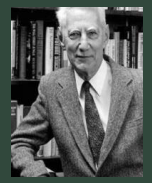
Практичні прийоми

Оцінювання ефективності коду

Від умови оптимальності — до порівняння Шеннона–Фано та Гаффмана.


Шеннон–Фано · Гаффман
Статистичне кодування

p_i	Код
0,500	0
0,250	10
0,125	110
0,125	111

6 прикладів / 2 алгоритми

План лекції

ДВА РОЗДІЛИ

01

Умова оптимальності коду. Теорема Фано.

Нижня межа середньої довжини та умови її досягнення.

02

Приклади оцінювання ефективності коду.

Надмірність → нерівномірний код → Шеннон–Фано → Гаффман.

Умова оптимальності коду

ТЕОРЕМА ФАНО • ФОРМУЛА (3.2)

Якщо дано джерело з обсягом алфавіту N та ентропією H(X), то, застосовуючи код з основою m, можливо забезпечити середнє число двійкових символів на літеру алфавіту джерела у таких межах:

$$\frac{H(X)}{\log_2 N} \leq n_{сер} < \frac{H(X)}{\log_2 N} + 1$$

СЕРЕДНЯ ДОВЖИНА КОДОВОГО СЛОВА

$n_{сер}$ — середнє число кодових символів на один символ джерела.

Двійковий випадок і оптимальний код

ЗОКРЕМА, ПРИ N = 2

$$H(X) \le n_{сер} < H(X) + 1$$

Середня кількість двійкових елементів коду принаймні дорівнює ентропії джерела; перевищення менше одиниці.

