OOMs, PSRs, S-MAs and a statistically efficient learning algorithm

Michael Thon
Herbert Jaeger, Ming-Jie Zhao

Jacobs University Bremen

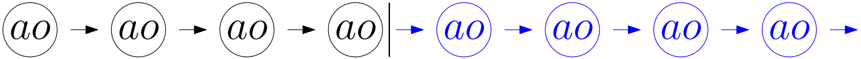
April 17, 2009

Outline

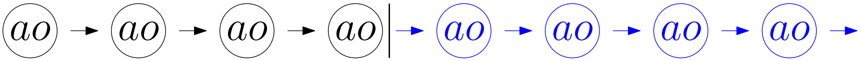
1. What are (IO)-OOMs?
   - Definitions and properties
2. Conversions
   - Conversions to and from PSRs
   - Conversions from HMMs/POMDPs
   - Relation to MAs and S-MAs
3. Learning OOMs from data
   - Learning equation
   - Error controlling idea and algorithm
   - Some initial results

JACOBS UNIVERSITY

Observable Operator Models

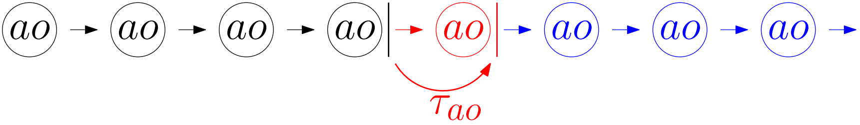


Observable Operator Models



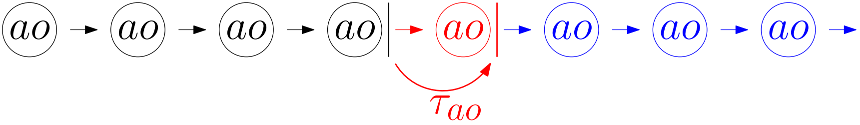
- State represented by *future prediction function* $f_{\bar{h}}(\bar{s}) = P(\bar{h}\bar{s})$
  $[F = \mathrm{span}\{f_{\bar{h}} : \bar{h} \in \Sigma^*\},\ w_{\bar{h}} \in F]$
  where $P(ao_1 \ldots ao_n) := \prod_{k=1}^{n} \Pr(o_k | ao_1 \ldots ao_{k-1} a_k)$

Observable Operator Models



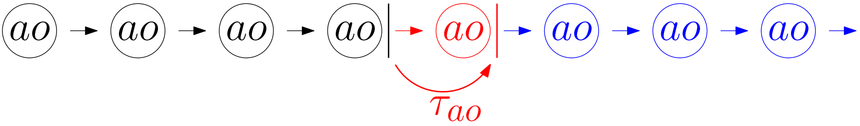
- State represented by *future prediction function* $f_{\bar{h}}(\bar{s}) = P(\bar{h}\bar{s})$
  $[F = \mathrm{span}\{f_{\bar{h}} : \bar{h} \in \Sigma^*\},\ w_{\bar{h}} \in F]$
  where $P(ao_1 \dots ao_n) := \prod_{k=1}^{n} \Pr(o_k | ao_1 \dots ao_{k-1} a_k)$
- State update by *observable operators* $\tau_{ao} : w_{\bar{h}} \mapsto w_{\bar{h}ao}$

Observable Operator Models



- State represented by *future prediction function* $f_{\bar{h}}(\bar{s}) = P(\bar{h}\bar{s})$
  $[F = \mathrm{span}\{\mathrm{f}_{\bar{\mathrm{h}}} : \bar{\mathrm{h}} \in \Sigma^*\},\ w_{\bar{h}} \in F]$
  where $P(ao_1 \ldots ao_n) := \prod_{k=1}^{n} \mathrm{Pr}(\mathrm{o_k}|\mathrm{ao_1} \ldots \mathrm{ao_{k-1}a_k})$
- State update by *observable operators* $\tau_{ao} : w_{\bar{h}} \mapsto w_{\bar{h}ao}$
- A *sigma functional* $\sigma^\top : w_{\bar{h}} \mapsto P(\bar{h})$ (Typically $\sigma^\top = (1, \ldots, 1)$).

