EXPECTED SIGNATURE STUFF

CIRM "Pathwise Stochastic Analysis & Applications"

March 2021

joint works with

HARALD OBERHAUSER
PATRIC BONNIER
CHONG LIU
ALEXANDER SCHELL

Datasig, ATI, OMI

For a $\mathbb{R}^d$-valued random variable X the sequence of moments

$$\left(\mathbb{E}[X^{\otimes m}]\right)_{m\geq 0} \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

is useful.

For a stochastic process $X=(X_t)_{t\geq 0}$ the sequence of signature moments

$$\left(\mathbb{E}\left[\int dX^{\otimes m}\right]\right)_{m\geq 0} \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

is useful.

I. INDEPENDENT COMPONENT ANALYSIS

II. ADAPTED TOPOLOGIES

ICA

A an invertible $(d \times d)$-matrix

X, Y $\mathbb{R}^d$-valued random variables

Mixing matrix

$$Y = A \cdot X$$

observations

sources

GOAL X has independent coordinates

Given (empirical) law of Y, recover X (resp. A)

→ 80's Hérault & Jutten, Comon, Cardoso, ...

POSSIBLE ?

For D diagonal matrix
P permutation matrix

$$Y = A \cdot X = \underbrace{A P^{-1} D^{-1}}_{\tilde{A}} \cdot \underbrace{D P X}_{\tilde{X}}$$

NEW GOAL

$\mathrm{Mon}(\mathbb{R}^d) := \{ M : (d \times d)\text{-matrix},\ M = DP$ for D diagonal, P Permutation $\}$

Given the (empirical) distribution of Y find a $\tilde{X}$ such that $\tilde{X} \equiv_{\mathrm{Mon}} X$

Theorem (Comon '94)

$X = (X^1, \dots, X^d)^T$ with $X^i \perp\!\!\!\perp X^j$, at most one X^i Gaussian

A an orthogonal (wlog) $d \times d$ matrix

$$Y = A \cdot X$$

Then

$BY \equiv_{Mon} X$ IFF BY has independent coordinates

Corollary Given

$$\Phi : \{\text{probability measures on } \mathbb{R}^d\} \longrightarrow [0, \infty)$$

such that

$$\Phi(\mu) = 0 \quad \text{IFF} \quad \mu = \mu^1 \otimes \dots \otimes \mu^d$$

Then $\hat{B} := \arg\min_B \Phi(\mathrm{Law}(BY))$ fulfills $\hat{B} Y \equiv_{Mon} X$

Theorem Let Y be a $\mathbb{R}^d$-valued random variable such that the sequence of moments

$$(\mu_Y^m)_{m\geq 0} := (\mathbb{E}[Y^{\otimes m}])_{m\geq 0} \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

characterizes the law of Y. The sequence of cumulants

$$(\kappa_Y^m)_{m\geq 0} := \pi_{sym}(\log \mathbb{E}[\exp(Y)]) \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

fulfills

i) $\langle \kappa_Y, e_{i_1} \otimes \cdots \otimes e_{i_m} \rangle = \sum_a (-1)^{|a|-1} (|a|-1)! \prod_i \langle \mu_Y, e_{a_i} \rangle$

with $\sum_a$ over all partitions $a = (a_1, \dots, a_k)$ of $(i_1, \dots, i_m)$

ii) $\kappa_Y \mapsto \mu_Y$ bijective $\langle \mu_Y, e_{i_1} \otimes \cdots e_{i_m} \rangle = \sum_a \prod_i \langle \kappa_Y, e_{a_i} \rangle$

iii) $I, J \subset \{1, \dots, d\}$. Then $(Y^i)_{i\in I} \perp\!\!\!\perp (Y^j)_{j\in J}$

IFF $\langle \kappa_Y, e_{\tau_1} \otimes e_{\tau_2} \rangle = 0 \quad \forall \tau_1 \in I^*, \tau_2 \in J^*$

$\Rightarrow$ T. SPEED publicized the combinatorial point of view in statistics

Take for $\Phi(\mathrm{Law}(Y)) := \sum_m \sum_{i_1,\dots,i_m} \langle K_Y, e_{i_1} \otimes \cdots e_{i_m} \rangle^2$

In practice : $\frac{1}{N} \sum_{j=1}^{N} \delta_{Y_j} \approx \mathrm{Law}(Y)$

POLYKAYS

$\rightarrow$ SOLVE ICA by optimization

$\rightarrow$ Many extensions, for example

$$Y_t = A \cdot X_t \qquad t = 0, 1, 2, \dots$$

(Blind Source Separation, cocktail party problem, ...)

