Risk-Aware Reinforcement Learning with Coherent Risk Measures and Non-Linear Function Approximation

Thanh Lam, Arun Verma, Bryan Kian Hsiang Low, Patrick Jaillet

ICLR 2023

Presenter: Wei-Chun Tsai

National Cheng Kung University

國立成功大學
National Cheng Kung University
1931

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

Introduction - Risk Issue & Risk Quantification

- Risk Issues in Reinforcement Learning (RL):
  - Traditional RL focuses on maximizing total rewards, often overlooking the risk of low rewards caused by environmental stochasticity.
  - **Objective:** To design RL algorithms that learn risk-aware strategies, minimizing the risk of low expected total rewards.
- Risk Quantification Methods
  - Entropic risk [1]
  - Value-at-Risk (VaR) [2]
  - Conditional Value-at-Risk (CVaR) [3]
  - Entropic Value-at-Risk (EVaR) [4]

[1] Hans Föllmer and Thomas Knispel. Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. Stochastics and Dynamics, 11(02n03):333–351, 2011.
[2] Michael Alan Howarth Dempster. Risk management: value at risk and beyond. Cambridge University Press, 2002.
[3] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. Journal of risk, 2:21–42, 2000.
[4] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. Journal of Optimization Theory and Applications, 155(3):1105–1123, 2012.

Introduction - Risk Issue & Risk Quantification

- Coherent Risk Measures
  - **Mathematical Consistency and Rationality**
    - Normalized, Monotonic, Super-additive, Positively Homogeneous, Translation Invariant [3]
  - **Time consistency:** risk preferences remain consistent across multi-stage decision-making processes.
  - Examples: Conditional Value-at-Risk (CVaR), Entropic Value-at-Risk (EVaR)

Introduction - Challenge

- Challenges in existing risk-aware RL methods:
  - Most methods assume a **tabular MDP**, requiring a complete traversal of the state space, which is infeasible for large or continuous state spaces. [5], [6]
  - Assume that MDP is known, eliminating the need for exploration or generalization, thus treating the problem as a planning task rather than a learning task. [7]
- Challenges in extending risk-neutral RL to risk-aware RL:
  - **Non-linear Bellman Equation:** The linearity property does not hold in the risk-aware RL setting.
  - **Unbiased Estimation:** In the risk-aware RL setting, it is not possible to construct an unbiased estimate of the Bellman update using a single sample of the next state.

[5] Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. Mathematical Methods of Operations Research, 74(3):361–379, 2011.

[6] Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse bayes-adaptive reinforcement learning. Advances in Neural Information Processing Systems, 34, 2021.

[7] Pengqian Yu, William B Haskell, and Huan Xu. Approximate value iteration for risk-aware markov decision processes. IEEE Transactions on Automatic Control, 63(9):3135–3142, 2018.

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

Related Work

- Fei et al. (2021) [8], Fei & Xu (2022) [9]
  - Consider the risk-aware reinforcement learning in the function approximation and regret minimization setting.
  - Both adopt the **entropic risk measure** to quantify risk.
- Tamar et al. (2015) [10]
  - Proposes an actor-critic algorithm for the entire class of **coherent risk measures.**
  - However, it does not provide a theoretical analysis of regret.

[8] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In International Conference on Machine Learning, pp. 3198–3207. PMLR, 2021.
[9] Yingjie Fei and Ruitu Xu. Cascaded gaps: Towards logarithmic regret for risk-sensitive reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 6392–6417. PMLR, 17–23 Jul 2022.
[10] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. Advances in neural information processing systems, 28, 2015.

