A Game-Theoretic Perspective on Risk-Sensitive Reinforcement Learning

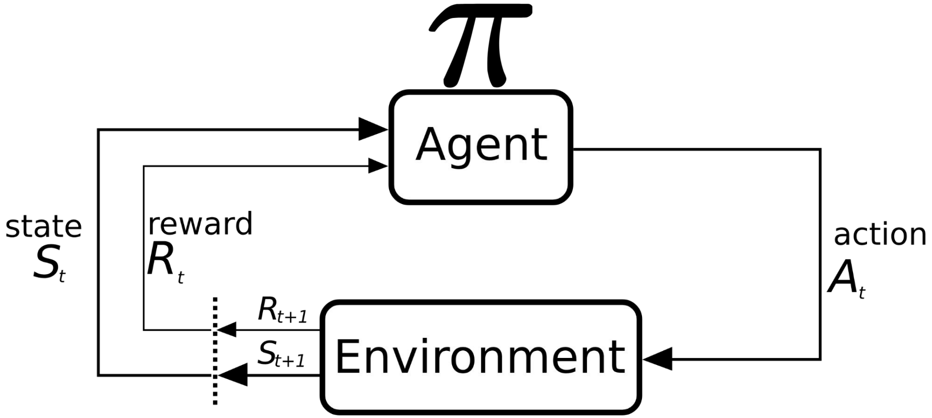
Mathieu Godbout, Maxime Heuillet, Sharath Chandra, Rupali Bhati, Audrey Durand

UNIVERSITÉ LAVAL

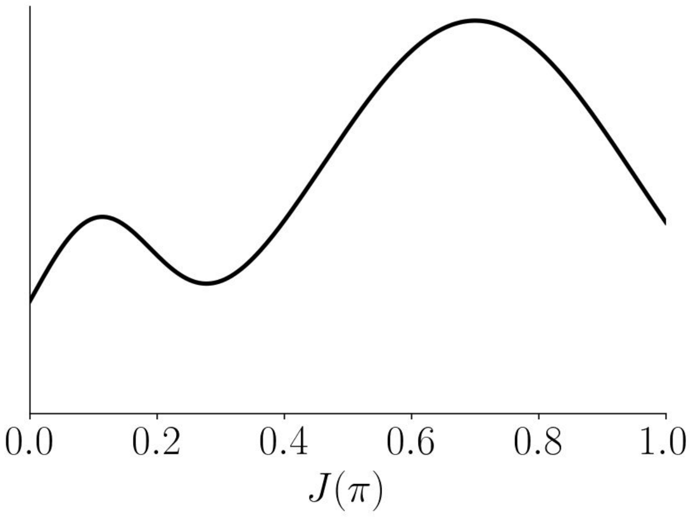
Institut intelligence et données

Mila

Reinforcement Learning (RL)

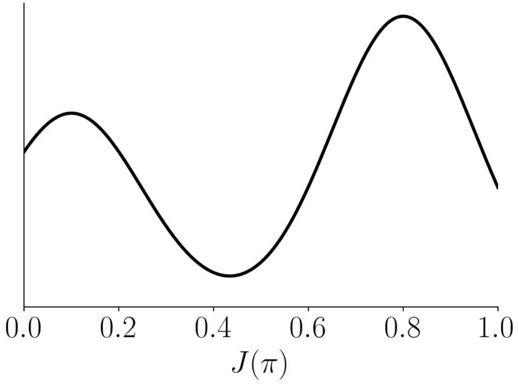


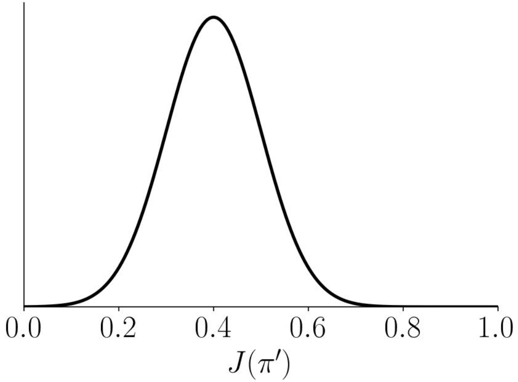
$$J(\pi) := \sum_{t=0}^{\infty} \gamma^t R_t$$



Classical RL objective: expectation maximization

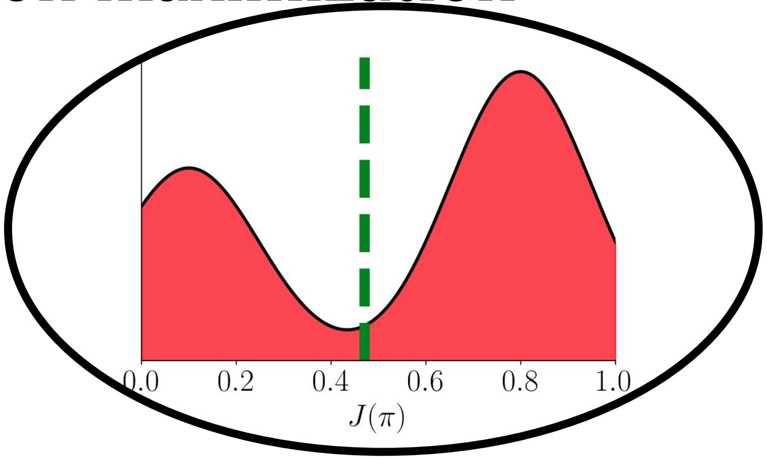
$$\pi^\star = \arg\max_\pi \mathbb{E}[J(\pi)]$$

