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
What is a Finite-State Channel (FSC)?

**Qualitative Definition**

- A FSC is a model for communication systems with memory
- A probabilistic mapping from a sequence of inputs to a sequence of outputs
- Each output depends only on the current input and channel state instead of the entire history of inputs and channel states

**Mathematical Definition**

- Channel input $X_i \in \mathcal{X}$, output $Y_i \in \mathcal{Y}$, state $S_i \in \mathcal{S}$, and law

$$Pr(Y^n_1 = y^n_1|X^n_1 = x^n_1, S_1 = s_1) = \sum_{s_2^{n+1} \in \mathcal{S}^n} \prod_{i=1}^{n} P(y_i, s_{i+1}|x_i, s_i)$$

- where $P(y, s'|x, s) \triangleq Pr(Y_i = y, S_{i+1} = s'|X_i = x, S_i = s)$
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

\[ X \in \{0,1\} \xrightarrow{H(z)=1-z} Y \xrightarrow{\text{Z}} \]

0

0/0

1/1

0/-1

1

1/0
Application: Fading Channels

Channel Properties

- Discrete memoryless channel associated with each state
- State evolves independently of inputs

The Gilbert-Elliot Channel

- Channel is either in a “good” state or a “bad” state
  - \( p_{GG} = 1 - p_{GB} \) and \( p_{BB} = 1 - p_{BG} \) are the transition probabilities
- Noise variance (or error rate) is larger in bad state
- Optimum detector increases data rate by inferring state from received signal
# 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
- 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

\[ \bar{C} \triangleq \lim_{n \to \infty} \frac{1}{n} \max_{s_0 \in S} \max \{ I(X_1^n; Y_1^n | S_0 = s_0) \} \]

\[ C \triangleq \lim_{n \to \infty} \frac{1}{n} \max_{s_0 \in S} \min \{ I(X_1^n; Y_1^n | S_0 = s_0) \} \]

Sufficient Conditions for \( \bar{C} = C = \underline{C} \)

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

### Markov Input Process (memory $m$)

\[
\Pr(X_i = x' | X^{i-1}_{i-m} = x) = R(x'|x)
\]

### A Sequence of Lower Bounds on Capacity

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

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

\[
C_{i.u.d.} \leq L_0 \leq L_1 \leq L_3 \leq \ldots \leq C
\]

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

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

**Forward/Backward State Probability**

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

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

**The Forward Recursion**

Randomly generate the sequence \(y_1, y_2, \ldots\) and compute

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

\[ \frac{1}{\psi_{i+1}} \sum_{q' \in Q} \sum_{x \in \mathcal{X}} \Pr(y, Y_i = y \in Y_i^{i-1} = y_i^{i-1}) \]

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

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

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

\[
\mu(A) \triangleq \sum_{q \in Q} \mu_q(A) = \lim_{i \to \infty} \Pr(\alpha_i \in A) \quad \text{(Forward Blackwell)}
\]

Entropy Rate

\[
H(Y) = \sum_{q \in Q} \int_{\mathcal{M}(Q)} \sum_{y \in Y} \frac{\Pr(Q=q, \alpha, Y=y)}{\sum_{q' \in Q} \alpha(q')f_{q'}(y)} \log \sum_{q' \in Q} \alpha(q')f_{q'}(y) \, d\mu_q(\alpha)f_q(y)
\]

\[
f_q(y) = \Pr(Y_i = y | Q_i = q)
\]
Consistency of the APP Estimate

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

$$
\Pr(X = x, A = a) = \Pr(A = a) \Pr(X = x|A = a) \\
= \Pr(A = a) \Pr(X = x|Y) \\
= \Pr(A = a)[a]_x
$$

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

HMP Relationship Between Blackwell and Furstenberg Measures

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

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 $A$: $[A]_x = \Pr(X = x | Y)$

$$\Pr(X = x, A = a) = \Pr(A = a) \Pr(X = x | A = a)$$
$$= \Pr(A = a) \Pr(X = x | Y)$$
$$= \Pr(A = a)[a]_x$$

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

HMP Relationship Between Blackwell and Furstenberg Measures

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

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 $\alpha_{i+1}$
Consistency of the APP Estimate

