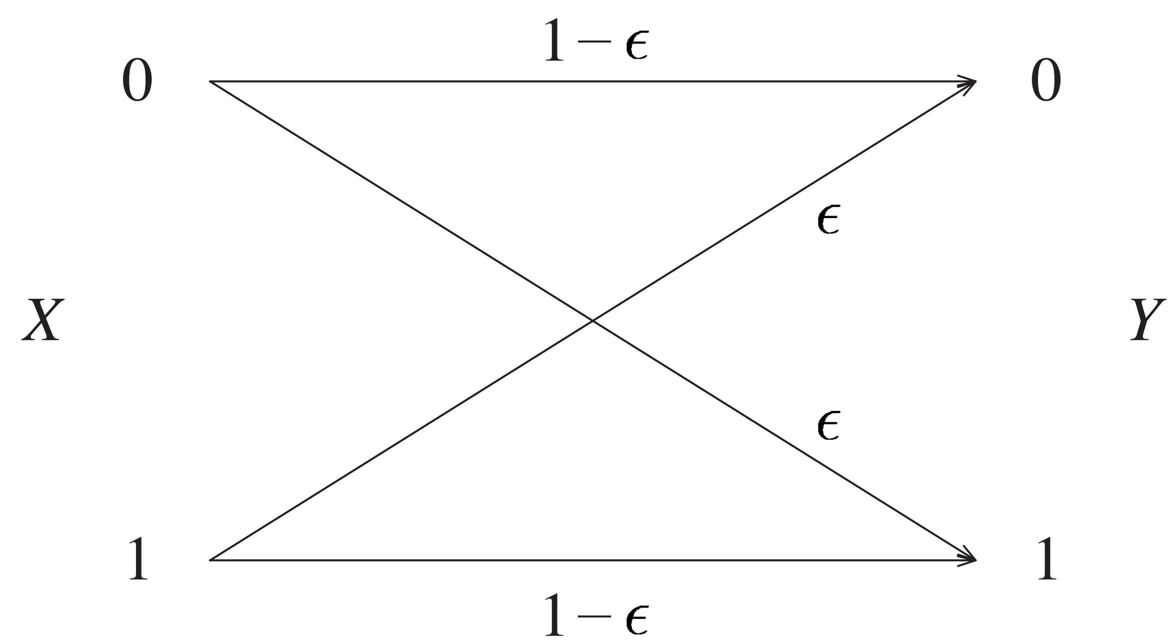
Chapter 7
Discrete Memoryless Channels

Department of Information Engineering
The Chinese University of Hong Kong

Binary Symmetric Channel



crossover probability $= \epsilon$

Repetition Channel Code

- Assume $\epsilon < 0.5$.
- $P_e = \epsilon$ if encode message A to 0 and message B to 1.
- To improve reliability, encode message A to $00\cdots0$ (n times) and message B to $11\cdots1$ (n times).
- $N_i = \#$ i's received, $i = 0, 1$.
- Receiver declares

$$\begin{cases} A & \text{if } N_0 > N_1 \\ B & \text{otherwise} \end{cases}$$

- If message is A, by WLLN, $N_0 \approx n(1-\epsilon)$ and $N_1 \approx n\epsilon$ w.p. $\to 1$ as $n \to \infty$.
- Decode correct w.p. $\to 1$ if message is A. Similarly if message is B.
- However, $R = \frac{1}{n} \log 2 \to 0$ as $n \to \infty$. :(

7.1 Definition and Capacity

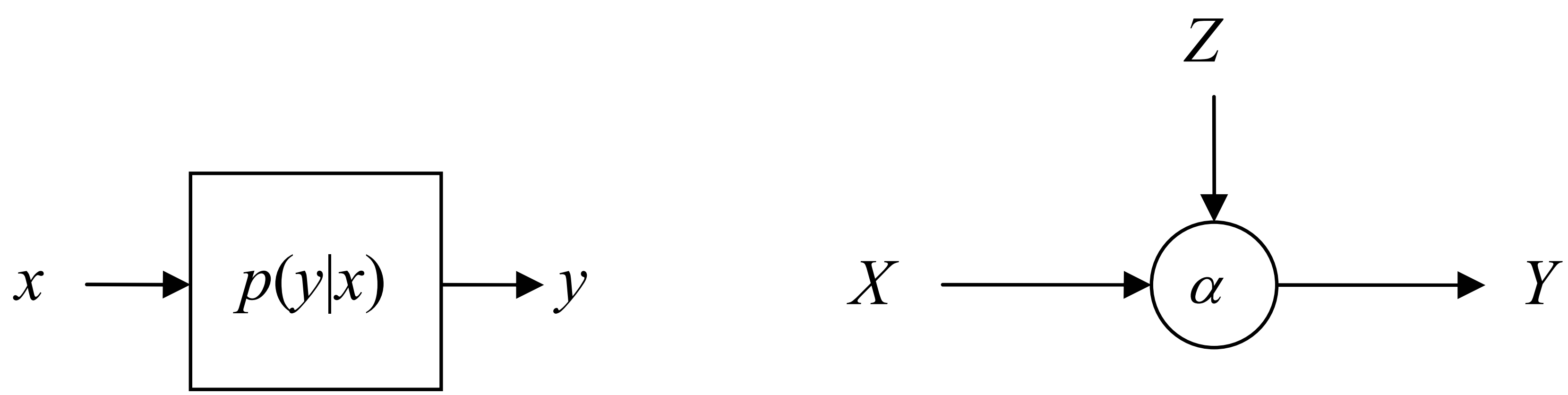
Definition 7.1 (Discrete Channel I) Let $\mathcal{X}$ and $\mathcal{Y}$ be discrete alphabets, and $p(y|x)$ be a transition matrix from $\mathcal{X}$ to $\mathcal{Y}$. A discrete channel $p(y|x)$ is a single-input single-output system with input random variable X taking values in $\mathcal{X}$ and output random variable Y taking values in $\mathcal{Y}$ such that

$$\Pr\{X = x, Y = y\} = \Pr\{X = x\}p(y|x)$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Definition 7.2 (Discrete Channel II) Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be discrete alphabets. Let $\alpha : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, and Z be a random variable taking values in $\mathcal{Z}$, called the noise variable. A discrete channel (α, Z) is a single-input single-output system with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. For any input random variable X, the noise variable Z is independent of X, and the output random variable Y is given by

$$Y = \alpha(X, Z).$$



(a) (b)

Two Equivalent Definitions for Discrete Channel

- II $\Rightarrow$ I: obvious
- I$\Rightarrow$ II:
  - Define r.v. Z_x with $\mathcal{Z}_x = \mathcal{Y}$ for $x \in \mathcal{X}$ such that $\Pr\{Z_x = y\} = p(y|x)$.
  - Assume Z_x, $x \in \mathcal{X}$ are mutually independent and also independent of X.
  - Define the noise variable $Z = (Z_x : x \in \mathcal{X})$.
  - Let $Y = Z_x$ if $X = x$, so that $Y = \alpha(X, Z)$.
  - Then

$$
\begin{aligned}
\Pr\{X = x, Y = y\} &= \Pr\{X = x\}\Pr\{Y = y|X = x\} \\
&= \Pr\{X = x\}\Pr\{Z_x = y|X = x\} \\
&= \Pr\{X = x\}\Pr\{Z_x = y\} \\
&= \Pr\{X = x\}p(y|x)
\end{aligned}
$$

Definition 7.3 Two discrete channels $p(y|x)$ and (α, Z) defined on the same input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ are equivalent if

$$\Pr\{\alpha(x, Z) = y\} = p(y|x)$$

for all x and y.

Some Basic Concepts

- A discrete channel can be used repeatedly at every time index $i = 1, 2, \cdots$.
- Assume the noise for the transmission over the channel at different time indices are independent of each other.
- To properly formulate a DMC, we regard it as a subsystem of a discrete-time stochastic system which will be referred to as “the system”.
- In such a system, random variables are generated sequentially in discrete-time.
- More than one random variable may be generated instantaneously but sequentially at a particular time index.

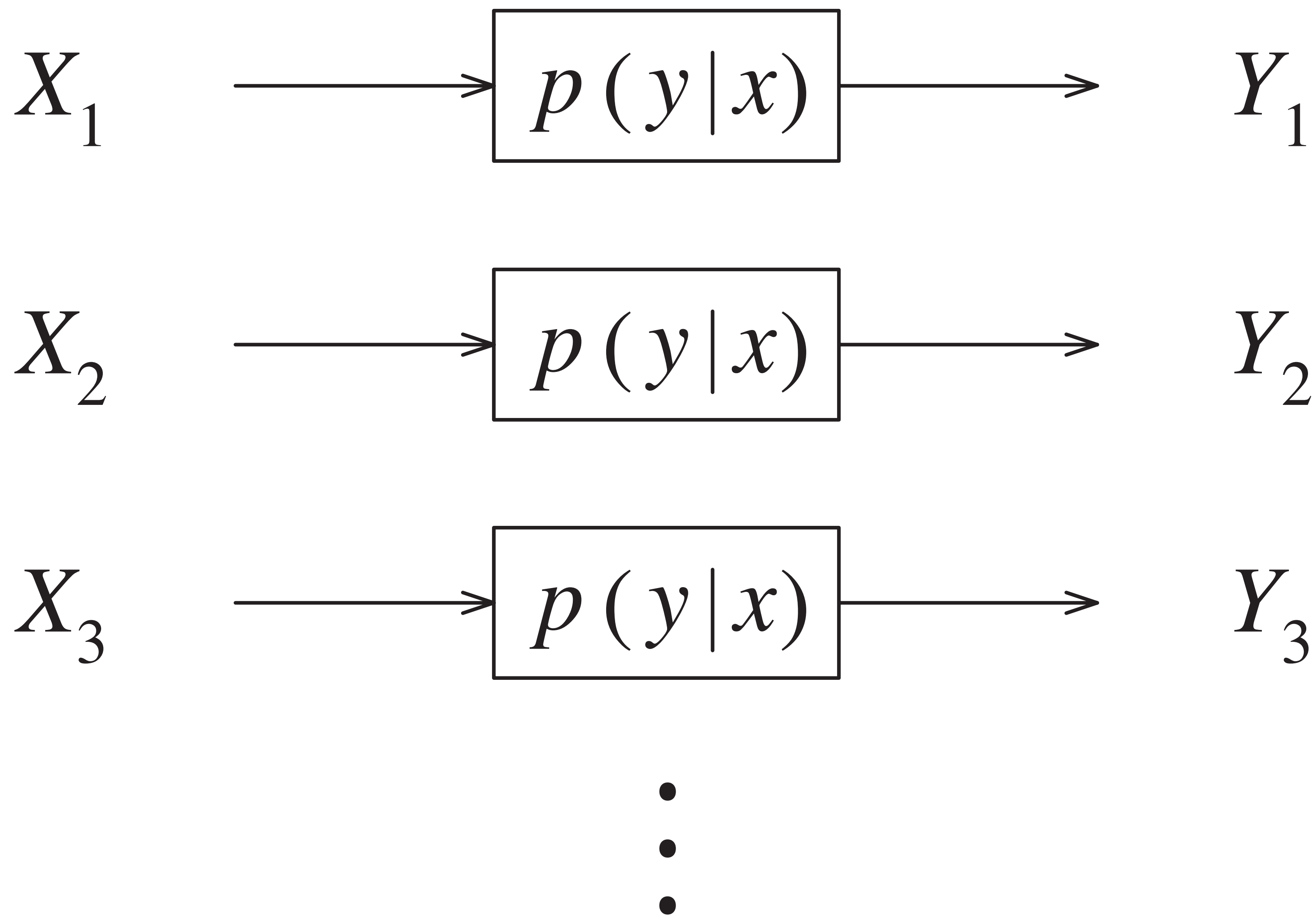
Definition 7.4 (DMC I) A discrete memoryless channel (DMC) $p(y|x)$ is a sequence of replicates of a generic discrete channel $p(y|x)$. These discrete channels are indexed by a discrete-time index i, where $i \geq 1$, with the ith channel being available for transmission at time i. Transmission through a channel is assumed to be instantaneous. Let X_i and Y_i be respectively the input and the output of the DMC at time i, and let T_{i-} denote all the random variables that are generated in the system before X_i. The equality

$$\Pr\{Y_i = y, X_i = x, T_{i-} = t\} = \Pr\{X_i = x, T_{i-} = t\}p(y|x)$$

holds for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{T}_{i-}$.

