Stochastically Dominant Distributional Reinforcement Learning

John D. Martin [1] **Michal Lyskawinski** [1] **Xiaohu Li** [1] **Brendan Englot** [1]

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 comprehensive evaluation of the environment's uncertainty. The necessary conditions for SSD require estimators to predict accurate second moments. To accommodate this, we map the distributional RL problem to a Wasserstein gradient flow, treating the distributional Bellman residual as a potential energy functional. We propose a particle-based algorithm for which we prove optimality and convergence. Our experiments characterize the algorithm's performance and demonstrate how uncertainty and performance are better balanced using an SSD policy than with other risk measures.

1. Introduction

Often in Reinforcement Learning (RL), agents select actions to maximize their expected sum of future (discounted) rewards. Many have pointed out how this strategy will sometimes lead to undesirable outcomes, particularly when the environment is stochastic (Heger, 1994; Tamar et al., 2013; Keramati et al., 2020). In these domains, an interesting subclass of problems require the agent to decide between several competing alternatives with the same expected outcome. These scenarios frequently arise in finance (Dentcheva & Ruszczyński, 2006a), where mutliple portfolios can lead to the same return but with different variability. In such settings, the expected value does not capture the full state of uncertainty, and it becomes prudent to employ the full distribution of outcomes.

[1]Stevens Institute of Technology, Hoboken, New Jersey, USA. Correspondence to: John D. Martin <jmarti3@stevens.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

The Conditional Value at Risk (CVaR_α) is a popular statistic that measures uncertainty with the expected outcome under an α-fraction of possibilities (Artzner et al., 1999). A great deal of recent RL research focuses on learning good CVaR policies (Chow & Ghavamzadeh, 2014; Tamar et al., 2015b; Dabney et al., 2018; Keramati et al., 2020). One point that remains unresolved is how to choose the right fraction of outcomes, i.e. the risk level α. It seems plausible that the best α-subset could vary across the environment. To our knowledge no one has considered this problem in RL, when uncertainty is used to evaluate competing actions. To address these issues, we introduce a distributional policy that simultaneously captures many risk levels, therefore removing the need to select one. Our policy is based on the Second Order Stochastic Dominance (SSD) relation.

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 cumulative CDFs, $F^{(2)}(\alpha) = \int_{-\infty}^{\alpha} F(x)dx$, satisfy (1), and we denote the relation $X \succeq_{(2)} Y$:

$$X \succeq_{(2)} Y \iff F_X^{(2)}(\alpha) \leq F_Y^{(2)}(\alpha)\ \forall \alpha \in \mathbb{R}. \quad (1)$$

The function $F^{(2)}$ defines the frontier of what is known as the *dispersion space* (Figure 1). The volume reflects the degree to which a random variable differs from its expected value, or its deterministic behavior. Outcomes that are disperse have more uncertainty and are considered risky in decision making settings. Indeed, a fundamental result from expected utility theory states that rational risk-averse agents prefer X to Y when $X \succeq_{(2)} Y$ (Dentcheva & Ruszczyński, 2006b). Drawing inspiration from this strategy, we apply SSD as an action selection method to reduce dispersion in the data distribution with which a policy is learned. Our paper offers the following contributions:

A distributional policy: Metrics such as variance (Sato et al., 2001; Tamar et al., 2013) and quantile statistics, like CVaR, (Chow & Ghavamzadeh, 2014; Dabney et al., 2018; Keramati et al., 2020) are prevalent in RL. A novel contribution of our work is the introduction of SSD for distributional action selection. As we will show, this relation eliminates the need to tune risk parameters. We apply the relation in settings where there are many solutions, and the agent wishes to select the least disperse (i.e. most certain) option.

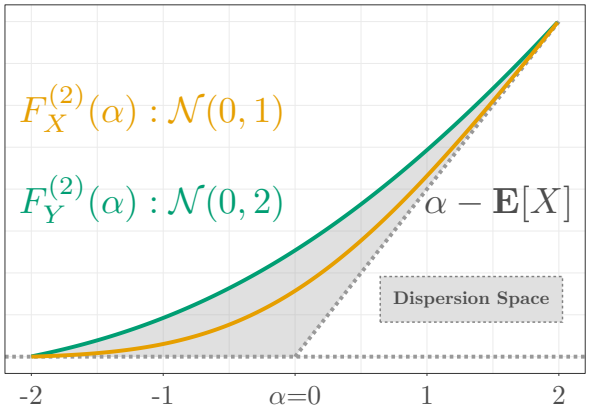


Figure 1. **Dispersion space:** The relative uncertainty of a random variable is shown as the space between its cumulative CDF $F_X^{(2)}$ and the asymptotes (dotted). Here, the line $\alpha - \mathbf{E}[X]$ defines the behavior of X as its uncertainty vanishes.

A new distributional RL algorithm: SSD implies an ordering on the first two moments of distributions. We propose a new learning algorithm based on Wasserstein gradient flows that is guaranteed to respect this, because its estimates converge in the first two moments.

We validate our theoretical claims with several targeted experiments. The main hypothesis we test is that the SSD policy induces the least-disperse data distribution from which optimality can be achieved when learning off-policy.