NONLINEAR ICA IN CONTINUOUS TIME

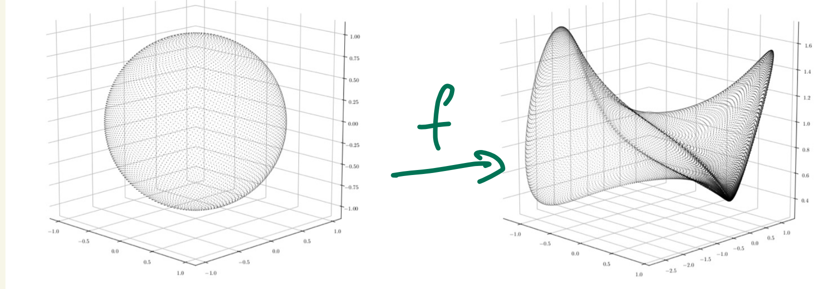
$$Y_t = f(X_t)$$

with $f : \mathbb{R}^d \to \mathbb{R}^d$ C^3 - diffeomorphism

$(X_t)_{t \geq 0}$ d-dimensional stochastic process with independent coordinates

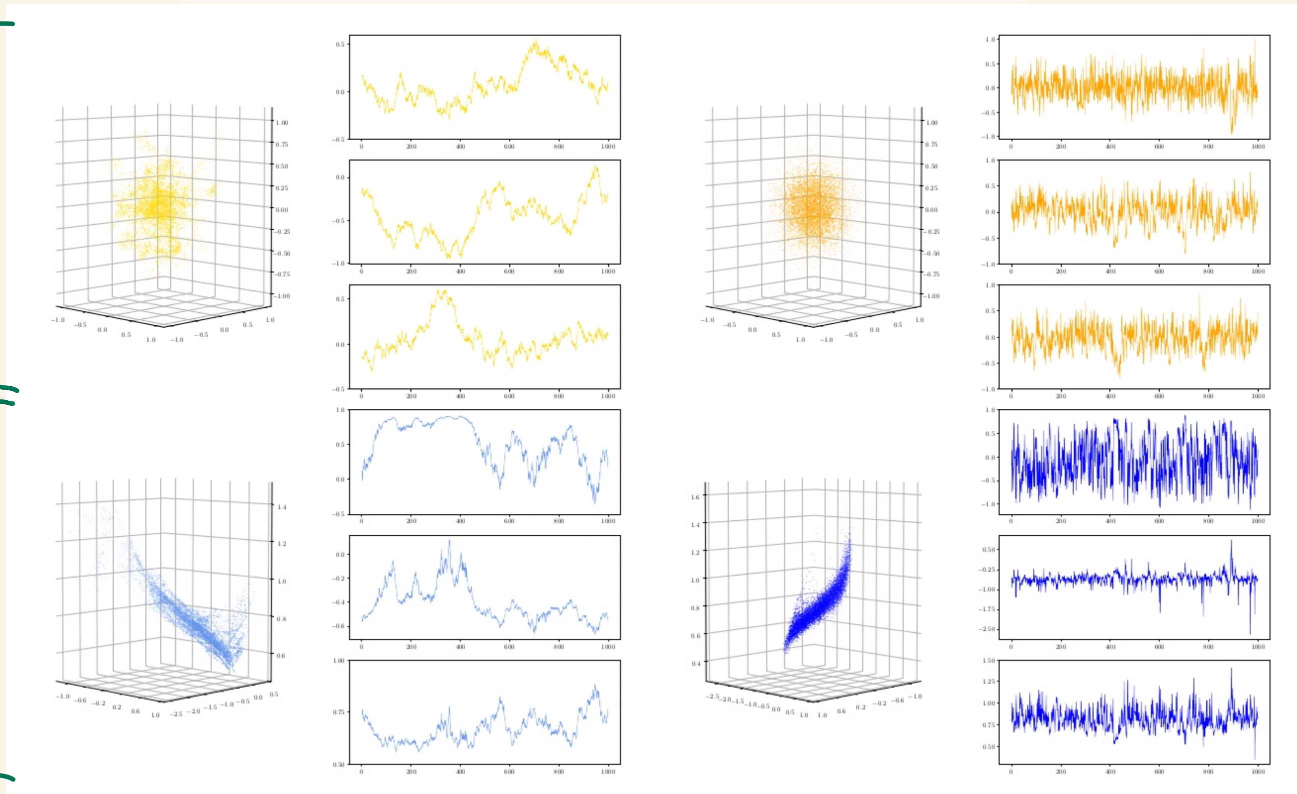
<u>Philosophy</u>: Exploit temporal structure to solve harder problem

