

Optimization of Achievable Information Rates and Number of Levels in Multilevel Flash Memories

Xiujie Huang*, Aleksandar Kavcic*, Xiao Ma†, Guiqiang Dong‡ and Tong Zhang‡

*Department of Electrical Engineering, University of Hawaii, Honolulu, HI 96822 USA

†Department of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, GD 510006 China

‡Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 USA

Email: xiujie@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 channel (m -LFMC) is modeled as an m -amplitude-modulation channel with input-dependent additive Gaussian noise whose standard deviation depends on the channel input. Then the capacity of an m -LFMC is given. The determination of the capacity 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.

Keywords—Channel capacity; input-dependent additive Gaussian noise (ID-AGN) channel; multilevel flash memory.

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. The first MLFM product was presented by Bauer *et al.* in [1], which had four levels and stored two bits in one memory cell. Later, MLFMs were investigated and/or designed by many researchers, such as the 4-level MLFM in [2] and the Intel StrataFlash™ 4-level memory in [3], the 8-level MLFMs in [4, 5] and the 16-level MLFMs in [6, 7].

As the number of levels increases, the capability of the MLFM could be enhanced and the reliability could be decreased. On one hand, due to the complexity of the configuration (including the programming and 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 [8, 9]. In [8], 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. On the other hand, to guarantee the reliability, two approaches, i.e., signal processing methods and error correcting codes (ECCs),

are investigated and applied in MLFMs. Two examples of signal processing methods are data postcompensation and data predistortion [10], which could tolerate cell-to-cell interference in MLFMs. Examples of ECCs include BCH codes [11], Reed-Solomon codes [12, 13], LDPC codes [8, 14], trellis coded modulation [11, 15, 16] and rank modulation [17, 18].

In this paper, we focus on the MLFM with m levels. To answer the question on the information-theoretic capability of the MLFM with m levels, we need to solve a key problem, i.e., channel modeling. The simplest model is the input-independent additive Gaussian noise channel, which is also called amplitude-limited AWGN channel [19, 20]. In [19, 20], 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 [3, 21]. The other is the cell-to-cell interference as mentioned in [10, 22]. In this paper, we consider only the former. In this case, we model the 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 input (The m -AM with ID-AGN channel can also be regarded as a constrained communication system [23, 24]). Then we give the channel capacity and present a numerical method to evaluate it.

Structure: The remainder of this paper is organized as follows. The channel model of the MLFM with m levels is introduced in Section II, and the channel capacity is given in Section III. Section IV presents an alternating iterative algorithm to evaluate a lower bound on the capacity. Numerical results and discussions are shown in Section V, followed by the conclusion in Section VI.

II. CHANNEL MODEL

For an MLFM with m levels, each level has an intended *threshold voltage* [1]. Affected by the configuration (including the programming and reading techniques and inter-cell interferences) 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 [1]. In this

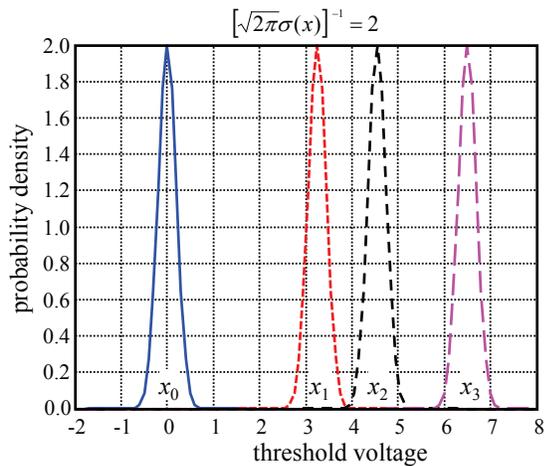


Fig. 1. A threshold voltage distribution model for a 4-LFMC, in which the noise at each level has the same variance $\sigma(x) = \frac{1}{2\sqrt{2\pi}}$.

