

Good-to-know mathematics to explore  
RL framework for classical,baysian or IB agents

ZKMathquant

Feb 2026

## Contents

<b>1</b>	<b>Mathematical Theory</b>	<b>3</b>
1.1	1. Formal Definitions	3
1.1.1	1.1 Policy-Dependent MDP	3
1.2	Credal Set	3
1.2.1	Infrabayesian Value Function	3
1.2.2	Credal Interval Update	3
1.2.3	2.2 Wasserstein Ball Update	4
1.3	Convergence Theorems	4
1.3.1	Policy Stability in Newcomb	4
1.3.2	Existence of Reflective Equilibrium	5
1.3.3	Robustness Under Misspecification	5
1.4	Comparison with Classical RL	5
1.4.1	Classical Q-Learning	5
1.4.2	Bayesian Q-Learning	6
1.4.3	Infrabayesian Q-Learning	6
1.5	Logical Dependence	6
1.5.1	Formal Model	6
1.5.2	Fixed Point Characterization	6
<b>2</b>	<b>Computational Complexity</b>	<b>6</b>
2.1	Worst-Case Value Computation	6
2.2	Bellman Backup	6
<b>3</b>	<b>Connection to Infra-Bayesianism</b>	<b>7</b>
3.1	Infra-Measures	7
3.2	Lower Previsions	7
3.3	Future Extensions	7
	<b>References</b>	<b>7</b>

## 1. Mathematical Theory

### 1.1. 1. Formal Definitions

#### 1.1.1 1.1 Policy-Dependent MDP

**Definition 1 (Policy-Dependent MDP):** A tuple  $M = (S, A, \Theta, T_\theta^\pi, R)$  where:

- $S$ : Finite state space
- $A$ : Finite action space
- $\Theta \subseteq \mathbb{R}^d$ : Compact parameter set
- $T_\theta^\pi : S \times A \times S \rightarrow [0, 1]$ : transition kernel depending on policy  $\pi$  and parameter  $\theta$
- $R : S \times A \rightarrow \mathbb{R}$ : bounded reward function

**Key property:**  $T_\theta^\pi(s'|s, a)$  depends on the agent's policy  $\pi$ , not just the state-action pair.

---

### 1.2. Credal Set

**Definition 2 (Credal Set):** A convex, closed set of probability distributions:  $\Theta_t = \{\theta \in \Theta : \text{constraints satisfied}\}$  (cf. [Kosoy n.d.])

In our implementation:

- 1D case:  $\Theta_t = [\theta_{\text{lower}}, \theta_{\text{upper}}]$
  - N-D case:  $\Theta_t = [\theta_{1,l}, \theta_{1,u}] \times \dots \times [\theta_{n,l}, \theta_{n,u}]$
- 

#### 1.2.1 Infrabayesian Value Function

**Definition 3 (IB Q-Function):** Utilizing the Maximin Expected Utility (MEU) framework (Gilboa & Schmeidler, 1989), the value is the worst-case expectation over the credal set:

$$Q_t^{IB}(s, a) = \min_{\theta \in \Theta_t} \mathbb{E}_\theta[R(s, a) + \gamma V_t(s') \mid s, a, \pi]$$

**Policy Selection:** The agent selects the action that maximizes this lower bound:

$$\pi_t(s) = \arg \max_a Q_t^{IB}(s, a)$$

---

## 2. Update Rules

### 1.2.2 Credal Interval Update

**Concentration Bound (Hoeffding):** To update the set  $\Theta_t$ , we use concentration bounds. Given  $n$  observations with empirical mean  $\hat{p}$ :

$$\epsilon_n = \sqrt{\frac{\log(2/\delta)}{2n}}$$

The updated credal set is:

$$\Theta_t = [\max(0, \hat{p} - \epsilon_n), \min(1, \hat{p} + \epsilon_n)]$$

**Theorem 1.1 (Credal Convergence).** *With probability  $\geq 1 - \delta$ :  $|\Theta_t| \rightarrow 0$  as  $t \rightarrow \infty$*

