Elements of Information Theory

Chapter 6 — Entropy, AEP, Shannon's Theorems, and Applications

W. Szpankowski

Average Case Analysis of Algorithms on Sequences, Ch. 6

Roadmap of the chapter

Foundations

- 6.1 Entropy, relative entropy, mutual information
- 6.2 Entropy rate, Shannon–McMillan–Breiman, Rényi rates
- 6.3 Asymptotic Equipartition Property (AEP)

Shannon's three theorems (6.4)

- Source coding
- Channel coding
- Rate distortion

Applications (6.5)

- Phrase length in Lempel–Ziv & suffix-tree depth
- Shortest common superstring
- Fixed-database lossy Lempel–Ziv

Guiding theme

Entropy and the AEP are the standard tools linking *information theory* to the *average-case analysis of algorithms on words*.

Information theory in one paragraph

Born in October 1948 with Shannon's *A Mathematical Theory of Communication*. Three fundamental problems:

- **Source coding** — shortest description of a message. Limit = entropy × length.
- **Channel coding** — send *as much* information as possible, reliably, over a noisy channel.
- **Rate distortion** — when exact description is impossible, how well can we *approximate* the source?

Two workhorses for the analysis of algorithms

The Shannon–McMillan–Breiman theorem and the random-coding technique. Rényi's entropy appears almost as often as Shannon's in problems on strings.

Motivation: typical sequences

I.i.d. binary sequence X_1^n, symbol probabilities p and $q = 1 - p$.

- If $p = q = \frac{1}{2}$: every sequence has probability 2^{-n}.
- If $p \neq q$: a *typical* sequence has $\approx np + O(\sqrt{n})$ symbols of one kind, hence

$$p^{np+O(\sqrt{n})}\, q^{nq+O(\sqrt{n})} \;=\; 2^{-n(-p\log p - q\log q)+O(\sqrt{n})} \;\sim\; 2^{-nh},$$

where $h = -p\log p - q\log q$ is the Bernoulli entropy.

Punchline

All typical sequences carry *approximately the same* probability 2^{-nh}. Making this rigorous is the content of the AEP.

Throughout: $\log \equiv \log_2$ (bits); $0 \log 0 = 0$.

6.1 Entropy, Relative Entropy, Mutual Information

Shannon entropy

For a random variable X over a finite alphabet $\mathcal{A}$ with $P(a) = \Pr\{X = a\}$:

$$h(X) \;=\; -\,\mathbb{E}[\log P(X)] \;=\; -\sum_{a\in\mathcal{A}} P(a)\log P(a). \tag{6.1}$$

- Measures the average uncertainty of X.
- Equals the minimum average number of bits needed to describe X.
- Uniform over 32 outcomes $\Rightarrow h(X) = 5$ bits.

We also write $h(P)$ or h, and sometimes $h(X) = \mathbb{E}_P[-\log P(X)]$ to stress the averaging measure.

Example 6.1 — entropy = description length

Let X take eight values with probabilities

$$\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{16}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}\right), \qquad h(X) = 2 \text{ bits}.$$

A naïve fixed code needs 3 bits. Assign instead

$$0,\ 10,\ 110,\ 1110,\ 111100,\ 111101,\ 111110,\ 111111.$$

- Shorter codewords to more likely outcomes.
- Average length $=$ 2 bits $= h(X)$.

Lesson

The entropy is exactly the average description length achievable by an optimal variable-length code.

The family of information quantities (Def. 6.1)

$$\text{Joint entropy} \quad h(X,Y) = -\sum_{a,a'} P(a,a') \log P(a,a') \tag{6.2}$$

$$\text{Conditional entropy} \quad h(Y \mid X) = -\sum_{a,a'} P(a,a') \log P(a' \mid a) \tag{6.3}$$

$$\text{Relative entropy (KL)} \quad D(P\|Q) = \sum_{a} P(a) \log \frac{P(a)}{Q(a)} \tag{6.4}$$

$$\text{Mutual information} \quad I(X;Y) = \sum_{a,a'} P(a,a') \log \frac{P(a,a')}{P(a)P(a')} \tag{6.5}$$

Conventions: $0 \log \frac{0}{Q} = 0, \quad P \log \frac{P}{0} = \infty$.
$I(X;Y)$ is the KL divergence between the joint law and the product of marginals.

