Monitoring of Fairness Properties in Markov Chains

Mahyar Karimi

Institute of Science and Technology Austria

September 2022

A Glance at Our Problem

- Fairness in a bank lending scenario[1].
- One fairness metric (*demographic parity*):

$$\mathbb{P}\,(\text{Acc.} \mid \text{G}_1) - \mathbb{P}\,(\text{Acc.} \mid \text{G}_2)$$

How can we approach this problem?

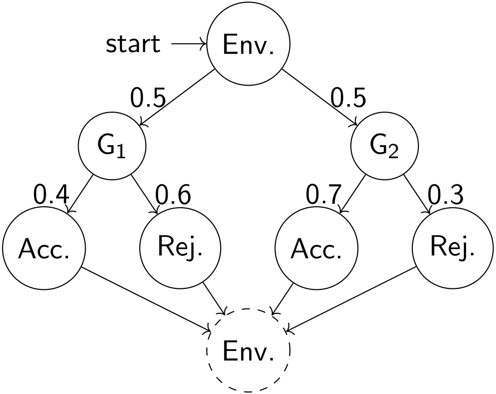


Figure 1: Lending problem, as a Markov chain

[1] Liu et al. "Delayed Impact of Fair Machine Learning", IJCAI 2019.

Our Contributions

- What we developed
  - Monitors for Markov chains
  - Monitoring fairness properties
  - Frequentist and Bayesian approaches: two well-known paradigms
- Estimate expressions of probabilities over an stochastic system
- Implemented some approaches as a Rust library

Table of Contents

Monitoring (1/2)

What monitoring is:

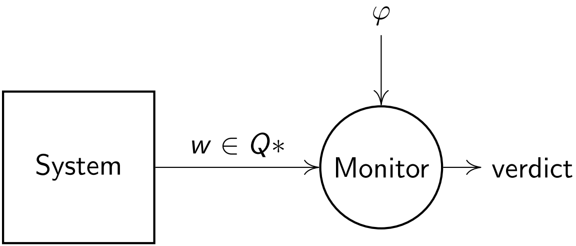


Figure 2: A Monitoring Setting.

What the monitor reports: (mainly) **true**/**false**.

Monitoring (2/2)

A monitor can also perform as a **numerical** estimator.

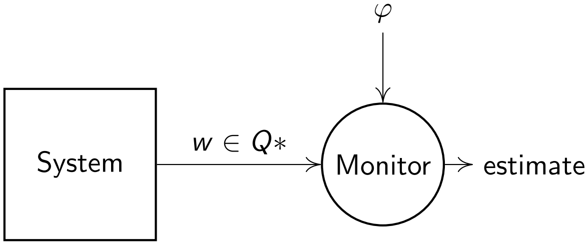


Figure 3: A Quantitative Monitoring Setting.

What the monitor reports: a **number** (e.g. confidence interval).

A Peek at Markov Chains

$$\mathcal{M} = (Q, \mathbb{P}(\cdot \mid \cdot), v)$$

- Q: State space
- $\mathbb{P}(\cdot \mid \cdot)$: Transition probabilities
- v: distribution of initial states

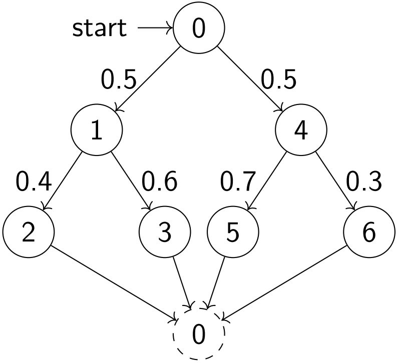


Figure 4: A Sample Markov Chain.

Table of Contents

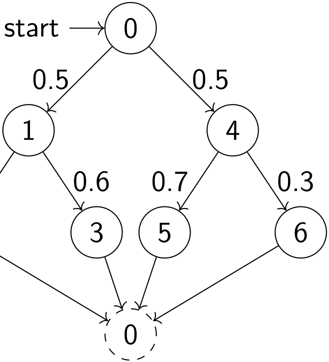
Problem Statement

Assume the following expression:

$$\mathbb{P}(2 \mid 1) - \mathbb{P}(5 \mid 4)$$

- How fair a bank gives loans to different ethnic groups.
- How well a load balancer distributes the load.

What can our monitor say about this expression?



