An Information-Theoretic Study of Communication Problems with Feedback and/or Feed-forward.

Ramji Venkataramanan

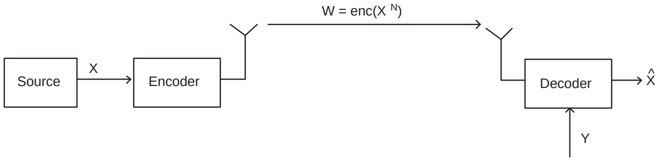
S. Sandeep Pradhan

Dept. of EECS, University of Michigan

Overview of research

1. Source Coding with Feed-forward
   - Rate-distortion function of a general source with feed-forward.
   - Error exponents of a general source with feed-forward.
2. Channel Coding with delayed feedback,state information
3. Evaluating the rate-distortion function and capacity

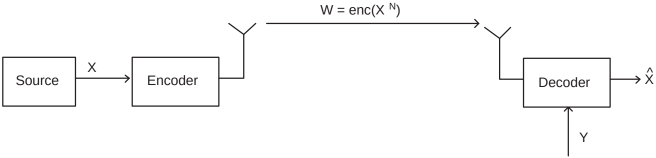
Source Coding with Side-Information



Example: Block length $N = 5$

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Side Info	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Decoder						$\hat{X}_1 - \hat{X}_5$				

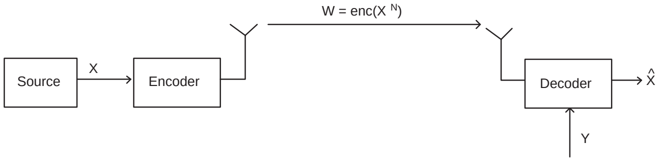
Source Coding with Side-Information



Example: Block length $N = 5$

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Side Info	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Decoder						$\hat{X}_1 - \hat{X}_5$				

Source Coding with Side-Information

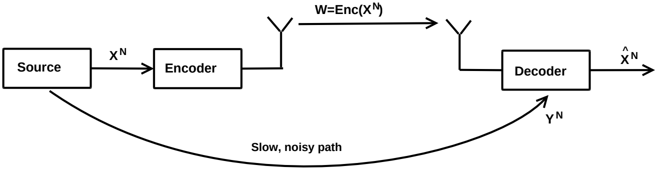


Example: Block length $N = 5$

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Side Info	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Side-Information with Delay

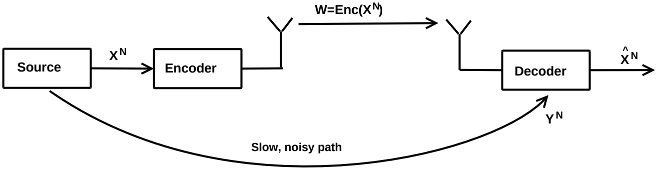
Suppose there is a delay in the side info available at the decoder.



Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Side Info	-	-	-	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Side-Information with Delay

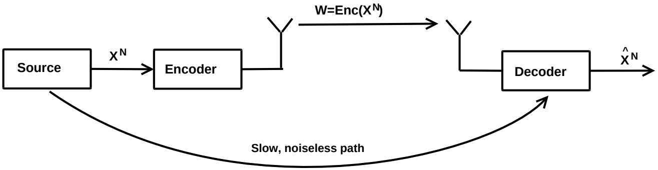
Suppose there is a delay in the side info available at the decoder.



Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Side Info	-	-	-	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

What is feed-forward?

The source field itself may be available in a delayed form at the decoder.

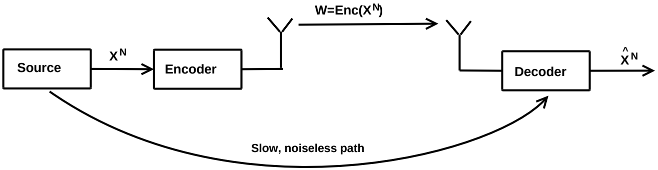


Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Extra info	-	-	-	-	-	-	X_1	X_2	X_3	X_4
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Here, block length= 5, delay is 6 time units.

What is feed-forward?

The source field itself may be available in a delayed form at the decoder.

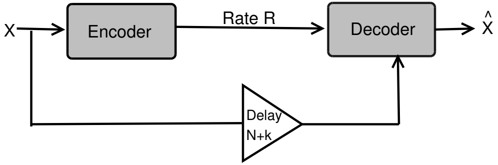


Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Extra info	-	-	-	-	-	-	X_1	X_2	X_3	X_4
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Here, block length $= 5$, delay is 6 time units.