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

Problem Setting - Risk-aware Episodic MDP

- Episodic finite-horizon Markov decision process (MDP)
  - **Objective:** To enable the agent to take actions in an uncertain environment, maximizing its cumulative rewards while considering the risks associated with returns.
- **The standard risk-neutral MDP:**
  - Maximize the expected value of cumulative rewards. (Eq. 1)
  - The risk-neutral objective DOES NOT account for the risk incurred due to the stochasticity in the state transitions and the agent's policy.

$$\max_{\pi} \mathbb{E}_{x_1}^{\pi}\left[\sum_{h=1}^{H} r_h(x_h, a_h)\right]. \tag{1}$$

Problem Setting - Risk-aware Episodic MDP

- **The risk-aware MDP**
  - Consider the randomness of state transitions and the risks induced by the policy.
  - Defines a new objective function. (Equation 2.)
    - ρ: coherent risk measure
  - Advantages:
    - Guarantees the existence of an optimal policy, and it’s Markovian.
    - Satisfies the time consistency property.

$$\max_{\pi} J^{\pi}(x_1),$$

$$\text{where} \quad J^{\pi}(x_1) := r_1(x_1, a_1) + \rho(r_2(x_2, a_2) + \rho(r_3(x_3, a_3) + \dots)), \qquad (2)$$

Problem Setting - Bellman Equation and Regret

- **Bellman Equation**
  - Based on the defined objective function, the risk-aware Bellman equation is developed.
  - Optimal Bellman Equation: (Eq. 3)

$$
\begin{aligned}
Q_h^*(x,a) &= (r_h + D_\rho^h V_{h+1}^*)(x,a), \\
V_h^*(x) &= \max_{a \in A} Q_h^*(x,a), \qquad (3) \\
V_{H+1}^*(x) &= 0.
\end{aligned}
$$

Problem Setting - Bellman Equation and Regret

- **Total Regret:** (Eq. 4)
    - The sum of regret across all episodes
    - Representing the total performance gap during the entire learning process. It reflects the total loss in returns caused by the agent not adopting the optimal policy during the learning period.

$$\mathfrak{R}_T(\rho) = \sum_{t=1}^{T} \Big[\underbrace{V_1^{\star}(x_1^t) - V_1^{\pi_t}(x_1^t)}_{\text{Regret}}\Big]. \tag{4}$$

Problem Setting - Weak Simulator Assumption

- Challenge
  - In risk-aware RL, a single sample is **insufficient** to reliably estimate future returns.
- Assumption
  - A weak simulator exists that allows drawing samples from the probability transition kernel to **generate multiple next-state samples**.
  - This assumption is much weaker than the archetypal simulator assumptions often seen in the RL literature.

Problem Setting - Estimating Non-linear Functions

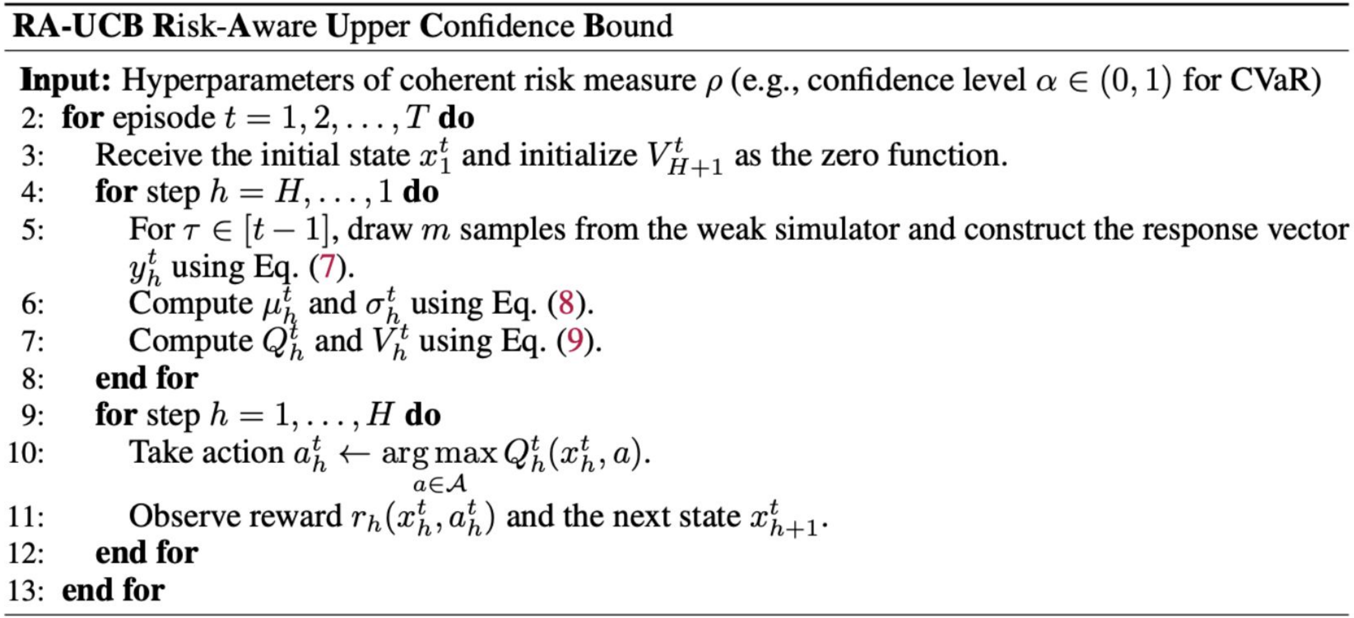
- Challenge
  - In risk-aware RL, the Bellman equation becomes non-linear, making standard linear function approximation techniques ineffective.
- Solution
  - Use the **Reproducing Kernel Hilbert Space (RKHS)** to represent the optimal action-value function as a non-linear function.

