Optimizing the Write Fidelity of MRAMs

Yongjune Kim*, Yoocharn Jeon*, Cyril Guyot*, and Yuval Cassuto*†
*Western Digital Research, Milpitas, CA, USA
Email: {yongjune.kim, yoocharn.jeon, cyril.guyot}@wdc.com
†Viterbi Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel
Email: ycassuto@ee.technion.ac.il

***Abstract*—Magnetic random-access memory (MRAM) is a promising memory technology due to its high density, nonvolatility, and high endurance. However, achieving high memory fidelity incurs significant write-energy costs, which should be reduced for the large-scale deployment of MRAMs. In this paper, we formulate an optimization problem to maximize the memory fidelity given energy constraints, and propose a biconvex optimization approach to solve it. The basic idea is to allocate non-uniform write pulses depending on the importance of each bit position. We consider the mean squared error (MSE) as a fidelity metric and propose an iterative *water-filling* algorithm to minimize the MSE. Although the iterative algorithm does not guarantee the global optimality, we can choose a proper starting point that decreases the MSE exponentially and guarantees fast convergence. For an 8-bit accessed word, the proposed algorithm reduces the MSE by a factor of 21.**

I. Introduction

Magnetic random access memory (MRAM) is a nonvolatile memory technology that has a potential to combine the speed of static RAM (SRAM) and the density of dynamic RAM (DRAM). Furthermore, MRAM technology is attractive since it provides high endurance and complementary metal-oxide-semiconductor (CMOS) compatibility [1], [2].

In spite of its attractive features, one of the main challenges is the high energy consumption to write information *reliably* in the memory element [1]–[3]. In an MRAM device, a memory state “1” or “0” is determined by the magnetic moment orientation of the memory element [1]. Switching the magnetic moment orientation requires high write current, which introduces write errors when the energy budget is limited [2]. In addition, high current injection through the tunneling barriers incurs a severe stress and leads to breakdown, which degrades the endurance of MRAM cells [3], [4]. Hence, one of the key directions of MRAM research has been toward providing reliable switching with limited energy cost. At the device level, new materials [5], [6] or new switching mechanisms [7], [8] have been explored. Several architectural techniques to reduce write energy can be found in [3], [9], [10].

However, prior efforts have not considered the differential importance of each bit position in error tolerant applications such as signal processing and machine learning (ML) tasks. In these applications, the impact of bit errors depends on bit position, i.e., most significant bits (MSBs) are more important than least significant bits (LSBs) [11], [12]. This differential importance has been leveraged to effectively optimize energy in major memory technologies such as SRAMs [13]–[16] and DRAMs [17], [18].

In this paper, we provide a *principled* approach to improving MRAM’s write fidelity. In error tolerant applications, the mean squared error (MSE) is a more meaningful fidelity metric than the write failure probability (or bit error rate). We formulate a *biconvex optimization* problem to minimize the MSE for a given write energy constraint. Since the write energy and the MSE depend on the write current and the write pulse duration, we attempt to optimize both parameters by solving the biconvex problem.

A biconvex problem is an optimization problem where the objective function and the constraint set are biconvex [19]. A common algorithm for solving biconvex problems is *alternate convex search (ACS)*, which updates each variable by fixing another and solving the corresponding convex problem in an iterative manner [20]. We propose an iterative algorithm based on ACS to optimize the write current and the write pulse duration. In addition, we show that the proposed iterative algorithm converges and the convergence speed can be very fast by choosing a proper starting point.

In general, ACS cannot guarantee the global optimal solution since biconvex problems may have a large number of local minima [19]. However, we prove that the proposed iterative algorithm can reduce the MSE exponentially by choosing a proper starting point. Furthermore, we show that this starting point guarantees the fastest convergence. We derive analytic expressions of the optimal solutions for each iteration. Since each iteration of the algorithm corresponds to solving convex problems, we rely on the Karush-Kuhn-Tucker (KKT) conditions to derive the optimal solutions. We also provide water-filling interpretations for each iteration.

Prior optimization studies on voltage swing of SRAMs [15], [16] and refresh operations of DRAMs [18] are similar in spirit, viz. minimizing the MSE for given resource constraints. However, the MRAM write optimization of this work is *non-convex* whereas the formulated problems in [15], [18] are convex. Hence, we propose the iterative algorithm and analyze convergence and improvement of the optimized MSE. To the best of our knowledge, our work is the first principled approach to optimization of write pulse parameters of MRAMs.