Source Coding with Feed-Forward

- Feed-forward$\Rightarrow$ Decoder knows some of the past source samples.

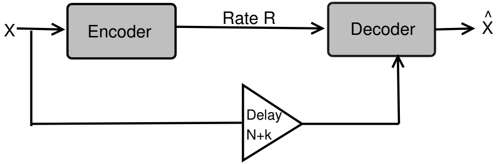


Feed-forward with delay k, block length N.

- To reconstruct X_n, the decoder knows index W and $(X_1, \ldots, X_{n-k})$.
- Applications in other areas too...

Source Coding with Feed-Forward

- Feed-forward$\Rightarrow$ Decoder knows some of the past source samples.

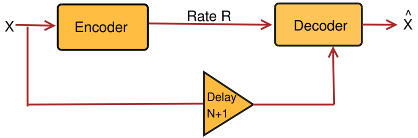


Feed-forward with delay k, block length N.

- To reconstruct X_n, the decoder knows index W and $(X_1, \ldots, X_{n-k})$.
- Applications in other areas too...

Feed-Forward: A Formal Definition

- First: Feed forward delay $= 1$
  [Weissman et al 03],[Pradhan 04],[Martinian et al 04]



- **Source** X: Alphabet $\mathcal{X}$, reconstruction alphabet $\widehat{\mathcal{X}}$
- **Encoder**: Rate R , $e : \mathcal{X}^N \to \{1, \dots, 2^{NR}\}$
- **Decoder**: knows all the past $(n-1)$ source samples to reconstruct nth sample.

$$g_n : \{1, \dots, 2^{NR}\} \times \mathcal{X}^{n-1} \to \widehat{\mathcal{X}}, \quad n = 1, \dots, N.$$

A Formal Definition (contd.)

- Distortion measure $d_N(X^N, \hat{X}^N)$.

GOAL

Given any source X, find the least R such that

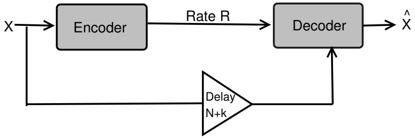
$$E[d_N(x^N, \hat{x}^N)] \leq D.$$

- Rate-Distortion function with Feed-forward!

Directed Information

- [Massey] The directed information flowing from A^N to B^N

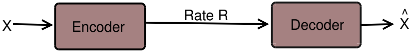
$$I(A^N \to B^N) = \sum_{n=1}^{N} I(A^n; B_n | B^{n-1}).$$



- Interestingly :

$$I(A^N \to B^N) = I(A^N; B^N) - \sum_{n=2}^{N} I(B^{n-1}; A_n | A^{n-1})$$

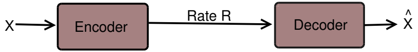
Interpretation of Directed Information



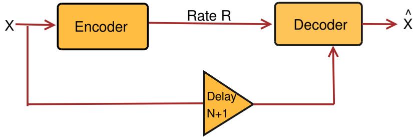
Without FF, need $I(\hat{X}^N; X^N)$ bits to represent X^N with $\hat{X}^N$.

- With feed-forward, to produce $\hat{X}_n$, the decoder knows X^{n-1}.
- Number of bits required is reduced by $I(\hat{X}_n; X^{n-1}|\hat{X}^{n-1})$.

Interpretation of Directed Information

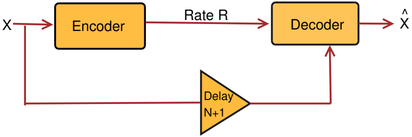


Without FF, need $I(\hat{X}^N; X^N)$ bits to represent X^N with $\hat{X}^N$.



- With feed-forward, to produce $\hat{X}_n$, the decoder knows X^{n-1}.
- Number of bits required is reduced by $I(\hat{X}_n; X^{n-1}|\hat{X}^{n-1})$.

Delay 1 feed-forward

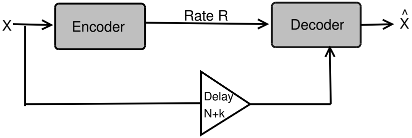


- With delay 1 feed-forward, we need

$$I(\hat{X}^N; X^N) - \sum_{n=2}^{N} I(\hat{X}_n; X^{n-1} | \hat{X}^{n-1}) \quad \text{bits.}$$

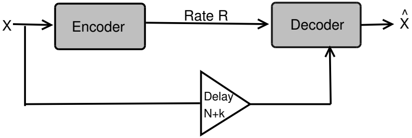
- Directed information from $\hat{X}^N$ to X^N !

