Stochastically Dominant Distributional Reinforcement Learning

John D. Martin, Michal Lyskawinski, Xiaohu Li, Brendan Englot
Stevens Institute of Technology, Hoboken NJ
Correspondence to jmarti3@stevens.edu, Full paper: https://arxiv.org/abs/1905.07318

Abstract

We describe a new approach for managing aleatoric uncertainty in the Reinforcement Learning (RL) paradigm. Instead of selecting actions according to a single statistic, we propose a distributional method based on the second-order stochastic dominance (SSD) relation. This compares the inherent dispersion of random returns induced by actions, producing a more comprehensive and robust evaluation of the environment's uncertainty. We propose a particle-based algorithm for which we prove optimality and convergence. Our experiments characterize the algorithm performance and demonstrate how uncertainty and performance are better balanced using an SSD policy than with other risk measures.

Distributional RL as WGF

Distributional RL can be cast as a Wasserstein Gradient Flow (WGF). The optimization landscape is shaped by the free energy functional:

$$E(\mu) = \underbrace{F(\mu)}_{\text{potential}} + \overbrace{\beta^{-1}}^{\text{temperature}} \underbrace{H(\mu)}_{\text{entropy}}.$$

The potential describes what it means to be optimal. This is in terms of the distributional Bellman operator (Bellemare et al., 2017):

$$F(\mu) = \frac{1}{2}\int \left(\mathcal{T}z^{(s,a)} - z^{(s,a)}\right)^2 d\mu = \int U(z)d\mu.$$

Discrete Time Solutions are obtained iteratively using the procedure of Jordan et al. (1998). The method discretizes time in steps of τ and applies the proximal operator

$$\text{Prox}^{\mathcal{W}}_{\tau E}(\mu_k) = \underset{\mu\in\mathcal{P}_2(\mu,\mu_k)}{\text{argmin}}\ \mathcal{W}_2^2(\mu,\mu_k) + 2\tau E(\mu).$$

Distance between probability measures is described with the Wasserstein distance

$$\mathcal{W}_2(\mu,\nu) = \inf_{\gamma\in\mathcal{P}_2(\mu,\nu)} \left\{\int_{\mathbb{R}^2} |x-y|^2 d\gamma(x,y)\right\}^{1/2}.$$

Quantile Regression vs WGF

Figure 1: Error comparison of moment estimates: Violin plots of the estimation error in the first two moments using quantile regression and Wasserstein Gradient Flow regression are shown. The number of regressed samples is shown in parentheses.

Measuring Risk through Stochastic Dominance

The SSD relation is defined using distribution functions and compared over the continuum of their realizable values. We say that X stochastically dominates Y in the second order when their respective CDFs integrate to satisfy

$$X \succeq_{(2)} Y \iff F_X^{(2)}(\alpha) \leq F_Y^{(2)}(\alpha)\ \forall \alpha \in \mathbb{R}.$$

The function $F^{(2)}(\alpha) = \int_{-\infty}^{\alpha} F(x)dx$ defines the frontier of what is known as the *dispersion space* (Dentcheva and Ruszczyński, 2006). The volume reflects the degree to which a random variable differs from its deterministic behavior - if it were simply a real number equal to its expected value. Disperse outcomes are considered risky. Rational risk-averse agents prefer X to Y when $X \succeq_{(2)} Y$.

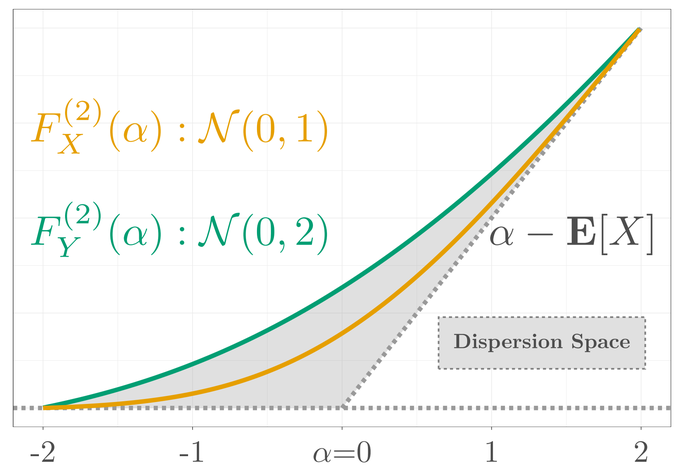


Figure 2: Dispersion space: The relative dispersion (i.e. uncertainty) of a random variable is shown as the space between its cumulative CDF $F_X^{(2)}$ and the asymptotes (dotted). Here, the line $\alpha - \mathbf{E}[X]$ defines the behavior of X as its uncertainty vanishes.