The rest of this paper is organized as follows. Section II explains the basics of MRAM and the challenges of high write energy consumption. Section III introduces the optimization metrics for MRAM write operations. Section IV formulates optimization problems and provides the iterative algorithm based on ACS. Section V provides theoretical analysis on

978-1-7281-6432-8/20/$31.00 ©2020 IEEE

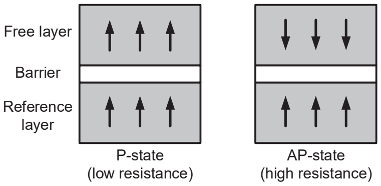

Fig. 1. P state and AP state of MTJ MRAM devices.

convergence and MSE reduction. Section VI gives numerical results and Section VII concludes.

II. Basic Principles of MRAMs

MRAM cells store information by controlling bistable magnetization of ferromagnetic material and retrieve information by sensing resistance of magnetic tunnel junctions (MTJs). An MTJ device consists of two ferromagnetic layers of reference layer (RL) and free layer (FL), separated by a very thin tunneling barrier. Since RL has a very stable magnetization, it maintains the magnetization throughout all operations. On the other hand, FL can be switched between two stable magnetization states by a moderate stimulus. The resistance of an MTJ depends on the relative orientation of the FL magnetization with respect to that of the RL (see Fig. 1). If the magnetizations of FL and RL are in the same direction (parallel- or P-state), then the corresponding resistance is low. The opposite direction (antiparallel- or AP-state) results in high resistance. The difference in tunneling currents between a P-state (low resistance) and a AP-state (high resistance) is utilized to encode binary data [1], [2].

Writing information into an MTJ is performed by driving a sufficient current through it. Depending on the current's direction, one can flip the magnetization of the FL into P- or AP-state. If a current flows from FL to RL (electrons from RL to FL), electrons are spin-polarized along the magnetization of RL while passing through the layer. The electrons transmitted from the RL interact and exchange the magnetic moments with ones in the FL. If the MTJ is in the AP-state and the current is sufficiently high, then the magnetization orientation is flipped to P-state. When the current is reversed, incoming electrons are polarized along the magnetization of FL. Since the RL's magnetization is parallel to the FL, the majority of the electrons tunnel the barrier while the minority that have antiparallel magnetizations are reflected. Because of this selective tunneling, the antiparallel spins are accumulated in the FL. If the enriched antiparallel spin dominates the FL, it flips the magnetization of the FL into the AP-state.

The magnetization switching between P state and AP state is not deterministic. The write (switching) failure probability depends on the magnitude and the duration of the write current pulse as follows [4, Eq. (26)]:

$$p(i,t) = 1 - \exp\left(-\frac{\Delta\pi^2(i-1)}{4\left\{i\exp(2(i-1)t)-1\right\}}\right), \tag{1}$$

where Δ denotes the thermal stability factor. The normalized current i is given by $i = \frac{I}{I_c}$ where I denotes the actual write current and I_c is the critical current. The normalized duration is given by $t = \frac{T}{T_c}$ where T denotes the actual write duration and T_c is the characteristic relaxation time. Note that Δ, I_c, and T_c are fabrication parameters [4], [21].

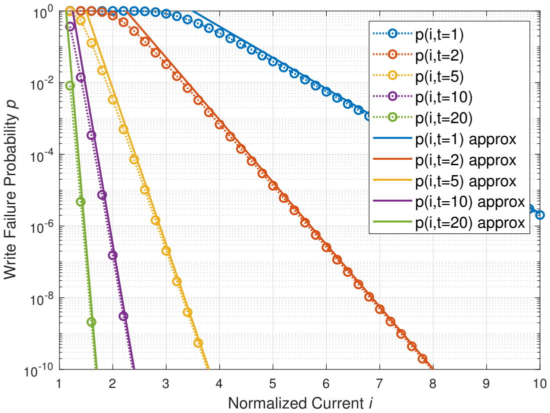

Fig. 2. Comparison of the write failure probability (1) and its approximation (2) ($\Delta = 60$ as in [4, Fig. 13]).

To ensure a low write failure probability, we should control the write current magnitude or the duration judiciously. A longer write duration may lower the write failure probability at the expense of longer write latency and higher energy consumption. Instead of increasing the write duration, we can adopt higher write current, which increases the write energy and the risk of dielectric breakdown of the MTJ.

The MRAM cells are arranged in arrays and each of the cells is selectively connected to the read/write circuits through a selector to access the data. Because of the write current requirement, most crossbar MRAM architectures allow *only one cell can be accessed at a time in each subarray*. Multiple subarrays are operated in parallel to match the required data bandwidth. This MRAM architecture provides an opportunity to write each bit in different conditions.

