2010 jwcc

santo stefano belbo

october 18, 2010

lossless data compression taught right

sergio verdú

princeton university

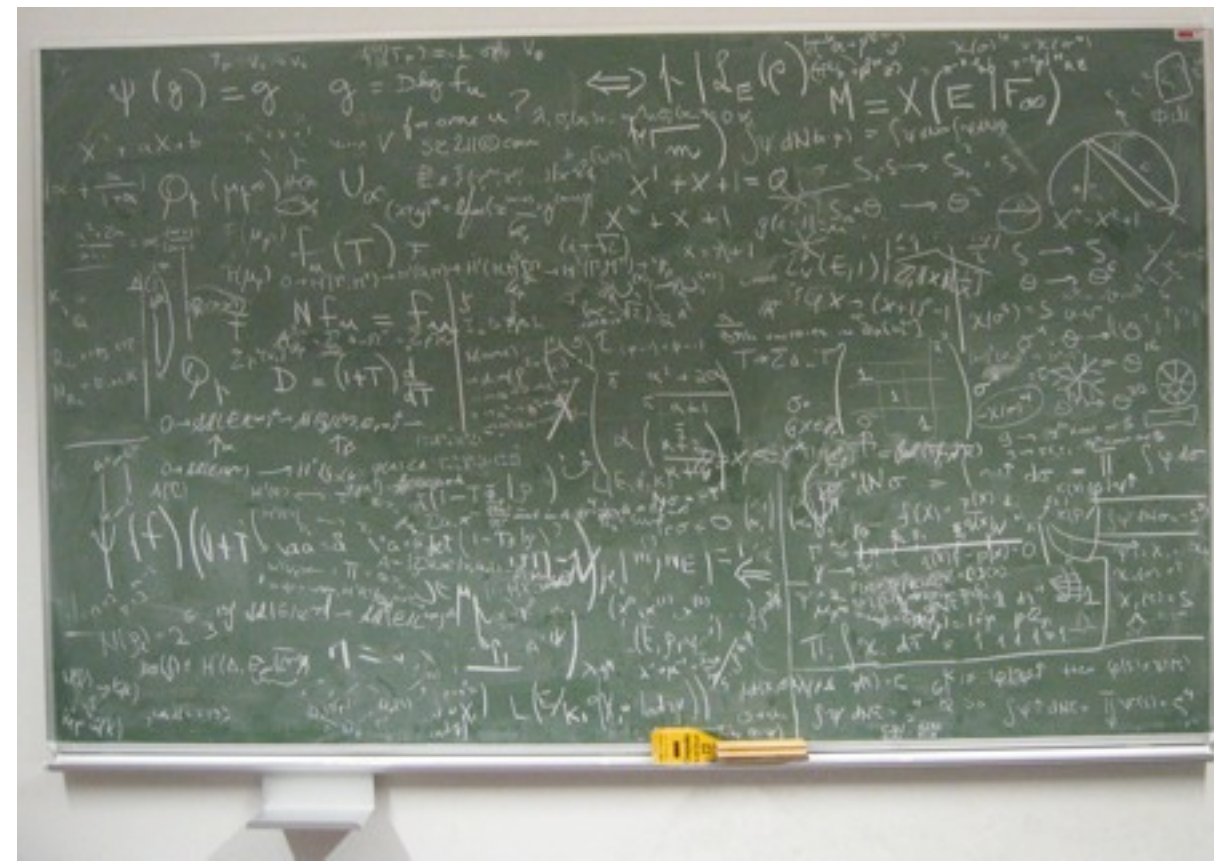
lossless data compression

- **Sources:** Data, Text, Images ...
- **Applications:** Computer files, pdf, software distribution, modems, fax, lossy data compression ("entropy coding")...
- **Goal:** Reduce transmission time / storage space by exploiting "redundancy" to compact data reversibly.

