Policy Iteration for Sum-Product Inferences

Pieter-Jan Vandaele, Jasper De Bock

{PieterJan.Vandaele,Jasper.DeBock}@UGent.be

GHENT UNIVERSITY

FLip, Belgium

Abstract

In this poster we take a closer look at the similarities between Markov Decision Processes and Imprecise Markov Chains. Using this connection, we apply the MDP-technique of policy iteration to sum-product inferences as known in IMC literature, leading to a new efficient algorithm.

Intro

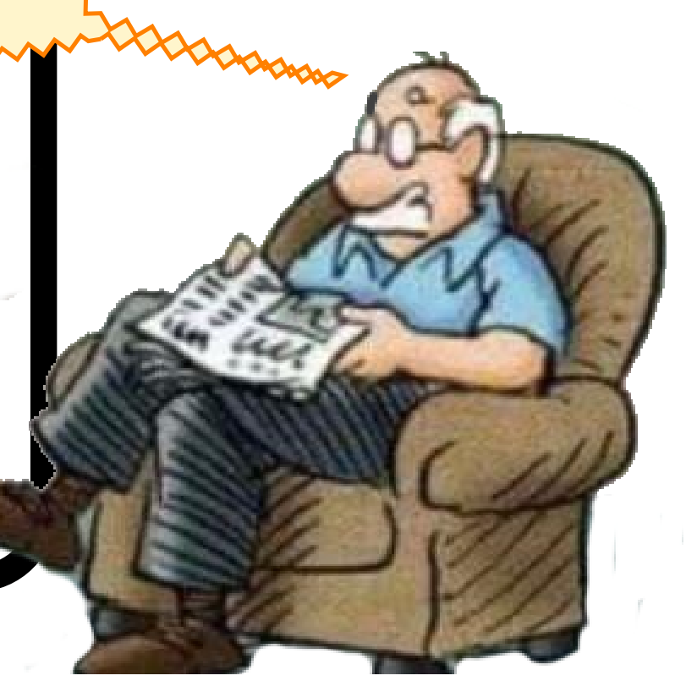
Does this problem sound familiar?

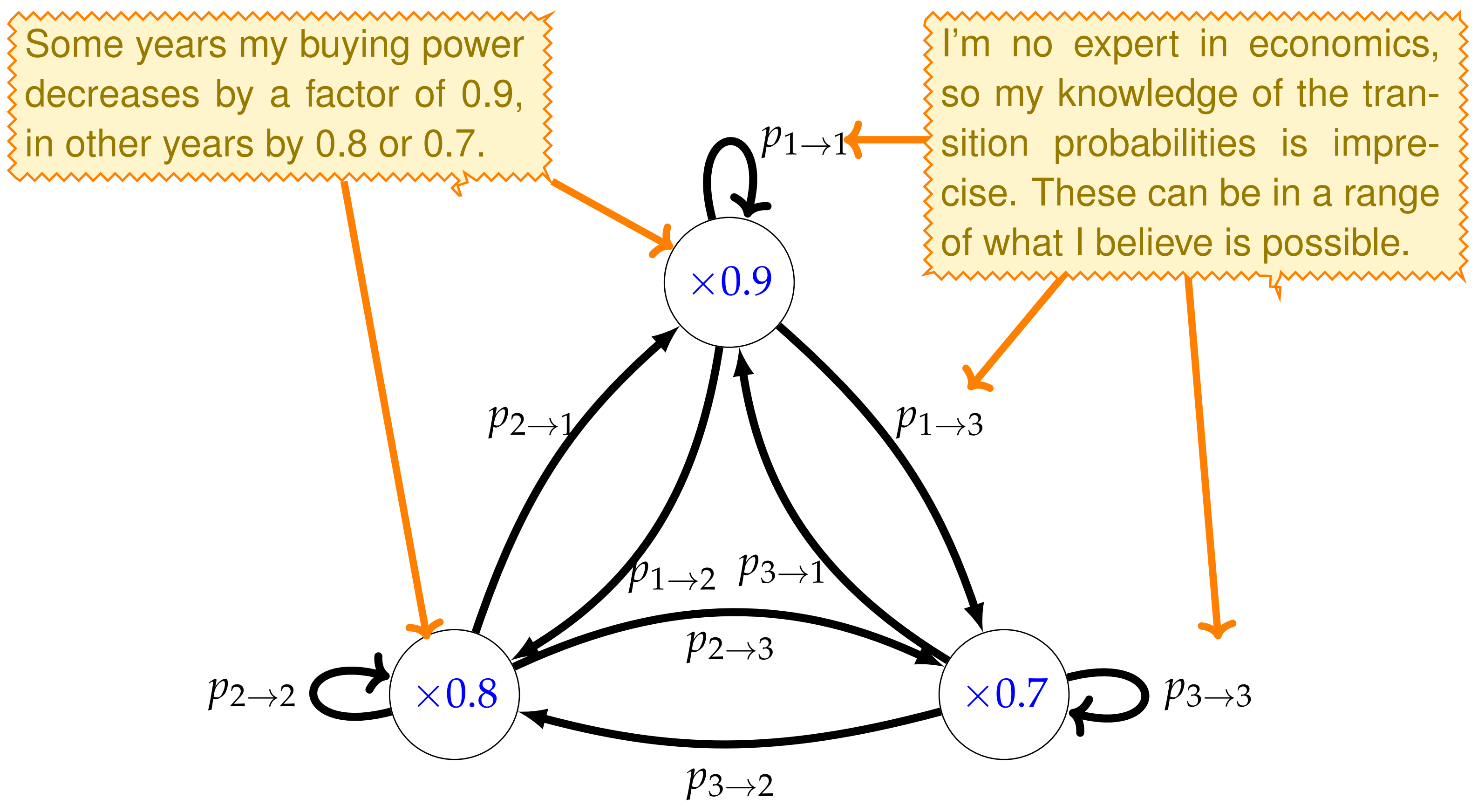
I only get paid a measly wage, and due to that darn ever-changing inflation, it is worth less and less!