Remark: $T_{i-} \to X_i \to Y_i$, or

Given X_i, Y_i is independent of everything in the past.

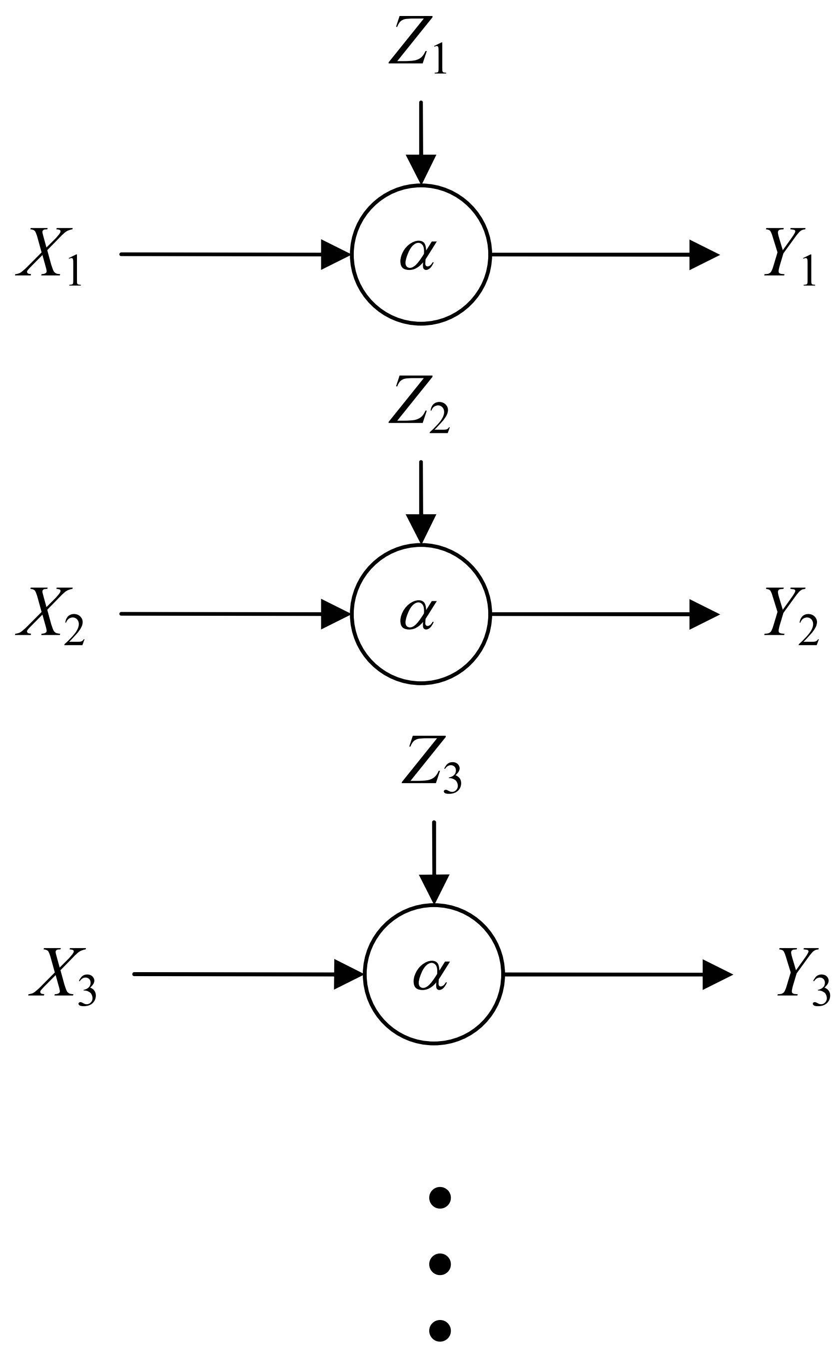


X_1
$p(y|x)$
Y_1
X_2
$p(y|x)$
Y_2
X_3
$p(y|x)$
Y_3

Definition 7.5 (DMC II) A discrete memoryless channel (α, Z) is a sequence of replicates of a generic discrete channel (α, Z). These discrete channels are indexed by a discrete-time index i, where $i \geq 1$, with the ith channel being available for transmission at time i. Transmission through a channel is assumed to be instantaneous. Let X_i and Y_i be respectively the input and the output of the DMC at time i, and let T_{i-} denote all the random variables that are generated in the system before X_i. The noise variable Z_i for the transmission at time i is a copy of the generic noise variable Z, and is independent of (X_i, T_{i-}). The output of the DMC at time i is given by

$$Y_i = \alpha(X_i, Z_i).$$

Remark: The equivalence of Definitions 7.4 and 7.5 can be shown. See textbook.



Z1
X1
α
Y1
Z2
X2
α
Y2
Z3
X3
α
Y3

Assume both $\mathcal{X}$ and $\mathcal{Y}$ are finite.

Definition 7.6 The capacity of a discrete memoryless channel $p(y|x)$ is defined as

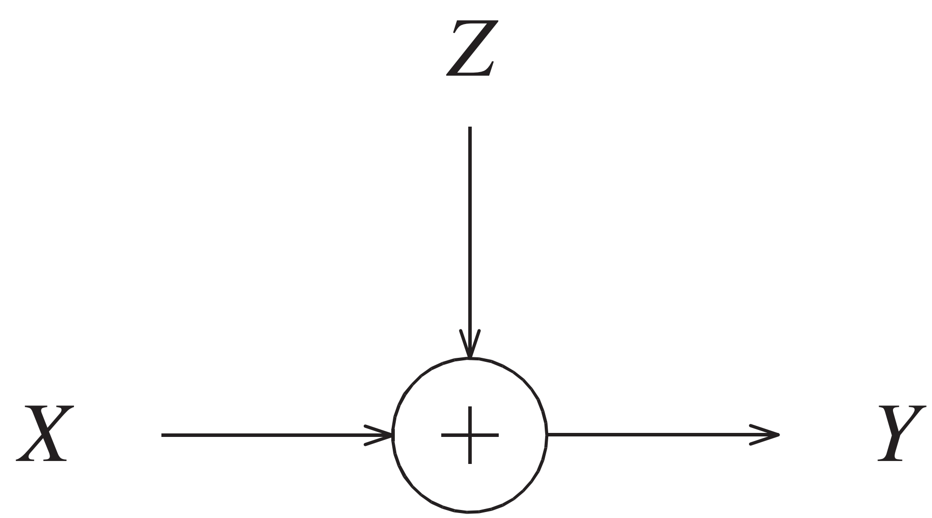
$$C = \max_{p(x)} I(X;Y),$$

where X and Y are respectively the input and the output of the generic discrete channel, and the maximum is taken over all input distributions $p(x)$.

Remarks:

- Since $I(X;Y)$ is a continuous functional of $p(x)$ and the set of all $p(x)$ is a compact set (i.e., closed and bounded) in $\Re^{|\mathcal{X}|}$, the maximum value of $I(X;Y)$ can be attained.
- Will see that C is in fact the maximum rate at which information can be communicated reliably through a DMC.
- Can communicate through a channel at a positive rate while $P_e \to 0$!

Example 7.7 (BSC)



Alternative representation of a BSC:

$$Y = X + Z \bmod 2$$

with

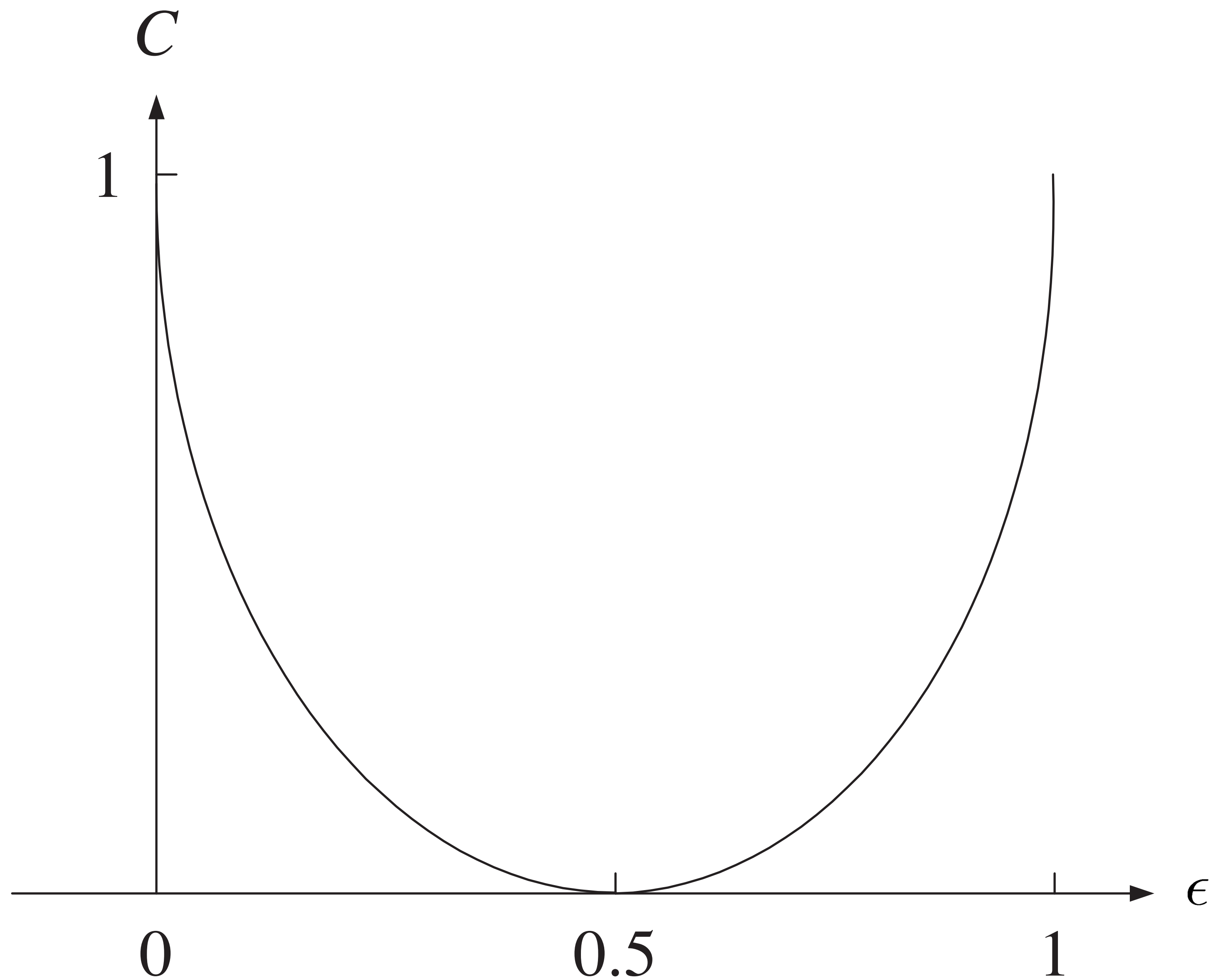
$$\Pr\{Z = 0\} = 1 - \epsilon \quad \text{and} \quad \Pr\{Z = 1\} = \epsilon$$

and Z is independent of X.

