Beyond Average Return in Markov Decision Processes

Alexandre Marthe[1], Aurélien Garivier[1], Claire Vernade[2]
[1] ENS de Lyon, [2] University of Tuebingen

Setting

Undiscounted Finite-Horizon Tabular Markov Decisions Processes $\mathcal{M}(\mathcal{X}, \mathcal{A}, P, R, H)$:

- $\mathcal{X}$ and $\mathcal{A}$ the finite *State Space* and *Action Space*
- P the *Transition Kernel*: $x_{h+1} \sim p_h^{a_h}(x_h, \cdot)$
- R the *Random Reward* of distribution ρ: $r_h \sim \rho_h^{(x_h,a_h)}$, bounded with Δ_R
- H the horizon

We write $Z_h^\pi(s,a) = \sum_{i=h}^H r_i \mid s_h = s, a_h = a, a_i \sim \pi(s_i)$ the return, and $\eta_{h,\pi}^{(s,a)}$ its distribution.

Distributional RL

→ Objective of Distributional Reinforcement Learning : Estimate the whole distribution of the return, instead of only the Expectation.

→ There exists a Distributional Bellman Equation:

$$\forall x, a, h, \quad \eta_{\pi,h}^{(x,a)} = \rho_h^{(x,a)} * \sum_{x'} p_h^a(x,x') \eta_{\pi,h+1}^{(x,\pi_{h+1}(x))} \;. \qquad (1)$$

→ In practice distributions are not tractable : they need to be parametrized.

Algorithm 1 Parametrized Policy Evaluation for Distributional RL

Input: model p, reward distributions ρ_h, policy π to evaluated, Π projection.
Data: $\eta \in \mathbb{R}^{H|\mathcal{X}||\mathcal{A}|N}$
$\forall x, a \in \mathcal{X} \times \mathcal{A}, \quad \eta_H^{(x,a)} = \delta_0$
for $h = H-1 \to 0$ **do**
 $\eta_h^{(x,a)} = \rho_h(x,a) * \sum_{x'} p_h^a(x,x') \eta_{h+1}^{(x',\pi_{h+1}(x'))} \quad \forall x, a \in \mathcal{X} \times \mathcal{A}$
 $\eta_h^{(x,a)} = \Pi\left(\eta_h^{(x,a)}\right) \quad \forall x, a \in \mathcal{X} \times \mathcal{A}$
end for
Output: $\eta_h^{(x,a)} \forall x, a, h$

Objective

(i) Which functionals can be exactly optimized through Bellman Dynamic Programming?

(ii) How accurately can we evaluate statistical functionals by using DistRL?

Exact Planning and Bellman Optimization

Algorithm 2 Pseudo-Algorithm: Exact Planning with Distributional RL

Input: model p, reward R, statistical functional s
Data: $\eta \in \mathbb{R}^{H|\mathcal{X}||\mathcal{A}|N}, \nu \in \mathbb{R}^{H|\mathcal{X}|N}$
$\forall x \in \mathcal{X}, \quad \nu_{H+1}^x = \delta_0$
for $h = H \to 1$ **do**
 $\eta_h^{(x,a)} = \rho_h^{(x,a)} * \sum_{x'} p_h^a(x,x') \nu_{h+1}^{x'} \quad \forall x, a \in \mathcal{X} \times \mathcal{A}$
 $\nu_h^x = \eta_h^{(x,a^*)}, \quad a^* \in \text{argmax}_a s(\eta_h^{(x,a)}) \quad \forall x \in \mathcal{X}$
end for
Output: $\eta_h^{(x,a)} \forall x, a, h$

→ Intuition : Dynamic Programming is used to compute distributions of the return. The actions are chosen to optimize the statistic at every timestep h.

Definition. A statistical functional s is said *Bellman Optimizable* if Algorithm 2 outputs an optimal distribution for s:

Results

A Bellman Optimizable statistical functional necessarily verifies 2 properties:

- *Independence Property*: If $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$ are such that $s(\nu_1) \geq s(\nu_2)$, then
$$\forall \nu_3 \in \mathscr{P}(\mathbb{R}), \forall \lambda \in [0,1], \quad s(\lambda\nu_1 + (1-\lambda)\nu_3) \geq s(\lambda\nu_2 + (1-\lambda)\nu_3)) \;.$$
- *Translation Property*: Let τ_c denote the translation on the set of distributions: $\tau_c \delta_x = \delta_{x+c}$. If $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$ are such that $s(\nu_1) \geq s(\nu_2)$, then
$$\forall c \in \mathbb{R}, \quad s(\tau_c \nu_1) \geq s(\tau_c \nu_2) \;.$$

Theorem 2. The only Bellman Optimizable statistical functionals are exponential utilities $U_{\exp} = \frac{1}{\lambda} \log \mathbb{E}\left[\exp(\lambda R)\right]$ for $\lambda \in \mathbb{R}$, with the special case of the expectation $\mathbb{E}\left[R\right]$ when $\lambda = 0$.

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

→ Policy-improvement-like algorithms may only work exactly with the exponential utilities.

