On Weighted Prefix Normal Words

Yannik Eikmeier
Kiel University, Germany
stu204329@mail.uni-kiel.de

Pamela Fleischmann
Kiel University, Germany
fpa@informatik.uni-kiel.de

Dirk Nowotka
Kiel University, Germany
dn@informatik.uni-kiel.de

Abstract

A prefix normal word is a binary word whose prefixes contain at least as many 1s as any of its factors of the same length. Introduced by Fici and Lipták in 2011 the notion of prefix normality is so far only defined for words over the binary alphabet. In this work we investigate possible generalisations for finite words over arbitrary finite alphabets, namely *weighted* and *subset prefix normality*. We prove that weighted prefix normality is more expressive than both binary and subset prefix normality and investigate the existence of a weighted prefix normal form. While subset prefix normality directly inherits most properties from the binary case, weighted prefix normality comes with several new peculiarities that did not already occur in the binary case. We characterise these issues and solve further questions regarding the weighted prefix normality and weighted prefix normal form.

2012 ACM Subject Classification Theory of computation → Formal languages and automata theory

Keywords and phrases Combinatorics on Words, Prefix Normal Words, Complexity Measures

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

Complexity measures of words, like for example the number of different factors of a word, are a central topic of investigation when dealing with properties of sequences. Characterising the maximum density of a particular letter in the set of factors of a given length, hence considering an abelian setting, falls into that category. Such characterisations inevitably prompt the search for and investigation of normal forms representing words with equivalent measures. Prefix normality is the concept considered in this paper and was firstly introduced by Fici and Lipták in 2011 [10] as a property describing the distribution of a designated letter within a binary word. A word over the binary alphabet $\{0, 1\}$ is *prefix normal* w.r.t. 1 if its prefixes contain at least as many 1s as any of its factors of the same lengths. For example the word 1101001 is 1-prefix normal but not 0-prefix normal, since 0 is a factor but not a prefix. In a sense, prefixes of 1-prefix normal words give an upper bound for the amount of 1s any other factor of the word may contain. For a given binary word w the *maximum-ones function* maps a length to the maximum amount of 1s, a factor of w of that length can have. This leads to the 1-*prefix normal form* of a binary word, which is the 1-prefix normal word with an identical maximum-ones function. For instance, the 1-prefix normal form of 1001101 is 1101001 because they have the same maximum-ones function and the second word is 1-prefix normal. The 0-prefix normal form is defined analogously. In [7] Burcsi *et al.* show that there exists exactly one 1-prefix normal form in the set of all binary words that have an identical maximum-ones function.

From an application point of view this complexity measure is directly connected to the *Binary Jumbled Pattern Matching Problem (BJPM)* (see e.g. [1, 3, 5]). The BJPM problem

is to determine whether a given finite binary word has factors containing given amounts of 1s and 0s. In [10] prefix normal forms were used to construct an index for the BJPM problem in $O(n)$ time where n is the given word's length. The best known algorithm for this problem has a runtime of $O(n^{1.864})$ (see [8]). In [2] Balister and Gerke showed that the number of prefix normal words of length n is $2^{n-\Theta(\log^2(n))}$ and the class of a given prefix normal word contains at most $2^{n-O(\sqrt{n\log(n)})}$ elements. In more theoretical settings, the language of binary prefix normal words has also been extended to infinite binary words [9]. Prefix normality has been shown to be connected to other fields of research within combinatorics on words, e.g. Lyndon words [10], and bubble languages [6]. Furthermore, efforts have been made to recursively construct prefix normal words, via the notions of extension critical words (collapsing words) and prefix normal palindromes [11, 6]. The goal therein was to learn more about the number of words with the same prefix normal form and the number of prefix normal palindromes. Very recently in [4] a Gray code for prefix normal words in amortized polylogarithmic time per word was generated. The sequences related to prefix normal words can be found in the On-Line Encyclopedia of Integer Sequences ([12]): A194850 (number of prefix normal words of length n), A238109 (list of prefix normal words over the binary alphabet), A238110 (maximum number of binary words of length n having the same prefix normal form), and A308465 (number of prefix normal palindromes of length n).

Our contribution. In this work, we investigate generalisations of prefix normality for finite words over arbitrary finite alphabets. We define a *weight measure*, which is a morphic function assigning a weight to every letter (and thus to every word) over an arbitrary finite alphabet. Based on those weights we can again compare factors and prefixes of words over this alphabet w.r.t. their weight. We define *weighted prefix normality* as follows: a word is prefix normal w.r.t a weight measure if no factor has a higher weight than the prefix of the same length as the factor. Note here, for some weight measures not every word has a unique prefix normal form. We prove basic properties of weight measures and weighted prefix normality, and give a characterisation of weight measures for which every word has a prefix normal form. Based on this, we define a generalised weight measure which only depends on the alphabetic order of the letters. Later we will also discuss the naïve approach to generalise binary prefix normality by selecting a subset X of the alphabet Σ: then every letter in X is treated like a 1 and the rest like 0s as in the binary case. We prove that this approach can already be obtained by the use of a weight measure with binary weights.

This paper is organized as follows: In Section 2, we define the basic terminology and notions needed to discuss weighted prefix normality. In Section 3, we prove that weighted prefix normality is indeed a proper generalisation of the binary case and present our results on the existence of a weighted prefix normal form. Following that, in Section 4, we further investigate special weight measures and their effect on weighted prefix normality. And finally, in Section 5, we present the aforementioned naïve *subset* approach to generalise prefix normality and compare it to weighted prefix normality.

2 Preliminaries

Let $\mathbb{N}$ denote the natural numbers $\{1, 2, 3, \dots\}$ and let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For all $k \in \mathbb{N}$ we define $\mathbb{N}_{>k} := \{n \in \mathbb{N} \mid n > k\}$, and analogously $\mathbb{N}_{<k}, \mathbb{N}_{\leq k}, \mathbb{N}_{\geq k}$. Let $\mathbb{Z}$ denote the integer numbers. We define the interval $[i, j] := \{n \in \mathbb{N} \mid i \leq n \leq j\}$, for $i, j \in \mathbb{N}$. Similarly we define $[n] := [1, n]$, for $n \in \mathbb{N}$, and $[n]_0 := [0, n]$. Let $\mathbb{P} \subset \mathbb{N}$ denote the prime numbers. We say a triple $(X, \cdot, \chi)$ is a *monoid* if X is a set and $\cdot : X^2 \to X$ is a associative binary operation with the neutral element χ, i.e. $a \cdot \chi = \chi \cdot a = a$ holds for all $a \in X$. Furthermore we say a

quadruple $(X, \cdot, \chi, \prec)$ is a *strictly totally ordered monoid* with the strict and total order $\prec$ on X. For monoids $(A, *, \varepsilon), (B, \circ, \lambda)$ a function $\mu : A \to B$ is a *morphism* if $\mu(x*y) = \mu(x) \circ \mu(y)$ holds for all $x, y \in A$. Notice, if the domain A is a free monoid over some set S, a morphism from $A \to B$ is sufficiently defined by giving a mapping from S to B.

An *alphabet* Σ is a finite set of letters. Let Σ_2 be the *binary alphabet* $\{\mathtt{0}, \mathtt{1}\}$. A *word* is a finite sequence of letters from a given alphabet. Σ^* denotes the set of all finite words over Σ, i.e. the free monoid over Σ. Let ε denote the *empty word* and set $\Sigma^+ := \Sigma^* \backslash \{\varepsilon\}$ as the free semi-group over Σ. We denote the length of a word $w \in \Sigma^*$ by $|w|$, i.e. the number of letters in w. Thus $|\varepsilon| = 0$ holds. Let w be a word of length $n \in \mathbb{N}$. Let $w[i]$ denote the i^{th} letter of w for $i \in [|w|]$, and set $w[i \dots j] = w[i] \dots w[j]$ for $i, j \in [|w|]$ and $i \leq j$. Let $w[i \dots j] = \varepsilon$ if $i > j$. The number of occurrences of a letter $\mathtt{a} \in \Sigma$ in w is denoted by $|w|_\mathtt{a} = |\{i \in [|w|] \mid w[i] = \mathtt{a}\}|$. We generalise this notation to $|w|_X := |\{i \in [|w|] \mid w[i] \in X\}|$ for $X \subseteq \Sigma$, i.e. $|w|_X$ is the number of letters of w that are elements of X. Notice that $|\cdot|_\mathtt{a} = |\cdot|_{\{\mathtt{a}\}}$ holds for any $\mathtt{a} \in \Sigma$. We say $x \in \Sigma^*$ is a *factor* of w if there exists $u, v \in \Sigma^*$ with $w = uxv$ holds and in this case u is called a *prefix*. We denote the set of w's factors (prefixes) by $\mathrm{Fact}(w)$ ($\mathrm{Pref}(w)$ resp.) and $\mathrm{Fact}_i(w)$ ($\mathrm{Pref}_i(w)$) denotes the set of all factors (prefix resp.) of length $i \in [|w|]$.

Given a total order $<$ over Σ let $<_{lex}$ denote the extension of $<$ to a lexicographic order over Σ^*. Fixing an strictly totally ordered alphabet $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \dots, \mathtt{a}_n\}$ with $\mathtt{a}_i < \mathtt{a}_j$ for $1 \leq i < j \leq n$, the *Parikh vector* of a word is defined by $p : \Sigma^* \to \mathbb{N}^n : w \mapsto (|w|_{\mathtt{a}_1}, |w|_{\mathtt{a}_2}, \dots, |w|_{\mathtt{a}_n})$. For some function $f : B \to C$ and $A \subseteq B$, A, B, C sets, we define $f(A) := \{f(a) \mid a \in A\}$ and the composition of functions (assumed $B = C$) by $f^i := f \circ f^{i-1}$ inductively with f^0 being the identity function for all $i \in \mathbb{N}$. For two sets of words $A, B \subseteq \Sigma^*$ we define their concatenation by $AB := \{vw \mid v \in A, w \in B\}$. For other common definitions and background on the topic of words see *Combinatorics on Words* [14].

Before we define the weight measures and weighted prefix normality we recall the definition for binary prefix normality as introduced by Fici and Lipták in [10].

▶ **Definition 1.** *([10]) Given $w \in \Sigma_2^*$ the* maximum-ones function f_w *and the* prefix-ones function p_w *are defined by*

$$f_w : [|w|]_0 \to \mathbb{N}_0,\ i \mapsto \max(|\operatorname{Fact}_i(w)|_1) \quad \textit{and} \quad p_w : [|w|]_0 \to \mathbb{N}_0,\ i \mapsto |\operatorname{Pref}_i(w)|_1.$$

The word w is called binary-prefix normal *if $f_w = p_w$ holds.*

Our generalisation of binary prefix normality is based on so called *weight measures*, i.e. we apply weights - represented by elements from a strictly totally ordered monoid - to every letter of the alphabet.

