CONTENTS

1 INTRODUCTION

2

COMPRESSION PRINCIPLES AND METHODS

Entropy, in essence, represents the minimal quantity of bits required to unequivocally distinguish an object within a set. Consequently, it serves as a foundational metric for the space utilization in compressed data representations. The ultimate aim of compressed data structures is to occupy space nearly equivalent to the entropy required for object identification, while simultaneously enabling efficient querying operations. This pursuit lies at the core of optimizing data compression techniques: achieving a balance between storage efficiency and query responsiveness.

WORST CASE ENTROPY

In its simplest form, entropy can be seen as the minimum number of bits required by identifiers (*codes*, see Section 2.2), when each element of a set U has a unique code of identical length. This is called the *worst case entropy* of U and it's denoted by $H_{wc}(U)$. The worst case entropy of a set U is given by the formula:

$$H_{wc}(U) = \log |U| \tag{1}$$

where $|U|$ is the number of elements in U.

Remark 2.1. *If we used codes of length $l < H_{wc}(U)$, we would have only $2^l \leqslant 2^{H_{wc}(U)} = |U|$ possible codes, which is not enough to uniquely identify all elements in U.*

The reason behind the attribute *worst case* is that if all codes are of the same length, then this length must be at least $\lceil \log |U| \rceil$ bits to be able to uniquely identify all elements in U. If they all have different lengths, the longest code must be at least $\lceil \log |U| \rceil$ bits long.

Example 2.2 (Worst-case entropy of $\mathcal{T}_n$)**.** *Let $\mathcal{T}_n$ denote the set of all general ordinal trees [5] with n nodes. In this scenario, each node can have an arbitrary number of children, and their order is distinguished. With n nodes, the number of possible ordinal trees is the $(n-1)$-th Catalan number, given by:*

$$|\mathcal{T}_n| = \frac{1}{n}\binom{2n-2}{n-1} \tag{2}$$

Using Stirling's approximation, we can estimate the worst-case entropy of $\mathcal{T}_n$ as:

$$|\mathcal{T}_n| = \frac{(2n-2)!}{n!(n-1)!} = \frac{(2n-2)^{2n-2} e^n e^{n-1}}{e^{2n-2} n^n (n-1)^{n-1} \sqrt{\pi n}} \left(1 + O\left(\frac{1}{n}\right)\right)$$

This simplifies to $\frac{4^n}{n^{3/2}} \cdot \Theta(1)$*, hence*

$$H_{wc}(\mathcal{T}_n) = \log |\mathcal{T}_n| = 2n - \Theta(\log n) \tag{3}$$

Thus, we have determined the minimum number of bits required to uniquely identify (encode) a general ordinal tree with n *nodes.*

2.1 ENTROPY

Let's introduce the concept of entropy as a measure of uncertainty of a random variable. A deeper explanation can be found in [19, 25, 8]

Definition 2.3 (Entropy of a Random Variable)**.** *Let* X *be a random variable taking values in a finite alphabet* $\mathcal{X}$ *with the probabilistic distribution* $P_X(x) = Pr\{X = x\}$ $(x \in \mathcal{X})$*. Then, the entropy of* X *is defined as*

This is also known as Shannon entropy, named after Claude Shannon, who introduced it in his seminal work [32]

$$H(X) = H(P_X) \stackrel{\text{def}}{=} E_{P_x}\{-\log P_X(x)\} = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) \tag{1}$$

Where E_P denotes the expectation with respect to the probability distribution P. The log is taken to the base 2 and the entropy is expressed in bits. It is then clear that the entropy of a discrete random variable will always be nonnegative[1].

Example 2.4 (Toss of a fair coin)**.** *Let* X *be a random variable representing the outcome of a toss of a fair coin. The probability distribution of* X *is* $P_X(0) = P_X(1) = \frac{1}{2}$*. The entropy of* X *is*

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \tag{2}$$

This means that the toss of a fair coin has an entropy of 1 bit.

Remark 2.5. *Due to historical reasons, we are abusing the notation and using* $H(X)$ *to denote the entropy of the random variable* X*. It's important to note that this is not a function of the random variable: it's a functional of the distribution of* X*. It does not depend on the actual values taken by the random variable, but only on the probabilities of these values.*

The concept of entropy, introduced in definition 2.3, helps us quantify the randomness or uncertainty associated with a random variable. It essentially reflects the average amount of information needed to identify a specific value drawn from that variable. Intuitively, we can think of entropy as the average number of digits required to express a sampled value.

1 The entropy is null if and only if $X = c$, where c is a costant with probability one

2.1.1 *Properties*

In the previous section 2.1, we have introduced the entropy of a single random variable X. What if we have two random variables X and Y? How can we measure the uncertainty of the pair (X, Y)? This is where the concept of joint entropy comes into play. The idea is to consider (X, Y) as a single vector-valued random variable and compute its entropy. This is the joint entropy of X and Y.

Definition 2.6 (Joint Entropy)**.** *Let (X, Y) be a pair of discrete random variables (X, Y) with a joint distribution $P_{XY}(x, y) = Pr\{X = x, Y = y\}$. The joint entropy of (X, Y) is defined as*

$$H(X, Y) = H(P_{XY}) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \log P_{XY}(x, y) \tag{3}$$

Which we can be extended to the joint entropy of n random variables $(X_1, X_2, \ldots, X_n)$ as $H(X_1, \ldots, X_n)$.

We also define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable. Given two random variables X and Y, we can define $W(y|x)$, with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, as the conditional probability of Y given X. The set W of those conditional probabilities is called *channel* with *input alphabet* $\mathcal{X}$ and *output alphabet* $\mathcal{Y}$.

Definition 2.7 (Conditional Entropy)**.** *Let (X, Y) be a pair of discrete random variables with a joint distribution $P_{XY}(x, y) = Pr\{X = x, Y = y\}$. The conditional entropy of Y given X is defined as*

$$\begin{aligned} H(Y|X) = H(W|P_X) &\stackrel{\text{def}}{=} \sum_x P_X(x) H(Y|x) && (4)\\ &= \sum_{x \in \mathcal{X}} P_X(x) \Big\{ -\sum_{y \in \mathcal{Y}} W(y|x) \log W(y|x) \Big\} && (5)\\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \log W(y|x) && (6)\\ &= E_{P_{XY}}\{-\log W(Y|X)\} && (7) \end{aligned}$$

Since entropy is always nonnegative, conditional entropy is likewise nonnegative; it has value zero if and only if Y can be entirely determined from X with certainty, meaning there exists a function $f(X)$ such that $Y = f(X)$ with probability one.

The connection between joint entropy and conditional is more evident when considering that the entropy of two random variables equals the entropy of one of them plus the conditional entropy of the other. This connection is formally proven in the following theorem.

Theorem 2.8 (Chain Rule). *Let* (X, Y) *be a pair of discrete random variables with a joint distribution* $P_{XY}(x, y)$*. Then, the joint entropy of* (X, Y) *can be expressed as*

This is also known as additivity of entropy.

$$H(X, Y) = H(X) + H(Y|X) \tag{8}$$

Proof. From the definition of conditional entropy (2.7), we have

$$\begin{aligned} H(X, Y) &= -\sum_{x,y} P_{XY}(x, y) \log W(y|x) \\ &= -\sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)} \\ &= -\sum_{x,y} P_{XY}(x, y) \log P_{XY}(x, y) + \sum_{x,y} P_X(x) \log P_X(x) \\ &= H(XY) + H(X) \end{aligned}$$

Where we used the relation

$$W(y|x) = \frac{P_{XY}(x, y)}{P_X(x)} \tag{9}$$

When $P_X(x) \neq 0$. □

Corollary 2.9.

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \tag{10}$$

Proof. The proof is analogous to the proof of the chain rule. □

Corollary 2.10.

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) \\ + \ldots + H(X_n|X_1, X_2, \ldots, X_{n-1}) \end{aligned} \tag{11}$$

Proof. We can apply the two-variable chain rule in repetition obtain the result. □

2.1.2 *Mutual Information*

Given two random variables X and Y, the mutual information between them quantifies the reduction in uncertainty about one variable due to the knowledge of the other. It is defined as the difference between the entropy and the conditional entropy. Figure 1 illustrates the concept of mutual information between two random variables.

