

# The Derivatives of Entropy Rate and Capacity for Finite-State Channels

Henry D. Pfister  
Texas A&M University, College Station

Entropy of Hidden Markov Processes  
and Connections to Dynamical Systems  
BIRS, Banff, Alberta, Canada  
October 5th, 2007

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing



# Taxonomy of Finite-State Channels

## Three Major Classes

- Deterministic State FSCs

$$\sum_{y \in \mathcal{Y}} P(y, s' | x, s) = \begin{cases} 1 & \text{if } f(x, s) = s' \\ 0 & \text{otherwise} \end{cases}$$

- Next state given by  $f(x, s)$  with current state  $s$  and input  $x$
- Example: Intersymbol Interference (ISI) Channels

- Independent State FSCs

$$\sum_{y \in \mathcal{Y}} P(y, s' | x, s) = \sum_{y \in \mathcal{Y}} P(y, s' | x', s) \text{ for all } x, x' \in \mathcal{X}$$

- Distribution of next state is independent of input
- Example: Fading Channels (e.g., Gilbert-Elliot Channel)

- General FSCs

- The next state is a random variable which depends on the input
- Example: Media noise in magnetic recording

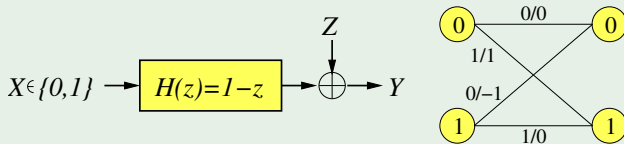
# Application: Magnetic Storage

## Channel Properties

- Strong write fields maximize reliability  $\implies$  binary-input
- Magnetization of nearby bits affects detector  $\implies$  ISI

## Simple Model: The Dicode Channel

- Discrete-time channel with linear response  $H(z) = 1 - z$
- AWGN used in general, or erasures for analysis
- State given by last input, edges labelled by input/output





# History of Finite-State Channels

## In the Beginning...

- The Entropy of a Function of a Finite-State Markov Chain (Blackwell 1957, +Breiman, Thomasian 1958, Birch 1962)

## Before Turbo

- Finite-State ISI Channels (Hirt 1988, Shamai et al. 1991)
- Finite-State Fading Channels (Goldsmith et al. 1994)

## Recent Work

- Monte Carlo Algorithms for Mutual Information (Arnold et al. 2001, Pfister et al. 2001, Sharma et al. 2001)
- A Generalized Blahut-Arimoto Algorithm (Kavčić 2001)
- An Upper Bound on Capacity (Vontobel et al. 2001)
- **and far too many to list in past 5 years**

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - **The Capacity of a Finite-State Channel**
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# Dependence on Initial State

## Upper and Lower Capacity

$$\bar{C} \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \max_{\Pr(X_1^n)} \max_{s_0 \in \mathcal{S}} I(X_1^n; Y_1^n | S_0 = s_0)$$

$$\underline{C} \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \max_{\Pr(X_1^n)} \min_{s_0 \in \mathcal{S}} I(X_1^n; Y_1^n | S_0 = s_0)$$

## Sufficient Conditions for $\bar{C} = C = \underline{C}$

- Indecomposable: Channel forgets initial state for all inputs
  - Namely  $|\Pr(S_n | X_1^n = x_1^n, S_0 = s) - \Pr(S_n | X_1^n = x_1^n, S_0 = s')| \rightarrow 0$
- Finite Memory: Channel state is a function of last  $\nu$  inputs

# Optimizing the Input Distribution

## Markov Input Process (memory $m$ )

$$\Pr(X_i = x' | X_{i-m}^{i-1} = \mathbf{x}) = R(x' | \mathbf{x})$$

## A Sequence of Lower Bounds on Capacity

Let  $\mathcal{M}_m(\mathcal{X})$  be the set of Markov input dist. memory  $m$  and

$$L_m = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{R \in \mathcal{M}_m(\mathcal{X})} I(X_1^n; Y_1^n)$$

$$C_{i.u.d.} \leq L_0 \leq L_1 \leq L_2 \leq \dots \leq C$$

## Treat As One Process by Combining Input and Channel State

- Joint Input-Channel State Set:  $\mathcal{Q} = \mathcal{S} \times \mathcal{X}^m$ 
  - Vector representation:  $(s, \mathbf{x}) \in \mathcal{Q}$  for  $s \in \mathcal{S}$  and  $\mathbf{x} \in \mathcal{X}^m$
  - Component- $\mathcal{S}$  projection: For  $q = (s, \mathbf{x})$ , we have  $\mathcal{S}(q) = s$
  - Component- $\mathcal{X}^m$  projection: For  $q = (s, \mathbf{x})$ , we have  $\mathcal{X}(q) = \mathbf{x}$

# The APP-BCJR Algorithm

## Forward/Backward State Probability

$$\alpha_i(q) \triangleq \Pr(Q_i = q | Y_1^{i-1} = y_1^{i-1}) \in \mathcal{M}(\mathcal{Q})$$

$$\beta_i(q) \triangleq \frac{1}{\pi_q} \Pr(Q_i = q | Y_i^n = y_i^n) \in \mathbb{R}_+^{|\mathcal{Q}|}$$

## The Forward Recursion

Randomly generate the sequence  $y_1, y_2, \dots$  and compute

$$\alpha_{i+1}(q) = \frac{1}{\psi_{i+1}} \sum_{q' \in \mathcal{Q}} \alpha_i(q') \sum_{x \in \mathcal{X}} \overbrace{P(y, \mathcal{S}(q) | x, \mathcal{S}(q'))}^{\Pr(Y_i=y, S_{i+1}=q | X_i=x, S_i=q')} \overbrace{R(x, \mathcal{X}(q'))}^{\Pr(X_i=x | X_1^{i-1})}$$

Normalization so  $\sum_q \alpha_{i+1}(q) = 1$  is  $\psi_{i+1} = \Pr(Y_i = y_i | Y_1^{i-1} = y_1^{i-1})$

$$-\frac{1}{n} \sum_{i=1}^n \log \psi_{i+1} = -\frac{1}{n} \log \Pr(y_1^n) \stackrel{a.s.}{\rightarrow} H(\mathcal{Y})$$

# Analysis of the Forward Recursion (1)

## Joint Process $\{Q_i, \alpha_i\}_{i \geq 1}$ Forms a Markov Chain

- Given  $q_i$ , choose  $q_{i+1}$ , generate  $y_i$ , and compute  $\alpha_{i+1}(\cdot)$

## Stationary Measures

$$\mu_q(A) \triangleq \lim_{i \rightarrow \infty} \Pr(Q_i = q, \alpha_i \in A) \quad (\text{Forward Furstenberg})$$

$$\mu(A) \triangleq \sum_{q \in \mathcal{Q}} \mu_q(A) = \lim_{i \rightarrow \infty} \Pr(\alpha_i \in A) \quad (\text{Forward Blackwell})$$