Let $X, Y$ be discrete r.v. with random vector $A$: 

$$ [A]_x = \Pr(X = x | Y) $$

$$ \Pr(X = x, A = a) = \Pr(A = a) \Pr(X = x | A = a) $$

$$ = \Pr(A = a) \Pr(X = x | Y) $$

$$ = \Pr(A = a)[a]_x $$

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

HMP Relationship Between Blackwell and Furstenberg Measures

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

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 $|Q| \times |Q|$ matrix defined by

$$
[M(y)]_{q,q'} \triangleq \Pr(Y_i = y, Q_{i+1} = q' | Q_i = q)
$$

$$
= \sum_{x \in \mathcal{X}} P(y, S(q') | x, S(q)) R(x, \mathcal{X}(q))
$$

Products compute the multi-step transition-observation matrix

$$
[M(y_j^k)]_{q,q'} \triangleq \left[ \prod_{i=j}^k M(y_i) \right]_{q,q'} = \Pr(Y_j^k = y_j^k, Q_{k+1} = q' | Q_j = q)
$$
The Forward/Backward Recursions

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

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

Matrix Form of the Forward/Backward Recursions

$$\alpha_{i+1} = \frac{\alpha_i M(y_i)}{\alpha_i M(y_i)1} \quad \beta_{i-1} = \frac{M(y_{i-1})\beta_i}{\pi M(y_{i-1})\beta_i}$$

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

The Entropy Rate as a Lyapunov Exponent

$$\lim_{n \to \infty} \frac{1}{n} \log \left\| \prod_{i=1}^{n} M(y_i) \right\| \overset{a.s.}{=} -H(\mathcal{Y})$$

Formulation gives many results: convergence, continuity, CLT, ...
What is the Measure of a Man?

- An HMP is also a matrix function of Markov chain
  - Markov $X_1, X_2, \ldots$ 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

Of Blackwell and Furstenberg
Mixing Conditions

Connectivity

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

A FSC is *indecomposable* if \( A(Y) \) 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
Introduction

1. Definition and Taxonomy of FSCs
2. The Capacity of a Finite-State Channel
3. Connections with Lyapunov Exponents

Derivatives

2. Motivating Example
   1. Entropy Rate of a Hidden Markov Processes
   2. Capacity of a FSC

Simple Example Channel

3. The Dicode Erasure Channel

Mixing Conditions and Forgetting

4. State Mixing vs. Process Mixing
# The Derivative of the Entropy Rate $H(Y)$

## 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

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

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

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

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

The Total Derivative Method

Let \( f_n : \mathbb{R} \to \mathbb{R} \) and \( g_n : \mathbb{R}^n \to \mathbb{R} \) be sequences such that

\[
f_n(\theta) = g_n(\theta, \ldots, \theta)
\]

The derivative can be broken into the sum of \( n \) terms

\[
\frac{d}{d\theta} f_n(\theta) = \sum_{i=1}^{n} \left. \frac{\partial}{\partial \theta_i} g_n(\theta_1, \ldots, \theta_n) \right|_{(\theta_1, \ldots, \theta_n) = (\theta, \ldots, \theta)}
\]

Connection to Log Spectral Radius

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

for non-neg. \( M \) implies

\[
\lim_{n \to \infty} f_n(\theta) = \log \rho (M(\theta))
\]
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 $M b = \rho(M) b$.

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

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

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

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

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

(A)
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 $M b = \rho(M) b$.

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

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

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

\[
f'_n(\theta^*) = \sum_{j=1}^{n} \frac{\partial}{\partial \theta_j} \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^{n} M(\theta_i) \right) v \right) \bigg|_{(\theta_1, \ldots, \theta_n) = (\theta^*, \ldots, \theta^*)}
= \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta_j} \log \left( u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^{n} M(\theta_i) \right) v \right) \bigg|_{\theta^n = (\theta^*, \ldots, \theta^*)}
\]

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

\[
= \frac{1}{n} \sum_{j=1}^{n} \frac{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M'(\theta_j) \left( \prod_{i=j+1}^{n} M(\theta_i) \right) v}{u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^{n} M(\theta_i) \right) v} \bigg|_{\theta^n = (\theta^*, \ldots, \theta^*)}
= \frac{1}{n} \sum_{j=1}^{n} \frac{u^T M^{j-1} M' M^{n-j} v}{u^T M^{j-1} M M^{n-j} v}
\]
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 $M b = \rho(M) b$.