▶ **Definition 2.** *Let $(A, \circ, \lambda, \prec)$ be a strictly totally ordered monoid. A morphism $\mu : \Sigma^* \to A$ is a* weight measure *over the alphabet Σ w.r.t. $(A, \circ, \lambda, \prec)$ if $\mu(vw) = \mu(wv)$ and $\mu(w) \prec \mu(wv)$ hold for all words $w \in \Sigma^*$ and $v \in \Sigma^+$. We refer to the second property as* increasing property. *We say the weights of the letters of Σ are* base weights *so $\mu(\Sigma)$ is the set of all* base weights.

▶ Remark 3. Notice that if there exists a weight measure μ w.r.t. some monoid $(A, \circ, \lambda, \prec)$ then $|A|$ is infinite, $\circ$ is commutative, and $\mu(\varepsilon) = \lambda$ holds. Moreover, the increasing property of weight measures ensures that only the neutral element of Σ^*, namely ε is mapped to the neutral element λ of A. Consequently we will see that our factor- and prefix-weight functions are strictly monotonically increasing in contrast to the functions defined in [10]. However, if we allow letters from Σ to also be assigned the neutral weight λ, we get the known results for binary alphabets.

▶ Remark 4. A weight measure μ can be defined for any alphabet Σ in two simple steps. First, choose some infinite commutative monoid with a total and strictly order, e.g. $(\mathbb{N}, +, 0, <)$. Second, assign a base weight that is greater than the neutral element to every letter in Σ. By the morphic property, the weight measure μ is well defined for all words in Σ^*.

In the following definition we introduce seven special weight measures.

▶ **Definition 5.** *A weight measure μ over the alphabet Σ w.r.t. $(A, \circ, \lambda, \prec)$. is*

- injective *if μ is injective on Σ,*
- alphabetically ordered *if Σ is strictly totally ordered by $<$ and $\mu(\mathtt{a}) \prec \mu(\mathtt{b})$ holds for all $\mathtt{a}, \mathtt{b} \in \Sigma$ with $\mathtt{a} < \mathtt{b}$,*
- binary *if $|\mu(\Sigma)| = 2$ holds, and* non-binary *if $|\mu(\Sigma)| > 2$ (the unary case is not of interest),*
- natural *if A is $\mathbb{N}_0$ or $\mathbb{N}$ with $\prec$ being the usual "less than" relation,*
- *a* sum weight measure *if it is natural with $\mathbb{N}_0$, $\circ = +$, and $\lambda = 0$,*
- *a* product weight measure *if it is natural with $\mathbb{N}$, $\circ = *$, and $\lambda = 1$,*
- prime *if it is a product weight measure and $\mu(\Sigma) \subseteq \mathbb{P}$ holds.*

Consider for example the alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. The weight measure μ over Σ with $\mu(\mathtt{a}) = 1$, $\mu(\mathtt{b}) = 2$, and $\mu(\mathtt{c}) = 3$ is *non-binary*, *natural*, and with the strictly totally ordered monoid $(\mathbb{N}_0, +, 0, <)$ it is a *sum weight measure*. It can not be a *product weight measure* with $(\mathbb{N}, *, 1, <)$ because then $\mu(\mathtt{a}) = 1$ would violate the increasing property of weight measures. However the weight measure μ over Σ w.r.t. $(\mathbb{N}, *, 1, <)$ with $\mu(\mathtt{a}) = 2$, $\mu(\mathtt{b}) = 3$, and $\mu(\mathtt{c}) = 5$ is not only a *product weight measure*, but also a *prime weight measure*.

▶ Remark 6. For the binary alphabet $\Sigma_2 = \{\mathtt{0}, \mathtt{1}\}$ a sum weight measure μ with $\mu(w) = |w|_1$ for all $w \in \Sigma_2^*$ can not exist since we would have $\mu(\mathtt{0}) = 0 = \mu(\varepsilon)$ which is a contradiction to the increasing property of weight measures. Later on we are going to circumvent this problem by setting $\mu(w) = |w|_1 + |w|$ for all $w \in \Sigma_2^*$ when implementing binary prefix normality via the usage of weight measures. Alternatively, we may relax the increasing property and allow $\mu(0) = 0$; this results in the binary case in exactly the same properties of [10].

▶ Remark 7. If μ is a injective prime weight measure then the weight of any word is characteristic for its Parikh vector by the uniqueness of the prime factorisation: By the Fundamental Theorem of Arithmetic [13] we know that every natural number can be represented uniquely up to order as the product of prime numbers (with 1 being the empty product), i.e. for every weight of a word there is a unique way to partition it into primes. Because the weight measure is injective, these primes have to be the exact weights of all the word's letters. In conclusion: any two words with the same weight under an injective prime weight measure must have the exact same letters, i.e. the same Parikh vectors.

We now define the analogons for the maximum-ones and prefix-ones function.

▶ **Definition 8.** *Let $w \in \Sigma^*$ and μ be a weight measure over the alphabet Σ w.r.t. $(A, \circ, \lambda, <)$. Define the* factor-weight function $f_{w,\mu}$ *and the* prefix-weight function $p_{w,\mu}$ *by*

$$f_{w,\mu} : [|w|]_0 \to A,\ i \mapsto \max(\mu(\mathrm{Fact}_i(w))) \text{ and } p_{w,\mu} : [|w|]_0 \to A,\ i \mapsto \mu(\mathrm{Pref}_i(w)).$$

(Notice for any word $w \in \Sigma^*$ and weight measure μ over Σ, $f_{w,\mu}(0) = p_{w,\mu}(0) = 0$ holds - we included the 0 only to have more convenience later on.)

Let μ be a sum weight measure, i.e. the target monoid A is given by $(\mathbb{N}_0, +, 0, <)$, with the base weights $\mu(\mathtt{a}) = 1$, $\mu(\mathtt{n}) = 2$, $\mu(\mathtt{b}) = 3$ for the alphabet $\Sigma = \{\mathtt{a}, \mathtt{n}, \mathtt{b}\}$. Now consider the words `banana` and `nanaba`. Table 1 shows the mappings of their prefix- and factor-weight

i	1	2	3	4	5	6
$p_{\texttt{nanaba},\mu}(i)$	2	3	5	6	9	10
$f_{\texttt{nanaba},\mu}(i)$	3	4	6	7	9	10
$p_{\texttt{banana},\mu}(i)$, $f_{\texttt{banana},\mu}(i)$	3	4	6	7	9	10

Table 1 Comparing `banana`'s and `nanaba`'s prefix- and factor-weight function.

functions. The factor-weight function of `nanaba` is realised by the factors `b`, `ab`, `nab`, `anab`, `nanab`, `nanaba`.

In the following we define a more general version of the binary position function defined in [10]. With the binary context in mind this function is defined to give the position of the i$^{\text{th}}$ `1` in the word w, i.e. $\text{pos}_w(i) := \min\{k \mid p_w(k) = i\}$ for all $i \in [p_w(|w|)]$ and $w \in \Sigma_2^*$. However, in the weighted context for most words not every weight corresponds to a prefix with exactly the same weight. Consequently we do not define a single exact position function, but two functions (maxpos and minpos) which together enclose the position within a word where a certain weight is reached. Only if both functions return the same position for some word and weight, that word's prefix up to that position has exactly that weight (see Lemma 11).

▶ **Definition 9.** *Let $w \in \Sigma^*$, and μ be a weight measure over Σ w.r.t. $(A, \circ, \lambda, <)$. We define the* max-position function *and* min-position function *by*

$$\text{maxpos}_{w,\mu} : A \to [|w|]_0,\ i \mapsto \max\{k \in [|w|]_0 \mid p_{w,\mu}(k) \leq i\},$$
$$\text{minpos}_{w,\mu} : A \to [|w|]_0,\ i \mapsto \min\{k \in [|w|]_0 \mid p_{w,\mu}(k) \geq i\}.$$

We now define a generalised approach for binary prefix normality, namely the *weighted prefix normality* for a given weight measure μ. As in the binary case, for a prefix normal word the factor-weight function and the prefix-weight function have to be identical, i.e. each factor is at most as *heavy* as the prefix of the same length.

▶ **Definition 10.** *Let $w \in \Sigma^*$, and μ be a weight measure over Σ w.r.t. $(A, \circ, \lambda, <)$. We say w is* μ-prefix normal *(or weighted prefix normal w.r.t. μ) if $p_{w,\mu} = f_{w,\mu}$ holds.*

Let μ be a sum weight measure over $\{\texttt{a}, \texttt{b}, \texttt{n}\}$ with the base weights $\mu(\texttt{a}) = 1$, $\mu(\texttt{n}) = 2$, and $\mu(\texttt{b}) = 3$. Then the word `banana` is μ-prefix normal but the word $w =$ `nanaba` is not μ-prefix normal witnessed by $f_{w,\mu}(1) = 3 \neq 2 = p_{w,\mu}(1)$.

3 Weighted Prefix Normal Words and Weighted Prefix Normal Form

In this section we show that *weighted prefix normality* is a proper generalisation of binary prefix normality and further investigate the weighted prefix normal form. We then examine special properties of weight measures and define *gapfree weight measures* for which every word has a weighted prefix normal equivalent.

Firstly we show some useful basic properties of the prefix- and factor-weight function, and the min- and max-position function which are direct generalisations of the binary case.

▶ **Lemma 11.** *Let μ be a weight measure over Σ w.r.t. $(A, \circ, \lambda, <)$, $w \in \Sigma^*$, $j, k \in [|w|]_0$ and $x, y \in A$. Then $p_{w,\mu}$ and $f_{w,\mu}$ and have the following properties:*

(1) *$j < k$ iff $f_{w,\mu}(j) < f_{w,\mu}(k)$ iff $p_{w,\mu}(j) < p_{w,\mu}(k)$,*

(2) $p_{w,\mu}(\text{maxpos}_{w,\mu}(x)) \leq x \leq p_{w,\mu}(\text{minpos}_{w,\mu}(x))$,
(3) $\text{maxpos}_{w,\mu}(p_{w,\mu}(k)) = \text{minpos}_{w,\mu}(p_{w,\mu}(k)) = k$,
(4) *if* $j > \text{maxpos}_{w,\mu}(x)$ *then* $p_{w,\mu}(j) > x$ *and if* $j < \text{minpos}_{w,\mu}(x)$ *then* $p_{w,\mu}(j) < x$,
(5) $\text{maxpos}_{w,\mu}(x) \leq \text{minpos}_{w,\mu}(x)$,
(6) $x < y$ *implies* $\text{maxpos}_{w,\mu}(x) \leq \text{maxpos}_{w,\mu}(y)$ *as well as* $\text{minpos}_{w,\mu}(x) \leq \text{minpos}_{w,\mu}(y)$.

