New Approaches for Modeling Rich Encodings of Long-Term Behavior

Thesis Proposal

March 2025

Harley Wiltzer
McGill University
harley.wiltzer@mail.mcgill.ca

Slides available at https://cs.mcgill.ca/~hwiltz/phd-proposal/.

Overview

In the study of reinforcement learning (RL), we are primarily interested in estimating the *long term* behavior of agents; particularly with regard to how they accumulate instantaneous *rewards* that reinforce certain behaviors. Predominant approaches to RL involve modeling the *value function* describing the total expected return earned by the agent for a given reward signal, enabling optimal control via simple greedy algorithms. However, such approaches effectively compress the long-term behavior of the agent into a single scalar value, eliminating any signal that could have shed light on the agent's behavior for alternative rewards, alternative utilities, and more. In this proposal, I explore novel *distributional* approaches for encoding long-term behavior, and present algorithms that enable predictions that generalize across time horizons, across reward functions, and across risk preferences.

I will begin by demonstrating fundamental challenges intrinsic to value-based RL as the agent's decision frequency (equivalently, the discount factor) increases. Such challenges have been foreseen for more than thirty years, but I will show that a distributional perspective on RL sheds light on complex statistical phenomena exhibited by *action gaps*—the degree to which different actions can be distinguished by their returns—and I use these insights to derive novel algorithms for *high-frequency* risk-sensitive control.

Following, I will introduce methods for enabling *zero-shot* prediction of return distributions from reward functions. Particularly, I will focus on distributional *successor features*, providing novel algorithms and convergence results for dynamic programming and temporal difference learning.

To conclude, I will propose methods by which such distributional models can enable robust behavior *fine-tuning*, via novel regularized distributional Bellman operators for constrained policy optimization, and via novel distributional objectives for learning from demonstrations.

1 Background and Related Work

In this section, I outline the relevant formalisms of the reinforcement learning setting, upon which the remainder of the proposal is based. We begin first with a review of probability metrics, which will be instrumental in our discussions of distributional reinforcement learning.

1.1 Probability Metrics

For any measurable space $(\mathcal{Y}, \Sigma)$, we denote by $\mathscr{P}(\mathcal{Y})$ the set of all probability measures $\mathbb{P} : \Sigma \to [0, 1]$. In this proposal, notions of *distance* between elements of $\mathscr{P}(\mathcal{Y})$ will be an important concept. While there are simply notions of distance on $\mathbb{N}, \mathbb{R}$, and even $\mathbb{R}^d$, distance between probability measures is much more ambiguous. We outline some common notions in this section.

1.1.1 Optimal Transport and the Wasserstein Metrics

A very useful class of probability metrics arose from the study of *optimal transport*, known as the Wasserstein distances. For a metric space (Ω, d), denoting by $\mathcal{K}(\mu, \nu)$ the set of all couplings between $\mu, \nu \in \mathscr{P}(\Omega)$, the Wasserstein distances W_p are given by

$$W_p(\mu, \nu) = \min_{\rho \in \mathcal{K}(\mu,\nu)} \left[\iint d^p(\omega_1, \omega_2) \rho(\mathrm{d}\omega_1 \, \mathrm{d}\omega_2) \right]^{\frac{1}{p}}, \quad p \in [1, \infty). \tag{1}$$

The minimizing coupling in Equation 1 is called the *optimal coupling*; it is well known that it exists under mild conditions, and that W_p defines a proper metric on $\mathscr{P}(\Omega)$ (Villani, 2009). Wasserstein distances are very useful for deriving upper bounds, establishing celebrated convergence rates in distributional RL (Bellemare et al., 2017a). However, sample-based estimators of Wasserstein distances beyond scalar distributions are generally biased (Bellemare et al., 2017b).

1.1.2 RKHS and Maximum Mean Discrepancy

Spaces of probability measures lack geometric structure that we often rely on, such as linearity and orthogonality. The work of Sriperumbudur et al. (2008) investigated methods for embedding probability measures into *reproducing kernel Hilbert spaces* (RKHS) to provide this geometric structure. The key insight of Gretton et al. (2012) was to leverage the embedding $\mathsf{M}_\kappa : \mu \mapsto \int \kappa(y, \cdot)\mu(\mathrm{d}y)$ to an RKHS with kernel κ, called *mean embeddings*, to induce a probability metric from the RKHS.

Definition 1 (Gretton et al. (2012)): Let $\mathcal{H}$ be an RKHS over a set $\mathcal{Y}$ with kernel κ. The *maximum mean discrepancy* (MMD) $\mathrm{MMD}_\kappa : \mathscr{P}(\mathcal{Y}) \times \mathscr{P}(\mathcal{Y}) \to \mathbb{R}_+$ is defined by

$$\mathrm{MMD}_\kappa(\mu, \nu) = \|\mathsf{M}_\kappa \mu - \mathsf{M}_\kappa \nu\|_{\mathcal{H}}. \tag{2}$$

When M_κ is injective , the kernel κ is said to be *characteristic*, and MMD_κ defines a proper metric on $\mathscr{P}(\mathcal{Y})$ (Gretton et al., 2012). The Hilbert space structure induces a simple estimator for the MMD.

Theorem 1.1 (Gretton et al. (2012)): Given independent samples $\{X_i\}_{i=1}^m$ drawn from μ and $\{Y_j\}_{j=1}^n$ drawn from ν, the following is an *unbiased estimator* of $\mathrm{MMD}_\kappa^2(\mu, \nu)$,

$$\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} \kappa(X_i, X_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \kappa(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} \kappa(X_i, Y_j). \tag{3}$$

1.2 Markov Decision Processes and Reinforcement Learning

This proposal is largely focused on analyzing and evaluating decision-making policies in Markov Decision Processes (MDPs). An MDP is defined as a tuple $(\mathcal{X}, \mathcal{A}, r, P, \rho_0)$ where $\mathcal{X}$ is a set of states (the *state space*), $\mathcal{A}$ is a set of actions (the *action space*), $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is a *reward function*, $P : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$ is a Markov kernel on $\mathcal{X}$, and $\rho_0 \in \mathscr{P}(\mathcal{X})$ is the initial state distribution. A Markov decision-making policy is a map $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ that (perhaps randomly) chooses actions in response to state observations. For a policy π, we define an averaged reward function $r^\pi : x \mapsto \mathbb{E}_{A \sim \pi(\cdot \mid x)}[r(x, A)]$. An agent behaving according to π induces a stochastic process on the state space $\{X_t\}_{t \geq 0}$, with

$$X_0 \sim \rho_0, \quad A_t \sim \pi(\cdot \mid X_t), \quad X_{t+1} \sim P^\pi(\cdot \mid X_t) := \int_{\mathcal{A}} P(\cdot \mid X_t, a)\pi(\mathrm{d}a \mid X_t). \tag{4}$$

Policies in reinforcement learning (RL) are evaluated according to how much reward they accumulate. More precisely, we define the (discounted) returns for policy π according to

$$Z^\pi(x, a) = \sum_{t \geq 0} \gamma^t r(X_t, A_t), \quad X_0 = x, A_0 = a, \text{ and } G^\pi(x) = \int_{\mathcal{A}} Z^\pi(x, a)\pi(\mathrm{d}a \mid x), \tag{5}$$

where $\gamma \in (0, 1)$ is a *discount factor* that discounts rewards earned further in the future.

1.2.1 Value Functions, Dynamic Programming, and Temporal Difference Learning

Traditionally, the RL literature has primarily evaluated policies based on the returns they earn *in expectation*. This is captured by the notion of *value functions* $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $V^\pi : \mathcal{X} \to \mathbb{R}$, defined by $Q^\pi(x, a) = \mathbb{E}[Z^\pi(x, a)]$ and $V^\pi(x) = \mathbb{E}[G^\pi(x)]$. In particular, value functions impose a partial order on the set of policies, such that $\pi_1 \succeq \pi_2 \Leftrightarrow V^{\pi_1} \geq V^{\pi_2}$. Notably, it is known that an *optimal* policy always exists – that is, a policy $\pi^\star$ for which $\pi^\star \succeq \pi$ for any other policy π.

A useful fact about value functions is that they satisfy a recurrence known as the *Bellman equation*,

$$V^\pi(x) = \mathbb{E}_{A \sim \pi(\cdot \mid x)}\left[\mathbb{E}_{X' \sim P(\cdot \mid x, A)}[r(x, A) + \gamma V^\pi(X')]\right] =: (\mathcal{T}^\pi V^\pi)(x). \tag{6}$$

The operator $\mathcal{T}^\pi : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is referred to as the *Bellman operator*. Crucially, $\mathcal{T}^\pi$ is a γ-contraction with respect to the ∞-norm, and consequently the Banach fixed point theorem certifies that it has a unique fixed point, namely V^π. This also suggests a *dynamic programming* (DP) approach to estimating V^π: one may start with an initial guess V_0 of the value function, and compute the iterates $V_{k+1} = \mathcal{T}^\pi V_k$; these converge to V^π as confirmed by the Banach fixed point theorem.

To estimate V^π via DP as discussed, one must have exact knowledge of r and P, and one also must evaluate potentially intractable expectations over the state space. Instead, practical implementations of RL employ a technique called *temporal difference learning* (TD learning), where the expectations in Equation 6 are substituted with samples drawn from the policy and transition kernel. Given a sequence of *transition tuples* $\{(X_k, A_k, R_k, X'_k)\}_{k \geq 0}$ with $A_k \sim \pi(\cdot \mid X_k)$, $R_k = r(X_k, A_k)$, and $X'_k \sim P(\cdot \mid X_k, A_k)$, one may compute the following,

$$\left(\hat{\mathcal{T}}_k^\pi V_k\right)(x) := \begin{cases} R_k + \gamma V_k(X'_k) \text{ if } X_k = x \\ V_k(x) \text{ otherwise} \end{cases} \tag{7}$$

Subsequently, TD-learning proceeds by computing the iterates $V_{k+1} = (1 - \alpha_k)V_k + \alpha_k \hat{\mathcal{T}}_k^\pi V_k$ for a sequence of step sizes $\{\alpha_k\}_{k \geq 0}$ and an initial guess V_0 of the value function. It is well known that, under some assumptions on the distribution of $\{X_k\}_{\{k \geq 0\}}$ and the step sizes, $V_k \to V^\pi$ with probability 1. Roughly, this result is an application of stochastic approximation theory to the Bellman equation, leveraging the realization that $\mathbb{E}\left[\left(\hat{\mathcal{T}}_k^\pi V_k\right)(x) \,\middle|\, X_k = x\right] = (\mathcal{T}^\pi V_k)(x)$.

1.2.2 Policy Optimization

So far, we have merely discussed the problem of estimating the value function for a given policy. Often, in RL, we are interested in the problem of finding an (approximately) *optimal* policy. It is well known that a sufficient condition for validating an optimal policy is that it is greedy with respect to its action-value function. Incorporating this into the Bellman equation (Equation 6), we have

$$Q^\star(x,a) = r(x,a) + \gamma \mathbb{E}_{X' \sim P(\cdot \mid x,a)}\left[\max_{a' \in \mathcal{A}} Q^\star(X',a')\right] =: (\mathcal{T}Q^\star)(x,a), \tag{8}$$

where $Q^\star$ is the action-value function for any optimal policy. The *Bellman optimality operator* $\mathcal{T}$ is also known to be a γ-contraction, enabling convergent DP. One can similarly apply TD learning to estimate $Q^\star$ from transition samples by substituting the expression for $\hat{\mathcal{T}}_k^\pi$ in Equation 7 with

$$\left(\hat{\mathcal{T}}_k^\star Q_k\right)(x,a) := \begin{cases} R_k + \gamma \max_{a' \in \mathcal{A}} Q_k(X_k', a') \text{ if } X_k = x \\ Q_k(x,a) \text{ otherwise} \end{cases}, \tag{9}$$

and computing the iterates $Q_{k+1} = (1-\alpha_k)Q_k + \alpha_k \hat{\mathcal{T}}_k^\star Q_k$. This procedure is known as *Q-learning*, and has similar convergence guarantees to temporal difference learning. Q-learning serves as a basis of many of the most successful applications of reinforcement learning in theory and practice.

