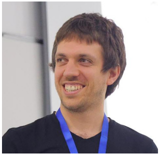
POMDPs and Blind MDPs: (Dis)continuity of Values and Strategies

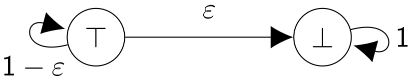


K. Chatterjee[1] **R. Saona**[1]

[1]Institute of Science and Technology Austria (ISTA)

Institute of Science and Technology Austria

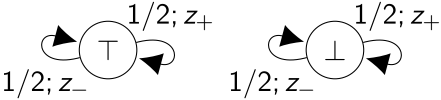
(Deterministic) (dynamic)



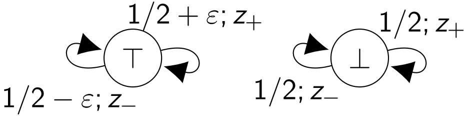
Similar? stochastic dynamic

Stochastic dynamics (MCs) must consider structure when analyzing continuity.

Continuity in Partially Observable Stochastic dynamics



(Static) partially observable (stochastic) dynamic



Similar? partially observable stochastic dynamic

Belief dynamics are fragile to structurally preserving changes.

Continuity concepts

- Value-continuity
  Value of similar POMDPs is close

- Weak strategy-continuity
  Some approximately-optimal strategy is still approximately-optimal in similar POMDPs

- Strong strategy-continuity
  All approximately-optimal strategies are approximately-optimal in similar POMDPs

Results

Model	Continuity		
	Value	Weak strategy	Strong strategy
Fully-observable MDPs	Yes	Yes	**No**
POMDPs	**No**	**No**	**No**
Blind MDPs	**Yes**	**Yes**	**Yes**

Theorem: Deciding whether a POMDP is continuous
is **algorithmically impossible**.

Remarks

- Blind MDPs are strictly more well-behaved than POMDPs
- Blind MDPs are strictly more well-behaved than MDPs

Model

A Partially-Observable Markov Decision Process (POMDP) is a tuple $\Gamma = (\mathrm{States}, \mathrm{Actions}, \mathrm{Zignals}, p_1, \delta)$ where

- States is a finite set of **states**;
- $\mathrm{Actions}$ is a finite set of **actions**;
- $\mathrm{Zignals}$ is a finite set of **signals**;
- $p_1 \in \Delta(\mathrm{States})$ is an **initial distribution**;
- $\delta \colon \mathrm{States} \times \mathrm{Actions} \to \Delta(\mathrm{States} \times \mathrm{Zignals})$ is a probabilistic transition function.

Special cases:

$$
\begin{aligned}
|\mathrm{Zignals}| = 1 \quad &\Rightarrow \quad \text{blind MDP} \\
\mathrm{zignal} = \mathrm{state} \quad &\Rightarrow \quad \text{(fully-observable) MDP}
\end{aligned}
$$

Model

- **strategy** $\sigma \colon \bigcup_{n \geq 0} (\mathrm{Actions} \times \mathrm{Zignals})^n \to \Delta(\mathrm{Actions})$
- **play** $\omega = (s_n, a_n, z_{n+1})_{n \geq 1} \subseteq \mathrm{States} \times \mathrm{Actions} \times \mathrm{Zignals}$
- probability $\mathbb{P}^{\sigma}_{p_1}[\Gamma]$ and expectation $\mathbb{E}^{\sigma}_{p_1}[\Gamma]$
- **belief**

$$\mathbb{P}^{\sigma}_{p_1}\left(S_m = \cdot \mid \forall i \in [m-1] \quad A_i = a_i, Z_{i+1} = z_{i+1}\right),$$

- **reward** $\mathrm{reward} \colon \mathrm{States} \times \mathrm{Actions} \to \mathbb{R}$
- **objective** $\mathrm{payoff}(\omega)$ is one of

$$\liminf_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathrm{reward}(s_i, a_i) \qquad \limsup_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathrm{reward}(s_i, a_i)$$

$$\liminf_{m \to \infty} \mathrm{reward}(s_m, a_m) \qquad \limsup_{m \to \infty} \mathrm{reward}(s_m, a_m)$$