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 $x_3 = 6.5$. By default, the threshold voltage distribution model of the manufactured 4-LFMC is shown in Fig. 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 [3, 21], 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 Fig. 2, where every level experiences more noise than in Fig. 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 Fig. 3, where every level has even more noise than in Fig. 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 Figs. 2 and 3, the dash-dot-dot curve $[\sqrt{2\pi}\sigma(x)]^{-1}$ is (approximately) the envelope of the peaks of the level-output-pdfs. In Fig. 1, the curve $\sigma(x)$ is assumed to be a constant, i.e., $\sigma(x) = \frac{1}{2\sqrt{2\pi}}$. \square

Models similar to Figs. 2 and 3 for the 4-LFMC were introduced in [3, 11, 14]. In [3, 11], the model of the 2 bits/cell (i.e., 4-level) NOR flash memory was presented, in which 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

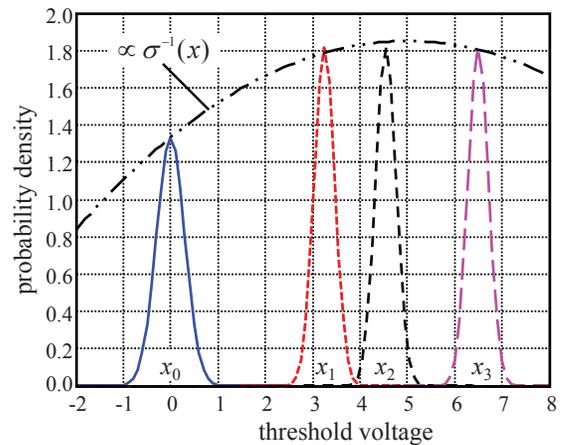


Fig. 2. A threshold voltage distribution model for a 4-LFMC, in which the first level x_0 is the most noisy level while the other three levels have roughly the same noise.

had almost the same noise variances. In [14], the model of a 4-level NAND flash memory was derived by accounting for the cell-to-cell interference, in which 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.

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:

$$Y = X + W. \quad (1)$$

- 2) The channel input X takes values from a finite alphabet $\mathcal{X}^{(m)} \triangleq \{x_0, x_1, \dots, x_{m-1}\}$ under the constraint

$$a \leq x_0 < x_1 < x_2 < \dots < x_{m-2} < x_{m-1} \leq b \quad (2)$$

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

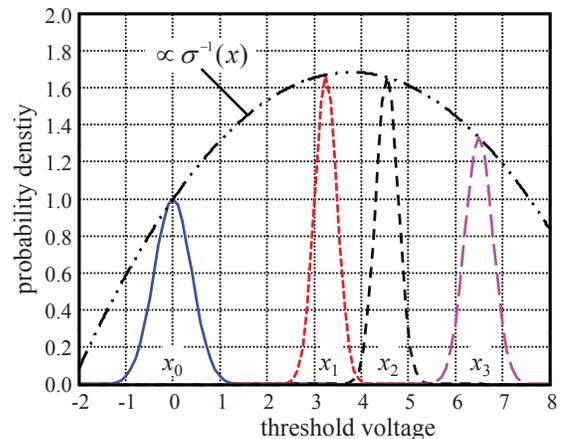


Fig. 3. A threshold voltage distribution model for a 4-LFMC, in which the first level x_0 and the last level x_3 are respectively the most noisy levels while the two middle levels x_1 and x_2 have roughly the same noise.

an m -amplitude-modulation (m -AM) signal set. Denote the collection of all such m -AM signal sets as $\mathcal{X}^{(m)}$, i.e., $\mathcal{X}^{(m)} \in \mathcal{X}^{(m)}$. In the following context, we also use the vector notation \underline{x} to denote the m levels, i.e., $\underline{x} = (x_1, x_2, \dots, x_{m-1})$.