Delay k feed-forward



- With delay k feed-forward, to produce $\hat{X}_n$, the decoder knows X^{n-k}.
- No. of bits: $I(\hat{X}^N; X^N) - \sum_{n=k+1}^{N} I(\hat{X}_n; X^{n-k} | \hat{X}^{n-1})$
- Not Directed Information- will denote it $I_k(\hat{X}^N \rightarrow X^N)$
- '$k-$directed information'.

Delay k feed-forward



- With delay k feed-forward, to produce $\hat{X}_n$, the decoder knows X^{n-k}.
- No. of bits: $I(\hat{X}^N; X^N) - \sum_{n=k+1}^{N} I(\hat{X}_n; X^{n-k} | \hat{X}^{n-1})$
- Not Directed Information- will denote it $I_k(\hat{X}^N \to X^N)$
- 'k−directed information'.

General source, general distortion measure

- Source could be non-stationary, non-ergodic
- Sequence of distortion functions $d_n(.,.)$

- Even when source is stationary and ergodic, with feed-forward, the optimal joint distribution may not be.
- Need to use information-spectrum methods [Han, Verdu]

General source, general distortion measure

- Source could be non-stationary, non-ergodic
- Sequence of distortion functions $d_n(.,.)$

- Even when source is stationary and ergodic, with feed-forward, the optimal joint distribution may not be.
- Need to use information-spectrum methods [Han, Verdu]

General source, general distortion measure

- Source could be non-stationary, non-ergodic
- Sequence of distortion functions $d_n(.,.)$

- Even when source is stationary and ergodic, with feed-forward, the optimal joint distribution may not be.
- Need to use information-spectrum methods [Han, Verdu]

Definitions

- $\mathbf{P_X} = \{P_{X_1}, P_{X^2}, \ldots, P_{X^N}, \ldots\}$
- $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}_1|X_1}, P_{\hat{X}^2|X^2}, \ldots, P_{\hat{X}^N|X^N}, \ldots\}$

$a_1, a_2, \ldots$: random sequence

- $\limsup_{in\ prob} a_n = \overline{a}$: Smallest number α such that

$$\lim_{n\to\infty} Pr(a_n > \alpha) = 0.$$

Definitions..

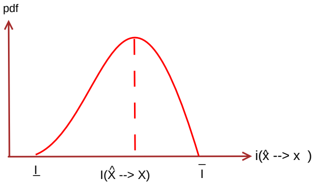
- We will need

$$i_k(\hat{x}^n \rightarrow x^n) = \frac{1}{n} \log \frac{P(x^n, \hat{x}^n)}{P(x^n) \cdot \prod_{i=1}^{n} P(\hat{x}_i | \hat{x}^{i-1}, x^{i-k})}$$

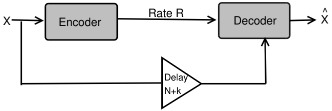
Definitions..

- We will need

$$i_k(\hat{x}^n \rightarrow x^n) = \frac{1}{n} \log \frac{P(x^n, \hat{x}^n)}{P(x^n) \cdot \prod_{i=1}^{n} P(\hat{x}_i | \hat{x}^{i-1}, x^{i-k})}$$



Rate-Distortion Theorem for a general source



Theorem

For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with feed-forward delay k, the infimum of all achievable rates at probability-1 distortion D, is given by

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} : \rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \leq D} \bar{I}_k(\hat{X} \to X),$$

where

$$\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n)$$

Stationary Ergodic Source

- Stationary, ergodic source: $\mathbf{P_X} = \{P_{X^N}\}_{N=1}^{\infty}$.
- Distortion measure d_N.

Generalization of the AEP

Lemma

$(\mathbf{X}, \mathbf{Y}) = \{X_n, Y_n\}$ *be a stationary, ergodic joint process characterized by* $P_{X^n, Y^n}, n = \cdots, -1, 0, 1, \cdots$. *Let*

$$P^k_{X^n|Y^n} \triangleq \prod_{i=1}^{n} P_{X_i|X^{i-1}, Y^{i-k}}.$$

Then

$$-\frac{1}{n} \log P^k_{X^n|Y^n} \to \lim_{n \to \infty} H(X_n|X^{n-1}, Y^{n-k}).$$

What happens as delay changes

- Compare delay 1 FF, delay k FF, no FF.
- Space of optimization remains the same-

  $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}$ that satisfy the distortion constraint.
- No constraints on the conditional distribution $\Rightarrow$

$$I(\hat{X}^N; X^N) - \sum I(\hat{X}_n; X^{n-k}|\hat{X}^{n-1}),$$

$$I(\hat{X}^N \to X^N),$$

$$I(\hat{X}^N; X^N)$$

- Are all different!