Rényi's entropy of order b (Def. 6.1(v))

For $-\infty \le b \le \infty$, $b \neq 0$:

$$h_b(X) \;=\; -\frac{\log \mathbb{E}[P^b(X)]}{b} \;=\; -\frac{1}{b}\log \sum_{a\in\mathcal{A}} P^{\,b+1}(a). \tag{6.6}$$

Limiting cases (Hilbert's inequalities on means):

$$h_{-\infty} = \min_{i\in\mathcal{A}} P(i) \ (\text{as } -\log), \qquad h_{\infty} = \max_{i\in\mathcal{A}} P(i) \ (\text{as } -\log),$$

$$h(X) = \lim_{b\to 0} h_b(X).$$

Why we care

Rényi entropies govern extremal parameters of digital trees: h_∞ and h_b appear in the *height*, $h_{-\infty}$ in the *shortest path* / fill-up level.

Identities — chain rules (Thm. 6.2)

$$h(X, Y) = h(X) + h(Y \mid X), \tag{6.9}$$

$$h(X_1, \ldots, X_n) = \sum_{i=1}^{n} h(X_i \mid X_{i-1}, \ldots, X_1), \tag{6.10}$$

$$I(X; Y) = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X), \tag{6.11}$$

$$I(X; X) = h(X). \tag{6.12}$$

Mutual information is the reduction in uncertainty about X once Y is observed — and it is symmetric.

Fundamental inequalities (Thm. 6.3)

$$D(P\|Q) \geq 0, \tag{6.13}$$

$$I(X;Y) \geq 0, \tag{6.14}$$

$$h(X) \geq h(X \mid Y), \tag{6.15}$$

$$h(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i). \tag{6.16}$$

Equalities: (6.13) iff $P = Q$; (6.14),(6.15) iff X, Y independent; (6.16) iff the X_i are independent.

One-line proof engine

$\log x \leq x - 1$ gives $-D(P\|Q) \leq \sum_a Q(a) - \sum_a P(a) = 0$; everything else follows.

6.2 Entropy Rate and Rényi Rates

Entropy rate — two definitions that agree

For a one-sided stationary sequence $\{X_k\}$:

$$h = -\lim_{n\to\infty} \frac{\mathbb{E}[\log P(X_1^n)]}{n}, \qquad h' = \lim_{n\to\infty} h(X_n \mid X_{n-1}, \ldots, X_1). \tag{6.17–6.18}$$

Conditional entropies are non-increasing in n (by (6.15) + stationarity), hence converge.

Theorem 6.4

For a stationary process both limits exist and $h = h'$.

Proof: chain rule + Cesàro mean of a convergent sequence.

Example 6.2 — memoryless and Markov rates

Memoryless source. $P(X_1^n) = P^n(X_1)$, so

$$h(X_1^n) = n\,\mathbb{E}[-\log P(X_1)] = n\,h(X_1).$$

Stationary Markov source (order 1). Using (6.18),

$$h \;=\; h(X_2 \mid X_1) \;=\; -\sum_{i,j\in\mathcal{A}} \pi_i\, p_{ij} \log p_{ij}, \tag{6.19}$$

where π is the stationary distribution and $P = \{p_{ij}\}$ the transition matrix.

The Shannon–McMillan–Breiman theorem

Theorem 6.5 (SMB)

For a stationary and ergodic sequence $\{X_k\}_{k\geq 1}$,

$$h = -\lim_{n\to\infty} \frac{\log P(X_1^n)}{n} \qquad \text{(a.s.)},$$

where h is the entropy rate.

Idea (Algoet–Cover "sandwich").

$$h^\infty \leq \liminf_n \tfrac{1}{n} \log \tfrac{1}{P(X_1^n)} \leq \limsup_n \tfrac{1}{n} \log \tfrac{1}{P(X_1^n)} \leq h^k$$

via the k-th order Markov approximation P^k; then $h^k \downarrow h^\infty$ by the martingale convergence theorem, and $h = h^\infty$ by Thm. 6.4.

- Removes the expectation $\mathbb{E}[\cdot]$ from the definition, replacing it with a.s. convergence of $\log P(X_1^n)$.