- 3) The probability mass function (pmf) of X over $\mathcal{X}^{(m)}$ is denoted by $\underline{p} = (p_0, p_1, \dots, p_{m-1})$ with $p_i = \Pr(X = x_i)$.
- 4) The noise \bar{W} is an input-dependent additive Gaussian noise (ID-AGN). That is, the noise W has mean zero and variance depending on the channel input $x \in \mathcal{X}^{(m)}$, i.e., $W \sim \mathcal{N}(0, \sigma^2(x))$. 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

$$f_{Y|X, \sigma(\cdot)}(y|x) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left\{-\frac{(y-x)^2}{2\sigma^2(x)}\right\}. \quad (3)$$

And the pdf of the channel output Y can be obtained as

$$f_{Y, \sigma(\cdot)}(y) = \sum_{i=0}^{m-1} p_i f_{Y|X, \sigma(\cdot)}(y|x_i). \quad (4)$$

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 Fig. 1. As the device ages, the noise standard deviations for different levels increase in different extents; see Figs. 2 and 3. That is, the noise standard deviations for an aged device are level-dependent.

III. CHANNEL CAPACITY

From the last section, 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 $\underline{p} = (p_0, p_1, \dots, p_{m-1})$ and the standard deviation function $\sigma(x)$. Therefore, to express the information-theoretic capacity of the m -LFMC, we introduce a new notation different slightly from the conventional one by inserting the subscript $(\mathcal{X}^{(m)}, \sigma(\cdot))$ into the mutual information expression, i.e.,

$$I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \triangleq \sum_{i=0}^{m-1} \int_{-\infty}^{\infty} p_i f_{Y|X, \sigma(\cdot)}(y|x_i) \log\left(\frac{f_{Y|X, \sigma(\cdot)}(y|x_i)}{f_{Y, \sigma(\cdot)}(y)}\right) dy. \quad (5)$$

Definition 1: The capacity of the m -LFMC with standard deviation function $\sigma(\cdot)$ is defined as

$$C_{m, \sigma(\cdot)} \triangleq \sup_{\mathcal{X}^{(m)}, \{\underline{p}\}} I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \quad (6)$$

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

$$a \leq x_0 < x_1 < \dots < x_{m-2} < x_{m-1} \leq b \quad (7)$$

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

$$p_i \geq 0, \text{ and } \sum_{i=0}^{m-1} p_i = 1. \quad (8)$$

□

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 [19, 20]. 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 [25], 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

$$I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) = h_{\mathcal{X}^{(m)}, \sigma(\cdot)}(Y) - \sum_{i=0}^{m-1} p_i \log \sigma(x_i) - \frac{1}{2} \log(2\pi e) \quad (9)$$

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 [26]. When \underline{p} is given, the composition of elementary functions in (5) is continuous and differentiable w.r.t. \underline{x} because $\sigma(x)$ is assumed to be continuous and differentiable. ■

IV. EVALUATION OF A LOWER BOUND ON CAPACITY

To evaluate the capacity (6) of the m -LFMC, we turn to a two-step optimization problem

$$C_{m, \sigma(\cdot)} = \sup_{\underline{x} \in [a, b]^m} \sup_{\underline{p} \in [0, 1]^m} I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \quad (10)$$

subject to $\begin{cases} a \leq x_0 < x_1 < \dots < x_{m-1} \leq b \\ p_i \geq 0, i \in \{0, 1, \dots, m-1\} \\ \sum_{i=0}^{m-1} p_i = 1 \end{cases}$

To solve the two-step optimization problem (10), we turn to two sub-problems.

Sub-problem I.

$$C(\underline{x}) = \max_{\underline{p} \in [0, 1]^m} I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \quad (11)$$

subject to $\begin{cases} p_i \geq 0, i \in \{0, 1, \dots, m-1\} \\ \sum_{i=0}^{m-1} p_i = 1 \end{cases}$

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) [27–29] can be used to solve Sub-problem I.

Sub-problem II.

$$C(\underline{p}) = \max_{\underline{x} \in [a,b]^m} I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \quad (12)$$

subject to $a \leq x_0 < x_1 < \dots < x_{m-1} \leq b$.

