Séminaire Lotharingien de Combinatoire **85B** (2021)
Article #81, 12 pp.

Proceedings of the 33rd Conference on Formal Power Series and Algebraic Combinatorics (Ramat Gan)

Spectral Analysis of Word Statistics

Chaim Even-Zohar[*1], Tsviqa Lakrec[†2], and Ran J. Tessler[‡3]

[1]*The Alan Turing Institute, London, United Kingdom*
[2]*Einstein Institute of Mathematics, The Hebrew University of Jerusalem, Israel*
[3]*Department of Mathematics, Weizmann Institute of Science, Rehovot, Israel*

Abstract. Given a random text over a finite alphabet, we study the frequencies at which fixed-length words occur as subsequences. As the data size grows, the joint distribution of word counts exhibits a rich asymptotic structure. We investigate all linear combinations of subword statistics, and fully characterize their different orders of magnitude using diverse algebraic tools.

Moreover, we establish the spectral decomposition of the space of word statistics of each order. We provide explicit formulas for the eigenvectors and eigenvalues of the covariance matrix of the multivariate distribution of these statistics. Our techniques include and elaborate on a set of algebraic word operators, recently studied and employed by Dieker and Saliola (2018).

Subword counts find applications in Combinatorics, Statistics, and Computer Science. We revisit special cases from the combinatorial literature, such as intransitive dice, random core partitions, and questions on random walk. Our structural approach describes in a unified framework several classical statistical tests. We propose further potential applications to data analysis and machine learning.

Remark. This is an extended abstract. A preprint of the full paper is available online at `https://arxiv.org/pdf/2012.00742` .

1 Word Statistics

Sequences over a finite alphabet are ubiquitous in pure and applied mathematics, and lie at the core of many probabilistic models. They may represent steps of a random walk, words of group generators, discrete-valued time series, DNA segments, or output of pseudorandom generators, to mention a few examples. In the analysis of such sequences, one often considers various numerical *statistics*, in order to capture their main

[*]chaim@ucdavis.edu C. E. was supported by the Lloyd's Register Foundation / Alan Turing Institute programme on Data-Centric Engineering.

[†]tsviqa@gmail.com T. L. was supported by the ISF grant 891/15 and ERC 2020 grant HomDyn 833423.

[‡]ran.tessler@weizmann.ac.il R. T. (incumbent of the Lillian and George Lyttle Career Development Chair) was supported by the ISF grant No. 335/19 and by a research grant from the Center for New Scientists of Weizmann Institute.

features, extract meaningful information, apply further processing, or take informed decisions. It is hence important to examine general families of such statistics, and thoroughly understand their expected behavior.

Subword counts give rise to a broad family of word statistics, which this work investigates. Given a finite alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \dots\}$, a pattern $u \in \Sigma^k$, and a longer text $w \in \Sigma^n$, we consider $\#u(w)$, the number of occurrences of u as a subsequence of w. The copies of u that we count do not have to appear consecutively in the text, nor to be disjoint. For example $\#\mathtt{fee}(\mathtt{referee}) = 3$. Many well-studied word statistics are special cases of these counts, or finite linear combinations of them.

Randomized models provide a natural setting to investigate words and their statistics. They help us analyze these fundamental objects via typical instances, and guide us in developing relevant tools for applications. Here are two basic models for a random word $w \in \Sigma^n$, that appear naturally in various contexts and applications.

- **One-Sample:** $\mathcal{W}(n, \mathbf{p})$ where $\mathbf{p} = (p_{\mathtt{a}}, p_{\mathtt{b}}, \dots) \in (0,1)^{|\Sigma|}$ and $\sum_{\mathtt{x}} p_{\mathtt{x}} = 1$
  The letters of w are independent, and every letter $w_i = \mathtt{x}$ with probability $p_{\mathtt{x}}$.
- **Multi-Sample:** $\mathcal{W}'(\mathbf{n})$ where $\mathbf{n} = (n_{\mathtt{a}}, n_{\mathtt{b}}, \dots) \in \mathbb{N}^{|\Sigma|}$ and $\sum_{\mathtt{x}} n_{\mathtt{x}} = n$
  Every word w with exactly $n_{\mathtt{x}} = \#\mathtt{x}(w)$ for every $\mathtt{x}$, is equally likely.

The word models $\mathcal{W}$ and $\mathcal{W}'$ parallel the two best-studied random graph models on n labeled vertices. For graphs, $\mathcal{G}(n,p)$ selects every edge with probability p independently, and $\mathcal{G}'(n,m)$ selects exactly m edges uniformly from all possible ways [9]. While the two kinds of models share many asymptotic properties, they differ in some important aspects, especially regarding subgraph counts, and in our models – subword counts.

Remark. This is an extended abstract. A preprint of the full paper is available online at [8], and includes all the details of the constructions, the proofs, and more applications.

2 Spaces of Subword Counts

We start with a general presentation of our approach to subword statistics. Some new results and special cases will be mentioned, but full formal statements are deferred to the subsequent §3.

Let $k \in \mathbb{N}$, and consider the random variables $\#u$, for all k-letter words $u \in \Sigma^k$. For the sake of this general discussion, the distribution of the underlying $w \in \Sigma^n$ may be either $\mathcal{W}(n, \mathbf{p})$ or $\mathcal{W}'(\mathbf{n})$. In the latter model, we let $n_{\mathtt{x}} = p_{\mathtt{x}} n$ and the same general statements apply up to minor changes.

