented to evaluate a lower bound

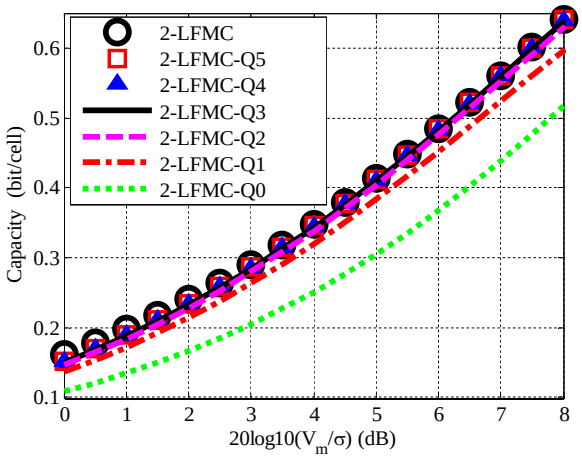

Figure 13. The capacities of different quantized DMCs for the 2-LFMC, where the channel inputs are fixed $\mathcal{X}^{(2)} = \{0, 6.5\}$.

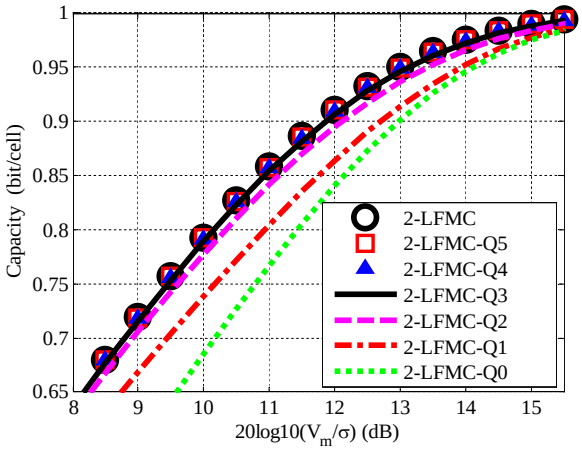

Figure 14. The capacities of different quantized DMCs for the 2-LFMC, where the channel inputs are fixed $\mathcal{X}^{(2)} = \{0, 6.5\}$.

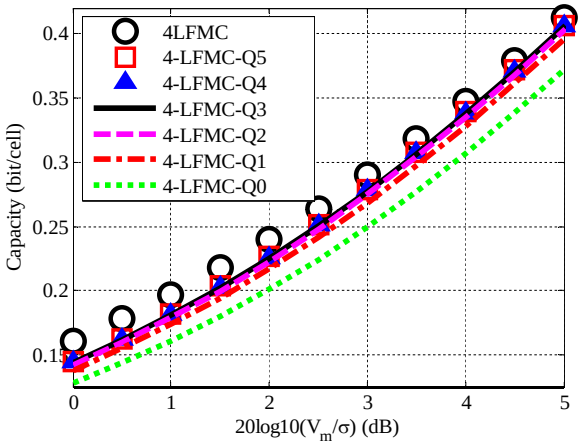

Figure 15. The capacities of different quantized DMCs for the 4-LFMC, where the channel inputs are fixed $\mathcal{X}^{(4)} = \{0, 3.25, 4.55, 6.5\}$.

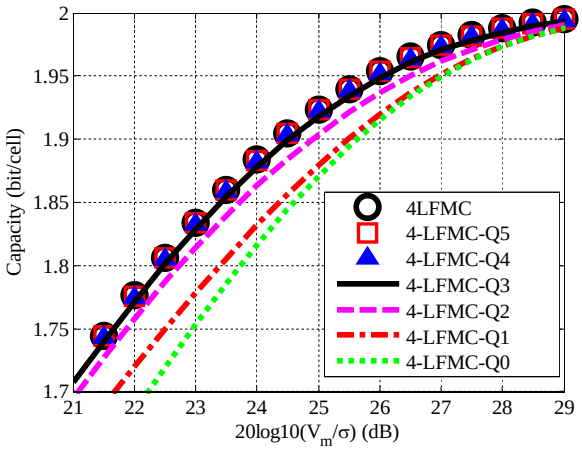

Figure 16. The capacities of different quantized DMCs for the 4-LFMC, where the channel inputs are fixed $\mathcal{X}^{(4)} = \{0, 3.25, 4.55, 6.5\}$.

