A Dual Process Theory of Optimistic Cognition

Peter Sunehag (peter.sunehag@anu.edu.au) and Marcus Hutter, Research School of Computer Science, Australian National University.

1 Abstract

Optimism is a prevalent bias in human cognition including variations like self-serving beliefs, illusions of control and overly positive views of one's own future. Further, optimism has been linked with both success and happiness. In fact, it has been described as a part of human mental well-being which has otherwise been assumed to be about being connected to reality. In reality, only people suffering from depression are realistic. Here we study a formalization of optimism within a dual process framework and investigate its usefulness beyond human needs in a way that also applies to artificial reinforcement learning agents. Optimism enables systematic exploration which is essential in an (partially) unknown world. The key property of an optimistic hypothesis is that if it is not contradicted when one acts greedily with respect to it, then one is well rewarded even if it is wrong.

2 General Reinforcement Learning

- An agent interacts with an environment in cycles
- The agent
  performs actions a_t from a finite set $\mathcal{A}$
  receives observations o_t from a finite set $\mathcal{O}$
  and rewards r_t from a finite set $\mathcal{R} \subset [0,1]$
- The result is a history $h_t := o_1 r_1 a_1, \ldots, o_t r_t \in \mathcal{H}$
- The value function: $V^\pi_\nu(h_{t-1}) := \mathbf{E}_{\nu(\cdot|\pi,h_{t-1})} \sum_{i=t}^\infty \gamma^{i-t} r_i$
- The optimal value function: $V^*_\nu(h_{t-1}) := \max_\pi V^\pi_\nu(h_{t-1})$

3 Rational and Optimistic Agents

- Given
  a countable class of environments $\mathcal{M}$
  strictly positive prior weights w_ν for all $\nu \in \mathcal{M}$
  we define the a-priori environment $\xi(\cdot) = \sum w_\nu \nu(\cdot)$
  A rational agent follows a policy

$$\pi^* \in \arg\max_\pi V^\pi_\xi(\epsilon).$$

- An optimist follows

$$\pi^\circ \in \arg\max_\pi \max_{\xi \in \Xi} V^\pi_\xi(\epsilon)$$

  for a finite set of beliefs (environments) Ξ.
- If Ξ has size one, the optimist is a rational agent.


- To achieve high rewards, it is not enough to predict well what is going to happen for a lazy unexplorative policy.
- A Bayesian RL agent sometimes fails, even with finite environment classes, to achieve asymptotic optimality (having an expected return arbitrarily close to optimal for the situation the agent is in).

4 Decision Functions

- A decision function $f : \mathbb{M} \to \mathcal{A}$ ($\mathbb{M}$ is the set of finite classes of environments) only depending on a class of environments $\mathcal{M}$.
- The decision function is independent of the history
- However, the class $\mathcal{M}$ fed to the decision function introduces an indirect dependence
- f is strictly rational for the class $\mathcal{M}$ if there are $\omega_\nu \geq 0$, $\nu \in \mathcal{M}, \sum_{\nu\in\mathcal{M}} w_\nu = 1$ such that $a = \pi(\varepsilon)$ for a policy

$$\pi \in \arg\max_\pi \sum_{\nu\in\mathcal{M}} \omega_\nu V^\pi_\nu \tag{1}$$

- A special case is when $|\mathcal{M}| = 1$ and (1) becomes

$$\pi \in \arg\max_\pi V^\pi_\nu$$

  where ν is the environment in $\mathcal{M}$.
- f is optimistic if $f(\mathcal{M}) = a$ implies that $a = \pi(\epsilon)$ for an optimistic policy π, i.e. for

$$\pi \in \arg\max_\pi \max_{\nu\in\mathcal{M}} V^\pi_\nu. \tag{2}$$

5 Hypothesis-Generating Functions

Given a decision function, what remains to create a complete agent is a hypothesis generating function $\Gamma(h) = \mathcal{M}$ that for any history $h \in \mathcal{H}$ produces a class of environments $\mathcal{M}$.

