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*
Problem Statement

We would like an agent to:

- Obtain scientifically valuable data
- Guarantee safety

Problem formulation

\[
\max \sum_{k=1}^{N-1} \gamma^{k-1} f_k(x_k)
\]

subject to \( g_k(x_k) > h \quad \forall k = [1 \ N - 1] \)

- \( x_k \): state vector
- \( f_k(x_k) \): reward function
- \( g_k(x_k) \): safety function
- \( h \): safety threshold
- \( \gamma \): discount rate

We assume that reward and safety are a priori unknown.
Our Contributions

Reinforcement learning

\[ \max \sum_{k=1}^{N} \gamma^{k-1} f_k(x_k) \]

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

Risk-Sensitive Decision Making

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

Add risk-sensitivity (Safety constraint)

Assume that safety is a priori unknown

Our research

\[ \max \sum_{k=1}^{N} \gamma^{k-1} f_k(x_k) \]

subject to \( g_k(x_k) > h \) \( \forall k = [1 \ N - 1] \)
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

Exploration of safety

Exploitation of reward

Collect already recognized reward

Seek higher reward

© NASA
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

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

Key point
How can we balance the three objectives?

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

- **Pessimistic** (Blue: safe, Red: uncertain or risky)
  - Time: 10%, 40%, 60%

- **True** (Blue: safe, Red: risky)

- **Optimistic** (Blue: safe or uncertain, Red: risky)
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

Difference is big → “exploration of safety” is necessary
Difference is small → it is OK to focus on reward
Overall Policy

\[ \pi_N(s, b^f, b^g) = \arg \max_{s \in S_{\text{safe}}} \left\{ \overline{J}(s, b^f, b^g) + \lambda \Delta J(s, b^f, b^g) \right\} \]

Value function for Optimistic case

Value function for Pessimistic case

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

Reasoning of reward and safety function via Gaussian Process Regression

Classification in terms of safety

Solve optimistic MDP

Solve pessimistic MDP

Obtain $\Delta J$

Execution of optimal action
Simulation result

Proposed method  Safety/reward known

M. Turchetta et al.  Safety known

: safe region
: Uncertain or unsafe region

Reward function
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.