ЯКЩО ЙМОВІРНІСТЬ ЗАДОВОЛЬНЯЄ УМОВІ

$$p(x_k) = N^{-n(x_k)}$$

ТО ДОСЯГАЄТЬСЯ НИЖНЯ МЕЖА, можна побудувати оптимальний код

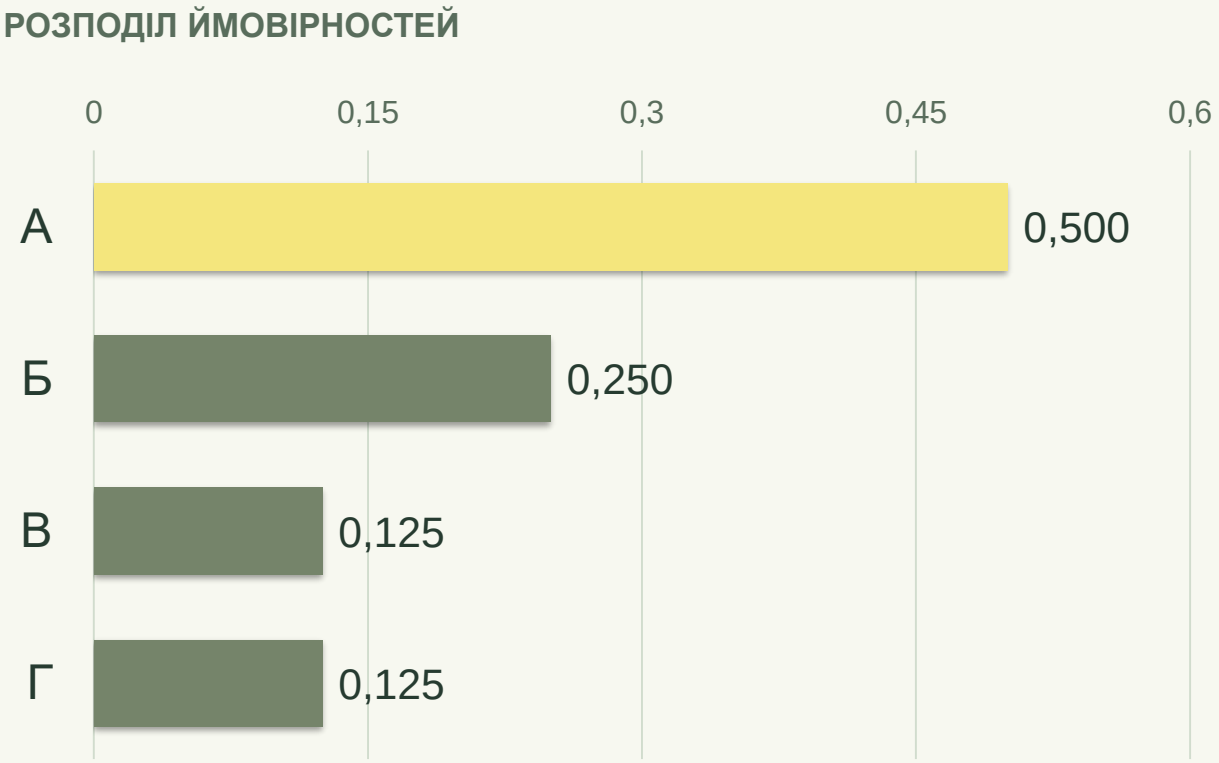
$$n_{сер} = \frac{H(X)}{\log_2 N}$$

Надмірність: алфавіт із чотирьох символів

ПРИКЛАД 1 • УМОВА

Джерело має алфавіт із символів А, Б, В, Г. Визначити надмірність повідомлень, складених із цього алфавіту.

Символ	Ймовірність
А	0,500
Б	0,250
В	0,125
Г	0,125



Приклад 1: ентропія джерела

РОЗВ'ЯЗОК • КРОК 1

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$= -0{,}5 \log_2 0{,}5 - 0{,}25 \log_2 0{,}25$$

$$- 0{,}125 \log_2 0{,}125 - 0{,}125 \log_2 0{,}125$$

$$= 0{,}5 + 0{,}5 + 0{,}375 + 0{,}375 = 1{,}75$$

Ентропія H(X) = 1,75 біт/символ.

Приклад 1: максимальна ентропія і надмірність

РОЗВ'ЯЗОК • КРОК 2

$$H_{max} = \log_2 4 = 2$$

$$D = \frac{H_{max}(X) - H(X)}{H_{max}(X)} = \frac{2 - 1{,}75}{2} = 0{,}125$$

ЕНТРОПІЯ

1,75

біт/символ

МАКСИМУМ

2

біти/символ

НАДМІРНІСТЬ

0,125

D

Відносні частоти букв українського алфавіту, таблиця 1

ПРИКЛАД 2

Визначити ентропію та середню надмірність тексту українською мовою.

Літера	Частота	Літера	Частота	Літера	Частота	Літера	Частота
–	0,138	Е	0,042	Я	0,019	Ю	0,008
О	0,086	С	0,037	Ь	0,016	Ж	0,007
Н	0,068	К	0,033	Б	0,013	Ш	0,005
А	0,064	М	0,029	Г	0,013	Є	0,005
И	0,055	У	0,027	Ч	0,011	Щ	0,004
В	0,046	Д	0,027	Х	0,011	Ф	0,003
Т	0,045	Л	0,027	Ї	0,010	Ґ	< 0,001
І	0,044	П	0,025	Ц	0,010		
Р	0,043	З	0,020	Й	0,009		

Приклад 2: ентропія та надмірність тексту

РОЗВ'ЯЗОК

$$H(X) = -\sum_{i=1}^{34} p(x_i) \log_2 p(x_i)$$

$$= -0{,}138 \log_2 0{,}138 - 0{,}086 \log_2 0{,}086$$

$$- \ldots - 0{,}003 \log_2 0{,}003$$

ЕНТРОПІЯ H(X)

4,5356

біт/символ

HMAX = LOG₂ 34

5,0874

біт/символ

$$D = \frac{5{,}0874 - 4{,}5356}{5{,}0874} = 0{,}1085$$

Усунення надмірності досягається використанням нерівномірних кодів. Частіші літери кодуються коротшими послідовностями, рідкісні — довшими.

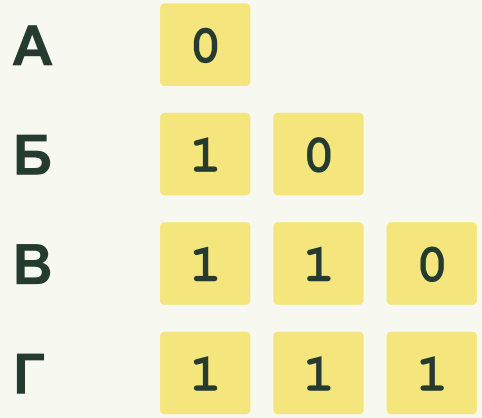
Нерівномірний код для чотирьох символів

ПРИКЛАД 3

Для джерела з прикладу 1 використовується код А → 0, Б → 10, В → 110, Г → 111. Визначити середнє число біт на символ і порівняти його з ентропією.

