On the Capacity of Multilevel NAND Flash Memory Channels

Yonglong Li[1] Aleksandar Kavčić[2] Guangyue Han[1]

[1]The University of Hong Kong [2]The University of Hawaii

August, 2016

On the Capacity of Multilevel NAND Flash Memory Channels

Yonglong Li[1] Aleksandar Kavčić[2] Guangyue Han[1]

[1]The University of Hong Kong [2]The University of Hawaii

August, 2016

Introduction

- NAND flash memories have been used widely in real-life applications such as storage devices for computers and cellphones.

Introduction

- NAND flash memories have been used widely in real-life applications such as storage devices for computers and cellphones.
- Flash memories have been more vulnerable to various device or circuit level noises due to the rapidly growing density.

Introduction

- NAND flash memories have been used widely in real-life applications such as storage devices for computers and cellphones.
- Flash memories have been more vulnerable to various device or circuit level noises due to the rapidly growing density.
- Various fault-tolerance techniques such as error correction coding and constrained coding have been proposed to overcome this problem.
  - Error correction codes: BCH codes Sun et al. 2007, LDPC codes by Wang et al. 2011 and Dong et al. 2011, rank modulation by Jiang et al. 2009;
  - Constrained codes: Qin et al. 2014 and Taranalli et al. 2015.

Introduction

Introduction

- The other direction is to model flash memory by communication channels and then analyze their theoretical information limits; Representative work includes Dong et al. 2011 and 2012, Cai et al. 2013, Li et al. 2014, Taranalli et al. 2015.

Introduction

- The other direction is to model flash memory by communication channels and then analyze their theoretical information limits; Representative work includes Dong et al. 2011 and 2012, Cai et al. 2013, Li et al. 2014, Taranalli et al. 2015.
- In 2014 based on Dong et al. 2011, Asadi et al. proposed a more mathematically tractable communication channel, which incorporates inter-symbol interference and output memory.

In this work, we mainly focus on Asadi et al.'s channel model.

Flash Memory Channel

Flash Memory Channel

Channel Model

$$Y_0 = X_0 + W_0 + U_0,$$
$$Y_n = X_n + A_n X_{n-1} + B_n(Y_{n-1} - E_{n-1}) + W_n + U_n, \ n \geq 1$$

Flash Memory Channel

Channel Model

$$Y_0 = X_0 + W_0 + U_0,$$
$$Y_n = X_n + A_n X_{n-1} + B_n(Y_{n-1} - E_{n-1}) + W_n + U_n, \ n \geq 1$$

Assumptions

(i) $\{X_i\}$ is the channel input process, taking values from a finite alphabet $\mathcal{X} \triangleq \{v_0, v_1, \cdots, v_{M-1}\}$.

Flash Memory Channel

Channel Model

$$Y_0 = X_0 + W_0 + U_0,$$
$$Y_n = X_n + A_n X_{n-1} + B_n(Y_{n-1} - E_{n-1}) + W_n + U_n, \; n \geq 1$$

Assumptions

(i) $\{X_i\}$ is the channel input process, taking values from a finite alphabet $\mathcal{X} \triangleq \{v_0, v_1, \cdots, v_{M-1}\}$.

(ii) $\{A_i\}$, $\{B_i\}$, $\{E_i\}$ and $\{W_i\}$ are i.i.d. Gaussian random processes with mean 0 and variance σ_A^2, $0 < \sigma_B^2 < 1$, σ_E^2 and 1, respectively;

Flash Memory Channel

Channel Model

$$Y_0 = X_0 + W_0 + U_0,$$
$$Y_n = X_n + A_n X_{n-1} + B_n(Y_{n-1} - E_{n-1}) + W_n + U_n,\ \ n \geq 1$$

Assumptions

(i) $\{X_i\}$ is the channel input process, taking values from a finite alphabet $\mathcal{X} \stackrel{\triangle}{=} \{v_0, v_1, \cdots, v_{M-1}\}$.

