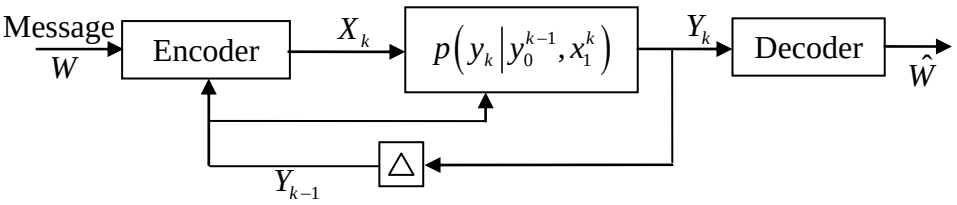
Directed Information and Channel with Feedback

Prapun Suksompong

December 10, 2006

In this article, we consider the problem of sending message via discrete channel with noiseless feedback as shown below.



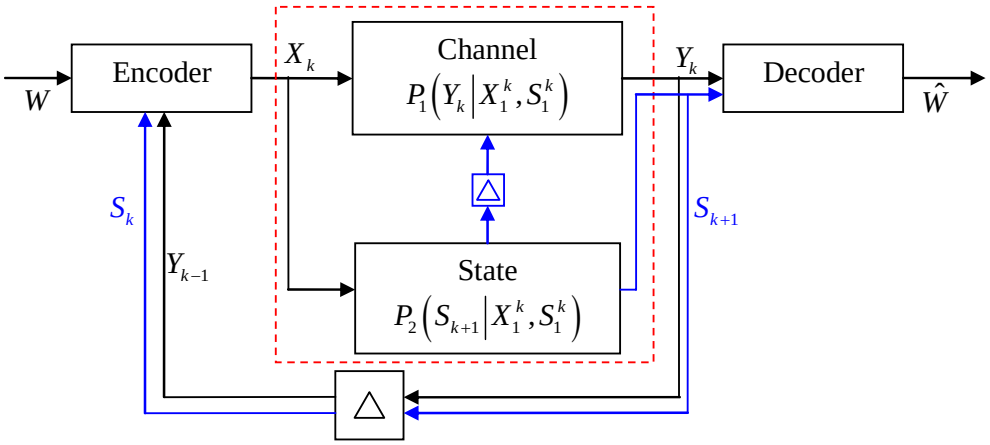
At time *k*, the encoder produce X_k from *W*, the previous encoder outputs X_1^{k-1}, and the fed-back channel outputs Y_0^{k-1}. The encoder can be a deterministic function; that is $X_k = f_k\left(X_1^{k-1}, Y_0^{k-1}\right)$. We also allow random encoder; that is X_k may be governed by the conditional distribution $p\left(x_k \middle| x_1^{k-1}, y_0^{k-1}\right)$. This X_k is put into channel. The channel output Y_k is generated according to the conditional distribution $p\left(y_k \middle| x_1^k, y_0^{k-1}\right)$ which not only depends on the current channel input X_k but may also depend on past channel inputs X_1^{k-1} and outputs Y_0^{k-1}.

We modify the classical Shannon's information measures (entropy, mutual information, and their conditional versions) so that they explicitly incorporate feedback. In particular, we shall focus on $I\left(X^N \to Y^N\right)$, a notation introduced by Massey [1990] to capture the directed information flowing from the length *N* sequence of random variables X^N to the length *N* sequence of random variables X^N. In fact, the idea of given direction to information has already been thought of by Marko [Marko 1973] whose paper define a quantity called directed transinformation. The direct information in [Massey 1990] refines this directed transinformation.

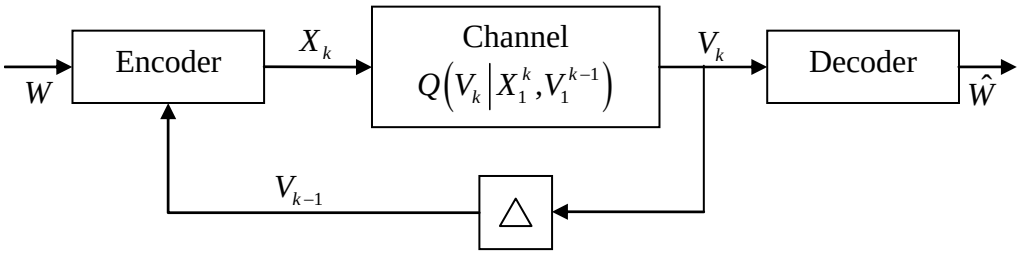
Equivalent System Models

In this section, we show three general models of the system we are interested in. We then prove that they are all equivalent in the sense that we can convert one into another. Hence, in some sense, it is sufficient to analyze only one of them. As usual, *k* is the time index. X_k and Y_k represent the channel input and output at time *k* respectively. S_k denotes the channel state at time *k*. In model 2, we use V_k instead of Y_k to represent the channel output. This notational difference is used so that the proof can be done more smoothly. Model 3 is introduced as an intermediate model to bridge model 1 and model 2. The definitions and proof are almost the same as those in [Chen, Suksompong, and Berger 2004]. What follows is conditioned on the initial state $S_1 = s_1$.

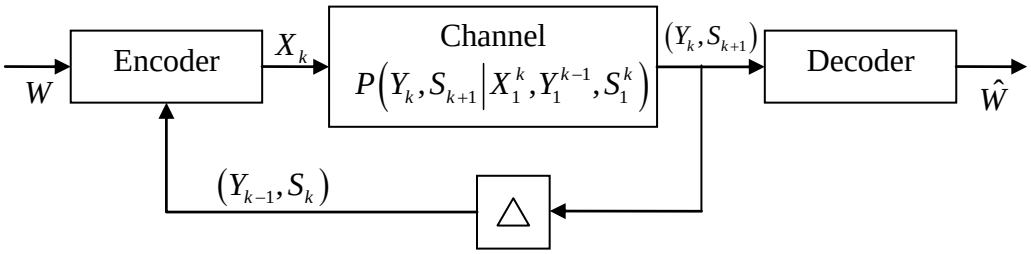
- Model 1:



- Model 2:



- Model 3:



- *Theorem*: Model 1, 2, and 3 are equivalent.

  Proof Let $\mathcal{M}_i$ be the set of systems that can be represented by Model *i*. Then,

  (1) $\mathcal{M}_1 \subset \mathcal{M}_3$ because we can let $P\left(Y_k, S_{k+1} \middle| X_1^k, Y_1^{k-1}, S_1^k\right) = P_1\left(Y_k \middle| X_1^k, S_1^k\right) P_2\left(S_{k+1} \middle| X_1^k, S_1^k\right)$.

  (2) $\mathcal{M}_3 \subset \mathcal{M}_2$ because we can let $V_k = \left(Y_k, S_{k+1}\right)$. Then, $V_1^{k-1} = \left(Y_1^{k-1}, S_1^k\right)$. Set $Q\left(V_k \middle| X_1^k, V_1^{k-1}\right) = P\left(Y_k, S_{k+1} \middle| X_1^k, Y_1^{k-1}, S_1^k\right)$.