Determination of C:

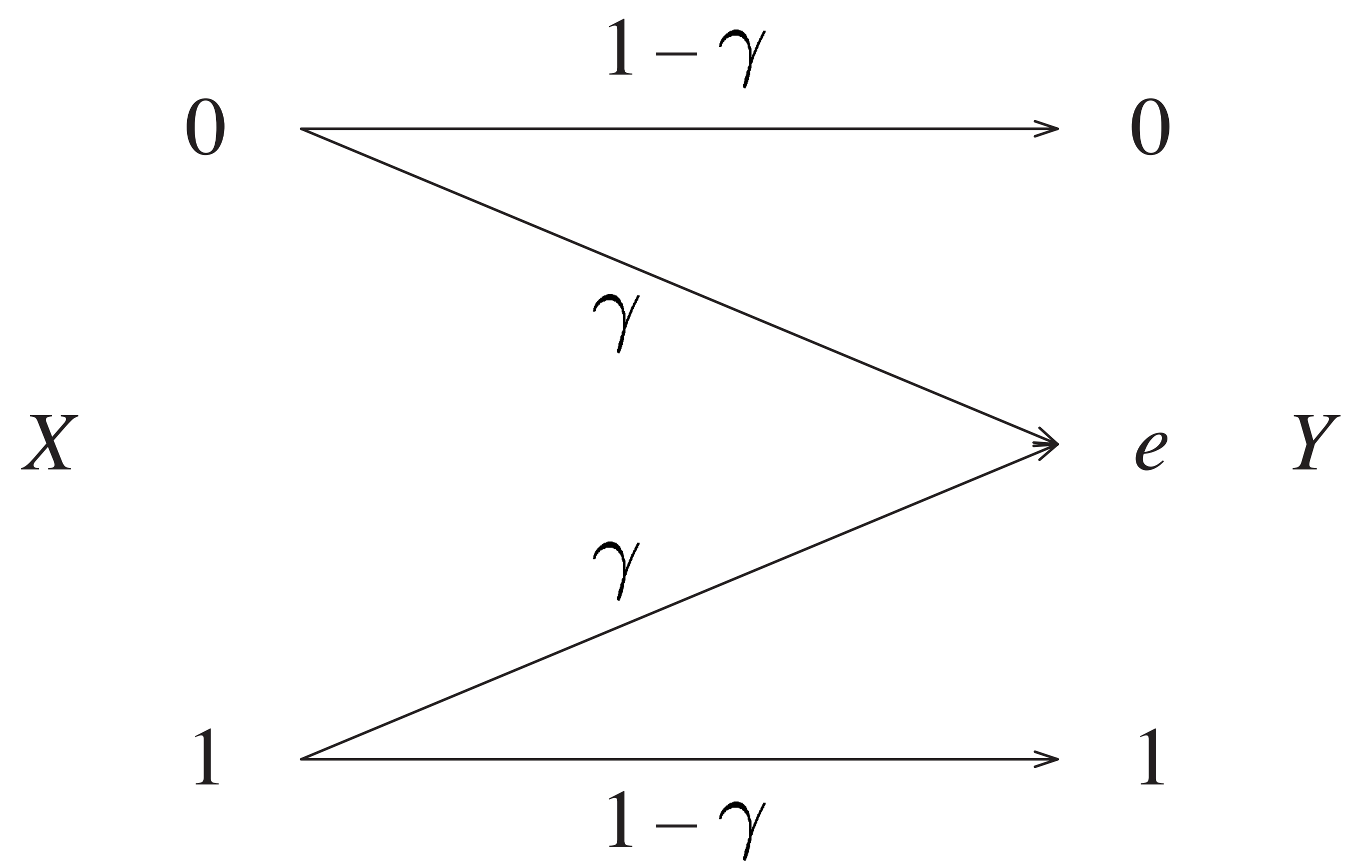
- $$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - \sum_x p(x) h_b(\epsilon) \\ &= H(Y) - h_b(\epsilon) \\ &\leq 1 - h_b(\epsilon) \end{aligned}$$

- So, $C \leq 1 - h_b(\epsilon)$.
- Tightness achieved by taking the uniform input distribution.
- Therefore, $C = 1 - h_b(\epsilon)$ bit per use.



C
1
ϵ
0
0.5
1

Example 7.8 (Binary Erasure Channel)



Erasure probability $= \gamma$; $C = (1-\gamma)$ bit per use

7.2 The Channel Coding Theorem

- **Direct Part** Information can be communicated through a DMC with an arbitrarily small probability of error at any rate less than the channel capacity.
- **Converse** If information is communicated through a DMC at a rate higher than the capacity, then the probability of error is bounded away from zero.

Definition of a Channel Code

Definition 7.9 An (n, M) code for a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined by an encoding function

$$f : \{1, 2, \cdots, M\} \rightarrow \mathcal{X}^n$$

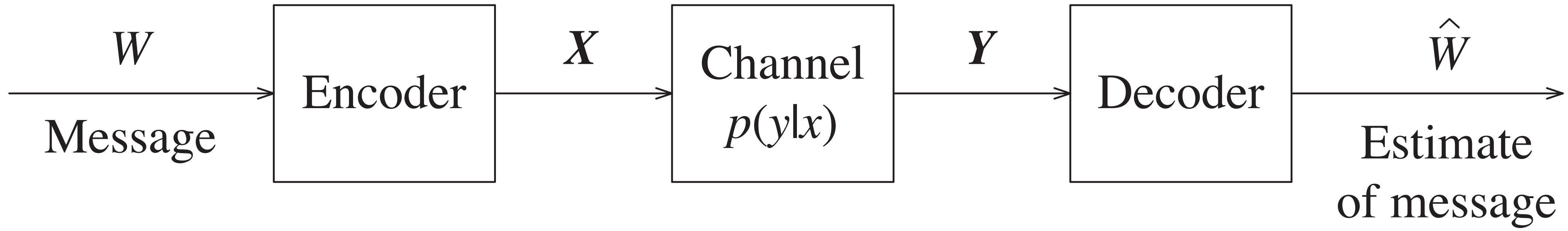
and a decoding function

$$g : \mathcal{Y}^n \rightarrow \{1, 2, \cdots, M\}.$$

- **Message Set** $\mathcal{W} = \{1, 2, \cdots, M\}$
- **Codewords** $f(1), f(2), \cdots, f(M)$
- **Codebook** The set of all codewords.

Assumptions and Notations

- W is randomly chosen from the message set $\mathcal{W}$, so $H(W) = \log M$.
- $\mathbf{X} = (X_1, X_2, \cdots, X_n)$; $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$
- Thus $\mathbf{X} = f(W)$.
- Let $\hat{W} = g(\mathbf{Y})$ be the estimate on the message W by the decoder.



Error Probabilities

Definition 7.10 For all $1 \leq w \leq M$, let

$$\lambda_w = \Pr\{\hat{W} \neq w | W = w\} = \sum_{\mathbf{y} \in \mathcal{Y}^n : g(\mathbf{y}) \neq w} \Pr\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = f(w)\}$$

be the conditional probability of error given that the message is w.

Definition 7.11 The maximal probability of error of an (n, M) code is defined as

$$\lambda_{max} = \max_w \lambda_w.$$

Definition 7.12 The average probability of error of an (n, M) code is defined as

$$P_e = \Pr\{\hat{W} \neq W\}.$$

P_e vs λ_{max}

- $$\begin{aligned} P_e &= \Pr\{\hat{W} \neq W\} \\ &= \sum_w \Pr\{W = w\} \Pr\{\hat{W} \neq W | W = w\} \\ &= \sum_w \frac{1}{M} \Pr\{\hat{W} \neq w | W = w\} \\ &= \frac{1}{M} \sum_w \lambda_w, \end{aligned}$$
- Therefore, $P_e \leq \lambda_{max}$.

Rate of a Channel Code

Definition 7.13 The rate of an (n, M) channel code is $n^{-1} \log M$ in bits per use.

Definition 7.14 A rate R is (asymptotically) achievable for a discrete memoryless channel if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\frac{1}{n} \log M > R - \epsilon$$

and

$$\lambda_{max} < \epsilon.$$

Theorem 7.15 (Channel Coding Theorem) A rate R is achievable for a discrete memoryless channel if and only if $R \leq C$, the capacity of the channel.

7.3 The Converse

- The communication system consists of the r.v.'s

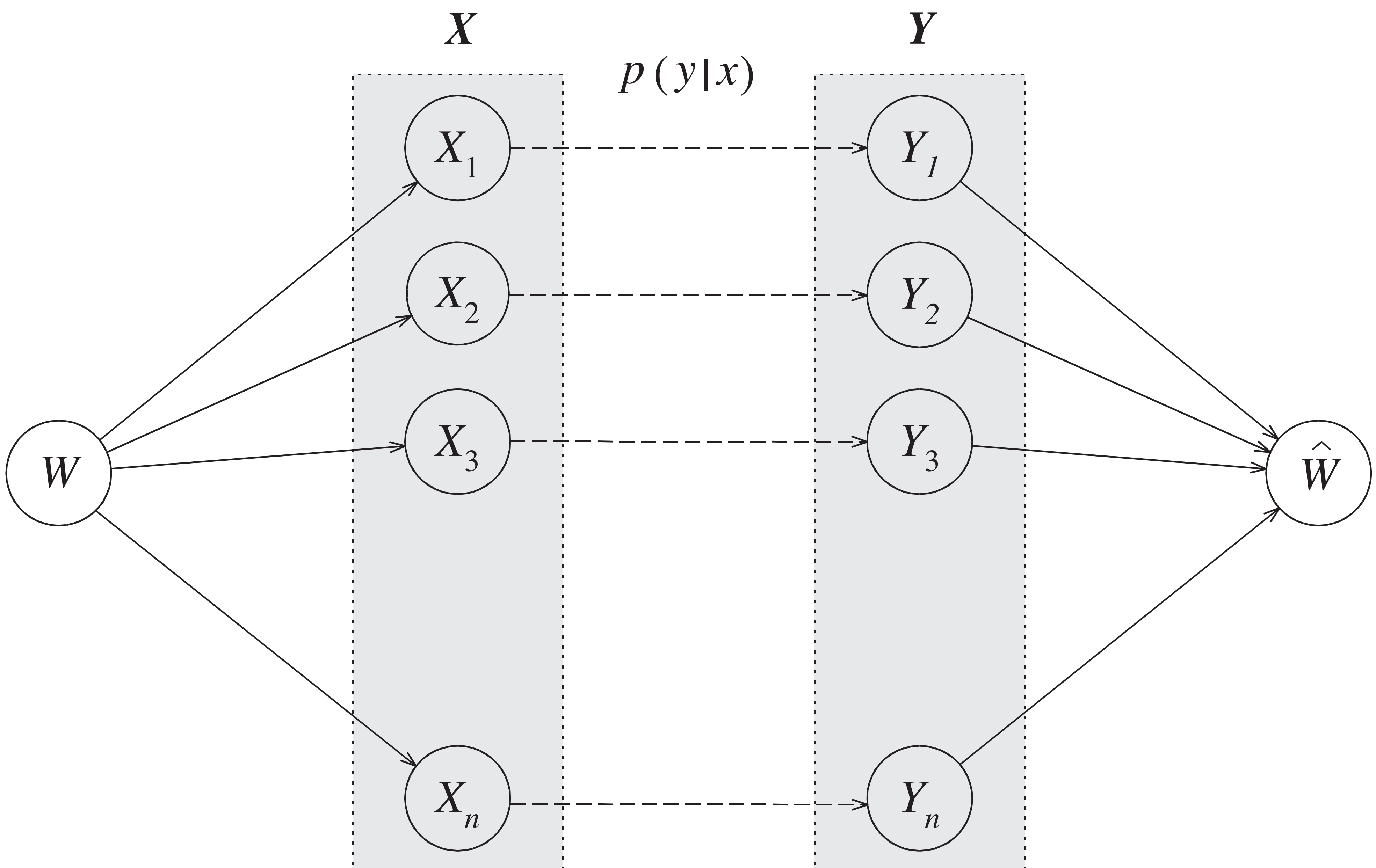
$$W, X_1, Y_1, X_2, Y_2, \cdots, X_n, Y_n, \hat{W}$$

