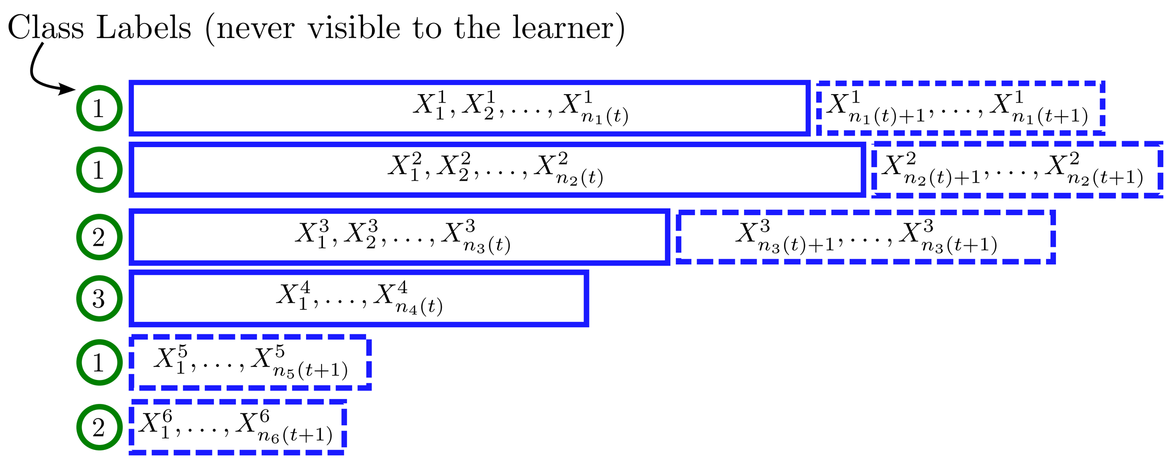
ONLINE CLUSTERING OF PROCESSES

AZADEH KHALEGHI, DANIIL RYABKO, JÉRÉMIE MARY, PHILIPPE PREUX

PROBLEM

Setup: We have a growing body of sequences of data. Each sequence is generated by on of k unknown discrete-time stochastic process. The number k of distributions is known.

Data are observed in an online fashion: → New samples arrive at every time-step; they either are continuations of previously received sequences or a new sequences.



Goal: Cluster the sequences at every time-step.

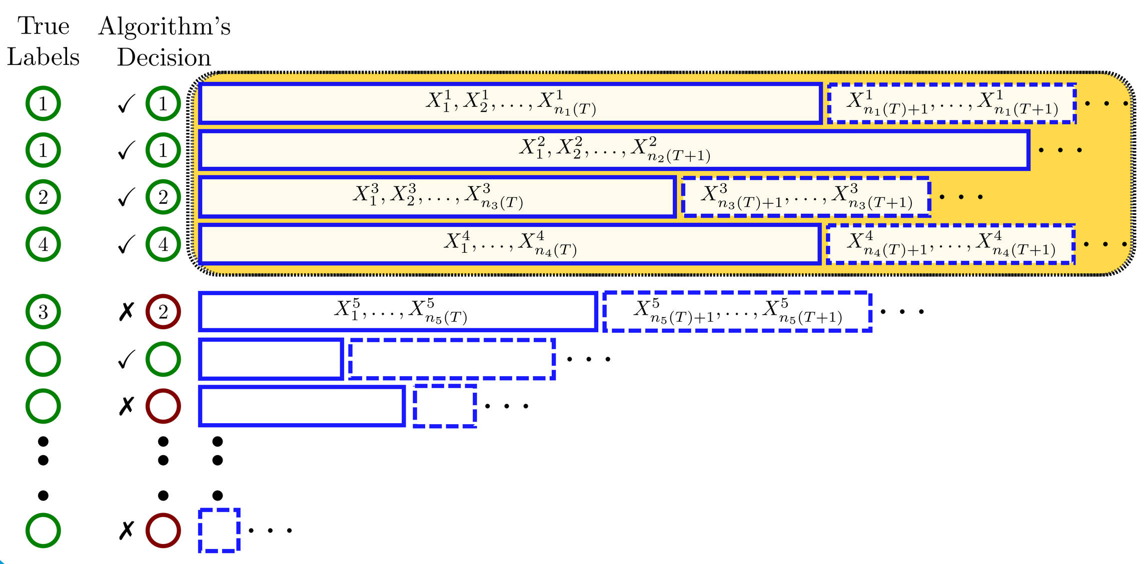
CONSISTENCY

In general it is hard to give a precise definition for "correct clustering".

