On the Capacity of Finite-State Channels[1]

Examples Via the Dicode Erasure Channel

H. D. Pfister
L'Ecole Polytechnique Fédérale de Lausanne

Technion Institute, Israel
March 31st, 2005

[1] Joint work with Joseph Soriaga and Paul Siegel.

Outline

History

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 (Kavcic 2001)
- An Upper Bound on Capacity (Vontobel et al. 2001)
- Connections with Random Matrices (Holliday et al. 2003)

What is a Finite-State Channel?

Qualitative Definition

- A FSC is a probabilistic mapping from a semi-infinite sequence of inputs to a semi-infinite sequence of outputs
- Each output depends only on the current input and channel state instead of the entire past

Mathematical Definition

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

$$Pr(Y_1^n \in y_1^n | X_1^n = x_1^n, S_1 = s_1) = \sum_{s_2^{n+1} \in \mathcal{S}^n} \prod_{i=1}^{n} P(x_i, y_i, s_i, s_{i+1})$$

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

Types of Finite-State Channels

Taxonomy

- Deterministic State FSCs: $P(x, \mathcal{Y}, s, s') = \delta_{f(x,s),s'}$
  - Next state given by $f(x, s)$ with current state s and input x
  - Example: Intersymbol Interference (ISI) Channels
- Independent State FSCs: $P(x, \mathcal{Y}, s, s') = P(x', \mathcal{Y}, s, s')$
  - Distribution of next state is independent of input
  - Example: Fading Channels (e.g., Gilbert-Elliot Channel)
- General FSCs
  - Distribution of next state depends on input
  - Example: Media noise in magnetic recording

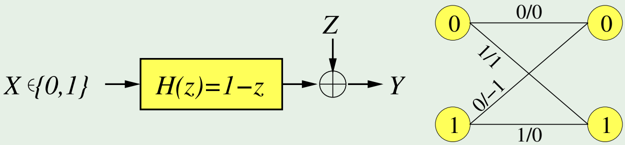
Application: Magnetic Storage

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 simplicity
- State given by last input, edges labelled by input/output



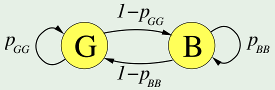
Application: Fading Channels

Channel Properties

- Discrete input alphabet can be used for any real system
- State evolves independently of inputs

The Gilbert-Elliot Channel

- Channel is either in a “good” state or a “bad” state
- Noise variance (or error rate) larger in bad state
- Detector must infer state from received signal



Dependence on Initial State

Upper and Lower Capacity

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

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

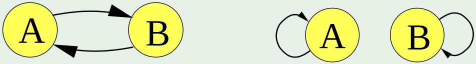
Sufficient Conditions for $\overline{C} = C = \underline{C}$

- Indecomposable: Channel forgets initial state for any input
    - Namely $|Pr(S_n|X_1^n = x_1^n, S_0) - Pr(S_n|X_1^n = x_1^n)| \to 0$
- Finite Memory: Channel state is a function of last ν inputs

Some FSCs with $\overline{C} \neq \underline{C}$

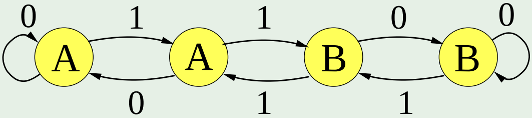
Classic Examples

- State diagram is periodic or decomposable



A Connected Deterministic Channel with $\overline{C} \neq \underline{C}$

- In this case, $\overline{C} = \underline{C}$ if the optimum input dist. includes “1 1”



Achievable Rates

Mutual Information Rate

$$I(\mathcal{X};\mathcal{Y}) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n | S_0 = s_0)$$

- For each s_0, rate is achievable via random coding
- Limit independent of s_0 if input-channel process is ergodic
  - In some cases, ergodicity depends on the input distribution

Symmetric Information Rate (SIR)

Max rate for independent uniformly distributed (i.u.d.) inputs

$$C_{i.u.d.} = \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n) \quad \text{with} \quad Pr(X_1^n) = |\mathcal{X}|^{-n}$$

Entropy Rates

Entropy Rate

For a process $\{Z_i\}_{i\geq 1}$ with $Z_i \in \mathcal{Z}$, we abuse notation and write

$$H(\mathcal{Z}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(Z_1^n)$$

Shannon-McMillan-Breiman Theorem

$$\lim_{n\to\infty} -\frac{1}{n} \log Pr(Z_1^n) \overset{a.s.}{=} H(\mathcal{Z})$$

Mutual Information Rate

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X})$$

The Input Distribution

Markov Input Process (memory m)

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

A Sequence of Lower Bounds on Capacity

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

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

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

Combine 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 $s \in S$ and $\mathbf{x} \in \mathcal{X}^m$
  - Scalar $\mathcal{S}$-projection: For $q = (s, \mathbf{x})$, we have $q^{\mathcal{S}} = s$
  - Scalar $\mathcal{X}^m$-projection: For $q = (s, \mathbf{x})$, we have $q^{\mathcal{X}} = \mathbf{x}$