*Proof.* By the Hoeffding inequality, for  $n$  independent observations of a random variable  $X \in [0, 1]$  with true mean  $p$  and empirical mean  $\hat{p}$ :

$$P(|\hat{p} - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Setting the right-hand side to  $\delta$ , we solve for the confidence radius  $\epsilon_n$ :

$$\epsilon_n = \sqrt{\frac{\log(2/\delta)}{2n}}$$

The credal interval is defined as  $\Theta_t = [\hat{p} - \epsilon_n, \hat{p} + \epsilon_n]$ . The diameter of this set is:

$$|\Theta_t| = (\hat{p} + \epsilon_n) - (\hat{p} - \epsilon_n) = 2\epsilon_n = 2\sqrt{\frac{\log(2/\delta)}{2n}}$$

As  $t \rightarrow \infty$ ,  $n \rightarrow \infty$ , which implies  $\epsilon_n \rightarrow 0$ . Thus,  $|\Theta_t| \rightarrow 0$ . Since  $\hat{p}$  converges to  $p$  almost surely by the Strong Law of Large Numbers,  $\Theta_t$  collapses to the singleton  $\{p\}$   $\square$

---

### 1.2.3 2.2 Wasserstein Ball Update

Alternatively, uncertainty can be modeled via a Wasserstein ball [Iyengar 2005; Nilim and El Ghaoui 2005]

**Definition 4 (Wasserstein Ball):**  $\mathcal{W}_\epsilon(P) = \{Q : W_1(P, Q) \leq \epsilon\}$ , where  $W_1$  is the 1-Wasserstein distance.

The worst-case expectation simplifies under the Lipschitz property:

$$\mathbb{E}_{\text{worst}}[V] = \mathbb{E}_P[V] - \epsilon \cdot \|V\|_{\text{Lip}}$$

The radius shrinks at a rate of  $\epsilon_t = O(1/\sqrt{t})$ .

---

## 1.3. Convergence Theorems

### 1.3.1 Policy Stability in Newcomb

**Theorem 1.2** (One-Boxing Convergence). *If the predictor's accuracy  $\theta_{\min} > 0.5$ , then for sufficiently small  $|\Theta_t|$ , the IB agent converges to  $\pi_t = \text{one-box}$ .*

*Proof.* In the Newcomb problem, let  $u_1$  be the utility of one-boxing and  $u_2$  be the utility of two-boxing. The IB agent selects the action  $a$  maximizing the lower prevision  $\underline{E}[R|a]$ .

The payoffs are:

One-boxing:  $10^6$  if predicted ( $P$ ), 0 if not.

Two-boxing:  $10^6 + 10^3$  if predicted ( $P$ ),  $10^3$  if not.

Calculating the worst-case (minimum) expectations over  $\Theta_t = [\theta_{\min}, \theta_{\max}]$ :

$$\underline{E}[R|\text{one-box}] = \min_{\theta \in \Theta_t} (\theta \cdot 10^6 + (1 - \theta) \cdot 0) = \theta_{\min} \cdot 10^6$$

$$\underline{E}[R|\text{two-box}] = \min_{\theta \in \Theta_t} (\theta \cdot 10^3 + (1 - \theta) \cdot (10^6 + 10^3))$$

The minimum for two-boxing occurs at the highest probability of the predictor failing to predict the action (since the "big prize" is predicated on the predictor being correct). Thus:

$$\underline{E}[R|\text{two-box}] = 10^3 + (1 - \theta_{\max}) \cdot 10^6$$

The agent chooses one-box if  $\underline{E}[R|\text{one-box}] > \underline{E}[R|\text{two-box}]$ :

$$\theta_{\min} \cdot 10^6 > 10^3 + (1 - \theta_{\max}) \cdot 10^6$$

Dividing by  $10^6$ :  $\theta_{\min} > 0.001 + 1 - \theta_{\max} \implies \theta_{\min} + \theta_{\max} > 1.001$ .

