# Global Estimation of Hidden Markov Models using Interval Arithmetic


**Tiago M. Montanher**

**Walter F. Mascarenhas**





Conselho Nacional de Desenvolvimento Científico e Tecnológico

## Outline

- The Problem
  - Statement
  - Objectives

- Interval arithmetic algorithm
  - Interval branch and bound
  - Evaluating the objective function
  - Discarding boxes

- Numerical experiments
  - Plotting solution sets
  - Basins of attraction

## What is a Hidden Markov Model?

### Dilma’s voting intention

| | High | Low | Initial State |
|---|---|---|---|
| High | 70% | 30% | 60% |
| Low | 50% | 50% | 40% |

### Petrobras Stock value PETR4($)

| | | |
|---|---|---|
| Increase | 30% | 60% |
| Unchanged | 50% | 30% |
| Decrease | 20% | 10% |

We observe only the Market performance
(Increase, Decrease, Unchanged, Decrease....)

- Let $q_t$ be a discrete time Markov chain assuming states $S_1, \ldots, S_N$.
- Let $O_t$ be a discrete time stochastic process assuming states $V_1, \ldots, V_M$ and satisfying

$$\mathbb{P}(O_t = V_i | q_t = S_j) = b_{ij}$$

- A Hidden Markov Model is a triple $\theta = (A, B, \pi)$ where

$$\sum_{j=1}^{N} a_{ij} = 1, \quad \forall i = 1, \ldots, N. \qquad a_{ij} \geq 0$$

$$\sum_{i=1}^{M} b_{ij} = 1, \quad \forall j = 1, \ldots, N. \qquad b_{ij} \geq 0$$

$$\sum_{i=1}^{N} \pi_i = 1. \qquad \pi_i \geq 0$$

## Problem Statement



We are given -

1. A set of observations $O_1, \ldots, O_T$.
2. The number of states $N$ for the hidden process.
3. The number of states $M$ for the observable process

$$
\begin{aligned}
\max\ z &= \mathbb{P}(\theta|O) \\
e_N^T A_i &= 1 \qquad \forall i = 1, \ldots, N. \\
e_M^T B^i &= 1 \qquad \forall i = 1, \ldots, N. \\
e_N^T \pi &= 1 \\
a_{ij}, b_{ij}, \pi_i &\geq 0.
\end{aligned}
$$

← Next slide

## Computing Probabilities



- We evaluate probabilities through the formula

$$\mathbb{P}(O|\theta) = \pi^T d(B_{O_1}) A d(B_{O_2}) \ldots A d(B_{O_T}) \mathbf{1}$$

where $d(B_{O_t})$ is the diagonal matrix given by line $O_t$.

- Function $\mathbb{P}(O|\theta)$ is non-convex.
- Since all parameters are non-negative, the objective function and their derivatives are non-negative.
- In order to make the evaluation easier we employ the backward recursion

$$\begin{aligned} \beta_T &= \mathbf{1} \\ \beta_t &= A d(B_{O_{t+1}}) \beta_{t+1} \qquad t = T-1, \ldots, 1 \\ \mathbb{P}(\theta|O) &= \pi^T d(B_{O_1}) \beta_1 \end{aligned}$$

## State of the Art

- Baum-Welch algorithm is a local method widely used to maximize $\mathbb{P}(\theta|O)$.
- It is a good option if the number of observations is large.

## Our Aims

- To develop a global optimization algorithm able to find all isolated global maxima of $\mathbb{P}(\theta|O)$.
- To study the convergence properties of Baum Welch-algorithm.

## Baum-Welch Algorithm



**Input –** An observation set O and an initial estimate $\theta^0$.

**Output –** A (local) approximation for the maximum likelihood estimator $\hat{\theta}$.

$$a_{ij}^{t+1} = \frac{a_{ij}^t \frac{\delta \mathbb{P}(\theta|O)}{\delta_{a_{ij}}}(\theta^t)}{\sum_{k=1}^{N} a_{ik}^t \frac{\delta \mathbb{P}(\theta|O)}{\delta_{a_{ik}}}(\theta^t)}$$

