Variable-length codes for channels with memory and feedback: fundamental limits and practical transmission schemes

Achilleas Anastasopoulos
anastas@umich.edu

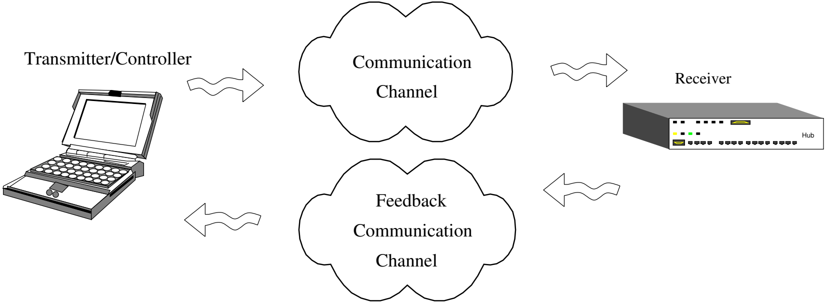
EECS Department
University of Michigan

April 6, 2017

- Joint work with Jui Wu

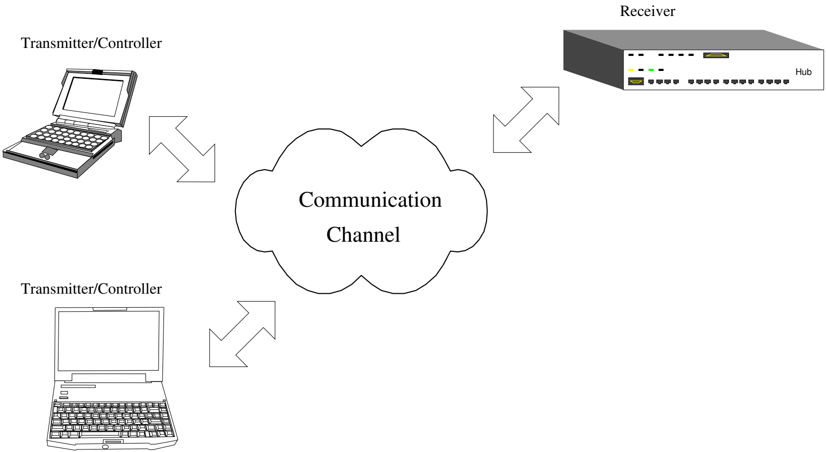
Problems in the intersection of Communications and Control

A) Viewing point-to-point communications as a Control problem



The act of transmitting a signal (partially) controls the overall communication system, with the hope of bringing it to a "desirable" state

B) Viewing multi-agent communications as a Control problem



Multiple agents (partially) control a communication network to bring it to a state beneficial for all (cooperatively/competitively)

C) More subtle: Viewing off-line optimization problems relevant to Information theory as control problems, e.g., Shannon capacity

$$C = \sup_{\{P_{X_t|X^{t-1},Y^{t-1}}(\cdot|\cdot,\cdot)\}_t} \frac{1}{T}\sum_{t=1}^{T} I(X_t \wedge Y_t|Y^{t-1})$$

No clear connection to Control:
Where is the controller?
where is the plant?
what is the observation/control action?

Overview

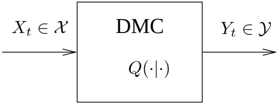
1. Discrete memoryless channels (DMCs)
   - DMC without feedback
   - DMC with feedback and fixed-length (FL) codes
   - DMC with feedback and variable-length (VL) codes
2. Channels with memory and feedback
   - Known capacity results
   - Recent results for error exponents of VL codes

Overview

1. Discrete memoryless channels (DMCs)
   - DMC without feedback
   - DMC with feedback and fixed-length (FL) codes
   - DMC with feedback and variable-length (VL) codes

2. Channels with memory and feedback
   - Known capacity results
   - Recent results for error exponents of VL codes

Discrete memoryless channels without feedback



- Discrete memoryless channel (DMC) $(\mathcal{X}, \mathcal{Y}, Q)$ without feedback

$$P(Y_t|X^t, Y^{t-1}) = Q(Y_t|X_t)$$

Discrete memoryless channels without feedback

- Discrete memoryless channel (DMC) $(\mathcal{X}, \mathcal{Y}, Q)$ without feedback

$$P(Y_t|X^t, Y^{t-1}) = Q(Y_t|X_t)$$

- Fixed-length (FL) code $\mathcal{C}$ with length n (channel uses) and size $M = 2^k$ (messages)

$$\text{encoder: } e : \{1, 2, \ldots, M\} \to \mathcal{X}^n \text{ with } e(W) = X^n \triangleq (X_1, \ldots, X_n)$$
$$\text{decoder: } d : \mathcal{Y}^n \to \{1, 2, \ldots, M\} \text{ with } d(Y^n) = \hat{W}$$

Discrete memoryless channels without feedback

- Discrete memoryless channel (DMC) $(\mathcal{X}, \mathcal{Y}, Q)$ without feedback

$$P(Y_t|X^t, Y^{t-1}) = Q(Y_t|X_t)$$

- Fixed-length (FL) code $\mathcal{C}$ with length n (channel uses) and size $M = 2^k$ (messages)

$$\text{encoder: } e : \{1, 2, \ldots, M\} \to \mathcal{X}^n \text{ with } e(W) = X^n \triangleq (X_1, \ldots, X_n)$$
$$\text{decoder: } d : \mathcal{Y}^n \to \{1, 2, \ldots, M\} \text{ with } d(Y^n) = \hat{W}$$

- Rate $R \triangleq \frac{\log M}{n} = \frac{k}{n}$ (info bits/channel use).
  Error probability $Pe \triangleq P(W \neq \hat{W})$

DMCs without feedback: basic results

- **Capacity [Shannon, 1948]:** The maximum transmission rate with arbitrarily low error probability is

$$C \triangleq \max_{P_X} I(X;Y) = \max_{P_X} \sum_{x,y} Q(y|x) P_X(x) \log \frac{Q(y|x)}{P_Y(y)}$$

DMCs without feedback: basic results

- **Capacity [Shannon, 1948]:** The maximum transmission rate with arbitrarily low error probability is

$$C \triangleq \max_{P_X} I(X;Y) = \max_{P_X} \sum_{x,y} Q(y|x) P_X(x) \log \frac{Q(y|x)}{P_Y(y)}$$

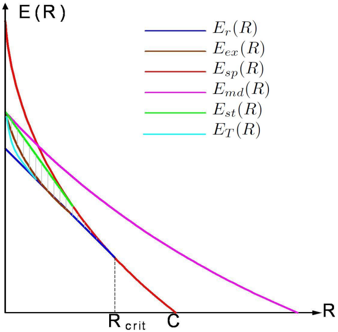
- **Error exponent [Fano, 1961, Gallager, 1965, Shannon et al., 1967]:** The error probability of the optimal codes decays exponentially with code length, n

$$Pe \approx 2^{-nE^*(R)}$$

where $E^*(R)$ is the (rate dependent) error exponent (a.k.a., channel reliability function).