  (3) $\mathcal{M}_2 \subset \mathcal{M}_1$ because we can set $S_k = V_{k-1}$, $P_2\left(S_{k+1} \middle| X_1^k, S_1^k\right) = Q\left(V_k \middle| X_1^k, V_1^{k-1}\right)$, and $Y_k = c$ a.s. (Equivalently, $\exists c \; \forall\left(x_1^k, s_1^k\right) \; P_1\left(Y_k = c \middle| X_1^k = x_1^k, S_1^k = s_1^k\right) = 1$.)

Henceforth, we shall focus on Model 2.

Discrete Channel

In this section, we consider the channel part of the system. Special cases of which are also defined.

- ***Discrete channel***: $\left(p\left(y_n \middle| x_1^n, y_1^{n-1}\right)\right)_{n=1}^{N}$.
  - [Tatikonda 2000] calls this nonanticipative channel.
- *Definition*: [Cover Thomas 1991] A channel is **memoryless** if $p\left(y_n \middle| x_1^n, y_1^{n-1}\right) = p\left(y_n \middle| x_n\right)$.

  Simple equivalent statements are

  $\equiv$ Given x_n, $\left(x_1^{n-1}, y_1^{n-1}\right)$ and y_n are independent;

  $\equiv$ $y_n - x_n - \left(x_1^{n-1}, y_1^{n-1}\right)$ forms a Markov chain;

  $\equiv$ $H\left(Y_n \middle| X_1^n, Y_1^{n-1}\right) = H\left(Y_n \middle| X_n\right)$;

  $\equiv$ $\forall i \;\; I\left(Y_j; \left(X_1^{j-1}, Y_1^{j-1}\right) \middle| X_j\right) = 0$.
- *Definition* [Massey 1990]: The channel is ***used without feedback*** if $p\left(x_n \middle| x_1^{n-1}, y_1^{n-1}\right) = p\left(x_n \middle| x_1^{n-1}\right)$.

  Equivalent statements are

  $\equiv$ $H\left(X_n \middle| X_1^{n-1}, Y_1^{n-1}\right) = H\left(X_n \middle| X_1^{n-1}\right)$;

  $\equiv$ $I\left(X_n; Y_1^{n-1} \middle| X_1^{n-1}\right) = 0$;

  $\equiv$ $Y_1^{n-1} - X_1^{n-1} - X_n$ forms a Markov chain;

  $\equiv$ [Ash 1965] for $\forall N > n$, $p\left(y_n \middle| x_1^N, y_1^{n-1}\right) = p\left(y_n \middle| x_1^n, y_1^{n-1}\right)$;

  (Note that this gives exactly the same condition as the definition above when $N = 2$.)

  $\equiv$ $H\left(Y_n \middle| X_1^N, Y_1^{n-1}\right) = H\left(Y_n \middle| X_1^n, Y_1^{n-1}\right)$;

  $\equiv$ $\forall i \;\; I\left(Y_i; X_{i+1}^N \middle| X_1^i, Y_1^{i-1}\right) = 0$;

  $\equiv$ $\forall i \;\; Y_i - \left(X_1^i, Y_1^{i-1}\right) - X_{i+1}^N$ forms a Markov chain.

  $\equiv$ $I\left(X_i; Y_j \middle| X_1^{i-1}, Y_1^{j-1}\right) = 0$ for any $j < i$.

  We shall show later that these are also equivalent to:

  $\equiv$ $I\left(X_1^N; Y_1^N\right) = I\left(X_1^N \to Y_1^N\right)$

  $\equiv$ $I\left(0 * Y_1^{N-1} \to X_1^N\right) = 0$.

  Interpretation: The choice of the next channel input digit, given all previous input digits, is not further related to the previous channel output digits.
- *Joint distribution*:

$$p\left(x_1^N, y_1^N\right) = \prod_{n=1}^{N} p\left(x_n, y_n \middle| x_1^{n-1}, y_1^{n-1}\right) = \prod_{n=1}^{N} p\left(x_n \middle| x_1^{n-1}, y_1^{n-1}\right) p\left(y_n \middle| x_1^n, y_1^{n-1}\right).$$

  - Discrete ***memoryless*** channel: $p\left(x_1^N, y_1^N\right) = \prod_{n=1}^{N} p\left(x_n \middle| x_1^{n-1}, y_1^{n-1}\right) p\left(y_n \middle| x_n\right)$.

- Discrete channel used ***without feedback***:

$$p\left(x_1^N, y_1^N\right) = \prod_{n=1}^{N} p\left(x_n \middle| x_1^{n-1}\right) p\left(y_n \middle| x_1^n, y_1^{n-1}\right) = \prod_{n=1}^{N} p\left(x_n \middle| x_1^{n-1}\right) \prod_{n=1}^{N} p\left(y_n \middle| x_1^n, y_1^{n-1}\right)$$

$$= p\left(x_1^N\right) \prod_{n=1}^{N} p\left(y_n \middle| x_1^n, y_1^{n-1}\right)$$

or equivalently,

$$p\left(y_1^N \middle| x_1^N\right) = \prod_{n=1}^{N} p\left(y_n \middle| x_1^n, y_1^{n-1}\right)$$

  - Used [Ash], $p\left(y_n \middle| x_1^N, y_1^{n-1}\right) = p\left(y_n \middle| x_1^n, y_1^{n-1}\right)$, we have

$$p\left(x_1^N, y_1^N\right) = p\left(x_1^N\right) \prod_{n=1}^{N} p\left(y_n \middle| x_1^N, y_1^{n-1}\right) = p\left(x_1^N\right) \prod_{n=1}^{N} p\left(y_n \middle| x_1^n, y_1^{n-1}\right);$$

    hence equivalent to Massey's.

- Discrete ***memoryless*** channel used ***without feedback***:

$p\left(x_1^N, y_1^N\right) = p\left(x_1^n\right) \prod_{n=1}^{N} p\left(y_n \middle| x_n\right)$, or equivalently $p\left(y_1^N \middle| x_1^N\right) = \prod_{n=1}^{N} p\left(y_n \middle| x_n\right)$.

  This is ***DMC used without feedback***.

  In this case,

  - $H\left(X_1^N, Y_1^N\right) = H\left(X_1^N\right) + \sum_{i=1}^{N} H\left(Y_i \middle| X_i\right)$.
  - $H\left(Y_1^N \middle| X_1^N\right) = \sum_{i=1}^{N} H\left(Y_i \middle| X_i\right)$

As an introduction to directed information, we present here one way to look at it. First we shall decompose mutual information into several conditional mutual information using chain rules. We associated each part with an arrow going between X_i and Y_j in a diagram representing the general channel in figure 1. The above specialized channels are then just the general channel with some of the arrows missing. Directed information takes only some of these parts and hence it is a part of mutual information. This idea was presented to Prof. T. Berger and L. Tong during the author's Q exam in June, 2005. This representation of directed information trivializes several of its properties.

- The definitions defined above can be represented using the diagrams.

