Some Random Words

Chaim Even-Zohar

Technion
Israel Institute of Technology

Tsviqa Lakrec

Universität Zürich UZH

Ran Tessler

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Random Word Models

Length n, finite $\Sigma=\{a,b,c,\ldots\}$

Random Word Models

Length n, finite $\Sigma=\{a,b,c,\ldots\}$

1. $W(n, (p_a,p_b,p_c,\ldots))$ iid

Random Word Models

Length n, finite $\Sigma=\{a,b,c,\ldots\}$

1. $W(n,(p_a,p_b,p_c,\ldots))$ iid

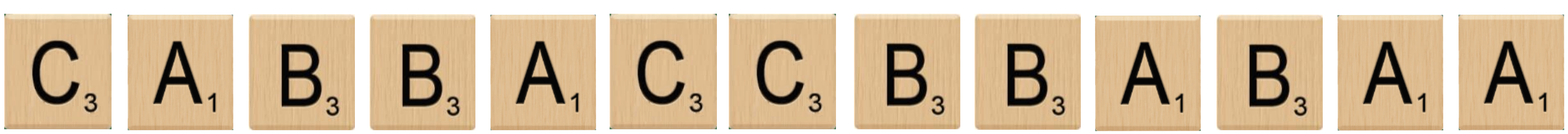
2.

Random Word Models

Length n, finite $\Sigma=\{a,b,c,\ldots\}$

1. $W(n,(p_a,p_b,p_c,\ldots))$ iid

2. $W'(n_a,n_b,n_c,\ldots)$ any order

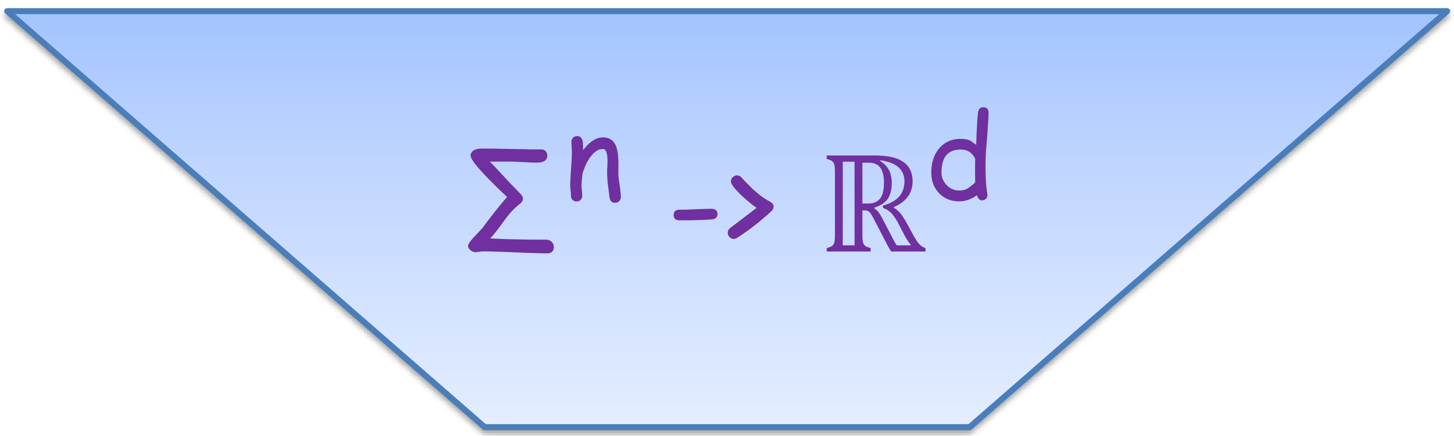


Word Statistics

CABCAABBACAACBCBBACBBBCCBBCABC

Word Statistics

CABCAABBACAACBCBBACBBBCCBBCABC



Statistics / Features

Subwords

CABCAABBACAACBCBBACBBBCCBBCABC

Subwords

CABCA<u>A</u>BBACAACBCBBACB<u>BB</u>CCBBC<u>A</u>BC

A B B A

Subwords

CABCAABBACAACBCBBACBBBCCBBCABC

Subword Statistics

#u(w) = how many occurrences of u in w

#FA(ALFALFA) = 3

Subword Statistics

#u(w) = how many occurrences of u in w

#FA(ALFALFA) = 3

ALFALFA

ALFALFA

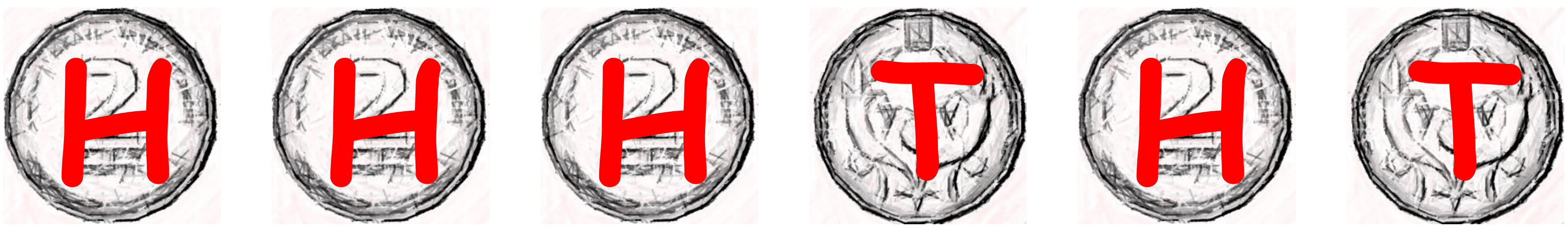
ALFALFA

Subword Spaces

Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$

Subword Spaces

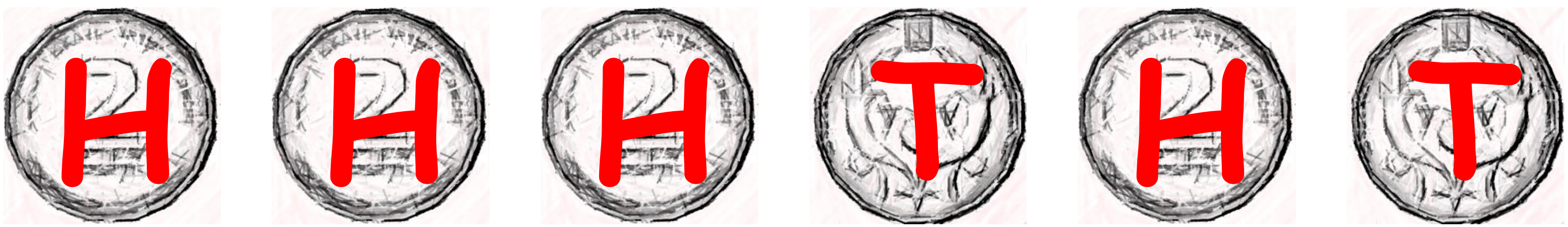
Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$



$$X_2 = \frac{1}{\binom{n}{2}} \begin{pmatrix} \#HH \\ \#HT \\ \#TH \\ \#TT \end{pmatrix} =$$