Bounds on the reliability function

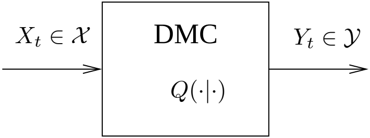
- What do we know about $E^*(R)$ for DMCs without feedback (after ~50 years of research)?



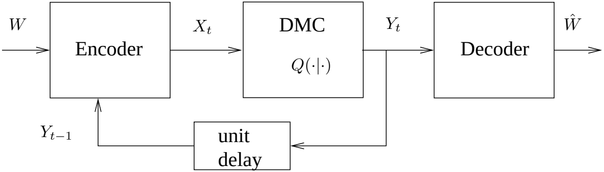
- Upper Bounds:
  - $E_{sp}(R)$
  - $E_{st}(R)$
  - $E_{md}(R)$
- Lower Bounds:
  - $E_r(R)$
  - $E_{ex}(R)$
  - $E_T(R)$

- Above R_{crit} the channel reliability function is known (matching bounds). Below R_{crit} we have bounds (not matching).

DMC with feedback



DMC with feedback



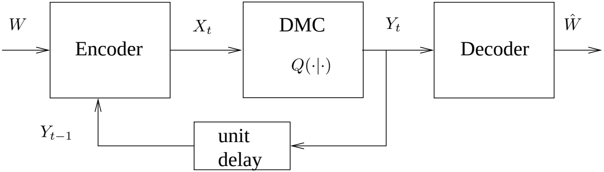
- Fixed-length (FL) code $\mathcal{C}$ with length n (channel uses) and size $M = 2^k$ (messages)

$$\text{encoder: } e = (e_t)_{t=1,\ldots,n}$$

$$e_t : \{1, 2, \ldots, M\} \times \mathcal{Y}^{t-1} \to \mathcal{X} \text{ with } X_t = e_t(W, Y^{t-1})$$

$$\text{decoder: } d : \mathcal{Y}^n \to \{1, 2, \ldots, M\} \text{ with } \hat{W} = d(Y^n)$$

DMC with feedback



- Fixed-length (FL) code $\mathcal{C}$ with length n (channel uses) and size $M = 2^k$ (messages)

$$\text{encoder: } e = (e_t)_{t=1,\ldots,n}$$

$$e_t : \{1, 2, \ldots, M\} \times \mathcal{Y}^{t-1} \to \mathcal{X} \text{ with } X_t = e_t(W, Y^{t-1})$$

$$\text{decoder: } d : \mathcal{Y}^n \to \{1, 2, \ldots, M\} \text{ with } \hat{W} = d(Y^n)$$

- Can also consider randomized encoders, e.g.,

$$X_t \sim e_t(\cdot|W, Y^{t-1}) \Leftrightarrow X_t = e_t(W, Y^{t-1}, V_t)$$

(with V_t some RV that induces the required randomness (possibly common information between Tx/Rx))

DMC with feedback: basic results

- **Capacity:** Capacity cannot be improved by feedback for DMCs!

$$C^{Feedback} = C^{NoFeedback}$$

DMC with feedback: basic results

- **Capacity:** Capacity cannot be improved by feedback for DMCs!

$$C^{Feedback} = C^{NoFeedback}$$

- **Error exponent for FL codes:** The error exponent with FL codes cannot be improved by feedback (at least for symmetric DMCs) above the critical rate! [Haroutunian, 1977]

$$E^{*,Feedback}(R) \leq E^{Feedback}_{Haroutunian}(R)\Big|_{\substack{\text{symmetric}\\ \text{DMCs}}} = E^{NoFeedback}_{sp}(R)$$

DMC with feedback: basic results

- **Capacity:** Capacity cannot be improved by feedback for DMCs!

$$C^{Feedback} = C^{NoFeedback}$$

- **Error exponent for FL codes:** The error exponent with FL codes cannot be improved by feedback (at least for symmetric DMCs) above the critical rate! [Haroutunian, 1977]

$$E^{*,Feedback}(R) \leq E^{Feedback}_{Haroutunian}(R)\Big|_{\substack{\text{symmetric}\\ \text{DMCs}}} = E^{NoFeedback}_{sp}(R)$$

- We can only hope for possible improvements in:
  1. non-symmetric DMCs (e.g., Z-channel)
  2. continuous-alphabet memoryless channels (e.g., Gaussian channels [Schalkwijk and Kailath, 1966])
  3. "third-order" performance improvements and/or simpler encoding/decoding schemes
  4. variable-length codes ?!?
  5. channels with memory

DMC with feedback and variable-length codes

- Variable-length (VL) code $\mathcal{C}$ with size $M = 2^k$ (messages)

$$
\begin{aligned}
\text{encoder: } & e = (e_t)_{t=1,2,\ldots} \\
& e_t : \{1, 2, \ldots, M\} \times \mathcal{Y}^{t-1} \to \mathcal{X} \text{ with } X_t = e_t(W, Y^{t-1}) \\
\text{decoder: } & d = (d_t)_{t=1,2,\ldots} \\
& d_t : \mathcal{Y}^t \to \{1, 2, \ldots, M\} \text{ with } \hat{W}_t = d_t(Y^t) \\
\text{stopping time: } & T \text{ with } \hat{W} = \hat{W}_T = d_T(Y^T)
\end{aligned}
$$

DMC with feedback and variable-length codes

- Variable-length (VL) code $\mathcal{C}$ with size $M = 2^k$ (messages)

$$\begin{aligned}
\text{encoder: } & e = (e_t)_{t=1,2,\dots} \\
& e_t : \{1,2,\dots,M\} \times \mathcal{Y}^{t-1} \to \mathcal{X} \text{ with } X_t = e_t(W, Y^{t-1}) \\
\text{decoder: } & d = (d_t)_{t=1,2,\dots} \\
& d_t : \mathcal{Y}^t \to \{1,2,\dots,M\} \text{ with } \hat{W}_t = d_t(Y^t) \\
\text{stopping time: } & T \text{ with } \hat{W} = \hat{W}_T = d_T(Y^T)
\end{aligned}$$

- Transmission time (code length), T, is a RV $\Longrightarrow$ "Variable-length codes".