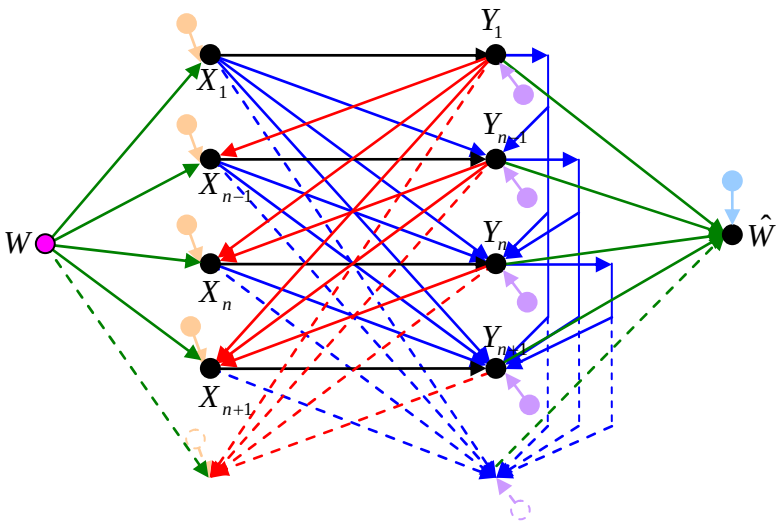


Figure 1 Discrete Channel

Figure 1 shows the discrete channel in its full generality. We can, in fact, decompose it into three parts which are shown in black, blue, and red. Consider, first, the part in black as shown in figure 2.

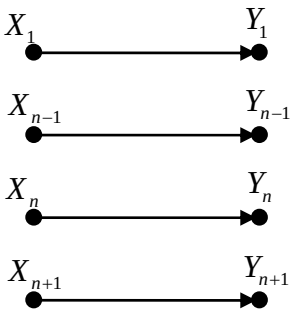


Figure 2 Discrete Memoryless Channel without Feedback

Figure 2 represents the discrete memoryless channel without feedback. We can add memory to channel by adding the blue part, as shown in figure 3. This allows y_k to depends on $\left(x_1^k, y_1^{k-1}\right)$, not just x_k.

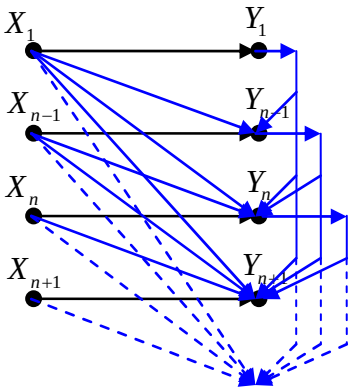


Figure 3 Discrete Channel without Feedback

The last part, with red color in figure 1, represents the feedback. Adding it to figure 3 leads us back the general discrete channel in figure 1.

- The mutual information $I\left(X_1^N;Y_1^N\right)$ delivered via the discrete channel described above can also be partitioned into three parts.

  First, by applying the chain rule for mutual information twice, we have

$$I\left(X_1^N;Y_1^N\right)=\sum_{i=1}^{N}I\left(X_i;Y_1^N\left|X_1^{i-1}\right.\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}I\left(X_i;Y_j\left|X_1^{i-1},Y_1^{j-1}\right.\right).$$

  Now, we separate the terms inside the sum above into three groups by asking whether *i* is equal to, less than, or grater than *j*. This gives

  1) $i=j:\ \sum_{i=1}^{N}I\left(Y_i;X_i\left|Y_1^{i-1},X_1^{i-1}\right.\right)$. This quantity will be defined as $I\left(X_1^N\leftrightarrow Y_1^N\right)$. We want to say that it relates to the "direct" paths in black above.

  2) $i<j:\ \sum_{j=1}^{N}\sum_{i=1}^{j-1}I\left(X_i;Y_j\left|X_1^{i-1},Y_1^{j-1}\right.\right)=\sum_{j=1}^{N}I\left(X_1^{j-1};Y_j\left|Y_1^{j-1}\right.\right)$. This quantity will be defined as $I\left(0*X_1^{N-1}\to Y_1^N\right)$. We want to say that it relates to the channel memory paths in blue above.

  3) $j<i:\ \sum_{i=1}^{N}\sum_{j=1}^{i-1}I\left(X_i;Y_j\left|X_1^{i-1},Y_1^{j-1}\right.\right)=\sum_{i=1}^{N}I\left(X_i;Y_1^{i-1}\left|X_1^{i-1}\right.\right)$. This quantity will be defined as $I\left(0*Y_1^{N-1}\to X_1^N\right)$. We want to say that it captures the feedback paths in red above.

  The directed information only adds up the terms that has $i\le j$ (all arrows that point from *X* to *Y*), that is

$$\begin{aligned}I\left(X_1^N\to Y_1^N\right)&=I\left(X_1^N\leftrightarrow Y_1^N\right)+I\left(0*X_1^{N-1}\to Y_1^N\right)\\&=I\left(X_1^N;Y_1^N\right)-I\left(0*Y_1^{N-1}\to X_1^N\right)\end{aligned}$$

  For the familiar DMC without feedback, we will show that

$$I\left(X_1^N\to Y_1^N\right)=I\left(X_1^N;Y_1^N\right)\le\sum_{i=1}^{N}I\left(Y_i;X_i\right)\le nC\ .$$

Directed Information and its properties

- [Kramer 1998] Assume that the sequences X_1^N and Y_1^N are "synchronized", i.e., that the n^{th} terms in the sequences occur "at the same time", and that the n^{th} terms occur "before" the $(n+1)^{\text{st}}$ terms.
- [Massey 1990] The **directed information** $I\left(X_1^N \to Y_1^N\right)$ from a sequence X_1^N to a sequence Y_1^N is defined by $I\left(X_1^N \to Y_1^N\right) = \sum_{i=1}^{N} I\left(X_1^i; Y_i \middle| Y_1^{i-1}\right)$.
  - *Remarks*:
    - $I\left(X_1^N; Y_1^N\right) = \sum_{i=1}^{N} I\left(X_1^N; Y_i \middle| Y_1^{i-1}\right)$ by chain rule. We add the information that Y_i tells about X_1^N which Y_1^{i-1} haven't already told.
    - For directed information, we use $I\left(X_1^n; Y_n \middle| Y_1^{n-1}\right)$: To eliminate feedback information, we ignore the information that Y_i may be giving about the future X_{i+1}^N.
    - $$I\left(X_1^N \to Y_1^N\right) = \sum_{i=1}^{N} \mathbb{E}\left[\log \frac{p\left(X_1^i, Y_i \middle| Y_1^{i-1}\right)}{p\left(X_1^i \middle| Y_1^{i-1}\right) p\left(Y_i \middle| Y_1^{i-1}\right)}\right]$$
      $$= \mathbb{E}\left[\log \frac{\prod_{i=1}^{N} p\left(Y_i \middle| X_1^i, Y_1^{i-1}\right)}{p\left(Y_1^N\right)}\right]$$
    - $I\left(Y_1^N \to X_1^N\right) = \sum_{i=1}^{N} I\left(Y_1^i; X_i \middle| X_1^{i-1}\right)$
- $I\left(X_1^N \to Y_1^N\right) \le I\left(X_1^N; Y_1^N\right)$ with equality iff channel is used without feedback

  Proof. $$I\left(X_1^N; Y_i \middle| Y_1^{i-1}\right) = H\left(Y_i \middle| Y_1^{i-1}\right) - H\left(Y_i \middle| X_1^N, Y_1^{i-1}\right)$$
  $$\ge H\left(Y_i \middle| Y_1^{i-1}\right) - H\left(Y_i \middle| X_1^i, Y_1^{i-1}\right)$$
  $$= I\left(X_1^i; Y_i \middle| Y_1^{i-1}\right)$$

  Hence,

  $$I\left(X_1^N; Y_1^N\right) = \sum_{i=1}^{N} I\left(X_1^N; Y_i \middle| Y_1^{i-1}\right) \ge \sum_{n=1}^{N} I\left(X_1^n; Y_n \middle| Y_1^{n-1}\right) = I\left(X_1^N \to Y_1^N\right).$$

  Equality occurs iff $\forall i \; H\left(Y_i \middle| X_1^N, Y_1^{i-1}\right) = H\left(Y_i \middle| X_1^n, Y_1^{i-1}\right)$. This is Ash's condition which is equivalent to the "used without feedback" condition.

