Pattern Matching, Entropy and Biological Sequence Analysis

Ioannis Kontoyiannis
Athens U of Econ & Business

Greek Stochastics α'

Lefkada, Greece, August 2009
AMX/EDT

Outline

I. Exact Pattern Matching & Lossless Data Compression

Waiting times and match lengths

Strong approximation

The AEP and its refinements

II. Approximate Pattern Matching & Lossy Data Compression

Large deviations

Finer asymptotics

The generalized AEP and its refinements

Example 1: Lossless Data Compression

message: $X_1\, X_2 \cdots X_n$

database: $Y_1\;\; Y_2\;\; Y_3\;\; \cdots\;\; Y_W\; Y_{W+1}\;\; \cdots\;\; Y_{W+n-1}\;\; \cdots$

Compression algorithm [Wyner-Ziv 89]: Describe $(X_1, X_2, \ldots, X_n)$ as the position W_n of its first appearance in the database $(Y_1, Y_2, \ldots)$

E.g. ($n = 5$ and $W_n = 15$):

$$\underbrace{1\,0\,1\,1\,0}$$

$$0\,1\,0\,1\,1\,1\,0\,1\,0\,0\,1\,1\,1\,0\,\underbrace{1\,0\,1\,1\,0}\,1\,1\,0\,0\,1\,1\,0\,1\,1\cdots$$

Question

What is the *rate* of this algorithm?

Answer

$$\approx \frac{\log W_n}{n} \to H, \quad \text{the entropy rate of } \{X_n\}, \quad \text{a.s}$$

Second Example: DNA Template Matching

template: $X_1 \; X_2 \; \cdots$

sequence: $Y_1 \; Y_2 \; Y_3 \; \cdots\cdots \; Y_m$

Matching algorithm: Find longest initial string $(X_1, X_2, \ldots, X_{L_m})$ matching somewhere into $(Y_1, Y_2, \ldots, Y_m)$ with $\leq 15\%$ mismathces

E.g. $(m = 18$ and $L_m = 8)$:

$$\underbrace{\text{A C C T A G T}}\;\text{A} \cdots$$

$$\text{C C A G C T}\;\underbrace{\text{A C C G A G T}}\;\text{G A G T C}$$

Question What is an "atypically" large L_m ?

Answer via Duality $L_m \geq n \quad$ iff $\quad \inf_{k \geq n} W_k \leq m$

$$\frac{\log W_n}{n} \to R \text{ a.s.} \qquad \Rightarrow \qquad \frac{L_m}{\log m} \to \frac{1}{R} \text{ a.s.}$$

Outline of Part I

Exact Pattern Matching & Lossless Data Compression

- $\rightsquigarrow$ **Waiting times** (and recurrence times)
- $\rightsquigarrow$ **Strong approximation**: $W_n \approx \frac{1}{P(X_1, X_2, \ldots, X_n)}$
- $\rightsquigarrow$ The Asymptotic Equipartition Property (**AEP**)
  - $\rightsquigarrow$ First-order asymptotics of W_n; optimality of LZ compression
- $\rightsquigarrow$ Refinements of the AEP
  - $\rightsquigarrow$ Second-order asymptotics of W_n; LZ optimality revisited
- $\rightsquigarrow$ **Duality** and **match lengths**
  - $\rightsquigarrow$ More realistic LZ compression and optimality
  - $\rightsquigarrow$ Second-order asymptotics for match lengths

The Setting

Let

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ be finite-valued, stationary, ergodic process with distribution P and values in A

$\boldsymbol{Y} = \{Y_1, Y_2, \ldots\}$ be finite-valued, stationary, ergodic process with distribution Q and values in A

Write

$$X_m^n = (X_m, X_{m+1}, \ldots, X_n), \quad 1 \le m \le n \le \infty$$

$$x_m^n = (x_m, x_{m+1}, \ldots, x_n), \quad 1 \le m \le n \le \infty, \text{ etc}$$

Define The **waiting time** $\boldsymbol{W_n = \inf\{k \ge 1 : X_1^n = Y_k^{k+n-1}\}}$

$$X_1\, X_2\, \cdots\, X_n$$

$$Y_1\;\; Y_2\;\; Y_3\;\; \cdots\cdots\;\; Y_W\;\; Y_{W+1}\;\; \cdots\;\; Y_{W+n-1}\;\; \cdots$$

Problem How does W_n behave as $n \to \infty$?

