Lossless Data Compression on Unbounded Integers

Patrick Peng, Samuel Tian, Alvin Xu
Mathworks at Texas State
Mentored by: Dr. Dan Tamir, *Texas State*

31 July 2019

TEXAS STATE UNIVERSITY

The rising STAR *of Texas*

Objectives

- Formulate compression algorithms that can be applied to the set of unbounded integers using symbol compression techniques
- Prove the universality and asymptotic optimality of these algorithms
- Demonstrate the practical advantages of using these algorithms on sets of integers

Definitions

Definition

A *lossless data compression algorithm* is one that encodes input into codewords that requires less storage and loses no information.

Definition

Entropy is a theoretical lower bound on the performance of a lossless compression algorithm. It is given by a function of the PMF (probability mass function) as shown:

$$H(P) = -\sum_i P_i \log P_i$$

Definitions

Definition

A *fixed-length code* (FLC) is one that uses a fixed number of bits to represent a symbol or integer - even if some bits are not necessary, they will be padded.

Definition

A *variable-length code* (VLC) is one that uses a variable number of bits to represent a symbol or integer, following the heuristic that higher-probability symbols have shorter codes and vice versa.

Definitions

Definition

A code is *universal* if the average code length is bounded by a multiple of the entropy, so

$$\frac{E_P(L_\rho)}{\max\{1, H(P)\}} \leq K_\rho.$$

Definition

A code is *asymptotically optimal* if the ratio between the average code length and the entropy is bounded by a function of entropy that converges to 1, so

$$\frac{E_P(L_\rho)}{\max\{1, H(P)\}} \leq R_\rho(H(P)) \leq K_\rho$$

with

$$\lim_{H \to \infty} R_\rho(H) = 1.$$

Definitions

Definition

In *Elias*-γ coding for integers, if the length of the binary representation of an integer x is N bits, then we prepend $N-1$ zeroes with the binary representation of that integer.

$$17 = \underbrace{0\ 0\ 0\ 0}_{\text{N-1 zeroes}}\ \overbrace{1\ 0\ 0\ 0\ 1}^{\text{x in binary}}$$

Definitions

Definition

In *Elias-δ* coding for integers, if the length of the binary representation of an integer x is N bits, then we prepend the Elias-γ encoding of N to the last $N-1$ bits of the binary representation of x.

$$17 = \underbrace{0\ 0\ 1\ 0\ 1}_{\gamma(N)} \overbrace{0\ 0\ 0\ 1}^{N-1 \text{ bits}}$$

Definitions

Definition

In *Shannon-Fano-Elias (SFE)* coding for a fixed set of symbols, where a source probability distribution is known, codewords for symbols are generated as the first $\left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1$ bits to the right of the decimal point in the binary form of

$$P(x) = \sum_{x_0 < x} P(x_0) + \frac{1}{2} P(x)$$

.

- Useful for encryption compared to other compression techniques due to different codewords for different symbol permutations
- The average code length is bounded between $1 + H(P)$ and $2 + H(P)$
- Only usable when source probability distribution is known

Results

Theorem

Given initially empty set AT (already transmitted), a symbol NYT (not yet transmitted) with constant probability $\frac{1}{n+1}$, where n is the number of symbols processed so far, and the FLC $R(x)$ for the symbol, Adaptive-SFE is defined as follows,

1. *Encode the symbol as $E(x)$, where:*

$$E(x) = \begin{cases} SFE(NYT) + R(x) & x \notin AT \\ SFE(x) & x \in AT \end{cases}$$

2. *Update AT and our probability distribution with the symbol in consideration.*

- Does not require a predetermined PMF

Results

Theorem

Canonical δ-SFE is defined as follows:

$$C\delta\text{-}SFE(x) = \begin{cases} SFE(NYT) + \delta(x) & x \notin AT \\ SFE(x) & x \in AT \end{cases}$$

The probability of the NYT element is kept constant at $\frac{1}{n+1}$ where n is the number of symbols processed.

Theorem

Canonical δ-SFE is universal, but not asymptotically optimal.

Results

Theorem

Increment δ-SFE follows the same encoding scheme of canonical δ-SFE, so

$$I\delta\text{-}SFE(x) = \begin{cases} SFE(NYT) + \delta(x) & x \notin AT \\ SFE(x) & x \in AT \end{cases}$$

However, the probability of the NYT element is incremented by a constant k for every occurrence of a new symbol.

Theorem

Increment δ-SFE is both universal and asymptotically optimal.

Results

Theorem

Flagged δ-SFE uses a binary flag to differentiate between occurrences of new or repeated symbols and is defined as follows:

$$F\delta\text{-}SFE(x) = \begin{cases} \delta(x+1) & x \notin AT \\ 1 + SFE(x) & x \in AT \end{cases}$$

Theorem

Flagged δ-SFE is both universal and asymptotically optimal.

γ-RNS is a combination of the Elias-γ encoding scheme and the Residue Number System.

Theorem

Given an integer j, the primes $p_1 = 2, p_2 = 3, ...p_k$ are used such that k is the minimum integer that satisfies $\prod_{n=1}^{k} p_k > j$. The residues $j \mod p_n$ are then calculated and expressed in binary. To ensure uniform lengths, we prepend 0s to the binary representation of each residue r_n until the length of each binary representation reaches $\lceil \log_2 p_n \rceil$. Then, we prepend $k - 1$ 0s and a 1 to the start of the bitstring.