1.2.3 Zero-Shot Transfer across Reward Functions

The RL methods described above synthesize a value function from an MDP, which is useful only for evaluating a policy for on a *single* reward function. This section will discuss methods for learning *representations* of a policy that enables immediate evaluation across a *class* of reward functions.

The **successor representation** for a policy π was introduced by Dayan (1993) in the setting where $\mathcal{X}$ is finite. By a simple rearrangement of Equation 6, we have $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi =: \mu^\pi r^\pi$. The matrix μ^π exists as a consequence of P^π being a stochastic matrix and $|\gamma| < 1$, and is called the successor representation (SR). Clearly, given knowledge of μ^π and any reward function r, one can immediately infer the value function. We call this ability **zero-shot policy evaluation**. Dayan (1993) noted that column i of μ^π is a value function for the reward function $r(x,a) = \mathbb{1}\{x = i\}$. Thus, DP and TD-learning techniques can be directly applied to learn the SR.

The work of Blier et al. (2021) formally studied the generalization of the SR to infinite state spaces. They define the **successor measure** (SM) m^π as a state-conditioned measure on the state space,

$$m^\pi(\cdot \mid x) = \sum_{t \geq 0} \gamma^t \mathbb{P}(X_t \in \cdot \mid X_0 = x), \quad X_{t+1} \sim P^\pi(\cdot \mid X_t), \tag{10}$$

where the expectation is taken over state sequences $\{X_t\}_{t \geq 0}$ sampled under the policy π. One can define a *normalized* successor measure $\Psi^\pi = (1-\gamma)m^\pi$, which satisfies $\Psi^\pi(\cdot \mid x) \in \mathscr{P}(\mathcal{X})$. The work of Janner et al. (2020) leverages this property to use TD-learning techniques for training Ψ^π as a conditional generative model of state, using common deep neural architectures for generative modeling; they refer to their resulting model as a γ-model. With an estimate of Ψ^π, one can estimate the value function V_r^π corresponding to any reward function r in a zero-shot manner,

$$V_r^\pi(x) = \frac{1}{1-\gamma} \mathbb{E}_{X' \sim \Psi^\pi(\cdot \mid x)}[r^\pi(X')] \approx \frac{1}{N(1-\gamma)} \sum_{i=1}^N r^\pi(X_i), \quad X_i \overset{\text{iid}}{\sim} \Psi^\pi(\cdot \mid x). \tag{11}$$

Successor features (Barreto et al., 2018, SFs) specialize the SM to a finite-dimensional vector space of reward functions, leading to a simpler representation. Given a *feature map* $\varphi : \mathcal{X} \to \mathbb{R}^d$, the SFs $\psi^\pi : \mathcal{X} \to \mathbb{R}^d$ are given by $\psi^\pi(x) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \varphi(X_t) \,\middle|\, X_0 = x\right]$, where $X_{t+1} \sim P^\pi(\cdot \mid X_t)$.

By linearity, it holds that for any r such that $r^\pi(x) = \langle \varphi(x), w_r \rangle$ for some $w_r \in \mathbb{R}^d$, we have $\langle \psi^\pi(x), w_r \rangle = V_r^\pi(x)$, enabling zero-shot policy evaluation. SFs, as "vector-valued value functions", are simpler to model than the SM, but generalize across a smaller space of reward functions.

1.3 Distributional Reinforcement Learning

The field of distributional RL (DRL) departs from the classic RL framework by modeling G^π and Z^π explicitly, providing insight into *risk-sensitive* statistics of returns. Bellemare et al. (2017a) established that the random returns satisfy *distributional* Bellman equations,

$$\begin{aligned} G^\pi(x) &=_{\text{law}} r(x, A) + \gamma G^\pi(X') =: (\mathcal{T}^\pi_{\mathsf{D:x}} G^\pi)(x), \quad A \sim \pi(\cdot \mid x), X' \sim P(\cdot \mid x, A) \\ Z^\pi(x) &=_{\text{law}} r(x, a) + \gamma Z^\pi(X', A') =: \left(\mathcal{T}^\pi_{\mathsf{D:x,a}} Z^\pi\right)(x, a), \quad A' \sim \pi(\cdot \mid X'). \end{aligned} \tag{12}$$

These random variables are not observable, so instead we model their distributions,

$$\eta^\pi(x) := \text{law}(G^\pi(x)) \quad \text{and} \quad \zeta^\pi(x, a) := \text{law}(Z^\pi(x, a)). \tag{13}$$

It is often convenient to express Equation 12 in terms of distributions entirely (Rowland et al., 2019),

$$\begin{aligned} \eta^\pi(x) &= \mathbb{E}_{A \sim \pi(\cdot \mid x)}\left[\mathbb{E}_{X' \sim P(\cdot \mid x, A)}\left[\left(\mathrm{b}_{r(x,A),\gamma}\right)_{\#} \eta^\pi(X')\right]\right] =: (\mathcal{T}^\pi_{\mathsf{D:x}} \eta^\pi)(x) \\ \zeta^\pi(x, a) &= \mathbb{E}_{X' \sim P(\cdot \mid x, a)}\left[\mathbb{E}_{A' \sim \pi(\cdot \mid X')}\left[\left(\mathrm{b}_{r(x,a),\gamma}\right)_{\#} \zeta^\pi(X', A')\right]\right] =: \left(\mathcal{T}^\pi_{\mathsf{D:x,a}} \zeta^\pi\right)(x, a), \end{aligned} \tag{14}$$

where $\mathrm{b}_{a,b}(z) = a + bz$ and $f_{\#}\mu = \mu \circ f^{-1}$. It is known that $\mathcal{T}^\pi_{\mathsf{D:x,a}}$ is a γ-contraction in the *supremal* Wasserstein metrics $\overline{W}_p : \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}} \times \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}} \to \mathbb{R}_+$, given by $\overline{W}_p(\zeta^1, \zeta^2) = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} W_p(\zeta^1(x, a), \zeta^2(x, a))$; once again, appealing to the Banach fixed point theorem, this certifies a convergent *distributional* dynamic programming algorithm for estimating ζ^π.

1.3.1 Tractable Distribution Representations

Distributional dynamic programming as presented above is highly infeasible: probability distributions over $\mathbb{R}$ have infinitely many degrees of freedom, so they cannot be represented in finite memory. Thus, to carry out distributional dynamic programming in practice, one must work with approximate representations of probability distributions. Two prevalent approximations include *categorical representations*, where one fixes a finite set of atom locations in $\mathbb{R}$ and models categorical distributions on those atoms, and *equally-weighted particle (EWP) representations*, where one models empirical distributions on a finite support. As I will discuss in the proposal, these representations each have their own major tradeoffs.

Rather than computing DP iterates of the form $\zeta_{k+1} = \mathcal{T}^\pi_{\mathsf{D:x,a}} \zeta_k$, we must project iterates of the distributional Bellman operator back onto a tractable class. This results in iterates $\zeta_{k+1} = \Pi_m \mathcal{T}^\pi_{\mathsf{D:x,a}} \zeta_k$ with m denoting the number of atoms in the approximation. It is a challenge to establish contractivity of $\Pi_m \mathcal{T}^\pi_{\mathsf{D:x,a}}$, as well as statistical properties necessary for convergence with sample-based TD updates. In general, Π_m and $\mathcal{T}^\pi_{\mathsf{D:x,a}}$ do not commute—so, the fixed point of $\Pi_m \mathcal{T}^\pi_{\mathsf{D:x,a}}$ is generally *not* equal to $\Pi_m \zeta^\pi$, and it is desirable to understand how well these fixed points approximate ζ^π.

In the case of categorical representations, Rowland et al. (2018) showed that the heuristic projection of Bellemare et al. (2017a) is a Hilbert projection under the *Cramér metric*. Then, this work established that the projected operator is a $\sqrt{\gamma}$-contraction in the supremal Cramér metric, whose fixed point approaches ζ^π at a rate of $m^{-\frac{1}{2}}$. As the projection is Hilbertian, the projected operator satisfies a similar unbiasedness property to $\mathcal{T}^\pi$, leading to a simple convergence proof for TD learning.

With regard to EWP projections, Dabney et al. (2017) established non-expansive projection $\Pi_{\text{EWP},m}$ in the W_1 metric (cf. Section 1.1.1), and therefore $\Pi_{\text{EWP},m} \mathcal{T}^\pi_{\mathsf{D:x,a}}$ is a γ-contraction in $\overline{W}_1$. In Bellemare et

al. (2023), it was shown that the fixed point of $\Pi_{\text{EWP},m}\mathcal{T}^{\pi}_{\text{D:x,a}}$ approaches ζ^{π} at a rate of $m^{-1/2}$. However, $\Pi_{\text{EWP},m}$ is not an affine operator, which substantially complicates the analysis of EWP TD algorithms. The solution came much later (Rowland et al., 2024), and required entirely novel analysis techniques.

1.3.2 Policy Optimization with Distributional Reinforcement Learning

So far, in the context of distributional RL, we have only discussed policy *evaluation*; the issue of learning return distributions for optimal policies has remained unaddressed.

The first reason for this is due to one of the most exciting opportunities unlocked by distributional RL: unlike expected returns, there is *freedom* over how one ranks return *distributions*. Aside from the expectation, other functionals can be applied to rank return distributions—these functionals are called *risk measures*. A popular example is the class of *distortion risk measures* (Dabney et al., 2018), which are linear functionals on the inverse CDF of a distribution—a special case is the Conditional Value at Risk at level $\alpha \in [0,1]$ (CVaR_{α}), induced by the funtional $\tau \mapsto \alpha^{-1}\mathbb{1}\{\tau \leq \alpha\}$.

On a more sour note, distributional Bellman operators are generally not as well behaved for policy optimization. This has been noted since the seminal paper on distributional RL, and formalized in Theorem 1.2. The operators $\mathcal{T}_{\text{D:x,a}}$ in this theorem generalize $\mathcal{T}$ for characterizing return distributions. The issue is that while many policies or actions may induce the same expected return (or the same risk measure of the return), the full distribution may still be different. As a consequence, this theorem tells us that we cannot ensure convergence of return distribution estimates with simple distributional-dynamic-programming-like methods. We will revisit this issue in Section 3.1.

Theorem 1.2 (Bellemare et al. (2017a), Lemma 4, Propositions 1, 2, and 3): Let $\rho : \mathscr{P}(\mathbb{R}) \to \mathbb{R}$ be a risk measure and define the set of *greedy policies* for ζ according to ρ by

$$\mathcal{G}^{\rho}_{\zeta} = \left\{\pi : \rho\Big(\mathbb{E}_{A\sim\pi(\cdot\,|\,x)}[\zeta(x,A)]\Big) = \max_{a\in\mathcal{A}} \rho(\zeta(x,a)) \quad \forall x \in \mathcal{X}\right\}. \tag{15}$$

Then, we define the distributional optimality operator $\mathcal{T}_{\text{D:x,a}}$ according to $\mathcal{T}_{\text{D:x,a}}\zeta = \mathcal{T}^{\mathsf{g}(\zeta)}_{\text{D:x}}\zeta$, where g is a *greedy selection rule*, mapping return distribution functions ζ to policies in $\mathcal{G}^{\rho}_{\zeta}$.