But, a natural notion for correct clustering exists in the considered setting:

Sequences generated by the same process distribution should be grouped together.

Asymptotic Consistency: *A clustering algorithm is (asymptotically) consistent if, with probability 1, for each $N \in \mathbb{N}$ from some time on, it clusters the first N observed sequences are clustered correctly.*



ASSUMPTIONS ON DATA

- Data revealed in an arbitrary fashion.
- Our only assumption is that the distributions generating the data are stationary-ergodic.

→ The samples are allowed to be dependent and the dependence can be arbitrary, or even adversarial. No such assumptions as iid, Markov etc.

Remark: In time-series literature, it is typically assumed that the distributions generating the data have a known form, ex. Gaussian, HMMs etc., and the samples are independent.

DISTANCE MEASURE

We measure the distance between two sequences $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{n_2}$ as

$$\hat{d}(\mathbf{x}_1, \mathbf{x}_2) := \sum_{m,l=1}^{\infty} 2^{-(m+l)} \sum_{B \in B^{m,l}} |\nu(\mathbf{x}_1, B) - \nu(\mathbf{x}_2, B)|$$

where $B^{m,l}$ $m, l \in \mathbb{N}$ is the set of all hypercubes of dimension m and edge-length 2^{-l} and $\nu(\mathbf{x}, B)$ is the frequency with which $\mathbf{x}$ crosses B.

Theorem: ($\hat{d}(\cdot,\cdot)$ is consistent) [1]

If $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by stationary-ergodic processes ρ_1 and ρ_2, then $\hat{d}(\mathbf{x}_1, \mathbf{x}_2)$ converges to the so-called distributional-distance:

$$d(\rho_1, \rho_2) := \sum_{m,l=1}^{\infty} 2^{-(m+l)} \sum_{B \in B^{m,l}} |\rho_1(B) - \rho_2(B)|$$

REFERENCES

[1] D. Ryabko. Clustering processes. ICML 2010.
[2] CMU graphics lab motion capture database.
[3] Lei Li and B. Aditya Prakash. Time series clustering: Complex is simpler! ICML 2011.
[4] T. Jebara, Y. Song, and K. Thadani. Spectral clustering and embedding with HMMs. ECML 2007.

MAIN THEORETICAL RESULT

Theorem: *There exists an online clustering algorithm that is asymptotically consistent provided that the distributions generating the data are stationary and ergodic.*

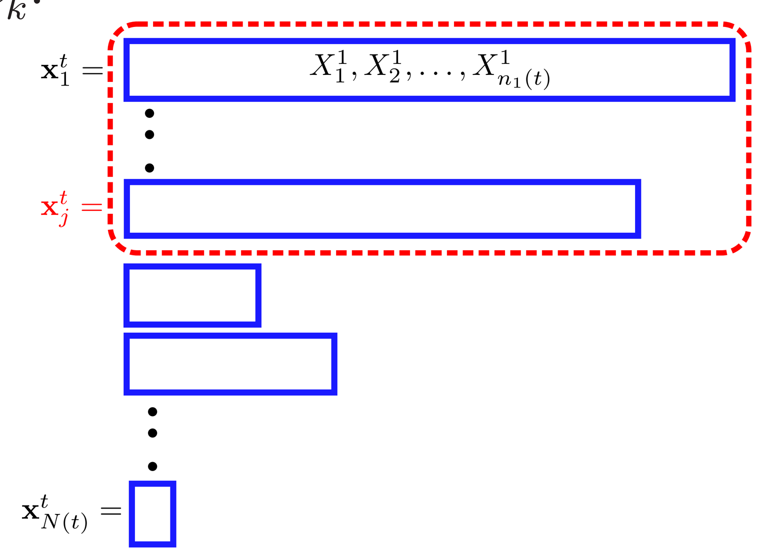
PROPOSED ALGORITHM

Key Idea:

Combine Batch Clusterings with Weights!

Algorithm

1. For $j = k..N(t)$, use a (consistent) batch algorithm on $\mathbf{x}_1^t, \ldots, \mathbf{x}_j^t$ to obtain k cluster centers: $c_1^j, \ldots, c_k^j$.



2. Calculate two sets of weights:

   i. $\gamma_j = \min_{i \neq i' \in 1..k} \hat{d}(c_i^j, c_{i'}^j)$ the min inter-cluster distance.

   ii. $w_j = j^{-2}$ the chronological weight.

