An OOM Tutorial

Herbert Jaeger

Jacobs University Bremen

Overview

1. From HMMs to OOMs
2. OOMs as sequence generators
3. Equivalence theorem
4. OOMs and HMMs
5. Interpretable OOMs
6. Basic learning algorithm
7. Efficient learning algorithms
8. The dreaded nonnegativity problem
9. From stochastic processes to OOMs
10. Historical notes
11. General OOM theory
12. Beginnings of a Hilbert space theory
13. Research topics

Some Shorthand Notation

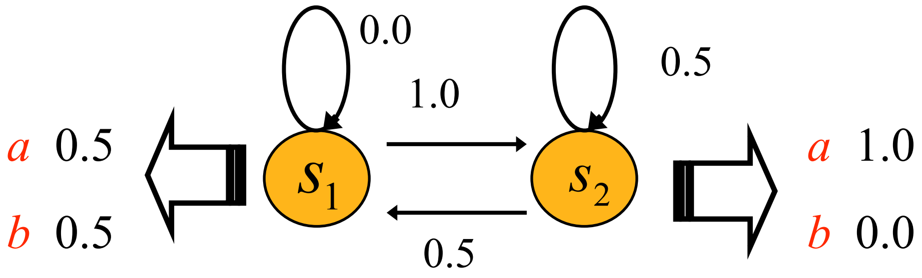
Let $(X_n)_{n=0,1,2,...}$ be a discrete-time stochastic process.

- For $P(X_0 = a_0,..., X_N = a_N)$ write $P(a_0...a_N)$ or $P(\overline{a})$.

- For $P(X_{N+1} = b_1,..., X_{N+M} = b_M \mid X_0 = a_0,..., X_N = a_N)$ write $P(b_{N+1}...b_{N+M} \mid a_0...a_N)$ or $P(\overline{b} \mid \overline{a})$.

1 From HMMs to OOMs

An HMM[1]):



$S = \{s_1, s_2\}$ hidden states

$O = \{a, b\}$ observable events

$$M = \begin{bmatrix} 0.0 & 1.0 \\ 0.5 & 0.5 \end{bmatrix} \qquad O_a = \begin{bmatrix} 0.5 & \\ & 1.0 \end{bmatrix} \qquad O_b = \begin{bmatrix} 0.5 & \\ & 0.0 \end{bmatrix}$$

$$M^T = \begin{bmatrix} 0.0 & 0.5 \\ 1.0 & 0.5 \end{bmatrix} \qquad M^T O_a = \begin{bmatrix} 0.0 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} =: T_a \qquad M^T O_b = \begin{bmatrix} 0.0 & 0.0 \\ 0.5 & 0.0 \end{bmatrix} =: T_b$$

w_0 invariant vector of $\mathbf{M}^\mathbf{T}$

$$\boxed{P(ab) = \mathbf{1} T_b T_a w_0}$$

$$\boxed{\mathrm{HMM} : (\mathbb{R}^2, \{T_a, T_b\}, w_0)}$$

1) Graphics from slide of K. Kretschmar

HMM:

- $\mathrm{HMM}: (\mathbb{R}^m, \{T_a, T_b\}, w_0)$
- $T_a + T_b = M^T$ where
  M is a Markov matrix
- W_0 is an invariant P-vector with component sum =1
- $P(ab) = \mathbf{1} T_b T_a w_0$
- non-negative entries only

τ

OMM:

- $\mathrm{OOM}: (\mathbb{R}^m, \{\boldsymbol{\tau}_a, \boldsymbol{\tau}_b\}, w_0)$
- $\boldsymbol{\tau}_a + \boldsymbol{\tau}_b = \mu,$ where
  μ has column sum = 1
- w_0 is an invariant vector with component sum =1
- $P(ab) = \mathbf{1} \boldsymbol{\tau}_b \boldsymbol{\tau}_a w_0$
- negative entries are permitted

Definition

An OOM is a structure $(\mathbb{R}^m, (\boldsymbol{\tau}_a)_{a\in\Sigma}, w_0)$, where $w_\mathbf{0} \in \mathbb{R}^m$, $\boldsymbol{\tau}_a : \mathbb{R}^m \to \mathbb{R}^m$ linear, such that

1. $\mathbf{1}\mu = \mathbf{1}\sum_{a\in\Sigma} \boldsymbol{\tau}_a = \mathbf{1}$
2. $\mathbf{1}w_0 = 1$
3. for every sequence $a_1 \ldots a_n \in O^n$: $\mathbf{1}\boldsymbol{\tau}_{a_n} \cdots \boldsymbol{\tau}_{a_1} w_0 \geq 0$

Note. A formally more general, but equivalent, definition replaces the all-ones row vector $\mathbf{1}$ by any row vector σ:

1. $\sigma \mu = \sigma$
2. $\sigma w_0 = 1$
3. for every sequence $a_1 \ldots a_n \in O^n : \boldsymbol{\sigma}\boldsymbol{\tau}_{a_n} \cdots \boldsymbol{\tau}_{a_1} w_0 \geq 0$

Theorem

An OOM $(\mathbb{R}^m, (\boldsymbol{\tau}_a)_{a \in \Sigma}, w_0)$ defines a stochastic process by putting

$$P(a_1 \cdots a_n) = \mathbf{1}\boldsymbol{\tau}_{a_n} \cdots \boldsymbol{\tau}_{a_1} w_0$$

for every sequence $a_1 ... a_n \in \Sigma^n$.

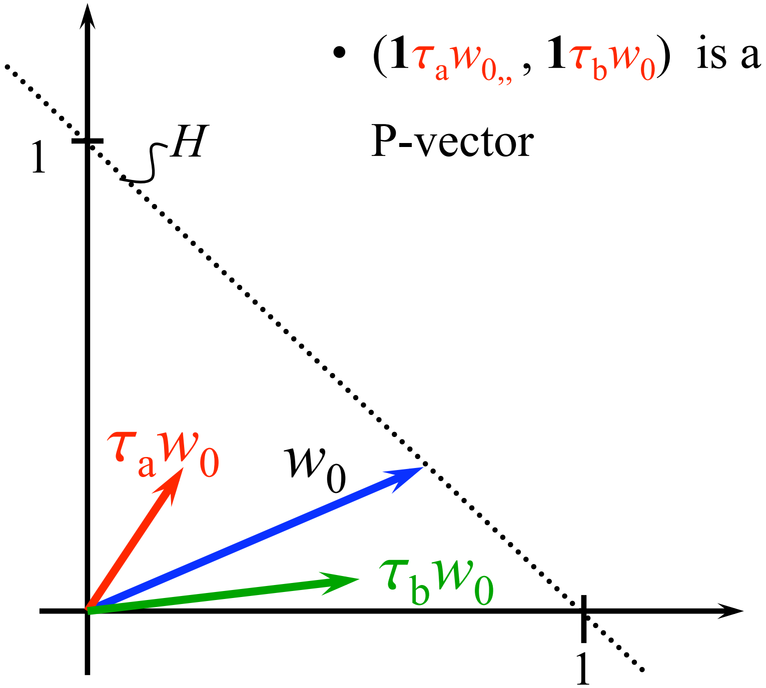
Note. The process is stationary iff $\mu w_0 = w_0$.

2 OOMs as sequence generators

$$\mathcal{A} = (\mathbb{R}^2, \{\tau_a, \tau_b\}, w_0) \qquad O = \{a, b\}$$

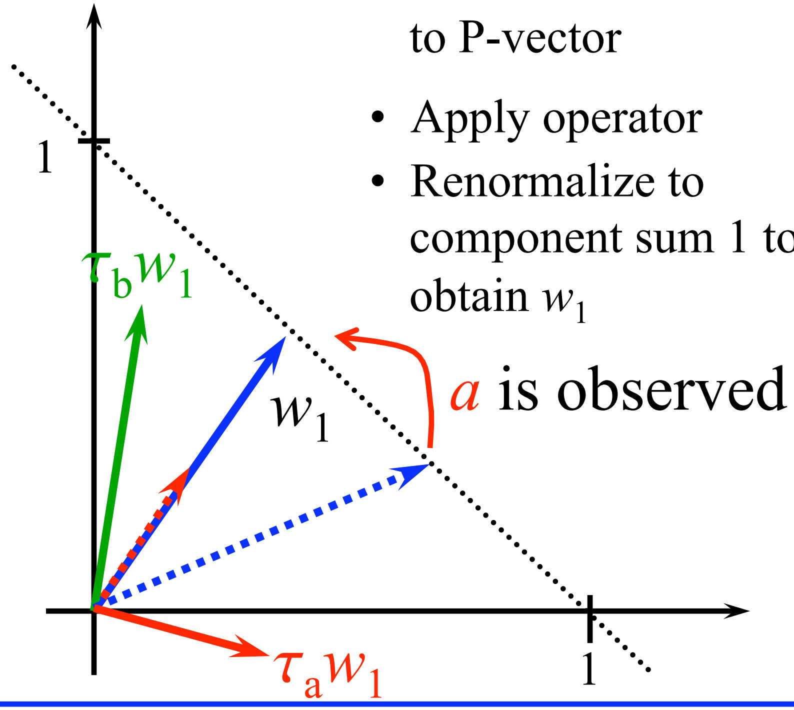
$t = 0$

- Compute $\mathbf{1}\tau_a w_0$, $\mathbf{1}\tau_b w_0$
- $(\mathbf{1}\tau_a w_{0,,}$, $\mathbf{1}\tau_b w_0)$ is a P-vector



$t = 1$

- Select a vs. b according to P-vector
- Apply operator
- Renormalize to component sum 1 to obtain w_1



JACOBS UNIVERSITY

3 Equivalence theorem

Two OOMs $A = (\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$, $B = (\mathbb{R}^m, (\tau'_a)_{a \in O}, w'_0)$, where m is minimal, generate the same process

iff

there exists a coordinate transformation ρ: $\mathbb{R}^m \rightarrow \mathbb{R}^m$, that preserves component sums of vectors, with

$$\tau'_a = \rho\, \tau_a \rho^{-1} \quad \text{for all } a \in O.$$

Corollary 1

For a given OOM $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$ there exist infinitely many different but equivalent OOMs of same dimension.

Proof: every coordinate transformation ρ: $\mathbb{R}^m \to \mathbb{R}^m$, that preserves component sums of vectors, yields a new version of $\mathcal{A}$ via $\tau'_a = \rho \, \tau_a \rho^{-1}$ for all $a \in O$.

Corollary 2

For two OOMs $A = (\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$, $B = (\mathbb{R}^{m'}, (\tau'_a)_{a \in O}, w'_0)$, it is decidable whether they are equivalent.

Proof: first transform them into minimal-dimensional versions (effective algorithm exists), then check whether $\tau'_a = \rho\, \tau_a \rho^{-1}$ for all $a \in O$, for some ρ .

JACOBS UNIVERSITY

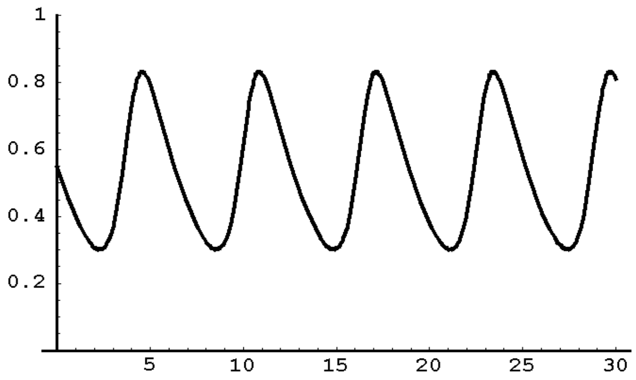
4 OOMs and HMMs

The OOM

$$\tau_a = \begin{pmatrix} 0.645 & -0.395 & 0.125 \\ 0.355 & 0.395 & -0.125 \\ 0 & 1 & 0 \end{pmatrix} \quad \tau_b = \begin{pmatrix} 0 & 0 & 0.218 \\ 0 & 0 & 0.329 \\ 0 & 0 & 0.452 \end{pmatrix}$$

generates/describes *aaaabaaaabaaabaaaabaa*...