Strong Approximation: $W_n \approx \frac{1}{Q(X_1^n)}$

Intuition

We expect W_n to be close to the reciprocal of the probability that the pattern X_1^n appears in $\boldsymbol{Y}$, i.e., $W_n \approx \frac{1}{Q(X_1^n)}$

Theorem 1: Strong Approximation [K 98][Dembo-K 99][Chi 01]

If $\boldsymbol{Y}$ has either $\psi(k) \to 0$ or $\sum_k \phi(k) < \infty$, then:

$$\log\left[\boldsymbol{W_n Q(X_1^n)}\right] = \boldsymbol{O(\log n)} \quad \text{a.s.}$$

$$\text{Recall:} \quad \psi(k) = \sup\left\{\left|\frac{Q(B|A)}{Q(B)} - 1\right| \;:\; B \in \sigma(Y_k^\infty),\ A \in \sigma(Y_{-\infty}^0),\ Q(A) > 0\right\}$$

$$\phi(k) = \sup\{|Q(B|A) - Q(B)| \;:\; B \in \sigma(Y_k^\infty),\ A \in \sigma(Y_{-\infty}^0),\ Q(A) > 0\}$$

Therefore, $\log W_n \approx -\log Q(X_1^n)$

But how does $-\log Q(X_1^n)$ behave?

Proof of Theorem 1

[LB] Under stationarity alone, a simple union bound yields

$$\Pr(\log[W_n Q(X_1^n)] < -2\log n | X_1^n = x_1^n) \;=\; \Pr\left(W_n < \frac{e^{-2\log n}}{Q(x_1^n)}\middle| X_1^n = x_1^n\right)$$

$$\leq \sum_{j=1}^{\frac{1}{n^2 Q(x_1^n)}} \Pr\left(W_n = j \middle| X_1^n = x_1^n\right) \;\leq\; \frac{1}{n^2 Q(x_1^n)} Q(x_1^n) \;=\; \frac{1}{n^2}$$

and the lower bound follows by Borel-Cantelli.

[UB] For the upper bound in the general case, blocking *a la* Ibragimov.

In the special case where both $\boldsymbol{X}, \boldsymbol{Y}$ are IID,

the probability $\Pr(\log[W_n Q(X_1^n)] > 3\log n | X_1^n = x_1^n)$ is

$$\Pr\left(W_n > K := \frac{n^3}{Q(X_1^n)}\middle| X_1^n = x_1^n\right)$$

$$\leq \Pr\left(Y_1^n \neq x_1^n,\; Y_{n+1}^{2n} \neq x_1^n,\; \ldots,\; Y_{K-n+1}^{K} \neq x_1^n\right)$$

$$\leq [1 - Q(x_1^n)]^{K/n} \;\leq\; \cdots \;\leq\; 2/n^2$$

and the upper bound again follows from Borel-Cantelli. □

Asymptotics of $-\log Q(X_1^n)$

Assume for the rest of part I that $\boldsymbol{X} \stackrel{\mathcal{D}}{=} \boldsymbol{Y}$

Simplest case when $\boldsymbol{X}, \boldsymbol{Y}$ both IID $\sim P$ on A. Then:

$$-\log P(X_1^n) = \sum_{i=1}^{n} [-\log P(X_i)]$$

Simple IID partial sums with:

$\Rightarrow$ mean $H = E[-\log P(X_1)] =$ **entropy of $\boldsymbol{X}$**

$\Rightarrow$ variance $\sigma^2 = \mathrm{Var}[-\log P(X_1)] =$ **minimal coding variance of $\boldsymbol{X}$**

More generally...

Asymptotics of $-\log P(X_1^n)$

LLN (Asymptotic Equipartition Property, or **AEP**,
or Shannon-McMillan-Breiman Theorem 1948-57)

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

CLT (Yushkevich 53, Ibragimov 62)

$$\frac{-\log P(X_1^n) - nH}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,\sigma^2)$$

LIL (Philipp & Stout 75)

$$\limsup_{n\to\infty} \frac{-\log P(X_1^n) - nH}{\sqrt{2n\log\log n}} = \sigma \quad \text{a.s.}$$

"Functional" versions, etc.

First-Order Asymptotics for W_n

Recall: the **entropy rate** of a stationary process $\boldsymbol{X}$ is:

$$H = \lim_{n\to\infty} \frac{1}{n} E[-\log P(X_1^n)]$$

Theorem 1 says: $\log W_n \approx -\log P(X_1^n) + O(\log n)$ a.s.
This together with the AEP imply:

Corollary 1 [Wyner-Ziv 89][Shields 93][Marton-Shields 95][K 98]

If $\boldsymbol{X} \stackrel{\mathcal{D}}{=} \boldsymbol{Y}$ has either $\psi(k) \to 0$ or $\sum_k \phi(k) < \infty$, then:

$$\frac{\log W_n}{n} \to H \quad \text{a.s.}$$

Idealized LZ compression algorihtm [Wyner-Ziv 89]: Describe X_1^n as W_n

message: $X_1 X_2 \cdots X_n$

database: $Y_1\ Y_2\ Y_3\ \cdots\ Y_W\ Y_{W+1}\ \cdots\ Y_{W+n-1}\ \cdots$

Questions What is the *rate* of this algorithm? How well does it compress?