  generated in this order.

- The memorylessness of the DMC imposes the following Markov constraint for each i:

$$(W, X_1, Y_1, \cdots, X_{i-1}, Y_{i-1}) \to X_i \to Y_i$$

- The dependency graph can be composed accordingly.



X
Y
p (y | x)
X1
X2
X3
Xn
Y1
Y2
Y3
Yn
W
Ŵ

- Use q to denote the joint distribution and marginal distributions of all r.v.'s.

- For all $(w, \mathbf{x}, \mathbf{y}, \hat{w}) \in \mathcal{W} \times \mathcal{X}^n \times \mathcal{Y}^n \times \hat{\mathcal{W}}$ such that $q(\mathbf{x}) > 0$ and $q(\mathbf{y}) > 0$,

$$q(w, \mathbf{x}, \mathbf{y}\, \hat{w}) = q(w) \left(\prod_{i=1}^{n} q(x_i|w) \right) \left(\prod_{i=1}^{n} p(y_i|x_i) \right) q(\hat{w}|\mathbf{y}).$$

- $q(w) > 0$ for all w so that $q(x_i|w)$ are well-defined.

- $q(x_i|w)$ and $q(\hat{w}|\mathbf{y})$ are deterministic.

- The dependency graph suggests the Markov chain $W \to \mathbf{X} \to \mathbf{Y} \to \hat{W}$.

- This can be formally justified by invoking Proposition 2.9.

Show that for $\mathbf{x}$ such that $q(\mathbf{x}) > 0$,

$$q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p(y_i|x_i)$$

First, for $\mathbf{x}$ and $\mathbf{y}$ such that $q(\mathbf{x}) > 0$ and $q(\mathbf{y}) > 0$,

$$\begin{aligned}
q(\mathbf{x}, \mathbf{y}) &= \sum_{w} \sum_{\hat{w}} q(w, \mathbf{x}, \mathbf{y}, \hat{w}) \\
&= \sum_{w} \sum_{\hat{w}} q(w) \left(\prod_{i} q(x_i|w) \right) \left(\prod_{i} p(y_i|x_i) \right) q(\hat{w}|\mathbf{y}) \\
&= \sum_{w} q(w) \left(\prod_{i} q(x_i|w) \right) \left(\prod_{i} p(y_i|x_i) \right) \sum_{\hat{w}} q(\hat{w}|\mathbf{y}) \\
&= \left[\sum_{w} q(w) \prod_{i} q(x_i|w) \right] \left[\prod_{i} p(y_i|x_i) \right]
\end{aligned}$$

Furthermore,

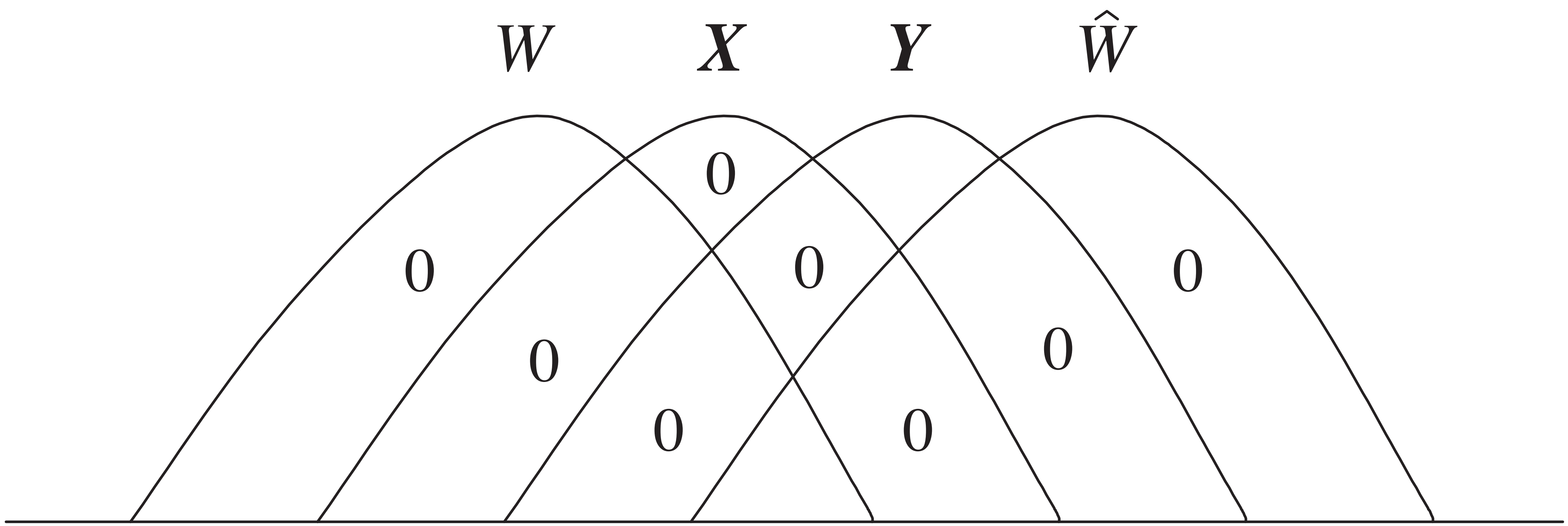
$$\begin{aligned}
q(\mathbf{x}) &= \sum_{\mathbf{y}} q(\mathbf{x}, \mathbf{y}) \\
&= \sum_{\mathbf{y}} \left[\sum_{w} q(w) \prod_{i} q(x_i|w) \right] \left[\prod_{i} p(y_i|x_i) \right] \\
&= \left[\sum_{w} q(w) \prod_{i} q(x_i|w) \right] \left[\sum_{y_1} \sum_{y_2} \cdots \sum_{y_n} \prod_{i} p(y_i|x_i) \right] \\
&= \left[\sum_{w} q(w) \prod_{i} q(x_i|w) \right] \prod_{i} \left(\sum_{y_i} p(y_i|x_i) \right) \\
&= \sum_{w} q(w) \prod_{i} q(x_i|w)
\end{aligned}$$

Therefore, for $\mathbf{x}$ such that $q(\mathbf{x}) > 0$,

$$q(\mathbf{y}|\mathbf{x}) = \frac{q(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})} = \prod_{i} p(y_i|x_i)$$

Why C is related to I(X;Y)?

- $H(\mathbf{X}|W) = 0$
- $H(\hat{W}|\mathbf{Y}) = 0$
- Since W and $\hat{W}$ are essentially identical for reliable communication, assume
$$H(\hat{W}|W) = H(W|\hat{W}) = 0$$
- Then from the information diagram for $W \to \mathbf{X} \to \mathbf{Y} \to \hat{W}$, we see that
$$H(W) = I(\mathbf{X}; \mathbf{Y}).$$
- This suggests that the channel capacity is obtained by maximizing $I(X;Y)$.



W
$\boldsymbol{X}$
$\boldsymbol{Y}$
$\hat{W}$
0
0
0
0
0
0
0
0
0
0

Building Blocks of the Converse

- For all $1 \leq i \leq n$,

$$I(X_i; Y_i) \leq C$$

- Then

$$\sum_{i=1}^{n} I(X_i; Y_i) \leq nC$$

- To be established in Lemma 7.16,

$$I(\mathbf{X}; \mathbf{Y}) \leq \sum_{i=1}^{n} I(X_i; Y_i)$$

- Therefore,

$$\begin{aligned}\frac{1}{n}\log M &= \frac{1}{n}H(W) \\ &= \frac{1}{n}I(\mathbf{X};\mathbf{Y}) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) \\ &\leq C\end{aligned}$$

Lemma 7.16 $I(\mathbf{X};\mathbf{Y}) \leq \sum_{i=1}^{n} I(X_i;Y_i)$

Proof

1. Establish

$$H(\mathbf{Y}|\mathbf{X}) = \sum_{i=1}^{n} H(Y_i|X_i)$$

2.

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i;Y_i) \end{aligned}$$

Formal Converse Proof

1. Let R be an achievable rate, i.e., for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\frac{1}{n}\log M > R - \epsilon \quad \text{and} \quad \lambda_{max} < \epsilon$$

2. Consider

$$\begin{aligned}
\log M &\stackrel{a)}{=} H(W) \\
&= H(W|\hat{W}) + I(W;\hat{W}) \\
&\stackrel{b)}{\leq} H(W|\hat{W}) + I(\mathbf{X};\mathbf{Y}) \\
&\stackrel{c)}{\leq} H(W|\hat{W}) + \sum_{i=1}^{n} I(X_i;Y_i) \\
&\stackrel{d)}{\leq} H(W|\hat{W}) + nC,
\end{aligned}$$

3. By Fano's inequality,

$$H(W|\hat{W}) < 1 + P_e \log M$$

4. Then,

$$\begin{aligned}\log M &< 1 + P_e \log M + nC \\ &\leq 1 + \lambda_{max} \log M + nC \\ &< 1 + \epsilon \log M + nC,\end{aligned}$$