Definition 2.11 (Mutual Information). *Let* (X, Y) *be a pair of discrete random variables with a joint distribution* $P_{XY}(x, y)$*. The mutual information between* X *and* Y *is defined as*

$$I(X; Y) = H(X) - H(X|Y) \tag{12}$$

Using the chain rule (2.8), we can rewrite it as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) && (13) \\ &= -\sum_x P_X(x)\log P_X(x) - \sum_y P_Y(y)\log P_Y(y) \\ &\quad + \sum_{x,y} P_{XY}(x,y)\log P_{XY}(x,y) && (14) \\ &= \sum_{x,y} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} && (15) \\ &= E_{P_{XY}}\left\{\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right\} && (16) \end{aligned}$$

It follows immediately that the mutual information is symmetric, $I(X;Y) = I(Y;X)$.

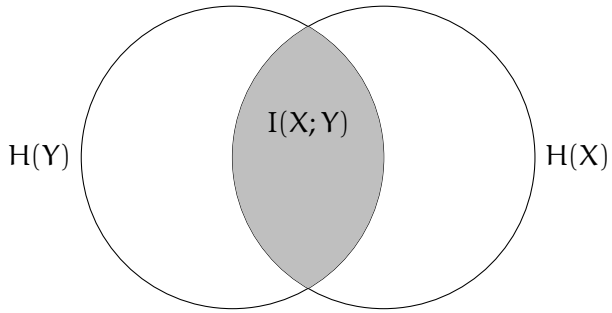

Figure 1: Mutual information between two random variables X and Y.

2.1.3 *Fano's inequality*

Information theory serves as a cornerstone for understanding fundamental limits in data compression. It not only allows us to prove the existence of encoders (Section 2.2) achieving demonstrably good performance, but also establishes a theoretical barrier against surpassing this performance. The following theorem, known as Fano's inequality, provides a lower bound on the probability of error in guessing a random variable X to it's conditional entropy $H(X|Y)$, where Y is another random variable[2].

Theorem 2.12 (Fano's Inequality)**.** *Let X and Y be two discrete random variables with X taking values in some discrete alphabet $\mathcal{X}$, we have*

$$H(X|Y) \leqslant Pr\{X \neq Y\}\log(|\mathcal{X}| - 1) + h(Pr\{X \neq Y\}) \quad (17)$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

2 We have seen in 2.7 that the conditional entropy of X given Y is zero if and only if X is a deterministic function of Y. Hence, we can estimate X from Y with zero error if and only if $H(X|Y) = 0$.

Proof. Let Z be a random variable defined as follows:

$$Z = \begin{cases} 1 & \text{if } X \neq Y \\ 0 & \text{if } X = Y \end{cases} \tag{18}$$

We can then write

$$\begin{aligned} H(X|Y) &= H(X|Y) + H(Z|XY) = H(XZ|Y) \\ &= H(X|YZ) + H(Z|Y) \\ &\leqslant H(X|YZ) + H(Z) \end{aligned} \tag{19}$$

The last inequality follows from the fact that conditioning reduces entropy. We can then write

$$H(Z) = h(\Pr\{X \neq Y\}) \tag{20}$$

Since $\forall y \in \mathcal{Y}$, we can write

$$H(X|Y = y, Z = 0) = 0 \tag{21}$$

and

$$H(X|Y = y, Z = 1) \leqslant \log(|\mathcal{X}| - 1) \tag{22}$$

Combining these results, we have

$$H(X|YZ) \leqslant \Pr\{X \neq Y\} \log(|\mathcal{X}| - 1) \tag{23}$$

From equations 19, 20 and 23, we have Fano's inequality. □

TODO: Add some conclusion to the section

2.2 SOURCE AND CODE

TODO: Some introduction about the source and coding, maybe with some very simple example like the morse code that uses a single dot to represent the most common symbol.

2.2.1 *Codes*

A source characterized by a random process generates symbols from a specific alphabet at each time step. The objective is to transform this output sequence into a more concise representation. This data reduction technique, known as *source coding* or *data compression*, utilizes a code to represent the original symbols more efficiently. The device that performs this transformation is termed an *encoder*, and the process itself is referred to as *encoding*. [19]

Definition 2.13 (Source Code)**.** *A source code for a random variable* X *is a mapping from the set of possible outcomes of* X*, called* $\mathcal{X}$*, to* $\mathcal{D}^*$*, the set of all finite-length strings of symbols from a* $\mathcal{D}$*-ary alphabet. Let* $C(X)$ *denote the codeword assigned to* x *and let* $l(x)$ *denote length of* $C(x)$

Definition 2.14 (Expected length)**.** *The expected length* $L(C)$ *of a source code* C *for a random variable* X *with probability mass function* $P_X(x)$ *is defined as*

$$L(C) = \sum_{x \in \mathcal{X}} P_X(x) l(x) \tag{1}$$

where $l(x)$ *is the length of the codeword assigned to* x*.*

Let's assume from now for simplicity that the $\mathcal{D}$-ary alphabet is $\mathcal{D} = \{0, 1, \ldots, D-1\}$.

Example 2.15. *Let's consider a source code for a random variable* X *with* $\mathcal{X} = \{a, b, c, d\}$ *and* $P_X(a) = 0.5$*,* $P_X(b) = 0.25$*,* $P_X(c) = 0.125$ *and* $P_X(d) = 0.125$*. The code is defined as*

$$C(a) = 0$$
$$C(b) = 10$$
$$C(c) = 110$$
$$C(d) = 111$$

The entropy of X *is*

$$H(X) = 0.5 \log 2 + 0.25 \log 4 + 0.125 \log 8 + 0.125 \log 8 = 1.75 \text{ bits}$$

The expected length of this code is also 1.75*:*

$$L(C) = 0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.125 \cdot 3 = 1.75 \text{ bits}$$

In this example we have seen a code that is optimal in the sense that the expected length of the code is equal to the entropy of the random variable.

Example 2.16 (Morse Code). *TODO from [8]*

Definition 2.17 (Nonsingular Code). *A code is nonsingular if every element of the range of* X *maps to a different element of* $\mathcal{D}^*$*. Thus:*

$$x \neq y \Rightarrow C(x) \neq C(y) \tag{2}$$

While a single unique code can represent a single value from our source X without ambiguity, our real goal is often to transmit sequences of these values. In such scenarios, we could ensure the receiver can decode the sequence by inserting a special symbol, like a "comma," between each codeword. However, this approach wastes the special symbol's potential. To overcome this inefficiency, especially when dealing with sequences of symbols from X, we can leverage the concept of self-punctuating or instantaneous codes. These codes possess a special property: the structure of the code itself inherently indicates the end of each codeword, eliminating the need for a separate punctuation symbol. The following definitions formalize this concept. [8]

Definition 2.18 (Extension of a Code). *The extension* C^* *of a code* C *is the mapping from finite-length sequences of symbols from* $\mathcal{X}$ *to finite-length strings of symbols from the* $\mathcal{D}$*-ary alphabet defined by*

$$C^*(x_1x_2\ldots x_n) = C(x_1)C(x_2)\ldots C(x_n) \tag{3}$$

where $C(x_1)C(x_2)\ldots C(x_n)$ *denotes the concatenation of the codewords assigned to* $x_1, x_2, \ldots, x_n$*.*

Example 2.19. *If* $C(x_1) = 0$ *and* $C(x_2) = 110$*, then* $C^*(x_1x_2) = 0110$*.*

Definition 2.20 (Unique Decodability). *A code* C *is uniquely decodable if its extension is nonsingular*

Thus, any encoded string in a uniquely decodable code has only one possibile source string that could have generated it.