Proof.
(1) With the increasing property of weight measures the equivalences follow from the definition of the factor-weight function as the maximum over all factors and the fact that every prefix itself is a prefix of every longer prefix.
(2) Directly follows by the definition of the max-position and min-position function as $\text{maxpos}_{w,\mu}(x) = \max\{i \in [|w|]_0 \mid p_{w,\mu}(i) \leq x\}$ and $\text{minpos}_{w,\mu}(x) = \min\{i \in [|w|]_0 \mid p_{w,\mu}(i) \geq x\}$.
(3) Follows by the definition of the max-position and min-position function and the fact that the prefix-weight function is strictly increasing (see (1)).
(4) Follows by the definition of the max-position function as a maximum and the min-position function as a minimum.
(5) Follows by (1) and (2).
(6) Suppose otherwise, so let $x < y$ but $\text{maxpos}_{w,\mu}(x) > \text{maxpos}_{w,\mu}(y)$ holds. With (4) we then have $p_{w,\mu}(\text{maxpos}_{w,\mu}(x)) > y$ and with (2) we have $p_{w,\mu}(\text{maxpos}_{w,\mu}(x)) \leq x$. Together these are a contradiction to $x < y$. Now suppose $\text{minpos}_{w,\mu}(x) > \text{minpos}_{w,\mu}(y)$ holds. Analogously to before with (4) we have $p_{w,\mu}(\text{minpos}_{w,\mu}(y)) < x$ and with (2) we have $p_{w,\mu}(\text{minpos}_{w,\mu}(y)) \geq y$. Which is again a contradiction to $x < y$.

◀

▶ **Lemma 12.** *For a weight measure* μ *over the alphabet* Σ *w.r.t.* $(A, \circ, \lambda, <)$ *and* $w \in \Sigma^*$ *there holds* $f_{w,\mu}(j) \leq f_{w,\mu}(i) \circ f_{w,\mu}(j-i)$ *for all* $i, j \in [|w|]_0$ *with* $i < j$.

Proof. Let $i, j \in [|w|]_0$ be indexes with $i < j$. Now suppose $f_{w,\mu}(j) > f_{w,\mu}(i) \circ f_{w,\mu}(j-i)$. Let $u \in \text{Fact}_j(w)$ be a factor of w with $\mu(u) = f_{w,\mu}(j)$. Then by the definition of the factor-weight function, $\mu(u[1 \dots i]) \leq f_{w,\mu}(i)$ and $\mu(u[(i+1) \dots j]) \leq f_{w,\mu}(j-i)$ both hold. And thus $f_{w,\mu}(j) > \mu(u[1 \dots i]) \circ \mu(u[(i+1) \dots j]) = \mu(u)$ holds. This is a contradiction because u was chosen with $\mu(u) = f_{w,\mu}(j)$, so the original claim follows. ◀

▶ **Proposition 13.** *For a weight measure* μ *over the alphabet* Σ *w.r.t.* $(A, \circ, \lambda, <)$ *and* $w \in \Sigma^*$ *the following properties are equivalent:*

(1) w *is* μ*-prefix normal,*
(2) $p_{w,\mu}(j) \leq p_{w,\mu}(i) \circ p_{w,\mu}(j-i)$ *for all* $i, j \in [|w|]_0$ *with* $i < j$,
(3) $\text{minpos}_{w,\mu}(\mu(v)) \leq |v|$ *for all* $v \in \text{Fact}(w)$,
(4) $\text{maxpos}_{w,\mu}(a) + \text{minpos}_{w,\mu}(b) \leq \text{minpos}_{w,\mu}(a \circ b)$ *for all* $a, b \in A$, $a \circ b \leq \mu(w)$.

Proof. (1)⇒(2). Follows by Lemma 12, since for any prefix normal word the prefix- and factor-weight function are equal by definition.

(2)⇒(3). Assume we have (2) but suppose there exists $v \in \text{Fact}(w)$ with $|v| < \text{minpos}_{w,\mu}(\mu(v))$. Now we write v as $v = w[i+1 \dots j]$ for some $i, j \in [|w|]_0$ with $i < j$. Then we have $p_{w,\mu}(j) = p_{w,\mu}(i) \circ \mu(v)$. And by Lemma 11 (8) we know that $\mu(v) > p_{w,\mu}(|v|)$ holds so in total we have $p_{w,\mu}(j) > p_{w,\mu}(i) \circ p_{w,\mu}(|v|)$. Which is a contradiction to (2), because we have $|v| = |w[i+1 \dots j]| = j - i$.

(3)⇒(1). Assume we have (3). Let $i \in [|w|]$ and let $v \in \mathrm{Fact}(w)$ with $\mu(v) = f_{w,\mu}(i)$ then we have $|v| \geq \mathrm{minpos}_{w,\mu}(\mu(v))$. By Lemma 11 (2,4) follows $p_{w,\mu}(|v|) \geq p_{w,\mu}(\mathrm{minpos}_{w,\mu}(\mu(v))) \geq \mu(v)$ from which (1) follows directly because we now have $p_{w,\mu}(i) \geq f_{w,\mu}(i)$.

(3)⇒(4). Let $a, b \in A$ with $a \circ b \leq \mu(w)$, $m = \mathrm{minpos}_{w,\mu}(a \circ b)$ and $n = \mathrm{maxpos}_{w,\mu}(a)$. Now consider w's factor $v = w[n+1 \dots m]$ which has a length of $m - n$. So $p_{w,\mu}(n) \circ \mu(v) = p_{w,\mu}(m)$ and $|v| = m - n$ each follow. By Lemma 11 (3,4) we know $a \circ \mu(v) \geq a \circ b$ and therefore $\mu(v) \geq b$ holds. Again by Lemma 11 (11) we get $\mathrm{minpos}_{w,\mu}(\mu(v)) \geq \mathrm{minpos}_{w,\mu}(b)$. So in total with (3) follows $\mathrm{minpos}_{w,\mu}(b) \leq \mathrm{minpos}_{w,\mu}(\mu(v)) \leq |v| = m - n = \mathrm{minpos}_{w,\mu}(a \circ b) - \mathrm{maxpos}_{w,\mu}(a)$.

(4)⇒(3). Let $v \in \mathrm{Fact}(w)$, we write v as $v = w[i+1 \dots j]$ for some $i, j \in [|w|]_0$. So with Lemma 11 we have $\mathrm{minpos}_{w,\mu}(p_{w,\mu}(i)) = i$ and $\mathrm{minpos}_{w,\mu}(p_{w,\mu}(i) \circ \mu(v)) = \mathrm{minpos}_{w,\mu}(p_{w,\mu}(j)) = j$. By (4) we then have $\mathrm{minpos}_{w,\mu}(\mu(v)) \leq \mathrm{minpos}_{w,\mu}(p_{w,\mu}(i) \circ \mu(v)) - \mathrm{minpos}_{w,\mu}(p_{w,\mu}(i)) = j - i = |v|$. ◀

Before we define the analogon to the prefix-equivalence for factor weights, we show that weighted prefix normality is more general and more expressive than binary prefix normality, i.e. every statement on binary prefix normality can be expressed by weighted prefix normality but not vice versa.

▶ **Theorem 14.** *Weighted prefix normality is a generalisation of binary prefix normality.*

Proof. W.l.o.g. consider just 1-prefix normality for the binary case. We construct a sum weight measure μ over the alphabet Σ_2. Let $\mu(1) = 2$ and $\mu(0) = 1$. Then $|w|_1 + |w| = \mu(w)$ holds for any binary word $w \in \Sigma_2$. It follows that $f_w(i) + |w| = \max(|\mathrm{Fact}_i(w)|_1) + |w| = \max(\mu(\mathrm{Fact}_i(w))) = f_{w,\mu}(i)$ and $p_w(i) + |w| = p_{w,\mu}(i)$ hold for all $i \in [|w|]$. Therefore, w is μ-prefix normal if and only if it is 1-prefix normal. So, with such a weight measure every statement on binary prefix normality can be transformed into an analogue using weighted prefix normality. ◀

Analogously to the binary case we define an equivalence relation on words based on equality of the factor-weight function.

▶ **Definition 15.** *Let μ be a weight measure over Σ w.r.t. $(A, \circ, \lambda, <)$. Two words $w, w' \in \Sigma^*$ are* factor-weight equivalent *($w \sim_\mu w'$) if $f_{w,\mu} = f_{w',\mu}$ holds.*

▶ Remark 16. The proof that $\sim_\mu$ is in fact an equivalence relation is straightforward. We denote the equivalence classes by $[w]_{\sim_\mu} := \{w' \in \Sigma^* \mid w \sim_\mu w'\}$.

We already saw (Table 1) that `banana` and `nanaba` over $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{n}\}$ (with the there given weight measure μ) are factor-weight equivalent. The complete equivalence class is given by $[\mathtt{banana}]_{\sim_\mu} = \{\mathtt{ananab}, \mathtt{anaban}, \mathtt{abanan}, \mathtt{nanaba}, \mathtt{nabana}, \mathtt{banana}\}$. Notice that all words in the class have the same Parikh vector. Also notice that only `banana` is μ-prefix normal. But for example if we were to add `c` to Σ and expand μ by $\mu(\mathtt{c}) = \mu(\mathtt{n}) = 2$ then $[\mathtt{banana}]_{\sim_\mu}$ would contain all words previously in it but also those where some `n`s are substituted by `c`. So now $[\mathtt{banana}]_{\sim_\mu}$ would contain four μ-prefix normal words, namely `banana`, `bacana`, `banaca`, and `bacaca`. Lastly, consider the sum weight measure ν over the same alphabet $\Sigma = \{\mathtt{a}, \mathtt{n}, \mathtt{b}\}$ and with the base weights $\mu(\mathtt{a}) = 1$, $\mu(\mathtt{n}) = 2$, $\mu(\mathtt{b}) = 4$. Now $[\mathtt{babn}]_{\sim_\nu}$ only contains `babn` and its reverse `nbab`. Interestingly none of the two words are ν-prefix normal, witnessed by $f_{\mathtt{babn},\nu} = f_{\mathtt{nbab},\nu} = (4, 6, 9, 11)$, $p_{\mathtt{babn},\nu} = (4, 5, 9, 11)$, and $p_{\mathtt{nbab},\nu} = (2, 6, 7, 11)$ (the functions are written as sequence for brevity). In a sense a letter with weight 3 is missing

to fill the gap between $f_{\mathtt{babn},\nu}(2) = 6$ and $f_{\mathtt{babn},\nu}(3) = 9$. For example with such a letter $\mathtt{x}$ in Σ with $\nu(\mathtt{x}) = 3$ the word $\mathtt{bnxn}$ is ν-prefix normal and in $[\mathtt{babn}]_{\sim_\nu}$. These examples show that factor-weight equivalence classes can contain words with different Parikh vectors, multiple prefix normal words, and even no prefix normal words at all.