  Alternative Proof.

  Use $I\left(X_1, X_2; Y \middle| Z\right) = I\left(X_1; Y\right) + I\left(X_2; Y \middle| X_1, Z\right)$. Then,

  $$I\left(X_1^N; Y_i \middle| Y_1^{i-1}\right) - I\left(X_1^i; Y_i \middle| Y_1^{i-1}\right) = I\left(X_{i+1}^N; Y_i \middle| X_1^i, Y_1^{i-1}\right) \ge 0$$

This proof give us the next property.

- $I\left(X_1^N;Y_1^N\right)-I\left(X_1^N\to Y_1^N\right)=\sum_{i=1}^{N}I\left(X_{i+1}^N;Y_i\middle|X_1^i,Y_1^{i-1}\right)$; in fact, the last term is 0; so

$$I\left(X_1^N;Y_1^N\right)-I\left(X_1^N\to Y_1^N\right)=\sum_{i=1}^{N-1}I\left(X_{i+1}^N;Y_i\middle|X_1^i,Y_1^{i-1}\right).$$

So, $I\left(X_1^N\to Y_1^N\right)\le I\left(X_1^N;Y_1^N\right)$ with equality iff $\forall i\ \ Y_i-\left(X_1^i,Y_1^{i-1}\right)-X_{i+1}^N$ forms a Markov chain.

- If $X_{i+1}^N-X_1^i-Y_1^i$ is a Markov chain $\left(\text{i.e. } p\left(X_{i+1}^N\middle|X_1^i,Y_1^i\right)=p\left(X_{i+1}^N\middle|X_1^i\right)\right)$, then $I\left(X_1^N;Y_1^N\right)=I\left(X_1^N\to Y_1^N\right)$.
  - The Markov chain condition above says that the future *X*'s are not influenced by the past and current *Y*'s when conditioned on the past and current *X*'s. In the context of channels this state that if there is no feedback, then the two different mutual information measures are equal. [Tatikonda 2000, p. 81]
  - Pearl [1988] call this the "weak union" property of conditional independence.

  Proof. $p\left(X_{i+1}^N\middle|X_1^i,Y_1^i\right)=p\left(X_{i+1}^N\middle|X_1^i\right)\Rightarrow\ p\left(X_{i+1}^N\middle|X_1^i,Y_1^i\right)=p\left(X_{i+1}^N\middle|X_1^i,Y_1^{i-1}\right)\Rightarrow$ $Y_i-\left(X_1^i,Y_1^{i-1}\right)-X_{i+1}^N$ Markov.

  (Recall that $p\left(Z\middle|V,U_1,U_2\right)=p\left(Z\middle|V\right)\Rightarrow$ $p\left(Z\middle|V,U_1,U_2\right)=p\left(Z\middle|V,U_1\right)=p\left(Z\middle|V,U_2\right)=p\left(Z\middle|V\right)$.)

- For DMC, $I\left(X_1^N\to Y_1^N\right)\le\sum_{i=1}^{N}I\left(X_i;Y_i\right)$ with equality iff Y_i are independent.

  Proof. $I\left(X_1^i;Y_i\middle|Y_1^{i-1}\right)=H\left(Y_i\middle|Y_1^{i-1}\right)-H\left(Y_i\middle|X_1^i,Y_1^{i-1}\right)$
  $=H\left(Y_i\middle|Y_1^{i-1}\right)-H\left(Y_i\middle|X_i\right)$; memoryless
  $\le H\left(Y_i\right)-H\left(Y_i\middle|X_i\right)$
  $=I\left(X_i;Y_i\right)$

  Recall that $H\left(Y_1^N\right)=\sum_{i=1}^{N}H\left(Y_i\middle|Y_1^{i-1}\right)\le\sum_{i=1}^{N}H\left(Y_i\right)$ with equality iff Y_i are independent.

- *Definition*: Denote the sequence $\left(0,Y_1,\ldots,Y_{N-1}\right)$ by DY_1^N [Kramer 1998], or $0*Y_1^{N-1}$ [Massey 1990].
  - The letter D represent delay by one time step (with discard of the last component).

- $I\left(0*Y_1^{N-1}\to X_1^N\right)=\sum_{i=1}^{N}I\left(0*Y_1^{i-1};X_i\middle|X_1^{i-1}\right)=\sum_{i=1}^{N}I\left(Y_1^{i-1};X_i\middle|X_1^{i-1}\right)$
$=\sum_{i=2}^{N}I\left(Y_1^{i-1};X_i\middle|X_1^{i-1}\right)$

- Define $I\left(X_1^N \leftrightarrow Y_1^N\right) = I\left(Y_1^N \leftrightarrow X_1^N\right) = \sum_{i=1}^{N} I\left(Y_i; X_i \middle| Y_1^{i-1}, X_1^{i-1}\right)$.
- $I\left(X_1^N \to Y_1^N\right) = I\left(0 * X_1^{N-1} \to Y_1^N\right) + I\left(X_1^N \leftrightarrow Y_1^N\right)$

  $I\left(Y_1^N \to X_1^N\right) = I\left(0 * Y_1^{N-1} \to X_1^N\right) + I\left(Y_1^N \leftrightarrow X_1^N\right)$

  Proof. $I\left(Y_1^i; X_i \middle| X_1^{i-1}\right) - I\left(Y_1^{i-1}; X_i \middle| X_1^{i-1}\right) = I\left(Y_i; X_i \middle| Y_1^{i-1}, X_1^{i-1}\right)$