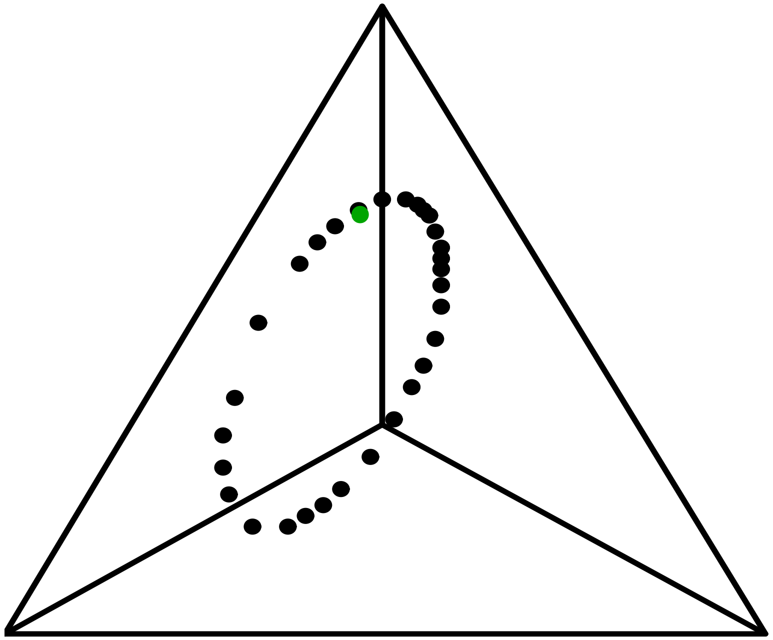
$P(a \mid ba^t)$



t

a "probability clock"

How the probability clock works



- τ_b is a projection: every state vector is mapped on ●.
- τ_a is a rotation: iterated applications yield states on a circle.
- This gives rise to oscillation of $P(a \mid ba^n)$.

Probability clocks cannot be modelled by HMMs, "because" rotation operators need negative entries.

Consequence

The processes that can be modelled by HMMs are a proper subclass of the processes that can be modelled by OOMs:

$$\text{HMM} \subsetneq \text{OOM}$$

5 Interpretable OOMs

Definition

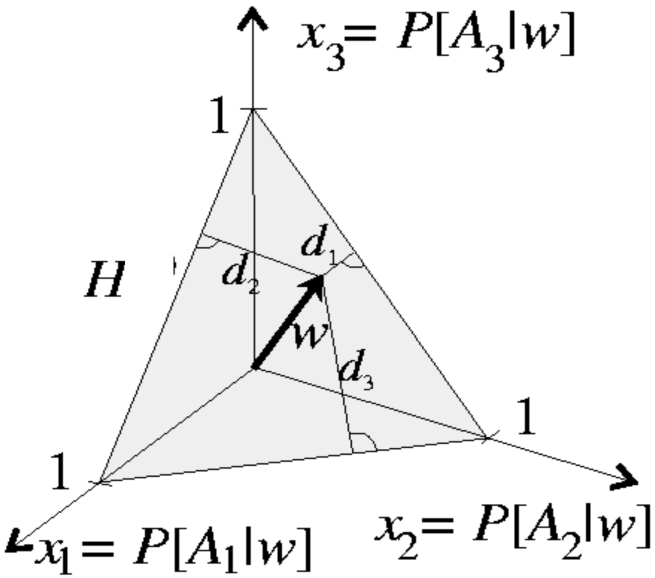
1. Let O be a finite set (alphabet) of observables, $k \geq 1$. A k-event is a nonempty subset of O^k.
2. Let furthermore $m \geq 1$. A partitioning $O^k = A_1 \cup \ldots \cup A_m$ into m disjoint nonempty k-events is a set of characteristic events (of length k and dimension m).

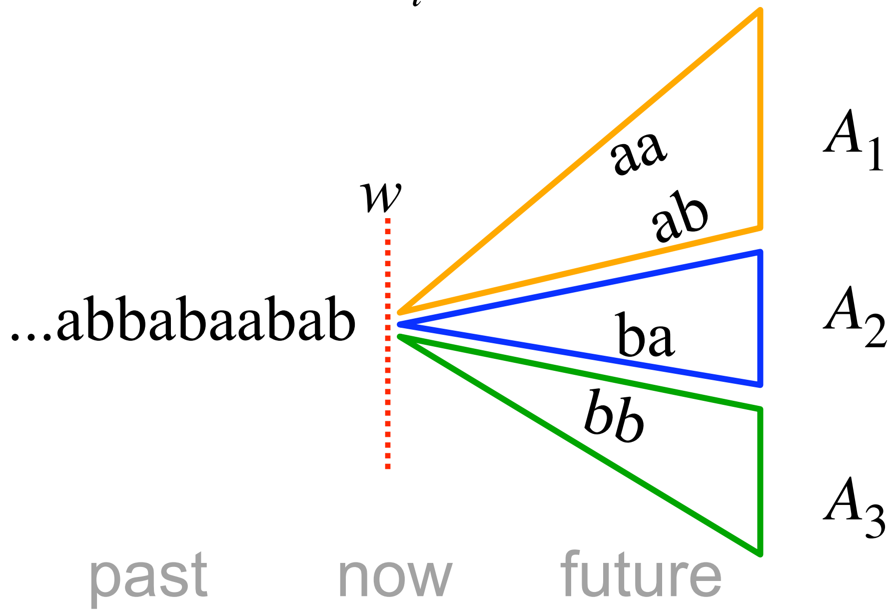
Example

$O = \{a,b\}$, $k = 2$, $m = 3$:

$A_1 = \{aa, ab\}$, $A_2 = \{ba\}$, $A_3 = \{bb\}$

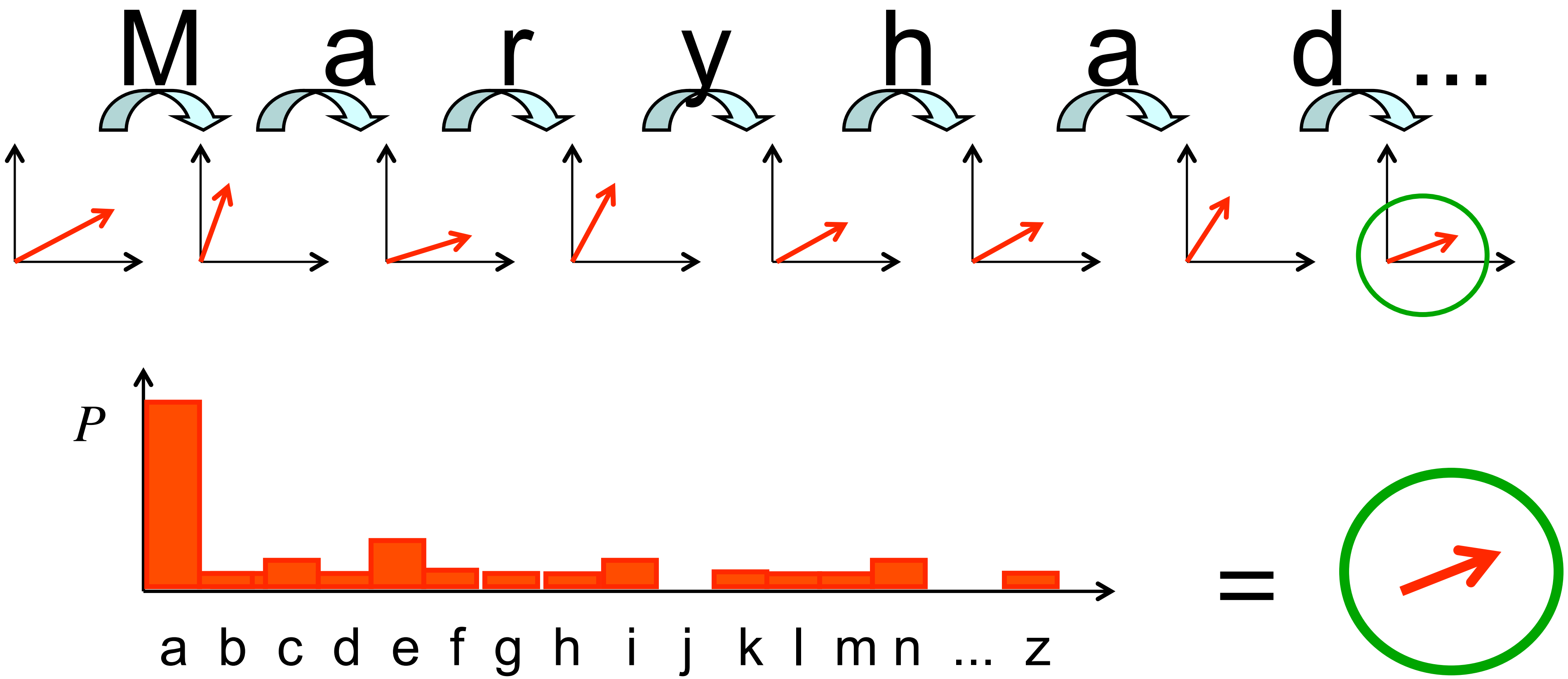
Explanation of concept. Consider a 3-dim OOM, and let A_1, A_2, A_3 be characteristic events of dim 3 and some length k (we don't care). Then this OOM is interpretable w.r.t. A_1, A_2, A_3, if the three components of state vectors = future probabilities of characteristic events A_i.





$$w = (P(A_1|\, w), P(A_2|\, w), P(A_3|\, w))$$

Example



Theorem

Let $A = (\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$ be a minimal-dimensional OOM.

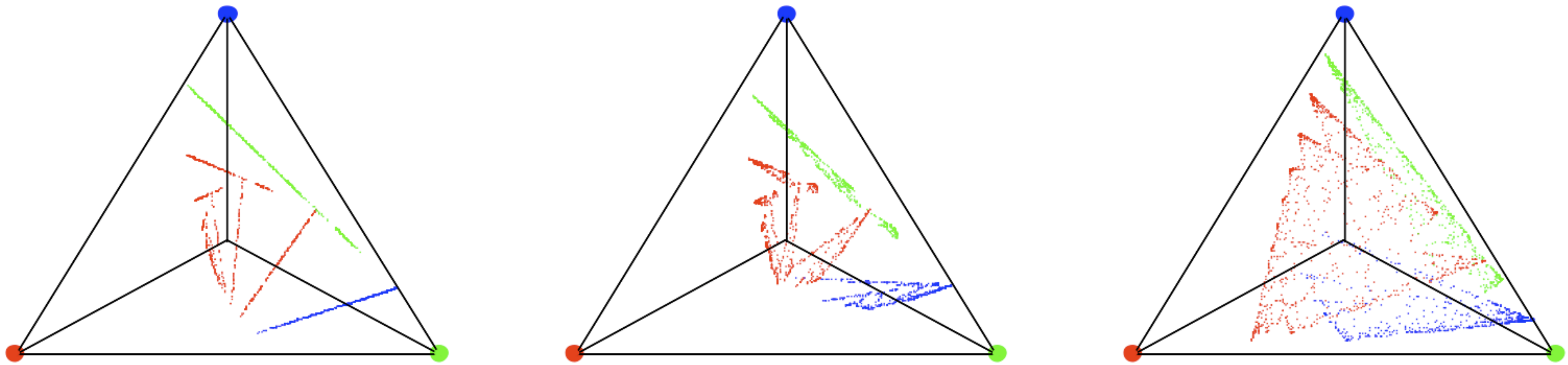
Let $O^k = \mathrm{A}_1 \cup ... \cup \mathrm{A}_m$ be characteristic events of dim m and some length k.

Then, generically, A can be effectively transformed into an equivalent OOM that is interpretable w.r.t. $\mathrm{A}_1, ..., \mathrm{A}_m$.

Proof: define a transformation ρ by $\rho(x) = (\mathbf{1}\boldsymbol{\tau}_{A_1} x, ..., \mathbf{1}\boldsymbol{\tau}_{A_m} x)$, where $\boldsymbol{\tau}_{A_i} = \sum_{a_1 ... a_k \in A_i} \boldsymbol{\tau}_{a_k} ... \boldsymbol{\tau}_{a_1}$. Then verify mechanically.

Application 1: visual comparison of OOMs

Given: two OOMs with same observables S. Both interpretable w.r.t. same characteristic events. If dim = 3, plot "fingerprints" immediately. If dim > 3, project on 3-dim subspace.



JACOBS UNIVERSITY

Application 2: Learning OOMs from data

Interpretable OOMs are at the core of efficient learning algorithms. It's so important that we will use a new section.

6 The basic learning algorithm

Core idea

- In an interpretable OOM:

$$w_0 = (P(A_1),\ldots,P(A_m)),$$
$$\tau_a w_0 = (P(aA_1),\ldots,P(aA_m)),$$
$$\tau_a \tau_b w_0 = (P(baA_1),\ldots,P(baA_m)),$$

etc.

- $w_0, \tau_a w_0, \tau_a \tau_b w_0,$ etc, can be estimated from data by counting frequencies

- Basic linear algebra: obtain τ_a from argument-value pairs

$$w_0 \rightarrow \tau_a w_0,$$
$$\tau_b w_0 \rightarrow \tau_a \tau_b w_0,$$

etc.