Policy Comparison Under Uncertainty

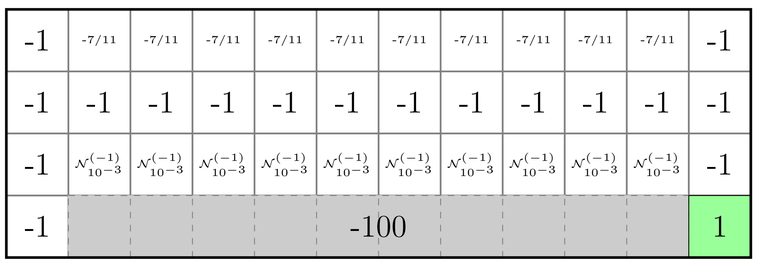
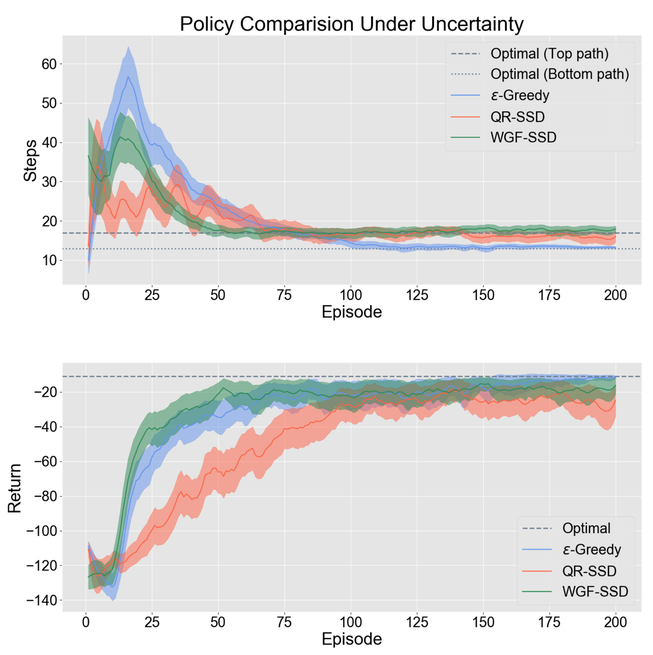


Figure 3: CliffWalk environment where two optimal policies exist in return. However, the top path has less uncertainty than the bottom path. We denote $\mathcal{N}_{\text{std}}^{\text{mean}}$.

The SSD behavior policy recovers the target policy using samples from the least-disperse data distribution: We compare the episodic step count and return using the SSD and ε-greedy policy. The two regression methods differ in their sample complexity but realize the same solutions.



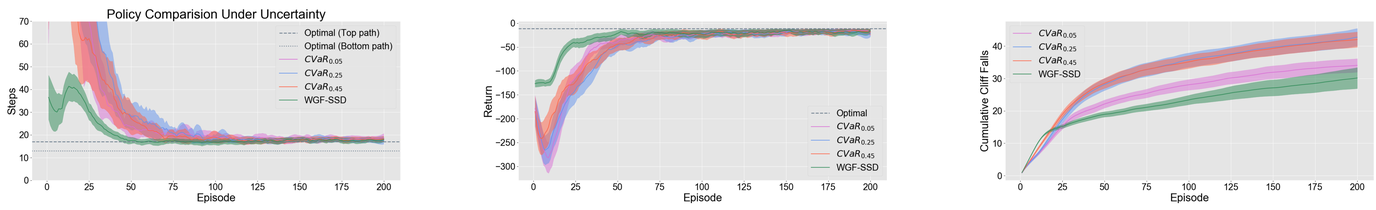


Figure 4: Using many risk levels can improve exploration: One risk level is not always appropriate for every state. Here, the CVaR policy leads the agent away from its goal, causing it to explore more than with the SSD policy, which uses many risk levels.

Function Approximation

In this experiment we test the hypothesis that WGF Fitted Q-iteration is scalable to function approximation in the control setting. We parameterize return distributions with a two-layer fully-connected neural network of 256 hidden units and use off-policy updates with bootstrapped targets.

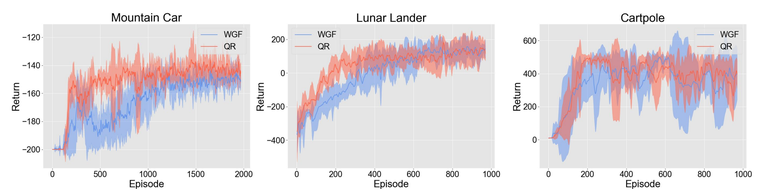


Figure 5: We find the WGF method matches the final average return of quantile regression.

Policy Evaluation

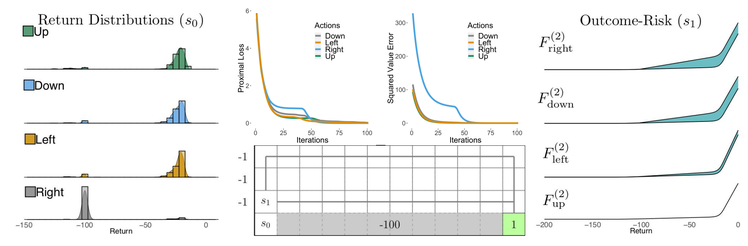


Figure 6: Distributional policy evaluation: The left plot shows the WGF estimated histograms of the smoothed target densities. Convergence of the proximal loss and the squared value error, shown in the top two plots, indicate the evaluation quality. This enables us to accurately recreate the outcome-risk diagrams, shown on the right, in descending order with respect to their dispersion space size.