The APP-BCJR Algorithm

Forward State Probability

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

The Forward Recursion

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

$$\alpha_{i+1}(q) = \frac{1}{a_{i+1}} \sum_{r \in \mathcal{Q}} \alpha_i(r) \sum_{x \in \mathcal{X}} P(x, y_i, r^{\mathcal{S}}, q^{\mathcal{S}}) W(x, r^{\mathcal{X}})$$

Normalization constant $a_{i+1} = Pr(Y_i = y_i | Y_1^{i-1} = y_1^{i-1})$

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

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 $\alpha_{i+1}(q)$

Stationary Distribution

$$\pi(A, B) = \lim_{i\to\infty} Pr(Q_i \in A, \alpha_i \in B)$$

Entropy Rate

$$H(\mathcal{Y}) = \sum_{q\in\mathcal{Q}} \int_{\mathcal{P}(\mathcal{Q})} \pi(q, d\alpha) \int_{\mathcal{Y}} f_q(dy) \log \sum_{q'\in\mathcal{Q}} \alpha(q') f_{q'}(dy)$$

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

Analysis of the Forward Recursion (2)

Consistency of the APP Estimate

Let X be a r.v., O be an observation, and $P_x = Pr(X = x|O)$

$$\begin{aligned} Pr(P, X = x) &= \textstyle\sum_O Pr(P, O) Pr(X = x|P, O) \\ &= \textstyle\sum_O Pr(P, O) P_x \\ &= Pr(P) P_x \end{aligned}$$

Stationary Distribution Determined by $Pr(\alpha)$

$$\pi(q, \alpha') = Pr(\alpha_\bullet = \alpha') Pr(Q_\bullet = q|\alpha_\bullet) = \pi(\mathcal{Q}, \alpha')\alpha'(q)$$

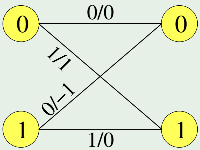
Simpler Yet Identical Markov Chain: $\{\alpha_i\}_{i\geq 1}$

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

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

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

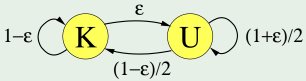
Evolution of the Markov Chain

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

The Forward Recursion for the DEC (2)

The Stationary Distribution

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



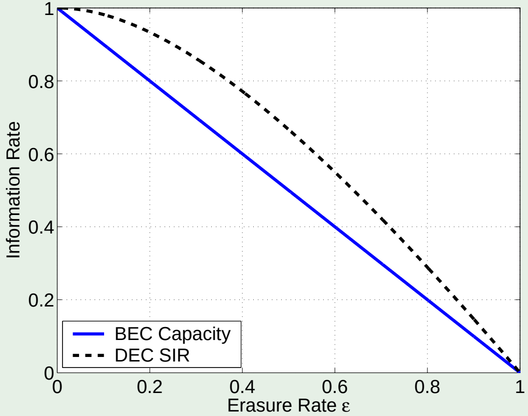
- Stationary distribution: $Pr(K) = \frac{1-\epsilon}{1+\epsilon}$ and $Pr(U) = \frac{2\epsilon}{1+\epsilon}$

Entropy Rate for Each Subset

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

The Symmetric Information Rate of the DEC

Results



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

Notes 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

Upper Bounds for Finite-Memory Channels

Simple Bound: Ignore Channel Memory Completely

$$H(Y_n|Y_1^{n-1}) \leq \max_{Pr(X_1^{\nu+1}) \in \mathcal{M}_\nu(\mathcal{X})} H(Y_{\nu+1})$$

- Bound appears to become tight as noise increases
  - Finite-ISI in AWGN: Gives the correct slope as rate $\rightarrow 0$
  - DEC: Gives the correct slope at $\epsilon = 1$
- Why: State is unpredictable and $H(Y_n|Y_1^{n-1}) \rightarrow H(Y_n)$

Vontobel-Arnold Upper Bound

- Based DMC upper bound and $H(Y_n|Y_1^{n-1}) \leq H(Y_n|Y_{n-L}^{n-1})$
- Max over all input distributions makes it non-obvious