Subword Spaces

Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$



$$X_2 = \frac{1}{\binom{n}{2}} \begin{pmatrix} \#HH \\ \#HT \\ \#TH \\ \#TT \end{pmatrix} = \begin{pmatrix} 6/15 \\ 7/15 \\ 1/15 \\ 1/15 \end{pmatrix}$$

In Random Words

Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$

LLN: $X_k \xrightarrow[\text{in prob}]{n \to \infty} E[X_k]$

In Random Words

Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$

LLN: $X_k \xrightarrow[\text{in prob}]{n \to \infty} E[X_k]$

CLT: $\sqrt{n}\,(X_k - E[X_k])$

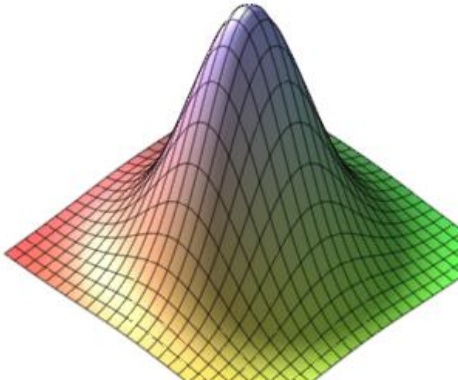
In Random Words

Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$

LLN: $X_k \xrightarrow[\text{in prob}]{n \to \infty} E[X_k]$

CLT: $\sqrt{n}\,(X_k - E[X_k]) \xrightarrow[\text{in dist}]{n \to \infty} \mathcal{N}(0, C)$

$C = \lim_{n \to \infty} (n\, COV[X_k])$



In Random Words

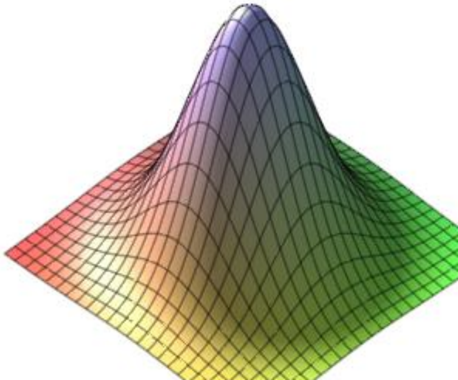
Subword frequencies: $X_k = \left\{ \frac{\#u}{\binom{n}{k}} \right\}_{u \in \Sigma^k}$

LLN: $X_k \xrightarrow[\text{in prob}]{n \to \infty} E[X_k]$

CLT: $\sqrt{n}\,(X_k - E[X_k]) \xrightarrow[\text{in dist}]{n \to \infty} \mathcal{N}(0, C)$

$C = \lim_{n \to \infty} (n\,\mathrm{COV}[X_k])$

rank $C = (|\Sigma| - 1)k$



Subword Spaces

Formal sums: $W_k = \mathbb{R}\Sigma^k$

Subword Spaces

Formal sums: $W_k = \mathbb{R}\Sigma^k$

Example: $W_2 = \mathbb{R}\{HH, HT, TH, TT\}$

$f = HH - HT - TH + TT \in W_2$

Smaller Order Stats

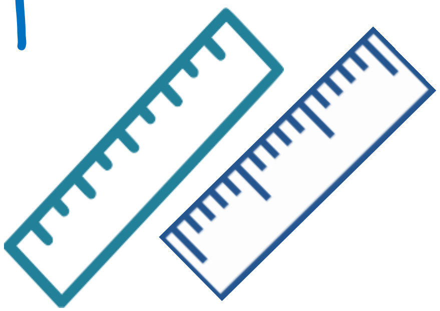
Formal sums: $W_k = \mathbb{R}\Sigma^k$

Example: $W_2 = \mathbb{R}\{HH, HT, TH, TT\}$

$f = HH - HT - TH + TT \in W_2$

Here $\#f = o(1/\sqrt{n})$ in prob by CLT

Actually $\#f$ scales as $1/n$

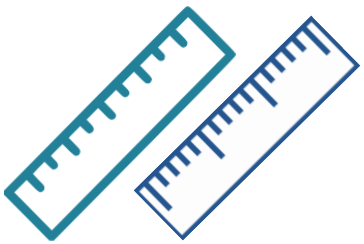


Goals

I.

Goals

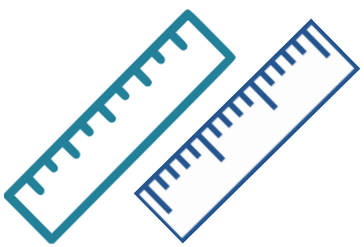
I. Scale every statistic $f \in W_k := \mathbb{R}\{a, b, \ldots, c\}^k$



$$n^r E[(\#f)^2] \rightarrow C_f > 0$$

Goals

I. Scale every statistic $f \in W_k := \mathbb{R}\{a,b,\dots,c\}^k$

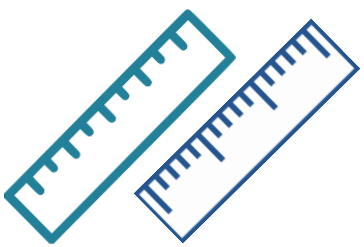


$n^r E[(\#f)^2] \to C_f > 0$

grading: $W_k = \bigoplus_r W_{kr}$

Goals

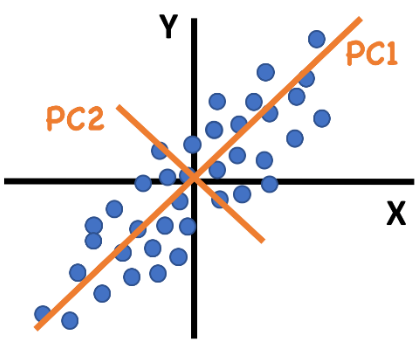
I. Scale every statistic $f \in W_k := \mathbb{R}\{a,b,\ldots,c\}^k$