Compression Performance

Corollary 1 says that the **rate** of this algorithm is:

$$\frac{\log W_n}{n} \;\to\; H \;\;\text{"bits/symbol," a.s., as } n \to \infty$$

Recall that a **compression algorithm** is a "nice" collection of **invertible** maps

$$C_n : A^n \longrightarrow \{0,1\}^* = \cup_{k\geq 1}\{0,1\}^k$$

with associated *length functions*

$$\ell_n(x_1^n) := \text{length of } C_n(x_1^n), \quad \text{bits}$$

In view of the following, the LZ algorithm above is compression-optimal

Pointwise Source Coding Theorem [Barron 85][Kieffer 91]

For any stationary ergodic process $\boldsymbol{X}$ and any compression algorithm:

$$\liminf_{n\to\infty} \frac{\ell_n(X_1^n)}{n} \geq H \quad \text{a.s.}$$

Second-Order Asymptotics for W_n

Recall: the **minimal coding variance** of a stationary process $\boldsymbol{X}$ is:

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}[-\log P(X_1^n)]$$

Combining Theorem 1, $\log W_n \approx -\log P(X_1^n) + O(\log n)$,
with the CLT/LIL refinements of the AEP yields:

Recall: $\gamma(k) = \max_{a\in A} E|\log P(X_0 = a|X_{-\infty}^0) - \log P(X_0 = a|X_{-k}^0)|$

Corollary 2

If $\boldsymbol{X} \stackrel{\mathcal{D}}{=} \boldsymbol{Y}$ has both $\psi(k), \gamma(k) \to 0$ "fast enough," then:

CLT [A.J. Wyner 93][K 98] $$\frac{\log W_n - nH}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,\sigma^2)$$

LIL [K 98] $$\limsup_{n\to\infty} \frac{\log W_n - nH}{\sqrt{2n\log\log n}} = \sigma \quad \text{a.s.}$$

Question How good is this in terms of compression?

Finer Compression Performance

Corollary 2 says that, for large n, the **rate** of this LZ algorithm is:

$$\frac{\log W_n}{n} \approx N\Big(H, \frac{\sigma^2}{n}\Big) \text{ bits/symbol}$$

In view of the following, this LZ algorithm is second-order compression-optimal

Second-order Source Coding Theorem [K 97]

If $\boldsymbol{X}$ has $\psi(k), \gamma(k) \to 0$ "fast enough," for any compression algorithm:

CLT There exist RVs Z_n such that

$$\liminf_{n\to\infty} \frac{\ell_n(X_1^n) - nH}{\sqrt{n}} - Z_n \geq 0, \quad \text{a.s.}$$

$$\text{and} \quad Z_n \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

LIL

$$\limsup_{n\to\infty} \frac{\ell_n(X_1^n) - nH}{\sqrt{2n \log\log n}} \geq \sigma \quad \text{a.s.}$$

Further Refinements

Same idea yields even more precise asymptotics for the waiting times W_n :

Functional CLT

Functional LIL

or even

$$\limsup_{n\to\infty} \frac{\sum_{k=1}^{n} |\log W_k - kH|}{\sqrt{2n^3 \log\log n}} = 3^{-1/2}\sigma \quad \text{a.s.}$$

Match Lengths and Duality

Recall template matching example:

template: $X_1\, X_2\, \cdots$

sequence: $Y_1\;Y_2\;Y_3\;\cdots\cdots\;Y_m$

Define

$L_m :=$ length of longest X_1^L appearing in Y_1^m

$$\underbrace{1\,0\,1\,1}\,0$$
$$0\,0\,1\,1\,1\,0\,\underbrace{1\,0\,1\,1}\,1\,0\,0\,1\,1$$

Duality: $\boldsymbol{L_m \geq n}$ iff $\boldsymbol{W_n \leq m}$

$\rightsquigarrow$ As in renewal theory, all results for W_n give corresponding results for L_m...

