Learning Temporal Properties from Event Logs via Sequential Analysis

Francesco Chiariello

IRIT, ANITI, University of Toulouse

Contact: francesco.chiariello@irit.fr

Table of Contents

Sequential Analysis

Statistical hypothesis testing approach with the following features:

- **Sequential Data Processing**: Data are examined in sequence, often as soon as they become available.
- **Step-by-Step Decision Making**: At each step, decide to accept, reject, or gather more data based on the evidence.
- **Data Efficiency**: More efficient than traditional approaches, which process data in large batches.
- **Robustness to Noise**: Based on statistical principles.

Sequential Analysis in Quality Control

Event Logs and Processes

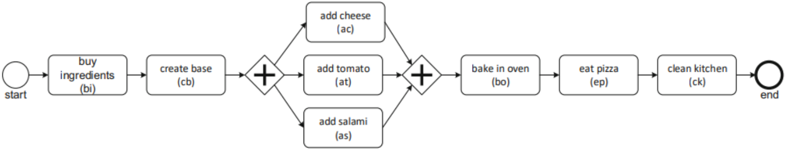


Figure: Pizza Process (from the Process Mining Handbook)

- A **process trace** is a sequence of activities from start to end.
- An **event log** is a sequence of traces.
- Traces may be affected by **noise**.

Linear Temporal Logic on Process Traces (LTL_p)

- An LTL_p formula φ over a set of activities Σ is defined by the following grammar:

$$\varphi ::= a \mid \neg\varphi \mid \varphi \wedge \varphi \mid X(\varphi) \mid \varphi U \varphi$$

- Common abbreviations are used: *true*, $\rightarrow$, $\vee$, F, G.

LTL_p semantics

- Given a formula φ, a trace $\pi = \pi_1, \pi_2, \ldots, \pi_{len(\pi)} \in \Sigma^+$, and a time instant i, with $1 \leq i \leq len(\pi)$, the semantics is defined as follows:
  - $\pi, i \models a$ iff $a = \pi_i$,
  - $\pi, i \models \neg\varphi$ iff $\pi, i \not\models \neg\varphi$,
  - $\pi, i \models \varphi_1 \wedge \varphi_2$ iff $\pi, i \models \varphi_1$ and $\pi, i \models \varphi_2$,
  - $\pi, i \models \mathsf{X}\varphi$ iff $i < len(\pi)$ and $\pi, i + 1 \models \varphi$,
  - $\pi, i \models \varphi_1 \mathsf{U} \varphi_2$ iff $\pi, j \models \varphi_2$ for some j, with $i \leq j \leq len(\pi)$, and $\pi, k \models \varphi_1$ for all $k = i, \ldots, j - 1$.
- We write $\pi \models \varphi$, and we say that π satisfies φ, if $\pi, 1 \models \varphi$.

Table of Contents

Learning Temporal Properties from Event Logs via Sequential Analysis

- **Problem**: learn temporal formulae from a given log

Learning Temporal Properties from Event Logs via Sequential Analysis

- **Problem**: learn temporal formulae from a given log
- **Solution:** Apply sequential analysis
  - Select a candidate formula.
  - Select the n-th trace.
  - Compute the number of defects of the trace w.r.t. the formula.
  - Add it to the cumulative error $defects_n$
  - If:
    - $defects_n \leq A_n$ we accept the formula.
    - $defects_n \geq R_n$ (with $R_n > A_n$) we reject the formula.
    - if $A_n < defects_n < R_n$ we select the $(n+1)$-th trace and repeat.

Learning Temporal Properties from Event Logs via Sequential Analysis

- **Problem**: learn temporal formulae from a given log
- **Solution:** Apply sequential analysis
  - Select a candidate formula.
  - Select the n-th trace.
  - Compute the number of defects of the trace w.r.t. the formula.
  - Add it to the cumulative error $defects_n$
  - If:
    - $defects_n \leq A_n$ we accept the formula.
    - $defects_n \geq R_n$ (with $R_n > A_n$) we reject the formula.
    - if $A_n < defects_n < R_n$ we select the $(n+1)$-th trace and repeat.
- **Question 1**: How do we select the thresholds?
- **Question 2**: How do we count the number of defects?

Sequential Probability Ratio Test (SPRT)

- Let p denote the probability of a defect, p_0 a tolerable value, and $p_1 > p_0$ an intolerable one.
- The null hypothesis is $H_0 : p = p_0$, while the alternative hypothesis is $H_1 : p = p_1$.
- SPRT defines A_n and R_n as:

$$A_n = \frac{\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)} + n\frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)},$$

$$R_n = \frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)} + n\frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)}$$

where the parameters α and β allows to control type I and type II errors, respectively.

SPRT (cont'd)

- We can therefore rewrite A_n and R_n as:

$$A_n = mn + c_A,$$
$$R_n = mn + c_R,$$

with $m > 0$, $c_A < 0$, and $c_R > 0$.

