Safe Exploration and Optimization of Constrained MDPs using Gaussian Processes

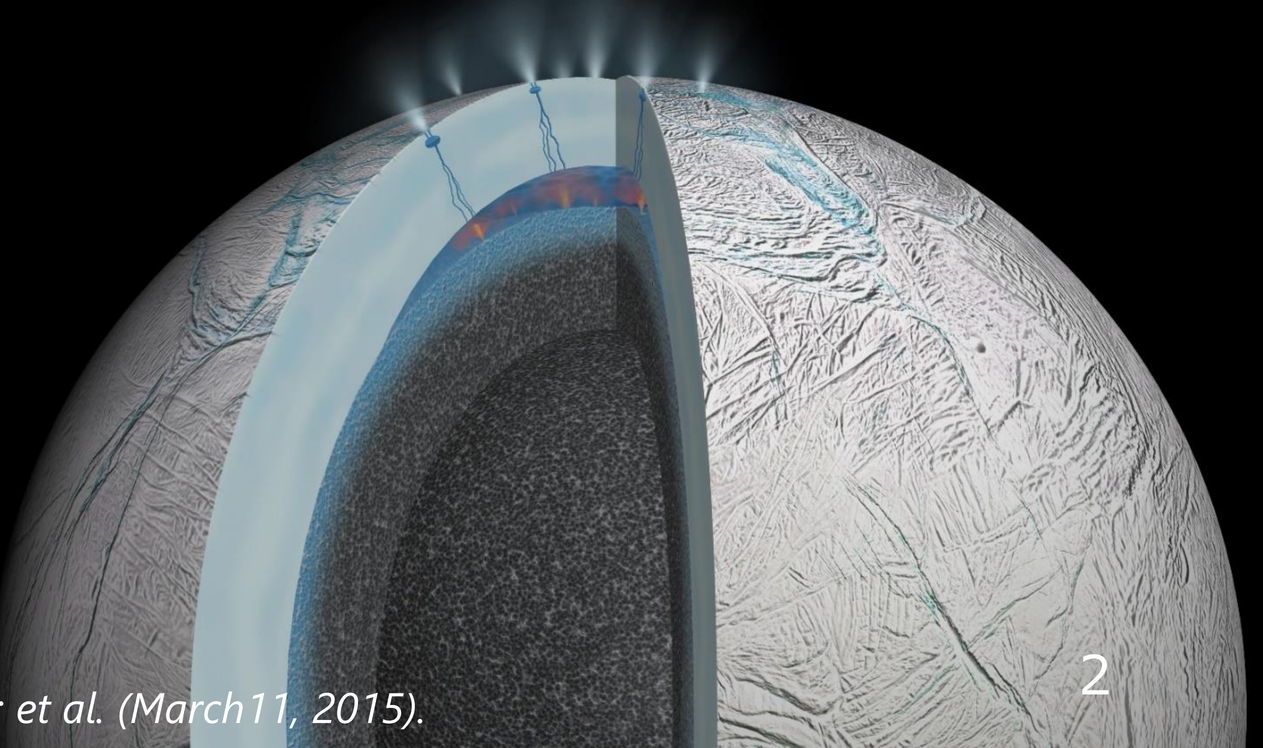
Akifumi Wachi (Univ. Tokyo), Yanan Sui (Caltech)
Yisong Yue (Caltech), Masahiro Ono (NASA/JPL)

February 5, 2018
AAAI Conference on Artificial Intelligence

Background

There is increasing need for automated exploration of the *unknown* environment

- *Unknown where is dangerous/safe*
- *Unknown where is scientifically worthwhile to visit*



Hsu, Hsiang-Wen; Postberg, Frank; et al. (March11, 2015).

Problem Statement

We would like an agent to:

- Obtain scientifically valuable data
- Guarantee safety

Problem formulation

$$\max \quad \sum_{k=1}^{N} \gamma^{k-1} f_k(\boldsymbol{x}_k)$$

$$\text{subject to} \quad g_k(\boldsymbol{x}_k) > h \quad \forall k = [1 \ N-1]$$

x_k : state vector

N : terminal time step

$f_k(\boldsymbol{x}_k)$: reward function

$g_k(\boldsymbol{x}_k)$: safety function

h : safety threshold

γ : discount rate

We assume that reward and safety are a priori unknown.