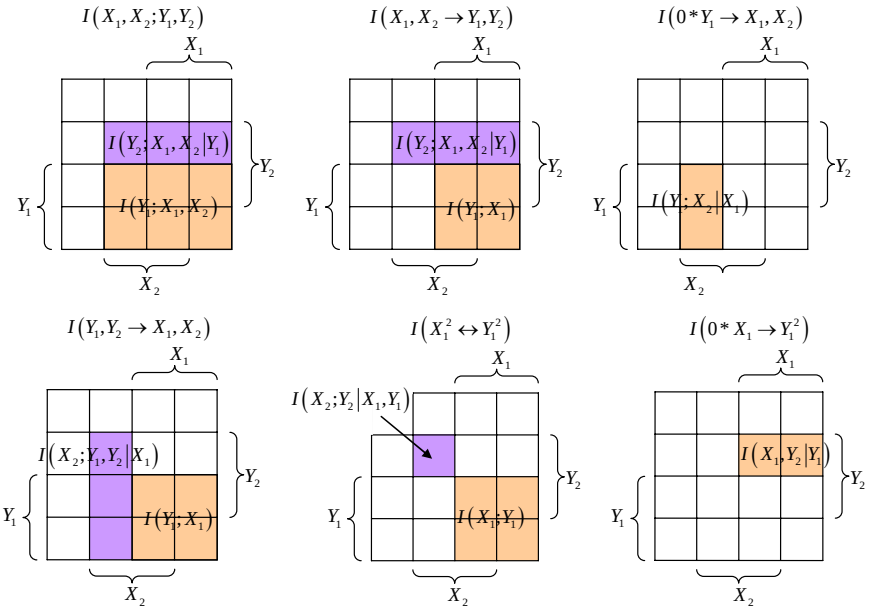
  $$I\left(Y_1^N \to X_1^N\right) - I\left(0 * Y_1^{N-1} \to X_1^N\right)$$
  $$= \sum_{i=1}^{N} I\left(Y_1^i; X_i \middle| X_1^{i-1}\right) - \sum_{i=1}^{N} I\left(Y_1^{i-1}; X_i \middle| X_1^{i-1}\right)$$
  $$= \sum_{i=1}^{N} I\left(Y_i; X_i \middle| Y_1^{i-1}, X_1^{i-1}\right)$$
- Conservation Law:

  $$I\left(X_1^N; Y_1^N\right) = I\left(X_1^N \to Y_1^N\right) + I\left(0 * Y_1^{N-1} \to X_1^N\right)$$
  $$= I\left(Y_1^N \to X_1^N\right) + I\left(0 * X_1^{N-1} \to Y_1^N\right)$$

  - Equivalently, $\sum_{i=2}^{N} I\left(Y_1^{i-1}; X_i \middle| X_1^{i-1}\right) = \sum_{i=1}^{N-1} I\left(X_{i+1}^N; Y_i \middle| X_1^i, Y_1^{i-1}\right)$.
  - *Interpretation*: By putting a 0 in the front,
    - We shift the delay position from the feedback to the channel. Before, X_i and Y_i are treated as if they happened during the same time step.
    - Now, the formula doesn't include the "middle" part which is $I\left(X_1^N \leftrightarrow Y_1^N\right)$.
- $I\left(X_1^N; Y_1^N\right) = I\left(X_1^N \to Y_1^N\right) + I\left(Y_1^N \to X_1^N\right) - I\left(X_1^N \leftrightarrow Y_1^N\right)$
  - In general, $I\left(X_1^N; Y_1^N\right) \neq I\left(X_1^N \to Y_1^N\right) + I\left(Y_1^N \to X_1^N\right)$.
    - This is obvious when $N = 1$ because

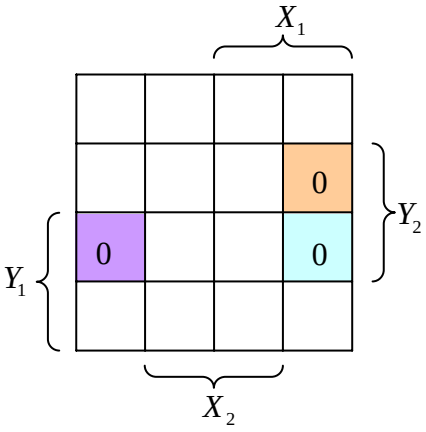
      $$I\left(X_1^N; Y_1^N\right) = I\left(X_1^N \to Y_1^N\right) = I\left(Y_1^N \to X_1^N\right).$$

- *Example*: (I-measure) $N = 2$

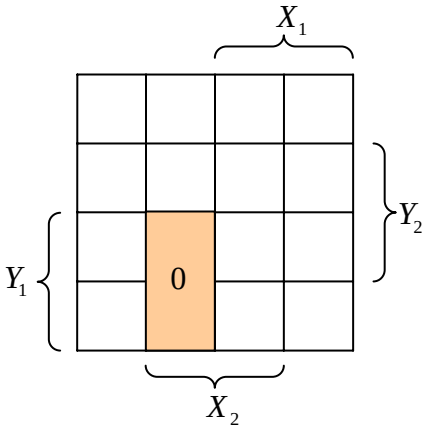


- DMC: $I(Y_2;(X_1,Y_1)|X_2)=0$ which equivalent to $I(Y_2;X_1|X_2,Y_1)=0$, $I(Y_2;Y_1|X_1,X_2)=0$, and $I(X_1;Y_1;Y_2|X_2)=0$.

DMC: $I(Y_2;(X_1,Y_1)|X_2)=0$



No feedback: $I(X_2;Y_1|X_1)=0$



We can clearly see now that when there is no feedback, $I(X_1^2;Y_1^2)=I(X_1^2 \to Y_1^2)$.

- We summarize the definitions and properties involving the directed information below:

- $I(X_1^N;Y_1^N)=\sum_{i=1}^{N} I(X_1^N;Y_i|Y_1^{i-1})$
- $I(X_1^N \to Y_1^N)=\sum_{i=1}^{N} I(X_1^i;Y_i|Y_1^{i-1})=H(Y_1^N)-\sum_{i=1}^{N} H(Y_i|X_1^i,Y_1^{i-1})$
- $I(0*Y_1^{N-1} \to X_1^N)=\sum_{i=2}^{N} I(Y_1^{i-1};X_i|X_1^{i-1})=\sum_{i=1}^{N-1} I(X_{i+1}^N;Y_i|X_1^i,Y_1^{i-1})$
  $=I(X_2;Y_1|X_1)+I(X_3;Y_1^2|X_1^2)+\cdots$

$$I\left(0*X_1^{N-1}\to Y_1^N\right)=\sum_{i=1}^{N}I\left(Y_i;X_1^{i-1}\middle|Y_1^{i-1}\right)$$
$$=I\left(X_1;Y_2\middle|Y_1\right)+I\left(X_1^2;Y_3\middle|Y_1^2\right)+\cdots$$

- $I\left(X_1^N\leftrightarrow Y_1^N\right)=I\left(Y_1^N\leftrightarrow X_1^N\right)=\sum_{i=1}^{N}I\left(Y_i;X_i\middle|Y_1^{i-1},X_1^{i-1}\right)$
- $I\left(X_1^N\to Y_1^N\right)=I\left(0*X_1^{N-1}\to Y_1^N\right)+I\left(X_1^N\leftrightarrow Y_1^N\right)$

  $I\left(Y_1^N\to X_1^N\right)=I\left(0*Y_1^{N-1}\to X_1^N\right)+I\left(Y_1^N\leftrightarrow X_1^N\right)$
- $I\left(X_1^N;Y_1^N\right)=I\left(X_1^N\to Y_1^N\right)+I\left(Y_1^N\to X_1^N\right)-I\left(X_1^N\leftrightarrow Y_1^N\right)$