$\longrightarrow$ Contrastive learning: Hyvärinen & Morioka 2016



$(X_t)_{t \geq 0}$

$(Y_t)_{t \geq 0}$



Theorem (Schell & O)

Let $(X_t)_{t \geq 0}$ be a d-dimensional α- or β- or γ- contrastive stochastic process with independent coordinates.

Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a C^3-diffeomorphism.

Then

$h \circ Y \equiv_{Mon} X$ IFF $h \circ Y$ has ind. coord.

__Corollary__ Let

$$\Phi : \{\text{Laws of stochastic processes}\} \to [0, \infty)$$

such that

$$\Phi(\mu) = 0 \quad \text{IFF} \quad \mu = \mu^1 \otimes \cdots \otimes \mu^d$$

Then

$$\hat{h} := \underset{h}{\operatorname{argmin}} \; \Phi(\operatorname{Law}(h \circ Y))$$

solves $\hat{h} \circ Y =_{\text{Mon}} X$

Theorem Let Y be a $\mathbb{R}^d$-valued random variable such that the sequence of moments

$$(\mu_Y^m)_{m\geq 0} := (\mathbb{E}[Y^{\otimes m}])_{m\geq 0} \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

characterizes the law of Y. The sequence of cumulants

$$(\kappa_Y^m)_{m\geq 0} := \pi_{sym}(\log \mathbb{E}[\exp(Y)]) \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

fulfills

i) $\langle \kappa_Y, e_{i_1} \otimes \cdots \otimes e_{i_m} \rangle = \sum_a (-1)^{|a|-1} (|a|-1)! \prod_i \langle \mu_Y, e_{a_i} \rangle$

with $\sum_a$ over all partitions $a = (a_1, \dots, a_k)$ of $(i_1, \dots, i_m)$

ii) $\kappa_Y \mapsto \mu_Y$ bijective $\langle \mu_Y, e_{i_1} \otimes \cdots e_{i_m} \rangle = \sum_a \prod_i \langle \kappa_Y, e_{a_i} \rangle$

iii) $I, J \subset \{1, \dots, d\}$. Then $(Y^i)_{i\in I} \perp\!\!\!\perp (Y^j)_{j\in J}$

IFF $\langle \kappa_Y, e_{\tau_1} \otimes e_{\tau_2} \rangle = 0 \quad \forall \tau_1 \in I^*, \tau_2 \in J^*$

Theorem (Bonnier & O)