Символ	p_i	Код	n_i
А	0,500	0	1
Б	0,250	10	2
В	0,125	110	3
Г	0,125	111	3

КОДОВІ СЛОВА



До прикладу 3: середня довжина дорівнює ентропії

РОЗРАХУНОК $n_{сер}$

$$n_{сер} = \sum_{i=1}^{4} p_i \cdot n_i$$

$$= 0{,}5 \cdot 1 + 0{,}25 \cdot 2 + 0{,}125 \cdot 3 + 0{,}125 \cdot 3$$

$$H(X) = n_{сер} = 1{,}75$$

Для цього джерела нерівномірний код є більш економним, ніж рівномірний, а також є оптимальним.

Приклад 3: коефіцієнт стиснення

РЕЗУЛЬТАТ

$$\mu = \frac{\log_2 N}{n_{сер}} = \frac{\log_2 4}{1{,}75} = \frac{2}{1{,}75} = 1{,}1429$$

Швидкість передачі по каналу зв’язку при використанні нерівномірного кодування може бути в 1,1429 раза більшою, ніж при рівномірному кодуванні.

КОЕФІЦІЄНТ μ

1,1429

Результат прикладу

Кодування букв українського алфавіту

ПРИКЛАД 4 • УМОВА

Символи українського алфавіту закодовані відповідно до таблиці 1. Визначити середню кількість двійкових символів, що припадають на один символ джерела.

Порівняти результат з ентропією джерела, визначеною в прикладі 2.

ОБСЯГ АЛФАВІТУ

34
символи

ЕНТРОПІЯ H(X)

4,5356
біт/символ

Таблиця 4.1: нерівномірне кодування • 1/2

ПРИКЛАД 4

Літера	p_i	Код	n_i
–	0,138	101	3
О	0,086	1111	4
Н	0,068	1001	4
А	0,064	1000	4
И	0,055	0110	4
В	0,046	0011	4
Т	0,045	0001	4
І	0,044	0000	4
Р	0,043	11101	5

Літера	p_i	Код	n_i
Е	0,042	11100	5
С	0,037	11001	5
К	0,033	11000	5
М	0,029	01110	5
У	0,027	01001	5
Д	0,027	01000	5
Л	0,027	01011	5
П	0,025	00101	5

Таблиця 4.1: нерівномірне кодування • 2/2 (продовження таблиці)

ПРИКЛАД 4 •

Літера	p_i	Код	n_i
З	0,020	110111	6
Я	0,019	110100	6
Ь	0,016	011110	6
Б	0,013	010100	6
Г	0,013	001001	6
Ч	0,011	001000	6
Х	0,011	1101101	7
Ї	0,010	1101100	7
Ц	0,010	1101011	7

Літера	p_i	Код	n_i
Й	0,009	0111111	7
Ю	0,008	0111110	7
Ж	0,007	0101011	7
Ш	0,005	11010101	8
Є	0,005	11010100	8
Щ	0,004	01010101	8
Ф	0,003	010101001	9
Ґ	0,0001	010101000	9

Для Ґ у таблиці 4.1 наведено p = 0,0001 (в таблиці 1 — «< 0,001»).

Приклад 4: середня кількість біт

РОЗВ'ЯЗОК

$$n_{сер} = \sum_{i=1}^{34} p_i \cdot n_i$$

РОЗГОРНУТИЙ ЗАПИС І РЕЗУЛЬТАТ

$$= 0{,}138 \cdot 3 + 0{,}086 \cdot 4 + \ldots$$
$$+ 0{,}002 \cdot 9 + 0{,}000 \cdot 9 = 4{,}5720$$

$$4{,}5356 < 4{,}572 < 5{,}0874$$

Середня довжина більша за H(X), але менша за максимальну ентропію Hmax.

Приклад 4: стиснення українського тексту

РЕЗУЛЬТАТ

$$\mu = \frac{\log_2 N}{n_{сер}} = \frac{\log_2 34}{4{,}572} = \frac{5{,}0875}{4{,}572} = 1{,}1128$$

Для зазначеного джерела нерівномірний код є більш економним, ніж рівномірний. Швидкість передачі при такому кодуванні може бути в 1,1128 раза більшою.