  Average rate $\bar{R} \triangleq \frac{\log M}{E[T]} = \frac{k}{E[T]}$

  Error probability $Pe \triangleq P(W \neq \hat{W}_T \cup T = \infty)$

DMC with feedback and VL codes: basic results

- The reliability function is known exactly [Burnashev, 1976]

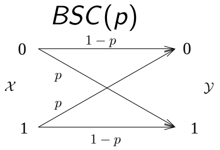
$$E^{*,VL}(\bar{R}) = C_1(1 - \frac{\bar{R}}{C})$$

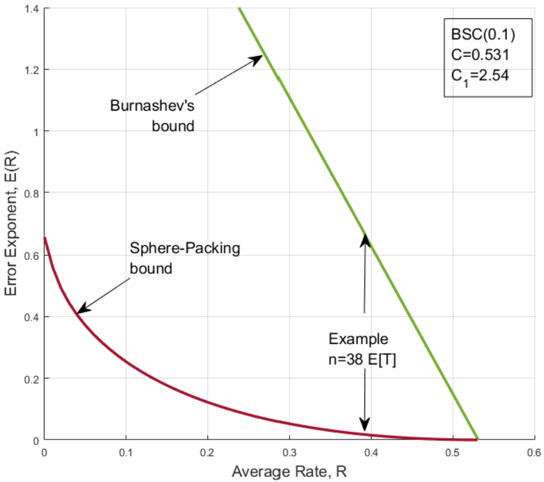
  where C_1 is a channel-dependent constant (max divergence)

$$C_1 \triangleq \max_{x \neq x'} \sum_{y \in \mathcal{Y}} Q(y|x) log \frac{Q(y|x)}{Q(y|x')} = D(Q(\cdot|x_0)||Q(\cdot|x_1))$$

- Transmission schemes achieving this bound are known:
  1. The Burnashev scheme [Burnashev, 1976]
  2. The Yamamoto-Itoh scheme [Yamamoto and Itoh, 1979]

Error exponent for VL codes: BSC





Upper bound derivation: basic concepts

Define the entropy of the posterior message distribution $\Pi_t(i) \triangleq P(W = i | Y^t)$

$$H_t \triangleq H(\Pi_t) = -\sum_{i=1}^{M} \Pi_t(i) \log \Pi_t(i)$$

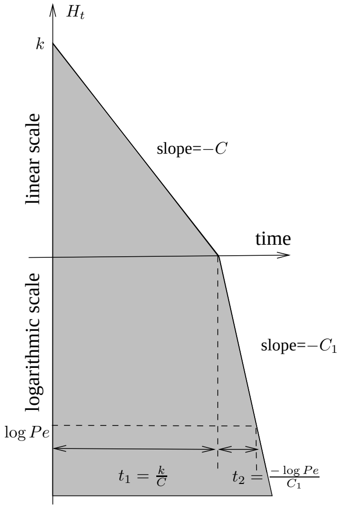
- Fano's inequality: connection between $Pe = P(W \neq \hat{W}_t)$ and H_t
- Study the rate of decay of H_t (drift analysis)

$$E\,[H_{t+1} - H_t | Y^t] \geq -C \qquad \text{(from converse)}$$

- When H_t becomes very small above result is useless.
  Instead study exponential bounds

$$E\,[\log(H_{t+1}) - \log(H_t) | Y^t] \geq -C_1, \qquad H_t \approx \text{small}$$

Upper bound derivation: basic concepts (cont.)



- Define a submartingale Ξ_t based on H_t for the two different regimes.
- Technical difficulty: stitch together the two regimes in one random process, Ξ_t, and perform drift analysis
- Intuition: total transmission time

$$E[T] \geq t_1 + t_2 = \frac{k}{C} + \frac{-\log Pe}{C_1}$$

$$\Rightarrow \frac{-\log Pe}{E[T]} \leq C_1(1 - \frac{k/E[T]}{C})$$

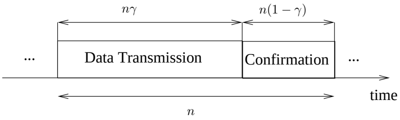
$$\Rightarrow E^{*,VL}(\bar{R}) \leq C_1(1 - \frac{\bar{R}}{C})$$

Upper bound derivation: the constant C_1

- The constant C_1 is intimately related to **binary hypothesis testing**
- When $H_t \approx$ small, one of the messages (say, $W = i$) has very high (≈ 1) posterior probability

$$
\begin{aligned}
&E\left[\log(H_{t+1}) - \log(H_t) | y^t\right] \\
&\quad \gtrapprox -\sum_{y_{t+1}} P(y_{t+1}|W=i,y^t) \log \frac{P(y_{t+1}|W=i,y^t)}{P(y_{t+1}|W\neq i,y^t)} \\
&\quad \geq -\max_{y^t,i,e_{t+1}} \sum_{y_{t+1}} Q(y_{t+1}|e_{t+1}(i,y^t)) \log \frac{Q(y_{t+1}|e_{t+1}(i,y^t))}{\sum_{j\neq i} Q(y_{t+1}|e_{t+1}(j,y^t)) \frac{\pi_t(j)}{1-\pi_t(i)}} \\
&\quad \geq -\max_{x\neq x'} \sum_{y} Q(y|x) \log \frac{Q(y|x)}{Q(y|x')} \\
&\quad = -\max_{x\neq x'} D(Q(\cdot|x)||Q(\cdot|x'))
\end{aligned}
$$

Yamamoto-Itoh transmission scheme

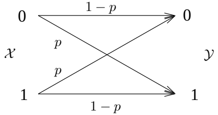


- Transmit in blocks of length n
- Two stages in each packet: (1) data transmission; (2) confirmation
- Data transmission: use a capacity-achieving **non-feedback** code (not specified).
- At end of data transmission both Tx/Rx know if message error occured
- Confirmation: Send 1 bit information whether decoded message was correct or not (x_0 if correct; x_1 if error)
- Repeat until correct confirmation is received (may take several blocks)
- Optimize $\gamma \in (0,1)$ for largest error exponent of $Pe^{data} \times P(e \rightarrow c)$.

$$E(R) = \max_{\gamma \in (0,1)} \gamma E^{NF}(R/\gamma) + (1-\gamma)C_1 \overset{\frac{R}{\gamma}=C}{=} 0 + (1 - \frac{R}{C})C_1$$

Burnashev transmission scheme

Example: BSC(p). Capacity-achieving input distribution $P_X(0) = P_X(1) = 0.5$



Keep track of the posterior distribution (pmf) of the message

$$\Pi_t(i) \stackrel{\triangle}{=} P(W = i | Y^t) \qquad i = 1, \ldots, M$$

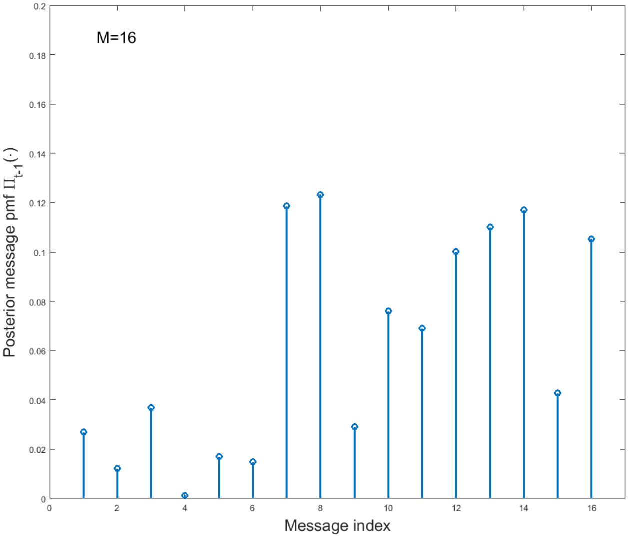
Randomized encoding (Y^{t-1} is summarized in Π_{t-1})

