Metareasoning for Planning Under Uncertainty

Christopher H. Lin*

University of Washington

Andrey Kolobov

Microsoft Research

Ece Kamar

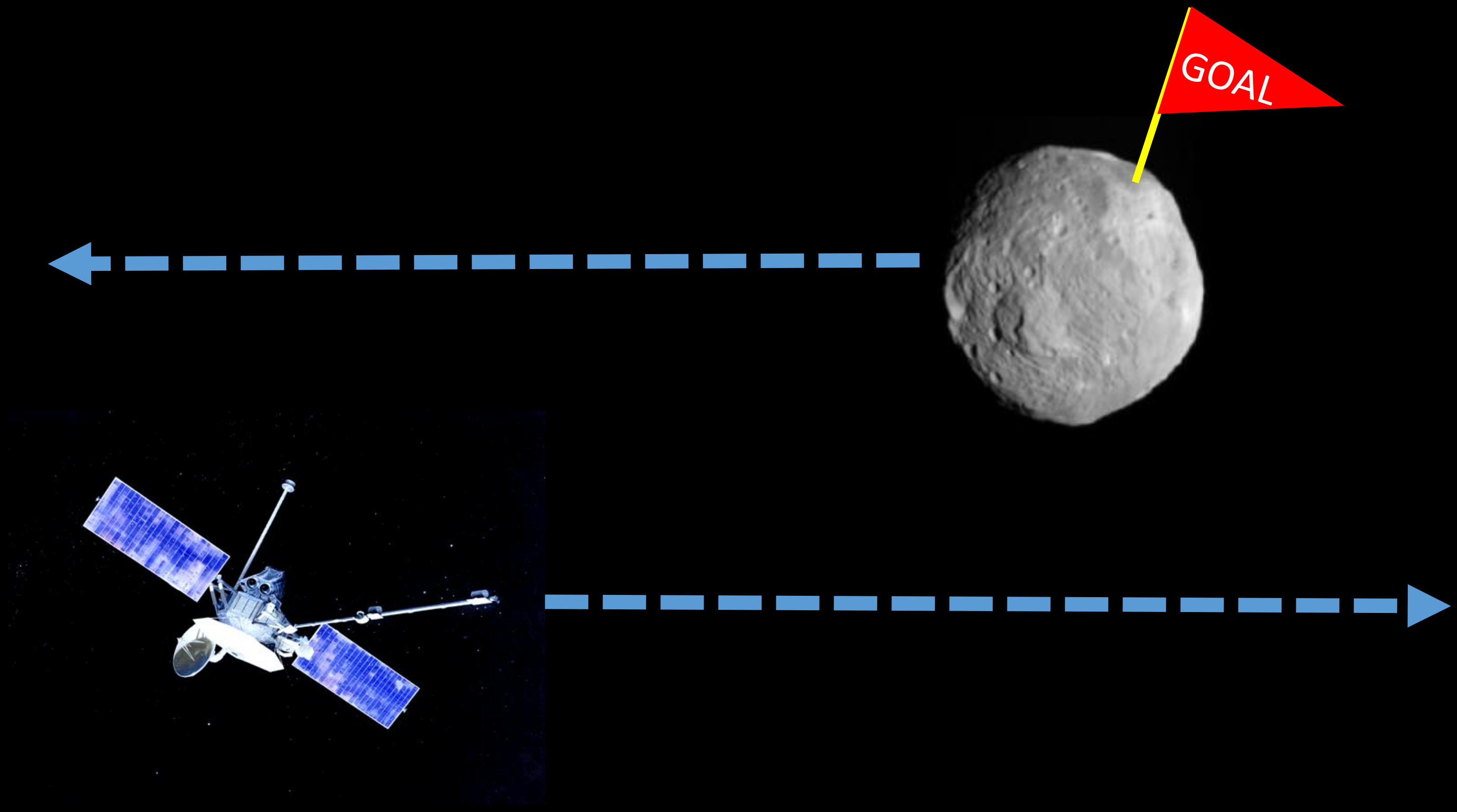
Microsoft Research

Eric Horvitz

Microsoft Research

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

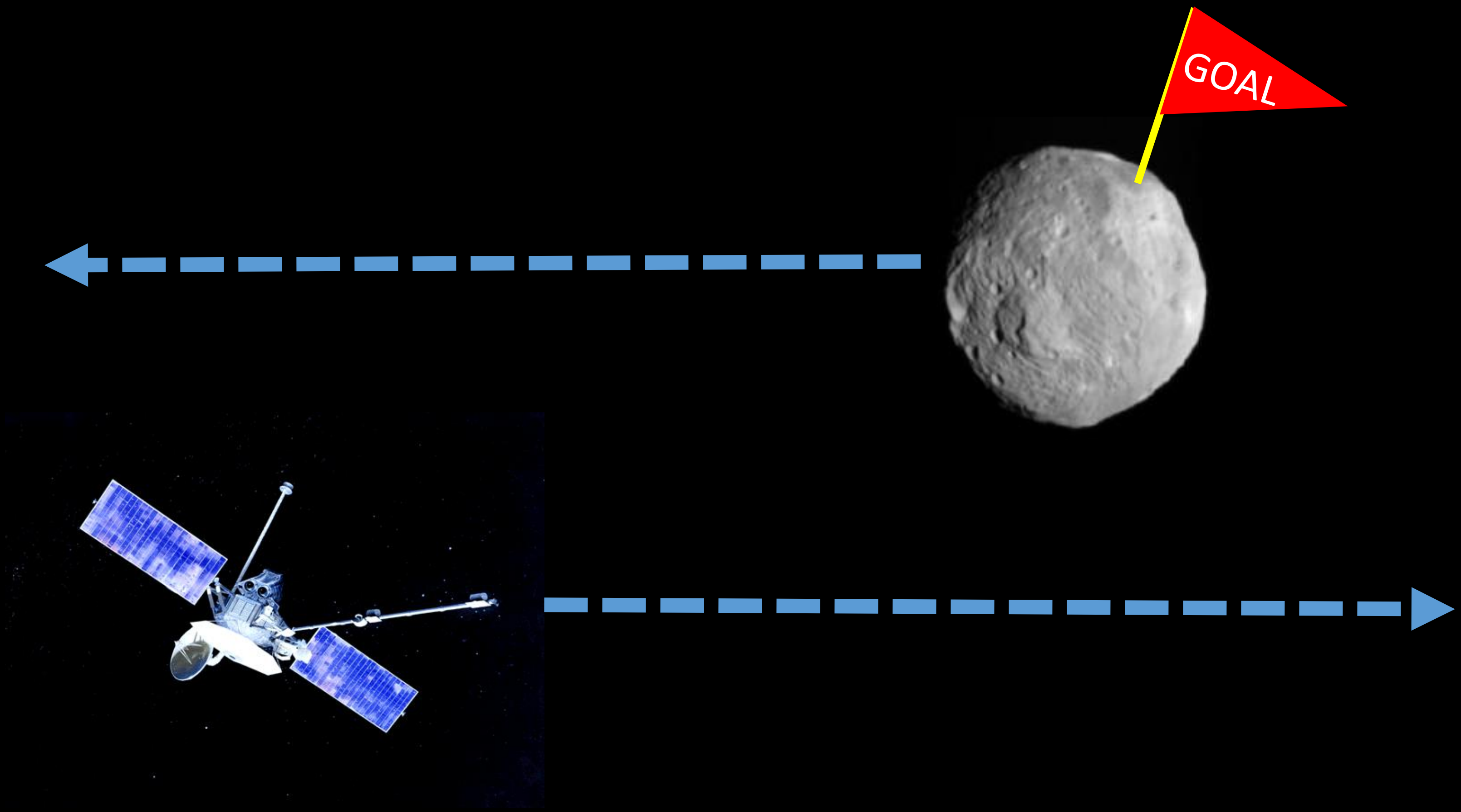
[https://xkcd.com/1445/]