Definition 2.21 (Prefix Code). *A code is a prefix code if no codeword is a prefix of any other codeword.*

Also called instantaneous code

Imagine receiving a string of coded symbols. An *instantaneous code* allows us to decode each symbol as soon as we reach the end of its corresponding codeword. We don't need to wait and see what comes next. Because the code itself tells us where each codeword ends, it's like the code "punctuates itself" with invisible commas separating the symbols. This let us decode the entire message by simply reading the string and adding commas between the codewords without needing to see any further symbols. Consider the example 2.15 seen a the beginning of this section, where the binary string `01011111010` is decoded as `0,10,111,110,10` because the code used naturally separates the symbols. [8]. Figure 2 shows the relationship between different types of codes.

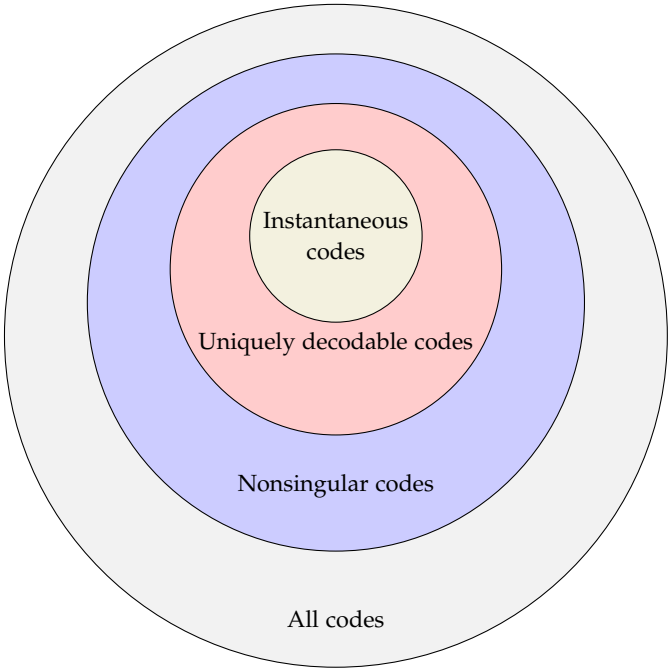

Figure 2: Relationship between different types of codes

2.2.2 *Kraft's Inequality*

We would like to construct instantaneous codes that are optimal in the sense that the expected length of the code is equal to the entropy of the random variable. However, we can't assign short codewords to all symbols and hope to be still prefix-free. Kraft's inequality provides a necessary and sufficient condition for the existence of a prefix code with given codeword lengths.

Let's denote the size of the source and code alphabets with $J = |\mathcal{X}|$ and $K = |\mathcal{D}|$, respectively. Different proofs of the following theorem can be found in [8, 19], here we report the one from [19], however the one proposed in [8] is also very interesting, based on the concept of a source tree.

Theorem 2.22 (Kraft's Inequality). *The codeword length* $l(x)$, $x \in \mathcal{X}$, *of any separable code* C *must satisfy the inequality*

$$\sum_{x \in \mathcal{X}} K^{-l(x)} \leqslant 1 \tag{4}$$

Proof. Consider the left hand side of the inequality 4 and consider its n-th power

$$\begin{aligned}\left(\sum_{x \in \mathcal{X}} K^{-l(x)}\right)^n &= \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \dots \sum_{x_n \in \mathcal{X}} K^{-l(x_1)} K^{-l(x_2)} \dots K^{-l(x_n)} \\ &= \sum_{x^n \in \mathcal{X}^\setminus} K^{-l(x^n)}\end{aligned} \tag{5}$$

Where $l(x^n) = l(x_1) + l(x_2) + \ldots + l(x_n)$ is the length of the concatenation of the codewords assigned to $x_1, x_2, \ldots, x_n$. If we consider the all the extended codewords of length m we have

$$\sum_{x^n \in \mathcal{X}^\backslash} K^{-l(x^n)} = \sum_{m=1}^{n l_{max}} A(m) K^{-m} \tag{6}$$

where $A(m)$ is the number source sequences of length n whose codewords have length m and l_{max} is the maximum length of the codewords in the code. Since the code is separable, we have that $A(m) \leqslant K^m$ and therefore each term of the sum is less than or equal to 1. Hence

$$\left(\sum_{x \in \mathcal{X}} K^{-l(x)} \right)^n \leqslant n l_{max} \tag{7}$$

That is

$$\sum_{x \in \mathcal{X}} K^{-l(x)} \leqslant (n l_{max})^{1/n} \tag{8}$$

Taking the limit as n goes to infinity and using the fact that $(n l_{max})^{1/n} = e^{1/n \log(n l_{max})} \to 1$ we have that

$$\sum_{x \in \mathcal{X}} K^{-l(x)} \leqslant 1 \tag{9}$$

That concludes the proof. □

2.2.3 *Source Coding Theorem*

Some introduction from [8, 32, 1, 19]

Theorem 2.23 (Source Coding Theorem). *TODO from [8, 19]*

Proof. TODO from [8, 19] □

2.3 EMPIRICAL ENTROPY

Before digging into the concept of empirical entropy, let's begin with the notion of binary entropy. Consider an alphabet $\mathcal{U}$, where $\mathcal{U} = \{0, 1\}$. Let's assume it emits symbols with probabilities p_0 and $p_1 = 1 - p_0$. The entropy of this source can be calculated using the formula:

$$H(p_0) = -p_0 \log_2 p_0 - (1 - p_0) \log_2 (1 - p_0)$$

We can extend this concept to scenarios where the elements are no longer individual bits, but sequences of these bits emitted by the source. Initially, let's assume the source is *memoryless* (or *zero-order*), meaning the probability of emitting a symbol doesn't depend on previously emitted symbols. In this case, we can consider chunks of n bits as our elements. Our alphabet becomes $\Sigma = \{0, 1\}^n$, and the Shannon Entropy of two independent symbols $x, y \in \Sigma$ will be the sum of their entropies. Thus, if the source emits symbols from an alphabet $\Sigma = [0, \sigma]$ where each symbol has a probability p_s, the entropy of the source becomes:

$$H(p_1, \ldots, p_\sigma) = -\sum_{s=1}^{\sigma} p_s \log p_s = \sum_{s=1}^{\sigma} p_s \log \frac{1}{p_s}$$

Remark 2.24. *If all symbols have a probability of* $p_s = 1$*, then the entropy is* 0*, and all other probabilities are* 0*. If all symbols have the same probability* $\frac{1}{\sigma}$*, then the entropy is* $\log \sigma$*. So given a sequence of* n *elements from an alphabet* Σ*, belonging to* $\mathcal{U} = \Sigma^n$*, its entropy is straightforwardly* $n\mathcal{H}(p_1, \ldots, p_\sigma)$

2.3.1 *Bit Sequences*

Let's consider a bit sequence, $B[1, n]$, which we aim to compress without access to an explicit model of a known bit source. Instead, we only have access to B. Although lacking a precise model, we may reasonably anticipate that B exhibits a bias towards either more 0s or more 1s. Hence, we might attempt to compress B based on this characteristic. Specifically, we say that B is generated by a zero-order source emitting 0s and 1s. Assuming m represents the count of 1s in B, it's reasonable to posit that the source emits 1s with a probability of $p = m/n$. This leads us to the concept of zero-order empirical entropy:

Definition 2.25 (Zero-order empirical entropy)**.** *Given a bit sequence* $B[1, n]$ *with* m 1s *and* $n - m$ 0s*, the zero-order empirical entropy of* B *is defined as:*

$$\mathcal{H}_0(B) = \mathcal{H}\left(\frac{m}{n}\right) = \frac{m}{n} \log \frac{n}{m} + \frac{n - m}{n} \log \frac{n}{n - m} \tag{1}$$

The concept of zero-order empirical entropy carries significant weight: it indicates that if we attempt to compress B using a fixed code $C(1)$ for 1s and $C(0)$ for 0s, then it's impossible to compress B to fewer than $\mathcal{H}_0(B)$ bits per symbol. Otherwise, we would have $m|C(1)| + (n-m)|C(0)| < n\mathcal{H}_0(B)$, which violates the lower bound established by Shannon entropy.