$$X_t = e_t(W, \Pi_{t-1}, V_t)$$

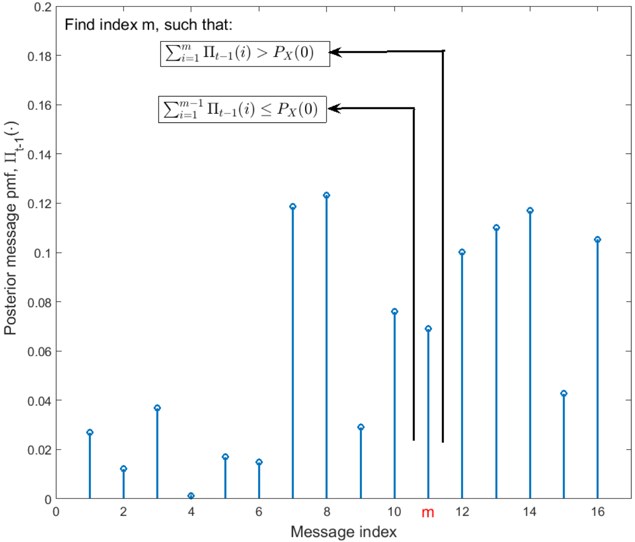
with V_t common randomness between Tx/Rx

Two distinct transmission stages...

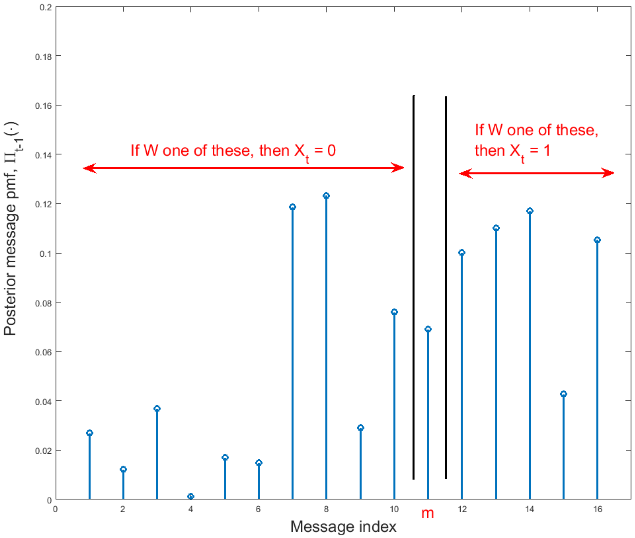
Burnashev scheme: stage 1



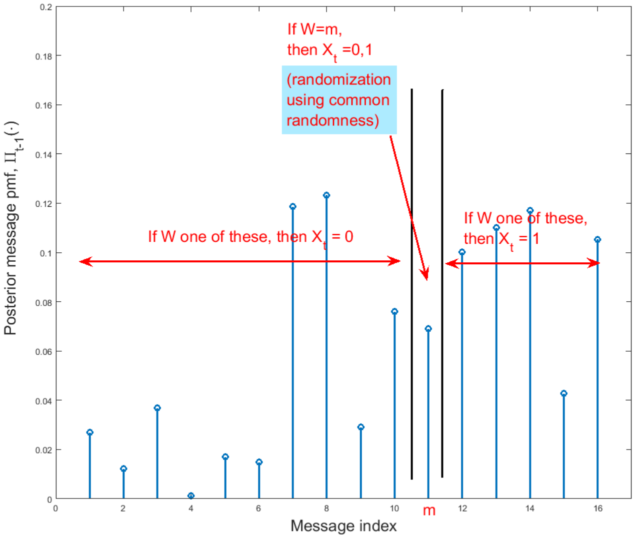
Burnashev scheme: stage 1



Burnashev scheme: stage 1



Burnashev scheme: stage 1



Burnashev scheme: stage 1

- Let's call this scheme TNGTNE (The noisy "guess the number" encoding)!

Burnashev scheme: stage 1

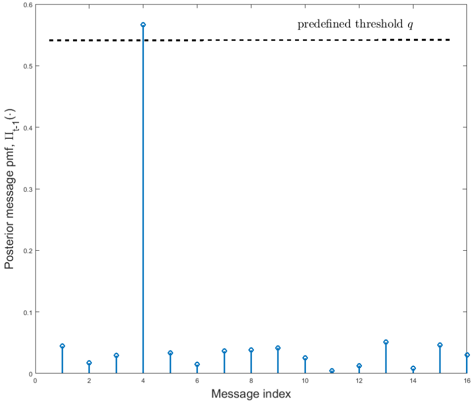
- Let's call this scheme TNGTNE (The noisy "guess the number" encoding)! well...maybe not...
- **Discrete randomized posterior matching (DRPM)**

$$X_t = DRPM(\Pi_{t-1}(\cdot), P_X(\cdot), W, V_t)$$

where
$\Pi_{t-1}(\cdot) \in \mathcal{P}(\{1, 2, \ldots, M\})$
$P_X(\cdot) \in \mathcal{P}(\mathcal{X})$
$W \in \{1, 2, \ldots, M\}$
$V_t \sim u([0, 1])$ (independent of all $W, X^{t-1}, Y^{t-1}, V^{t-1}$)

Burnashev scheme: stage 1

We expect that after a number of steps doing DRPM we get something like



- Continue in stage 1 with DRPM until $\max_i \Pi_{t_0}(i) > q$
- Provisional message estimate $\hat{W}_{t_0} = \arg\max_i \Pi_{t_0}(i)$

Burnashev scheme: stage 2

Stage 2:

- If $W = \hat{W}_{t_0}$ (hypothesis h0) keep sending the predefined symbol $X_t = x_0$
- If $W \neq \hat{W}_{t_0}$ (hypothesis h1) keep sending the predefined symbol $X_t = x_1$
- Continue in stage 2 until
  either
  $\max_i \Pi_t(i) > 1 - Pe$ (say at time T) and declare $\hat{W}_T = \arg\max_i \Pi_T(i)$
  or
  $\max_i \Pi_t(i)$ drops below threshold q, and go back to stage 1

Burnashev scheme: stage 2

Stage 2:

- If $W = \hat{W}_{t_0}$ (hypothesis h0) keep sending the predefined symbol $X_t = x_0$
- If $W \neq \hat{W}_{t_0}$ (hypothesis h1) keep sending the predefined symbol $X_t = x_1$
- Continue in stage 2 until
  either
  $\max_i \Pi_t(i) > 1 - Pe$ (say at time T) and declare $\hat{W}_T = \arg\max_i \Pi_T(i)$
  or
  $\max_i \Pi_t(i)$ drops below threshold q, and go back to stage 1