AVAILABLE ACTIONS

Plan

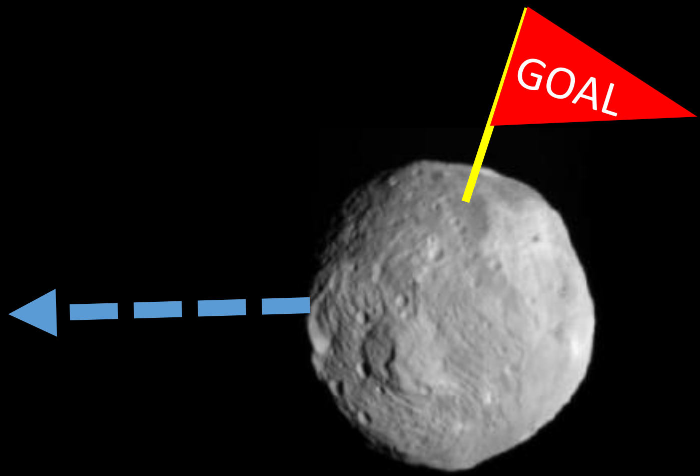
Act

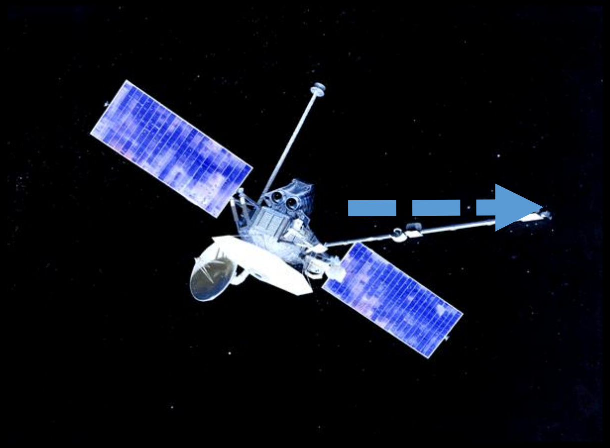


AVAILABLE ACTIONS

Plan

Act

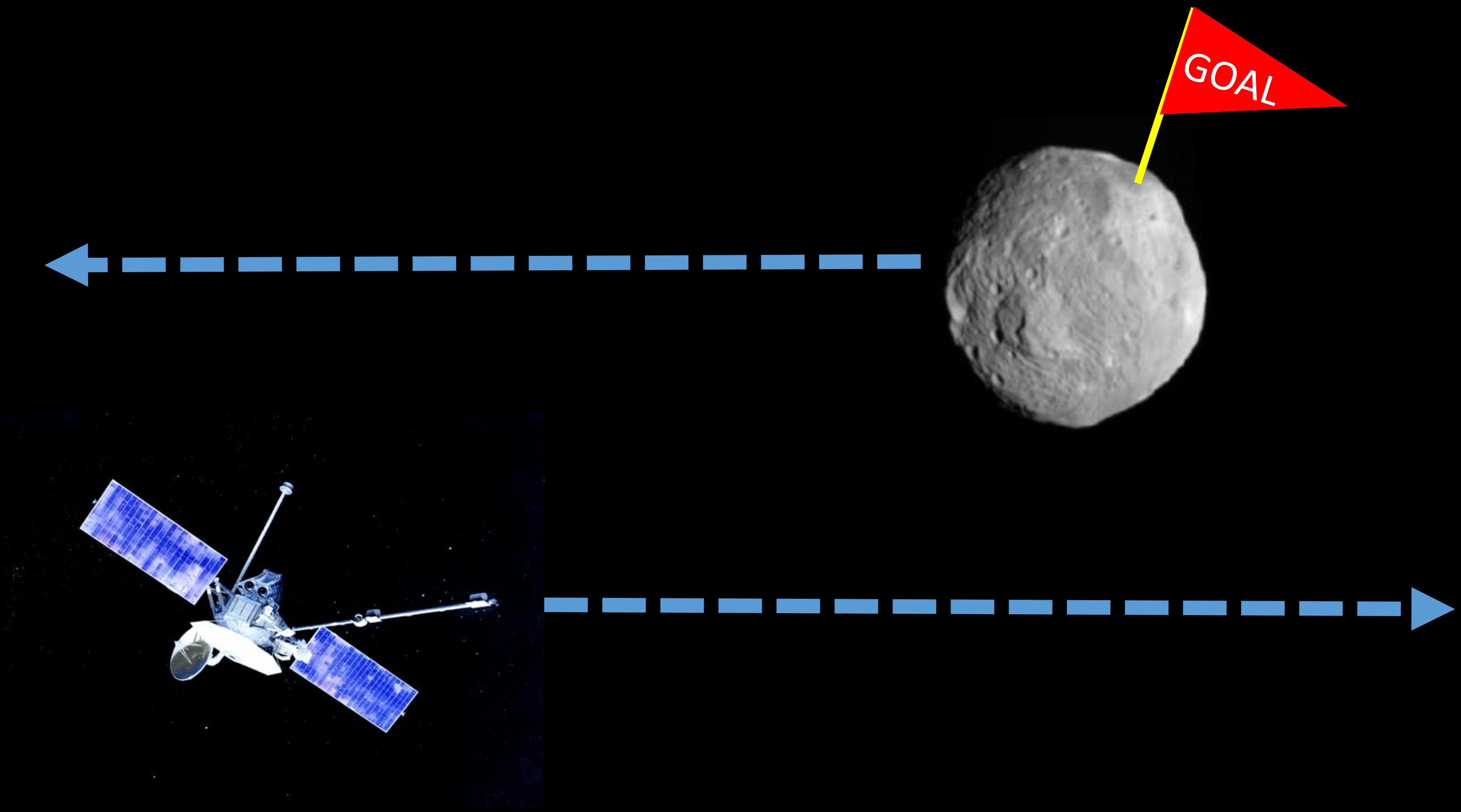




AVAILABLE ACTIONS

Plan

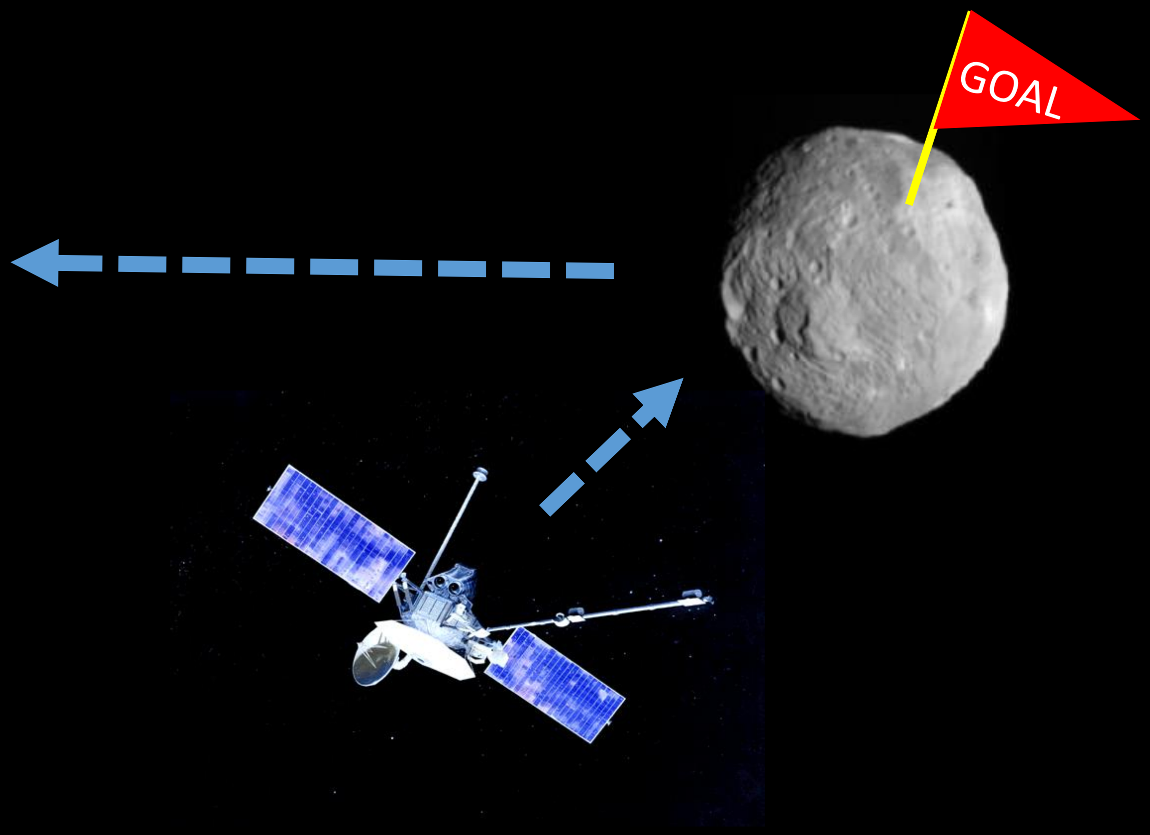
Act



AVAILABLE ACTIONS

Plan

Act



AVAILABLE ACTIONS

Plan

Act

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: States

A: Actions

T(s, a, s'): Transition Function

C(s, a, s'): Cost Function

s_0: Initial State

s_g: Goal State

Base SSP MDP

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

NOP Thinking/Planning Action

World doesn't PAUSE!

Base SSP MDP

S: States

A: Actions

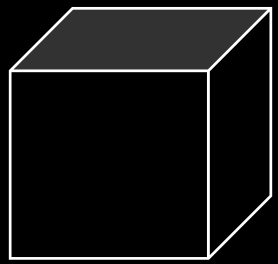
T(s, a, s’): Transition Function

C(s, a, s’): Cost Function

s_0: Initial State

s_g: Goal State

Black Box Online Planner



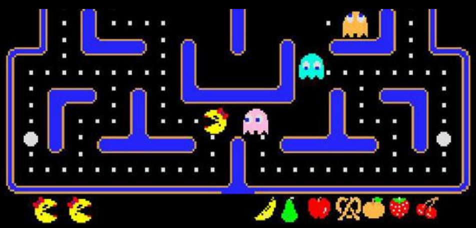
X: Internal State of the Planner (State of Mind)

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