Rényi's entropy rates

For $-\infty \le b \le \infty$, with $P(w) = \Pr\{X_1^n = w\}$:

$$h_b \;=\; \lim_{n\to\infty} \frac{-\log \mathbb{E}[P^b(X_1^n)]}{bn} \;=\; \lim_{n\to\infty} \frac{-\log\Big(\sum_{w\in\mathcal{A}^n} P^{\,b+1}(w)\Big)^{1/b}}{n}. \tag{6.20}$$

Memoryless (alphabet size V): $h_b = -\frac{1}{b}\log\big(p_1^{b+1} + \cdots + p_V^{b+1}\big)$, $h_{-\infty} = -\log p_{\min}$, $h_{\infty} = -\log p_{\max}$.

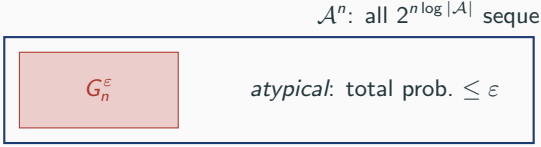
Markovian: $h_b = -\frac{1}{b}\log \lambda_{[b+1]}$, where $\lambda_{[b]}$ is the largest eigenvalue of the Schur product $P_{[b]} = P \circ \cdots \circ P$; $h_{\pm\infty}$ solved via min/max mean cycle in a weighted digraph (Karp).

6.3 Asymptotic Equipartition Property

Typical sequences (Def. 6.6)

For $\varepsilon > 0$, the typical set G_n^ε w.r.t. measure P is

$$G_n^\varepsilon = \left\{ x_1^n \in \mathcal{A}^n : \; 2^{-n(h(P)+\varepsilon)} \leq P(x_1^n) \leq 2^{-n(h(P)-\varepsilon)} \right\}. \tag{6.25}$$



All typical sequences carry *almost the same* probability $\approx 2^{-nh}$ — a direct corollary of SMB.

Asymptotic Equipartition Property (Thm. 6.7)

For a stationary, ergodic X_1^n and $\varepsilon > 0$, partition $\mathcal{A}^n = B_n^\varepsilon \cup G_n^\varepsilon$ so that for $n \geq N_\varepsilon$:

$$2^{-nh(1+\varepsilon)} \leq P(x_1^n) \leq 2^{-nh(1-\varepsilon)}, \qquad x_1^n \in G_n^\varepsilon, \tag{6.26}$$

$$\Pr\{B_n^\varepsilon\} \leq \varepsilon, \tag{6.27}$$

$$(1-\varepsilon)\, 2^{nh(1-\varepsilon)} \leq |G_n^\varepsilon| \leq 2^{nh(1+\varepsilon)}. \tag{6.28}$$

Three-in-one

Typical sequences (i) are *equiprobable*, (ii) carry *almost all* the probability, (iii) number $\approx 2^{nh}$.

Jointly typical sequences (Def. 6.8, Thm. 6.9)

(x_1^n, y_1^n) jointly typical: x_1^n, y_1^n individually typical *and* $2^{-n(h(X,Y)+\varepsilon)} \le P(x_1^n, y_1^n) \le 2^{-n(h(X,Y)-\varepsilon)}$.

Joint AEP

$\Pr\{(X_1^n, Y_1^n) \in G_n^\varepsilon\} \ge 1-\varepsilon, \quad |G_n^\varepsilon| \doteq 2^{nh(X,Y)}$, and for *independent* $\widetilde{X}_1^n, \widetilde{Y}_1^n$ with the same marginals

$$\Pr\{(\widetilde{X}_1^n, \widetilde{Y}_1^n) \in G_n^\varepsilon\} \doteq 2^{-n\,I(X;Y)}. \tag{6.32}$$

Channel-coding preview. About $2^{nh(X)}$ typical X's, $2^{nh(Y)}$ typical Y's, but only $2^{nh(X,Y)}$ jointly typical pairs $\Rightarrow$ one inspects $\approx 2^{nI(X;Y)}$ inputs per output.

AEP for biased distributions (Thm. 6.10)

Motivating problem. A 19th-century Polish manuscript surfaces in Paris — is it a lost Mickiewicz? Compare unknown source P against the *known* distribution Q of authentic samples.

Divergence rate: $D(P\|Q) = -\lim_n \frac{1}{n}\,\mathbb{E}_P\big[\log Q(X_1^n)/P(X_1^n)\big]$ (needs joint mixing, (JMX)).

Biased AEP

Under joint mixing, for $x_1^n \in G_n^\varepsilon(Q)$:

$$Q(x_1^n) \doteq 2^{-n(h(P)+D(P\|Q))}, \qquad |G_n^\varepsilon(Q)| \doteq 2^{n(h(P)+D(P\|Q))}.$$

The "cost" of scoring P-data with the wrong model Q is exactly the divergence $D(P\|Q)$.