III. Metrics for MRAM Write Operations

The write failure probability expression of (1) is too complicated to formulate an optimization problem. Fortunately, we can use the following approximation instead of (1):

$$p(i,t) \approx c\exp\left(-2(i-1)t\right). \tag{2}$$

where $c = \frac{\Delta\pi^2}{4}$. This is a slightly modified approximation of [4, Eq. (27)] so as to formulate a optimization problem. Fig. 2 shows that the approximated write failure probability (2) is very close to (1), especially for lower p. The write failure probability can be controlled by the normalized current i and the normalized write duration t. The fabrication parameters such as Δ does not affect the optimized i and t.

The normalized energy for writing a single bit is given by

$$\mathsf{E}(i,t) = i^2 t. \tag{3}$$

As shown in (2) and (3), the write current i and the write duration t are key knobs to control the trade-off between write failure probability and the write energy. If we allocate different write currents and durations depending on the importance of each bit position, then the corresponding current and duration assignments are given by

$$\mathbf{i} = (i_0, \ldots, i_{B-1}), \quad \mathbf{t} = (t_0, \ldots, t_{B-1}) \tag{4}$$

where i_0 and t_0 define the write pulse for least significant bit (LSB) and i_{B-1} and t_{B-1} are the write pulse parameters for most significant bit (MSB).

We define metrics for energy, latency, and fidelity for writing a B-bit word.

Definition 1 (Normalized Energy): The normalized energy of writing a B-bit word is $\mathsf{E}(\mathbf{i}, \mathbf{t}) = \sum_{b=0}^{B-1} i_b^2 t_b$.

Definition 2 (Normalized Latency): The normalized latency of writing a B-bit word depends on the maximum write duration among $\mathbf{t} = (t_0, \ldots, t_{B-1})$, i.e., $\mathsf{L}(\mathbf{t}) = \max\{t_0, \ldots, t_{B-1}\}$.

Note that $\mathsf{E}(\mathbf{i}, \mathbf{t})$ and $\mathsf{L}(\mathbf{t})$ are resource metrics. As a fidelity metric, we consider mean squared error (MSE).

Definition 3: The MSE of B-bit words is given by

$$\mathsf{MSE}(\mathbf{i}, \mathbf{t}) = \sum_{b=0}^{B-1} 4^b p(i_b, t_b) \tag{5}$$

where $p(i_b, t_b)$ is given in (2) and the weight 4^b represents the differential importance of each bit position. The derivation can be found in [14], [15].

IV. Optimizing Parameters of Write Operations

In this section, we investigate optimization of write operation parameters. First, the optimized current and duration for a single bit will be discussed and then we provide biconvex optimization problems for a B-bit word.

A. Optimized Parameters for Single Bit Write

First, we note that the normalized current should be greater than 1 for a successful write in (2). It shows that the write current should be greater than the critical current (i.e., $I > I_c$) so as to switch the direction of magnetization [4], [21]. Then, we can formulate the following optimization problem for single-bit (also multi-bit uniform) write:

$$\begin{aligned} \underset{i,t}{\text{minimize}} \quad & p(i,t) = c\exp\left(-2(i-1)t\right) \\ \text{subject to} \quad & i^2 t \le \mathcal{E}, \quad i \ge 1+\epsilon, \quad t \ge 0, \end{aligned} \tag{6}$$

where $\mathcal{E}$ is a constant corresponding to the given write energy budget. We introduce $\epsilon > 0$ to guarantee $i > 1$. This optimization problem is equivalent to

$$\begin{aligned} \underset{i,t}{\text{maximize}} \quad & (i-1)t \\ \text{subject to} \quad & i^2 t \le \mathcal{E}, \quad i \ge 1+\epsilon, \quad t \ge 0. \end{aligned} \tag{7}$$

Note that the objective function $(i-1)t$ is not concave. However, we can obtain the optimal i^* and t^* as follows.

Lemma 4: The optimized current and duration for single bit write are $i^* = 2$ and $t^* = \frac{\mathcal{E}}{4}$, respectively. The corresponding write failure probability is given by

$$p(i^*, t^*) = c\exp\left(-\frac{\mathcal{E}}{2}\right). \tag{8}$$

Proof: The proof is given in [22]. ■

Note that the write failure probability is an exponentially decaying function of $\mathcal{E}$.