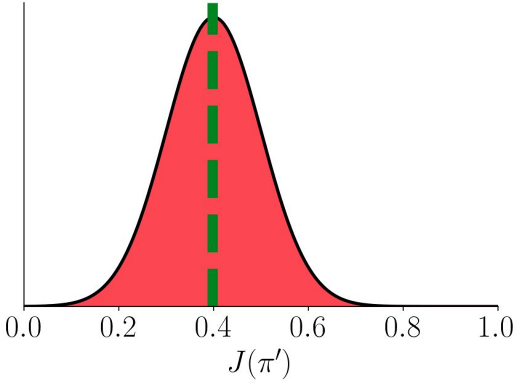




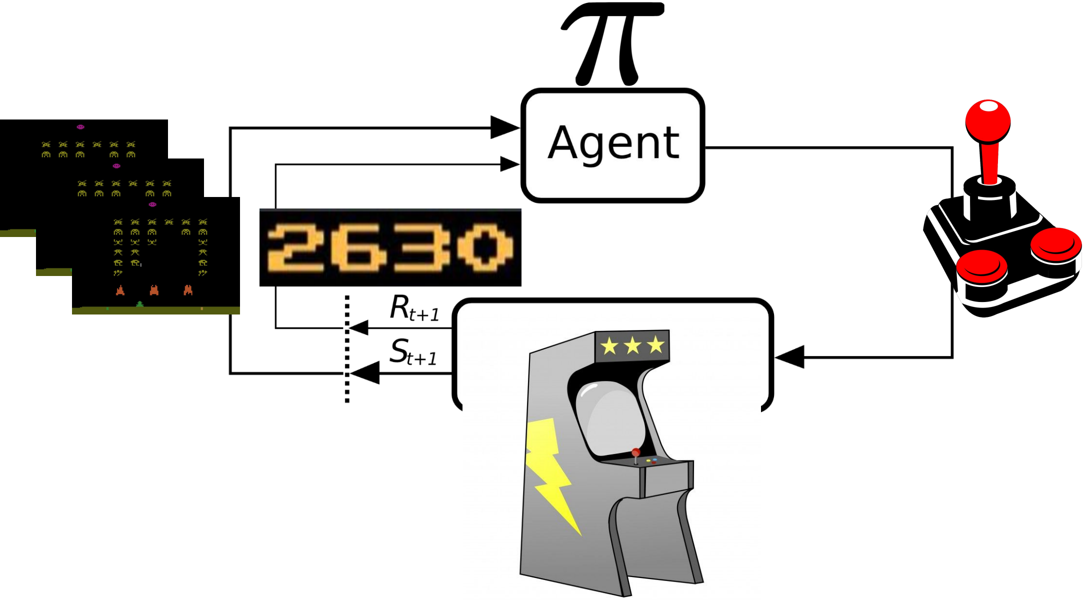
Classical RL objective: expectation maximization

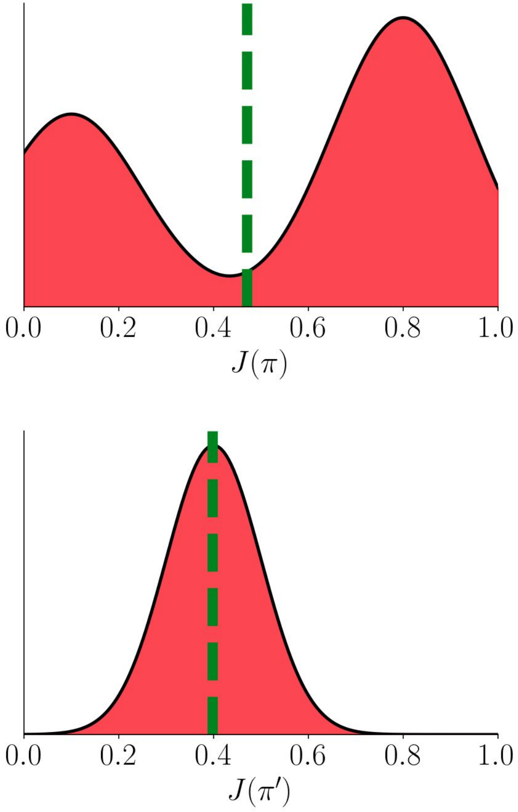
$$\pi^\star = \arg\max_\pi \mathbb{E}[J(\pi)]$$