Model

- **value**

$$\mathsf{val}(\Gamma) := \sup_{\sigma} \mathbb{E}^{\sigma}_{p_1}(\mathrm{payoff}(\omega))$$

- ε**-optimal strategy** $\mathbb{E}^{\sigma}_{p_1}(\mathrm{payoff}(\omega)) \geq \mathsf{val}(\Gamma) - \varepsilon$

Special concepts

- **structural equivalence**

$$\mathsf{supp}(\delta(s,a)) = \mathsf{supp}\left(\widetilde{\delta}(s,a)\right)$$

- ξ**-similar POMDPs** For all s, a, s', z,

$$\frac{1}{1+\xi}\,\delta(s,a)(s',z) \;\leq\; \widetilde{\delta}(s,a)(s',z) \;\leq\; (1+\xi)\,\delta(s,a)(s',z)$$

Results

Model	Continuity		
	Value	Weak strategy	Strong strategy
Fully-observable MDPs	Yes	Yes	**No**
POMDPs	**No**	**No**	**No**
Blind MDPs	**Yes**	**Yes**	**Yes**

Theorem: Deciding whether a POMDP is value-, weakly strategy-, or strongly strategy-continuous is **algorithmically impossible**.

Blind MDPs:
no signals
guarantees continuity

Previous results

Theorem (Stability of invariant distribution, O'Cinneide 1993)

Consider an irreducible stochastic matrix Δ.
Computing the stable distribution

$$p^\top = p^\top \Delta$$

is a stable operation.

The proof is by induction on the dimension of Δ, possible thanks to a characterization of the limit

Previous result

Theorem (Stability of discounted occupation times, Solan 2003)

Consider a Markov Chain with a starting state.
The ratio of λ-discounted occupation times at a state s is a rational function of the transition probabilities, i.e., for $\lambda > 0$

$$\delta \mapsto \frac{time_\lambda(s, \delta)}{time_\lambda(s, \widetilde{\delta})} = \frac{poly_s(\delta, \widetilde{\delta})}{poly_s(\delta, \widetilde{\delta})} .$$

From this result, we conclude value- and weak strategy-continuity for (fully-observable) MDPs (and zero-sum stochastic games).

Blind MDPs: Belief dynamic

The belief update in blind MDPs is directly given by the transition. For each action a, define the matrix

$$(M_a)_{s,s'} := \delta(s,a)(s') \,.$$

After playing actions a, b, a, ..., the beliefs are

$$p_1^\top \qquad p_1^\top M_a \qquad p_1^\top M_a M_b \qquad p_1^\top M_a M_b M_a \qquad \ldots$$

For similar matrices $\widetilde{M}_a$, the beliefs in the corresponding similar blind MDP are

$$p_1^\top \qquad p_1^\top \widetilde{M}_a \qquad p_1^\top \widetilde{M}_a \widetilde{M}_b \qquad p_1^\top \widetilde{M}_a \widetilde{M}_b \widetilde{M}_a \qquad \ldots$$

How different can they be?

Belief-continuity is enough

Definition (Belief-continuity)

A blind MDP is belief-continuous if, for every $\varepsilon > 0$, there exists $\xi > 0$ such that, for all $\widetilde{\Gamma}$ such that $\mathrm{dist}(\Gamma, \widetilde{\Gamma}) \leq \xi$,

$$\sup_{\substack{n \geq 1 \\ (a(i))_{i \in [n]}}} \mathrm{dist}\left(M_{a(1)} \cdot \ldots \cdot M_{a(n)}, \widetilde{M}_{a(1)} \cdot \ldots \cdot \widetilde{M}_{a(n)} \right) \leq \varepsilon \,.$$

Lemma

If a blind MDP is belief-continuous, then it is XXXX continuous.

Belief-continuity

Theorem

Every blind MDP is belief continuous.