(ii) $\{A_i\}$, $\{B_i\}$, $\{E_i\}$ and $\{W_i\}$ are i.i.d. Gaussian random processes with mean 0 and variance σ_A^2, $0 < \sigma_B^2 < 1$, σ_E^2 and 1, respectively;

(iii) $\{A_i\}$, $\{B_i\}$, $\{E_i\}$, $\{W_i\}$, $\{U_i\}$ and $\{X_i\}$ are mutually independent;

Flash Memory Channel

Channel Model

$$Y_0 = X_0 + W_0 + U_0,$$
$$Y_n = X_n + A_n X_{n-1} + B_n(Y_{n-1} - E_{n-1}) + W_n + U_n,\ \ n \geq 1$$

Assumptions

(i) $\{X_i\}$ is the channel input process, taking values from a finite alphabet $\mathcal{X} \stackrel{\triangle}{=} \{v_0, v_1, \cdots, v_{M-1}\}$.

(ii) $\{A_i\}$, $\{B_i\}$, $\{E_i\}$ and $\{W_i\}$ are i.i.d. Gaussian random processes with mean 0 and variance σ_A^2, $0 < \sigma_B^2 < 1$, σ_E^2 and 1, respectively;

(iii) $\{A_i\}$, $\{B_i\}$, $\{E_i\}$, $\{W_i\}$, $\{U_i\}$ and $\{X_i\}$ are mutually independent;

(iv) $\{U_i\}$ is an i.i.d. random process with the uniform distribution over (α_1, α_2).

Channel Model

Remarks

Channel Model

Remarks

- Flash memory channel is not stationary in the sense that the output process is not stationary even if the input is stationary.

Channel Model

Remarks

- Flash memory channel is not stationary in the sense that the output process is not stationary even if the input is stationary.
- Flash memory channels is a channel which possesses input and output memory.

Channel Model

Remarks

- Flash memory channel is not stationary in the sense that the output process is not stationary even if the input is stationary.
- Flash memory channels is a channel which possesses input and output memory.
- Flash memory channel is a channel with infinite states if (x_i, y_i) is regarded as the state for the channel at time $i + 1$.

Capacity

Shannon Capacity

$$C_{Shannon} = \lim_{n\to\infty} \frac{1}{n+1} \sup_{p(x_0^n)} I(X_0^n; Y_0^n).$$

Capacity

Shannon Capacity

$$C_{Shannon} = \lim_{n\to\infty} \frac{1}{n+1} \sup_{p(x_0^n)} I(X_0^n; Y_0^n).$$

m-th Order Markov Capacity

$$C_{Markov}^{(m)} = \sup I(X; Y),$$

where the supremum is taken over all m-th order stationary Markov chains X.

Capacity

Shannon Capacity

$$C_{Shannon} = \lim_{n\to\infty} \frac{1}{n+1} \sup_{p(x_0^n)} I(X_0^n; Y_0^n).$$

m-th Order Markov Capacity

$$C_{Markov}^{(m)} = \sup I(X; Y),$$

where the supremum is taken over all m-th order stationary Markov chains X.

Stationary Capacity

$$C_S = \sup I(X; Y) = \sup \lim_{n\to\infty} \frac{1}{n+1} I(X_0^n; Y_0^n),$$

where the supremum is taken over all stationary processes X.

Main Result

Main Theorem

Let C be the operational capacity of the flash memory channel, that is, C is the supremum of the achievable rates. Then

$$C = C_{Shannon} = C_S = \lim_{m \to \infty} C^{(m)}_{Markov}.$$

Markov Approximation

- In general, for channels with memory or states there is no closed-form characterization of channel capacity.

Markov Approximation

- In general, for channels with memory or states there is no closed-form characterization of channel capacity.
- A natural idea is using the so-called *Markov approximation* scheme to numerically compute $C_{Markov}^{(m)}$ and its capacity achieving distribution.

Markov Approximation

- In general, for channels with memory or states there is no closed-form characterization of channel capacity.
- A natural idea is using the so-called *Markov approximation* scheme to numerically compute $C_{Markov}^{(m)}$ and its capacity achieving distribution.
- Two known algorithms:

Markov Approximation

- In general, for channels with memory or states there is no closed-form characterization of channel capacity.
- A natural idea is using the so-called *Markov approximation* scheme to numerically compute $C_{Markov}^{(m)}$ and its capacity achieving distribution.
- Two known algorithms:
  - P. O. Vontobel, A. Kavčić, D. M. Arnold, and H. A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1887–1918, May 2008.
  - G. Han, "A randomized algorithm for the capacity of finite-state channels," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3651-3669, July 2015.