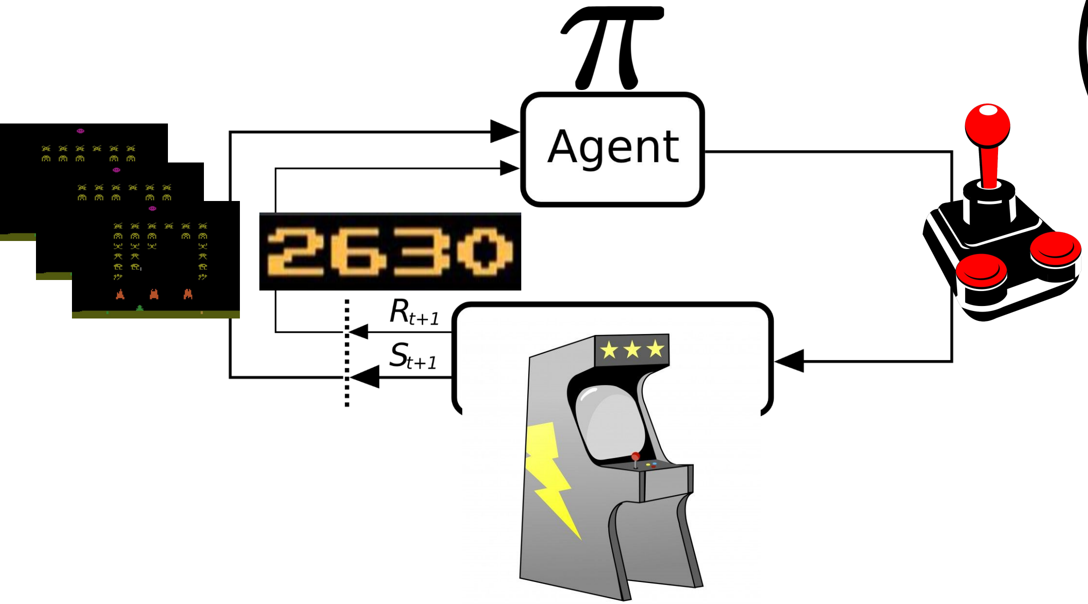


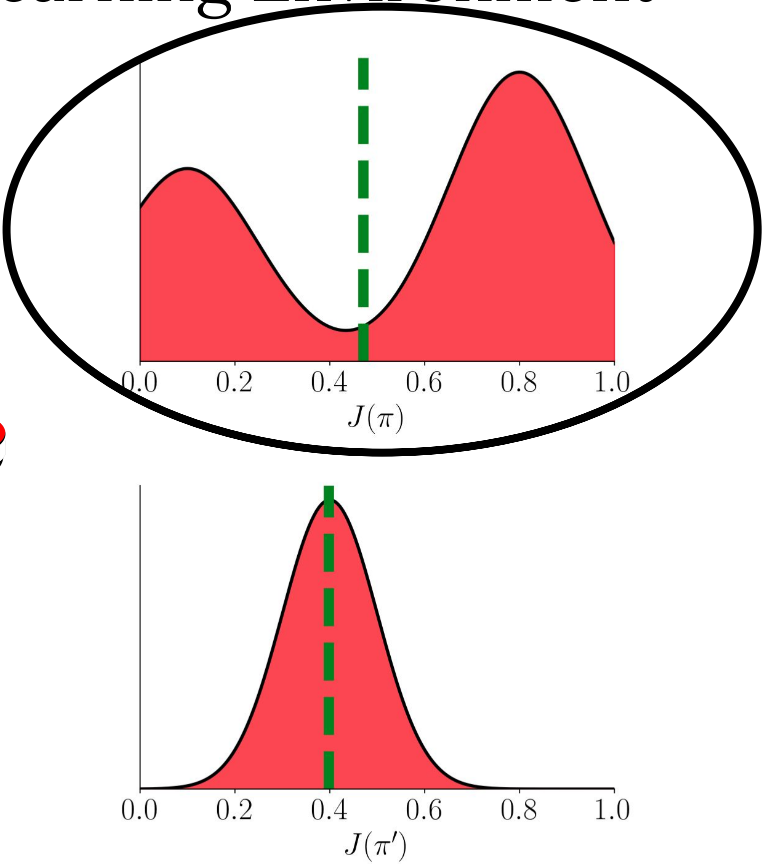
Good objective for the Arcade Learning Environment (Bellemare et al., 2013)



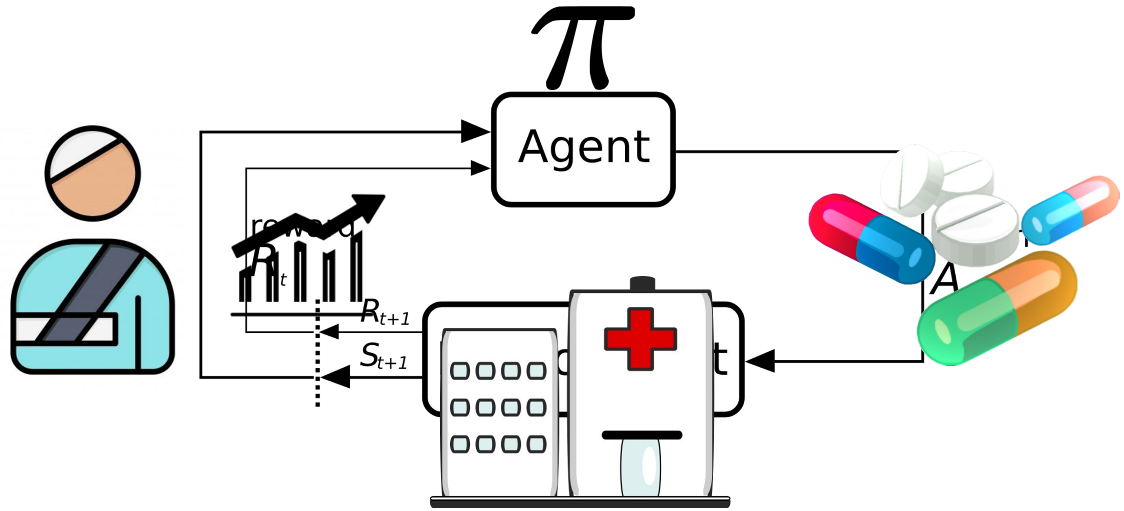


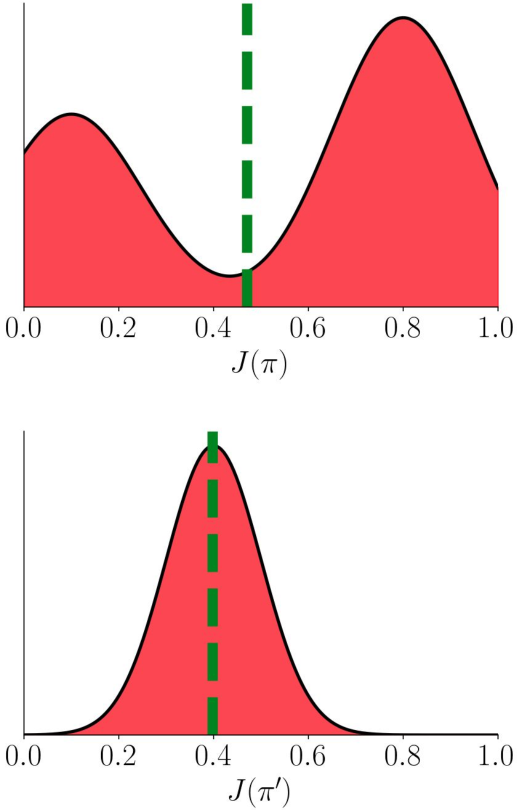
Good objective for the Arcade Learning Environment (Bellemare et al., 2013)