   Therefore,

$$R - \epsilon < \frac{1}{n} \log M < \frac{\frac{1}{n} + C}{1 - \epsilon}$$

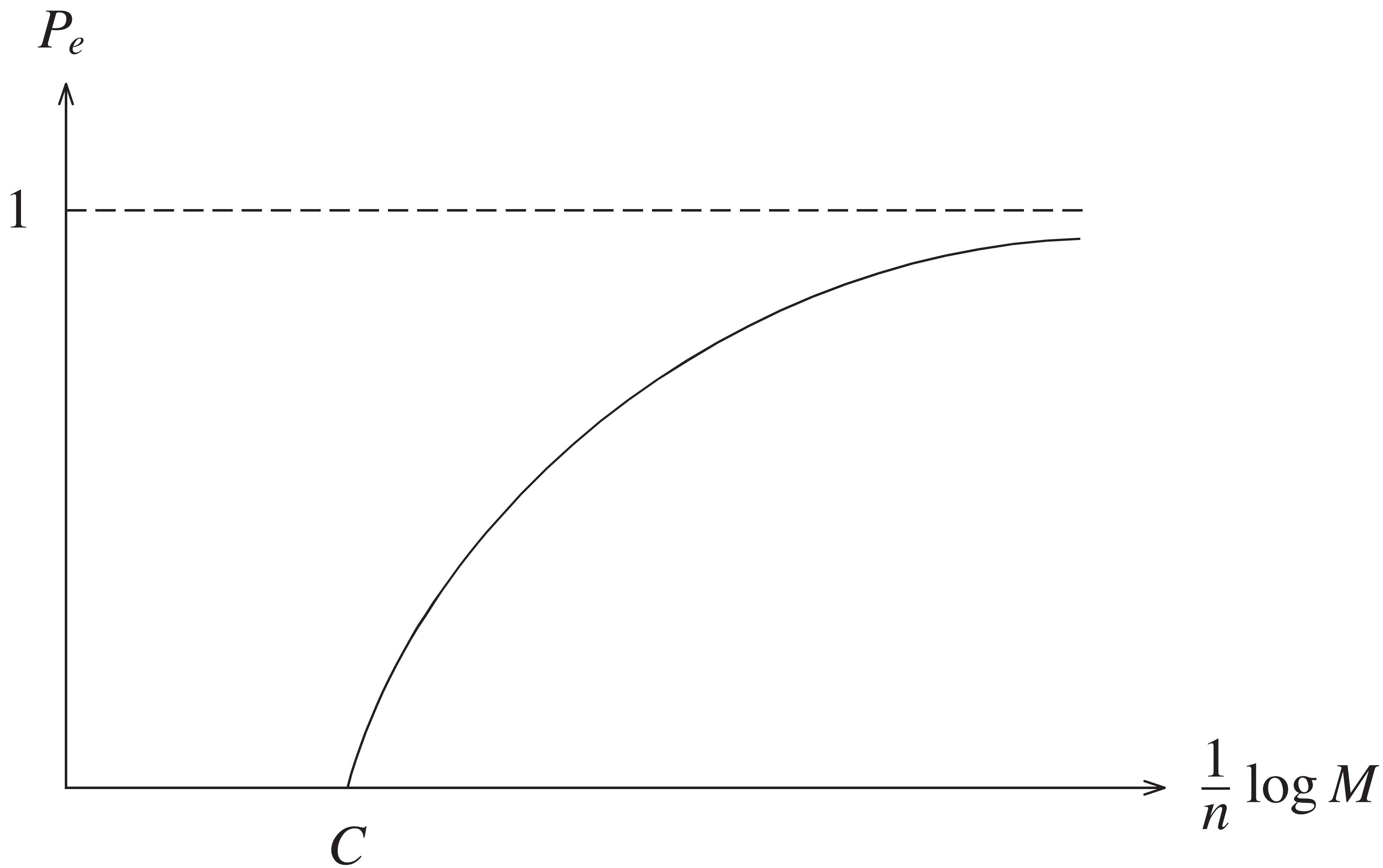
5. Letting $n \to \infty$ and then $\epsilon \to 0$ to conclude that $R \leq C$.

Asymptotic Bound for P_e: Weak Converse

- For large n,

$$P_e \geq 1 - \frac{1+nC}{\log M} = 1 - \frac{\frac{1}{n}+C}{\frac{1}{n}\log M} \approx 1 - \frac{C}{\frac{1}{n}\log M}$$

- $\frac{1}{n}\log M$ is the actual rate of the channel code.
- If $\frac{1}{n}\log M > C$, then $P_e > 0$ for large n.
- This implies that if $\frac{1}{n}\log M > C$, then $P_e > 0$ for all n.



P_e
1
$\frac{1}{n}\log M$
C

Strong Converse

- If there exists an $\epsilon > 0$ such that $\frac{1}{n} \log M \geq C + \epsilon$ for all n, then $P_e \to 1$ as $n \to \infty$.

7.4 Achievability

- Consider a DMC $p(y|x)$.
- For every input distribution $p(x)$, prove that the rate $I(X;Y)$ is achievable by showing for large n the existence of a channel code such that
  1. the rate of the code is arbitrarily close to $I(X;Y)$;
  2. the maximal probability of error λ_{max} is arbitrarily small.
- Choose the input distribution $p(x)$ to be one that achieves the channel capacity, i.e., $I(X;Y) = C$.

Lemma 7.17 Let $(\mathbf{X}', \mathbf{Y}')$ be n i.i.d. copies of a pair of generic random variables (X', Y'), where X' and Y' are independent and have the same marginal distributions as X and Y, respectively. Then

$$\Pr\{(\mathbf{X}', \mathbf{Y}') \in T^n_{[XY]\delta}\} \leq 2^{-n(I(X;Y)-\tau)},$$

where $\tau \to 0$ as $\delta \to 0$.

Proof of Lemma 7.17

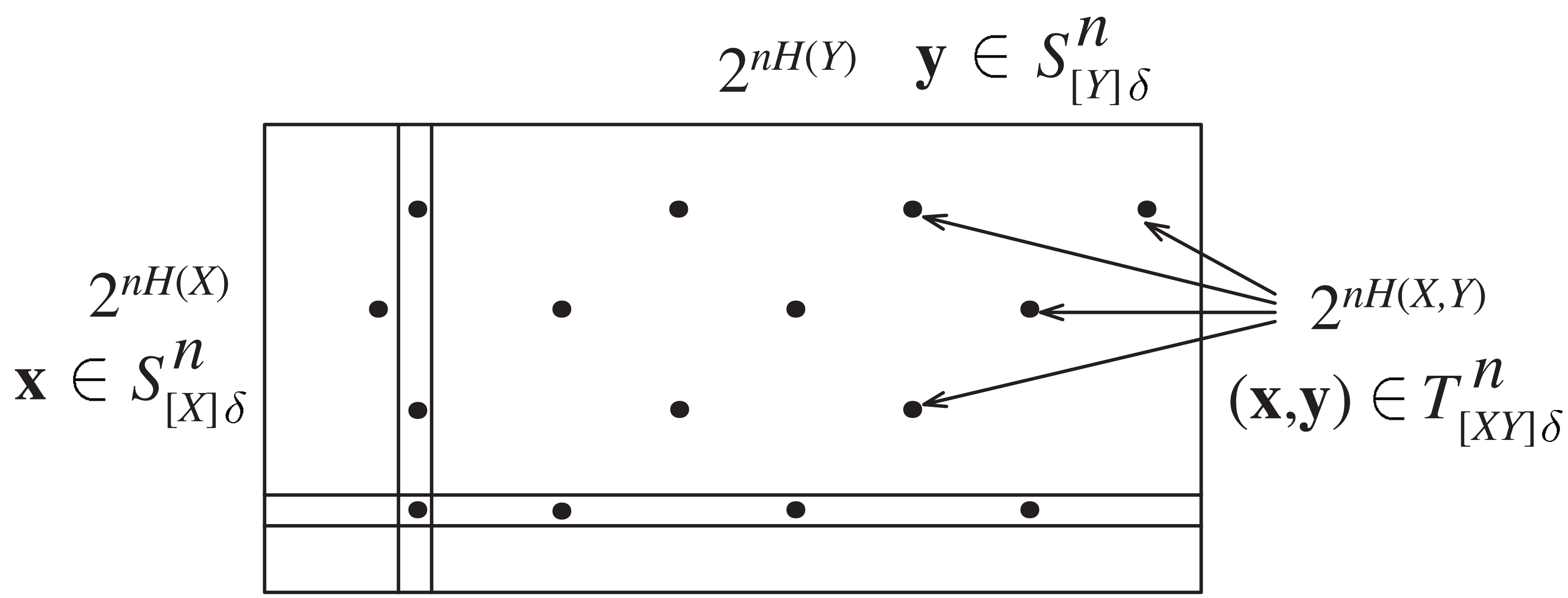
- Consider

$$\Pr\{(\mathbf{X}', \mathbf{Y}') \in T^n_{[XY]\delta}\} = \sum_{(\mathbf{x},\mathbf{y}) \in T^n_{[XY]\delta}} p(\mathbf{x})p(\mathbf{y})$$

- Consistency of strong typicality: $\mathbf{x} \in T^n_{[X]\delta}$ and $\mathbf{y} \in T^n_{[Y]\delta}$.
- Strong AEP: $p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}$ and $p(\mathbf{y}) \leq 2^{-n(H(Y)-\zeta)}$.
- Strong JAEP: $|T^n_{[XY]\delta}| \leq 2^{n(H(X,Y)+\xi)}$.
- Then

$$\begin{aligned}
&\Pr\{(\mathbf{X}', \mathbf{Y}') \in T^n_{[XY]\delta}\} \\
&\quad \leq \quad 2^{n(H(X,Y)+\xi)} \cdot 2^{-n(H(X)-\eta)} \cdot 2^{-n(H(Y)-\zeta)} \\
&\quad = \quad 2^{-n(H(X)+H(Y)-H(X,Y)-\xi-\eta-\zeta)} \\
&\quad = \quad 2^{-n(I(X;Y)-\xi-\eta-\zeta)} \\
&\quad = \quad 2^{-n(I(X;Y)-\tau)}
\end{aligned}$$

An Interpretation of Lemma 7.17



- Randomly choose a row with uniform distribution and randomly choose a column with uniform distribution.
-

$$\Pr\{\text{Obtaining a jointly typical pair}\} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI((X;Y)}$$