Consider the iterates $\zeta_{k+1} = \mathcal{T}_{\text{D:x,a}}\zeta_k$ for an arbitrary $\zeta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. The following hold:

1. If $\rho = \mathbb{E}$, then $\mathbb{E}_{Z_k\sim\zeta_k(x,a)}[Z_k] \to Q^{\star}(x,a)$ uniformly and at a geometric rate; *but* (!)
2. $\mathcal{T}_{\text{D:x,a}}$ is not a contraction in $\overline{W}_p$ and may not have a fixed point;
3. The existence of a fixed point of $\mathcal{T}_{\text{D:x,a}}$ is insufficient to guarantee that $\{\zeta_k\}_{k\geq 0}$ converges.

2 Current Progress

This section outlines contributions of my PhD that have been completed. Henceforth, yellow blocks refer to novel concepts or results, pink blocks refer to concepts that I plan to investigate in future work, and author with names marked with asterisks jointly lead corresponding research projects with me.

2.1 Action Gaps and Advantages in Continuous-Time Distributional RL

This section describes the results of our paper entitled *Action Gaps and Advantages in Continuous-Time Distributional Reinforcement Learning* (Wiltzer et al., 2024a), published in NeurIPS 2024. This projected was completed in collaboration with Marc G. Bellemare, David Meger, Patrick Shafto, and Yash Jhaveri*.

In this work, we investigate the behavior of action-conditioned return distributions (i.e., ζ^{π}) as a function of *decision frequency*— that is, the amount of decisions made by the agent per unit of time. While decision

frequency is not a parameter that is often studied or discussed in the RL literature, it was hypothesized in the early work of Baird (1993) that Q-learning methods can fail at high decision frequency. Baird's intuition was that, when actions persist for such a short period of time, individual actions have minimal influence on the return, and $Q^\pi(x,a) \approx V^\pi(x)$; thus, the estimated ranking of actions by their Q-values is dominated by approximation error in the Q-function. Later, Tallec et al. (2019) proved this in the setting of deterministic dynamics, and Jia and Zhou (2023) extended the result to stochastic MDPs. Ultimately, their results indicate that for policies that make decisions every h units of time, we have $Q^\pi(x,a) = V^\pi(x) + O(h)$. As such, the common approach to resolving this issue (and the one suggested even by Baird (1993)) is to learn the *rescaled advantage* $A_h^\pi(x,a) = (Q^\pi(x,a) - V^\pi(x))/h$ instead of Q^π – this maintains a non-negligible *action gap*, which is more robust to function approximation error.

Our work shows that, while Baird's hypothesis holds when estimating ζ^π instead of Q^π, the situation is much more complicated. Crucially, our work proves that while all statistics of the action-conditioned returns collapse to those of η^π as $h \to 0$, *different statistics collapse at different rates.*

2.1.1 Problem Setup

For a probability measure ν, we denote by F_ν its CDF, and by $F_\nu^{-1} : \tau \mapsto \operatorname{arginf}\{z \in \mathbb{R} : \tau \le F_\nu(z)\}$ its *quantile function*. This study considers the setting where the environment evolves continuously in time. In this setting, we define action-conditioned returns parameterized by a time duration h (known as the decision period), which describes the return achieved after holding a given action for h units of time and following some policy thereafter. Thus, Z^π and Q^π will be substituted with Z_h^π and Q_h^π, defined as

$$\begin{aligned} Z_h^\pi(x,a) &=_{\text{law}} \int_0^h r(X_t)\,\mathrm{d}t + \gamma^h G^\pi(X_h), \quad X_0 = x, A_{0:h} = a \\ \zeta_h^\pi(x,a) &:= \text{law}(Z_h^\pi(x,a)), \text{ and } Q_h^\pi(x,a) := \mathbb{E}[Z_h^\pi(x,a)]. \end{aligned} \tag{16}$$

This is closely related to the setting of *continuous-time distributional RL* (Wiltzer, 2021; Wiltzer et al., 2022), with the major difference being that we limit the decision frequency of the policy to $h^{-1} < \infty$. As hypothesized by Baird (1993) and later proved by Tallec et al. (2019) and Jia and Zhou (2023), Q_h^π has *low action gap* in the following sense,

$$\mathsf{gap}(Q_h^\pi) := \sup_{x \in \mathcal{X}} \max_{a_1 \neq a_2} |Q_h^\pi(x,a_1) - Q_h^\pi(x,a_2)| \lesssim h. \tag{17}$$

On the other hand, the *rescaled advantage* $A_h^\pi = (Q_h^\pi - V^\pi)/h$ proposed by Baird (1993) yields $\mathsf{gap}(A_h^\pi) \in \Theta(1)$. The Advantage Updating (Baird, 1993) and Deep Advantage Updating (Tallec et al., 2019, DAU) algorithms learn A_h^π directly rather than Q_h^π, and demonstrate much faster learning across decision frequencies than their Q-learning counterparts. Naturally, one might expect that the same phenomenon exists in distributional RL, and that a similar rescaling technique is a solution. In this work, we demonstrate that the story is in fact not so simple.

2.1.2 Distributional Action Gap

To investigate the robustness of "distributional Q-learning" algorithms to decision frequency, we begin by defining a distributional notion of action gap. Our definition is the following,

Definition 2 (Distributional Action Gap): For any $\rho : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ and probability metric $d : \mathscr{P}(\mathbb{R}) \times \mathscr{P}(\mathbb{R}) \to \mathbb{R}_+$, the *distributional action gap* $\mathsf{distgap}_d(\rho)$ is defined as

$$\mathsf{distgap}_d(\rho) = \sup_{x \in \mathcal{X}} \max_{a_1 \neq a_2} d(\rho(x,a_1), \rho(x,a_2)). \tag{18}$$

Unlike the case of the scalar action gap, there is no canonical metric for comparing return distributions; one must be careful when choosing an appropriate metric. Consider a deterministic MDP with a deterministic policy π. Then the return distributions contain no more information than their expected returns. But, as long as $\mathsf{gap}(Q_h^\pi) > 0$, we will have $\mathsf{distgap}_{d_{\mathrm{TV}}}(\zeta_h^\pi) = 1$. Despite the fact that return distributions encode only expected returns, the total variation distance cannot identify that the action gaps are vanishing as shown by Tallec et al. (2019). Thus, we will primarily focus on analyzing $\mathsf{distgap}_{W_p}(\cdot)$. We show that W_p distributional action gaps do vanish in stochastic MDPs.

Theorem 2.1: If r is bounded, then $\lim_{h\downarrow 0} \mathsf{distgap}_{W_p}(\zeta_h^\pi) = 0$. Moreover, if r is Lipschitz and episode horizons are finite (but possibly random), it holds that $\mathsf{distgap}_{W_p}(\zeta_h^\pi) \lesssim \sqrt{h}$.

Note that for small h, $\sqrt{h} \geq h$. So, on the surface, Theorem 2.1 appears to be loose, since the decay rate of the action gap is faster. However, surprisingly, we show that this bound is *tight*.

Theorem 2.2: Under the clauses of Theorem 2.1, there are MDPs with $\mathsf{distgap}_{W_p}(\zeta_h^\pi) \gtrsim \sqrt{h}$.

Theorem 2.2 has some deep consequences. Notably, this result establishes that not all statistics of action-conditioned return distributions decay at the same rate: indeed, the mean return decays at a rate of h, and the variance, for example, decays at a rate of $\sqrt{h}$. In particular, if we applied the h-rescaling techniques from existing work to distributional RL algorithms, we would have unbounded distributional action gaps, resulting in distributions with unbounded variance.

This result also sheds light on a statistical challenge in the estimation of A_h^π: this requires estimating the mean of a random variable with unbounded variance. As we will show in Section 2.1.5, this proves to be problematic in stochastic MDPs in practice.

2.1.3 Distributional Superiority

In order to design distributional-action-gap-preserving algorithms, we begin by introducing a distributional notion of the advantage. Concretely, this should represent the distribution of the *difference* in return should an agent execute a given action for h units of time rather than following its policy. Mathematically, this is expressed by the random variable $S_h^\pi(x,a) = Z_h^\pi(x,a) - G^\pi(x)$. However, note that this *is not* equality in law, this expression is an equality in *random variables* and we cannot observe Z_h^π or G^π as random variables directly. Instead, we define the superiority axiomatically.

Axiom 1: For each $(x,a) \in \mathcal{X} \times \mathcal{A}$, there exist random variables $Z_h^{x,a}, G^x$ such that $\mathrm{law}(Z_h^{x,a}) = \zeta^\pi(x,a)$, $\mathrm{law}(G^x) = \eta^\pi(x)$, and $\mathrm{law}(S_h^\pi(x,a)) = \mathrm{law}(Z_h^{x,a} - G^x)$.

Axiom 2: Whenever $\zeta_h^\pi(x,a) = \eta^\pi(x)$, $\psi_h^\pi(x,a) = \delta_0$.

Axiom 1 simply imposes that the superiority is distributed by the difference between $Z \sim \zeta_h^\pi$ and $G \sim \eta^\pi$ under *some* correlation of these random variables—naturally this includes the "oracle" $Z_h^\pi - G^\pi$. The intuition behind Axiom 2 is similar to the property that $A_h^\pi(x,a) = 0$ when the policy takes action a in state x. Here, we additionally require that the difference in (random) returns is identically 0 when the the query action a achieves the same return distribution as the policy. The following result gives a concrete characterization of the superiority distribution.

Theorem 2.3: The function $\psi : (x, a) \mapsto \Delta_{\#}\rho^\star(x, a)$, where $\rho^\star(x, a)$ is the optimal coupling between $\zeta_h^\pi(x, a)$ and $\eta^\pi(x)$, satisfies Axiom 1 and Axiom 2. This is the only such mapping that satisfies the axioms generally (for all MDPs and policies). This mapping can be expressed via

$$\psi(x, a) = \left(F^{-1}_{\zeta_h^\pi(x,a)} - F^{-1}_{\eta^\pi(x)}\right)_{\#} \text{Uniform}(0, 1). \tag{19}$$

Based on Theorem 2.3, we define the superiority distribution function by $\psi_h^\pi(x, a) = \Delta_{\#}\rho^\star(x, a)$. The recent work of Mesnard et al. (2023) employed this object for variance-reduced policy gradients, but our work is the first to justify this as a principaled notion of distributional superiority.

2.1.4 Preserving Distributional Action Gaps with Transformed Superiority

In order to increase the *distributional* action gap, we consider *rescaled* superiority distributions $\psi_{h;q}^\pi$ given by $\psi_{h;q}^\pi(x, a) = (h^{-q}\ \mathsf{id})_{\#}\psi_h^\pi$, where $q > 0$ will be referred to as the *rescaling factor*.

Theorem 2.4:

1. Under the conditions of Theorem 2.1, for all MDPs and policies, $\mathsf{distgap}_{W_p}\left(\psi_{h;q}^\pi\right) \lesssim h^{\frac{1}{2}-q}$.
2. Under the conditions of Theorem 2.2, there exist MDPs where $\mathsf{distgap}_{W_p}\left(\psi_{h;q}^\pi\right) \gtrsim h^{\frac{1}{2}-q}$.

By Theorem 2.4, we see that only the rescaling factor $q = \frac{1}{2}$ leads to order-1 distributional action gaps. Our next theorem shows that rescaling the superiority this way also preserves certain risk-sensitive action rankings, motivating its utility for high-frequency risk-sensitive control.