Remark

This theorem justifies the effectiveness of Markov approximation for multilevel NAND flash memory channels.

Asymptotic Mean Stationarity

One of the main tools that will be used in this work is the so-called asymptotic mean stationarity (AMS).

- Let $T : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ be the *left shift operator* defined by

$$Tx = (x_1, x_2, \cdots) \text{ for } x = (x_0, x_1, x_2, \cdots) \in \mathbb{R}^{\mathbb{N}}.$$

- A probability measure μ on $\mathbb{R}^{\mathbb{N}}$ is said to be *asymptotically mean stationary* if there exists a probability measure $\bar{\mu}$ such that for any Borel set $A \subset R^{\mathbb{N}}$,

$$\bar{\mu}(A) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mu(T^{-i}A); \tag{1}$$

  And $\bar{\mu}$ in (1), if it exists, is said to be the *stationary mean* of μ.
- The process $\{Y_n\}$ is said to be *asymptotically mean stationary* if the associated measure P_Y is *asymptotically mean stationary*.

Asymptotic Mean Stationarity

The following theorem gives an analog of Birkhoff's ergodic theorem for asymptotically mean stationary processes.

Theorem

Suppose that P_Y is asymptotically mean stationary with stationary mean $\bar{P}_Y$. If $\mathbf{E}_{\bar{P}_Y}[|Y_0|] < \infty$, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \textit{exists} \;\; P_Y - a.s.$$

Asymptotic Mean Stationarity

The following two theorems relate convergences with respect to the measure P_Y and its stationary mean $\bar{P}_Y$.

Theorem

If P_Y is symptotically mean stationary with stationary mean $\bar{P}_Y$, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ exists } P_Y\text{−a.s. if and only if } \quad \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ exists } \bar{P}_Y\text{−a.s.}$$

Also, if the limiting function as above is integrable (with respect to P_Y or $\bar{P}_Y$), then

$$\mathbf{E}_{P_Y}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \mathbf{E}_{\bar{P}_Y}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i\right].$$

Asymptotic Mean Stationarity

Theorem (A.Barron,1985)

Suppose that P_Y is asymptotically mean stationary with stationary mean $\bar{P}_Y$, and suppose that for each n, there exists $k = k(n)$ such that $I_{P_Y}(Y_1^n; Y_{k+n+1}^{\infty} | Y_{n+1}^{n+k})$ is finite. If for some shift invariant random variable Z (i.e., $Z = Z \circ T$),

$$\lim_{n\to\infty} \frac{1}{n} \log \bar{f}(Y_1^n) = Z, \quad \bar{P}_Y - a.s.,$$

then we have

$$\lim_{n\to\infty} \frac{1}{n} \log f(Y_1^n) = Z, \quad P_Y - a.s.$$

Indecomposability

The flash memory channel is "indecomposable" in the following sense.

Lemma

a) For any $k \le n$, x_k^n, y_k and $\tilde{y}_k$, we have

$$\int_{-\infty}^{\infty} \left| f_{Y_n|X_k^n,Y_k}(y_n|x_k^n, y_k) - f_{\tilde{Y}_n|X_k^n,\tilde{Y}_k}(y_n|x_k^n, \tilde{y}_k) \right| dy_n \le \sigma_B^{2(n-k)}(y_k^2 + \tilde{y}_k^2).$$

b) For any k, n, x_n and y_n and $\hat{x}_0^n$, we have

$$\int_{-\infty}^{\infty} \left| f_{Y_n|X_0^n}(\hat{y}|\hat{x}_0^n) - f_{Y_{n+k+1}|X_{k+1}^{n+k+1},X_k,Y_k}(\hat{y}|\hat{x}_0^n, x_k, y_k) \right| d\hat{y} \le \sigma_B^{2n}(\sigma_A^2 x_k^2 + 2\sigma_B^2(y_k^2 + \sigma_E^2)).$$