We now investigate the question which equivalence classes contain a single prefix normal word that can be used to represent that class, i.e. a *normal form*, as it always exists for the binary case [10].

▶ **Definition 17.** *For $w \in \Sigma^*$ and a weight measure μ over Σ we define the μ-prefix normal subset of the factor-weight equivalence class of w by $\mathcal{P}_\mu(w) := \{v \in [w]_{\sim_\mu} \mid p_{v,\mu} = f_{v,\mu}\}$.*

In the example above, multiple prefix normal words in a single class are a result of ambiguous base weights, i.e. multiple letters having the same weight. That is because all letters with the same weights are interchangeable in any word with no effect on the weight of that word, so there exist multiple prefix normal words for a word w if w is affected by the *non-injectivity* of the weight measure. By choosing an injective weight measure we can avoid this behaviour. However, the problematic case where some equivalence classes contain no prefix normal words at all still remains. We now give a characterisation of special, so called *gapfree*, weight measures and show that they guarantee the existence of a prefix normal word in every equivalence class of the factor-weight equivalence. Before we prove the just stated claims we define formally the previous observations of *gaps*.

▶ **Definition 18.** *We say a weight measure μ over the alphabet Σ w.r.t. $(A, \circ, \lambda, \prec)$ is* gapfree*, if for all words $w \in \Sigma^*$ and all $i \in [|w|]$ there exists an $\mathtt{a} \in \Sigma$ such that $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{a})$ holds. Otherwise, if for some word $w \in \Sigma^*$ and $i \in [|w|]$ there exists no $\mathtt{a} \in \Sigma$ such that $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{a})$ holds we say μ is* gapful *and has a gap over the word w at the index i.*

Let $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ and let μ be a sum weight measure over Σ with $\mu(\mathtt{a}) = 2$, $\mu(\mathtt{b}) = 4$, and $\mu(\mathtt{c}) = 6$. We show that μ is gapfree by proving the existence of a letter in Σ with weight $x \in \mathbb{N}$ such that $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ x$ holds for all $w \in \Sigma^*$ and $i \in [|w|]$. The factor-weight function is defined as a maximum consequently x can at most be equal to $\mu(\mathtt{c}) = 6$. On the other hand x has to be at least $\mu(\mathtt{a}) = 2$ because the factor-weight function is strictly increasing. And because all the base weights $\mu(\Sigma) = \{2, 4, 6\}$ are even, so are $f_{w,\mu}(i)$ and $f_{w,\mu}(i-1)$. Thus x has to be even as well. In total we then have $x \in \{2, 4, 6\} = \mu(\Sigma)$. So there is a letter in Σ with the correct weight to fill every possible gap, i.e. μ is gapfree. On the other hand the sum weight measure ν over Σ with $\nu(\mathtt{a}) = 1$, $\nu(\mathtt{b}) = 3$, and $\nu(\mathtt{c}) = 4$ is gapful. Consider the word $w = \mathtt{bcac}$ then ν has a gap over w at the index 3 since $f_{w,\nu}(3) = 9$ (witnessed by the factor $\mathtt{cac}$) and $f_{w,\nu}(2) = 7$ (witnessed by the factor $\mathtt{bc}$).

▶ **Theorem 19.** *Let μ be a weight measure over Σ. Then*
- *there exists a $w \in \Sigma^*$ such that $|\mathcal{P}_\mu(w)| = 0$ iff μ is gapful,*
- *there exists a $w \in \Sigma^*$ such that $|\mathcal{P}_\mu(w)| > 1$ iff μ is not injective, and*
- *for all $w \in \Sigma^*$ we have $|\mathcal{P}_\mu(w)| = 1$ iff μ gapfree and injective.*

Proof. Let μ be a weight measure over Σ w.r.t. $(A, \circ, \lambda, \prec)$. For the first equivalence consider firstly that μ is gapful. Then there exists some word $w \in \Sigma^*$ and an index $i \in [|w|]$ such that w has a gap at i. Thus there exists no $n \in A$ for which $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ n$ holds. Now suppose there exists some word $w' \in \mathcal{P}_\mu(w)$. For such a word $p_{w',\mu}(i) = f_{w',\mu}(i) = f_{w,\mu}(i)$ and $p_{w',\mu}(i-1) = f_{w',\mu}(i-1) = f_{w,\mu}(i-1)$ both must hold. Thus we get $f_{w,\mu}(i-1) \circ \mu(w'[i]) = f_{w',\mu}(i-1) \circ \mu(w'[i]) = p_{w',\mu}(i-1) \circ \mu(w'[i]) = p_{w',\mu}(i) = f_{w',\mu}(i) = f_{w,\mu}(i)$. Which is

a contradiction to the gap, so $\mathcal{P}_\mu(w) = \emptyset$ holds. For the second direction choose $w \in \Sigma^*$ with $\mathcal{P}_\mu(w) = \emptyset$. Suppose μ is gapfree, so $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{a})$ holds for all $i \in [|w|]$ and appropriate $\mathtt{a} \in \Sigma$. Then we have a contradiction by constructing a word $w' \in \mathcal{P}_\mu(w)$ as follows: Choose $w'[1] \in \Sigma$ with $\mu(w'[1]) = f_{w,\mu}(1)$, which is possible according to the assumption for $i = 1$. And for $i \in [|w|]$ we can inductively choose $w'[i] \in \Sigma$ with $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(w'[i])$, which is also possible according to the assumption. Now $p_{w',\mu} = f_{w',\mu}$ and $p_{w',\mu} = f_{w,\mu}$ hold by construction, so $w' \in \mathcal{P}_\mu(w)$ holds.

For the second claim let μ be not injective. Now there exist some distinct letters $\mathtt{a}, \mathtt{b} \in \Sigma$ which have the same weight $\mu(\mathtt{a}) = \mu(\mathtt{b}) = i \in A$. So $[\mathtt{a}]_{\sim_\mu} = [\mathtt{b}]_{\sim_\mu}$ and $p_{\mathtt{a},\mu} = f_{\mathtt{b},\mu}$ both hold directly. From which $\{\mathtt{a}, \mathtt{b}\} \subseteq \mathcal{P}_\mu(\mathtt{a})$ follows. For the second direction consider now $w, u, v \in \Sigma^*$ with $u \neq v$ and $\{u, v\} \subseteq \mathcal{P}_\mu(w)$. Now by the definition of $\mathcal{P}_\mu(w)$, the prefix-weight function of u and v are both equal to the factor-weight function of w. So $p_{u,\mu} = p_{v,\mu}$ holds, and therefore $\mu(u[j]) = \mu(v[j])$ holds for all $j \in [|u|]$. On the other hand because u and v are different words, there exists some $i \in [|u|]$ with $u[i] \neq v[i]$. In other words, μ is not injective.

The third claim follows directly from the first two. ◀

▶ **Definition 20.** *Let μ be a gapfree and injective weight measure over Σ and $w \in \Sigma^*$. Then $|\mathcal{P}_\mu(w)| = 1$ holds and $w' \in \mathcal{P}_\mu(w)$ (or short $\mathcal{P}_\mu(w)$) is the μ-prefix normal form of w.*

With the alphabet $\Sigma = \{\mathtt{a}, \mathtt{n}, \mathtt{b}, \mathtt{x}, \mathtt{y}\}$ and the sum weight measure μ over Σ with base weights $\mu(\mathtt{a}) = 1$, $\mu(\mathtt{n}) = 2$, $\mu(\mathtt{b}) = 3$, $\mu(\mathtt{x}) = 4$, and $\mu(\mathtt{y}) = 4$ we have $\mathcal{P}_\mu(\mathtt{nanaba}) = \{\mathtt{banana}\}$ and $\mathcal{P}_\mu(\mathtt{bbax}) = \{\mathtt{xnnb}, \mathtt{ynnb}\}$. So `banana` is the μ-prefix normal form of `nanaba` and in the second case we see that `bbax` has no *true* μ-prefix normal form. Also if `n` were not in Σ we see μ would have a gap over `bbax` and $\mathcal{P}_\mu(\mathtt{bbax})$ would be empty. Additionally, ignoring the `y`, `bbax` is an example of a word whose prefix normal form differs in its Parikh vector from the original word.

We now give a naïve inductive construction for a word's weighted prefix normal form.

▶ Remark 21. Let μ be a gapfree and injective weight measure over the alphabet Σ w.r.t. $(A, \circ, 0, \prec)$ and $w \in \Sigma^*$. Then the μ-prefix normal form $w' = \mathcal{P}_\mu(w)$ can be constructed inductively over $|w|$ as follows:

$$w'[1] = \mathtt{a} \in \Sigma \text{ ,where } f_{w,\mu}(1) = \mu(\mathtt{a}),$$
$$w'[i] = \mathtt{a} \in \Sigma \text{ ,where } f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{a}) \text{ and } i \in [|w|],\ i > 1.$$

For a gapfree non-injective weight measure this inductive construction can be used to non-deterministically construct any prefix normal word within the factor-weight equivalence class of a word. We can see that this construction always results in a prefix normal word by the following arguments.

First, such a w' exists because μ is gapfree and therefore for all $i \in [|w|]$ there exists an $\mathtt{a} \in \Sigma$ with $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{a})$. Second, w' is unambiguous because μ is injective and therefore there exists exactly one $\mathtt{a} \in \Sigma$ for any specific weight. And third, the word w' is in $[w]_{\sim_\mu}$ and is μ-prefix normal because its prefixes are constructed to have exactly the weight of w's corresponding factor with the maximum weight. So we have $w' = \mathcal{P}_\mu(w)$. We show this by induction over $i \in [|w|]$. For $i = 1$, we directly have $p_{w',\mu}(1) = \mu(w'[1]) = f_{w,\mu}(1)$ by construction. Now assuming the claim holds for some $i \in [|w| - 2]$, we then have $p_{w',\mu}(i+1) = \mu(w'[1 \ldots i+1]) = p_{w',\mu}(i) \circ \mu(w'[i+1])$ which is equal to $f_{w,\mu}(i) \circ \mu(w'[i+1]) = f_{w,\mu}(i+1)$ by our inductive assumption.

4 Grapfree and Injective Weight Measures

In this section we investigate the behaviour of gapfree and injective weight measures in more detail. First of all, by their definition we can infer that every binary weight measure is gapfree and prime weight measures are in general gapful.