Theorem 2.5: Let $q, h > 0$ and let ρ be the conditional value at risk, or more generally, any distortion risk measure. Then, for each state x such that $\rho(\eta^\pi(x)) < \infty$,

$$\operatorname*{argmax}_{a\in\mathcal{A}} \rho\left(\psi_{h;q}^\pi(x, a)\right) = \operatorname*{argmax}_{a\in\mathcal{A}} \rho(\zeta_h^\pi(x, a)). \tag{20}$$

Modeling the rescaled advantage

While $\psi_{h;\frac{1}{2}}^\pi$ preserves the distributional action gap, the results of Jia and Zhou (2023) suggest that $h^{-\frac{1}{2}}A_h^\pi(x, a) \to 0$ as $h \to 0$. So, for the sake of risk-neutral control, $\psi_{h;\frac{1}{2}}^\pi$ may struggle to identify the optimal action in the presence of approximation error. We consider two alternatives, where we model:

1. $\psi_{h;1}^\pi$ (instead of $\psi_{h;\frac{1}{2}}^\pi$) for risk-neutral control;
2. *Shifted* superiority distributions $\vartheta_{h;\frac{1}{2}}^\pi(x, a)$, where $F^{-1}_{\vartheta^\pi_{h;\frac{1}{2}}(x,a)} = A_h^\pi(x, a)\left(1 - h^{\frac{1}{2}}\right) + F^{-1}_{\psi^\pi_{h;\frac{1}{2}}(x,a)}$.

While both solutions should provide order-1 (expected) action gaps for risk-neutral control, we hypothesize that neither is sufficient. The former option will suffer from unbounded distribution tails, which will raise difficulties in function approximation. The latter still involves estimating A_h^π, which is the mean of a random variable with unbounded variance, which we anticipate to a major statistical challenge. Moreover, Theorem 2.5 would *not* be satisfied with the shifted superiority distribution, making the approach specific to the risk-neutral setting.

These issues highlight the technical challenges associated with the phenomena that we uncovered in this work. We identify the following open problem as a very interesting direction for future work.

Research Question 1: Determine a map $\Upsilon_h^\pi : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ and a class $\mathsf{C} \supsetneq \{\mathbb{E}\}$ of risk measures for which
1. $\mathsf{distgap}_{W_p}(\Upsilon_h^\pi) \in \Theta(1)$ and $\mathsf{gap}\big((x,a) \mapsto \mathbb{E}_{Z\sim\Upsilon_h^\pi(x,a)}[Z]\big) \in \Theta(1)$;
2. $\operatorname{argmax}_{a\in\mathcal{A}} \Upsilon_h^\pi(x,a) = \operatorname{argmax}_{a\in\mathcal{A}} \zeta_h^\pi(x,a)$ for each $x \in \mathcal{X}$ and $\rho \in \mathsf{C}$.

Algorithms for learning distributional superiority

Leveraging results from continuous-time distributional RL (Wiltzer, 2021), it is simple to show that

$$F^{-1}_{\zeta_h^\pi(x,a)} = F^{-1}_{\eta^\pi(x)} + h^q F^{-1}_{\psi^\pi_{h;q}(x,a)} \text{ by Theorem 2.3, and} \tag{21}$$

$$\begin{aligned} Z_h^\pi(X_t,a) &=_{\text{law}} hr(X_t) + \gamma^h G^\pi(X_{t+h}) + o(h) \\ \therefore F^{-1}_{\zeta_h^\pi(X_t,a)} &= hr(X_t) + \gamma^h F^{-1}_{\eta^\pi(X_{t+h})} + o(h) \approx hr(X_t) + \gamma^h F^{-1}_{\eta^\pi(X_{t+h})}. \end{aligned} \tag{22}$$

As such, we
1. Estimate the quantile function of $\eta^\pi(x)$ using standard QR-DQN (Dabney et al., 2017);
2. Learn the quantile function of $\psi^\pi_{h;q}(x,a)$ by quantile regression between the RHS of Equation 21 and Equation 22; without propagating gradients through $\psi^\pi_{h;q}$ in the latter.

Axiom 1 is satisfied implicitly here by Equation 21. We enforce Axiom 2 by representing $F^{-1}_{\psi^\pi_{h;q}}$ by

$$F^{-1}_{\psi^\pi_{h;q}(x,a)} = \phi(x,a) - \phi(x,a^\star), \quad a^\star \in \operatorname*{argmax}_{a\in\mathcal{A}} \rho(\phi(x,a)), \tag{23}$$

where ρ is a distortion risk measure, and ϕ denotes the function approximator being tuned. This structure ensures that $\psi^\pi_{h;q}(x,a^\star) = \delta_0$, which satisfies Axiom 2 for a ρ-greedy policy π. In our experiments, we refer to this algorithm as $\mathsf{DSUP}(q)$ — we consider only $q \in \{\frac{1}{2}, 1\}$, encompassing our proposed rate ($q = \frac{1}{2}$) and that of Baird ($q = 1$).

We also consider a similar algorithm for learning the shifted superiority distribution $\vartheta^\pi_{h;\frac{1}{2}}$ in risk-neutral control. To do so, we learn A_h^π using the Deep Advantage Updating algorithm (Tallec et al., 2019, DAU), with parameter sharing between the function approximators for A_h^π and ϕ to leverage the representation learning benefits of distributional RL for estimating the advantage. Then, we may act greedily according to A_h^π. The resulting algorithm is referred to as $\mathsf{DAU{+}DSUP}(\frac{1}{2})$.

2.1.5 Experiments

Next, we estimate the rescaled (and shifted) superiority, and examine the qualitative characteristics of the resulting distributions, as well as its performance in optimal (risk-sensitive) control.

Monte Carlo Estimation

We begin by estimating the superiority and its transformations by Monte Carlo in a simple continuous MDP; the resulting distributions are shown in Figure 1 across a range of decision frequencies.

As predicted by our theory, ζ_h^π tends towards the Dirac at 0 for high decision frequencies. Moreover, the 1-rescaled distribution superiority has unbounded variance in this limit (see the horizontal axis). When scaling the superiority by our proposed $h^{\frac{1}{2}}$ rate, we find that the superiority distributions do not collapse or explode. Interestingly, we see that the shifted superiority $\vartheta^\pi_{h;\frac{1}{2}}$ is in fact roughly centered at the ground-truth mean of 100, however the centers are quite noisy: we interpret this as a consequence of estimating the mean of a very high-variance object (the h-rescaled advantage).

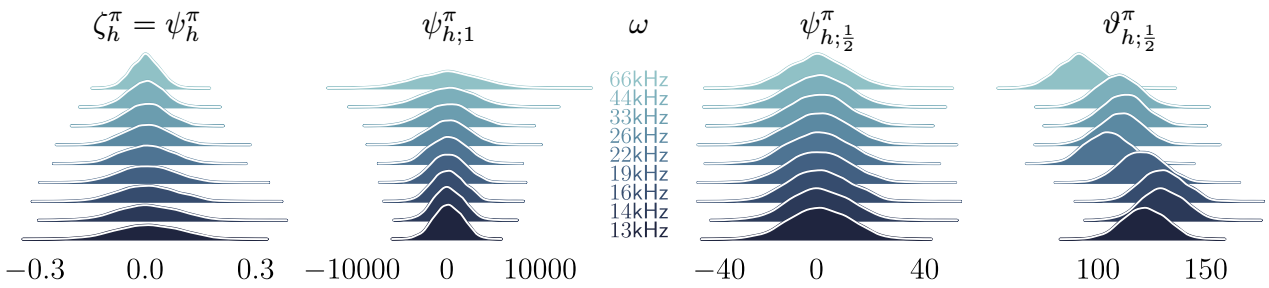

Figure 1: Monte-Carlo estimates of $\psi_{h;q}^\pi$ for $q \in \{0, 1, \frac{1}{2}\}$, and $\vartheta_{h;\frac{1}{2}}^\pi$, as a function of decision frequencies $\omega = \frac{1}{h}$.

High-Frequency Distributional TD-Learning

Next, we employ DSUP(q), DAU+DSUP($\frac{1}{2}$), and baselines QR-DQN (Dabney et al., 2017) and DAU (Tallec et al., 2019) to learn a risk-neutral policy for American Option Trading, using the benchmark of Lim and Malik (2022), for $q \in \{\frac{1}{2}, 1\}$. This benchmark estimates the parameters of a financial model from stock data to sample arbitrarily-many transitions during training, and then testing on held-out data. Using the estimated financial model, we may increase the time resolution to evaluate our DRL algorithms at arbitrarily high decision frequency. Our results are shown in Figure 2.

We see that only DSUP($\frac{1}{2}$) reliably maintains performance as the decision frequency increases: it is the only algorithm that is robust to decision frequency. DSUP(1), DAU+DSUP($\frac{1}{2}$), and DAU perform well at select high decision frequencies, but they are ultimately unreliable. This is due to the difficulty of estimating statistics of distributions with unbounded variance.

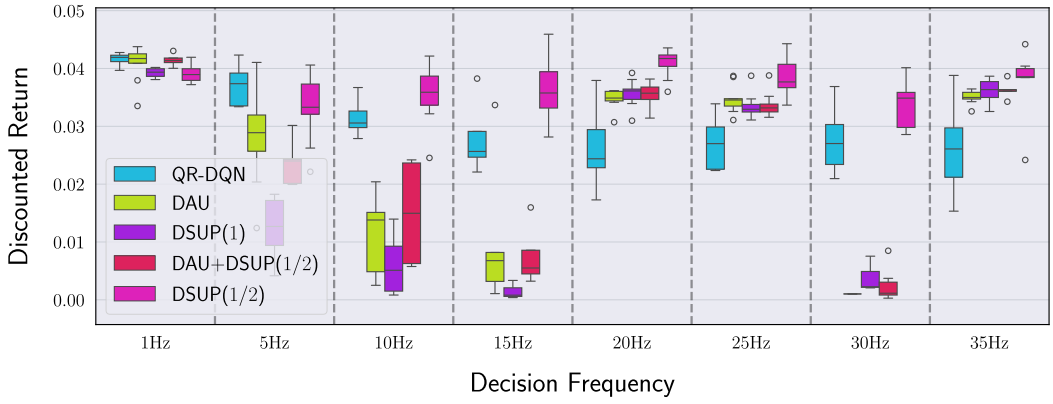

Figure 2: Risk-neutral algorithms for high-frequency option-trading as a function of decision frequency.

Finally, we evaluate DSUP($\frac{1}{2}$) and QR-DQN for the purpose of risk-sensitive control in this domain.

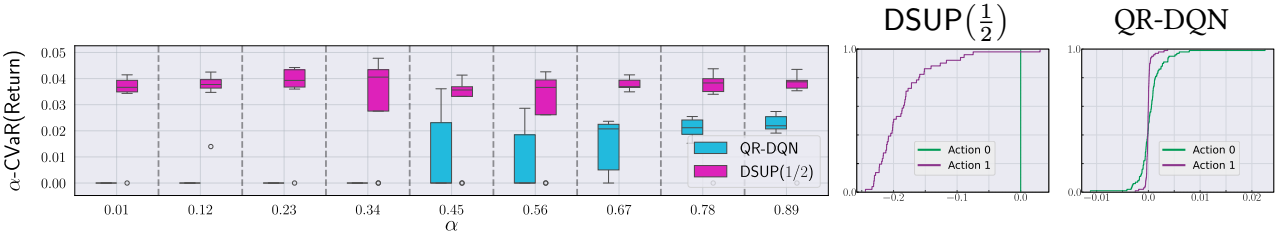

Figure 3: Performance of CVaR$_\alpha$-optimizing agents in American Option Trading at $\omega = 35$Hz.