КОЕФІЦІЄНТ μ

1,1128

Шеннон–Фано для дев’яти повідомлень

ПРИКЛАД 5 • УМОВА

Побудувати код Шеннона–Фано для дев’яти повідомлень. Визначити максимальну ентропію Hmax, ентропію джерела H(X), середнє число біт на символ $n_{сер}$ та коефіцієнт стиснення μ.

Символ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
$p(x_i)$	0,35	0,15	0,13	0,09	0,09	0,08	0,05	0,04	0,02

ПОСЛІДОВНІСТЬ ОЦІНЮВАННЯ

01 Побудова коду

02 H(X) та $n_{сер}$

03 Hmax та μ

Таблиця 5.1: побудова коду Шеннона–Фано

ПРИКЛАД 5 • РОЗБИТТЯ НА ГРУПИ ТА КОДОВІ СЛОВА

x_i	$p(x_i)$	Біт 1	Біт 2	Біт 3	Біт 4	Біт 5	Код	n_i	$n_i p_i$
x_1	0,35	1	1	—	—	—	11	2	0,7
x_2	0,15	1	0	—	—	—	10	2	0,3
x_3	0,13	0	0	1	—	—	001	3	0,39
x_4	0,09	0	0	0	—	—	000	3	0,27
x_5	0,09	0	1	1	1	—	0111	4	0,36
x_6	0,08	0	1	1	0	—	0110	4	0,32
x_7	0,05	0	1	0	0	—	0100	4	0,2
x_8	0,04	0	1	0	1	1	01011	5	0,2
x_9	0,02	0	1	0	1	0	01010	5	0,1

«—» означає, що кодове слово завершене. Кожен рядок відтворює розбиття й код з оригіналу.

Приклад 5: ентропія повідомлень

РОЗВ'ЯЗОК • H(X)

$$H(X) = -\sum_{i=1}^{9} p(x_i) \log_2 p(x_i)$$

$$= -(0{,}35 \log_2 0{,}35 + 0{,}15 \log_2 0{,}15 + 0{,}13 \log_2 0{,}13$$

$$+ 0{,}09 \log_2 0{,}09 + 0{,}09 \log_2 0{,}09 + 0{,}08 \log_2 0{,}08$$

$$+ 0{,}05 \log_2 0{,}05 + 0{,}04 \log_2 0{,}04 + 0{,}02 \log_2 0{,}02)$$

$$H(X) = 2{,}75$$

Одиниця вимірювання в лекції — біт/символ повідомлення

Приклад 5: код близький до оптимального

СЕРЕДНЯ ДОВЖИНА, МАКСИМАЛЬНА ЕНТРОПІЯ, СТИСНЕННЯ

$$n_{сер} = \sum_{i=1}^{9} p(x_i)\, n(x_i)$$

$H(X) = 2{,}75$

$$= 0{,}7 + 0{,}3 + 0{,}39 + 0{,}27 + 0{,}36 + 0{,}32 + 0{,}2 + 0{,}2 + 0{,}1 = 2{,}84$$

$$H_{max} = \log_2 N = \log_2 9 = 3{,}1699$$

$$\mu = \frac{\log_2 N}{n_{сер}} = \frac{\log_2 9}{2{,}84} = \frac{3{,}1699}{2{,}84} = 1{,}1162$$

H(X) ≈ nсер: умова Фано виконується, але її ліва межа не досягається.

Порівняння Шеннона–Фано та Гаффмана

ПРИКЛАД 6

Побудувати коди Шеннона–Фано і Гаффмана для восьми елементів повідомлення. Визначити Hmax, H(X), $n_{сер}$ та μ; порівняти ефективність кодів.

Позначення елементів	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$p(x_i)$	0,39	0,16	0,13	0,11	0,09	0,06	0,03	0,03

Літери розташовано за спаданням імовірностей; відповідність $x_1 \ldots x_8$ взято з пояснення до прикладу.

Таблиця 6.1: код Шеннона–Фано

ПРИКЛАД 6 • ПЕРШИЙ АЛГОРИТМ

x_i	$p(x_i)$	Біт 1	Біт 2	Біт 3	Біт 4	Біт 5	Код	n_i	$n_i p_i$
x_1	0,39	1	1	—	—	—	11	2	0,78
x_2	0,16	1	0	—	—	—	10	2	0,32
x_3	0,13	0	1	1	—	—	011	3	0,39
x_4	0,11	0	1	0	—	—	010	3	0,33
x_5	0,09	0	0	0	—	—	000	3	0,27
x_6	0,06	0	0	1	1	—	0011	4	0,24
x_7	0,03	0	0	1	0	1	00101	5	0,15
x_8	0,03	0	0	1	0	0	00100	5	0,15

Останній стовпець — внесок кожного кодового слова у середню довжину.