Results

δ-RNS is a combination of the Elias-δ encoding scheme and the Residue Number System.

Theorem

Taking the binary representation of the residues (generated using the same method as for γ-RNS), we instead prepend the binary representation k of the length of the residues + 1. Then, we prepend the unary representation in 0s of the length of k - 1.

Experimental Setup

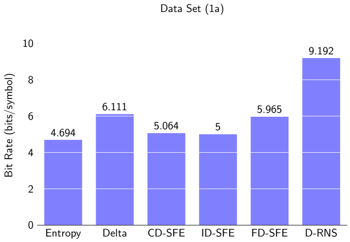
Bit Rate

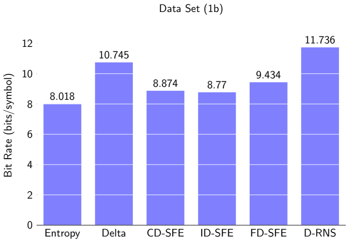
- All data sets consist of 10,000 integers following some distribution
- Data Set 1: geometric PMF distributions with $p = 0.1, 0.01$
- Data Set 2: Poisson distribution with $\lambda = 128$
- Data Set 3: pseudo-randomly generated integers from 1 to 1000
- Compare with entropy and Elias-δ

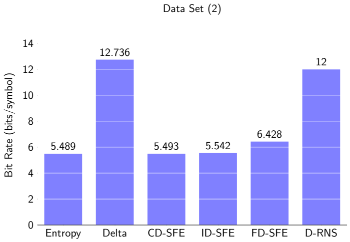
δ-RNS independent testing

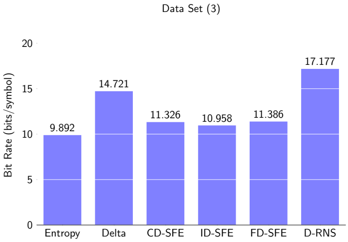
- δ-RNS creates codes of individual symbols
- Depicts performance on integers up to 2^{32}
- Compare with γ-RNS and Elias-δ

Results

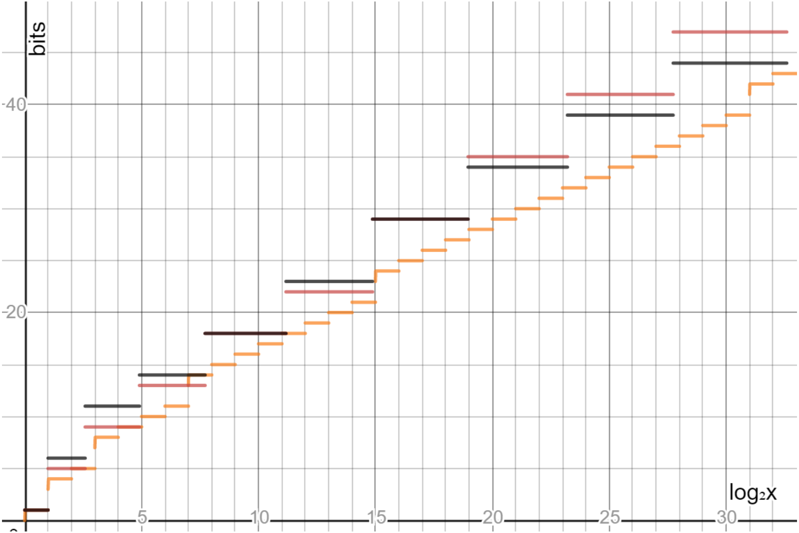

Data Set (1a)
Bit Rate (bits/symbol)
0
2
4
6
8
10
4.694
6.111
5.064
5
5.965
9.192
Entropy
Delta
CD-SFE
ID-SFE
FD-SFE
D-RNS


Data Set (1b)
Bit Rate (bits/symbol)
0
2
4
6
8
10
12
8.018
10.745
8.874
8.77
9.434
11.736
Entropy
Delta
CD-SFE
ID-SFE
FD-SFE
D-RNS


Data Set (2)
Bit Rate (bits/symbol)
0
2
4
6
8
10
12
14
5.489
12.736
5.493
5.542
6.428
12
Entropy
Delta
CD-SFE
ID-SFE
FD-SFE
D-RNS


Data Set (3)
Bit Rate (bits/symbol)
0
5
10
15
20
9.892
14.721
11.326
10.958
11.386
17.177
Entropy
Delta
CD-SFE
ID-SFE
FD-SFE
D-RNS

Results



Conclusion and Future Work

- δ-SFE versions operate at high compression ratios near entropy for unbounded integers
- Applications of prime numbers to compression yield high efficiency gains, especially for large integers
- Investigate similar applications of Elias-δ to other symbol compression techniques such as arithmetic coding
- Applications
  - Compression of inverse indexes used by databases and search engines
  - Extension of super-exponential decryption time of SFE to infinite alphabets
  - Efficient operations on large integers represented in δ-RNS

Acknowledgements

We would like to thank:

Dr. Dan Tamir for his support and encouragement,

2019 HSMC Mathworks for support,

Mathworks research coordinators Jenny Lu, Danika Luo, and Eric Wu, for their feedback,

and the audience for their kind attention.

TEXAS STATE UNIVERSITY

The rising STAR *of Texas*