Lossy generalization of the AEP (6.3.4)

Distortion $d : \mathcal{A} \times \hat{\mathcal{A}} \to \mathbb{R}_+$, per-symbol $d_n(x_1^n, \hat{X}_1^n) = \frac{1}{n}\sum_i d(x_i, \hat{x}_i)$; ball $B_D(c_1^k) = \{x_1^k : d_k \le D\}$.

Generalized Rényi entropy (Def. 6.12)

$$r_b(D) = -\lim_{k\to\infty} \frac{\log \mathbb{E}[P^b(B_D(X_1^k))]}{bk}, \qquad r_0(D) = -\lim_{k\to\infty} \frac{\mathbb{E}[\log P(B_D(X_1^k))]}{k}.$$

Lossy AEP (Thms. 6.13, 6.16). For weakly mixing sources, $r_b(D)$ exists, $-\frac{1}{k}\log P(B_D(X_1^k)) \to r_0(D)$ a.s., and $P(B_D(x_1^n)) \doteq 2^{-nr_0(D)}$ on the good set — the AEP with balls in place of points.

6.4 Three Theorems of Shannon

Source coding: codes and Kraft's inequality

A source code $\phi : \mathcal{A} \to \{0,1\}^*$; extension $\phi_n : \mathcal{A}^n \to \{0,1\}^*$.

- **Uniquely decodable**: one-to-one.
- **Prefix / instantaneous**: no codeword is a prefix of another.

Theorem 6.17 (Kraft)

Prefix codeword lengths $\ell_1, \dots, \ell_m$ satisfy

$$\sum_{i=1}^{m} 2^{-\ell_i} \le 1,$$

and conversely any lengths obeying this build a prefix code.

Proof: a prefix code is a set of leaves in a binary tree; count descendants at depth $\ell_{\max}$.

You cannot beat the entropy (Thm. 6.18)

Average length $\bar{\ell}_n(C^n) = \sum_{x_1^n} \ell(x_1^n) P(x_1^n)$. Then for any prefix code

$$\bar{\ell}_n(C^n) \;\geq\; h_n(X_1^n).$$

Proof. With $K = \sum_{x_1^n} 2^{-\ell(x_1^n)} \leq 1$,

$$\bar{\ell}_n - h_n = \underbrace{\sum_{x_1^n} P(x_1^n) \log \frac{P(x_1^n)}{2^{-\ell(x_1^n)}/K}}_{\geq 0 \text{ (divergence)}} - \underbrace{\log K}_{\leq 0} \;\geq\; 0.$$

Redundancy

$\bar{r}_n(C^n) = \bar{\ell}_n(C^n) - h_n(X_1^n) \geq 0$; the average bits per symbol are bounded below by the entropy rate — even if errors are allowed.

Source Coding Theorem (Thm. 6.19, Shannon 1948)

Bit rate $R = \frac{\log M}{n}$, code $(2^{nR}, n)$, error $P_E = \Pr\{X_1^n \neq \hat{X}_1^n\}$. Universal source coding: smallest R with $P_E < \varepsilon$, *regardless* of the unknown P.

Theorem 6.19

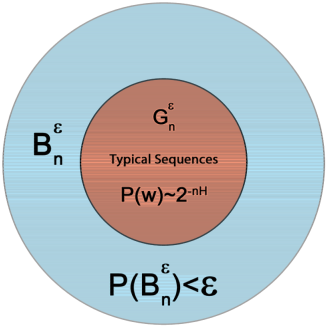
(i) There exist $(2^{nR}, n)$ universal codes with any $R > h(P)$ and $P_E < \varepsilon$ for large n.

(ii) Conversely, every code with $P_E \to 0$ must have $R > h(P)$.

Proof via AEP. Encode only the typical set G_n^ε ($P(G_n^\varepsilon) \geq 1 - \varepsilon$); its size gives $R = \frac{1}{n} \log |G_n^\varepsilon| \in [h(P) - \varepsilon, h(P) + \varepsilon]$. Converse: $R < h$ forces $|C'| < 2^{n(h-2\delta)}$, so $P_E > 1 - \varepsilon$.

Source Coding Explained

Theorem 1: AEP and Typical Sequences



Shannon-McMilan-Breiman:

$-\frac{1}{n}\log P(X_1^n) \to H(X) \quad (\text{pr.})$

$H(X)$ is the entropy rate.