B. Optimized Parameters for B-bit Word Writes

We formulate an optimization problem to determine the currents and durations. For a given write energy constraint, we seek to minimize MSE as follows.

$$\begin{aligned} \underset{\mathbf{i},\mathbf{t}}{\text{minimize}} \quad & \sum_{b=0}^{B-1} 4^b \exp(-2(i_b-1)t_b) \\ \text{subject to} \quad & \sum_{b=0}^{B-1} i_b^2 t_b \le \mathcal{E} \\ & i_b \ge 1+\epsilon, \quad t_b \ge 0, \quad b = 0, \ldots, B-1 \end{aligned} \tag{9}$$

We may include additional constraints such as $\mathsf{L}(\mathbf{t}) \le \delta$ to guarantee a required write speed performance. Note that $\mathsf{L}(\mathbf{t}) \le \delta$ is a convex constraint.

Although the optimization problem (9) is not convex, we show that (9) is a *biconvex* optimization problem. Hence, we can find suboptimal solutions via effective algorithms such as *alternate convex search (ACS)* [19].

Definition 5 (Biconvex Set [19]): Let $S \subseteq X \times Y$ where $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ denote two non-empty and convex sets. The set S is defined as a *biconvex set* on $X \times Y$, if for every fixed $\mathbf{x} \in X$, $S_{\mathbf{x}} \triangleq \{\mathbf{y} \in Y \mid (\mathbf{x}, \mathbf{y}) \in S\}$ is a convex set in Y and for every fixed $\mathbf{y} \in Y$, $S_{\mathbf{y}} \triangleq \{\mathbf{x} \in X \mid (\mathbf{x}, \mathbf{y}) \in S\}$ is a convex set in X.

Definition 6 (Biconvex Function [19]): A function $f : S \to \mathbb{R}$ is defined as a *biconvex function* on S, if for every fixed $\mathbf{x} \in X$, $f_{\mathbf{x}}(\cdot) = f(\mathbf{x}, \cdot) : S_{\mathbf{x}} \to \mathbb{R}$ is a convex function on $S_{\mathbf{x}}$, and for every fixed $\mathbf{y} \in Y$, $f_{\mathbf{y}}(\cdot) = f(\cdot, \mathbf{y}) : S_{\mathbf{y}} \to \mathbb{R}$ is a convex function on $S_{\mathbf{y}}$.

Definition 7 (Biconvex Problem [19]): An optimization problem of the following form:

$$\text{minimize}\,\{f(\mathbf{x}, \mathbf{y}) \mid (\mathbf{x}, \mathbf{y}) \in S\} \tag{10}$$

is defined as a *biconvex problem*, if the feasible set S is biconvex on $X \times Y$ and the objective function f is biconvex on S.

Theorem 8: The optimization problem (9) is biconvex.

Proof: First, we show that $\sum_{b=0}^{B-1} i_b^2 t_b \le \mathcal{E}$ is a biconvex set. Note that $i_b^2 t_b$ is a convex function of i_b for every fixed $t_b \ge 0$. In addition, $i_b^2 t_b$ is a convex function for every fixed $i_b \ge 1+\epsilon$. Hence, $\sum_{b=0}^{B-1} i_b^2 t_b \le \mathcal{E}$ is a biconvex set.

It is clear that $\exp(-2(i_b-1)t_b)$ is a biconvex function of i_b and t_b. Since the positive weight 4^b preserves convexity, the objective function is biconvex. ■

Since (9) is a biconvex problem, ACS can effectively find a suboptimal solution [19], [20]. It alternatively updates variables by fixing one of them and solving the corresponding convex optimization problem. We propose Algorithm 1 to optimize the write current $\mathbf{i}$ and the write duration $\mathbf{t}$ of the biconvex optimization problem (9) by using ACS.

Algorithm 1 ACS algorithm to solve (9)

```
1: Choose a starting point $\mathbf{i}^{(0)}$ from the feasible set $S$ and set $k = 0$.
2: For fixed $\mathbf{i}^{(k)}$, find $\mathbf{t}^{(k+1)}$ by solving the following convex problem:
       $\underset{\mathbf{t}}{\text{minimize}} \quad \sum_{b=0}^{B-1} 4^b \exp\left(-2\left(i_b^{(k)} - 1\right) t_b\right)$
       $\text{subject to} \quad \sum_{b=0}^{B-1} (i_b^{(k)})^2 t_b \le \mathcal{E}$                    (11)
       $t_b \ge 0, \quad b = 0, \ldots, B-1$
3: For fixed $\mathbf{t}^{(k+1)}$, find $\mathbf{i}^{(k+1)}$ by solving the following convex problem.
       $\underset{\mathbf{i}}{\text{minimize}} \quad \sum_{b=0}^{B-1} 4^b \exp\left(-2(i_b - 1) t_b^{(k+1)}\right)$
       $\text{subject to} \quad \sum_{b=0}^{B-1} i_b^2 t_b^{(k+1)} \le \mathcal{E}$                    (12)
       $i_b \ge 1 + \epsilon, \quad b = 0, \ldots, B-1$
4: If the point $(\mathbf{i}^{(k+1)}, \mathbf{t}^{(k+1)})$ satisfies a stopping criterion, then stop. Otherwise, $k := k + 1$ and go back to line 2.
```