the conventional way: two separate worlds

- Variable-Length Symbol-by-Symbol Lossless
- Fixed-Length Almost-Lossless

data compression



lossless
variable-length

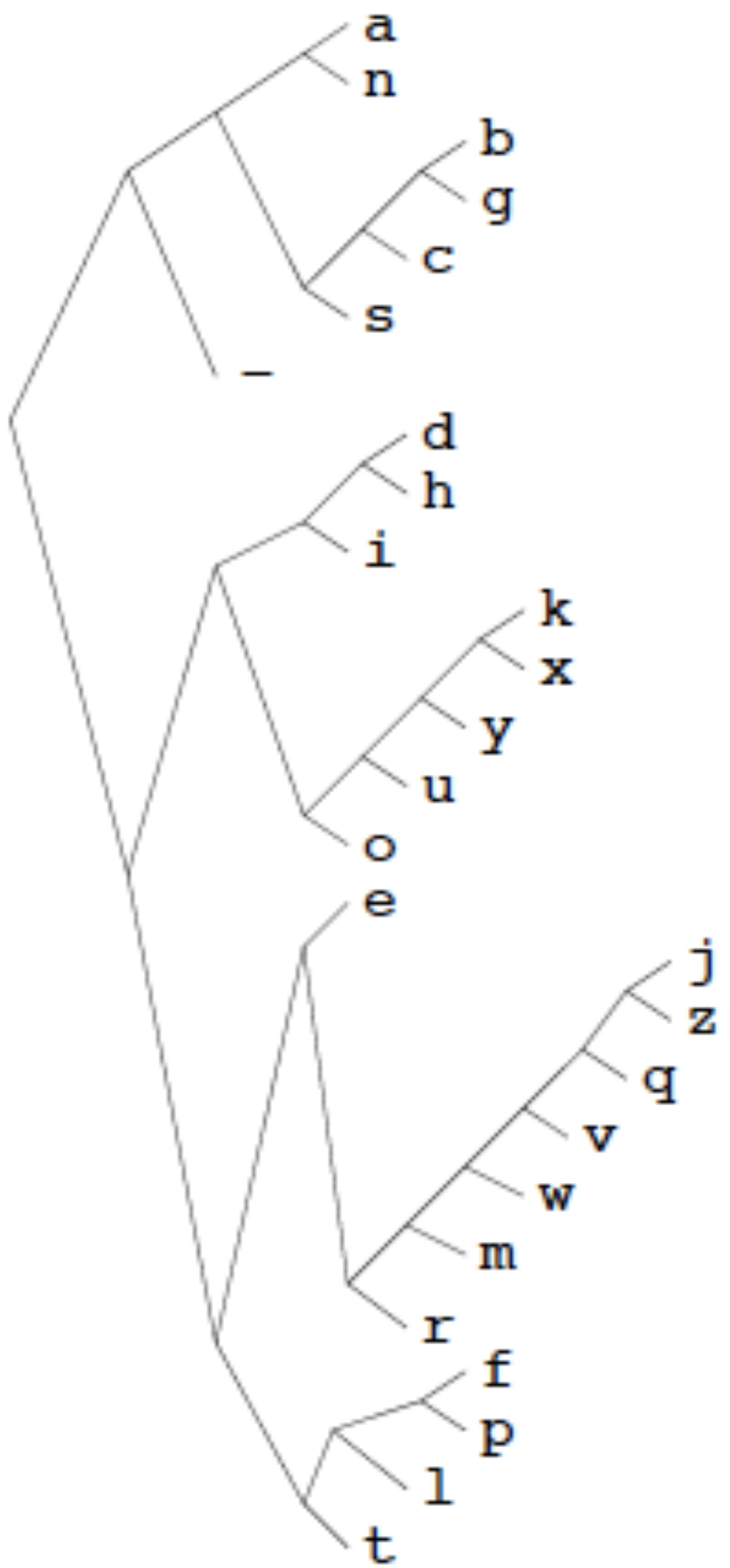
almost lossless
fixed-length

lossless data compression: the conventional way

- Prefix Codes
- "Uniquely Decodable" Codes
- Kraft's Inequality
- Bounds on Average Length
- Huffman Codes