Dual Results for L_m

With H and σ^2 as before:

Theorem 2 [K 98]

Under the corresponding assumptions in Corollaries 1, 2:

LLN

$$\frac{L_m}{\log m} \to \frac{1}{H} \quad \text{a.s.}$$

CLT

$$\frac{L_m - \frac{\log m}{H}}{\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0, \sigma^2 H^{-3})$$

LIL

$$\limsup_{n\to\infty} \frac{L_m - \frac{\log m}{H}}{\sqrt{2 \log m \log \log \log m}} = \sigma H^{-3/2} \quad \text{a.s.}$$

Outline of Part II

Approximate Pattern Matching & Lossy Data Compression

- $\rightsquigarrow$ **Waiting times**
- $\rightsquigarrow$ **Strong approximation**: $W_n(D) \approx \frac{1}{Q(B(X_1^n,D))}$
- $\rightsquigarrow$ The **generalized AEP**
  - $\rightsquigarrow$ First-order asymptotics of $W_n(D)$
- $\rightsquigarrow$ Refinements of the generalized AEP
  - $\rightsquigarrow$ Second-order asymptotics of $W_n(D)$
- $\rightsquigarrow$ **Duality** and **match lengths**
  - $\rightsquigarrow$ Asymptotics for match lengths
- $\rightsquigarrow$ **A short course on lossy data compression**
  - $\rightsquigarrow$ Optimality, waiting times, and lossy LZ compression
  - $\rightsquigarrow$ Practical LZ compression

The General Setting

Let $\boldsymbol{X} = \{X_1, X_2, \ldots\}$, $\boldsymbol{Y} = \{Y_1, Y_2, \ldots\}$ be stationary, ergodic processes with distributions P, Q and values in the *general alphabets* $A, \hat{A}$, resp.

Fix an arbitrary distortion measure $d : A \times \hat{A} \to [0, \infty)$, let

$$d(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i), \qquad x_1^n \in A^n,\ y_1^n \in \hat{A}^n$$

and write $\boldsymbol{B(x_1^n, D) \;=\; \{y_1^n \in \hat{A}^n \;:\; d(x_1^n, y_1^n) \leq D\}}$

Define the **waiting time** $\boldsymbol{W_n(D) = \inf\{k \geq 1 \;:\; Y_k^{k+n-1} \in B(X_1^n, D)\}}$

$$X_1\, X_2\, \cdots\, X_n$$
$$Y_1\;\; Y_2\;\; Y_3\;\; \cdots\cdots\;\; Y_W\;\; Y_{W+1}\;\; \cdots\;\; Y_{W+n-1}\;\; \cdots$$

Problem: How does $W_n(D)$ behave as $n \to \infty$?

Strong Approximation: $W_n(D) \approx \frac{1}{Q(B(X_1^n, D))}$

Intuition

Again we expect W_n to be close to the reciprocal of the probability that the pattern X_1^n appears in $\boldsymbol{Y}$, within distortion D,i.e., $W_n \approx \frac{1}{Q(B(X_1^n,D))}$

Theorem 3: Strong Approximation [Dembo-K 99][Chi 01]

If $\boldsymbol{Y}$ has either $\psi(k) \to 0$ or $\sum_k \phi(k) < \infty$
and $Q(B(X_1^n, D)) > 0$ ev. a.s., then:

$$\log [W_n(D) Q(B(X_1^n, D))] = O(\log n) \quad \text{a.s.}$$

Therefore, $\log W_n(D) \approx -\log Q(B(X_1^n, D))$

But how does $-\log Q(B(X_1^n, D))$ behave?

Proof of Theorem 3

[LB] Under stationarity alone, same argument as before

[UB] For the upper bound in the general case, use the "second moment method" + a blocking *a la* Ibragimov.

In the special case where $\boldsymbol{X}, \boldsymbol{Y}$ are IID, fix a "good" realization x_1^∞, and let

$$S_n = \sum_{j=1}^{n^2/Q(B(x_1^n,D))} \mathbb{I}\{Y_{jn+1}^{(j+1)n} \in B(x_1^n, D)\}$$

so that

$$\Pr(\log[W_n(D)Q(B(X_1^n, D))] > 3\log n | X_1^n = x_1^n) \leq \Pr(S_n = 0) \leq \frac{\mathrm{Var}(S_n)}{(E[S_n])^2}.$$

By stationarity,

$$E[S_n] = \frac{n^2}{Q(B(x_1^n, D))} Q(B(x_1^n, D)) = n^2$$

and by independence, $\mathrm{Var}(S_n) = n^2$ too; therefore, as before,

$$\Pr(\log[W_n(D)Q(B(X_1^n, D))] > 3\log n | X_1^n = x_1^n) \leq 1/n^2$$

and the upper bound again follows from Borel-Cantelli □

Asymptotics of $-\log Q(B(X_1^n, D))$

Recall that $\log W_n(D) \approx -\log Q(B(X_1^n, D))$
but how does $-\log Q(B(X_1^n, D))$ behave?