2. Background

Reinforcement Learning describes a sequential decision making problem, whereby an agent learns to act optimally from rewards collected after taking actions. At each time step, the agent selects an action $A \in \mathcal{A}$ based on its current state $S \in \mathcal{S}$, then transitions to the next state $S' \in \mathcal{S}$ and collects a reward $R \in \mathbb{R}$. The interaction is formally modeled as a Markov Decision Process (MDP) (Putterman, 1994), which we denote $\langle \mathcal{S}, \mathcal{A}, p, \gamma \rangle$. The transition kernel $p\colon \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R} \times \mathcal{S})$ defines a joint distribution over the reward and next state, given the current state-action pair. Here, $\mathcal{S}$ and $\mathcal{A}$ are measurable Borel subsets of complete and separable metric spaces, which we take as finite. And for each state $s \in \mathcal{S}$, the set $\mathcal{A}_s \subset \mathcal{A}$ is a measurable set indicating the feasible actions from s. The *random return* is

$$Z_\pi^{(s,a)} \triangleq \sum_{t=0}^{\infty} \gamma^t R^{(S_t, A_t)} \,\Big|\, S_0 = s, A_0 = a. \qquad (2)$$

Returns reflect the outcome of a decision sequence that starts after taking action a in state s then following the policy $\pi \in \Pi$ thereafter. Polices are stationary distributions over actions, coming from the set $\Pi = \{\pi | \pi\colon \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$. Here, $\gamma \in [0, 1)$ is a discount factor, and $R^{(S_t, A_t)}$ is the real-valued random reward associated with the state and action at time $t \in \mathbb{N}$.

2.1. Bellman's Equations

Bellman (1966) showed that the expected return (2), also known as the *value function*, has a recursive decomposition:

$$Q_\pi^{(s,a)} \triangleq \mathbf{E}_{\mu_\pi^{(s,a)}}[Z_\pi^{(s,a)}] = \mathbf{E}_{p,\pi}[R + \gamma Q_\pi^{(S',A')}],$$
$$Q_*^{(s,a)} \triangleq \mathbf{E}_{\mu^{(s,a)}}[Z^{(s,a)}] = \mathbf{E}_p[R + \gamma \max_{a' \in \mathcal{A}_{S'}} Q_*^{(S',a')}].$$

Here the *value*, $Q_\pi^{(s,a)}$, is defined as the expected return (Sutton & Barto, 2018) of policy π. We denote $\mu_\pi^{(s,a)}$ to be the corresponding distribution of returns under π from (s, a). Returns under $\mu^{(s,a)}$ follow the greedy policy, $\pi_*(s) = \arg\max_{a \in \mathcal{A}_s} \mathbf{E}_{\mu^{(s,a)}}[Z^{(s,a)}]$ for all $s \in \mathcal{S}$. When clear from context, we will drop the superscripts and refer to a single measure $\mu^{(s,a)}$ as μ. Viewed as a functional operator, these equations are known to contract to a unique fixed point. The contractive property motivates algorithms that update representations of Q_* to minimize the difference formed between both sides. Two popular methods for model-free learning are Sarsa (Rummery & Niranjan, 1994) and Q-learning (Watkins & Dayan, 1992). These algorithms use samples of the form (s, a, r, s') to iteratively update value estimates. Sarsa is an on-policy algorithm because it evaluates the policy it uses to gather data:

$$Q_\pi^{(s,a)} \leftarrow Q_\pi^{(s,a)} + \alpha(r - \gamma \mathbf{E}_\pi[Q_\pi^{(s',A')}] - Q_\pi^{(s,a)}).$$

Q-learning is *off-policy* because it gathers data with a separate *behavior policy* to learn the target greedy policy:

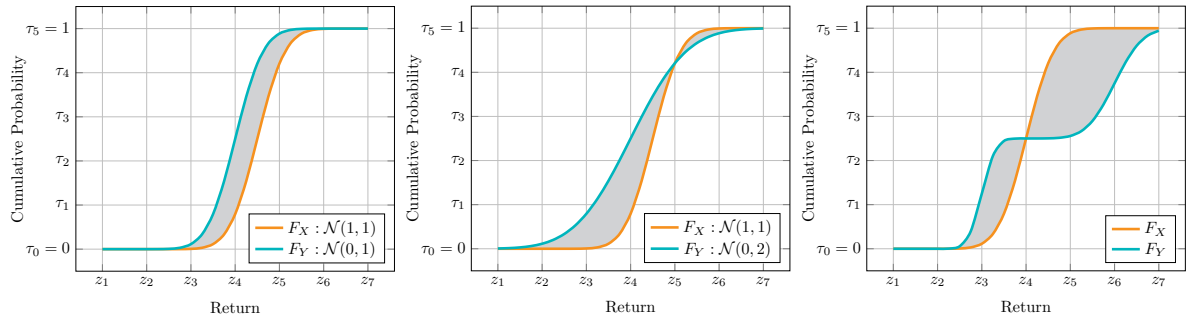
$$Q_*^{(s,a)} \leftarrow Q_*^{(s,a)} + \alpha(r - \gamma \max_{a' \in \mathcal{A}_{s'}} Q_*^{(s',A')} - Q_*^{(s,a)}).$$

2.2. Distributional Bellman Operators

The return distribution, $\mu_\pi^{(s,a)}$, also satisfies a distributional variant of Bellman's equation $\mathcal{T}^\pi \mu_\pi^{(s,a)} \triangleq$

$$\int_{\mathbb{R}} \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} f_\sharp^{(r,\gamma)} \mu_\pi^{(s',a')} \pi(a'|s') p(dr, s'|s, a). \qquad (3)$$

Here, $\mathcal{T}^\pi$ is the distributional Bellman operator. It embeds a measurable mapping that reflects the Bellman action: $f^{(r,\gamma)}(x) \triangleq r + \gamma x$, where the push-forward measure is $f_\#(\mu(A)) \triangleq \mu(f_\#^{-1}(A)) = \nu(A)$, for all Borel measurable sets A. Just as the standard Bellman equation is the focus of standard value-based RL, the distributional Bellman operator (3) plays the central role in Distributional RL (DRL); it motivates algorithms which attempt to represent $\mu_\pi^{(s,a)}$ and approximate it by repeated application of the update