prefix codes

a_i	p_i	$\log_2 \frac{1}{p_i}$	l_i	$c(a_i)$
a	0.0575	4.1	4	0000
b	0.0128	6.3	6	001000
c	0.0263	5.2	5	00101
d	0.0285	5.1	5	10000
e	0.0913	3.5	4	1100
f	0.0173	5.9	6	111000
g	0.0133	6.2	6	001001
h	0.0313	5.0	5	10001
i	0.0599	4.1	4	1001
j	0.0006	10.7	10	1101000000
k	0.0084	6.9	7	1010000
l	0.0335	4.9	5	11101
m	0.0235	5.4	6	110101
n	0.0596	4.1	4	0001
o	0.0689	3.9	4	1011
p	0.0192	5.7	6	111001
q	0.0008	10.3	9	110100001
r	0.0508	4.3	5	11011
s	0.0567	4.1	4	0011
t	0.0706	3.8	4	1111
u	0.0334	4.9	5	10101
v	0.0069	7.2	8	11010001
w	0.0119	6.4	7	1101001
x	0.0073	7.1	7	1010001
y	0.0164	5.9	6	101001
z	0.0007	10.4	10	1101000001
–	0.1928	2.4	2	01



uniquely decodable codes

$$c({\color{red}\spadesuit}) = 0$$
$$c({\color{yellow}\spadesuit}) = 10$$
$$c({\color{green}\spadesuit}) = 11$$

$$c({\color{red}\clubsuit}) = 0$$
$$c({\color{yellow}\clubsuit}) = 1$$
$$c({\color{green}\clubsuit}) = 10$$

Kraft's inequality

1. A uniquely decodable binary code with codewords of lengths $\{\ell_a, a \in \mathcal{A}\}$ must satisfy:
$$\sum_{a \in \mathcal{A}} 2^{-\ell_a} \leq 1 \qquad (1)$$