Random Coding Scheme

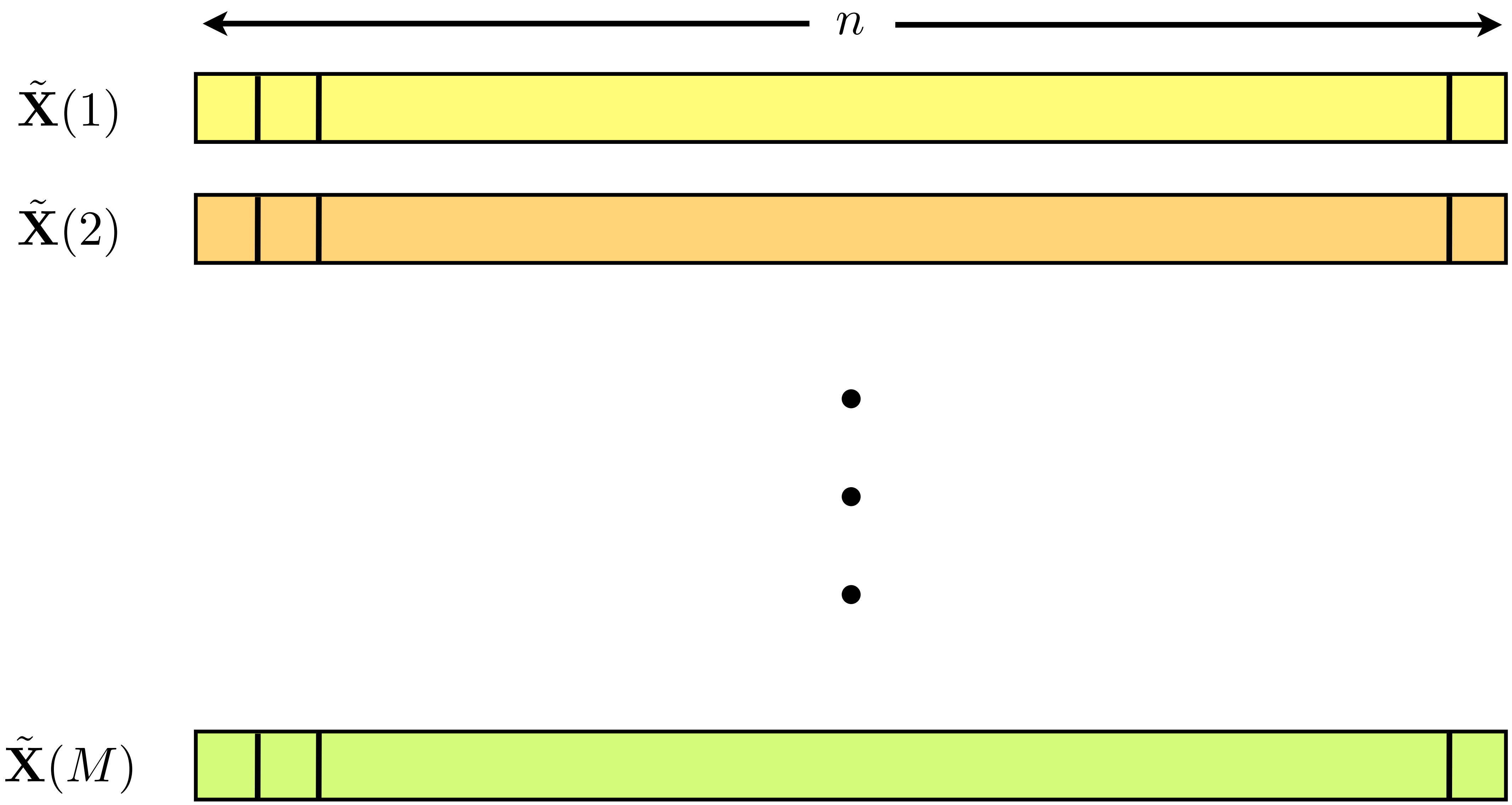
- Fix $\epsilon > 0$ and input distribution $p(x)$. Let δ to be specified later.
- Let M be an even integer satisfying

$$I(X;Y) - \frac{\epsilon}{2} < \frac{1}{n}\log M < I(X;Y) - \frac{\epsilon}{4},$$

  where n is sufficiently large, i.e., $M \approx 2^{nI(X;Y)}$.

The random coding scheme:

1. Construct the codebook $\mathcal{C}$ of an (n, M) code by generating M codewords in $\mathcal{X}^n$ independently and identically according to $p(x)^n$. Denote these codewords by $\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \cdots, \tilde{\mathbf{X}}(M)$.
2. Reveal the codebook $\mathcal{C}$ to both the encoder and the decoder.
3. A message W is chosen from $\mathcal{W}$ according to the uniform distribution.
4. Transmit $\mathbf{X} = \tilde{\mathbf{X}}(W)$ through the channel.



- Generate each component according to $p(x)$.
- There are a total of $|\mathcal{X}|^{Mn}$ possible codebooks that can be constructed.
- Regard two codebooks whose sets of codewords are permutations of each other as two different codebooks.

5. The channel outputs a sequence $\mathbf{Y}$ according to

$$\Pr\{\mathbf{Y} = \mathbf{y} | \tilde{\mathbf{X}}(W) = \mathbf{x}\} = \prod_{i=1}^{n} p(y_i | x_i)$$

6. The sequence $\mathbf{Y}$ is decoded to the message w if

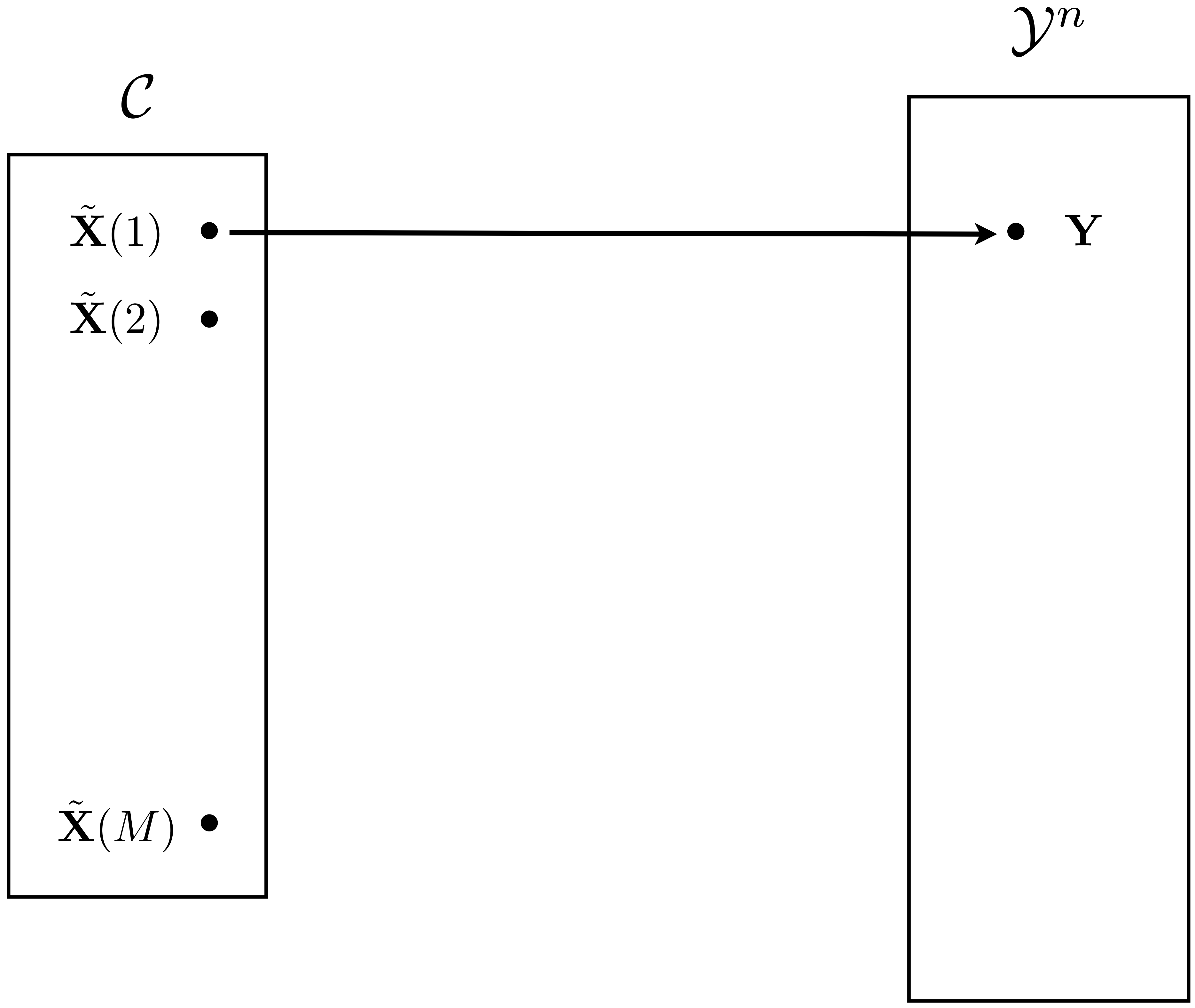
   - $(\tilde{\mathbf{X}}(w), \mathbf{Y}) \in T^n_{[XY]\delta}$, and
   - there does not exists $w' \neq w$ such that $(\tilde{\mathbf{X}}(w'), \mathbf{Y}) \in T^n_{[XY]\delta}$.

   Otherwise, $\mathbf{Y}$ is decoded to a constant message in $\mathcal{W}$. Denote by $\hat{W}$ the message to which $\mathbf{Y}$ is decoded.

Performance Analysis

- To show that $\Pr\{Err\} = \Pr\{\hat{W} \neq W\}$ can be arbitrarily small.
- $$\begin{aligned} \Pr\{Err\} &= \sum_{w=1}^{M} \Pr\{Err|W = w\}\Pr\{W = w\} \\ &= \Pr\{Err|W = 1\} \sum_{w=1}^{M} \Pr\{W = w\} \\ &= \Pr\{Err|W = 1\} \end{aligned}$$
- For $1 \leq w \leq M$, define the event

$$E_w = \{(\tilde{\mathbf{X}}(w), \mathbf{Y}) \in T^n_{[XY]\delta}\}$$



$\mathcal{C}$
$\tilde{\mathbf{X}}(1)$
$\tilde{\mathbf{X}}(2)$
$\tilde{\mathbf{X}}(M)$
$\mathcal{Y}^n$
$\mathbf{Y}$

- If E_1 occurs but E_w does not occur for all $2 \leq w \leq M$, then no decoding error. Therefore,

$$\Pr\{Err^c|W=1\} \geq \Pr\{E_1 \cap E_2^c \cap E_3^c \cap \cdots \cap E_M^c|W=1\}$$

-

$$\begin{aligned}
\Pr\{Err|W=1\} &= 1 - \Pr\{Err^c|W=1\} \\
&\leq 1 - \Pr\{E_1 \cap E_2^c \cap E_3^c \cap \cdots \cap E_M^c|W=1\} \\
&= \Pr\{(E_1 \cap E_2^c \cap E_3^c \cap \cdots \cap E_M^c)^c|W=1\} \\
&= \Pr\{E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_M|W=1\}
\end{aligned}$$

- By the union bound,

$$\Pr\{Err|W=1\} \leq \Pr\{E_1^c|W=1\} + \sum_{w=2}^{M} \Pr\{E_w|W=1\}$$

- By strong JAEP,

$$\Pr\{E_1^c|W=1\} = \Pr\{(\tilde{\mathbf{X}}(1), \mathbf{Y}) \notin T^n_{[XY]\delta}|W=1\} < \nu$$

- Conditioning on $\{W=1\}$, for $2 \leq w \leq M$, $(\tilde{\mathbf{X}}(w), \mathbf{Y})$ are n i.i.d. copies of the pair of generic random variables (X', Y'), where X' and Y' have the same marginal distributions as X and Y, respectively.

- Since a DMC is memoryless, X' and Y' are independent because $\tilde{\mathbf{X}}(1)$ and $\tilde{\mathbf{X}}(w)$ are independent and the generation of $\mathbf{Y}$ depends only on $\tilde{\mathbf{X}}(1)$. See textbook for a formal proof.

- By Lemma 7.17,

$$\begin{aligned} \Pr\{E_w|W=1\} &= \Pr\{(\tilde{\mathbf{X}}(w), \mathbf{Y}) \in T^n_{[XY]\delta}|W=1\} \\ &\leq 2^{-n(I(X;Y)-\tau)} \end{aligned}$$

where $\tau \to 0$ as $\delta \to 0$.

- $$\frac{1}{n}\log M < I(X;Y) - \frac{\epsilon}{4} \quad \Longleftrightarrow \quad M < 2^{n(I(X;Y)-\frac{\epsilon}{4})}$$

- Therefore,

$$\begin{aligned}\Pr\{Err\} &< \nu + 2^{n(I(X;Y)-\frac{\epsilon}{4})} \cdot 2^{-n(I(X;Y)-\tau)} \\ &= \nu + 2^{-n(\frac{\epsilon}{4}-\tau)}\end{aligned}$$

- ϵ is fixed. Since $\tau \to 0$ as $\delta \to 0$, we can choose δ to be sufficiently small so that

$$\frac{\epsilon}{4} - \tau > 0$$

- Then $2^{-n(\frac{\epsilon}{4}-\tau)} \to 0$ as $n \to \infty$.