How is the subword count $\#u$ distributed as the text size n grows? By summation over all $\binom{n}{k}$ potential occurrences, one can see that the expected value and variance are

$$\mathbb{E}\left[\#u\right] \;=\; \frac{p_{u_1} \cdots p_{u_k}}{k!}\, n^k \pm O(n^{k-1})\,, \qquad \mathrm{V}\left[\#u\right] \;=\; O\left(n^{2k-1}\right)$$

It follows that the vector of *subword frequencies*, $\#u/\binom{n}{k}$ for $u \in \Sigma^k$, satisfies a law of large numbers:

$$\mathbf{X}_k \;:=\; \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u\in\Sigma^k} \quad \xrightarrow[\text{in probability}]{n\to\infty} \quad \mathbb{E}\left[\mathbf{X}_k\right] \;=\; \mathbf{p}^{\otimes k}$$

It is then natural to study interactions between different subword counts. In general, there is a nonzero correlation between $\#u$ and $\#v$, even in the limit as $n \to \infty$. These correlations are encoded in the following $|\Sigma|^k$-dimensional central limit theorem, as we will see later on.

$$\sqrt{n}\left(\mathbf{X}_k - \mathbb{E}\left[\mathbf{X}_k\right]\right) \quad \xrightarrow[\text{in distribution}]{n\to\infty} \quad \mathcal{N}\left(\mathbf{0},\, \lim_{n\to\infty} n\,\mathrm{Cov}\left[\mathbf{X}_k\right]\right)$$

However, the multivariate Gaussian limit reveals only a small part of the asymptotic picture. It turns out that the rank of the limiting covariance matrix is much lower than $|\Sigma|^k$, so that the limit law is supported on a low-dimensional subspace. In terms of linear combinations of the form $\sum_u f_u \#u$ with $f_u \in \mathbb{R}$, many of those are significantly more concentrated than their individual constituents, and should be scaled differently.

Let $\mathbb{R}\Sigma^k$ denote the space of formal linear combinations of k-letter words over Σ. Every $f = \sum_u f_u u \in \mathbb{R}\Sigma^k$ defines a scalar random variable $\#f$ by linearity. One desirable goal is to find the typical order of magnitude of all $\#f$. The first step in our approach is *grading* the space of all subword combinations. This grading provides an orthogonal decomposition $\mathbb{R}\Sigma^k = \bigoplus_r V_r$ such that $\mathbb{E}[(\#f/n^k)^2] = \Theta(1/n^r)$ for every nonzero $f \in V_r$.

The next goal is to analyze the random variables within each component, that is, $n^{r/2}\mathbf{X}_k$ projected onto V_r. The spaces of word statistics in our models come with natural inner product structures. The most fundamental and most practical objective is a basis of statistics that diagonalizes the covariance matrix of this multivariate distribution, in the spirit of principal component analysis, PCA. Thus, the second step is a *spectral* decomposition of each component V_r.

Having a full explicit decomposition of this form, one can readily obtain the precise leading term of the variance $\mathrm{V}[\#f]$ for any feature f which is a scalar projection of $\mathbf{X}_k$. It lets one identify and compare the various "modes" of the joint distribution, which reveals much of its structure.

Our main contribution is the implementation of this plan. We provide gradings of word statistics by scale, and diagonalizations by second moments, as stated below in §3. These are demonstrated on diverse examples in §5. Several previously-studied word statistics naturally arise as special cases, including some of order smaller than $1/\sqrt{n}$. Also new families of word statistics constructed this way seem to be meaningful and useful.

The analysis of multivariate statistical features of ordered or sequential data is a direct practical application of our work. Linear decompositions of data on combinatorial structures have been studied since the seminal monograph by Diaconis [4, §8], that

introduced the use of algebraic tools such as representations of the symmetric group. However, the crucial issue of choosing bases for components has mostly been left arbitrary, depending on matters of convenience, or ad hoc interpretations. Our proposed approach, which turns to the second moment structure of typical data distributions, aims to provide a systematic treatment that seems very natural from a practical perspective. In fact, the random word models we use make it particularly well-suited for extracting features in the high noise regime.

3 Main Results

We now present the scaling decompositions and the spectral decompositions of the subword statistics of random words. The one-sample model $\mathcal{W}(n, \mathbf{p})$, where letters are independent, is treated in Theorems 1 and 2. Theorems 3 and 4 concern the more involved setting of the multisample model $\mathcal{W}'(\mathbf{n})$, with randomly ordered letters.

All the components can be obtained by straightforward elementary computations, using Gaussian elimination and combinatorial manipulations on words. These constructions are very briefly sketched in the next section, while detailed accounts are available in the full version at [8].