(a) $X \succeq_{(1)} Y$ and $X \succeq_{(2)} Y$ (b) $X \not\succeq_{(1)} Y$ and $X \succeq_{(2)} Y$ (c) $X \not\succeq_{(1)} Y$ and $X \not\succeq_{(2)} X$

Figure 2. **Stochastic dominance action selection:** Imagine X and Y are returns induced by different actions. Our action selection rule can be visualized with plots of the CDF. In Fig. 2a $X \succeq_{(2)} Y$, because X places more mass on points larger than α. In Fig. 2b, the area left of z_5 is greater than the area to its right, hence $X \succeq_{(2)} Y$. In Fig. 2c, neither variable dominates, because for $\alpha \geq z_4$, the enclosed area is larger than the other region. In these cases, we select from among the competing actions at random.

$\mu^{(s,a)}_{\pi,t+1} \leftarrow \mathcal{T}^{\pi}\mu^{(s,a)}_{\pi,t}$, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. The optimality operator is realized under the greedy policy:

$$\mathcal{T}\mu^{(s,a)} \triangleq \int_{\mathbb{R}} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} f^{(r,\gamma)}_{\sharp}\mu^{(s',a^*)}p(dr,s'|s,a). \quad (4)$$

3. Distributional Action Selection

We consider scenarios where a DRL agent often encounters several outcomes which all appear equivalent under the expected return. The uncertainty-sensitive decision problem is to select from among these choices, the option that minimizes uncertainty. For this we propose the SSD policy, whose comparisons are visualized in Figure 2.

At each state s, the agent makes a point-wise comparison of the distribution functions $F^{(2)}_{Z^{(s,a)}}(\alpha)$, for all $a \in \mathcal{A}_s$. The dominating action is the one whose cumulative CDF achieves the lowest value for every $\alpha \in \mathbb{R}$:

$$\mathcal{A}^{(2)}_s \triangleq \{a \in \mathcal{A}_s : Z^{(s,a)} \succeq_{(2)} Z^{(s,a')}, \forall\, a' \in \mathcal{A}_s \setminus \{a\}\}.$$

3.1. Numerically tractable comparisions

Evaluating SSD appears to be a numerically intractable task, involving point-wise comparisons over an infinite domain. Fortunately, we can circumvent this problem by using cumulative quantile functions (Dentcheva & Ruszczyński, 2006b): $F^{-2}(\tau) = \int_0^\tau F^{-1}(t)dt$. Now the SSD relation becomes:

$$X \succeq_{(2)} Y \iff F_X^{-2}(\tau) \geq F_Y^{-2}(\tau)\ \forall\, \tau \in (0,1), \quad (5)$$

where we assume that $F_Y^{-2}(0) = 0$, and $F_Y^{-2}(1) = \infty$.

Notice that $F_X^{-2}(\tau)/\tau = \mathbf{E}[X|X \leq \xi^{(\tau)}]$ is the Conditional Value at Risk for level τ. Thus, the SSD relation can be interpreted as a continuum of CVaR comparisons for all $\tau \in (0,1)$. From this we can surmise that points along the boundary of dispersion space (Figure 1) represent unconditional Values at Risk (VaR). Furthermore, this connection suggests a numerically-tractable way to compute F^{-2}.

Lemma 1. *Let $\tau \in (0,1)$ and consider $\xi^{(\tau)} = F_X^{-1}(\tau)$. Then $F_X^{-2}(\tau) = \mathbf{E}[X \leq \xi^{(\tau)}]$.*

Lemma 1 makes it possible to compare total expectations on subsets of the return space, instead of dealing with probability integrals over an unbounded domain.

Computations simplify even further when we consider discrete measure approximations to the return distribution. We propose a Lagrangian (particle-based) discretization, where the measure is supported on $N \in \mathbb{N}$ equally-likely diracs:

$$\mu^{(s,a)} \approx \frac{1}{N}\sum_{i=1}^{N} \delta^{(s,a)}_{z^{(i)}}.$$

Values are straightforward to compute from the corresponding samples: $Q^{(s,a)} = \frac{1}{N}\sum_{i=1}^{N} z^{(i)}$.

To apply (5), denote the ordered coordinates of a return distribution to be $z^{[1]} \leq z^{[2]} \leq \cdots \leq z^{[N]}$. Then given particle sets for two random returns induced by the actions a_1 and a_2, we have the following result.

Proposition 1. *$Z^{(s,a_1)} \succeq_{(2)} Z^{(s,a_2)}$ if, and only if*

$$\sum_{i=1}^{j} z^{[i]}_{a_1} \geq \sum_{i=1}^{j} z^{[i]}_{a_2},\ \forall\, j = 1,\cdots,N. \quad (6)$$

The SSD policy is executed at each step by constructing $\mathcal{A}^{(2)}_s$ using (6) and a discrete representation of $\mu^{(s,a)}$. In some cases $\mathcal{A}^{(2)}_s$ will be empty (Figure 2c), indicating that total dominance cannot be established. There are several heuristics that could handle this outcome, including a next-best strategy, or an additional decision criterion. We choose

to sample the greedy actions uniformly at random. This increases uniform exploration when dominance cannot be established and still constitutes a strict enhancement of the greedy policy when multiple solutions are present.

3.2. Necessary conditions for SSD

When is it possible to apply the SSD policy? The following result from Fishburn (1980) implies an ordering on the first two moments of the distributions under SSD.

