Optimizing Write Voltages for Independent, Equal-Rate Pages in Flash Memory

Semira Galijasevic
Department of Electrical and Computer Engineering
University of California
Los Angeles, USA
semiragali@g.ucla.edu

Richard D. Wesel
Department of Electrical and Computer Engineering
University of California
Los Angeles, USA
wesel@ucla.edu

***Abstract*—This paper uses a mutual-information maximization paradigm to optimize the voltage levels written to cells in a Flash memory. To enable low-latency, each page of Flash memory stores only one coded bit in each Flash memory cell. For example, three-level cell (TL) Flash has three bit channels, one for each of three pages, that together determine which of eight voltage levels are written to each cell. Each Flash page is required to store the same number of data bits, but the various bits stored in the cell typically do not have to provide the same mutual information. A modified version of dynamic-assignment Blahut-Arimoto (DAB) moves the constellation points and adjusts the probability mass function for each bit channel to increase the mutual information of a worst bit channel with the goal of each bit channel providing the same mutual information. The resulting constellation provides essentially the same mutual information to each page while negligibly reducing the mutual information of the overall constellation. The optimized constellations feature points that are neither equally spaced nor equally likely. However, modern shaping techniques such as probabilistic amplitude shaping can provide coded modulations that support such constellations.**

***Index Terms*—Flash memory, mutual information, coded modulation, constellation design, constellation shaping**

I. Introduction

A. Background

Flash memory is a non-volatile storage medium invented in the 1980s that quickly became a prominent segment within the semiconductor industry [1] . With fast reads and no moving parts, flash memory is widely used for storage and data transfer in consumer devices such as phones, digital cameras, SD cards, tablets and laptops, enterprise systems, data centers and industrial applications. Flash memory can retain data for a long period of time regardless of whether a flash-equipped device is powered on or off. It is small, reliable, and inexpensive which makes it attractive for mobile and miniature products, two major market demands for electronic devices.

To meet demands for information density, flash memory has introduced multilevel cell techniques and technology scaling which degrades the Flash read channel [2]. Sometimes, the values manufacturers provide in publicly available datasheets differ from the actual Flash device performance, which is significantly worse and highly variable as demonstrated in [2]. These trends exacerbate inter-cell interference and program/erase (P/E) cycling effects. A P/E cycle includes writing (programming) the data to the cells in a page, reading the data from a page (possibly multiple times) and then erasing the data. The Flash memory read channel degrades over time as a function of number of P/E cycles or the cumulative amount of charge that is written into and subsequently erased from the memory cell [3], [4]. Multilevel cell techniques store multiple bits in each cell. For example, triple-level-cell (TLC) Flash stores three bits per cell by using eight levels of write voltage. However, to minimize the read latency, each bit in the same cell is mapped to a different page. In this manner the bits corresponding to the same cell are encoded independently.

While nearly all practical devices encode each bit in the cell independently as part of a different page, the academic literature sometimes explores Flash memory capacity for the joint-encoding case where all the bits belonging to the same cell are jointly encoded as part of one codeword. Example papers considering joint encoding for Flash memory include [1], [5], [6], [7], [8], [9], [10]. Various methods for maximizing mutual information for encoding and decoding Flash read channels were explored in [6], [7], [8], [9], [10] and [11]. Recent work considering independent encoding in comparison with joint encoding include [12], [13], and [14]. In [12] the benefit of joint encoding for multi-level cell (MLC) is examined for traditional hard decoding and enhanced precision decoding in terms of mutual information (MI). It is shown that for MLC the difference in sum-rate MI between joint and independent encoding is small for hard decoding and negligible for enhanced precision decoding as long as the pages are permitted to have different rates.

In [13], the focus is the differing performance of the possible binary labelings of the cell levels for independent encoding for MLC and TLC flash for various decoding schemes. The results obtained in [13] show that for P/E cycling model the largest sum rate is achieved with Gray labeling for Treating Interference as Noise (TIN) decoding. Very recently, [14] explored both joint and independent encoding for TLC. They found that joint encoding increases the mutual information and improves hard-input error correction performance with LDPC codes [14].

B. Contributions.

While joint encoding provides an information-theoretic benefit, this paper focuses exclusively on the practical scenario of independent encoding of Flash under the constraint that all pages must convey the same number of bits. We use a mutual information maximizing paradigm similar to [15] to adapt the locations and probabilities of write levels to increase the mutual information of the weakest bit channel and hence improve the reliability of its corresponding page. In this way, we seek a constellation of write levels that delivers the same amount of mutual information to the bit channel for each page. For simplicity, we restrict attention to TLC Flash with a simplistic additive white Gaussian noise (AWGN) channel model.

C. Organization

The remaining parts of this paper are organized into four main sections. Sec. II describes Flash paradigm for writing and reading information. Sec. III introduces our problem statement from the mutual information perspective. Sec. IV describes two algorithms for maximizing the minimum information rate in under power and symmetry constraints. Alg. 1 adapts only the positions of write levels in the constellation, seeking to maximize the minimum rate. Alg. 2 adapts both the position and the probability of each levels, again seeking to maximize the minimum rate. Sec. V concludes the paper.