Our Contributions

<u>Reinforcement learning</u>

[Sutton and Barto, 1998; Bertsekas and Tsitsikils, 1995]

$$\max \quad \sum_{k=1}^{N} \gamma^{k-1} f_k(\boldsymbol{x}_k)$$

<u>Risk-Sensitive Decision Making</u>

[Schwarm and Nikolaou, 1999; Blackmore et al., 2010]

<u>Add risk-sensitivity (Safety constraint)</u> ⬇

⬇ **<u>Assume that safety is a priori unknown</u>**

<u>Our research</u>

$$\max \quad \sum_{k=1}^{N} \gamma^{k-1} f_k(\boldsymbol{x}_k)$$

$$\text{subject to} \quad g_k(\boldsymbol{x}_k) > h \quad \forall k = [1 \ N-1]$$

Exploration & Exploitation

If you were a treasure hunter, what would you do?

- Collect treasure which has been already identified?
- Seek more valuable treasure?
- Try to search safe region ?

Such a problem is called *exploration/exploitation problem*

Three-way trade-off

Check where is safe/dangerous

Collect already recognized reward

Seek higher reward

How should we deal with safety?

It is difficult to guarantee safety because ...

It is too late to realize the hazard after hitting it!

Just single mistake may result in failure.

We must predict hazards before actually visiting unsafe states.

How should we deal with safety?

Parameters in natural environments have some regularity

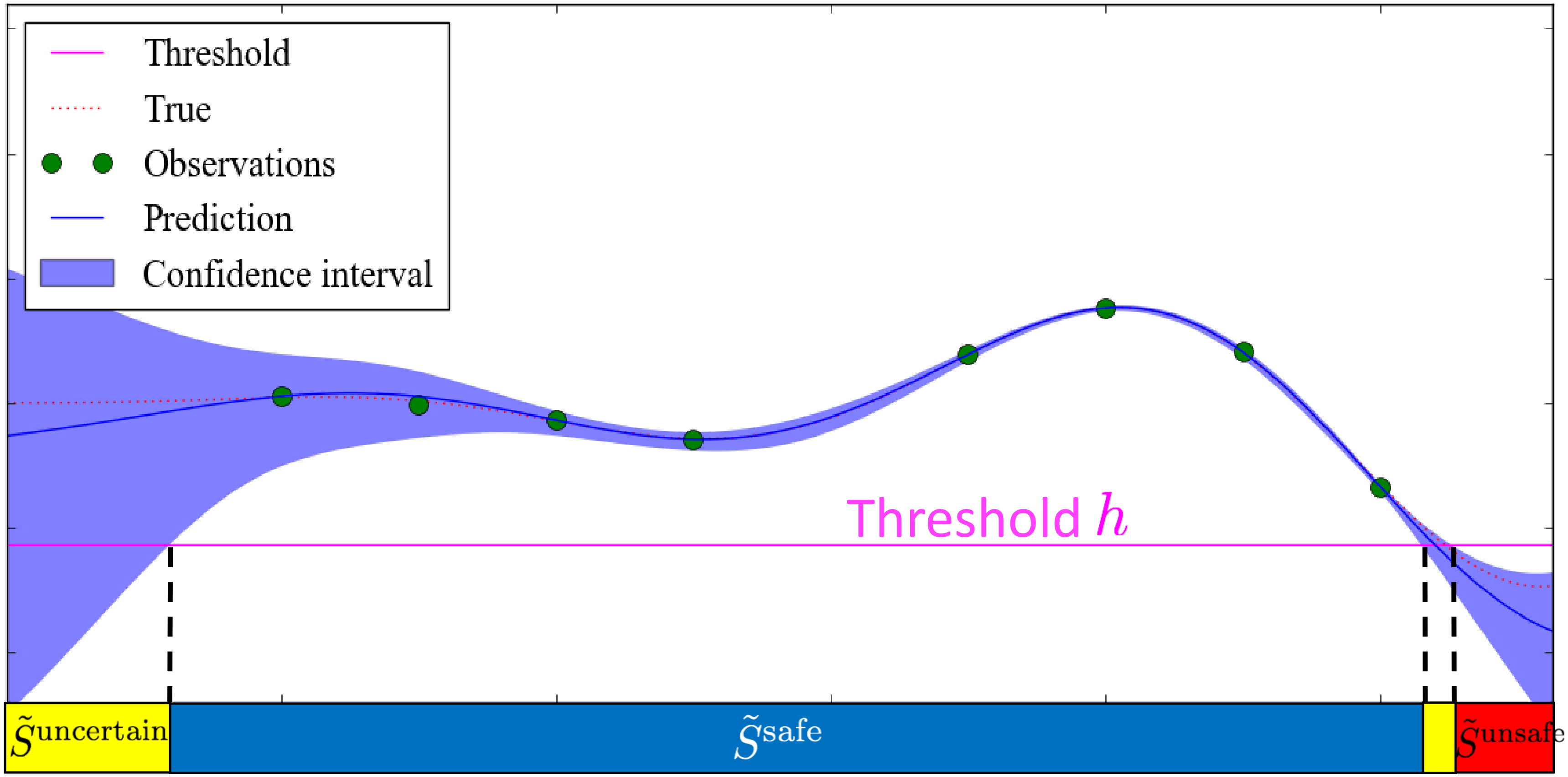
We assume that similar states have similar values of safety function.

