UNIVERSITÉ DE LYON

ENS DE LYON

Beyond Average Return in Markov Decision Processes

Alexandre Marthe[1], Aurélien Garivier[1], Claire Vernade[2]

[1] ENS de Lyon, [2] University of Tuebingen

FRANCE 2030 PROGRAMME DE RECHERCHE INTELLIGENCE ARTIFICIELLE

cnrs

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Setting

We consider the setting of **Undiscounted Finite-Horizon** Tabular Markov Decision Processes $\mathcal{M}(\mathcal{X}, \mathcal{A}, P, R, H)$.
We are interested in the **Return** seen as a random variable :

$$\mathcal{Z}^\pi(x) = \sum_{h=0}^{H} R_h, \quad X_0 = x$$

Most of the literature focuses on optimizing the **Expected Return** :

$$\max_\pi \mathbb{E}[\mathcal{Z}^\pi(x)]$$

→ can we optimize other *risk-aware* functionals ψ of the Return, such as CVaR, quantiles, etc?

$$\max_\pi \psi(\mathcal{Z}^\pi(x))$$

Motivation

For some environments with a complex Return distribution, optimizing the mean might be arbitrary. The functional ψ can be a risk measure of the Return, leading to risk-dependent strategies.

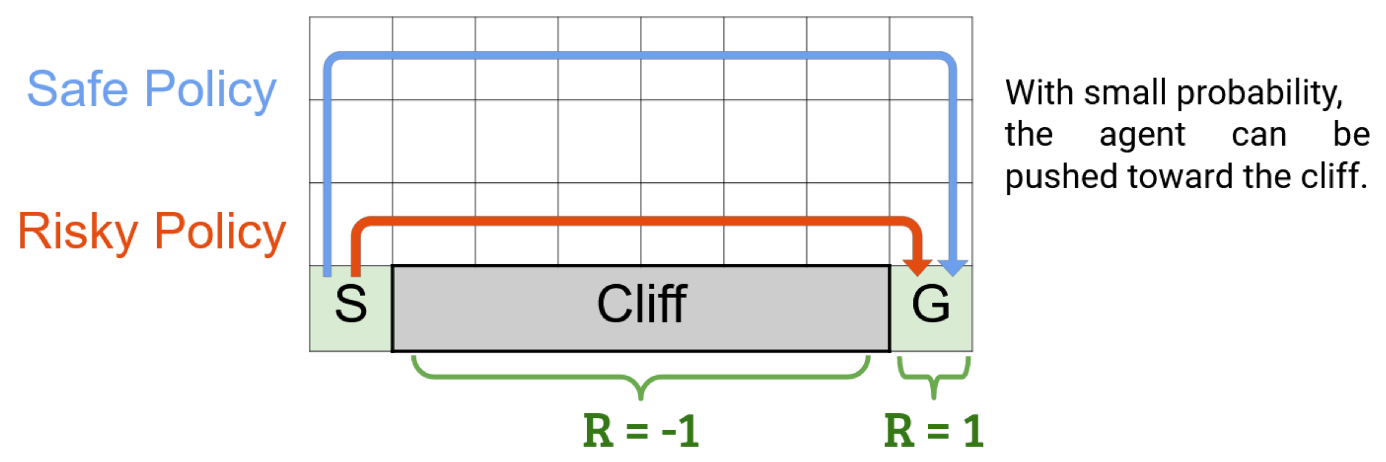

Figure 1: Cliff MDP where the agent should go from S to G.