x_0 : Planner's initial internal state

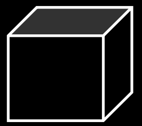
$f : S, X \rightarrow A$: Map from Base State and state of mind to Base Level Action

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

×



X: Internal State of the Planner
T(x, x’): Transition Function
x_0 : Planner’s initial internal state

S^m: S x X

A^m: A

T^m : Restricted to two actions – NOP and f(s,x)

C^m : As you would expect

s_0^m: (s_0, x_0)

s_g^m: (s_g, x)

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 =

$$Q(s^m, f(s,x)) - Q(s^m, NOP)$$

IF VOC > 0	PLAN
ELSE	ACT

Value of Computation =

$$Q(s^m, f(s,x)) - Q(s^m, NOP)$$

The value of taking the currently recommended action

Value of Computation =

$$Q(s^m, f(s,x)) - Q(s^m, NOP)$$

The value of
taking a NOP

Assumption 1
Metamyopic Assumption
[Russell and Wefald, 1991]

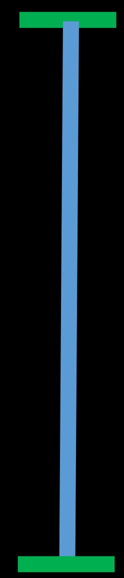
In any state, after the current step, the agent will never again think, and hence never change its policy