Observable Operator Models



- State represented by *future prediction function* $f_{\bar{h}}(\bar{s}) = P(\bar{h}\bar{s})$
  $[F = \mathrm{span}\{f_{\bar{h}} : \bar{h} \in \Sigma^*\},\ w_{\bar{h}} \in F]$
  where $P(ao_1 \ldots ao_n) := \prod_{k=1}^{n} \Pr(o_k | ao_1 \ldots ao_{k-1} a_k)$
- State update by *observable operators* $\tau_{ao} : w_{\bar{h}} \mapsto w_{\bar{h}ao}$
- A *sigma functional* $\sigma^\top : w_{\bar{h}} \mapsto P(\bar{h})$ (Typically $\sigma^\top = (1, \ldots, 1)$).

$$P(ao_1 \ldots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

JACOBS UNIVERSITY

Observable Operator Models

A d-dimensional OOM $(\sigma, \{\tau_{ao}\}_{ao \in \Sigma}, w_0 \in \mathbb{R}^d)$ defines a function $f_{OOM} : \Sigma^* \to \mathbb{R}$ via

$$f_{OOM}(ao_1 \dots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

Observable Operator Models

A d-dimensional OOM $(\sigma, \{\tau_{ao}\}_{ao\in\Sigma}, w_0 \in \mathbb{R}^d)$ defines a function $f_{OOM} : \Sigma^* \longrightarrow \mathbb{R}$ via

$$f_{OOM}(ao_1 \ldots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

For f_{OOM} to represent a controlled process

1. $f(\varepsilon) = 1$
2. $\forall\, a, \bar{s} :\ f(\bar{s}) = \sum_o f(\bar{s}ao)$
3. $\forall\, \bar{s} :\ f(\bar{s}) \in [0, 1]$

Observable Operator Models

A d-dimensional OOM $(\sigma, \{\tau_{ao}\}_{ao \in \Sigma}, w_0 \in \mathbb{R}^d)$ defines a function $f_{OOM} : \Sigma^* \to \mathbb{R}$ via

$$f_{OOM}(ao_1 \dots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

For f_{OOM} to represent a controlled process

1. $f(\varepsilon) = 1$
2. $\forall\, a, \bar{s} :\ f(\bar{s}) = \sum_o f(\bar{s}ao)$
3. $\forall\, \bar{s} :\ f(\bar{s}) \in [0, 1]$

So a valid (IO)-OOM must satisfy

1. $\sigma^\top w_0 = 1$
2. $\forall\, a :\ \sigma^\top \sum_o \tau_{ao} = \sigma^\top$
3. $\forall \overline{ao} :\ \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0 \geq 0$

Observable Operator Models

A d-dimensional OOM $(\sigma, \{\tau_{ao}\}_{ao \in \Sigma}, w_0 \in \mathbb{R}^d)$ defines a process

$$P(ao_1 \dots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

Observable Operator Models

A d-dimensional OOM $(\sigma, \{\tau_{ao}\}_{ao \in \Sigma}, w_0 \in \mathbb{R}^d)$ defines a process

$$P(ao_1 \dots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

- Two OOMs are *equivalent* ($\equiv$) iif they define the same process
- $\boxed{\rho \in \mathrm{GL}(\mathrm{d}, \mathbb{R}) \Rightarrow (\sigma, \{\tau_{\mathrm{ao}}\}, \mathrm{w}_0) \equiv ((\rho^{-1})^\top \sigma, \{\rho \tau_{\mathrm{ao}} \rho^{-1}\}, \rho \mathrm{w}_0)}$

Observable Operator Models

A d-dimensional OOM $(\sigma, \{\tau_{ao}\}_{ao \in \Sigma}, w_0 \in \mathbb{R}^d)$ defines a process

$$P(ao_1 \ldots ao_n) := \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0$$

- Two OOMs are *equivalent* $(\equiv)$ iif they define the same process
- $\rho \in \mathrm{GL}(\mathrm{d}, \mathbb{R}) \Rightarrow (\sigma, \{\tau_{\mathrm{ao}}\}, \mathrm{w}_0) \equiv ((\rho^{-1})^\top \sigma, \{\rho \tau_{\mathrm{ao}} \rho^{-1}\}, \rho \mathrm{w}_0)$
- An OOM is *minimal* if no equivalent OOM with lower dimension exists [$\Leftrightarrow \{\tau_{\bar{s}} w_0 : \bar{s} \in \Sigma^*\} \cap \{\tau_{\bar{s}}^\top \sigma : \bar{s} \in \Sigma^*\} = \mathbb{R}^d$]
- One can convert an OOM to an equivalent minimal OOM
- Two minimal OOMs are equivalent iff they are conjugated
- Equivalence of OOMs can be efficiently decided

JACOBS UNIVERSITY

Conversions

Conversions

- *Normalize*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, \ldots, 1)$

Conversions

- *Normalize*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, \ldots, 1)$
- $\rightarrow$ *MA*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, 0, \ldots, 0)$