Since  $|\Theta_t| \rightarrow 0$ ,  $\theta_{\min}$  and  $\theta_{\max}$  both approach the true parameter  $\theta^*$ . If  $\theta^* > 0.5005$ , the inequality is satisfied for large  $t$   $\square$

---

### 1.3.2 Existence of Reflective Equilibrium

**Definition 5 (Reflective Equilibrium):** A policy  $\pi^*$  is a reflective equilibrium if:

$$\pi^* \in \arg \max_{\pi} \min_{\theta \in \Theta} \mathbb{E}_{\theta} [V \mid \pi, P(\pi)]$$

**Theorem 1.3 (Existence).** *Under compactness of  $\Pi$  and continuity of  $T$ , a reflective equilibrium  $\pi^*$  exists.*

*Proof.* We define the equilibrium as a fixed point of the best-response correspondence  $\mathcal{B} : \Pi \rightarrow \Pi$ .

$$\mathcal{B}(\pi) = \arg \max_{\pi' \in \Pi} \left( \min_{\theta \in \Theta} \mathbb{E}_{\theta} [V \mid \pi', P(\pi)] \right)$$

The policy space  $\Pi$  is a probability simplex, which is a non-empty, compact, convex subset of a Euclidean space.

The function  $g(\pi', \pi) = \min_{\theta \in \Theta} \mathbb{E}_{\theta} [V \mid \pi', P(\pi)]$  is continuous in  $\pi$  and  $\pi'$  by the Maximum Theorem, given that the transition kernel  $T$  is continuous and  $\Theta$  is compact.

The correspondence  $\mathcal{B}(\pi)$  is upper hemi-continuous and, because the objective is linear/concave in  $\pi'$ , the values are convex sets.

By Kakutani's Fixed-Point Theorem, any such correspondence from a compact convex set to itself has at least one fixed point  $\pi^* \in \mathcal{B}(\pi^*)$   $\square$

### 1.3.3 Robustness Under Misspecification

[Iyengar 2005; Nilim and El Ghaoui 2005]

**Theorem 1.4 (Robust Performance).** *If the true parameter  $\theta^* \notin \Theta_t$  but  $\text{dist}(\theta^*, \Theta_t) \leq \delta$ , then:*

$$|V^{IB}(\pi_t) - V^*(\pi^*)| \leq C \cdot \delta$$

where  $C$  depends on reward scale and the Lipschitz constant of the value function.

**Interpretation:** IB agents degrade gracefully under misspecification.

*Proof.* The IB value function  $V^{IB}$  is the lower envelope of a family of linear functions (expectations). By the properties of the minimum of Lipschitz continuous functions, the operator  $\min_{\theta \in \Theta} E_{\theta}[V]$  is itself Lipschitz with respect to the Hausdorff distance between parameter sets. Let  $f(\theta) = E_{\theta}[R + \gamma V]$ . Since  $R$  is bounded and the transition  $T$  is linear in  $\theta$ ,  $f$  has a Lipschitz constant  $L$ .

$$| \min_{\theta \in \Theta_t} f(\theta) - f(\theta^*) | \leq L \cdot \inf_{\theta \in \Theta_t} \|\theta - \theta^*\| = L \cdot \delta$$

Defining  $C$  to account for the geometric series of the discount factor  $\gamma$ , the total value error is bounded by  $C\delta$ , where  $C = \frac{L}{1-\gamma}$ . This demonstrates the "graceful degradation" of Infrabayesian agents compared to the potential "brittleness" of point-estimate Bayesian agents  $\square$

## 1.4. Comparison with Classical RL

### 1.4.1 Classical Q-Learning

**Update:**  $Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$

**Assumption:** Single true environment model.

**Failure mode:** In policy-dependent environments, exploration causes the predictor to change, violating the stationarity assumption.