II. Writing and Reading Flash Information

A basic cell of a flash memory is composed of a floating gate transistor. The amount of charge in the floating gate controls the voltage at which the transistor turns on. We often refer to the write-level as a voltage because the read process will apply a voltage and learn whether the transistor is on or not via a sense-amp comparator.

The data stored in the cell is represented by cell voltage levels. The number of voltage levels corresponds to the number of bits stored in the cell. There are 2^m voltage levels when m coded bits are stored in each cell. Flash technologies are named based on the number of bits they can store. For example, Single Level Cell (SLC) stands for one bit per cell, Multi Level Cell (MLC) indicates two bits per cell, Triple Level Cell (TLC) indicates three bits per cell, etc. Cells are grouped into pages which are grouped into blocks. Pages are the smallest unit for write and read operations.

Since bits belonging to the same cell are mapped independently to different pages, the read process does not provide the actual voltage level, but a single bit at a time checking a threshold and reporting whether the voltage is above or below that threshold. Fig. 1 (following [13]) illustrates the eight voltage levels and Gray labeling for a cell in TLC flash memories. During each read process the information obtained is whether the voltage is on the left or the right side of the threshold depicted by vertical lines. We notice that for decoding three pages a total of 7 reads are needed. However, bit B_1 requires only one threshold check, B_2 requires two checks, and B_4 requires four thresholds to be checked. Additional threshold reads can provide soft information [12].

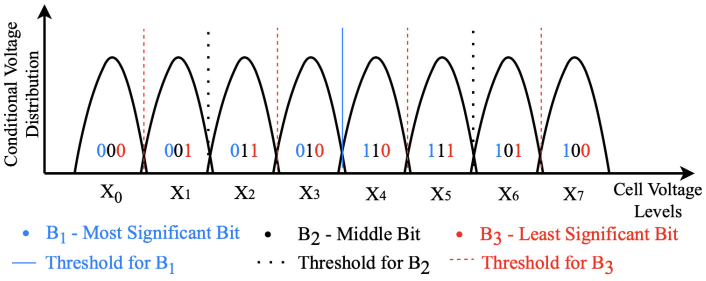

Fig. 1. The eight voltage levels and Gray labeling for a TLC flash memory. B_1, B_2, B_3 represent three bits corresponding to three different pages. $X_0, ..., X_7$ are cell voltage levels. Vertical lines represent thresholds for each bit. The read process provides a single bit at a time checking a threshold and reports whether the voltage is above or below that threshold.

Consider a TLC flash memory with eight voltage levels as illustrated in Fig. 1. The three bits B_1, B_2 and B_3 written to a cell for the three independent pages together cause the threshold voltage X to be written to the Flesh cell. As noted in [13], this is analogous to a multiple access channel (MAC) with three users as illustrated in Fig. 2. When the cell position is read, distortion causes the actual threshold voltage at the time of reading to be Y. Therefore, the system model is given by following equation:

$$Y = X + Z, \quad Z \sim \mathcal{N}(0, N) \tag{1}$$

The noise Z is assumed to be independent of the signal X.

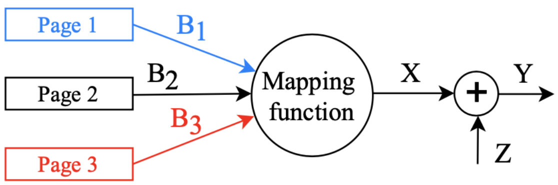

Fig. 2. Flash Memory Layout as Multiple Access Channel (MAC). The noise Z is drawn i.i.d. from a Gaussian distribution with variance N.

We can model the Flash write levels as M-ary pulse amplitude modulation (M-PAM). For TLC, we use 8-PAM constellations to store 3 bits per page. We will consider equally spaced equally likely (ESEL) 8-PAM as illustrated in the Table I as a baseline for comparison, although practical Flash write levels are not equally spaced.

TABLE I
8-PAM ESEL Constellations and Gray Labeling for TLC Flash

Voltage Levels	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
ESEL	-7	-5	-3	-1	1	3	5	7
B_1	0	0	0	0	1	1	1	1
B_2	0	0	1	1	1	1	0	0
B_3	0	1	1	0	0	1	1	0

III. The Mutual Information Perspective

A. Mutual Information for Three Independent Pages

Let p_1 denote the probability of bit B_1 being equal to 0, i.e. $p_1 = P(B_1 = 0)$. Similarly, let p_2 and p_3 denote probability of B_2 and B_3 being equal to 0, respectively. Let the 8-PAM alphabet $\mathcal{X} = \{x_0, x_1, ..., x_7\}$ and let ${\mathcal{X}_\mathcal{B}}^{(0)}$ be a subset of alphabet $\mathcal{X}$ for which $B_j = 0$ and ${\mathcal{X}_\mathcal{B}}^{(1)}$ be a subset of alphabet $\mathcal{X}$ for which $B_j = 1$, where $j = 1, 2, 3$. We use AWGN for channel model, i.e., $Z \sim \mathcal{N}(0, N)$ and i.i.d.