Приклад 6: ентропія джерела

РОЗВ'ЯЗОК • ЕНТРОПІЯ ОДНАКОВА ДЛЯ ОБОХ АЛГОРИТМІВ

$$H(X) = -\sum_{i=1}^{8} p(x_i) \log_2 p(x_i)$$

$$= -(0{,}39 \log_2 0{,}39 + 0{,}16 \log_2 0{,}16 + 0{,}13 \log_2 0{,}13$$

$$+ 0{,}11 \log_2 0{,}11 + 0{,}09 \log_2 0{,}09 + 0{,}06 \log_2 0{,}06$$

$$+ 0{,}03 \log_2 0{,}03 + 0{,}03 \log_2 0{,}03)$$

$$H(X) = 2{,}5455$$

H(X) = 2,5455 біт/елем. повідомлення

Приклад 6: характеристики Шеннона–Фано

СЕРЕДНЯ ДОВЖИНА ТА КОЕФІЦІЄНТ СТИСНЕННЯ

$$n_{серш} = \sum_{i=1}^{8} p(x_i)\, n(x_i)$$

РОЗГОРТАННЯ ЗА ДАНИМИ ТАБЛИЦІ 6.1

$$= 0{,}78 + 0{,}32 + 0{,}39 + 0{,}33$$
$$+ 0{,}27 + 0{,}24 + 0{,}15 + 0{,}15 = 2{,}63$$

$$H_{max} = \log_2 N = \log_2 8 = 3$$

$$\mu = \frac{\log_2 N}{n_{сер}} = \frac{\log_2 8}{2{,}63} = \frac{3}{2{,}63} = 1{,}1407$$

Таблиця 6.2: побудова коду Гаффмана • 1/2

ПРИКЛАД 6 • ПОСЛІДОВНЕ ОБ’ЄДНАННЯ НАЙМЕНШИХ ІМОВІРНОСТЕЙ

КРОК 1

Вузол	p
x_1	0,39
x_2	0,16
x_3	0,13
x_4	0,11
x_5	0,09
x_6	0,06
x_7	0,03
x_8	0,03

КРОК 2

Вузол	p
x_1	0,39
x_2	0,16
x_3	0,13
x_4	0,11
x_5	0,09
x_6	0,06
x_{78}	0,06

КРОК 3

Вузол	p
x_1	0,39
x_2	0,16
x_3	0,13
x_{678}	0,12
x_4	0,11
x_5	0,09

КРОК 4

Вузол	p
x_1	0,39
x_{45}	0,20
x_2	0,16
x_3	0,13
x_{678}	0,12

Жовтим виділено пару з найменшими ймовірностями для об’єднання на наступному кроці.

