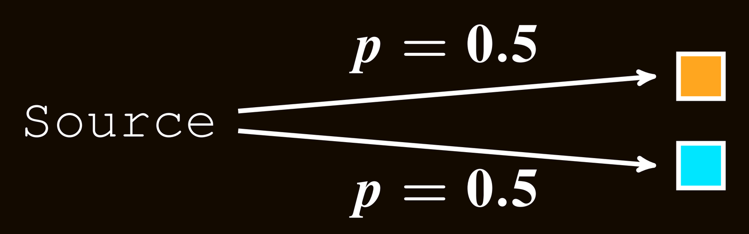
Compressing multisets using tries

Vincent Gripon, Michael Rabbat, Vitaly Skachek and Warren J. Gross (Télécom Bretagne, Brest - McGill University, Montréal)

1. Motivation



How to efficiently encode binary words produced by the source while disregarding order?

2. About order and sequences

2.1. Toy example

Encode sequence :

2.1.1. With order

Cannot do better!

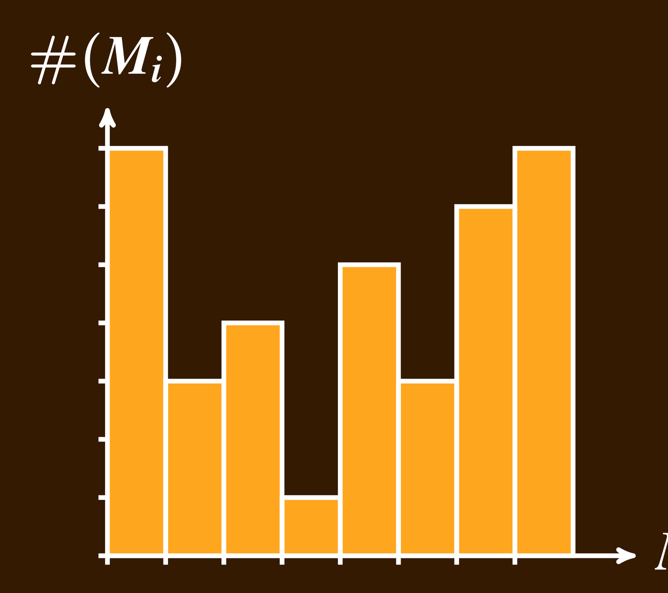
2.1.2. Without order

2.2. Previous work ([Varshney & Goyal],[Reznik])

Multiset $M = \{M_1, M_2 \dots M_m\}$ obtained by drawing m elements of n bits with repetition (contains at most m distinct elements).

2.2.1. $2^n = o(m)$

Encode histogram of cardinalities:



2.2.2. $m = o(2^n)$

Ignore repetitions $\Rightarrow$ elements are drawn uniformly.

Problem: under these conditions, asymptotic entropies of sequences with or without order are the same.

$\Rightarrow$ disregarding order has a negligible impact on compression.

2.3. Our contribution

We derive a lower bound and present an algorithm that asymptotically achieves this bound within a constant factor for the regime where

$$c \triangleq \frac{2^n}{m}, \quad c \geq 2 \text{ and fixed} .$$

3. Lower bound

As a consequence of the Kraft inquality, the expected minimum number of bits to encode such a multiset is its entropy:

$$H(M) \underset{n\to\infty}{\sim} m(n - \log_2(m)) \underset{n\to\infty}{\sim} \frac{2^n(\log_2(c))}{c} .$$