2. If $\{\ell_a, a \in \mathcal{A}\}$ satisfy (1), then there exists a prefix binary code with those lengths.

entropy and average length: converse

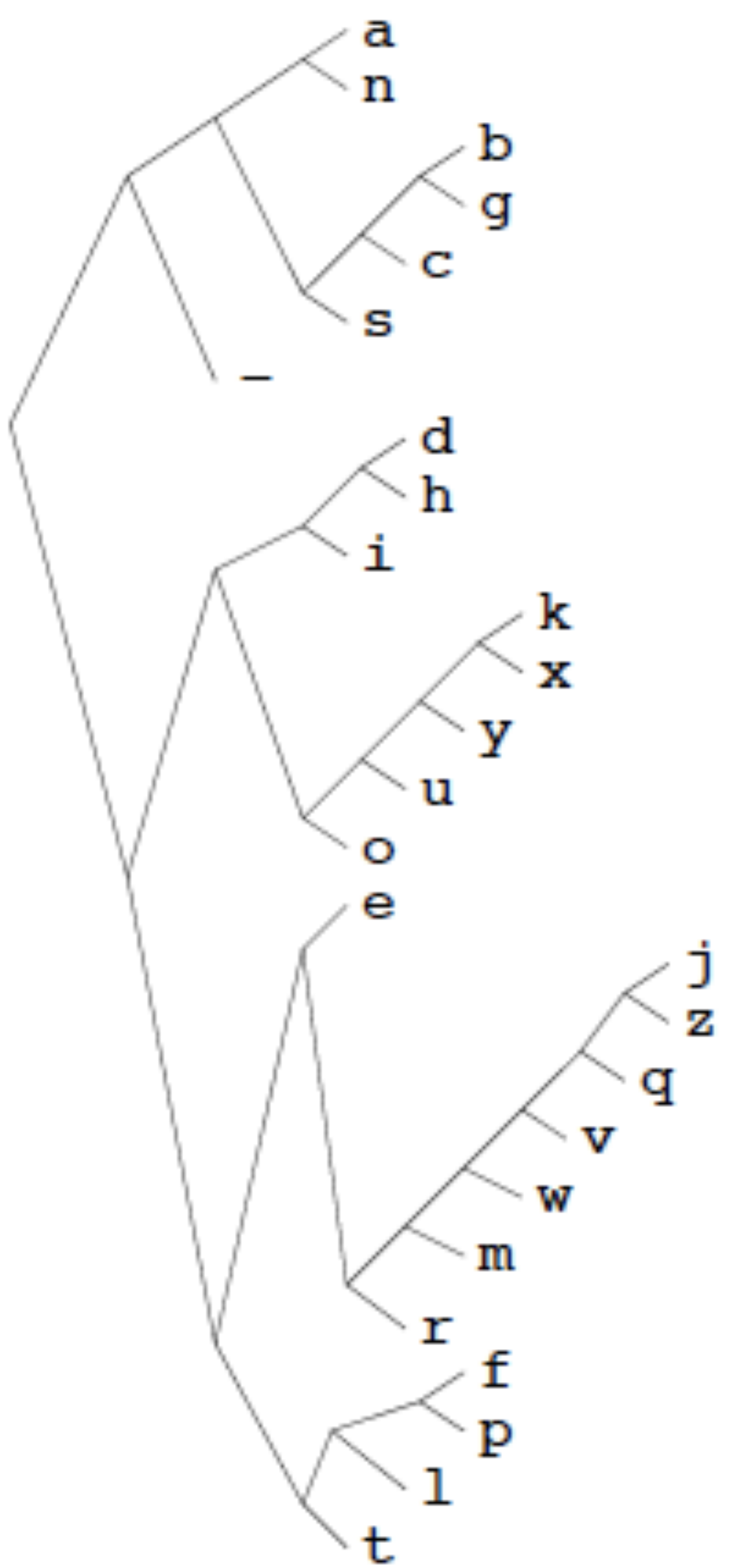
$$H(X) \leq \mathbb{E}[\ell(\mathrm{f}(X))]$$

entropy and average length: achievability

$$\min_{f \in F_{\mathcal{A}}} \mathbb{E}[\ell(f(X))] < H(X) + 1$$

Huffman code

a_i	p_i	$\log_2 \frac{1}{p_i}$	l_i	$c(a_i)$
a	0.0575	4.1	4	0000
b	0.0128	6.3	6	001000
c	0.0263	5.2	5	00101
d	0.0285	5.1	5	10000
e	0.0913	3.5	4	1100
f	0.0173	5.9	6	111000
g	0.0133	6.2	6	001001
h	0.0313	5.0	5	10001
i	0.0599	4.1	4	1001
j	0.0006	10.7	10	1101000000
k	0.0084	6.9	7	1010000
l	0.0335	4.9	5	11101
m	0.0235	5.4	6	110101
n	0.0596	4.1	4	0001
o	0.0689	3.9	4	1011
p	0.0192	5.7	6	111001
q	0.0008	10.3	9	110100001
r	0.0508	4.3	5	11011
s	0.0567	4.1	4	0011
t	0.0706	3.8	4	1111
u	0.0334	4.9	5	10101
v	0.0069	7.2	8	11010001
w	0.0119	6.4	7	1101001
x	0.0073	7.1	7	1010001
y	0.0164	5.9	6	101001
z	0.0007	10.4	10	1101000001
–	0.1928	2.4	2	01



not symbol-by-symbol

$$H(X_1 \dots X_n)_{\texttt{bits}} \leq \mathbb{E}[\ell_{X_1 \dots X_n}] \leq H(X_1 \dots X_n)_{\texttt{bits}} + 1$$

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E}[\ell_{X_1 \dots X_n}] = \lim_{n \to \infty} \frac{1}{n} H(X_1 \dots X_n)$$

almost-lossless data compression: the conventional way

- Asymptotic Equipartition Property
- Entropy rate at vanishing probability
- Link with Lossless

asymptotic equipartition property

3.1 ASYMPTOTIC EQUIPARTITION PROPERTY THEOREM

The asymptotic equipartition property is formalized in the following theorem.