Remark 9 (Starting Point): Since biconvex optimization problems may have a large number of local minima [19], a starting point $\mathbf{i}^{(0)}$ can affect the final solution. We can choose $\mathbf{i}^{(0)} = (2, \ldots, 2)$ as a starting point, which minimizes the uniform write failure probability (see Lemma 4). In Corollary 16, we show that this starting point guarantees the fastest convergence.

Remark 10 (Stopping Criterion [19]): There are several ways to define the stopping criterion in Algorithm 1. For example, we can consider the absolute values of the differences between $(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})$ and $(\mathbf{i}^{(k+1)}, \mathbf{t}^{(k+1)})$ or the difference between $\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})$ and $\mathsf{MSE}(\mathbf{i}^{(k+1)}, \mathbf{t}^{(k+1)})$. Alternatively, we can set a maximum number of iterations.

V. Analysis of Alternate Convex Search for MRAM Write Parameters

A. Optimal Solutions for Each Iteration

In this subsection, we present the optimal solutions for (11) and (12). Since these problems are convex, we exploit the structure of the problems to derive the optimal solutions analytically using the KKT conditions.

Theorem 11: For fixed $\mathbf{i}^{(k)} = \mathbf{i}$, the optimal $\mathbf{t}^{(k+1)} = \mathbf{t}^*$ of (11) is given by

$$t_b^* = \begin{cases} 0, & \text{if } \nu \ge \frac{2 \cdot 4^b (i_b - 1)}{i_b^2}; \\ \frac{\log\left(\frac{1}{\nu} \cdot \frac{2 \cdot 4^b (i_b - 1)}{i_b^2}\right)}{2(i_b - 1)}, & \text{otherwise} \end{cases} \tag{13}$$

where ν is a dual variable of corresponding KKT conditions. Note that ν depends on the energy budget $\mathcal{E}$.

Proof: The proof is given in [22]. ■

Theorem 12: For fixed $\mathbf{t}^{(k+1)} = \mathbf{t}$, the optimal $\mathbf{i}^{(k+1)} = \mathbf{i}^*$ of (12) is given by

$$i_b^* = \begin{cases} 1 + \epsilon, & \text{if } \nu' \ge \frac{4^b}{1+\epsilon} e^{-2 t_b \epsilon}; \\ \frac{1}{2 t_b} W\left(\frac{2 \cdot 4^b t_b e^{2 t_b}}{\nu'}\right), & \text{otherwise} \end{cases} \tag{14}$$

where ν' is a dual variable. Also, $W(\cdot)$ denotes the *Lambert W function* (i.e., the inverse function of $f(x) = x e^x$) [23].

Proof: The proof is given in [22]. ■

Remark 13: The solutions of (13) and (14) can be interpreted as water-filling (see [22]). Each bit position can be regarded as an individual channel among B parallel channels as in [15], [16]. The ground levels depend on the importance of bit positions; hence, larger current or longer duration are assigned to more significant bit positions.

B. Convergence of MSE

We show that Algorithm 1 guarantees convergence to a locally optimal MSE. The converged MSE depends on a starting point.

Lemma 14: The sequence $\left\{\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})\right\}_{k \in \mathbb{N}}$ obtained by Algorithm 1 is monotonically decreasing, i.e., $\mathsf{MSE}(\mathbf{i}^{(k+1)}, \mathbf{t}^{(k+1)}) \le \mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})$ for all $k \in \mathbb{N}$.

Proof: Note that $\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k+1)}) \le \mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})$ and $\mathsf{MSE}(\mathbf{i}^{(k+1)}, \mathbf{t}^{(k+1)}) \le \mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k+1)})$ because of (11) and (12), respectively. Hence, $\mathsf{MSE}(\mathbf{i}^{(k+1)}, \mathbf{t}^{(k+1)}) \le \mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})$. ■

Theorem 15: The sequence $\left\{\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})\right\}_{k \in \mathbb{N}}$ obtained by Algorithm 1 converges monotonically.