Asymptotic Equipartition Property: Sequences of length n can be partitioned into

good set	G_n^ε	$P(w) \sim 2^{-nH(X)}, \quad w \in G_n^\varepsilon$
bad set	B_n^ε	$P(B_n^\varepsilon) < \varepsilon.$

Also, $|G_n^\varepsilon| \sim 2^{nH(X)}$.

Channel coding: the discrete memoryless channel



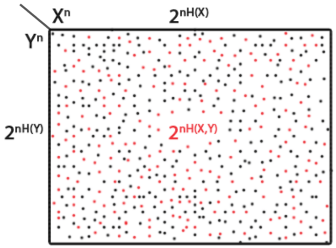
DMC without feedback: $P(y_1^n \mid x_1^n) = \prod_i P(y_i \mid x_i)$; (M, n) code, $M = \lceil 2^{nR} \rceil$, average error $P_E = \Pr\{W \neq \hat{W}\}$.

Channel capacity

$$C = \max_{P(X)} I(X;Y). \tag{6.60}$$

Divide $\approx 2^{nh(Y)}$ typical outputs into $2^{nI(X;Y)}$ distinguishable groups $\Rightarrow$ at most $2^{nI(X;Y)}$ reliable messages.

Theorem 2: Shannon Random Decoding Rule



There are $2^{nH(X)}$ X-typical sequences

There are $2^{nH(Y)}$ Y-typical sequences

There are $2^{nH(X,Y)}$ jointly X,Y-typical pair of sequences

Decoding Rule: Declare that sequence sent X is the one that is jointly typical with the received sequence Y provided there is unique X satisfying this property!

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y).

2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$. Indeed:
- there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n, respectively.
- there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y).

The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{nH(X)+H(Y)}} = 2^{-nI(X,Y)}.$$

3. Probability of error when 2^{nR} messages are sent is approximately

$$\min P(\text{error}) \sim 2^{-n(\sup_{P(X)} I(X,Y)-R)} = 2^{-n(C-R)}.$$

4. In conclusion:

$$R < C \qquad P(\text{error}) \sim 2^{-n\delta}$$

Converse via Fano's inequality

Goal: if $R > C$ then P_E is bounded away from 0. Two ingredients:

$$I(X_1^n; Y_1^n) \;\le\; nC \qquad \text{(single-letterization over the DMC)},$$

$$h(W \mid Y_1^n) \;\le\; 1 + nR\,P_E \qquad \text{(Fano)}. \tag{6.63}$$

Combining with $h(W \mid Y_1^n) \ge h(W) - I(X_1^n; Y_1^n) \ge n(R - C)$ yields

$$\boxed{P_E \;\ge\; 1 - \frac{C}{R} - \frac{1}{nR}} \tag{6.61}$$

so reliable transmission ($P_E < \varepsilon$) forces $R < C$.

Channel Coding Theorem (Thm. 6.20)

Theorem 6.20 (Shannon)

(i) For every $\varepsilon > 0$ and $R < C$ there is a sequence of $(2^{nR}, n)$ codes with $P_E < \varepsilon$ for large n.

(ii) Conversely, any code sequence with $P_E \to 0$ has $R < C$.

Achievability = random coding. Draw 2^{nR} codewords i.i.d. from the capacity-achieving P; decode by joint typicality. Then

$$P_E \;\le\; \varepsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\varepsilon)} \;\le\; \varepsilon + 2^{-n(I(X;Y)-R-3\varepsilon)} \;\le\; 2\varepsilon$$

for $R < I(X;Y) - 3\varepsilon$; choose P achieving C.

Rate distortion: setup

When exact reconstruction is impossible or undesirable (continuous signals, lossy compression): *minimum rate to achieve a given distortion?*

Code $\phi_n : \mathcal{A}^n \to \{1, \dots, 2^{nR}\}$, reproduction $\psi_n : \{1, \dots, 2^{nR}\} \to \hat{\mathcal{A}}^n$, $\hat{X}_1^n = \psi_n(\phi_n(X_1^n))$; bounded distortion $d_{\max} < \infty$.

- (R, D) is ε-achievable if $\Pr\{d_n(X_1^n, \hat{X}_1^n) \le D + \delta\} \ge 1 - \varepsilon$.
- Equivalent (asymptotically) to $\mathbb{E}[d_n] \le D$.
- $R(D) = \lim_{\varepsilon \to 0} R_\varepsilon(D) =$ rate distortion function.