4. An algorithm based on tries

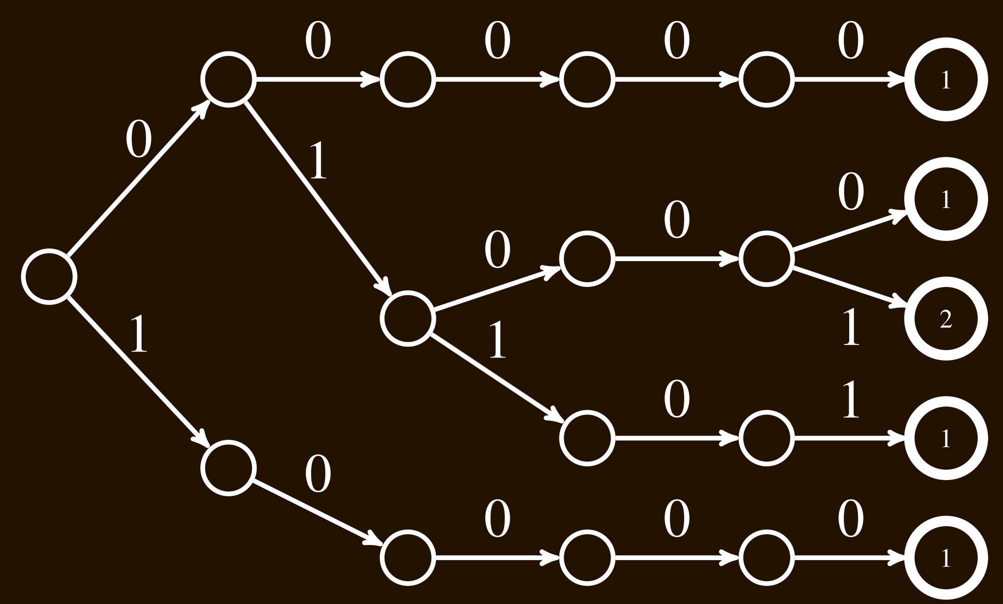
4.1. Tries

R. de La Briandais, "File searching using variable length keys," in *Proc. Western Joint Computer Conference*, 1959, pp. 295–298.

E. Fredkin, "Trie memory," *Communications of the ACM*, vol. 3, no. 9, pp. 490–499, 1960.

M = {00000,
01000,
10000,
degree 2 { 01001,
01001,
01101}

$\Rightarrow$



Idea: use tries to factorize prefixes.

5. Encoding & Decoding

5.1. Encoding

1 List all branches in lexicographic order,
[`00000`:**1**, `01000`:**1**, `01001`:**2**, `01101`:**1**, `10000`:**1**]
2 Remove duplicate consecutive prefixes,
[`00000`:**1**, `1000`:**1**, `1`:**2**, `101`:**1**, `10000`:**1**]
3 Replace all `01` with `0101`
[`00000`:**1**, `1000`:**1**, `1`:**2**, `10101`:**1**, `10000`:**1**]
4 Add `01` at the ends,
[`0000001`:**1**, `100001`:**1**, `101`:**2**, `1010101`:**1**, `1000001`:**1**]
5 Add as many `0` as the degree of the branch (0 for degree 1),
[`0000001`, `100001`, `10100`, `1010101`, `1000001`]
6 Concatenate.
`00000011000011010010101011000001`

5.2. Decoding

0 `000000110000110100110101011000001`
1 Split after maximal occurences of `(01)`$^{2i+1}$`0(0)`$^+$,
[`0000001`, `100001`, `10100`, `1010101`, `1000001`]
2 Count ending `0`'s and make it the degree (degree 1 for 0),
[`0000001`:**1**, `100001`:**1**, `101`:**2**, `1010101`:**1**, `1000001`:**1**]
3 Remove last `01` for each branch,
[`00000`:**1**, `1000`:**1**, `1`:**2**, `10101`:**1**, `10000`:**1**]
4 Replace maximum `(0101)`^{i} by `(01)`^{i},
[`00000`:**1**, `1000`:**1**, `1`:**2**, `101`:**1**, `10000`:**1**]
5 Duplicate missing prefixes.
[`00000`:**1**, `01000`:**1**, `01001`:**2**, `01101`:**1**, `10000`:**1**]

5.3. Remarks

- Lossless encoding,
- Limited complexity:
  - Encoding: $O(m(n + \log(m)))$,
  - Decoding: $O(mn)$.

6. Performance & Conclusions

6.1. Asymptotic analysis

Let L be the expected length of the encoding of a multiset using the trie technique.

6.1.2. Comparison to encoding ordered sequence

Let $H(\mathcal{M})$ be the entropy of the ordered sequence obtained using the source ($H(\mathcal{M}) = nm$), then, for c fixed:

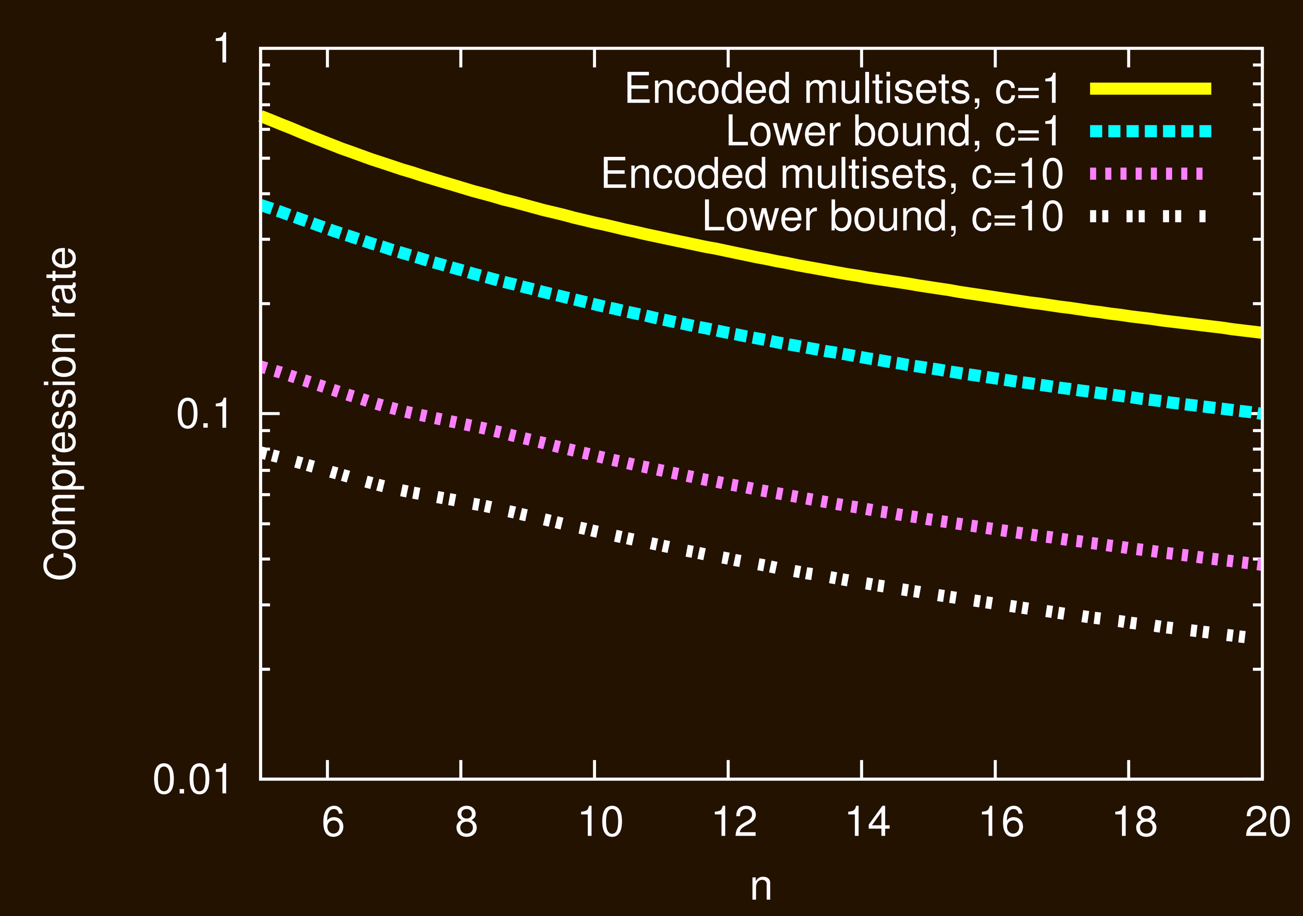
$$\frac{L}{H(\mathcal{M})} \leq \frac{10c}{3n} + \frac{4}{3n} + \frac{c}{3 \cdot 2^n} + \frac{2c}{3n \cdot 2^n} \to 0 .$$

6.1.3. Comparison to the lower bound

Let ℓ be a non-negative integer, and $c = 2^\ell$. Then for any $\epsilon > 0$, there exists a positive integer n_0 such that for any $n \geq n_0$ we have

$$\frac{L}{H(M)} \leq \frac{5}{3}\left[\frac{2}{\log_2(c)} + c(1 - e^{-\frac{1}{c}})\right] + \epsilon \underset{c\to\infty}{\longrightarrow} \frac{5}{3} + \epsilon .$$

6.2. Simulations



6.3. Conclusions

- Fast algorithm to encode multisets,
- At a constant factor of $5/3$ from the lower bound.

6.4. Open questions

- What if the source is not uniform?
- What about non-Bernouilli sources?
- Can the encoded version be efficiently used to apply set operations on multisets (union, intersection...)?
- How to get closer to the lower bound?

References

L. Varshney and V. Goyal, "Ordered and disordered source coding," in Proc. ITA, San Diego, CA, Feb. 2006.
L. Varshney and V. Goyal, "On universal coding of unordered data," in Proc. ITA, San Diego, CA, Jan. 2007.
Y. Reznik, "Codes for unordered sets of words," in Proc. IEEE ISIT, St. Petersburg, Russia, Jul. 2011.