▶ **Lemma 22.** *All binary weight measures are gapfree.*

Proof. Let μ be a binary weight measure over Σ w.r.t. $(A, \circ, \lambda, \prec)$ and with the two base weights $\mu(\Sigma) = \{x, y\}$, where $x \prec y$. W.l.o.g. let μ be injective, so Σ is binary as well. Furthermore w.l.o.g. let $\Sigma = \Sigma_2$ and $\mu(\mathtt{0}) = x$, $\mu(\mathtt{1}) = y$. Now let $w \in \Sigma^*$ and $i \in [|w|]$. Then $f_{w,\mu}(i)$ is realised by some factor $u \in \mathrm{Fact}_i(w)$ with $\mu(u) = f_{w,\mu}(i)$ and $f_{w,\mu}(i-1)$ is realised by some factor $v \in \mathrm{Fact}_i(v)$ with $\mu(v) = f_{w,\mu}(i-1)$. Now $|v|_1 - |u|_1 \in \{0, 1\}$ holds because otherwise $\mu(v)$ or $\mu(u)$ would not be the maximum weight a factor of length i or $i-1$ could have. In total either $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{1})$ or $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{0})$ holds. Therefore μ is gapfree. ◀

▶ Remark 23. With the above we see that when modelling binary prefix normality by means of weighted prefix normality (e.g. in the proof of Theorem 14) we have the existence of a unique binary prefix normal form as expected.

▶ **Lemma 24.** *All non-binary prime weight measures are gapful.*

Proof. By the definition of the prime numbers none of them can be obtained by multiplying two other natural numbers. Let μ be a non-binary prime weight measure over an alphabet Σ and let $\mathtt{a}, \mathtt{b}, \mathtt{c} \in \Sigma$ with $\mu(\mathtt{a}) < \mu(\mathtt{b}) < \mu(\mathtt{c})$. Consider the word $w =: \mathtt{bcac} \in \Sigma^*$ then $f_{w,\mu}(3) = \mu(\mathtt{cac})$ and $f_{w,\mu}(2) = \mu(\mathtt{bc})$ both hold. But there cannot exist an $a \in \mu(\Sigma)$ such that $\mu(\mathtt{cac}) = \mu(\mathtt{bc}) * a$ would hold, because all base weights are prime numbers and if such an a existed the uniqueness of the prime number factorisation would be violated. Consequently μ has a gap over w at the index 3 and therefore μ is gapful. ◀

We now give an alternative condition under which a weight measure is gapfree. In most cases this condition is easier to prove than the original statement on gapfree weight measures. We will also see that this condition is a proper characterisation for gapfree weight measures w.r.t. certain kinds of monoids.

▶ **Definition 25.** *Let $(A, \circ, \lambda, \prec)$ be a strictly totally ordered monoid. A* step function *is a function $\sigma : A \to A$ for which there exists an $s \in A$ (the* step*) such that we have $\sigma(a) = a \circ s$ for all $a \in A$. A weight measure μ over Σ w.r.t $(A, \circ, \lambda, \prec)$. has* stepped base weights *if there exists a step function σ such that $\mu(\Sigma) = \{\sigma^i(\min(\mu(\Sigma))) \mid i \in [0, |\mu(\Sigma)| - 1]\}$ holds.*

▶ Remark 26. In other words a weight measure μ has stepped base weights if every base weight $\mu(\Sigma)$ is of the form $\sigma^i(\min(\mu(\Sigma))) = \min(\mu(\Sigma)) \circ s \cdots \circ s$ where σ is some step function and s its step. Additionally for each step from the minimum up to the maximum in $\mu(\Sigma)$ there has to exist a letter with that exact weight.

In the previous example for $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ the weight measure μ over Σ with $\mu(\mathtt{a}) = 2$, $\mu(\mathtt{b}) = 4$, and $\mu(\mathtt{c}) = 6$ has stepped base weights with the step of 2. In contrast the sum weight measure ν over Σ with $\nu(\mathtt{a}) = 1$, $\nu(\mathtt{b}) = 3$, and $\nu(\mathtt{c}) = 4$ does not, because the difference between $\nu(\mathtt{a})$ and $\nu(\mathtt{b})$ is 2 but between $\nu(\mathtt{b})$ and $\nu(\mathtt{c})$ it is only 1. Consequently there cannot be any one step $s \in \mathbb{N}$ such that we would have $\nu(\mathtt{a}) + s = \nu(\mathtt{b})$ and $\nu(\mathtt{c}) + s = \nu(\mathtt{c})$.

The following two propositions characterise the relation between gapfree weight measures and stepped base weights.

▶ **Proposition 27.** *All non-binary weight measures with stepped base weights are gapfree.*

Proof. Let μ be a non-binary weight measure over Σ w.r.t $(A, \circ, \lambda, \prec)$ with stepped base weights. W.l.o.g let μ be injective and let Σ be of the form $\{\mathtt{a}_0, \mathtt{a}_1, \ldots, \mathtt{a}_{n-1}\}$, where the letters are in ascending order of their weight, so $\mu(\mathtt{a}_i) \prec \mu(\mathtt{a}_j)$ holds for all $i < j$, $i, j \in [0, n-1]$.

There exists a step function σ with the step $s \in A$ such that $\mu(\Sigma)$ is of the form $\{\sigma^i(\min(\mu(\Sigma))) \mid i \in [0, |\mu(\Sigma)| - 1]\}$. Consequently the weight of every letter in Σ is $\mu(\mathtt{a}_i) = \sigma^i(\min(\mu(\Sigma)))$ for all $i \in [0, n-1]$. In particular we have $\mu(\mathtt{a}_0) = \min(\mu(\Sigma))$. Now consider some word $w \in \Sigma^*$ and index $l \in [|w|]$, then $f_{w,\mu}(l)$ is realised by some factor $\mathtt{a}_{p_1} \ldots \mathtt{a}_{p_l} \in \mathrm{Fact}_l(w)$ with some sequence $p_1, \ldots, p_l \in [0, n-1]$. Therefore $f_{w,\mu}(l)$ is of the form $\sigma^{p_1}(\mu(\mathtt{a}_0)) \circ \cdots \circ \sigma^{p_l}(\mu(\mathtt{a}_0))$. And similarly $f_{w,\mu}(l-1)$ is realised by some other factor $\mathtt{a}_{q_1} \ldots \mathtt{a}_{q_{l-1}} \in \mathrm{Fact}_{l-1}(w)$ for some sequence $q_1, \ldots, q_{l-1} \in [0, n-1]$, and we have $f_{w,\mu}(l-1) = \sigma^{q_1}(\mu(\mathtt{a}_0)) \circ \cdots \circ \sigma^{l_i}(\mu(\mathtt{a}_0))$. Now let $m = \sum_{i=1}^{l}(p_i)$ and $o = \sum_{i=1}^{l-1}(q_i)$ be the number of steps in the weights $f_{w,\mu}(l)$ and $f_{w,\mu}(l-1)$. So they are the maximum number of steps any of w's factors of length l and $l-1$ can have in their weight.

First of all $m - o \geq 0$ holds because we know the factor-weight function is strictly increasing and otherwise m would not be the maximum for length l. We also know $m - o < n$ holds because if $m - o \geq n$ held, o would not be the maximum number of steps in a factor of length $l-1$. We now know $f_{w,\mu}(l)$ and $f_{w,\mu}(l-1)$ only differ by $k := m - o$ steps where $0 \leq k < n$ holds. Consequently $f_{w,\mu}(l) = f_{w,\mu}(l-1) \circ \sigma^k(\mathtt{a}_0)$ holds and because μ has stepped base weights there exists such a letter $\mathtt{a}_k \in \Sigma$ with $\mu(\mathtt{a}_k) = \sigma^k(\mathtt{a}_0)$. Thus μ is gapfree. ◀

For the converse statement to hold we unfortunately need the additional prerequisite that there exist elements in the chosen monoid that are able to fulfil the property of stepped base weights. For every two weights $a < b$ we need an element x such that $b = a \circ x$ holds. To see that this is necessary, consider for example the alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ and the strictly totally ordered monoid $(A, \circ, \lambda, \prec)$ where A is the set $\{\binom{a}{b} \mid a, b \in \mathbb{N}_0\}$, $\circ$ is the usual addition on vectors, $\lambda = \binom{0}{0}$, and $\prec$ is the order obtained by the lexicographical expansion of the usual *less than* onto vectors. So for example $\binom{0}{0} \prec \binom{0}{2} \prec \binom{1}{1} \prec \binom{2}{0}$ holds. Now the weight measure μ over Σ with the base weights $\mu(\mathtt{a}) = \binom{0}{2}$, $\mu(\mathtt{b}) = \binom{1}{1}$ and $\mu(\mathtt{c}) = \binom{2}{0}$ is gapfree. For seeing this, let $u \in \mathrm{Fact}_{i+1}(w)$ and $v \in \mathrm{Fact}_{i+1}(w)$ determine $f_{w,\mu}(i+1)$ and $f_{w,\mu}(i)$ respectively. If v is a prefix or a suffix of u, the claim holds immediately. Thus consider that v is neither a prefix nor suffix of u; indeed we can also assume that they are overlap-free since the overlap would count for $\mu(u)$ and $\mu(v)$ in the same way. If we fix the different amount of $\mathtt{a}$ and $\mathtt{c}$ between v and u, namely r and t, the amount of $\mathtt{b}$ in both u and v is given by their length and therefore we assume $|u|_\mathtt{b} = 0$. Now by combining $\mu(u) \succ \mu(v)$, $\mu(v) \succ \mu(u[1..|u|-1]), \mu(u[2..|u|])$, and $|v| + 1 = |u|$ we get that $t - r \in \{-1, 0\}$. Evaluating both cases gives us $f_{w,\mu}(i+1) = f_{w,\mu}(i) + \mu(\mathtt{a})$ and $f_{w,\mu}(i+1) = f_{w,\mu}(i) + \mu(\mathtt{b})$ respectively. (A full prove can be found in the appendix in Proposition 41). But even though μ is gapfree, $\mu(\Sigma) = \{\binom{0}{2}, \binom{1}{1}, \binom{2}{0}\}$ can never be of the form $\{\sigma^i(\binom{0}{0}) \mid i \in [0, 2]\}$ for any step function $\sigma(a) = a \circ \binom{x}{y}$. This is the case because there exists no $\binom{x}{y} \in A$ such that $\binom{1}{1} = \binom{0+x}{2+y}$, $\binom{2}{0} = \binom{1+x}{1+y}$, or $\binom{2}{0} = \binom{0+x}{2+y}$ would hold.

▶ **Proposition 28.** *Let μ be a non-binary, gapfree weight measure over Σ w.r.t $(A, \circ, \lambda, \prec)$. If additionally, for all $\mathtt{a}, \mathtt{b} \in \Sigma$ with $\mu(\mathtt{a}) \prec \mu(\mathtt{b})$ there exists an $s \in A$ such that $\mu(\mathtt{b}) = \mu(\mathtt{a}) \circ s$ then μ has stepped base weights.*

Proof. W.l.o.g let μ be injective and let Σ be of the form $\{\mathtt{a}_0, \mathtt{a}_1, \ldots, \mathtt{a}_{n-1}\}$, where the letters are in ascending order of their weight, so $\mu(\mathtt{a}_i) \prec \mu(\mathtt{a}_j)$ holds for all $i < j$, $i, j \in [0, n-1]$.

