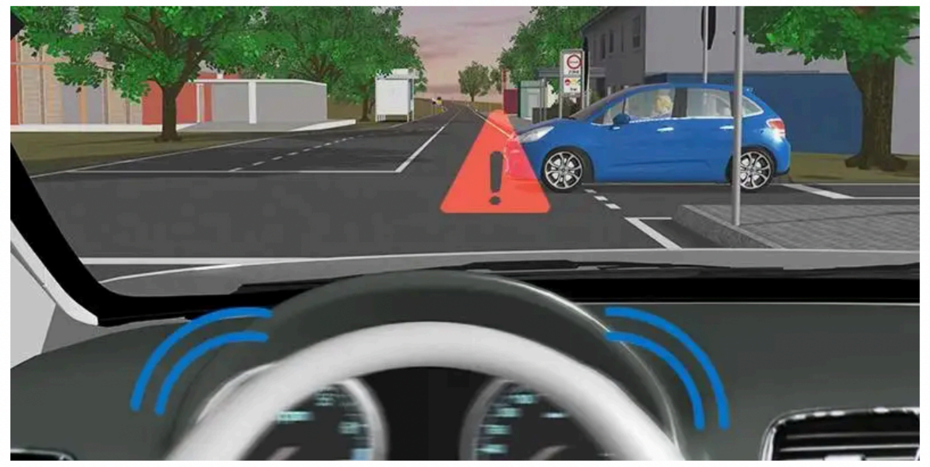
Online Learning in risk-sensitive constrained MDP

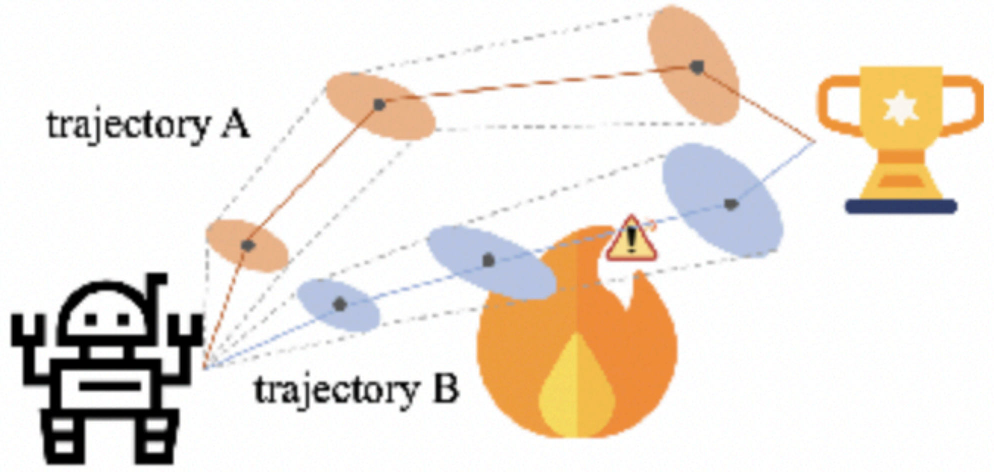
Arnob Ghosh, Assistant Professor at NJIT

Equal Contribution from Mehrdad Moharrami, Assistant Professor at University of Iowa

Decisions in Complex System

- Some Examples:
  - Safe Autonomous vehicle: Reach destination while maintaining safety;
  - Safe Robot navigation: Reach the goal state with minimum steps while avoiding obstacles.
  - Finance: Maximize return while ensuring the portfolio balance is above a certain threshold.





Need to satisfy constraints

Why risk-neutrality is not enough?

- Existing works model as constrained MDP (CMDP): $\{S, A, H, P, r, g\}$
  $\max_{\pi} V^{\pi}_{r,1}(s)$, subject to $V^{\pi}_{g,1}(s) \geq B$
  $V^{\pi}_{r,1}$: $\mathbb{E}_{\pi}[\sum_h r_h(s_h, a_h) \,|\, s_1 = s]$, Expected cumulative Reward,

- $V^{\pi}_{g,1}$: $\mathbb{E}_{\pi}[\sum_h g_h(s_h, a_h) \,|\, s_1 = s]$Expected cumulative utility

- **Markovian Optimal Policy:**
  Common way to achieve a policy, considering Lagrangian:
  $\min_{\lambda} \max_{\pi} V^{\pi}_{r,1}(s_0) + \lambda(V^{\pi}_{g,1}(s_0) - B)$
    - For a given λ, simply solve a RL problem with reward $r + \lambda g$. Tune the dual-variable then.
      Stong duality exists if Slater's condition holds.