Let w be a random word in the model $\mathcal{W}(n, \mathbf{p})$. Recall that $\Sigma = \{\mathtt{a}, \mathtt{b}, \cdots\}$ is a finite alphabet, so that $d := |\Sigma| \geq 2$, and the characters of w are independent and distributed with $\mathbf{p} = (p_{\mathtt{a}}, p_{\mathtt{b}}, \cdots) \in \mathbb{R}^d$. We count subwords $u \in \Sigma^k$ occurring in $w \in \Sigma^n$, and study the normalized statistic,

$$\bar{\#}u(w) \;:=\; \frac{\#u(w)}{\binom{n}{k}} \;\in\; [0,1]$$

Moreover, we study all linear combinations of the random variables $\bar{\#}u$ for $u \in \Sigma^k$. Every formal sum $f = \sum_u f_u u$ in the d^k-dimensional space $W_k := \mathbb{R}\Sigma^k$ defines such statistics $\#f$ and $\bar{\#}f$ by linearity.

Working with a single length $k \in \mathbb{N}$ is not a real restriction. Indeed, in §2.4 of [8] we define compatible linear embeddings $W_k \hookrightarrow W_{k+1}$. Therefore, every W_k contains all W_j for $j < k$. The space of all subword statistics is thus denoted $W := \bigcup_{k\in\mathbb{N}} W_k$.

In order to establish the scaling of $\#f$ for every $f \in W$, we study the structure of every W_k. In §2.1 of [8] we define a grading on the spaces W_k, which will yield a well-defined grading on W. Every space W_k decomposes into $k+1$ subspaces, denoted

$$\begin{aligned} W_k \;&=\; W_{k0} \oplus W_{k1} \oplus \ldots \oplus W_{kk} \\ \dim W_{kr} \;&=\; \tbinom{k}{r}(d-1)^r \end{aligned}$$

This primary decomposition depends on the probability vector $\mathbf{p}$. The following theorem asserts that it determines the order of magnitude in n of any statistic in W_k, and different components are uncorrelated.

Theorem 1 (Grading under $\mathcal{W}(n,\mathbf{p})$)**.**
Let $k \in \mathbb{N}$ and $r \in \{0,1,\dots,k\}$. For every nonzero statistic $f \in W_{kr}$ there exists $C_{f,\mathbf{p}} > 0$ such that

$$n^r \ \mathbb{E}_{\mathcal{W}(n,\mathbf{p})}\left[(\bar{\#}f)^2\right] \ \xrightarrow[n\to\infty]{} \ C_{f,\mathbf{p}}.$$

Moreover, for every $r' \neq r$ and $f' \in W_{kr'}$, $\mathbb{E}_{\mathcal{W}(n,\mathbf{p})}[\,\bar{\#}f\ \bar{\#}f'\,] = 0$.

Remark. This decomposition also has the property that $W_{k0},\dots,W_{kk}$ are pairwise orthogonal. Here we work with the inner product on W_k, naturally induced from the measure $\mathcal{W}(k,\mathbf{p})$, and denoted $\langle f, f'\rangle_{\mathbf{p}}$.

We further refine each component W_{kr} into $k-r+1$ orthogonal subspaces. For every $k \geq r \geq 1$, the following decomposition is given in §2.3 of [8].

$$\begin{aligned} W_{kr} \ &= \ W_{kr0} \oplus W_{kr1} \oplus \dots \oplus W_{kr(k-r)} \\ \dim W_{krm} \ &= \ \tbinom{r+m-1}{m}(d-1)^r \end{aligned}$$

This secondary decomposition yields a full asymptotic diagonalization of the covariance of W_k, as follows.

Theorem 2 (Spectrum under $\mathcal{W}(n,\mathbf{p})$)**.**
Let $k \in \mathbb{N}$, $r \in \{1,\dots,k\}$, and $m, m' \in \{0,\dots,k-r\}$. For every $f \in W_{krm}$ and $f' \in W_{krm'}$

$$\mathbb{E}_{\mathcal{W}(n,\mathbf{p})}\left[\left(n^{r/2}\,\bar{\#}f\right)\left(n^{r/2}\,\bar{\#}f'\right)\right] \ \xrightarrow[n\to\infty]{} \ \frac{(k!)^2\ \langle f, f'\rangle_{\mathbf{p}}}{(k+m)!(k-r-m)!}$$

In particular, if $m' \neq m$ then this limit is $\langle f, f'\rangle_{\mathbf{p}} = 0$.

In §2.6 of the full version [8] we present a concise and practical description of the spaces W_{krm}, which provides insight into their structure. We establish an explicit isomorphism between W_{krm} and $U_{krm} \otimes (\mathbb{R}^{d-1})^{\otimes r}$, where U_{krm} are spaces of *multivariate orthogonal polynomials on the discrete simplex*.

Remark. As we discuss in §2.12 of [8], if $f \in W_{kr}$ then $\bar{\#}f$ is a so-called *U-statistic of rank r*. This fact provides additional information on the distribution of these random variables.

We now turn to the other model $\mathcal{W}'(\mathbf{n})$ where the random word w has a prescribed *composition* $\mathbf{n} = (n_{\mathbf{a}}, n_{\mathbf{b}}, n_{\mathbf{c}}, \dots)$, meaning $\#\mathbf{x}(w) = n_{\mathbf{x}}$ for every letter $\mathbf{x} \in \Sigma$. Denote the set of such words by $\binom{\Sigma}{\mathbf{n}}$, and denote their length by $n = |\mathbf{n}| := \sum_{\mathbf{x}} n_{\mathbf{x}}$. The number of words in the set $\binom{\Sigma}{\mathbf{n}}$ is the multinomial coefficient $\binom{n}{\mathbf{n}} = n!/(n_{\mathbf{a}}!n_{\mathbf{b}}!\cdots)$, and each one is equally likely in $\mathcal{W}'(\mathbf{n})$.