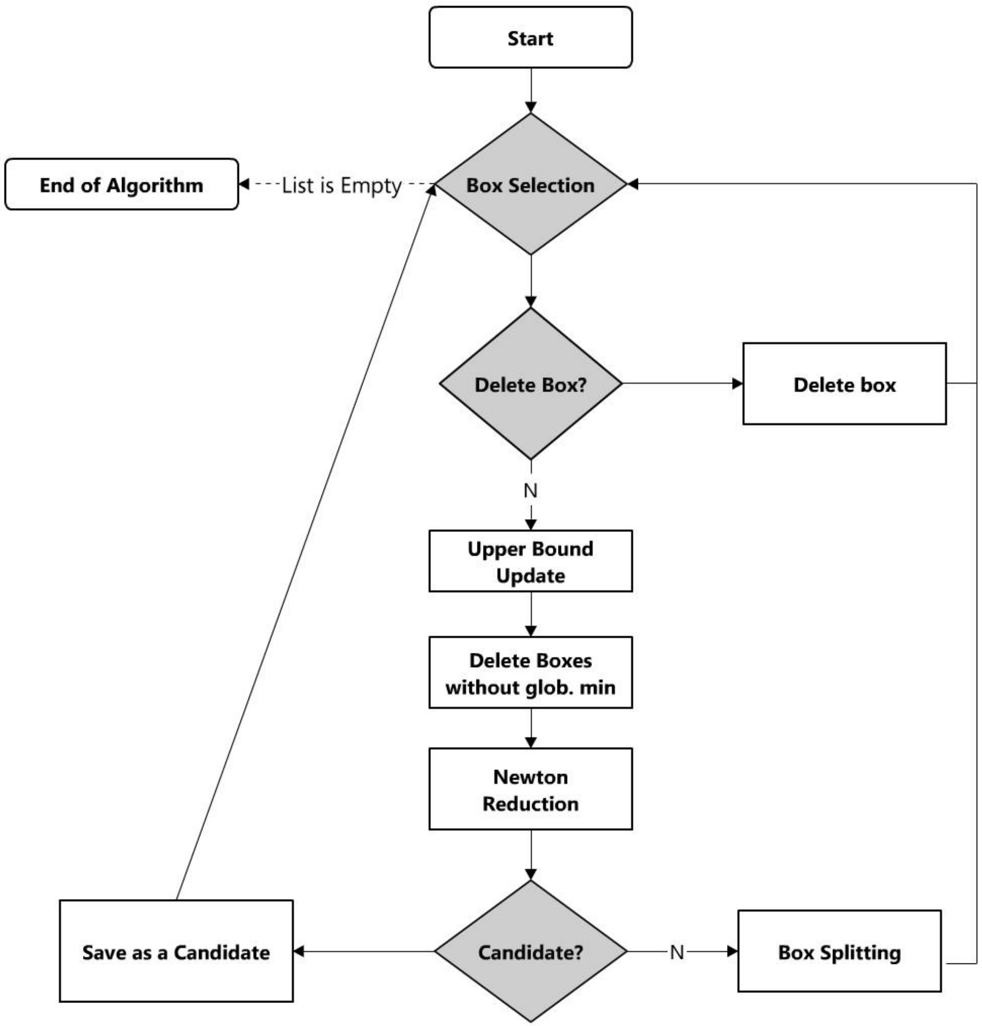
$$b_{ij}^{t+1} = \frac{b_{ij}^t \frac{\delta \mathbb{P}(\theta|O)}{\delta_{b_{ij}}}(\theta^t)}{\sum_{k=1}^{M} b_{kj}^t \frac{\delta \mathbb{P}(\theta|O)}{\delta_{b_{kj}}}(\theta^t)}$$

$$\pi_{i}^{t+1} = \frac{\pi_{i}^t \frac{\delta \mathbb{P}(\theta|O)}{\delta_{\pi_{i}}}(\theta^t)}{\sum_{k=1}^{N} \pi_{k}^t \frac{\delta \mathbb{P}(\theta|O)}{\delta_{\pi_{k}}}(\theta^t)}$$

## Interval Branch and Bound

**Input –** An interval extension $\mathbb{P}(\boldsymbol{\theta}|O)$ of $\mathbb{P}(\theta|O)$.

**Output –** A list of boxes which contains all global maximum likelihood estimator



## Computing probabilities – Interval Arithmetic



- If we just replace the floating point operations by interval operations we have a natural extension

$$\mathbb{P}(\boldsymbol{\theta}|O) = \boldsymbol{\pi}^{\top} \boldsymbol{d}(\boldsymbol{B}_{\mathsf{O}_1})\boldsymbol{\beta}_{\mathbf{1}}$$