### 1.4.2 Bayesian Q-Learning

**Belief:**  $P(\theta | D) \propto P(D | \theta)P(\theta)$

**Action selection:** Thompson sampling or posterior mean.

**Failure mode:** Converges to a point estimate and loses robustness. In Newcomb, converges to two-boxing.

---

### 1.4.3 Infrabayesian Q-Learning

**Belief:** Credal set  $\Theta_t$  (set of distributions)

**Action selection:** Worst-case optimization [Kosoy n.d.]

**Success:** Maintains robustness and converges to one-boxing in Newcomb.

---

## 1.5. Logical Dependence

### 1.5.1 Formal Model

**Definition 5 (Logical Predictor):**  $P : \Pi \rightarrow A$

where  $\Pi$  is the space of policies.

[Garraabrant et al. 2016]

**Key property:** Predictor inspects policy representation, not just samples actions.

**Implementation:** `predicted_action = P( $\pi$ .greedy_action())`

This creates logical dependence:  $T_{\theta}^{\pi}(s'|s, a)$  depends on  $\pi$ .

This creates logical dependence:  $T_{\theta}^{\pi}(s'|s, a)$  depends on  $\pi$ .

---

### 1.5.2 Fixed Point Characterization

**Definition 6 (Reflective Equilibrium):**

A policy  $\pi^*$  is a reflective equilibrium if:  $\pi^* \in \arg \max_{\pi} \min_{\theta \in \Theta} \mathbb{E}_{\theta} [V | \pi, P(\pi)]$

**Theorem 4 (Existence):** Under compactness and continuity, reflective equilibrium exists. [Garraabrant et al. 2016] **Proof sketch:** Kakutani fixed-point theorem.  $\square$

---

## 2. Computational Complexity

### 2.1. Worst-Case Value Computation

[iyengar; nilim]

**1D Credal Interval:**

- Evaluate at endpoints:  $O(1)$

**N-D Credal Rectangle:**

- Evaluate at  $2^N$  vertices:  $O(2^N)$

**Wasserstein Ball:**

- Closed-form for discrete:  $O(|S|)$
  - General case: LP with  $O(|S|^3)$  complexity
- 

### 2.2. Bellman Backup

**Classical:**  $O(|S||A|)$

**IB with 1D credal:**  $O(|S||A|)$

**IB with Wasserstein:**  $O(|S|^2|A|)$

**Scalability:** Tractable for tabular settings, requires approximation for large state spaces.

---

### 3. Connection to Infra-Bayesianism

#### 3.1. Infra-Measures

**Full infra-Bayes:** Convex sets of semimeasures (mass  $\leq 1$ )

**Our framework:** Convex sets of probability measures (mass = 1)

**Relationship:** Our framework is a special case (normalized infra-distributions).

---

#### 3.2. Lower Previsions

**Infra-Bayes:** Lower expectation functional

**Our framework:** min over credal set

**Equivalence:** For finite credal sets, these coincide.

---

#### 3.3. Future Extensions

To reach full infra-Bayes:

- Allow semimeasures (unnormalized)
- Infinite credal sets (via constraints)
- Non-additive uncertainty (Choquet integration)
- Logical induction dynamics

### References

- [1] Garrabrant, Scott et al. (2016). *Logical Induction*. arXiv preprint (cit. on p. 6).
- [2] Iyengar, Garud N. (2005). “Robust Dynamic Programming”. In: *Mathematics of Operations Research* (cit. on pp. 4, 5).
- [3] Kosoy, Vanessa (n.d.). *Infra-Bayesian Decision Theory*. LessWrong sequence (cit. on pp. 3, 6).
- [4] Nilim, Arnab and Laurent El Ghaoui (2005). “Robust Control of Markov Decision Processes”. In: *Operations Research* (cit. on pp. 4, 5).

---

:: YOU HAVE REACHED THE END OF THE DOCUMENT, THANK YOU FOR READING ::

---