Diverse Projection Ensembles for Distributional Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In contrast to classical reinforcement learning, distributional RL algorithms aim to learn the distribution of returns rather than their expected value. Since the nature of the return distribution is generally unknown a priori or arbitrarily complex, a common approach finds approximations within a set of representable, parametric distributions. Typically, this involves a *projection* of the unconstrained distribution onto the set of simplified distributions. We argue that this projection step entails a strong inductive bias when coupled with neural networks and gradient descent, thereby profoundly impacting the generalization behavior of learned models. In order to facilitate reliable uncertainty estimation through diversity, this work studies the combination of several different projections and representations in a distributional ensemble. We establish theoretical properties of such *projection ensembles* and derive an algorithm that uses ensemble disagreement, measured by the average 1-Wasserstein distance, as a bonus for deep exploration. We evaluate our algorithm on the behavior suite benchmark and find that diverse projection ensembles lead to significant performance improvements over existing methods on a wide variety of tasks with the most pronounced gains in directed exploration problems.

1 Introduction

In reinforcement learning (RL), agents interact with an unknown environment, aiming to acquire policies that yield high cumulative rewards. In pursuit of this objective, agents must engage in a trade-off between information gain and reward maximization, a dilemma known as the exploration/exploitation trade-off. In the context of model-free RL, many algorithms designed to address this problem efficiently rely on a form of the *optimism in the face of uncertainty* principle (Auer, 2002) where agents act according to upper confidence bounds of value estimates. When using high-capacity function approximators (e.g., neural networks) the derivation of such confidence bounds for value estimates is non-trivial. One popular approach employs an ensemble of learned approximations (Dietterich, 2000; Lakshminarayanan et al., 2017) with the intuition that independent models generalize differently to unseen states or actions, yielding a measure of ensemble disagreement as a proxy for uncertainty. Only upon visiting a state-action region sufficiently often are ensemble members expected to converge to almost equal predictions. This notion of reducible uncertainty is also known as *epistemic uncertainty* (Hora, 1996; Der Kiureghian and Ditlevsen, 2009).

A concept somewhat orthogonal to epistemic uncertainty is *aleatoric uncertainty*, that is the uncertainty associated with the inherent irreducible randomness of an event. The latter is the subject of the recently popular *distributional* branch of RL (Bellemare et al., 2017), which aims to approximate the distribution of returns, as opposed to only its mean. While distributional RL naturally lends itself to risk-sensitive learning, several empirical results show significant improvements over classical RL even when distributions are used only to recover the mean (Bellemare et al., 2017; Dabney et al.,

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

2018b; Rowland et al., 2019; Yang et al., 2019; Nguyen-Tang et al., 2021). Since the space of potential return distributions is infinite-dimensional in general, many recent advancements in this field rely on novel methods to *project* the unconstrained return distribution onto a set of representable distributions.

In this paper, we study the combination of different *projections* and *representations* in an ensemble of distributional value learners. We argue that the inductive bias imposed by the projection step leads to distinct generalization behaviors when joined with neural function approximation and gradient descent. We thus deem distributional projections instrumental to the construction of diversely generalizing ensembles and subsequently to effective epistemic uncertainty quantification and exploration. Fig. 1 illustrates the distinct generalization signatures of two such projection methods on a toy regression problem.

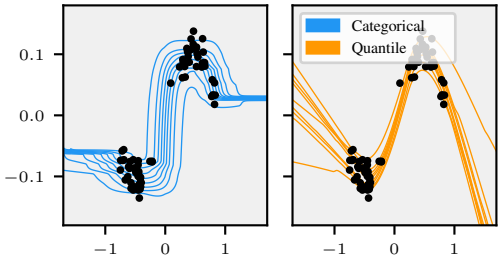

Figure 1: Fit and generalization on a toy 1-D regression task. Black dots are training data points. Shown here are 9 quantile lines of learned distributions based on categorical and quantile projections.

Our main contributions are as follows:

(1) We introduce distributional *projection ensembles* and analyze their properties theoretically. In our setting, each model is iteratively updated toward the projected mixture over ensemble return distributions. We describe such use of distributional ensembles formally through a *projection mixture operator* and establish several of its properties, including contractivity and residual approximation errors.

(2) In connection with the projection mixture operator, we derive a *propagation* method for epistemic uncertainty that reconciles the use of a joint target distribution with directed exploration. In particular, we relate distributional temporal difference (TD) errors, as measured by the 1-Wasserstein distance, to errors regarding the true return distribution. We leverage this insight to devise an optimism-based algorithm for directed exploration that leverages projection ensembles as an intrinsic signal.

(3) We implement the above-described algorithmic components in a deep RL setting and evaluate the resulting agent on the behavior suite (Osband et al., 2019a), a benchmark collection consisting of a wide range of tasks including a total of 468 environments. Our experiments show that *projection ensembles* aid reliable uncertainty estimation and learning stability, outperforming baselines on most tasks, even when compared to significantly larger ensemble sizes.

2 Related Work

Our work builds on a swiftly growing body of literature in distributional RL (Rösler, 1992; Morimura et al., 2010; Bellemare et al., 2017). In particular, several of our theoretical results rely on previous work by Rowland et al. (2018) and Dabney et al. (2018b), who first provided contraction properties of categorical and quantile projections in distributional RL respectively. Numerous recently proposed algorithms (Dabney et al., 2018a; Rowland et al., 2019; Yang et al., 2019; Nguyen-Tang et al., 2021) are based on novel representations and projections. To the best of our knowledge, our work is the first to study the combination of different projection operators and representations in the context of distributional RL. Several works, however, have applied ensemble techniques to distributional approaches. For example, Eriksson et al. (2022) and Hoel et al. (2023) use a bootstrapped ensemble of distributional models to derive aleatoric and epistemic risk measures. Nikolov et al. (2018) combine a bootstrapped deterministic DQN ensemble with a distributional categorical model for information-directed sampling. In a broader sense, the use of deep ensembles for value estimation and exploration is widespread (Osband et al., 2016, 2019b; Flennerhag et al., 2020; Fellows et al., 2021; Chen et al., 2017). A notable distinction between such algorithms is whether independence between ensemble members is maintained or whether a joint bootstrapped TD target is used. Our work falls into the latter category, necessitating a propagation mechanism that reestablishes value uncertainty rather than uncertainty in TD targets (Fellows et al., 2021; Moerland et al., 2017; Janz et al., 2019). Our proposed propagation scheme establishes a temporal consistency between *distributional* temporal difference errors and errors in the true return distributions in a similar fashion to the uncertainty Bellman equations (O'Donoghue et al., 2018).

3 Background

Throughout this work, we consider a finite Markov Decision Process (MDP) (Bellman, 1957) of the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma, P, \mu)$ as the default problem framework, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ is the immediate reward distribution, $\gamma \in [0, 1]$ is the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition kernel, and $\mu : \mathscr{P}(\mathcal{S})$ is the start state distribution. Here, we write $\mathscr{P}(\mathcal{X})$ to indicate the space of probability distributions defined over some space $\mathcal{X}$. Given a state S_t at time t, agents draw an action A_t from a stochastic policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$ to be presented the random immediate reward $R_t \sim \mathcal{R}(\cdot|S_t, A_t)$ and the successor state $S_{t+1} \sim P(\cdot|S_t, A_t)$. Under policy π and transition kernel P, the discounted return is a random variable given by the discounted cumulative sum of random rewards according to $Z^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t R_t$, where $R_t \sim \mathcal{R}(S_t, A_t)$ and $S_0 = s, A_0 = a$. Note that our notation will generally use uppercase letters to indicate random variables. Furthermore, we write $\mathcal{D}(Z^\pi(s, a)) \in \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ to denote the distribution of the random variable Z^π, that is a state-action-dependent distribution residing in the space of probability distributions $\mathscr{P}(\mathbb{R})$. For explicit referrals, we label this distribution $\eta^\pi(s, a) = \mathcal{D}(Z^\pi(s, a))$. The expected value of $Z^\pi(s, a)$ is known as the state-action value $Q^\pi(s, a) = \mathbb{E}[Z^\pi(s, a)]$ and adheres to a temporal consistency condition described by the Bellman equation (Bellman, 1957)

$$Q^\pi(s, a) = \mathbb{E}_{P,\pi}[R_0 + \gamma Q^\pi(S_1, A_1)|S_0 = s, A_0 = a]\,, \tag{1}$$

where $\mathbb{E}_{P,\pi}$ indicates that successor states and actions are drawn from P and π respectively. Moreover, the Bellman operator $T^\pi Q(s, a) := \mathbb{E}_{P,\pi}[R_0 + \gamma Q(S_1, A_1)|S_0 = s, A_0 = a]$ has the unique fixed point $Q^\pi(s, a)$.

3.1 Distributional reinforcement learning

The *distributional* Bellman operator $\mathcal{T}^\pi$ (Bellemare et al., 2017) is a probabilistic generalization of T^π and considers return distributions rather than their expectation. For notational convenience, we first define P^π to be the transition operator according to

$$P^\pi Z(s, a) :\overset{D}{=} Z(S_1, A_1), \qquad \text{where} \qquad S_1 \sim P(\cdot|S_0 = s, A_0 = a), \quad A_1 \sim \pi(\cdot|S_1), \tag{2}$$

and $\overset{D}{=}$ indicates an equality in distributional law (White, 1988). In this setting, the distributional Bellman operator is defined as

$$\mathcal{T}^\pi Z(s, a) :\overset{D}{=} R_0 + \gamma P^\pi Z(s, a). \tag{3}$$

Similarly to the classical Bellman operator, the distributional counterpart $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ has the unique fixed point $\mathcal{T}^\pi Z^\pi = Z^\pi$, that is the true return distribution Z^π. In the context of iterative algorithms, we will also refer to the identity $\mathcal{T}^\pi Z(s, a)$ as a bootstrap of the distribution $Z(s, a)$. For the analysis of many properties of $\mathcal{T}^\pi$, it is helpful to define a distance metric over the space of return distributions $\mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$. Here, the supremum p-Wasserstein metric $\bar{w}_p : \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \times \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \to [0, \infty]$ has proven particularly useful. In the univariate case, $\bar{w}_p$ is given by

$$\bar{w}_p(\nu, \nu') = \sup_{s,a\in\mathcal{S}\times\mathcal{A}} \left(\int_0^1 |F^{-1}_{\nu(s,a)}(\tau) - F^{-1}_{\nu'(s,a)}(\tau)|^p d\tau\right)^{\frac{1}{p}}, \tag{4}$$

where $p \in [0, \infty)$, ν, ν' are any two state-action return distributions, and $F_{\nu(s,a)} : \mathbb{R} \to [0, 1]$ is the cumulative distribution function (CDF) of $\nu(s, a)$. For notational brevity, we will use the notation $w_p(\nu(s, a), \nu'(s, a)) = w_p(\nu, \nu')(s, a)$ for the p-Wasserstein distance between distributions ν, ν', evaluated at (s, a). One of the central insights of previous works in distributional RL is that the operator $\mathcal{T}^\pi$ is a γ-contraction in $\bar{w}_p$ (Bellemare et al., 2017), meaning that we have $\bar{w}_p(\mathcal{T}^\pi\nu, \mathcal{T}^\pi\nu') \leq \gamma\bar{w}_p(\nu, \nu')$, a property that allows us (in principle) to construct convergent value iteration schemes in the distributional setting.