- With FF, the reduction in rate is due to *smaller objective function.*
- Space of optimization remains the same!

What happens as delay changes

- Compare delay 1 FF, delay k FF, no FF.
- Space of optimization remains the same-

  $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}$ that satisfy the distortion constraint.
- No constraints on the conditional distribution $\Rightarrow$

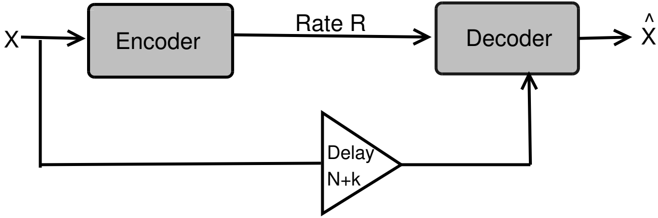
$$I(\hat{X}^N; X^N) - \sum I(\hat{X}_n; X^{n-k}|\hat{X}^{n-1}),$$
$$I(\hat{X}^N \rightarrow X^N),$$
$$I(\hat{X}^N; X^N)$$

- Are all different!

- With FF, the reduction in rate is due to *smaller objective function.*
- Space of optimization remains the same!

Source Coding Optimization

- Source X with distribution $\mathbf{P_X}$.



- Multi-letter optimization- difficult!

Source Coding Optimization

- Given source $\mathbf{P_X} = \{P_{X^n}\}$
- Pick a conditional distribution $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}^n|X^n}\}$
- For what sequence of distortion measures d_n does $\mathbf{P_{\hat{X}|X}}$ achieve the infimum in the rate-distortion formula ?
- $\mathbf{P_{\hat{X}|X}}$ has to minimize $\overline{I}_k(\hat{X} \to X)$ over the set

$$\mathcal{Q}(D) = \{\mathbf{W_{\hat{X}|X}} : \limsup_{\text{in prob } PW} d_n(X^n, \hat{X}^n) \leq D\}.$$

- Approach- similar in spirit to [Csiszar and Korner], [Gastpar et al], [Pradhan et al]

Source Coding Optimization

- Given source $\mathbf{P_X} = \{P_{X^n}\}$
- Pick a conditional distribution $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}^n|X^n}\}$
- For what sequence of distortion measures d_n does $\mathbf{P_{\hat{X}|X}}$ achieve the infimum in the rate-distortion formula ?
- $\mathbf{P_{\hat{X}|X}}$ has to minimize $\overline{I}_k(\hat{X} \to X)$ over the set

$$\mathcal{Q}(D) = \{\mathbf{W_{\hat{X}|X}} : \limsup_{\textit{in prob PW}} d_n(X^n, \hat{X}^n) \leq D\}.$$

- Approach- similar in spirit to [Csiszar and Korner], [Gastpar et al], [Pradhan et al]

Structure of Distortion Function

Theorem

A stationary, ergodic source X characterized by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$ with feed-forward delay k. $\mathbf{P_{\hat{X}|X}} = \{P_{X^n|X^n}\}_{n=1}^{\infty}$ is a conditional distribution such that the joint distribution is stationary and ergodic. Then $\mathbf{P_{\hat{X}|X}}$ achieves the rate-distortion function if for all sufficiently large n, the distortion measure satisfies

$$d_n(x^n, \hat{x}^n) = -c \cdot \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}(x^n, \hat{x}^n)}{\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n)} + d_0(x^n),$$

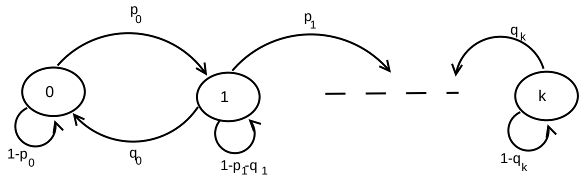
where

$$\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \prod_{i=1}^{n} P_{\hat{X}_i|X^{i-k},\hat{X}^{i-1}}(\hat{x}_i|x^{i-k}, \hat{x}^{i-1})$$

and c is any positive number and $d_0(.)$ is an arbitrary function.

Stock Market Example

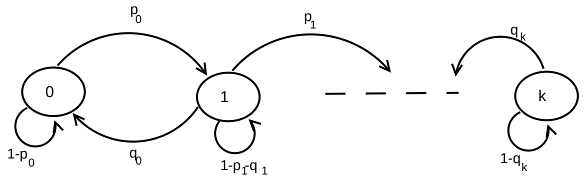
- Wish to observe the behavior of a particular stock over an N-day period.
- Value of the stock- modeled as a $k+1$-state Markov chain.