- However:
    - **Humans are risk-averse:** Natural to consider risk-averse constraints.
    - For real-life implementation, needs to avoid high-cost (or, low utility) events even when they are rare as they can be catastrophic (e.g, autonomous driving, navigating after natural disaster).

Risk-Constrained MDP

- We consider a risk-constrained MDP.
- $\max_{\pi} V_r^{\pi}(s), \quad$ subject to $V_{g,1}^{\pi}(s) \geq B$,
- Entropic Risk Measure: $V_{g,1}^{\pi}(s) = \frac{1}{\alpha} \log \left[\mathbb{E}_{\pi} e^{\alpha \sum_{h=1}^{H} g_h(s_h, a_h)} \mid s_1 = s \right]$, Risk-aversion $\alpha < 0$:

Key Question: How do you solve the problem?
In the online learning—> Can you minimize Regret while being close to feasibility?

- Challenges:
  - Our result: Markovian Policy on the original state-space is no-longer optimal.
  - The value function is not linear in state-action occupancy measure—> Primal-Dual does not work.
  - Stong Duality may no longer hold.

Our Approach

- Consider Optimized Certainty equivalence (OCE) Representation
  $\mathrm{OCE}_{u,\pi}(s) = \sup_{\tau}\{\tau + \mathbb{E}_{\pi} u(\sum_{h} g_h(s_h, a_h) - \tau)\}$,
- $u(t) = \frac{1}{\alpha}(e^{\alpha t} - 1)$
- For $\alpha < 0$, $\mathrm{OCE}_{u,\pi}(s) = V^{\pi}_{g,1}(s)$.
- Augment the state-space $c_h = \tau - \sum_{h'=1}^{h-1} g_h(s_h, a_h)$, τ—> initial budget.
- Consider Markovian policy with respect to the augmented-space (s_h, c_h).
- $V^{\pi}_{g,1}(s, \tau)$: only depends on the last-state value, $c_{H+1} = \tau - \sum_{h=1}^{H} g_h(s_h, a_h)$, $V^{\pi}_{g,1}(s_1, \tau) = u(-c_{H+1})$.

Augmented Risk-constrained MDP

- $\max_{\pi} V^{\pi}_{r,1}(s, \hat{\tau}),$, subject to $\hat{\tau} = \arg\max\{\tau + V^{\pi}_{g,1}(s, \tau)\}, V_{g,1}(s, \hat{\tau}) \geq B$.
- How do you solve it?
  $\min_{\lambda} \max_{\tau} \max_{\pi} V^{\pi}_{r,1}(s, \tau) + \lambda(\tau + V^{\pi}_{g,1}(s, \tau) - B),$
- Challenge: Continuous augmented state-space as c_h is continuous, problem is not convex in τ.
  - Discretize the space over τ (initial budget) and available budget c_h, and iterate over all possible values of τ to find the maximum.
- How do you update the dual-variable?
  - Gradient-descent: $\lambda \leftarrow \max\{\min\{\lambda + \eta(B - V^{\pi}_{g,1}(s, \tau)), \xi\}, 0\}$

Results

- **Assumption:** There is a Markovian optimal policy on the augmented state-space.
- Regret$(K) = \sum_{k=1}^{K} (V_{r,1}^{\pi^*}(s, \tau^*) - V_{r,1}^{\pi_k}(s, \tau_k))$, Violation$(K) = \sum_{k=1}^{K} (B - \max_{\tau}(\tau + V_{g,1}^{\pi_k}(s, \tau)))$.

With Probability $1 - \delta$, our proposed Algorithm achieves

$$\text{Regret}(K) = \tilde{\mathcal{O}}\left(V_{g,max}K^{3/4} + \sqrt{H^4S^2A\log(1/\delta)K^{3/4}}\right),$$

$$\text{Violation}(K) = \tilde{\mathcal{O}}\left(V_{g,max}K^{3/4}\sqrt{H^3S^2A\log(1/\delta)}\right)$$

- First such result for risk-constrained MDP.
- Regret and Violation bounds are $\tilde{\mathcal{O}}(K^{3/4})$, worse than the CMDP ($\tilde{\mathcal{O}}(K^{1/2})$).
  - **Open Question:** Can we improve it?
- $V_{g,max} = \frac{1}{|\alpha|}\exp(|\alpha|H)$

Simulation Environment

5×5 Grid World

Table 2. Reward matrix $r(i, j)$ for state (i, j)

Row \ Col	0	1	2	3	4
0	0.0	0.1	0.2	0.2	0.1
1	0.5	0.1	1.5	0.5	0.3
2	0.1	0.1	0.4	0.3	0.2
3	0.1	0.1	0.3	0.1	0.6
4	0.1	0.2	0.3	0.1	0.0

Table 3. Utility matrix $u(i, j)$ for state (i, j).