## Entropy Rate

$$H(\mathcal{Y}) = \sum_{q \in \mathcal{Q}} \int_{\mathcal{M}(\mathcal{Q})} \sum_{y \in \mathcal{Y}} \overbrace{d\mu_q(\alpha) f_q(y)}^{\Pr(Q=q, \alpha, Y=y)} \log \overbrace{\sum_{q' \in \mathcal{Q}} \alpha(q') f_{q'}(y)}^{\Pr(Y|\alpha)}$$

$$f_q(y) = \Pr(Y_i = y | Q_i = q)$$

# Analysis of the Forward Recursion (2)

## Consistency of the APP Estimate

Let  $X, Y$  be discrete r.v. with random vector  $\mathbf{A}$ :  $[\mathbf{A}]_x = \Pr(X = x|Y)$

$$\begin{aligned} \Pr(X = x, \mathbf{A} = \mathbf{a}) &= \Pr(\mathbf{A} = \mathbf{a}) \Pr(X = x|\mathbf{A} = \mathbf{a}) \\ &= \Pr(\mathbf{A} = \mathbf{a}) \Pr(X = x|Y) \\ &= \Pr(\mathbf{A} = \mathbf{a})[\mathbf{a}]_x \end{aligned}$$

$\mathbf{A}$  is a sufficient statistic for  $X$  because  $Y \rightarrow \mathbf{A} \rightarrow X$

## HMP Relationship Between Blackwell and Furstenberg Measures

$$\mu_q(\mathbf{a}) = \Pr(\alpha_* = \mathbf{a}) \Pr(Q_* = q|\alpha_* = \mathbf{a}) = \mu(\mathbf{a})[\mathbf{a}]_q$$

Simpler Yet Identical Simulation Strategy: Markov Chain  $\{\alpha_i\}_{i \geq 1}$

- Pick  $q_i \sim \alpha_i(\cdot)$ , choose  $q_{i+1}$ , generate  $y_i$ , and compute  $\alpha_{i+1}$

# Analysis of the Forward Recursion (2)

## Consistency of the APP Estimate

Let  $X, Y$  be discrete r.v. with random vector  $\mathbf{A}$ :  $[\mathbf{A}]_x = \Pr(X = x|Y)$

$$\begin{aligned} \Pr(X = x, \mathbf{A} = \mathbf{a}) &= \Pr(\mathbf{A} = \mathbf{a}) \Pr(X = x|\mathbf{A} = \mathbf{a}) \\ &= \Pr(\mathbf{A} = \mathbf{a}) \Pr(X = x|Y) \\ &= \Pr(\mathbf{A} = \mathbf{a}) [\mathbf{a}]_x \end{aligned}$$

$\mathbf{A}$  is a sufficient statistic for  $X$  because  $Y \rightarrow \mathbf{A} \rightarrow X$

## HMP Relationship Between Blackwell and Furstenberg Measures

$$\mu_q(\mathbf{a}) = \Pr(\alpha_* = \mathbf{a}) \Pr(Q_* = q|\alpha_* = \mathbf{a}) = \mu(\mathbf{a})[\mathbf{a}]_q$$

Simpler Yet Identical Simulation Strategy: Markov Chain  $\{\alpha_i\}_{i \geq 1}$

- Pick  $q_i \sim \alpha_i(\cdot)$ , choose  $q_{i+1}$ , generate  $y_i$ , and compute  $\alpha_{i+1}$

# Analysis of the Forward Recursion (2)

## Consistency of the APP Estimate

Let  $X, Y$  be discrete r.v. with random vector  $\mathbf{A}$ :  $[\mathbf{A}]_x = \Pr(X = x|Y)$

$$\begin{aligned}\Pr(X = x, \mathbf{A} = \mathbf{a}) &= \Pr(\mathbf{A} = \mathbf{a}) \Pr(X = x|\mathbf{A} = \mathbf{a}) \\ &= \Pr(\mathbf{A} = \mathbf{a}) \Pr(X = x|Y) \\ &= \Pr(\mathbf{A} = \mathbf{a})[\mathbf{a}]_x\end{aligned}$$

$\mathbf{A}$  is a sufficient statistic for  $X$  because  $Y \rightarrow \mathbf{A} \rightarrow X$

## HMP Relationship Between Blackwell and Furstenberg Measures

$$\mu_q(\mathbf{a}) = \Pr(\alpha_* = \mathbf{a}) \Pr(Q_* = q|\alpha_* = \mathbf{a}) = \mu(\mathbf{a})[\mathbf{a}]_q$$

## Simpler Yet Identical Simulation Strategy: Markov Chain $\{\alpha_i\}_{i \geq 1}$

- Pick  $q_i \sim \alpha_i(\cdot)$ , choose  $q_{i+1}$ , generate  $y_i$ , and compute  $\alpha_{i+1}$

# Outline

- 1 **Introduction**
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - **Connections with Lyapunov Exponents**
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# A Matrix Perspective

## The Transition-Observation Matrix

For any  $y \in \mathcal{Y}$ , let  $M(y)$  be the  $|\mathcal{Q}| \times |\mathcal{Q}|$  matrix defined by

$$\begin{aligned} [M(y)]_{q,q'} &\triangleq \Pr(Y_i = y, Q_{i+1} = q' | Q_i = q) \\ &= \sum_{x \in \mathcal{X}} P(y, \mathcal{S}(q') | x, \mathcal{S}(q)) R(x, \mathcal{X}(q)) \end{aligned}$$

Products compute the multi-step transition-observation matrix

$$[M(\mathbf{y}_j^k)]_{q,q'} \triangleq \left[ \prod_{i=j}^k M(y_i) \right]_{q,q'} = \Pr(Y_j^k = \mathbf{y}_j^k, Q_{k+1} = q' | Q_j = q)$$

# The Forward/Backward Recursions

## Markov Chain: Transition Matrix $P$ and Stationary Distribution $\pi$

$$[P]_{q,q'} \triangleq \Pr(Q_{i+1} = q' | Q_i = q) = \sum_{y \in \mathcal{Y}} M(y) \quad \pi P = \pi$$

## Matrix Form of the Forward/Backward Recursions

$$\alpha_{i+1} = \frac{\alpha_i M(\mathbf{y}_i)}{\alpha_i M(\mathbf{y}_i) \mathbf{1}} \quad \beta_{i-1} = \frac{M(\mathbf{y}_{i-1}) \beta_i}{\pi M(\mathbf{y}_{i-1}) \beta_i}$$

$\pi, \alpha$  row vectors,  $\beta$  column vector,  $\pi \mathbf{1} = 1$ ,  $\alpha_{i+1} \mathbf{1} = 1$ , and  $\pi \beta = 1$

## The Entropy Rate as a Lyapunov Exponent

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \left\| \prod_{i=1}^n M(\mathbf{y}_i) \right\| \stackrel{a.s.}{=} -H(\mathcal{Y})$$

- Formulation gives many results: convergence, continuity, CLT, ...