Expand

$$
\begin{aligned}
Q(B(x_1^n, D)) &= \Pr\{d(X_1^n, Y_1^n) \le D \mid X_1^n = x_1^n\} \\
&= \Pr\left\{\frac{1}{n}\sum_{i=1}^n d(x_i, Y_i) \le D\right\}
\end{aligned}
$$

Intuition

Given $X_1^n = x_1^n$, the prob $Q(B(X_1^n, D))$ is a *large deviations probability* for the non-stationary process $\{(x_i, Y_i)\}$ (when D is small enough)

Assume

From now on that $d(\cdot,\cdot)$ is bounded
and that $D_{\min} := E[\operatorname{ess\,inf}_{Y_1} d(X_1, Y_1)] < D < D_{\mathrm{av}} := E[d(X_1, Y_1)]$

The Generalized AEP

Write

P_n, Q_n for the nth order marginals of $\boldsymbol{X}, \boldsymbol{Y}$, resp.
$H(\mu\|\nu) := \int \log(\frac{d\mu}{d\nu})d\mu$ for the relative entropy

Theorem 4: Generalized AEP [Dembo-K 99][Chi 01]

If $\boldsymbol{Y}$ has $\psi(k) \to 0$, then:

$$-\frac{1}{n} \log Q(B(X_1^n, D)) \to R(P, Q, D) \quad \text{a.s.}$$

where $R(P,Q,D) = \lim_n \frac{1}{n} R_n(P_n, Q_n, D)$ and $R_n(P_n, Q_n, D)$ is the "large deviations exponent"

$$R_n(P_n, Q_n, D) \;=\; \inf \int H(\nu_n(\cdot|x_1^n)\|Q_n(\cdot))dP_n(x_1^n)$$

where the infimum is over all measures ν_n on $A^n \times \hat{A}^n$ s.t. the A^n-marginal of ν is P_n, and $\int d(x_1^n, y_1^n)\, d\nu_n(x_1^n, y_1^n) \leq D$

Proof Outline of The Generalized AEP

Recall: $$Q(B(X_1^n, D)) = \Pr\left\{ \frac{1}{n}\sum_{i=1}^{n} d(X_i, Y_i) \leq D \,\middle|\, X_1^n \right\}$$

Step 1: Upper bound. Easy, *a la* Chernov bound

Step 2: Lower bound. Parameter dependent change of measure
+ blocking argument for the LLN of the twisted measure

Step 3: Identification of the rate function. Convex duality
+ blocking argument for regularity and convexity of $\Lambda^* = R$ □

First-Order Asymptotics for $W_n(D)$

Thm 3 $\Rightarrow$ $\qquad \log W_n(D) \; \approx \; -\log Q(B(X_1^n, D))$

Thm 4 $\Rightarrow$ $\qquad -\log Q(B(X_1^n, D)) \; \approx \; nR(P, Q, D)$

Combining, yields:

Corollary 3 [Luczak-Szpankowski 97][Yang-Kieffer 98][Dembo-K 99][Chi 01]

If $\boldsymbol{Y}$ has $\psi(k) \to 0$ then:

$$\frac{\log W_n(D)}{n} \to R(P, Q, D) \quad \text{a.s.}$$

Questions

Does this have any implications for compression? [Later]

Finer asymptotics? Where to start...?

Finer Large Deviations for $Q(B(X_1^n, D))$

Assume

From now on that $\boldsymbol{Y}$ is IID, $Q_n = Q^n$ for some distr Q on $\hat{A}$

Write

$\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for the empirical measure induced by X_1^n on A
$R(\hat{P}_n) = R_1(\hat{P}_n, Q, D)$ and $R(P) = R_1(P_1, Q, D)$

Theorem 5: Large Deviations [Dembo-K 99][Yang-Zhang 99]

If $\boldsymbol{Y}$ is IID:

$$-\log Q(B(X_1^n, D)) - nR(\hat{P}_n) \;=\; \frac{1}{2}\log n + O(1) \quad \text{a.s.}$$

Proof: Upper bound: Easy argument *a la* Chernov bound.
Lower bound: parameter dependent change of measure + CLT
a la Bahadur-Rao, with Berry-Esséen bound □

“Differentiability” of $R(\cdot)$

So far we've shown

$$\log W_n(D) \approx -\log Q(B(X_1^n, D)) \approx nR(\hat{P}_n)$$

probability $\rightsquigarrow$ analysis!