Do you also wonder the following?

Last year, I could buy 100 Pokémon cards with my salary. Now I can only buy 90! I really wonder how big my collection will be when I retire.

Well, wonder no more, because policy iteration for sum-product inferences is here! This is an example of a system that can be modelled with a Sum-Product Inference. In particular, for a sum-product gamble f, we are interested in the limit $\underline{\mathrm{E}}_\infty(f|x) = \lim_{n\to\infty} E[f(X_{1:n})|X_1 = x]$.





Some years my buying power decreases by a factor of 0.9, in other years by 0.8 or 0.7.

I'm no expert in economics, so my knowledge of the transition probabilities is imprecise. These can be in a range of what I believe is possible.

Imprecise Markov Chains (IMCs) ≈ Markov Decision Processes (MDPs)

IMCs in one sentence: When facing an uncertain process, what is the maximal or minimal outcome that I could expect from a given gamble?

MDPs in one sentence: Given a range of possible actions, what should I do to maximise or minimise the expected reward of some process?

Main Features

IMCs		MDPs
Discrete time •	↔	• Decision epochs $T = 1, 2, 3\dots$
State space $\mathcal{X}$ •	↔	• State space S (here assumed to be finite)
Set of transition matrices $\mathcal{T}$, in this poster we assume this set to be closed and has separately defined rows. •	↔	• Each state has an action space $\mathcal{A}_{s,t}$. • A policy $\pi(s)$, which dictates what action to take at state s • Transition probabilites $P_{t,a}$
Gamble $f(x)$ •	↔	• Reward $r_t(s,a)$
Inferences can be calculated with recursive schemes •	↔	• The maximal expected reward can be calculated using the value iteration algorithm.

Different types

An IMC can be seen as a set of stochastic processes, of which three are particularly interesting.

MDP literature distinguishes different types of policies, based on how a policy can choose which action to take.

IMCs		MDPs
1. **Repetition independence (RI)**: contains all stationary precise Markov chains compatible with the imprecise probabilities.	↔	1. **Stationary deterministic (SD)**: regardless of time, the policy will always choose the same action in a given state.
2. **Complete independence (CI)**: contains all precise Markov chains compatible with the imprecise probabilities.	↔	2. **Markovian deterministic (MD)**: the action is chosen based solely on the current state, but the decision strategy can change over time.
3. **Epistemic irrelevance (EI)**: contains all compatible stochastic processes, but the individual processes are not neccesarily Markovian.	↔	3. **History random (HR)**: the chosen policy chooses an action based on the previous actions and states.