3.2 Categorical and quantile distributional RL

In practice, implementing an iteration scheme based strictly on the distributional Bellman operator $\mathcal{T}^\pi$ is impossible without significant restrictions on the return distributions that occur in an MDP. This is because we can not, in general, represent arbitrary probability distributions in $\mathscr{P}(\mathbb{R})$ and instead resort to parametric models capable of representing a subset $\mathscr{F}$ of $\mathscr{P}(\mathbb{R})$. Following Bellemare et al. (2023), we refer to $\mathscr{F}$ as a *representation* and define it to be the set of parametric distributions P_θ with $\mathscr{F} = \{P_\theta \in \mathscr{P}(\mathbb{R}) : \theta \in \Theta\}$. Furthermore, we define the *projection operator* $\Pi : \mathscr{P}(\mathbb{R}) \to \mathscr{F}$ to be a mapping from the space of probability distributions $\mathscr{P}(\mathbb{R})$ to the representation $\mathscr{F}$. Recently,

two particular choices for representation and projection have proven highly performant in deep RL: the *categorical* and *quantile* model.

The **categorical representation** (Bellemare et al., 2017; Rowland et al., 2018) assumes a weighted mixture of K Dirac delta distributions $\delta_{z_k}(z)$ with support at evenly spaced locations $z_k \in [z_1, ..., z_K]$. The categorical representation $\mathscr{F}_C$ is thus given by

$$\mathscr{F}_C = \big\{ \sum_{k=1}^{K} \theta_k \delta_{z_k}(z) \,\Big|\, \theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1 \big\}. \tag{5}$$

The corresponding categorical projection operator Π_C maps a distribution ν from $\mathscr{P}(\mathbb{R})$ to a distribution in $\mathscr{F}_C$ by assigning probability mass inversely proportional to the distance to the closest z_k in the support $[z_1, ..., z_K]$ for every point in the support of ν. For example, for a single Dirac distribution $\delta_x(z)$ and assuming $z_k \leq x \leq z_{k+1}$ the projection is given by

$$\Pi_C \delta_x(z) = \tfrac{z_{k+1} - x}{z_{k+1} - z_k} \delta_{z_k}(z) + \tfrac{x - z_k}{z_{k+1} - z_k} \delta_{z_{k+1}}(z). \tag{6}$$

The corner cases are defined such that $\Pi_C \delta_x(z) = \delta_{z_1}(z) \, \forall x \leq z_1$ and $\Pi_C \delta_x(z) = \delta_{z_K}(z) \, \forall x \geq z_K$. It is straightforward to extend the above projection step to finite mixtures of Dirac distributions through $\Pi_C \sum_i p_i \delta_{z_i}(z) = \sum_i p_i \Pi_C \delta_{z_i}(z)$.

The **quantile representation** (Dabney et al., 2018b), like the categorical representation, comprises mixture distributions of Dirac deltas $\delta_{\theta_k}(z)$, but in contrast parametrizes their locations rather than probabilities. Assuming equal weighting among the Dirac distributions, this yields the representation

$$\mathscr{F}_Q = \big\{ \sum_{k=1}^{K} \tfrac{1}{K} \delta_{\theta_k}(z) \,\Big|\, \theta_k \in \mathbb{R} \big\}. \tag{7}$$

For some distribution $\nu \in \mathscr{P}(\mathbb{R})$, the quantile projection $\Pi_Q \nu$ is a mixture of K Dirac delta distributions with the particular choice of locations that minimizes the 1-Wasserstein distance between $\nu \in \mathscr{P}(\mathbb{R})$ and the projection $\Pi_Q \nu \in \mathscr{F}_Q$. The parametrization θ_k with minimal 1-Wasserstein distance is given by the evaluation of the inverse CDF F_ν^{-1} at midpoint quantiles $\tau_k = \frac{2k-1}{2K}$, $k \in [1, ..., K]$, s.t. $\theta_k = F_\nu^{-1}(\frac{2k-1}{2K})$. Equivalently, θ_k is the minimizer of the *quantile regression loss* (QR) (Koenker and Hallock, 2001), which is more amenable to gradient-based optimization. The loss is given by

$$\mathcal{L}_Q(\theta_k, \nu) = \mathbb{E}_{Z \sim \nu}[\rho_{\tau_k}(Z - \theta_k)], \tag{8}$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}_{\{u \leq 0\}}(u))$ is an error function that assigns asymmetric weight to over- or underestimation errors. $\mathbb{1}$ here denotes the indicator function.

4 Exploration with distributional projection ensembles

This paper is foremost concerned with leveraging diverse ensembles that rely on different representations and projection operators. To introduce the concept of distributional projection ensembles and their properties, we begin by outlining the main algorithmic components in a formal setting that foregoes sample-based stochastic approximation or function approximation and defer a more realistic RL setting to Section 5.

Consider an ensemble $E = \{\eta_i(s, a) \,|\, i \in [1, ..., M]\}$ of M member distributions $\eta_i(s, a)$, each associated with a representation $\mathscr{F}_i$ and a projection operator Π_i. In this setting, we assume that each member distribution $\eta_i(s, a) \in \mathscr{F}_i$ is an element of the associated representation $\mathscr{F}_i$ and the projection operator $\Pi_i : \mathscr{P}(\mathbb{R}) \to \mathscr{F}_i$ maps any distribution $\nu \in \mathscr{P}(\mathbb{R})$ to $\mathscr{F}_i$ such that $\Pi_i \nu \in \mathscr{F}_i$. The uniform mixture distribution over E given by $\eta_M(s, a) = 1/M \sum_{i=1}^{M} \eta_i(s, a)$ then has support over the union of representations $\mathscr{F}_M = \cup_{i=1}^{M} \mathscr{F}_i$. We can now define a central identity in this paper, the projection mixture operator $\Omega_M : \mathscr{P}(\mathbb{R}) \to \mathscr{F}_M$, as follows:

$$\Omega_M \eta(s, a) = \tfrac{1}{M} \sum_{i=1}^{M} \Pi_i \eta(s, a). \tag{9}$$

Joining Ω_M with the distributional Bellman operator $\mathcal{T}^\pi$ yields the combined operator $\Omega_M \mathcal{T}^\pi$. Intuitively, $\Omega_M \mathcal{T}^\pi$ performs the following steps: It first applies the distributional Bellman operator $\mathcal{T}^\pi$ to a return distribution η, then projects the resulting distribution with the individual projection operators Π_i onto M different representations $\mathscr{F}_i$, and finally recombines the ensemble members into a mixture model in $\mathscr{F}_M$ (illustrated graphically in Fig. 2). In connection with iterative algorithms, we are often interested in the contractivity of the combined operator $\Omega_M \mathcal{T}^\pi$ to establish convergence.

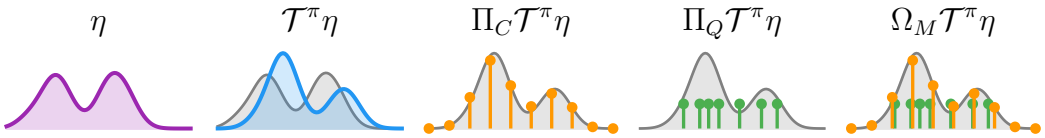

Figure 2: Illustration of the projection mixture operator with quantile and categorical projections.

Proposition 1 delineates conditions under which we can combine individual projections Π_i such that the resulting combined operator $\Omega_M\mathcal{T}^\pi$ is a contraction mapping.

Proposition 1 *Let Π_i, $i \in [1, ..., M]$ be projection operators $\Pi_i : \mathscr{P}(\mathbb{R}) \to \mathscr{F}_i$ mapping from the space of probability distributions $\mathscr{P}(\mathbb{R})$ to representations $\mathscr{F}_i$ and denote the projection mixture operator $\Omega_M : \mathscr{P}(\mathbb{R}) \to \cup_{i=1}^M \mathscr{F}_i$ as defined in Eq. (9). Furthermore, assume that for some $p \in [0, \infty]$ each projection Π_i is bounded in the p-Wasserstein metric in the sense that for any two return distributions η, η' we have $w_p\big(\Pi_i\eta, \Pi_i\eta'\big)(s,a) \leq c_i w_p\big(\eta, \eta'\big)(s,a)$ for a constant c_i. Then, the combined operator $\Omega_M\mathcal{T}^\pi$ is bounded in the supremum p-Wasserstein distance $\bar{w}_p$ by*

$$\bar{w}_p(\Omega_M\mathcal{T}^\pi\eta, \Omega_M\mathcal{T}^\pi\eta') \leq \bar{c}\gamma\bar{w}_p(\eta, \eta')$$

and is accordingly a contraction so long as $\bar{c}\gamma < 1$, where $\bar{c} = \frac{1}{M}\sum_{i=1}^M c_i$.

The proof is deferred to Appendix A. The contraction condition in Proposition 1 is naturally satisfied for example if all projections Π_i are non-expansions in a joint metric w_p. It is, however, more permissive in the sense that it only requires the average modulus $\bar{c}$ to be limited, allowing for expanding operators in the ensemble as well. A contracting combined operator $\Omega_M\mathcal{T}^\pi$ allows us to formulate a simple convergent iteration scheme where in a sequence of steps k, ensemble members are moved toward the projected mixture distribution according to $\hat{\eta}_{i,k+1} = \Pi_i\mathcal{T}^\pi\hat{\eta}_{M,k}$, yielding the $(k+1)$-th mixture distribution $\hat{\eta}_{M,k+1} = \frac{1}{M}\sum_{i=1}^M \hat{\eta}_{i,k+1}$. This procedure can be compactly expressed by

$$\hat{\eta}_{M,k+1} = \Omega_M\mathcal{T}^\pi\hat{\eta}_{M,k}, \quad \text{for} \quad k = [0, 1, 2, 3, ...] \tag{10}$$

and has a unique fixed point which we denote $\eta_M^\pi = \hat{\eta}_{M,\infty}$.