Figure Figure 3 shows that, as α decreases (the risk-measure becomes more pessimistic / risk-averse), DSUP($\frac{1}{2}$) becomes far more performant. The two rightmost plots depict the rescaled superiority distributions for DSUP($\frac{1}{2}$) and action-conditioned return distributions for QR-DQN when $\alpha = 0.12$: notably, the superiority distributions clearly depict that action 1 (corresponding to selling the stock) can only produce worse outcomes, while the action-conditioned return distributions learned by QR-DQN are much more similar to each other. This highlights the benefit of learning the rescaled superiority in place of action-conditioned return distributions for high-frequency control.

2.2 Zero-Shot Distributional Transfer in RL

In this section, I describe the results of our paper entitled *Foundations of Multivariate Distributional Reinforcement Learning* (Wiltzer et al., 2024b), which was presented at NeurIPS 2024. This work was completed in collaboration with Jesse Farebrother, Arthur Gretton, and Mark Rowland. This paper contributes results leading to a better theoretical understanding of our ICML 2024 work on the *Distributional Successor Measure* (Wiltzer et al., 2024c), another contribution of my PhD, with Jesse Farebrother*, Arthur Gretton, Yunhao Tang, André Barreto, Will Dabney, Marc G. Bellemare, and Mark Rowland.

This project rigorously analyzes algorithms for learning *distributional successor features*. Recall that successor features (cf. Section 1.2.3) are value functions for a multivariate reward function $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. Distributional successor features (DSFs), which we denote $\daleth^\pi$, can be thought of as return distribution functions corresponding to multivariate reward functions,

$$\daleth^\pi(x) = \operatorname{law}\left(\sum_{t\geq 0}\gamma^t\varphi(X_t)\,\middle|\, X_0 = x\right), \quad X_{t+1}\sim P^\pi(\cdot \mid X_t). \tag{24}$$

For any reward function of the form $r(x) = \langle \varphi(x), w\rangle$ for $w \in \mathbb{R}^d$, we have that $(z \mapsto \langle z, w\rangle)_\# \daleth^\pi(x) = \eta_r^\pi(x)$, enabling zero-shot distributional policy evaluation. If $\varphi(x)$ is a one-hot encoding of x (or a Dirac located at x for continuous state spaces), $\daleth^\pi$ is called the *distributional successor measure* (DSM), and it maps any bounded measurable reward function to its corresponding return distribution function under π—we derive an algorithm for estimating the DSM from data in Wiltzer et al. (2024c).

Our work is not the first to study DSFs. The work of Gimelfarb et al. (2021) presented a method for learning DSFs under the assumption that each dimension of the features is statistically independent — crucially, this prohibits zero-shot distributional policy evaluation. The works of Freirich et al. (2019) and Zhang et al. (2021) presented algorithms for learning (proper) DSFs, but without providing any convergence analysis for them. Moreover, Wu et al. (2023) provided convergence analysis for learning DSFs, but only with access to intractable maximum likelihood estimation oracles.

While our work (Wiltzer et al., 2024c) presented some conceptual results about the DSM as well as an algorithmic framework for estimating it, it left a notable gap between the algorithms with theoretical guarantees and the practical framework introduced. Namely,

1. It studied only the convergence of DP, which requires complete knowledge of the MDP dynamics;
2. It studied only the *nonparametric* distributional Bellman operator – that is, the operator that applies to intractable distribution representations.

In this work, we present several computationally tractable algorithms, with provable convergence guarantees, that approximate $\daleth^\pi$. In the sequel, we assume that the image of φ is contained in $[0, R_{\max}]^d$, and we define $\mathcal{R} = [0, (1-\gamma)^{-1}R_{\max}]^d$ (so that $\daleth^\pi(x) \in \mathscr{P}(\mathcal{R})$).

2.2.1 Dynamic Programming in the Right Metric Space

Existing works that have studied distributional dynamic programming for DSFs have demonstrated that the multivariate distributional Bellman operator[1] is a contraction in $\overline{W}_p$ (Zhang et al., 2021), however, optimization of Wasserstein metrics for multivariate variables from samples is known to be problematic or intractable (Bellemare et al., 2017b). As such, we extend a result of Nguyen et al. (2020) and demonstrate that the multivariate distributional Bellman operator is contractive for a class of maximum mean discrepancies, which admit unbiased sample estimators. I will specialize the results in this text to MMD with *energy distance kernels* $\kappa_\alpha(z_1, z_2) = \frac{1}{2}\left(\|z_1\|_2^\alpha + \|z_2\|_2^\alpha - \|z_1 - z_2\|_2^\alpha\right)$ for $\alpha \in (0, 2)$, however the results hold more broadly (Wiltzer et al., 2024b).

[1] This is the same as Equation 14, allowing the reward function to have multidimensional image.

Theorem 2.6: The distributional Bellman operator $\mathcal{T}^\pi_{\mathsf{DSF}}$ is a $\gamma^{\frac{\alpha}{2}}$-contraction in $\overline{\mathrm{MMD}}_{\kappa_\alpha}$.

This result firstly justifies the use of certain MMD metrics for multivariate distributional RL, however we have still not dealt with the issue of finitely-parameterizing DSF representations or model-free learning.

2.2.2 Multivariate Distributional Dynamic Programming with Finite Parameterizations

As discussed in Section 1.3.1, in any tractable algorithm, it is necessary to pick a finitely-parameterized class of return distribution approximations, and study *projected operators* that maintain distributions among this class. Our work studies both EWP and categorical representations, presenting the first algorithms for tractable multivariate distributional dynamic programming.

2.2.2.1 EWP Projected Distributional Dynamic Programming

The projection we consider onto the class of EWP representations is simply that projection which minimizes the MMD. That is, the *EWP MMD projection* $\Pi^m_{\mathrm{EWP},\kappa} : \mathscr{P}(\mathcal{R})^{\mathcal{X}} \to \mathscr{C}_{\mathrm{EWP},m}$ is given by

$$\left(\Pi^m_{\mathrm{EWP},\kappa} \daleth\right)(x) \in \operatorname*{arginf}_{p \in \mathscr{C}^\circ_{\mathrm{EWP},m}} \mathrm{MMD}_\kappa(p, \daleth(x)). \tag{25}$$

Unfortunately, this projection is not even uniquely-defined in the scalar ($d = 1$) case (Rowland et al., 2024), and is generally *not* a non-expansion in $\overline{\mathrm{MMD}}_\kappa$ as shown by Lhéritier and Bondoux (2022) and Rowland et al. (2024) – this precludes standard techniques from providing convergence proofs. Moreover, the optimization problem in Equation 25 is non-convex, precluding efficient computation.

In our work, we solve these issues by demonstrating that a notion of *approximate projection* suffices to ensure that dynamic programming iterates closely approximate $\daleth^\pi$. Ultimately, we introduce the *randomized* operator $\mathsf{BootProj}^\pi_{\kappa,m}$ as a proxy for $\Pi_{\mathrm{EWP},m}\mathcal{T}^\pi_{\mathsf{DSF}}$, and it is defined as follows,

$$\left(\mathsf{BootProj}^\pi_{\kappa,m} \daleth\right)(x) = \frac{1}{m}\sum_{i=1}^m \delta_{\varphi(x)+\gamma Z_i}, \quad Z_i \sim \daleth(X_i), \quad X_i \overset{\text{iid}}{\sim} P^\pi(\cdot \mid x). \tag{26}$$

Using this randomized operator, we can achieve accurate approximations of $\daleth^\pi$.

Theorem 2.7: Denote by κ_α the energy distance kernel with parameter $\alpha \in (0, 2)$. For any $\daleth_0 \in \mathscr{P}(\mathcal{R})^{\mathcal{X}}$ for which $\overline{\mathrm{MMD}}_{\kappa_\alpha}(\daleth_0, \daleth^\pi) \leq D < \infty$, sample a sequence of iterates $\daleth_{k+1} = \mathsf{BootProj}^\pi_{\kappa_\alpha,m}\daleth_k$. Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\overline{\mathrm{MMD}}_{\kappa_\alpha}(\daleth_K, \daleth^\pi) \in \tilde{O}\left(\frac{d^{\frac{\alpha}{2}} R^\alpha_{\max}}{(1-\gamma^{-\frac{\alpha}{2}})(1-\gamma)^\alpha\sqrt{m}} \log\left(\frac{|\mathcal{X}|\delta^{-1}}{\log \gamma^{-\frac{\alpha}{2}}}\right)\right) \tag{27}$$

where $K = \left\lceil \frac{\log m}{\alpha \log \gamma^{-1}} \right\rceil$, and $\tilde{O}$ omits logarithmic factors in m.

This is the first known algorithm to produce a tractable approximation of $\daleth^\pi$, with $m \in \mathsf{polylog}\left(d, \frac{1}{\varepsilon}\right)$ for an ε-approximation of $\daleth^\pi$ in the case of $d > 1$. This bound matches the convergence rate for approximating scalar return distribution functions with EWP representations given in Bellemare et al. (2023), generalizing the result to a wider collection of kernels and higher-dimensional reward functions.

2.2.2.2 Categorical Projected Distributional Dynamic Programming

For categorical representations, we show that a MMD projection is more well-behaved. We consider a class $\mathscr{C}_{\mathrm{C}}^{\mathsf{S}}$ induced by a *support map* $\mathsf{S} : \mathcal{X} \to \mathcal{R}^*$, which associates to each state x a fixed support, and we denote $N(x) = |\mathsf{S}(x)|$. Then, we consider the projection $\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} : \mathscr{P}(\mathcal{R})^{\mathcal{X}} \to \mathscr{C}_{\mathrm{C}}^{\mathsf{S}}$ given by

$$\left(\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} \daleth\right)(x) = \operatorname*{arginf}_{p \in \mathscr{P}(\mathsf{S}(x))} \mathrm{MMD}_{\kappa}(p, \daleth(x)). \tag{28}$$

First, we establish that dynamic programming with the projected operator $\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi}$ is convergent.

Theorem 2.8: For any energy distance kernel κ_α, $\Pi_{\mathrm{C},\kappa_\alpha}^{\mathsf{S}}$ is uniquely-defined, the operator $\Pi_{\mathrm{C},\kappa_\alpha}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi}$ is a $\gamma^{\frac{\alpha}{2}}$-contraction in $\overline{\mathrm{MMD}}_{\kappa_\alpha}$, and for any $\daleth_0 \in \mathscr{C}_{\mathrm{C}}^{\mathsf{S}}$ with $\overline{\mathrm{MMD}}_{\kappa_\alpha}(\daleth_0, \daleth^\pi) \leq D < \infty$, the iterates $\daleth_{k+1} = \Pi_{\mathrm{C},\kappa_\alpha}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi} \daleth_k$ converge in $\overline{\mathrm{MMD}}_{\kappa_\alpha}$ to the fixed point $\daleth_{\mathrm{C},\kappa_\alpha}^{\pi}$.

Our analysis yields a tractable algorithm for computing $\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi}$ via quadratic programming (i.e., using a quadratic programming solver $\mathsf{QPSolve}$ such as OSQP (Stellato et al., 2020)), shown in Algorithm 1.