Technical execution

- Assume $S = a_1 a_2 \ldots a_N$ is generated by $(\mathbb{R}^m, (\tau_\mathrm{a})_{\mathrm{a} \in O}, w_0)$.
- Task: get estimate $\tilde{\tau}_a$ from S.

Algorithm

- Choose m.
- Choose characteristic events $A_1, \ldots, A_m$.
- Count occurences $\#(A_i A_j)$ and $\#(A_i a A_j)$ and put them into matrices $\mathrm{V} = (\#(A_i A_j))$ and $\mathrm{W}_\mathrm{a} = (\#(A_i a A_j))$.
- Obtain estimate $\tilde{\tau}_a = W_a V^{-1}$.
- Do this for all operators.

Example

Given: *aabbabbbaabbabaabbba*

Step 0: estimate model dim and choose characteristic events. Here: dim = 2, $A_1 = \{\mathrm{a}\}$, $A_2 = \{\mathrm{b}\}$.

Step 1: perform frequency counts of characteristic events:

$$V = \begin{pmatrix} \#aa & \#ba \\ \#ab & \#bb \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix}$$

$$W_a = \begin{pmatrix} \#aaa & \#baa \\ \#aab & \#bab \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 3 & 2 \end{pmatrix}$$

Step 2 and finish:

$$\tilde{\tau}_a = W_a V^{-1} = \begin{pmatrix} 5 & -3 \\ -4 & 3 \end{pmatrix}$$

(do the same for observable b)

Two good properties of learning algorithm

If process is generated by m-dimensional OOM, and m is estimated correctly, the learning algorithm...

- is asymptotically correct (= yields correct model as size of training sequence goes to infinity) regardless of choice of characteristic events,
- is constructive and computationally efficient with $\mathrm{O}(N + |O|\, m^3 / p)$, where p is degree of parallelization.

Standard HMM learning via EM algorithm has neither property 1 nor 2.

Two bad properties of learning algorithm

The algorithm

- depends in its statistical efficiency crucially on the choice of characteristic events – the statistical efficiency problem.

- will often yield a set of operator matrices which violate the condition $\mathbf{1}\tau_{a_n} \cdots \tau_{a_1} w_0 \geq 0$, i.e., the model will assign negative "probabilities" to some sequences – the non-negativity problem.

The first of these two problems has prevented a practical use of OOMs for a long time, and the second has driven at least three people I know almost crazy.

7 Statistically efficient learning algorithms

Characterizers

Definition. Let $k \geq 1$, and $\bar{b}_1 \ldots \bar{b}_\kappa$ be the alphabetical enumeration of O^k. Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$ be an OOM of some process with distribution P and states $w_{\bar{a}}$.

Let $C \in Mat_{m \times \kappa}$ have unit column sums. Then C is a **characterizer** of length k of $\mathcal{A}$ iff for all $\bar{a} \in O^*$:

$$w_{\bar{a}} = C \begin{pmatrix} P(\bar{b}_1 \mid \bar{a}) \\ \vdots \\ P(\bar{b}_\kappa \mid \bar{a}) \end{pmatrix}$$

Intuitive Interpretation

$$w_{\bar{a}} = C \begin{pmatrix} P(\bar{b}_1 \mid \bar{a}) \\ \vdots \\ P(\bar{b}_\kappa \mid \bar{a}) \end{pmatrix}$$

A characterizer C transforms the future distribution after initial history $\bar{a}$ (as represented by the probs $P(\bar{b}_i \mid \bar{a})$) into the OOM state $w_{\bar{a}}$.

Some Properties of Characterizers

1. Every OOM has characterizers of length k for sufficiently large k.
2. Characteristic events, as introduced before, are a special case of characterizers. Example:

   $O = \{\mathrm{a,b}\},\ k = 2,\ m = 3,$

   Characteristic events $\mathrm{A}_1 = \{\mathrm{aa,\ ab}\}$, $\mathrm{A}_2 = \{\mathrm{ba}\}$, $\mathrm{A}_3 = \{\mathrm{bb}\}$:

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} P(aa \mid \bar{c} \\ P(ab \mid \bar{c} \\ P(ba \mid \bar{c} \\ P(bb \mid \bar{c} \end{pmatrix} = \begin{pmatrix} P(A_1 \mid \bar{c}) \\ P(A_2 \mid \bar{c}) \\ P(A_2 \mid \bar{c}) \end{pmatrix} = w_{\bar{c}}\,.$$

Learning Equation

Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$ be an OOM of some process with distribution P with characterizer C. Let

$$V = \begin{pmatrix} P(\bar{b}_1 \mid \bar{a}_1) & \cdots & P(\bar{b}_1 \mid \bar{a}_\kappa) \\ \vdots & & \vdots \\ P(\bar{b}_\kappa \mid \bar{a}_1) & \cdots & P(\bar{b}_\kappa \mid \bar{a}_\kappa) \end{pmatrix}, \quad W_a = \begin{pmatrix} P(a\bar{b}_1 \mid \bar{a}_1) & \cdots & P(a\bar{b}_1 \mid \bar{a}_\kappa) \\ \vdots & & \vdots \\ P(a\bar{b}_\kappa \mid \bar{a}_1) & \cdots & P(a\bar{b}_\kappa \mid \bar{a}_\kappa) \end{pmatrix}.$$

Then

$$\boxed{\tau_a = C W_a (C V)^+}$$

Generalized Learning Algorithm

1. Choose a characterizer C.
2. Estimate (by obvious frequency counting from data)

$$\hat{V} = \begin{pmatrix} \hat{P}(\bar{b}_1 \mid \bar{a}_1) & \cdots & \hat{P}(\bar{b}_1 \mid \bar{a}_\kappa) \\ \vdots & & \vdots \\ \hat{P}(\bar{b}_\kappa \mid \bar{a}_1) & \cdots & \hat{P}(\bar{b}_\kappa \mid \bar{a}_\kappa) \end{pmatrix}, \quad \hat{W}_a = \begin{pmatrix} \hat{P}(a\bar{b}_1 \mid \bar{a}_1) & \cdots & \hat{P}(a\bar{b}_1 \mid \bar{a}_\kappa) \\ \vdots & & \vdots \\ \hat{P}(a\bar{b}_\kappa \mid \bar{a}_1) & \cdots & \hat{P}(a\bar{b}_\kappa \mid \bar{a}_\kappa) \end{pmatrix}.$$

3. Compute $\hat{\tau}_a = C\hat{W}_a(C\hat{V})^+$.

Properties of General Learning Algorithm(s)

1. Yields asymptotically correct estimates $\hat{\tau}_a$ with any characterizer C.

2. Model variance (statistical efficiency) depends crucially on choice of C.

3. Search for "good" (low model variance, i.e. high statistical efficiency) learning algorithms boils down to optimizing C.

Algorithms for characterizer optimizing on the market today

1. Error controlling algorithm: M. Zhao, H. Jaeger, M. Thon (2009): **A Bound on Modeling Error in Observable Operator Models and an Associated Learning Algorithm.** Neural Computation, posted online 6/2009, doi: 10.1162/neco.2009.01-08-687
2. An unnamed, PCA based algorithm: Rosencrantz, M., Gordon, G., Thrun, S. (2004): **Learning Low Dimensional Predictive Representations**. Proc. 21st Int. Conf. on Machine Learning (ICML), Banff, Canada, 2004
3. Efficiency sharpening algorithm: H. Jaeger, M. Zhao, K. Kretzschmar, T. Oberstein, D. Popovici, A. Kolling (2006): **Learning observable operator models via the ES algorithm.** In: S. Haykin, J. Principe, T. Sejnowski, J. McWhirter (eds.), New Directions in Statistical Signal Processing: from Systems to Brain. MIT Press, Cambridge, MA., 417-464

Notes on algorithms 1 & 2

- Algorithms 1 and 2 yield equivalent results (M. Thon, in preparation)
- Core idea: set $C = L_m^{\mathsf{T}}$, where L_m is made from the first m singular vectors of $\hat{V}$ (i.e., $C\hat{V}$ is the PCA-transform of $\hat{V}$).
- Theory (M. Zhao 2007, 2009): this C minimizes an upper bound on the relative error e of estimated operators $\hat{\tau}$ over the true τ:

$$e = \left\|\hat{T} - T\right\|_{Frob} / \left\|T\right\|_{Frob}, \quad \text{where } \hat{T} = \left(\hat{\tau}_{a_1} \ldots \hat{\tau}_{a_l}\right), T = \left(\tau_{a_1} \ldots \tau_{a_l}\right)$$

- Resource problem: algorithms have time and space complexity that scales with $m\,N^3$ in the worst case, where m is model dimension and N training data length.
- Both algorithms are constructive.

Notes on algorithm 3

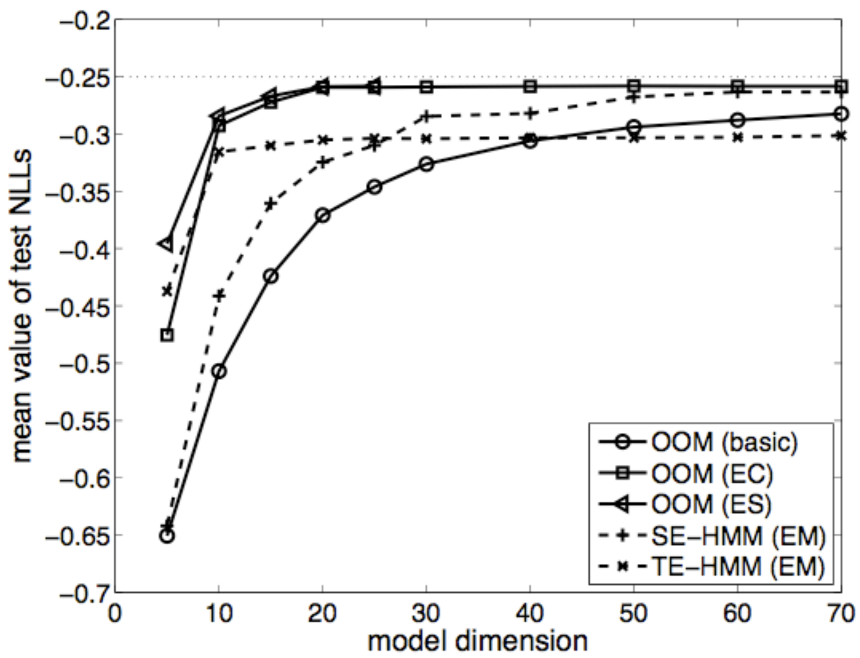
- Core idea: exploit a certain algebraic (!) characterization of the statistically maximally efficient characterizer $C_{max\text{-}eff}$. Approximate this precious $C_{max\text{-}eff}$ by an iterative re-estimation method.
- About 2-5 iterations usually suffice.
- One iteration has time and space cost scaling with $m^2 N$ (as opposed to worst-case $m N^3$ for algorithms 1 & 2).
- Algorithm does not necessarily converge (can jitter around terminal value). For too large assumed model dimensions prone to numerical instability. Iteration dynamics is not understood.