# Of Blackwell and Furstenberg

## What is the Measure of a Man?

- An HMP is also a matrix function of Markov chain
  - Markov  $X_1, X_2, \dots$  with  $Y_i = \phi(X_i)$  gives matrices  $M(\phi(Y_i))$
  - The matrix sequence is a function of a FS Markov chain
  - The Lyapunov exponent of the matrix sequence is  $-H(\mathcal{Y})$
- Lyapunov exp. can be computed via the **Furstenberg measure**
  - Proposed for i.i.d. in 1963 and for Markov (with Kifer) in 1983
  - Use measure on state  $X_i$  and vector  $v_i = M_i v_{i-1} / \|M_i v_{i-1}\|$
  - Easy to see that  $\{X_i, v_i\}_{i \geq 1}$  forms a Markov process
  - Lyapunov exp. via integration against 1-step kernel
  - Simplifies to **Blackwell's approach** using APP consistency
- Extension to the relative entropy (or divergence)
  - $D(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -H(P) - E_P [\log q(x)]$
  - Modified matrix function  $\tilde{M}(\cdot)$  allows  $E_P [\log q(x)]$  for two HMPs

# Mixing Conditions

## Connectivity

$$[A(\mathcal{Y})]_{s,s'} \triangleq \begin{cases} 1 & \text{if } P(y, s' | x, s) > 0 \forall x \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$$

A FSC is *indecomposable* if  $A(\mathcal{Y})$  is irreducible and aperiodic

## Rank 1 Condition (R1)

- There exists  $k < \infty$  and  $y_1^k$  s.t.  $\Pr(y_1^k) > 0$  and  $M(y_1^k)$  is rank 1
- Receiving  $y_1^k$  resets the process and marks a renewal time

## Strong Mixing Condition (S1)

- There exists  $k < \infty$  s.t.  $M(y_1^k) > 0$  for all  $y_1^k$
- All sample paths forget the past exponentially fast

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - **Motivating Example**
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# The Derivative of the Entropy Rate $H(\mathcal{Y})$

## What Form Does the Derivative Take?

- Consider the derivative w.r.t. parameters (e.g.,  $M(\cdot)$ ) of the HMP
- Using the Blackwell integral approach
  - Can express it in terms of the derivative of the Blackwell Measure
  - But, the derivative of the Blackwell Measure is non-trivial

## Is There a Simpler Formula?

- Yes, there is a formula without derivatives of measures
- Warning: It is perhaps known and published much earlier
  - Ex. Vontobel et al. have same result with a more tedious proof
- Generalizes naturally to Lyapunov exponents
- Can be motivated with a very simple example

# The Log Spectral Radius Example

## Exponential Growth Rate Formulation

Let  $M$  be a  $d \times d$  non-neg. matrix with spectral radius  $\rho(M)$ , then

$$\log \rho(M) = \lim_{n \rightarrow \infty} \frac{1}{n} \log (u^T M^n v)$$

for any positive real vectors  $u, v \in \mathbb{R}_+^d$ .

## Well-known Formula for the Derivative of the Log Spectral Radius

Let  $M(\theta) \in \mathbb{R}^{d \times d}$  have a simple top eigenvalue for  $\theta \in D \subset \mathbb{R}$ , then

$$\left. \frac{d}{d\theta} \log \rho (M(\theta)) \right|_{\theta=\theta^* \in D} = \frac{a^T M'(\theta^*) b}{a^T M(\theta^*) b},$$

$M'(\theta^*)$  element-wise derivative,  $a, b$  left/right eigenvectors of  $M(\theta^*)$

# Method of Proof

## The Total Derivative Method

Let  $f_n : \mathbb{R} \rightarrow \mathbb{R}$  and  $g_n : \mathbb{R}^n \rightarrow \mathbb{R}$  be sequences such that

$$f_n(\theta) = g_n(\theta, \dots, \theta)$$

The derivative can be broken into the sum of  $n$  terms

$$\frac{d}{d\theta} f_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta_i} g_n(\theta_1, \dots, \theta_n) \Big|_{(\theta_1, \dots, \theta_n) = (\theta, \dots, \theta)}$$

## Connection to Log Spectral Radius

$$g_n(\theta_1, \dots, \theta_n) = \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^n M(\theta_i) \right) v \right)$$

for non-neg.  $M$  implies

$$\lim_{n \rightarrow \infty} f_n(\theta) = \log \rho(M(\theta))$$

# Proof of the LSR Derivative Formula (1)

Let  $M = M(\theta^*)$  be non-neg. and  $M' = M'(\theta^*)$  for  $\theta^* \in D$ . Let  $a, b$  be left/right eigenvectors satisfying  $a^T M = \rho(M)a^T$  and  $Mb = \rho(M)b$ .

$$\begin{aligned}
 f'_n(\theta^*) &= \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^n M(\theta_i) \right) v \right) \Bigg|_{(\theta_1, \dots, \theta_n) = (\theta^*, \dots, \theta^*)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \log \left( u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v \right) \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)} \\
 &\quad \text{note: } \frac{d}{d\theta} x^T M(\theta) y = \sum_{k,l} x_k \frac{d}{d\theta} M_{k,l}(\theta) y_l = x^T M'(\theta) y \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M'(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v}{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v} \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{u^T M^{j-1} M' M^{n-j} v}{u^T M^{j-1} M M^{n-j} v} \tag{A}
 \end{aligned}$$

# Proof of the LSR Derivative Formula (1)

Let  $M = M(\theta^*)$  be non-neg. and  $M' = M'(\theta^*)$  for  $\theta^* \in D$ . Let  $a, b$  be left/right eigenvectors satisfying  $a^T M = \rho(M)a^T$  and  $Mb = \rho(M)b$ .

$$\begin{aligned}
 f'_n(\theta^*) &= \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^n M(\theta_i) \right) v \right) \Bigg|_{(\theta_1, \dots, \theta_n) = (\theta^*, \dots, \theta^*)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \log \left( u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v \right) \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)}
 \end{aligned}$$

$$\text{note: } \frac{d}{d\theta} x^T M(\theta) y = \sum_{k,l} x_k \frac{d}{d\theta} M_{k,l}(\theta) y_l = x^T M'(\theta) y$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{j=1}^n \frac{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M'(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v}{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v} \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{u^T M^{j-1} M' M^{n-j} v}{u^T M^{j-1} M M^{n-j} v} \tag{A}
 \end{aligned}$$

# Proof of the LSR Derivative Formula (1)

Let  $M = M(\theta^*)$  be non-neg. and  $M' = M'(\theta^*)$  for  $\theta^* \in D$ . Let  $a, b$  be left/right eigenvectors satisfying  $a^T M = \rho(M)a^T$  and  $Mb = \rho(M)b$ .