As before, we count the occurrences of subwords $u \in \Sigma^k$ and analyze the random variables $\#u$, or $\#f$ for linear combinations $f = \sum_u f_u u$. However, in this model it is sufficient to consider words $u \in \binom{\Sigma}{\boldsymbol{\kappa}}$, fixing the composition $\boldsymbol{\kappa} = (k_{\mathbf{a}}, k_{\mathbf{b}}, \dots)$ of u. Indeed,

in §2.7 of [8] we show how subwords of different compositions reduce to this case. We therefore work in the linear space of formal sums of words of composition $\boldsymbol{\kappa}$, denoted $W_{\boldsymbol{\kappa}} = W_{(k_{\mathbf{a}},k_{\mathbf{b}},\dots)} := \mathbb{R}\binom{\Sigma}{\boldsymbol{\kappa}}$. Note that $\dim W_{\boldsymbol{\kappa}} = \binom{k}{\boldsymbol{\kappa}}$ where $k = |\boldsymbol{\kappa}|$. For $u \in \binom{\Sigma}{\boldsymbol{\kappa}}$, a natural choice of normalization is

$$\tilde{\#}u \;:=\; \frac{\#u}{\prod_{\mathbf{x}\in\Sigma}\binom{n_{\mathbf{x}}}{k_{\mathbf{x}}}} \;\in\; [0,1]$$

extended to $\tilde{\#}f$ for linear combinations $f = \sum_u f_u u \in W_{\boldsymbol{\kappa}}$. Without loss of generality we assume $k_{\mathbf{a}} \geq k_{\mathbf{b}} \geq \cdots > 0$ unless stated otherwise.

Our primary decomposition of $W_{\boldsymbol{\kappa}}$ is based on representations of the symmetric group S_k. The space $W_{\boldsymbol{\kappa}}$ admits an action of S_k by reordering all k-letter words in its basis. The implied decomposition of $W_{\boldsymbol{\kappa}}$ as a direct sum of simple S_k representations is well-studied. In §2.8-2.9 of [8], we briefly review this decomposition and use it to describe the following $k - k_{\mathbf{a}} + 1$ components of word statistics.

$$W_{\boldsymbol{\kappa}} \;=\; W_{\boldsymbol{\kappa}0} \oplus W_{\boldsymbol{\kappa}1} \oplus \cdots \oplus W_{\boldsymbol{\kappa}(k-k_{\mathbf{a}})}$$

The next theorem asserts that the word statistics in $W_{\boldsymbol{\kappa}r}$ have order of magnitude $n^{-r/2}$, and that different components $W_{\boldsymbol{\kappa}r}$ and $W_{\boldsymbol{\kappa}r'}$ are asymptotically uncorrelated. By writing $\mathbf{n}/n \to \mathbf{p}$ we denote the assumption that the parameters $\mathbf{n} = (n_{\mathbf{a}}, n_{\mathbf{b}}, \dots)$ grow such that $n_{\mathbf{x}}/n \to p_{\mathbf{x}} > 0$ as $n = |\mathbf{n}| \to \infty$, for every $\mathbf{x}$.

Theorem 3 (Grading under $\mathcal{W}'(n_{\mathbf{a}}, n_{\mathbf{b}}, n_{\mathbf{c}}, \dots)$)**.**
Let $f \in W_{\boldsymbol{\kappa}r}$ be a nonzero statistic of composition $\boldsymbol{\kappa} = (k_{\mathbf{a}}, k_{\mathbf{b}}, \dots)$ where $r \in \{0, \dots, |\boldsymbol{\kappa}| - k_{\mathbf{a}}\}$, and suppose that $\mathbf{n}/n \to \mathbf{p}$. Then, there exists $C'_{f,\mathbf{p}} > 0$ such that

$$n^r \;\mathbb{E}_{\mathcal{W}'(\mathbf{n})}\left[\left(\tilde{\#}f\right)^2\right] \;\xrightarrow[n\to\infty]{}\; C'_{f,\mathbf{p}}\,.$$

Moreover, for every $r' \neq r$ and $f' \in W_{\boldsymbol{\kappa}r'}$

$$\mathbb{E}_{\mathcal{W}'(\mathbf{n})}\left[\left(n^{r/2}\,\tilde{\#}f\right)\left(n^{r'/2}\,\tilde{\#}f'\right)\right] \;\xrightarrow[n\to\infty]{}\; 0\,.$$

Remark. The components $W_{\boldsymbol{\kappa}0}, W_{\boldsymbol{\kappa}1}, W_{\boldsymbol{\kappa}2}, \dots$ are pairwise orthogonal with respect to the standard inner product of $W_{\boldsymbol{\kappa}}$, denoted $\langle -, - \rangle$.

Remark. Similar to the first random model, the random variables $\tilde{\#}f$ are *generalized U-statistics of rank r*. See §2.13 in the full version [8].