Proof of the Main Theorem

Proof of $C_S \leq C$

- We prove the AEP for flash memory channel model. Let X be a stationary and ergodic input process and Y be the output by passing X through the flash memory channel. Then (X, Y) is asymptotic mean stationary and ergodic and also with probability 1,

$$-\frac{1}{n+1}\log f(Y_0^n) \to H(Y);$$

$$\frac{1}{n+1}\log \frac{f(Y_0^n|X_0^n)}{f(Y_0^n)} \to I(X;Y).$$

- For any rate $R < C_S$ and $\varepsilon > 0$, choose a stationary ergodic input process X such that $R < I(X;Y) - \varepsilon$. As shown above, $\{X, Y\}$ satisfies the AEP, we can complete the proof of the achievability by going through the usual random coding argument.

Proof of the Main Theorem

P_Y is AMS

- $P(Y_k \in A) = \sum_{x_0^k} p_{X_0^k}(x_0^k) p_{Y_k|X_0^k}(A|x_0^k)$.

Proof of the Main Theorem

P_Y is AMS

- $P(Y_k \in A) = \sum_{x_0^k} p_{X_0^k}(x_0^k) p_{Y_k|X_0^k}(A|x_0^k)$.
- $P(Y_{k+1} \in A) = \sum_{\tilde{x}_0, x_0^k} \Big\{ p_{X_1^{k+1}}(x_0^k) p_{X_0|X_1^{k+1}}(\tilde{x}_0|x_0^k)$

  $\times \int f_{Y_0|X_0}(\tilde{y}|\tilde{x}_0)\, p_{Y_{k+1}|X_1^{k+1}, X_0, Y_0}(A|x_0^k, \tilde{x}_0, \tilde{y}) d\tilde{y} \Big\}$

Proof of the Main Theorem

P_Y is AMS

- $P(Y_k \in A) = \sum_{x_0^k} p_{X_0^k}(x_0^k) p_{Y_k|X_0^k}(A|x_0^k)$.
- $P(Y_{k+1} \in A) = \sum_{\tilde{x}_0, x_0^k} \Big\{ p_{X_1^{k+1}}(x_0^k) p_{X_0|X_1^{k+1}}(\tilde{x}_0|x_0^k)$

  $\times \int f_{Y_0|X_0}(\tilde{y}|\tilde{x}_0)\, p_{Y_{k+1}|X_1^{k+1},X_0,Y_0}(A|x_0^k, \tilde{x}_0, \tilde{y}) d\tilde{y} \Big\}$
- $|P(Y_{k+1} \in A) - P(Y_k \in A)| \leq M\sigma_B^{2k}$.

Proof of the Main Theorem

P_Y is AMS

- $P(Y_k \in A) = \sum_{x_0^k} p_{X_0^k}(x_0^k) p_{Y_k|X_0^k}(A|x_0^k)$.
- $P(Y_{k+1} \in A) = \sum_{\tilde{x}_0, x_0^k} \Big\{ p_{X_1^{k+1}}(x_0^k) p_{X_0|X_1^{k+1}}(\tilde{x}_0|x_0^k)$

  $\times \int f_{Y_0|X_0}(\tilde{y}|\tilde{x}_0)\, p_{Y_{k+1}|X_1^{k+1}, X_0, Y_0}(A|x_0^k, \tilde{x}_0, \tilde{y}) d\tilde{y} \Big\}$
- $|P(Y_{k+1} \in A) - P(Y_k \in A)| \leq M\sigma_B^{2k}$.
- $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} P(Y_k \in A)$ exists.

Proof of the Main Theorem

Existence of $H(Y)$

- Uniform integrability of $\{Y_n^2\}$ under P_Y, together with $P_{Y_n}(\cdot) \to \bar{P}_Y(\cdot)$, implies that

$$\mathbf{E}_{\bar{P}_Y}[Y_0^2] = \lim_{n\to\infty} \mathbf{E}[Y_n^2] < \infty.$$

- Under P_Y, with probability 1, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n} Y_n^2$ exists.
- $|\log f(Y_0^n)| \leq M_0 + M_1 \sum_{i=0}^{n} Y_i^2$.
- $\mathbf{E}\left[-\frac{1}{n+1}\log f(Y_0^n)\right] \to H(Y)$.