Proposition 2 (Fishburn (1980)). *Assume μ has two finite moments. Then $X \succeq_{(2)} Y$ implies $\mu_X^{(1)} \geq \mu_Y^{(1)}$ or $\mu_X^{(1)} = \mu_Y^{(1)}$ and $\mu_X^{(2)} \leq \mu_Y^{(2)}$, where $(\cdot)$ denotes a particular moment of the distribution μ.*

The ordering indicates that the dominating distribution either has the greatest mean, or it has the smallest second moment when means are equal. Given infinite precision and random initialization, the chances of more than one action having the same value may seem unlikely. However, in cases where finite precision is used (e.g. finance), or cases where a tolerance is applied to comparisons, equivalence arises often. Proposition 2 imposes a necessary requirement on estimates of the return distribution. Namely, the estimates must be accurate enough to respect the ordering. Moving forward we seek distributional learning algorithms that we know converge in the first two moments.

4. Wasserstein Gradient Flows for RL

In this section we describe how return distributions can be obtained from the solution of a Wasserstein Gradient Flow (WGF). We detail the solution procedure and show how it naturally integrates into the fitted value iteration paradigm (Gordon, 1995). We expand on the WGF theory to show that solutions converge in the first two moments, as we need to respect Proposition 2.

4.1. Wasserstein convergence

The k-th order Wasserstein distance for any two univariate measures $\mu, \nu \in \mathcal{P}_k(\mathbb{R})$, is defined as

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

where $\mathcal{P}_k(\mu, \nu)$ is the set of all joint distributions with marginals μ and ν having k finite moments. The distance describes an optimal transport problem, where one seeks to transform μ to ν with minimum cost; here the cost is $|x-y|^k$. The $\mathcal{W}_k$ distance is appealing as a distributional learning objective, because its convergence implies convergence in the first k moments (Villani, 2008).

4.2. Distributional RL as free-energy minimization

We cast the distributional RL problem as a free-energy minimization in terms of the functional:

$$E(\mu) \triangleq F(\mu) + \beta^{-1} H(\mu). \tag{7}$$

Here, we have dropped the superscript notation. F is the potential and H is the entropy of a single probability measure; $\beta \in \mathbb{R}_+$ is an inverse temperature parameter.

The potential energy defines what it means for a distribution to be optimal. We choose the low-energy equilibrium to coincide with minimum expected Bellman error, formed from (4). Energy is minimized when the mapping $\mathcal{T}$ reaches its fixed point, $\mathcal{T}\mu^{(s,a)} = \mu^{(s,a)}$ for some (s, a). Given a transition sample (s, a, r, s'), we compute the distributional targets $\mathcal{T}z^{(s,a)}$, which denote realizations of $\mathcal{T}\mu^{(s,a)}$, and define Bellman's potential energy as

$$F(\mu) \triangleq \frac{1}{2} \int \left(\mathcal{T}z^{(s,a)} - z^{(s,a)} \right)^2 d\mu = \int U(z) d\mu. \tag{8}$$

The optimal probability measure for these models is known to be the Gibbs measure: $\mu_*(z) = \mathcal{Z}^{-1} \exp\{-\beta U(z)\}$, where $\mathcal{Z} = \int \exp\{-\beta U(z)\} dz$. Energy-based models have been applied for policy optimization (Haarnoja et al., 2017; Zhang et al., 2018), but to our knowledge, they have not appeared in value-based methods for DRL.

4.3. The Fokker-Planck Equation

We would like to understand the convergence behavior of return distributions as the free-energy is minimized. Systems of this nature are typically modeled as continuous-time stochastic diffusion processes, where the distributions $\{\mu_t\}_{t\in[0,1]}$ evolve over a smooth manifold of probability measures from $\mathcal{P}_2(\mathbb{R})$. The dynamics of μ_t is known to obey a diffusive partial differential equation called the Fokker-Planck equation (Risken, 1984):

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla (\frac{\delta E}{\delta \mu_t}) \right). \tag{9}$$

Here, the sub-gradient with respect to time is denoted ∂_t, and the first variation (Gâteaux derivative) of free energy $\frac{\delta E}{\delta \mu}$. The Fokker-Plank equation plays a central role in statistical physics, chemistry, and biology. In optimization, it defines the solution path, or gradient flow, of μ as it evolves over the manifold of probability measures.

Proposition 3 (Ambrosio (2005)). *Let $\{\mu_t\}_{t\in[0,1]}$ be an absolutely-continuous curve in $\mathcal{P}_2(\mathbb{R})$. Then for $t \in [0, 1]$, the vector field $\mathbf{v}_t = \nabla(\frac{\delta E}{\delta t}(\mu))$ defines a gradient flow on $\mathcal{P}_2(\mathbb{R})$ as $\partial_t \mu_t = -\nabla \cdot (\mu_t \mathbf{v}_t)$, where $\nabla \cdot \mathbf{u}$ is the divergence of the vector $\mathbf{u}$.*

Intuitively, the free-energy E characterizes the diffusion process and thus, the optimization landscape of our new

distributional RL problem. Convergence to an optimal point can be guaranteed provided E is convex, which we know to be the case for (7), which is quadratic and logarithmic in μ.

4.4. Discrete Time Solutions

To approximately solve (9), we adopt an iterative procedure due to Jordan et al. (1998). The method discretizes time in steps of $h \in \mathbb{R}_+$ and applies the proximal operator

$$\text{Prox}^{\mathcal{W}}_{hE}(\mu_k) \triangleq \underset{\mu \in \mathcal{P}_2(\mu, \mu_k)}{\arg\min}\ \mathcal{W}_2^2(\mu, \mu_k) + 2hE(\mu). \quad (10)$$

For every step $k \in \mathbb{N}$, the operator generates a path of distributions $\{\mu_t\}_{t=1}^K$ such that $\mu_{k+1} = \text{Prox}^{\mathcal{W}}_{hE}(\mu_k)$ is equivalent to μ_K. In contrast with DRL methods that apply (4), we apply the proximal operator to minimize a free energy with a $\mathcal{W}_2^2$-regularizer via (semi-)gradient steps. And because E is convex, this method obtains the unique solution to (9).