$$Q(s^m, f(s, x)) - Q(s^m, NOP)$$

Assumption 2

The planner is BRTDP

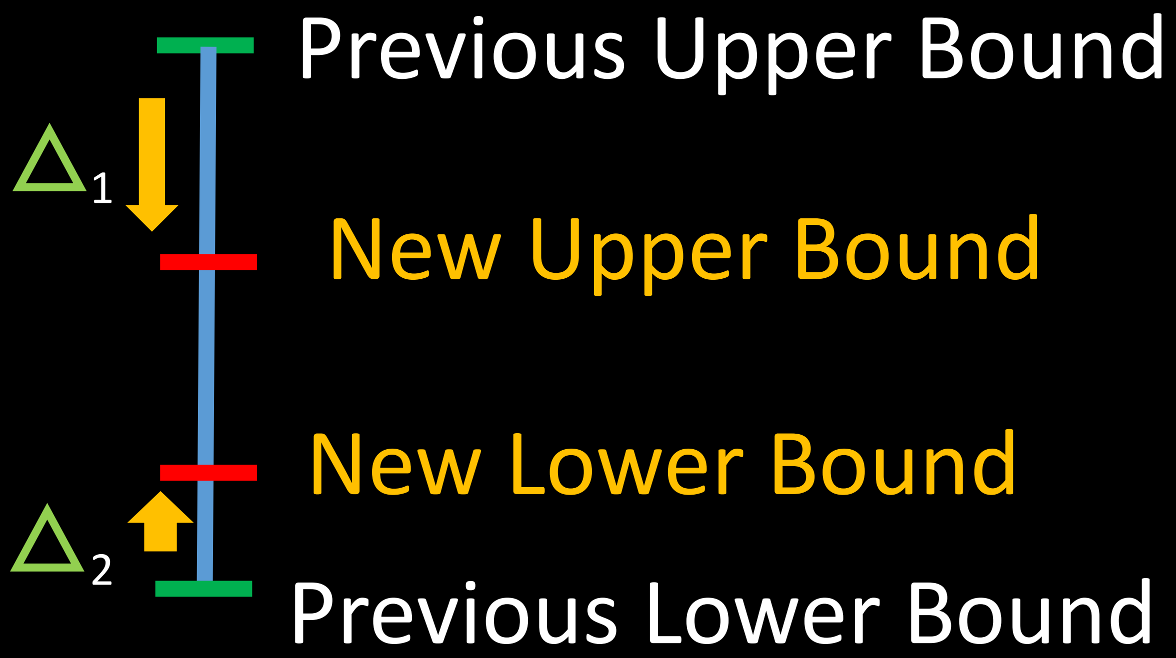
BRTDP Maintains Two Bounds



Upper Bound

Lower Bound

Q-value Function



Q-value Function

Previous Upper Bound

Q-value Function ~

Previous Lower Bound

Value of Computation =

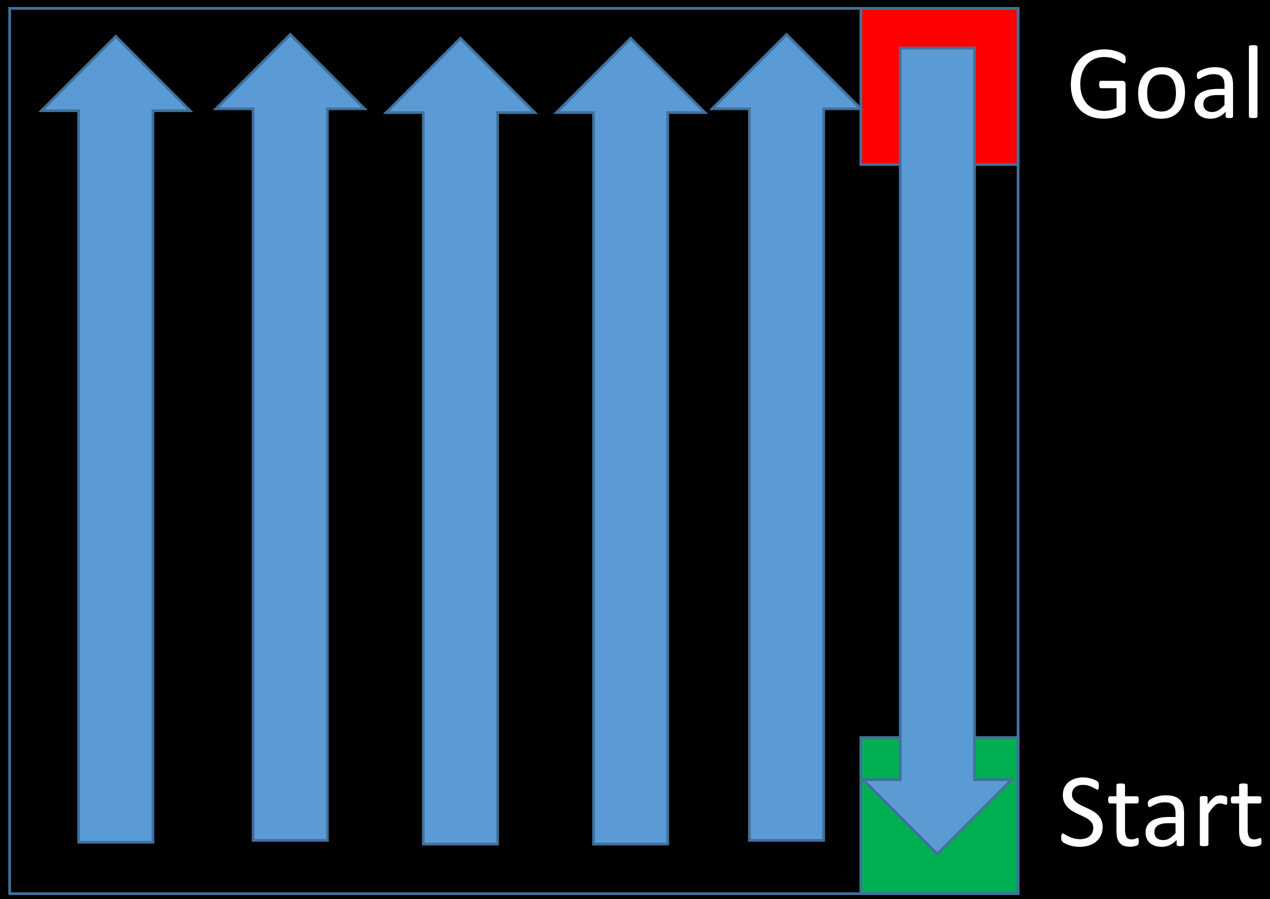
$$Q(s^m, f(s, x)) - Q(s^m, NOP)$$