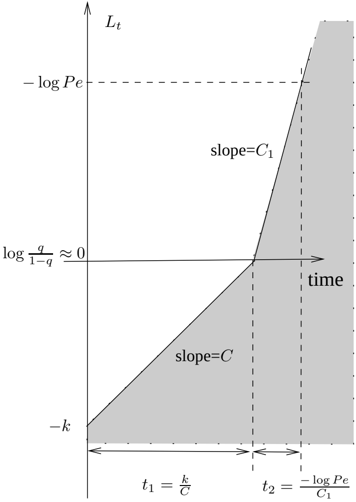
- No codebook to store at Tx/Rx; simple decoding at Rx

Burnashev scheme: Analysis

[Burnashev, 1976]:

- Would like to analyze how fast $\log \frac{\max_i \Pi_t(i)}{1-\max_i \Pi_t(i)}$ (log-likelihood ratio of **best message** posterior probability) grows towards the threshold $\log \frac{1-Pe}{Pe}$
- Instead, analyze the process $L_t \triangleq \log \frac{\Pi_t(W)}{1-\Pi_t(W)}$ (log-likelihood ratio of **true message** posterior probability)
- $E\left[L_{t+1} - L_t | Y^t\right] \geq C$
- $E\left[L_{t+1} - L_t | Y^t\right] \geq C_1 > C$ (if $L_t > \log \frac{q}{1-q}$, for appropriately defined q)
- Create a submartingale Z_t from L_t and apply optional stopping theorem
- Intuition: geometric picture

Burnashev scheme: Analysis, Intuition



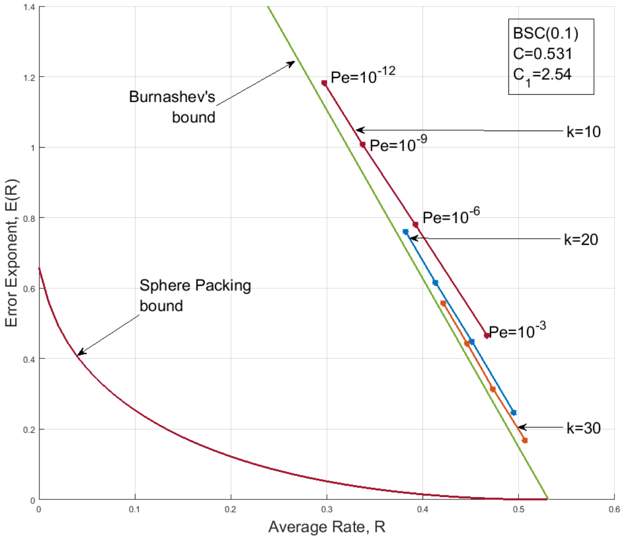
- Intuition: total transmission time

$$E[T] \le t_1 + t_2 = \frac{k}{C} + \frac{-\log Pe}{C_1}$$

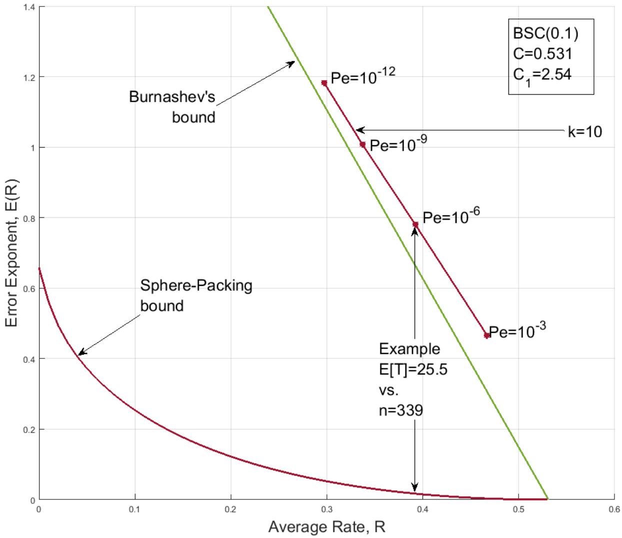
$$\Rightarrow \frac{-\log Pe}{E[T]} \ge C_1\left(1 - \frac{k/E[T]}{C}\right)$$

$$\Rightarrow E^{*,VL}(\bar{R}) \ge C_1\left(1 - \frac{\bar{R}}{C}\right)$$

Error exponent for VL codes: BSC simulation



Error exponent for VL codes: BSC simulation

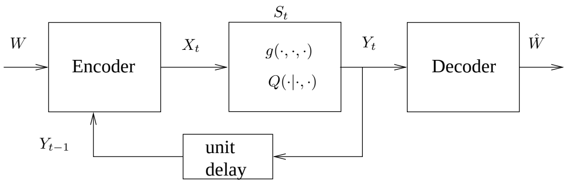


Overview

1. Discrete memoryless channels (DMCs)
   - DMC without feedback
   - DMC with feedback and fixed-length (FL) codes
   - DMC with feedback and variable-length (VL) codes

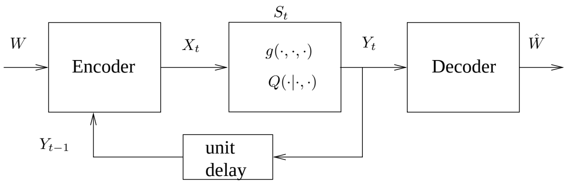
2. Channels with memory and feedback
   - Known capacity results
   - Recent results for error exponents of VL codes

Unifilar channel with feedback



- Information message $W \in \{1, 2, \ldots, M\}$
- Transmitted symbols $X_t \in \mathcal{X}$, $t = 1, 2, \ldots$
- Channel state $S_t \in \mathcal{S}$, $t = 1, 2, \ldots$
- Received symbols $Y_t \in \mathcal{Y}$, $t = 1, 2, \ldots$
- Input/output conditional distribution $Q(Y_t|X_t, S_t)$
- Deterministic state update $S_{t+1} = g(S_t, X_t, Y_t)$
- Encoding functions $X_t = e_t(W, Y^{t-1}, S_1, V_t)$, $t = 1, 2, \ldots$
- Decoding function $\hat{W}_t = d_t(Y^t)$ (together with a stopping time T)

Unifilar channel with feedback



- Information message $W \in \{1, 2, \ldots, M\}$
- Transmitted symbols $X_t \in \mathcal{X}$, $t = 1, 2, \ldots$
- Channel state $S_t \in \mathcal{S}$, $t = 1, 2, \ldots$
- Received symbols $Y_t \in \mathcal{Y}$, $t = 1, 2, \ldots$
- Input/output conditional distribution $Q(Y_t|X_t, S_t)$
- Deterministic state update $S_{t+1} = g(S_t, X_t, Y_t)$
- Encoding functions $X_t = e_t(W, Y^{t-1}, S_1, V_t)$, $t = 1, 2, \ldots$
- Decoding function $\hat{W}_t = d_t(Y^t)$ (together with a stopping time T)

- State known to Tx but not to Rx (Tx knows S_1, X^{t-1}, $Y^{t-1} \Rightarrow S^t$)!