Proof: It is clear that $\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)}) \ge 0$ for all $k \in \mathbb{N}$ by (2) and (5). Then, $\left\{\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})\right\}_{k \in \mathbb{N}}$ is monotonically decreasing and bounded below, $\left\{\mathsf{MSE}(\mathbf{i}^{(k)}, \mathbf{t}^{(k)})\right\}_{k \in \mathbb{N}}$ converges because of *monotone convergence theorem*. ■

Corollary 16: By setting $\mathbf{i}^{(0)} = (2, \ldots, 2)$, we obtain

$$\lim_{k \to \infty} \left(\mathbf{i}^{(k)}, \mathbf{t}^{(k)}\right) = \left(\mathbf{i}^{(0)}, \mathbf{t}^{(1)}\right), \tag{15}$$

if $t_b^{(1)} \ne 0$ for all $b \in [0, B-1]$.

Proof: The proof is given in [22]. ■

Corollary 16 means that the starting point $\mathbf{i}^{(0)} = (2, \ldots, 2)$ guarantees the *fastest convergence*.

C. Starting Point of $\mathbf{i}^{(0)} = (2, \ldots, 2)$

In this subsection, we show that $\mathbf{i}^{(0)} = (2, \ldots, 2)$ is a good starting point, in the sense that it reduces the MSE exponentially with B.

Suppose that the starting point is $\mathbf{i}^{(0)} = (2, \ldots, 2)$. By Theorem 11 and Corollary 16, Algorithm 1 provides the following optimized write durations $\mathbf{t}^{(1)} = \widetilde{\mathbf{t}} = (\widetilde{t}_0, \ldots, \widetilde{t}_{B-1})$ where

$$\widetilde{t}_b = \begin{cases} 0, & \text{if } \nu \ge \frac{4^b}{2}; \\ \frac{1}{2} \log\left(\frac{1}{\nu} \cdot \frac{4^b}{2}\right), & \text{otherwise.} \end{cases} \tag{16}$$

Lemma 17: If $\mathcal{E} > 2B(B-1)\log 2$, then $\widetilde{t}_b > 0$ for all $b \in [0, B-1]$ and

$$\widetilde{t}_b = \frac{\mathcal{E}}{4B} + \left(b - \frac{B-1}{2}\right) \cdot \log 2. \tag{17}$$

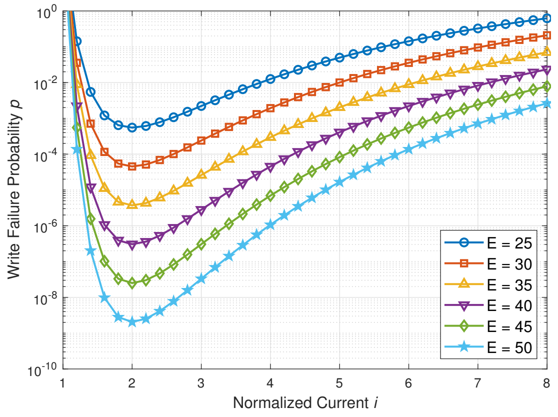

Fig. 3. Normalized write current to minimize the write failure probability (see Lemma 4) for several energy constraints.

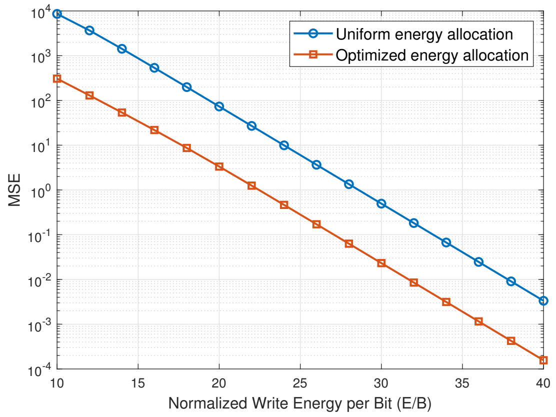

Fig. 4. Comparison of the conventional uniform energy allocation and the optimized energy allocation by Algorithm 1 ($B = 8$).

Proof: The proof is given in [22]. ■

Theorem 18: If $\mathcal{E} > 2B(B-1)\log 2$, then the MSE reduction ratio by Algorithm 1 is given by

$$\gamma = \frac{\mathsf{MSE}\left(\mathbf{i}^{(0)}, \widetilde{\mathbf{t}}\right)}{\mathsf{MSE}\left(\mathbf{i}^{(0)}, \mathbf{t}^{(0)}\right)} = \frac{3B}{2} \cdot \frac{2^B}{4^B - 1} \approx \frac{3B}{2} \cdot 2^{-B} \tag{18}$$