The value of taking the currently recommended action

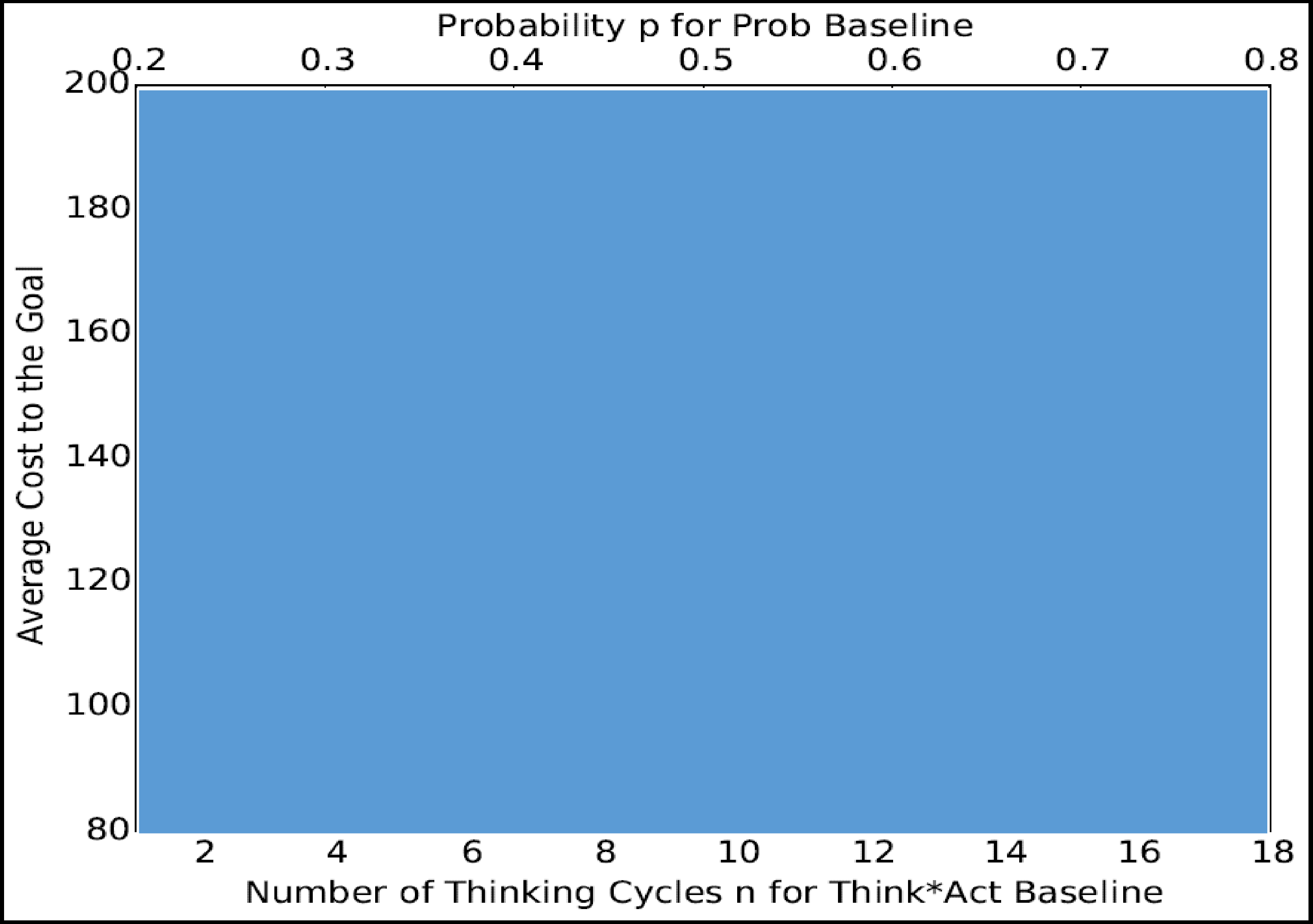
The value of taking a NOP

$O(K|A|^2)$

Experiments

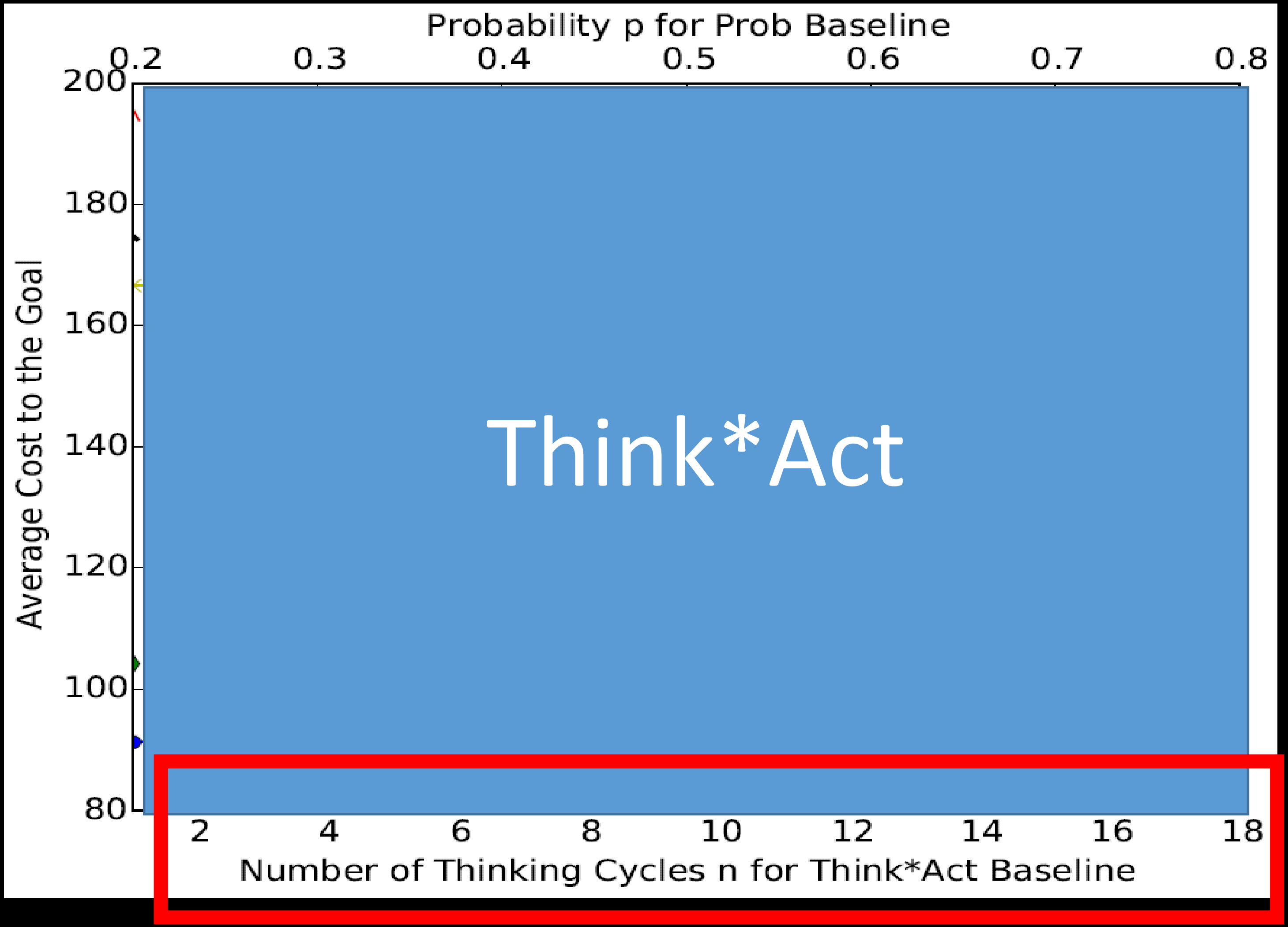

Goal
Start

Baselines



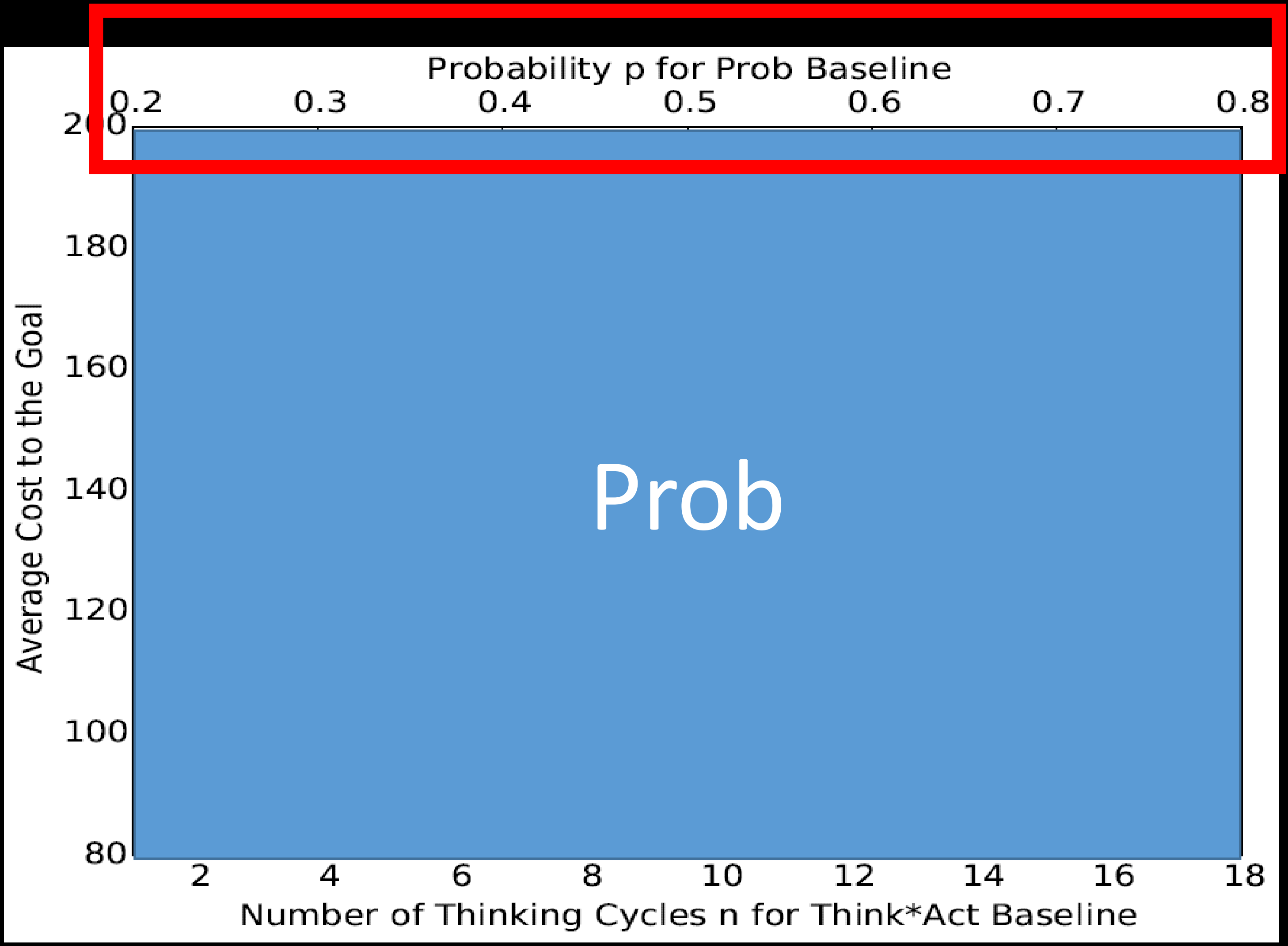
Probability p for Prob Baseline
0.2
0.3
0.4
0.5
0.6
0.7
0.8
200
180
160
140
120
100