CONNECTION WITH WORST CASE ENTROPY TBD if to add this paragraph, from [25] 2.3.1

2.3.2 *Entropy of a Text*

The zero-order empirical entropy of a string $S[1, n]$, where each symbol s occurs n_s times in S, is similarly determined by the Shannon entropy of its observed probabilities:

Definition 2.26 (Zero-order empirical entropy of a text)**.** *Given a text $S[1, n]$ with n_s occurrences of symbol s, the zero-order empirical entropy of S is defined as:*

$$\mathcal{H}_0(S) = \mathcal{H}\left(\frac{n_1}{n}, \ldots, \frac{n_\sigma}{n}\right) = \sum_{s=1}^{\sigma} \frac{n_s}{n} \log \frac{n}{n_s} \tag{2}$$

Example 2.27. *Let $S =$ "abracadabra". We have that $n = 11$, $n_a = 5$, $n_b = 2$, $n_c = 1$, $n_d = 1$, $n_r = 2$. The zero-order empirical entropy of S is:*

$$\mathcal{H}_0(S) = \frac{5}{11} \log \frac{11}{5} + 2 \cdot \frac{2}{11} \log \frac{11}{2} + 2 \cdot \frac{1}{11} \log \frac{11}{1} \approx 2.04$$

Thus, we could expect to compress S to $nH_0(S) \approx 22.44$ bits, which is lower than the $n \log \sigma = 11 \cdot \log 5 \approx 25.54$ bits of the worst-case entropy of a general string of length n over an alphabet of size $\sigma = 5$.

However, this definition falls short because in most natural languages, symbol choices aren't independent. For example, in English text, the sequence "*don*'" is almost always followed by "*t*". Higher-order entropy (Section 2.4) is a more accurate measure of the entropy of a text, as it considers the probability of a symbol given the preceding symbols

2.4 HIGHER ORDER ENTROPY

TODO: A bit of introduction

Definition 2.28 (Redundacy)**.** *TODO: Give a formal definition of redundacy: informally is a measure of the distance between the source's entropy and the compression ration, and can thereby be seen as a measure of how fast the algorithm reaches the entropy of the source.*

While using measures like 2.28 are certainly intriguing, their actual usability is questionable due to the inherent challenge of determining the entropy of the source generating the string we aim to compress. To address this issue, an alternative empirical approach is the concept of the *k-th order empirical entropy* of a string S, denoted as $\mathcal{H}_k(S)$. In statistical coding (Section 2.6), we will see a scenario where $k = 0$, relying on symbol frequencies within the string. Now, with $\mathcal{H}_k(S)$, our objective is to extend the entropy concept by examining the frequencies of k-grams in string S. This requires analyzing subsequences of symbols with a length of k, thereby capturing the *compositional structure* of S. [10]

Let S be a string over the alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_n\}$. Denote with n_ω the number of occurrences of the k-gram ω in S. [3]

Definition 2.29 (k-th Order Empirical Entropy)**.** *The* k-th order empirical entropy *of a string* S *is defined as*

$$\mathcal{H}_k(S) = \frac{1}{|S|} \sum_{\omega \in \Sigma^k} \left(\sum_{i=1}^{h} n_{\omega\sigma_i} \log \left(\frac{n_\omega}{n_{\omega\sigma_i}} \right) \right) \tag{1}$$

where $|S|$ *is the length of the string* S*.*

When considering a sequence $S[1, n]$ we can compute the *empirical k-th entropy* of S by considering the frequencies of symbols depending on the k preceding symbols.

$$\mathcal{H}_k(S) = \sum_{\omega \in \Sigma^k} \frac{|S_\omega|}{n} \cdot \mathcal{H}_\prime(S_\omega) \tag{2}$$

where S_ω is a string formed by collecting the symbol that follows each occurrence of the k-gram $\omega = \sigma_1 \ldots \sigma_k$ in S.

Example 2.30. *Consider the example 2.27, where* $S =$ *"abracadabra" and* $\Sigma = \{a, b, c, d, r\}$*. The zero-order empirical entropy of* S *is* $\mathcal{H}_0(S) \approx 2.04$*. Now, let's calculate the first-order empirical entropy of* S*. We have that* $S_a =$ *"bcdb\$" (where* \$ *is the end-of-string symbol),* $S_b =$ *"rr",* $S_c =$ *"a",*

3 We will use the notation $\omega \in \Sigma^k$ to denote a k-gram, i.e., a subsequence of k symbols in the string S.

S_d = *"a"*, *and* S_r = *"aa"*. *Thus,* $H_0(S_a) \approx 1.922$, $H_0(S_b) = H_0(S_c) = H_0(S_d) = H_0(S_r) = 0$. *Therefore, the first-order empirical entropy of* S *is:*

$$\mathcal{H}_1(S) = \frac{5}{11} \cdot \mathcal{H}_0(S_a) \approx 0.874$$

That is much lower than the zero-order empirical entropy of S.

The quantity $n\mathcal{H}_k(S)$ serves as a lower bound for the minimum number of bits attainable by any encoding of S, under the condition that the encoding of each symbol may rely on itself and the k symbols preceding it in S. Consistently, any compressor that surpasses this threshold would also have the capability to compress symbols originating from the related kth-order source to a level lower than its Shannon entropy.

Remark 2.31. *As* k *grows large (up to* $k = n - 1$*, and often sooner), the* k*-th order empirical entropy of* S *reaches null, given that each* k*-gram appears only once. This renders our model ineffective as a lower bound for compressors. Even before reaching the* k *value where* $\mathcal{H}_k(S) = 0$*, compressors face practical difficulties in achieving the target of* $n\mathcal{H}_k(S)$ *bits, particularly for high* k *values. This is due to the necessity of storing the set of* σ^{k+1} *probabilities or codes, adding complexity to compression. Likewise, adaptive compressors must incorporate* σ^{k+1} *escape symbols into the compressed file, further complicating the process. In theory, it is commonly assumed that* S *can be compressed up to* $n\mathcal{H}k(S) + o(n)$ *bits for any* $k + 1 \leqslant \alpha \log_\sigma n$ *and any constant* $0 < \alpha < 1$*. In such cases, storing* σ^{k+1} *numbers within the range* $[1, n]$ *(such as the frequencies of the* k*-grams) requires* $\sigma^{k+1} \log n \leqslant n^\alpha \log n = o(n)$ *bits. [25]*

Definition 2.32 (Coarsely Optimal Compression Algorithm)**.** *A compression algorithm is coarsely optimal if, for every value of* k*, there exists a function* $f_k(n)$ *that tends to zero as the length of the sequence* n *approaches infinity, such that for all sequences* S *of increasing length, the compression ratio achieved by the algorithm remains within* $\mathcal{H}_k(S) + f_k(|S|)$.

The *Lempel-Ziv* algorithm (LZ78) serves as an example of a coarsely optimal compression technique, as outlined by Plotnik et al. in [27]. This algorithm relies on the idea of dictionary-based compression. However, as highlighted by Manzini and Korařaju [22], the notion of coarse optimality doesn't necessarily guarantee the effectiveness of an algorithm. Even when the entropy of the string is extremely low, the algorithm might still perform inadequately due to the presence of the supplementary term $f_k(|S|)$.

FURTHER COMMENTS ON LZ77 AND LZ78 TBD if to include this section, but I think it's not relevant for the thesis. If included, it should discuss very briefly the LZ77 and LZ78 algorithms, and the differences between them. [10]. And then prove two lemmas: one about the compression ration achieved by LZ78 and the other about LZ77 not being coarsely optimal. [10], end of chapter 13.

2.5 INTEGER CODING

Most important references for this section: [10, 30, 25]

TODO: Introduce the following problem: given $S = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{N}$, we want to represent the integers of S as a sequence of bits that are self-delimiting. The goal is to minimize the space occupancy of the representation [10]. Add here some examples of where this problem appears in practice [33]

The central concern in this section revolves around formulating an efficient binary representation method for an indefinite sequence of integers. Our objective is to minimize bit usage while ensuring that the encoding remains prefix-free. In simpler terms, we aim to devise a binary format where the codes for individual integers can be concatenated without ambiguity, allowing the decoder to reliably identify the start and end of each integer's representation within the bit stream and thus restore it to its original uncompressed state.