The next result elaborates on the two-sample random model $\mathcal{W}'(n_{\mathbf{a}}, n_{\mathbf{b}})$. Here w is a uniformly random word of length $n = n_{\mathbf{a}} + n_{\mathbf{b}}$ with $\#\mathbf{a}(w) = n_{\mathbf{a}}$ and $\#\mathbf{b}(w) = n_{\mathbf{b}}$, and we count all subwords of composition $\boldsymbol{\kappa} = (k_{\mathbf{a}}, k_{\mathbf{b}})$ with $k_{\mathbf{a}} \geq k_{\mathbf{b}} \geq 1$, where $k = |\boldsymbol{\kappa}| = k_{\mathbf{a}} + k_{\mathbf{b}}$. The primary decomposition of $W_{\boldsymbol{\kappa}}$ already gives the components $W_{\boldsymbol{\kappa}r}$ for $r \in \{0, \dots, k_{\mathbf{b}}\}$.

The full decomposition of $W_{\boldsymbol{\kappa}}$ is defined in §2.11 in [8], where we refine every $W_{\boldsymbol{\kappa}r}$ into $(k-2r+1)r$ orthogonal subspaces as follows:

$$W_{\boldsymbol{\kappa}r} \;=\; \bigoplus_{i=0}^{k-2r} \bigoplus_{j=0}^{r-1} W_{\boldsymbol{\kappa}rij} \qquad\qquad r \in \{1,\dots,k_{\mathbf{b}}\}$$

$$\dim W_{\boldsymbol{\kappa}rij} \;=\; \frac{(k-2r-i+j+1)\,(k-i-j-2)!}{(k-i-r)!\,(r-j-1)!}$$

We do not consider the case $r=0$, because $W_{\boldsymbol{\kappa}0}$ is simply the 1-dimensional space of constant statistics.

This decomposition yields the following full asymptotic diagonalization of the covariance matrix. In writing $f \in W_{\boldsymbol{\kappa}rij}$ it is implied that r,i,j are any numbers in the applicable ranges $r \in \{1,\dots,k_{\mathbf{b}}\}$, $i \in \{0,\dots,k-2r\}$, and $j \in \{0,\dots,r-1\}$, where as usual $k = k_{\mathbf{a}} + k_{\mathbf{b}}$ and $n = n_{\mathbf{a}} + n_{\mathbf{b}}$.

Theorem 4 (Spectrum under $\mathcal{W}'(n_{\mathbf{a}}, n_{\mathbf{b}})$)**.**
Let $\boldsymbol{\kappa} = (k_{\mathbf{a}}, k_{\mathbf{b}})$. For every two word statistics $f \in W_{\boldsymbol{\kappa}rij}$ and $f' \in W_{\boldsymbol{\kappa}r'i'j'}$

$$\mathbb{E}_{\,w\in\mathcal{W}(n_{\mathbf{a}},n_{\mathbf{b}})}\left[\left(\left(\tfrac{n_{\mathbf{a}}n_{\mathbf{b}}}{n}\right)^{r/2}\tilde{\#}f\right)\left(\left(\tfrac{n_{\mathbf{a}}n_{\mathbf{b}}}{n}\right)^{r'/2}\tilde{\#}f'\right)\right] \quad\xrightarrow[n_{\mathbf{a}},n_{\mathbf{b}}\to\infty]{}\quad \Lambda_{\boldsymbol{\kappa}rij}\left\langle f, f'\right\rangle$$

where

$$\Lambda_{\boldsymbol{\kappa}rij} \;:=\; \frac{(k_{\mathbf{a}}!)^2\,(k_{\mathbf{b}}!)^2\,(k-2r)!\,(k-2r+1)!}{(k_{\mathbf{a}}-r)!\,(k_{\mathbf{b}}-r)!\,i!\,(2k-r-i-j)!\,(k-2r+1+j)!}$$

In particular, if $(r',i',j') \neq (r,i,j)$ then this limit is $\langle f, f'\rangle = 0$.

Remark. This spectral decomposition of $W_{\boldsymbol{\kappa}r}$ does not depend on $p_{\mathbf{a}}$ and $p_{\mathbf{b}}$, if these are respectively the limits of $n_{\mathbf{a}}/n$ and $n_{\mathbf{b}}/n$ as in Theorem 3. This remarkable property is not true in general, in the case of three samples or more.

In fact, the limit in this theorem is taken with respect to any $n_{\mathbf{a}}$ and $n_{\mathbf{b}}$ such that $\min(n_{\mathbf{a}}, n_{\mathbf{b}}) \to \infty$. This is a relaxation of the assumption of Theorem 3 that $n_{\mathbf{x}}/n$ converges to a positive constant $p_{\mathbf{x}}$ for every $\mathbf{x} \in \Sigma$. For this reason, the formulation of Theorem 4 restates the case $r \neq r'$.

4 Decompositions

The purpose of this short section is to give some general idea what goes into the decompositions of Theorems 1-4. Again, all the details are available in the full version [8].

In Theorem 1, the primary components W_{kr}, giving the scaling of the statistics in W_k, rely on an orthogonal decomposition $\mathbb{R}\Sigma = \mathbb{R}\mathbf{1} \oplus (\mathbf{1}^{\perp})$ depending on $\mathbf{p}$, and the structure

it induces on $W_k \cong (\mathbb{R}\Sigma)^{\otimes k}$. In Theorem 3, the primary decomposition of $W_{\boldsymbol{\kappa}}$ uses the structure induced by the S_k action, where the different components $W_{\boldsymbol{\kappa} r}$ correspond to submodules of the different Young diagram "widths".