  $=I\left(0*X_1^{N-1}\to Y_1^N\right)+I\left(0*Y_1^{N-1}\to X_1^N\right)+I\left(X_1^N\leftrightarrow Y_1^N\right)$

- [Tatikonda 2000] If the process $\left\{p\left(x_1^N,y_1^N\right)\right\}_{N=1}^{\infty}$ is information stable [p. 89], then $\lim_{N\to\infty}\frac{1}{N}I\left(X_1^N\to Y_1^N\right)$ exists and we can work directly with $I\left(X_1^N\to Y_1^N\right)$ (instead of liminf in probability as defined in [Tatikonda 2000 p. 89] and [Verdú 1994].
- *Definition*: [Kramer's] **causal conditioning**
  - We see above that the channel outputs are given by $p\left(y_1^N\middle\|x_1^N\right)=\prod_{n=1}^{N}p\left(y_n\middle|x_1^n,y_1^{n-1}\right)$. This lead us to define

    $H\left(Y_1^N\middle\|X_1^N\right)=-\mathbb{E}\left[p\left(Y_1^N\middle\|X_1^N\right)\right]=\sum_{n=1}^{N}H\left(Y_n\middle|X_1^n,Y_1^{n-1}\right).$
    - Again, $H\left(Y_1^N\middle\|X_1^N\right)$ differs from the conditional entropy $H\left(Y_1^N\middle|X_1^N\right)$ only in that X_1^n replaces X_1^N.
    - The term "causal" reflect the conditioning on past and present values of the sequence X_1^N only.
    - It differs from "free information" [Marko 1973] only in that X_n is included in the conditioning.
  - $p\left(x_1^N\middle\|\circ\, y_1^N\right)=\prod_{n=1}^{N}p\left(x_n\middle|x_1^{n-1},y_1^{n-1}\right);$

    $H\left(X_1^N\middle\|\circ\, Y_1^N\right)=-\mathbb{E}\left[\log p\left(X_1^N\middle\|\circ\, Y_1^N\right)\right]=\sum_{n=1}^{N}H\left(X_n\middle|X_1^{n-1},Y_1^{n-1}\right).$

    Note the ***asymmetry*** in the definitions above.
  - $p\left(x_1^N,y_1^N\right)=p\left(x_1^N\middle\|\circ\, y_1^N\right)p\left(y_1^N\middle\|x_1^N\right).$

    $H\left(X_1^N,Y_1^N\right)=H\left(X_1^N\middle\|\circ\, Y_1^N\right)+H\left(Y_1^N\middle\|X_1^N\right).$
  - $\tilde{p}\left(x_1^N,y_1^N\right)=p\left(x_1^N\middle\|\circ\, y_1^N\right)p\left(y_1^N\right)=p\left(y_1^N\right)\prod_{n=1}^{N}p\left(x_n\middle|x_1^{n-1},y_1^{n-1}\right)$

- $p\left(x_1^N \middle\| y_1^N\right) = \prod_{n=1}^{N} p\left(x_n \middle| x_1^{n-1}, y_1^n\right)$

  $H\left(X_1^N \middle\| Y_1^N\right) = -\mathbb{E}\left[\log p\left(X_1^N \middle\| Y_1^N\right)\right] = \sum_{n=1}^{N} H\left(X_n \middle| X_1^{n-1}, Y_1^n\right).$

- $H\left(X_1^N \middle\| Y_1^N\right) \le H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right)$

  Proof. Conditioning only reduces entropy.

- $H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right) - H\left(X_1^N \middle\| Y_1^N\right) = I\left(X_1^N \leftrightarrow Y_1^N\right)$

  Proof. $H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right) - H\left(X_1^N \middle\| Y_1^N\right) = \sum_{n=1}^{N}\left(H\left(X_n \middle| X_1^{n-1}, Y_1^{n-1}\right) - H\left(X_n \middle| X_1^{n-1}, Y_1^n\right)\right)$

  $= \sum_{n=1}^{N}\left(I\left(X_n; Y_n \middle| X_1^{n-1}, Y_1^{n-1}\right)\right)$

  $= I\left(X_1^N \leftrightarrow Y_1^N\right)$

- $I\left(X_1^N \to Y_1^N\right) = H\left(Y_1^N\right) - H\left(Y_1^N \middle\| X_1^N\right).$

  Proof. $I\left(X_1^N \to Y_1^N\right) = \sum_{i=1}^{N} I\left(X_1^i; Y_i \middle| Y_1^{i-1}\right) = H\left(Y_1^N\right) - \sum_{i=1}^{N} H\left(Y_i \middle| X_1^i, Y_1^{i-1}\right)$

  $= H\left(Y_1^N\right) - H\left(Y_1^N \middle\| X_1^N\right)$

  Alternative Proof. More directly,

  $$I\left(X_1^N \to Y_1^N\right) = \mathbb{E}\left[\log \frac{\prod_{i=1}^{N} p\left(Y_i \middle| X_1^i, Y_1^{i-1}\right)}{p\left(Y_1^N\right)}\right] = \mathbb{E}\left[\log \frac{p\left(Y_1^N \middle\| X_1^N\right)}{p\left(Y_1^N\right)}\right].$$

- $I\left(X_1^N \to Y_1^N\right) = \mathbb{E}\left[\log \frac{p\left(Y_1^N \middle\| X_1^N\right)}{p\left(Y_1^N\right)}\right] = \mathbb{E}\left[\log \frac{p\left(X_1^N, Y_1^N\right)}{p\left(X_1^N \middle\| {}^{\circ} Y_1^N\right) p\left(Y_1^N\right)}\right]$

  $= D\left(p\left(x_1^N, y_1^N\right) \middle\| p\left(x_1^N \middle\| {}^{\circ} y_1^N\right) p\left(y_1^N\right)\right)$

- $I\left(0 * Y_1^{N-1} \to X_1^N\right) = \sum_{i=1}^{N} I\left(Y_1^{i-1}; X_i \middle| X_1^{i-1}\right) = H\left(X_1^N\right) - H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right)$

  Proof. $I\left(0 * Y_1^{N-1} \to X_1^N\right) = \sum_{i=1}^{N} I\left(Y_1^{i-1}; X_i \middle| X_1^{i-1}\right).$

  $I\left(Y_1^{i-1}; X_i \middle| X_1^{i-1}\right) = H\left(X_i \middle| X_1^{i-1}\right) - H\left(X_i \middle| Y_1^{i-1}, X_1^{i-1}\right).$

- $I\left(X_1^N \to Y_1^N\right) = H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right) - H\left(X_1^N \middle| Y_1^N\right)$

  Proof. $I\left(X_1^N \to Y_1^N\right) = H\left(Y_1^N\right) - H\left(Y_1^N \middle\| X_1^N\right)$

  $= H\left(Y_1^N\right) - \left(H\left(X_1^N, Y_1^N\right) - H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right)\right)$

  $= H\left(X_1^N \middle\| {}^{\circ} Y_1^N\right) - H\left(X_1^N \middle| Y_1^N\right)$