Proposition 4 (Jordan et al. (1998))**.** *Let $\mu_0 \in \mathcal{P}_2(\mathbb{R})$ have finite free energy $E(\mu_0) < \infty$, and for a given $h > 0$, let $\{\mu_t^{(h)}\}_{t=1}^K$ be the solution of the discrete-time variational problem (10), with measures restricted to $\mathcal{P}_2(\mathbb{R})$, the space with finite second moments. Then as $h \to 0$, $\mu_K^{(h)} \to \mu_T$, where μ_T is the unique solution of (9) at $T = hK$.*

Furthermore, one can evaluate the free-energy (8) over the solution sequence and observe it becomes a decreasing function of time (i.e. a Lyapunov function). The result implies that the expected distributional Bellman residual is minimized when using the JKO approach.

Proposition 5. *Let $\{\mu_k^{(h)}\}_{k=0}^K$ be the solution of the discrete-time variational problem (10), with measures restricted to $\mathcal{P}_2(\mathbb{R})$, the space with finite second moments. Then $E(\mu_k)$ is a decreasing function of time.*

Finally, we can show that as β is annealed, the output of our free-energy optimization (10) is equivalent to the solution obtained from the distributional Bellman operator (4).

Theorem 1. *If $\mathcal{T}\mu = \mu$, then $\text{Prox}^{\mathcal{W}}_{hE}(\mu) = \mu$ as $\beta \to \infty$.*

4.5. Discrete Measure Solutions

Given an initial set of particles at some state-action pair $z(s,a) = \{z^{(1)}, \cdots, z^{(N)}\}$, we evolve them forward in time with steps of h to obtain the solution at $t+h$. We apply a finite number of gradient steps to approximate the convergence limit $T = hK$. Finally we consider an entropic-regulated form of $\mathcal{W}_2^2$ (Cuturi, 2013) for two finite distributions $\mu = \sum_{i=1}^N \mu_i \delta_{x^{(i)}}$ and $\nu = \sum_{j=1}^N \nu_j \delta_{y^{(j)}}$:

$$\mathcal{W}_\beta(\mu, \nu) \triangleq \inf_{P \in \mathbb{R}_+^{N \times N}} \langle P, C \rangle + \beta \mathsf{KL}(P | \mu \otimes \nu),$$
$$\text{s.t.} \sum_{j=1}^N P_{ij} = \mu_i, \sum_{i=1}^N P_{ij} = \nu_j.$$

Here, $\langle P, C \rangle$ denotes the Frobenius norm between the joint P and the square Euclidean cost $C_{ij} = (x_i - y_j)^2$, and $\mathsf{KL}(P|\mu \otimes \nu) = \sum_{i,j}[P_{ij} \log(P_{ij}/\mu_i \nu_j) - P_{ij} + \mu_i \nu_j]$. The entropic term promotes numerical stability by acting as a barrier function in the positive octant. JKO stepping under this new distance is denoted

$$\text{Prox}^{\mathcal{W}_\beta}_{hF}(\mu_k) \triangleq \underset{\mu \in \mathcal{P}_2(\mu, \mu_k)}{\arg\min}\ \mathcal{W}_\beta(\mu, \mu_k) + 2hF(\mu). \quad (11)$$

One can compute the entropic-regularized distance, $\mathcal{W}_\beta$, using Sinkhorn iterations (Sinkhorn, 1967). This procedure (detailed in the appendix) is differentiable, which allows us to update represent particle locations with parametric models and update their predictions with gradient steps computed through auto-differentiation.

4.6. Online WGF Fitted Q-iteration

We are now ready to describe Online WGF Fitted Q-iteration (Alg. 1). The algorithm combines the solution of (11) into a Fitted Q-iteration framework to repeatedly fit return distributions. The loss is computed with Alg. 2. The principles apply in both the on-policy and off-policy settings. Here, we consider the off-policy case to compare different behavior policies. Both distributional policies and those based on point estimates are represented with the operator $\mathcal{B}\colon \mathcal{P}(\mathbb{R})^{|\mathcal{A}|} \to \mathcal{P}(\mathcal{A})$. Given a set of return distributions, this outputs a distribution over actions.

Algorithm 1 Online WGF Fitted Q-iteration

```
# Initialize particles
z(s,a) = {z^(i)}_{i=1}^N ∀ (s,a) ∈ S × A
for t = 1, 2, ··· do
    # Explore with the behavior policy
    s', r ~ p(·|s,a) with a ~ Bz(s,:)
    # Exploit with the greedy target policy
    a* ← arg max_{a∈A} {(1/N) Σ_{i=1}^N z^(i)(s',a)}
    Tz^[i] ← r + γ z^[i](s',a*) ∀ i ∈ [N]
    # Update particles with proximal step
    z(s,a) ← arg min_z L^{W_β}_{hF_T}(z, z(s,a))
end for
```

Algorithm 2 Proximal Loss

```
input: Source and target particles z, z_0; Tz
F_T(z) ← (1/2N) Σ_{i=1}^N [Tz^[i] − z^[i]]^2
W_β(z, z_0) ← Sinkhorn_β(z, z_0)
# Output JKO loss
output: L^{W_β}_{hF_T} = W_β(z, z_0) + 2hF_T(z)
```

5. Connections with Related Work

Modeling Risk for RL: Many have employed measures of uncertainty to replace or regulate the optimization ob-