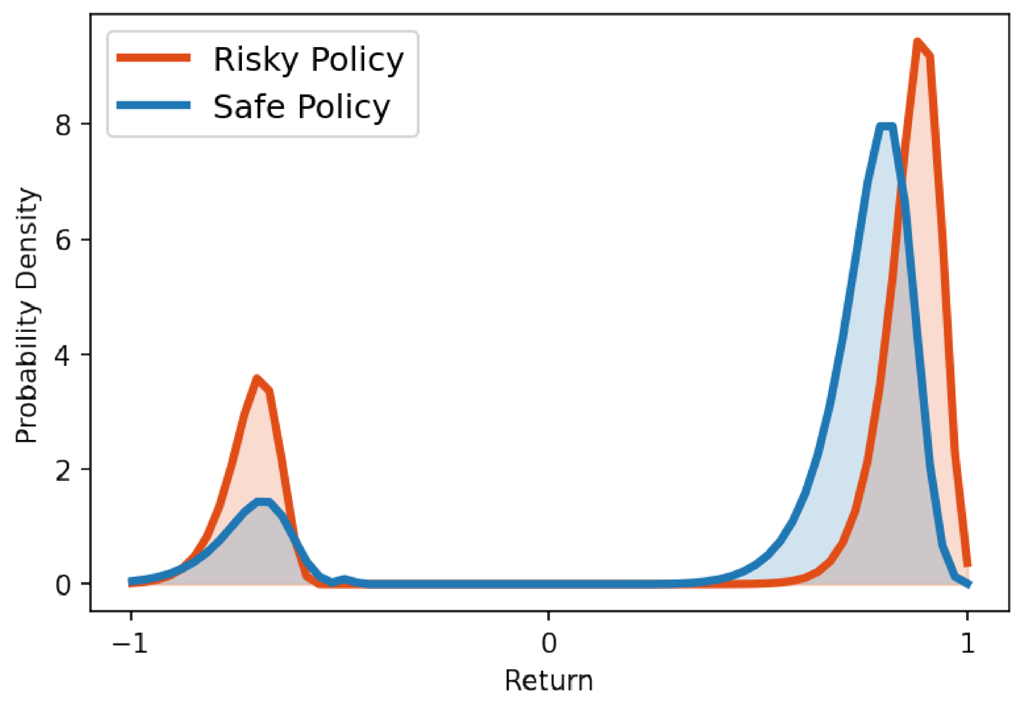

Figure 2: Distribution of the return for the policy in Cliff.

Can such policy be computed through dynamic programming ?

Distributional RL

→ Objective of Distributional RL: Estimate the whole distribution of the return, instead of only its expectation.

→ There exists a Distributional Bellman Equation:

$$Z_h^\pi(x,a) = R_h + Z_{h+1}^\pi(X', A')$$

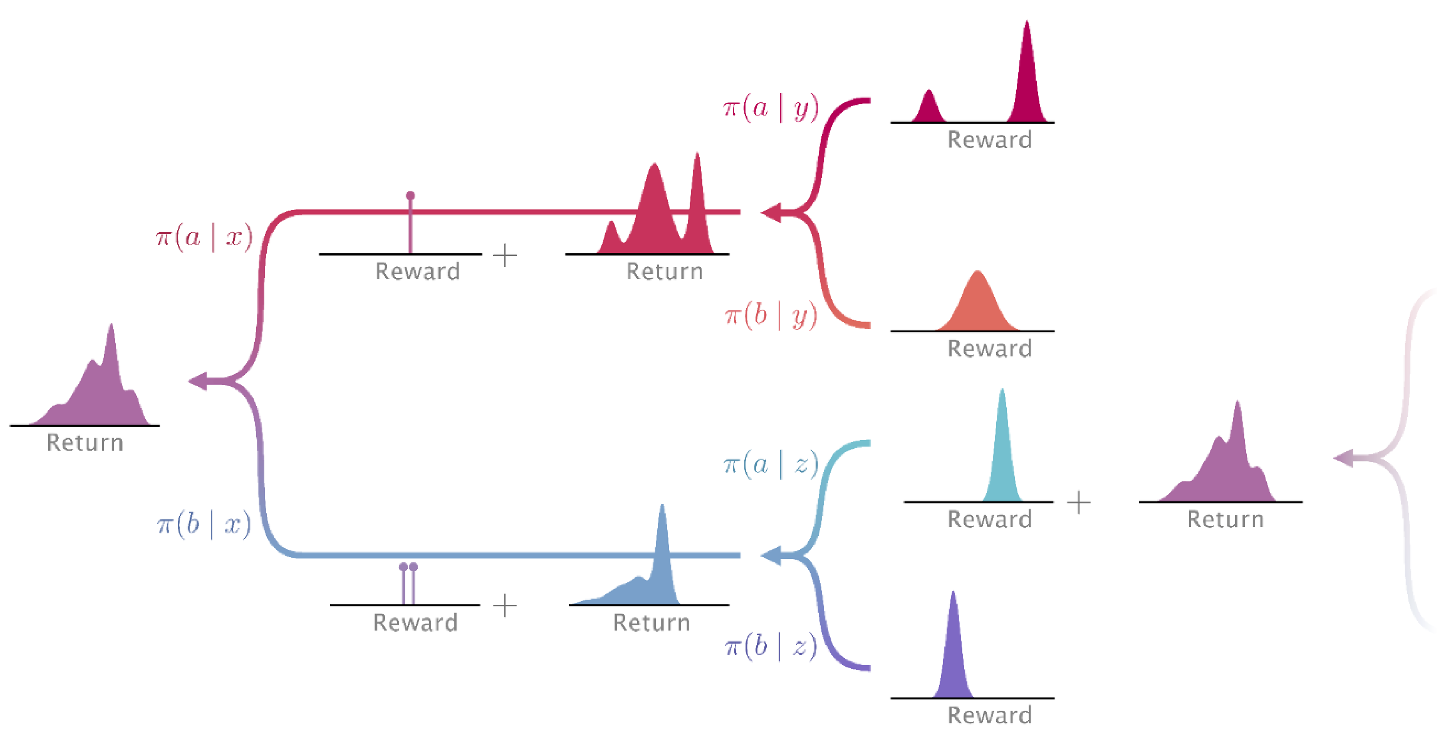

Figure 3: Distributional Dynamic Programming. [Bellemare et. al, 2023]

→ Convenient tool for Dynamic Programming: $\forall h, x$, choose $a^* = \mathrm{argmax}_a \psi(Z_h^\pi(x,a))$. From $h = H$ down to 0, compute recursively a policy π_{DP}

Question

What functionals of the return can be *optimized exactly* through Dynamic Programming ?