2.5.1 *Unary Code*

We begin by examining the unary code, a straightforward encoding method that represents a positive integer $x \geqslant 1$[4] using x bits. It represents x as a sequence of $x-1$ zeros followed by a single one. The correctness of this encoding is straightforward to verify: the decoder can identify the end of the integer by detecting the first one in the sequence, and the number of zeros preceding it determines the value of x.

This coding method requires x bits to represent the integer x, which is way more than the $\lceil \log_2(x) \rceil$ bits needed by a fixed-length binary code. In fact, it is very efficient for small values of x but becomes increasingly inefficient as x grows. This is a direct consequence of Theorem 2.23, which states that the ideal code length $L(c)$ for a symbol c is $-\log_2 P(c)$, where $P(c)$ is the probability of symbol c. In the case of the unary code, where we are considering positive integers, the ideal code for x would be $-\log_2 P(x) = -\log_2 2^{-x} = x$ bits. The following theorem formalizes this observation. [10]

Theorem 2.33. *The unary code of a positive integer x takes x bits, and thus it is optimal for the distribution $P(x) = 2^{-x}$.*

It is important to note that implementing a unary code requires a lot of bit shifts and bitwise operations, which are computationally expensive on modern processors. This makes the unary code impractical for large values of x

4 This is not a strict condition, but we will assume it for clarity

2.5.2 *Elias Codes*

First introduced by Levenstein in the 1960s and later refined by Elias [9] in the 1970s, the γ and δ codes are two of the most popular *universal codes* for integers. The term *universal code* refers to the characteristic of these codes to have fixed-length of $O(\log x)$ for any integer x. Compared to the binary code that requires $\lceil \log_2(x+1) \rceil$ bits, the γ and δ codes are just a constant factor away from it, while having the advantage of being prefix-free.

GAMMA CODE The γ code represents a positive integer x is divided into two parts: given $|B(x)|$ as the number of bits needed to represent x in binary, the first part is a sequence of $|B(x)|-1$ zeros followed by the binary representation of x. The γ code of x is then the concatenation of these two parts, delimited by the first one bit. The decoding process is therefore very simple: the decoder reads the bits until it finds the first one, and the number of zeros preceding it determines the length of the binary representation of x. From Shannon's condition of ideal codes (Theorem 2.23), we can see that the γ code is optimal for the distribution $P(x) \approx 1/x^2$. [10]

Theorem 2.34. *The γ code of a positive integer x takes $2\lceil \log_2(x+1) \rceil - 1$ bits, and thus it is optimal for the distribution $P(x) = 1/x^2$. This is within a factor of two from the bit length $|B(x)| = \lceil \log_2(x) \rceil$ of the fixed-length binary code.*

The γ code is inefficient due to the large number of zeros that need to be stored in the prefix, that becomes increasingly large as x grows. The δ code, introduced by Elias in 1975, addresses this issue by using a more efficient prefix.

DELTA CODE The δ code is a variation of the γ code that uses a more efficient prefix. It represents a positive integer x by first encoding the binary length of x using the γ code (we can write it as $\gamma(|B(x)|)$) and then appending the binary representation of x itself. The δ code is thus the concatenation of these two parts that do not share any bits. The decoding process is similar to the γ code: the decoder reads the bits until it finds the first one, and the number of zeros preceding it determines the length of the binary representation of x, then we fetch the next $|B(x)|$ bits to get the binary representation of x. This code takes $|\gamma(|B(x)|)| + |B(x)| = 2\lceil \log_2(|B(x)|+1) \rceil - 1 + |B(x)| \approx 2\log\log x + 1 + \log x$ bits, which is $1 + o(1)$ factor away from the bit length of the fixed-length binary code. [10]

Theorem 2.35. *The δ code of $x \geqslant 0$ takes $1 + \log_2 x + 2\log_2\log_2 x$ bits, and thus it is optimal for the distribution $P(x) \approx 1/x(\log x)^2$. This is within a factor of $1 + o(1)$ from the bit length $|B(x)| = \lceil \log_2(x) \rceil$ of the fixed-length binary code. [10]*

As for the Unary Code, implementing these codes requires a lot of bit shifts during the decoding process, making them impractical for large values of x.

2.5.3 *Rice Code*

Rice codes [29] are a family of codes parameterized by a positive integer k. Their representation is the concatenation of $(1 + x/2^k)$ as a unary code and the binary representation of the integer $(x \mod 2^k)$. Rice codes are very efficient when the values iof x are close to 2^k. When dealing with large values of x, the efficiency of the γ and δ codes decreases, and Rice codes become a better alternative providing fast compression and decompression.

The fist part takes $1 + \lceil \log_2(x/2^k) \rceil$ bits, and the second part takes k bits (since it's in the range $[0, 2^k)$). Thus, the first part is encoded in variable-length unary code, and the second part is encoded in fixed-length binary code. The closer the values of x are to 2^k, the shorter the first part becomes, making the decoding process faster.

2.5.4 *Elias-Fano Code*

TBD if to include this section or not. If included, it should be a brief mention of the Elias-Fano code and its use in integer compression [10].

2.6 STATISTICAL CODING

This section explores a technique called *statistical coding*: a method for compressing a sequence of symbols (*texts*) drawn from a finite alphabet Σ. The idea is to divide the process in two key stages: modeling and coding. During the modeling phase, statistical characteristics of the input sequence are analyzed to construct a model. In the coding phase, this model is utilized to generate codewords for the symbols of Σ, which are then employed to compress the input sequence. We will focus on two popular statistical coding methods: Huffman coding and Arithmetic coding.

2.6.1 *Huffman Coding*

Compared to the methods seen in Section 2.5, Huffman Codes (introduced by Huffman in his landmark paper [20] in the 1950s) offer a broader applicability as they do not require any specific assumptions about the probability distribution, only that all probabilities are non-zero. This versatility makes them suitable for all distributions, including those where there is no clear relationship between symbol number and probability, such as in text data.

For example, in text, characters typically range from "*a*" to "*z*" and are often mapped to a contiguous range, such as 0 to 25 or 97 to 122 in ASCII. However, there is no direct correlation between a symbol's number and its frequency rank.

TODO: Talk about construction, encoding and decoding, canonical Huffman codes [31], and the optimality of Huffman codes. [10, 30].

CONSTRUCTION OF HUFFMAN CODES The construction of Huffman codes is a greedy algorithm based on the idea of building a binary tree, where each leaf corresponds to a symbol in the alphabet Σ. The tree is built in a bottom-up fashion, starting with the symbols as leaves (we define their *size* as the number of occurrences) and iteratively merging the two nodes with the smallest probabilities until a single node is left. The code for each symbol is then obtained by traversing the tree from the root to the leaf, assigning a 0 for each left branch and a 1 for each right branch (or vice-versa). The resulting code is the path from the root to the leaf. More details can be found in [10, 30, 19, 8]

Example 2.36. *TODO: Classic example of Huffman coding with for example* $\mathcal{X} = \{a, b, c, d, e\}$ *and* $P(a) = 0.25$, $P(b) = 0.25$, $P(c) = 0.2$,

$P(d) = 0.15$, $P(e) = 0.15$. *Do a nice tree and show the encoding of each symbol.*

Let $L_C = \Sigma_{\sigma \in \Sigma} L(\sigma) \cdot P[\sigma]$ be the average length of the codewords produced by a prefix-free code C, that encodes every symbol $\sigma \in \Sigma$ with a codeword of length $L(\sigma)$. The Huffman coding produces optimal prefix codes (not in the sense that produces an optimal encoding, but in the sense that no prefix code can have a smaller average length). This is formalized in the following theorem.

Theorem 2.37 (Optimality of Huffman Codes)**.** *Let C be an Huffman Code and L_C is the shortest possible average length among all prefix-free codes C'. That is, $L_C \leqslant L_{C'}$*

This can also be interpreted as the *the minimality of the average depth* of the Huffman tree. A proof can be found in most information theory books [10, 30, 19, 8].