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

Risk-aware RL Algorithm with Coherent Risk Measures

- **Risk-Aware Upper Confidence Bound (RA-UCB)**
  - Built upon the foundation of the *Value Iteration Algorithm.* [11]



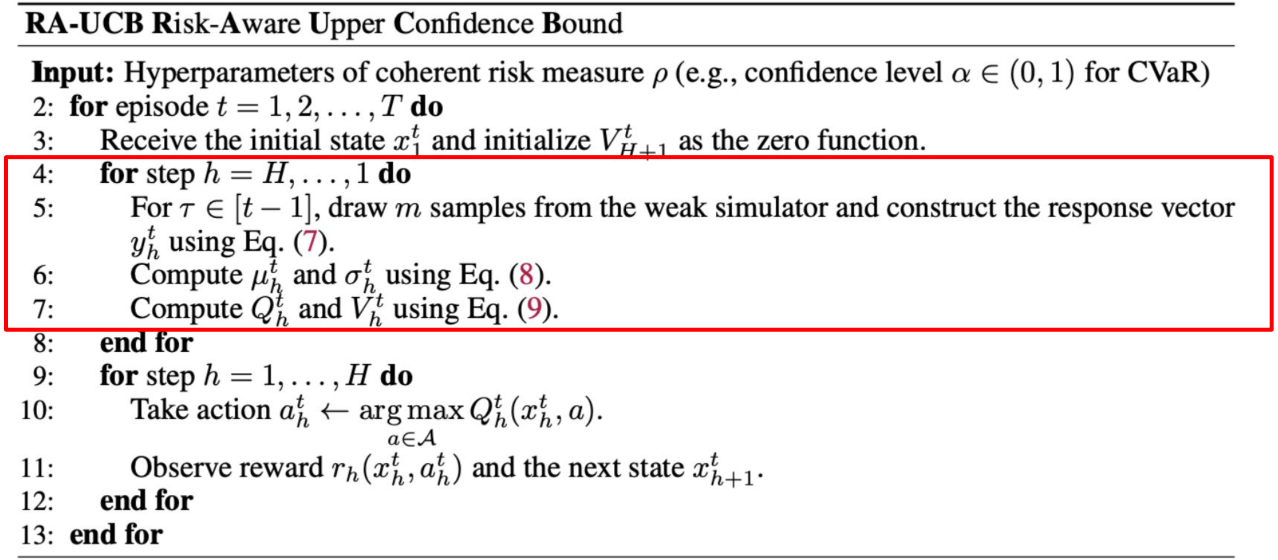
RA-UCB Risk-Aware Upper Confidence Bound

```
Input: Hyperparameters of coherent risk measure $\rho$ (e.g., confidence level $\alpha \in (0, 1)$ for CVaR)
for episode $t = 1, 2, \dots, T$ do
    Receive the initial state $x_1^t$ and initialize $V_{H+1}^t$ as the zero function.
    for step $h = H, \dots, 1$ do
        For $\tau \in [t-1]$, draw $m$ samples from the weak simulator and construct the response vector $y_h^t$ using Eq. (7).
        Compute $\mu_h^t$ and $\sigma_h^t$ using Eq. (8).
        Compute $Q_h^t$ and $V_h^t$ using Eq. (9).
    end for
    for step $h = 1, \dots, H$ do
        Take action $a_h^t \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^t(x_h^t, a)$.
        Observe reward $r_h(x_h^t, a_h^t)$ and the next state $x_{h+1}^t$.
    end for
end for
```