The Karush-Kuhn-Tucker (KKT) conditions [30] of the sub-problem are that there exists $\mathbf{v}^* = (\underline{x}^*, \lambda^*, \mu^*)$ such that

$$\left\{ \begin{array}{l} \frac{\partial I}{\partial x_0} \Big|_{\mathbf{v}^*} = -\lambda^*, \\ \frac{\partial I}{\partial x_{m-1}} \Big|_{\mathbf{v}^*} = \mu^*, \\ \frac{\partial I}{\partial x_i} \Big|_{\mathbf{v}^*} = 0, \quad i \in \{1, 2, \dots, m-2\} \\ x_0^* \geq a, \\ x_{m-1}^* \leq b, \\ x_{i-1}^* < x_i^*, \quad i \in \{1, 2, \dots, m-1\} \\ \lambda^* \geq 0, \\ \mu^* \geq 0, \\ \lambda^*(x_0^* - a) = 0, \\ \mu^*(x_{m-1}^* - b) = 0. \end{array} \right. \quad (13)$$

Note that the solution of (13) may be sub-optimal (a local solution) since the concavity of the mutual information w.r.t. \underline{x} is unknown; see the Appendix for a method to solve (13).

Based on the two sub-problems, an alternating iterative scheme is presented to solve problem (10). At each iteration, the two-stage alternating strategy shown below is employed.

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

$$\underline{p}^* = \arg \max_{\underline{p}} I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y). \quad (14)$$

Stage 2. Fix \underline{p} . Solve (13) to obtain a better \underline{x}^* such that

$$I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \Big|_{\underline{x}^*} \geq I_{\mathcal{X}^{(m)}, \sigma(\cdot)}(X; Y) \Big|_{\underline{x}}. \quad (15)$$

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 numerically compute lower bounds on capacities of different m -LFMCs using the alternating iterative scheme of Section IV. We also interpret the results and put them in context with respect to prior work [19, 20].

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 serves 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 Fig. 4. We consider three different standard deviation functions $\sigma(x)$, denoted as

$$\sigma_i(x) = q_i(x) \cdot \sigma, \quad \text{where } i \in \{1, 2, 3\}. \quad (16)$$

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

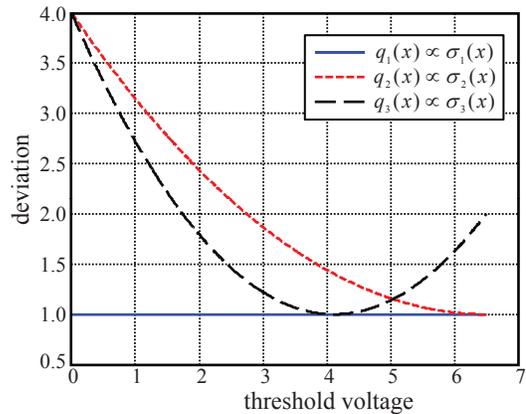


Fig. 4. The standard deviation functions $\sigma_i(x)$ of the input-dependent Gaussian noise W : $\sigma_i(x) \propto q_i(x)$, where $i \in \{1, 2, 3\}$.

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 Figs. 5, 6 and 7, respectively.

From Fig. 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 15.5 dB, 3-LFMC, 4-LFMC and 5-LFMC have the same rates.
- 3) When the VDR is less than 18.5 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 15.5 dB, the optimized lower bound is achieved with $m^* = 3$ levels, even if, say, the constraint allows up to $m = 5$ levels. This is consistent with prior work [19, 20] which showed that for the amplitude-limited AWGN channel, the capacity is achieved by a discrete channel input distribution over a *finite* alphabet. In other words, for a fixed VDR, there is an optimal number of levels m^* for a given multilevel flash memory channel. Increasing the number of levels m beyond m^* does not further increase the capacity (nor the computed lower bound.)