In the worst case, an Huffman Code can have a length of $|\Sigma| - 1$ bits, which is the same as the number of internal nodes in the tree. However, its length is limited also by $\lfloor \log_\Phi \frac{1}{p_{min}} \rfloor$, where p_{min} is the smallest probability in the set and Φ is the golden ratio. [25]. Thus, if the probabilities come from the observed frequencies of the symbols in the text, let's say n symbols, then $p_{min} \geqslant \frac{1}{n}$ and the maximum length of the code is $\log_\Phi n$. In particular, the encoding process is linear in the size of the input text [5].

The decoding process uses the Huffman Tree. It starts by reading consecutive bits from the stream and traversing the tree from the root towards a leaf based on the read bits. Upon reaching a leaf, we output the symbol it represents and then reset back to the root of the tree. Consequently, the overall decoding duration scales proportionally with the length of the compressed sequence in bits, denoted as $O(n(H(Pr) + 1))$. Since the codes are of length $O(\log n)$, it follows that any symbol can be decoded within $O(\log n)$ time.

Theorem 2.38. *Let H be the entropy of a source emitting the symbols of an alphabet Σ, hence $H = \sum_{\sigma \in \Sigma} P(\sigma) \log_2 \left(\frac{1}{P(\sigma)} \right)$. Then, the average length of the Huffman code is bounded by $H < L_H < H + 1$, where L_H is the average length of the Huffman code.*

Proof. The first inequality comes from Shannon's source coding theorem (Theorem 2.23). Let's define $l_\sigma = \lceil -\log_2 P(\sigma) \rceil$ as the length of the code for symbol σ, which is the smallest integer such upper bounding Shannon's optimal codeword length. We can easily derive that $\sum_{\sigma \in \Sigma} 2^{-l_\sigma} \leqslant 1$. Thus, recalling Kraft's inequality (Theorem 2.22), we have that exists a binary tree with $|\Sigma|$ leaves and depths l_σ for

5 In the RAM model, $O(\log n)$ bits can be manipulated in $O(1)$, so the this is true also in practice

each leaf. This tree is a prefix code, and its average codeword length is $L_C = \sum_{\sigma \in \Sigma} P(\sigma) \cdot l_\sigma$. By optimality of the Huffman code (2.37), we have that $L_H \leqslant L_C$; thus from the definition of entropy H and from the inequality $l_\sigma < 1 + \log_2\left(\frac{1}{P(\sigma)}\right)$, we have that $H < L_H < H + 1$. □

2.6.2 *Arithmetic Coding*

TODO: Do a brief introduction to Arithmetic Coding, explaining the idea behind it (Elias Code from 1960s) and the main differences with Huffman Coding. Section 12.2 from [10], section 4.2 fron [19], chapter 5 from [30].

Example 2.39. *TODO: Begin with an example to underline this differences*

Decoding Process

TODO: Talk about the compression algorithm and make an example of encoding a sequence of symbols. Add a pseudo code of the algorithm.

Decoding Process

TODO: Talk about the decompression algorithm and make an example of decoding a sequence of symbols. Add a pseudo code of the algorithm.

Efficiency of Arithmetic Coding

Theorem 2.40. *The number of bits emitted by arithmetic coding for a sequence S of n symbols is at most $2 + n\mathcal{H}$, where $\mathcal{H}$ is the empirical entropy of the sequence S.*

Proof. TODO: from [10], page 228-229. □

NOTE: this theorem requires a lemma and corollary to be proven first.

FURTER COMMENTS ON ARITHMETIC CODING TBD if to include this section. If so, it should show some other techniques such as range coding and prediction by partial matching.

3

WAVELET TREES

3.1 BITVECTORS

Consider the following problem [10]: imagine a dictionary $\mathcal{D}$ containing n strings from an alphabet Σ. We can merge all strings in $\mathcal{D}$ into a single string $T[1, m]$, without any separators between them, where m is the total length of the dictionary. The task is to handle the following queries:

- `Access_string(i)`: retrieve the i-th string in $\mathcal{D}$.
- `Which_string(x)`: find the starting position of the string in T, including the character $T[x]$.

The conventional solution involves employing an array of pointers $A[1, n]$ to the strings in $\mathcal{D}$, represented by their offsets in $T[1, m]$, requiring $\Theta(n \log n)$ bits. Consequently, `Access_string(i)` simply returns $A[i]$, while `Which_string(x)` involves locating the predecessor of x in A. The first operation is instantaneous, whereas the second one necessitates $O(\log n)$ time using binary search.

We can address the problem by employing a compressed representation of the offsets in A via a binary array $B[1, m]$ of m bits, where $B[i] = 1$ if and only if i is the starting position of a string in T. In this case then `Access_string(i)` searches for the i-th 1 in B, while `Which_string(x)` counts the number of 1s in the prefix $B[1, x]$.

In modern literature this two operations are well known as *rank* and *select* queries, respectively.

Definition 3.1 (Rank and Select)**.** *Given $B[1, n]$ a binary array of n bits (a bitvector), we define the following operations:*

- *The **rank** of an index i in B relative to a bit b is the number of occurrences of b in the prefix $B[1, i]$. We denote it as $\texttt{rank}_1(i) = \sum_{j=1}^{i} B[j]$. Similarly we can compute $\texttt{rank}_0(i) = i - \texttt{rank}_1(i)$ in constant time.*
- *The **select** of the i-th occurrence of a bit b in B is the index of the i-th occurrence of b in B. We denote it as $\texttt{select}_b(i)$. Opposite to rank, we can't derive select of 0 from select of 1 in constant time.*

Example 3.2. *TODO: Maybe add a simple example of a bitvector and show how to compute rank and select.*

In the following sections, our aim is to built structures of size $o(n)$ bits that can be added on top either the bit array or the compressed representation of B to facilitate rank and select operations.

TBD: Do I talk also about how to compress the bitvector to $n\mathcal{H}_0(B) + o(n)$ bits with those extra $o(n)$ to support rank and select operations still in $O(1)$ time? There is also the compressed solution via Elias Fano, when the numbers of 1s in B is much smaller than n, where I can achieve $n\mathcal{H}_0(B) + O(m)$ bits of space occupancy. This approach poses a much lower overhead over the entropy. It supports $\mathtt{Select}$ in constant time, while $\mathtt{access}$ and $\mathtt{Rank}$ in $O(\log \frac{n}{m})$ [25, 10]. On a further note, in [13] Ferragina and Venturini achieved higher order compression for general strings, supporting access in constant time (so adding rank and select for bitvectors are natural extensions), do I talk about this as well?

3.1.1 *Rank*

TODO:

- Explain the tree level succinct data structure supporting the rank operation as in RRR [28]
- Proof of the space occupancy in bits of the structure. [10]
- Explain how to answer rank queries upon this data structure [10, 25]

Theorem 3.3. *The space occupancy of the Rank data structure is* $o(m)$ *bits, and thus it is asymptotically sublinear in the size of the binary array* $B[1, m]$. *The Rank algorithm takes constant time in the worst case, and accesses the array* B *only in read-mode*

TBD: Do I talk about [18]?

Example 3.4. *TODO: numeric small example to show how to answer rank queries.*

3.1.2 *Select*

TODO:

- Explain how to the implementation of the Select operation mainly follows the three-level design of the Rank data structure, with the algorithmic twist that here the binary array B is not split into big and small blocks of fixed length, but the splitting is driven by the number of bits set to 1. [10, 25]
- Talk about the space occupancy in bits of the structure. [10, 25]

Theorem 3.5. *The space occupancy of the* $\texttt{Select}_1$ *data structure is* $o(m)$ *bits, and thus it is asymptotically sublinear in the size of the binary array* $B[1, m]$. *The* $\texttt{Select}_1$ *algorithm takes constant time in the worst case, and accesses the array* B *only in read-mode. The same time and space bounds hold for the* $\texttt{Select}_0$ *algorithm. [10]*

3.2 WAVELET TREES

TODO: Improve this introduction, just a draft.