Mutual information rates $I(B_1; Y), I(B_2; Y)$ and $I(B_3; Y))$ for independent encoding of pages are calculated as follows:

$$I(B_j; Y) = \int_y f(B_j = 0, y) \log_2 \left(\frac{f(B_j = 0, y)}{P(B_j = 0) f(y)} \right) dy$$
$$+ \int_y f(B_j = 1, y) \log_2 \left(\frac{f(B_i = 1, y)}{P(B_i = 1) f(y)} \right) dy$$

where $P(B_j = 0) = p_j, \quad P(B_j = 1) = 1 - p_j$ and $j = 1, 2, 3$.

We can expend the terms inside the integrals as follows:

$$f(B_j = 0, y) = P(B_j = 0) f(y|B_j = 0)$$
$$= p_j f(y|B_j = 0) \tag{2}$$
$$f(B_j = 1, y) = P(B_j = 1) f(y|B_j = 1)$$
$$= (1 - p_j) f(y|B_j = 1) \tag{3}$$
$$f(y|B_j = 0) = \sum_{x_i \in {\mathcal{X}_\mathcal{B}}^{(0)}} p(x_i) f(y|x_i) \tag{4}$$
$$f(y|B_j = 1) = \sum_{x_i \in {\mathcal{X}_\mathcal{B}}^{(1)}} p(x_i) f(y|x_i) \tag{5}$$
$$f(y) = \sum_{x_i \in \mathcal{X}} p(x_i) f(y|x_i)$$
$$= \sum_{x_i \in \mathcal{X}} p(x_i) \frac{1}{\sqrt{2\pi N}} \exp \frac{-(y - x_i)^2}{2N} \tag{6}$$

Where $p(x_i) = P(X = x_i)$ or equivalently,

$$p(x_i) = P(B_1 = b_1(i)) P(B_2 = b_2(i)) P(B_3 = b_3(i)) \tag{7}$$

where i ranges from 0 to 7 and $b_1(i)$ is the most significant bit of the binary representation of i, i.e. the blue bit in Fig. 1. Similarly, $b_2(i)$ is the middle bit and $b_3(i)$ is the least significant bit.

The penalty for decoding the three pages independently (rather than jointly) is calculated as follows:

$$I(B_1, B_2, B_3; Y) - I(B_1; Y) - I(B_2; Y) - I(B_3; Y)$$
$$= I(B_1; B_2|Y) + I(B_1, B_2; B_3|Y) \geq 0 \tag{8}$$

B. Joint Vs Independent Encoding for ESEL Constellations

For ease of notation let $I_1 = I(B_1; Y)$, $I_2 = I(B_2; Y)$ and $I_3 = I(B_3; Y)$. Plots of three independent rates I_1, I_2 and I_3 as a function of signal-to-noise ratio (SNR) for ESEL constellations are given in Fig. 3. Fig. 4 illustrates the mutual information $I(X; Y) = I(B_1, B_2, B_3; Y)$ for the overall ESEL constellations and $3 \times \min_j (I_j)$ for $j = 1, 2, 3$ as a function of SNR. Observing the Fig. 4 we note that there would be an improvement if we could find a way to equalize the rates so that $3 \times \min_j (I_j)$ is closer to $I(X; Y)$.

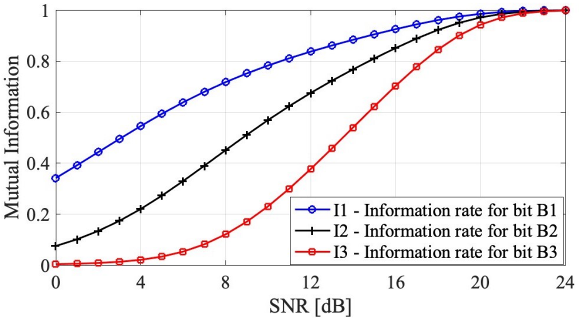

Fig. 3. Independent MI rates for Equally Spaced Equally Likely (ESEL) Constellations as a function of SNR.

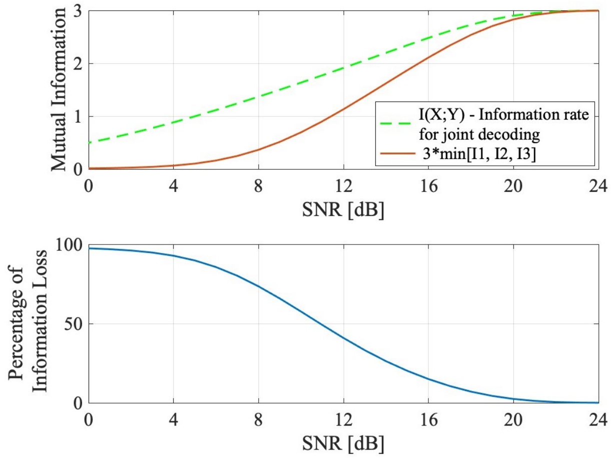

Fig. 4. Upper subplot: Comparison of Information rate $I(X; Y) = I(B_1, B_2, B_3; Y)$ for joint decoding depicted by green curve and $3 \times \min\{I_1, I_2, I_3\}$ where $I_1 = I(B_1; Y), I_2 = I(B_2; Y), I_3 = I(B_3; Y)$ for Equally Spaced Equally Likely (ESEL) points. Lower subplot: Percentage of Mutual Information loss for ESEL points with equal rate constraint.