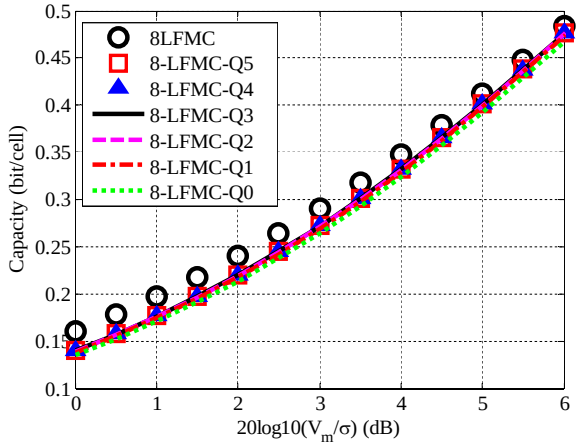

Figure 17. The capacities of different quantized DMCs for the 8-LFMC, where the channel inputs are fixed $\mathcal{X}^{(8)} = \{0, 1, 2, 3, 4, 5, 6, 6.5\}$.

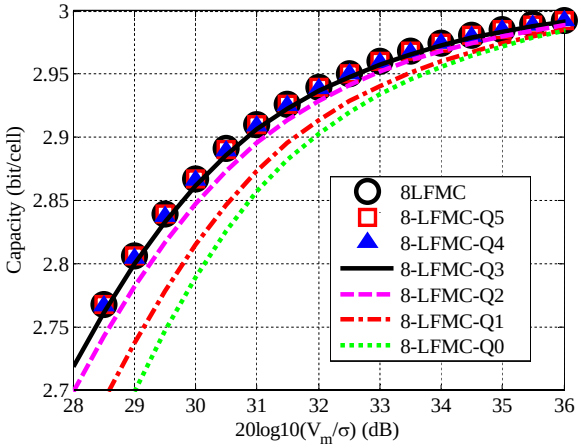

Figure 18. The capacities of different quantized DMCs for the 8-LFMC, where the channel inputs are fixed $\mathcal{X}^{(8)} = \{0, 1, 2, 3, 4, 5, 6, 6.5\}$.

on the capacity of the m-LFMC. This algorithm delivered not only the optimal distribution of channel inputs but also the optimal values of channel inputs. Numerical results showed that at any given VDR there exists an optimal (i.e., minimal) value m^* such that the capacity (or its lower bound) is achieved by an m^*-LFMC, and that increasing the number of levels m above m^* does not further increase the information rate for a fixed VDR. Numerical results also showed that if $(m * 2^k)$-

level quantizers with $k = 3$ are used at the channel output, the quantized capacity almost matches the capacity of the m-LFMC. Moreover, using the optimal coding rates as shown in the numerical results, we will design proper codes for the MLFM, which is one of our future works.

ACKNOWLEDGMENT

This work was supported by the collaborative NSF grants ECCS-1128705 and ECCS-1128148, and by the NSFC (No. 61172082).

REFERENCES

[1] X. Huang, A. Kavcic, X. Ma, G. Dong, and T. Zhang, "Optimization of achievable information rates and number of levels in multilevel flash memories," in *ICN 2013: The Twelfth International Conference on Networks*, Seville, Spain, Jan. 27-Feb. 1 2013, pp. 125–131.

[2] M. Bauer, R. Alexis, and et al., "A multilevel-cell 32Mb flash memory," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, Feb. 1995, pp. 132–133, 351.

[3] T.-S. Jung, Y.-J. Choi, and et al., "A 117-mm2 3.3-V only 128-Mb multilevel NAND flash memory for mass storage applications," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1575–1583, Nov. 1996.

[4] G. Atwood, A. Fazio, D. Mills, and B. Reaves, "Intel StrataFlash$^{\mathrm{TM}}$ memory technology overview," *Intel Technology Journal*, pp. 1–8, 4th Quarter 1997.

[5] Y. Li, S. Lee, and et al., "A 16Gb 3b/cell NAND flash memory in 56nm with 8MB/s write rate," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, Feb. 2008, pp. 506–507.

[6] T. Futatsuyama, N. Fujita, and et al., "A 113mm2 32Gb 3b/cell NAND flash memory," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, Feb. 2009, pp. 242–243.

[7] N. Shibata, H. Maejima, and et al., "A 70nm 16Gb 16-level-cell NAND flash memory," in *IEEE VLSI Circuits*, 2007, pp. 190–191.

[8] C. Trinh, N. Shibata, and et al., "A 5.6MB/s 64Gb 4b/cell NAND flash memory in 43nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, Feb. 2009, pp. 246–247, 247a.