Unifilar channel with feedback: capacity

Capacity is the result of the **off-line** optimization problem [Permuter et al., 2008] over infinitely many conditional distributions on $\mathcal{X}$

$$C = \lim_{N\to\infty} \sup_{\{P(X_t|S_t,Y^{t-1})\}_{t\geqslant 1}} \frac{1}{N}\sum_{i=1}^{N} I(X_t, S_t; Y_t|Y^{t-1}).$$

- Observe: $P_{X_t|S_t,Y^{t-1}} \in \mathcal{S}\times\mathcal{Y}^{t-1} \to \mathcal{P}(\mathcal{X})$, so its domain increases with t

Unifilar channel with feedback: capacity

Capacity is the result of the **off-line** optimization problem [Permuter et al., 2008] over infinitely many conditional distributions on $\mathcal{X}$

$$C = \lim_{N\to\infty} \sup_{\{P(X_t|S_t,Y^{t-1})\}_{t\geqslant 1}} \frac{1}{N}\sum_{i=1}^{N} I(X_t, S_t; Y_t|Y^{t-1}).$$

- Observe: $P_{X_t|S_t,Y^{t-1}} \in \mathcal{S}\times\mathcal{Y}^{t-1} \to \mathcal{P}(\mathcal{X})$, so its domain increases with t
- How can we utilize Control theory to solve this problem?

Unifilar channel with feedback: capacity

Capacity is the result of the **off-line** optimization problem [Permuter et al., 2008] over infinitely many conditional distributions on $\mathcal{X}$

$$C = \lim_{N\to\infty} \sup_{\{P(X_t|S_t,Y^{t-1})\}_{t\geqslant 1}} \frac{1}{N}\sum_{i=1}^{N} I(X_t, S_t; Y_t|Y^{t-1}).$$

- Observe: $P_{X_t|S_t,Y^{t-1}} \in \mathcal{S}\times\mathcal{Y}^{t-1} \to \mathcal{P}(\mathcal{X})$, so its domain increases with t

- Define posterior belief of the state (Tx/Rx can evaluate it)

$$B_t(s) \triangleq P(S_{t+1} = s|Y^t)$$

- $\{B_t\}_t$ forms a (controlled) Markov process, which can be (partially) controlled by $P(X_t|S_t, Y^{t-1})$
- Utilize theory of Markov Decision Processes (MDPs) to derive a single-letter expression [Permuter et al., 2008]

$$C = \sup_{P(X_t|S_t,B_{t-1})} I(X_t, S_t; Y_t|B_{t-1})$$

Error exponents for VL coding: upper bound

- How can we generalize Burnashev's analysis to channels with memory?

Error exponents for VL coding: upper bound

- How can we generalize Burnashev's analysis to channels with memory?

- **Basic idea #1**: analyze multi-step drift (to capture memory effects)
  For any $\epsilon > 0$ there exists a large enough step N, s.t.

$$\frac{1}{N}E[H_{t+N} - H_t | Y^t = y^t, S_1 = s_1]$$
$$\geq -\frac{1}{N}\sum_{k=t}^{t+N-1} I(X_{k+1}, S_{k+1}; Y_{k+1} | Y_{t+1}^k, Y^t = y^t, S_1 = s_1)$$
$$\geq -(C + \epsilon), \qquad \text{(from ergodicity of } \{B_t\}_t)$$

and similarly (for the case of small H_t)

$$\frac{1}{N}E[\log(H_{t+N}) - \log(H_t) | Y^t, S_1] \geqslant -(C_1 + \epsilon) \qquad a.s.$$

Error exponents for VL coding: the C_1 constant

- What is C_1 in this case?

$$C_1 = \max_{s_1, y^t, i} \limsup_{N \to \infty} \max_{\{e_\tau\}_{\tau=t+1}^{t+N}} \frac{1}{N} \sum_{Y_{t+1}^{t+N}} P(Y_{t+1}^{t+N} | W = i, y^t, s_1) \log \frac{P(Y_{t+1}^{t+N} | W = i, y^t, s_1)}{P(Y_{t+1}^{t+N} | W \neq i, y^t, s_1)}.$$

- It relates to a binary hypothesis testing problem with
  h0: $W = i$
  h1: $W \neq i$

Error exponents for VL coding: the C_1 constant

- **Basic idea #2**: Define $X_t^i = e_t(i, Y^{t-1}, S_1)$ and $S_t^i = g_t(i, Y^{t-1}, S_1)$ which are the input and the state at time t, conditioned on $W = i$. Then,

$$P(Y_{t+1}^{t+N} | W = i, y^t, s_1) = \prod_{\tau=t+1}^{t+N} Q(Y_\tau | S_\tau^i, X_\tau^i)$$

- Define $X_t^{\bar{i}}(x|s)$, as the induced input distribution at time t, conditioned on $S_t = s$ and $W \neq i$
- Define $B_{t-1}^1(s) \triangleq P(S_t = s | W \neq i, Y^{t-1}, S_1)$ as the posterior state belief at time t, conditioned on $W \neq i$. Then,

$$P(Y_{t+1}^{t+N} | y^t, s_1, W \neq i) = \prod_{\tau=t+1}^{t+N} \left[\sum_{x,s} Q(Y_\tau | x, s) X_\tau^{\bar{i}}(x|s) B_{\tau-1}^1(s) \right]$$

Error exponents for VL coding: the C_1 constant

- C_1 relates to the average reward per unit time of an MDP with:
  state: $(S_t^0, B_{t-1}^1) \in \mathcal{S} \times \mathcal{P}(\mathcal{S})$,
  action: $(X_t^0, X_t^1) \in \mathcal{X} \times (\mathcal{S} \to \mathcal{P}(\mathcal{X}))$,
  instantaneous reward: $R(S_t^0, B_{t-1}^1; X_t^0, X_t^1)$,
  transition kernel:

$$P(S_{t+1}^0, B_t^1 | S_t^0, B_{t-1}^1, X_t^0, X_t^1) = \sum_y \delta_{g(S_t^0, X_t^0, y)}(S_{t+1}^0) \delta_{\phi(B_{t-1}^1, X_t^1, y)}(B_t^1) Q(y | X_t^0, S_t^0).$$

Error exponents for VL coding: the C_1 constant

- C_1 relates to the average reward per unit time of an MDP with:
  state: $(S_t^0, B_{t-1}^1) \in \mathcal{S} \times \mathcal{P}(\mathcal{S})$,
  action: $(X_t^0, X_t^1) \in \mathcal{X} \times (\mathcal{S} \to \mathcal{P}(\mathcal{X}))$,
  instantaneous reward: $R(S_t^0, B_{t-1}^1; X_t^0, X_t^1)$,
  transition kernel:

$$P(S_{t+1}^0, B_t^1 | S_t^0, B_{t-1}^1, X_t^0, X_t^1) = \sum_y \delta_{g(S_t^0, X_t^0, y)}(S_{t+1}^0) \delta_{\phi(B_{t-1}^1, X_t^1, y)}(B_t^1) Q(y | X_t^0, S_t^0).$$