[11] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

Risk-aware RL Algorithm with Coherent Risk Measures

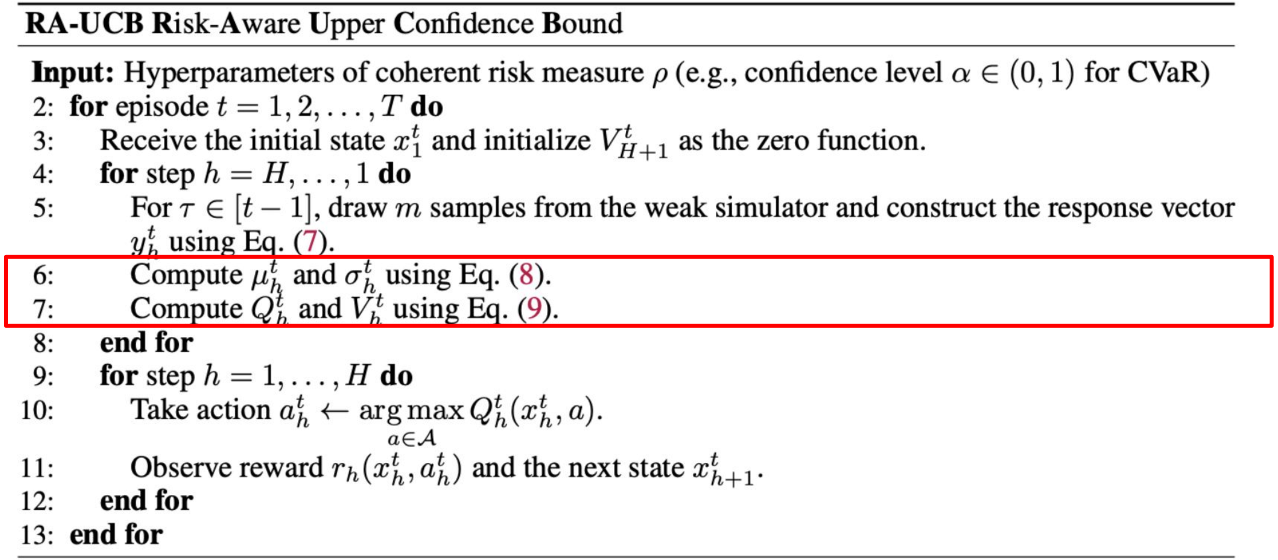
- **Step 1: Value Function Estimation using Kernel Least-Square Regression**
  - 5: Sample multiple next states & construct the response vector.
  - 6: Compute μ and σ of the value function
  - 7: Further compute Q and V to complete the value function estimation.



```
RA-UCB Risk-Aware Upper Confidence Bound
Input: Hyperparameters of coherent risk measure $\rho$ (e.g., confidence level $\alpha \in (0,1)$ for CVaR)
 2: for episode $t = 1, 2, \dots, T$ do
 3:     Receive the initial state $x_1^t$ and initialize $V_{H+1}^t$ as the zero function.
 4:     for step $h = H, \dots, 1$ do
 5:         For $\tau \in [t-1]$, draw $m$ samples from the weak simulator and construct the response vector
            $y_h^t$ using Eq. (7).
 6:         Compute $\mu_h^t$ and $\sigma_h^t$ using Eq. (8).
 7:         Compute $Q_h^t$ and $V_h^t$ using Eq. (9).
 8:     end for
 9:     for step $h = 1, \dots, H$ do
10:         Take action $a_h^t \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^t(x_h^t, a)$.
11:         Observe reward $r_h(x_h^t, a_h^t)$ and the next state $x_{h+1}^t$.
12:     end for
13: end for
```

Risk-aware RL Algorithm with Coherent Risk Measures

- **Step 2: Compute Optimistic Bonus**
    - This is reflected in the formulas for calculating the value functions Q and V, where the σ represents the uncertainty of actions or states.



