Compression and Estimation Over Large Alphabets

Alon Orlitsky

Narayana P. Santhanam

Krishnamurthy Viswanathan

Junan Zhang

UCSD

Compression [Sh 48]

Setup: $\mathcal{A}$ — alphabet

p — p.d. over $\mathcal{A}^n$

random sequence $\sim p$

$L_q \stackrel{\text{def}}{=}$ expected # bits of encoder q

Question: $L \stackrel{\text{def}}{=} \min_q L_q = ?$

Answer: $L \approx H(p)$

Problem: p not known

Solution: Universal compression

Univeral Compression [Sh 48] [Fi 66, Da 73]

Setup: $\mathcal{A}$ — alphabet

$\mathcal{P}$ — collection of p.d.'s over $\mathcal{A}^n$

random sequence $\sim p \in \mathcal{P}$ (unknown)

$L_q \stackrel{\text{def}}{=}$ expected # bits of encoder q

Redundancy: $R_q \stackrel{\text{def}}{=} \max_p L_q - H(p)$

Question: $R \stackrel{\text{def}}{=} \min_q \ R_q =?$

if $R/n \to 0$, Universally Compressible

Answer: iid, markov, cxt tree, stnr ergd — UC

iid: $R \approx \frac{1}{2}(|\mathcal{A}| - 1)\log n$

Problem: $|\mathcal{A}| \approx$ or $> n$ (text, images)

[Kief. 78]: As $|\mathcal{A}| \to \infty$, $R/n \to \infty$

Solution: Several

Solutions

Theoretical: Constrain distributions

Monotone: [Els 75], [GPM 94], [FSW 02]

Bounded moments: [UK 02,03]

Others: [YJ 00], [HY 03]

Concern: May not apply

Practical: Convert to bits

Lempel Ziv

Context-tree weighting

Concern: May lose context

Change the question

Why ∞?

Alphabet: $\mathcal{A} \stackrel{\text{def}}{=} \mathbb{N}$

Collection: $\mathcal{P} \stackrel{\text{def}}{=} \{p_k : k \in \mathbb{N}\}$

p_k: constant-k distribution

$$p_k(\overline{x}) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \overline{x} = k \ldots k \\ 0 & \text{otherwise} \end{cases}$$

If k is known: $H(p_k) = 0$

0 bits

Universally: must describe k

∞ bits (for worst k)

$R = \infty$

Conclusion: Describe elts & pattern separately

Patterns

Replace each symbol by its order of appearance

Sequence:	a	b	r	a	c	a	d	a	b	r	a
Pattern:	1	2	3	1	4	1	5	1	2	3	1

Convey

pattern: 12314151231

dictionary:

1	2	3	4	5
a	b	r	c	d

Compress pattern and dictionary separately

Related application (PPM): [ÅSS 97]

Main result

Patterns of iid distributions over any alphabet (large, infinite, uncountably infinite, unknown) can be universally compressed (sequentially and efficiently).

Details

Block: $R \le \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}$

Sequential (super-poly): $R \le \left(\frac{4\pi}{3(2-\sqrt{2})}\right)\sqrt{n}$

Sequential (linear): $R \le 10\, n^{2/3}$

In all: $R/n \to 0$

Additional results

R_m: redundancy for m-symbol patterns

Identical technique

For $m \le o(n^{1/3})$,

$$R_m \le \log\left(\binom{n-1}{m-1}\frac{1}{m!}\right)$$

Similar average-problem when alphabet assumed to contain no unseen symbols consequently considered by [Sh 03]

Proof technique

Compression = probability estimation

Estimate distributions over large alphabets

Considered by I.J. Good and A. Turing

Good-Turing estimator is good, not optimal

View as set partitioning

Construct optimal estimators

Use results by Hardy and Ramanujan

Probability estimation

Safari preparation

Observe sample of animals

3 giraffes, 1 hippopotamus, 2 elephants

Probability estimation?

Species	Prob
giraffe	3/6
hippo	1/6
elephant	2/6

Problem?

Lions!

Laplace estimator

Add one, including to new

3+1 giraffes, 1+1 hippopotamus,

2+1 elephants, 0+1 new

Species	Prob
giraffe	4/10
hippo	2/10
elephant	3/10
new	1/10

Many add-constant variations

Krichevsky-Trofimov estimator

Add half

Achieves Jeffreys' prior

Best for fixed alphabet, length $\to \infty$

Are add-constant estimators good?

DNA

n samples (n large)

All different

Probability estimation?

For each observed: $1 + 1 = 2$

For new: $0 + 1 = 1$

Sample	Probability
observed	$2/(2n+1)$
new	$1/(2n+1)$

Problem?