Information rate distortion function:

$$R_I(D) = \min_{P(\hat{x}|x):\, \mathbb{E}[d(X,\hat{X})] \le D} I(X; \hat{X}). \tag{6.68}$$

Rate Distortion Theorem (Thm. 6.24)

Theorem 6.24 (Shannon, 1959)

For memoryless sources with bounded distortion,

$$R(D) \;=\; R_I(D) \;=\; \min_{P(\hat{x}|x):\, \mathbb{E}[d(X,\hat{X})]\le D} I(X;\hat{X}). \tag{6.69}$$

$R_I(D)$ is non-increasing, convex, continuous (Lemma 6.23).

Operational view (Def. 6.25, Thm. 6.26)

With $N_\varepsilon(n, D)$ the minimum number of D-balls covering a $(1-\varepsilon)$-mass set,

$$R(D) = R_O(D) = \lim_{\varepsilon\to 0} \lim_{n\to\infty} \tfrac{1}{n} \log N_\varepsilon(n, D).$$

Rate distortion = covering the typical set with distortion balls.

6.5 Applications to Algorithms on Words

Three applications

1. **Phrase length in Lempel–Ziv = suffix-tree depth.** Lossy and lossless; depth $\sim \log n / r_0(D)$.
2. **Shortest common superstring.** Greedy vs. optimal overlap; both $\sim \frac{1}{h}\, n \log n$ — greedy is asymptotically optimal.
3. **Fixed-database lossy Lempel–Ziv.** Average bit rate $\to r_0(D)$, the generalized Shannon entropy.

Common thread

Each parameter is controlled by an (ordinary, Rényi, or generalized) *entropy*, obtained through the AEP and typicality arguments.

App. 1 — LZ phrase length & suffix-tree depth

Database $\hat{X}_1^n$ (possibly distorted, measure $\hat{P}$); source X_{n+1}^∞ with measure P; $\{\hat{X}_k, X_k\}$ ψ-strongly mixing; Hamming distortion, fixed $D > 0$.

- D_n: longest prefix of X_{n+1}^∞ within distance D of some length-K substring of $\hat{X}_1^n$.
- $D_n(M)$: the M-th such (depth of the M-th suffix).
- At $D = 0$ these are exactly the depth of insertion and the M-th depth in a suffix tree over $\hat{X}_1^n$.

Generalized (lossy) Shannon entropy driving the growth:

$$\hat{r}_0(D) = -\lim_{k\to\infty} \frac{\mathbb{E}_P[\log \hat{P}(B_D(X_1^k))]}{k}. \tag{6.75}$$

App. 1 — the depth is $\log n / \hat{r}_0(D)$

Theorem 6.27 (Łuczak & Szpankowski, 1997)

For a ψ-strongly mixing $\{\hat{X}_k, X_k\}$ with $\hat{r}_0(D) > 0$,

$$\lim_{n\to\infty} \frac{D_n}{\log n} = \lim_{n\to\infty} \frac{D_n(M)}{\log n} = \frac{1}{\hat{r}_0(D)} \qquad \text{(pr.)}.$$

- Lossless case $D = 0$: recovers the classic suffix-tree depth $\sim \frac{1}{h} \log n$.
- The Lempel–Ziv phrase length inherits the same asymptotics — the compression "look-ahead" grows logarithmically, at rate set by the (lossy) entropy.

App. 2 — shortest common superstring

n independent strings $X_1^\ell(1), \ldots, X_1^\ell(n)$, length $\ell \to \infty$. Superstring = string containing all as substrings. Maximize overlap $O_n = n\ell -$ (superstring length).

- Optimal O_n^{opt} is NP-hard (Hamiltonian path).
- **RGREEDY**: repeatedly attach the string with the largest overlap with the current superstring, giving O_n^{gr}.

Overlaps are distributed like trie depths: $D_n(i) \sim \frac{1}{h} \ln n$, and $H_n = \max_i D_n(i) \sim \frac{2}{h_1} \ln n$ with $h_1 = -\ln(p_1^2 + \cdots + p_V^2)$.