Theorem 3.1.1 (*AEP*) *If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \qquad \text{in probability.} \tag{3.2}$$

asymptotic equipartition property
link with lossless data compression

Theorem 3.2.1 *Let X^n be i.i.d. $\sim p(x)$. Let $\epsilon > 0$. Then there exists a code that maps sequences x^n of length n into binary strings such that the mapping is one-to-one (and therefore invertible) and*

$$E\left[\frac{1}{n}l(X^n)\right] \leq H(X) + \epsilon \tag{3.23}$$

for n sufficiently large.

reflections on the conventional way

- why obsession with symbol-by-symbol codes?
- if not symbol-by-symbol, why prefix condition?
- why expected length only?
- why teach almost-lossless fixed-length codes?

variable-length lossless code

Compressor: $\mathsf{f}\colon \mathcal{X} \mapsto \{0,1\}^*$

Decompressor: $\mathsf{g}\colon \mathsf{f}(\mathcal{X}) \subset \{0,1\}^* \mapsto \mathcal{X}$

$$\mathsf{g}(\mathsf{f}(a)) = a, \;\; \forall a \in \mathcal{X}$$

$$\{0,1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}$$

optimum lossless code

$$P_X(1) \geq P_X(2) \geq \ldots .$$

$$\begin{aligned}
\mathsf{f}^*(1) &= \emptyset \\
\mathsf{f}^*(2) &= \mathsf{0} \\
\mathsf{f}^*(3) &= \mathsf{1} \\
\mathsf{f}^*(4) &= \mathsf{00} \\
\mathsf{f}^*(5) &= \mathsf{01} \\
\mathsf{f}^*(6) &= \mathsf{10} \\
\mathsf{f}^*(7) &= \mathsf{11} \\
\mathsf{f}^*(8) &= \mathsf{000} \\
\mathsf{f}^*(9) &= \mathsf{001} \\
\mathsf{f}^*(10) &= \mathsf{010} \\
&\vdots
\end{aligned}$$