Unique to IMCs

- Considers more general inferences such as hitting times, hitting probabilities or 'time-bounded until' events. These appear under the umbrella of

Sum-Product Inferences ❶.

Unique to MDPs

- Typically considers 'discounted' reward processes
- Rewards (or costs) can depend on the chosen action
- Not only interested in the extremal value possible, but also the path that it takes to get there
- Has a wider variety of algorithms, such as

policy iteration ❷

Sum-Product Inferences

We consider so-called sum-product gambles $f(X_{1:n}) \in \mathcal{L}(\mathcal{X}^n)$. Their value for a trajectory $x_{1:n}$ is given by

$$f(x_{1:n}) := \sum_{k=1}^{n} g(x_k) \prod_{l=1}^{k-1} h(x_l). \qquad (1)$$

Here g can be any gamble, and $g(x)$ seen as a reward collected when entering state x. The gamble h should be non-negative and represents a multiplicative modifier for future rewards. When looking at sum-product gambles, the following holds:

$$\underline{\mathrm{E}}^{CI}[f(X_{1:n})] = \underline{\mathrm{E}}^{EI}[f(X_{1:n})]. \qquad (2)$$

That is, there is no distinction in lower (or upper) expectation when considering an IMC under EI or CI. This value, however, does not neccesarily coincide with the expectation of the IMC under RI.

Existing method for calculating Sum-Product inferences

When considering IMCs under repetition independence or epistemic irrelevance, the upper and lower expectation at time instant n can be calculated by means of a recursive scheme:

$$\underline{\mathrm{E}}(f(X_{1:n})) = \underline{\mathrm{L}}^n 0, \qquad (3)$$

with the lower progression function $\underline{\mathrm{L}} : \mathcal{X} \to \mathcal{X}$ defined as

$$\underline{\mathrm{L}}\pi = h\underline{\mathrm{T}}\pi + g \qquad (4)$$

This approach, however, leaves two problems:

- In the limit for n going to infinity, the recursion should be done infinitely many times.

I have to this calculation an infinite amount of times!? Nobody's got time for that!

- This does not work in the case of repetition independence, where all possible precise Markov chains should be compared.

I should compare ALL compatible precise chains?

❶ + ❷ = Our novel algorithm:

The algorithm

If all $h(x)$ are smaller than unity, it is possible to use a policy-iteration like scheme to calculate the limit expectations. The procedure is as follows:

1. Initialisation: choose a transition matrix T_0 from $\mathcal{T}$
2. Evaluation: solve the linear equation for f_n
$$(I - h \cdot T_n)f_n = g \qquad (5)$$
3. Update: Choose T_{n+1} such that
$$T_{n+1}f_n = \underline{\mathrm{T}}f_n \qquad (6)$$
choosing $T_{n+1} = T_n$ if possible.
4. Stop if $T_{n+1} = T_n$, then $\underline{\mathrm{E}}_1(f_n) = \underline{\mathrm{E}}_\infty(f)$

In the limit, all types of IMCs are equivalent!

We have shown that when looking at the limit $\underline{\mathrm{E}}_\infty(f|x)$ the type of IMC considered does not matter. This means we can limit ourselves to looking at stochastic processes in the set of IMCs under repetition independence, the most restrictive set. This is the basic idea of policy iteration.

Time Complexity

The main computational expense of each iteration consists of solving the linear equation, which can be done in $\mathcal{O}(|\mathcal{X}|^\omega)$, where ω is the matrix multiplication constant (in pratical algorithms this is about 2.8).

Assuming that $\mathcal{T}$ is defined by a finite amount of linear bounds - which is doable in most practical settings. Then the algorithm finishes in **a finite amount of iterations**.

Question: what is the IMC equivalent of history random MDPs?

¿¿¿FLiP quiz???

References:

[1] Martin L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, 2005.

[2] De Bock, Erreygers, Krak. Sum-product laws and efficient algorithms for imprecise Markov chains. PMLR, 2021.