Trace Alignment

- **Trace Alignment** is the problem of aligning a trace π with a formula φ, producing a new trace π' satisfying φ.
- Actions:
  - `add(?activity)`, of cost 1,
  - `del`, of cost 1,
  - `read`, of cost 0.
- Repair a trace can be reduced to cost-optimal planning.

Example

$\varphi = \mathsf{G}(a \rightarrow \mathsf{F}b)$ $\quad\quad \pi = aba \quad\quad \pi' = aba$

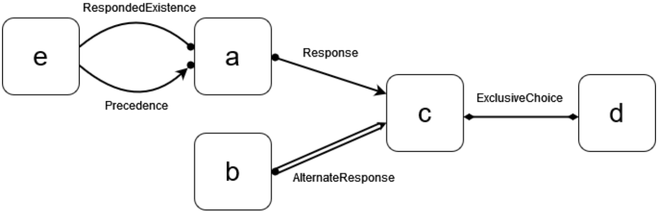
1. `read`: a|ba
2. `read`: ab|a
3. `del`: ab|

Table of Contents

Declarative Model

We consider the declarative model over $\Sigma = \{a, b, c, d, e\}$ specified by the following 5 constraints:

$$
\begin{aligned}
&(C1) & \mathit{ExclusiveChoice}(c, d) &\equiv \mathsf{F}(c \vee d) \wedge \neg(\mathsf{F}c \wedge \mathsf{F}d)\\
&(C2) & \mathit{Response}(a, b) &\equiv \mathsf{G}(a \rightarrow \mathsf{F}b)\\
&(C3) & \mathit{RespondedExistence}(a, e) &\equiv \mathsf{F}a \rightarrow \mathsf{F}e\\
&(C4) & \mathit{Precedence}(e, a) &\equiv (\neg a)\mathsf{W}e\\
&(C5) & \mathit{AlternateResponse}(b, c) &\equiv \mathsf{G}(b \rightarrow \mathsf{X}(\neg b \mathsf{U} c))
\end{aligned}
$$



Log Generation

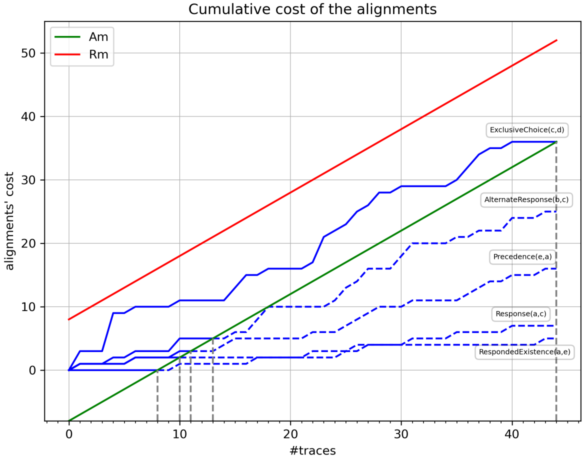
- We generate a log of 100 traces of length varying from 6 to 15. In particular, for each length, we generate 5 positive traces and 5 negative ones.

Table: Number of traces violating the constraints arranged according to the cost of repairs.

	1	2	3+	#traces	total cost
C1	19	13	6	38	68
C2	14	0	0	14	14
C3	9	0	0	9	9
C4	31	0	0	31	31
C5	16	15	2	33	52

Results

- We choose $m = 1, c_R = 8, c_A = -8$.



Results

- We choose $m = 1, c_R = 8, c_A = -8$.

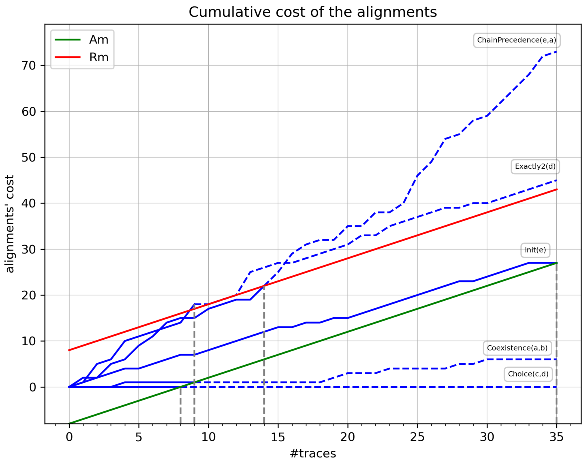


Table of Contents

Conclusion and Future Work

Contributions:

- Learn LTL formulae with sequential analysis.
- Quantify defects using trace alignment.
- Approach **robust** and **data-efficient**.

Future Directions:

- Select candidate formulae.
- Learn full model.
- Quantitative semantics.

Thank you for your attention!

Financé par
GOUVERNEMENT
Liberté
Égalité
Fraternité

FRANCE RELANCE

Financé par
l'Union européenne
NextGenerationEU