Theorem 6: Uniform Approximation [Dembo-K 99, 03]

If $\boldsymbol{X}$ has $\phi(k) \to 0$ fast enough and $\boldsymbol{Y}$ is IID, then,
for an explicitly identified, zero-mean $f : A \to \mathbb{R}$:

$$nR(\hat{P}_n) \;=\; nR(P) \;+\; \sum_{i=1}^{n} f(X_i) \;+\; O(\log\log n) \quad \text{a.s.}$$

Combining Theorems 3, 5 and 6:

$$\log W_n(D) \approx -\log Q(B(X_1^n, D)) \approx nR(\hat{P}_n) \approx nR(P) + \sum_{i=1}^{n} f(X_i)$$

i.e. $\boldsymbol{\log W_n(D) - nR(P) \;\approx\; \sum_{i=1}^n f(X_i)}$

Proof Outline

Letting $$\Lambda(x;\lambda) = \log E\left[e^{\lambda d(Y_1,x)}\right], \qquad x \in A,\ \lambda \in \mathbb{R}$$

we note that $R(P)$ can be expressed

$$R(P) \;=\; \sup_{\lambda \le 0}\Big[\lambda D - E[\Lambda(X_1;\lambda)]\Big] \;=\; \lambda^* D - E[\Lambda(X_1;\lambda^*)]$$

where $\lambda^* < 0$ is s.t.

$$\frac{d}{d\lambda}E[\Lambda(X_1;\lambda)]\bigg|_{\lambda=\lambda^*} = D$$

For n large enough, the difference $n[R(P) - R(\hat{P}_n)]$ can be expressed as a supremum over a small neighborhood around λ^*, in terms of $E_{\hat{P}_n}[\Lambda(X;\lambda)]$ alone, which is itself an IID partial sum.

The uniform LLN then yields the result, upon defining:

$$f(\cdot) = -\left(\Lambda(\cdot;\lambda^*) - E[\Lambda(X_1;\lambda^*)]\right)$$

NOTE: f depends on all of P, Q, D □

Second-Order Asymptotics for $W_n(D)$

Recall

$\rightsquigarrow$ $R(P)$ can be expressed as $\lambda^* D - E[\Lambda(X_1; \lambda^*)]$

$\rightsquigarrow$ Theorems 3,5,6 $\Rightarrow$ $\log W_n(D) - nR(P) \approx \sum_{i=1}^n f(X_i)$

Define the D, Q-**coding variance** of $\boldsymbol{X}$ as:

$$\sigma^2 = \sigma^2_{P,Q,D} = \mathrm{Var}\big(\Lambda(X_1, \lambda^*)\big) = \mathrm{Var}(f(X_1))$$

Combining the above approx with the CLT/LIL:

Corollary 4 [Dembo-K 99, 03]

If $\boldsymbol{X}$ has $\phi(k) \to 0$ fast enough and $\boldsymbol{Y}$ is IID, then:

CLT $$\frac{\log W_n(D) - nR(P)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

LIL $$\limsup_{n\to\infty} \frac{\log W_n(D) - nR(P)}{\sqrt{2n \log\log n}} = \sigma \quad \text{a.s.}$$

Approximate Match Lengths and Duality

Template matching example:

template: $X_1 \, X_2 \, \cdots$

sequence: $Y_1 \; Y_2 \; Y_3 \; \cdots\cdots \; Y_m$

Define

$$\boldsymbol{L_m(D)} \;:=\; \text{length of longest } X_1^L \text{ appearing in } Y_1^m \text{ with distortion} \leq D$$
$$= \max\{L \geq 1 \;:\; Y_j^{j+L-1} \in B(X_1^L, D) \text{ for some } 1 \leq j \leq m - L + 1\}$$

E.g. $D =$ "agree in $\geq 70\%$ of all positions", $m = 15$, $L_m(D) = 4$

$$\underbrace{1\,0\,1\,1}\,0$$
$$0\,0\,1\,1\,\underbrace{1\,0\,0\,1}\,1\,0\,0\,1\,0\,0\,1$$

Duality?

Here: $L_m(D) \geq n \quad \Leftarrow \quad W_n(D) \leq m$ but NOT conversely!

Modified Duality and Asymptotics for $L_m(D)$

Modified duality: $\boldsymbol{L_m(D) \geq n}$ iff $\boldsymbol{\inf_{k \geq n} W_k(D) \leq m}$

Again, all results for $W_n(D)$ give corresponding results for $L_m(D)$
but we have to work for them!

Theorem 7 [Dembo-K 99, 03]

If $\boldsymbol{Y}$ is IID, then with $R(P) = R_1(P_1, Q, D)$ and $\sigma^2 = \sigma^2_{P,Q,D}$ as before:

LLN $$\frac{L_m(D)}{\log m} \to \frac{1}{R(P)} \quad \text{a.s.}$$

If, in addition $\boldsymbol{X}$ has $\phi(k) \to 0$ fast enough :

CLT $$\frac{L_m(D) - \frac{\log m}{R(P)}}{\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0, \sigma^2 R(P)^{-3})$$

LIL $$\limsup_{n \to \infty} \frac{L_m(D) - \frac{\log m}{R(P)}}{\sqrt{2 \log m \log \log \log m}} = \sigma R(P)^{-3/2} \quad \text{a.s.}$$