We prove inductively the existence of $s \in A$ by extending a word that contains the possible combinations of letters that define $f_{w,\mu}$. Consider the word $w = \mathtt{a}_1\mathtt{a}_2\mathtt{a}_0\mathtt{a}_2 \in \Sigma^*$.

Then $f_{w,\mu}(2) = \mu(\mathtt{a}_1\mathtt{a}_2)$ and $f_{w,\mu}(3) = \mu(\mathtt{a}_2\mathtt{a}_0\mathtt{a}_2)$ both hold. Because μ is gapfree there exists some $\mathtt{a}_x \in \Sigma$ with $x \in [0, n-1]$ for which $\mu(\mathtt{a}_2\mathtt{a}_0\mathtt{a}_2) = \mu(\mathtt{a}_1\mathtt{a}_2) \circ \mu(\mathtt{a}_x)$ holds. Consequently $\mu(\mathtt{a}_2\mathtt{a}_0) = \mu(\mathtt{a}_1\mathtt{a}_x)$ holds. And together with $\mu(\mathtt{a}_0) \prec \mu(\mathtt{a}_1) \prec \mu(\mathtt{a}_2)$ follows that $\mu(\mathtt{a}_0) \prec \mu(\mathtt{a}_x) \prec \mu(\mathtt{a}_2)$ holds. In other words we now have $0 < x < 2$. So $x = 1$ and $\mu(\mathtt{a}_2\mathtt{a}_0) = \mu(\mathtt{a}_1\mathtt{a}_1)$ hold. Now by the additional prerequisite there exists an $s \in A$ such that $\mu(\mathtt{a}_1) = \mu(\mathtt{a}_0) \circ s$ and so $\mu(\mathtt{a}_2\mathtt{a}_0) = \mu(\mathtt{a}_1\mathtt{a}_0) \circ s$ holds. Consequently also $\mu(\mathtt{a}_2) = \mu(\mathtt{a}_1) \circ s$ and $\mu(\mathtt{a}_1) = \mu(\mathtt{a}_0) \circ s$ hold.

So we already know that $\sigma(a) = a \circ s$ is the step function we need. And finally, by inductively repeating the argument from above for the words $\mathtt{a}_{1+i}\mathtt{a}_{2+i}\mathtt{a}_{0+i}\mathtt{a}_{2+i} \in \Sigma^*$ for all $i \in [1, n-3]$ it follows that μ has stepped base weights with the step function σ. For instance, from μ being gapfree over the word $\mathtt{a}_2\mathtt{a}_3\mathtt{a}_1\mathtt{a}_2 \in \Sigma^*$ it follows that also $\mu(\mathtt{a}_3) = \mu(\mathtt{a}_2) \circ s$ holds. ◀

▶ **Corollary 29.** *For all natural, non-binary weight measures the additional property holds and consequently for such a weight measure μ we have that it is gapfree iff it has stepped base weights.*

Proof. The claim follows directly by the structure of $\mathbb{N}$. ◀

Following this characterisation we get another generalisation for weight measures.

▶ **Theorem 30.** *For a strictly totally ordered alphabet Σ, all weight measures that are natural, alphabetically ordered, and gapfree behave similarly to one another, i.e. for two such weight measures μ and ν the following holds for all words $v, w \in \Sigma^*$ with $|v| = |w|$*

$$\mu(w) = \mu(v) \text{ iff } \nu(w) = \nu(v) \quad \text{and} \quad \mu(w) < \mu(v) \text{ iff } \nu(w) < \nu(v).$$

Proof. Let μ and ν be alphabetically ordered, natural, gapfree weight measures over the strictly totally ordered alphabet $\Sigma = \{\mathtt{a}_0, \mathtt{a}_1, \ldots, \mathtt{a}_{n-1}\}$ for some $n \in \mathbb{N}$ and with $\mu(\mathtt{a}_i) < \mu(\mathtt{a}_j)$ for all $i < j$, $i, j \in [0, n-1]$. And let $v, w \in \Sigma^*$ with $|v| = |w|$. Then μ and ν are both injective because they both are alphabetically ordered.

In case Σ is binary, so are μ and ν. Therefore if $\mu(v) = \mu(w)$ holds $|v|_{\mathtt{a}_0} = |w|_{\mathtt{a}_0}$ and $|v|_{\mathtt{a}_1} = |w|_{\mathtt{a}_1}$ must hold, and so we have $\nu(v) = \nu(w)$. The converse and the same for "$<$" hold analogously.

In the non-binary case both μ and ν have stepped base weights by Proposition 28. Thus two step functions σ_μ for μ and σ_ν for ν exist. Notice, the number of steps in any letter $\mathtt{a}_i \in \Sigma$ is always i for both of those step functions. Therefore the weights of all words have the same number of steps in their weight regardless of which one of the two weight measures is considered. Now assume $\mu(v) = \mu(w)$ holds, then both words have the same number of steps in their μ-weight. Thus also their ν-weight must have this number of steps and so $\nu(v) = \nu(w)$ holds. Again the converse and the same for "$<$" hold analogously. ◀

Theorem 30 essentially shows that the monoid and the individual base weights of any natural, alphabetically ordered, and gapfree weight measure do not impact the relative behaviour of such weight measures, e.g. the prefix normal form of any word is the same w.r.t. different such natural, alphabetically ordered, and gapfree weight measures. This behaviour leads us to the definition of a generalised weight measure that only depends on the alphabetic order of the letters. By Theorem 30 this *alphabetic weight measure* behaves like every other natural, alphabetically ordered, and gapfree weight measure in most ways.

▶ **Definition 31.** *Let $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_n\}$ be a strictly totally ordered alphabet, with $n \in \mathbb{N}$. We define μ_Σ as the alphabetically ordered sum weight measure with base weights: $\mu(\mathtt{a}_i) = i$ for all $i \in [n]$. We say μ_Σ is the* alphabetic weight measure *for Σ.*

▶ **Lemma 32.** *The alphabetic weight measure is gapfree.*

Proof. Follows directly by Proposition 27 with $\sigma(n) = n + 1$ with $n \in \mathbb{N}$. ◀

▶ Conjecture 1. We conjecture that the statement of Theorem 30 holds not only for natural, alphabetically ordered, and gapfree weight measures (or weight measures with the aforementioned composition property) but also for any alphabetically ordered, gapfree weight measures.

In the remaining part of this section we investigate injective weight measures, specifically we provide a construction that can be used to transform any weight measure into an injective one. This enables us to consider injective weight measures w.l.o.g most of the time.

▶ **Definition 33.** *Let μ be a weight measure over Σ w.r.t. $(A, \circ, \lambda, \prec)$. First, we define the* μ-projected alphabet *as $\Sigma_\mu := \{[\mathtt{a}]_\mu \mid \mathtt{a} \in \Sigma\}$, where $[\mathtt{a}]_\mu$ is the set of all letters with the same weight as $\mathtt{a} \in \Sigma$, i.e. $[\mathtt{a}]_\mu := \{\mathtt{b} \in \Sigma \mid \mu(\mathtt{b}) = \mu(\mathtt{a})\}$ for $\mathtt{a} \in \Sigma$. Then we define μ's* projected weight measure *as the weight measure $\hat{\mu}$ over Σ_μ w.r.t. $(A, \circ, \lambda, \prec)$ and with the base weights $\hat{\mu}([\mathtt{a}]_\mu) = \mu(\mathtt{a})$. Finally for a word $w \in \Sigma^*$ we construct its* μ-projection *$w_\mu \in \Sigma_\mu^*$ with $w_\mu := [w[1]]_\mu \dots [w[|w|]]_\mu$.*

▶ **Lemma 34.** *For a weight measure μ over an alphabet Σ and a word $w \in \Sigma^*$ we have $\hat{\mu}(w_\mu) = \mu(w)$ and the projected weight measure $\hat{\mu}$ is injective on Σ_μ.*

Proof. The first statement holds directly by the construction of the projected weight measure with the base weights $\hat{\mu}([\mathtt{a}]_\mu) = \mu(\mathtt{a})$ for all $\mathtt{a} \in \Sigma$. For the second claim choose $[\mathtt{a}]_\mu, [\mathtt{b}]_\mu \in \Sigma_\mu$ with $\hat{\mu}([\mathtt{a}]_\mu) = \hat{\mu}([\mathtt{b}]_\mu)$. Then $\mu(\mathtt{a}) = \mu(\mathtt{b})$ holds and therefore both $\mathtt{b} \in [\mathtt{a}]_\mu$ and $\mathtt{a} \in [\mathtt{b}]_\mu$ hold. In other words $[\mathtt{a}]_\mu = [\mathtt{b}]_\mu$, so $\hat{\mu}$ is injective on Σ_μ. ◀

▶ Remark 35. With this construction a word w and its μ-projection w_μ for some weight measure μ behave the same way under any functions that are based on the weights of the letters in the words, e.g. $f_{w,\mu} = f_{w_\mu,\hat{\mu}}$, $p_{w,\mu} = p_{w_\mu,\hat{\mu}}$, $\operatorname{maxpos}_{w,\mu} = \operatorname{maxpos}_{w_\mu,\hat{\mu}}$, and $\operatorname{minpos}_{w,\mu} = \operatorname{minpos}_{w_\mu,\hat{\mu}}$ all hold. Analogously, all other statements that depend on those functions follow for the μ-projection words and alphabets. For example μ is gapfree iff $\hat{\mu}$ is gapfree. In some sense the word w_μ represents all words within the set $\{v \in \Sigma^* \mid v[i] \in w[i] \text{ for all } i \in [|w|]\}$.

The following theorem essentially shows that the prefix normal form of a projected word like in Definition 33 represents the set of prefix normal words that are factor-weight equivalent to the original word. In other words, for some $w \in \Sigma^*$ the sets $\mathcal{P}_\mu(w)$ and $\mathcal{P}_{\hat{\mu}}(w_\mu)$ represent the same prefix normal words over Σ that are equivalent to w. Thus, also in the non-injective case we are able to obtain one prefix normal form by considering projections.

▶ **Theorem 36.** *Let μ be a gapfree weight measure over Σ w.r.t. $(A, \circ, \lambda, \prec)$ and $w \in \Sigma^*$. Then with $w' = \mathcal{P}_{\hat{\mu}}(w_\mu)$ we have $\mathcal{P}_\mu(w) = \{v \in \Sigma^* \mid v[i] \in w'[i] \text{ for all } i \in [|w|]\}$.*

Proof. Let w' be the prefix normal form of w's μ-projection w_μ, so $w' := \mathcal{P}_{\hat{\mu}}(w_\mu)$. This is possible because $\hat{\mu}$ is gapfree and injective by Lemma 34. Notice here that $|w| = |w_\mu| = |w'|$ holds by construction. We now show the claim by two subset-proofs.