⇩

Evaluate safety function using Gaussian Processes (GPs)

Gaussian Processes

Safety constraint: $g_k(x_k) \geq h$



An agent can guarantee safety with high probability.
(For more detail, see our paper.)

Exploration/Exploitation Problem

Exploration of safety

M. Turchetta et al. (2016) focuses on exploration of safety.

Key point

How can we balance the three objectives?

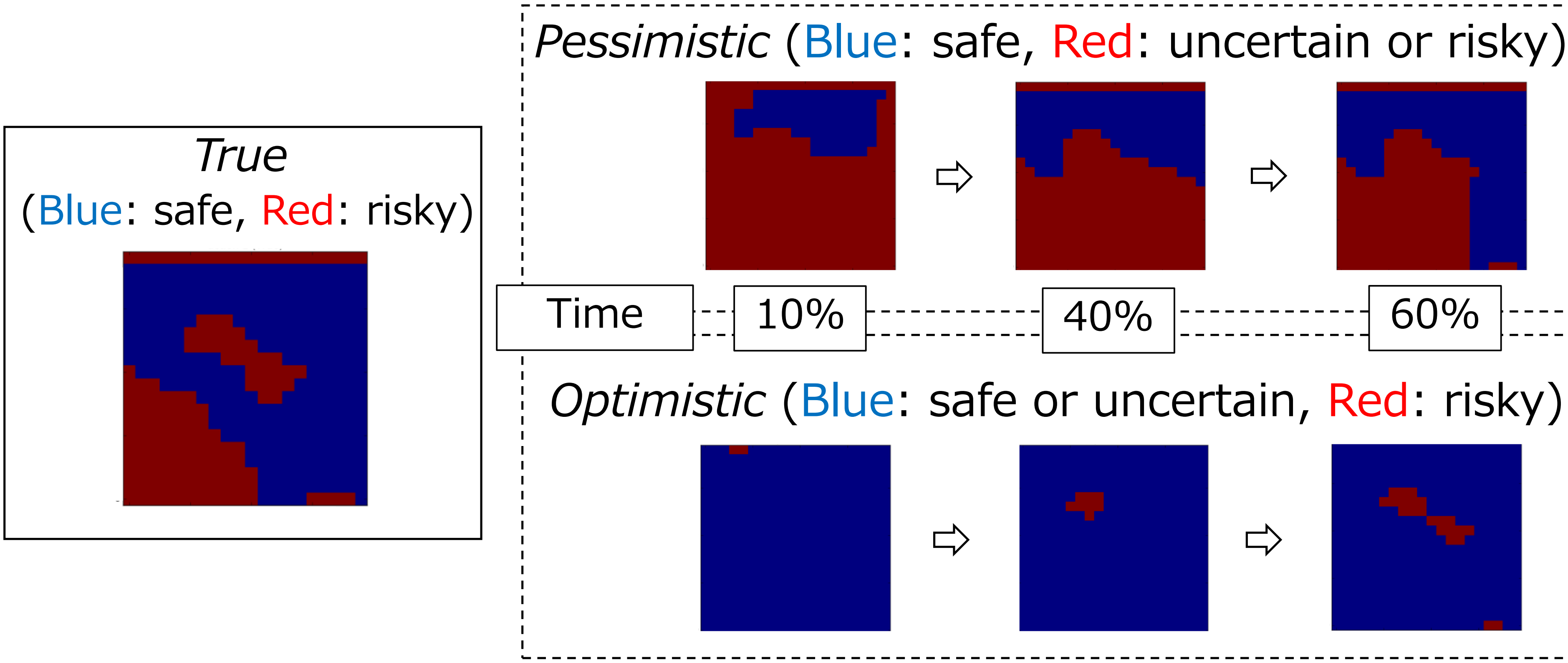
Exploitation of reward

Exploration of reward

A great deal of previous work in the field of reinforcement learning has worked on this problem