C. Maximizing the minimum page mutual information

We formulate following optimization problem to maximize the minimum page mutual information to seek a solution for which all three page rates in TLC Flash memory are equal:

$$\max_{x \in \mathcal{X}, p_2, p_3} \min_j I(B_j; Y) \quad j = 1, 2, 3.$$
$$\text{s.t.} \quad \sum_i p(x_i) x_i^2 = E_{ESEL}, \quad i = 0, ..., M - 1$$
$$x_k = -x_{M-1-k}, \quad k = 0, 1, 2, \frac{M}{2} - 1$$
$$p(x_i) = P(B_1 = b_1(i)) P(B_2 = b_2(i)) P(B_3 = b_3(i))$$
$$\sum_i p(x_i) x_i = 0$$

With $M = 8$ for TLC Flash and $\mathcal{X} = \{x_0, x_1, ..., x_7\}$ having the ESEL values shown in Table I, E_{ESEL}, the average power for equally spaced equally likely constellation points given by:

$$E_{ESEL} = \frac{\sum_{i=0,...,M-1} x_i^2}{M} = 21 \tag{9}$$

We seek to design an alphabet $\mathcal{X}$ subject to symmetry and the power constraint $EX^2 \leq E_{ESEL}$ that maximizes the minimum MI between binary source B_j and a real valued output Y.

IV. Maximizing the minimum rate

In this section we present two algorithms for solving optimization problem in Sec. III-C. The first algorithm retains equally likely points but moves points to increase the minimum value of $I(B_j; Y)$. The second algorithm additionally adjusts the probability of the constellation points x_i under the constraint that the three bits that determine the constellation label are independent as described by (7), as required for independent encoding. Alg. 1 enumerates the steps for maximizing the minimum rate through dynamic assignment of the write level positions. Alg. 2 enumerates the steps for maximizing the minimum rate by optimizing both positions and probabilities of the write levels.

A. Dynamic Assignment of Write Level Positions

Alg. 1 maximizes the minimum rate under the constraint that the constellation points are equally likely, i.e.,

$$P(B_j = 0) = P(B_j = 1) = \frac{1}{2} \tag{10}$$

for $j = 1, 2, 3$. The details of the steps in Alg. 1 are as follows:

Initialization: Select the modulation number M = 8 for 8-PAM TLC Flash memory. Tolerance ϵ is the maximum acceptable distance from any two independent rates $I(B_j; Y)$ for $j = 1, 2, 3$. Initialize constellation (point) locations to $\mathcal{X} = \{-7, -5, -3, -1, 1, 3, 5, 7\}$ and average power $E = 21$ for ESEL constellations. In Alg. 1 we will optimize only point locations and therefore we fix the input PMFs to be $p_1 = p_2 = p_3 = 0.5$.

Iterations:

Adjust Write-level Positions $\mathcal{X}$: The process at each iteratoin is similar to Step 4 in the power-constrained AWGN dynamic-assignment Blahut-Arimmoto (DAB) algorithm introduced by Xiao in [15]. Each iteration begins by identifying the minimum MI $I(B_j; Y)$ and optimizing one symmetric pair of constellation points to increase the selected minimum MI $I(B_j; Y) = I_j(\mathcal{X}, p_1, p_2, p_3)$ calculated using MATLAB integral() function.

For each symmetric pair of constellation points we select the direction $\tilde{\mathcal{D}}^{(k)} = e_l - e_r$ with e_j being a j-th standard basis vector as described in [15]. Define l and r such that $r = |\mathcal{X}^{(k)}| - l + 1$. Thus, for the two symmetric points, l corresponds to the left location being moved and r corresponds to the right location being moved.

Algorithm 1 Dynamic Write-Level Position Assignment

```
Initialization:
$M \leftarrow 8$
$\epsilon = 10^{-4}$
$\mathcal{X}^{(1)} = \{x_1, x_2, ..., x_{|\mathcal{X}^{(1)}|}\}$    /* Initial ESEL points */
$E = \frac{1}{M} \sum_{x \in \mathcal{X}^{(1)}} x^2$
$p_1 = p_2 = p_3 = 0.5$
$i \leftarrow 1$
Iterations:
$I_j = \min(I_1, I_2, ..., I_{log_2 M})$
 1) Optimize constellation locations $\mathcal{X}$:
        a) Determine direction vector $\tilde{\mathcal{D}}^{(k)}$ to adjust $\mathcal{X}$
        b) Compute
              $\mathcal{X}^{(k+1)} = \tilde{d}\left(\mathcal{X}^{(k)} + \lambda^* \tilde{\mathcal{D}}^{(k)}\right),$    (11)
    where $\tilde{d} = diag(d_1, ..., d_M)$ ensures the power constraint,
    and the optimal movement factor $\lambda^*$ is determined as
    $\lambda^* = \arg\max_\lambda I_j \left(\tilde{d}\left(\mathcal{X}^{(k)} + \lambda^* \tilde{\mathcal{D}}^{(k)}\right), p_1^{(k)}, p_2^{(k)}, p_3^{(k)}\right)$
    where $I_j(\mathcal{X}, \mathcal{P})$ is the mutual information $I(B_j; Y)$ calculated from input PMF characterized by $\mathcal{X}$ and $p_1, p_2, p_3$.
    Note that in the maximization that yields $\lambda^*$, the resulting new values must satisfy $I_i(\mathcal{X}^{(k+1)}, \mathcal{P}) \geq I_j(\mathcal{X}^{(k)}, \mathcal{P})$ for all $i \neq j$, $0 \ \leq i \leq 2$.
    Stop If $(I_1 - I_j) \leq \epsilon, (I_2 - I_j) \leq \epsilon, \ldots (I_{log_2 |\mathcal{X}^{(1)}|}) - I_j \leq \epsilon$
    Or if maximum number of iterations is reached
    $i \leftarrow i + 1$
```