Notes on all three algorithms

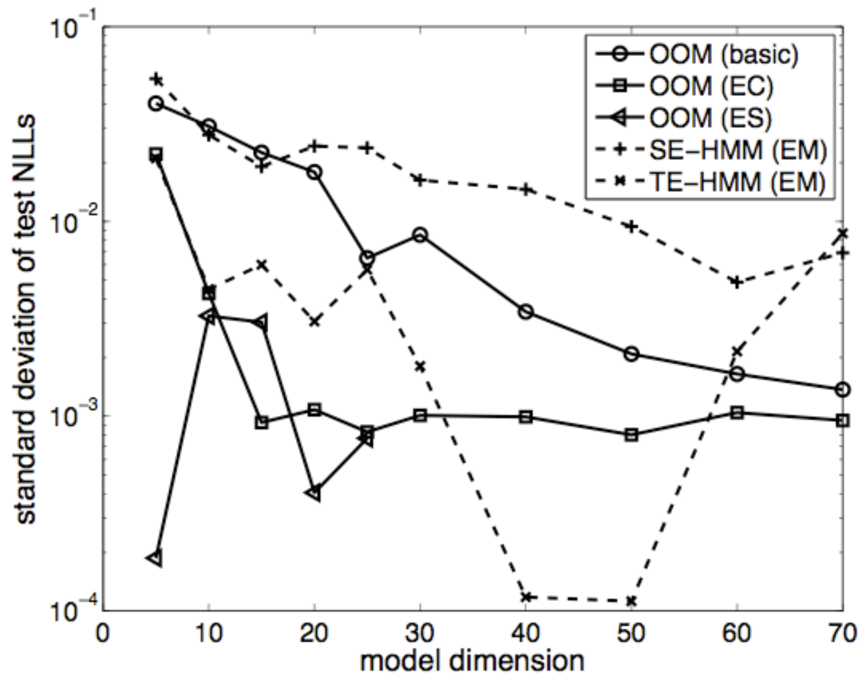
- These algorithms are rooted in the OOM-typical translation of stochastic concepts into algebraic ones:
  - algorithm 1 exploits algebraic characterization of maximal statistical efficiency,
  - the other two minimize estimation error bound on metric distance between estimated and true model matrices.
- All algorithms are technically involved and need care when implementing them in space/time efficient ways.
- Model quality is empirically found similar for algorithms 1 & 2 vs. algorithm 3
- Model accuracy (statistical efficiency) is far superior to EM-trained HMMs
- All algorithms by design are insensitive to overfitting (test performance does not decrease when model dim is chosen too big)
- Computational cost of algorithm 3 is about 10 times less than EM-learning of HMMs due to low number of iterations
- Average cost of algorithms 1 & 2 appears to be much less than that of algorithm 3 (worst-case cost is however much higher), depends much on nature of process, needs analysis

Demo 1: logistic chaos process[1)]

- Data from 16-bin discretized logistic process $x(n+1) = 4\,x(n)\,(1 - x(n))$, which is strongly chaotic (max. Lyapunov exponent = 2)



a.



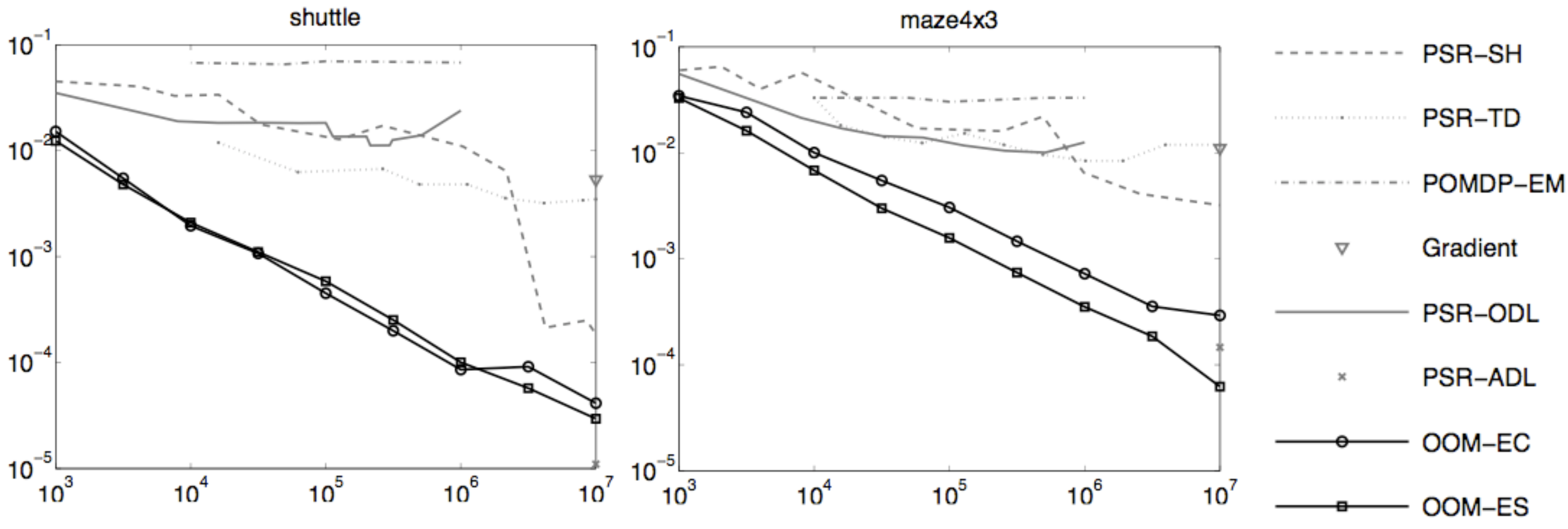
b.

- CPU times here were about 1:10 of algorithm 3 vs. EM-HMM, and again 1:10 of algorithm 1 vs. algorithm 3

1) From Zhao/Jaeger/Thon, Neural Computation, forthcoming

Demo 2: some standard benchmarks from the PSR community[1)]

- These are input-output processes; OOMs can accomodate
- 2 out of 7 examples shown, others are similar
- Figures show average 1-step prediction error vs. training data length



1) From Zhao/Jaeger/Thon, Neural Computation, forthcoming

Efficient learning algorithms: summary

- The problem of finding statistically efficient versions of the basic OOM learning algorithm has essentially been solved.
- Algorithms starkly outperform EM-HMM in accuracy and cost.
- Algorithms use novel learning principles:
  - Algorithms 1 & 2: minimizing error bound on model parameters
  - Algorithm 3: optimizing statistical efficiency of asymptotically correct estimator
- More research needed:
  - Algorithms 1 & 2: improving worst-case cost
  - Algorithm 3: analysis of iteration dynamics and numerical stability
- Algorithms are much more complicated than EM-HMM.
- Overview and analytic comparison/unification paper (M. Thon) is in preparation.

8 The dreaded nonnegativity problem

- Recall defining conditions of OOM $(\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$:

  1. $\mu = \sum_{a\in\Sigma} \tau_a$ has column sums = 1
  2. $\mathbf{1} w_0 = 1$
  3. for every sequence $a_1 \ldots a_n \in O^n$ it holds that

$$\mathbf{1}\tau_{a_n} \cdots \tau_{a_1} w_0 \geq 0$$

- Conditions 1. and 2. are easy to check; the non-negativity condition 3. isn't.
- Learnt models often (even typically, for nontrivial data) violate non-negativity.
- Utterly desirable: method to check for non-negativity condition; method to transform invalid learnt OOM into "closest" valid one.
- Every OOM researcher I know has burnt lots of lifetime on this problem, aging prematurely in the process.

Three solutions to the dreaded problem

1. **Empirical workaround**: when an invalid model is used in prediction / generation, and invalid (negative-probability) states occur, renormalize them on the fly.
   - A recommended method is detailed in [1]

2. **Emphatic anti-solution**: it is undecidable whether a set of candidate operator matrices satisfies the nonnegativity condition.
   - Proof by E. Wiewora [2], by adaptation of a related proof by Denis and Esposito [3]

3. **Emperor's solution**: disallow non-negativity by using norm-OOMs, which are built around the idea to set

$$P(a_1 \dots a_n) = \left\| \boldsymbol{\tau}_{a_n} \dots \boldsymbol{\tau}_{a_1} w_0 \right\|^2$$

   - Introduced by M. Zhao [4,5], including a general stochastic framework and a basic learning algorithm.

1. H. Jaeger, M. Zhao, K. Kretzschmar, T. Oberstein, D. Popovici, A. Kolling (2006): Learning observable operator models via the ES algorithm. In: S. Haykin, J. Principe, T. Sejnowski, J. McWhirter (eds.), New Directions in Statistical Signal Processing: from Systems to Brain. MIT Press, Cambridge, MA., 417-464
2. Wiewora, E. W. 2008. Modeling probability distributions with predictive state representations. PhD thesis, Dpt. of Computer Science, Univ. of California, San Diego
3. Denis, F. and Esposito, Y., 2004. Learning Classes of Probabilistic Automata. In: Learning Theory: Springer LNCS 3120, 124-139
4. M. Zhao, H. Jaeger (2007): **Norm observable operator models.** Jacobs University technical report Nr. 8
5. Zhao, M. and Jaeger, H. (2010). Norm Observable Operator Models. Neural Computation, to appear

JACOBS UNIVERSITY

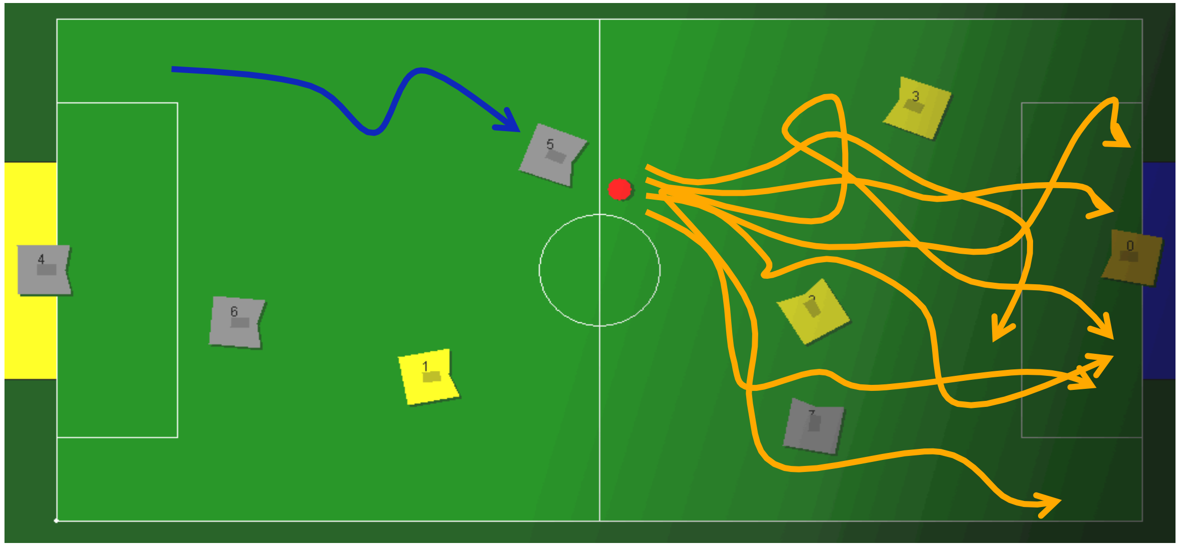
7 From stochastic processes to OOMs

So far, we introduced OOMs as generalizations of HMMs.

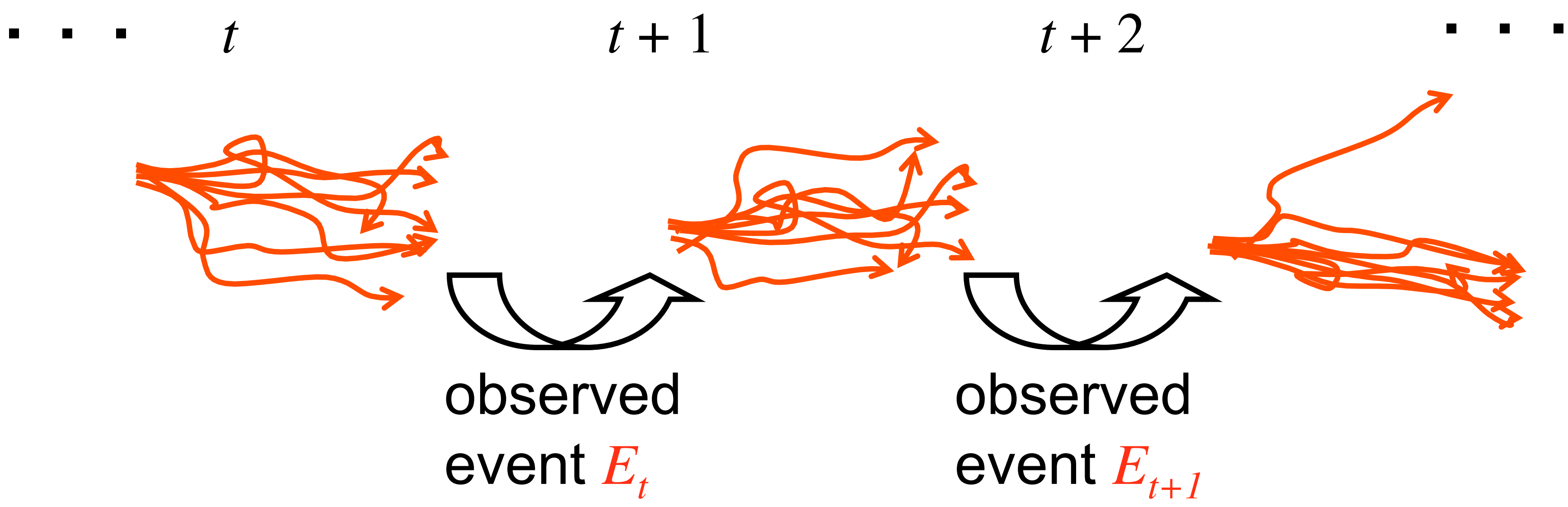
Now we will re-introduce OOMs in a very different way, starting from stochastic processes and showing that (basically) every stochastic process has an OOM.