$$\begin{aligned}
 f'_n(\theta^*) &= \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^n M(\theta_i) \right) v \right) \Bigg|_{(\theta_1, \dots, \theta_n) = (\theta^*, \dots, \theta^*)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \log \left( u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v \right) \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)}
 \end{aligned}$$

$$\text{note: } \frac{d}{d\theta} x^T M(\theta) y = \sum_{k,l} x_k \frac{d}{d\theta} M_{k,l}(\theta) y_l = x^T M'(\theta) y$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{j=1}^n \frac{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M'(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v}{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v} \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{u^T M^{j-1} M' M^{n-j} v}{u^T M^{j-1} M M^{n-j} v} \tag{A}
 \end{aligned}$$

# Proof of the LSR Derivative Formula (1)

Let  $M = M(\theta^*)$  be non-neg. and  $M' = M'(\theta^*)$  for  $\theta^* \in D$ . Let  $a, b$  be left/right eigenvectors satisfying  $a^T M = \rho(M)a^T$  and  $Mb = \rho(M)b$ .

$$f'_n(\theta^*) = \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^n M(\theta_i) \right) v \right) \Bigg|_{(\theta_1, \dots, \theta_n) = (\theta^*, \dots, \theta^*)}$$

$$= \frac{1}{n} \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \log \left( u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v \right) \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)}$$

note:  $\frac{d}{d\theta} x^T M(\theta) y = \sum_{k,l} x_k \frac{d}{d\theta} M_{k,l}(\theta) y_l = x^T M'(\theta) y$

$$= \frac{1}{n} \sum_{j=1}^n \frac{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M'(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v}{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^n M(\theta_i) \right) v} \Bigg|_{\theta_1^n = (\theta^*, \dots, \theta^*)}$$

$$= \frac{1}{n} \sum_{j=1}^n \frac{u^T M^{j-1} M' M^{n-j} v}{u^T M^{j-1} M M^{n-j} v} \tag{A}$$

# Proof of the LSR Derivative Formula (2)

Since  $M$  has a simple top eigenvalue,  $\exists \gamma < 1$  such that

$$\frac{u^T M^{j-1}}{\|u^T M^{j-1}\|} = a^T + O(\gamma^{j-1}) \quad \frac{M^{n-j} v}{\|M^{n-j} v\|} = b + O(\gamma^{n-j})$$

Focusing on the interior values of the sum in (A) gives

$$f'_n(\theta^*) = O\left(\frac{\lfloor (\ln n)^2 \rfloor}{n} \frac{\|M'\| (a^T b)}{\rho(M) \frac{u^T b}{\|u\|} \frac{a^T v}{\|v\|}}\right) + \frac{1}{n} \sum_{j=\lfloor (\ln n)^2 \rfloor + 1}^{n - \lfloor (\ln n)^2 \rfloor} \frac{a^T M^j b + O(\gamma^{(\ln n)^2}) \|M'\|}{a^T M^j b + O(\gamma^{(\ln n)^2}) \|M\|}$$

Since  $f_n(\theta)$  and  $f'_n(\theta)$  converge uniformly for all  $\theta \in D$ , we find that

$$\left. \frac{d}{d\theta} \log \rho(M(\theta)) \right|_{\theta=\theta^* \in D} = \frac{a^T M'(\theta^*) b}{a^T M(\theta^*) b}.$$

## Proof of the LSR Derivative Formula (2)

Since  $M$  has a simple top eigenvalue,  $\exists \gamma < 1$  such that

$$\frac{u^T M^{j-1}}{\|u^T M^{j-1}\|} = a^T + O(\gamma^{j-1}) \quad \frac{M^{n-j}v}{\|M^{n-j}v\|} = b + O(\gamma^{n-j})$$

Focusing on the interior values of the sum in (A) gives

$$f'_n(\theta^*) = O\left(\frac{\lfloor (\ln n)^2 \rfloor}{n} \frac{\|M'\| (a^T b)}{\rho(M) \frac{u^T b}{\|u\|} \frac{a^T v}{\|v\|}}\right) + \frac{1}{n} \sum_{j=\lfloor (\ln n)^2 \rfloor + 1}^{n - \lfloor (\ln n)^2 \rfloor} \frac{a^T M^j b + O(\gamma^{(\ln n)^2}) \|M'\|}{a^T M^j b + O(\gamma^{(\ln n)^2}) \|M\|}$$

Since  $f_n(\theta)$  and  $f'_n(\theta)$  converge uniformly for all  $\theta \in D$ , we find that

$$\left. \frac{d}{d\theta} \log \rho(M(\theta)) \right|_{\theta=\theta^* \in D} = \frac{a^T M'(\theta^*) b}{a^T M(\theta^*) b}.$$

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# The Backward Blackwell Measure

## Properties of the Backward Recursion

$$[\beta_{i-1}]_q \triangleq \frac{1}{\pi_q} \Pr(Q_{i-1} = q | Y_{i-1}^n) = \frac{[M(Y_{i-1})\beta_i]_q}{\pi M(Y_{i-1})\beta_i}$$

- Consider the backward stationary measures:

$$\nu_q(B) \triangleq \lim_{i \rightarrow -\infty} \Pr(Q_i = q, \beta_i \in B) \quad (\text{Backward Furstenberg})$$

$$\nu(B) \triangleq \sum_{q \in \mathcal{Q}} \nu_q(A) = \lim_{i \rightarrow -\infty} \Pr(\beta_i \in B) \quad (\text{Backward Blackwell})$$

- Since  $\beta$  is defined slightly differently than  $\alpha$ , consistency gives

$$\nu_q(\mathbf{b}) = \Pr(\beta_* = \mathbf{b}) \Pr(Q_* = q | \beta_* = \mathbf{b}) = \nu(\mathbf{b}) \pi_q[\mathbf{b}]_q$$

## Backward Simulation Process: Markov Chain $\{\beta_i\}_{i \leq n}$

- Pick  $q_i \sim [\beta_i]_q \pi_q$ , choose  $q_{i-1}$ , generate  $y_{i-1}$ , and compute  $\beta_{i-1}$

# Simple Formula for the Entropy Rate Derivative

## Using the Derivative Method

$$\begin{aligned}
 g_n(\theta_1, \dots, \theta_n) &= -\frac{1}{n} \sum_{y_1^n \in \mathcal{Y}^n} \Pr(Y_1^n = y_1^n; \theta_1^n) \log \Pr(Y_1^n = y_1^n; \theta_1^n) \\
 &= -\frac{1}{n} \sum_{y_1^n \in \mathcal{Y}^n} \pi \left( \prod_{i=1}^n M_{\theta_i}(y_i) \right) \mathbf{1} \cdot \log \left[ \pi \left( \prod_{i=1}^n M_{\theta_i}(y_i) \right) \mathbf{1} \right]
 \end{aligned}$$

implies that  $\lim_{n \rightarrow \infty} f_n(\theta) = H(\mathcal{Y}; \theta)$  and  $\frac{d}{d\theta} H(\mathcal{Y}; \theta)$  is

$$\int d\mu(\alpha) \int d\nu(\beta) \sum_{y \in \mathcal{Y}} [\alpha^T M'(y) \beta \log(\alpha^T M(y) \beta) + \alpha^T M'(y) \beta]$$