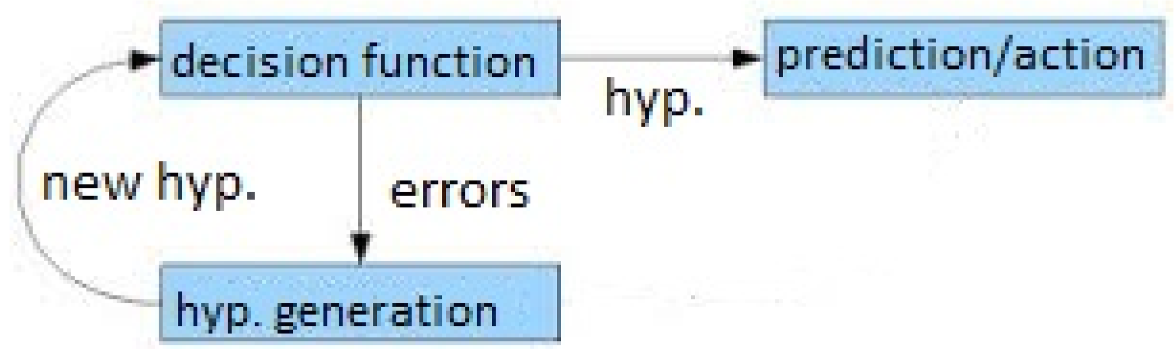
- A special case is defined by combining the initial $\Gamma(\epsilon) = \mathcal{M}_0$ with an update function $\psi(\mathcal{M}_{t-1}, h_t) = \mathcal{M}_t$.
- A hypothesis generating function satisfies Epicurus principle if the update function is such that it might add new environments in any way while removing environments if a hypothesis is implausible (likely to be false) in light of the observations made.
- Given $0 < \varepsilon < 1$, we define the number of ε-inconfidence points in the history h to be

$$n(h,\varepsilon) := |\{i \leq l(h) \mid \max_{\nu_1,\nu_2\in\Gamma(h_i)} |V^{\pi^*}_{\nu_1} - V^{\pi^*}_{\nu_2}| > \varepsilon\}|$$

  where $\pi^* := \arg\max_\pi \max_{\nu\in\Gamma(h_t)} V^\pi_\nu$. In the $\gamma = 0$ case studied here, we can equivalently use $a^* := \arg\max_a \max_{\nu\in\Gamma(h_t)} V^a_\nu$ instead of π^*.

- We define a hypothesis generating function from a countable enumerated class $\mathcal{M}$ based on a budget function for ε-inconfidence that is increasing and unbounded.
- When the number of ε-inconfidence points is below budget we introduce the next environment in the class.
- This form of hypothesis generating function enables bounds on the number of errors made by optimistic agents and it implements the intuition that the agent should not introduce more environments when the existing ones are very contradictory.

6 Agents in a Dual Process Framework



An agent, i.e. a function from histories to actions, is defined from a hypothesis generating function Γ and a decision function f by choosing action $a = f(\Gamma(h))$ after seeing history h.

Example 1. *Suppose that $\mathcal{M}$ is a finite class of deterministic environments and let $\Gamma(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}$ consistent with $h\}$. If we combine Γ with the optimistic decision function we have defined the optimistic agents for finite classes of deterministic environments. We here extend the analysis to infinite classes by letting $\Gamma(h_t)$ contain new environments that were not in $\Gamma(h_{t-1})$.*

Example 2. *The Model Based Interval Estimation (MBIE) method for Markov Decision Processes (MDPs) defines $\Gamma(h)$ as a set of MDPs (for a given state space) with transition probabilities in confidence intervals calculated from h. This is combined with the optimistic decision function.*

Given $0 \leq \varepsilon < 1$, we define the number of ε-errors in history h to be

$$m(h,\varepsilon) = |\{i \leq \ell(h) \mid V^{a_i}_\mu(h_i) < V^*_\mu(h_i) - \varepsilon\}|$$

where μ is the true environment, $\ell(h)$ is the length of h, a_i is the i:th action and $V^*_\mu(h) = \arg\max_a V^a_\mu(h)$.

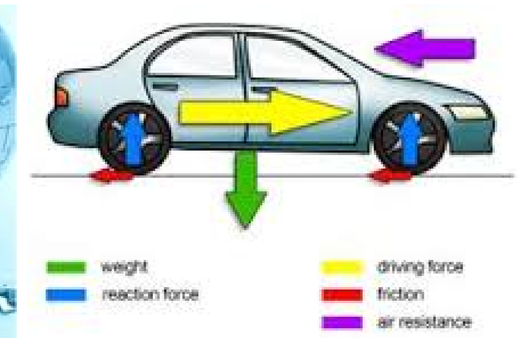
Theorem 3 (Finite error bound)**.** *Following π° (optimistic decision function with hypothesis-generating function that excludes inconsistent environments but does not add) for $\mu \in \mathcal{M}$ (finite deterministic class),*

$$V^{\pi^\circ}_\mu(h_t) \geq \max_{\pi\in\Pi} V^\pi_\mu(h_t) - \varepsilon,\ 0 < \varepsilon < 1/(1-\gamma)$$

for all but at most $K^{\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}} \leq |\mathcal{M}-1|^{\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}}$ time steps t where K is the number of times that some environment is contradicted.