- *Definition*: $I\left(X_1^N \to Y_1^N \middle\| Z_1^N\right) = H\left(Y_1^N \middle\| Z_1^N\right) - H\left(Y_1^N \middle\| X_1^N, Z_1^N\right)$.
  - $I\left(X_1^N \to Y_1^N \middle\| Z_1^N\right) = \sum_{n=1}^{N} I\left(Y_n; X_1^n \middle| Y_1^{n-1}, Z_1^n\right)$

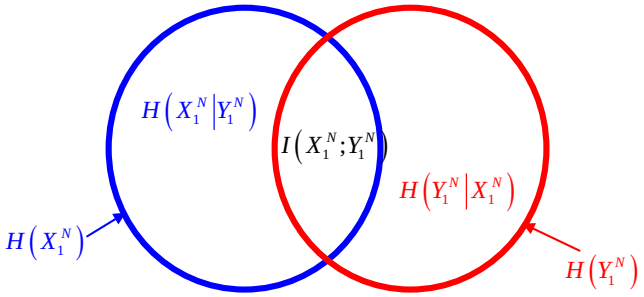
    Proof. $I\left(X_1^N \to Y_1^N \middle\| Z_1^N\right) = \sum_{n=1}^{N} H\left(Y_n \middle| Y_1^{n-1}, Z_1^n\right) - H\left(Y_n \middle| Y_1^{n-1}, X_1^n, Z_1^n\right)$

    $= \sum_{n=1}^{N} H\left(Y_n \middle| Y_1^{n-1}, Z_1^n\right) - H\left(Y_n \middle| Y_1^{n-1}, X_1^n, Z_1^n\right)$

    $= \sum_{n=1}^{N} I\left(Y_n; X_1^n \middle| Y_1^{n-1}, Z_1^n\right)$
  - Let $Z_1^N = 0 * X_1^{N-1}$, then $I\left(X_1^N \to Y_1^N \middle\| 0 * X_1^{N-1}\right) = I\left(X_1^N \leftrightarrow Y_1^N\right)$.

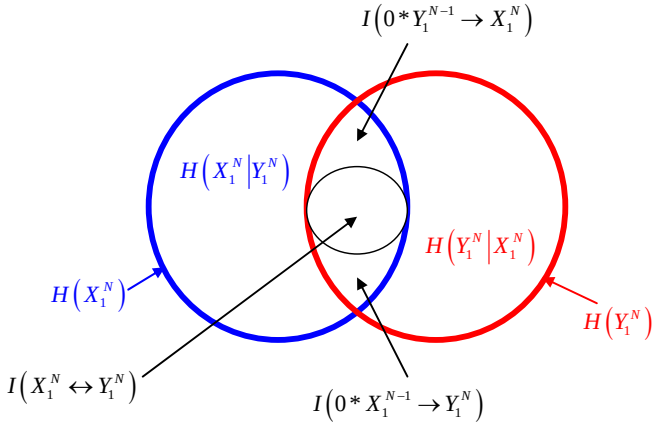
    Proof. $I\left(X_1^N \to Y_1^N \middle\| 0 * X_1^{N-1}\right) = \sum_{n=1}^{N} I\left(Y_n; X_1^n \middle| Y_1^{n-1}, X_1^{n-1}\right)$

    $= \sum_{n=1}^{N} I\left(Y_n; X_n \middle| Y_1^{n-1}, X_1^{n-1}\right)$
  - Similarly, by symmetry, we have $I\left(Y_1^N \to X_1^N \middle\| 0 * Y_1^{N-1}\right) = I\left(X_1^N \leftrightarrow Y_1^N\right)$.
- We conclude this section with diagrams. All the diagrams below are the same diagram. First, we have the familiar diagram:

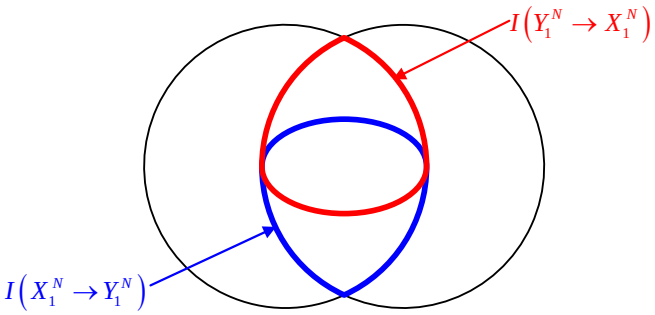


Next, the mutual information $I\left(X_1^N; Y_1^N\right)$ is partitioned into three subsets:

$$I\left(X_1^N; Y_1^N\right) = I\left(0 * X_1^{N-1} \to Y_1^N\right) + I\left(0 * Y_1^{N-1} \to X_1^N\right) + I\left(X_1^N \leftrightarrow Y_1^N\right).$$

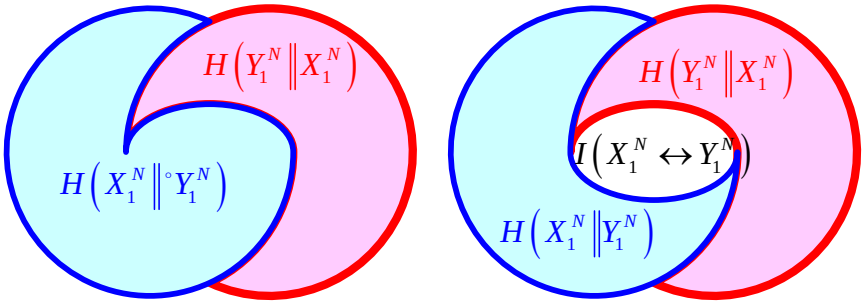


$I\left(X_1^N \to Y_1^N\right)$ and $I\left(Y_1^N \to X_1^N\right)$ are parts of $I\left(X_1^N; Y_1^N\right)$:

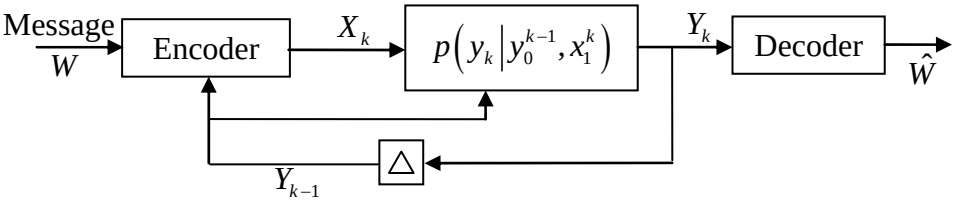


Finally, $H\left(X_1^N, Y_1^N\right) = H\left(X_1^N \,\|^{\circ}\, Y_1^N\right) + H\left(Y_1^N \,\|\, X_1^N\right)$

$= H\left(X_1^N \,\|\, Y_1^N\right) + H\left(Y_1^N \,\|\, X_1^N\right) + I\left(X_1^N \leftrightarrow Y_1^N\right)$



Converse Channel Coding Theorem



- ***Assume*** that the system is causal; that is $p\left(y_i \,|\, x_1^i, y_1^{i-1}, w\right) = p\left(y_i \,|\, x_1^i, y_1^{i-1}\right)$.

  The idea is that the source output sequence should be thought of a specified prior to the process of sending sequences over channels and the channel should be aware of such sequences only via its past inputs and outputs and its current input.