- For example, if $N = 2$ and $M = 3$ HMM with $O = \{1, 3\}$ and a box $\boldsymbol{x} = [0.25, 0.75]^{12}$ then

$$\mathbb{P}(\boldsymbol{\theta}|O) = [0.015625, 1.26562]$$

- We propose an extension based on the simplices structure of the problem. Let $\boldsymbol{c} = \boldsymbol{d}(\boldsymbol{\beta}_{\boldsymbol{O}_{t+1}})\boldsymbol{\beta}_{\boldsymbol{t+1}}$ then

$$\beta_t(i) = \begin{bmatrix} \min \underline{\boldsymbol{c}}^T A_i & \max \overline{\boldsymbol{c}}^T A_i \\ e^T A_i = 1 \;, & e^T A_i = 1 \\ a_{ij} \in \boldsymbol{a_{ij}} & a_{ij} \in \boldsymbol{a_{ij}} \end{bmatrix}.$$

$$\mathbb{P}(\boldsymbol{\theta}|O) = [0.0625, 0.5625]$$

## Computing probabilities – Interval Arithmetic

$$([\underline{a_1}, \overline{a_1}], \ldots, [\underline{a_N}, \overline{a_N}]) \begin{bmatrix} [\underline{b_1}, \overline{b_1}] \\ \vdots \\ [\underline{b_N}, \overline{b_N}] \end{bmatrix} \rightarrow$$

$4 * N + 3$ rounding mode switching per backward iteration

Since we have only non-negative coefficients

$$\begin{pmatrix} \underline{a_1} \\ \underline{a_N} \\ \vdots \\ \overline{a_1} \\ \overline{a_N} \end{pmatrix} \begin{pmatrix} \underline{a_1} \\ \underline{a_N} \\ \vdots \\ \overline{a_1} \\ \overline{a_N} \end{pmatrix} \rightarrow$$

3 rounding mode switching to evaluate the objective function and their derivatives

## KKT Conditions

- If $\hat{\theta}$ is a solution then there exists $(\hat{\lambda}, \hat{\mu})$ such that

$$
\begin{aligned}
\frac{\delta \mathbb{P}(\theta|O)}{\delta a_{ij}}(\hat{\theta}) - \hat{\lambda}_{A_i} - \hat{\mu}_{a_{ij}} &= 0 \qquad \forall i, j = 1, \ldots, N. \\
\frac{\delta \mathbb{P}(\theta|O)}{\delta b_{ij}}(\hat{\theta}) - \hat{\lambda}_{B^i} - \hat{\mu}_{b_{ij}} &= 0 \qquad \forall i = 1, \ldots, N \text{ and } j = 1, \ldots, M. \\
\frac{\delta \mathbb{P}(\theta|O)}{\delta \pi_i}(\hat{\theta}) - \hat{\lambda}_{\pi} - \hat{\mu}_{\pi_i} &= 0 \qquad \forall i, j = 1, \ldots, N. \\
e^T \hat{\theta}_{A_i} &= 1 \qquad \forall i = 1, \ldots, N. \\
e^T \hat{\theta}_{B^i} &= 1 \qquad \forall i = 1, \ldots, N. \\
e^T \hat{\theta}_{\pi} &= 1 \\
\mu_{a_{ij}} \hat{\theta}_{a_{ij}} &= 0 \qquad \forall i, j = 1, \ldots, N. \\
\mu_{b_{ij}} \hat{\theta}_{b_{ij}} &= 0 \qquad \forall i = 1, \ldots, N \text{ and } j = 1, \ldots, M. \\
\mu_{\pi_i} \hat{\theta}_{\pi_i} &= 0 \qquad \forall i, j = 1, \ldots, N. \\
\mu_{a_{ij}}, \mu_{b_{ij}}, \mu_{\pi_i} &\geq 0
\end{aligned}
$$

## Discarding Boxes



- From KKT conditions, we have

$$\begin{aligned} \frac{\delta \mathbb{P}(\theta|O)}{\delta a_{ij}}(\hat{\theta}) - \hat{\lambda}_{A_i} - \hat{\mu}_{a_{ij}} &= 0 \\ \frac{\delta \mathbb{P}(\theta|O)}{\delta a_{ik}}(\hat{\theta}) - \hat{\lambda}_{A_i} - \hat{\mu}_{a_{ik}} &= 0 \end{aligned} \quad \Rightarrow$$

$$\frac{\delta \mathbb{P}(\theta|O)}{\delta a_{ij}}(\hat{\theta}) - \frac{\delta \mathbb{P}(\theta|O)}{\delta a_{ik}}(\hat{\theta}) - \hat{\mu}_{a_{ij}} + \hat{\mu}_{a_{ik}} = 0.$$