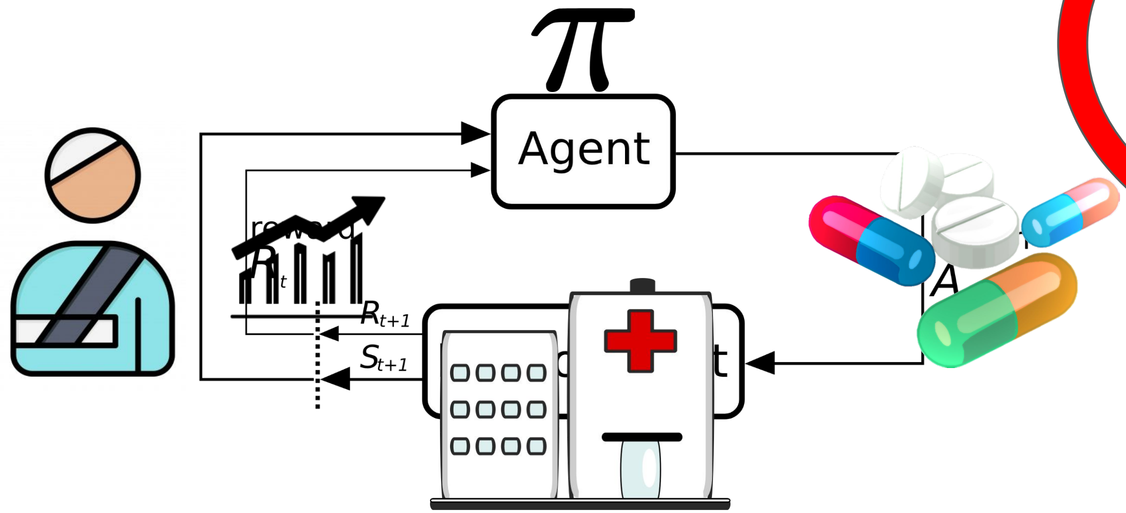


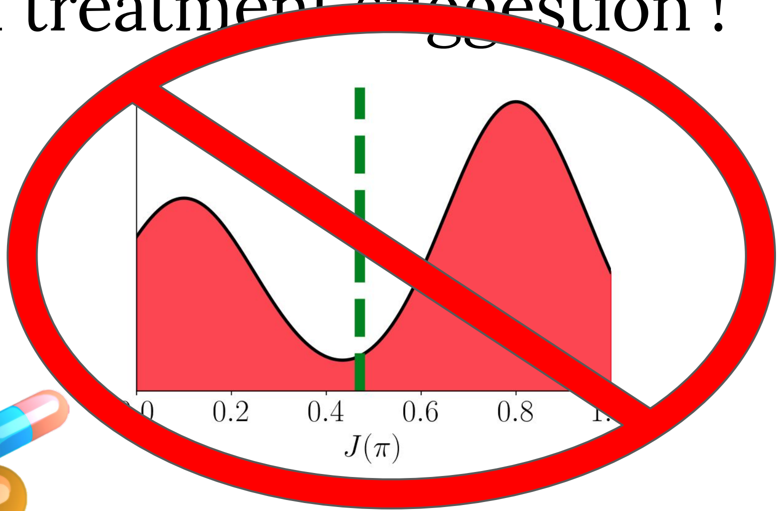
Unfit for a risky task like clinical treatment suggestion !

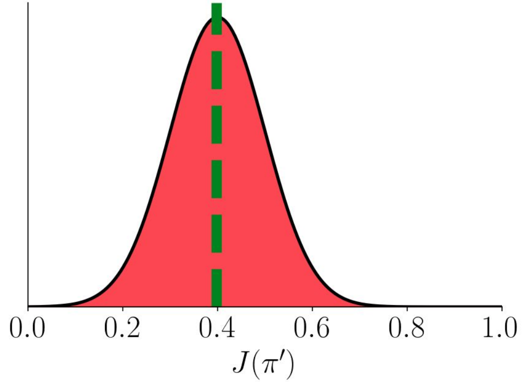




Unfit for a risky task like clinical treatment suggestion !



