Reasoning for Reinforcement Learning

Silviu Pitis • Georgia Institute of Technology • silviu.pitis@gmail.com • silviupitis.com

Key Idea: One can explicitly reason to alternative representations of a value function by composing more specific measures (in the form of general value functions) hierarchically, according to known rules of reasoning. This would allow reasoning-enhanced agents to explicitly justify their value judgments and actions, reason to more accurate estimates of value, and apply consistency-based learning.

3-Part Framework for Reasoning

Implicit knowledge represented by black box functions

→ Comparable to human intuition

Measures related by rules of reasoning

1

Learned Primitives

Table 1: Primitive measures (distances & densities)

Name	Notation	Recursive definition
Time similarity	$\Gamma(s_t, g)$	$\max_a \mathbb{E}^a[\gamma \cdot \Gamma'(s_{t+1}, g)]$, where $\Gamma'(s, g) := \Gamma(s, g)$ unless $s = g$, in which case $\Gamma'(g, g) := 1$.
Value distance†	$\mathtt{vdist}(s_t, g)$	$\mathbb{E}^{\pi(s_t)}[r_t + \gamma \cdot \mathtt{vdist}'(s_{t+1}, g)]$.
Event distance†*	$\mathtt{xdist}(s_t, x, g)$	$I(s_t, x) + \mathbb{E}^{\pi(s_t)}[\gamma \cdot \mathtt{xdist}'(s_{t+1}, x, g)]$.
Event density††*	$\mathtt{maxsr}(s_t, x)$	$I(s_t, x) + \max_a \mathbb{E}^a[\gamma \cdot \mathtt{maxsr}(s_{t+1}, g)]$.
Value	$V(s_t)$	$\max_a \mathbb{E}^a[r_t + \gamma \cdot V(s_{t+1})]$.

* $I(s, x)$ is 1 if $s = x$ and 0 otherwise. To generalize in the same way as [1] generalizes the SR to successor features, define x as a feature index (rather than a state) and redefine $I(s, x)$ as the xth feature of s.

† In each case, $\pi(s)$ is greedy with respect to $\mathbb{E}[\Gamma(s_{t+1}, g)]$ and $f' := f$ unless $s_{t+1} = g$, in which case $f' := 0$.

†† Technically, this should be *maximum* event density. The sr in maxsr stands for successor representation [5].

Value functions of derivative MDPs

→ "General Value Functions" (GVFs)

→ Can be learned by Horde or with UVFAs

(Sutton et al. 2011, Schaul et al. 2015)

Allow for hierarchical decomposition of value function

2

Rules of Reasoning

Table 2: Landmark and factor-based rules of reasoning

Landmark lower bound, V	$V(s)$	$\geq$	$\mathtt{vdist}(s, \ell) + \Gamma(s, \ell) \cdot V(\ell)$
Landmark lower bound, Γ	$\Gamma(s, g)$	$\geq$	$\Gamma(s, \ell) \cdot \Gamma(\ell, g)$
Landmark upper bound, V	$V(s)$	$\leq$	$(V(l) - \mathtt{vdist}(l, s))/\Gamma(l, s)$
Landmark upper bound, xdist	$\mathtt{xdist}(s, x, g)$	$\leq$	$\mathtt{maxsr}(s, x) - \Gamma(s, g) \cdot \mathtt{maxsr}(g, x)$
Event-reward decomposition	$\mathtt{vdist}(s, g)$	$\approx$	$\sum_{x_i \in X \subseteq S} \mathtt{xdist}(s, x_i, g) \cdot r(x_i)$
Distance-density duality	$\mathtt{maxsr}(s, g)$	$=$	$I(s, g) + \Gamma(s, g) \cdot (1/(1 - \Gamma(g, g)))$
Distance-density lower bound	$V(s)$	$\geq$	$\mathtt{maxsr}(s, s) \cdot \mathtt{vdist}(s, s)$

Landmarks decompose measures temporally

→ Cf. interruptible options (Sutton et al. 1999)

Factors decompose values into sources of rewards

Composed xdist segments → successor representation

(Dayan 1993)

3

Metacognitive Controller

Algorithm Sketch 1 Recursive reasoning

```
function REASON_TO(prim, budget)
    if CONFIDENT(prim, budget) then
        return prim
    end if
    estimates ← {prim}
    for rule in GENERATE(prim, budget) do
        REASON_TO required primitives
        Compute estimate using rule and
            add it to estimates
    end for
    return RESOLVE(estimates, budget)
end function

function CONFIDENT(prim, budget):
    Returns true if agent sufficiently confident in
    prim given budget.

function GENERATE(prim, budget):
    Generates rules (e.g., landmark bounds) that
    might produce useful representations of prim.

function RESOLVE(estimates, budget):
    Returns single estimate of prim based on the set
    estimates (not necessarily a member thereof).
    Optionally invokes consistency-based learning.
```

Requires knowledge of primitive distributions

→ E.g., variance primitives + rules for composition

(Sobel 1982, Engel et al. 2005, Bellemare et al. 2017)

Requires generator of applicable reasoning rules, including, e.g., the relevant landmarks and factor sets

→ Modular; can be rule-based or ML-based

Benefits

Interpretation / Justification

Although primitives are black boxes, their decomposition is explicit

Allows for justification (even *ex post*) of actions:

"I took action 2 because it takes me in the direction of landmark A, which is a good place to be; further, the path to A has a reasonable chance of encountering event X, which has been very rewarding."

Reasoning → Better Representations

Some representations are better than others

→ partial supervision (verbal instruction) can be applied to segments

→ values of small segments easier to learn than long-run values

→ focused exploration leads to familiar landmarks

→ shifting rewards change values but not event distances

Consistency-based Learning

Multiple representations *should* be consistent

→ can optimize for consistency, toward more reliable estimates

→ can generalize partial supervision across measures

→ E.g., actor-critic methods; TD learning

Ultimate goal: all primitives *implicitly* satisfy the rules of reasoning

→ may be unachievable, so that explicit reasoning necessary; but even if achieved, explicit reasoning still useful for learning, justification

Extensions

State and Event Abstraction

Primitives and rules should ideally be reformulated with respect to:

→ sets of states (esp. for landmarks) (Li et al. 2006)

→ temporally extended events (esp. for factors)

Would entail a rich set of definition (=), negation (¬), inclusion (∈,⊆), exclusion (∉,⊈), preference (≻) and composition (∪,∩) rules

Autonomous Rule Acquisition

Rules might be acquired through explicit programming, via verbal interaction, or by evolutionary methods

Could agents learn rules of reasoning on their own?

An autonomous process might be inspired by the scientific method:

"First, we guess [a rule]; then we compute the consequences of the guess; and then we compare those computation results to nature, or experiment, or experience ... if [the rule] disagrees with experiment, it's wrong." -- Richard Feynman