Let Y be a bounded variation process such that the sequence of signature moments

$$(\mu_Y^m)_{m\geq 0} := \left(\mathbb{E}\left[\int dY^{\otimes m}\right]\right)_{m\geq 0} \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

characterizes the law of Y. The sequence of signature cumulants

$$(\kappa_Y^m)_{m\geq 0} := \log\left(\mathbb{E}\left[\int dY^{\otimes m}\right]\right)_{m\geq 0}] \in \prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$$

fulfills

i) $\langle \kappa_Y, e_{i_1} ⧢ \cdots ⧢ e_{i_m} \rangle = \sum_a (-1)^{|a|-1} \frac{a!}{|a|} \mu_X(a)$

with $\sum_a$ over all ordered partitions of $(i_1, \dots, i_m)$

ii) $\kappa_Y \mapsto \mu_Y$ bijective $\langle \mu_Y, e_{i_1} \otimes \cdots e_{i_m} \rangle = \sum_a \frac{1}{|a|!} \kappa_X(a)$

iii) $I, J \subset \{1, \dots, d\}$. Then $(Y^i)_{i\in I} \perp\!\!\!\perp (Y^j)_{j\in J}$

IFF $\langle \kappa_Y, e_{\tau_1} ⧢ e_{\tau_2} \rangle = 0 \quad \forall \tau_1 \in I^*, \tau_2 \in J^*$

IN PRACTICE

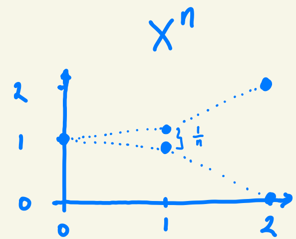
$$\underset{h}{\operatorname{argmin}} \quad \Phi(\, \mathrm{Law}(h \circ Y))$$

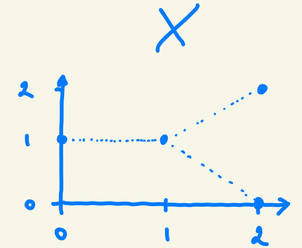
- Law ($h \circ Y$) estimated by empirical measure
  ergodicity
- Φ signature polykeys
- argmin over NN h + SGD

II] Adapted Topologies

WEAK TOPOLOGY ... IS WEAK

Example 1





$X^n \to X$ weakly as $n \to \infty$

Example 2

$(\Omega, \mathcal{G}, \mathbb{P}, X) \longmapsto \sup_{\tau} \mathbb{E}[L_\tau]$ is not continuous

→ ALDOUS, RÜSCHENDORF, VERSHIK, HOOVER & KEISLER & MANY OTHERS

→ BACKHOFF, BARTL, BEIGLBÖCK, EDER

$I = \{0, \dots, T\}$ "time"

U "state space", compact subset of linear space V

$\underset{\sim}{X} = (\Omega, \mathcal{G}, \mathbb{P}, X)$ adapted stochastic process on a filtered probability space $(\Omega, \mathcal{G}, \mathbb{P})$ with coordinate process $X = (X_t)_{t \in I}$

$\mathcal{S}(U)$ the set of adapted processes with state space U

WHAT TOPOLOGY ON $\mathcal{S}(U)$?

<u>Recall</u>: Weak convergence says

$$\underset{\sim}{X}^n \xrightarrow[n \to \infty]{} \underset{\sim}{X} \quad \text{IFF} \quad \mathbb{E}[f(X^n_{t_1}, \dots, X^n_{t_m})] \to \mathbb{E}[f(X_{t_1}, \dots, X_{t_m})]$$

$$\forall f \in C_b(U^m, \mathbb{R}) \qquad t_1, \dots, t_m \in I$$

The set of adapted functionals AF consists of maps that map an element of $\mathcal{S}(U)$ to a real-valued random variable. It is defined inductively:

i for $t_1, \dots, t_n \in I$ and $f \in C_b(U^n, \mathbb{R})$

$$\underline{X} \longmapsto f(X_{t_1}, \dots, X_{t_n}) \quad \in AF$$

ii for $f_1, \dots, f_n \in AF$ and $f \in C_b(\mathbb{R}^n, \mathbb{R})$

$$\underline{X} \longmapsto f(f_1(\underline{X}), \dots, f_n(\underline{X})) \quad \in AF$$

iii for $t \in I$ and $f \in AF$

$$\underline{X} \longmapsto \mathbb{E}[f(\underline{X}) \mid \mathcal{G}_t] \quad \in AF$$

Denote with $AF_r \subset AF$ the subset that can be built with at most r conditional expectations (Point iii).