Row \ Col	0	1	2	3	4
0	0.1	0.1	0.2	0.1	0.1
1	0.4	0.2	0.1	0.0	0.0
2	0.3	0.4	1.0	0.0	0.1
3	0.2	0.5	0.4	0.2	0.1
4	0.1	0.1	0.4	0.2	0.0

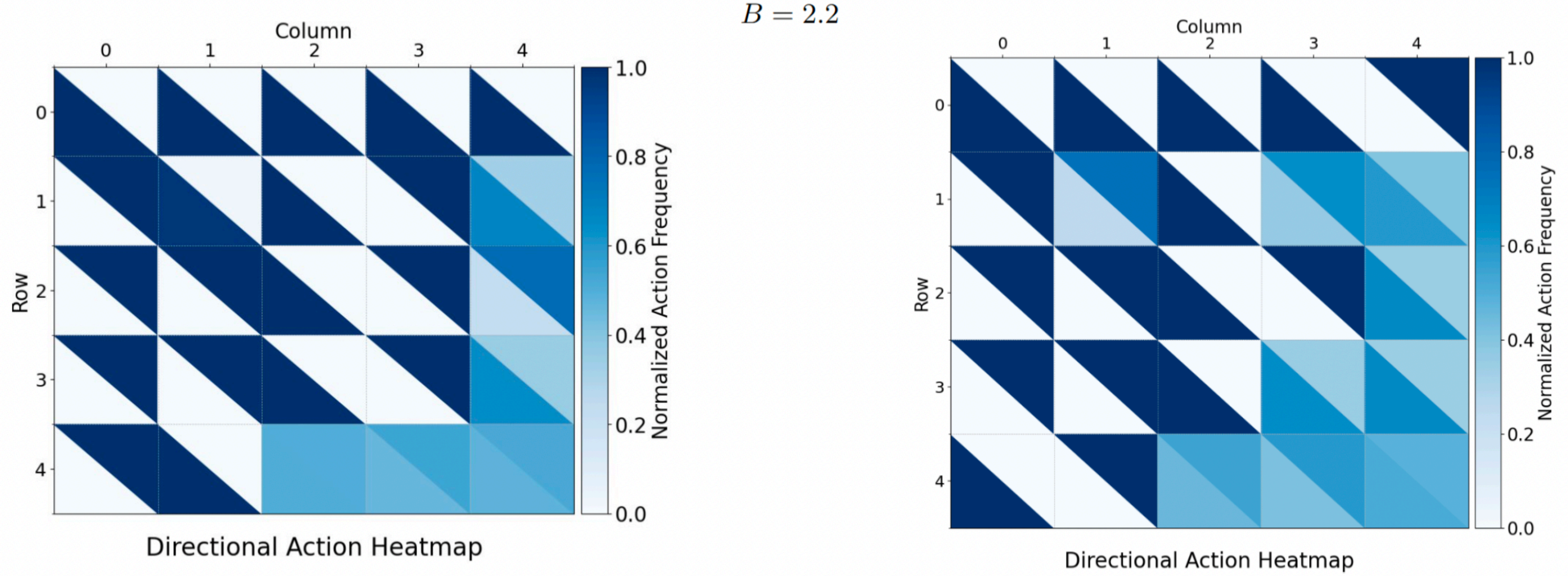
Table 4. Probability matrix $p(i, j)$ representing the likelihood that the action taken in state (i, j) will occur.

Row \ Col	0	1	2	3	4
0	0.9	0.9	0.7	0.5	1.0
1	0.9	0.9	0.5	0.5	1.0
2	0.7	0.9	0.9	0.6	1.0
3	0.9	0.8	0.8	0.5	1.0
4	1.0	1.0	1.0	1.0	1.0

Simulation Results

- α, less negative—> **closer to risk-neutrality, tends to take more riskier option to get a higher reward**.
-

$B = 2.2$



$\alpha = -0.01$ $\alpha = -0.0001$

Summary and Open question

- Risk-constrained MDP is important for practical implementation of RL.
- However, we may not have Markovian optimal policy; can not apply the primal-dual algoritm .
- Augmented state-space and OCE representation can address those problems.
- **Open questions:**
  - Can we extend to other risk-measures?
  - Can we achieve result for stricter violation metrics?