where $\mathsf{MSE}(\mathbf{i}^{(0)}, \widetilde{\mathbf{t}})$ (i.e., the optimized MSE by Algorithm 1) is given by $\mathsf{MSE}\left(\mathbf{i}^{(0)}, \widetilde{\mathbf{t}}\right) = c \cdot \frac{B}{2} \cdot 2^B \exp\left(-\frac{\mathcal{E}}{2B}\right)$ where the optimized $\widetilde{\mathbf{t}}$ is given by (16). In addition, $\mathsf{MSE}(\mathbf{i}^{(0)}, \mathbf{t}^{(0)})$ (i.e., the MSE by uniform energy allocation) is given by $\mathsf{MSE}\left(\mathbf{i}^{(0)}, \mathbf{t}^{(0)}\right) = c \cdot \frac{4^{B}-1}{3} \exp\left(-\frac{\mathcal{E}}{2B}\right)$ where $\mathbf{t}^{(0)}$ is the uniform value to satisfy the energy constraint (i.e., $\mathbf{t}^{(0)} = \frac{\mathcal{E}}{4B} \cdot (1, \ldots, 1)$).

Proof: The proof is given in [22]. ■

$\mathsf{MSE}\left(\mathbf{i}^{(0)}, \mathbf{t}^{(0)}\right)$ is the MSE corresponding to the parameters minimizing the write failure probability (see Lemma 4).

Remark 19: By setting $\mathbf{i}^{(0)} = (2, \ldots, 2)$, Algorithm 1 reduces the MSE exponentially with B, compared to the parameters optimized for write failure probability. Although we cannot guarantee that $(\mathbf{i}^{(0)}, \mathbf{t}^{(1)} = \widetilde{\mathbf{t}})$ is globally optimal, $(\mathbf{i}^{(0)}, \mathbf{t}^{(1)})$ decreases the MSE exponentially by solving (11) once (see Corollary 16). Furthermore, the solution of (11) can be efficiently computed by Lemma 17.

VI. Numerical Results

We evaluate the solutions to optimize the write failure probability for single bits as well as the MSE for B-bit words. The critical current I_c and the characteristic relaxation time T_c do not affect the numerical results because the normalized values $i = \frac{I}{I_c}$ and $t = \frac{T}{T_c}$ are considered. As in [4], we set $\Delta = 60$ for the thermal stability factor.

Fig. 3 shows that $i^* = 2$ and $t^* = \frac{\mathcal{E}}{4}$ minimize the write failure probability as proved in Lemma 4. The corresponding minimal write failure probability decreases exponentially with the write energy as shown in (8).

Fig. 4 shows numerical results by solving (9). Fig. 4 compares the MSEs of the uniform write energy allocation and the optimized energy allocation by Algorithm 1. We set a starting point $\mathbf{i}^{(0)} = (2, \ldots, 2)$. As shown in Theorem 18, the MSE reduction ratio is $\gamma \approx \frac{3B}{2} \cdot 2^{-B} = 0.0469$ for $B = 8$.

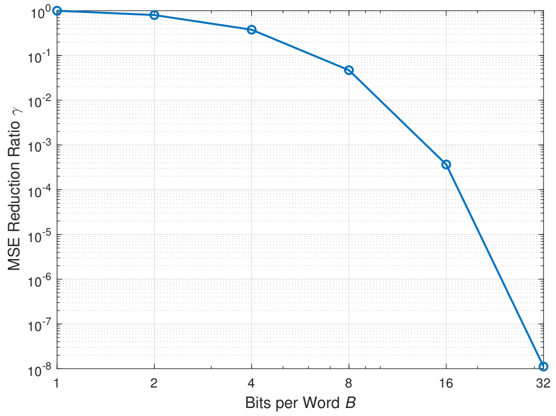

Fig. 5. The MSE reduction ratio γ by Theorem 18.

Fig. 5 shows that the MSE reduction ration improves exponentially with B (as derived in Theorem 18). Although we cannot guarantee the optimality, the proposed Algorithm 1 is very effective to reduce the MSE. Note that $\gamma = 3.66 \times 10^{-4}$ for $B = 16$ and $\gamma = 1.12 \times 10^{-8}$ for $B = 32$.

VII. Conclusion

We proposed a principled approach to improving MRAM's write energy efficiency. After formulating the biconvex optimization problem, we proposed the iterative algorithm to solve the biconvex problem, which attempts to minimize the MSE under a write energy budget. Also, we proved that the proposed algorithm converges and it can reduce the MSE exponentially. The proposed optimization scheme can be extended in future work to coded information representations, where redundancy is added to the written values to further improve the fidelity.

Acknowledgment

The work of Yuval Cassuto was partly supported by the US-Israel Binational Science Foundation, and by the Israel Science Foundation.