$$n^r E[(\#f)^2] \to C_f > 0$$

grading: $W_k = \bigoplus_r W_{kr}$

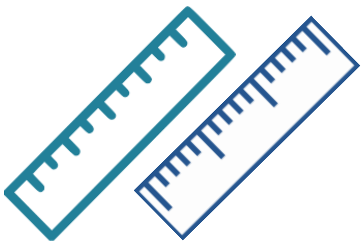
II. Diagonalize the covariance in every W_{kr}



$$n^r E[\#f_i \, \#f_j] \to \lambda_i \delta_{ij}$$

Goals

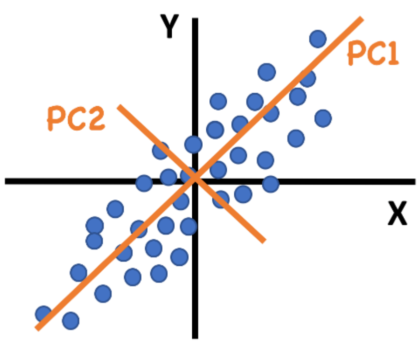
I. Scale every statistic $f \in W_k := \mathbb{R}\{a,b,\ldots,c\}^k$



$$n^r E[(\#f)^2] \to C_f > 0$$

grading: $W_k = \bigoplus_r W_{kr}$

II. Diagonalize the covariance in every W_{kr}



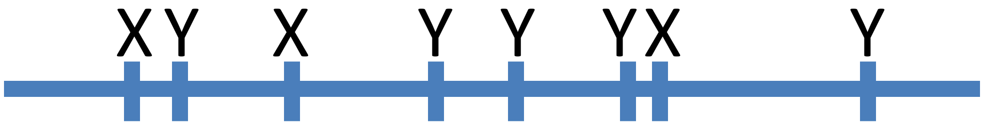
$$n^r E[\#f_i \, \#f_j] \to \lambda_i \delta_{ij}$$

components: $W_{kr} = \bigoplus_i W_{kri}$

Word Patterns Everywhere

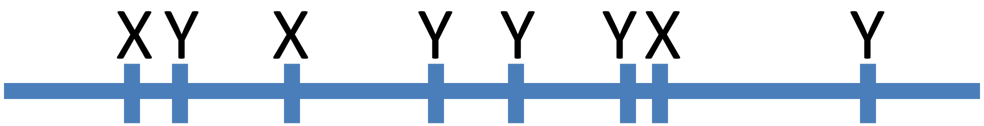
Word Patterns Everywhere

2-sample Statistical Tests #XY, #XYX+#YXY

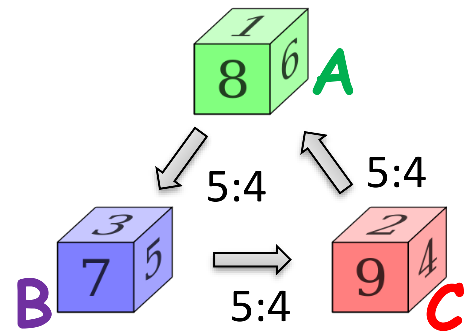


Word Patterns Everywhere

2-sample Statistical Tests #XY, #XYX+#YXY

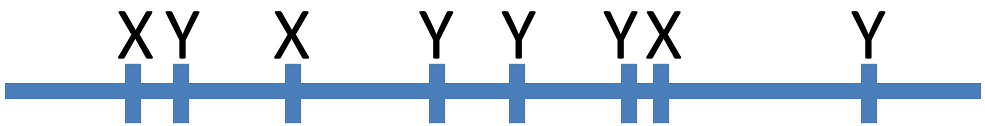


Intransitive Dice #AB + #BC + #CA

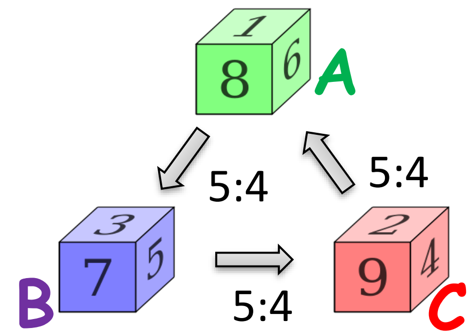


Word Patterns Everywhere

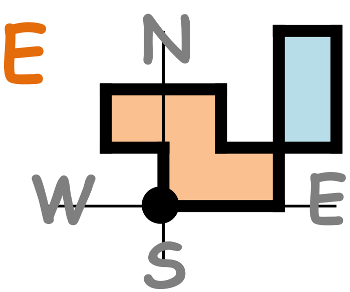
2-sample Statistical Tests #XY, #XYX+#YXY



Intransitive Dice #AB + #BC + #CA

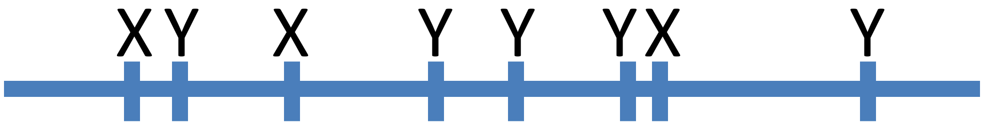


Discrete Lévy Area #NE-#NW+#SW-#SE

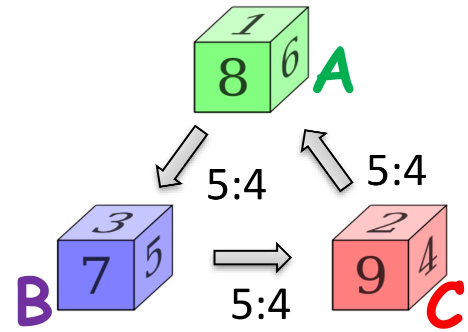


Word Patterns Everywhere

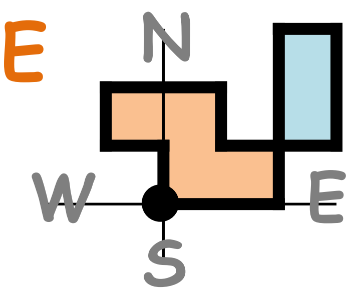
2-sample Statistical Tests #XY, #XYX+#YXY