Once the direction is found we perform a line search routine (e.g. using fminbnd in MATLAB) to find the λ^* that maximizes minimum rate $I_j(\mathcal{X}, p_1, p_2, p_3)$ where $\mathcal{X}$ is a scaled version of $\mathcal{X}^{(k)} + \lambda \tilde{\mathcal{D}}^{(k)}$. The interval for line search constrained so that other two non-minimum rates stay above $I_j(\mathcal{X}, p_1, p_2, p_3)$.

To enforce the power constraint, the constellation needs to be scaled to increase power if the two points move closer to the origin or to decrease power if the two points move further from the origin. To do this, we define a diagonal matrix $\tilde{d} = diag(d_1, ..., d_M)$ to scale the remaining points to meet the power constraint as follows:

Let x_i be the i-th element of $\mathcal{X}^{(k)} + \lambda \tilde{\mathcal{D}}^{(k)}$. Since MATLAB is used as an optimization tool, we will consider the indices of constellation points to start at 1, i.e. $i = 1, ..., 8$. Similar to above, define the symmetric pair of positions being moved to be l and r such that $r = |\mathcal{X}^{(k)}| - l + 1$.

If $l = 1$ or $l = \frac{M}{2}$ the following $\tilde{d}$ satisfies the power constraint:

$$d_i \in \tilde{d} \text{ and } d_i = 1 \text{ if } i \in \{l, r\}, \text{ o.w. } d_i = \alpha \tag{12}$$

such that $P_{\tilde{d}} = E_{ESEL}$, where

$$P_{\tilde{d}} = p(x_l)x_l^2 + p(x_r)x_r^2 + \sum_{i \notin [l,r]} p(x_i)\alpha^2 x_i^2 \,. \tag{13}$$

Otherwise, only the outermost points are scaled as follows:

$$d_i \in \tilde{d} \text{ and } d_i = \alpha \text{ if } i \in \{1, M\}, \text{ o.w. } d_i = 1 \tag{14}$$

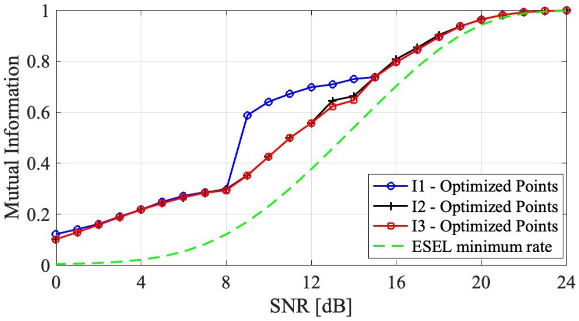

Fig. 5. Three 8-PAM independent MI rates corresponding to bits B_1, B_2 and B_3 respectively for optimized points locations using Alg. 1. Green dotted curve is minimum Information rate for Equally Spaced Equally Likely Points

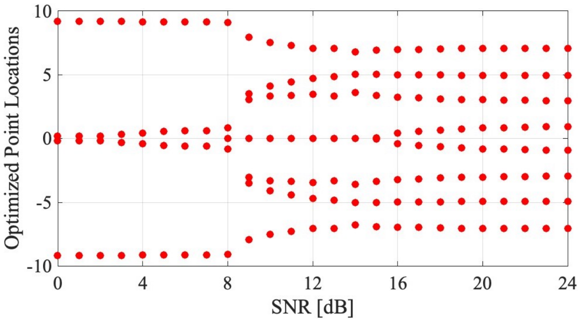

Fig. 6. 8-PAM Optimized constellations locations as a function of SNR.

such that $P_{\tilde{d}} = E_{ESEL}$, where

$$P_{\tilde{d}} = p(x_1)\alpha^2 x_1^2 + p(x_M)\alpha^2 x_M^2 + \sum_{1<i<M} p(x_i)x_i^2 \,. \tag{15}$$

Note that in this section $p(x_i) = \frac{1}{8}$ since we are only optimizing point locations and keep PMFs the same.

All three rates are calculated with updated alphabet consisting of optimized symmetric pair of points and new minimum rate is determined to continue iteration process. The algorithm stops if all three rates are within epsilon distance from each other or if we reach maximum number of iterations. For most SNR values maximum minimum rate is achieved for less than 50 iterations.

Fig. 5 shows optimized independent rates with respect to point locations as a function of SNR. For reference we also show the minimum rate for the ESEL constellation. The difference between the lowest of the optimized rates and the minimum rate for the ESEL constellation represents the an improvement in page rate for the actual system. Fig. 6 shows optimized point locations as a function of SNR.