References

[1] J. Zhu, “Magnetoresistive random access memory: The path to competitiveness and scalability,” *Proc. IEEE*, vol. 96, no. 11, pp. 1786–1798, Nov. 2008.

[2] J. Kim *et al.*, “Spin-based computing: Device concepts, current status, and a case study on a high-performance microprocessor,” *Proc. IEEE*, vol. 103, no. 1, pp. 106–130, Jan. 2015.

[3] Y. Kim, S. K. Gupta, S. P. Park, G. Panagopoulos, and K. Roy, “Write-optimized reliable design of STT MRAM,” in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, Jul.-Aug. 2012, pp. 3–8.

[4] A. V. Khvalkovskiy *et al.*, “Basic principles of STT-MRAM cell operation in memory arrays,” *J. Phys. D: Appl. Phys*, vol. 46, no. 7, p. 074001, Feb. 2013.

[5] S. Ikeda *et al.*, “A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction,” *Nature Mater.*, vol. 9, no. 9, pp. 721–724, Jul. 2010.

[6] H. Meng and J.-P. Wang, “Spin transfer in nanomagnetic devices with perpendicular anisotropy,” *Appl. Phys. Lett.*, vol. 88, no. 17, p. 172506, Apr. 2006.

[7] T. Nozaki, Y. Shiota, M. Shiraishi, T. Shinjo, and Y. Suzuki, “Voltage-induced perpendicular magnetic anisotropy change in magnetic tunnel junctions,” *Appl. Phys. Lett.*, vol. 96, no. 2, p. 022506, Jan. 2010.

[8] W.-G. Wang, M. Li, S. Hageman, and C. L. Chien, “Electric-field-assisted switching in magnetic tunnel junctions,” *Nature Mater.*, vol. 11, no. 1, pp. 64–68, Nov. 2012.

[9] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, “Energy reduction for STT-RAM using early write termination,” in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2009, pp. 264–268.

[10] A. Ranjan, S. Venkataramani, X. Fong, K. Roy, and A. Raghunathan, “Approximate storage for energy efficient spintronic memories,” in *Proc. Design Autom. Conf. (DAC)*, Jun. 2015, pp. 1–6.

[11] S. Mittal, “A survey of techniques for approximate computing,” *ACM Comput. Surv.*, vol. 48, no. 4, pp. 62:1–62:33, Mar. 2016.

[12] M. Alioto, “Energy-quality scalable adaptive VLSI circuits and systems beyond approximate computing,” in *Proc. Design Autom. Test Europe (DATE)*, Mar. 2017, pp. 127–132.

[13] F. Frustaci, D. Blaauw, D. Sylvester, and M. Alioto, “Approximate SRAMs with dynamic energy-quality management,” *IEEE Trans. VLSI Syst.*, vol. 24, no. 6, pp. 2128–2141, Jun. 2016.

[14] X. Yang and K. Mohanram, “Unequal-error-protection codes in SRAMs for mobile multimedia applications,” in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2011, pp. 21–27.

[15] Y. Kim, M. Kang, L. R. Varshney, and N. R. Shanbhag, “Generalized water-filling for source-aware energy-efficient SRAMs,” *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4826–4841, Oct. 2018.

[16] ——, “SRAM bit-line swings optimization using generalized waterfilling,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1670–1674.

[17] K. Cho, Y. Lee, Y. H. Oh, G.-c. Hwang, and J. W. Lee, “eDRAM-based tiered-reliability memory with applications to low-power frame buffers,” in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, Aug. 2014, pp. 333–338.

[18] Y. Kim, W. H. Choi, C. Guyot, and Y. Cassuto, “On the optimal refresh power allocation for energy-efficient memories,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[19] J. Gorski, F. Pfeuffer, and K. Klamroth, “Biconvex sets and optimization with biconvex functions: A survey and extensions,” *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, Jun. 2007.

[20] R. E. Wendell and A. P. Hurter, “Minimization of a non-separable objective function subject to disjoint constraints,” *Oper. Res.*, vol. 24, no. 4, pp. 643–657, Jul.-Aug. 1976.

[21] W. H. Butler *et al.*, “Switching distributions for perpendicular spin-torque devices within the macrospin approximation,” *IEEE Trans. Magn.*, vol. 48, no. 12, pp. 4684–4700, Dec. 2012.

[22] Y. Kim, Y. Jeon, C. Guyot, and Y. Cassuto, “Optimizing the write fidelity of MRAMs,” *arXiv preprint arXiv:2001.03803*, Jan. 2020. [Online]. Available: https://arxiv.org/abs/2001.03803

[23] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, “On the Lambert W function,” *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.