Algorithm 1: Categorical MMD Projection

Require: Support map S, kernel κ, transition kernel P^π, feature map φ, discount γ, DSF $\daleth \in \mathscr{C}_{\mathrm{C}}^{\mathsf{S}}$
for $x \in \mathcal{X}$ **do**
  $(\mathcal{T}_{\mathsf{DSF}}^{\pi} \daleth)(x) \leftarrow \sum_{x' \in \mathcal{X}} \sum_{\xi \in \mathsf{S}(x')} P^\pi(x' \mid x)[\daleth(x')](\{\xi\}) \delta_{\varphi(x) + \gamma\xi}$
  $K_{i,j}^x \leftarrow \kappa(\xi_i, \xi_j)$ for each $(\xi_i, \xi_j) \in \mathsf{S}(x)^2$
  $q_j^x \leftarrow \sum_{\xi \in \mathrm{supp}((\mathcal{T}_{\mathsf{DSF}}^{\pi} \daleth)(x))} [(\mathcal{T}_{\mathsf{DSF}}^{\pi} \daleth)(x)](\{\xi\}) \kappa(\xi_j, \xi)$ for each $\xi_j \in \mathsf{S}(x)$
  $p \leftarrow \mathsf{QPSolve}\left(\min_{p \in \mathbb{R}^{N(x)}} [p^\top K^x p - 2p^\top q] \text{ subject to } p \succeq 0, \sum_i p_i = 1\right)$
  $\left(\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi} \daleth\right)(x) \leftarrow \sum_{i=1}^{N(x)} p_i \delta_{\mathsf{S}(x)_i}$
end for
Return $\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi} \daleth$

While we have demonstrated a tractable algorithm for approximating the fixed point $\daleth_{\mathrm{C},\kappa}^{\pi}$ of $\Pi_{\mathrm{C},\kappa}^{\mathsf{S}} \mathcal{T}_{\mathsf{DSF}}^{\pi}$, we have not yet established that this is a reasonable approximation of $\daleth^\pi$. Our work demonstrates that the quality of the approximation $\daleth_{\mathrm{C},\kappa}^{\pi}$ is controlled by the most favorable partition of $\mathcal{R}$ in a sense we will define below. For a given support map S and any state x, we consider the class of partitions $\mathcal{P}_{\mathsf{S}(x)}$ of sets $P \subset 2^{\mathcal{R}}$ such that

1. $|P| = N(x), \bigcup_{p \in P} p = \mathcal{R}$, and each $p_i \in P$ contains exactly one element $\xi_i \in \mathsf{S}(x)$;
2. For any $p_1, p_2 \in P$, either $p_1 \cap p_2 = \emptyset$ or $p_1 = p_2$;

For any partition $P \in \mathcal{P}_{\mathsf{S}(x)}$, we define a notion of *mesh size* that will measure the quality of $\daleth_{\mathrm{C},\kappa}^{\pi}$.

Definition 3: For any finite partition P of $\mathcal{R}$, we define the *mesh size* $\mathsf{mesh}(P; \alpha)$ according to

$$\mathsf{mesh}(P; \alpha) = \max_{p \in P} \sup_{\xi_1, \xi_2 \in p} \|y_1 - y_2\|_2^\alpha. \tag{29}$$

Now, we can bound the discrepancy between $\daleth_{\mathrm{C},\kappa}^{\pi}$ and $\daleth^\pi$.

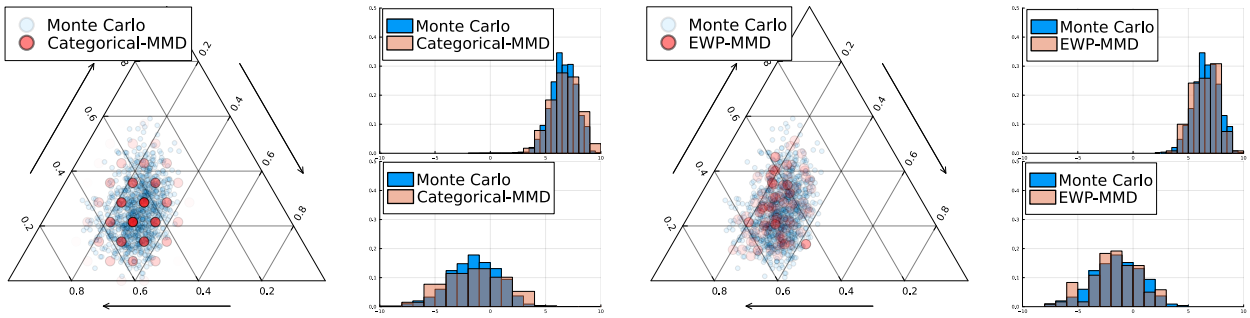

Figure 4: DSMs learned by the EWP and categorical DP methods for a tristate MDP. **Simplex plots**: Points in the simplex correspond to random occupancy measures; these plots show distributions over random occupancy measures learned by the projected DP methods in red, with MC estimates of the DSM in blue. **Histograms**: Histograms show results of zero-shot distributional evaluation for two randomly-sampled reward functions (corresponding to rows).

Theorem 2.9: Let κ_α for $\alpha \in (0, 2)$ denote an energy distance kernel. Then,

$$\overline{\mathrm{MMD}}_{\kappa_\alpha}\left(\daleth^\pi_{\mathrm{C},\kappa_\alpha}, \daleth^\pi\right) \leq \frac{1}{1-\gamma^{\frac{\alpha}{2}}} \sup_{x\in\mathcal{X}} \inf_{P\in\mathcal{P}_{\mathsf{S}(x)}} \sqrt{\mathsf{mesh}(P;\alpha)}. \tag{30}$$

When S uniformly discretizes $\mathcal{R}$ with m atoms at each state, this materializes to

$$\overline{\mathrm{MMD}}_{\kappa_\alpha}\left(\daleth^\pi_{\mathrm{C},\kappa_\alpha}, \daleth^\pi\right) \leq \frac{1}{(1-\gamma^{\frac{\alpha}{2}})(1-\gamma)^{\frac{\alpha}{2}}} \frac{d^{\frac{\alpha}{4}} R_{\max}^{\frac{\alpha}{2}}}{\left(m^{\frac{1}{d}}-2\right)^{\frac{\alpha}{2}}}. \tag{31}$$

This matches the $O\left(m^{-\frac{1}{2}}\right)$ rate of Rowland et al. (2018) in the $d = 1$ case, again generalizing the result to a family of kernels and multidimensional rewards.

2.2.2.3 Simulation

To illustrate the convergence of the proposed algorithms and the qualitative properties of the EWP and categorical representations, we used the algorithms above to learn the DSM of a tristate MDP. The results are shown in Figure 4, along with examples of zero-shot return distribution predictions.

2.2.3 Multivariate Distributional Temporal Difference Learning

Analysis of TD-learning is challenging because it involves tracking a dynamical system with a stochastically perturbed version, which can lead to quick divergence. Generally, to ensure that this does not happen, it suffices to show (roughly) that the expectation of the perturbed dynamics of the system equal the unperturbed dynamics, as well as a variance bound on the random perturbations. In our work, we focused on analyzing the convergence of multivariate distributional TD-learning under categorical representations of the form described in Section 2.2.2.2. This is because the categorical DP algorithm that we studied produces iterates evolving under a deterministic dynamical system, unlike the randomized EWP DP algorithm of Section 2.2.2.1.

Unfortunately, however, we identified a severe issue: the MMD projection $\Pi^{\mathsf{S}}_{\mathrm{C},\kappa}$ *does not* in general commute with the expectation of successor states when $d > 1$ (unlike the case with $d = 1$, (Rowland et al., 2018)). Thus, it *is not* true that the projection of a sampled distributional Bellman backup is equal in expectation to $\Pi^{\mathsf{S}}_{\mathrm{C},\kappa}\mathcal{T}^\pi_{\mathsf{DSF}}$. To handle this issue, rather than representing our DSFs as categorical distributions, we represent them as mass-1 *signed measures* on supports dictated by S.

Algorithmically, this involves replacing $\Pi^{\mathsf{S}}_{\mathrm{C},\kappa}$ with a novel *signed* categorical projection $\Pi^{\mathsf{S}}_{\mathrm{SC},\kappa} : \mathscr{M}_1\left(\mathbb{R}^d\right)^{\mathcal{X}} \to \mathscr{C}^{\mathsf{S}}_{\mathrm{SC}}$, where $\mathscr{M}_1(\mathcal{Y})$ denotes the set of signed measures μ on a set $\mathcal{Y}$ with $\mu(\mathcal{Y}) = 1$, and $\mathscr{C}^{\mathsf{S}}_{\mathrm{SC}}$ is the space of signed categorical representations given by

$$\mathscr{C}^{\mathsf{S}}_{\mathrm{SC}} = \left\{ x \mapsto \sum_{i=1}^{N(x)} p_i(x) \mathsf{S}(x)_i : \sum_i p_i(x) = 1 \right\}. \tag{32}$$

Notably, to compute this projection, it suffices to simply remove the constraint that $p \succeq 0$ in Algorithm 1. In our work, we showed that dynamic programming is still convergent with this operator, and that the price paid by modeling signed measures is relatively benign.

Theorem 2.10: Let κ be a kernel satisfying the conditions of Theorem 2.9. Then the operator $\Pi^{\mathsf{S}}_{\mathrm{SC},\kappa} \mathcal{T}^{\pi}_{\mathsf{DSF}} : \mathscr{C}^{\mathsf{S}}_{\mathrm{SC}} \to \mathscr{C}^{\mathsf{S}}_{\mathrm{SC}}$ has unique fixed point $\daleth^{\pi}_{\mathrm{SC},\kappa}$, and iterates $\daleth_{k+1} = \Pi^{\mathsf{S}}_{\mathrm{SC},\kappa} \mathcal{T}^{\pi}_{\mathsf{DSF}} \daleth_k$ converge geometrically with rate $\gamma^{\frac{c}{2}}$ to $\daleth^{\pi}_{\mathrm{SC},\kappa}$. Moreover, we have that

1. $\overline{\mathrm{MMD}}_{\kappa_\alpha}\left(\daleth^{\pi}_{\mathrm{SC},\kappa}, \daleth^{\pi}\right) \leq \frac{1}{1-\gamma^{\frac{c}{2}}} \sup_{x \in \mathcal{X}} \inf_{P \in \mathcal{P}_{\mathsf{S}(x)}} \sqrt{\mathsf{mesh}(P; \alpha)}$; and
2. $\overline{\mathrm{MMD}}_{\kappa_\alpha}\left(\Pi^{\mathsf{S}}_{\mathrm{C},\kappa} \daleth^{\pi}_{\mathrm{SC},\kappa}, \daleth^{\pi}\right) \leq \left(1 + \frac{1}{1-\gamma^{\frac{c}{2}}}\right) \sup_{x \in \mathcal{X}} \inf_{P \in \mathcal{P}_{\mathsf{S}(x)}} \sqrt{\mathsf{mesh}(P; \alpha)}$.

The second part of Theorem 2.10 shows that we can approximate $\daleth^{\pi}$ among the class of signed measures, and then project this approximation back onto the class of probability measures without ruining the approximation bound.

Finally, we showed that TD-learning is convergent with signed categorical representations.

Theorem 2.11: Let κ be a kernel satisfying the conditions of Theorem 2.9. Given a sequence $\{T_t\}_{\{t \geq 0\}}$ of *transitions* $T_t = (X_t, A_t, R_t, X'_t)$ such that $A_t \sim \pi(\cdot \mid X_t), R_t = \varphi(X_t), X'_t \sim P^{\pi}(\cdot \mid X_t)$, define the sequence of operators

$$\left(\hat{\mathcal{T}}^{\pi}_{\mathsf{DSF},t} \daleth\right)(x) = \begin{cases} \left(\mathrm{b}_{R_t,\gamma}\right)_{\#} \daleth(X'_t) \text{ if } x = X_t \\ \daleth(x) \text{ otherwise} \end{cases}. \tag{33}$$

With probability 1, the iterates $\hat{\daleth}_{k+1} = (1-\alpha_t)\hat{\daleth}_k + \alpha_t \Pi^{\mathsf{S}}_{\mathrm{SC},\kappa} \hat{\mathcal{T}}^{\pi}_{\mathsf{DSF},t} \hat{\daleth}_k$ converge to $\daleth^{\pi}_{\mathrm{SC},\kappa}$.