Outline of Part II revisited

Approximate Pattern Matching & Lossy Data Compression

- *Waiting times*
- *Strong approximation:* $W_n(D) \approx \frac{1}{Q(B(X_1^n,D))}$
- *The generalized AEP*
  - *First-order asymptotics of* $W_n(D)$
- *Refinements of the generalized AEP*
  - *Second-order asymptotics of* $W_n(D)$
- *Duality and match lengths*
  - *Asymptotics for match lengths*
- **A short course on lossy data compression**
  - Optimality, waiting times, and lossy LZ compression
  - Practical LZ compression

Lossy Compression in More Detail

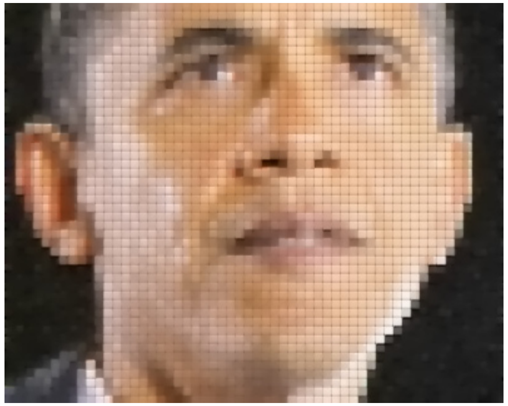
Data: $X_1^n = X_1, X_2, \ldots, X_n$ IID $\sim P_n = P^n$ on A^n

Quantizer: $\mathcal{K}_n : A^n \to$ codebook $B_n \subset \hat{A}^n$

Encoder: $\mathcal{E}_n : B_n \to \{0,1\}^*$ "uniquely decodable"

Length function: $\ell_n(X_1^n) =$ length of $\mathcal{E}_n(\mathcal{K}_n(X_1^n))$ bits

$\xrightarrow{\mathcal{K}_n}$



$\xrightarrow{\mathcal{E}_n}$ 0010111010110 101101000 . . .

$\xleftarrow{\mathcal{E}_n^{-1}}$

Distortion requirement

With a distortion measure $d(x_1^n, y_1^n)$ as before

the code $(\mathcal{K}_n, \mathcal{E}_n, \ell_n)$ operates at distortion level $D > 0$

if $d(x_1^n, \mathcal{K}_n(x_1^n)) \leq D$ for all x_1^n

Fundamental Limits of Lossy Compression

Question

For a code $(\mathcal{K}_n, \mathcal{E}_n, \ell_n)$ operating at distortion level D
on data generated by the IID "source" $\boldsymbol{X} = \{X_1, X_2, \ldots\}$
what is the best (=smallest) achievable compression rate,

$$\textbf{compression rate} := \lim_{n\to\infty} \frac{\ell_n(X_1)}{n} \quad \text{bits/symbol ?}$$

Recall

For any prob distr Q on $\hat{A}$: $R_1(P,Q,D) = \inf \int H(\nu(\cdot|x)\|Q(\cdot))dP(x)$
where the infimum is over all measures ν on $A\times\hat{A}$ s.t.
the A-marginal of ν is P and $\int d(x,y)\, d\nu(x,y) \le D$

Answer The optimal compression rate is given
by the **rate-distortion function** of $\boldsymbol{X}$:

$$R(D) := R_1(P,Q^*,D) = \inf_Q R_1(P,Q,D)$$

Fundamental Limits of Lossy Compression

Fix

IID random source $\boldsymbol{X}$ with distr P on the source alphabet A

Optimal distr Q^* on the reproduction alphabet $\hat{A}$

Single-letter distortion measure $d(x_1^n, y_1^n)$ as before

Distortion values D in the interesting range $D_{\min} < D < D_{\text{av}}$

Pointwise Source Coding Theorem [Kieffer 91][K 00]

For any code $(\mathcal{K}_n, \mathcal{E}_n, \ell_n)$ operating at distortion level D:

$$\liminf_{n\to\infty} \frac{\ell_n(X_1^n)}{n} \geq R(D) \quad \text{bits/symbol, a.s.}$$

$\rightsquigarrow$ Can we/How can we achieve this lower bound?!

Idealized Lossy LZ Compression

Describe X_1^n as $W_n(D)$, as before:

message: $X_1\, X_2 \cdots X_n$

database: $Y_1\ Y_2\ Y_3\ \cdots\ Y_W\ \cdots\ Y_{W+n-1}\ \cdots$ **IID** $\sim \boldsymbol{Q}$

In view of Corollary 3, the **rate** of this algorithm is:

$$\text{compression rate} \approx \frac{\log W_n(D)}{n} \to R_1(P, Q, D) \quad \text{bits/symbol, a.s.}$$

In particular, if we take $Q = Q^*$ as in the definition of the rate-distortion function, the compression rate is *optimal*:

$$\textbf{compression rate} \approx \boldsymbol{\frac{W_n(D)}{n}} \to \boldsymbol{R(D)} \quad \text{bits/symbol, a.s.}$$

$\leadsto$ How about finer optimality properties?