These observations imply that 2-LFMC, 3-LFMC and 4-LFMC can achieve the capacity

$$C_{\sigma_1(\cdot)} = \max_m C_{m, \sigma_1(\cdot)}$$

in the cases of $\text{VDR} \leq 10.5$ dB, $10.5 \text{ dB} < \text{VDR} \leq 15.5$ dB and $15.5 \text{ dB} < \text{VDR} \leq 18.5$ dB, respectively. In other words, in the view of capacity, at a given VDR less than 10.5 dB, a 2-LFMC is “optimal”; at a given VDR less than 15.5 dB, a 3-LFMC is “optimal”; at a given VDR less than 18.5 dB, a 4-LFMC is “optimal”. Naturally, as the VDR increases, the optimal number of levels doesn’t decrease.

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 Fig. 7, the optimal number of levels is $m^* = 4$ in the VDR regime between 13 dB and

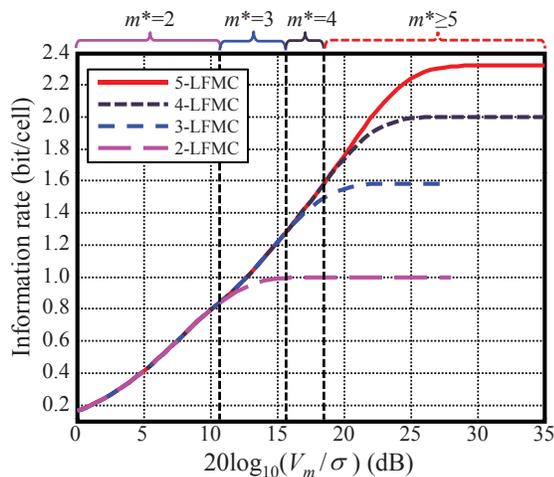


Fig. 5. The information rates of m -LFMCs with $m \in \{2, 3, 4, 5\}$ when the standard deviation function is $\sigma_1(x)$. The numbers m^* on the top of the figure indicate that 2-LFMC, 3-LFMC and 4-LFMC can achieve the (computed) maximum rates in the cases of $VDR \leq 10.5$ dB, 10.5 dB $< VDR \leq 15.5$ dB and 15.5 dB $< VDR \leq 18.5$ dB, respectively.

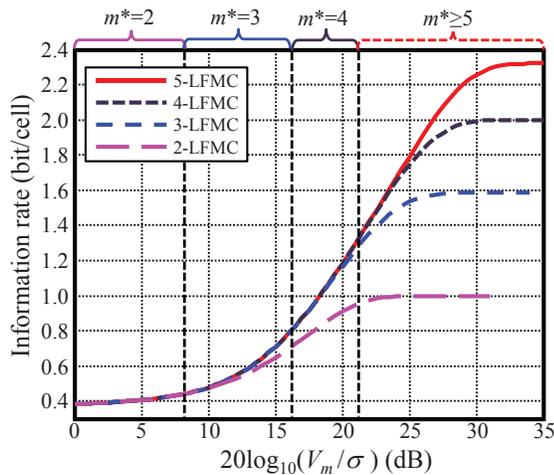


Fig. 6. The information rates of m -LFMCs with $m \in \{2, 3, 4, 5\}$ when the standard deviation function is $\sigma_2(x)$. The numbers m^* on the top of the figure indicate that 2-LFMC, 3-LFMC and 4-LFMC can achieve the (computed) maximum rates in the cases of $VDR \leq 8.5$ dB, 8.5 dB $< VDR \leq 16$ dB and 16 dB $< VDR \leq 21$ dB, respectively.

17.5 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 Fig. 8. Again, this is consistent with the literature [19, 20] for the amplitude-limited AWGN channel, even though in m -LFMC the noise standard deviation $\sigma(x)$ is input-dependent.