- Investor has this stock over an $N-$day period, needs to be forewarned whenever the value drops.
- There is an insider with *a priori* knowledge about the behavior of the stock.
- Can give information to the investor at a cost c/bit of info.

Stock Market Example

- Wish to observe the behavior of a particular stock over an N-day period.
- Value of the stock- modeled as a $k+1$-state Markov chain.



- Investor has this stock over an $N-$day period, needs to be forewarned whenever the value drops.
- There is an insider with *a priori* knowledge about the behavior of the stock.
- Can give information to the investor at a cost c/bit of info.

Stock Price Model

- Value of the stock: Markov source $\{X_n\}$
- Decision of investor on day n: $\hat{X}_n$
- $\hat{X}_n = 1 \Rightarrow$ price is going to drop from day $n-1$ to n, $\hat{X}_n = 0$ means otherwise.
- Hamming distortion:
  Distortion 1 when investor fails to predict drop, or falsely predicts.
- Before day n, investor knows all the previous values of the stock X^{n-1}, has to make the decision $\hat{X}_n$ - *feed-forward!*

Stock Price Model

- Value of the stock: Markov source $\{X_n\}$
- Decision of investor on day n: $\hat{X}_n$
- $\hat{X}_n = 1 \Rightarrow$ price is going to drop from day $n-1$ to n, $\hat{X}_n = 0$ means otherwise.
- Hamming distortion:
  Distortion 1 when investor fails to predict drop, or falsely predicts.
- Before day n, investor knows all the previous values of the stock X^{n-1}, has to make the decision $\hat{X}_n$ - *feed-forward!*

Stock Market Example

- $R_{ff}(D)$: Minimum information (in bits/sample) the investor needs to predict drops in value with distortion D.
- Try first-order Markov conditional distribution.

Proposition

For the stock-market problem described above,

$$R_{ff}(D) = \sum_{i=1}^{k-1} \pi_i \left[h(p_i, q_i, 1 - p_i - q_i) - h(\epsilon, 1 - \epsilon) \right] + \pi_k \left(h(q_k, 1 - q_k) - h(\epsilon, 1 - \epsilon) \right),$$

where $h()$ is the entropy function, $\left[\pi_0, \pi_1, \cdots, \pi_k\right]$ is the stationary distribution of the Markov chain and $\epsilon = \frac{D}{1-\pi_0}$.

Stock Market Example

- $R_{ff}(D)$: Minimum information (in bits/sample) the investor needs to predict drops in value with distortion D.
- Try first-order Markov conditional distribution.

Proposition

For the stock-market problem described above,

$$R_{ff}(D) = \sum_{i=1}^{k-1} \pi_i \left[h(p_i, q_i, 1 - p_i - q_i) - h(\epsilon, 1-\epsilon)\right] + \pi_k \left(h(q_k, 1-q_k) - h(\epsilon, 1-\epsilon)\right),$$

where $h()$ is the entropy function, $\left[\pi_0, \pi_1, \cdots, \pi_k\right]$ is the stationary distribution of the Markov chain and $\epsilon = \frac{D}{1-\pi_0}$.

Cost function for feedback channels

Theorem

Suppose we are given a channel $P^{ch}_{\mathbf{Y}|\mathbf{X}}$ with $k-$delay feedback and an input distribution $P^{k}_{\mathbf{X}|\mathbf{Y}}$ such that the joint process $P_{\mathbf{X},\mathbf{Y}}$ is stationary, ergodic. Then the input distribution $P^{k}_{\mathbf{X}|\mathbf{Y}}$ achieves the $k-$delay feedback capacity of the channel if for all sufficiently large n, the cost measure satisfies

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)} + d_0,$$

where λ is any positive number and d_0 is an arbitrary constant.

Cost function for feedback channels

Theorem

Suppose we are given a channel $P^{ch}_{\mathbf{Y}|\mathbf{X}}$ with $k-$delay feedback and an input distribution $P^{k}_{\mathbf{X}|\mathbf{Y}}$ such that the joint process $P_{\mathbf{X},\mathbf{Y}}$ is stationary, ergodic. Then the input distribution $P^{k}_{\mathbf{X}|\mathbf{Y}}$ achieves the $k-$delay feedback capacity of the channel if for all sufficiently large n, the cost measure satisfies

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)} + d_0,$$

where λ is any positive number and d_0 is an arbitrary constant.

Summary

1. Source Coding with Feed-forward
   - Rate-distortion function of a general source- why we get k-directed information!
   - Error exponents of a general source with feed-forward.
2. Channel Coding with delayed feedback,state information
3. Evaluating the rate-distortion function and capacity