First, let $v \in \{u \in \Sigma^* \mid u[i] \in w'[i] \text{ for all } i \in [|w|]\}$. To show $v \in \mathcal{P}_\mu(w)$ we show that v is in $[w]_{\sim_\mu}$ and v is μ-prefix normal. Firstly, $f_{v,\mu} = f_{w',\hat{\mu}}$ holds by the choice of v, $f_{w',\hat{\mu}} = f_{w_\mu,\hat{\mu}}$ holds by the choice of w' as the prefix normal form of w_μ, and $f_{w_\mu,\hat{\mu}} = f_{w,\mu}$ holds by the construction of w_μ and $\hat{\mu}$. In total $f_{v,\mu} = f_{w,\mu}$ holds and we have $v \in [w]_{\sim_\mu}$. Secondly, also $p_{v,\mu} = p_{w',\mu'}$ holds by the choice of v and also $p_{w',\hat{\mu}} = f_{w_\mu,\hat{\mu}}$ holds by the choice of w' as the prefix normal form of w_μ. In total $p_{v,\mu} = f_{w_\mu,\hat{\mu}} = f_{w,\mu}$ holds and so v is μ-prefix normal.

For the second part of the proof, let $v \in \mathcal{P}_\mu(w)$. To show $v \in \{u \in \Sigma^* \mid u[i] \in w'[i] \text{ for all } i \in [|w|]\}$ we prove that $v[i] \in w'[i]$ for all $i \in [|v|]$. Let $i \in [|v|]$. By the construction of the prefix normal form (Proposition 21) we know that $v[i] = \mathtt{a}$ for some $\mathtt{a} \in \Sigma$ such that $f_{w,\mu}(i) = f_{w,\mu}(i-1) \circ \mu(\mathtt{a})$ holds. We also know $w'[i] = \mathtt{x}$ for some $\mathtt{x} \in \Sigma_\mu$ such that $f_{w_\mu,\hat{\mu}}(i) = f_{w_\mu,\hat{\mu}}(i-1) \circ \hat{\mu}(\mathtt{x})$ holds. And because $f_{w_\mu,\hat{\mu}} = f_{w,\mu}$ again holds by the construction of w_μ and $\hat{\mu}$, we have $\hat{\mu}(\mathtt{x}) = \mu(\mathtt{a})$. So in total, because $\mathtt{x}$ is the subset of Σ with all the letters of weight $\hat{\mu}(\mathtt{x})$, by the construction of Σ_μ, we have $\mathtt{a} \in \mathtt{x}$, i.e. $v[i] \in w'[i]$. ◀

With Theorem 36 we can also accurately calculate the cardinality of $\mathcal{P}_\mu(w)$ for some word $w \in \Sigma^*$ and a non-injective weight measure μ.

▶ **Corollary 37.** *Let μ be gapfree weight measure over the alphabet Σ, $w \in \Sigma^*$, and $w' = \mathcal{P}_{\hat{\mu}}(w_\mu)$. Then $|\mathcal{P}_\mu(w)| = |w'[1]| * |w'[2]| \dots |w'[|w|]| = \prod_{i=1}^{|w|} |w'[i]|$ holds.*

Proof. Follows directly by Proposition 36. ◀

We conclude this section by revisiting an example w.r.t. the projected weight measure. Again consider the sum weight measure μ over $\Sigma = \{\mathtt{a}, \mathtt{n}, \mathtt{c}, \mathtt{b}\}$ with the base weights $\mu(\mathtt{a}) = 1$, $\mu(\mathtt{n}) = \mu(\mathtt{c}) = 2$, and $\mu(\mathtt{b}) = 3$. Then μ's projected weight measure $\hat{\mu}$ is a weight measure over the alphabet $\Sigma_\mu = \{\{\mathtt{a}\}, \{\mathtt{n}, \mathtt{c}\}, \{\mathtt{b}\}\}$ with the base weights $\hat{\mu}(\{\mathtt{a}\}) = 1$, $\hat{\mu}(\{\mathtt{n}, \mathtt{c}\}) = 2$, and $\hat{\mu}(\{\mathtt{b}\}) = 3$ and we see that $\hat{\mu}$ is injective on Σ_μ. We already know that $\mathtt{nanaba}$ has multiple factor-weight equivalent words that are prefix normal, specifically we have $\mathcal{P}_\mu(\mathtt{nanaba}) = \{\mathtt{banana}, \mathtt{bacana}, \mathtt{banaca}, \mathtt{bacaca}\}$. With the μ-projection of $\mathtt{nanaba}$ being $(\mathtt{nanaba})_\mu = \{\mathtt{n}, \mathtt{c}\}\{\mathtt{a}\}\{\mathtt{n}, \mathtt{c}\}\{\mathtt{a}\}\{\mathtt{b}\}\{\mathtt{a}\}$ this word now has the prefix normal form $\mathcal{P}_{\hat{\mu}}((\mathtt{nanaba})_\mu) = \{\mathtt{b}\}\{\mathtt{a}\}\{\mathtt{n}, \mathtt{c}\}\{\mathtt{a}\}\{\mathtt{n}, \mathtt{c}\}\{\mathtt{a}\}$. All of $\mathtt{nanaba}$'s factor-weight equivalent and prefix normal words are represented by this word when reading it as a non-deterministic concatenation of letters, like shown in Theorem 36. In other words, we have $\mathcal{P}_\mu(\mathtt{nanaba}) = \{v \in \Sigma^* \mid v[i] \in \mathcal{P}_{\hat{\mu}}((\mathtt{nanaba})_\mu)[i], i \in [6]\}$.

5 Subset Prefix Normal Words

In this section we briefly investigate a naïve approach to generalise binary prefix normality and prove that it is already covered by the weight measure approach. The main idea is if Σ is a finite alphabet to take a subset $X \subseteq \Sigma$ and instead of counting the amount of $\mathtt{1}$ or $\mathtt{0}$ resp. we count how many letters in a prefix or factor are contained in X.

▶ **Definition 38.** *Let $w \in \Sigma^*$ and $X \subseteq \Sigma$. We define the* prefix-X-function $p_{w,X}$ *and the* maximum-X-factor function $f_{w,X}$ *by*

$$p_{w,X} : \quad [|w|]_0 \to \mathbb{N}, \quad i \mapsto |\operatorname{Pref}_i(w)|_X \text{ and } f_{w,X} : \quad [|w|]_0 \to \mathbb{N}, \quad i \mapsto \max(|\operatorname{Fact}_i(w)|_X).$$

We say that w is X-prefix normal *(or subset prefix normal w.r.t X) if $p_{w,X} = f_{w,X}$ holds.*

We now show that subset prefix normality is indeed a generalisation of binary prefix normality, and also that subset prefix normality can already be expressed by means of weighted prefix normality. However this is not possible the other way around. So in total we see that weighted prefix normality is more expressive and therefore a more useful generalisation.

▶ **Theorem 39.** *Subset prefix normality is a generalisation of binary prefix normality.*

Proof. W.l.o.g. consider just 1-prefix normality for the binary case. We choose $X \subseteq \Sigma$ with $X = \{1\}$. Then $|w|_1 = |w|_X$ holds for any binary word $w \in \Sigma_2$. It follows that $f_w(i) = \max(|\operatorname{Fact}_i(w)|_1) = \max(|(\operatorname{Fact}_i(w)|_X) = f_{w,X}(i)$ and $p_w(i) = p_{w,X}(i)$ hold for all $i \in [|w|]$. Therefore, w is X-prefix normal if and only if it is 1-prefix normal. So, with such an X every statement on binary prefix normality can be transformed into an analogue using subset prefix normality. ◀

In other words, in the context of the binary alphabet $\{1\}$-prefix normality and 1-prefix normality are the same.

▶ **Theorem 40.** *Weighted prefix normality is a generalisation of subset prefix normality.*

Proof. Let Σ be an alphabet and let $X \subseteq \Sigma$. We construct a sum weight measure μ over Σ. Let $\mu(\mathtt{x}) = 2$ and $\mu(\mathtt{y}) = 1$ for every $\mathtt{x} \in X$ and $\mathtt{y} \in \Sigma \backslash X$. Then $|w|_1 + |w| = \mu(w)$ holds for any word $w \in \Sigma$. It follows that $f_{w,X}(i) + |w| = \max(|Fact_i(w)|_X) + |w| = \max(\mu(Fact_i(w))) = f_{w,\mu}(i)$ and $p_{w,X}(i) + |w| = p_{w,\mu}(i)$ hold for all $i \in [|w|]$. Therefore, w is μ-prefix normal if and only if it is X-prefix normal. So, with such a weight measure every statement on subset prefix normality can be transformed into an analogue using weighted prefix normality. ◀

By Theorem 40 we immediately see that subset prefix normality behaves exactly like weighted prefix normality when using a binary weight measure, which we know by Lemma 22 is gapfree.

6 Conclusions

In this work we presented the generalisation of prefix normality on binary alphabets as introduced by [10] to arbitrary alphabets by applying weights to the letters and comparing the weight of a factor with the weight of the prefix of the same length. Since one of the main properties of binary prefix normality, namely the existence of a unique prefix normal form does not hold in general for weighted prefix normality, we investigated necessary restrictions to obtain a unique prefix normal form even in the generalised setting. Here, it is worth noticing that we did not only generalise the size of the alphabet but also the weights are rather general: they belong to any (totally ordered) monoid. This is of interest because some peculiarities do not occur if $\mathbb{N}$ or $\mathbb{N}_0$ are chosen. In Section 3 we have proved that there always exists a unique prefix normal form if the weight measure is chosen gapfree and injective. We further characterised natural, gapfree weight measures to have stepped base weights. This led to a generalisation of such gapfree weight measures, the alphabetic weight measure, which allows for more convenience when working with weighted prefix normality. We have also shown how in the case of a non-injective weight measure the alphabet and the weight measure can be altered to group the prefix normal forms to obtain a unique representative. In the last section we briefly investigated a naïve approach of generalising the binary prefix normality by subsets of the alphabets and showed that this generalisation can be expressed by weighted prefix normality.

However, the exact behaviour of the weighted prefix normal form, or generally factor-weight equivalent words, regarding changes in their Parikh vectors remains an open problem. Also our Conjecture 1 that all alphabetically ordered and gapfree weight measures behave relatively the same way is an open problem that requires further investigation. Moreover a reconnection of weighted prefix normality to the initial problem of indexed jumbled pattern matching would be of some interest and might prove useful when investigating pattern matching problems w.r.t. a non-binary alphabet.