## Conditions Required for Hidden Markov Process

- Let the HMP parameters vary smoothly with the parameter  $\theta$
- F/B Blackwell Measures converge exponentially for  $\theta \in D \subset \mathbb{R}$ 
  - Easily shown under (R1) or (S1) conditions above

# Proof of Entropy Rate Derivative (1)

For  $\theta^* \in D$ , we have

$$f'_n(\theta^*) = -\frac{1}{n} \sum_{j=1}^n \frac{d}{d\theta_j} \sum_{\mathbf{y}_1^n \in \mathcal{Y}^n} \pi \left( \prod_{i=1}^n M_{\theta_i}(\mathbf{y}_i) \right) \mathbf{1} \cdot \log \left[ \pi \left( \prod_{i=1}^n M_{\theta_i}(\mathbf{y}_i) \right) \mathbf{1} \right]$$

$$\stackrel{(a)}{=} -\frac{1}{n} \sum_{j=1}^n \frac{d}{d\theta_j} \sum_{\mathbf{y}_1^n \in \mathcal{Y}^n} \pi M(\mathbf{y}_1^n) \mathbf{1} \cdot \log \left[ \frac{\pi M(\mathbf{y}_1^{j-1}) M(\mathbf{y}_j) M(\mathbf{y}_{j+1}^n) \mathbf{1}}{(\pi M(\mathbf{y}_1^{j-1}) \mathbf{1}) (\pi M(\mathbf{y}_{j+1}^n) \mathbf{1})} \right],$$

where (a) follows from the fact that

$$\frac{d}{d\theta_j} \sum_{\mathbf{y}_1^n \in \mathcal{Y}^n} \pi M(\mathbf{y}_1^n) \mathbf{1} \cdot \log \left[ \pi M(\mathbf{y}_1^{j-1}) \mathbf{1} \right] = \frac{d}{d\theta_j} \sum_{\mathbf{y}_1^{j-1} \in \mathcal{Y}^{j-1}} \pi M(\mathbf{y}_1^{j-1}) \mathbf{1} \cdot \log \left[ \pi M(\mathbf{y}_1^{j-1}) \mathbf{1} \right] = 0$$

# Proof of Entropy Rate Derivative (2)

Let  $U_j(A) \triangleq \left\{ y_1^{j-1} \in \mathcal{Y}^{j-1} \mid \alpha_j = \frac{\pi M(y_1^{j-1})}{\pi M(y_1^{j-1})\mathbf{1}} \in A \right\}$  and  $\mu^{(j)}(A) \triangleq \Pr(Y_1^{j-1} \in U_j(A))$

Let  $V_j(B) \triangleq \left\{ y_j^n \in \mathcal{Y}^{n-j+1} \mid \beta_j = \frac{\pi M(y_j^n)}{\pi M(y_j^n)\mathbf{1}} \in B \right\}$  and  $\nu^{(j)}(B) \triangleq \Pr(Y_j^n \in V_j(B))$

$\mu^{(j)}(\cdot), \nu^{(j)}(\cdot)$  are probability measures on  $E = \mathbb{R}^{|\mathcal{Q}|}$  for  $\alpha_j, \beta_j$

$$\begin{aligned}
 &= -\frac{1}{n} \sum_{j=1}^n \frac{d}{d\theta_j} \sum_{y_1^n \in \mathcal{Y}^n} \overbrace{\pi M(y_1^{j-1})}^{\alpha_j [\pi M(y_1^{j-1})\mathbf{1}]} M(y_j) \overbrace{M(y_{j+1}^n)\mathbf{1}}^{\beta_{j+1} [\pi M(y_{j+1}^n)\mathbf{1}]} \log \left[ \frac{\overbrace{\pi M(y_1^{j-1})}^{\alpha_j}}{\pi M(y_1^{j-1})\mathbf{1}} M(y_j) \frac{\overbrace{M(y_{j+1}^n)\mathbf{1}}^{\beta_{j+1}}}{\pi M(y_{j+1}^n)\mathbf{1}} \right] \\
 &= -\frac{1}{n} \sum_{j=1}^n \frac{d}{d\theta_j} \int_E d\mu^{(j)}(\alpha) \int_E d\nu^{(j+1)}(\beta) \sum_{y_j \in \mathcal{Y}} \alpha M(y_j) \beta \log(\alpha M(y_j) \beta) \\
 &= -\frac{1}{n} \sum_{j=1}^n \int_E d\mu^{(j)}(\alpha) \int_E d\nu^{(j+1)}(\beta) \sum_{y_j \in \mathcal{Y}} \left[ \alpha^T M'(y_j) \beta \log(\alpha^T M(y_j) \beta) + \alpha^T M'(y_j) \beta \right]
 \end{aligned}$$

# Proof of Entropy Rate Derivative (3)

- The final step ignores terms within  $(\ln n)^2$  of the edges
  - Exponential convergence gives:  $\int |d\mu(\alpha) - d\mu^{(j)}(\alpha)| f(\alpha) \leq C\gamma^j$
  - So, error decays faster than polynomial:  $\gamma^{(\ln n)^2} = n^{\ln n \cdot \ln \gamma}$

$$-\frac{1}{n} \sum_{j=1}^n \int_E d\mu^{(j)}(\alpha) \int_E d\nu^{(j+1)}(\beta) \sum_{y_j \in \mathcal{Y}} [\alpha^T M'(y_j) \beta \log(\alpha^T M(y_j) \beta) + \alpha^T M'(y_j) \beta]$$

converges to

$$\frac{d}{d\theta} H(\mathcal{Y}; \theta) = \int_E d\mu(\alpha) \int_E d\nu(\beta) \sum_{y \in \mathcal{Y}} [\alpha^T M'(y) \beta \log(\alpha^T M(y) \beta) + \alpha^T M'(y) \beta]$$

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - **Capacity of a FSC**
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# The Derivative of Capacity for a FSC

A family of FSCs which varies smoothly in parameter  $\theta$

- The achievable rate depends on the input distribution
  - A Markov- $m$  input dist. is defined by vector  $\vec{P}$  of  $|\mathcal{X}|^m$  values
  - Let the optimal input distribution be  $\vec{P}(\theta)$