Main Result

The only continuous **Bellman Optimizable** functionals are:

- The Expected Return : $\mathbb{E}[Z]$
- The Exponential Utilities : $\mathbb{E}[\exp(\lambda Z)]$

→ The statistics optimizable through Distributional RL are the same as with Classical RL.

→ Exponential utilities allow for risk-dependant strategies
e.g. if $Z \sim \mathcal{N}(\mu, \sigma)$, $\quad \lambda^{-1} \ln \mathbb{E}[\exp(\lambda Z)] = \mu + \lambda\sigma$

Definition : Bellman Optimizable

A functional ψ is said to be *Bellman Optimizable* if π_{DP} is optimal for any Markov Decision Processes.

Dynamic Programming

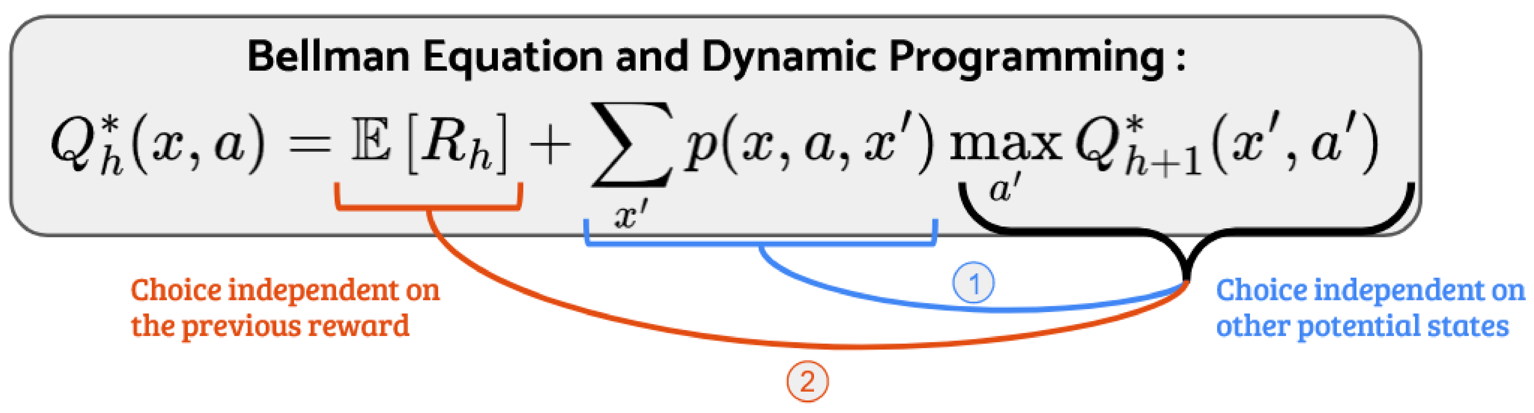

Figure 4: Dynamic Programming for Expected Return

The distributional form, using $Z_h^{\pi_{\text{DP}}}(x,a) \sim \eta_h^{\pi_{\text{DP}}}(x,a)$ and $R_h(x,a) \sim \varrho_h(x,a)$ is:

$$\psi(\eta_h^{\pi_{\text{DP}}}(x,a)) = \psi\left(\varrho_h(x,a) * \sum_{x'} p_h(x,a,x') \eta_{h+1}^{\pi_{\text{DP}}}(x', a^*_{h+1,x'})\right)$$

Bellman Optimizable Properties

A Bellman Optimizable functional ψ necessarily verifies 2 properties:

- *1. Independence Property*: Mixing in other distributions should not change the choice of action.

$$\psi(\nu_1) \geq \psi(\nu_2) \implies \forall \nu_3, \forall \lambda, \psi(\lambda\nu_1 + (1-\lambda)\nu_3) \geq \psi(\lambda\nu_2 + (1-\lambda)\nu_3))$$

→ allows to apply the Expected Utility Theorem: ψ can be written in the form $E[f(\cdot)]$ for some f.

- *2. Translation Property*: Translating by a constant should not change the choice of action.

$$\psi(\nu_1) \geq \psi(\nu_2) \implies \forall c, \quad \psi(\nu_1(\cdot + c)) \geq \psi(\nu_2(\cdot + c))$$

→ allows to find a differential equation verified by f. The solutions are the Exponential and Linear functions.

Important References

- Bellemare, M. G., Dabney, W., & Rowland, M. (2023). *Distributional reinforcement learning*. MIT Press.
- R. A. Howard and J. E. Matheson. *Risk-Sensitive Markov Decision Processes*. 1972.

NEURAL INFORMATION PROCESSING SYSTEMS

anr agence nationale de la recherche