RA-UCB Risk-Aware Upper Confidence Bound

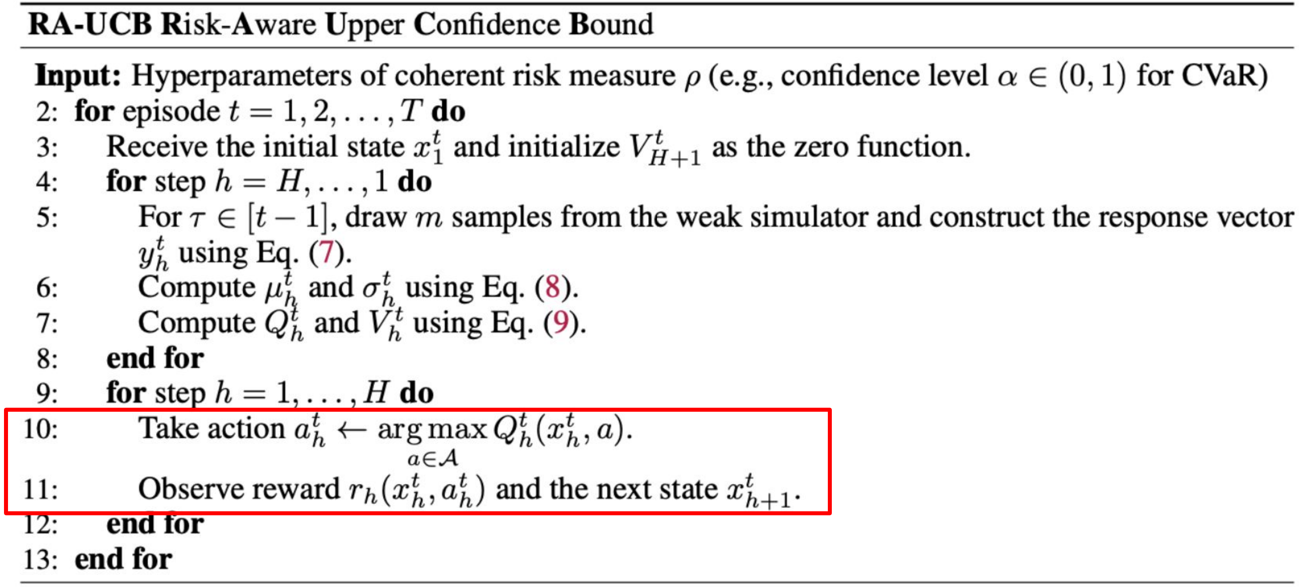
```
Input: Hyperparameters of coherent risk measure $\rho$ (e.g., confidence level $\alpha \in (0,1)$ for CVaR)
for episode $t = 1, 2, \ldots, T$ do
    Receive the initial state $x_1^t$ and initialize $V_{H+1}^t$ as the zero function.
    for step $h = H, \ldots, 1$ do
        For $\tau \in [t-1]$, draw $m$ samples from the weak simulator and construct the response vector
        $y_h^t$ using Eq. (7).
        Compute $\mu_h^t$ and $\sigma_h^t$ using Eq. (8).
        Compute $Q_h^t$ and $V_h^t$ using Eq. (9).
    end for
    for step $h = 1, \ldots, H$ do
        Take action $a_h^t \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^t(x_h^t, a)$.
        Observe reward $r_h(x_h^t, a_h^t)$ and the next state $x_{h+1}^t$.
    end for
end for
```

Eq. (9)

$$Q_h^t(x,a) := \min\left\{\boxed{\mu_h^t(x,a) + \beta \cdot \sigma_h^t(x,a)}, H - h + 1\right\}, \quad V_h^t(x) := \max_{a \in \mathcal{A}} Q_h^t(x,a),$$

Risk-aware RL Algorithm with Coherent Risk Measures

- **Step 3: Execute the Greedy Policy**
  - 10: Select the action that maximizes the value function for the given state.
  - 11: Execute the action and observe the reward as well as the next state.