\[
f'_n(\theta^*) = \sum_{j=1}^{n} \frac{\partial}{\partial \theta_j} \frac{1}{n} \log \left( u^T \left( \prod_{i=1}^{n} M(\theta_i) \right) \nu \right) \bigg|_{(\theta_1, \ldots, \theta_n) = (\theta^*, \ldots, \theta^*)}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta_j} \log \left( u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M(\theta_j) \left( \prod_{i=j+1}^{n} M(\theta_i) \right) \nu \right) \bigg|_{\theta_1^n = (\theta^*, \ldots, \theta^*)}
\]

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

\[
= \frac{1}{n} \sum_{j=1}^{n} u^T \left( \prod_{i=1}^{j-1} M(\theta_i) \right) M'(\theta_j) \left( \prod_{i=j+1}^{n} M(\theta_i) \right) \nu \bigg|_{\theta_1^n = (\theta^*, \ldots, \theta^*)}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \frac{u^T M^{j-1} M' M^{n-j} \nu}{u^T M^{j-1} M M^{n-j} \nu}
\]

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

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

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

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

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

$$\frac{d}{d\theta} \log \rho(M(\theta)) \bigg|_{\theta=\theta^* \in D} = \frac{a^T M'(\theta^*) b}{a^T M(\theta^*) b}.$$
Proof of the LSR Derivative Formula (2)

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

$$
\frac{u^T M^{j-1}}{\|u^T M^{j-1}\|} = a^T + O\left(\gamma^{j-1}\right)
$$

$$
\frac{M^{n-j} v}{\|M^{n-j} v\|} = b + O\left(\gamma^{n-j}\right)
$$

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

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

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

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

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

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

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

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

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

\[ g_n (\theta_1, \ldots, \theta_n) = - \frac{1}{n} \sum_{y_1^n \in Y^n} \Pr (Y_1^n = y_1^n; \theta_1^n) \log \Pr (Y_1^n = y_1^n; \theta_1^n) \]

\[ = - \frac{1}{n} \sum_{y_1^n \in Y^n} \pi \left( \prod_{i=1}^{n} M_{\theta_i}(y_i) \right) \mathbf{1} \cdot \log \left[ \pi \left( \prod_{i=1}^{n} M_{\theta_i}(y_i) \right) \mathbf{1} \right] \]

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

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

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

$$f_n'(\theta^*) = -\frac{1}{n} \sum_{j=1}^{n} \frac{d}{d \theta_j} \sum_{y_1^n \in \mathcal{Y}^n} \pi \left( \prod_{i=1}^{n} M_{\theta_i}(y_i) \right) 1 \cdot \log \left[ \pi \left( \prod_{i=1}^{n} M_{\theta_i}(y_i) \right) 1 \right]$$

$$(a) \quad -\frac{1}{n} \sum_{j=1}^{n} \frac{d}{d \theta_j} \sum_{y_1^n \in \mathcal{Y}^n} \pi M(y_1^n) 1 \cdot \log \left[ \frac{\pi M(y_1^{j-1}) M(y_j) M(y_{j+1}^n) 1}{(\pi M(y_1^{j-1}) 1) (\pi M(y_{j+1}^n) 1)} \right],$$

where $(a)$ follows from the fact that

$$\frac{d}{d \theta_j} \sum_{y_1^n \in \mathcal{Y}^n} \pi M(y_1^n) 1 \cdot \log \left[ \pi M(y_1^{j-1}) 1 \right] = \frac{d}{d \theta_j} \sum_{y_1^{j-1} \in \mathcal{Y}^{j-1}} \pi M(y_1^{j-1}) 1 \cdot \log \left[ \pi M(y_1^{j-1}) 1 \right] = 0$$
Proof of Entropy Rate Derivative (2)