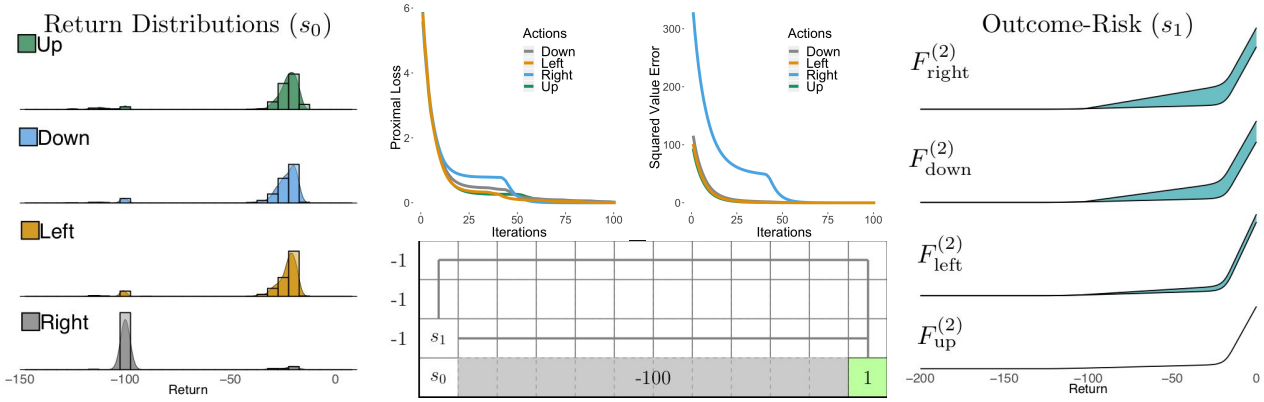


Figure 3. **Policy evaluation in the CliffWalk domain:** The left plot shows WGF estimates of the smoothed target distributions. Convergence of the proximal loss and the squared value error are shown in the top two plots. Outcome-risk diagrams (right) derived from distribution estimates illustrate the relative dispersion space size at s_1. The two inflections represent the bimodality of the distribution.

jective in RL using the Markowitz mean-variance model (Markowitz, 1952). Among these include policy gradient methods (Tamar et al., 2015a), actor-critics (Tamar & Mannor, 2013) and TD methods (Sato et al., 2001; Tamar et al., 2013; Keramati et al., 2020). Constraint techniques have also been considered using CVaR within a policy gradient and actor-critic framework (Chow et al., 2017). In contrast to methods that directly constrain the policy parameters, we constrain the data distribution with action selection using SSD among the return distributions. Dabney et al. (2018) trains risk-averse and risk-seeking agents from return distributions sampled from various distortion risk measures. However, they do not address problems involving multiple solutions. Furthermore, is it unclear how to sample from SSD-equivalent distortions when total dominance cannot be established. This investigation is left for future work.

Distributional RL: Our learning algorithm is inspired by the class of DRL algorithms (Bellemare et al., 2017). These methods model a distribution over the return, whose mean is the familiar value function, and use it to evaluate and optimize a policy (Barth-Maron et al., 2018; Hessel et al., 2018). Bellemare et al. (2017) first showed the distributional Bellman operator contracts in the supremal Wasserstein distance. They proposed a discrete-measure approximation algorithm (C51) using a fixed mesh in return space and later showed it converges in the Cramer distance (Rowland et al., 2018). Particle-based methods that use Quantile Regression (QR), have shown encouraging progress on empirical benchmarks (Dabney et al., 2017; 2018). However, understanding their convergence beyond the first moment has been more challenging. By casting the optimization problem as free-energy minimization in the space of probability measures, we show that DRL can be modeled as the evolution of a WGF. Updates in this framework have well-defined dynamics, permitting us to better understand convergence and optimality.

Wasserstein Gradient Flows in RL: To our knowledge WGF solutions have only been applied to policy gradient algorithms. Zhang et al. (2018) models stochastic policy inference as free-energy minimization, and applies the JKO scheme to derive a policy gradient algorithm. Their method is couched within the Soft-Q learning paradigm (Haarnoja et al., 2017; 2018). These algorithms train a deep neural network to sample from a target Gibbs density using Stein Variational Gradient Descent (Liu & Wang, 2016). Our algorithm learns distributions of the underlying return and thus can be considered value-based. Furthermore, we are concerned with decision making in the presence of aleatoric uncertainty, and when the agent must select the most certain outcome from among many alternatives.

6. Experiments

In this section we verify several prior assertions. Namely, we test the hypothesis that WGF regression produces two accurate moment estimates. Next we show WGF solutions from Alg. 1 can recover the latent return distribution in a policy evaluation setting. We extend these results to the control setting with bootstrapped off-policy updates under function approximation. In our final experiment, we quantify an agent's ability to mitigate uncertainty while gathering training data with the SSD behavior policy. Details of each experiment can be found in the Appendix.

6.1. Regression Comparison

Given that standard quantile regression learns samples from a uniform mesh in probability space, theory suggests accuracy improvements can be gained with a non-uniform mesh produced from the solution of a WGF. To evaluate this hypothesis, we compared the root mean squared error on a five component one-dimensional Gaussian mixture model,

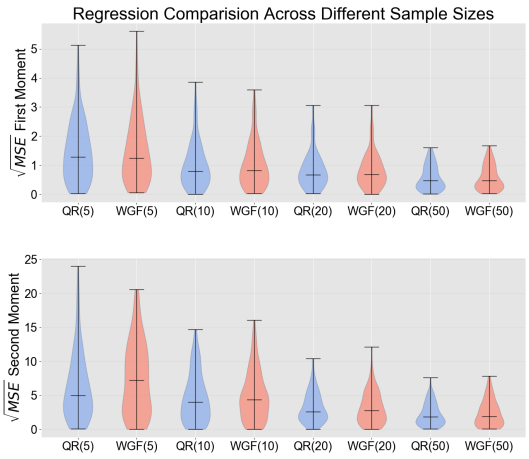


Figure 4. **Distributions of moment estimation error:** Quantile regression and WGF regression produce similar estimates in one dimension. The number of support samples is shown in parentheses.

intended to be representative of a geometrically-complex return distribution. Ablations informed the parameterization of the proximal loss (See appendix). We collected data over 100 independent trials, varying the number of samples each method regressed. Our data shows there to be *no statistical difference* between QR and WGF regression (Figure 4).