3. Assign points to clusters: For every sequence $\mathbf{x}$, choose the index $i \in 1..k$, s.t. i minimizes,

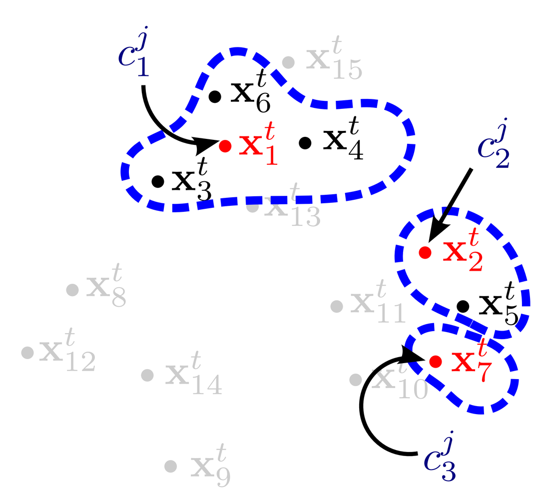
$$\frac{1}{\eta} \sum_{j=1}^{N(t)} w_j \gamma_j \hat{d}(\mathbf{x}, c_i^j)$$

where, $\eta := \sum_{j=1}^{N(t)} w_j \gamma_j$ is the normalization factor.

IDEA OF THE PROOF

1. **The distance $\hat{d}(\cdot,\cdot)$ is consistent:**

→ The performance weight γ_j converges to 0, when the cluster-centers are obtained from sequences generated by less than k processes.



2. **The batch algorithm is consistent [1]:**

→ Once samples from all k clusters are observed, from some time on, the cluster-centers $c_1^j, \ldots, c_k^j$ are consistently chosen to each, uniquely represent one of the k distributions.

3. **Algorithm is not confused by "bad" points:**

Sets of sequences $\mathbf{x}_1^t, \ldots, \mathbf{x}_j^t$ for larger j contain potential "bad" points: newly formed sequences, with inaccurate distance estimates. Decisions based on earlier sequences are more reliable.

→ The chronological weight w_j gives precedence to cluster-centers $c_1^j, \ldots, c_k^j$ produced earlier, i.e. smaller j.