The Achievable Rate is  $\mathcal{I}(\theta, \vec{P})$  for input  $\vec{P}$

- Expanding this function, with gradient vector  $\mathcal{I}'_P(\theta, P)$ , gives

$$d\mathcal{I}(\theta, \vec{P}) = \mathcal{I}'_{\theta}(\theta, \vec{P}) d\theta + \mathcal{I}'_P(\theta, P) \cdot d\vec{P}$$

- The optimality of  $\vec{P}(\theta)$  implies  $\mathcal{I}'_P(\theta, \vec{P}(\theta)) \cdot d\vec{P} = 0$  for any  $d\vec{P}$  satisfying  $d\vec{P} \cdot \mathbf{1} = 0$  (i.e., the sum of  $\vec{P}(\theta)$  is a constant)
- So, the derivative of capacity is a derivative of entropy rates

$$\frac{d}{d\theta} \mathcal{C}(\theta) = \frac{d}{d\theta} \mathcal{I}(\theta, \vec{P}(\theta)) = \mathcal{I}'_{\theta}(\theta, \vec{P})$$

# Example: BSC( $\varepsilon$ ) with (0,1) RLL Constraint

## High Noise Regime $\varepsilon = 1/2 - \sqrt{\theta}$

- Standard binary symmetric channel with an input constraint
  - Input cannot have two 1s in a row (i.e., two-state input process)
  - $\Pr(X_{t+1}=j|X_t=i) = p_{ij}$  with  $p_{11} = 0$ ,  $\Pr(X_t=i) = \pi_i$  with  $\pi_0 = \frac{1}{2-p_{00}}$
- Rate is zero at  $\theta = 0$ , so  $\mathcal{I}(\theta, p_{00}) = \mathcal{I}'_{\theta}(0, p_{00})\theta + o(\theta)$
- Using  $H(\mathcal{Y}|\mathcal{X}) = h(\varepsilon)$  and the derivative of  $H(\mathcal{Y})$  gives

$$\mathcal{I}'_{\theta}(0, p_{00}) = \frac{8}{\ln 2} \left[ \frac{1-p_{00}}{2-p_{00}} - \left( \frac{1-p_{00}}{2-p_{00}} \right)^2 \right]$$

## Optimal Input Distribution

- Slope of expansion (at  $p_{00} = 0$ ) matches the unconstrained BSC
  - **Markov-1 achieves capacity** and opt. slope is  $\mathcal{I}'_{\theta}(0, 0) = \frac{2}{\ln 2}$
  - Codewords are alternating 01 sequences with an occasional 00

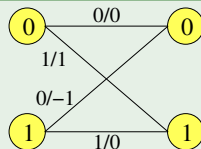
# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 Mixing Conditions and Forgetting
  - State Mixing vs. Process Mixing

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



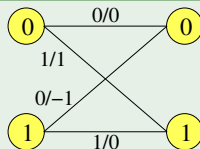
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$							
$\tilde{Y}_i$							
$Y_i$							
$Pr(X_i = 0   Y_1^{i-1}, X_0)$							
$Pr(X_i = 1   Y_1^{i-1}, X_0)$							

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



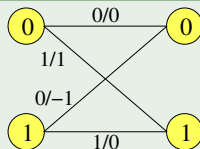
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0						
$\tilde{Y}_i$							
$Y_i$							
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1						
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0						

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



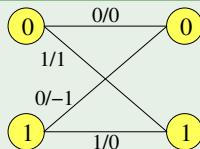
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0	1					
$\tilde{Y}_i$		1					
$Y_i$		1					
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1	0					
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0	1					

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



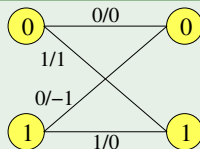
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0	1	1				
$\tilde{Y}_i$		1	0				
$Y_i$		1	?				
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1	0	$\frac{1}{2}$				
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0	1	$\frac{1}{2}$				

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



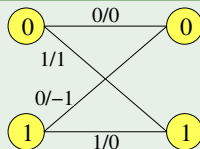
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0	1	1	0			
$\tilde{Y}_i$		1	0	-1			
$Y_i$		1	?	-1			
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1	0	$\frac{1}{2}$	1			
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0	1	$\frac{1}{2}$	0			

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



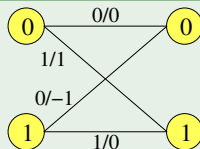
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0	1	1	0	0		
$\tilde{Y}_i$		1	0	-1	0		
$Y_i$		1	?	-1	0		
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1	0	$\frac{1}{2}$	1	1		
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0	1	$\frac{1}{2}$	0	0		

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



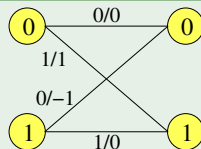
## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0	1	1	0	0	1	
$\tilde{Y}_i$		1	0	-1	0	1	
$Y_i$		1	?	-1	0	?	
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1	0	$\frac{1}{2}$	1	1	$\frac{1}{2}$	
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0	1	$\frac{1}{2}$	0	0	$\frac{1}{2}$	

# The Dicode Erasure Channel (DEC)

## Not Very Realistic, but Solvable

- Filter Output:  $\tilde{Y}_i = X_i - X_{i-1}$
- Channel Output:  $Y_i = \begin{cases} \tilde{Y}_i & \text{Prob. } 1 - \epsilon \\ ? & \text{Prob. } \epsilon \end{cases}$



## Example Sequences

$i$	0	1	2	3	4	5	6
$X_i$	0	1	1	0	0	1	0
$\tilde{Y}_i$		1	0	-1	0	1	-1
$Y_i$		1	?	-1	0	?	-1
$Pr(X_i = 0   Y_1^{i-1}, X_0)$	1	0	$\frac{1}{2}$	1	1	$\frac{1}{2}$	1
$Pr(X_i = 1   Y_1^{i-1}, X_0)$	0	1	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0

# The Forward Recursion for the DEC (1)

## The Markov Chain for $\{\alpha_i\}$

- Each  $\alpha_i = [\alpha_i(0) \ \alpha_i(1)]$  satisfies  $\alpha_i(1) = 1 - \alpha_i(0)$ 
  - Distribution  $Pr(\alpha_i(0) = \alpha')$  supported on  $\alpha' \in \{0, \frac{1}{2}, 1\}$
  - Symmetry:  $Pr(\alpha_i(0) = 1) = Pr(\alpha_i(0) = 0)$

## Evolution of the Markov Chain

- Let  $\alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \implies$  "state known at decoder"
- Let  $\alpha_i \in U = [\frac{1}{2} \ \frac{1}{2}] \implies$  "state unknown at decoder"
- If  $Y_i = ?$  then  $\alpha_{i+1} \in U$  regardless of  $\alpha_i$
- If  $Y_i = 0$  then  $\alpha_{i+1} \in K$  iff  $\alpha_i \in K$
- If  $Y_i \in \{-1, 1\}$  then  $\alpha_{i+1} \in K$  regardless of  $\alpha_i$