- From MDP theory: optimal action only function of current state: $X^0[S_t^0, B_{t-1}^1]$ and $X^1[S_t^0, B_{t-1}^1](\cdot|\cdot)$

Error exponents for VL coding: the C_1 constant

- Intuition gained: C_1 relates to a binary hypothesis test over a channel with memory, with a weird twist!
  - Under h0, input is a deterministic symbol (function of current state S_t^0 and belief B_{t-1}^1)
    $X_t = X^0[S_t^0, B_{t-1}^1]$
  - Under h1, input is a **random** symbol (function of **hypothesized state under h0,** S_t^0, belief B_{t-1}^1, and state $S_t \sim B_{t-1}^1(\cdot)$)
    $X_t \sim X^1[S_t^0, B_{t-1}^1](\cdot|S_t)$

Error exponents for VL coding: the C_1 constant

- Intuition gained: C_1 relates to a binary hypothesis test over a channel with memory, with a weird twist!
  - Under h0, input is a deterministic symbol (function of current state S_t^0 and belief B_{t-1}^1)
    $X_t = X^0[S_t^0, B_{t-1}^1]$
  - Under h1, input is a **random** symbol (function of **hypothesized state under h0,** S_t^0, belief B_{t-1}^1, and state $S_t \sim B_{t-1}^1(\cdot)$)
    $X_t \sim X^1[S_t^0, B_{t-1}^1](\cdot|S_t)$
- Why so complicated? dual objective
  (1) resolving the hypothesis by transmitting the most distinguishable symbols and
  (2) partially controlling the channel state evolution

Error exponents for VL coding: the C_1 constant

- Intuition gained: C_1 relates to a binary hypothesis test over a channel with memory, with a weird twist!
  - Under h0, input is a deterministic symbol (function of current state S_t^0 and belief B_{t-1}^1)
    $X_t = X^0[S_t^0, B_{t-1}^1]$
  - Under h1, input is a **random** symbol (function of **hypothesized state under h0,** S_t^0, belief B_{t-1}^1, and state $S_t \sim B_{t-1}^1(\cdot)$)
    $X_t \sim X^1[S_t^0, B_{t-1}^1](\cdot|S_t)$
- Why so complicated? dual objective
  (1) resolving the hypothesis by transmitting the most distinguishable symbols and
  (2) partially controlling the channel state evolution
- Challenge: turn this into an actual transmission scheme!
  Last part of this talk...

A Burnashev-like VL coding scheme

- Keep track of the posterior distribution (pmf) of the message

$$\Pi_t(i) \triangleq P(W = i|\mathcal{F}_t) \qquad i = 1, \ldots, M$$

and the vector of states

$$\underline{S}_t = (S_t^1, S_t^2, \ldots, S_t^M),$$

where S_t^i is the hypothesized state at time t conditioned on $W = i$.

A Burnashev-like VL coding scheme

- Keep track of the posterior distribution (pmf) of the message

$$\Pi_t(i) \triangleq P(W = i|\mathcal{F}_t) \qquad i = 1, \ldots, M$$

and the vector of states

$$\underline{S}_t = (S_t^1, S_t^2, \ldots, S_t^M),$$

where S_t^i is the hypothesized state at time t conditioned on $W = i$.
- Calculate posterior beliefs

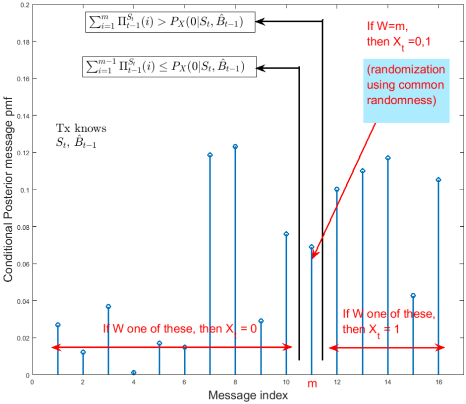
$$\hat{B}_{t-1}(s) = \sum_{i=1}^{M} \Pi_{t-1}(i) 1_{\{S_t^i = s\}} = P(S_t = s|\mathcal{F}_{t-1})$$

$$\Pi_{t-1}^s(i) = \frac{\Pi_{t-1}(i) 1_{\{S_t^i = s\}}}{\hat{B}_{t-1}(s)} = P(W = i|S_t = s, \mathcal{F}_{t-1}),$$

Two distinct transmission stages...

A Burnashev-like VL coding scheme: stage 1

Stage 1: $X_t = DRPM(\Pi_{t-1}^{S_t}(\cdot), P_{X|SB}(\cdot|S_t, \hat{B}_{t-1}), W, V_t)$



Continue in stage 1 until $\max_i \Pi_{t_0}(i) > q$

Provisional message estimate $\hat{W} = \arg\max_i \Pi_{t_0}(i)$

A Burnashev-like VL coding scheme: stage 2

Stage 2:

- Calculate new posteriors (conditioned on h1: $W \neq \hat{W}_{t_0}$)

$$\hat{B}^1_{t-1}(s) = \frac{\sum_{i \neq \hat{W}} \Pi_{t-1}(i) 1_{\{S^i_t = s\}}}{1 - \Pi_{t-1}(\hat{W})} = P(S_t = s | \mathcal{F}_{t-1}, h1)$$

$$\Pi^{1,s}_{t-1}(i) = \frac{\Pi_{t-1}(i) 1_{\{i \neq \hat{W}\}} 1_{\{S^i_t = s\}}}{\hat{B}^1_{t-1}(s)(1 - \Pi_{t-1}(\hat{W}))} = P(W = i | S_t = s, \mathcal{F}_{t-1}, h1)$$

A Burnashev-like VL coding scheme: stage 2

Stage 2:

- Calculate new posteriors (conditioned on h1: $W \neq \hat{W}_{t_0}$)

$$\hat{B}^1_{t-1}(s) = \frac{\sum_{i \neq \hat{W}} \Pi_{t-1}(i) 1_{\{S^i_t = s\}}}{1 - \Pi_{t-1}(\hat{W})} = P(S_t = s | \mathcal{F}_{t-1}, h1)$$

$$\Pi^{1,s}_{t-1}(i) = \frac{\Pi_{t-1}(i) 1_{\{i \neq \hat{W}\}} 1_{\{S^i_t = s\}}}{\hat{B}^1_{t-1}(s)(1 - \Pi_{t-1}(\hat{W}))} = P(W = i | S_t = s, \mathcal{F}_{t-1}, h1)$$

- If $W = \hat{W}$ (hypothesis h0) transmit $X_t = X^0[S^{\hat{W}}_t, \hat{B}^1_{t-1}]$

