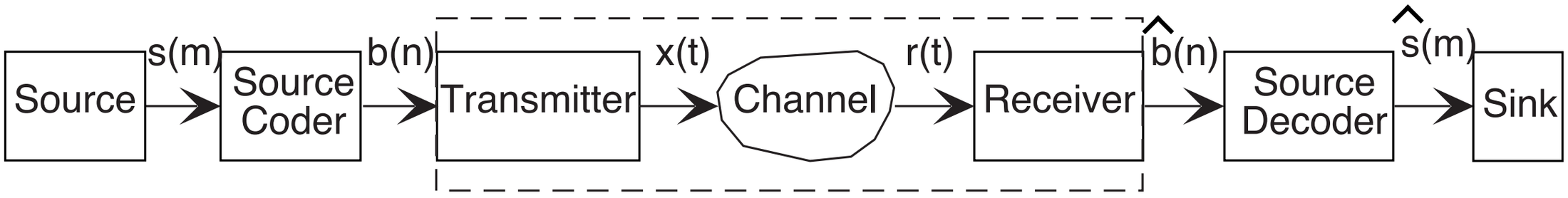
Fundamentals of Electrical Engineering

Representing digital sources with bits

- Revised digital channel model
- Source Coding Theorem
- Source compression

RICE

Digital Communication Model



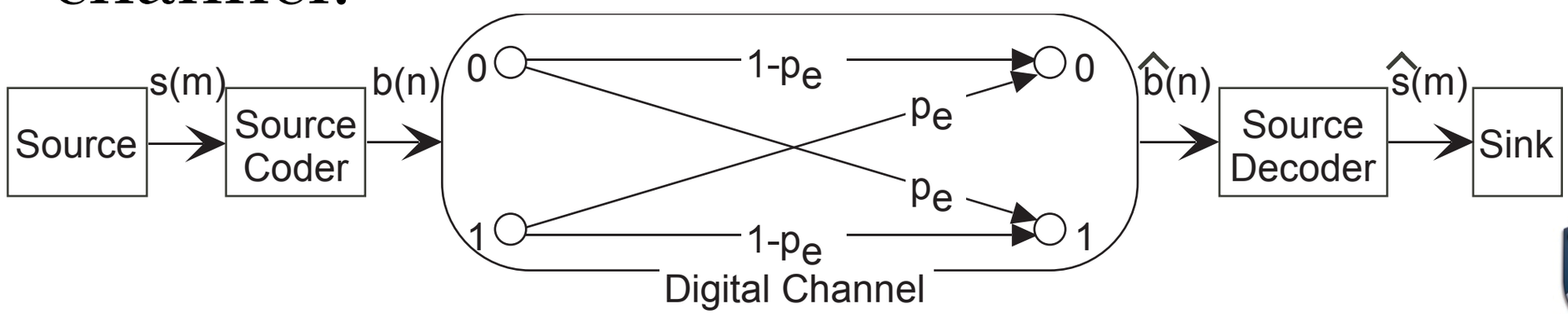
Information Theory

- Source emits *letters* in a probabilistic fashion

$$s(m) \in \{a_1, \ldots, a_K\} = A$$

$$0 \leq \Pr[a_k] \leq 1,\ k = 1, \ldots, K \qquad \sum_{k=1}^{K} \Pr[a_k] = 1$$

- Alphabet A known by source and sink
- Issues
  - ✳ How should letters be represented by bits?
  - ✳ Can we mitigate the errors introduced by the channel?



RICE

Source Coding Theorem

Define source entropy $H(A)$

$$H(A) \equiv -\sum_k \Pr[a_k] \log_2 \Pr[a_k] \text{ bits}$$

Entropy maximized when letters are
equally likely $(\Pr[a_k] = 1/K \implies H(A) = \log_2 K)$

Entropy minimized when one letter occurs
with probability one $(\Pr[a_k] = 1 \implies H(A) = 0)$

RICE

Entropy Calculation

Example

$$\Pr[a_1] = \frac{1}{2} \quad \Pr[a_2] = \frac{1}{4} \quad \Pr[a_3] = \frac{1}{8} \quad \Pr[a_4] = \frac{1}{8}$$

$$H(A) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}\right)$$

$$= -\left(\frac{1}{2}(-1) + \frac{1}{4}(-2) + \frac{1}{8}(-3) + \frac{1}{8}(-3)\right)$$

$$= 1.75 \text{ bits}$$

RICE

Source Coding Theorem

$B(a_k) =$ number of bits assigned to a_k

$\bar{B}(A) = \sum_{k=1}^{K} \Pr[a_k] B(a_k)$ average number of bits to represent A