What's a future?

For a robot, or anybody else modelling stochastic processes, the future is a probability distribution over possible future developments.



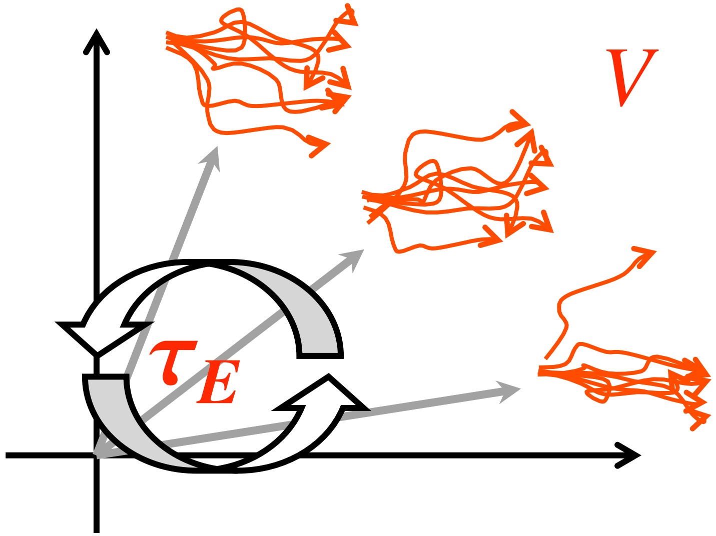
Dynamics of future distributions



- Observations update expectations, that is, future distributions.
- ... E_t , E_{t+1} , ... : observations. Formally, events in observation space s-algebra.

The basic idea

observable events = operators that change distributions



- set of all distributions is a (functional) vector space V
- for every event E an "observable operator" τ_E
- observable operators operate on V

Characterizing SD processes 1

The distribution of a stationary, discrete-valued process (SD process) is fully characterized by its conditioned continuation probabilities

$$P(X_{n+1} = b_1, \ldots, X_{n+m} = b_m \mid X_1 = a_1, \ldots, X_n = a_n)$$

$$=: P(b_1 \ldots b_m \mid a_1 \ldots a_n)$$

$$=: P(\overline{b} \mid \overline{a})$$

where $m \geq 1, n \geq 0.$

Special case $n = 0$: $P(\overline{b} \mid \varepsilon) = P(\overline{b})$ would suffice.

JACOBS UNIVERSITY

Characterizing SD processes 2

SD process
$\cong$ all $P(\bar{b} \mid \bar{a})$

Consider the vector space of all numerical functions on finite sequences,

$$\mathcal{D} = \{d : O^* \rightarrow \mathbb{R}\}$$

For each antecedent $\bar{a}$, define a predictor function

$$g_{\bar{a}} : O^* \rightarrow \mathbb{R}, \; g_{\bar{a}}(\bar{b}) = P(\bar{b} \mid \bar{a})$$

Shorthand: $g_{\bar{a}} = P(\cdot \mid \bar{a})$

The set of all such predictor functions,

$$\{g_{\bar{a}} \mid \bar{a} \in O^*\} \subset \mathcal{D}$$

describes all $P(\bar{b} \mid \bar{a})$ and thus characterizes the process.

Characterizing SD processes 3

SD process

$\cong$

all $P(\bar{b} \mid \bar{a})$

$\cong$

$\{g_{\bar{a}} \mid \bar{a} \in \Sigma^*\}$

Consider the linear subspace spanned by all predictor functions $g_{\bar{a}}$,

$$\mathcal{G} = [\{g_{\bar{a}} \mid \bar{a} \in \Sigma^*\}]_{\mathcal{D}}$$

Let $t_a : \mathcal{G} \to \mathcal{G}$ be a linear mapping satisfying

$$t_a(g_{\bar{c}}) = P(a \mid \bar{c})\, g_{\bar{c}a}$$

for all $a \in O, \bar{c} \in O^*$. (They exist!)

Let

$$g_\varepsilon : O^* \to \mathbb{R},\;\; g_\varepsilon(\bar{b}) = P(\bar{b} \mid \varepsilon) = P(\bar{b})$$

Let $\mathbf{1} : \mathcal{G} \to \mathbb{R}$ be a linear mapping satisfying $\mathbf{1} g_{\bar{c}} = 1$ for all $\bar{c} \in O^*$. (exists!)

Characterizing SD processes 4

SD process $\cong$ $\{g_{\bar{a}} \mid \bar{a} \in \Sigma^*\}$
$\mathcal{G} =$ $[\{g_{\bar{a}} \mid \bar{a} \in O^*\}]_{\mathcal{D}}$
$t_a(g_{\bar{c}}) =$ $P(a \mid \bar{c}) g_{\bar{c}a}$
$g_\varepsilon(\bar{b}) = P(\bar{b})$
$\mathbf{1} g_{\bar{c}} = 1$

Theorem.
For any $a_1 \ldots a_n \in O^*$ it holds that

$$P(a_1 \ldots a_n) = \mathbf{1} t_{a_n} \cdots t_{a_1} g_\varepsilon$$

Compare:

$$P(a_1 \cdots a_n) = \mathbf{1} \boldsymbol{\tau}_{a_n} \cdots \boldsymbol{\tau}_{a_1} w_0$$

Definition. $\dim(\mathcal{G})$ is the dimension of the process.

Corollary. A finite-dimensional process of dimension m has a "matrix" OOM

$$(\mathbb{R}^m, (\tau_a)_{a \in \Sigma}, w_0) \cong (\mathcal{G}, (t_a)_{a \in \Sigma}, g_\varepsilon).$$

Characterizing SD processes 5

SD process $\cong$ $\{g_{\bar{a}} \mid \bar{a} \in \Sigma^*\}$
$\mathcal{G} = [\{g_{\bar{a}} \mid \bar{a} \in O^*\}]_{\mathcal{D}}$
$t_a(g_{\bar{c}}) = P(a \mid \bar{c}) g_{\bar{c}a}$
$g_\varepsilon(\bar{b}) = P(\bar{b})$
$\mathbf{1} g_{\bar{c}} = 1$
$(\mathcal{G}, (t_a)_{a \in \Sigma}, g_\varepsilon)$

Every SD process has an "abstract" OOM $(\mathcal{G}, (t_a)_{a \in \Sigma}, g_\varepsilon)$.

These abstract OOMs are unique ("coordinate-free representation").

The dimension of a process may be infinite.

Abstract OOMs are needed for proving the equivalence theorem.

8 General OOM theory 1

Theorem. Let $(X_t)_{t \geq 0}$ be a process with values in $(B, \mathcal{B})$, not necessarily stationary. Then there exists an OOM

$$(\mathbb{R}^K, (\boldsymbol{\tau}_{A,t})_{A \in \mathcal{B}, t>0}, w_0)$$

such that

$$P(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n)$$

$$= \mathbf{1}\boldsymbol{\tau}_{A_n, t_n - t_{n-1}} \cdots \boldsymbol{\tau}_{A_1, t_1} w_0.$$

Furthermore, it holds that

(1) $$\boldsymbol{\tau}_{\bigcup_n A_n, t} = \sum_n \boldsymbol{\tau}_{A_n, t}$$

(2) $$\boldsymbol{\tau}_{A, t_1 + t_2} = \boldsymbol{\tau}_{A, t_2} \boldsymbol{\tau}_{B, t_1}$$

Decomposing OOMs 1

Recall: observable operators of OOMs derived from HMMs have the form

$$T_a = M^{\mathrm{T}}\, \mathrm{O}_a$$

where M is the transition matrix of a Markov chain and O_a is a (diagonal) observation matrix containing emission probabilities of a.

Decomposing OOMs 2

Theorem. Let $(X_t)_{t\geq 0}$ be a process with values in $(B,\mathcal{B})$, not necessarily stationary. Then there exists an OOM with

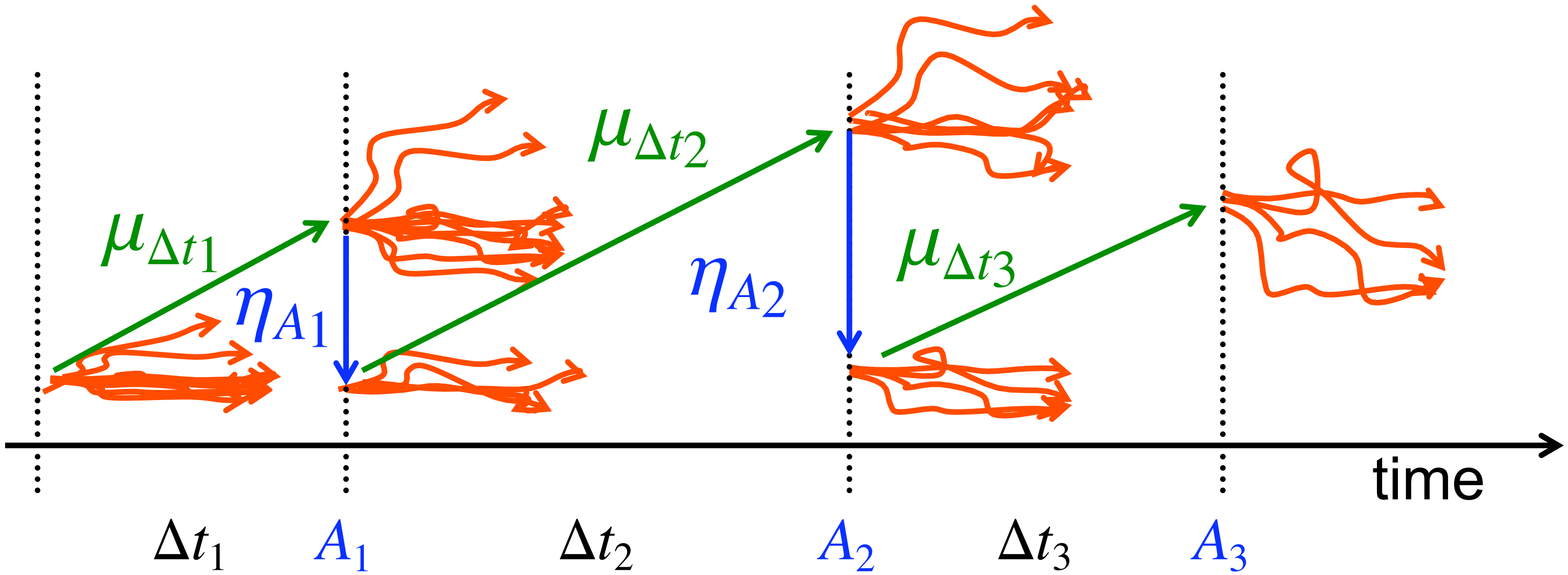
evolution operators $(\mu_r)_{r>0}$
observation operators $(\eta_A)_{A\in\mathrm{B}}$

such that

$$P(X_0 \in A_0, X_{t_1} \in A_1, \ldots, X_{t_{n-1}} \in A_{n-1})$$
$$= \mathbf{1}\eta_{A_{n-1}}\mu_{t_{n-1}-t_{n-2}} \cdots \eta_{A_2}\mu_{t_2-t_1}\eta_{A_1}\mu_{t_1}\eta_{A_0}w_0.$$

Decomposing OOMs 3

Visualization of evolution operators $\mu_{\Delta t}$ and observation operators η_A



From linear algebra back to processes

Recall: in an abstract OOM $(\mathbb{R}^K, (\tau_{A,t})_{A\in\mathcal{B}, t>0}, w_0)$ we obtain

$$P(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = \mathbf{1}\boldsymbol{\tau}_{A_n, t_n - t_{n-1}} \cdots \boldsymbol{\tau}_{A_1, t_1} w_0.$$

Theorem. Let $(B, \mathcal{B})$ be a polish measure space, V a real vector space with basis E, $w_0 \in V$, $(\boldsymbol{\tau}_{A,t})_{A\in\mathrm{B}, t>0}$ a family of linear operators on V, V be generated by the vectors $\boldsymbol{\tau}_{A_n, t_n - t_{n-1}} \cdots \boldsymbol{\tau}_{A_1, t_1} w_0$ and the numerical function $P : (\mathcal{B} \times \mathbb{R}^+)^* \rightarrow \mathbb{R}$ be defined by $P((A_1, t_1), \ldots, (A_n, t_n)) = \mathbf{1}\boldsymbol{\tau}_{A_n, t_n - t_{n-1}} \cdots \boldsymbol{\tau}_{A_1, t_1} w_0.$