  - So, $p\left(w, x_1^N, y_1^N\right) = p(w)\prod_{n=1}^{N} p\left(x_n \,|\, x_1^{n-1}, y_1^{n-1}, w\right) p\left(y_n \,|\, x_1^n, y_1^{n-1}\right)$.

- $I\left(W; Y_0^N\right) \le I\left(X_1^N \to Y_1^N \,|\, Y_0\right)$ which is $\le I\left(X_1^N; Y_1^N \,|\, Y_0\right)$.

  Proof. $I\left(W; Y_0^N\right) = H\left(Y_0^N\right) - H\left(Y_0^N \,|\, W\right) = \sum_{k=1}^{N}\left(H\left(Y_k \,|\, Y_0^{k-1}\right) - H\left(Y_k \,|\, Y_0^{k-1}, W\right)\right)$.

  Because X_k is produced by $\left(Y_0^{k-1}, W\right)$, we have X_1^k is produced by $f\left(Y_0^{k-1}, W\right)$, and therefore

  $$H\left(Y_k \,|\, Y_0^{k-1}, W\right) = H\left(Y_k \,|\, Y_0^{k-1}, X_1^k, W\right) \overset{(a)}{=} H\left(Y_k \,|\, Y_0^{k-1}, X_1^k\right),$$

  where (*a*) comes from the causality assumption.

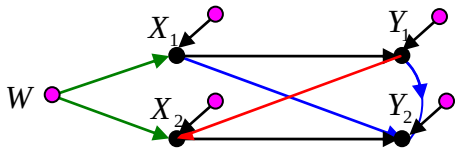
  A more general X_k gives

$$H\left(Y_k \middle| Y_0^{k-1}, W\right) \geq H\left(Y_k \middle| Y_0^{k-1}, X_1^k, W\right) \overset{(a)}{=} H\left(Y_k \middle| Y_0^{k-1}, X_1^k\right).$$

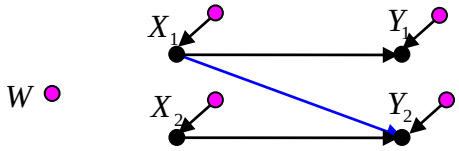
In any case,

$$I\left(W; Y_0^N\right) \leq \sum_{k=1}^{N}\left(H\left(Y_k \middle| Y_0^{k-1}\right) - H\left(Y_k \middle| Y_0^{k-1}, X_1^k\right)\right) = \sum_{k=1}^{N} I\left(Y_k; X_1^k \middle| Y_0^{k-1}\right).$$

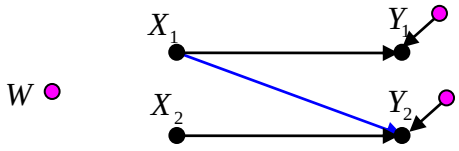
- *Remark*: Let's reconsider the equality of $H\left(Y_k \middle| Y_0^{k-1}, W\right)$ and $H\left(Y_k \middle| Y_0^{k-1}, X_1^k, W\right)$ using functional dependence graphs as in [Kramer 1998]. Suppose $k = 2$, then the relevant graph is shown below:



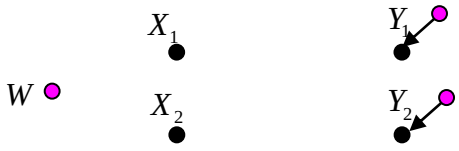
We then remove the arrows coming out of (W, Y_1). The resulted graph shows that (W, Y_1) does not *d*-separate Y_2 from X_1^2. Since all secondary random variables have incoming branches, we also know that (W, Y_1) does not *fd*-separate Y_2 from X_1^2.



Now, instead of generating X_n by conditional distribution $p\left(x_n \middle| x_1^{n-1}, y_1^{n-1}\right)$, we use deterministic encoder; that is $X_n = f_n\left(X_1^{n-1}, Y_1^{n-1}\right)$. Then, after removing the arrows coming out of (W, Y_1), the secondary random variables X_1 and X_2 have no incoming branches.



Hence, we can further delete the arrows coming out of X_1 and X_2.



We conclude that (W, Y_1) *fd*-separate Y_2 from X_1^2 in the case of deterministic encoder.

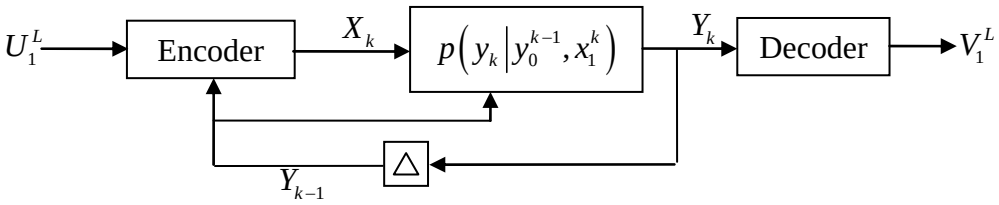
- Feedback does not increase the capacity of DMC.

Proof. For DMC, we know that $I\left(X_1^N \to Y_1^N \middle| Y_0\right) \leq \sum_{i=1}^{N} I\left(X_i; Y_i \middle| Y_0\right)$. Hence,

$$I\left(X_1^N \to Y_1^N \middle| Y_0\right) \le \sum_{i=1}^{N} I\left(X_i;Y_i\right) \le NC_{\text{DMC w/o feedback}}.$$

- This first result in the information theory of feedback channel is due to Shannon [1948].

- In fact, if we replace W by U_1^L and replace $\hat{W}$ by V_1^L.



Then,

- For any discrete channel, we have $I\left(U_1^L;Y_0^N\right) \le I\left(X_1^N \to Y_1^N \middle| Y_0\right)$.
- Furthermore, if the channel is memoryless, we have
  $I\left(X_1^N \to Y_1^N \middle| Y_0\right) \le \sum_{i=1}^{N} I\left(X_i;Y_i \middle| Y_0\right)$.

References

- C. Shannon, “The zero error capacity of a noisy channel,” 1956.
- Ash, Robert B., Information theory. Corrected reprint of the 1965 original. Dover Publications, Inc., New York, 1990. xii+339 pp.
- J. L. Massey, “Causality, Feedback and Directed Information,” pp. 303-305 in Proc. 1990 Int. Symp. on Info. Th. & its Appls., Hawaii, USA, Nov. 27-30, 1990.
- H. Marko, “The Bidirectional Communication Theory A Generalization of Information Theory”, IEEE Trans. Commun., vol. COM-21, pp. 1345-1351, Dec. 1973.
- S. Tatikonda, “Control Under Communication Constraints.” MIT Ph.D. thesis, August 2000.
- G. Kramer, Directed Information for Channels with Feedback, ETH Series in Inform. Proc., vol. 11. Konstanz: Hartung--Gorre, 1998. (PhD Thesis)
  - G. Kramer, “Capacity Results for the Discrete Memoryless Network,” IEEE Trans. Inform. Th., vol. IT-49, pp. 4-21, Jan. 2003.
- J. Chen, P. Suksompong and T. Berger, "Communication through a Finite-State Machine with Markov Property", CISS 2004