References

1. Amihood Amir, Timothy Chan, Moshe Lewenstein, and Noa Lewenstein. On hardness of jumbled indexing. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming*, pages 114–125, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.
2. P. Balister and S. Gerke. The asymptotic number of prefix normal words. *Jour. Comb. Theo.*, 2019.
3. Péter Burcsi, Ferdinando Cicales, Gabriele Fici, and Zsuzsanna Lipták. Algorithms for jumbled pattern matching in strings. *International Journal of Foundations of Computer Science*, 23(02):357–374, Feb 2012. doi:10.1142/s0129054112400175.
4. Péter Burcsi, Gabriele Fici, Zsuzsanna Lipták, Rajeev Raman, and Joe Sawada. Generating a Gray code for prefix normal words in amortized polylogarithmic time per word. *CoRR*, abs/2003.03222, 2020. URL: https://arxiv.org/abs/2003.03222.
5. Péter Burcsi, Gabriele Fici, Zsuzsanna Lipták, Frank Ruskey, and Joe Sawada. Normal, abby normal, prefix normal. In *Fun with Algorithms*, pages 74–88, Cham, 2014. Springer International Publishing.
6. Péter Burcsi, Gabriele Fici, Zsuzsanna Lipták, Frank Ruskey, and Joe Sawada. On combinatorial generation of prefix normal words. *Lecture Notes in Computer Science*, page 60–69, 2014. doi:10.1007/978-3-319-07566-2_7.
7. Péter Burcsi, Gabriele Fici, Zsuzsanna Lipták, Frank Ruskey, and Joe Sawada. On prefix normal words and prefix normal forms. *Theoretical Computer Science*, 659:1–13, Jan 2017. doi:10.1016/j.tcs.2016.10.015.
8. Timothy M. Chan and Moshe Lewenstein. Clustered integer 3sum via additive combinatorics. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 31–40, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2746539.2746568.
9. Ferdinando Cicalese, Zsuzsanna Lipták, and Massimiliano Rossi. On infinite prefix normal words. In Barbara Catania, Rastislav Královič, Jerzy Nawrocki, and Giovanni Pighizzini, editors, *SOFSEM 2019: Theory and Practice of Computer Science*, page 122–135, Cham, 2019. Springer International Publishing.
10. Gabriele Fici and Zsuzsanna Lipták. On prefix normal words. In Giancarlo Mauri and Alberto Leporati, editors, *Developments in Language Theory*, pages 228–238, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
11. Pamela Fleischmann, Mitja Kulczynski, Dirk Nowotka, and Danny Bøgsted Poulsen. On collapsing prefix normal words. In Alberto Leporati, Carlos Martín-Vide, Dana Shapira, and Claudio Zandron, editors, *Language and Automata Theory and Applications - 14th International Conference, LATA 2020, Milan, Italy, March 4-6, 2020, Proceedings*, volume 12038 of *Lecture Notes in Computer Science*, pages 412–424. Springer, 2020. doi:10.1007/978-3-030-40608-0_29.
12. OEIS Foundation Inc. The on-line encyclopedia of integer sequencess, 2019. URL: http://oeis.org/.
13. Calvin T. Long. *Elementary introduction to number theory(2nd ed.)*. D. C. Heath and Company, 1972.
14. M. Lothaire. *Combinatorics on Words*. Cambridge Mathematical Library. Cambridge University Press, 1997. doi:10.1017/CBO9780511566097.

Appendix

▶ **Proposition 41.** *Let $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ and let $(A, \circ, \lambda, \prec)$ be the strictly totally ordered monoid where A is the set $\{\binom{a}{b} \mid a, b \in \mathbb{N}_0\}$, $\circ$ is the usual addition on vectors, $\lambda = \binom{0}{0}$, and $\prec$ is the order obtained by the lexicographical expansion of the usual* less than *onto vectors, e.g $\binom{0}{0} \prec \binom{0}{2} \prec \binom{1}{1} \prec \binom{2}{0}$ holds. The weight measure μ over Σ with base weights $\mu(\mathtt{a}) = \binom{0}{2}$, $\mu(\mathtt{b}) = \binom{1}{1}$, and $\mu(\mathtt{c}) = \binom{2}{0}$ is gapfree.*

Proof. For $i \in [|w|]$ let $u \in \mathrm{Fact}_{i+1}(w)$ be the factor determining $f_{w,\mu}(i+1)$ and $v \in \mathrm{Fact}_i(w)$ be the factor determining $f_{w,\mu}(i)$ such that i is minimal with u and v not overlapping (if they overlap, the non-overlapping parts are taken as u and v respectively). Now choose $r, s, t \in \mathbb{Z}$ with $r = |u|_\mathtt{a} - |v|_\mathtt{a}$, $s = |u|_\mathtt{b} - |v|_\mathtt{b}$, and $t = |u|_\mathtt{c} - |v|_\mathtt{c}$. Thus we have $r + s + t = 1$ by

$$r + s + t = |u|_\mathtt{a} - |v|_\mathtt{a} + |u|_\mathtt{b} - |v|_\mathtt{b} + |u|_\mathtt{c} - |v|_\mathtt{c} = |u| - |v| = i + 1 - i = 1.$$

Moreover we have

$$\mu(u) = \binom{|u|_\mathtt{b} + 2|u|_\mathtt{c}}{2|u|_\mathtt{a} + |u|_\mathtt{b}} = \binom{|v|_\mathtt{b} + s + 2|v|_\mathtt{c} + 2t}{2|v|_\mathtt{a} + 2r + |v|_\mathtt{b} + s} = \mu(v) + \binom{s + 2t}{2r + s}.$$

And with $|v|_\mathtt{b} = |u|_\mathtt{b} - s = |u|_\mathtt{b} + r + t - 1$ we get

$$\mu(u) = \binom{|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t}{|u|_\mathtt{b} + 2|v|_\mathtt{a} + 2r} \text{ and } \mu(v) = \binom{|u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}}{2|v|_\mathtt{a} + |u|_\mathtt{b} + r + t - 1}.$$

By evaluating $f_{w,\mu}(i+1) = \mu(u) \succ \mu(v) = f_{w,\mu}(i)$ we get the following two cases: if $|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t = |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ and $|u|_\mathtt{b} + 2|v|_\mathtt{a} + 2r > 2|v|_\mathtt{a} + |u|_\mathtt{b} + r + t - 1$ hold we get $t = r - 1$ and $r > t - 1$, thus $t + 1 = r$. If $|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t > |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ holds we get $t + 1 > r$. Hence, in general we know $t + 1 \geq r$ must hold. Now set $u' = u[2..|u|]$ (the case $u' = u[1..|u| - 1]$ is symmetric). By the assumption that v and u do not overlap we have $\mu(u') \prec \mu(v)$. We now evaluate this inequality in a similar fashion but also considering the three possible letters for $u[1]$.

case 1: $u[1] = \mathtt{a}$
We have $\mu(u') = \binom{|u'|_\mathtt{b} + 2|u'|_\mathtt{c}}{2|u'|_\mathtt{a} + |u'|_\mathtt{b}} = \binom{|u|_\mathtt{b} + 2|u|_\mathtt{c}}{2(|u|_\mathtt{a} - 1) + |u|_\mathtt{b}} = \binom{|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t}{2|v|_\mathtt{a} + 2r - 2 + |u|_\mathtt{b}}$. By $\mu(u') \prec \mu(v)$ we have either $|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t = |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ and $2|v|_\mathtt{a} + 2r - 2 + |u|_\mathtt{b} < 2|v|_\mathtt{a} + |u|_\mathtt{b} + r + t - 1$ which implies $t = r - 1$ and $r - 1 < t$, which is a contradiction, or $|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t < |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ which implies $t < r - 1$, which is a contradiction to $t + 1 \geq r$. Hence we get $u[1] \neq \mathtt{a}$
case 2: $u[1] = \mathtt{b}$
We have $\mu(u') = \binom{|u'|_\mathtt{b} + 2|u'|_\mathtt{c}}{2|u'|_\mathtt{a} + |u'|_\mathtt{b}} = \binom{|u|_\mathtt{b} - 1 + 2|u|_\mathtt{c}}{2|u|_\mathtt{a} + |u|_\mathtt{b} - 1} = \binom{|u|_\mathtt{b} - 1 + 2|v|_\mathtt{c} + 2t}{2|v|_\mathtt{a} + 2r + |u|_\mathtt{b} - 1}$. By $\mu(u') \prec \mu(v)$ we have either $|u|_\mathtt{b} - 1 + 2|v|_\mathtt{c} + 2t = |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ and $2|v|_\mathtt{a} + 2r + |u|_\mathtt{b} - 1 < 2|v|_\mathtt{a} + |u|_\mathtt{b} + r + t - 1$ which gives again a contradiction by $t = r$ and $r < t$, or $|u|_\mathtt{b} - 1 + 2|v|_\mathtt{c} + 2t < |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ which implies $t < r$.
case 3: $u[1] = \mathtt{c}$
We have $\mu(u') = \binom{|u'|_\mathtt{b} + 2|u'|_\mathtt{c}}{2|u'|_\mathtt{a} + |u'|_\mathtt{b}} = \binom{|u|_\mathtt{b} + 2(|u|_\mathtt{c} - 1)}{2|u|_\mathtt{a} + |u|_\mathtt{b}} = \binom{|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t - 2}{2|v|_\mathtt{a} + 2r + |u|_\mathtt{b}}$. By $\mu(u') \prec \mu(v)$ we have here either $|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t - 2 = |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ and $2|v|_\mathtt{a} + 2r + |u|_\mathtt{b} < 2|v|_\mathtt{a} + |u|_\mathtt{b} + r + t - 1$ which leads to the contradiction $t - 1 = r$ and $r < t - 1$, or $|u|_\mathtt{b} + 2|v|_\mathtt{c} + 2t - 2 < |u|_\mathtt{b} + r + t - 1 + 2|v|_\mathtt{c}$ which implies $t < r + 1$.
Hence, in all cases we get $t < r + 1$ and by $t + 1 \geq r$ we know $t = r - 1$ or $t = r$ must hold. We can now prove that μ is gapfree by distinguishing these cases.
case 1: $t = r - 1$

By $r+s+t=1$ we get $s=-2t$ and consequently

$$\begin{aligned} f_{w,\mu}(i+1) &= f_{w,\mu}(i) + \binom{s+2t}{2r+s} = f_{w,\mu}(i) + \binom{0}{2r-2t} \\ &= f_{w,\mu}(i) + \binom{0}{2(r-t)} = f_{w,\mu}(i) + \binom{0}{2} \\ &= f_{w,\mu}(i) + \mu(\mathtt{a}). \end{aligned}$$

case 2: $t=r$
By $r+s+t=1$ we get $s=1-2t$ and consequently

$$\begin{aligned} f_{w,\mu}(i+1) &= f_{w,\mu}(i) + \binom{s+2t}{2r+s} = f_{w,\mu}(i) + \binom{1}{1} \\ &= f_{w,\mu}(i) + \mu(\mathtt{b}). \end{aligned}$$

Thus in both cases exists an $x \in \Sigma$ with $f_{w,\mu}(i+1) = f_{w,\mu}(i) + \mu(x)$. ◀