Then P can be extended to the distribution of a stochastic process iff

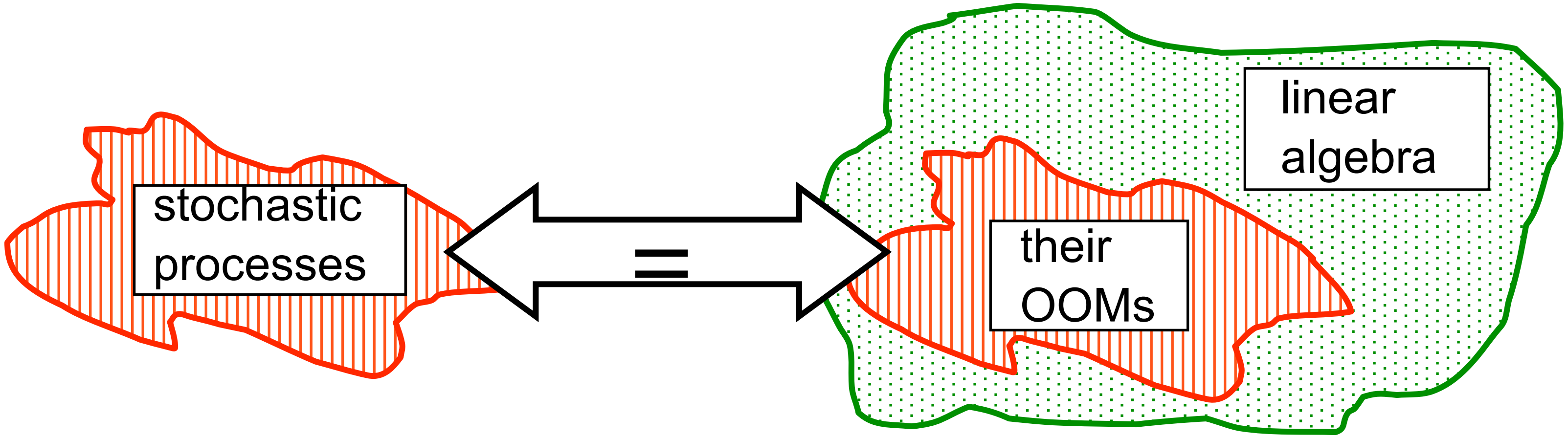
(1) $\mathbf{1} w_0 = 1$ (2) $\mathbf{1}\boldsymbol{\tau}_{(B,t)} e = 1$ for all basis vectors e and times t

(3) $\mathbf{1}\boldsymbol{\tau}_{A_n, t_n - t_{n-1}} \cdots \boldsymbol{\tau}_{A_1, t_1} w_0 \geq 0$ for all τ sequences

(4) $\boldsymbol{\tau}_{\dot{\bigcup}_n A_n, t} = \sum_n \boldsymbol{\tau}_{A_n, t}$ (5) $\boldsymbol{\tau}_{A, t_1 + t_2} = \boldsymbol{\tau}_{A, t_2} \boldsymbol{\tau}_{B, t_1}$

From processes to linear algebra and back to processes

$$P(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = \mathbf{1}\boldsymbol{\tau}_{A_n, t_n - t_{n-1}} \cdots \boldsymbol{\tau}_{A_1, t_1} w_0.$$



The theory of distributions of stochastic processes (with polish measure spaces and real or discrete time) becomes a subtheory of linear algebra.

Historical notes / related approaches

1957 - 1970	A long series of investigations in mathematical probability theory concerning the question when two HMMs are equivalent (overviews in [1] [2])
1984	Ito/Amari/Kobayashi [1] solve problem by embedding HMM processes in OOM-like processes. Further refinements and extensions in [3] [4]
... - 1969	A Roumanian school of probability theory develops theory to describe stochastic processes by observable operators (although it is not recognized that they can always be chosen linear) [5]
1997	Upper [6] and Jaeger [7][8] independently find that OOM processes can be learnt by estimating linear operators. Jaeger introduces OOM formalism.
2001	Littman/Sutton/Singh [9] introduce predictive state representations (PSRs) for input-driven processes, unaware of IO-OMMs described earlier by Jaeger [10].
1969	Schützenberger [11] introduces multiplicity automata (MAs), which are equivalent to finite-dimensional OOMs, expressed in a context of automata theory.
1980's - present	A series of investigations in statistical learning theory and stochastic languages on learnability and decidability issues concerning MAs. Among other, it is found that the non-negativity problem is undecidable [12][13]
antiquity - present	Ancient idea in quantum mechanics, information theory [14] and statistical physics [15]: the state of a physical system is that which contains all information about the future

1. H. Ito, S.-I. Amari, and K. Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. IEEE trans- actions on information theory, 38(2):324–333, 1992
2. H. Jaeger, M. Zhao, K. Kretzschmar, T. Oberstein, D. Popovici, A. Kolling (2006): Learning observable operator models via the ES algorithm. In: S. Haykin, J. Principe, T. Sejnowski, J. McWhirter (eds.), New Directions in Statistical Signal Processing: from Systems to Brain. MIT Press, Cambridge, MA., 417-464
3. V. Balasubramanian. Equivalence and reduction of Hidden Markov models. A.I. Technical Report 1370, MIT AI Lab, 1993
4. H. Ito. An algebraic study of discrete stochastic systems. Phd thesis, Dpt. of Math. Engineering and Information Physics, 1992.
5. M. Iosifescu and R. Theodorescu. Random Processes and Learning, volume 150 of Die Grundlagen der mathematischen Wissenschaften in Einzel- darstel lungen. Springer Verlag, 1969
6. D.R. Upper. Theory and algorithms for Hidden Markov models and Generalized Hidden Markov models. Phd thesis, Univ. of California at Berkeley, 1997.
7. H. Jaeger. Observable operator models and conditioned continuation rep- resentations. Arbeitspapiere der GMD 1043, GMD, Sankt Augustin, 1997.
8. H. Jaeger. Observable operator models II: Interpretable models and model induction. Arbeitspapiere der GMD 1083, GMD, Sankt Augustin, 1997
9. M. L. Littman, R. S. Sutton, and S. Singh. Predictive representation of state. In Advances in Neural Information Processing Systems 14 (Proc. NIPS 01), pages 1555–1561, 2001
10. H. Jaeger. Discrete-time, discrete-valued observable operator models: a tutorial. GMD Report 42, GMD, Sankt Augustin, 1998
11. Schützenberger, M. P. 1961. On the definition of a family of automata. Inf. Control 4, 245–270
12. Denis, F. and Esposito, Y., 2004. Learning Classes of Probabilistic Automata. In: Learning Theory: Springer LNCS 3120, 124-139
13. Wiewora, E. W. 2008. Modeling probability distributions with predictive state representations. PhD thesis, Dpt. of Computer Science, Univ. of California, San Diego
14. Zadeh, L.A. (1969): The Concept of System, Aggregate, and State in System Theory. In: Zadeh, L.A. and Polak, E. (eds.), System Theory, McGraw-Hill, New York 1969
15. Shalizi, C. R. and Crutchfield, J. P. (2001). Computational Mechanics: Pattern and Prediction, Structure and Simplicity. J. Statistical Mechanics 104(314), 817-879

Norm OOMs [1,2]

- Motivation: avoid the non-negativity problem of standard linear OOMs
- Approach: keep basic structure of OOMs: $(\mathbb{R}^m, (\boldsymbol{\tau}_a)_{a \in O}, w_0)$, but compute probabilities from states by

$$P(a_1 \ldots a_n) = \left\| \boldsymbol{\tau}_{a_n} \ldots \boldsymbol{\tau}_{a_1} w_0 \right\|^2 .$$

- This avoids the non-negativity problem by design.

Norm OOMs, cont'd

Definition. Let $O = \{a_1, ..., a_k\}$ be a finite set of observables, and let E be a real vector space with an inner product (and hence, a norm). Let $w_0 \in E$ and for each $a \in O$, let τ_a be a linear map on E, and $\tau_a{}^*$ its adjoint operator (i.e., $\langle \tau_a{}^* u, v \rangle = \langle u, \tau_a v \rangle \; \forall u,v \in E$). Then $(E,(\tau_a)_{a\in O}, w_0)$ is a norm-OOM, if

1. $\|w_0\| = 1,$ 2. $\sum_{a\in O} \tau_a{}^* \tau_a = id_E.$

Theorem. If $(E,(\tau_a)_{a\in O}, w_0)$ is a norm-OOM, then the prescription

$$P(a_1 \dots a_n) = \left\| \tau_{a_n} \dots \tau_{a_1} w_0 \right\|^2$$

describes the distribution of a stochastic process.

Theorem. Every stochastic process with observables $O = \{a_1, ..., a_k\}$ has a norm-OOM $(E,(\tau_a)_{a\in O}, w_0)$ which describes the distribution of the process by the above formula.

Norm OOMs, notes

- Mingjie Zhao [2] found a constructive, asymptotically correct learning algorithm for norm-OOMs.
- This algorithm is computationally prohibitively expensive. Mingjie explores tractable versions.
- Mingjie also has found another, iterative, EM-based learning algorithm (manuscript in preparation).
- Unlike the deplorable case of linear OOMs, it is decidable whether a structure $(\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$ is a norm-OOM.
- Every finite-dimensional norm-OOM $(\mathbb{R}^m, (\tau_a)_{a \in O}, w_0)$ can be effectively transformed into an equivalent (higher-dimensional) linear OOM. It appears that processes obtained from randomly generated norm-OOMs are generically non-HMM.
- It is unknown whether finite-dim HMM processes are a subclass of finite-dim norm-OOM processes. All that is known is that m-state Markov Chains have an m-dimensional norm-OOM.

1. M. Zhao, H. Jaeger (2007): **Norm observable operator models.** Jacobs University technical report Nr. 8
2. M. Zhao, H. Jaeger (2010): **Norm Observable Operator Models** Neural Computation, to appear

9 Beginnings of a Hilbert space theory 1

Consider a K-dimensional process $(\Omega, \mathcal{A}, P, X_t)$ with discrete values and one of its OOMs $(\mathbb{R}^K, (\tau_a)_{a \in \Sigma}, w_0)$.

For every $x \in \mathbb{R}^K$, one can construct a P-measurable function $\gamma(x) : \Omega \to \mathbb{R}$ where $\gamma(w_0) \equiv 1$, such that the following holds:

Theorem. (i) $\forall x \in \mathbb{R}^K : \gamma(x) \in \mathcal{L}^\infty(P)$

(ii) $\langle x, y \rangle := \int_\Omega \gamma(x)\gamma(y)\,dP$ defines an inner product and thereby a norm on $\mathbb{R}^K$

(iii) The operators $(\tau_a)_{a \in \Sigma}$ are continuous w.r.t. this norm

Hilbert space theory 2: construction of γ

1. For every $w = \boldsymbol{\tau}_{\bar{b}} w_0 \,/\, \mathbf{1}\boldsymbol{\tau}_{\bar{b}} w_0$, where $P(\bar{b}) > 0$,

   we obtain a measure μ_w on $(\Omega, \mathcal{A})$ by extending

$$\mu_w(a_1 \ldots a_n) = \mathbf{1}\boldsymbol{\tau}_{a_n} \ldots \boldsymbol{\tau}_{a_1} w$$

2. For every such w, $\gamma(w)$ is defined as the density of μ_w w.r.t. P.

3. There exists a basis of $\mathbb{R}^K$ consisting of such w's.

   For $x \in \mathbb{R}^K$, $x = \sum_{i=1}^{k} \alpha_i w_i$, define

$$\gamma(x) = \sum \alpha_i \gamma(w_i)$$

Hilbert space theory 3: the open issue

Open question[1]**:** it is not clear under which conditions the metric space $\mathbb{R}^K$ (where the metric is the one induced by $\langle x, y \rangle := \int_\Omega \gamma(x)\gamma(x)dP$) is complete.

If we had a complete vector space, it would be a Hilbert space and we could develop an approximation theory (of infinite-dimensional operators by finite-dimensional).