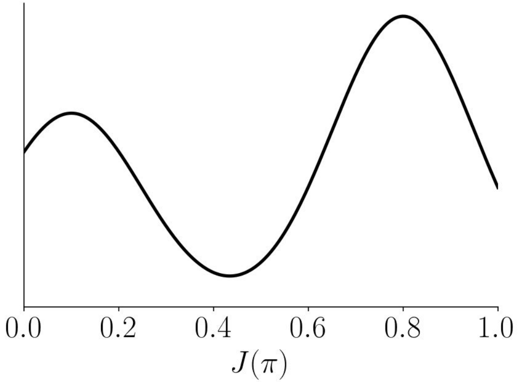


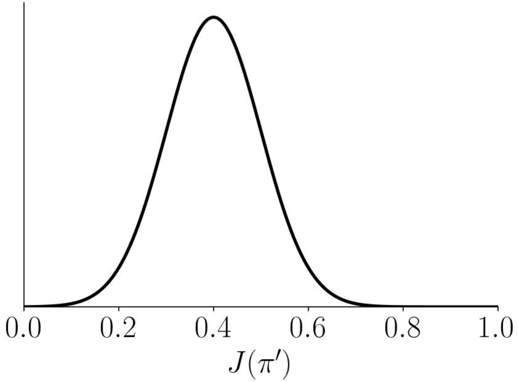


Conditional-value-at-risk RL: a risk-sensitive objective

$$\mathrm{CVaR}_\alpha(Z) = \mathbb{E}[z \mid z \leq \mathrm{VaR}_\alpha(Z)]$$

$$\pi^\star = \arg\max_\pi \mathrm{CVaR}_\alpha[J(\pi)]$$

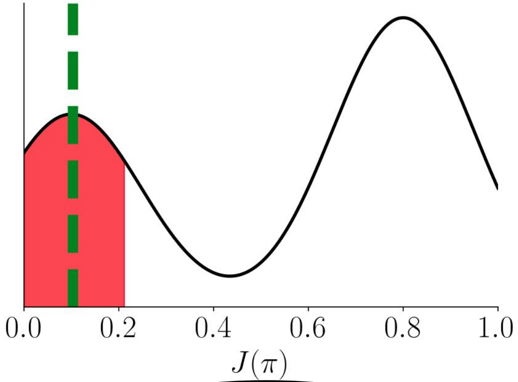


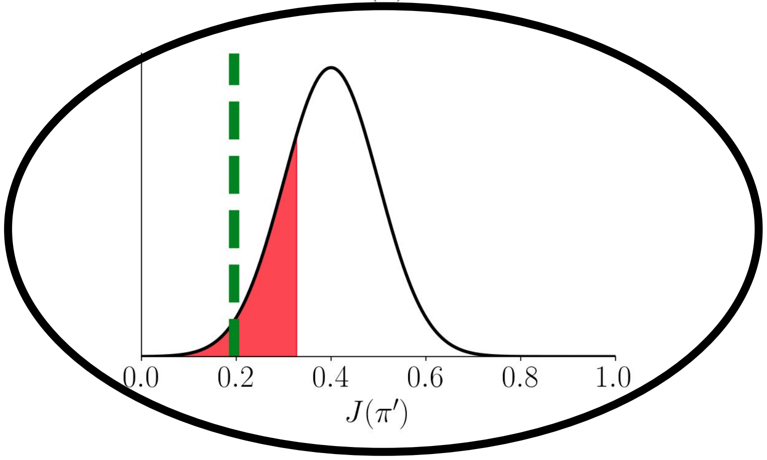


Conditional-value-at-risk RL: a risk-sensitive objective

$$\mathbf{CVaR}_\alpha(Z) = \mathbb{E}[z \mid z \leq \mathbf{VaR}_\alpha(Z)]$$

$$\pi^\star = \arg\max_\pi \mathbf{CVaR}_\alpha[J(\pi)]$$

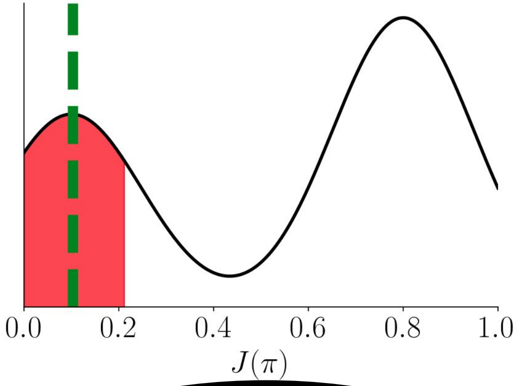


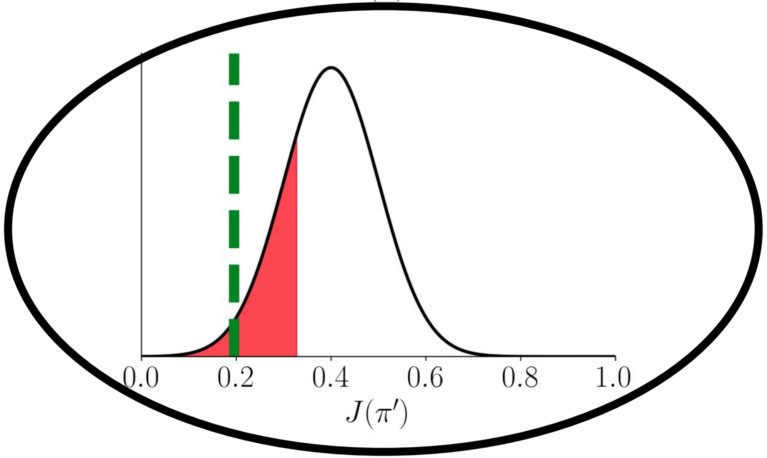


Conditional-value-at-risk RL: a risk-sensitive objective

$$\mathrm{CVaR}_\alpha(Z) = \mathbb{E}[z \mid z \leq \mathrm{VaR}_\alpha(Z)]$$

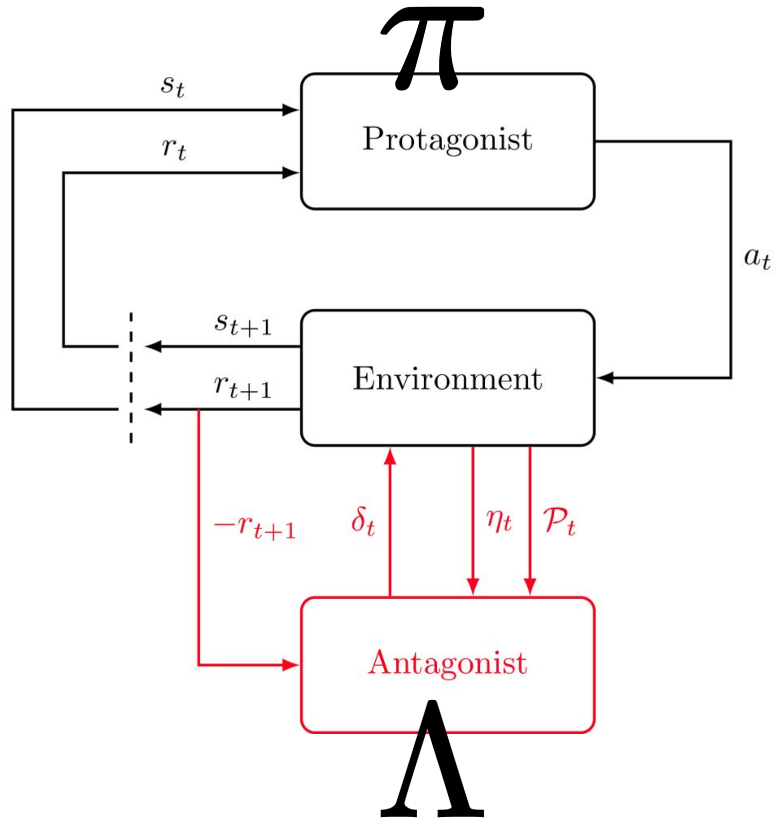
$$\pi^\star = \arg\max_\pi \mathrm{CVaR}_\alpha[J(\pi)]$$