- Let $\boldsymbol{g} = \frac{\delta f}{\delta a_{ij}}(\boldsymbol{\theta}) - \frac{\delta f}{\delta a_{ik}}(\boldsymbol{\theta})$, by complementary conditions:

1. $0 < \underline{\boldsymbol{g}} \Longrightarrow \hat{\mu}_{a_{ij}} > 0$ and $\theta_{a_{ij}} = 0$
2. $0 > \overline{\boldsymbol{g}} \Longrightarrow \hat{\mu}_{a_{ik}} > 0$ e $\theta_{a_{ik}} = 0.$

## Discarding Boxes



- We can re-label the observations in order that the most frequent is labeled by 1 the second most frequent 2, . . .
- After re-labeling we have that

$$\begin{aligned} b_{11} &\geq b_{1j} \qquad j = 2, \ldots, N. \\ b_{11} &\geq \frac{1}{N}. \end{aligned}$$

And, in general

$$b_{ii} \geq b_{ij} \qquad j = i + 1, \ldots, N.$$

- We discard any box that do not satisfy the constraints above.
- We also apply constraint propagation using these inequalities in order to reduce boxes.

## Implementation details

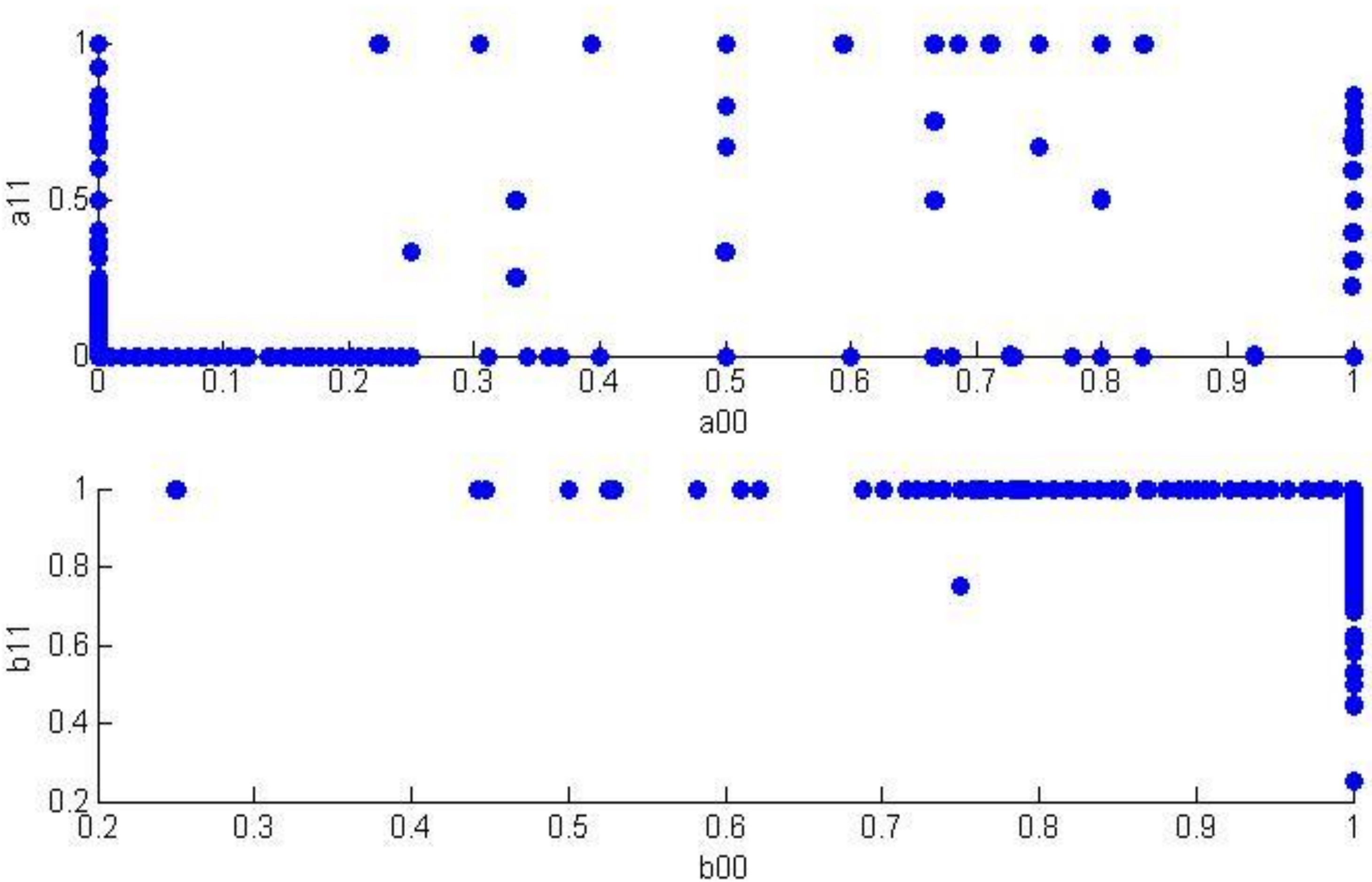
- Algorithm in C++11.
- We developed our own interval arithmetic library based on cfenv.h library.
- Avoid round mode switching. Only three per function or gradient evaluation.
- To avoid underflow/overflow we use the functions frexp and scalbln at each vector operation.
- Interval analysis methods -