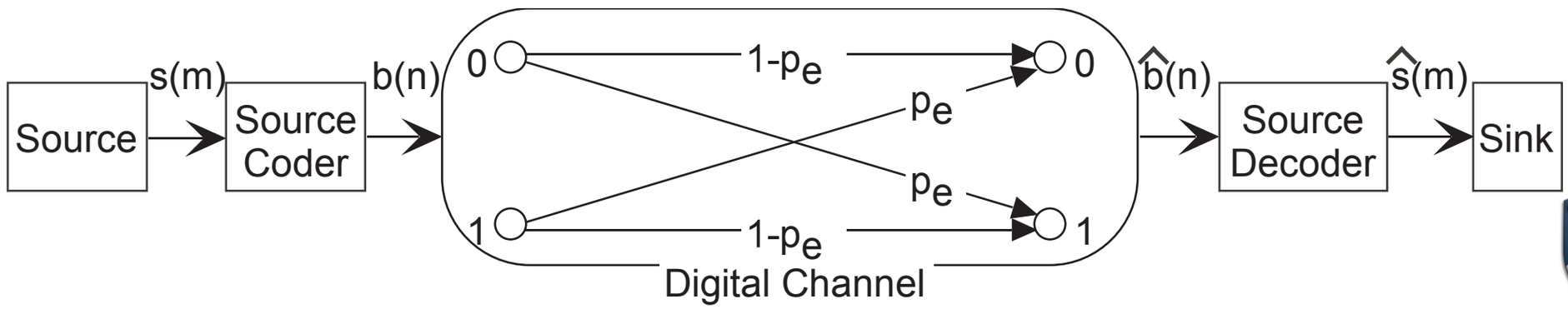
Shannon (1948)

There exists a source coder and decoder that can represent and retrieve the source's alphabet if

$$H(A) \leq \bar{B}(A) < H(A) + 1$$ Lossless compression

Furthermore, *no* error-free source coder/decoder exists if

$$\bar{B}(A) < H(A)$$ Lossy compression



RICE

Compression Schemes

Lossless Compression:
LZW (Lempel-Ziv-Welch), zip, Huffman, simple codes

Lossy Compression:
JPEG, MPEG, MP3,…

Compression Example

$$\Pr[a_1] = \frac{1}{2} \quad \Pr[a_2] = \frac{1}{4} \quad \Pr[a_3] = \frac{1}{8} \quad \Pr[a_4] = \frac{1}{8}$$

$$H(A) = 1.75 \text{ bits} \qquad 1.75 \leq \bar{B}(A) < 2.75 \text{ bits}$$

Simple binary code: $B(a_k) = \lceil \log_2 K \rceil = \bar{B}(A)$

$a_1 \leftrightarrow 00 \quad a_2 \leftrightarrow 01 \quad a_3 \leftrightarrow 10 \quad a_4 \leftrightarrow 11 \quad \bar{B}(A) = 2.0 \text{ bits}$

Better binary code: *Variable length code*

$a_1 \leftrightarrow 0 \quad a_2 \leftrightarrow 10 \quad a_3 \leftrightarrow 110 \quad a_4 \leftrightarrow 111 \quad \bar{B}(A) = 1.75 \text{ bits}$