The Frequentist, The Bayesian

Two well-known paradigms of this problem:

- The *frequentist* approach
  - Fixed underlying Markov chain
  - Probabilities $\equiv$ Long-run frequencies
- The *Bayesian* approach
  - Markov chains themselves are sampled
  - User-specified distribution over Markov chains

Each approach gives us a different expression to estimate.

Table of Contents

The Frequentist Approach (1/2)

Prior work in this direction:

- Albarghouthi and Vinitsky[2]: Estimating single frequencies, aggregating estimates (and errors).
- Programming framework for specifying fairness properties.
- Output is a confidence interval, from Hoeffding's inequality.
  - Not assuming the underlying system is a Markov chain.
  - Error propagation might result in error blow-up.

[2] Albarghouthi and Vinitsky. "Fairness-Aware Programming", FAT* '19

The Frequentist Approach (2/2)

Our approach:

- Translate the expression into an expression tree.
- Construct "proxy" random variables for each sub-tree.
- Compute estimate at tree root $\rightarrow$ only one place of error generation!
- More complex expressions such as $\mathbb{P}(2 \mid 1) \times \mathbb{P}(5 \mid 4)$ or $(\mathbb{P}(2 \mid 1))^{-1}$ can be monitored as well.

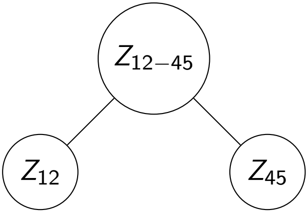


Figure 5: RV's for $\mathbb{P}(2 \mid 1) - \mathbb{P}(5 \mid 4)$

A Frequentist Example (1/2)

- Assume just one probability $\mathbb{P}(2 \mid 1)$ ($\equiv$ Bernoulli with $p = 0.4$).
- Sequence of trials ($\equiv$ observations): $X_1, X_2, \cdots$.

$$Z_{12}$$

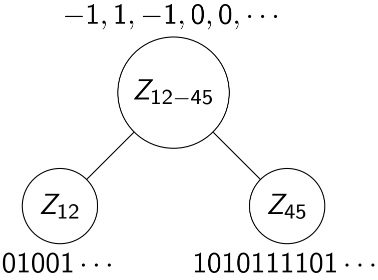
$$01001\cdots$$

- Estimating p: already known technique.
- Hoeffding's inequality:

$$\mathbb{P}\left((X_1 + \cdots + X_n) - np \geq \varepsilon\right) \leq e^{-\frac{2\varepsilon^2}{n}}$$

A Frequentist Example (1/2)

What about an expression like $\mathbb{P}(2 \mid 1) - \mathbb{P}(5 \mid 4)$?



- We construct a new sequence.
- Preserving expected value:

$$\mathbb{E}(Z_{12-45}) = \mathbb{E}(Z_{12}) - \mathbb{E}(Z_{45})$$

Our Frequentist Monitor

- To construct intermediate RVs, we use queues to keep observed data.
- To get a confidence interval, we again use Hoeffding's inequality.

We can formally define our frequentist monitor as follows:

Definition

For a given Markov chain $\mathcal{M}$, a property φ, and a confidence level δ, a frequentist monitor $\mathcal{A} : \Sigma * \to I_{-\infty,\infty}$ will, after observing a word w, provide an interval $\mathcal{A}(w)$ such that:

$$\mathbb{P}([\![\varphi]\!] \in \mathcal{A}(w)) \geq 1 - \delta$$

Table of Contents

The Bayesian Approach (1/2)

- In the Bayesian case, we assume the Markov chain **itself** is sampled from a distribution.
- The distribution and its parameters are set by the operator.
- To the best of our knowledge, no prior work exists in this direction.

The Bayesian Approach (2/2)

Simple solution to our problem: integrate over all Markov chains. This solution is **infeasible**! The space of all Markov chains is not finite.

The Bayesian Approach (2/2)

Simple solution to our problem: integrate over all Markov chains. This solution is **infeasible**! The space of all Markov chains is not finite.

Better solution: do the same thing! Just smarter.

The Bayesian Approach (2/2)

Simple solution to our problem: integrate over all Markov chains. This solution is **infeasible**! The space of all Markov chains is not finite.

Better solution: do the same thing! Just smarter.

- Fix the size of Markov chains.
- Assume each row is sampled from a Dirichlet distribution, with given parameters.

The Bayesian Approach (2/2)

Simple solution to our problem: integrate over all Markov chains. This solution is **infeasible**! The space of all Markov chains is not finite.