2.2.3.1 Simulation

We simulated our signed categorical TD-learning algorithm with the EWP TD-learning algorithm of Zhang et al. (2021), for which no convergence guarantees are known. Our results are shown in Figure 5. Both methods achieve reasonable accuracy. Notably, the EWP representation contains particles that lie outside the true support of the DSFs, which is likely a consequence of convergence to a local optimum.

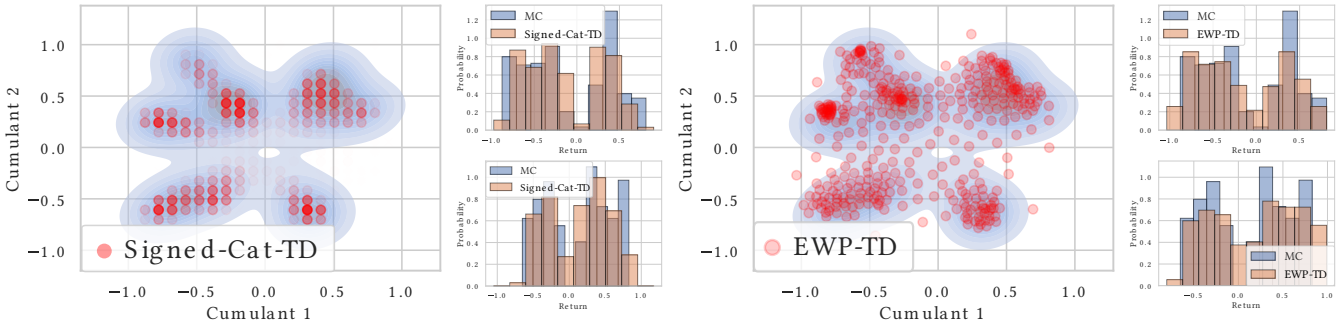

Figure 5: Distributional SFs learned by signed categorical TD (cf. Theorem 2.11) and EWP TD (cf. (Zhang et al., 2021)). **Scatter plots**: distributions over multivariate returns (DSFs) in red, with blue mass depicting the DSFs estimated by Monte Carlo. **Histograms**: predicted return distributions (red) for two randomly sampled reward functions in the span of the features, with blue histograms denoting the return distributions estimed by Monte Carlo.

3 Ongoing and Future Work

This section will outline two research projects that I plan on completing over the next year. First, I will discuss recent results concerning convergence of return distributions in policy optimization, providing directions for future work. Second, I will propose a framework under which distributional RL, and particularly distributional SFs (cf. Section 2.2), can be leveraged for robust imitation learning.

3.1 Convergent Control in Distributional Reinforcement Learning

In Section 1.3.2, we saw that standard approaches to policy optimization in DRL do not induce convergent return distributions. In this section, I propose a framework under which we can achieve convergence of return distributions under approximate policy optimization via regularization.

The reason for the lack of convergence is that, while we can ensure convergence of the value function (the expected returns) under policy improvement, we *cannot* ensure convergence of the policy. Bellemare et al. (2023, Proposition 7.7) gives an explicit example of a MDP and a risk-neutral greedy selection rule (cf. Theorem 1.2) under which return distributions oscillate indefinitely.

While this example illustrates a bold result, it is relatively contrived: greedy selection rule is not reminiscent of any policy optimization algorithm (distrbutional or otherwise) seen in practice. To begin, we illustrate that the conclusion (that return distributions do not converge under policy optimization) holds much more generally, even with standard approaches to distributional policy optimization. In particular, return distributions do not converge as long as the rewards are (even slightly) stochastic. We demonstrate this phenomenon in Figure 6.

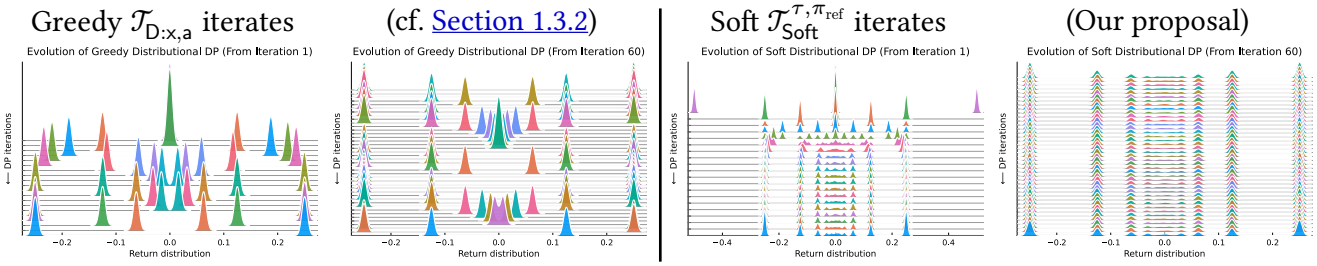

Figure 6: Dynamics of return distributions with standard updates (left) and our regularized updates (right). Plots show evolution of DP iterates from top to bottom. Colors distinguish distributions learned at different DP iterations.

Figure 6 illustrates the evolution of return distributions at the initial state of the MDP of Bellemare et al. (2023, Proposition 7.7), but with independent $\mathcal{N}(0, 10^{-5})$ noise added to the rewards (which has much smaller scale than the reward function). The plots on the left depict the return distributions under the distributional optimality operator with the greedy selection rule g given by

$$(\mathsf{g}(\zeta))(x) = \min \operatorname*{argmax}_{a \in \mathcal{A}} \mathbb{E}_{Z_a \sim \zeta(x,a)}[Z_a]. \tag{34}$$

The outer minimum simply selects the "first" of the mean-return-maximizing actions under some ordering of the action space. This is equivalent to the greedy selection rule implicit in common DRL algorithms such as C51 (Bellemare et al., 2017a) as well as standard value-based RL algorithms such as DQN, due to the ubiquitous convention in software libraries to implement argmax as min argmax. Notably, this greedy selection rule is "too nice" to exhibit non-convergence in the example of Bellemare et al. (2023, Proposition 7.7), but we see in Figure 6 that even a miniscule amount of noise induces oscillation. The idea here is that this noise, even nearly convergence, is enough to change the set of actions chosen by g, preventing convergence of the policy (and its return distributions), even though the mean returns converge. The issue with the distributional optimality operator is that it is *not continuous*: small changes to return distributions can induce a different ordering over statistics of the action-conditioned returns, possibly resulting in large changes to the estimated greedy policy.

The right side of Figure 6 demonstrates the dynamics of our proposed approach, which I will outline next. Broadly, the idea is to introduce a *continuous* greedy selection rule, so that convergence of the statistics of interest of the action-conditioned return distributions will induce convergence of the return distributions themselves.

3.1.1 Proposed Approach

Bellemare et al. (2023, Theorem 7.9) provides a condition under which risk-neutral distributional policy optimization *does* produce convergent return distributions: uniqueness of the optimal policy. The idea is that, since the expected returns *do* converge, after enough dynamic programming iterations, the error in estimated expected returns will be small enough to correctly identify the optimal action at each state. After this point, the policy selected by any greedy selection rule will remain the same, and so the procedure reduces to distributional policy evaluation, which provides convergent return distributions.

Inspired by this result, our approach employs a strongly convex regularizer to ensure the uniqueness of the optimum. The standard objective in risk-neutral RL is

$$\max_{\mu} J(\mu) \coloneqq \max_{\mu} \int_{\mathcal{X}} \mu r \, \mathrm{d}\nu_0 : \mu \text{ is a SR for some policy} \tag{35}$$

where $\nu_0 \in \mathscr{P}(\mathcal{X})$ is some initial state distribution. We instead consider the *maximum entropy RL* objective (Ziebart et al., 2008),

$$\max_{\mu} J_\tau(\mu) \coloneqq \max_{\mu} \int_{\mathcal{X}} \left[(\mu r)(x) + \tau \int_{\mathcal{X}} D_{\mathsf{KL}}(\pi^\mu(\cdot \mid x') \parallel \pi_{\text{ref}}(\cdot \mid x')) \mu(\mathrm{d}x' \mid x) \right] \nu_0(\mathrm{d}x), \tag{36}$$

where π^μ is the policy with SR μ, $\tau > 0$ is a regularization weight referred to as the *temperature*, and $\pi_{\text{ref}} : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ is any *reference policy* (for instance, the uniform random policy). Due to the strong convexity of D_{KL}, J_τ is in fact strongly convex, and consequently it has a unique optimal occupancy measure $\mu_\tau^\star$. Given the correspondence between occupancy measures and policies, there is a unique optimal policy π_τ for this regularized RL objective.

Our intuition is that the uniqueness of the optimal policy will enable stable distributional policy optimization. Towards this end, we introduce the following *soft distributional Bellman operator*,

$$\begin{aligned}
\left(\mathcal{T}_{\mathsf{Soft}}^{\tau,\pi_{\text{ref}}} \zeta\right)(x,a) &\coloneqq \int_{\mathcal{X}} \sum_{a' \in \mathcal{A}} \left[\left(\mathrm{b}_{\tilde{r}_\tau^{\pi_{\text{ref}}}(x,x'),\gamma} \right)_{\#} \zeta^\pi(x',a') \right] (\mathcal{B}^\tau \zeta)(a' \mid x') P(\mathrm{d}x' \mid x,a), \\
\text{where } \tilde{r}_\tau^{\pi_{\text{ref}}}(x,x') &\coloneqq r(x) + \gamma \tau D_{\mathsf{KL}}((\mathcal{B}^\tau \zeta)(\cdot \mid x') \parallel \pi_{\text{ref}}(\cdot \mid x')), \quad \text{and} \\
(\mathcal{B}^\tau \zeta)(x) &\coloneqq \text{Categorical}\left(\text{softmax}\left(\left\{ \mathbb{E}_{Z_a \sim \zeta(x,a)} \left[\frac{Z_a}{\tau} \right] \right\}_{a \in \mathcal{A}} \right) \right).
\end{aligned} \tag{37}$$

We have shown the following result concerning this operator,

Theorem 3.1: For any $\tau > 0$, there exists $C_\tau \in \mathbb{R}$ and $\zeta^\tau : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ such that the iterates $\{\zeta_n\}_{\{n \geq 0\}} \subset (\mathscr{P}(\mathbb{R}))^{\mathcal{X} \times \mathcal{A}}$ given by $\zeta_{n+1} = \mathcal{T}_{\mathsf{Soft}}^{\tau,\pi_{\text{ref}}} \zeta_n$ satisfy

$$\overline{W}_p(\zeta_n, \zeta^\tau) \leq C_\tau n \gamma^n \tag{38}$$

for any $\zeta_0 : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$, and $\zeta_n \to \zeta^\tau$ in $\overline{W}_p$. Additionally, $\mathcal{B}^\tau \zeta^\tau$ (cf. Equation 37) is the policy that optimizes the objective of Equation 36.

Ultimately, Theorem 3.1 establishes an algorithm for approximating the optimal policy for the regularized RL objective via return distribution iterates that *converge*.

3.1.2 Looking Forward

Theorem 3.1 unlocks several interesting research directions. I will outline three of them below.

Improved training dynamics in deep RL We have already seen that both distributional RL and regularized RL (e.g., with entropy regularization) have lead to stronger performance when training agents with nonlinear function approximators. Is there still performance to be gained by combining the two? Our existing results are promising: since we can now ensure convergence of return distributions, it may be easier to learn the fixed point of $\mathcal{T}_{\mathsf{Soft}}^{\tau,\pi_{\mathrm{ref}}}$ than $\mathcal{T}_{\mathsf{D:x,a}}$, since the target distributions will eventually stabilize (in theory). Moreover, applications of $\mathcal{T}_{\mathsf{Soft}}^{\tau,\pi_{\mathrm{ref}}}$ involve computing a mixture distribution over all "next actions", thereby using information from each action-conditioned return distribution estimate in each target distribution—this hedges action-conditioned return estimates, which can promote robustness in training, and provides learning signal to each action-conditioned return distribution on every gradient. Finally, in *streaming* RL settings (Elsayed et al., 2024), the convergence of iterates under $\mathcal{T}_{\mathsf{Soft}}^{\tau,\pi_{\mathrm{ref}}}$ may be critical.