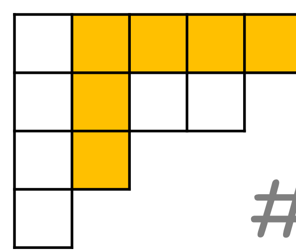


Intransitive Dice #AB + #BC + #CA



Discrete Lévy Area #NE-#NW+#SW-#SE

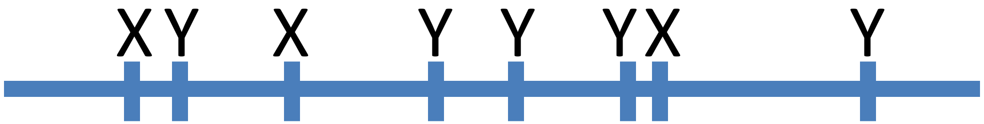


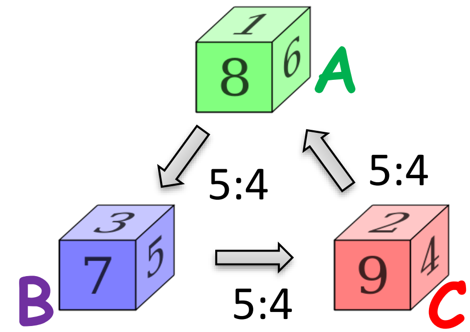


Simultaneous Core Partitions #STST+#TSTS

Word Patterns Everywhere

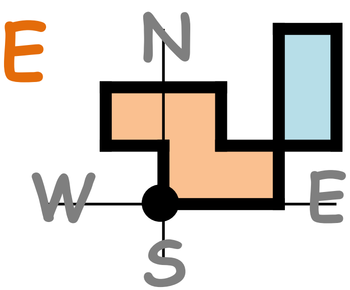
2-sample Statistical Tests #XY, #XYX+#YXY

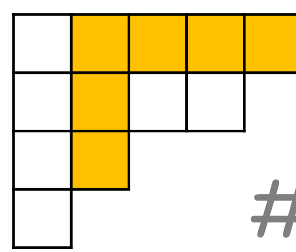




Intransitive Dice #AB + #BC + #CA

Discrete Lévy Area #NE-#NW+#SW-#SE



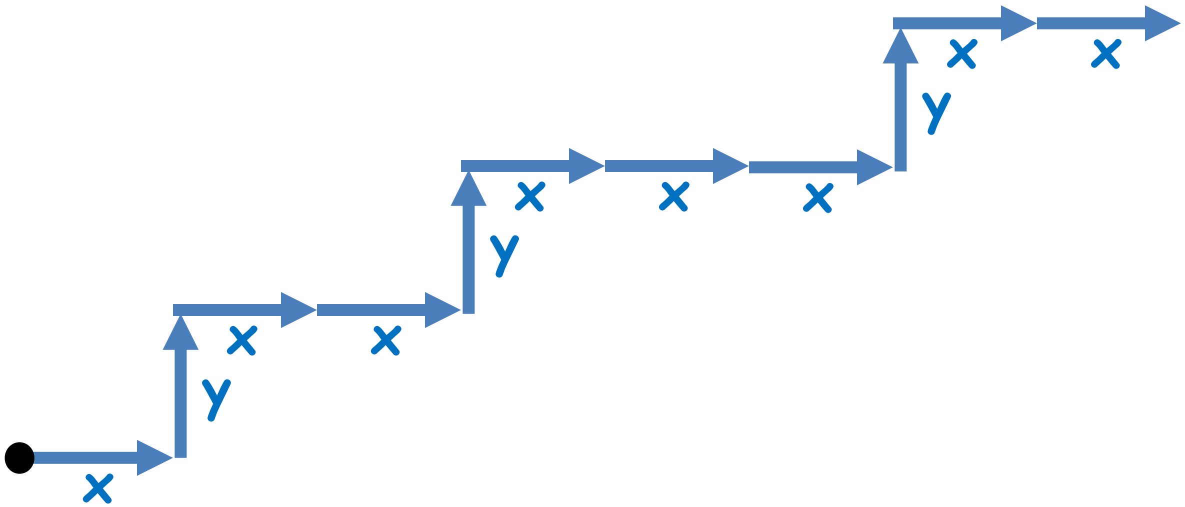


Simultaneous Core Partitions #STST+#TSTS

Funcs on Hypercube $\{0,1\}^n$ #001-#011-#100+#110

Signature Method Chen (1958) Lyons (1998)

Discrete stream from Σ^n -> a path in $\mathbb{R}^{|\Sigma|}$

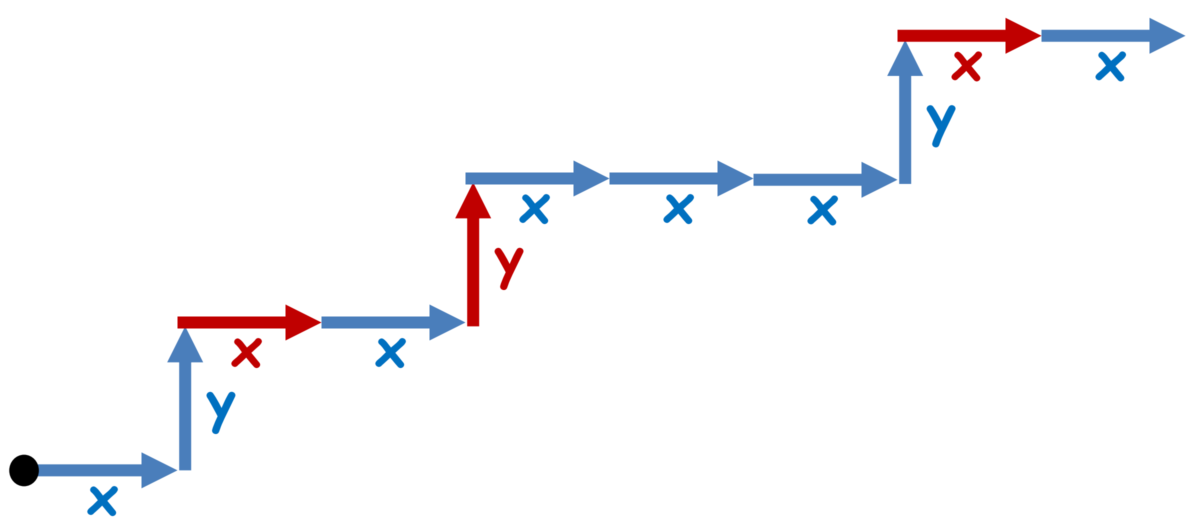


Signature Method Chen (1958) Lyons (1998)

Discrete stream from Σ^n -> a path in $\mathbb{R}^{|\Sigma|}$

Counting subwords from Σ^k -> k-fold integrals, eg:

$$\#xyx \;\rightarrow\; \int\int\int_{r<s<t} dx_r\, dy_s\, dx_t$$



Signature Method Chen (1958) Lyons (1998)

Discrete stream from Σ^n -> a path in $\mathbb{R}^{|\Sigma|}$

Counting subwords from Σ^k -> k-fold integrals, eg:

$$\#xyx \;\to\; \int\int\int_{r<s<t} dx_r\, dy_s\, dx_t$$

Machine learning application: Use the level-k signature as characteristic features of paths.