A Burnashev-like VL coding scheme: stage 2

Stage 2:

- Calculate new posteriors (conditioned on h1: $W \neq \hat{W}_{t_0}$)

$$\hat{B}^1_{t-1}(s) = \frac{\sum_{i \neq \hat{W}} \Pi_{t-1}(i) 1_{\{S^i_t = s\}}}{1 - \Pi_{t-1}(\hat{W})} = P(S_t = s | \mathcal{F}_{t-1}, h1)$$

$$\Pi^{1,s}_{t-1}(i) = \frac{\Pi_{t-1}(i) 1_{\{i \neq \hat{W}\}} 1_{\{S^i_t = s\}}}{\hat{B}^1_{t-1}(s)(1 - \Pi_{t-1}(\hat{W}))} = P(W = i | S_t = s, \mathcal{F}_{t-1}, h1)$$

- If $W = \hat{W}$ (hypothesis h0) transmit $X_t = X^0[S^{\hat{W}}_t, \hat{B}^1_{t-1}]$
- If $W \neq \hat{W}$ (hypothesis h1)

$$X_t = DRPM(\Pi^{1,S^W_t}_{t-1}(\cdot), X^1[S^{\hat{W}}_t, \hat{B}^1_{t-1}](\cdot | S_t), W, V_t)$$

- Continue in stage 2 until either
  $\max_i \Pi_t(i) > 1 - Pe$ (say at time T) and declare $\hat{W}_T = \arg\max_i \Pi_T(i)$
  or
  $\max_i \Pi_t(i)$ drops below threshold q, and go back to stage 1

A Burnashev-like VL coding scheme: Analysis

- Analyze the one-step drift of the process $L_t \triangleq \log \frac{\Pi_t(W)}{1-\Pi_t(W)}$ and use ergodicity to get multi-step results

A Burnashev-like VL coding scheme: Analysis

- Analyze the one-step drift of the process $L_t \triangleq \log \frac{\Pi_t(W)}{1-\Pi_t(W)}$ and use ergodicity to get multi-step results

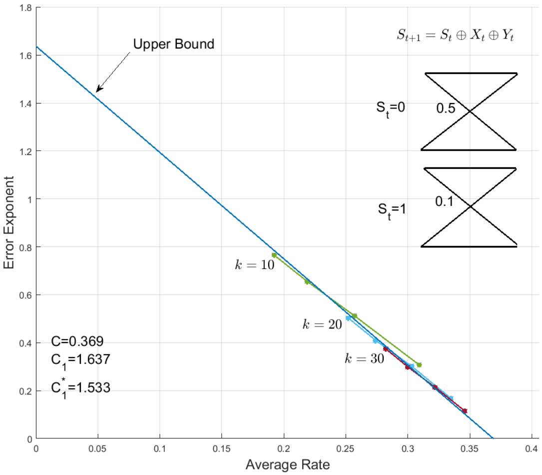
- **Unresolved issue:** The defined process $\{\hat{B}_t\}_t$ does not have the same statistics as $\{B_t\}_t$ (related to capacity expression)
- $B_{t-1}(s) = P(S_t = s|Y^{t-1}, S_1)$ vs $\hat{B}_{t-1}(s) = P(S_t = s|Y^{t-1}, V^{t-1}, S_1)$
- This is because of the introduction of common randomness (RVs V_t)!
- In fact $\{\hat{B}_t\}_t$ is not a Markov chain (but $\hat{B}_{t-1}$ is measurable wrt a "bigger" Markov chain $\{(\underline{S}_t, \Pi_{t-1})\}_t$)

A Burnashev-like VL coding scheme: Analysis

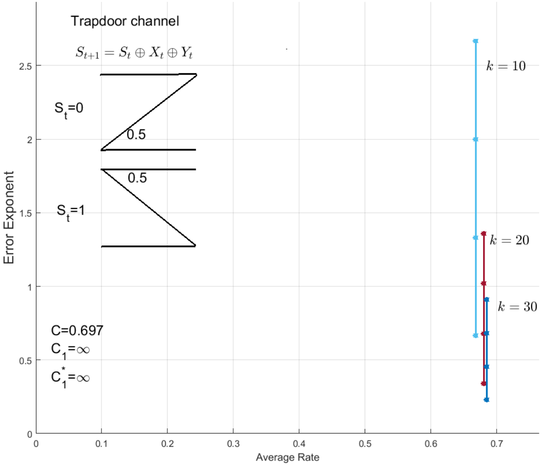
- Analyze the one-step drift of the process $L_t \triangleq \log \frac{\Pi_t(W)}{1-\Pi_t(W)}$ and use ergodicity to get multi-step results

- **Unresolved issue:** The defined process $\{\hat{B}_t\}_t$ does not have the same statistics as $\{B_t\}_t$ (related to capacity expression)
- $B_{t-1}(s) = P(S_t = s | Y^{t-1}, S_1)$ vs $\hat{B}_{t-1}(s) = P(S_t = s | Y^{t-1}, V^{t-1}, S_1)$
- This is because of the introduction of common randomness (RVs V_t)!
- In fact $\{\hat{B}_t\}_t$ is not a Markov chain (but $\hat{B}_{t-1}$ is measurable wrt a "bigger" Markov chain $\{(\underline{S}_t, \Pi_{t-1})\}_t$)
- Some kind of "concentration" result is at play here. **Any ideas?**

Numerical/Simulation results



Numerical/Simulation results



Thank you!

Burnashev, M. V. (1976).
Data transmission over a discrete channel with feedback. Random transmission time.
Problemy Peredachi Informatsii, 12(4):10–30.

Fano, R. M. (1961).
Transmission of information: A statistical theory of communications.
American Journal of Physics, 29(11):793–794.

Gallager, R. G. (1965).
A simple derivation of the coding theorem and some applications.
IEEE Trans. Information Theory, 11(1):3–18.

Haroutunian, E. A. (1977).
Lower bound for error probability in channels with feedback.
Problemy Peredachi Informatsii, 13:36–44.

Permuter, H., Cuff, P., Roy, B. V., and Weissman, T. (2008).
Capacity of the trapdoor channel with feedback.
IEEE Trans. Information Theory, 54(7):3150–3165.

Schalkwijk, J. and Kailath, T. (1966).

A coding scheme for additive noise channels with feedback–I: No bandwidth constraint.
IEEE Trans. Inform. Theory, 12(2):172–182.

Shannon, C. E. (1948).
A mathematical theory of communication.
Bell System Tech. J., 27:379–423, 623–656.

Shannon, C. E., Gallager, R. G., and Berlekamp, E. R. (1967).
Lower bounds to error probability for coding on discrete memoryless channels.
Inform. and Control, 10:65–103 (Part I), 522–552 (Part II).

Yamamoto, H. and Itoh, K. (1979).
Asymptotic performance of a modified schalkwijk-barron scheme for channels with noiseless feedback (corresp.).
IEEE Transactions on Information Theory, 25(6):729–733.