RA-UCB Risk-Aware Upper Confidence Bound

```
Input: Hyperparameters of coherent risk measure $\rho$ (e.g., confidence level $\alpha \in (0,1)$ for CVaR)
2:  for episode $t = 1, 2, \ldots, T$ do
3:      Receive the initial state $x_1^t$ and initialize $V_{H+1}^t$ as the zero function.
4:      for step $h = H, \ldots, 1$ do
5:          For $\tau \in [t-1]$, draw $m$ samples from the weak simulator and construct the response vector
            $y_h^t$ using Eq. (7).
6:          Compute $\mu_h^t$ and $\sigma_h^t$ using Eq. (8).
7:          Compute $Q_h^t$ and $V_h^t$ using Eq. (9).
8:      end for
9:      for step $h = 1, \ldots, H$ do
10:         Take action $a_h^t \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^t(x_h^t, a)$.
11:         Observe reward $r_h(x_h^t, a_h^t)$ and the next state $x_{h+1}^t$.
12:     end for
13: end for
```

Risk-aware RL Algorithm with Coherent Risk Measures

Main Theoretical Results

- **Sub-linear Regret**
    - The total regret of RA-UCB is **sub-linear**
    - The learning efficiency of the algorithm is theoretically guaranteed.
- **Applicability to Coherent Risk Measures**
    - The theoretical results are not limited to a single risk measure (e.g., CVaR) but **cover the entire class of coherent risk measures**.
    - RA-UCB can be broadly applied to various types of risk preference scenarios.

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
    - Synthetic Experiment: Robot Navigation
    - Real-world Experiment: Trading
- Conclusion

Experiments - Synthetic Experiment: Robot Navigation

- Problem Setup
  - **Object.** The robot needs to navigate inside a room full of obstacles to reach the target destination.
  - 4 actions: {up, down, left, right}
  - Movement direction is perturbed, with angles following a uniform distribution.
  - **Reward.** +10 for reaching the target; negative rewards increase exponentially near obstacles.

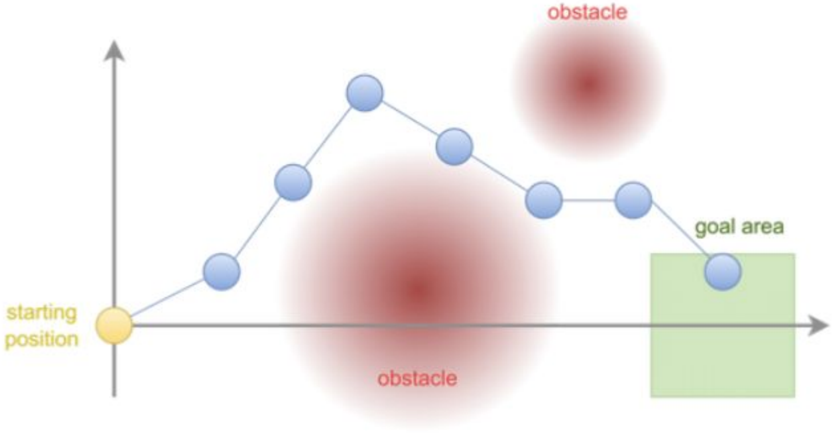


Figure 1. Illustration of the continuous version of the cliff walking problem.

Experiments - Synthetic Experiment: Robot Navigation

- Experimental Setup
  - Analyzed cumulative rewards over 50 episodes using the learned policy.
  - Generated policies with risk parameters: {0.9, 0.5, 0.1}.
- Experimental Results
  - Smaller risk parameters reduced tail risk (very low rewards).
  - Larger risk parameters increased average rewards but allowed occasional very low rewards.