The secondary decompositions that diagonalize the covariances are defined in terms of a set of linear operators in the words algebra, that has been studied and developed in a recent line of work [18, 5]. The operators ∂, $\sqcup\!\sqcup$, $\mathcal{L}$, and Θ respectively perform deletion, insertion, lifting, and replacement, of letters in formal sums of words. The decompositions of Theorems 2 and 4 can be expressed as

$$\begin{aligned} W_{krm} &= \sqcup\!\sqcup_{\mathtt{1}}^{k-r-m} \ker\left(\partial_{\mathtt{1}}\big|_{W_{(r+m)r}}\right) \\ W_{\boldsymbol{\kappa} rij} &= \Theta_{\mathtt{ab}}^{k_b-r}\, \mathcal{L}_{\mathtt{b}}^{j}\, \sqcup\!\sqcup_{\mathtt{a}}^{i} \ker\left(\partial_{\mathtt{a}}\big|_{W_{(k_{\mathtt{a}}+k_{\mathtt{b}}-r-i,r-j),r-j}}\right) \end{aligned}$$

The relation of operators in the words algebra to terms of a certain order in the covariance of subword statistics is not immediate, and the translation requires considerable combinatorial effort. The resulting expressions for the leading terms are often quite intricate, and our analysis combines the existing knowledge on the words algebra and representations with some new techniques and ideas.

5 Examples

We list a variety of examples for subword statistics, as special cases of our treatment. We examine how they are scaled and classified by the scheme of Theorems 1-4. All the computations of the decompositions and second moments are straightforward from the definitions and can be done automatically.

We keep the discussion brief as our main purpose is not to study these particular examples, but demonstrate how various statistics from diverse contexts unify under one framework. Nevertheless, in several cases our perspective sheds a new light on them, or points to potential generalizations. The full version at [8] discusses additional examples from the combinatorial literature, such as the *discrete Lévy area* of a random walk on the grid, and the problem of *intransitive dice*.

Example 1. *Warm Up: Coin Flips*

A sequence of n tosses of a fair coin gives a word in $\{\mathtt{H},\mathtt{T}\}^n$, distributed by $\mathcal{W}(n,(\frac{1}{2},\frac{1}{2}))$. The decomposition for $k=1$ gives $W_{10} = \mathrm{span}\{\mathtt{H}+\mathtt{T}\}$ and $W_{11} = \mathrm{span}\{\mathtt{H}-\mathtt{T}\}$. As Theorem 1 claims, the former yields $\bar{\#}(\mathtt{H}+\mathtt{T}) \equiv 1$ of constant order. The latter, of order $1/\sqrt{n}$, is the "observed bias" of the coin under the fairness hypothesis.

Computing for $k=2$: $W_{20} = \mathrm{span}\{\mathtt{HH}+\mathtt{HT}+\mathtt{TH}+\mathtt{TT}\}$, $W_{210} = \mathrm{span}\{\mathtt{HH}-\mathtt{TT}\}$, $W_{211} = \mathrm{span}\{\mathtt{HT}-\mathtt{TH}\}$, $W_{220} = \mathrm{span}\{\mathtt{HH}+\mathtt{TT}-\mathtt{HT}-\mathtt{TH}\}$.

The first two come from W_{10} and W_{11} via the embedding $W_1 \hookrightarrow W_2$. The new statistic $\bar{\#}(\mathtt{HT} - \mathtt{TH})$ may be interpreted as the tendency of tails to occur after heads. It also scales as $1/\sqrt{n}$, but Theorem 2 implies that its variance is $\frac{1}{3}$ of that of $\bar{\#}(\mathtt{HH} - \mathtt{TT})$, and these two statistics are uncorrelated. By §2.13 in [8], their joined distribution is asymptotically binormal. The fourth statistic scales as $1/n$ and leads to the next example.

Example 2. *Pearson's χ^2 Test Statistic*

The following holds up to a *constant* correction of smaller order in n:

$$\bar{\#}(\mathtt{HH} + \mathtt{TT} - \mathtt{HT} - \mathtt{TH}) \;=\; 2\bar{\#}(\mathtt{HH} + \mathtt{TT}) - 1 \;\approx\; \frac{(\bar{\#}\mathtt{H} - 0.5)^2}{0.5} + \frac{(\bar{\#}\mathtt{T} - 0.5)^2}{0.5}$$

This is the classical Pearson's χ^2 test statistic for fitting the frequencies of $\mathtt{H}$ and $\mathtt{T}$ to the distribution $(0.5, 0.5)$ [15]. This fact extends to any finite-dimensional distribution vector $\mathbf{p}$. The combination $\sum_{\mathbf{x}} \bar{\#}\mathbf{x}\mathbf{x}/p_{\mathbf{x}} - 1$, which is essentially Pearson's χ^2 statistic, always lies in W_{220}.