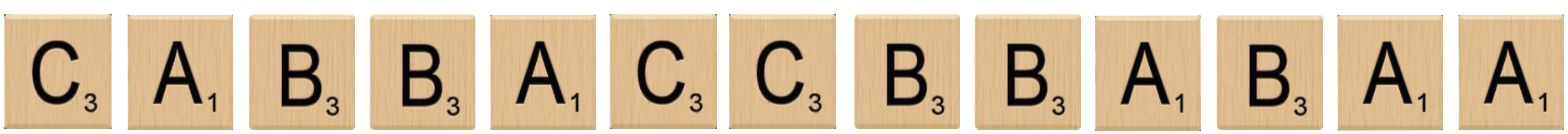
Levin Lyons Ni (2013) Chevyrev Kormilitzin (2016)

Random Word Models

Length n, finite $\Sigma=\{a,b,c,\ldots\}$

1. $W(n, (p_a,p_b,p_c,\ldots))$ iid

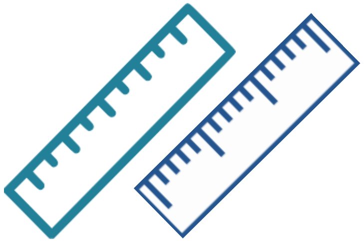
2. $W'(n_a,n_b,n_c,\ldots)$ any order



Results E Lakrec Tessler 2020

One-sample model $W(n,(p_a,p_b,p_c,\ldots))$

I. Grading: $W_k = W_{k0} \oplus \cdots \oplus W_{kk}$

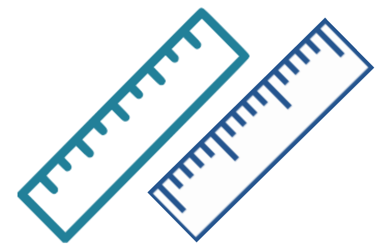


$\#f/n^k$ from W_{kr} has order $n^{-r/2}$

Results E Lakrec Tessler 2020

One-sample model $W(n,(p_a,p_b,p_c,\ldots))$

I. Grading: $W_k = W_{k0} \oplus \cdots \oplus W_{kk}$



$\#f/n^k$ from W_{kr} has order $n^{-r/2}$

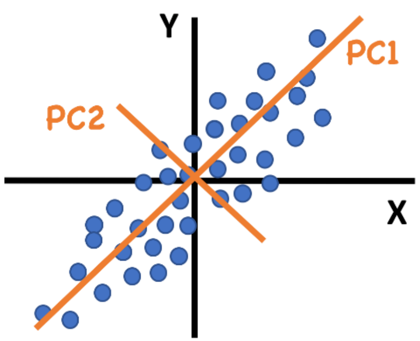
II.

$$W_{kr} = W_{kr0} \oplus W_{kr1} \oplus \ldots \oplus W_{kr(k-r)}$$

$$W_{krm} = \text{Ш}_1^{k-r-m} \ker \left(\partial_1 \Big|_{W_{(r+m)r}} \right)$$

$$\dim W_{krm} = \binom{r+m-1}{m} (|\Sigma| - 1)^r$$

$$\lambda_{krm} = \frac{(k!)^2}{(k+m)!(k-r-m)!}$$



The Words Algebra

Shuffle / Insertion operator

$$Ш_a \, bbc = abbc + babc + bbac + bbca$$

e.g. [Dieker Saliola 2018]

The Words Algebra

Shuffle / Insertion operator

$$Ш_a \, bbc = abbc + babc + bbac + bbca$$

Deletion operator

$$\partial_a \, (abac - baab) = bac + abc - 2\, bab$$

e.g. [Dieker Saliola 2018]

The Words Algebra

Shuffle / Insertion operator

$$\text{Ш}_a\, bbc = abbc + babc + bbac + bbca$$

Deletion operator

$$\partial_a\,(abac - baab) = bac + abc - 2\,bab$$

Replacement operator

$$\Theta_{ab}\; sababa = sbbaba + sabbba + sababb$$

e.g. [Dieker Saliola 2018]

The Words Algebra

Random to random card shuffling [DS 2018]

$$R = ш_a \partial_a + ш_b \partial_b + ш_c \partial_c + ш_d \partial_d + \dots$$

$$R\, bbc = 5\, bbc + 3\, bcb + 1\, cbb$$

The Words Algebra

Random to random card shuffling [DS 2018]

$$R = \shuffle_a \partial_a + \shuffle_b \partial_b + \shuffle_c \partial_c + \shuffle_d \partial_d + \ldots$$

$$R\ bbc = 5\ bbc + 3\ bcb + 1\ cbb$$

Merging operator [ELT 2020]

$$M_a = \mathrm{Id} + \shuffle_a \partial_a + \frac{1}{2!^2} \shuffle_a{}^2 \partial_a{}^2 + \frac{1}{3!^2} \shuffle_a{}^3 \partial_a{}^3 + \ldots$$

$$M_b\ bbc = 6\ bbc + 3\ bcb + 1\ cbb$$

Primary Decomposition

For $\Sigma=\{a,b,c,\ldots\}$ consider $1,2 \in \mathbb{R}\Sigma$

$$1 = a + b + c + \ldots$$

2 such that $\langle 1,2\rangle_p = 0$ $\quad p=(p_a,p_b,\ldots)$

Primary Decomposition

For $\Sigma=\{a,b,c,\ldots\}$ consider $1,2 \in \mathbb{R}\Sigma$

$$1 = a + b + c + \ldots$$

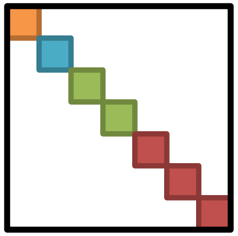
$$2 \text{ such that } \langle 1,2\rangle_p = 0 \qquad p=(p_a,p_b,\ldots)$$