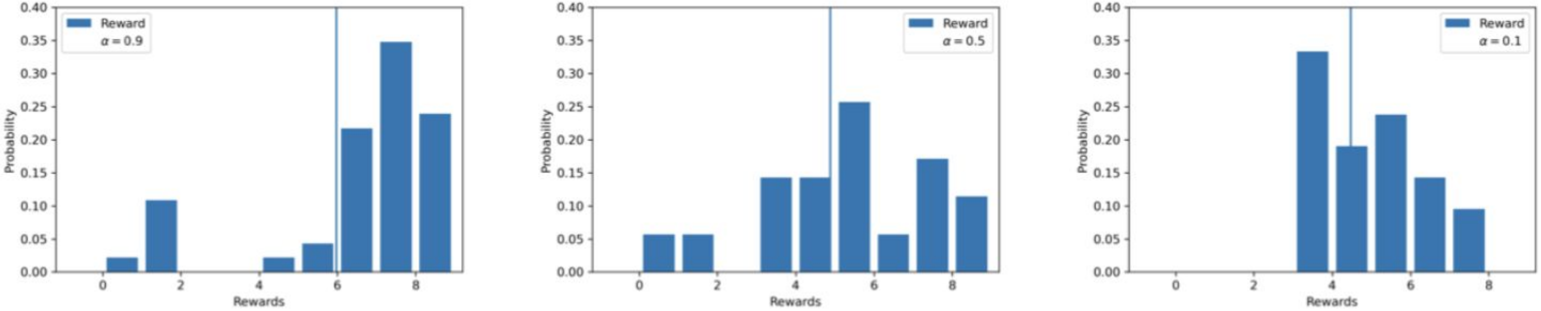


Figure 2: Estimated distribution of the cumulative reward for different risk parameters

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

Experiments - Real-world Experiment: Trading

- Problem Setup
  - A simplified foreign exchange trading environment using real EUR/USD exchange rates and volumes from 2017.
  - Trade volume fixed at 10,000 per hour.
  - 2 actions: {buy, sell}
  - State: Current position and signal features (recent historical prices and volumes).

Experiments - Real-world Experiment: Trading

- Experimental Results
  - Fig. 3 shows cumulative terminal wealth over 100 episodes for different risk parameters: {0.9, 0.5, 0.1}
  - Smaller α reduces tail risk

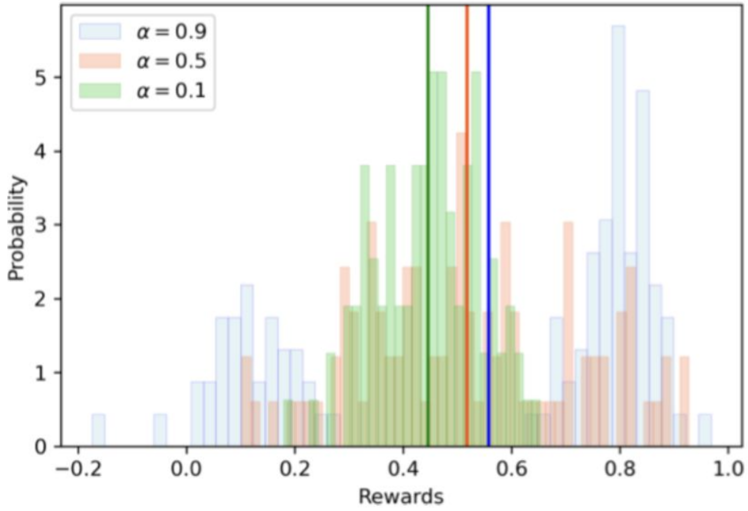


Figure 3. Estimated distribution of the normalized terminal wealth for different risk parameters.

OUTLINE

- Introduction
- Related Work
- Problem Setting
- Risk-aware RL Algorithm with Coherent Risk Measures
- Experiments
  - Synthetic Experiment: Robot Navigation
  - Real-world Experiment: Trading
- Conclusion

Conclusion

- Contribution
  - Proposed a risk-aware RL algorithm called **RA-UCB** that integrates coherent risk measures and non-linear function approximations.
  - Provided a finite-sample regret upper bound guarantee for RA-UCB.
  - Demonstrated the effectiveness of RA-UCB in robot navigation and forex trading environments.
- Future Directions
  - **Relaxing the Weak Simulator Assumption:** Develop methods that do not require a weak simulator
  - **Episodic MDPs with Variable Horizons:** Explore scenarios where episodes have varying lengths or even infinite horizons.