Approximate Policy Evaluation

- s a statistic of the form $s(\eta) = \mathbb{E}_{Z\sim\eta}\left[f(Z)\right]$ or $s(\eta) = \mathbb{E}_{\tau\sim\mathcal{U}(0,1)}\left[\beta'(\tau) F_\eta^{-1}(\tau)\right]$, β or f L-Lipschitz.
- The Quantile Parametrization with Resolution N : $\Pi(\eta) = \frac{1}{N}\sum \delta_{z_i}$ with $z_i \in F_\eta^{-1}(\frac{2i+1}{2N})$

Theorem 1. Let $\hat{\eta}_\pi$ be the approximated return distribution computed with Algorithm 1. Then, the error with the computed statistic is bounded:

$$\sup_{x,a,h} |s(\hat{\eta}_{\pi,h}^{(x,a)}) - s(\eta_{\pi,h}^{(x,a)})| \leq LH^2 \frac{\Delta_R}{2N} \;.$$

→ The error is up to *quadratic* in the horizon.

Experimental Validation

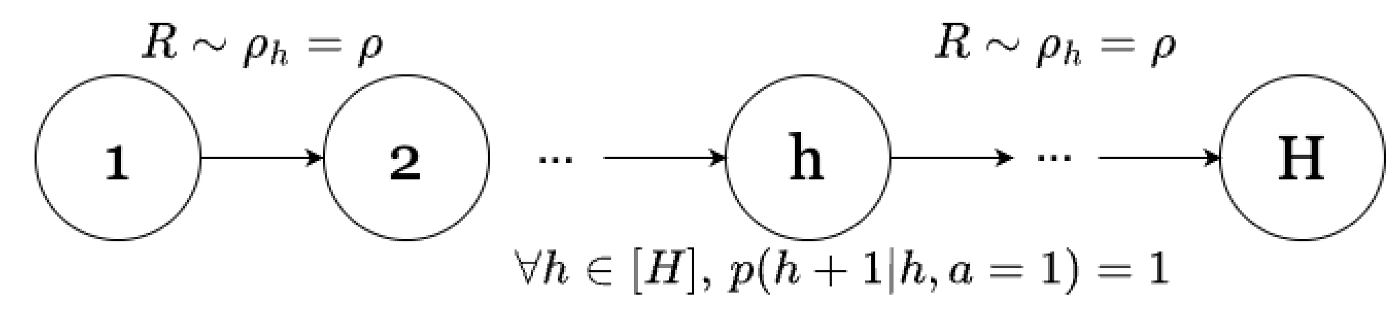

Figure 1: Chain MDP with a stationnary reward $\mathcal{B}(0.5)$.

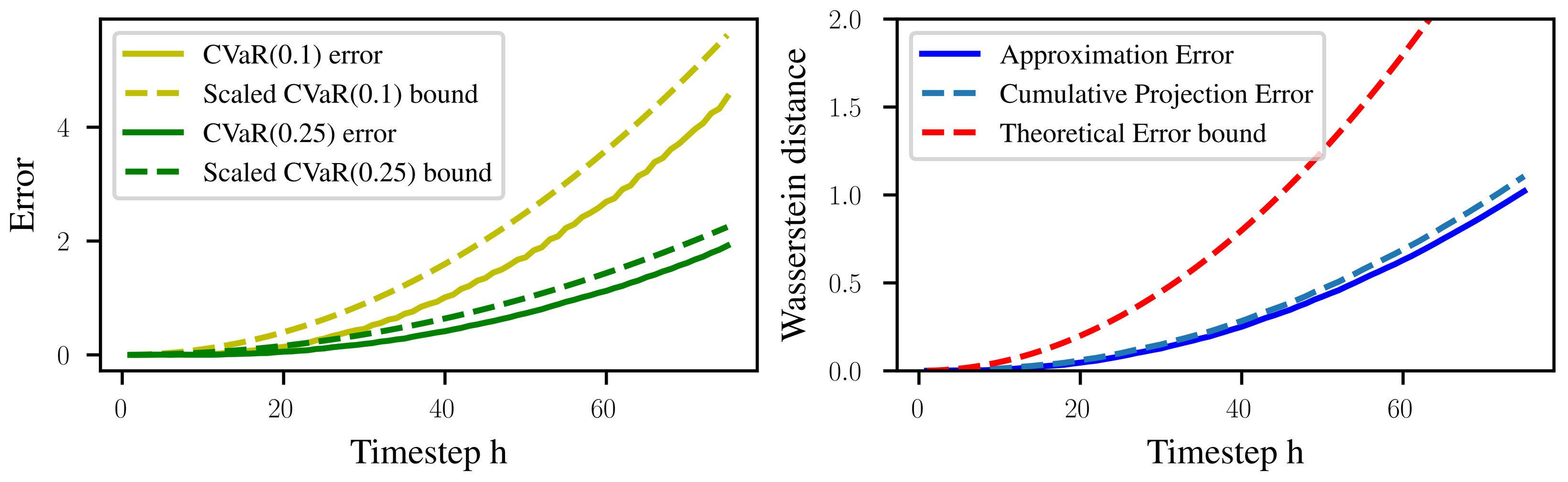

Figure 2: Right : The Wasserstein Error is the sum of the successive Projection Errors. Left : The CVaR Error is quadratic in the horizon.

Important References

- Bellemare, M. G., Dabney, W., Rowland, M. (2023). *Distributional reinforcement learning*. MIT Press.
- Von Neumann, J., Morgenstern, O. (2007). *Theory of games and economic behavior (60th Anniversary Commemorative Edition)*. Princeton university press.