The covariance of Σ^k is composed of blocks,

$$M_1 : V_{kr} \to V_{kr}$$

$$V_{kr} = \text{span}\{e : \#1(e)=k-r,\ \#2(e)=r\}$$

Full Decomposition



Over {1,2}, the merging operator

$$M_1 : V_{kr} \to V_{kr}$$

$$M_1 = \mathrm{Id} + Ш_1 \partial_1 + \frac{1}{2!^2} Ш_1{}^2 \partial_1{}^2 + \frac{1}{3!^2} Ш_1{}^3 \partial_1{}^3 + \ldots$$

admits the eigendecomposition

$$V_{krm} = (\ker \partial_1{}^{k-r-m+1}) \cap (\ker \partial_1{}^{k-r-m})^{\perp}$$

$$\mu_{krm} = \binom{2k-r}{k+m} \qquad \dim V_{krm} = \binom{m+r-1}{m}$$

Structure

E Lakrec Tessler 2020

One-sample model $W(n,(p_a,p_b,p_c,\dots))$

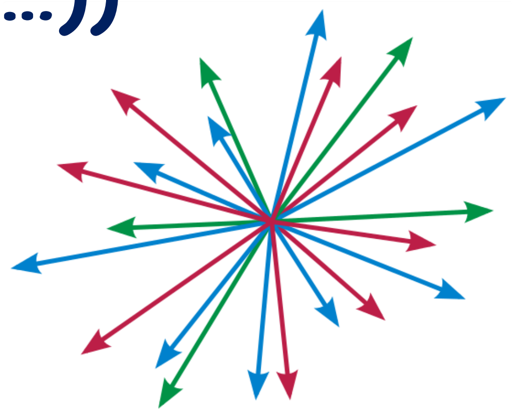
$$W = \cup_{k=1}^{\infty} W_k$$

$$W_1 \hookrightarrow W_2 \hookrightarrow W_3 \dots$$

#-invariant

Structure E Lakrec Tessler 2020

One-sample model $W(n,(p_a,p_b,p_c,\ldots))$

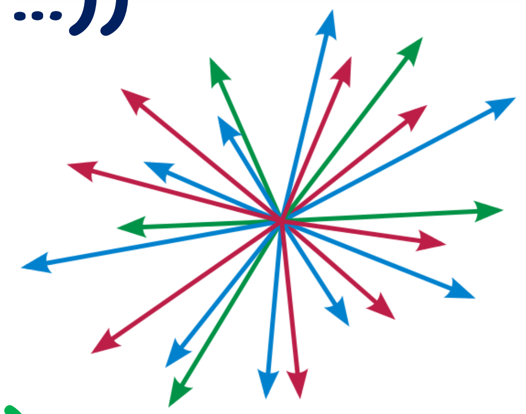


$$W = \bigcup_{k=1}^{\infty} W_k$$

$$\cong \bigoplus_{r,m} W_{krm} \qquad k \geq r+m$$

Structure E Lakrec Tessler 2020

One-sample model $W(n,(p_a,p_b,p_c,\ldots))$

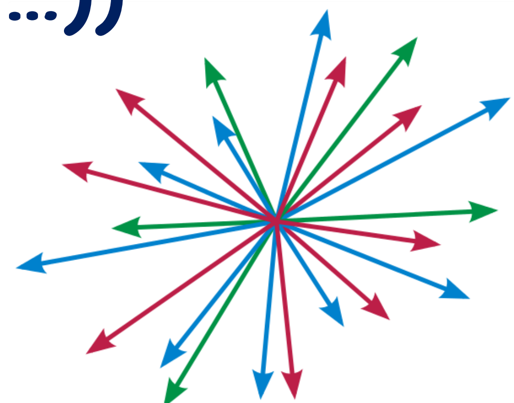


$$W = \bigcup_{k=1}^{\infty} W_k$$

$$\cong \bigoplus_{r,m}(U_{rm} \otimes \mathbb{R}^{|\Sigma|-1})$$

Structure E Lakrec Tessler 2020

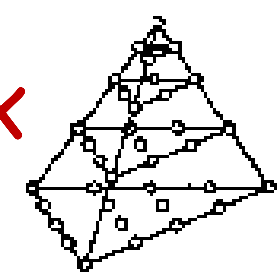
One-sample model $W(n,(p_a,p_b,p_c,\ldots))$



$$W = \bigcup_{k=1}^{\infty} W_k$$

$$\cong \bigoplus_{r,m}\left(U_{rm} \otimes \mathbb{R}^{|\Sigma|-1}\right)$$

spaces of r-variate orthogonal polynomials of degree m wrt the discrete simplex

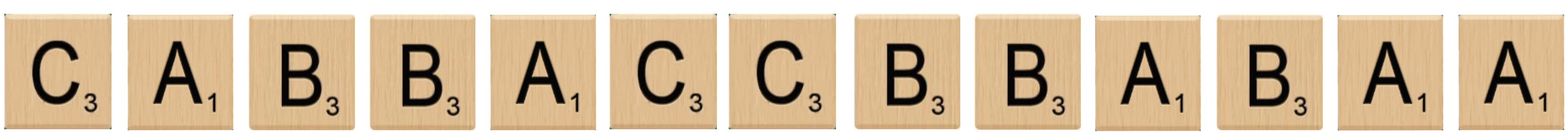


Random Word Models

Length n, finite $\Sigma=\{a,b,c,\dots\}$

1. $W(n,(p_a,p_b,p_c,\dots))$ iid

2. $W'(n_a,n_b,n_c,\dots)$ any order



More Subword Spaces

Partition **k** into **κ = (k_a, k_b, …)**

W_κ = sums of words with **(#a,#b,…) = κ**

$W_{(2,2)}$ = $\mathbb{R}\{SSTT, STST, STTS, TSST, TSTS, TTSS\}$

$W_{(1,1,1)}$ = $\mathbb{R}\{ABC, ACB, BAC, BCA, CAB, CBA\}$

More Subword Spaces

Partition $\mathbf{k}$ into $\boldsymbol{\kappa} = (k_a, k_b, \ldots)$

W_κ = sums of words with $(\#a, \#b, \ldots) = \kappa$

$W_{(2,2)} = \mathbb{R}\{SSTT, STST, STTS, TSST, TSTS, TTSS\}$

$W_{(1,1,1)} = \mathbb{R}\{ABC, ACB, BAC, BCA, CAB, CBA\}$

$$W_\kappa \hookrightarrow W_k \hookrightarrow W_{(k,\ldots,k)}$$

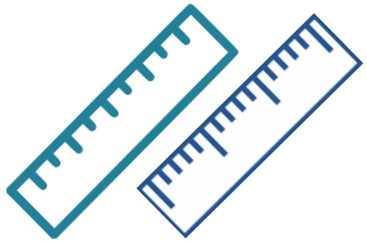
#-invariant

Results   E Lakrec Tessler 2020

Multi-sample model $W'(n_a, n_b, n_c, \ldots)$

III. Grading: $W_\kappa = W_{\kappa 0} \oplus \cdots \oplus W_{\kappa(|\kappa| - \max \kappa)}$

$\#f/n^{|\kappa|}$ from $W_{\kappa r}$ has order $n^{-r/2}$

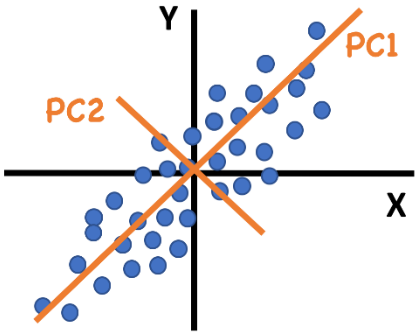


Assumption: $\forall x$ $n_x/n \to p_x > 0$

Results E Lakrec Tessler 2020

Two-sample model W'(n_a,n_b)