  Midpoint test,

  Constraints propagation(linear and Taylor based),

  Taylor evaluation of objective function.

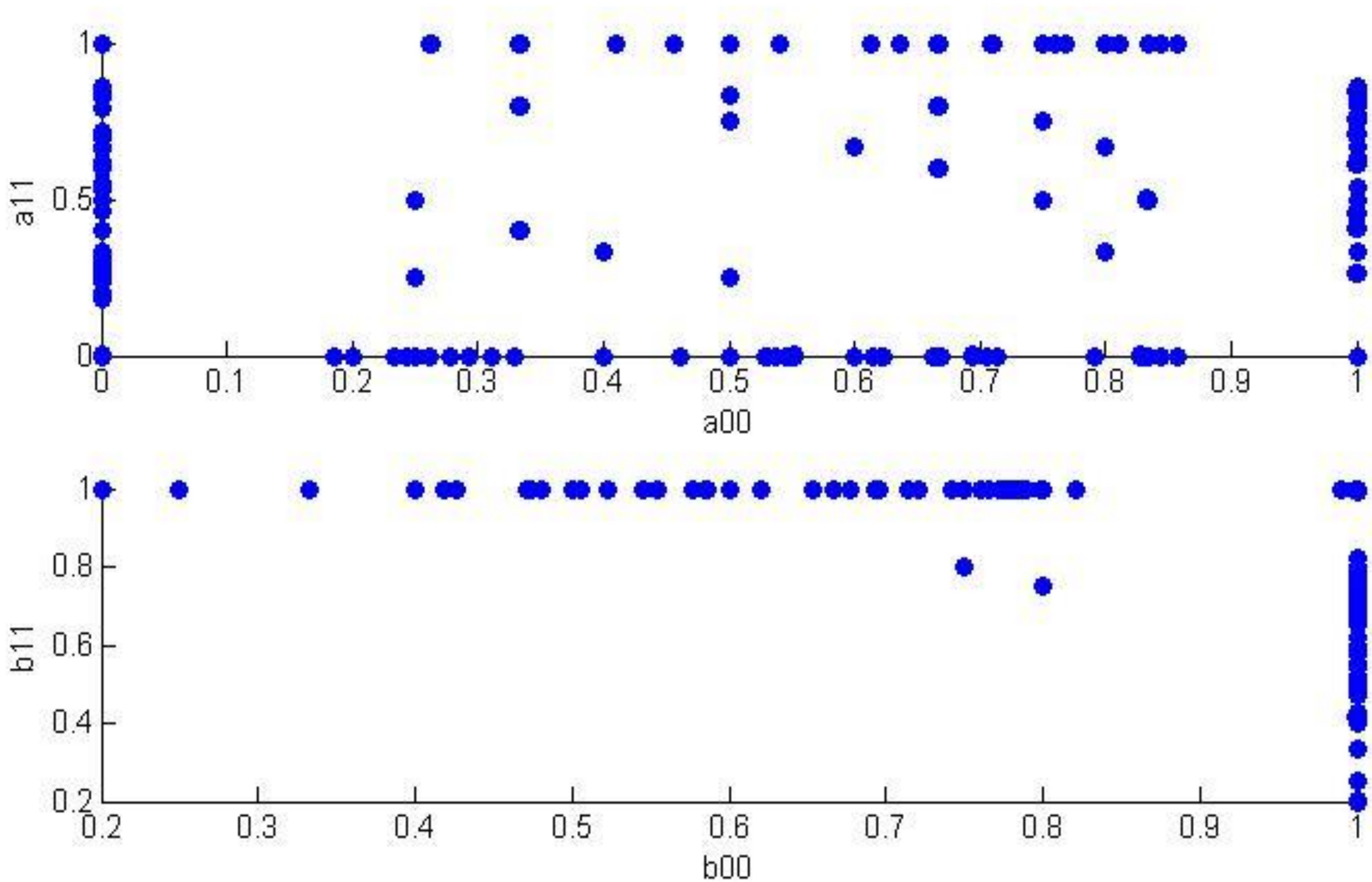