[9] G. Dong, S. Li, and T. Zhang, "Using data postcompensation and predistortion to tolerate cell-to-cell interference in MLC NAND flash memory," *IEEE Trans. Circuits Syst.–I: Reg. Papers*, vol. 57, no. 10, pp. 2718–2728, Oct. 2010.

[10] J. Wang, T. Courtade, H. Shankar, and R. D. Wesel, "Soft information for LDPC decoding in flash: mutual-information optimized quantization," in *Proc. IEEE GLOBECOM 2011*, Houston, Texas, USA, Dec. 2011.

[11] S. Gregori, A. Cabrini, O. Khouri, and G. Torelli, "On-chip error correcting techniques for new-generation flash memories," *Proceedings of the IEEE*, vol. 91, no. 4, pp. 602–616, Apr. 2003.

[12] F. Sun, S. Devarajan, K. Rose, and T. Zhang, "Design of on-chip error correction systems for multilevel NOR and NAND flash memories," *IET Circuits Devices Syst.*, vol. 1, no. 3, pp. 241–249, 2007.

[13] J. Chen and P. H. Siegel, "Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels," *IEEE Trans. Inform. Theory*, vol. 54, no. 3, pp. 1295–1303, Mar. 2008.

[14] B. M. Kurkoshi, "The E8 lattice and error correction in multi-level flash memory," in *Proc. IEEE International Conference on Communications*, Kyoto, Japan, June 5-9 2011, pp. 1–5.

[15] G. Dong, N. Xie, and T. Zhang, "On the use of soft-decision error-correction codes in NAND flash memory," *IEEE Trans. Circuits Syst.–I: Reg. Papers*, vol. 58, no. 2, pp. 429–439, Feb. 2011.

[16] H. Lou and C. Sundberg, "Increasing storage capacity in multilevel memory cells by means of communications and signal processing techniques," *IEE Proc.-Circuits Devices Syst.*, vol. 147, no. 4, pp. 229–236, Aug. 2000.

[17] S. Soldà, D. Vogrig, A. Bevilacqua, A. Gerosa, and A. Neviani, "Analog decoding of trellis coded modulation for multi-level flash memories," in *Proc. the 2008 IEEE International Symposium on Circuits and Systems (ISCAS 2008)*, Seattle, U.S.A., May 18-21 2008, pp. 744–747.

[18] A. Jiang, R. Mateescu, M. Schwartz, and J. Bruck, "Rank modulation for flash memories," *IEEE Trans. Inform. Theory*, vol. 55, no. 6, pp. 2659–2673, Jun. 2009.

[19] Z. Wang and J. Bruck, "Partial rank modulation for flash memories," in *Proc. IEEE Intern. Symp. on Inform. Theory*, Austin, Texas, U.S.A., June 13-18 2010, pp. 864–868.

[20] D. Park and J. Lee, "Floating-gate coupling canceller for multi-level cell NAND flash," *IEEE Trans. Magn.*, vol. 47, no. 3, pp. 624–628, Mar. 2011.

[21] J. G. Smith, "On the information capacity of peak and average power constrained gaussian channels," Ph.D. dissertation, University of California, Berkeley, California, Dec. 1969.

[22] ——, "The information capacity of amplitude-and variance-constrained scalar Gaussian channels," *Information and Control*, vol. 18, pp. 203–219, 1971.

[23] Kingston, "Flash memory guide," Kingston, Tech. Rep., 2011.

[24] J.-D. Lee, S.-H. Hur, and J.-D. Choi, "Effects of floating-gate interference on NAND flash memory cell operation," *IEEE Electron Device Letters*, vol. 23, no. 5, pp. 264–266, May 2002.

[25] M. Asadi, X. Huang, A. Kavcic, and N. P. Santhanam, "Optimal detector for multilevel NAND flash memory channels with intercell interference," Nov. 2013, accepted by IEEE Journal on Seleted Areas in Communcation (J-SAC).

[26] S. Shamai, "Information theoretic aspects of constrained systems," in *MSRI Workshop on Information Theory*, Berkeley, California, U.S.A., Feb. 25 - Mar. 1 2002.

[27] ——, "Information theoretic aspects of constrained cell-sites cooperation," in *IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Nov. 17-20 2010, p. 000086.

[28] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc, 1991.

[30] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[31] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.

[32] A. Kavčić, "On the capacity of Markov sources over noisy channels," in *Proc. IEEE GLOBECOM 2001*, vol. 5, San Antonio, TX, USA, Nov. 25-29 2001, pp. 2997–3001.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.