Таблиця 6.2: побудова коду Гаффмана • 2/2 продовження таблиці

ПРИКЛАД 6 • ПОСЛІДОВНЕ ОБ’ЄДНАННЯ НАЙМЕНШИХ ІМОВІРНОСТЕЙ

КРОК 5

Вузол	p
x_1	0,39
x_{3678}	0,25
x_{45}	0,20
x_2	0,16

КРОК 6

Вузол	p
x_1	0,39
x_{245}	0,36
x_{3678}	0,25

КРОК 7

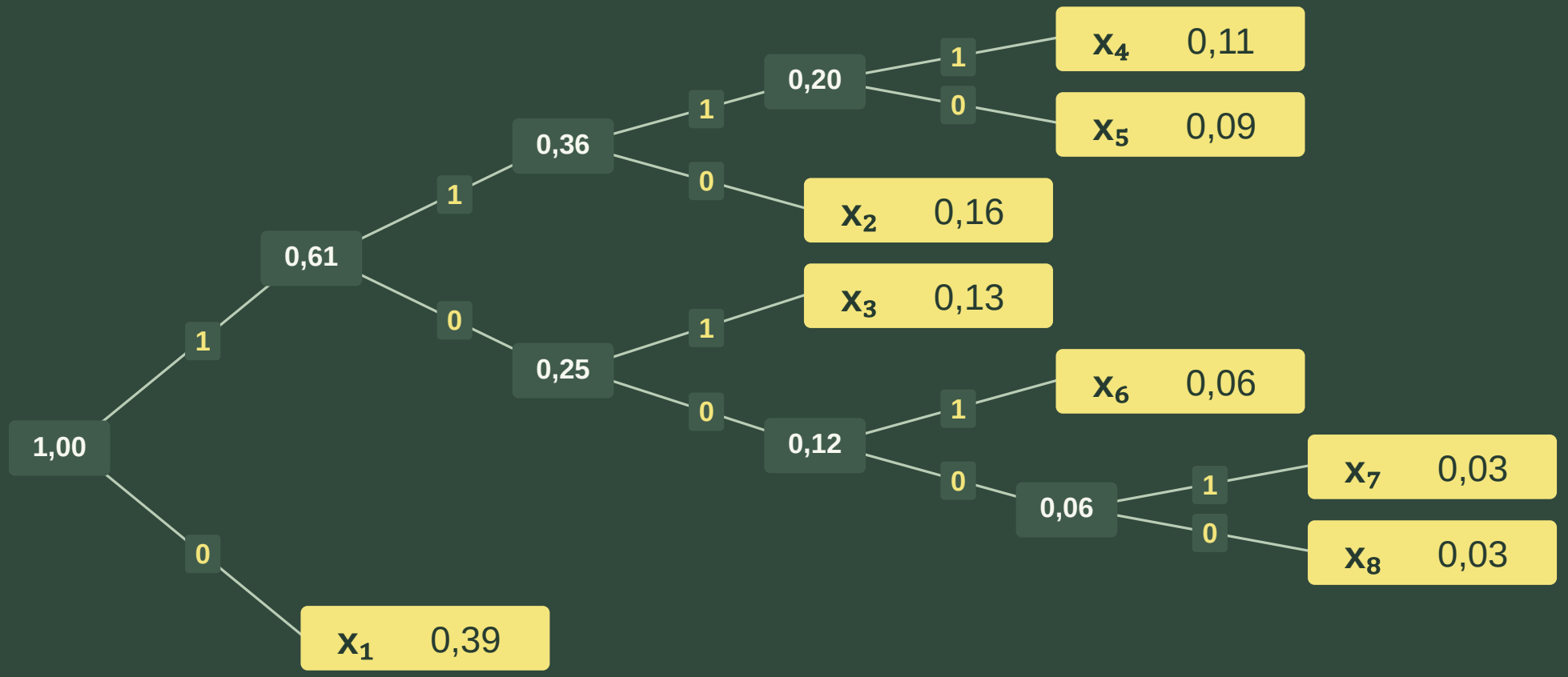
Вузол	p
$x_{2453678}$	0,61
x_1	0,39

$$0{,}61 + 0{,}39 = 1{,}00$$

Жовтим виділено пару з найменшими ймовірностями для об’єднання на наступному кроці.

Кодове дерево Гаффмана

ПРИКЛАД 6 • РИСУНОК 6.1



У вузлах — імовірності, на гілках — біти 0/1. Код читаємо від кореня до листка.

Кодові слова, отримані з дерева

ПРИКЛАД 6 • КОД ГАФФМАНА

Літера	Символ	$p(x_i)$	Код	n_i
Ґ	x_1	0,39	0	1
А	x_2	0,16	110	3
Д	x_3	0,13	101	3
Б	x_4	0,11	1111	4
В	x_5	0,09	1110	4
Є	x_6	0,06	1001	4
Е	x_7	0,03	10001	5
Г	x_8	0,03	10000	5