4.1 From distributional approximations to confidence bounds

We proceed to describe how distributional ensembles can be leveraged for exploration. Our setting considers exploration strategies based on the upper-confidence-bound (UCB) algorithm (Auer, 2002). In the context of model-free RL, provably efficient algorithms often rely on the construction of a bound, that overestimates the true state-action value with high probability (Jin et al., 2018, 2019). In other words, we are interested in finding an optimistic value $\hat{Q}^+(s,a)$ such that $\hat{Q}^+(s,a) \geq Q^\pi(s,a)$ with high probability. To this end, Proposition 2 relates an estimate $\hat{Q}(s,a)$ to the true value $Q^\pi(s,a)$ through a distributional error term.

Proposition 2 *Let $\hat{Q}(s,a) = \mathbb{E}[\hat{Z}(s,a)]$ be a state-action value estimate where $\hat{Z}(s,a) \sim \hat{\eta}(s,a)$ is a random variable distributed according to an estimate $\hat{\eta}(s,a)$ of the true state-action return distribution $\eta^\pi(s,a)$. Further, denote $Q^\pi(s,a) = \mathbb{E}[Z^\pi(s,a)]$ the true state-action, where $Z^\pi(s,a) \sim \eta^\pi(s,a)$. We have that $Q^\pi(s,a)$ is upperbounded by*

$$\hat{Q}(s,a) + w_1\big(\hat{\eta}, \eta^\pi\big)(s,a) \geq Q^\pi(s,a) \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where w_1 is the 1-Wasserstein distance metric.

The proof follows from the definition of the Wasserstein distances and is given in Appendix A. Proposition 2 implies that, for a given distributional estimate $\hat{\eta}(s,a)$, we can construct an optimistic upper bound on $Q^\pi(s,a)$ by adding a bonus of the 1-Wasserstein distance between an estimate $\hat{\eta}(s,a)$ and the true return distribution $\eta^\pi(s,a)$, which we define as $b^\pi(s,a) = w_1(\hat{\eta}, \eta^\pi)(s,a)$ in the following. By adopting an optimistic action-selection with the guaranteed upper bound on $Q^\pi(s,a)$ according to

$$a = \arg\max_{a \in \mathcal{A}} [\hat{Q}(s,a) + b^\pi(s,a)], \tag{11}$$

we maintain that the resulting policy inherits efficient exploration properties of known optimism-based exploration methods. Note that in a convergent iteration scheme, we expect the bonus $b^\pi(s,a)$ to almost vanish in the limit of infinite iterations. We thus refer to $b^\pi(s,a)$ as a measure of the epistemic uncertainty of the estimate $\hat{\eta}(s,a)$.

4.2 Propagation of epistemic uncertainty through distributional errors

As is, the previously described bonus does not inform a practicable algorithm, since we do not assume knowledge of the true return distribution $\eta^\pi(s,a)$ and accordingly can not access $b^\pi(s,a)$. Unfortunately, the ensemble E does not straightforwardly lend itself as an estimator of $b^\pi(s,a)$ either. This is because our iteration procedure's use of the mixture operator Ω_M precludes the independence of individual members with model estimates moving toward a joint target at every iteration. This well-known issue (described extensively in the Bayesian setting by Fellows et al. (2021)) causes the ensemble to represent uncertainty w.r.t. the bootstrap $\Omega_M \mathcal{T}^\pi \hat{\eta}_M(s,a)$ rather than the true return distribution $\eta^\pi(s,a)$. While maintaining strict independence between ensemble members can resolve this issue (Osband et al., 2019b; Chen et al., 2017), numerous works demonstrate that joining ensemble predictions into a single bootstrap alleviates overestimation bias and learning instability (Hasselt, 2010; Van Hasselt et al., 2018; Kuznetsov et al., 2020; Chen et al., 2021).

In this paper, we aim for the latter and make the assumption that iterations based on the combined operator $\Omega_M \mathcal{T}^\pi$ and a well-behaved distributional ensemble E yield an estimate of the one-step uncertainty $w_{\text{avg}}(s,a) \approx w_1\big(\hat{\eta}_M, \Omega_M \mathcal{T}^\pi \hat{\eta}_M\big)(s,a)$. Here, $w_{\text{avg}}(s,a)$ is simply the average disagreement between ensemble members in 1-Wasserstein according to

$$w_{\text{avg}}(s,a) = \tfrac{1}{M(M-1)} \sum_{i,j=1}^{M} w_1\big(\hat{\eta}_i, \hat{\eta}_j\big)(s,a)\,. \tag{12}$$

To establish a bonus that allows for optimistic action selection we derive a propagation scheme for epistemic uncertainty in the distributional setting. More specifically, we find that an upper bound on the bonus $b^\pi(s,a)$ satisfies a temporal consistency condition, similar to the Bellman equations, that relates the distributional return error $w_1(\hat{\eta}, \eta^\pi)(s,a)$ to a *one-step* error $w_1(\hat{\eta}, \Omega_M \mathcal{T}^\pi \hat{\eta})(s,a)$ that is more amenable to estimation.

Theorem 3 *Let $\hat{\eta}(s,a) \in \mathscr{P}(\mathbb{R})$ be an estimate of the true return distribution $\eta^\pi(s,a) \in \mathscr{P}(\mathbb{R})$, and denote the projection mixture operator $\Omega_M : \mathscr{P}(\mathbb{R}) \to \cup_{i=1}^M \mathscr{F}_i$ with members Π_i and bounding moduli c_i as defined in Proposition 1. Furthermore, assume $\Omega_M \mathcal{T}^\pi$ is a contraction mapping with fixed point η_M^π. We then have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$*

$$w_1\big(\hat{\eta}, \eta_M^\pi\big)(s,a) \;\leq\; w_1\big(\hat{\eta}, \Omega_M \mathcal{T}^\pi \hat{\eta}\big)(s,a) + \bar{c}\,\gamma\, \mathbb{E}\Big[w_1\big(\hat{\eta}, \eta_M^\pi\big)(S_1, A_1) \big| S_0 = s, A_0 = a\Big],$$

where $S_1 \sim P(\cdot|S_0 = s, A_0 = a)$ and $A_1 \sim \pi(\cdot|S_1)$.

The proof is given in Appendix A and exploits the triangle inequality property of the Wasserstein distances. It may be worth noting that Theorem 3 is a general result that is not restricted to the use of projection ensembles. It is, however, a natural complement to the iteration described in Eq. (10) in that it allows us to reconcile the benefits of bootstrapping diverse ensemble mixtures with optimistic action selection for directed exploration. To this end, we devise a separate iteration procedure aimed at finding an approximate upper bound on $w_1(\hat{\eta}, \eta_M^\pi)(s,a)$. Denoting the k-th iterate of the bonus estimate $\hat{b}_k(s,a)$, we have by Theorem 3 that the iteration

$$\hat{b}_{k+1}(s,a) = w_1\big(\hat{\eta}, \Omega_M \mathcal{T}^\pi \hat{\eta}\big)(s,a) + \bar{c}\gamma \mathbb{E}_{P,\pi}\big[\hat{b}_k(S_1, A_1) \big| S_0 = s, A_0 = a\big]\; \forall (s,a) \in \mathcal{S} \times \mathcal{A}\,,$$

converges to an upper bound on $w_1(\hat{\eta}, \eta_M^\pi)(s,a)$[1].

We conclude this section with the remark that the use of projection ensembles as described here clashes with the intuition that the epistemic uncertainty $w_{\text{avg}}(s,a)$ should vanish in convergence. This is because each estimate $\hat{\eta}_i$ inherits irreducible approximation errors from the projections Π_i. In Appendix A, we provide general bounds for these errors and show that residual errors can be controlled through the number of atoms K in the specific example of an ensemble based on the quantile and categorical projections.

5 Deep distributional reinforcement learning with projection ensembles

Section 4 has introduced the concept of projection ensembles in a formal, sanitized setting. In this section, we aim to transcribe the previously derived algorithmic components into a practicable deep

[1] To see the convergence, note that the sequence is equivalent to an iteration with T^π in an MDP with the deterministic immediate reward $w_1\big(\hat{\eta}, \Omega_M \mathcal{T}^\pi \hat{\eta}\big)(s,a)$.

RL algorithm that departs from several of the previous assumptions. Specifically, this includes 1) control with a greedy policy, 2) sample-based stochastic approximation, 3) nonlinear function approximation, and 4) gradient-based optimization. This sets the following section firmly apart from the results in Section 4, yet it is exactly in this scenario that we hypothesize diverse projection ensembles to bring to bear their benefits. The central underlying idea is that distributional projections and the functional constraints they entail offer an effective tool to impose diverse generalization behaviors on an ensemble of learners. In particular, we implement the above-described algorithm with a neural ensemble comprising the models of the two popular deep RL algorithms QR-DQN (Dabney et al., 2018b) and C51 (Bellemare et al., 2017).

5.1 Do different distribution projections lead to different generalization behavior?

As formal descriptions of the generalization behavior of deep NN architectures remain elusive, we provide empirical support for the claim in question and motivate the use of projection ensembles with a preliminary experiment.

We modify a recent distributional RL algorithm that relies on predictive variance as an exploration signal: under the premise that predictions of a QR model produce high variance predictions when encountering unfamiliar state-action regions, DLTV-QR (Mavrin et al., 2019) employs an exploratory action-selection rule akin to[2]

$$a = \arg\max_{a \in \mathcal{A}} \left(\mathbb{E}[Z] + \beta \mathbb{V}[Z] \right), Z \sim \hat{\eta}_{\text{QR}}(s, a).$$

We replace the quantile representations in the original formulation with the categorical architecture of C51. Fig. 3 shows the results of this study on the deep exploration benchmark *deep sea*, where agents are tasked with finding a rare rewarding state. These results strongly suggest that C51 in this case does not follow the presumption that unseen state-action tuples are associated with high-variance predictions, resulting in a counterproductive exploration signal.