80
Average Cost to the Goal
2
4
6
8
10
12
14
16
18
Number of Thinking Cycles n for Think*Act Baseline

Baselines

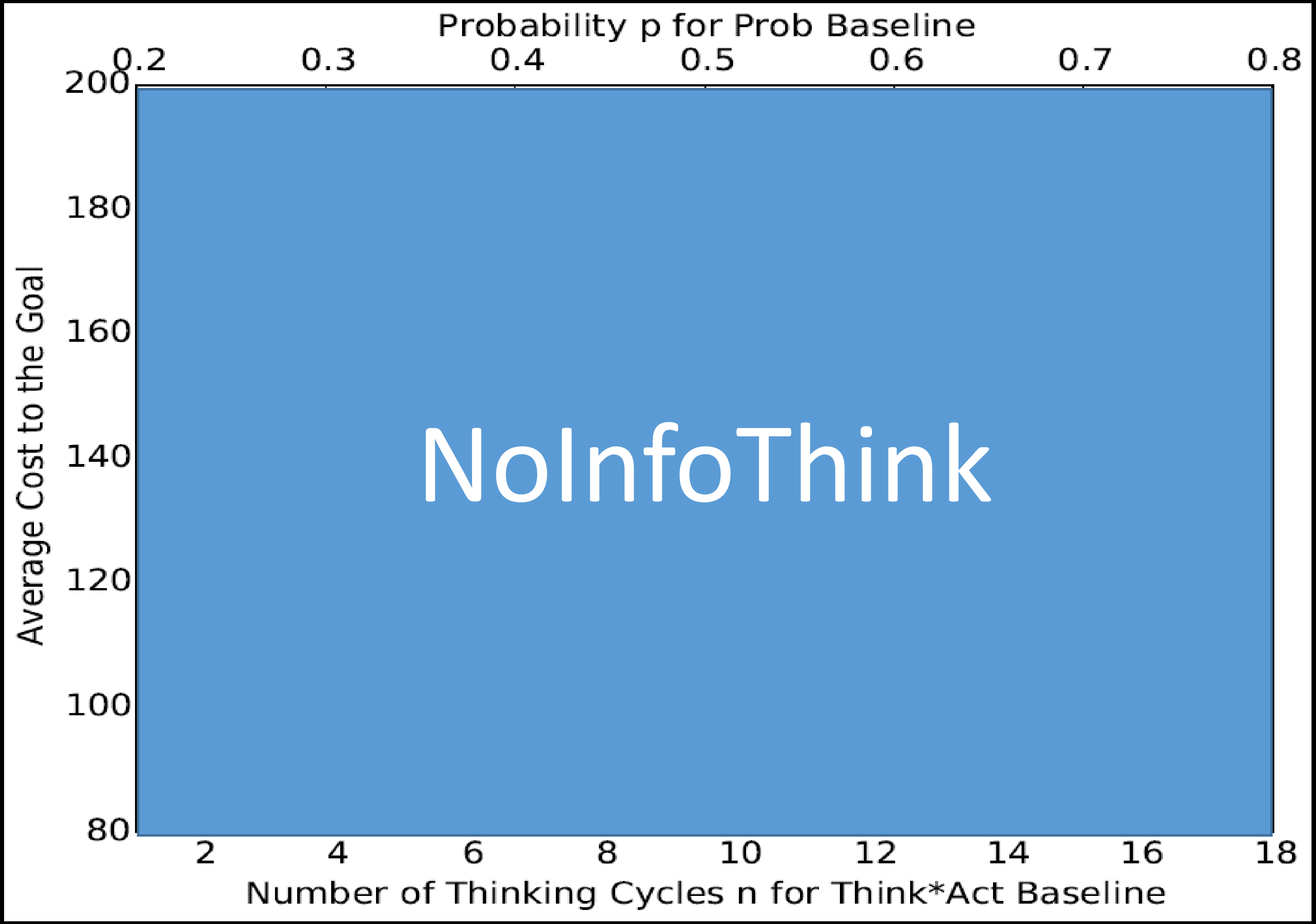


Probability p for Prob Baseline
0.2 0.3 0.4 0.5 0.6 0.7 0.8
Average Cost to the Goal
200 180 160 140 120 100 80
Think*Act
2 4 6 8 10 12 14 16 18
Number of Thinking Cycles n for Think*Act Baseline