## Numerical Experiments



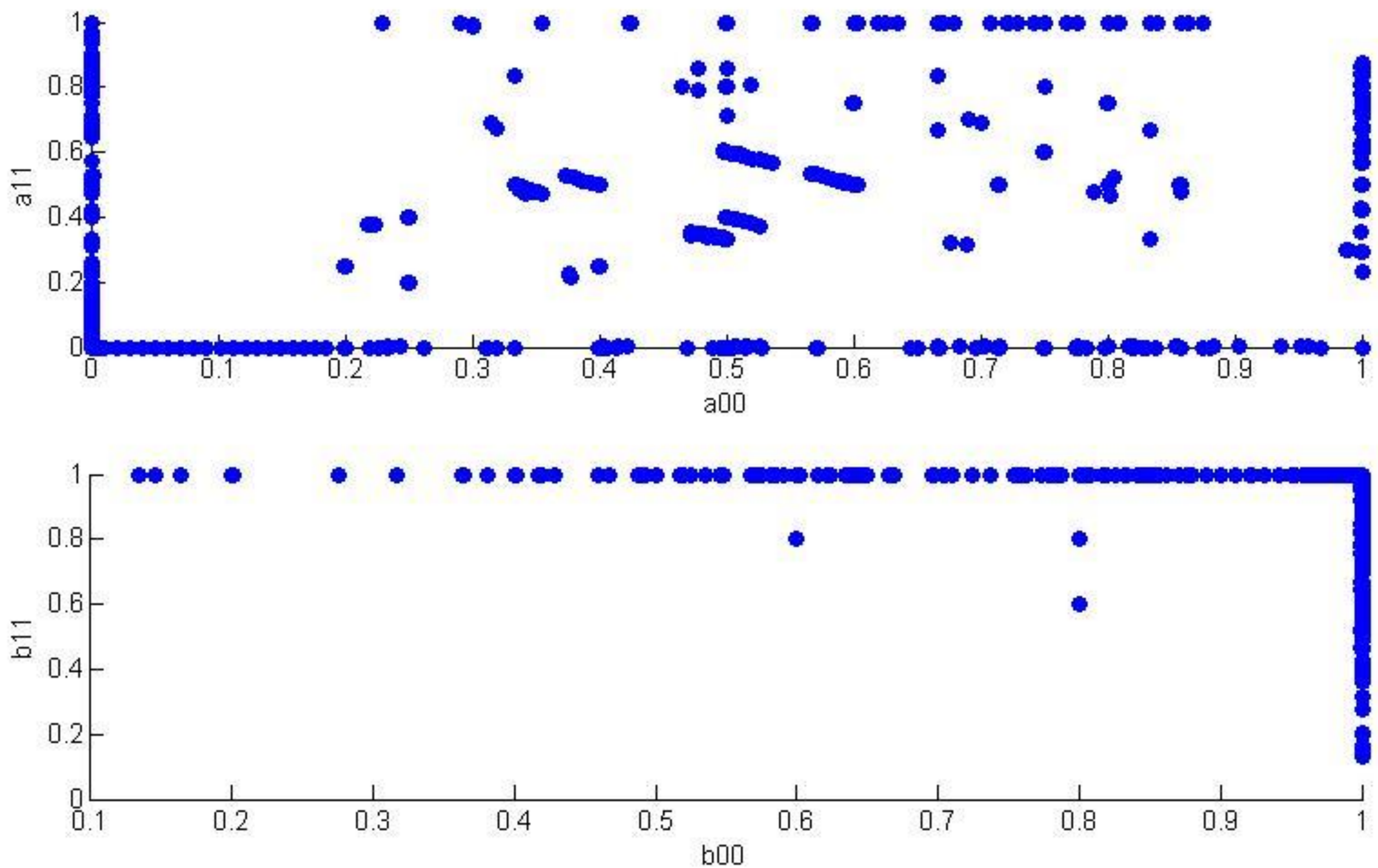

Solution set for all instances with N = 2, M = 2 and T = 8 (254 instances)