We interpret the observed insignificance as a consequence of using low-dimensional data. The error from a uniform grid is expected to become more pronounced as dimensionality increases. Given that we are concerned with one-dimensional return distributions, however, these results inform different message within our problem setting. Namely, the distributions regressed through QR may be reasonably employed for SSD action selection. We believe practitioners will find this result valuable when choosing a regression method where two accurate moment estimates are required.

6.2. WGF Policy Evaluation

Proposition 5 argues that repeated application of the proximal step (10) produces a decreasing function of time, implying that the Bellman free energy is minimized at convergence. Here, we verify this is indeed the case by learning the return distribution in a policy evaluation setting. The problem is set within the CliffWalk domain (Fig. 3). The transition dynamics follow those in Sutton & Barto (2018). However, we include a five-percent chance of falling off the cliff from adjacent states. We used fixed Monte Carlo (MC) targets from the optimal greedy policy.

Figure 3 shows the convergence of the proximal loss and the mean square value error from the start state. As we can see, the estimated distribution (the histogram) accurately captures the target's features: the near certainty of walking off the cliff when moving right, the added chance of doing the same when choosing left or down, and finally the most profitable choice, moving up.

6.3. WGF in the Control Setting

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. We use off-policy updates with bootstrapped targets and compare performance results with an agent trained using the QR loss (Dabney et al., 2017) on three common control tasks from the OpenAI Gym (Brockman et al., 2016): MountainCar, CartPole, and LunarLander. The results in Figure 5 show that the WGF method matches the performance of QR.

6.4. Control in the Presence of Uncertainty

This experiment studies how aleatoric uncertainty is handled during training. Specifically, we compare different policies for selecting among a multiplicity of competing solutions. We consider the ε-greedy, SSD, and CVaR$_\alpha$ behavior policies for $\alpha \in \{0.05, 0.25, 0.45\}$. Each policy gathers data to update a greedy target policy. Different data distributions arise from the way each measures uncertainty.

We expect the data distribution under the SSD policy to favor outcomes with higher certainty, because SSD compares the expected outcome over all represented risk levels. CVaR policies consider the expected outcome for a single risk level. Uncertainty drives action selection only when the specified risk level captures the true risk in the current state. Otherwise, we expect CVaR policies to become risk neutral.

We revisit the CliffWalk domain with a modified reward structure (See appendix). Traversing the top and bottom rows have equal value. Each path has different reward uncertainty; the top row is deterministic, whereas the bottom row samples rewards from the Gaussian $\mathcal{N}(-1, 10^{-3})$. Under these conditions, we expect the SSD policy to prefer the top path and risk neutral methods to prefer the bottom row, since it will be more likely under a risk neutral policy.

Figure 6 shows the average episodic step count and return, along with their 95% confidence intervals computed from 50 trials. The step count data confirms our hypothesis that the SSD policy induces the least-disperse data distribution, since it takes the top path on average. We can also confirm that the ε-greedy policy chooses the bottom path, which is more likely under the sampling distribution from s_0 and incidentally more dispersed. We observe similar behavior between QR to WGF Q iteration, consistent with results in Figure 4. Both methods induce similar data distributions over the top path at around the 75th episode. And in this domain, the WGF method learns the quickest.

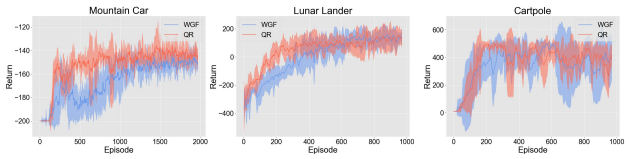


Figure 5. **Performance on control problems:** The WGF method matches the final average return of quantile regression.

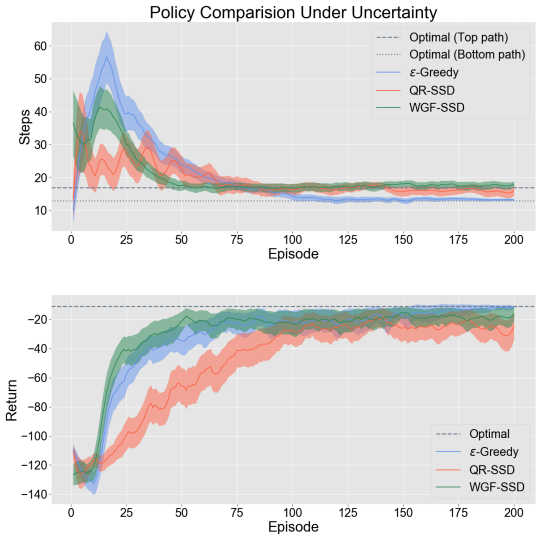


Figure 6. **The SSD behavior policy recovers the optimal target policy using samples from the least-disperse data distribution:** We compare the episodic step count and return using the SSD and ε-greedy policy. The distributional methods differ in their sample complexity but realize the same final solution.