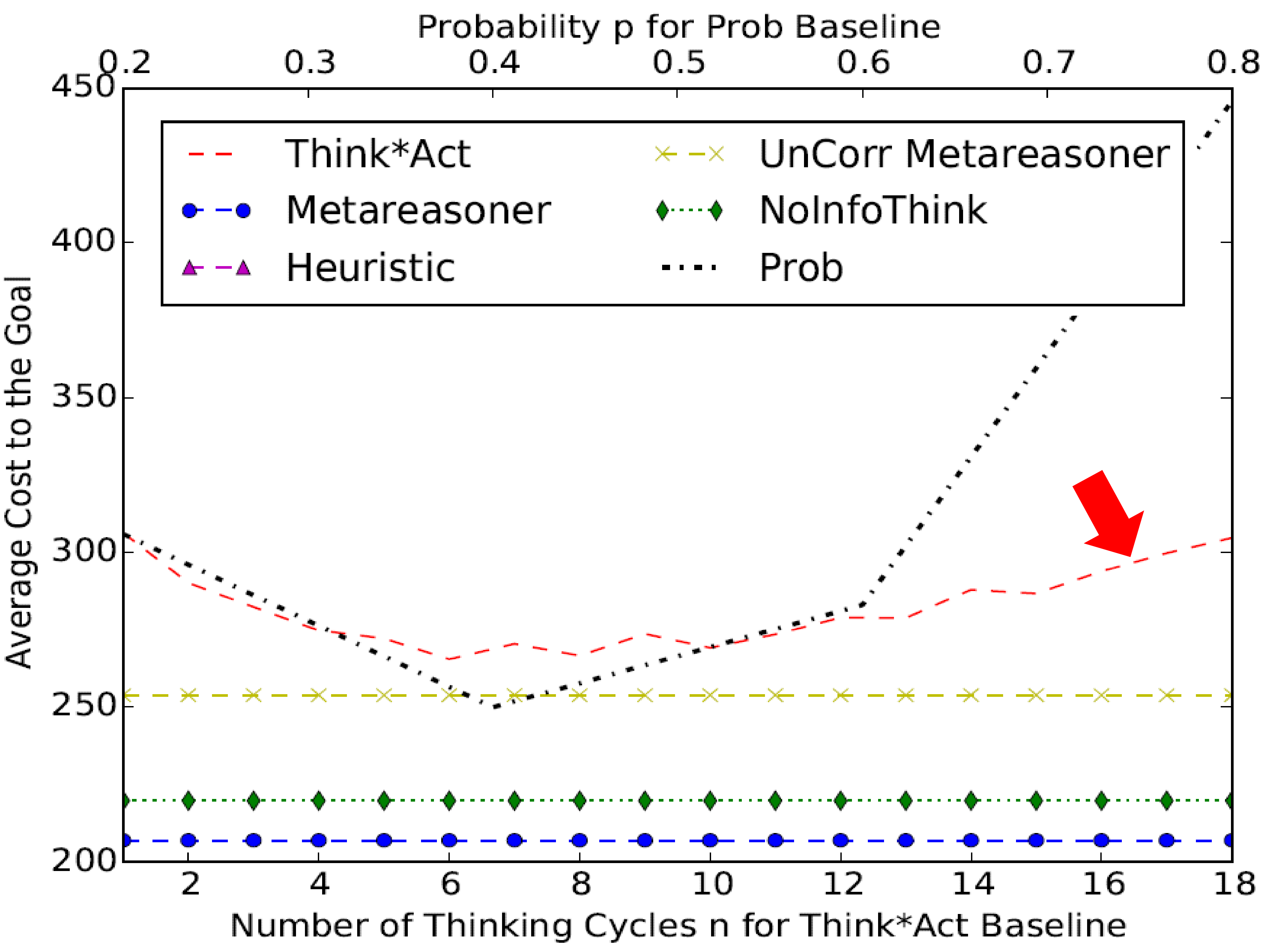
Baselines

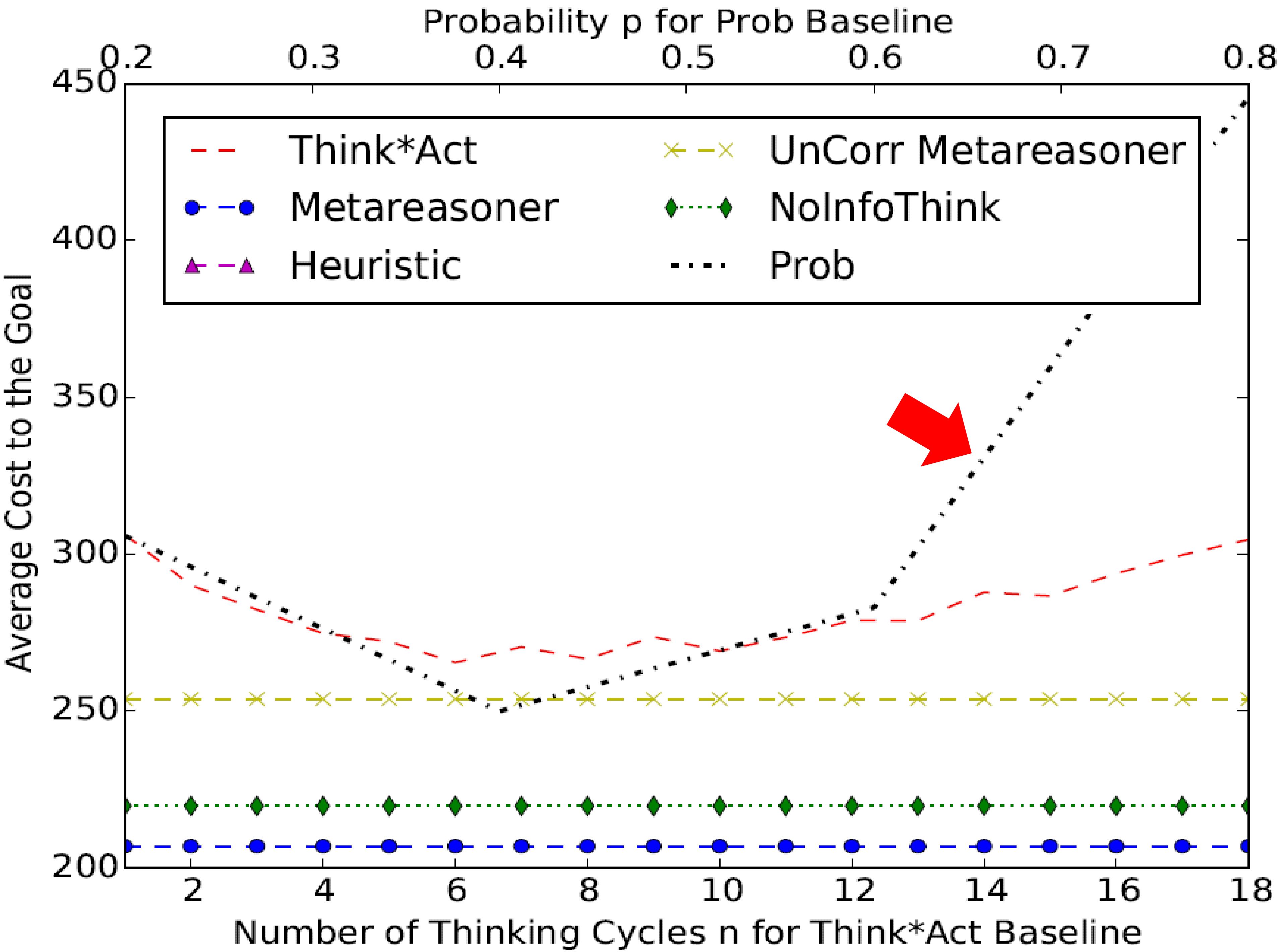

Probability p for Prob Baseline
0.2
0.3
0.4
0.5
0.6
0.7
0.8
200
180
160
140
120
100
80
Average Cost to the Goal
Prob
2
4
6
8
10
12
14
16
18
Number of Thinking Cycles n for Think*Act Baseline

Baselines

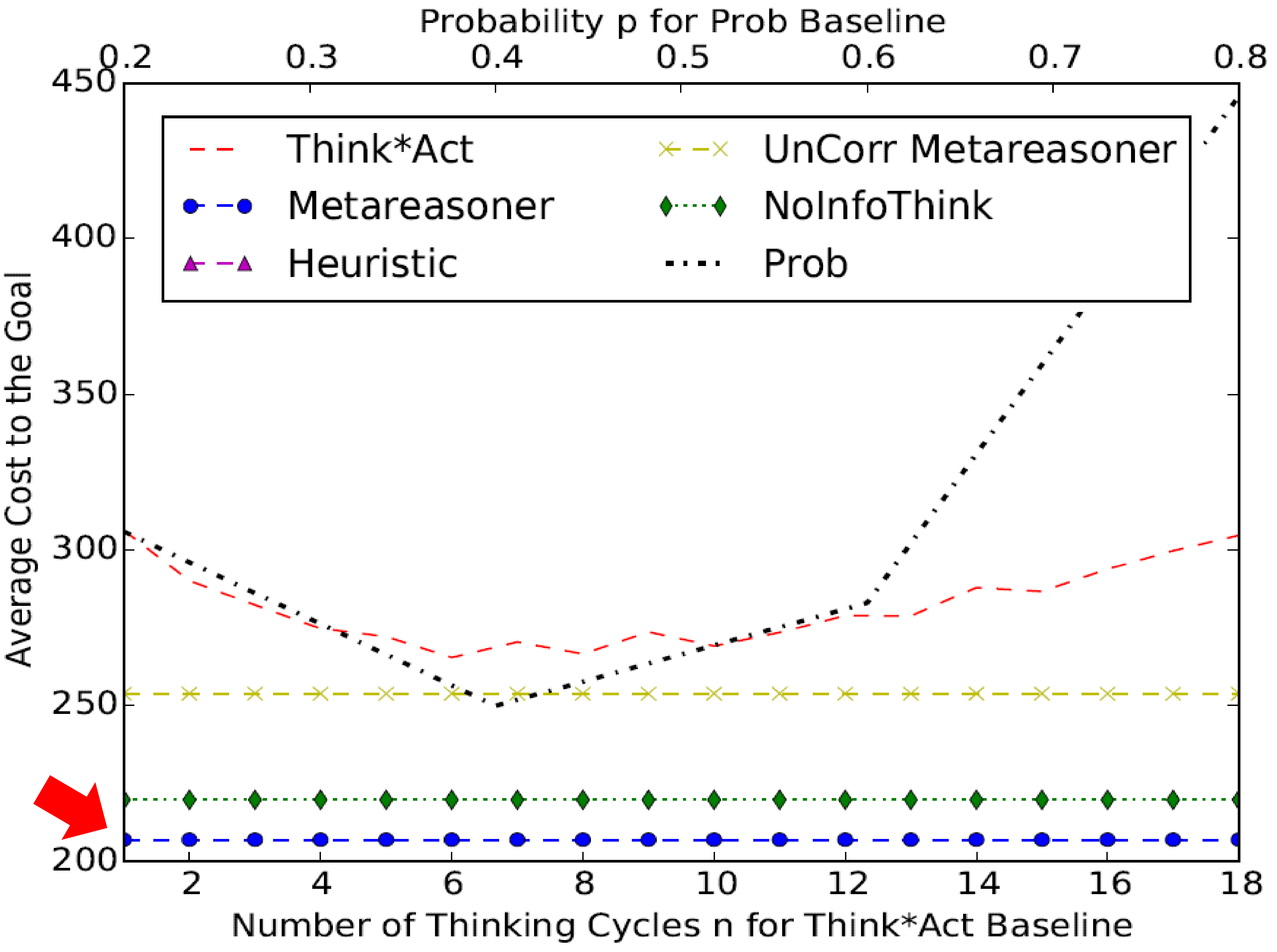


Probability p for Prob Baseline
0.2 0.3 0.4 0.5 0.6 0.7 0.8
200 180 160 140 120 100 80
Average Cost to the Goal
NoInfoThink
2 4 6 8 10 12 14 16 18
Number of Thinking Cycles n for Think*Act Baseline


Probability p for Prob Baseline
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Think*Act
Metareasoner
Heuristic
UnCorr Metareasoner
NoInfoThink
Prob
Average Cost to the Goal
450
400
350
300
250
200
2
4
6
8
10
12
14
16
18
Number of Thinking Cycles n for Think*Act Baseline



Probability p for Prob Baseline
0.2
0.3
0.4
0.5
0.6
0.7
0.8
450
400
350
300
250
200
Average Cost to the Goal
Think*Act
Metareasoner
Heuristic
UnCorr Metareasoner
NoInfoThink
Prob
2
4
6
8
10
12
14
16
18
Number of Thinking Cycles n for Think*Act Baseline



Probability p for Prob Baseline
0.2
0.3
0.4
0.5
0.6
0.7
0.8
450
400
350
300
250
200
Average Cost to the Goal
2
4
6
8
10
12
14
16
18
Number of Thinking Cycles n for Think*Act Baseline
Think*Act
Metareasoner
Heuristic
UnCorr Metareasoner
NoInfoThink
Prob

Metareasoning for Planning Under Uncertainty: Conclusions

Formalization and complexity analysis

Fast algorithms for approximate metareasoning with no hyperparameters!