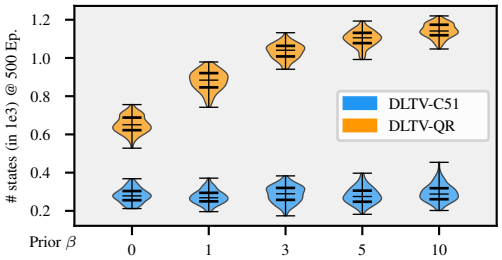

Figure 3: State exploration on deep-sea with DLTV based on categorical and quantile projections. Higher (number of visited states) is better. The horizontal axis represents different factors of randomized prior functions. Bars represent medians and interquartile ranges of 100 seeds.

5.2 Deep quantile and categorical projection ensembles for exploration

In this section, we introduce projection ensemble DQN (PE-DQN), a deep RL algorithm that combines the quantile and categorical projections of QR-DQN (Dabney et al., 2018b) and C51 (Bellemare et al., 2017) into a diverse ensemble to drive exploration and learning stability.

Parametric model. We begin by introducing our parametric model of the mixture distribution $\eta_{M,\theta}$. We construct $\eta_{M,\theta}$ as an equal mixture between a quantile and a categorical representation, each parametrized through a NN with K output logits where we use the notation θ_{ik} to mean the k-th logit of the network parametrized by the parameters θ_i of the i-th model in the ensemble. In the following, we will consider a sample transition (s, a, r, s', a') where a' is chosen greedily according to $\mathbb{E}_{Z \sim \eta_{M,\theta}(s',a')}[Z]$. Dependencies on (s, a) are hereafter dropped for conciseness by writing $\theta_{ik}(s, a) = \theta_{ik}$ and $\theta_{ik}(s', a') = \theta'_{ik}$. The full mixture model $\eta_{M,\theta}$ is then given by

$$\eta_{M,\theta} = \frac{1}{2} \sum_{i=1}^{M=2} \sum_{k=1}^{K} p(\theta_{ik}) \delta_{z(\theta_{ik})}(z), \quad \text{with} \quad \begin{array}{l} p(\theta_{1k}) = \frac{1}{K},\ z(\theta_{1k}) = \theta_{1k}, \\ p(\theta_{2k}) = \sigma(\theta_{2k}),\ z(\theta_{2k}) = z_k, \end{array} \tag{13}$$

where $\sigma(x_i) = e^{x_i} / \sum_j e^{x_j}$ is the softmax transfer function. Consequently, this representation comprises a total of $2K$ atoms, K of which parametrize locations in the quantile model, and the remaining K parametrizing probabilities in the categorical representation.

Projection losses. Next, we assume that bootstrapped return distributions are generated by a set of delayed parameters $\tilde{\theta}$, as is common (Mnih et al., 2015). The stochastic (sampled) version of the

[2]For notational simplicity we have here simplified the original formulation, which uses a decaying left-truncated version of the variance as a bonus. All our experiments make use of the original formulation.

distributional Bellman operator $\hat{\mathcal{T}}^{\pi}$, applied to the target ensemble's mixture distribution $\eta_{M,\tilde{\theta}}$ yields

$$\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}} = \frac{1}{2}\sum_{i=1}^{M=2}\sum_{k=1}^{K} p(\tilde{\theta}'_{ik})\,\delta_{r+\gamma z(\tilde{\theta}'_{ik})}(z). \tag{14}$$

Instead of applying the projection mixture Ω_M analytically, as done previously in Section 4, the parametric estimates $\eta_{M,\theta}$ are moved incrementally towards a projected target distribution through gradient descent on a loss function.

In the *quantile* representation, we augment the classical quantile regression loss (Koenker and Hallock, 2001) with an importance-sampling ratio $Kp(\tilde{\theta}'_{ij})$ to correct for the non-uniformity of atoms from the bootstrapped distribution $\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}}$. For a set of fixed quantiles τ_k, the loss $\mathcal{L}_1$ is given by

$$\mathcal{L}_1\big(\eta_{\theta_1}, \Pi_Q\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}}\big) = \sum_{i=1}^{M=2}\sum_{k,j=1}^{K} Kp(\tilde{\theta}'_{ij})\Big(\rho_{\tau_k}\big(r+\gamma z(\tilde{\theta}'_{ij}) - \theta_{1k}\big)\Big). \tag{15}$$

The *categorical* model minimizes the Kullback-Leibler (KL) divergence between the projected bootstrap distribution $\Pi_C\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}}$ and an estimate η_{θ_2}. The corresponding loss is given by

$$\mathcal{L}_2\big(\eta_{\theta_2}, \Pi_C\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}}\big) = D_{KL}\big(\Pi_C\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}} \| \eta_{\theta_2}\big). \tag{16}$$

As $\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}}$ is a mixture of Dirac distributions, the definition of the projection Π_C according to Eq. (6) can be applied straightforwardly to obtain the projected bootstrap distribution $\Pi_C\hat{\mathcal{T}}^{\pi}\eta_{M,\tilde{\theta}}$.

Uncertainty Propagation. Given a state-action tuple (s, a), the local uncertainty estimate $w_{\text{avg}}(s, a)$ as defined in Eq. (12) can be computed deterministically from a distributional ensemble E. To obtain a parametric bonus estimate $b_\phi(s, a)$ we reproduce the steps above with a model of parameters ϕ and an alternate tuple $(s, a, w_{\text{avg}}, s', a'_\epsilon)$, where we replaced the immediate reward with the ensemble disagreement. a'_ϵ is an exploratory action chosen greedily according to the rule

$$a = \arg\max_{a\in\mathcal{A}} \big(\mathbb{E}_{Z\sim\eta_{M,\theta}(s,a)}[Z] + \beta\, b_\phi(s,a)\big), \quad \text{where} \quad b_\phi(s,a) = \mathbb{E}_{B\sim\eta_{M,\phi}(s,a)}[B]. \tag{17}$$

Here, β is a hyperparameter to control the policy's drive towards exploratory actions. Finally, the action selection in Eq. (17) also constitutes our agents' behavioral policy for exploration.

6 Experimental results

6.1 The behaviour suite

While our algorithm design has an emphasis on hard exploration problems, we intend to assess how the use of projection ensembles affects the learning process of agents in various aspects on a wide range of tasks. To this end, we evaluate PE-DQN on the behavior suite (bsuite) (Osband et al., 2019a), a battery of benchmark problems constructed to assess key properties of RL algorithms. The suite consists of 22 tasks with up to 22 variations in size or seed, totaling 468 environments. Our code is available at [anonymized URL].

Experimental setup. We compare PE-DQN to several baselines: the popular Bootstrapped DQN with prior functions (BDQN+P) (Osband et al., 2019b), the aforementioned DLTV QR-DQN (Mavrin et al., 2019), and information-directed sampling (IDS-C51) (Nikolov et al., 2018). BDQN+P approximates posterior sampling of a parametric value function by combining statistical bootstrapping with additive prior functions in an ensemble of DQN base learners. IDS-C51 builds on the BDQN+P architecture but acts according to an information-gain ratio for which Nikolov et al. (2018) estimate aleatoric uncertainty (noise) with the categorical model of C51. In contrast, DLTV QR-DQN employs a distributional value approximation based on the quantile representation and follows a decaying exploration bonus of the left-truncated variance (for more details, see Mavrin et al. (2019)). We aimed to keep as many hyperparameters between the implementations equal, up to algorithm-specific parameters, network architecture, prior factors, and learning rate. We tuned each algorithm with a grid search on a selected subset of problems in the bsuite. Our implementations make slight adjustments to DLTV and IDS in their favor and generally outperformed the vanilla baselines where available. Further details on the experimental design and implementation are provided in Appendix B.

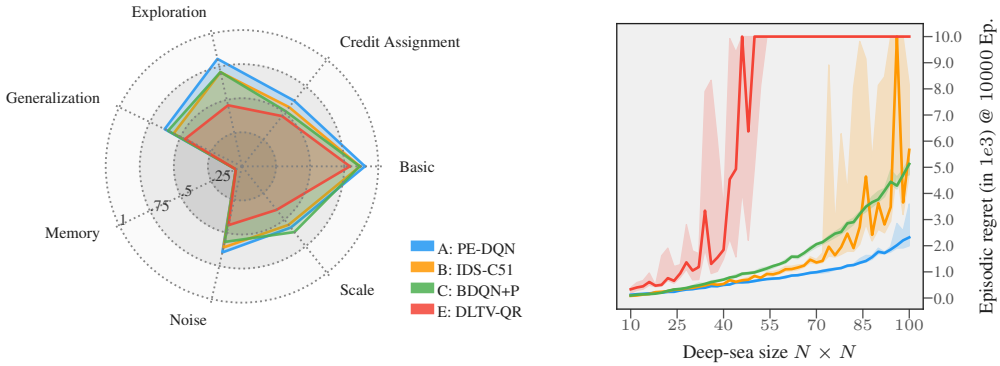

Figure 4: (a) Summary of bsuite experiments. Wider is better. (b) Median episodic regret over deep sea size. Lower is better. Shaded regions are the interquartile range of 10 seeds.

Comparative evaluation. Fig. 4 (a) shows the results of the entire suite experiment, summarized in seven *core capabilities*. These capability scores are computed as proposed by Osband et al. (2019a) and follow a handcrafted scoring function per environment. For example, exploration capability is scored by the average regret in the sparse reward environments *deep sea*, *stochastic deep sea*, and *cartpole swingup*. The full set of results is provided in Appendix B. Perhaps unsurprisingly, PE-DQN has its strongest performance in the exploration category but we find that it improves upon baselines in several more categories. Note here, that PE-DQN uses substantially fewer models than the baselines, with a total of 4 distributional models compared to the 20 DQN models used in the ensembles of both BDQN+P and IDS, where the latter requires an additional C51 model.

6.2 The deep-sea environment

The deep sea environment is one of the exploration problems in the behavior suite and has recently gained popularity as an exploration benchmark (Osband et al., 2019b; Janz et al., 2019; Flennerhag et al., 2020). Deep sea is a sparse reward environment where agents can reach the single rewarding state at the bottom right of an $N \times N$ grid only through a unique sequence of actions in an exponentially growing trajectory space. We ran an additional experiment on deep sea with grid sizes up to 100; double the maximal size in the behavior suite. Fig. 4 (b) shows a summary of this experiment where we evaluated episodic regret, that is the number of non-rewarding episodes with a maximum budget of 10000 episodes. PE-DQN scales more gracefully to larger sizes of the problem than the baselines, reducing the median regret by roughly half. Additional experimental results provided in Appendix C show that the use of diverse projection ensembles is crucial in attaining this performance, outperforming variations of PE-DQN that rely on ensembles of equal architecture by a large margin.