Example 3. *Functions on the Boolean Hypercube*

Consider a binary stream $w \in \{\mathtt{0}, \mathtt{1}\}^n$, distributed with $\mathcal{W}(n, (p, q))$. The subword statistics of w correspond to $\mathbb{R}\{\mathtt{0}, \mathtt{1}\}^k$, or equivalently Boolean functions $f : \{\mathtt{0}, \mathtt{1}\}^k \to \mathbb{R}$, so they take the form $\sum_u f(u)\,\bar{\#}u$. The primary decomposition of $\mathbb{R}\{\mathtt{0}, \mathtt{1}\}^k$ follows the so-called "slices" of the Fourier basis of Boolean functions. Namely, we expand all the "monomials" with $k - r$ times $(\mathtt{0} + \mathtt{1})$ and r times $(q\,\mathtt{0} - p\,\mathtt{1})$ to obtain $\binom{k}{r}$ combinations that span W_{kr}. For example, the expansion of $(q\,\mathtt{0} - p\,\mathtt{1})^k$ from W_{kk} gives the most concentrated statistic, with variance $\sim k!/n^k$ by Theorems 1-2. For $p = q = \frac{1}{2}$, it is the bias of the parities of k-bit subwords of w. The diagonalization in each slice introduces a finer decomposition into orthogonal subspaces, corresponding to the above-mentioned special orthogonal polynomials. For example, in order $1/\sqrt{n}$ we obtain the basis $\{P_i(1)f_1 + \cdots + P_i(k)f_k\}_{0 \le i < k}$, where $f_1, \ldots, f_k$ are so-called "dictatorship" functions, and $P_i(x)$ are the orthogonal polynomials of the uniform measure on $\{1, \ldots, k\}$. As n grows, these k statistics tend to independent Gaussian distributions.

Example 4. *Two-Sample Statistical Tests*

Consider two real-valued samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ drawn independently from unknown continuous distributions, denoted by the random variables X and Y. The relative order of the observations induces a word w over $\{\mathtt{x}, \mathtt{y}\}$ of length $n + m$. For example, if $X_2 < Y_3 < Y_1 < X_1 < Y_2$ then $w = \mathtt{xyyxy}$. If the two distributions coincide, $X \sim Y$, then w is exactly as in the random model $\mathcal{W}'(n, m)$. This is the null hypothesis of several nonparametric tests for comparing two distributions.

Persson [16] represents several two-sample test statistics in terms of subword counts in w. We review these statistics below.

Mann–Whitney U [14]. This test statistic, $U = \#\mathbf{yx}$ estimates how much $P(Y < X)$ deviates from $1/2$ for randomly selected X and Y. The equivalent combination $u = (\mathbf{yx} - \mathbf{xy})/2$ lies in the component $W_{\boldsymbol{\kappa}1}$ where $\boldsymbol{\kappa} = (1,1)$. The null distribution of $\tilde{\#}u$ is asymptotically normal with variance $\frac{1}{12}(\frac{1}{m} + \frac{1}{n})$, reproduced by Theorems 3-4.

Cramér–von Mises criterion [11]. The above U might fail to detect $X \not\sim Y$ when the probability of $X < Y$ happens to be exactly $1/2$. However, given four independent replications X, X', Y, and Y', the probability that $\max(X, X') < \min(Y, Y')$ or $\max(Y, Y') < \min(X, X')$ is $1/3$ if and only if $X \sim Y$. Otherwise, it is greater than $1/3$ by an L^2 difference between the distribution functions F_X and F_Y. This difference can be estimated by $2\,\tilde{\#}t$ for the following centralized combination in $W_{(2,2)}$:

$$t \;=\; \tfrac{1}{3}(\mathbf{xxyy} + \mathbf{yyxx}) - \tfrac{1}{6}(\mathbf{xyyx} + \mathbf{yxxy} + \mathbf{xyxy} + \mathbf{yxyx})$$

Theorems 3-4 give $t \in W_{(2,2)201}$ and $\mathrm{V}[\tilde{\#}t] \;\sim\; \frac{1}{45}(\frac{1}{m} + \frac{1}{n})^2$ in agreement with [1].

Watson's U^2 [19]. Now suppose that $\{X_i\}$ and $\{Y_i\}$ are samples on the circle S^1. In this case, the previous test for $X \sim Y$ depends on an arbitrary choice of a starting point. Another notion of difference by Watson can be estimated by $\tilde{\#}s$, for the following *rotation invariant* combination.

$$s \;=\; \tfrac{1}{12}(\mathbf{xxyy} + \mathbf{yyxx} + \mathbf{xyyx} + \mathbf{yxxy}) - \tfrac{1}{6}(\mathbf{xyxy} + \mathbf{yxyx})$$

This is not a principal direction of the covariance, but $s = v + \frac{1}{4}t$ for $v \in W_{(2,2)200}$ and so $W_{(2,2)2} = \mathrm{span}\{s, t\}$. By Theorem 4, $\mathrm{V}[\tilde{\#}v] \sim \frac{1}{720}(\frac{1}{m} + \frac{1}{n})^2$, so $\mathrm{V}[\tilde{\#}s] \sim \frac{1}{360}(\frac{1}{m} + \frac{1}{n})^2$.

Remark. A surprising connection between Watson's test statistic and the size distribution of (s, t)-core partitions has been discovered in [6], and further unfolded in [7] based on this perspective of subword counts.