$P(\text{new}) = 1/(2n+1) \approx 0$

$P(\text{observed}) = 2n/(2n+1) \approx 1$

Opposite more accurate

Good-Turing problem

Enigma cipher

Captured German book of keys

Had previous decryptions

Looked for distribution of key pages

Similar as # pages large compared to data

Good-Turing estimator

Surprising and complicated

Works well for infrequent elements

Used in a variety of applications

Suboptimal for frequent elements

Modifications: empirical for frequent elements

Several explanations

Some evaluations

Evaluation

Observe sequence:

$$x_1, x_2, x_3, \ldots$$

Successively estimate prob given previous:

$$q(x_i | x_1^{i-1})$$

Assign probability to whole sequence:

$$q(x_1^n) = \prod_{i=1}^{n} q(x_i | x_1^{i-1})$$

Compare to highest possible $p(x_1^n)$

Cf. compression, online algorithms/learning

Precise definitions require patterns

Pattern of a sequence

Replace symbol by order of appearance

g,h,g,e,e,g

giraffe — 1, hippo — 2, elephant — 3

1,2,1,3,3,1

Can enumerate, assign probabilities

Sequence = pattern

Example: q_{+1}

Sequence: ghge $\rightarrow$ NNgN

$$q_{+1}(ghge) = q_{+1}(N) \cdot q_{+1}(N|g) \cdot q_{+1}(g|gh) \cdot q_{+1}(N|ghg)$$
$$= \frac{1}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{6}$$
$$= \frac{1}{45}$$

Pattern: 1213

$$q_{+1}(1213) = q_{+1}(1) \cdot q_{+1}(2|1) \cdot q_{+1}(1|12) \cdot q_{+1}(3|121)$$
$$= \frac{1}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{6}$$
$$= \frac{1}{45}$$

Patterns

Strings of positive ingeters

First appearance of $i > 2$ follows that of $i - 1$

Patterns: 1, 11, 12, 121, 122, 123

Not patterns: 2, 21, 132

Ψ^n — length-n patterns

Pattern probability

$\mathcal{A}$ — alphabet

p — distribution over $\mathcal{A}$

$\overline{\psi}$ — pattern in Ψ^n

$p^{\Psi}(\overline{\psi}) \stackrel{\text{def}}{=} p\{\overline{x} \in \mathcal{A}^n \text{ with pattern } \overline{\psi}\}$

Example

$\mathcal{A} = \{a, b\}$

$p(a) = \alpha$, $p(b) = \overline{\alpha}$

$p^{\Psi}(11) = p\{aa, bb\} = \alpha^2 + \overline{\alpha}^2$

$p^{\Psi}(12) = p\{ab, ba\} = 2\alpha\overline{\alpha}$

Maximum pattern probability

Highest probability of pattern

$$\hat{p}^{\Psi}(\overline{\psi}) \stackrel{\text{def}}{=} \max_{p} \; p^{\Psi}(\overline{\psi})$$

Examples

$\hat{p}^{\Psi}(11) = 1$ [constant distributions]

$\hat{p}^{\Psi}(12) = 1$ [continuous distributions]

In general, difficult

$\hat{p}^{\Psi}(112) = 1/4$ [$p(a) = p(b) = 1/2$]

$\hat{p}^{\Psi}(1123) = 12/125$ [$p(a) = ... = p(e) = 1/5$]

General results

Obtained several results

m: # symbols appearing

μ_i: # times i appears

μ_{min}, μ_{max}: smallest, largest μ_i

Example: 111223, $\mu_1 = 3$, $\mu_{\text{min}} = 1$, $\mu_{\text{max}} = 3$

$\widehat{k}$: # symbols in maximizing distribution

Upper bound: $\widehat{k} \le m + \frac{m-1}{2^{\mu_{\text{min}}}-2}$

Lower bound: $\widehat{k} \ge m - 1 + \frac{\sum 2^{-\mu_i} - 2^{-\mu_{\text{max}}}}{2^{\mu_{\text{max}}}-2}$

Attenuation

Attenuation of q for ψ_1^n

$$R(q, \psi_1^n) \stackrel{\text{def}}{=} \frac{\widehat{p}^{\Psi}(\psi_1^n)}{q(\psi_1^n)}$$

Worst-case sequence attenuation of q (n symb)

$$R_n(q) \stackrel{\text{def}}{=} \max_{\psi_1^n} R(q, \psi_1^n)$$

Worst-case attenuation of q

$$R^*(q) \stackrel{\text{def}}{=} \limsup_{n \to \infty} \left(R_n(q)\right)^{1/n}$$