Alg. 1 was able to achive essentially equal values for I_1, I_2 and I_3 for SNR values above 15 dB and below 8 dB. However, it was unable to achieve equal values in the region between 8 dB and 15 db. Seeking to improve on the performance Alg. 1, the next section introduces Alg. 2 which adds the capability to adapt the probability of each write level so that the constellation points are no longer equally likely.

B. Optimizing Write-Level Positions and Probabilities

This section presents Alg. 2, which optimizes both the positions and the probabilities of the write levels. The constellation point probabilities must still be symmetric, which means that p_1 remains fixed at $\frac{1}{2}$. Functionally, the main difference between Alg. 2 and Alg. 1 is that Alg. 2 contains one additional step that optimizes the PMFs corresponding to bit channels B_2 and B_3, which are defined by the parameters p_2 and p_3. Gradient descent adjusts p_2 and p_3 while scaling the constellation points to maintain the power constraint.

Let I_j be the minimum rate. Since objective function is maximizing I_j, let $f = -I_j(\mathcal{X}, p_1, p_2, p_3)$ so that we change the problem into minimizing f with respect to p_2 and p_3. With this notation, gradient descent method is as follows:

$$\triangle p := -\nabla f \tag{16}$$

$$p = p + \alpha \triangle p \,. \tag{17}$$

We used $\alpha = 0.01$ as our gradient step to optimize PMFs p_2 and p_3. Referring back to Fig.1 and Table I, note that the PMF of the write levels $P_X = p(x)$ can be directly computed from p_1, p_2 and p_3 as follows:

$$\begin{aligned} P_X = [&p_1 p_2 p_3 & (18)\\ &p_1 p_2(1-p_3) & (19)\\ &p_1(1-p_2)(1-p_3) & (20)\\ &p_1(1-p_2)p_3 & (21)\\ &(1-p_1)(1-p_2)p_3 & (22)\\ &(1-p_1)(1-p_2)(1-p_3) & (23)\\ &(1-p_1)p_2(1-p_3) & (24)\\ &(1-p_1)p_2 p_3] & (25)\end{aligned}$$

The gradient descent is complicated by the need to maintain the power constraint. Let Q_X be the previous constellation PMF used in step 1) of Alg. 2. We perform two separate steps in the power constrained gradient descent to separately adjust p_2 and p_3 as follows:

Step 1: Adjust p_2 by using (16) and (17) with $p = p_2$. Use the new p_2 to compute the new P_X according to (18)-(25). Scale the constellation points to satisfy power constraint as follows:

$$a_i = \frac{Q_X(i)}{P_X(i)}, i = 1, .., M. \tag{26}$$

$$x_i^{(k+1)} = \sqrt{a_i} x_i^{(k)} \tag{27}$$

After a gradient step in direction of p_2 is performed and points are scaled recalculate all three rates and check if they are all greater or equal to minimum rate. If this condition is not satisfied adopt initial p_2 and constellation points and move to Step 3. Otherwise, keep the solution and move to Step 3.

Step 2: Adjust p_3. This step is identical to Step 1 except applied to p_3 instead of p_2. We fix p_2 obtained from Step 2) and perform the gradient with respect to p_3 while scaling points to maintain the power constraint. Step 1 and Step 2 are repeated until there is no more improvement in maximizing

Algorithm 2 Optimizing Positions and Probabilities

```
Initialization:
$M \leftarrow 8$
$\alpha \leftarrow 0.01$    /*gradient step*/
$\epsilon = 10^{-4}$
$\mathcal{X}^{(1)} = \{x_1, x_2, ..., x_{|\mathcal{X}^{(1)}|}\}$ /* Initial ESEL points */
$E = \frac{1}{M} \sum_{x \in \mathcal{X}^{(1)}} x^2$
$p_1 = p_2 = p_3 = 0.5$
$i \leftarrow 1$
Iterations:
1) Optimize constellation locations $\mathcal{X}$:
for index = $1 : \frac{M}{2}$ do
    $I_j = min(I_1, I_2, ..., I_{log_2 M})$
        a) Determine direction vector $\tilde{\mathcal{D}}^{(k)}$ to adjust $\mathcal{X}$
        b) Compute
            $$\mathcal{X}^{(k+1)} = \tilde{d}\left(\mathcal{X}^{(k)} + \lambda^* \tilde{\mathcal{D}}^{(k)}\right), \quad (28)$$
    where $\tilde{d} = diag(d_1, ..., d_M)$ ensures the power constraint,
    and the optimal movement factor $\lambda^*$ is determined as
    $\lambda^* = \arg\max_\lambda I_j\left(\tilde{d}\left(\mathcal{X}^{(k)} + \lambda^* \tilde{\mathcal{D}}^{(k)}\right), p_1^{(k)}, p_2^{(k)}, p_3^{(k)}\right)$
    where $I_j(\mathcal{X}, \mathcal{P})$ is the mutual information $I(B_j; Y)$
    calculated from input PMF characterized by $\mathcal{X}$ and
    $p_1, p_2, p_3$. Note that in the maximization that yields $\lambda^*$,
    the resulting new values must satisfy $I_i(\mathcal{X}^{(k+1)}, \mathcal{P}) \geq$
    $I_j(\mathcal{X}^{(k)}, \mathcal{P})$ for all $i \neq j$, $0 \leq i \leq 2$.
end for
$I = \min(I_1, I_2, ..., I_{log_2 M})$
2) Optimize PMFs $p_2$ and $p_3$:
    Use gradient descent to optimize $p_2$ and $p_3$ while
    satisfying power constraint by scaling constellations.
Stop If $(I_1 - I_2) \leq \epsilon$  &&...&&  $(I_1 - I_{log_2|\mathcal{X}^{(1)}|}) \leq \epsilon$
Or if maximum number of iterations is reached
$i \leftarrow i + 1$
```

the minimum rate.