Enhancing risk-sensitive policy optimization As we saw, existing policy optimization methods in distributional RL induce oscillatory return distribution estimates. While the return distribution means converge, their other statistics do not. Does this make risk-sensitive control fundamentally more difficult? Can we leverage the stable dynamics under $\mathcal{T}_{\mathsf{Soft}}^{\tau,\pi_{\mathrm{ref}}}$ updates to solve this problem?

Distributionally-reinforced fine-tuning In recent years, regularized RL has been highly influential as a tool to fine-tune a pre-trained policy—in such settings, the pre-trained policy takes the place of π_{ref}, and we aim to optimize returns without straying too far from π_{ref} (precisely like in Equation 36). We have introduced the first distributional RL approach to regularized RL; how can we use distributional RL to ameliorate fine-tuning methods?

3.2 Alignment by Distributional Feature Matching

In *imitation learning* and *inverse reinforcement learning*, a central goal is to synthesize a policy that behaves similarly to given *demonstrations*, as opposed to maximizing returns for a given reward function. A prominent approach to achieving this is to learn the reward function that maximally characterizes the difference in returns between demonstration trajectories (sampled from an *expert policy* π_{E}) and rollouts from the learned policy π. It is simple to see that this is equivalent to approximating the SR of the demonstration policy in *operator norm*—assuming a deterministic initial state x_0 for simplicity,

$$\|\mu^{\pi}(\cdot \mid x_0) - \mu^{\pi_{\mathrm{E}}}(\cdot \mid x_0)\|_{\mathcal{R}^*} \equiv \sup_{r \in \mathcal{R}} \frac{|(\mu^{\pi} r)(x_0) - (\mu^{\pi_{\mathrm{E}}} r)(x_0)|}{\|r\|_{\mathcal{R}}} = \sup_{r \in \mathcal{R}} \frac{|V_r^{\pi_{\mathrm{E}}}(x_0) - V_r^{\pi}(x_0)|}{\|r\|_{\mathcal{R}}}, \quad (39)$$

where $\mathcal{R}$ is a hypothesis class of reward functions, and $\mathcal{R}^*$ is its algebraic dual. The insight here is that rather than finding the reward function r above that maximally distinguishes the policies, the same effect can be achieved by learning a policy π whose SR approximates that of the expert.

However, matching the SR on a continuous (and maybe high-dimensional) state space can be quite challenging if $\mathcal{R}$ has little structure. Thus, it is common (Ziebart et al., 2008) to impose structure on $\mathcal{R}$, particularly, by assuming it is contained in a finite-dimensional space (say, a bounded subset of $\mathbb{R}^d$). Under this assumption, the LHS of Equation 39 is precisely the difference in *successor features* between the learned policy and the expert, for a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ satisfying

$$\mathcal{R} = \{x \mapsto \langle \varphi(x), w \rangle : w \in \mathbb{R}^d\}. \quad (40)$$

One can simply take policy gradients through the error $\|\psi^{\pi}(x_0) - \psi^{\pi_{\mathrm{E}}}(x_0)\|_2$ in successor features from the initial state x_0, which presents an attractive framework for imitation learning (Jain et al., 2025).

However, naturally, some information has been lost: by constraining the hypothesis class of reward functions to a finite dimensional space, there will generally be expert behavior that cannot be described by a reward function in $\mathcal{R}$.

Following the framework of Jain et al. (2025), we can learn an imitation policy by taking policy gradients through a distributional SF loss between the learned policy and the expert data. Rather than minimizing the ℓ_2 norm between SFs of the learner and the experts, we would minimize the maximum mean discrepancy (cf. Section 1.1.2) between distributional SFs, using techniques from Wiltzer et al. (2024b) to estimate the distributional SFs of the learner policy π and the expert policy π_{E},

$$\min_{\pi} \mathrm{MMD}_{\kappa_\alpha}(\daleth^{\pi}(x_0), \daleth^{\pi_{\mathrm{E}}}(x_0)). \tag{41}$$

Our proposal is that by matching *distributional* successor features (Wiltzer et al. (2024b) or Section 2.2), we may substantially broaden the scope of expert behavior that can be captured with a fixed feature map φ. We briefly investigated this in Wiltzer et al. (2024b, Appendix F). Suppose we wish to imitate an expert parking a car—there is a parking spot directly in front of a car, but a median blocks its path. Moreover, suppose the features satisfy $\varphi(x) = (x_1, \mathbb{1}\{x \text{ is in parking configuration}\})^{\top}$, where x_1 indicates the lateral component of the car's position x.

Consider a scenario where we learn from a diverse dataset of expert demonstrations. The SFs $\psi^{\pi_{\mathrm{E}}}$ will be roughly proportional to $(0, 1)^{\top}$, since by symmetry, experts should spend an equal amount of time navigating leftward around the median as they do navigating around the median on the right. So, it is possible to match the expert's SF by driving straight through the median! We immediately see that such a catastrophe is avoidable by *distributional* SF matching. Indeed, the *DSF* of the expert will have *no* probability mass on the line $z \mapsto (0, z)^{\top}$: consequently, matching the DSF of the expert requires an understanding that one must navigate around the median to the left or right (and have roughly equal probability mass assigned to the two).

This example is a certificate of DSF matching offering benefits for imitation learning. Can we quantify when DSF matching is beneficial, and establish in what sense it is beneficial? Moreover, does DSF matching improve training dynamics at scale (such as imitation learning from camera observations)?

4 Timeline and Conclusion

In this proposal, I presented two completed works developing theory and algorithms for enriching distributional representations of future behavior for high-frequency long-horizon control, and for zero-shot generalization across rewards and utility functions. In addition to these papers, I jointly lead the research resulting in (Wiltzer et al., 2024c), which I discussed briefly in connection to my work on distributional successor features.

To round off my thesis, I proposed two additional projects. The first project, encompassing the novel convergent regularized distributional control algorithm and its applications to RL fine-tuning, will nominally be submitted to NeurIPS 2025 in May. The second project will employ our distributional SF techniques for robust learning from demonstrations, and will nominally be submitted to ICLR 2026 in October. Finally, I plan to begin writing my thesis following my submission to NeurIPS, and aim to complete the thesis prior to January 2026.

Bibliography

Leemon C. Baird. *Advantage Updating*, 1993. https://doi.org/10.21236/ADA280862

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor Features for Transfer in Reinforcement Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A Distributional Perspective on Reinforcement Learning. *International Conference on Machine Learning (ICML)*, 2017a.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. The MIT Press, 2023. https://doi.org/10.7551/mitpress/14207.001.0001

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramér Distance as a Solution to Biased Wasserstein Gradients. *CoRR*, 2017b. http://arxiv.org/abs/1705.10743

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning Successor States and Goal-Dependent Values: A Mathematical Viewpoint. *CoRR*, 2021. http://arxiv.org/abs/2101.07123

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. *Implicit Quantile Networks for Distributional Reinforcement Learning* (Issue arXiv:1806.06923). arXiv, 2018. https://doi.org/10.48550/arXiv.1806.06923

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional Reinforcement Learning with Quantile Regression. *AAAI Conference on Artificial Intelligence*, 2017.

Peter Dayan. Improving Generalisation for Temporal Difference Learning: The Successor Representation. *Neural Computation*, *5*(4), 613–624, 1993.

Mohamed Elsayed, Gautham Vasan, and A. Rupam Mahmood. Streaming Deep Reinforcement Learning Finally Works. *CoRR*, 2024.

Dror Freirich, Tzahi Shimkin, Ron Meir, and Aviv Tamar. Distributional Multivariate Policy Evaluation and Exploration with the Bellman GAN. *International Conference on Machine Learning (ICML)*, 2019.

Michael Gimelfarb, André Barreto, Scott Sanner, and Chi-Guhn Lee. Risk-Aware Transfer in Reinforcement Learning Using Successor Features. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, *13*(25), 723–773, 2012.

Arnav Kumar Jain, Harley Wiltzer, Jesse Farebrother, Irina Rish, Glen Berseth, and Sanjiban Choudhury. Non-Adversarial Inverse Reinforcement Learning via Successor Feature Matching. *International Conference on Learning Representations (ICLR)*, 2025.

Michael Janner, Igor Mordatch, and Sergey Levine. Gamma-Models: Generative Temporal Difference Learning for Infinite-Horizon Prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yanwei Jia and Xun Yu Zhou. q-Learning in Continuous Time. *Journal of Machine Learning Research*, *24*(161), 1–61, 2023.

Alix Lhéritier and Nicolas Bondoux. A Cramér Distance perspective on Quantile Regression based Distributional Reinforcement Learning. *Artificial Intelligence and Statistics (AISTATS)*, 2022.

Shiau Hong Lim and Ilyas Malik. Distributional Reinforcement Learning for Risk-Sensitive Policies. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Thomas Mesnard, Wenqi Chen, Alaa Saade, Yunhao Tang, Mark Rowland, Theophane Weber, Clare Lyle, Audrunas Gruslys, Michal Valko, Will Dabney, Georg Ostrovski, Eric Moulines, and Remi Munos. Quantile Credit Assignment. *International Conference on Machine Learning (ICML)*, 24517–24531, 2023.

Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional Reinforcement Learning via Moment Matching. *AAAI Conference on Artificial Intelligence*, 2020.

Mark Rowland, Marc G. Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An Analysis of Categorical Distributional Reinforcement Learning. *Artificial Intelligence and Statistics (AISTATS)*, 2018.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and Samples in Distributional Reinforcement Learning. *International Conference on Machine Learning (ICML)*, 2019.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G. Bellemare, and Will Dabney. An Analysis of Quantile Temporal-Difference Learning. *Journal of Machine Learning Research*, *25*(163), 1–47, 2024.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Scholkopf. Injective Hilbert Space Embeddings of Probability Measures. *Conference on Learning Theory (COLT)*, 2008.

B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, *12*(4), 637–672, 2020. https://doi.org/10.1007/s12532-020-00179-2

Corentin Tallec, Léonard Blier, and Yann Ollivier. Making Deep Q-learning Methods Robust to Time Discretization. *International Conference on Machine Learning (ICML)*, 2019.

Cédric Villani. *Optimal transport: old and new* (Vol. 338). Springer, 2009.

Harley Wiltzer. *On the Evolution of Return Distributions in Continuous-Time Reinforcement Learning*, 2021.

Harley Wiltzer, Marc G. Bellemare, David Meger, Patrick Shafto, and Yash Jhaveri. Action Gaps and Advantages in Continuous-Time Distributional Reinforcement Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of Multivariate Distributional Reinforcement Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, Yunhao Tang, André Barreto, Will Dabney, Marc G. Bellemare, and Mark Rowland. A Distributional Analogue to the Successor Representation. *International Conference on Machine Learning (ICML)*, 2024c.

Harley Wiltzer, David Meger, and Marc G. Bellemare. Distributional Hamilton-Jacobi-Bellman Equations for Continuous-Time Reinforcement Learning. *International Conference on Machine Learning (ICML)*, 2022.

Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional Offline Policy Evaluation with Predictive Error Guarantees. *International Conference on Machine Learning (ICML)*, 2023.

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional Reinforcement Learning for Multi-Dimensional Reward Functions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum Entropy Inverse Reinforcement Learning. *AAAI Conference on Artificial Intelligence*, 2008.