The adapted topology of rank r on $\mathcal{S}(U)$ is defined by

$$\boxed{\lim \underline{X}_n = \underline{X} \quad \text{IFF} \quad \lim \mathbb{E}[f(\underline{X}_n)] = \mathbb{E}[f(\underline{X})] \quad \forall f \in AF_r}$$

Another natural quantity to associate with an adapted process $\underline{X}$ is the so-called prediction process $\hat{X}^r$ of rank r defined inductively as

$$\hat{X}^0_t := X_t$$

$$\hat{X}^{r+1}_t := \mathbb{P}(\hat{X}^r \in \cdot \mid \mathcal{F}_t)$$

Note: $\hat{X}^0_t$ takes values in U

$\hat{X}^1_t$ — " — $\mathcal{M}(I \to U)$

$\hat{X}^2_t$ — " — $\mathcal{M}(I \to \mathcal{M}(I \to U))$

$\vdots$ $\quad$ $\vdots$

$\hat{X}^{r+1}_t$ — " — $\mathcal{M}(I \to \mathcal{M}(\ldots \mathcal{M}(I \to U) \ldots))$

$$\boxed{\lim_n \underline{X}_n = \underline{X} \quad \text{IFF} \quad \hat{X}^{n\,r} \xrightarrow[n\to\infty]{\text{weakly}} \hat{X}^r}$$

Both, AF & Prediction Process, give a topology for $\mathcal{S}(U)$.

WHICH SHOULD WE USE?

Theorem Let $\underset{\sim}{X}, \underset{\sim}{Y} \in \mathcal{G}(U)$. TFAE for every $r \geq 0$

i) $\mathbb{E}[f(\underset{\sim}{X})] = \mathbb{E}[f(\underset{\sim}{Y})] \quad \forall f \in AF_r$

ii) $\mathrm{Law}(\hat{X}^r) = \mathrm{Law}(\hat{Y}^r)$

We call the resulting topology on $\mathcal{G}(U)$ the adapted topology of rank r

HOW DO WE DESCRIBE THE LAW OF A PROCESS (SUCH AS $\hat{X}^r$) ?

$\longrightarrow$ SIGNATURE MOMENTS

$\boxed{\text{CAVEAT}}$ STATE SPACE OF $\hat{X}^r$ IS PRETTY BIG !

LET'S START AT $r=0$ AND WORK OUR WAY UP

PATHS IN PATHS IN ...

U

$U_0 := U$

$U_1 := (I \longrightarrow U_0)$

$U_2 := (I \longrightarrow U_1) \equiv I \rightarrow (I \rightarrow U)$

$\vdots$

$U_r := (I \rightarrow U_{r-1}) \equiv \underbrace{I \rightarrow (I \rightarrow (\cdots \underbrace{(I \rightarrow \underbrace{(I \rightarrow U)}_{U_1}) \cdots)}_{U_2})}_{U_r}$

Paths of rank r

RANDOM PATHS IN RANDOM PATHS IN...

$\mathcal{M}_0 := U$

$\mathcal{M}_1 := \mathcal{M}(I \rightarrow \mathcal{M}_0) \equiv \mathcal{M}(I \rightarrow U)$

$\mathcal{M}_2 := \mathcal{M}(I \rightarrow \mathcal{M}_1) \equiv \mathcal{M}(I \rightarrow \mathcal{M}(I \rightarrow U))$