7 Conclusion

In this work, we have introduced projection ensembles for distributional reinforcement learning, a method combining diversely generalizing models based on different parametric representations and projections of return distributions. We have provided a theoretical analysis that establishes convergence conditions and bounds on residual approximation errors that apply to general compositions of such projection ensembles. Furthermore, we have introduced a general propagation method that reconciles one-step distributional errors with optimism-based exploration. Our empirical analysis of PE-DQN, a deep RL algorithm based on projection ensembles, demonstrates the efficacy of diversely generalizing projection ensembles on hard exploration tasks and showed significant performance improvements on a wide range of tasks.

We believe our work opens up a number of promising avenues for future research. For example, we have only considered the use of uniform mixtures over distributional ensembles in this work. A logical continuation of this approach may aim to use a diverse collection of models less conservatively, aiming to exploit the strengths of particular models flexibly where beneficial. Naturally, a more carefully curated composition of ensembles and a more exploitative usage of their characteristics will require a deeper understanding of how the choice of projection and parametrization affects the generalization behavior of neural networks.

References

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

P. Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002. ISSN ISSN 1533-7928.

M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

M. G. Bellemare, W. Dabney, and M. Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. `http://www.distributional-rl.org`.

R. Bellman. A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957. ISSN 0095-9057.

R. Y. Chen, S. Sidor, P. Abbeel, and J. Schulman. UCB Exploration via Q-Ensembles, Nov. 2017.

X. Chen, C. Wang, Z. Zhou, and K. Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018a.

W. Dabney, M. Rowland, M. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2): 105–112, 2009.

T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer, 2000.

H. Eriksson, D. Basu, M. Alibeigi, and C. Dimitrakakis. Sentinel: taming uncertainty with ensemble based distributional reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 631–640. PMLR, 2022.

M. Fellows, K. Hartikainen, and S. Whiteson. Bayesian bellman operators. *Advances in Neural Information Processing Systems*, 34:13641–13656, 2021.

S. Flennerhag, J. X. Wang, P. Sprechmann, F. Visin, A. Galashov, S. Kapturowski, D. L. Borsa, N. Heess, A. Barreto, and R. Pascanu. Temporal difference uncertainties as a signal for exploration. *arXiv preprint arXiv:2010.02255*, 2020.

H. Hasselt. Double q-learning. *Advances in neural information processing systems*, 23, 2010.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

C.-J. Hoel, K. Wolff, and L. Laine. Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

S. C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996.

D. Janz, J. Hron, P. Mazur, K. Hofmann, J. M. Hernández-Lobato, and S. Tschiatschek. Successor uncertainties: exploration and uncertainty in temporal difference learning. *Advances in Neural Information Processing Systems*, 32, 2019.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably Efficient Reinforcement Learning with Linear Function Approximation, Aug. 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

A. Kuznetsov, P. Shvechikov, A. Grishin, and D. Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pages 5556–5566. PMLR, 2020.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

E. Mariucci and M. Reiß. Wasserstein and total variation distance between marginals of Lévy processes. *Electronic Journal of Statistics*, 12(2):2482 – 2514, 2018.

B. Mavrin, H. Yao, L. Kong, K. Wu, and Y. Yu. Distributional Reinforcement Learning for Efficient Exploration. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4424–4434. PMLR, May 2019.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

T. M. Moerland, J. Broekens, and C. M. Jonker. Efficient exploration with Double Uncertain Value Networks. *arXiv:1711.10789 [cs, stat]*, Nov. 2017.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806, 2010.

T. Nguyen-Tang, S. Gupta, and S. Venkatesh. Distributional Reinforcement Learning via Moment Matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9144–9152, May 2021. ISSN 2374-3468.

N. Nikolov, J. Kirschner, F. Berkenkamp, and A. Krause. Information-directed exploration for deep reinforcement learning. *arXiv preprint arXiv:1812.07544*, 2018.

I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep Exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

I. Osband, Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepesvari, S. Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019a.

I. Osband, B. Van Roy, D. J. Russo, Z. Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019b.

B. O'Donoghue, I. Osband, R. Munos, and V. Mnih. The uncertainty bellman equation and exploration. In *International Conference on Machine Learning*, pages 3836–3845, 2018.

U. Rösler. A fixed point theorem for distributions. *Stochastic Processes and their Applications*, 42 (2):195–214, 1992.

M. Rowland, M. Bellemare, W. Dabney, R. Munos, and Y. W. Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

M. Rowland, R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, and W. Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

H. Van Hasselt, Y. Doron, F. Strub, M. Hessel, N. Sonnerat, and J. Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

D. J. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications*, 56(1):1–29, Jan. 1988. ISSN 1573-2878. doi: 10.1007/BF00938524.

D. Yang, L. Zhao, Z. Lin, T. Qin, J. Bian, and T.-Y. Liu. Fully Parameterized Quantile Function for Distributional Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A Appendix

This section provides proofs for the theoretical claims and establishes further results on the residual approximation error incurred by our method.

A.1 Proof of Proposition 1

Before stating supporting lemmas and proofs of the results in section 4, we recall several basic properties of the p-Wasserstein distances which we will find useful in the subsequent proofs. Derivations of these properties can for example be found in an overview by Mariucci and Reiß (2018) .

P.1 The p-Wasserstein distances satisfy the *triangle inequality*, that is

$$w_p(X,Y) \leq w_p(X,Z) + w_p(Z,Y)\,.$$

P.2 For random variables X and Y and an auxiliary variable Z independent of X and Y, the p-Wasserstein metric satisfies the inequality

$$w_p(X+Z, Y+Z) \leq w_p(X,Y)\,.$$

P.3 For a real-valued scalar $a \in \mathbb{R}$, we have

$$w_p(aX, aY) = |a| w_p(X,Y)\,.$$

Lemma 4 *Let $X_1, X_2, X_3, ..., X_n$ and $Y_1, Y_2, Y_3, ..., Y_n$ be a collection of independent random variables. Furthermore, let $a_1, a_2, a_3, ..., a_n$ be scalars $a_i \in \mathbb{R}$. Then, the p-Wasserstein metric satisfies the inequality*

$$w_p(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} a_i Y_i) \leq \sum_{i=1}^{n} |a_i| w_p(X_i, Y_i)\,.$$

Proof. The case for $n = 1$ is trivially satisfied by Property P.3. The proof then follows by induction for the case $n = 2$. Consider an auxiliary random variable $\hat{X}_2$ that is equal in distribution law to X_2. We have by the triangle inequality that

$$\begin{aligned} w_p(a_1X_1 + a_2X_2, a_1Y_1 + a_2Y_2) \leq\; & w_p(a_1X_1 + a_2X_2, a_1Y_1 + a_2\hat{X}_2) \\ & + w_p(a_1Y_1 + a_2\hat{X}_2, a_1Y_1 + a_2Y_2)\,. \end{aligned}$$

By Property P.2, it follows that

$$\begin{aligned} w_p(a_1Y_1 + a_2\hat{X}_2, a_1Y_1 + a_2Y_2) &\leq w_p(a_2\hat{X}_2, a_2Y_2) = |a_2| w_p(X_2, Y_2), \\ w_p(a_1X_1 + a_2X_2, a_1Y_1 + a_2\hat{X}_2) &\leq w_p(a_1X_1, a_1Y_1) = |a_1| w_p(X_1, Y_1), \end{aligned}$$

and hence $w_p(\sum_{i=1}^{n=2} a_iX_i, \sum_{i=1}^{n=2} a_iY_i) \leq \sum_{i=1}^{n=2} |a_i| w_p(X_i, Y_i)$.

Proposition 1 *Let Π_i, $i \in [1, ..., M]$ be projection operators $\Pi_i : \mathscr{P}(\mathbb{R}) \to \mathscr{F}_i$ mapping from the space of probability distributions $\mathscr{P}(\mathbb{R})$ to representations $\mathscr{F}_i$ and denote the projection mixture operator $\Omega_M : \mathscr{P}(\mathbb{R}) \to \cup_{i=1}^{M} \mathscr{F}_i$ as defined in Eq. (9). Furthermore, assume that for some $p \in [0, \infty]$ each projection Π_i is bounded in the p-Wasserstein metric in the sense that for any two return distributions η, η' we have $w_p\big(\Pi_i\eta, \Pi_i\eta'\big)(s,a) \leq c_i w_p\big(\eta, \eta'\big)(s,a)$ for a constant c_i. Then, the combined operator $\Omega_M \mathcal{T}^\pi$ is bounded in the supremum p-Wasserstein distance $\bar{w}_p$ by*

$$\bar{w}_p(\Omega_M \mathcal{T}^\pi \eta, \Omega_M \mathcal{T}^\pi \eta') \leq \bar{c}\gamma \bar{w}_p(\eta, \eta')$$

and is accordingly a contraction so long as $\bar{c}\gamma < 1$, where $\bar{c} = \frac{1}{M}\sum_{i=1}^{M} c_i$.

Proof. Due to the assumption of the proposition, we have $w_p(\Pi_i\nu, \Pi_i\nu') \leq c_i w_p(\nu, \nu')$. With Lemma 4 and the γ-contractivity of $\mathcal{T}^\pi$, the statement follows directly according to

$$\begin{aligned} \bar{w}_p(\Omega_M \mathcal{T}^\pi \eta, \Omega_M \mathcal{T}^\pi \eta') &= \bar{w}_p(\sum_{i=1}^{M} \tfrac{1}{M} \Pi_i \mathcal{T}^\pi \eta, \sum_{i=1}^{M} \tfrac{1}{M} \Pi_i \mathcal{T}^\pi \eta') \\ &\leq \tfrac{1}{M} \sum_{i=1}^{M} \bar{w}_p(\Pi_i \mathcal{T}^\pi \eta, \Pi_i \mathcal{T}^\pi \eta') \\ &\leq \tfrac{1}{M} \sum_{i=1}^{M} c_i \bar{w}_p(\mathcal{T}^\pi \eta, \mathcal{T}^\pi \eta') \\ &\leq \tfrac{1}{M} \sum_{i=1}^{M} c_i \gamma \bar{w}_p(\eta, \eta') \\ &= \bar{c}\gamma \bar{w}_p(\eta, \eta')\,. \end{aligned}$$

A.2 Proof of Proposition 2

Proposition 2 *Let $\hat{Q}(s,a) = \mathbb{E}[\hat{Z}(s,a)]$ be a state-action value estimate where $\hat{Z}(s,a) \sim \hat{\eta}(s,a)$ is a random variable distributed according to an estimate $\hat{\eta}(s,a)$ of the true state-action return distribution $\eta^\pi(s,a)$. Further, denote $Q^\pi(s,a) = \mathbb{E}[Z^\pi(s,a)]$ the true state-action, where $Z^\pi(s,a) \sim \eta^\pi(s,a)$. We have that $Q^\pi(s,a)$ is upperbounded by*

$$\hat{Q}(s,a) + w_1\big(\hat{\eta}, \eta^\pi\big)(s,a) \geq Q^\pi(s,a) \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where w_1 is the 1-Wasserstein distance metric.