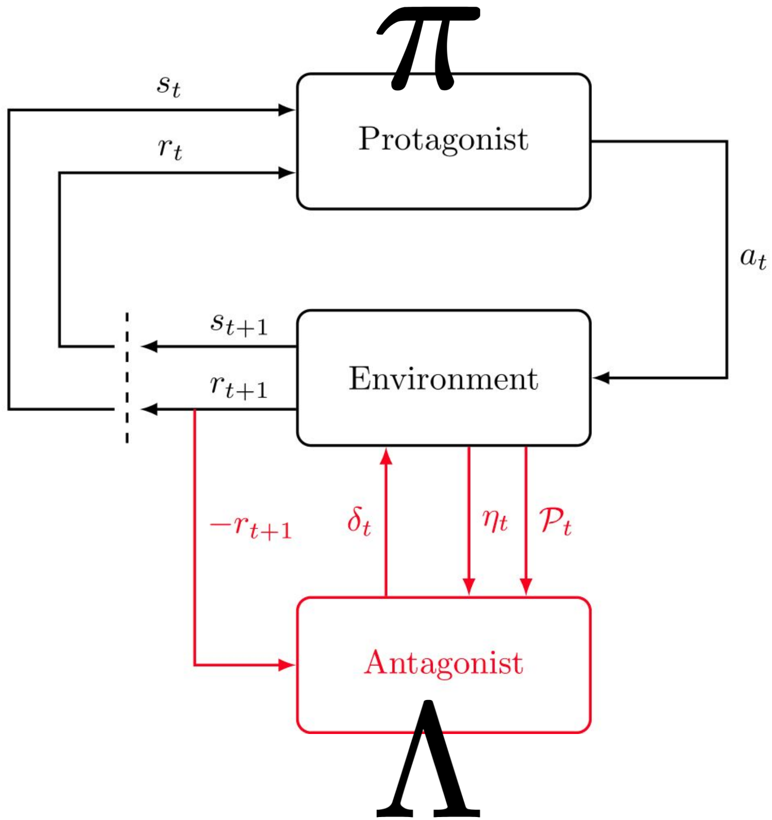
All existing approaches require distributional RL methods.

Game Structure



Antagonist produces perturbed next state transitions $\hat{P}_t$ to minimize the protagonist's rewards

Game Structure

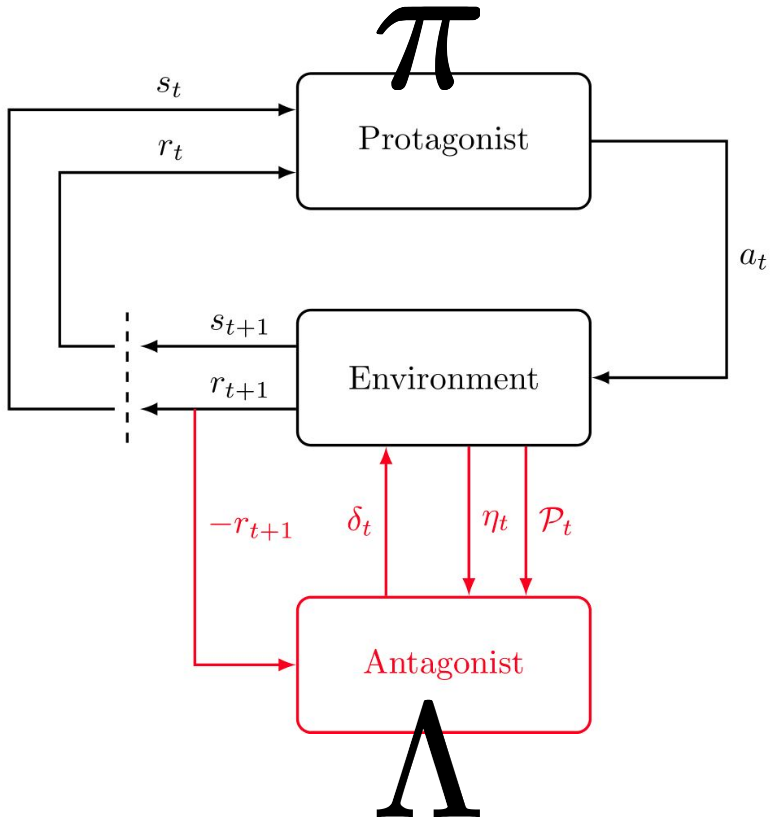


Antagonist produces perturbed next state transitions $\hat{P}_t$ to minimize the protagonist's rewards

Perturbations are multiplicative

$$\hat{P}_t = P_t \circ \delta_t$$

Game Structure

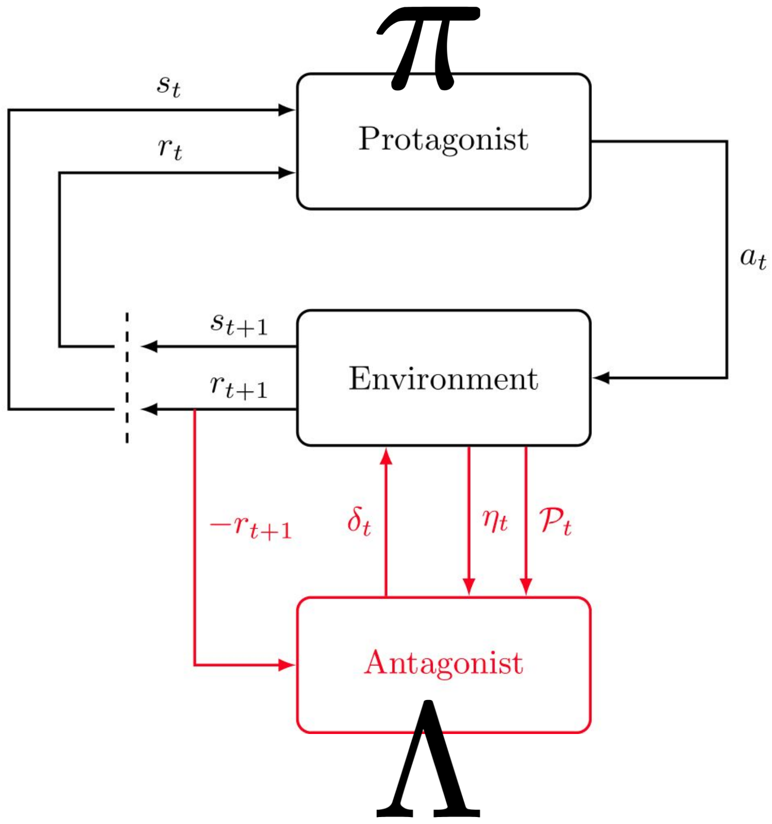


Antagonist produces perturbed next state transitions $\hat{P}_t$ to minimize the protagonist's rewards