IV. Diagonalization of covariance



$$W_{\kappa r} = \bigoplus_{i=0}^{k-2r} \bigoplus_{j=0}^{r-1} W_{\kappa r i j} \qquad \kappa = (k_\mathsf{a}, k_\mathsf{b}), \;\; r \in \{1, \ldots, k_\mathsf{b}\}$$

$$\dim W_{\kappa r i j} = \frac{(k_\mathsf{a} + k_\mathsf{b} - 2r - i + j + 1)\,(k_\mathsf{a} + k_\mathsf{b} - i - j - 2)!}{(k_\mathsf{a} + k_\mathsf{b} - i - r)!\,(r - j - 1)!}$$

$$W_{\kappa r i j} = \Theta_\mathsf{ab}^{k_b - r}\,\mathcal{L}_\mathsf{b}^j\,Ш_\mathsf{a}^i\,\ker\left(\partial_\mathsf{a}\Big|_{W_{(k_\mathsf{a}+k_\mathsf{b}-r-i,r-j),r-j}}\right) \qquad \mathcal{L}_\mathsf{b}\Big|_{W_{(a,b)}} = Ш_\mathsf{b} - \frac{\Theta_\mathsf{ab} Ш_\mathsf{a}}{a-b+1}$$

$$\lambda_{\kappa r i j} = \frac{(k_\mathsf{a}!)^2 (k_\mathsf{b}!)^2 (k_\mathsf{a} + k_\mathsf{b} - 2r)!(k_\mathsf{a} + k_\mathsf{b} - 2r + 1)!}{(k_\mathsf{a} - r)!(k_\mathsf{b} - r)!(i)!(2k_\mathsf{a} + 2k_\mathsf{b} - r - i - j)!(k_\mathsf{a} + k_\mathsf{b} - 2r + 1 + j)!}$$

Decompositions

$S_k \circlearrowright$ k-letter words

Decompositions

$S_k \circlearrowright$ k-letter words

(aabc + 8 cbaa) (2341) =

Decompositions

$S_k \circlearrowright$ k-letter words

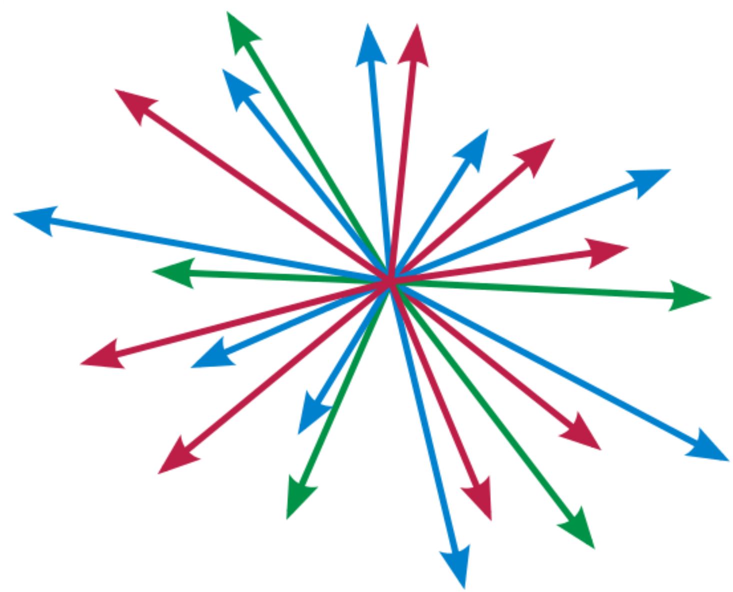
$(aabc + 8\ cbaa)\ (2341) =$

$(abca + 8\ baac)$

Decompositions

S_k ↻ k-letter words

$W_k \cong \bigoplus$ simple S_k-representations



Decompositions

$S_k \circlearrowright$ k-letter words

$W_\kappa \cong \bigoplus$ simple S_k-representations

$\cong W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2} \oplus \ldots$

$W_{\kappa r} = \bigoplus$representations of width $k-r$

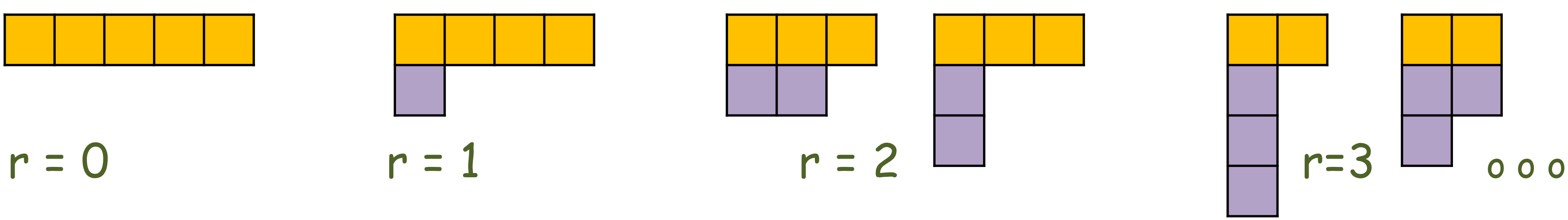
Decompositions

$S_k \circlearrowright$ k-letter words

$W_\kappa \cong \bigoplus$ simple S_k-representations

$\cong W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2} \oplus \ldots$

$W_{\kappa r} = \bigoplus$ representations of width $k-r$



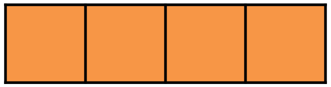
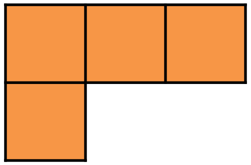
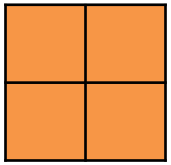
Example: $\kappa = (2,2)$