Wavelet trees, introduced in 2003 by Grossi, Gupta, and Vitter [15] are a self indexing data structure: meaning they can answer rank and select queries, while still allowing to access the text. This combination makes them particularly useful for compressed full-text indexes like the FM-index [12]. In such indexes, wavelet trees are employed to efficiently answer rank queries during the search process.

TBD if to add: Upon closer examination, one can recognize that the wavelet tree is a slight extension of an older (1988) data structure by Chazelle [7], commonly used in Computational Geometry. This structure represents points on a two-dimensional grid, undergoing a reshuffling process to sort them by one coordinate and then by the other. Kärkkäinen (1999) [21] was the first to apply this structure to text indexing, although the concept and usage differed from Grossi et al.'s proposal four years later.

Wavelet Trees can be seen in different ways: (i) as sequence representation, (ii) as a permutation of elements, and (iii) as grid point representation. Since 2003, these perspectives and their interconnections have proven valuable across diverse problem domains, extending beyond text indexing and computational geometry, where the structure originated [26, 17, 11].

An introduction to the problem

Consider a sequence $S[1, n]$ as a generalization of bitvectors whose elements $S[i]$ are drawn from an alphabet Σ[1]. The problem is to support the following queries:

- `Access(i)`: return the i-th element of S.
- `Rank(c,i)`: return the number of occurrences of character c in the prefix $S[1, i]$.
- `Select(c,i)`: return the position of the i-th occurrence of character c in S.

However, dealing with sequences is much more complex than dealing with bitvectors (as we have seen in Section 3.1) TODO: further expand this concept: show an attemp to solve this problem with a

1 The size of the alphabet varies depending on the application. For example, in DNA sequences, the alphabet is $\Sigma = \{A, C, G, T\}$ (in Chapter 4 we will focus more on this specific case), while in other case it could be of millions of characters, such as in natural language processing.

naive approach, and see its shortcomings. The problem will be space occupancy, since if we use the solution for bitvectors, then the total space occupancy will be $n\sigma + o(n\sigma)$ bits, which is way too much. [25]

From now on, let $S[1, n] = s_1 s_2 \dots s_n$ be a sequence of length n over an alphabet Σ that for simplicity we write as $\Sigma = \{1, \dots, \sigma\}$. In this way, the string can be represented using $n\lceil \log \sigma \rceil = n \log \sigma + o(n)$ bits in plain form.

3.2.1 *Structure*

In the beginning of this section we showed that storing one bitvector per symbol is not space-efficient. The wavelet tree is a data structure that solves this problem by using a recursive hierarchical partitioning of the alphabet. Consider the subset $[a, b] \subset [1, \dots, \sigma]$, then a wavelet tree over $[a, b]$ is a balanced binary tree with $b - a + 1$ leaves[2]. The root node v_{root} is associated with the whole sequence $S[1, n]$, and stores a bitmap $B_{v_{root}}[1, n]$ defined as follows: $B_{v_{root}}[i] = 0$ if $S[i] \leqslant (a + b)/2$ and $B_{v_{root}}[i] = 1$ otherwise. The tree is then recursively built by associating the subsequence $S_0[1, n_0]$ of elements in $[a, \dots, \lfloor (a + b)/2 \rfloor]$ to the left child of v, and the subsequence $S_1[1, n_1]$ of elements in $[\lfloor (a + b)/2 \rfloor + 1, \dots, b]$ to the right child of v. This process is repeated until the leaves are reached. In this way the left child of the root node, is a wavelet tree for $S_0[1, n_0]$ over the alphabet $[a, \dots, \lfloor (a + b)/2 \rfloor]$, and the right child is a wavelet tree for $S_1[1, n_1]$ over the alphabet $[\lfloor (a + b)/2 \rfloor + 1, \dots, b]$. [26]

Example 3.6. *TODO: Add an example of a wavelet tree for a sequence with small alphabet.*

Remark 3.7. *The wavelet tree described has σ leaves and $\sigma - 1$ internal nodes, and the height of the tree is $\lceil \log \sigma \rceil$. The space occupancy of each level it's exactly n bits, while we have at most n bits for the last level. The total number of bits stored by the wavelet tree in then upper bounded by $n\lceil \log \sigma \rceil$ bits. [26]. This however is not sufficient, if we want to also store the topology of the tree, we need to add extra $O(\sigma \log n)$ bits.*

Tracking symbols

We have seen how the wavelet tree serves as a representation for a string S, but more than that it is a succinct data structure for the string. Thus, it takes space asymptotically close to the plain representation of the string and allows us to access the i-th symbol of the string in $O(\log \sigma)$ time. In algorithm 1 we show how extract the i-th symbol of the string S using a wavelet tree T, this operation is called `Access`.

2 if $a = b$ then the tree is just a leaf

Algorithm 1 Answering `Access` queries on a wavelet tree

Require: Sequence S (as a wavelet tree T), position i
Ensure: The i-th symbol of S, i.e the output of `Access(S,i)`

```
[a, b] ← [1, σ]
while a ≠ b do
    if access(v.B, i) = 0 then                ▷ i-th bit of the bitmap of v
        i ← rank_0(v.B, i)
        v ← v.left                            ▷ move to the left child of node v
        b ← ⌊(a + b)/2⌋
    else
        i ← rank_1(v.B, i)
        v ← v.right                           ▷ move to the right child of node v
        a ← ⌊(a + b)/2⌋ + 1
    end if
end while
return a
```

In order to find $S[i]$, we first look at the bitmap associated with the root node of the wavelet tree, and depending on the value of the i-th bit of the bitmap, we move to the left or right child of the root node and continue recursively. However, the problem is to determine where our i has been mapped to: if we move to the left child, then we need to find the i-th 0 in the bitmap of the left child, and if we move to the right child, then we need to find the i-th 1 in the bitmap of the right child. This is done by the $\mathtt{rank}_0$ and $\mathtt{rank}_1$ functions, respectively. We continue this process until we reach a leaf node, and then we return the value of the leaf node.

Example 3.8. *TODO and TBD: add a figure with the wavelet tree and the process of finding the i-th symbol of the string.*

In addition to retrieving the i-th symbol of the string, we might also need to perform the inverse operation. That is, given a symbol's position at a leaf node, we aim to determine the position of the symbol in the string. This operation is referred to as `Select` and is outlined in Algorithm **??**..

TODO: add the select algorithm and explain the process of finding the position of a symbol in the string. The references are Navarro's [26, 25]

TODO: Show how to generalize the rank and select operations to the wavelet tree. So with a query $\mathtt{rank}_c(S, i)$, instead of moving from the root toward the leaf corresponding to $S[i]$, we move toward the leaf of c. The idea is similar for select. [25] 6.2.2

TODO: Add some pratical consideration about using constant-time $\mathtt{rank}$ and $\mathtt{select}$ on the bitvectors. [25] 6.2

TBD if to add: Reducing redundancy

TODO: In 3.7 we mentioned that storing the topology of the wavelet tree requires $O(\sigma \log n)$ bits. This may be critical for large alphabets, and in this section we will show that this term can be removed by slightly altering the balanced wavelet tree shape. [24, 23].

3.2.2 *Construction*

TODO: Explain how to build a wavelet tree in $O(n \log n)$ time (and also talk about space occupancy), the procedure is well described in [25] 6.2.3.

Algorithm 2 Building a wavelet tree

Require: Sequence $S[1, n]$ over alphabet $\Sigma = \{1, \ldots, \sigma\}$
Ensure: Wavelet tree T for S
TODO: add the algorithm to build the wavelet tree

3.3 COMPRESSED WAVELET TREES

TODO: An introduction on why we need to compress wavelet trees. From [25] 6.2.4

3.3.1 *Entropy Coding*

Compressing the bitvectors, from [26] 3.1 and [25] 6.2.4

3.3.2 *Huffman-Shaped Wavelet Trees*

TODO: An alternative to obtain nearly zero-order compressed space while using plain bitvectors is to give a Huffman shape to the wavelet tree, instead of using a balanced tree of height $\log \sigma$. From [25] 6.2.4, [26] 3.2

3.3.3 *Higher Order Entropy Coding*

TODO: In 3.3 from [26] there is a good explanation on how to compress the bitvectors using higher-order entropy coding. It references papers as [16]. There is also section 5 from [11] that gives a more technical explanation

4 SUBSET WAVELET TREES

4.1 INTRODUCTION: DEGENERATE STRINGS

Brief introduction to degenerate strings [14, 4]

4.2 STRUCTURE OF THE SUBSET WAVELET TREE

TODO: Describe the structure of the subset wavelet tree, from [3]

4.3 SUBSET-RANK AND SUBSET-SELECT

TODO: Describe the problem of subset-rank and subset-select, from [3]. Add pseudocode for the algorithms.