Perturbations are multiplicative

$$\hat{P}_t = P_t \circ \delta_t$$

Game Structure



Antagonist produces perturbed next state transitions $\hat{P}_t$ to minimize the protagonist's rewards

Perturbations are multiplicative

$$\hat{P}_t = P_t \circ \delta_t$$

Antagonist has limited budget

$$\delta_0(s_1)\delta_1(s_2)\cdots\delta_{T-1}(s_T) \leq \eta$$

Retrieving CVaR RL optimal policies

Max-min objective: $$\max_{\pi} \min_{\Lambda} \; \mathbb{E}[J^{\eta}(\pi, \Lambda)]$$

Retrieving CVaR RL optimal policies

Max-min objective: $$\max_{\pi} \min_{\Lambda} \mathbb{E}[J^{\eta}(\pi, \Lambda)]$$

The solution is the equilibrium point $(\pi^{\star}, \Lambda^{\star})$, for which we have (Chow et al., 2015):

$$\pi^{\star} = \arg\max_{\pi} \mathrm{CVaR}_{\frac{1}{\eta}}[J(\pi)]$$

Retrieving CVaR RL optimal policies

Max-min objective:

$$\max_{\pi} \min_{\Lambda} \mathbb{E}[J^{\eta}(\pi, \Lambda)]$$

The solution is the equilibrium point $(\pi^\star, \Lambda^\star)$, for which we have (Chow et al., 2015):

$$\pi^\star = \arg\max_{\pi} \mathrm{CVaR}_{\frac{1}{\eta}}[J(\pi)]$$

CARL: Game properties

The objective for both the agent and the adversary is to maximize their **expected** rewards.

CARL: Game properties

The objective for both the agent and the adversary is to maximize their **expected** rewards.

Risk tolerance is based on a single hyperparameter and is easy to interpret.

Stackelberg games for gradient updates

Updating each player naively is unstable due to the non-stationarity of games (Fiez et al., 2019)

Stackelberg games for gradient updates

Updating each player naively is unstable due to the non-stationarity of games (Fiez et al., 2019)

Stackelberg game: a leader (π) takes for granted that its follower (Λ) is optimal with respect to itself.

$$\pi^\star = \arg\max_\pi \left\{ \mathbb{E}\left[J^\eta\left(\pi,\, \Lambda'\right)\right] \text{ s.t. } \Lambda' = \arg\max_\Lambda \mathbb{E}[J^\eta(\pi,\, \Lambda)] \right\}$$

$$\Lambda^\star = \arg\max_\Lambda \mathbb{E}[J^\eta(\pi,\, \Lambda)]$$

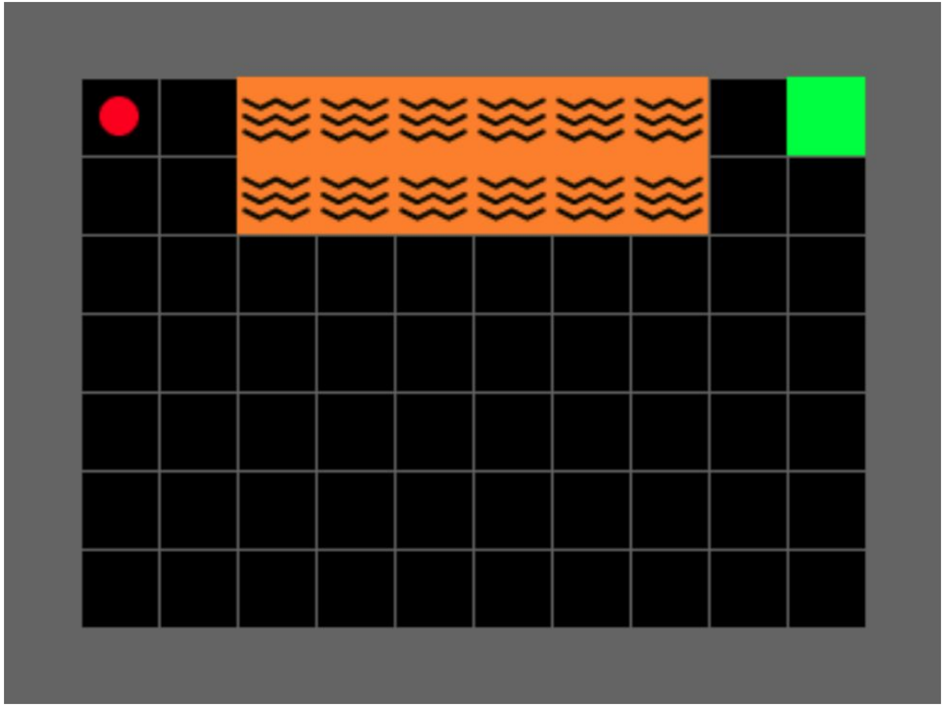
Practical Stackelberg-based algorithm

Algorithm 1: CVaR Adversarial Stackelberg Algorithm

Require: π_θ (protagonist), Λ_ω (antagonist), η (perturbation budget), K_{ant} (number of intermediate antagonist steps)

$N_{\text{updates}} = 0$
while training not done **do**
  Get initial state s_t
  $\eta_\tau = \eta$ ▷ Remaining antagonist budget
  while s_t not terminal **do**
    $a_t \sim \pi_\theta(s_t), \mathcal{P}_t = \mathcal{P}(s_t, a_t)$
    $\delta_t = \Lambda_\omega(\mathcal{P}_t, \eta_\tau)$
    $\hat{\mathcal{P}}_t = \mathcal{P}_t \circ \delta$
    $s_{t+1} \sim \hat{\mathcal{P}}_t, r_{t+1} \sim \mathcal{R}(s_{t+1})$
    $\eta_\tau = \frac{\eta_\tau}{\delta_t(s_{t+1})}$ ▷ Update remaining budget
  end while
  Update θ or ω according to N_{updates} and K_{ant}.
  $N_{\text{updates}} = N_{\text{updates}} + 1$
end while