VI. CONCLUSIONS

In this paper, the m -LFMC was modeled as an m -AM channel with ID-AGN, in which the standard deviation of

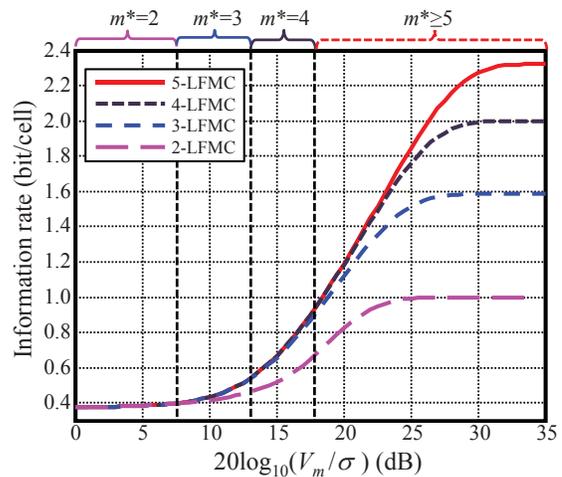


Fig. 7. The achievable rates of m -LFMCs with $m \in \{2, 3, 4, 5\}$ when the standard deviation function is $\sigma_3(x)$. The numbers m^* on the top of the figure indicate that 2-LFMC, 3-LFMC and 4-LFMC can achieve the (computed) maximum rates in the cases of $VDR \leq 7.5$ dB, 7.5 dB $< VDR \leq 13.5$ dB and 13.5 dB $< VDR \leq 18$ dB, respectively.

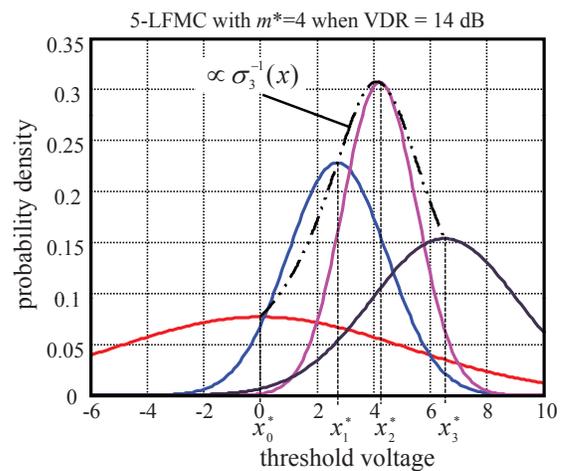


Fig. 8. The pdfs of channel output distributions around the optimal threshold voltage levels when the m -LFMC with the standard deviation function $\sigma_3(x)$ and $m = 5$ is used at VDR = 14 dB. The optimal number of levels m^* is 4 with assignment $x_0^* = 0$, $x_1^* \approx 2.718$, $x_2^* \approx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

noise depends on the channel input. The capacity of the m -LFMC was given. The determination of the capacity is an optimization problem, which can be transformed into two optimization sub-problems. One can be solved by 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 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 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.

APPENDIX: SOLVING KKT CONDITIONS (13)

For convenience, we denote the pdfs $f_{Y|X,\sigma(\cdot)}(y|x_i)$ and $f_{Y,\sigma(\cdot)}(y)$ and the mutual information $I_{X^{(m)},\sigma(\cdot)}(X;Y)$ 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 compute partial derivatives of the transition pdf $f(y|x_i)$ in (3) and the output pdf $f(y)$ in (4) as, for all $i \in \{0, 1, \dots, m-1\}$,

$$\frac{\partial f(y|x_i)}{\partial x_i} = \begin{cases} f(y|x_i) \left[-\frac{\sigma'(x_i)}{\sigma(x_i)} + \frac{y-x_i}{\sigma^2(x_i)} + \frac{(y-x_i)^2 \sigma'(x_i)}{\sigma^3(x_i)} \right], & \text{if } i=j \\ 0, & \text{if } i \neq j \end{cases} \quad (17)$$

where $\sigma'(x_i) \triangleq \frac{d\sigma(x)}{dx_i}$ denotes the derivative of $\sigma(x_i)$ w.r.t. x_i . Then, according to (9), partial derivatives of the mutual information are obtained as