EXPERIMENTAL RESULTS

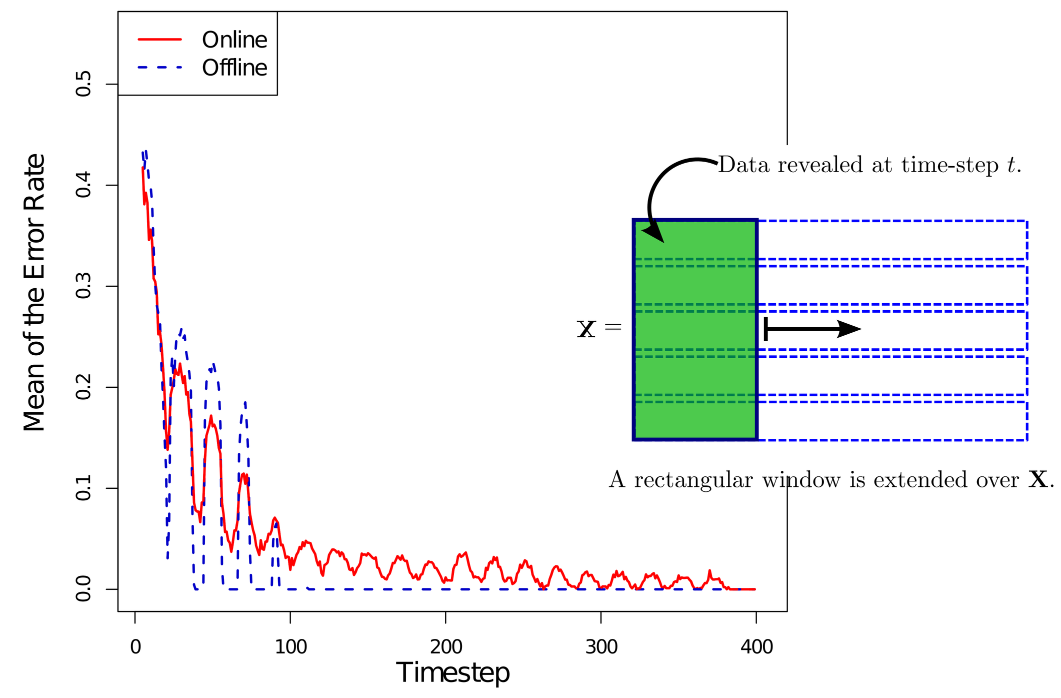
1. Synthetic Data

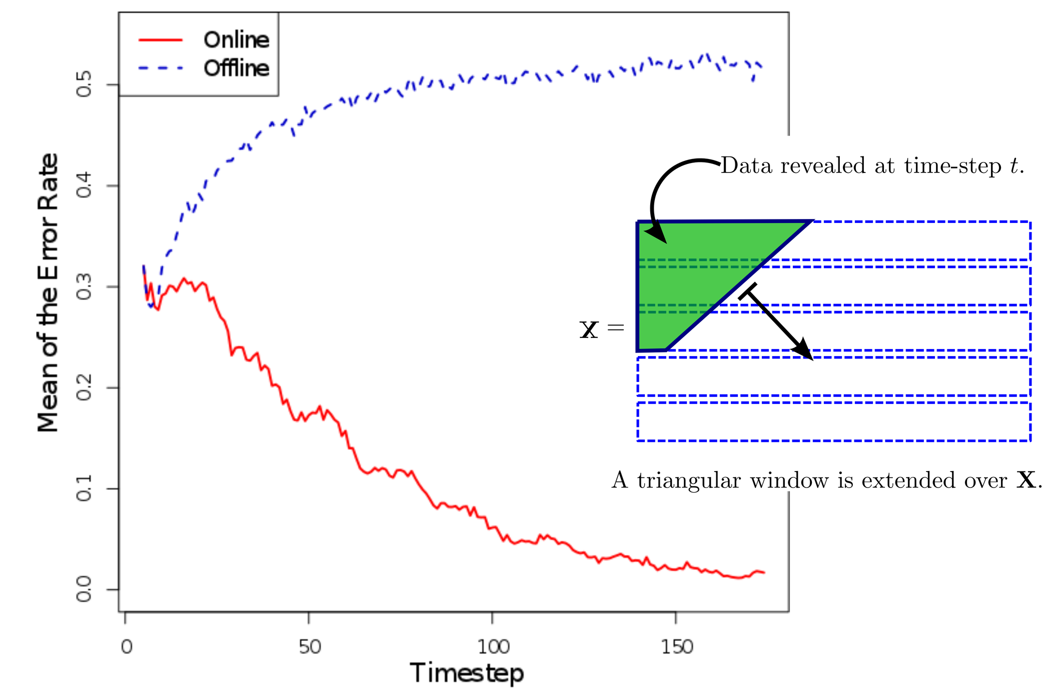
Setup: We generated a data matrix $\mathbf{X}$, where each row a sequence generated by one of the five processes, $k = 5$.

Batch Simulation: Data revealed via a rectangular window extended over $\mathbf{X}$.

Online Simulation: Data revealed via a triangular window extended over $\mathbf{X}$.

Remark: We use processes that, while being stationary-ergodic do not belong to any "simpler" class. They cannot be modeled as a hidden Markov process with a countable set of states.





Top: error-rate vs. sequence-length in batch setting (both algorithms are consistent). Bottom: error-rate vs. # of samples in online setting (the offline algorithm is constantly confused by the new sequences).

2. Real Data:

(Clustering Motion Capture Sequences)

Setup: We used time-series data from [2] representing human locomotion; sequences are marker positions tracked spatially through time.

Objective: Cluster the video sequences based on the activity they represent, ex. Walking, Running, etc.

We compare against [3] and [4].

Dataset	[3]	$f(\cdot,\cdot)$
Walk vs. Run (#35)	0.1015	**0**
Walk vs.Run (#16)	0.3786	**0.2109**

Dataset	[4]	$f(\cdot,\cdot)$
Ergodic Motions		
Run vs. Run/Jog	**100%**	**100%**
Walk vs. Run/Jog	95%	**100%**
Non-ergodic Motions		
Jump vs. Jump fwd.	87%	**100%**
Jump vs. Jump fwd.	**66%**	60%

Top: Comparison against [3]; (performance measure: entropy of the true labeling with respect to the prediction) Bottom: Comparison against [4]; (performance measure: the percentage of correct classification). The numerical of [3, 4] results are taken directly from their corresponding articles.; the same sets of sequences, and means of evaluation are used.