optimum lossless code length

$$\ell(\mathsf{f}^*(x)) = \lfloor \log_2 x \rfloor$$

theorem

$$\ell(\mathsf{f}^*(x)) \leq \imath_X(x) = \log_2 \frac{1}{P_X(x)}$$

proof

$$P_X(k) \leq \frac{1}{k}$$

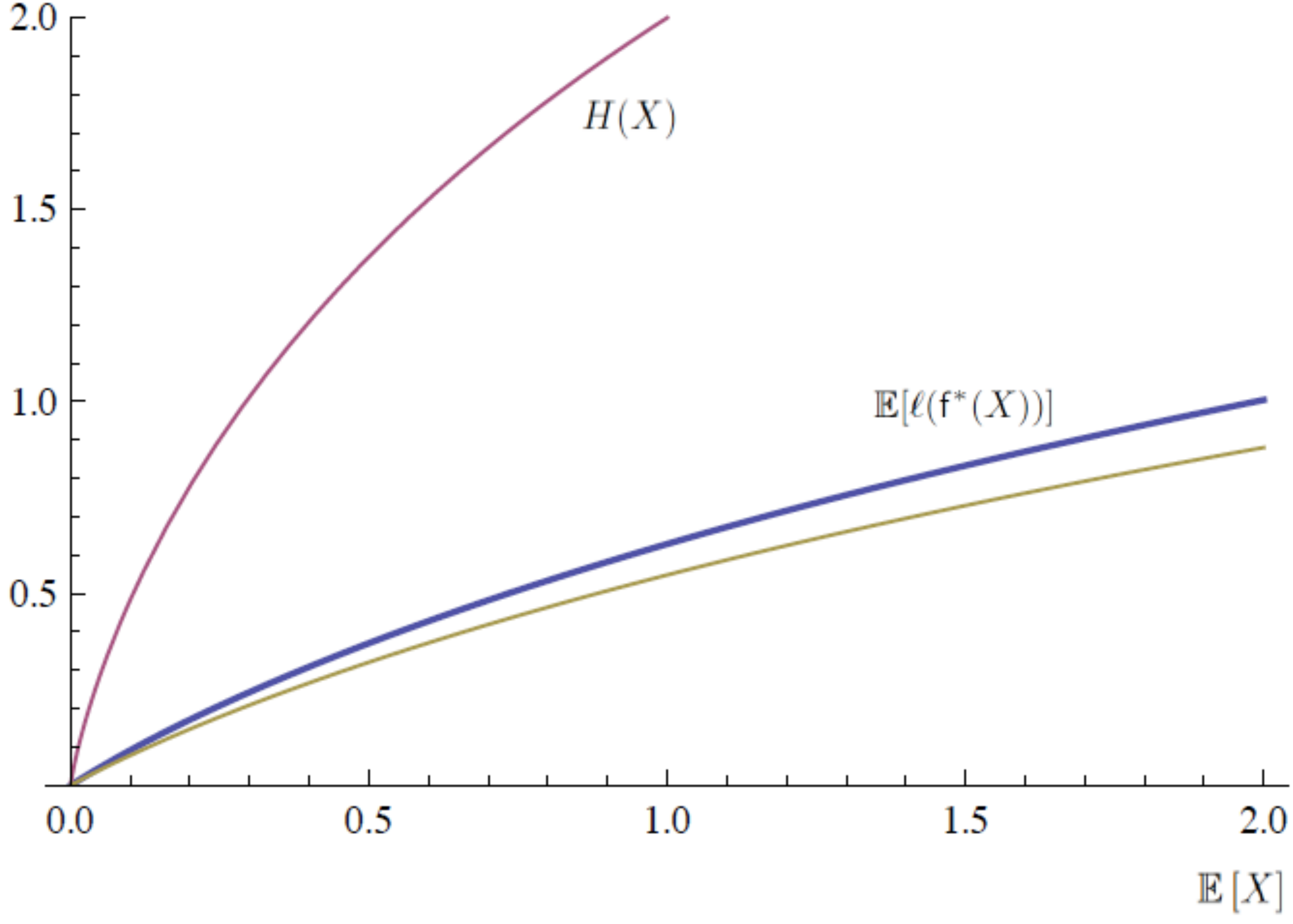
minimal average length

$$\mathbb{E}[\ell(\mathsf{f}^*(X))] \quad = \quad \sum_{k=1}^{\infty} \mathbb{P}[X \geq 2^k]$$