[$\leadsto$ What if we don't know Q^* ?]

Finer Compression Performance

Idealized lossy LZ algorithm with $Q = Q^*$

Given $\boldsymbol{X}, \boldsymbol{Y}$ with distr P, Q^*, resp., and $D > 0$,
Theorems 3,5 and 6 $\Rightarrow$ there exists a zero-mean, bounded $f : A \longrightarrow \mathbb{R}$ s.t.

$$\begin{aligned} \mathsf{LZ}_n(X_1^n) &= \log W_n(D) + O(\log n) \\ &= nR(D) + \sum_{i=1}^{n} f(X_i) + O(\log n) \quad \text{bits, a.s.} \end{aligned}$$

Finer Compression Performance

Idealized lossy LZ algorithm with $Q = Q^*$

Given $\boldsymbol{X}, \boldsymbol{Y}$ with distr P, Q^*, resp., and $D > 0$,
Theorems 3,5 and 6 $\Rightarrow$ there exists a zero-mean, bounded $f : A \to \mathbb{R}$ s.t.

$$\begin{aligned} \mathsf{LZ}_n(X_1^n) &= \log W_n(D) \;+\; O(\log n) \\ &= nR(D) \;+\; \sum_{i=1}^n f(X_i) \;+\; O(\log n) \quad \text{bits, a.s.} \end{aligned}$$

$\rightsquigarrow$ In view of the following, this is optimal up to a very fine scale!

Second-order Source Coding Theorem [K 00]

For ANY seq of codes $(\mathcal{K}_n, \mathcal{E}_n, \ell_n)$ operating at distortion level D

$$\ell_n(X_1^n) \;\geq\; nR(D) + \sum_{i=1}^n f(X_i) - 2\log n \quad \text{bits, ev. a.s.}$$

Properties of Lossless LZ Schemes

Lossless Lempel-Ziv schemes are *extremely successful* in practice. Why?

A. Compression Optimality/Universality

Can be deduced from studying the "idealized" scheme

B. Convergence speed: Bad!

$$O\left(\frac{\log\log m}{\log m}\right)$$

C. Complexity/Implementation: Superb

- efficient string matching algorithms
- the algorithm is *tunable*

Aside: The AEP and its Generalizations

Let $\boldsymbol{X} \sim P$ be stationary ergodic

The classical AEP

If A is finite:

$$-\frac{1}{n}\log P_n(X_1^n) \;\to\; H(P) \quad \text{a.s.}$$

Barron's extension

If $Q_n = Q^n$ is IID on A:

$$-\frac{1}{n}\log \frac{dP_n}{dQ^n}(X_1^n) \;\to\; -H(P\|Q) \quad \text{a.s.}$$

Theorem 4

If $Q_n = Q^n$ is IID on $\hat{A}$ and $d(\cdot,\cdot)$ is bounded:

$$-\frac{1}{n}\log Q^n(B(X_1^n, D)) \;\to\; R(P,Q,D) \quad \text{a.s.}$$

Densities vs Balls?

Let $\boldsymbol{X} \sim P$ be IID, Q be an IID measure on $\hat{A}$ with $P \ll Q$ and $d(\cdot,\cdot)$ be bounded. With "probability one":

$$
\begin{aligned}
-H(P\|Q) &\leftarrow -\frac{1}{n}\log\frac{dP^n}{dQ^n}(X_1^n) \\
&\leftarrow -\frac{1}{n}\log\frac{P^n(B(X_1^n,D))}{Q^n(B(X_1^n,D))} \\
&= -\frac{1}{n}\log P^n(B(X_1^n,D)) \;+\frac{1}{n}\log Q^n(B(X_1^n,D)) \\
&\to R(P,P,D) - R(P,Q,D) \\
&\to -H(P\|Q)
\end{aligned}
$$

Further Extensions, Generalizations

Applications

- $\rightsquigarrow$ Lossy Minimun Description Length (MDL) compression
- $\rightsquigarrow$ Entropy estimation
- $\rightsquigarrow$ Realistic lossy data compression

Theory

- $\rightsquigarrow$ Sphere covering and measure concentration converses
- $\rightsquigarrow$ Error exponents
- $\rightsquigarrow$ Uniform generalized AEP and refinements
- $\rightsquigarrow$ Random fields
- $\rightsquigarrow$ Small balls and the Brin-Katok theorem