## Numerical Experiments





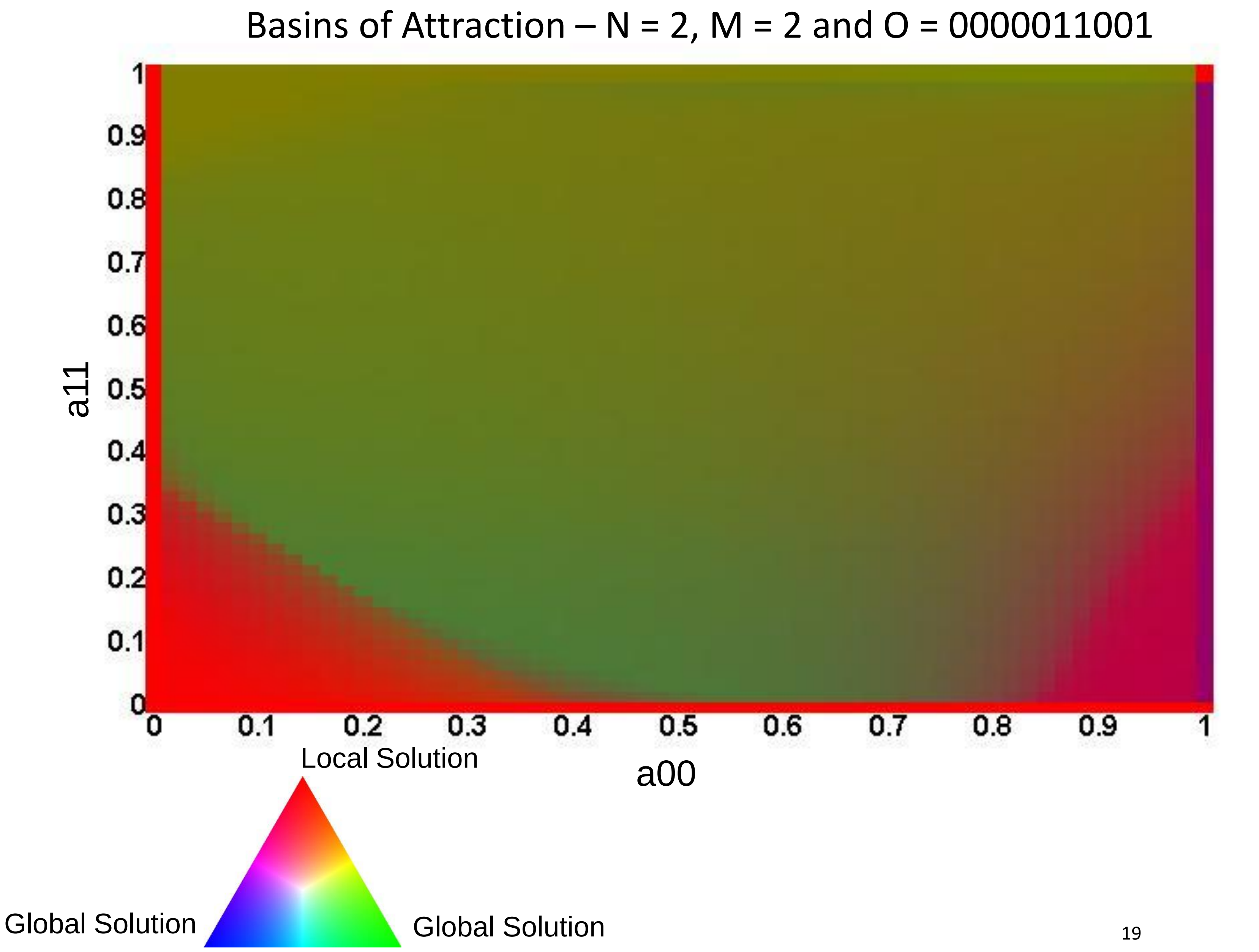

Solution set for all instances with N = 2, M = 2 and T = 9 (510 instances)




Solution set for all instances with N = 2, M = 2 and T = 10 (1022 instances)