Proof. We begin by stating a property that relates the expected value $\mathbb{E}[X]$ to the CDF of X under the condition that the expectation $\mathbb{E}[X]$ is well-defined and finite. Let $X \sim \nu$ and write F_ν for the CDF of ν, then:

$$\mathbb{E}[X] = \int_0^\infty \big(1 - F_\nu(x)\big)dx - \int_{-\infty}^0 F_\nu(x)dx\,.$$

Now, suppose an auxiliary variable X' is distributed according to the law ν'. It then follows that

$$\begin{aligned}
\big|\mathbb{E}[X] - \mathbb{E}[X']\big| &= \Big| \int_0^\infty \big(F_{\nu'}(x) - F_\nu(x)\big)dx - \int_{-\infty}^0 \big(F_\nu - F_{\nu'}(x)\big)dx \Big| \\
&= \Big| \int_{-\infty}^\infty F_{\nu'}(x) - F_\nu(x)dx \Big| \\
&\leq \int_{-\infty}^\infty \big|F_{\nu'}(x) - F_\nu(x)\big|dx \\
&= w_1(\nu, \nu'),
\end{aligned}$$

where the last step was obtained by a change of variables in the definition of the 1-Wasserstein distance:

$$\begin{aligned}
w_1(\nu, \nu') &= \int_0^1 |F_\nu^{-1}(\tau) - F_{\nu'}^{-1}(\tau)|d\tau \\
&= \int_{\mathbb{R}} |F_\nu(x) - F_{\nu'}(x)|dx.
\end{aligned}$$

The result of Proposition 2 is obtained by rearranging.

A.3 Proof of Theorem 3

Before stating the proof of Theorem 3, we formalize the notion of a pushforward distribution which will be useful in a more explicit description of the distributional Bellman operator $\mathcal{T}^\pi$. Our notation here follows the detailed exposition by Bellemare et al. (2023) .

Definition 5 *For a function $f : \mathbb{R} \to \mathbb{R}$ and a random variable Z with distribution $\nu = \mathcal{D}(Z)$, $\nu \in \mathscr{P}(\mathbb{R})$, the pushforward distribution $f_\# \nu \in \mathscr{P}(\mathbb{R})$ of ν through f is defined as*

$$f_\# \nu(B) = \nu(f^{-1}(B)), \quad \forall B \in \mathbb{R}\,,$$

with B a Borel set in $\mathbb{R}$.

Equivalently to Definition 5, we may write $f_\# \nu = \mathcal{D}(f(Z))$. By defining a bootstrap transformation $b_{r,\gamma} : \mathbb{R} \to \mathbb{R}$ with $b_{r,\gamma} = r + \gamma x$, we can state a more explicit definition of the distributional Bellman operator $\mathcal{T}^\pi$ according to Definition 6.

Definition 6 *[Distributional Bellman Operator (Bellemare et al., 2017)] The distributional Bellman operator $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ is given by*

$$\big(\mathcal{T}^\pi \eta\big)(s,a) = \mathbb{E}\big[\big(b_{R_0,\gamma}\big)_\# \eta(S_1, A_1)\big| S_0 = s, A_0 = a\big]\,,$$

where $S_1 \sim P(\cdot|S_0 = s, A_0 = a), \quad A_1 \sim \pi(\cdot|S_1)$.

Lemma 7 *Let $(b_{r,\gamma})_\# \nu \in \mathscr{P}(\mathbb{R})$ be the pushforward distribution of $\nu \in \mathscr{P}(\mathbb{R})$ through $b_{r,\gamma} : \mathbb{R} \to \mathbb{R}$. Then we have for two distributions ν, ν' and the 1-Wasserstein distance w_1 that*

$$w_1\big((b_{r,\gamma})_\# \nu, (b_{r,\gamma})_\# \nu'\big) = \gamma w_1(\nu, \nu').$$

Proof. The proof follows from the definition of the 1-Wasserstein distance. Let $Z \sim \nu$ and $Z' \sim \nu'$ be two independent random variables, then

$$\begin{aligned} w_1\big((b_{r,\gamma})_\# \nu, (b_{r,\gamma})_\# \nu'\big) &= w_1\big(\mathcal{D}(r+\gamma Z), \mathcal{D}(r+\gamma Z')\big) \\ &= \int_0^1 \big|F^{-1}_{(b_{0,\gamma})_\#\nu}(\tau) - F^{-1}_{(b_{0,\gamma})_\#\nu'}(\tau)\big| d\tau \\ &= |\gamma| w_1(\nu, \nu')\,. \end{aligned}$$

Theorem 3 *Let $\hat{\eta}(s,a) \in \mathscr{P}(\mathbb{R})$ be an estimate of the true return distribution $\eta^\pi(s,a) \in \mathscr{P}(\mathbb{R})$, and denote the projection mixture operator $\Omega_M : \mathscr{P}(\mathbb{R}) \to \cup_{i=1}^M \mathscr{F}_i$ with members Π_i and bounding moduli c_i as defined in Proposition 1. Furthermore, assume $\Omega_M \mathcal{T}^\pi$ is a contraction mapping with fixed point η^π_M. We then have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$*

$$w_1\big(\hat{\eta}, \eta^\pi_M\big)(s,a) \;\le\; w_1\big(\hat{\eta}, \Omega_M \mathcal{T}^\pi \hat{\eta}\big)(s,a) + \bar{c}\,\gamma\, \mathbb{E}\Big[w_1\big(\hat{\eta}, \eta^\pi_M\big)(S_1, A_1)\big| S_0 = s, A_0 = a\Big],$$

where $S_1 \sim P(\cdot|S_0 = s, A_0 = a)$ and $A_1 \sim \pi(\cdot|S_1)$.

Proof. Since $\eta^\pi_M(s,a)$ is the fixed point of the combined operator $\Omega_M \mathcal{T}^\pi$, we have that $\Omega_M \mathcal{T}^\pi \eta^\pi_M(s,a) = \eta^\pi_M(s,a)$. From the triangle inequality it follows that

$$w_1\big(\hat{\eta}, \eta^\pi_M\big)(s,a) \le w_1\big(\hat{\eta}, \Omega_M \mathcal{T}^\pi \hat{\eta}\big)(s,a) + w_1\big(\Omega_M \mathcal{T}^\pi \hat{\eta}, \Omega_M \mathcal{T}^\pi \eta^\pi_M\big)(s,a). \tag{18}$$

Furthermore, for the second term on the r.h.s. in Eq. (18) the following holds:

$$\begin{aligned} w_1\big(\Omega_M \mathcal{T}^\pi \hat{\eta}, \Omega_M \mathcal{T}^\pi \eta^\pi_M\big)(s,a) &= w_1\big(\tfrac{1}{M}\sum_{i=1}^M \Pi_i \mathcal{T}^\pi \hat{\eta}, \tfrac{1}{M}\sum_{i=1}^M \Pi_i \mathcal{T}^\pi \eta^\pi_M\big)(s,a) \\ &\le \tfrac{1}{M}\sum_{i=1}^M c_i w_1\big(\mathcal{T}^\pi \hat{\eta}, \mathcal{T}^\pi \eta^\pi_M\big)(s,a) \\ &= \bar{c} w_1\big(\mathcal{T}^\pi \hat{\eta}, \mathcal{T}^\pi \eta^\pi_M\big)(s,a). \end{aligned}$$

Under slight abuse of the assumptions in Section 3, we here consider an immediate reward distribution with finite support on $\mathcal{R}$ to simplify the following derivation. In this case, we can write out the expectation in Definition 6 as

$$\big(\mathcal{T}^\pi \hat{\eta}\big)(s,a) = \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{S}} Pr(R_0 = r, A_1 = a', S_1 = s' | S_0 = s, A_0 = a)\big((b_{r,\gamma})_\# \hat{\eta}(s', a')\big),$$

where $Pr(\cdot)$ is the joint probability distribution given by the transition kernel $P(\cdot|s,a)$, the immediate reward distribution $\mathcal{R}(\cdot|s,a)$, and the policy $\pi(\cdot|S')$. Thus, by Lemma 4 and Lemma 7 it follows that

$$\begin{aligned} &\bar{c} w_1\big(\mathcal{T}^\pi \hat{\eta}, \mathcal{T}^\pi \eta^\pi_M\big)(s,a) \\ &\quad\le \bar{c}\mathbb{E}\big[w_1\big((b_{R_0,\gamma})_\# \hat{\eta}(S_1, A_1), (b_{R_0,\gamma})_\# \eta^\pi_M(S_1, A_1)\big)\big| S_0 = s, A_0 = a\big] \\ &\quad= \bar{c}\gamma \mathbb{E}\big[w_1\big(\hat{\eta}, \eta^\pi_M\big)(S_1, A_1)\big| S_0 = s, A_0 = a\big]\,, \end{aligned}$$

where $S_1 \sim P(\cdot|S_0 = s, A_0 = a)$ and $A_1 \sim \pi(\cdot|S')$. The proof is completed by plugging in the intermediate results.

A.4 Residual epistemic uncertainty

Due to a limitation to finite-dimensional representations and the use of varying projections, our algorithm incurs residual approximation errors which may not vanish even in convergence. In the context of epistemic uncertainty quantification, this is unfortunate as it can frustrate exploration or lead to overconfident predictions. Specifically, the undesired properties are twofold: 1) Even in convergence, the fixed point η^π_M does not equal the true return distribution (bias). 2) Even in the fixed point ν^π_M, the ensemble disagreement w_{avg} does not vanish. Often, however, we may be able to upper bound and control the error incurred due to the projections Π_i. In this case, Propositions 8 and 9 provide upper bounds on both types of errors as a function of bounded projection errors.