$W_\kappa = \mathbb{R}\{AABB, ABAB, ABBA, BAAB, BABA, BBAA\}$

Example: $\kappa = (2,2)$

$$W_\kappa = \mathbb{R}\{AABB, ABAB, ABBA, BAAB, BABA, BBAA\}$$

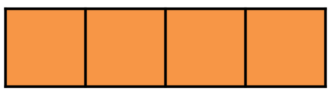
$$= W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2}$$

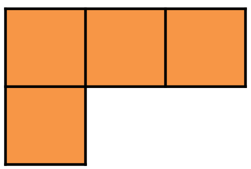




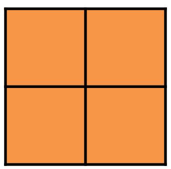
Example: $\kappa = (2,2)$

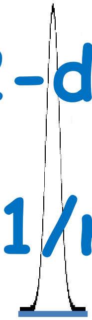
$$W_\kappa = \mathbb{R}\{AABB, ABAB, ABBA, BAAB, BABA, BBAA\}$$

$$= W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2}$$







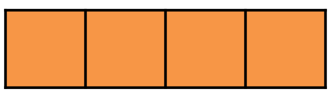


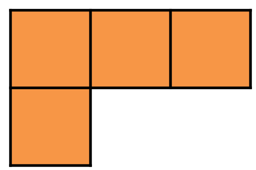
	$W_{\kappa 0}$	$W_{\kappa 1}$	$W_{\kappa 2}$
dim:	1-dim	3-dim	2-dim
scale:	1	$1/\sqrt{n}$	$1/n$
dist:	const	normal	Σ normal^2

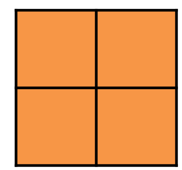
Example: $\kappa = (2,2)$

$W_\kappa = \mathbb{R}\{AABB, ABAB, ABBA, BAAB, BABA, BBAA\}$

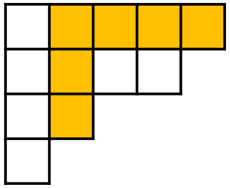
$= W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2}$







s,t core



	$W_{\kappa 0}$	$W_{\kappa 1}$	$W_{\kappa 2}$
dim:	1-dim	3-dim	2-dim
scale:	1	$1/\sqrt{n}$	$1/n$
dist:	const	normal	Σ normal^2

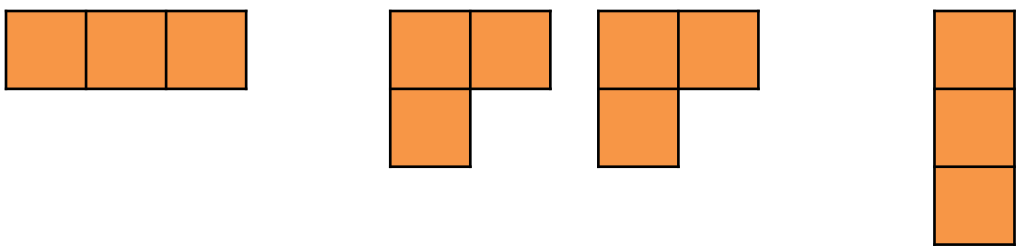
Example: $\kappa = (1,1,1)$

$W_\kappa = \mathbb{R}\{ABC, ACB, BAC, BCA, CAB, CBA\}$

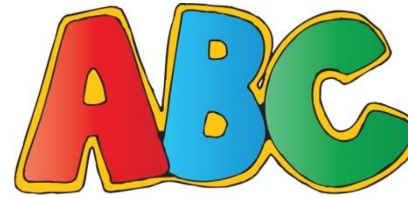
Example: $\kappa = (1,1,1)$

$$W_\kappa = \mathbb{R}\{ABC, ACB, BAC, BCA, CAB, CBA\}$$

$$= W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2}$$

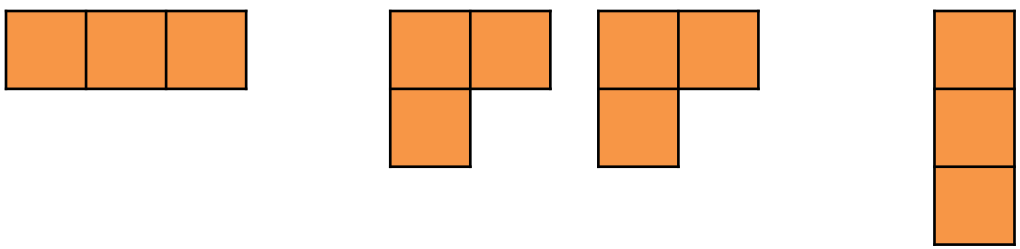


Example: $\kappa = (1,1,1)$



$W_\kappa = \mathbb{R}\{ABC, ACB, BAC, BCA, CAB, CBA\}$

$= W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2}$

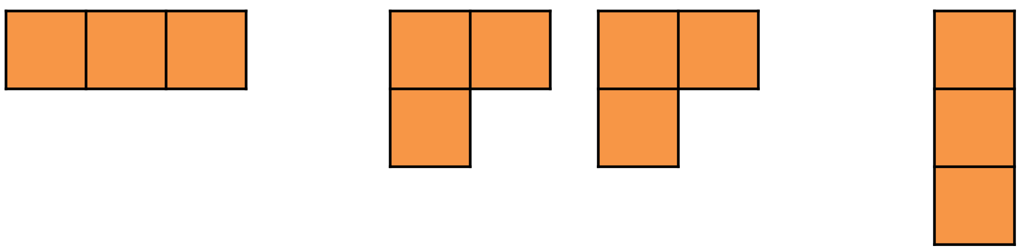


	$W_{\kappa 0}$	$W_{\kappa 1}$	$W_{\kappa 2}$
dim:	1-dim	4-dim	1-dim
scale:	1	$1/\sqrt{n}$	$1/n$
dist:	const	normal	Σ normal^2

Example: $\kappa = (1,1,1)$

$$W_\kappa = \mathbb{R}\{ABC, ACB, BAC, BCA, CAB, CBA\}$$

$$= W_{\kappa 0} \oplus W_{\kappa 1} \oplus W_{\kappa 2}$$



Gepner

dim:	1-dim	4-dim	1-dim
scale:	1	$1/\sqrt{n}$	$1/n$
dist:	const	normal	Σ normal^2

THANK YOU!