- Given
  a countable class of deterministic environments $\mathcal{M}$
  Γ excluding contradicted environments from finite initial class
  a budget function $N : \mathbb{N} \to \mathbb{N}$, accuracy $\varepsilon = 0$
  π° is defined by combining Γ with an optimistic decision function.
- The number of 0-errors $m(h_t, 0)$ is at most $n(h_t, 0) + C$ for some constant $C > 0$ (dependent on choice of budget function N but not on t) that is the time at which the truth is included.
- $\forall i \in \mathbb{N}$ there is $t_i \in \mathbb{N}$ such that $t_i < t_{i+1}$ and $n(h_{t_i}, 0) < N(t_i)$.

7 Combining Laws into Environments

- Observations of the form of a feature vector
- $o = \vec{x} = (x_j)_{j=1}^m \in \mathcal{O} = \times_{j=1}^m \mathcal{O}_j$, $\mathcal{O}_\perp = \times_{j=1}^m (\mathcal{O}_j \cup \{\perp\})$
- $\perp$ means that there is no prediction for this feature.
- A law is a function $\tau : \mathcal{H} \times \mathcal{A} \to \mathcal{O}_\perp$.
- A set of laws $\tilde{\mathcal{T}}$ is complete and coherent if for each h, a and j exactly one prediction is made by laws in the class.
- Let $\mathcal{C}(\mathcal{T})$ denote the complete and coherent subsets of $\mathcal{T}$.
- The class of environments generated by $\mathcal{T}$ is

$$\mathcal{M}(\mathcal{T}) := \{\nu(\tilde{\mathcal{T}}) \mid \tilde{\mathcal{T}} \in \mathcal{C}(\mathcal{T})\}.$$

Theorem 4 (Finite error bound when using laws)**.** *Suppose that $\mathcal{T}$ is a finite class of deterministic laws and let $\Gamma(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}(\{\tau \mid \tau \in \mathcal{T}$ consistent with $h\})\}$. We define $\bar{\pi}$ by combining g with the optimistic decision function. Following $\bar{\pi}$ for a finite class of deterministic laws $\mathcal{T}$ in an environment $\mu \in \mathcal{M}(\mathcal{T})$, we have for any $0 < \varepsilon < \frac{1}{1-\gamma}$ that*

$$V^{\bar{\pi}}_\mu(h_t) \geq \max_\pi V^\pi_\mu(h_t) - \varepsilon \tag{5}$$

for all but at most $|\mathcal{T} - l|^{\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}}$ time steps t where l is the minimum number of laws from $\mathcal{T}$ needed to define a complete environment.

Example 3 (Deterministic laws for fixed vector)**.** *Consider an environment with a constant binary feature vector of length m. There are 2^m such environments. Every such environment can be defined by combining m out of a class of $2m$ laws. Each law says what the value of one of the features is, one law for 0 and one for 1. In this example, a coherent set of laws is simply one feature for each coefficient. The generated environment is the constant vector defined by that vector and the set of all the generated environments is the full set of 2^m environments.*

8 Conclusions

- Optimism enables sufficient exploration for short-sighted agents to achieve optimality. Strict rationality fails to guarantee this.
- Viewing environments as combinations of laws can improve bounds exponentially
- Outlook: Milder form better related to human's with Reward Modulated-Inference as in reward-modulated spike-timing plasticity

Paper1: *A Dual Process Theory of Optimistic Cognition*,
Peter Sunehag and Marcus Hutter, CogSci'2014
Paper2: *Optimistic Agents are Asymptotically Optimal*,
Peter Sunehag and Marcus Hutter, AusAI'2012
Paper3: *Optimistic AIXI*,
Peter Sunehag and Marcus Hutter, AGI'2012
Paper4: *Axioms for Rational Reinforcement Learning*,
Peter Sunehag and Marcus Hutter, ALT'2011

Australian National University

ANU College of
Engineering & Computer Science