App. 2 — greedy is asymptotically optimal

Theorem 6.30 (Alexander 1996; Frieze & Szpankowski 1998)

For n memoryless sources over $\mathcal{A}$, $|\mathcal{A}| = V$,

$$\lim_{n\to\infty} \frac{O_n^{\mathrm{opt}}}{n \log n} = \lim_{n\to\infty} \frac{O_n^{\mathrm{gr}}}{n \log n} = \frac{1}{h} \qquad \text{(pr.)},$$

provided $\ell > \dfrac{4}{h_1} \ln n$, $h_1 = -\ln(p_1^2 + \cdots + p_V^2)$.

- Both grow like $\frac{1}{h}\, n \log n$: the simple greedy overlap matches the optimum to leading order.
- Caveat: SCS is a *poor* compressor — it exploits only prefix/suffix redundancy, so $c_n \to 1$ when ℓ grows faster than $\log n$.

App. 3 — fixed-database lossy Lempel–Ziv

Common database $\hat{X}_1^n \sim \hat{P}$ (Markov), source $X_1^M \sim P$ (Markov), independent. Partition X_1^M into phrases $Z^1, \ldots, Z^{\Pi_n}$:

$$D_n^1 = \max\{k : \ d(\hat{X}_i^{i+k-1}, X_1^k) \le D, \ 1 \le i \le n-k+1\}.$$

Each phrase $\to$ (pointer, length), cost $\log n + \Theta(\log L)$; bit rate

$$r_n(X_1^n) = \frac{1}{M} \sum_{i=1}^{\Pi_n} \big(\log n + \Theta(\log L_n^i) \big). \tag{6.97}$$

Phrase length concentrates: $D_n^1 \doteq \frac{1}{\hat{r}_0(D)} \log n$ (Thm. 6.27 arguments).

App. 3 — bit rate converges to $\hat{r}_0(D)$

Theorem 6.31

Fixed-database model, both sources Markovian with all transition probabilities positive. The average bit rate satisfies

$$\lim_{n\to\infty} \lim_{M\to\infty} \mathbb{E}[r_n(X_1^M)] \;=\; \hat{r}_0(D),$$

where $\hat{r}_0(D) = \lim_{n\to\infty} \dfrac{\mathbb{E}_P[-\log \hat{P}(B_D(X_1^n))]}{n}$.

- The lossy compression rate is exactly the *generalized Shannon entropy* of the ball — the lossy analogue of "rate = entropy".
- Proof: split into "long" / "short" phrases (Wyner–Ziv, Kontoyiannis) and apply the lossy AEP.

Summary

Key results at a glance

Concept	Statement	Role
SMB (Thm. 6.5)	$-\frac{1}{n}\log P(X_1^n) \to h$ a.s.	foundation of AEP
AEP (Thm. 6.7)	$\lvert G_n^\varepsilon\rvert \doteq 2^{nh}$, equiprobable	typicality method
Source coding (6.19)	best rate $= h(P)$	optimal compression
Channel coding (6.20)	best rate $= C = \max I(X;Y)$	reliable comms
Rate distortion (6.24)	$R(D) = \min I(X;\hat{X})$	lossy limits

Why this matters for the analysis of algorithms

- **Entropy is the universal constant** for word problems: depths, heights, phrase lengths, overlaps are all $\Theta(\log n)$ or $\Theta(n \log n)$ with entropy-valued constants.
- **Rényi entropies** refine the picture: h_1 for maxima (heights), $h_{\pm\infty}$ for extremes (shortest paths).
- **The AEP / typicality method** converts a.s. probabilistic statements into tight combinatorial counts.
- **Lossy generalizations** ($r_0(D)$, balls B_D) carry the same program to approximate pattern matching and lossy compression.

Bridge

Chapter 6 supplies the information-theoretic constants; Part III (generating functions, Mellin, (de)Poissonization) supplies the machinery to compute them precisely.

Selected references

- C. E. Shannon, *A Mathematical Theory of Communication*, Bell Syst. Tech. J., 1948.
- T. Cover & J. Thomas, *Elements of Information Theory*, Wiley — primary source for §6.1–6.4.
- I. Csiszár & J. Körner, *Information Theory*.
- T. Łuczak & W. Szpankowski, *A suboptimal lossy data compression based on approximate pattern matching*, 1997 (Thms. 6.27, 6.31).
- A. Frieze & W. Szpankowski, *Greedy algorithms for the shortest common superstring*, 1998 (Thm. 6.30).
- P. Algoet & T. Cover, sandwich proof of SMB (Thm. 6.5).

Entropy is the price of description,
the AEP is how we pay it,
and on words that price is a running time.