$\vdots$

$\mathcal{M}_r := \mathcal{M}(I \rightarrow \mathcal{M}_{r-1}) \equiv \ldots$

Measures of rank r

For a linear space V denote $\boxed{T(V) := \prod_{m\geq 0} V^{\otimes m}}$

$T(V)$ is a linear space, and we can repeat this construction

$$\boxed{T_0(V) := V \qquad T_{r+1}(V) := T(T_r(V))}$$

(I'm cheating here)
→ Kurusch & Patras

EXAMPLE

$$x \in (I \longrightarrow (I \longrightarrow \underbrace{(I \longrightarrow U)}))) \cong U_3$$

$$\downarrow s$$

$$(I \longrightarrow \underbrace{(I \longrightarrow T(V))})$$

$$\downarrow s$$

$$\underbrace{(I \longrightarrow T(T(V)))}$$

$$\downarrow s$$

$$T(T(T(V)))$$

HIGHER RANK SIGNATURES

Define the family of maps $S_r : U_r \to T_r(V)$ inductively by setting

$$S_1 := S \qquad S_{r+1}(x) := S(x^* S_r)$$

where $x^* S_r$ is the pullback of S_r along x^*.

<u>PROPOSITION</u> The maps $S_r : U_r \longrightarrow T_r(V)$ are injections.

HIGHER RANK EXPECTED SIGNATURES

Define the family of maps $\bar{S}_r : \mathcal{M}_r \to T_r(V)$ inductively by setting

$$\bar{S}_1(\mu) := \int S(x)\,\mu(dx) \qquad \bar{S}_{r+1}(\mu) := \int S(x^* \bar{S}_r)\,\mu(dx)$$

<u>PROPOSITION</u> For every $r \geq 1$ the maps

$$S_r : U_r \longrightarrow T_r(V)$$

$$\bar{S}_r : \mathcal{M}_r \longrightarrow T_r(V)$$

are injective.

For $r = 1$, S_1 is the usual signature

$\bar{S}_1$ is the usual expected signature

THEOREM For every $r \geq 0$ the map

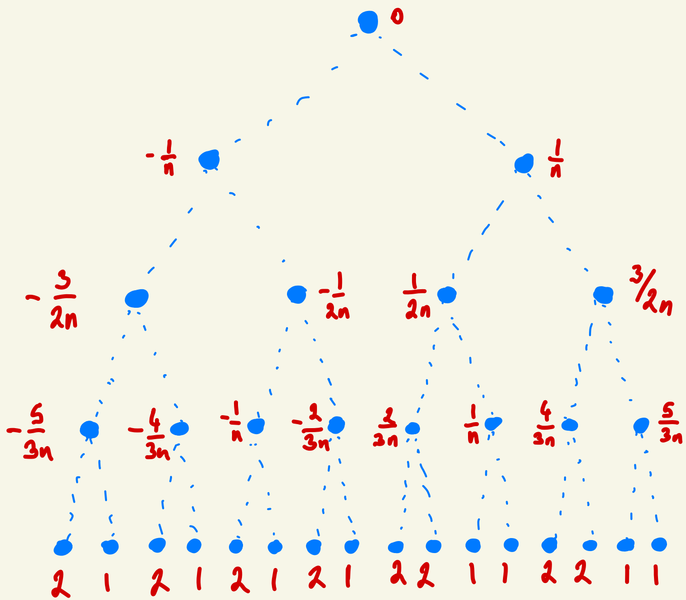
$$d_r : \mathcal{FP}(U) \times \mathcal{FP}(U) \longrightarrow [0,\infty)$$
$$(\underline{X}, \underline{Y}) \longmapsto \Big\| \underbrace{\bar{S}_{r+1}(\mathrm{Law}(\hat{\underline{X}}^r))}_{\Phi(\underline{X})} - \underbrace{\bar{S}_{r+1}(\mathrm{Law}(\hat{\underline{Y}}^r))}_{\Phi(\underline{Y})} \Big\|_{T_{r+1}(V)}$$

is a metric that metrizes the adapted topology of rank r.