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

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

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

\[
\begin{align*}
&= -\frac{1}{n} \sum_{j=1}^{n} \frac{d}{d\theta_j} \left( \sum_{y_1^n \in \mathcal{Y}^n} \alpha_j \frac{\pi M(y_1^{j-1})1}{\pi M(y_1^{j-1}) M(y_j) M(y_{j+1})1} \log \left( \frac{\alpha_j}{\pi M(y_1^{j-1})1} M(y_j) \frac{\beta_{j+1}}{\pi M(y_j^{n+1})1} \right) \right) \\
&= -\frac{1}{n} \sum_{j=1}^{n} \frac{d}{d\theta_j} \int_E d\mu^{(j)}(\alpha) \int_E d\nu^{(j+1)}(\beta) \sum_{y_j \in \mathcal{Y}} \alpha M(y_j) \beta \log (\alpha M(y_j) \beta) \\
&= -\frac{1}{n} \sum_{j=1}^{n} \int_E d\mu^{(j)}(\alpha) \int_E d\nu^{(j+1)}(\beta) \sum_{y_j \in \mathcal{Y}} [\alpha^T \mathcal{M}'(y_j) \beta \log (\alpha^T \mathcal{M}(y_j) \beta) + \alpha^T \mathcal{M}'(y_j) \beta]
\end{align*}
\]
Proof of Entropy Rate Derivative (3)

- The final step ignores terms within \((\ln n)^2\) of the edges
  - Exponential convergence gives: \(\int \left| d\mu(\alpha) - d\mu^{(j)}(\alpha) \right| f(\alpha) \leq C\gamma^j\)
  - 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

\[
\frac{d}{d\theta} H(\mathcal{Y}; \theta) = \int_E d\mu(\alpha) \int_E d\nu(\beta) \sum_{y \in \mathcal{Y}} [\alpha^T M'(y)\beta \log(\alpha^T M(y)\beta) + \alpha^T M'(y)\beta]
\]
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 $|X|^m$ 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
  $$d\mathcal{I} \left( \theta, \vec{P} \right) = \mathcal{I}'_\theta \left( \theta, \vec{P} \right) d\theta + \mathcal{I}'_P \left( \theta, P \right) \cdot d\vec{P}$$

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

  - So, the derivative of capacity is a derivative of entropy rates
  $$\frac{d}{d\theta} C(\theta) = \frac{d}{d\theta} \mathcal{I} \left( \theta, \vec{P}(\theta) \right) = \mathcal{I}'_\theta \left( \theta, \vec{P} \right)$$
Example: BSC(ε) 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 $I(\theta, p_{00}) = I_0'(0, p_{00}) \theta + o(\theta)$
- Using $H(Y|X) = h(\varepsilon)$ and the derivative of $H(Y)$ gives
  $$I_0'(0, p_{00}) = \frac{8}{\ln 2} \left[ \frac{1 - p_{00}}{2 - p_{00}} - \left( \frac{1 - p_{00}}{2 - p_{00}} \right)^2 \right]$$

Optimal Input Distribution

- Slope of expansion (at $p_{00} = 0$) matches the unconstrained BSC
  - Markov-1 achieves capacity and opt. slope is $I_0'(0, 0) = \frac{2}{\ln 2}$
  - 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

<table>
<thead>
<tr>
<th>( i )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \tilde{Y}_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \tilde{Y}_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Pr(X_i = 0</td>
<td>\tilde{Y}_{1}^{i-1}, X_0) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Pr(X_i = 1</td>
<td>\tilde{Y}_{1}^{i-1}, X_0) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i$</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{Y}_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 0</td>
<td>Y_1^{i-1}, X_0)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 1</td>
<td>Y_1^{i-1}, X_0)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>( i )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_i )</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \tilde{Y}_i )</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y_i )</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Pr(X_i = 0</td>
<td>Y_1^{i-1}, X_0) )</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Pr(X_i = 1</td>
<td>Y_1^{i-1}, X_0) )</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>( i )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_i )</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \tilde{Y}_i )</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y_i )</td>
<td>1</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Pr(X_i = 0</td>
<td>Y_{1}^{i-1}, X_0) )</td>
<td>1</td>
<td>0</td>
<td>( \frac{1}{2} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Pr(X_i = 1</td>
<td>Y_{1}^{i-1}, X_0) )</td>
<td>0</td>
<td>1</td>
<td>( \frac{1}{2} )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{Y}_i$</td>
<td></td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_i$</td>
<td></td>
<td></td>
<td>?</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 0</td>
<td>Y_1^{i-1}, X_0)$</td>
<td>1</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 1</td>
<td>Y_1^{i-1}, X_0)$</td>
<td>0</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{Y}_i$</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_i$</td>
<td>1</td>
<td>?</td>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 0</td>
<td>Y_1^{i-1}, X_0)$</td>
<td>1</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 1</td>
<td>Y_1^{i-1}, X_0)$</td>
<td>0</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$\tilde{Y}_i$</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_i$</td>
<td>1</td>
<td>?</td>
<td>-1</td>
<td>0</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Pr(X_i = 0</td>
<td>Y_{i-1}^1, X_0)$</td>
<td>1</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$Pr(X_i = 1</td>
<td>Y_{i-1}^1, X_0)$</td>
<td>0</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
</tr>
</tbody>
</table>
The Dicode Erasure Channel (DEC)