minimal average length

$$\mathbb{E}[\ell(\mathsf{f}^*(X))] \leq H(X)$$

example: geometric distribution

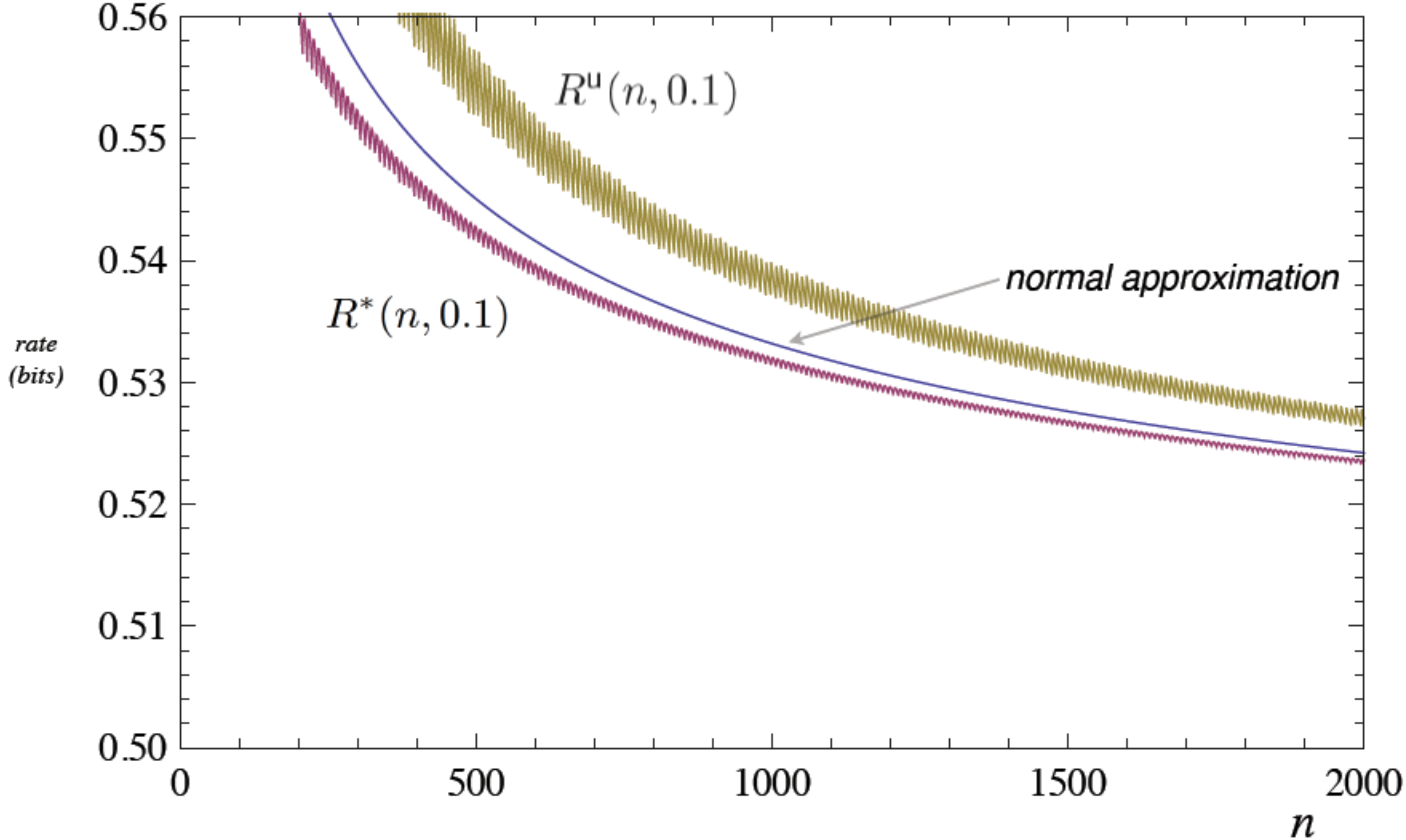


asymptotics

For an information-stable stationary source with entropy rate H,

$$\frac{1}{n}\ell(\mathsf{f}^*(X_1,\ldots,X_n)) \xrightarrow{i.p.} H$$

example: Bernoulli-0.11



almost-lossless fixed length
vs.
lossless variable length

$$\epsilon_X(2^k) = \mathbb{P}\left[\ell(\mathsf{f}^*(X)) \geq k\right]$$

↑

error probability of k-bit fixed-length code

reflections on the conventional way

- why obsession with symbol-by-symbol codes?
- if not symbol-by-symbol, why prefix condition?
- why expected length only?
- why teach almost-lossless fixed-length codes?