1) Pointed out by A. Schönhuth

10 Research topics

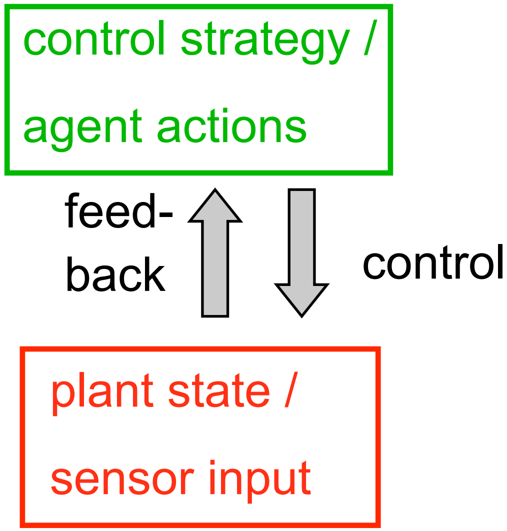
- Algebraic characterization of OOM matrices
- Characterization of OOMs that are HMMs
- Recovery of discrete "hidden" event structure from observation sequences
- Learning nonstationary OOMs
- Online learning algorithms
- Spatio-temporal OOMs, "Bayesian network" OOMs
- OOM and quantum mechanics
- OOMs in speech processing and biosequence modeling
- Efficient "direct" and online learning algorithms for input-output OOMs
- Incorporating prior knowledge into learning
- Learning with missing values and unequal observation intervals
- Characterization of standard processes
- Development of Hilbert space theory

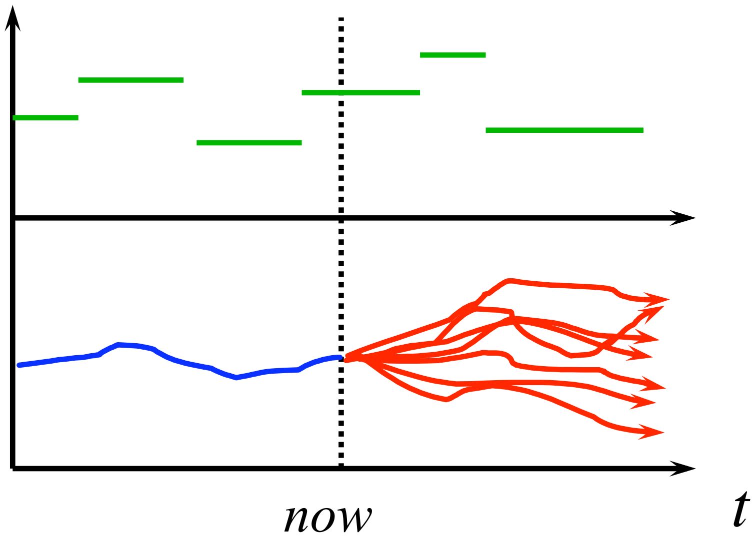
11 A survey of further results

11.1 Input-output OOMs and Predictive State Representations

Controlled stochastic processes

- In open systems, future distributions depend on input.

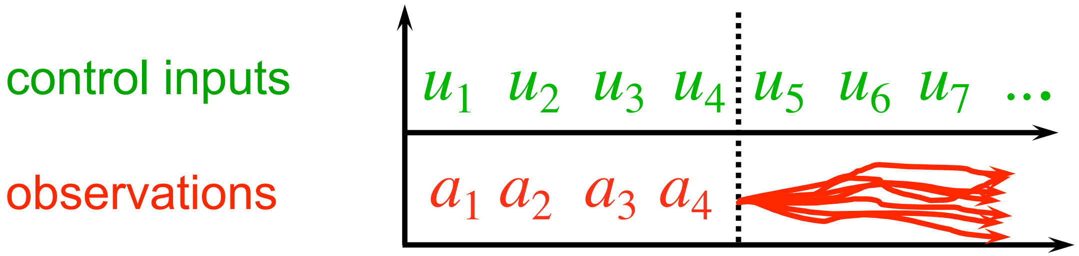




- Formally, a *controlled stochastic process* [1] is defined by conditional probabilities of the kind

$$P(\boxed{X_{n+1} = a} | \boxed{X_{n-k} = a_0, \ldots, X_n = a_k}, \boxed{U_{n-k} = u_0, \ldots, U_n = u_k, U_{n+1} = u})$$

Input-Output OOMs (IO-OOMs)



$$(\mathbb{R}^m, (\tau_a^u)_{a \in O,\ u \in U}, \sigma, w_0)$$

- An IO-OOM [2] is essentially a set of OOMs of same dimension; these OOMs are indexed by possible inputs; input switches between them.
- IO-OOMs standardly use σ, not $\mathbf{1}$, for projection of states on probabilities.
- If at time n the IO-OOM state is w_n, the probability to observe a at time $n+1$, given that input at time $n+1$ is u, is

$$P(X_{n+1} = a | w_n) = \sigma \tau_a^u w_n .$$

- The probability to see observations $a_1, \ldots, a_n$, given control input $u_1, \ldots, u_n$, is

$$P(X_1 = a_1, \ldots X_n = a_n \mid U_1 = u_1, \ldots, U_n = u_n) = \sigma \tau_{a_n}^{u_n} \cdots \tau_{a_1}^{u_1} w_n .$$

Learning IO-OOMs[1)]

$$(\mathbb{R}^m, (\tau_a^u)_{a\in O,\, u\in U}, \sigma, w_0)$$

Given: training sequence $u_1 a_1 \ldots u_N a_N$.

1. Choose κ indicative & characteristic sequences, typically $(U \times O)^l$.
2. Let $\hat{V} = \left(\hat{P}(\bar{q}^{\,j}\bar{c}^{\,i})\right)_{i,j}$ and $\hat{W}_{ua} = \left(\hat{P}(\bar{q}^{\,j}\, ua\,\bar{c}^{\,i})\right)_{i,j}$ and $\hat{c} = \left(\hat{P}(\bar{c}^{\,i})\right)_i$ and $\hat{q}^{\mathsf{T}} = \left(\hat{P}(\bar{q}^{\,j})\right)_j$.

   $$\text{Note:}\quad \hat{P}(u_1 a_1 \cdots u_l a_l) = \prod_{n=1,\ldots,l} \frac{\# u_1 a_1 \cdots u_n a_n}{\# u_1 a_1 \cdots u_{n-1} a_{n-1} u_n}.$$
3. Estimate dimension m of IO-OOM as $numrank(\hat{V})$.
4. Scale colums of $\hat{V}$ and $\hat{W}_{ua}$ by $\sqrt{\# \bar{q}_j}$.
5. Choose characterizer $C \in \mathbb{R}^{m\times\kappa}$ such that $C\hat{V}$ is invertible.
6. Set
   $$\hat{\tau}_a^u = C\hat{W}_{ua}(C\hat{V})$$
   $$\hat{w}_0 = C\hat{c}$$
   $$\hat{\sigma}^{\mathsf{T}} = \hat{q}^{\mathsf{T}}(C\hat{V})^{-1}$$

Note: step 5 is where all the effort and quality lies. Re-use the efficient OOM-learning algorithms here.

1) adapted from a slide of M. Thon

Predictive state representations (PSR)

- PSRs are equivalent to IO-OOMs, using a slightly different formalism.
- Independently discovered by Littman, Sutton and Singh 2001 [3]. Context: modeling action selection of agents in stochastic environments; PSRs introduced as generalizations of POMDPs. Now a fertile field of research (try Google).
- Basic concept: tests. A test t is any sequence $u_1 a_1 \ldots u_l a_l$ of input/observation pairs.
- For an m-dimensional (in the sense of IO-OOMs) controlled stochastic process, there exist m core tests $t_1, \ldots, t_m$, s.th. for any history $h = u_1 a_1 \ldots u_N a_N$, the predictive state $p(h) = (P(t_1|h), \ldots, P(t_m|h))^{\mathsf{T}}$ – i.e., an m-dimensional column vector – is a sufficient statistic of the future distribution of the process.
- This amounts to the following. For every history h, next input u and observation a, one can compute from $p(h)$ the probability $P(a \mid h, u)$ to see a under this input, by

$$P(a \mid h, u) = m_{ua}\, p(h),$$

  where m_{ua} is an m-dimensional row vector which depends only on u and a.
- PSRs amount thus to IO-OOMs whose states are interpretable w.r.t. the core tests.

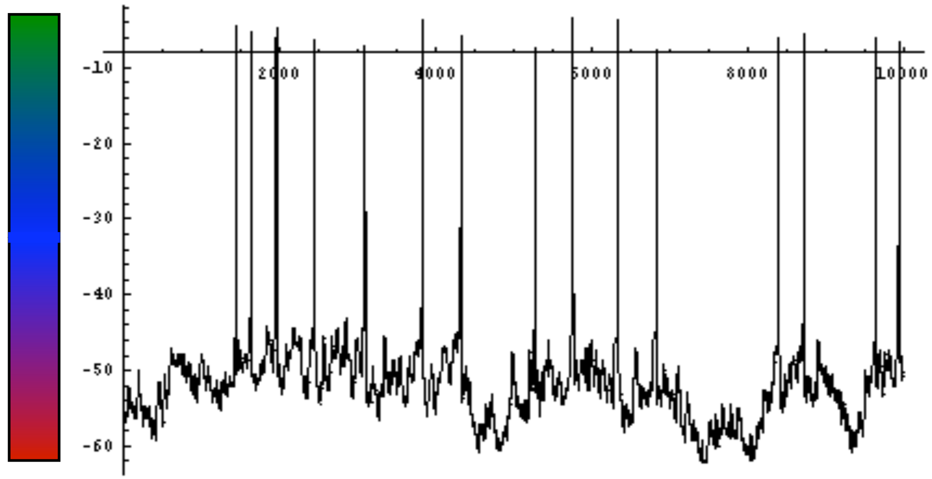
JACOBS UNIVERSITY

1. Gihman, I.I. and Skorohod, A. V. (1979): Controlled Stochastic Processes. Springer Verlag 1979
2. H. Jaeger (1998): **Observable operator models of stochastic processes: a tutorial.** GMD Report 42, German National Research Center for Information Technology 1998. (Section 10)
3. Littman, M. L., Sutton, R. S., Singh, S. (2001), Predictive representation of state. In: Advances in Neural Information Processing Systems 14 (Proc. NIPS 01), 1555-1561

11.2 Mixture OOMs

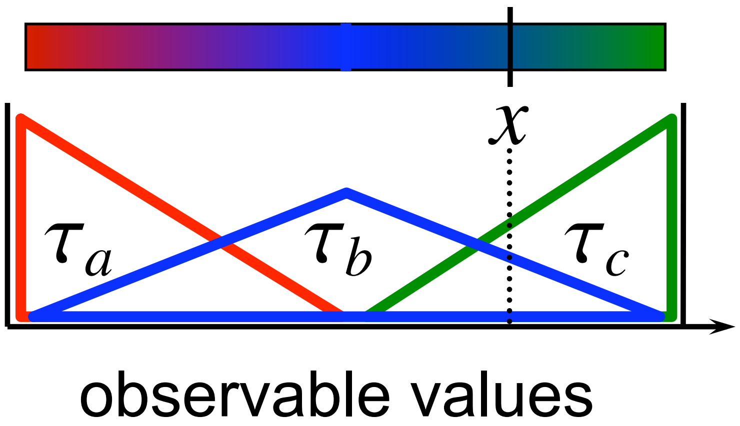
Problem:

Continuous-valued processes have continuously many observable operators.



A solution [1]**:**

Combine observable operators from finite number of basis operators through membership functions.



$\tau_x =$
$0.3\ \tau_b$
$+0.7\ \tau_c$

Mixture OOMs, Results 1

- Fundamental equation transfers

$$\boxed{P(X_1 \in I_1, \ldots, X_k \in I_k) = \sigma \circ \tau_{I_k} \circ \cdots \circ \tau_{I_1} v_0}$$

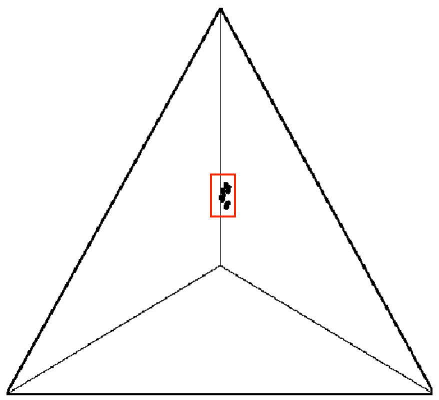
where $\tau_I = \int_I \sum_{a \in E} \nu_a(x) \tau_a \, dx$.

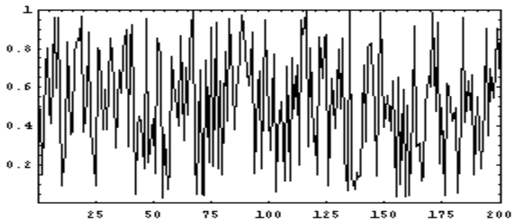
- When membership functions are fixed, basic learning algorithms transfer.