# The Forward Recursion for the DEC (1)

## The Markov Chain for $\{\alpha_i\}$

- Each  $\alpha_i = [\alpha_i(0) \ \alpha_i(1)]$  satisfies  $\alpha_i(1) = 1 - \alpha_i(0)$ 
  - Distribution  $Pr(\alpha_i(0) = \alpha')$  supported on  $\alpha' \in \{0, \frac{1}{2}, 1\}$
  - Symmetry:  $Pr(\alpha_i(0) = 1) = Pr(\alpha_i(0) = 0)$

## Evolution of the Markov Chain

- Let  $\alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \implies$  “state known at decoder”
- Let  $\alpha_i \in U = [\frac{1}{2} \ \frac{1}{2}] \implies$  “state unknown at decoder”
- If  $Y_i = ?$  then  $\alpha_{i+1} \in U$  regardless of  $\alpha_i$
- If  $Y_i = 0$  then  $\alpha_{i+1} \in K$  iff  $\alpha_i \in K$
- If  $Y_i \in \{-1, 1\}$  then  $\alpha_{i+1} \in K$  regardless of  $\alpha_i$

# The Forward Recursion for the DEC (1)

## The Markov Chain for $\{\alpha_i\}$

- Each  $\alpha_i = [\alpha_i(0) \ \alpha_i(1)]$  satisfies  $\alpha_i(1) = 1 - \alpha_i(0)$ 
  - Distribution  $Pr(\alpha_i(0) = \alpha')$  supported on  $\alpha' \in \{0, \frac{1}{2}, 1\}$
  - Symmetry:  $Pr(\alpha_i(0) = 1) = Pr(\alpha_i(0) = 0)$

## Evolution of the Markov Chain

- Let  $\alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \implies$  “state known at decoder”
- Let  $\alpha_i \in U = [\frac{1}{2} \ \frac{1}{2}] \implies$  “state unknown at decoder”
- If  $Y_i = ?$  then  $\alpha_{i+1} \in U$  regardless of  $\alpha_i$
- If  $Y_i = 0$  then  $\alpha_{i+1} \in K$  iff  $\alpha_i \in K$
- If  $Y_i \in \{-1, 1\}$  then  $\alpha_{i+1} \in K$  regardless of  $\alpha_i$

# The Forward Recursion for the DEC (1)

## The Markov Chain for $\{\alpha_i\}$

- Each  $\alpha_i = [\alpha_i(0) \ \alpha_i(1)]$  satisfies  $\alpha_i(1) = 1 - \alpha_i(0)$ 
  - Distribution  $Pr(\alpha_i(0) = \alpha')$  supported on  $\alpha' \in \{0, \frac{1}{2}, 1\}$
  - Symmetry:  $Pr(\alpha_i(0) = 1) = Pr(\alpha_i(0) = 0)$

## Evolution of the Markov Chain

- Let  $\alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \implies$  “state known at decoder”
- Let  $\alpha_i \in U = [\frac{1}{2} \ \frac{1}{2}] \implies$  “state unknown at decoder”
- If  $Y_i = ?$  then  $\alpha_{i+1} \in U$  regardless of  $\alpha_i$
- If  $Y_i = 0$  then  $\alpha_{i+1} \in K$  iff  $\alpha_i \in K$
- If  $Y_i \in \{-1, 1\}$  then  $\alpha_{i+1} \in K$  regardless of  $\alpha_i$

# The Forward Recursion for the DEC (1)

## The Markov Chain for $\{\alpha_i\}$

- Each  $\alpha_i = [\alpha_i(0) \ \alpha_i(1)]$  satisfies  $\alpha_i(1) = 1 - \alpha_i(0)$ 
  - Distribution  $Pr(\alpha_i(0) = \alpha')$  supported on  $\alpha' \in \{0, \frac{1}{2}, 1\}$
  - Symmetry:  $Pr(\alpha_i(0) = 1) = Pr(\alpha_i(0) = 0)$

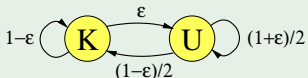
## Evolution of the Markov Chain

- Let  $\alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \implies$  “state known at decoder”
- Let  $\alpha_i \in U = [\frac{1}{2} \ \frac{1}{2}] \implies$  “state unknown at decoder”
- If  $Y_i = ?$  then  $\alpha_{i+1} \in U$  regardless of  $\alpha_i$
- If  $Y_i = 0$  then  $\alpha_{i+1} \in K$  iff  $\alpha_i \in K$
- If  $Y_i \in \{-1, 1\}$  then  $\alpha_{i+1} \in K$  regardless of  $\alpha_i$

# The Forward Recursion for the DEC (2)

## The Stationary Distribution

- $Pr(\alpha_{i+1} \in U | \alpha_i \in K) = Pr(Y_i = ? | \alpha_i \in K) = \epsilon$
- $Pr(\alpha_{i+1} \in K | \alpha_i \in U) = Pr(Y_i \in \{-1, 1\} | \alpha_i \in U) = (1 - \epsilon)/2$



- Stationary distribution:  $Pr(K) = \frac{1-\epsilon}{1+\epsilon}$  and  $Pr(U) = \frac{2\epsilon}{1+\epsilon}$

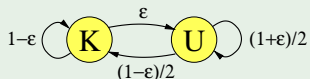
## Entropy Rate for Each Subset

- $H(Y_i | \alpha_i \in K) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{2}, 0, \frac{1}{2}\right) = h(\epsilon) + (1 - \epsilon)$
- $H(Y_i | \alpha_i \in U) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = h(\epsilon) + \frac{3}{2}(1 - \epsilon)$
- $h(p_1, \dots, p_k) \triangleq -\log\left(\sum_{i=1}^k p_i \log p_i\right)$  and  $h(p) \triangleq h(p, 1 - p)$

# The Forward Recursion for the DEC (2)

## The Stationary Distribution

- $Pr(\alpha_{i+1} \in U | \alpha_i \in K) = Pr(Y_i = ? | \alpha_i \in K) = \epsilon$
- $Pr(\alpha_{i+1} \in K | \alpha_i \in U) = Pr(Y_i \in \{-1, 1\} | \alpha_i \in U) = (1 - \epsilon)/2$



- Stationary distribution:  $Pr(K) = \frac{1-\epsilon}{1+\epsilon}$  and  $Pr(U) = \frac{2\epsilon}{1+\epsilon}$

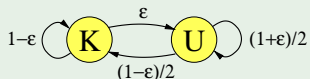
## Entropy Rate for Each Subset

- $H(Y_i | \alpha_i \in K) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{2}, 0, \frac{1}{2}\right) = h(\epsilon) + (1 - \epsilon)$
- $H(Y_i | \alpha_i \in U) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = h(\epsilon) + \frac{3}{2}(1 - \epsilon)$
- $h(p_1, \dots, p_k) \triangleq -\log\left(\sum_{i=1}^k p_i \log p_i\right)$  and  $h(p) \triangleq h(p, 1 - p)$