Conversions

- *Normalize*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, \dots, 1)$
- $\to$ *MA*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, 0, \dots, 0)$

Definition (Interpretable OOMs)

Let $A_1, \dots, A_d \subset \Sigma^*$. An OOM is *interpretable (w.r.t. these events)* if

$$w_{\bar{h}} = \begin{pmatrix} P(\bar{h}A_1) \\ \vdots \\ P(\bar{h}A_d) \end{pmatrix} \quad \forall\, \bar{h} \in \Sigma^*.$$

JACOBS UNIVERSITY

Conversions

- *Normalize*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, \ldots, 1)$
- $\rightarrow$ *MA*: Choose ρ s.t. $\sigma' = (\rho^{-1})^\top \sigma = (1, 0, \ldots, 0)$

Definition (Interpretable OOMs)

Let $A_1, \ldots, A_d \subset \Sigma^*$. An OOM is *interpretable (w.r.t. these events)* if

$$w_{\bar{h}} = \begin{pmatrix} P(\bar{h}A_1) \\ \vdots \\ P(\bar{h}A_d) \end{pmatrix} \quad \forall\, \bar{h} \in \Sigma^*.$$

- *Interpret*: Choose $\rho = \begin{pmatrix} \sigma^\top \tau_{A_1} \\ \vdots \\ \sigma^\top \tau_{A_d} \end{pmatrix}$. $\tau_A = \sum_{\bar{s} \in A} \tau_{s_n} \cdots \tau_{s_1}$

  Note: ρ must be non-singular!

  One may have to convert to a minimal OOM first.

JACOBS UNIVERSITY

Conversions

A PSR $(\{m_{ao}\}, \{M_{ao}\}, w_0)$ defines a process via

$$P(ao_1 \dots ao_n) = m_{ao_n}^{\top} M_{ao_{n-1}} \cdots M_{ao_1} w_0$$

- The states of a PSR are always interpretable w.r.t. *core tests*

JACOBS UNIVERSITY

Conversions

A PSR $(\{m_{ao}\}, \{M_{ao}\}, w_0)$ defines a process via

$$P(ao_1 \ldots ao_n) = m_{ao_n}^\top M_{ao_{n-1}} \cdots M_{ao_1} w_0$$

- The states of a PSR are always interpretable w.r.t. *core tests*
- $\rightarrow$ *PSR*: Interpret w.r.t. core tests. Set $m_{ao}^\top = \sigma^\top \tau_{ao}, M_{ao} = \tau_{ao}$.

Conversions

A PSR $(\{m_{ao}\}, \{M_{ao}\}, w_0)$ defines a process via

$$P(ao_1 \ldots ao_n) = m_{ao_n}^\top M_{ao_{n-1}} \cdots M_{ao_1} w_0$$

- The states of a PSR are always interpretable w.r.t. *core tests*
- $\rightarrow$ *PSR*: Interpret w.r.t. core tests. Set $m_{ao}^\top = \sigma^\top \tau_{ao}, M_{ao} = \tau_{ao}$.
- *(T)PSR* $\rightarrow$ *OOM*: Set $\sigma = \sum_{o \in O} m_{ao}, \tau_{ao} = M_{ao}$. (Normalize)

$$\left[\begin{array}{rcl} \sigma^\top \tau_{ao_n} \cdots \tau_{ao_1} w_0 & = & \sum_{o \in O} m_{ao}^\top M_{ao_n} \cdots M_{ao_1} w_0 \\ & = & \sum_{o \in O} P(ao_1 \ldots ao_n ao) \\ & = & P(ao_1 \ldots ao_n) \qquad \square \end{array}\right]$$

Conversions

A POMDP can be written as $(b \in \mathbb{R}^d, \{T_a \in \mathbb{R}^{d\times d}\}, \{O_{ao} \in \mathbb{R}^d\})$

Let $O'_{ao} = \mathrm{diag}(\mathrm{O_{ao}})$. Then the POMDP models a process via:

$$P(ao_1 \dots ao_n) = (1, \dots, 1)(O'_{ao_n} T_{a_n}) \cdots (O'_{ao_1} T_{a_1})b.$$

Conversions

A POMDP can be written as $(b \in \mathbb{R}^d, \{T_a \in \mathbb{R}^{d\times d}\}, \{O_{ao} \in \mathbb{R}^d\})$

Let $O'_{ao} = \mathrm{diag(O_{ao})}$. Then the POMDP models a process via:

$$P(ao_1 \dots ao_n) = (1, \dots, 1)(O'_{ao_n} T_{a_n}) \cdots (O'_{ao_1} T_{a_1}) b.$$

- *POMDP → OOM*: Set $w_0 = b, \tau_{ao} = (O'_{ao} T_a), \sigma^\top = (1, \dots, 1)$.