$$\begin{aligned} \frac{\partial}{\partial x_i} I(X;Y) &= - \int_{-\infty}^{\infty} \frac{\partial}{\partial x_i} (f(y) \ln f(y)) dy - \frac{p_i \sigma'(x_i)}{\sigma(x_i)} \\ &= \left[-\frac{p_i \sigma'(x_i)}{\sigma^3(x_i)} \int_{-\infty}^{\infty} f(y|x_i) \ln f(y) dy \right] \cdot x_i^2 \\ &\quad + \left[\frac{2p_i \sigma'(x_i)}{\sigma^3(x_i)} \int_{-\infty}^{\infty} y f(y|x_i) \ln f(y) dy \right. \\ &\quad \left. + \frac{p_i}{\sigma^2(x_i)} \int_{-\infty}^{\infty} f(y|x_i) \ln f(y) dy \right] \cdot x_i \\ &\quad + \left[-\frac{p_i \sigma'(x_i)}{\sigma^3(x_i)} \int_{-\infty}^{\infty} y^2 f(y|x_i) \ln f(y) dy \right. \\ &\quad \left. - \frac{p_i}{\sigma^2(x_i)} \int_{-\infty}^{\infty} y f(y|x_i) \ln f(y) dy \right. \\ &\quad \left. + \frac{p_i \sigma'(x_i)}{\sigma(x_i)} \int_{-\infty}^{\infty} f(y|x_i) \ln f(y) dy - \frac{p_i \sigma'(x_i)}{\sigma(x_i)} \right] \\ &\triangleq A_i x_i^2 + B_i x_i + C_i. \end{aligned} \quad (18)$$

Solving KKT conditions (13) is equivalent to finding quantities $(\underline{x}, \lambda, \mu)$ that satisfy the equalities

$$\begin{cases} A_0 x_0^2 + B_0 x_0 + (C_0 + \lambda) = 0 \\ \lambda(x_0 - a) = 0 \end{cases}, \quad (19a)$$

$$\begin{cases} A_{m-1} x_{m-1}^2 + B_{m-1} x_{m-1} + (C_{m-1} - \mu) = 0 \\ \mu(x_{m-1} - b) = 0 \end{cases}, \quad (19b)$$

$$A_i x_i^2 + B_i x_i + C_i = 0, \quad i \in \{1, 2, \dots, m-2\}, \quad (19c)$$

and the inequalities

$$\begin{cases} \lambda \geq 0 \\ \mu \geq 0 \\ a \leq x_0 < x_1 < \dots < x_{m-2} < x_{m-1} \leq b \end{cases}. \quad (20)$$

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 (19) by an iterative method, we assume that quantities A_i , B_i and C_i are independent of x_i . Then Eqns. (19) have at most $9 \times 2^{m-2}$ solutions. Moreover, under the full constraints in (20), the number of solutions may be much less than $9 \times 2^{m-2}$ (This happens in our numerical computations). Based

on (19) and (20), 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. (19). Pick those solutions that satisfy all constraints in (20), and from them choose the one with the highest information rate as the improved input vector $\underline{x}^{(k+1)}$.

ACKNOWLEDGMENT

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

REFERENCES

- [1] 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.
- [2] 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.
- [3] G. Atwood, A. Fazio, D. Mills, and B. Reaves, "Intel StrataFlash™ memory technology overview," *Intel Technology Journal*, pp. 1–8, 4th Quarter 1997.
- [4] 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.
- [5] 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.
- [6] N. Shibata, H. Maejima, and et al., "A 70nm 16Gb 16-level-cell NAND flash memory," in *IEEE VLSI Circuits*, 2007, pp. 190–191.
- [7] 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.
- [8] 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.
- [9] A. Jiang, H. Li, and J. Bruck, "On the capacity and programming of flash memories," *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1549–1564, Mar. 2012.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.

- [19] 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.
- [20] —, "The information capacity of amplitude-and variance-constrained scalar Gaussian channels," *Information and Control*, vol. 18, pp. 203–219, 1971.
- [21] Kingston, "Flash memory guide," Kingston, Tech. Rep., 2011.
- [22] 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.
- [23] S. Shamai, "Information theoretic aspects of constrained systems," in *MSRI Workshop on Information Theory*, Berkeley, California, U.S.A., Feb. 25 - Mar. 1 2002.
- [24] —, "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.
- [25] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.
- [26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc, 1991.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.