Finite-Memory Channels with Erasures

Mutual Information Decomposition

- Since $X_1^n \implies \widetilde{Y}_{\nu+1}^n$, we have $H(\mathcal{Y}|\mathcal{X}) = h(\epsilon)$
- If Y_n is not erased, then it equals $\widetilde{Y}_n$, so

$$H(Y_n|Y_1^{n-1}) = h(\epsilon) + (1-\epsilon)H(\widetilde{Y}_n|Y_1^{n-1})$$

- which implies that

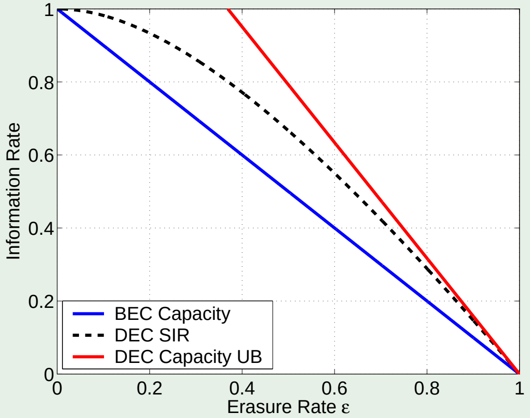
$$I(\mathcal{X};\mathcal{Y}) = (1-\epsilon) \lim_{n\to\infty} H(\widetilde{Y}_n|Y_1^{n-1}) \geq (1-\epsilon) \lim_{n\to\infty} H(\widetilde{Y}_n|\widetilde{Y}_1^{n-1})$$

Apply Simple Upper Bound

$$(1-\epsilon)H(\widetilde{Y}_n|Y_1^{n-1}) \leq (1-\epsilon) \max_{Pr(X_1^{\nu+1})\in\mathcal{M}_\nu(\mathcal{X})} H(\widetilde{Y}_{\nu+1})$$

Capacity Bounds for the DEC

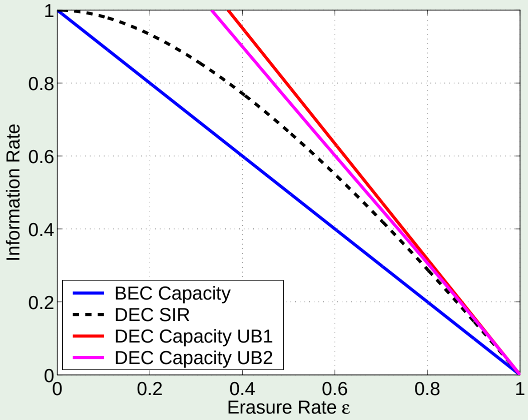
Input with Transition Probability p



$$C \leq (1-\epsilon) \max_{0\leq p\leq 1} h\left(\frac{p}{2}, 1-p, \frac{p}{2}\right) = (1-\epsilon)\log 3$$

Vontobel-Arnold Bound for the DEC

Improvement Over Simple Bound is Small for $L = 1$



Notes and Open Problems

Note: Performance of the Vontobel-Arnold

- For many channels, bound becomes good for moderate L
- But, not for channels with ambiguous output sequences
  - Dicode: All zero output gives no info even without noise
  - DC-free channels always confuse all 0 and all 1 inputs

Open: Capacity Bounds for Almost Perfect Channels

- Can we expand capacity about "zero" noise?
  - Is the first non-zero term the same as $C_{i.u.d}$?
- For the DEC, the expansion is $C_{i.u.d.} = 1 - 2\epsilon^2 + O(\epsilon^3)$
  - Any single erasure on the DEC is corrected by the BCJR

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

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

Products compute the multi-step transition-observation matrix

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

Similarly, $[M(x, y)]_{r,q} = Pr(X_i = x, Y_i = y, Q_{i+1} = q | Q_i = r)$

The Forward Recursion Revisited

Matrix Form of the Forward Recursion

$$\alpha_{i+1} = \frac{\alpha_i M(y_i)}{\|\alpha_i M(y_i)\|_1} \qquad a_{i+1} = \|\alpha_i M(y_i)\|_1$$

The Entropy Rate as a Lyapunov Exponent

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

- Advantage of this formulation is a wealth of results on convergence, continuity, central limit, etc...

Mixing Conditions

Connectivity

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

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

Rank 1 Condition (R1)

- There exists some 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 a k s.t. $M(y_1^k) > 0$ for all y_1^k
- Each sample path forgets the past exponentially fast

Central Limit Theorem

Entropy Estimates are Asymptotically Gaussian

- $\mathrm{Var}\left(-\ln A_i\right) < \infty$ under mild conditions on $Pr(Y|Q)$
- Under (R1), both Markov chain and Renewal CLTs apply
- Under (S1), a CLT for products of random matrices applies
- Conjecture: CLT holds in general if variance is finite
  - Mean/variance may depend on initial state

Related Work

- Exponential Forgetting in HMMs (Le Gland,Mevel 2000)
- On Entropy and Lyapunov Exponents (Holliday et al.)

Summary

The Capacity of Finite-State Channels

- Inrtroduced the definition and some applications
  - The “formula” has been known since Shannon’s time
- Until recently, it has infeasible to evaluate it
  - New methods offer computational approaches
  - Sufficient effort usually allows tight bounds

The Dicode Erasure Channel

- Allows “worked” examples without simulations
- Exact expressions are fairly simple in some cases
- As simple as it is, the capacity is unknown
  - Curse of infinite dimensionality offers little hope