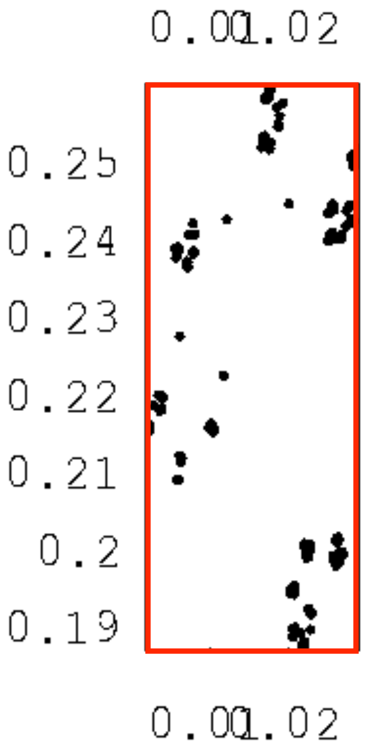
Blended OOMs, results 2

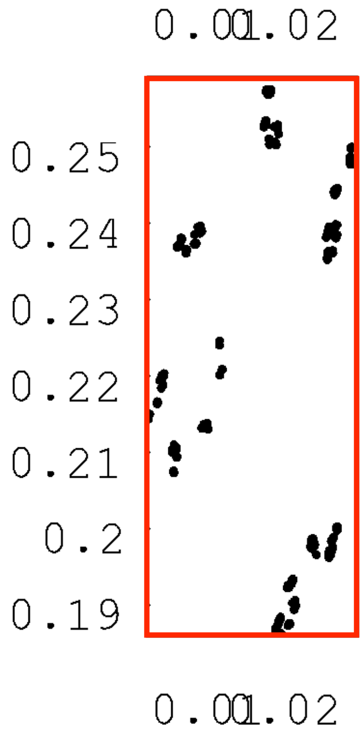
Learning algorithm adapted:

- Example: learning a continuous-valued version of the probability clock - an almost white-noise process









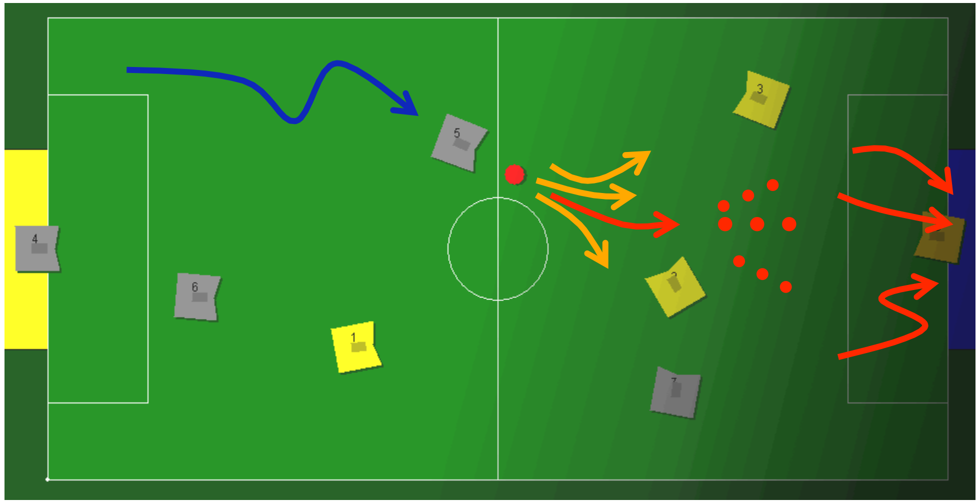
JACOBS UNIVERSITY

1. H. Jaeger(2001): **Modeling and learning continuous-valued stochastic processes with OOMs.** GMD Report 102, German National Research Center for Information Technology, 2001

11.3 Optimal decision making

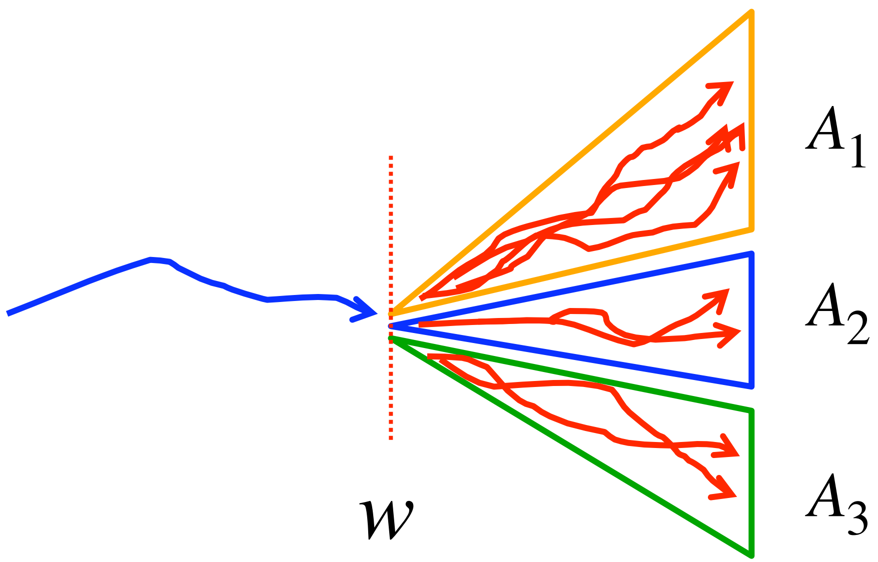
Setup

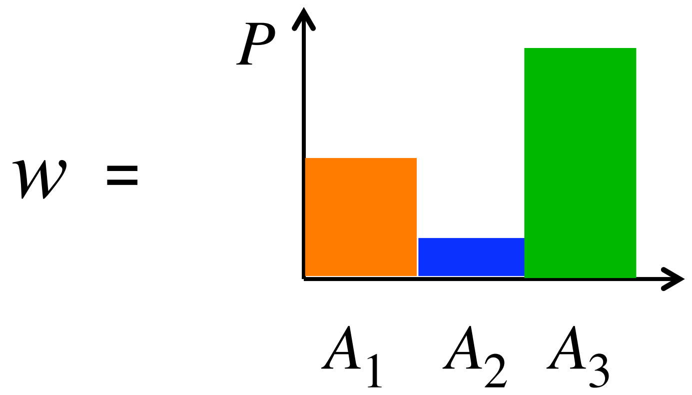
- reward delayed for uncertain time up to time horizon h
- stochasticity in sensing, acting, environment



Optimal decision making 2

A quick recap of interpretable OOMs:

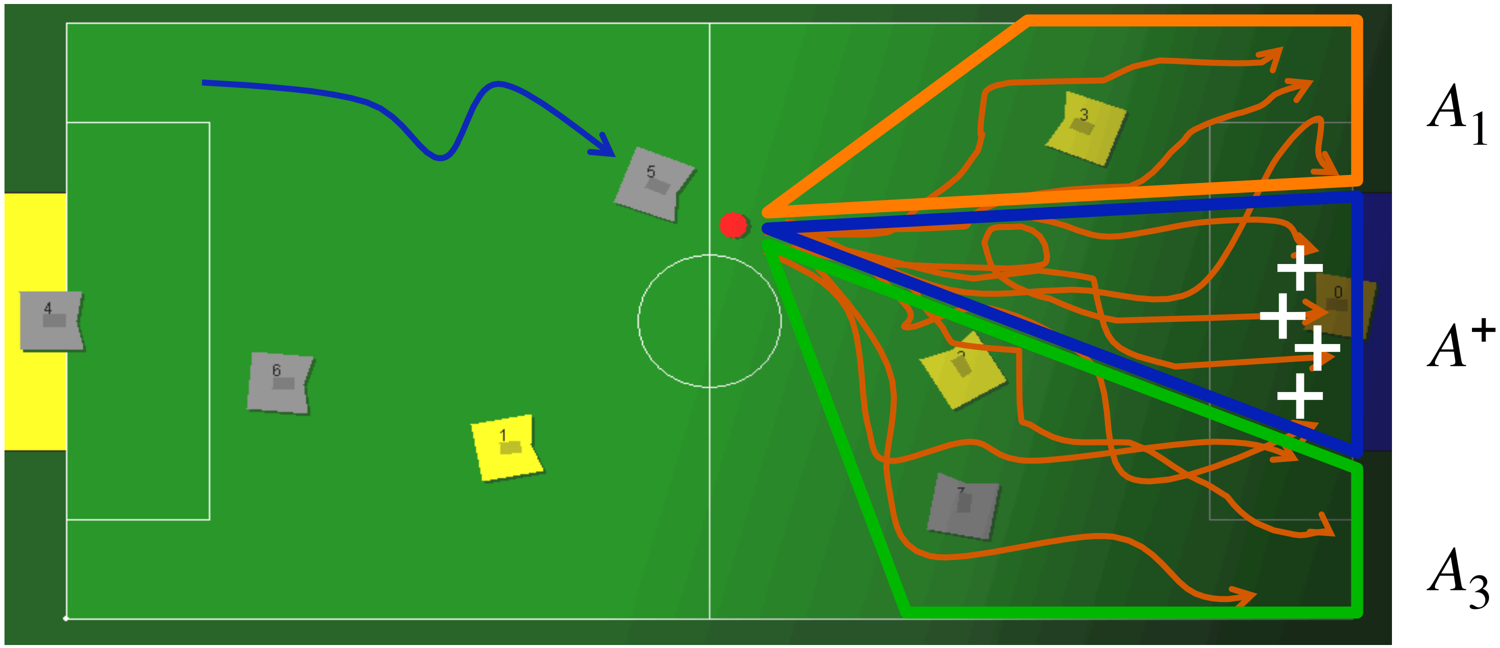




state components = probabilities of characteristic events

Optimal decision making 3

Approach [1]: Merge into characteristic event A^+ all futures which yield reward within h



Optimal decision making 4

- Assume agent has a "self-and-world-OOM" $\mathcal{S}$ of how it acts and how the world reacts.
- Let $U = \{u_1,...,u_k\}$ and $A = \{a_1,...,a_l\}$ be the actions and world (sensed) observables. Let $O = U \times A$. Then $\mathcal{S} = (\mathbb{R}^m, (\tau_{ua})_{ua \in O}, w_0)$.
- Put $\boldsymbol{\tau}_u = \sum_{a \in A} \boldsymbol{\tau}_{au}$.
- In non-deliberate mode, agent acts and updates OOM state w_n as follows:
  1. Choose action u_i according to probabilities $\mathbf{1}\ \tau_{u_i} w_n$.
  2. Execute chosen action u_i and observe world sensor feedback a_j.
  3. Update state $w_{n+1} = \tau_{u_i a_j} w_n / \mathbf{1}\tau_{u_i a_j} w_n$.
- If $\mathcal{S}$ models world feedback correctly, $P(a_j \mid w_n, u_i) = \mathbf{1}\boldsymbol{\tau}_{u_i a_j} w_n / \mathbf{1}\boldsymbol{\tau}_{u_i} w_n$.

Optimal decision making 5

- Recall from previous slide: $\mathcal{S} = (\mathbb{R}^m, (\tau_{ua})_{ua \in O}, w_0)$, $P(a_j \mid w_n, u_i) = \mathbf{1}\boldsymbol{\tau}_{u_i a_j} w_n / \mathbf{1}\boldsymbol{\tau}_{u_i} w_n$.
- Assume $\mathcal{S}$ is interpretable w.r.t. characteristic events A_i, where $A_1 = A^+$. Then the agent knows that the probability $P(+, h \mid w_n)$ to get a reward within horizon h, when the current state is w_n, is the first component $w_n[1]$ of w_n. This is subject to the condition that the agent continues operating in non-deliberative mode.
- The agent may want to do better than this, by switching to a deliberated action. That is, it would be advantageous to deliberately use action u at time n, if $P(+, h-1 \mid w_{n,}\ u) > P(+, h \mid w_n)$.
- $P(+, h-1 \mid w_{n,}\ u)$ can be computed cheaply:
    - Let $B^+ = \{u_1 a_1 \ldots u_{h\text{-}1} a_{h\text{-}1} \mid u_1 a_1 \ldots u_{h\text{-}1} a_{h\text{-}1} \textit{ contains a reward}\}$, and let $t_u = \mathbf{1}\boldsymbol{\tau}_{B^+}\boldsymbol{\tau}_u$.
    - Then, $P(+, h-1 | w_n, u) = t_u w_n$.

1. H. Jaeger (1999): **Action selection for delayed, stochastic reward.** Proc. 4th Annual Conf. of the German Cognitive Science Society (KogWis99), Infix Verlag, 213-219.

Thank you.