Proposition 8 *Let Ω_M be a projection mixture operator with individual projections Π_i defined as in Eq. (9) . Further, assume each projection Π_i is upper bounded by $w_p(\Pi_i \nu, \nu) \le d_i$ for some $p \in [0, \infty]$. Then, the p-Wasserstein distance between the fixed point $\eta^\pi_M(s,a) = \Omega_M \mathcal{T}^\pi \eta^\pi_M(s,a)$ and the true return distribution $\eta^\pi(s,a) = \mathcal{T}^\pi \eta^\pi(s,a)$ satisfies*

$$w_p(\eta^\pi_M, \eta^\pi)(s,a) \;\le\; \tfrac{\bar{d}}{1-\bar{c}\gamma} \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad \textit{where} \qquad \bar{d} \;=\; \tfrac{1}{M}\sum_{i=1}^M d_i\,.$$

Proof. To show the desired property, we will make use of Proposition 1 and Lemma 4. We omitted the dependency on (s, a) in this section for brevity. It follows then from the triangle inequality that

$$\begin{aligned} w_p(\eta_M^\pi, \eta^\pi) &\leq w_p(\Omega_M \mathcal{T}^\pi \eta_M^\pi, \Omega_M \eta^\pi) + w_p(\Omega_M \eta^\pi, \eta^\pi) \\ &= w_p(\Omega_M \mathcal{T}^\pi \eta_M^\pi, \Omega_M \mathcal{T}^\pi \eta^\pi) + w_p(\Omega_M \eta^\pi, \eta^\pi) \\ &\leq \bar{c}\gamma w_p(\eta_M^\pi, \eta^\pi) + w_p(\tfrac{1}{M} \sum_{i=1}^{M} \Pi_i \eta^\pi, \eta^\pi) \\ &\leq \bar{c}\gamma w_p(\eta_M^\pi, \eta^\pi) + \tfrac{1}{M} \sum_{i=1}^{M} w_p(\Pi_i \eta^\pi, \eta^\pi)\,. \end{aligned}$$

Per the assumption of Proposition 8 and by rearranging we obtain the desired result.

Proposition 9 *Let w_{avg} be the average ensemble disagreement defined according to Eq.* (12) *and assume individual projections Π_i are bounded by $w_p(\Pi_i \nu, \nu) \leq d_i$. For an ensemble E whose mixture distribution equals exactly the fixed point $\eta_M^\pi(s, a) = \Omega_M \mathcal{T}^\pi \eta_M^\pi(s, a)$, the average ensemble disagreement w_{avg} satisfies the inequality*

$$w_{avg}(s, a) \;\leq\; \tfrac{2M}{M-1}\bar{d} \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \qquad \textit{where} \qquad \bar{d} \;=\; \tfrac{1}{M} \sum_{i=1}^{M} d_i\,.$$

Proof. In the fixed point $\eta_M^\pi(s, a) = \Omega_M \mathcal{T}^\pi \eta_M^\pi(s, a)$, the distributional error estimated by $w_{\text{avg}}(s, a)$ does not vanish, unlike the ground truth error $w_1\big(\eta_M^\pi, \Omega_M \mathcal{T}^\pi \eta_M^\pi\big)(s, a) = 0$. The shown property upper bounds this mismatch and is a direct consequence of the assumption $w_p(\Pi_i \nu, \nu) \leq d_i$ which postulates an upper bound on the error introduced by the projection Π_i in terms of the p-Wasserstein distance. The average disagreement is given by

$$w_{\text{avg}}(s, a) = \tfrac{1}{M(M-1)} \sum_{i,j=1}^{M} w_p(\hat{\eta}_i, \hat{\eta}_j)(s, a)\,.$$

The proof is given by applying the triangle inequality and the assumption of the proposition with

$$\begin{aligned} w_p(\hat{\eta}_i, \hat{\eta}_j) &= w_p(\Pi_i \eta_M^\pi, \Pi_j \eta_M^\pi) \\ &\leq w_p(\Pi_i \eta_M^\pi, \eta_M^\pi) + w_p(\eta_M^\pi, \Pi_j \eta_M^\pi) \\ &\leq d_i + d_j\,. \end{aligned}$$

Plugging in and rearranging yields the desired result.

Lemma 10 *[Projection error of the categorical projection (Rowland et al., 2018)] For any distribution $\nu \in \mathscr{P}([z_{\min}, z_{\max}])$ with support on the interval $[z_{\min}, z_{\max}]$ and a categorical projection as defined in Eq.* (6) *with K atoms $z_k \in [z_1, ..., z_K]$ s.t. $z_1 \geq Z_{\min}$ and $z_K \leq z_{\max}$, the error incurred by the projection Π_C is upper bounded in the* 1*-Wasserstein distance by the identity*

$$w_1(\Pi_C \nu, \nu) \leq \big[\sup_{1 \leq k \leq K} (z_{k+1} - z_k)\big]\,.$$

Proof (restated). The proof uses the duality between the 1-Wasserstein distance and the 1-Cramér distance stating

$$l_1(\nu, \nu') = \int_{\mathbb{R}} |F_\nu(x) - F_{\nu'}(x)| dx = \int_0^1 |F_\nu^{-1}(\tau) - F_{\nu'}^{-1}(\tau)| d\tau = w_1(\nu, \nu')\,,$$

and can be obtained by a change of variables. The l_1 formulation simplifies the analysis of the categorical projection, yielding

$$\begin{aligned} w_1(\Pi_C \nu, \nu) &= \int_{\mathbb{R}} |F_{\Pi_C \nu}(x) - F_\nu(x)| dx \\ &\leq \sum_{k=1}^{K-1} (z_{k+1} - z_k) |F_{\Pi_C \nu}(z_k) - F_\nu(z_k)| \\ &\leq \sum_{k=1}^{K-1} (z_{k+1} - z_k) |F_\nu(z_{k+1}) - F_\nu(z_k)| \\ &\leq \big[\sup_{1 \leq k \leq K} (z_{k+1} - z_k)\big] \sum_{k=1}^{K-1} |F_\nu(z_{k+1}) - F_\nu(z_k)| \\ &\leq \big[\sup_{1 \leq k \leq K} (z_{k+1} - z_k)\big]\,. \end{aligned}$$

Table 1: Hyperparameter search space

Hyperparameter	Values
Neural net architecture	$[[64, 64], [128, 128], [512]]$
Learning rate	$[5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}]$
Prior function scale	$[0.0, 5.0, 20.0]$
Heads K	[51, 101]
Bonus β	[0.5, 5.0, 50.0]

Lemma 11 *[Projection error of the quantile projection (Dabney et al., 2018b)] For any distribution $\nu \in \mathscr{P}([z_{\min}, z_{\max}])$ with support on the interval $[z_{\min}, z_{\max}]$ and a quantile projection defined according to Eq.* (8) *with K equally weighted locations $\theta_k \in [\theta_1, ..., \theta_K]$, the error incurred by the projection Π_Q is bounded in the* 1*-Wasserstein distance by the identity*

$$w_1(\Pi_Q \nu, \nu) \leq \frac{z_{\max} - z_{\min}}{K} .$$

Proof (restated). The projection Π_Q is given by

$$\Pi_Q \nu = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \delta_{F_\nu^{-1}(\tau_k)}(z) , \qquad \text{where } \tau_k = \tfrac{2k-1}{2K} .$$

The desired identity $w_1(\Pi_Q \nu, \nu)$ is accordingly given by the continuous integral

$$w_1(\Pi_Q \nu, \nu) = \int_0^1 |F^{-1}_{\Pi_Q \nu}(\tau) - F^{-1}_\nu(\tau)| d\tau ,$$

and can be rewritten in terms of a sum of piecewise expectations

$$w_1(\Pi_Q \nu, \nu) = \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{X \sim \nu} \big[|X - F_\nu^{-1}(\tfrac{2k-1}{2K})| \big| F_\nu^{-1}(\tfrac{k-1}{K}) < X \leq F_\nu^{-1}(\tfrac{k}{K}) \big] .$$

From this, it follows that

$$\begin{aligned} w_1(\Pi_Q \nu, \nu) &\leq \tfrac{1}{K} (F_\nu^{-1}(1) - F_\nu^{-1}(0)) \\ &\leq \frac{z_{\max} - z_{\min}}{K} . \end{aligned}$$

Corollary 12 *Let $\eta_M^\pi(s, a)$ be the fixed point return distribution for an ensemble of the categorical and quantile projections with the mixture operator $\Omega_M \eta(s, a) = 1/2\, \Pi_Q \eta(s, a) + 1/2\, \Pi_C \eta(s, a)$. Furthermore, suppose the return distribution $\eta_M^\pi(s, a)$ has bounded support on the interval $(R_{\max} - R_{\min})/(1 - \gamma)$ where $R_{\max}$ and $R_{\min}$ denote the maximum and minimum immediate reward of the MDP. The average ensemble disagreement $w_{avg}(s, a)$ is then bounded by*

$$w_{avg}(s, a) \leq \frac{4(R_{\max} - R_{\min})}{(1 - \gamma)K} .$$

Proof. The result follows straightforwardly from Proposition 9 and Lemmas 10, 11.

B Experimental Details

We provide a detailed exposition of our experimental setup, including the hyperparameter search procedure, hyperparameter settings, algorithmic details, and the full bsuite experimental results.

B.1 Hyperparameter settings

In our experiments, we aimed to keep most hyperparameters between different implementations equal to maintain comparability between the analyzed methods. Algorithm-specific hyperparameters were optimized over a search space of hyperparameters using Optuna (Akiba et al., 2019). The total search space is given in Table 1, where the *Heads K* parameter only applies to distributional algorithms, and in the case of IDS, hyperparameters for the DQN ensemble and the distributional model were searched jointly. C51 requires us to define return ranges, which we defined manually and can be found in the online code repository. All algorithms use the Adam optimizer (Kingma and Ba, 2015).