Risky Gridworld: experimental setting

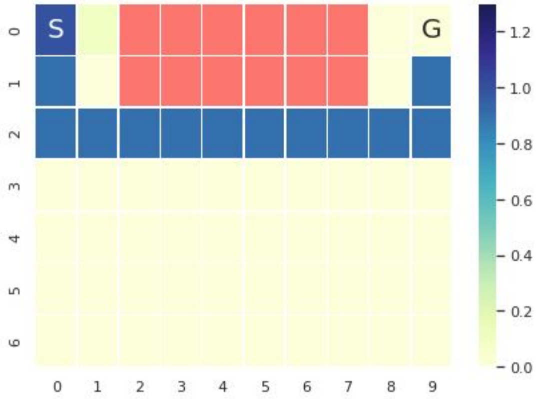


5 % chance that the environment executes a random action.

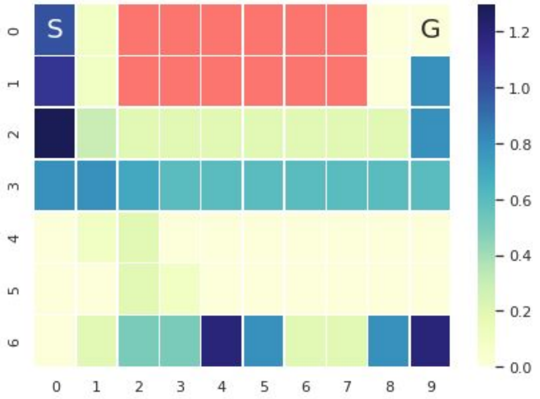
The agent's degree of caution is represented by its willingness to move lower on the grid to distance itself from the lava tiles.

Empirical results

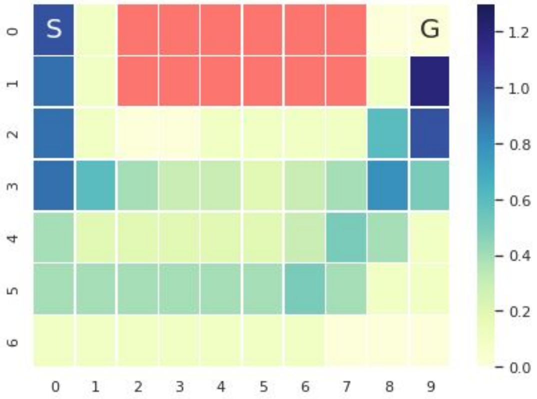
$\eta = 1$



$\eta = 25$

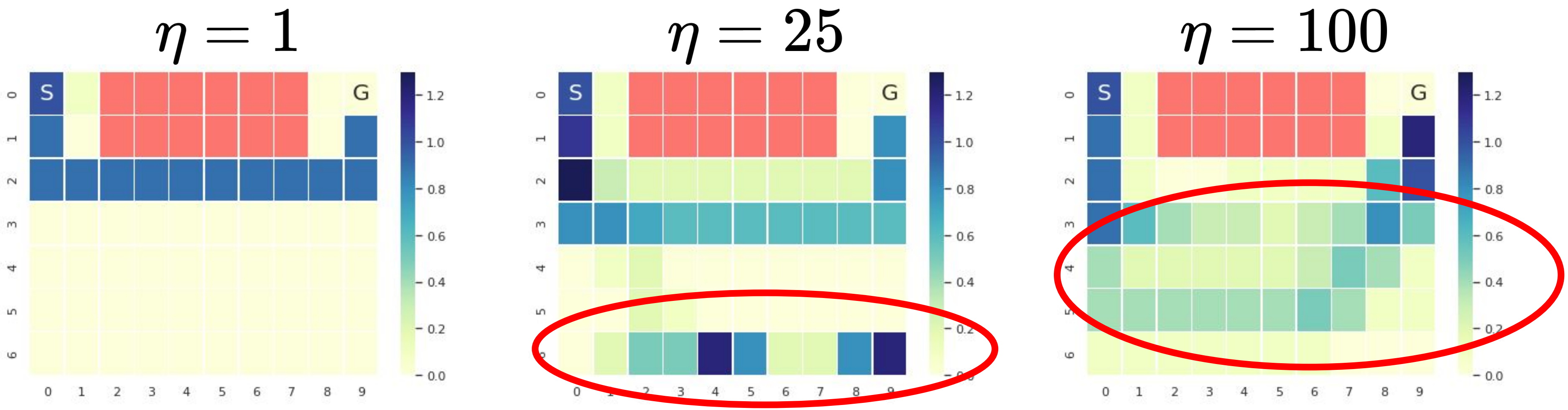


$\eta = 100$



Increasing the adversary’s budget leads to an increasingly cautious agent.

Empirical results



Increasing the adversary's budget leads to an increasingly cautious agent.

There appears to be instability issues in the training procedure.

Conclusion

We proposed a new risk-sensitive RL method for the CVaR risk measure which does not require distributional RL algorithms.

Conclusion

We proposed a new risk-sensitive RL method for the CVaR risk measure which does not require distributional RL algorithms.

We estimate that our proposal can serve as a building block because it paves the way to incorporate results from the Game Theory litterature to risk-sensitivity in RL.

References

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47, 253-279.

Chow, Y., Tamar, A., Mannor, S., & Pavone, M. (2015). Risk-sensitive and robust decision-making: a cvar optimization approach. arXiv preprint arXiv:1506.02188.

Fiez, T., Chasnov, B., & Ratliff, L. J. (2019). Convergence of learning dynamics in stackelberg games. arXiv preprint arXiv:1906.01217.