Conversions

A POMDP can be written as $(b \in \mathbb{R}^d, \{T_a \in \mathbb{R}^{d\times d}\}, \{O_{ao} \in \mathbb{R}^d\})$

Let $O'_{ao} = \mathrm{diag}(\mathrm{O_{ao}})$. Then the POMDP models a process via:

$$P(ao_1 \dots ao_n) = (1, \dots, 1)(O'_{ao_n} T_{a_n}) \cdots (O'_{ao_1} T_{a_1})b.$$

- *POMDP → OOM*: Set $w_0 = b, \tau_{ao} = (O'_{ao} T_a), \sigma^\top = (1, \dots, 1)$.
- *OOM → POMDP*: map onto "closest" POMDP, ?!?

JACOBS UNIVERSITY

Multiplicity automata

Definition

A *multiplicity automata* (MA) is a structure $(\sigma, \{\tau_s\}_{s\in\Sigma}, w_0 \in \mathbb{R}^d)$ with $\sigma^\top = (1, 0, \ldots, 0)$ that specifies a function via

$$f_{MA}(\bar{s}) = (1, 0, \ldots, 0)\tau_{s_n} \cdots \tau_{s_1} w_0$$

- Such a function is then called a *rational language*
- If f_{MA} is a probability distribution function on $\Sigma^* \rightarrow$ *stochstic MA* $[f_{MA} \geq 0$ and $\sum_{\bar{s}\in\Sigma^*} f_{MA}(\bar{s}) = 1]$
- f_{SMA} is called a rational *stochastic* language
- Is any given MA a SMA? $\rightarrow$ undecidable!

JACOBS UNIVERSITY

Multiplicity automata

Definition

A *multiplicity automata* (MA) is a structure $(\sigma, \{\tau_s\}_{s\in\Sigma}, w_0 \in \mathbb{R}^d)$ with $\sigma^\top = (1, 0, \dots, 0)$ that specifies a function via

$$f_{MA}(\bar{s}) = (1, 0, \dots, 0)\tau_{s_n} \cdots \tau_{s_1} w_0$$

- Such a function is then called a *rational language*
- If f_{MA} is a probability distribution function on $\Sigma^* \rightarrow$ *stochstic MA* $[f_{MA} \geq 0$ and $\sum_{\bar{s}\in\Sigma^*} f_{MA}(\bar{s}) = 1]$
- f_{SMA} is called a rational *stochastic* language
- Is any given MA a SMA? $\rightarrow$ undecidable!
- *SMA* $\rightarrow$ *OOM*: (Minimize) and set $(\sigma')^\top = [\sigma^\top \sum_{k=0}^{\infty}(\sum_{s\in\Sigma} \tau_s)^k \ , \ 1] = [\sigma^\top(I - \sum_{s\in\Sigma} \tau_s)^{-1} \ , \ 1]$, $\tau'_s = \begin{bmatrix} \tau_s & 0 \\ \frac{1}{|\Sigma|}\sigma & \frac{1}{|\Sigma|} \end{bmatrix}$ and $w'_0 = \begin{bmatrix} w_0 \\ 0 \end{bmatrix}$.

OOM learning

For a given sequence $\bar{s}$ (which we assume is generated by an OOM):

OOM learning

For a given sequence $\bar{s}$ (which we assume is generated by an OOM):

1. Choose indicative/characteristic sequences $\{\bar{q}_j\}$ and $\{\bar{c}_i\}$ (e.g. Σ^l)
2. Let $\hat{V} = [\hat{P}(\bar{q}_j, \bar{c}_i)]_{i,j}$ ("*big V*") and $\hat{W}_{ao} = [\hat{P}(\bar{q}_j(ao)\bar{c}_i)]_{i,j}$,
   $\hat{c} = [\hat{P}(\bar{c}_i)]_i$, $\hat{q}^\top = [\hat{P}(\bar{q}_j)]_j$
   $$[\hat{P}(ao_1 \dots ao_n) = \prod_{k=1}^{n} \frac{\#ao_1 \dots ao_k}{\#ao_1 \dots ao_{k-1}a_k}]$$
3. Estimate the dimension d of the OOM as the numerical rank of $\hat{V}$
4. Scale the columns of $\hat{V}$ and $\hat{W}_{ao}$ by $\sqrt{\#\bar{q}_j}$
5. Choose $C, Q^\top \in \mathbb{R}^{d \times |\hat{V}|}$ such that $C\hat{V}Q$ is invertible.
6. Set
   - $\boxed{\hat{\tau}_{ao} = (C\hat{W}_{ao}Q)(C\hat{V}Q)^{-1}}$ or solve $\boxed{\hat{\tau}_{ao}C\hat{V} = C\hat{W}_{ao}}$ $(Q = (C\hat{V})^\dagger)$
   - $\hat{w}_0 = C\hat{c}$
   - $\hat{\sigma}^\top = \hat{q}^\top Q(CVQ)^{-1}$