Приклад 6: характеристики коду Гаффмана

СЕРЕДНЯ ДОВЖИНА ТА КОЕФІЦІЄНТ СТИСНЕННЯ

$$n_{серх} = \sum_{i=1}^{8} p(x_i)\, n(x_i)$$

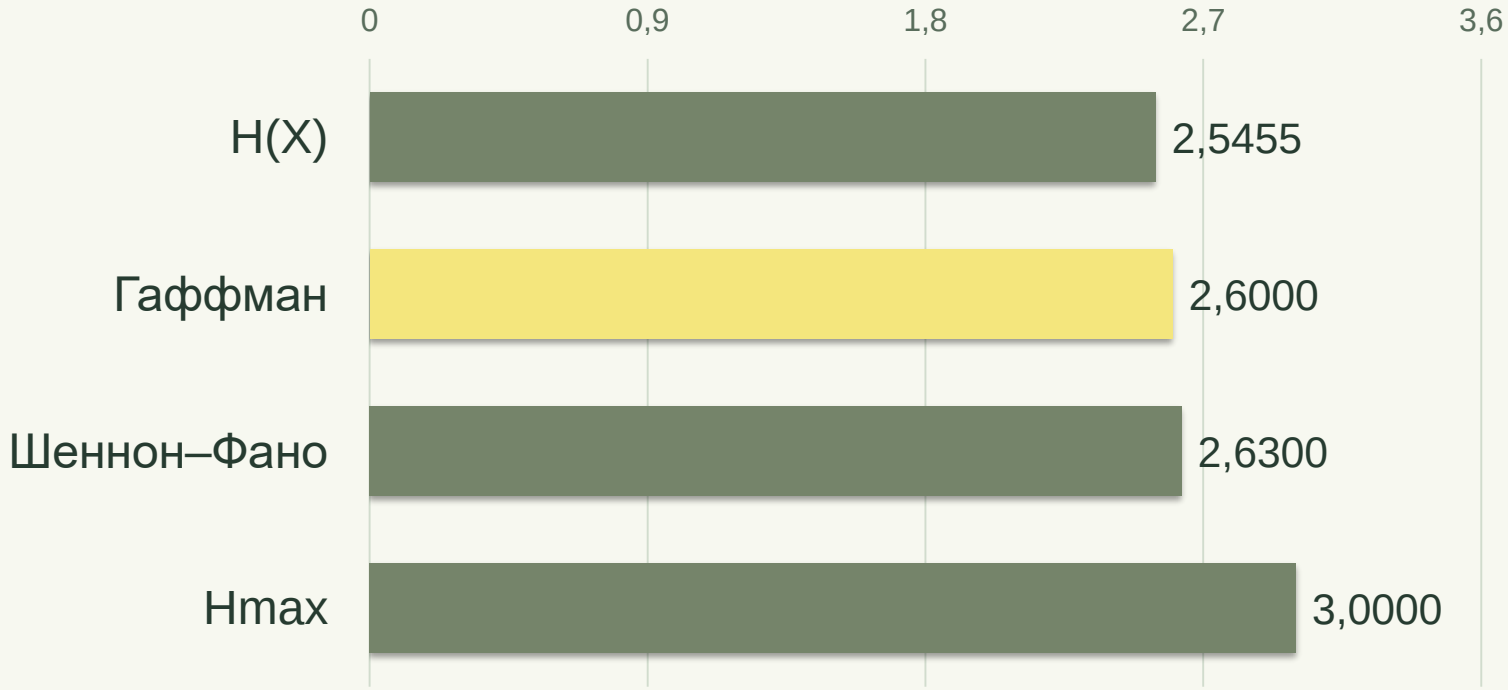
$$= 0{,}39 + 0{,}48 + 0{,}39 + 0{,}44$$
$$+ 0{,}36 + 0{,}24 + 0{,}15 + 0{,}15 = 2{,}60$$

$$\mu = \frac{\log_2 N}{n_{сер}} = \frac{\log_2 8}{2{,}60} = \frac{3}{2{,}60} = 1{,}1538$$

Середня кількість біт на символ для коду Гаффмана: 2,60 біт/символ.

Порівняння ефективності двох кодів

ПІДСУМОК ПРИКЛАДУ 6



μ · ГАФФМАН

1,1538

Коефіцієнт стиснення

μ · ШЕННОН–ФАНО

1,1407

Коефіцієнт стиснення

$$H(X) \le n_{серх} \le n_{серш} \le H_{max}$$

Для наведеного джерела: 2,5455 ≤ 2,60 ≤ 2,63 ≤ 3.

Логіка оцінювання ефективності коду

УЗАГАЛЬНЕННЯ МАТЕРІАЛУ ЛЕКЦІЇ

01	Ентропія джерела	$H(X) = -\sum p(x_i) \log_2 p(x_i)$
02	Середня довжина коду	$n_{сер} = \sum p_i\, n_i$
03	Порівняння з межею	$H(X) \le n_{сер} < H(X) + 1$
04	Коефіцієнт стиснення	$\mu = \frac{\log_2 N}{n_{сер}}$

Надмірність повідомлень: D = (Hmax − H) / Hmax.