- Let $\nu < \frac{\epsilon}{3}$ to obtain

$$\Pr\{Err\} < \frac{\epsilon}{2}$$

for sufficiently large n.

Idea of Analysis

- Let n be large.
- $\Pr\{\tilde{\mathbf{X}}(1)$ jointly typical with $\mathbf{Y}\} \to 1$.
- For $i \neq 1$, $\Pr\{\tilde{\mathbf{X}}(i)$ jointly typical with $\mathbf{Y}\} \approx 2^{-nI(X;Y)}$.
- If $|\mathcal{C}| = M$ grows at a rate $< I(X;Y)$, then

$$\Pr\{\tilde{\mathbf{X}}(i) \text{ jointly typical with } \mathbf{Y} \text{ for some } i \neq 1\ \}$$

  can be made arbitrarily small.
- Then $\Pr\{\hat{W} \neq W\}$ can be made arbitrarily small.

Existence of Deterministic Code

- According to the random coding scheme,

$$\Pr\{Err\} = \sum_{\mathcal{C}} \Pr\{\mathcal{C}\}\Pr\{Err|\mathcal{C}\}$$

- Then there exists at least one codebook $\mathcal{C}^*$ such that

$$P_e = \Pr\{Err|\mathcal{C}^*\} \leq \Pr\{Err\} < \frac{\epsilon}{2}$$

- By construction, this codebook has rate

$$\frac{1}{n}\log M > I(X;Y) - \frac{\epsilon}{2}$$

Code with $\lambda_{max} < \varepsilon$

- We want a code with $\lambda_{max} < \epsilon$, not just $P_e < \epsilon/2$.
- Technique: Discard the worst half of the codewords in $\mathcal{C}^*$.
- Consider
$$\frac{1}{M}\sum_{w=1}^{M}\lambda_w < \frac{\epsilon}{2} \iff \sum_{w=1}^{M}\lambda_w < \left(\frac{M}{2}\right)\epsilon$$
- Observation: the conditional probabilities of error of the better half of the M codewords are $< \epsilon$ (M is even).

- After discarding the worse half of $\mathcal{C}^*$, the rate of the code becomes

$$\begin{aligned}\frac{1}{n}\log\frac{M}{2} &= \frac{1}{n}\log M - \frac{1}{n} \\ &> \left(I(X;Y) - \frac{\epsilon}{2}\right) - \frac{1}{n} \\ &> I(X;Y) - \epsilon\end{aligned}$$

- Here we assume that the decoding function is unchanged, so that deletion of worst half of the codewords does not affect the conditional probabilities of error of the remaining codewords.

7.5 A Discussion

- The channel coding theorem says that an indefinitely long message can be communicated reliably through the channel when the block length $n \to \infty$. This is much stronger than BER $\to 0$.

- The direct part of the channel coding theorem is an existence proof (as opposed to a constructive proof).

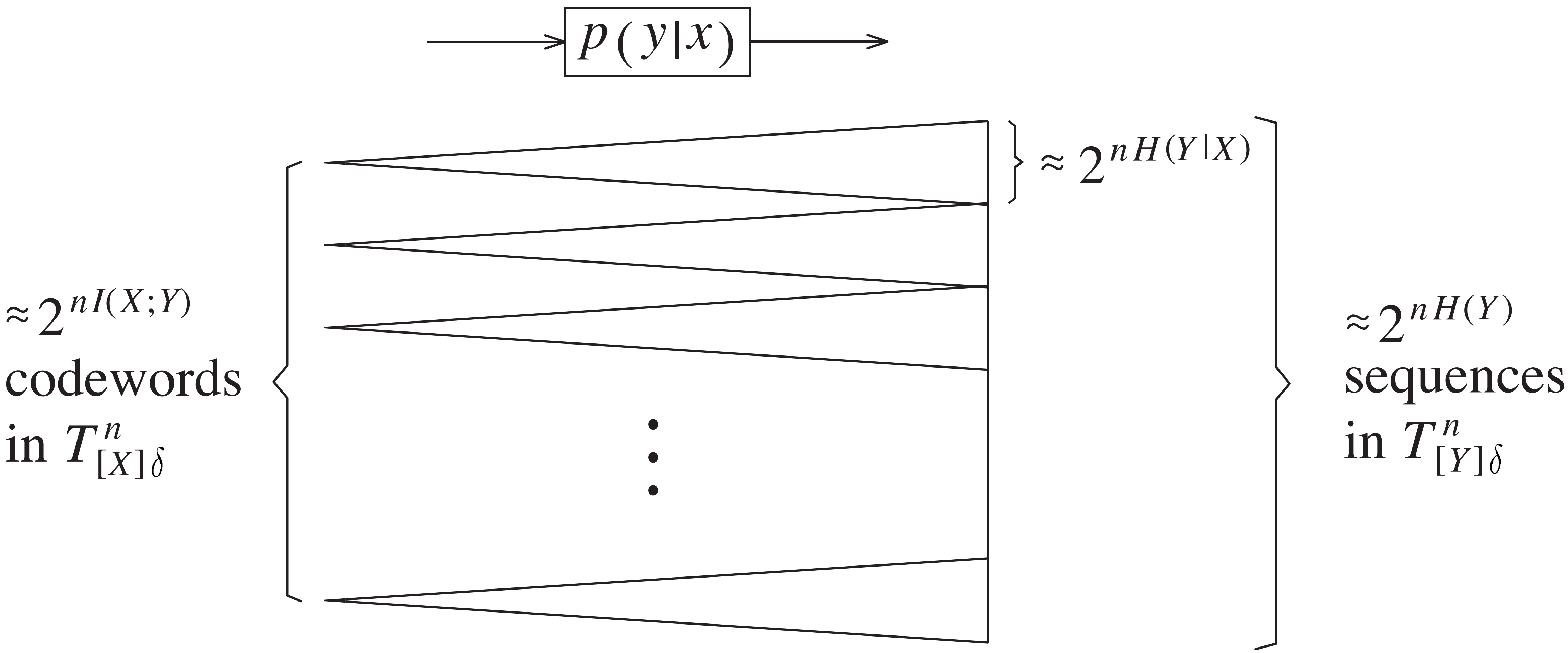
- A randomly constructed code has the following issues:

  - Encoding and decoding are computationally prohibitive.
  - High storage requirements for encoder and decoder.

- Nevertheless, the direct part implies that when n is large, if the codewords are chosen randomly, most likely the code is good (Markov lemma).

- It also gives much insight into what a good code would look like.

- In particular, the repetition code is not a good code because the numbers of '0' and '1's in the codewords are not roughly the same.



The number of codewords cannot exceed about

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)} = 2^{nC}.$$

Channel Coding Theory

- Construction of codes with efficient encoding and decoding algorithms falls in the domain of channel coding theory.

- Performance of a code is measured by how far the rate is away from the channel capacity.

- All channel codes used in practice are linear: efficient encoding and decoding in terms of computation and storage.

- Channel coding has been widely used in home entertainment systems (e.g., audio CD and DVD), computer storage systems (e.g., CD-ROM, hard disk, floppy disk, and magnetic tape), computer communication, wireless communication, and deep space communication.

- The most popular channel codes used in existing systems include the Hamming code, the Reed-Solomon code, the BCH code, and convolutional codes.

- In particular, turbo code, a kind of convolutional code, is "capacity achieving."

7.6 Feedback Capacity

- Feedback is common in practical communication systems for correcting possible errors which occur during transmission.
- Daily example: phone conversation.
- Data communication: the receiver may request a packet to be retransmitted if the *parity check* bits received are incorrect (Automatic Repeat-reQuest).
- The transmitter can at any time decide what to transmit next based on the feedback so far
- Can feedback increase the channel capacity?
- Not for DMC, even with complete feedback!

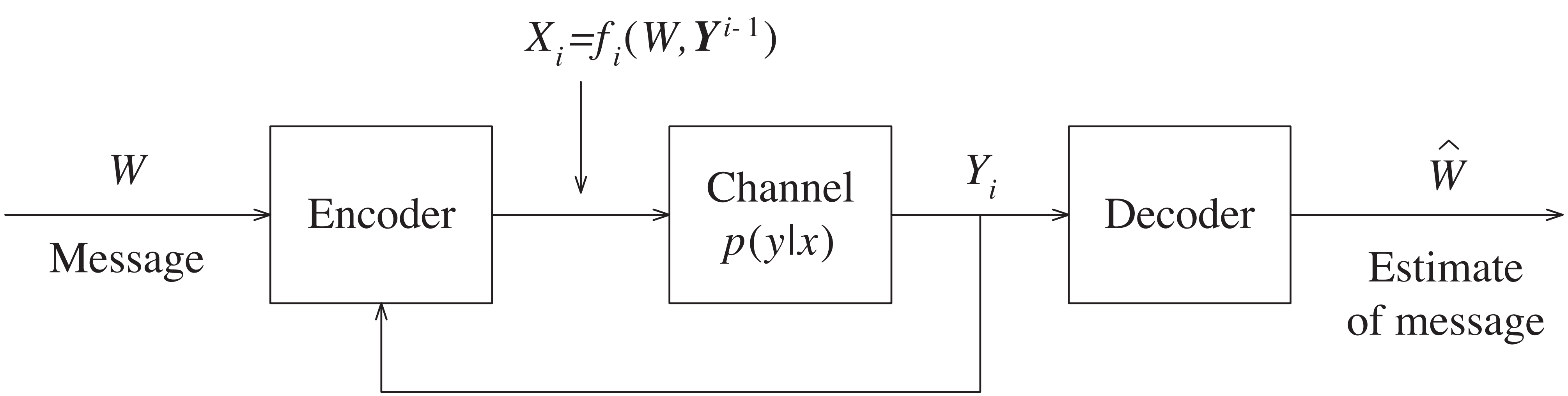
Definition 7.18 An (n, M) code with complete feedback for a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined by encoding functions

$$f_i : \{1, 2, \cdots, M\} \times \mathcal{Y}^{\,i-1} \to \mathcal{X}$$

for $1 \le i \le n$ and a decoding function

$$g : \mathcal{Y}^n \to \{1, 2, \cdots, M\}.$$

Notations: $\mathbf{Y}^i = (Y_1, Y_2, \cdots, Y_i)$, $X_i = f_i(W, \mathbf{Y}^{i-1})$



Definition 7.19 A rate R is achievable with complete feedback for a discrete memoryless channel $p(y|x)$ if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code with complete feedback such that