4.4 TBD WHAT TO DO FROM HERE

In [3] they compare 5 methods that supports rank and rank-pairs queries on small alphabet sequences (they need it to answer those queries on the sequences stored at the nodes of the subset wavelet tree). The methods are

- Wavelet Trees
  - Wavelet Trees with un-compressed bitvectors
  - Wavelet Trees with compressed bitvectors (RRR [28])
- Scanning Rank
- Sequence Splitting
- Generalized RRR

TBD: Do I have to talk about them? They developed this methods to support fast membership queries on de Bruijn graphs (Do I care about this in the thesis?). Further more, the danish in [6] say that:

> we show that any structure supporting either subset-rank or subset-select must use at least $N \log \sigma - o(N \log \sigma)$ bits in the worst case (Theorem 2). By plugging a standard rank-select data structure into Theorem 1 we, in many cases, match this bound to within lower order terms, while simultaneously matching the query time of the fastest known rank-select data structures (see below). Note that any lower

> bound for rank-select queries also holds for subset rank-select queries since any string is also a degenerate string. All our results hold on a word RAM with logarithmic word-size. Finally, we provide implementations of the reductions and compare them to the implementations of the Subset Wavelet Tree provided in [3], and the implementations of the reductions provided in [2]. Our most compact structure matches the space of their most compact structure while answering queries twice as fast. We also provide a structure using vector processing features that matches the space of the most compact structure while improving query time by a factor four to seven, remaining competitive with the fast structures for queries.

Of course it remains the open problem stated in [3]:

> The main open problem we leave is to find a tighter analysis of the space required by subset wavelet trees when entropy compression is applied to their node sequences. In particular, can the size of the resulting structure be related in some way to the entropy of the subset sequence

So a part of this chapter will focus on trying to answer this question. It's TBD how to structure all this and what to keep and what to leave out.

BIBLIOGRAPHY

[1] N. Vereshchagin A. Shen V. A. Uspensky. *Kolmogorov Complexity and Algorithmic Randomness*. Mathematical Surveys and Monographs. Amer Mathematical Society, 2017.

[2] Jarno N Alanko, Simon J Puglisi, and Jaakko Vuohtoniemi. "Small searchable κ-spectra via subset rank queries on the spectral burrows-wheeler transform." In: *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA23)*. SIAM. 2023, pp. 225–236.

[3] Jarno N. Alanko et al. "Subset Wavelet Trees." In: *21st International Symposium on Experimental Algorithms (SEA 2023)*. Ed. by Loukas Georgiadis. Vol. 265. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, 4:1–4:14.

[4] Mai Alzamel et al. "Degenerate string comparison and applications." In: *WABI 2018-18th Workshop on Algorithms in Bioinformatics*. Vol. 113. 2018, pp. 1–14.

[5] David Benoit et al. "Representing trees of higher degree." In: *Algorithmica* 43 (2005), pp. 275–292.

[6] Philip Bille, Inge Li Gørtz, and Tord Stordalen. *Rank and Select on Degenerate Strings*. 2023.

[7] Bernard Chazelle. "A Functional Approach to Data Structures and Its Use in Multidimensional Searching." In: *SIAM Journal on Computing* 17.3 (1988), pp. 427–462.

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012.

[9] P. Elias. "Universal codeword sets and representations of the integers." In: *IEEE Transactions on Information Theory* 21.2 (1975), pp. 194–203.

[10] P. Ferragina. *Pearls of Algorithm Engineering*. Cambridge University Press, 2023.

[11] Paolo Ferragina, Raffaele Giancarlo, and Giovanni Manzini. "The myriad virtues of Wavelet Trees." In: *Information and Computation* 207.8 (2009), pp. 849–866.

[12] Paolo Ferragina and Giovanni Manzini. "Opportunistic data structures with applications." In: *Proceedings 41st annual symposium on foundations of computer science*. IEEE. 2000, pp. 390–398.

[13] Paolo Ferragina and Rossano Venturini. "A simple storage scheme for strings achieving entropy bounds." In: *Theoretical Computer Science* 372.1 (2007), pp. 115–121. ISSN: 0304-3975.

[14] Michael J Fischer and Michael S Paterson. "String-matching and other products." In: (1974).

[15] Roberto Grossi, Ankur Gupta, and Jeffrey Vitter. "High-Order Entropy-Compressed Text Indexes." In: *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* (Nov. 2002).

[16] Roberto Grossi, Ankur Gupta, Jeffrey Scott Vitter, et al. "When indexing equals compression: experiments with compressing suffix arrays and applications." In: *SODA*. Vol. 4. 2004, pp. 636–645.

[17] Roberto Grossi, Jeffrey Scott Vitter, and Bojian Xu. "Wavelet Trees: From Theory to Practice." In: *2011 First International Conference on Data Compression, Communications and Processing*. 2011, pp. 210–221.

[18] Roberto Grossi et al. *More Haste, Less Waste: Lowering the Redundancy in Fully Indexable Dictionaries*. 2009.

[19] T.S. Han and K. Kobayashi. *Mathematics of Information and Coding*. Fields Institute Monographs. American Mathematical Society, 2002.

[20] David A. Huffman. "A Method for the Construction of Minimum-Redundancy Codes." In: *Proceedings of the IRE* 40.9 (1952), pp. 1098–1101.

[21] J Kärkkäinen. "Repetition-based text indexing." PhD thesis. Ph. D. thesis, Department of Computer Science, University of Helsinki, Finland, 1999.

[22] S Rao Kosaraju and Giovanni Manzini. "Compression of low entropy strings with Lempel–Ziv algorithms." In: *SIAM Journal on Computing* 29.3 (2000), pp. 893–911.

[23] Veli Mäkinen and Gonzalo Navarro. "Position-Restricted Substring Searching." In: *LATIN 2006: Theoretical Informatics*. Ed. by José R. Correa, Alejandro Hevia, and Marcos Kiwi. Springer Berlin Heidelberg, 2006, pp. 703–714.

[24] Veli Mäkinen and Gonzalo Navarro. "Rank and select revisited and extended." In: *Theoretical Computer Science* 387.3 (2007), pp. 332–347.

[25] G. Navarro. *Compact Data Structures: A Practical Approach*. Cambridge University Press, 2016.

[26] Gonzalo Navarro. "Wavelet trees for all." In: *Journal of Discrete Algorithms* 25 (2014). 23rd Annual Symposium on Combinatorial Pattern Matching, pp. 2–20. ISSN: 1570-8667.

[27] Eli Plotnik, Marcelo J Weinberger, and Jacob Ziv. "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm." In: *IEEE transactions on information theory* 38.1 (1992), pp. 66–72.

[28] Rajeev Raman, Venkatesh Raman, and Srinivasa Rao Satti. "Succinct indexable dictionaries with applications to encoding k-ary trees, prefix sums and multisets." In: *ACM Transactions on Algorithms* 3.4 (Nov. 2007), p. 43.

[29] Robert F Rice. *Some practical universal noiseless coding techniques*. Tech. rep. 1979.

[30] K. Sayood. *Lossless Compression Handbook*. Communications, Networking and Multimedia. Elsevier Science, 2002, pp. 55–64.

[31] Eugene S Schwartz and Bruce Kallick. "Generating a canonical prefix encoding." In: *Communications of the ACM* 7.3 (1964), pp. 166–169.

[32] C. E. Shannon. "A mathematical theory of communication." In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[33] Ian H Witten, Alistair Moffat, and Timothy C Bell. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.