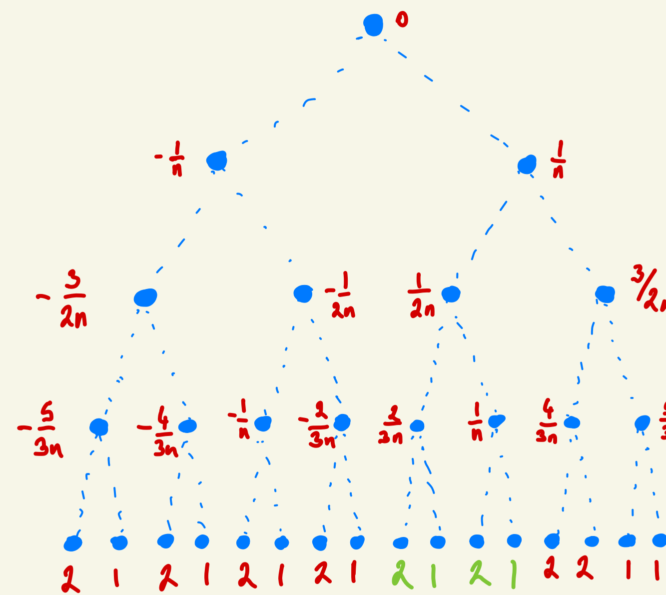
$\boxed{\text{$\Phi(\underline{X})$ yields much more than a metric. It gives us a (multi-) graded description of the law of $\underline{X}$ and how it interacts with the filtration.}}$

Example (Hoover-Keisler 84): FILTRATIONS ≃ TREES

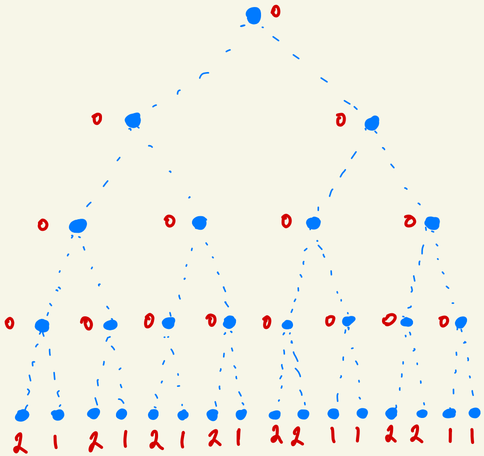
& WHY WE NEED HIGHER RANK TOPOLOGIES



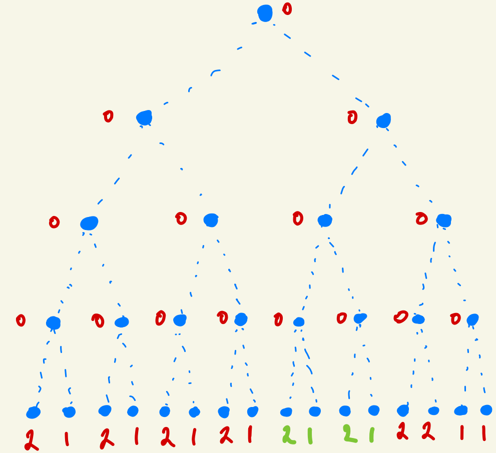
$\underline{X}^n$



$\underline{Y}^n$



$\underline{X}^n \to \underline{X}$



$\underline{Y}^n \to \underline{Y}$

$\underline{X} = \underline{Y}$ in $r = 0, 1$ adapted topology

BUT

$$\mathbb{E}[\,\mathbb{E}[X_4 \mid \mathcal{F}_3]^2 \mid \mathcal{F}_1] = \begin{cases} 5/2 \\ 9/4 \end{cases} \text{with probability } \tfrac{1}{2}$$

$$\mathbb{E}[\,\mathbb{E}[Y_4 \mid \mathcal{F}_3]^2 \mid \mathcal{F}_1] = \frac{19}{8}$$

HENCE $\underline{X}$, $\underline{Y}$ are <u>not</u> close in rank $r = 2$ topology

EXECUTIVE SUMMARY

i structured description of STOCHASTIC PROCESS AND ITS FILTRATION in terms of (multi)-graded tensors

ii natural feature map for sequential decision making
induces the adapted topology of rank r for any $r \geq 0$

iii interesting new structures in finite discrete time
discontinuity not due to roughness

THANKS FOR YOUR TIME!

→ SCHELL, O "Nonlinear ICA in continuous time"

→ BONNIER, LIU, O "Higher Rank Signatures and Adapted Topologies"