Classification of State Space

- Classify the state space into three regions using GP
  → *safe space, uncertain space, risky space*



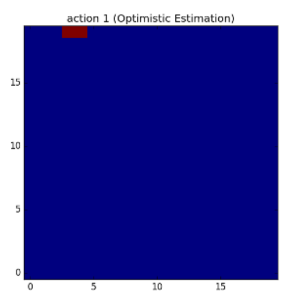
Introduction of Delta-J

Difference of cumulative reward for Optimistic and Pessimistic cases means the need for exploration of safety

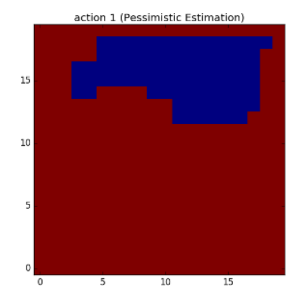
$$\Delta J(s) = \hat{J}(s) - \bar{J}(s)$$

Optimistic Pessimistic

Optimistic



Pessimistic



Difference is big → *"exploration of safety" is necessary*
Difference is small → *it is OK to focus on reward*

Overall Policy

exploration/exploitation of reward

exploration of safety

$$\pi_N(s, b^f, b^g) = \arg\max_{s \in S^{\mathrm{safe}}} \{\boxed{\bar{J}(s, b^f, b^g)} + \lambda \boxed{\Delta J(s, b^f, b^g)}\}$$