But there is an issue…

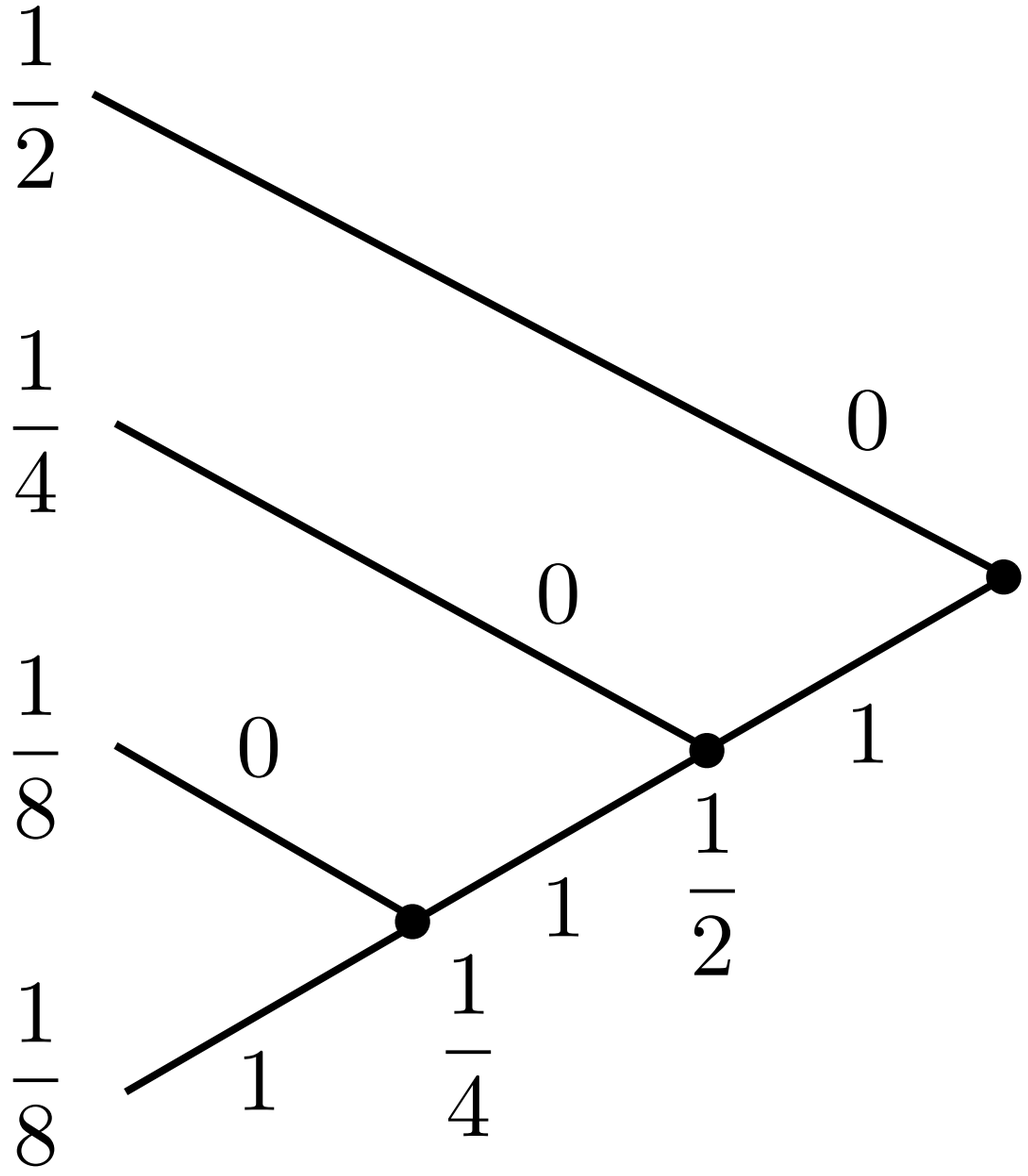
$a_1 a_1 a_3 a_1 a_2 a_1 a_4 \leftrightarrow 001100100111$

RICE

Huffman Codes

Build a coding tree according to specific rules

Letter	Probability	Source Code
a_1	$\frac{1}{2}$	0
a_2	$\frac{1}{4}$	10
a_3	$\frac{1}{8}$	110
a_4	$\frac{1}{8}$	111



Huffman Code Properties

- *Always* yields a prefix code
- May not produce a code "at the entropy limit" but…
- No other lossless source code can have a shorter average code length, which means it is optimal!
- Very sensitive to channel-induced errors

$$a_1 \leftrightarrow 0 \quad a_2 \leftrightarrow 10 \quad a_3 \leftrightarrow 110 \quad a_4 \leftrightarrow 111$$

$$a_1 a_1 a_3 a_1 a_2 a_1 a_4 \leftrightarrow 001100100111$$

RICE

Morse Had the Right Idea

	%	**Morse Code**	**Huffman Code**
A	6.22	.-	1011
B	1.32	-...	010100
C	3.11	-.-.	10101
D	2.97	-..	01011
E	10.53	.	001
F	1.68	..-.	110001
G	1.65	–.	110000
H	3.63	....	11001
I	6.14	..	1001
J	0.06	.—	01010111011
K	0.31	-.-	01010110
L	3.07	.-..	10100
M	2.48	–	00011
N	5.73	-.	0100
O	6.06	—	1000
P	1.87	.–.	00000
Q	0.10	–.-	0101011100
R	5.87	.-.	0111
S	5.81	...	0110
T	7.68	-	1101
U	2.27	..-	00010
V	0.70	...-	0101010
W	1.13	.–	000011
X	0.25	-..-	010101111
Y	1.07	-.–	000010
Z	0.06	–..	0101011101011

$$H(A) = 4.14 \text{ bits}$$

Morse Code: $\bar{B}(A) = 5.56$ bits

Huffman Code: $\bar{B}(A) = 4.35$ bits

RICE

Source Coding (Compression)

- Most data sources can be compressed (lossless) with fewer bits than a simple code, thereby lowering the source datarate
- Lossy compression can *always* be used, but want to control the nature of the errors
- Decompression is sensitive to channel errors

RICE