# The Forward Recursion for the DEC (2)

## The Stationary Distribution

- $Pr(\alpha_{i+1} \in U | \alpha_i \in K) = Pr(Y_i = ? | \alpha_i \in K) = \epsilon$
- $Pr(\alpha_{i+1} \in K | \alpha_i \in U) = Pr(Y_i \in \{-1, 1\} | \alpha_i \in U) = (1 - \epsilon)/2$



- Stationary distribution:  $Pr(K) = \frac{1 - \epsilon}{1 + \epsilon}$  and  $Pr(U) = \frac{2\epsilon}{1 + \epsilon}$

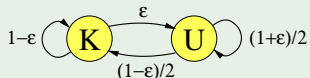
## Entropy Rate for Each Subset

- $H(Y_i | \alpha_i \in K) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{2}, 0, \frac{1}{2}\right) = h(\epsilon) + (1 - \epsilon)$
- $H(Y_i | \alpha_i \in U) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = h(\epsilon) + \frac{3}{2}(1 - \epsilon)$
- $h(p_1, \dots, p_k) \triangleq -\log\left(\sum_{i=1}^k p_i \log p_i\right)$  and  $h(p) \triangleq h(p, 1 - p)$

# The Forward Recursion for the DEC (2)

## The Stationary Distribution

- $Pr(\alpha_{i+1} \in U | \alpha_i \in K) = Pr(Y_i = ? | \alpha_i \in K) = \epsilon$
- $Pr(\alpha_{i+1} \in K | \alpha_i \in U) = Pr(Y_i \in \{-1, 1\} | \alpha_i \in U) = (1 - \epsilon)/2$



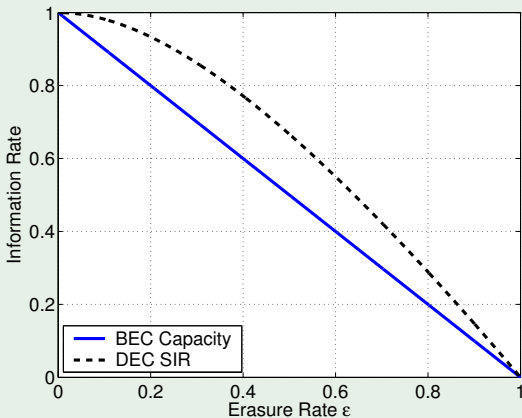
- Stationary distribution:  $Pr(K) = \frac{1-\epsilon}{1+\epsilon}$  and  $Pr(U) = \frac{2\epsilon}{1+\epsilon}$

## Entropy Rate for Each Subset

- $H(Y_i | \alpha_i \in K) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{2}, 0, \frac{1}{2}\right) = h(\epsilon) + (1 - \epsilon)$
- $H(Y_i | \alpha_i \in U) = h(\epsilon) + (1 - \epsilon)h\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = h(\epsilon) + \frac{3}{2}(1 - \epsilon)$
- $h(p_1, \dots, p_k) \triangleq -\log\left(\sum_{i=1}^k p_i \log p_i\right)$  and  $h(p) \triangleq h(p, 1 - p)$

# The Symmetric Information Rate of the DEC

## Results



$$C_{i.u.d.} = \frac{1-\epsilon}{1+\epsilon} (1-\epsilon) + \frac{2\epsilon}{1+\epsilon} \left( \frac{3(1-\epsilon)}{2} \right) = 1 - \frac{2\epsilon^2}{(1+\epsilon)}$$

# Extensions and Open Problems

## Note: Markov-1 Rate for the DEC

- $\{\alpha_i\}$  process becomes a countably infinite Markov chain
- Exact expression for rate is given by a infinite sum

## Open: Exact Expressions for Two-State Channels

- State probabilities represented by  $\alpha_i(0)$ ,  $\log \frac{\alpha_i(0)}{\alpha_i(1)}$ , etc...
- Let  $f_i(x)$  be the density time  $i$ 
  - Closed form recursion for  $f_{i+1}(x)$  in terms of  $f_i(x)$
- Solving this recursion enables closed form expressions
- For example: Dicode channel in AWGN

# Extensions and Open Problems

## Note: Markov-1 Rate for the DEC

- $\{\alpha_i\}$  process becomes a countably infinite Markov chain
- Exact expression for rate is given by a infinite sum

## Open: Exact Expressions for Two-State Channels

- State probabilities represented by  $\alpha_i(0)$ ,  $\log \frac{\alpha_i(0)}{\alpha_i(1)}$ , etc...
- Let  $f_i(x)$  be the density time  $i$ 
  - Closed form recursion for  $f_{i+1}(x)$  in terms of  $f_i(x)$
- Solving this recursion enables closed form expressions
- For example: Dicode channel in AWGN

# Outline

- 1 Introduction
  - Definition and Taxonomy of FSCs
  - The Capacity of a Finite-State Channel
  - Connections with Lyapunov Exponents
- 2 Derivatives
  - Motivating Example
  - Entropy Rate of a Hidden Markov Processes
  - Capacity of a FSC
- 3 Simple Example Channel
  - The Dicode Erasure Channel
- 4 **Mixing Conditions and Forgetting**
  - **State Mixing vs. Process Mixing**

# Mixing

## State vs. Process Mixing

- An HMP is “state mixing” if
  - The belief process  $\alpha_i$  forgets initial belief
  - Note: this is typically proven using Birkhoff contraction results
- An HMP is “process mixing” if
  - Conditional distributions forget initial state exponentially ( $|\gamma| < 1$ )

$$\left| \Pr(Y_i = y_i | Y_1^{i-1} = y_1^{i-1}) - \Pr(Y_i = y_i | Y_1^{i-1} = y_1^{i-1}, X_1) \right| \leq \gamma^i$$

- Mixing can occur in two ways
  - For all  $y_1^n$  (e.g., S1 strong mixing condition)
  - For almost all  $y_1^n$  (e.g., R1 rank 1 condition)
- It is easy for “state mixing” to fail, when process mixing occurs

# A Counterexample of Kaijser for State Mixing

$$P = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Let  $X_t$  the Markov chain associated with  $P$

- Mapping function  $\phi(x)$  maps states 1, 2, 3, 4 to new states 1, 2
  - $Y_t = \phi(X_t)$  is a function of a finite-state Markov chain
  - Defined by  $\phi(1) = \phi(2) = 1$  and  $\phi(3) = \phi(4) = 2$
- Analysis shows that all sequences are equally likely
  - Therefore, the output gives no information about state
  - Prior information on state is dominant and persists forever
  - But, process is immediately “Process Mixing”!

# Conclusions

- Introduction to Finite State Channels
- The Derivative of the Entropy Rate and Capacity
- The Dicode Erasure Channel
- State Mixing versus Process Mixing