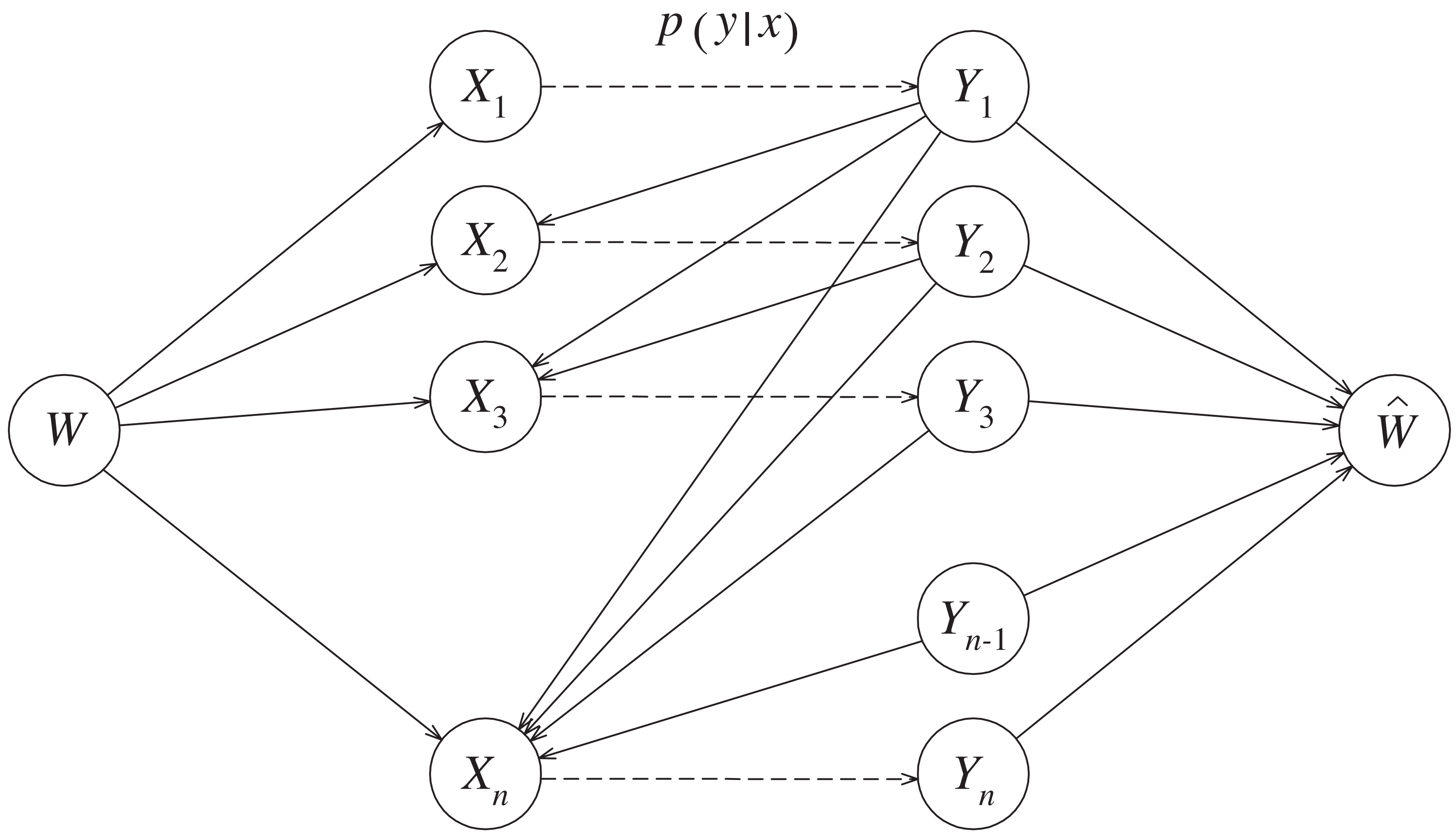
$$\frac{1}{n} \log M > R - \epsilon$$

and

$$\lambda_{max} < \epsilon.$$

Definition 7.20 The feedback capacity, C_{FB}, of a discrete memoryless channel is the supremum of all the rates achievable by codes with complete feedback.

Proposition 7.21 The supremum in the definition of C_{FB} in Definition 7.20 is the maximum.



- The above is the dependency graph for a channel code with feedback, from which we obtain

$$q(w, \mathbf{x}, \mathbf{y}, \hat{w}) = q(w) \left(\prod_{i=1}^{n} q(x_i | w, \mathbf{y}^{i-1}) \right) \left(\prod_{i=1}^{n} p(y_i | x_i) \right) q(\hat{w} | \mathbf{y})$$

for all $(w, \mathbf{x}, \mathbf{y}, \hat{w}) \in \mathcal{W} \times \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{W}$ such that $q(w, \mathbf{y}^{i-1}), q(x_i) > 0$ for $1 \le i \le n$ and $q(\mathbf{y}) > 0$, where $\mathbf{y}^i = (y_1, y_2, \cdots, y_i)$.

Lemma 7.22 For all $1 \leq i \leq n$,

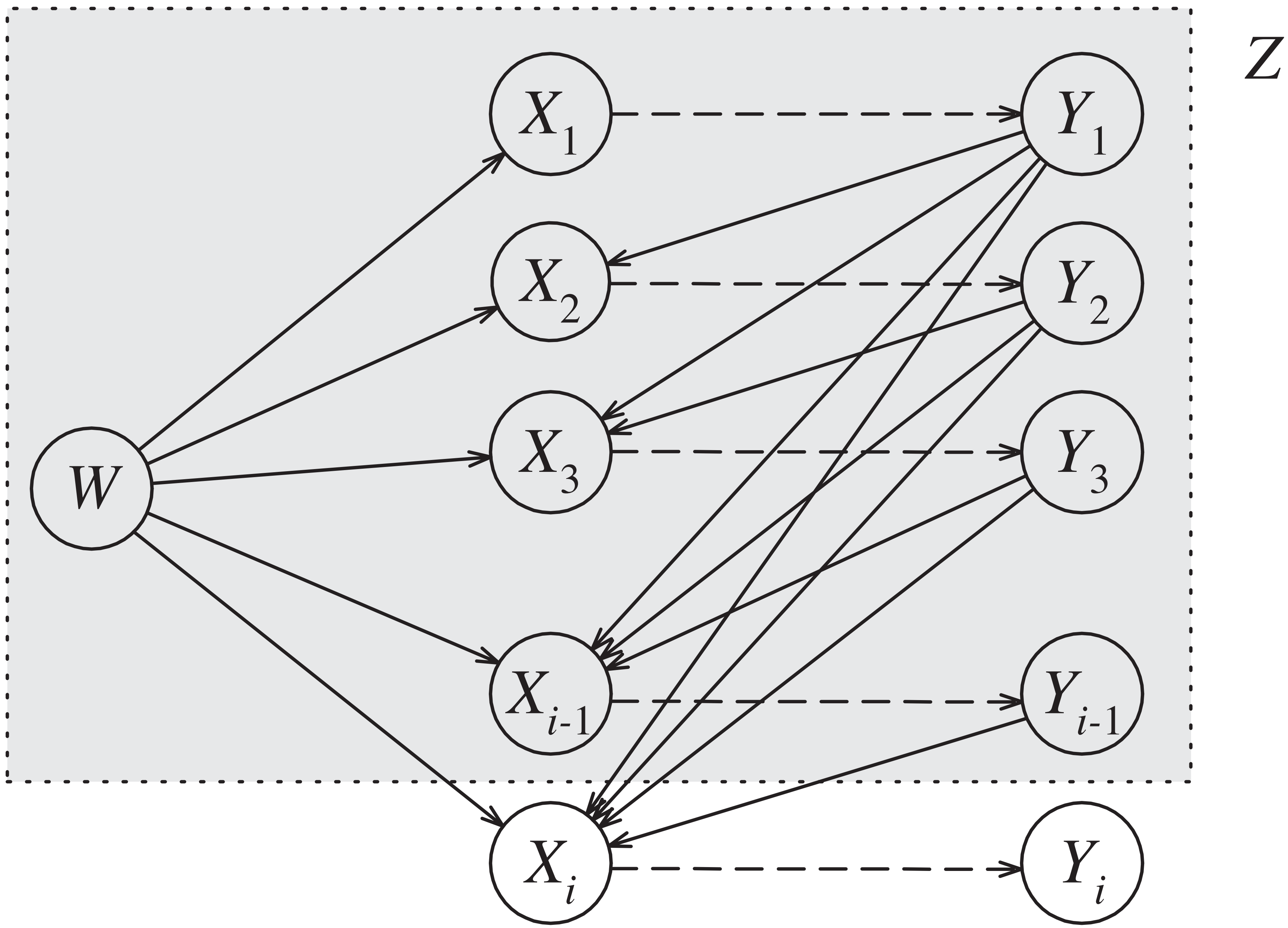
$$(W, \mathbf{Y}^{i-1}) \to X_i \to Y_i$$

forms a Markov chain.

Proof First establish the Markov chain

$$(W, \mathbf{X}^{i-1}, \mathbf{Y}^{i-1}) \to X_i \to Y_i$$

by Proposition 2.9 (see the dependency graph for $W, \mathbf{X}^i$, and $\mathbf{Y}^i$).


Z
W
X_1
X_2
X_3
X_{i-1}
X_i
Y_1
Y_2
Y_3
Y_{i-1}
Y_i

- Consider a code with complete feedback.
- Consider
$$\log M = H(W) = I(W;\mathbf{Y}) + H(W|\mathbf{Y}).$$
- First,

$$\begin{aligned}
I(W;\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|W) \\
&= H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|\mathbf{Y}^{i-1}, W) \\
&\overset{a)}{=} H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|\mathbf{Y}^{i-1}, W, X_i) \\
&\overset{b)}{=} H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&= \sum_{i=1}^{n} I(X_i;Y_i) \\
&\leq nC,
\end{aligned}$$

- Second,
$$H(W|\mathbf{Y}) = H(W|\mathbf{Y}, \hat{W}) \le H(W|\hat{W})$$
- Then upper bound $H(W|\hat{W})$ by Fano's inequality.
- Filling in the ϵ's and δ's, we conclude that
$$R \le C$$

Remark

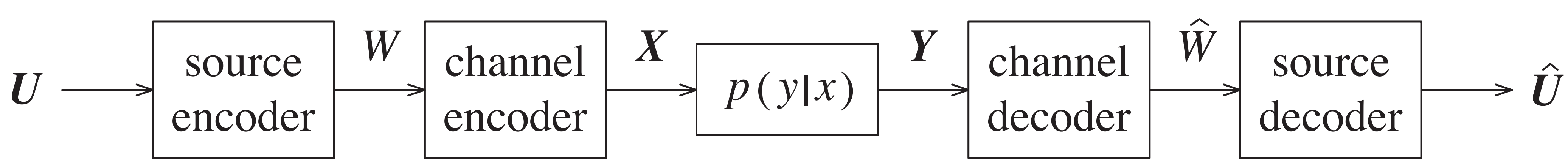
1. Although feedback does not increase the capacity of a DMC, the availability of feedback often makes coding much simpler. See Example 7.23.
2. In general, if the channel has memory, feedback can increase the capacity.

7.7 Separation of Source and Channel Coding

- Consider transmitting an information source with entropy rate H reliably through a DMC with capacity C.
- If $H < C$, this can be achieved by separating source and channel coding without using feedback.
- Specifically, choose R_s and R_c such that

$$H < R_s < R_c < C$$

- It can be shown that even with complete feedback, reliable communication is impossible if $H > C$.



The separation theorem for source and channel coding has the following engineering implications:

- asymptotic optimality can be achieved by separating source coding and channel coding
- the source code and the channel code can be designed separately without losing asymptotic optimality
- only need to change the source code for different information sources
- only need to change the channel code for different channels

Remark For finite block length, the probability of error generally can be reduced by using joint source-channel coding.