Not Very Realistic, but Solvable

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

Example Sequences

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\widetilde{Y}_i$</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>$Y_i$</td>
<td>1</td>
<td>?</td>
<td>-1</td>
<td>0</td>
<td>?</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

$Pr(X_i = 0|Y_1^{i-1}, X_0)$

|       | 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 \( \alpha_i \in K = \{[1 \; 0], [0 \; 1]\} \) \( \rightarrow \) “state known at decoder”
- Let \( \alpha_i \in U = \left[ \frac{1}{2} \; \frac{1}{2} \right] \) \( \rightarrow \) “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 $\alpha_i \in K = \{[1 \; 0], [0 \; 1]\} \implies \text{“state known at decoder”}$
- Let $\alpha_i \in U = \left[\frac{1}{2} \; \frac{1}{2}\right] \implies \text{“state unknown at decoder”}$
- 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 \( \alpha_i = \begin{bmatrix} \alpha_i(0) \\ \alpha_i(1) \end{bmatrix} \) 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 \( \alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \mapsto \) “state known at decoder”
- Let \( \alpha_i \in U = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \mapsto \) “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 \(\alpha_i \in K = \{[1 \ 0], [0 \ 1]\}\) \(\rightarrow\) “state known at decoder”
- Let \(\alpha_i \in U = [\frac{1}{2} \ \frac{1}{2}]\) \(\rightarrow\) “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 \( \alpha_i \in K = \{[1 \ 0], [0 \ 1]\} \implies \text{“state known at decoder”} \)
- Let \( \alpha_i \in U = \left[\frac{1}{2} \ \frac{1}{2}\right] \implies \text{“state unknown at decoder”} \)
- 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

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

![Diagram](K_U_diagram.png)

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

Entropy Rate for Each Subset

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

The Stationary Distribution

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

\[
\begin{align*}
K & \quad \text{K} \\
\varepsilon & \quad \text{U} \\
(1-\varepsilon)/2 & \quad (1+\varepsilon)/2
\end{align*}
\]

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

Entropy Rate for Each Subset

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

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

The Stationary Distribution

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

\[ \begin{array}{c}
K \\
(1-\epsilon)/2 \\
\epsilon \\
(1+\epsilon)/2 \\
U \\
\end{array} \]

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

Entropy Rate for Each Subset

- \( H(Y_i | \alpha_i \in K) = h(\epsilon) + (1 - \epsilon)h \left( \frac{1}{2}, 0, \frac{1}{2} \right) = h(\epsilon) + (1 - \epsilon) \)
- \( H(Y_i | \alpha_i \in U) = h(\epsilon) + (1 - \epsilon)h \left( \frac{1}{4}, \frac{1}{2}, \frac{1}{4} \right) = h(\epsilon) + \frac{3}{2}(1 - \epsilon) \)
- \( h(p_1, \ldots, p_k) \triangleq -\log \left( \sum_{i=1}^{k} p_i \log p_i \right) \) 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 an 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 an 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
Introduction
- Definition and Taxonomy of FSCs
- The Capacity of a Finite-State Channel
- Connections with Lyapunov Exponents

Derivatives
- Motivating Example
- Entropy Rate of a Hidden Markov Processes
- Capacity of a FSC

Simple Example Channel
- The Dicode Erasure Channel

Mixing Conditions and Forgetting
- State Mixing vs. Process 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

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”!

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

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