We find the greatest differences between the SSD and CVaR policies occur in the transient phase of learning (Figure 7). The CVaR agent takes more exploratory steps as a result of using a single uniform risk level. In high-stakes settings, the consequence of exploration can vary from undesirable to catastrophic. Here a cliff fall models a very costly outcome. Figure 7 shows the number of cliff falls encountered throughout learning. Using the SSD policy results in a significantly lower number of these experiences. We interpret this as positive evidence to suggest that SSD provides a more comprehensive measure of uncertainty than CVaR.

7. Conclusion

This paper argues for the use of SSD to select among a multiplicity of competing solutions. This can be useful in settings where one wishes to minimize exposure to uncertainty. We presented a convergent, online algorithm for learning return distributions (WGF Fitted Q-iteration). Our simulations demonstrated the algorithm can learn good policies, and that it scales up to function approximation. Based on our

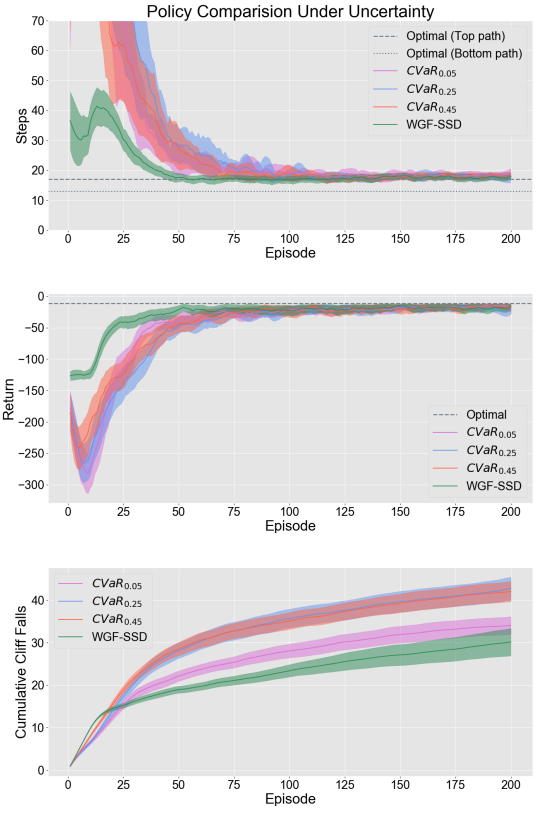


Figure 7. **Using many risk levels can improve exploration:** One risk level is not always appropriate for every state. Here, the CVaR policy leads the agent away from its goal, causing it to explore more than with the SSD policy, which uses many risk levels.

experimental results, we concluded that the SSD behavior policy can reduce dispersion in the data distribution and improve exploration in the presence of uncertainty.

Acknowledgements

The authors wish to thank the anonymous reviewers for their feedback. A special thanks goes to Marc G. Bellemare and to Shruti Mishra for their thoughtful reviews of earlier drafts.

This work relates to Department of Navy award N00014-20-1-2570 issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. This work was also supported in part by the National Science Foundation, grant number IIS-1652064, by the Robert Crooks Stanley Graduate Fellowship in Engineering and Science, and by the U.S. Department of Homeland Security under Cooperative Agreement No. 2014-ST-061-ML0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

Ambrosio, L. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zurich, 2005.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., Dhruva, T., Muldal, A., Heess, N., and Lillicrap, T. Distributed Distributional Deterministic Policy Gradients. In *International Conference on Learning Representations (ICLR)*, 2018.

Bellemare, M., Dabney, W., and Munos, R. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Bellman, R. *Dynamic programming*, volume 153. American Association for the Advancement of Science, 1966.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym, 2016.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR Optimization in MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-Constrained Reinforcement Learning with Percentile Risk Criteria. *Journal of Machine Learning Research (JMLR)*, 2017.

Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. Distributional Reinforcement Learning with Quantile Regression. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit Quantile Networks for Distributional Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Dentcheva, D. and Ruszczyński, A. Portfolio optimization with stochastic dominance constraints. *Journal of Banking & Finance*, 30(2):433–451, 2006a.

Dentcheva, D. and Ruszczyński, A. Inverse stochastic dominance constraints and rank dependent expected utility theory. *Mathematical Programming*, 108(2-3):297–311, 2006b.

Fishburn, P. C. Stochastic dominance and moments of distributions. *Mathematics of Operations Research*, 5(1): 94–100, 1980.

Gordon, G. J. Stable function approximation in dynamic programming. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, 1995.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Heger, M. Consideration of risk in reinforcment learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML)*, 1994.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. Being optimistic to be conservative: Quickly learning a CVaR policy. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2016.

Markowitz, H. Portfolio Selection. *Journal of Finance*, 12: 77–91, 1952.

Putterman, M. *Markov Decision Processes*. John Wiley & Sons, 1994.

Risken, H. *Fokker-Planck Equation*. Springer Berlin Heidelberg, 1984.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., and Teh, Y. An Analysis of Categorical Distributional Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Rummery, G. A. and Niranjan, M. On-line Q-Learning using connectionist systems. Technical report, Cambridge University, 1994.

Sato, M., Kimura, H., and Kobayashi, S. TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001.

Sinkhorn, R. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tamar, A. and Mannor, S. Variance adjusted actor critic algorithms. *arXiv:1310.3697*, 2013.

Tamar, A., Castro, D. D., and Mannor, S. Temporal difference methods for the variance of the reward to go. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems (NeurIPS) 28*, 2015a.

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the CVaR via sampling. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015b.

Villani, C. *Optimal Transport: Old and New*. Springer, 2008.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Zhang, R., Chen, C., Li, C., and Carin, L. Policy optimization as Wasserstein gradient flows. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.