Fig. 7 shows the independent rates achieved by the optimized point locations and input PMFs obtained using Alg. 2. Fig. 8 shows optimized point positions and input PMFs resulting from using Alg. 2 to achieve independent rates in Fig. 7. In Fig. 9, the joint mutual information resulting from optimized points and PMFs in Alg. 2 is compared to MI of DAB optimized input PMFs given in [15] as well as to 3 times minimum information rate and sum of independent rates for optimized points and PMFs. Fig. 10 shows three losses for flash system:

1) Shaping loss: $I(X;Y)_{DAB} - I(B_1, B_2, B_3; Y)$
2) Independent encoding loss: $I(B_1, B_2, B_3; Y) - I(B_1 : Y) - I(B_2; Y) - I(B_3; Y)$
3) Equal rate constraint loss: $I(B_1; Y) + I(B_2; Y) + I(B_3; Y) - 3 * \min_j\{I(B_j : Y)\}$

The results in Fig. 10 show that equal rate constraint loss is negligible for nearly all SNRs. Independent encoding loss is higher for very low SNRs but zero after 15 dB. Shaping loss is evident mostly from 10 to 22 dB with highest value slightly less than 0.15 bits. Shaping loss could be improved by explicitly adding a step to improve the joint mutual information.

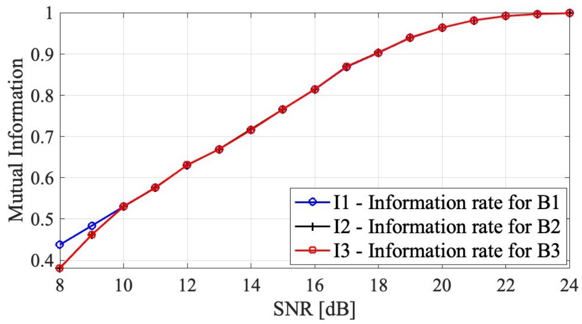

Fig. 7. 8-PAM Independent rates for optimized constellation positions and input PMFs.

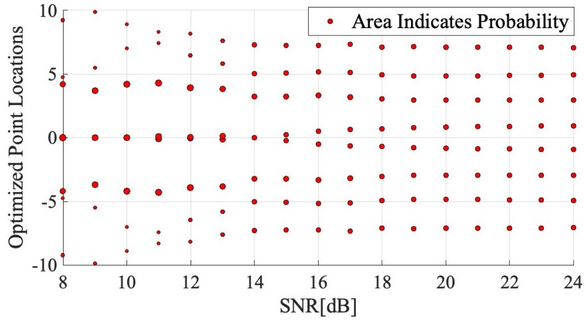

Fig. 8. 8-PAM Optimized constellations locations using Alg. 2 as a function of SNR.

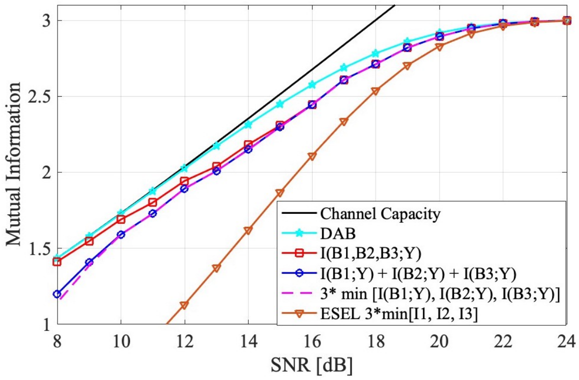

Fig. 9. Joint mutual information rate $I(B_1, B_2, B_3; Y)$ for optimized points and PMFs is compared to the sum of independent rates, 3 times minimum independent rate and joint MI for DAB optimized input PMFs. Equal rate constraint loss is nearly zero for all SNRs. Independent encoding loss is negligible for $SNR = 15$ dB and higher. The most evident loss comes from shaping.

V. Conclusion

This paper focuses on the practical scenario of independent encoding of Flash under the constraint that all pages have

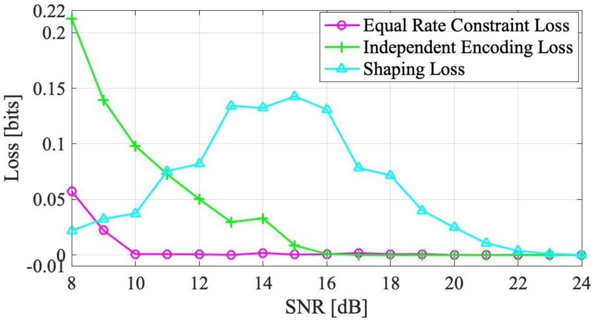

Fig. 10. Equal Rate Constraint Loss: $I(B1 : Y) + I(B2;Y) + I(B3;Y) - 3 * min[I(B1 : Y), I(B2;Y), I(B3;Y)]$, where $I(B1 : Y), I(B2;Y), I(B3;Y)$ are information rates for optimized constellations and input PMFs. Independent encoding loss: $I(B1, B2, B3; Y) - (I(B1 : Y) + I(B2;Y) + I(B3;Y))$ given in Eqn. (9). Shaping Loss: Difference between DAB Mutual Information Rate $I(X;Y)$ given in [15] and $I(B1, B2, B3; Y)$ for optimized points and PMFs.