$\boxed{0 \leq \lambda \leq 1}$

$$\Downarrow$$

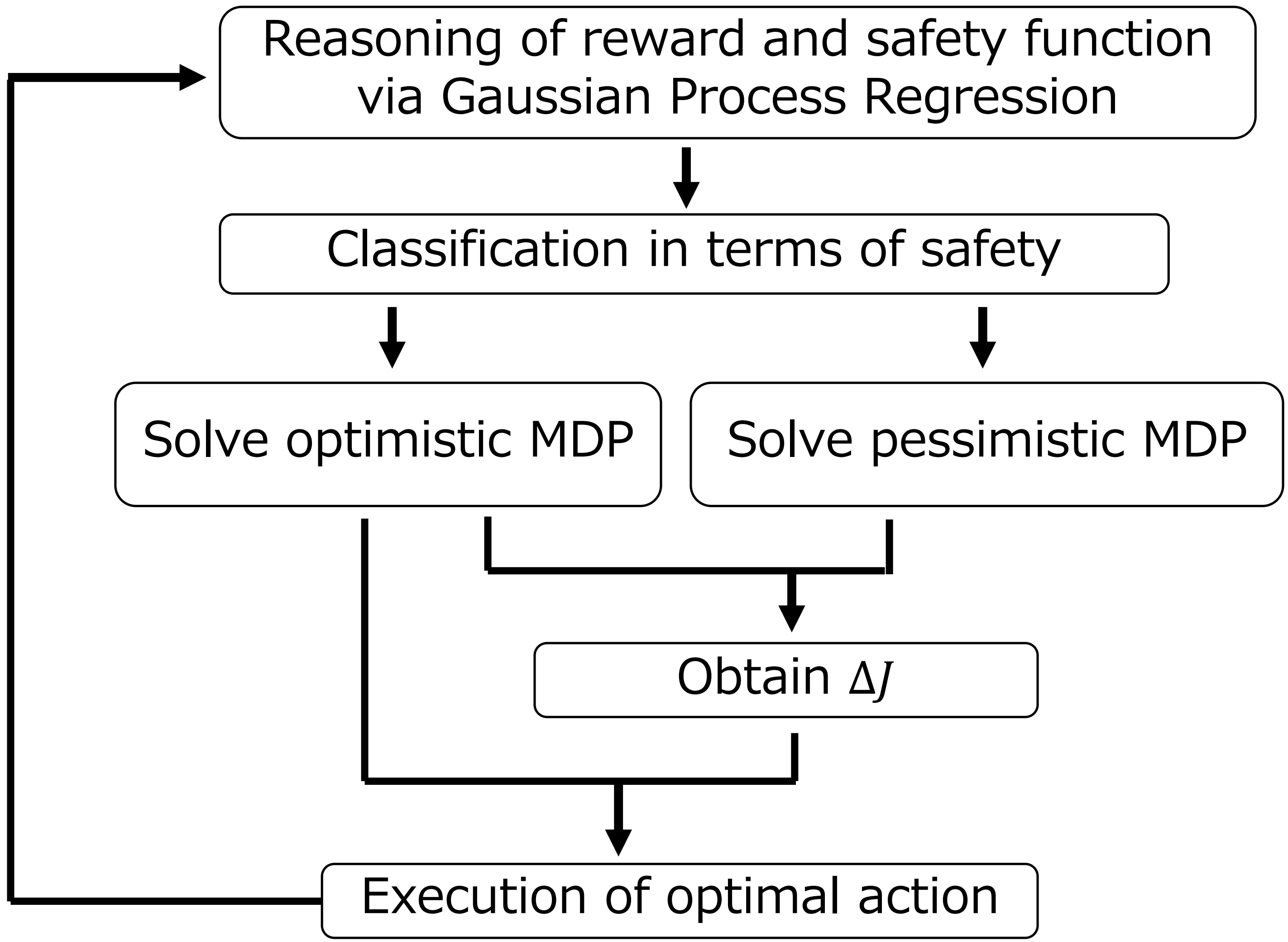
value function for *Optimistic case*

value function for *Pessimistic case*

$$\pi_N(s, b^f, b^g) = \arg\max_{s \in S^{\mathrm{safe}}} \{\lambda \boxed{\hat{J}(s, b^f, b^g)} + (1 - \lambda) \boxed{\bar{J}(s, b^f, b^g)}\}$$

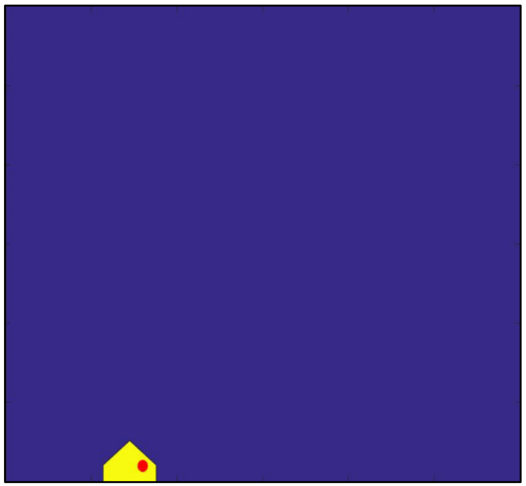
An agent should choose a state with the maximum value of the dividing point of optimistic and pessimistic cases.