Choosing good C, Q by the *error controlling* principle

Bound on the error of the learnt observable operators $\hat{\tau}_{ao}$

$$\frac{||\tau - \hat{\tau}||}{||\tau||} < k \cdot ||C|| \cdot ||Q(C\hat{V}Q)^{-1}||$$

- this holds for normalized OOMs where $\sigma^\top = (1, \ldots, 1)$.
- we may set $Q' = Q(C\hat{V}Q)^{-1}$. Then $C\hat{V}Q' = I$.

JACOBS UNIVERSITY

Choosing good C, Q by the *error controlling* principle

Bound on the error of the learnt observable operators $\hat{\tau}_{ao}$

$$\frac{||\tau - \hat{\tau}||}{||\tau||} < k \cdot ||C|| \cdot ||Q(C\hat{V}Q)^{-1}||$$

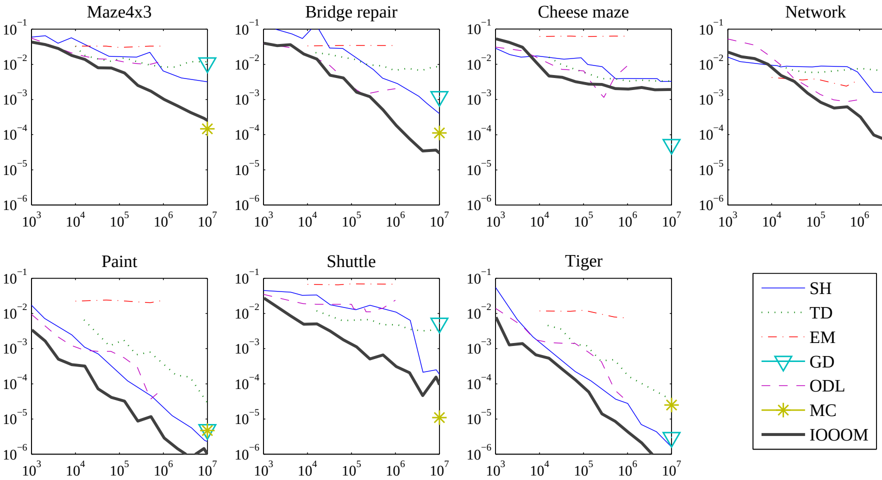
- this holds for normalized OOMs where $\sigma^\top = (1, \ldots, 1)$.
- we may set $Q' = Q(C\hat{V}Q)^{-1}$. Then $C\hat{V}Q' = I$.

Optimization problem

$$(C, Q) = \underset{(C,Q)}{\operatorname{argmin}}\{||C|| \cdot ||Q|| : C\hat{V}Q = I, \hat{q}^\top Q = (1, \ldots, 1)\}$$

JACOBS UNIVERSITY

Results on simple POMDP problems



Comments

- Will use a suffix-tree representation of the data $\bar{s}$ to pre-select characteristic and indicative sequences based on count numbers, and avoiding redundant information.
- This in principle allows for much longer characteristic/indicative sequences
- $\hat{P}(\overline{ao}) = \prod_{k=1}^{n} \hat{\mathrm{Pr}}(o_k | ao_1 \ldots ao_{k-1} a_k) = \prod_{k=1}^{n} \frac{\# ao_1 \ldots ao_k}{\# ao_1 \ldots a_k O}$
  What if a conditioning action never occurs?

Comments

- Will use a suffix-tree representation of the data $\bar{s}$ to pre-select characteristic and indicative sequences based on count numbers, and avoiding redundant information.
- This in principle allows for much longer characteristic/indicative sequences
- $\hat{P}(\overline{ao}) = \prod_{k=1}^{n} \hat{\Pr}(o_k | ao_1 \dots ao_{k-1} a_k) = \prod_{k=1}^{n} \frac{\# ao_1 \dots ao_k}{\# ao_1 \dots a_k O}$
  What if a conditioning action never occurs?

Thanks you all for the great week in Barbados!

JACOBS UNIVERSITY