We mention the possibility of similarly analyzing Cramér–von Mises type tests for the classical K-sample problem [10, 17]. It is also possible to study such functionals with higher L_p norms of $(F_X - F_Y)$ as word statistics. This may be interesting because the infinity norm gives another popular two-sample test by Kolmogorov–Smirnov.

Example 5. *Path Signature and Machine Learning*

Finally, we describe a potential application to machine learning, which will be investigated in future work. In many application areas the data takes the form of a long random-like text over a finite alphabet. This may either be a stream of symbols, that comes with a natural ordering or "time" parameter, or a mixture of d samples of real-valued data points, as in Example 4. Suppose that one wishes to classify, model, estimate a parameter, or learn a function of such sequences, say by applying a neural network.

Then, the input sequence first has to be summarized as a vector of *characteristic features*, of reasonable length.

The *signature method* is a generic way of extracting feature sets for sequential data. The basic idea is to embed the data as a path $[0,1] \to \mathbb{R}^d$, and then to use features from its *signature*, which is the graded sequence of its iterated integrals. The coordinates of the signature are definite integrals of the path $(x_t, y_t, \dots)_{0\le t\le 1}$ such as $\int_t dx_t$, $\int_t dy_t$, $\int_{t<s} dx_t dx_s$, $\int_{t<s} dx_t dy_s$, and so on. This method has achieved success in several recent machine learning applications to financial data, clinical symptoms, handwriting recognition, and more [12, 3, for overviews]. The notion of path signature originates in the fundamental theory of rough paths [2, 13].

Though the signature method has mostly been applied to vector-valued time series and spatial data, also a text over d symbols naturally embeds as a path in $\mathbb{R}^d$. Every appearance of a letter **x** contributes a unit step along the axis that corresponds to **x**. The signature of the resulting path is essentially the set of subword statistics in the given text, where the kth level corresponds to subwords of k letters.

Now, our results on the diagonalization of the space of subword statistics provide a suitable choice of basis for feature selection in the signature. Such a basis may be crucial for addressing several important challenges, such as how and where to truncate the coordinates of the signature, how to adjust input parameters in specific applications, how to interpret the contribution of the various characteristic features, etc.

One prediction we would like to make is that our suggested basis of attributes will actually be most beneficial for *highly noisy* data, since our decomposition diagonalizes the joint distribution under randomness. Then, it seems particularly preferable to use a basis of uncorrelated features that distinguishes between statistics that scale differently with the data length.

References

[1] T. W. Anderson. "On the distribution of the two-sample Cramér-von Mises criterion". *The Annals of Mathematical Statistics* (1962), pp. 1148–1159.

[2] K.-T. Chen. "Integration of paths – A faithful representation of paths by noncommutative formal power series". *Transactions of the American Mathematical Society* **89**.2 (1958), pp. 395–407.

[3] I. Chevyrev and A. Kormilitzin. "A primer on the signature method in machine learning" (2016). arXiv:1603.03788.

[4] P. Diaconis. "Group representations in probability and statistics". *Lecture notes – monograph series* **11** (1988), pp. i–192.

[5] A. B. Dieker and F. V. Saliola. "Spectral analysis of random-to-random Markov chains". *Advances in Mathematics* **323** (2018), pp. 427–485.

[6] S. B. Ekhad and D. Zeilberger. "Explicit expressions for the variance and higher moments of the size of a simultaneous core partition and its limiting distribution" (2015). arXiv:1508.07637.

[7] C. Even-Zohar. "Sizes of Simultaneous Core Partitions" (2020). arXiv:2003.13671.

[8] C. Even-Zohar, T. Lakrec, and R. J. Tessler. "Spectral Analysis of Word Statistics" (2020). arXiv:2012.00742.

[9] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*. Vol. 45. John Wiley & Sons, 2011.

[10] J Kiefer. "K-sample analogues of the Kolmogorov–Smirnov and Cramér–v. Mises tests". *The Annals of Mathematical Statistics* (1959), pp. 420–447.

[11] E. L. Lehmann. "Consistency and unbiasedness of certain nonparametric tests". *The annals of mathematical statistics* (1951), pp. 165–179.

[12] D. Levin, T. Lyons, and H. Ni. "Learning from the past, predicting the statistics for the future, learning an evolving system" (2013). arXiv:1309.0260.

[13] T. J. Lyons. "Differential equations driven by rough signals". *Revista Matemática Iberoamericana* **14**.2 (1998), pp. 215–310.

[14] H. B. Mann and D. R. Whitney. "On a test of whether one of two random variables is stochastically larger than the other". *The annals of mathematical statistics* (1947), pp. 50–60.

[15] K. Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**.302 (1900), pp. 157–175.

[16] T. Persson. "A new way to obtain Watson's U^2". *Scandinavian Journal of Statistics* (1979), pp. 119–122.

[17] M. L. Puri. "Some distribution-free k-sample rank tests of homogeneity against ordered alternatives". *Communications on Pure and Applied Mathematics* (1965).

[18] V. Reiner, F. Saliola, and V. Welker. *Spectra of symmetrized shuffling operators, Vol. 228, No. 1072*. American Mathematical Society, 2014.

[19] G. S. Watson. "Goodness-of-fit tests on a circle. II". *Biometrika* **49**.1/2 (1962), pp. 57–63.