Laplace estimator

Pattern: $123\ldots n$

$$\widehat{p}^{\Psi}(123\ldots n) = 1$$

$$q_{+1}(123\ldots n) = \frac{1}{1\cdot 3\cdot\ldots\cdot(2n+1)}$$

$$R_n(q_{+1}) \geq \frac{\widehat{p}^{\Psi}(123...n)}{q_{+1}(123...n)} = 1\cdot 3\cdots(2n+1) \approx \left(\frac{2n}{e}\right)^n$$

$$R^*(q_{+1}) = \limsup_{n\to\infty} \frac{2n}{e} = \infty$$

Good-Turing estimator

Multiplicity of $\psi \in \mathbb{Z}^+$ in ψ_1^n

$$\mu_\psi \stackrel{\text{def}}{=} |\{1 \le i \le n : \psi_i = \psi\}|$$

Prevalence of multiplicity μ in ψ_1^n

$$\varphi_\mu \stackrel{\text{def}}{=} |\{\psi : \mu_\psi = \mu\}|$$

Increased multiplicity

$$r \stackrel{\text{def}}{=} \mu_{\psi_{n+1}}$$

Good-Turing estimator

$$q(\psi_{n+1}|\psi_1^n) = \begin{cases} \frac{\varphi'_1}{n}, & r = 0 \\ \frac{r+1}{n}\frac{\varphi'_{r+1}}{\varphi'_r}, & r \ge 1 \end{cases}$$

φ'_μ — smoothed version of φ_μ

Performance of Good Turing

Analyzed three versions

Simple: $1.39 \leq R^*(q_{\mathsf{sgt}}) \leq 2$

Church-Gale: experimatnatally > 1

Common-sense: same

Diminishing attenuation

$c[n] = \lceil n^{1/3} \rceil$

$f_{c[n]}(\varphi) \stackrel{\text{def}}{=} \max(\varphi, c[n])$

$$q_{\frac{1}{3}}(\psi_{n+1}|\psi_1^n) = \frac{1}{S_{c[n]}(\psi_1^n)} \cdot \begin{cases} f_{c[n]}(\varphi_1 + 1) & r = 0 \\ (r+1)\frac{f_{c[n]}(\varphi_{r+1}+1)}{f_{c[n]}(\varphi_r)} & r > 0 \end{cases}$$

$S_{c[n]}(\psi_1^n)$ is a normalization factor

$R_n(q_{\frac{1}{3}}) \leq 2^{\mathcal{O}(n^{2/3})}$, constant ≤ 10

$R^*(q_{\frac{1}{3}}) \leq 2^{\mathcal{O}(n^{-1/3})} \to 1$

Proof: Potential functions

Low-attenuation estimator

t_n — largest power of 2 that is $\le n$

$$\Psi^{2t_n}(\psi_1^n) \stackrel{\text{def}}{=} \{y_1^{2t_n} \in \Psi^{2t_n} : y_1^n = \psi_1^n\}$$

$$\tilde{p}(\psi_1^n) \stackrel{\text{def}}{=} \frac{\prod_{\mu=1}^n \mu!^{\varphi_\mu} \varphi_\mu!}{n!}$$

$$q_{\frac{1}{2}}(\psi_{n+1}|\psi_1^n) = \frac{\sum_{\overline{y} \in \Psi^{2t_n}(\psi_1^{n+1})} \tilde{p}(\overline{y})}{\sum_{\overline{y} \in \Psi^{2t_n}(\psi_1^n)} \tilde{p}(\overline{y})}$$

$$R_n(q_{\frac{1}{2}}) \le \exp\left(\frac{4\pi}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}\right)$$

$$R^*(q_{\frac{1}{2}}) \le \exp\left(\frac{4\pi}{\sqrt{3}(2-\sqrt{2})\sqrt{n}}\right) \to 1$$

Proof: Integer partitions, Hardy-Ramanujan

Lower bound

$R_n(q_{\frac{1}{3}}) \leq 2^{\mathcal{O}(n^{2/3})}$

$R_n(q_{\frac{1}{2}}) \leq 2^{\mathcal{O}(n^{1/2})}$

For any q,

$$R_n(q) \geq 2^{\Omega(n^{1/3})}$$

Proof: Generating functions and Hayman's thm

"Test"

$aaaa\ldots$ $\quad q(\text{new}) = \Theta(\frac{1}{n})$

$abab\ldots$ $\quad q(\text{new}) = \Theta(\frac{1}{n})$

$abcd\ldots$ $\quad q(\text{new}) = 1 - \Theta(\frac{1}{n^{2/3}})$

$aabbcc\ldots$ $\quad q(\text{new}) =$ Possible guess: 1/2

$q(\text{new}) = 1/4$ after even, 0 after odd

"Explanation": likely $|\alpha\beta| = 0.62n$

$p(\text{new}) \approx 0.2$