to store the same amount of information. By optimizing the positions of the write levels in the constellation and input probabilities to maximize a mutual information rate of the bit channel with worst information rate, we were able to significantly improve the minimum page rate. The independent encoding loss and equal rate constraint loss are negligible (zero) for operational SNRs. Shaping loss is the most significant loss in an interesting range between 10 and 22 dB, with the largest loss approaching 0.15 bits. This loss may be reduced by an additional optimization step, which is the subject of future research. We note that 8-PAM constellations with points that are not equally likely can be supported by coded modulation techniques such as probabilistic amplitude shaping [16]–[19].

Acknowledgment

We thank Derek Xiao, an author of paper referenced in [15], for providing DAB algorithm MATLAB code and DAB optimized PMFs and points needed for DAB rate in Fig. 9.

References

[1] L. Dolecek and Y. Cassuto, "Channel coding for nonvolatile memory technologies: Theoretical advances and practical considerations," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1705–1724, 2017.

[2] L. M. Grupp, A. M. Caulfield, J. Coburn, S. Swanson, E. Yaakobi, P. H. Siegel, and J. K. Wolf, "Characterizing flash memory: Anomalies, observations, and applications," in *2009 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 24–33.

[3] T.-Y. Chen, A. R. Williamson, and R. D. Wesel, "Increasing flash memory lifetime by dynamic voltage allocation for constant mutual information," in *2014 Information Theory and Applications Workshop (ITA)*, 2014, pp. 1–5.

[4] Q. Li, A. Jiang, and E. F. Haratsch, "Noise modeling and capacity analysis for nand flash memories," in *2014 IEEE International Symposium on Information Theory*, 2014, pp. 2262–2266.

[5] C. Schoeny, F. Sala, and L. Dolecek, "Analysis and coding schemes for the flash normal-laplace mixture channel," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 2101–2105.

[6] J. Wang, K. Vakilinia, T.-Y. Chen, T. Courtade, G. Dong, T. Zhang, H. Shankar, and R. Wesel, "Enhanced precision through multiple reads for ldpc decoding in flash memories," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 5, pp. 880–891, 2014.

[7] J. Wang, T. Courtade, H. Shankar, and R. D. Wesel, "Soft information for ldpc decoding in flash: Mutual-information optimized quantization," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, 2011, pp. 1–6.

[8] C. A. Aslam, Y. L. Guan, and K. Cai, "Dynamic write-level and read-level signal design for mlc nand flash memory," in *2014 9th International Symposium on Communication Systems, Networks Digital Sign (CSNDSP)*, 2014, pp. 336–341.

[9] C. Duangthong, W. Phakphisut, and P. Supnithi, "Capacity enhancement of asymmetric multi-level cell (mlc) nand flash memory using write voltage optimization," in *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2019, pp. 1–4.

[10] ——, "Read voltage optimization in mlc nand flash memory via the density evolution," in *2019 26th International Conference on Telecommunications (ICT)*, 2019, pp. 361–365.

[11] F. J. C. Romero and B. M. Kurkoski, "Ldpc decoding mappings that maximize mutual information," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 9, pp. 2391–2401, 2016.

[12] N. Wong, E. Liang, H. Wang, S. V. S. Ranganathan, and R. D. Wesel, "Decoding flash memory with progressive reads and independent vs. joint encoding of bits in a cell," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[13] P. Huang, P. H. Siegel, and E. Yaakobi, "Performance of multilevel flash memories with different binary labelings: A multi-user perspective," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 9, pp. 2336–2353, 2016.

[14] D. N. Bailon, S. Shavgulidze, and J. Freudenberger, "Cell-wise encoding and decoding for tlc flash memories," in *2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin)*, 2022, pp. 1–6.

[15] D. Xiao, L. Wang, D. Song, and R. D. Wesel, "Finite-support capacity-approaching distributions for awgn channels," in *2020 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–5.

[16] L. Wang, D. Song, F. Areces, and R. D. Wesel, "Achieving short-blocklength rcu bound via crc list decoding of tcm with probabilistic shaping," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 2906–2911.

[17] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. on comm.*, vol. 63, no. 12, pp. 4651–4665, 2015.

[18] G. Böcherer, P. Schulte, and F. Steiner, "Probabilistic shaping and forward error correction for fiber-optic communication systems," *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 230–244, 2019.

[19] M. C. Coşkun, G. Durisi, T. Jerkovits, G. Liva, W. Ryan, B. Stein, and F. Steiner, "Efficient error-correcting codes in the short blocklength regime," *Physical Communication*, vol. 34, pp. 66–79, 2019.