Metareasoning for Planning Under Uncertainty

Christopher H. Lin*
University of Washington
Andrey Kolobov
Microsoft Research
Ece Kamar
Microsoft Research
Eric Horvitz
Microsoft Research

*This work performed while the author was an intern at Microsoft Research
TIME COST

STRATEGY A

STRATEGY B

ANALYZING WHETHER STRATEGY A OR B IS MORE EFFICIENT

THE REASON I AM SO INEFFICIENT

[https://xkcd.com/1445/]
AVAILABLE ACTIONS

Plan

Act

GOAL
Plan

Act

AVAILABLE ACTIONS

GOAL
Related Work

- **Type II rationality** [Good, 1971]
- Metalevel reflection for **one-shot decisions** [Horvitz 1987/89, Hansen and Zilberstein 2001]
- Guiding sequences of **single actions** in search [Russell and Wefald 1991, Burns et al 2013]
- Maximizing policy reward under **computational constraints** [Kolobov et al 2012]
- **Time allocation** [Zilberstein and Russell 1993/96]
- Optimizing **portfolios of planning strategies** [Dean et al 1995]
Metareasoning for Planning Under Uncertainty: Contributions

Formalization and complexity analysis

Fast algorithms for approximate metareasoning with no hyperparameters!
Base Stochastic Shortest Path (SSP)
Markov Decision Process (MDP)

\[ S: \text{States} \]
\[ A: \text{Actions} \]
\[ T(s, a, s'): \text{Transition Function} \]
\[ C(s, a, s'): \text{Cost Function} \]
\[ s_0: \text{Initial State} \]
\[ s_g: \text{Goal State} \]
S: States
A: Actions Contains a NOP action
T(s, a, s’): Transition Function
C(s, a, s’): Cost Function
s_0: Initial State
s_g: Goal State
NOP  Thinking/Planning Action
World doesn’t PAUSE!
Base SSP MDP

S: States
A: Actions
T(s, a, s’): Transition Function
C(s, a, s’): Cost Function
s₀: Initial State
s₉: Goal State
Black Box Online Planner

$X$: Internal State of the Planner (State of Mind)

$T(x, x')$: Transition Function between states of mind

$x_0$: Planner’s initial internal state
f : S,X → A: Map from Base State and state of mind to Base Level Action
Meta-MDP

$S$: States
$A$: Actions
$T(s, a, s')$: Transition Function
$C(s, a, s')$: Cost Function
$s_0$: Initial State
$s_g$: Goal State

$S^m$: $S \times X$
$A^m$: $A$
$T^m$: Restricted to two actions – NOP and $f(s, x)$
$C^m$: As you would expect
$s^m_0$: $(s_0, x_0)$
$s^m_g$: $(s_g, x)$

$X$: Internal State of the Planner
$T(x, x')$: Transition Function
$x_0$: Planner’s initial internal state
Theoretical Properties
Theorem 1. If

1) the base MDP is an SSP MDP, and
2) the planner halts on the base MDP with a proper policy,

then the Metareasoning MDP is an SSP MDP.
Theorem 2. Solving the Metareasoning MDP is at most polynomially harder than solving the base MDP in the size of the base MDP.
Theorem 3. The Metareasoning problem is P-complete under NC-reduction.
Challenges of Exact Metareasoning

Don’t have planner’s transition function

Infinite Regress
Algorithms for Metareasoning
Value of Computation =

\[ Q(s^m, f(s,x)) - Q(s^m, \text{NOP}) \]
IF VOC > 0

PLAN

ELSE

ACT
Value of Computation =

\[ Q(s^m, f(s,x)) - Q(s^m, \text{NOP}) \]

The value of taking the currently recommended action
Value of Computation =

\[ Q(s^m, f(s,x)) - Q(s^m, \text{NOP}) \]
Assumption 1
Metamyopic Assumption
[Russell and Wefald, 1991]

In any state, after the current step, the agent will never again think, and hence never change its policy

\[ Q(s^m, f(s, x)) - Q(s^m, \text{NOP}) \]
Assumption 2
The planner is BRTDP
BRTDP Maintains Two Bounds

Q-value Function

Upper Bound

Lower Bound
Q-value Function

Previous Upper Bound

$\Delta_1$

New Upper Bound

$\Delta_2$

New Lower Bound

Previous Lower Bound
Value of Computation =

\[ Q(s^m, f(s, x)) - Q(s^m, \text{NOP}) \]

The value of taking the currently recommended action

\[ O(K |A|^2) \]

The value of taking a NOP
Experiments
Baselines

![Graph showing the relationship between probability p for Prob Baseline, number of thinking cycles n for Think*Act Baseline, and average cost to the goal. The x-axis represents the number of thinking cycles ranging from 2 to 18, the y-axis represents the average cost ranging from 0 to 200, and the z-axis represents the probability ranging from 0.2 to 0.8.](image-url)
Baselines

Think*Act
Baselines

**Probability $p$ for Prob Baseline**

- $p = 0.2$ to $0.8$

**Number of Thinking Cycles $n$ for Think*Act Baseline**

- $n = 2$ to $18$

**Average Cost to the Goal**

- From 80 to 200
Baselines

NoInfoThink
Metareasoning for Planning Under Uncertainty: Conclusions

Formalization and complexity analysis

Fast algorithms for approximate metareasoning with no hyperparameters!