Better solution: do the same thing! Just smarter.

- Fix the size of Markov chains.
- Assume each row is sampled from a Dirichlet distribution, with given parameters.

With these new assumptions, integrating over all Markov chains becomes analytically solvable.

Our Bayesian Monitor

The Bayesian monitor estimates a **posterior** value for a property, after a finite word is observed.

Definition

For a given property φ, a prior distribution with parameters O, a Bayesian monitor $\mathcal{A} : \Sigma* \to \mathbb{R}$ will, after observing a word w, provide a value for $\mathbb{E}\left(\llbracket\varphi\rrbracket_{\mathcal{M}} \mid w\right)$.

A Taste of The Bayesian

Consider the same problem:

$$\mathbb{P}(2 \mid 1) - \mathbb{P}(5 \mid 4)$$

Assume we use the Bayesian monitor with a uniform prior. We have shown that, after observing a word w, we have:

$$\mathbb{E}\left(\mathbb{P}(2 \mid 1) - \mathbb{P}(5 \mid 4) \mid w\right) = \frac{1 + \#_w(12)}{2 + \#_w(1)} - \frac{1 + \#_w(45)}{2 + \#_w(4)}$$

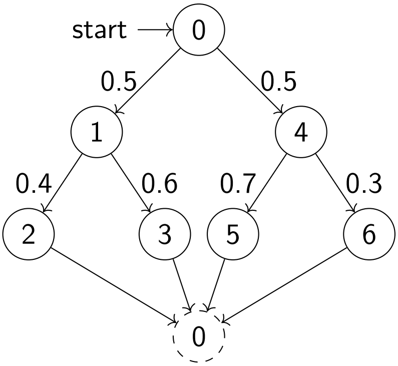


Table of Contents

Monitoring Regular Expressions

- Our results show we can monitor polynomial expressions over transition probabilities.
- Monitoring a fixed-size regular expression can also be reduced as monitoring a polynomial expression of probabilities.

$\Longrightarrow$ We can also monitor fixed-size regular expressions.

Table of Contents

Implementation and Results (1/2)

We have implemented the two monitors we have developed in the Rust programming language.

- We have tested this implementation with the Markov chain in figure 6.
- We observed 10000 iterations of the chain in each experiment.
- For the Bayesian monitor, we use uniform prior.

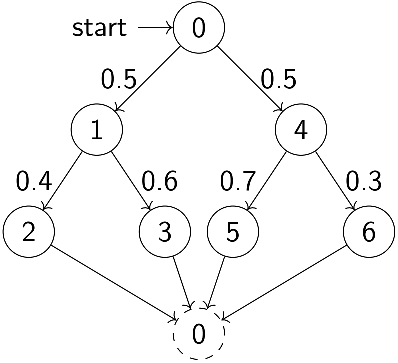


Figure 6: Our test Markov chain.

Implementation and Results (2/2)

Results with
$\varphi = \mathbb{P}(2 \mid 1) - \mathbb{P}(5 \mid 4)$:

- The frequentist monitor:

  $$\mathbb{P}([\![\varphi]\!] \in [-0.359, -0.226]) \geq 0.95$$

- The Bayesian monitor:

  $$\mathbb{E}([\![\varphi]\!] \mid w) = -0.297$$

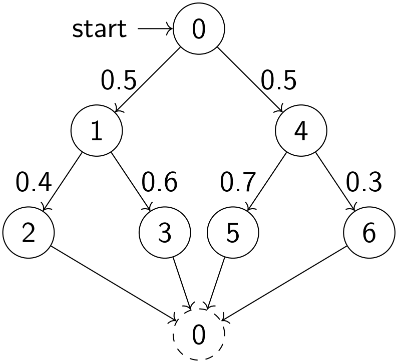


Table of Contents

Future Steps

- Improve the use of queues in the frequentist.
- Unifying result form for the two monitors.
  - Using other inequalities for confidence bounds (such as Chebyshev's).
  - Other metrics, such as p-value.
- What about **time-varying** systems?
  - Many real-world systems vary over time.
  - To what extent can we use our current monitors?

Table of Contents

Summary

- Monitoring fairness metrics; we assume our system is a Markov chain.
- Using the two established paradigms in this direction:
  - The frequentist $\rightarrow$ Confidence intervals.
  - The Bayesian $\rightarrow$ (Exact) posterior expectation.
- Implementation and experiments with examples from the literature.