Proof of the Main Theorem

Proof of $C_S \geq \lim_{m\to\infty} C_{Markov}^{(m)}$

This can be shown by observing that stationary Markov processes is a subclass of stationary processes.

Proof of the Main Theorem

Proof of $C_S \geq \lim_{m\to\infty} C_{Markov}^{(m)}$

This can be shown by observing that stationary Markov processes is a subclass of stationary processes.

Proof of $C_S \leq \lim_{m\to\infty} C_{Markov}^{(m)}$

- It suffices to show that for any $\varepsilon > 0$ and any stationary and ergodic process X, one can find an m-th order Markov chain $\tilde{X}$ such that
$$I(\tilde{X};\tilde{Y}) \geq I(X;Y) - \varepsilon.$$
- Given X, construct the m-th order Markov chain $\tilde{X}$ by setting
$$P(\tilde{X}_0^m = x_0^m) = P(X_0^m = x_0^m).$$

Proof of Main Theorem

Proof of $C_S \leq \lim_{m\to\infty} C_{Markov}^{(m)}$

$$
\begin{aligned}
H(\tilde{X}|\tilde{Y}) &\leq \lim_{s\to\infty} \frac{1}{s(m+1)} \sum_{i=0}^{s-1} \left\{ H(\tilde{X}_{im+i}^{(i+1)m+i}) - I(\tilde{X}_{im+i}^{(i+1)m+i}; \tilde{Y}_{im+i}^{(i+1)m+i}) \right\} \\
&\leq \lim_{s\to\infty} \frac{1}{s(m+1)} \sum_{i=0}^{s-1} \left\{ H(\tilde{X}_0^m) - I(\tilde{X}_0^m; \tilde{Y}_0^m) + \varepsilon \right\} \\
&= \frac{1}{m+1} H(\tilde{X}_0^m|\tilde{Y}_0^m) + \frac{\varepsilon}{m+1} \\
&= \frac{1}{m+1} H(X_0^m|Y_0^m) + \frac{\varepsilon}{m+1}.
\end{aligned}
$$

Proof of Main Theorem

Proof of $C_S \leq \lim_{m\to\infty} C_{Markov}^{(m)}$

$$
\begin{aligned}
I(\tilde{X};\tilde{Y}) &= H(\tilde{X}) - H(\tilde{X}|\tilde{Y}) \geq H(\tilde{X}) - \frac{1}{m+1} H(\tilde{X}_0^m|\tilde{Y}_0^m) \\
&= H(\tilde{X}_m|\tilde{X}_0^{m-1}) - \frac{1}{m+1} H(X_0^m|Y_0^m) - \frac{\varepsilon}{m+1} \\
&= H(X_m|X_0^{m-1}) - \frac{1}{m+1} H(X_0^m|Y_0^m) - \frac{\varepsilon}{m+1} \\
&\geq H(X) - \frac{1}{m+1} H(X_0^m|Y_0^m) - \frac{\varepsilon}{m+1} \\
&\geq I(X;Y) - \varepsilon.
\end{aligned}
$$

Conclusion and Future Work

Conclusion

(a) For a multilevel NAND flash memory channel under mild assumptions, we prove that such a channel is indecomposable and it features asymptotic equipartition property;

Conclusion and Future Work

Conclusion

(a) For a multilevel NAND flash memory channel under mild assumptions, we prove that such a channel is indecomposable and it features asymptotic equipartition property;

(b) We prove equalities among operational capacity, Shannon capacity, Stationary capacity and Markov capacity.

Conclusion and Future Work

Conclusion and Future Work

Future Work

(a) Investigate the concavity of $I(X; Y)$ with respect to the parameters of the input Markov chain.

(b) Numerically compute the Markov capacity and its capacity achieving distributions by generalizing the GBAA or Han's randomized algorithm.

(c) Investigate the effectiveness of Markov approximation for the two dimensional flash memory channel.

Thanks for Your Attention!