Focus on the n-th step. Define

$$
\begin{aligned}
p^\top &:= p_1^\top M_{a(1)} \cdot \ldots \cdot M_{a(n)} \\
\widetilde{p}^\top &:= p_1^\top \widetilde{M}_{a(1)} \cdot \ldots \cdot \widetilde{M}_{a(n)}
\end{aligned}
$$

We would like that, for all $\varepsilon > 0$, we can choose $\xi > 0$ so that, for all actions a,

$$
\operatorname{dist}\left(p^\top, \widetilde{p}^\top\right) \leq \varepsilon \qquad \text{and} \qquad \operatorname{dist}\left(p^\top M_a, \widetilde{p}^\top \widetilde{M}_a\right) \leq \varepsilon
$$

A stronger notion is the **invariant**

$$
\operatorname{dist}\left(p^\top, \widetilde{p}^\top\right) \leq \varepsilon \qquad \Rightarrow \qquad \operatorname{dist}\left(p^\top M_a, \widetilde{p}^\top \widetilde{M}_a\right) \leq \varepsilon
$$

Blind MDPs: Belief dynamic

Consider stochastic matrices $\{M_i : i \in \mathcal{I}\}$
where the smallest strictly positive transition is uniformly bounded

$$M_i(s, s') > 0 \quad \Rightarrow \quad M_i(s, s') > \delta_{\min} \,.$$

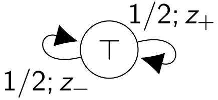
Consider similar matrices $\{\widetilde{M}_i : i \in \mathcal{I}\}$.

$$\begin{array}{ccccc} p_1^\top & p_1^\top M_{i(1)} & p_1^\top M_{i(1)} M_{i(2)} & p_1^\top M_{i(1)} M_{i(2)} M_{i(3)} & \dots \\ p_1^\top & p_1^\top \widetilde{M}_{i(1)} & p_1^\top \widetilde{M}_{i(1)} \widetilde{M}_{i(2)} & p_1^\top \widetilde{M}_{i(1)} \widetilde{M}_{i(2)} \widetilde{M}_{i(3)} & \dots \end{array}$$

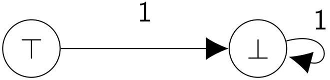
They can not differ by much!

Thank you!

Motivating example



Action win



Action lose

Result: This POMDP is not weakly strategy-continuous.
Proof: There is a fragile approximately optimal strategy.

Proof: Fragile approximately optimal strategy

Consider $t \geq 1$ large enoughand the strategy that plays
$A_1 = A_2 = \ldots = A_t = \mathit{win}$, and,
if *lose* has been played, then $A_{m+1} = \mathit{win}$,
if only win has been played, for $m \geq t$,

$$A_{m+1} = \mathit{lose} \quad \Leftrightarrow \quad |\{i \in [2..(m+1)] : Z_i = z_+\}| \geq \left(1 + m^{-1/4}\right) \frac{m}{2} \,.$$

Proof: Fragile approximately optimal strategy

Lemma (Approximate optimality)

Consider Γ *the previous POMDP. Then,*

$$\mathbb{P}^{\sigma}_{p_1}[\Gamma](\exists m \geq 1,\ A_m = lose) \leq \varepsilon\,.$$

Lemma (Fragility)

Consider $\widetilde{\Gamma}$ *a small perturbation of* Γ. *Then,*

$$\mathbb{P}^{\sigma}_{p_1}[\widetilde{\Gamma}](\exists m \geq 1,\ A_m = lose) = 1\,.$$

Extending discontinuity

Theorem

There exists a POMDP for each of the following combinations.

Example	*Continuity*		
	Value	*Weak strategy*	*Strong strategy*
#1	*Yes*	*Yes*	*No*
#2	*No*	*Yes*	*No*
#3	*No*	*No*	*No*

Remarks:

- All continuities are different
- The exact relationship between the continuity concepts is not fully characterized.

Characterizing continuity of POMDPs

Theorem (Mathematical characterization, open)

A POMDP is XXXX continuous if and only if ???

Theorem (Algorithmic impossibility)

*The problem of deciding whether
a given POMDP is XXXX continuous is undecidable.*