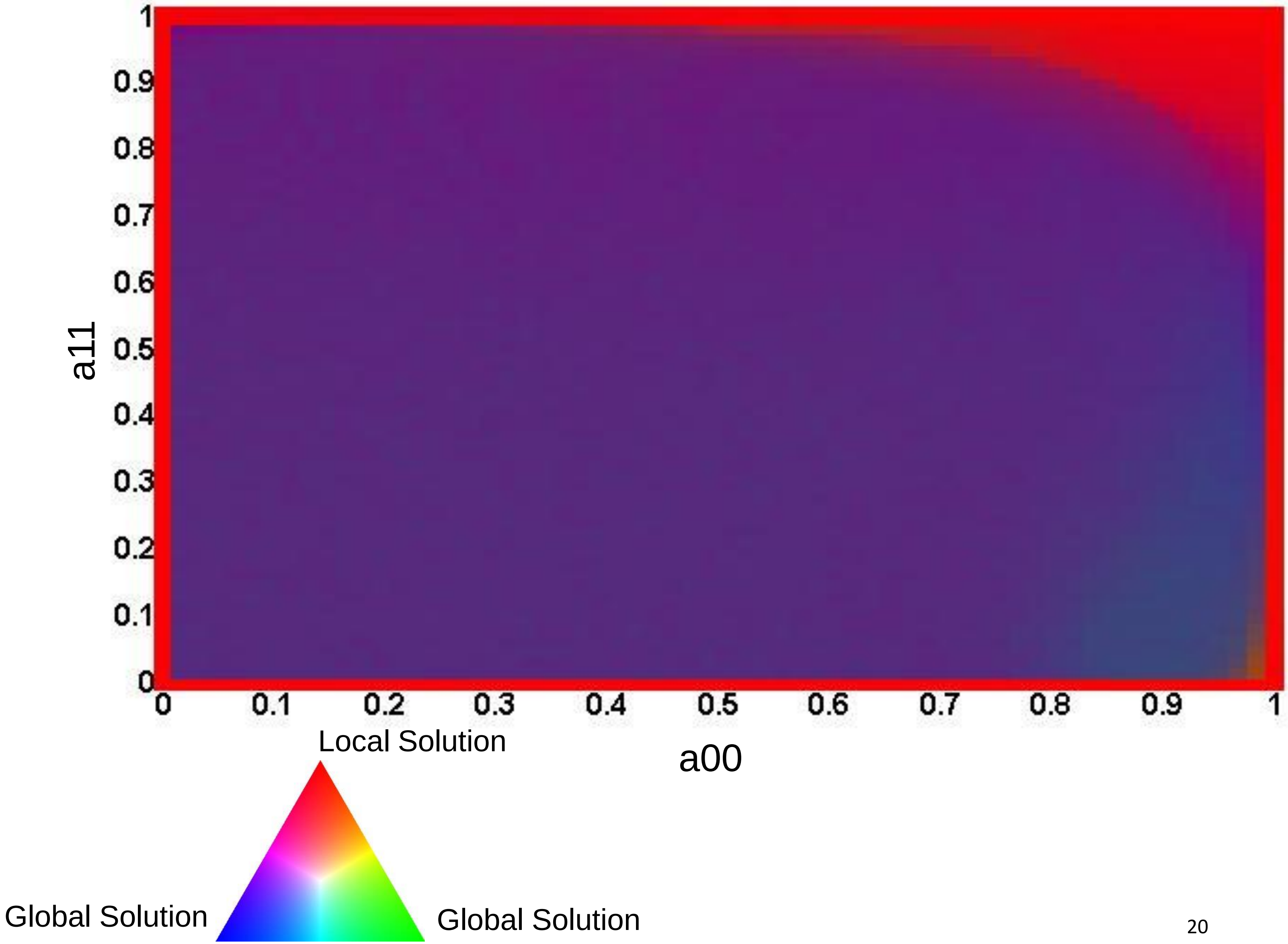
## Basins of Attraction





## Basins of Attraction

Basins of Attraction – N = 2, M = 2 and O = 1001000100001001000101111



## Final Remarks

- We provide an alternative to estimate parameters of Hidden Markov Models.
- We derive a new interval extension and discarding tests based on the structure of the problem.
- Only few models can be correctly predicted when we have a small set of observations.
- The Baum Welch algorithm does not find a global maximum likelihood estimator for more than 50% of initial points.

## Questions?