Algorithm flow

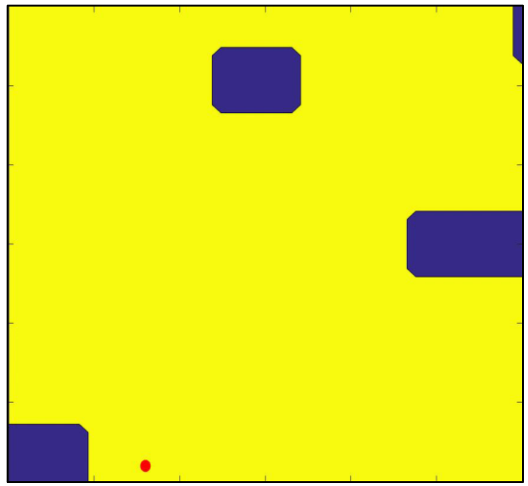


Simulation result

Proposed method

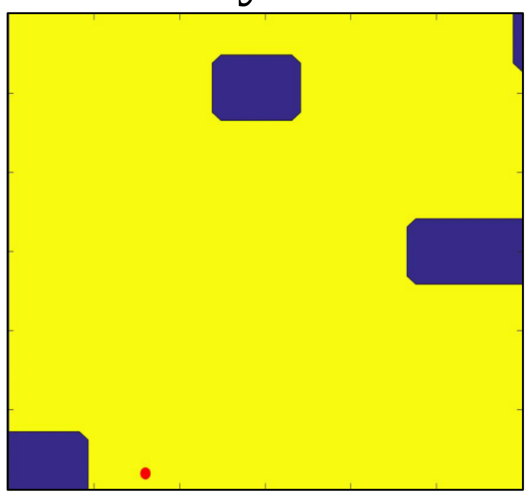


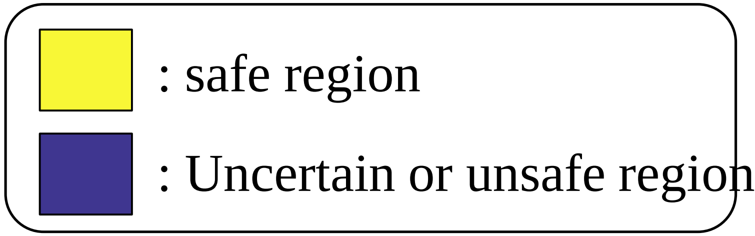
Safety/reward known

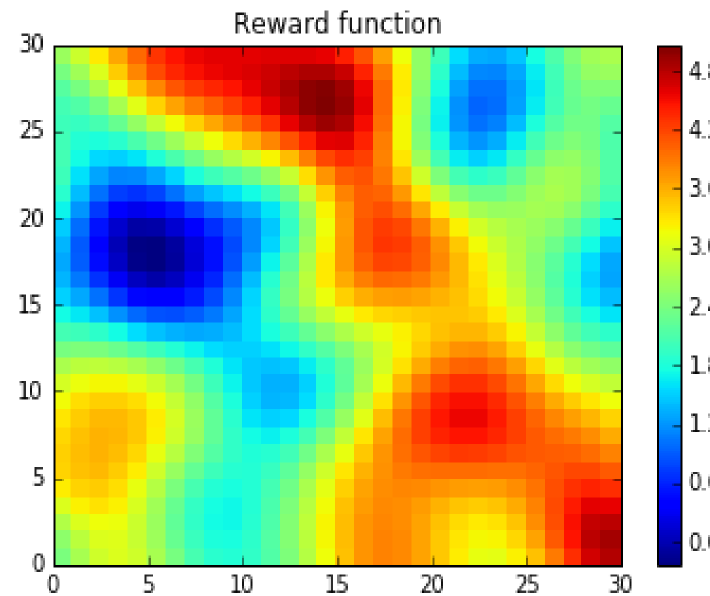


M. Turchetta et al.

Safety known







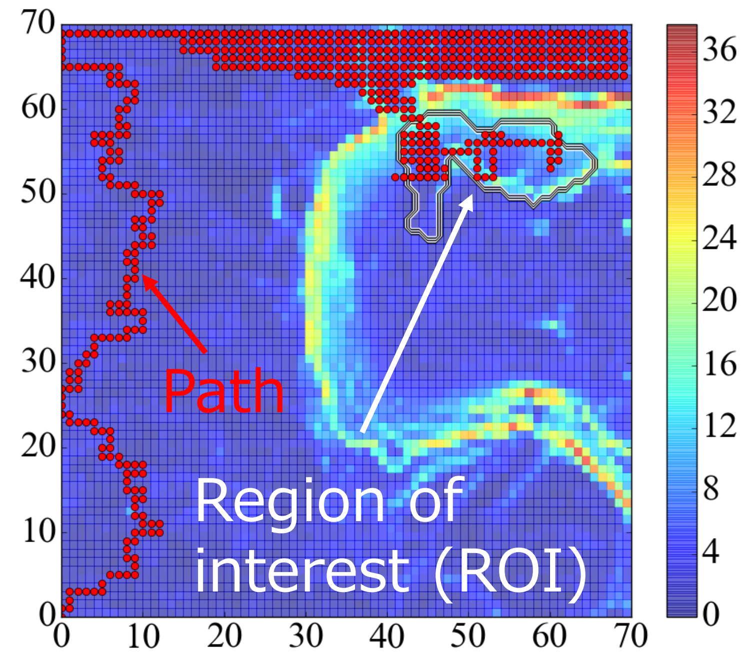
Mars Surface Exploration Example

Conducted a simulation using Mars terrain data based on real Mars surface exploration scenario.

- Safety function: slope angle
- Threshold: 25deg
- Reward function: binary (one within in ROI)

Result

- Succeeded in arriving at ROI while guaranteeing safety.
- Proved that our proposed method can be applied to practical applications



Conclusion

1. Formulate a problem in which a MDP is safely explored and optimized with a priori unknown reward and safety

2. Propose an algorithm to balance exploration of reward, exploitation of reward, and exploration of safety

3. Demonstrate the effectiveness of our proposed method by the three types of simulation including one using real Martian data