Table 2: Hyperparameter search environments

Environment ID	Horizon in no. of episodes	Scoring function f
deep_sea/20	500	$\sum_{(s,a)} \mathbb{1}_{\text{visited}}(s,a)$
deep_sea_stochastic/20	1500	$\sum_{(s,a)} \mathbb{1}_{\text{visited}}(s,a)$
mountain_car/19	100	$\sum_0^t(-1)$

Table 3: Hyperparameter settings

Hyperparameter	BDQN+P	DLTV	IDS (DQN/C51)	PE-DQN (QR/C51)
Net architecture	$[64, 64]$	$[512]$	$[64, 64]$ / $[512]$	$[512]$
Adam Learning rate	1×10^{-3}	1×1^{-3}	1×10^{-3} / 5×10^{-4}	5×10^{-4}
Prior function scale	5.0	20.0	20.0 / 5.0	20.0 / 0.0
Heads K	1	101	1 / 101	$101/101$
Ensemble size	20	1	$20/1$	$2/2$
Initial bonus β_{init}	n/a	5.0	5.0	5.0
Discount			0.99	
Buffer size			$10,000$	
Adam epsilon			$0.001/\text{batch size}$	
Initialization			He truncated normal (He et al., 2015)	
Update frequency			1	
Target update step size			1.0	
Target update frequency			4	
Batch size			128	

The hyperparameter search was conducted on a subselection of environments of the bsuite, as shown in Table 2. For each environment, we evaluate a set of hyperparameters by means of a scoring function. A particular set of hyperparameters is evaluated every $T/5$ episodes with a maximum training horizon of T episodes. The "continuous" scoring functions make the hyperparameter search more amenable to pruning, for which we use the median pruner of Optuna, reducing the computational burden of the combinatorial search space significantly.

$\sum_{(s,a)} \mathbb{1}_{\text{visited}}(s,a)$ here is the count of visited state-action tuples and $\sum_0^t(-1)$ is simply the negative number of total environment interactions. For every hyperparameter configuration ζ_i, the scores $f(\zeta_i)$ are calibrated to facilitate a meaningful comparison between different environments. The calibrated score function we use is given by

$$f_c(\zeta_i) = \exp\left(0.693 \frac{f(\zeta_i) - \mu_\zeta}{\sup_i f(\zeta_i) - \mu_\zeta}\right), \tag{19}$$

where μ_ζ is the average score of all hyperparameter configurations $\mu_\zeta = \sum_i^N 1/N f(\zeta_i)$, and $\sup_i f(\zeta_i)$ is the maximal score achieved. The calibration function in Eq. (19) was chosen heuristically to have an intuitive interpretation: it assigns a score of 1 to the best-performing hyperparameter configuration, 0.5 to configurations that achieve exact average performance, and decays exponentially according to score. The final score assigned to a hyperparameter configuration ζ_i is the sum of all scores of the tested environments. Table 3 shows the full set of hyperparameters used for every algorithm.

B.2 Implementation details

Randomized prior functions are added to all baselines and PE-DQN. Specifically, we add the output of a fixed, randomly initialized neural network of the same architecture as the main net, scaled by a hyperparameter, to the main network's logits. In the case of C51, the prior function is added pre softmax. To the best of our knowledge, DLTV-QR does not use prior functions in its original formulation but we find it to be crucial in improving exploration performance. Fig. 5 (b) shows an experiment assessing the exploration performance of DLTV-QR with randomized prior functions and prior scale 20 (*DLTV [rpf20]*) compared to the vanilla implementation without priors (*DLTV [rpf0]*).

Information-gain in our IDS implementation is computed in a slightly modified way compared to the vanilla version. Nikolov et al. (2018) compute the information gain function $I(s, a)$ with

$$I(s, a) = \log\left(1 + \frac{\sigma^2(s, a)}{\rho^2(s, a)}\right) + \epsilon_2 \,,$$

where $\sigma^2(s, a)$ is the empirical variance of BDQN+P predictions, $\epsilon_2 = 1 \times 10^{-5}$ is a zero-division protection, and $\rho^2(s, a)$ is the clipped action-space normalized return variance

$$\rho(s, a)^2 = \max\left(\frac{\text{Var}(Z(s, a))}{\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \text{Var}(Z(s, a))}, 0.25\right). \tag{20}$$

$\text{Var}(Z(s, a))$ here is the variance of the distributional estimate provided by C51. We replace the clipping in Eq. (20) by adding a small constant $\epsilon_1 = 1 \times 10^{-4}$ to $\text{Var}(Z(s, a))$, s.t.

$$\rho_\epsilon(s, a)^2 = \frac{\text{Var}(Z(s, a)) + \epsilon_1}{\epsilon_1 + \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \text{Var}(Z(s, a))} \,.$$

Fig. 5 (b) shows the effect of clipping as in the vanilla version (*IDS-C51 [clip]*) compared to our variation (*IDS-C51 [noclip]*) on the deep sea environment.

Intrinsic reward priors are a computational method we implement with PE-DQN, which leverages the fact that we can compute the one-step uncertainty estimate $w_{\text{avg}}(s, a)$ deterministically from a parametric ensemble given a state-action tuple. This obviates the need to learn it explicitly in the bonus estimation step. We thus add $w_{\text{avg}}(s, a)$ automatically to the forward pass of the bonus estimator $b_\phi(s, a)$ as a sort of "prior" mechanism according to

$$b_\phi(s, a) := b_\phi^{\text{raw}}(s, a) + w_{\text{avg}}(s, a) \,,$$

where b_ϕ^{raw} is the raw output of the bonus estimator NN of parameters ϕ.

Bonus decay is the decaying of the exploratory bonus during action selection. It is well-known that the factor β is a sensitive parameter for UCB-type exploration algorithms, enabling efficient exploration when chosen correctly but simultaneously preventing proper convergence when chosen wrongly. Due to the variety of tasks included in the bsuite, we opted for a fixed schedule by which β is linearly decayed to 0.0 over one third of the total training horizon. In the bsuite experiments, we apply this schedule to all tested baselines where applicable and chose the initial β_{init} value according to the hyperparameter search.

Ensemble size is a central parameter in IDS and BDQN+P. We used a size of 20 as in the implementation by Osband et al. (2019a) , who find that increasing the ensemble size beyond 20 did not lead to significant performance improvements on the bsuite. Fig. 5 (a) shows a comparison of the influence of ensemble size in BDQN+P compared to PE-DQN.

The **computational resources** we used to conduct the bsuite experiments were supplied by cluster A [anonymized reference] and cluster B [anonymized reference]. We deployed the environments in 16 parallel jobs to be executed on 8 NVIDIA Tesla V100S 32GB GPUs, 16 Intel XEON E5-6248R 24C 3.0GHz CPUs, and 64GB of memory in total. In this setup, the execution of one seed on the entire suite experiment took approximately 38 hours for DLTV, 72 hours for PE-DQN, and 80 hours for IDS. Due to the narrower network architecture of BDQN+P, we in this case parallelized environments over 64 Intel XEON E5-6248R 24C 3.0GHz CPUs, taking approximately 76 hours wall-clock time for the entire suite.

B.3 Full results of bsuite experiments

Fig. 7 shows the averaged undiscounted episodic return for all bsuite tasks. Each curve represents the average over approximately 20 variations of the same task (Osband et al. (2019a) provide a detailed account of the task variations) where results were taken from a separate evaluation episode using a greedy action-selection rule. In the "scale" environments, evaluation results were rescaled to the original reward range to maintain a sensible average. Bold titles indicate environments tagged as hard exploration tasks.

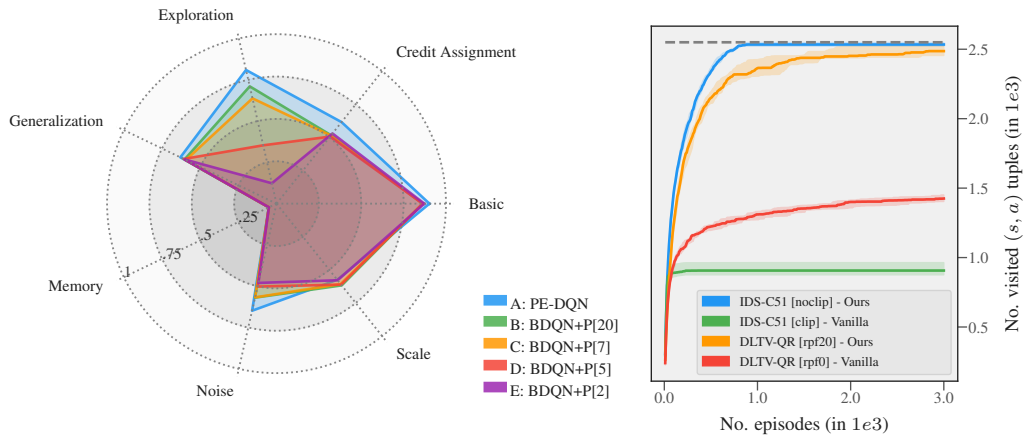

Figure 5: (a) Summary of bsuite experiments. Comparison between BDQN+P with different ensemble sizes and PE-DQN (total ensemble size 4). (b) Deep sea comparison between our implementations and vanilla implementations of baseline algorithms. Shown are median state-action visitation counts over number of episodes on the deep sea environment with size 50. Shaded regions represent the interquartile range of 10 seeds. Higher is better.

C Additional Experimental Results

We conduct an additional ablation study on PE-DQN to better understand the practical merits and limitations of its algorithmic components. Specifically, we aim to address the following questions:

(1) What is the effect of using diverse projection ensembles as opposed to ensembles composed of equivalent models?

(2) What is the effect of bootstrapping mixture distributions as opposed to fully independent training of ensemble members?

Fig. 6 shows the results of a study with several variations of PE-DQN on the extended deep sea benchmark outlined in Section 6.2. Here, PE-DQN [QR/QR] and PE-DQN [C51/C51] are variants of PE-DQN where we replace all models in the ensemble with all quantile-based or all categorical-based models respectively. While we find that the quantile-based ensemble is able to solve several levels of the deep sea environment, it scales significantly worse to larger problem sizes than the vanilla PE-DQN version. Unfortunately, we were not able to implement a well-performing ensemble based on categorical models, which consistently collapsed to zero predictions everywhere despite the use of prior functions. Conversely, PE-DQN [Ind.] consists of both a quantile and a categorical model but unlike PE-DQN trains these independently with models bootstrapping strictly their own target networks' predictions. In this case, the ensemble disagreement w_{avg} is added directly as a bonus to exploratory action selection. While PE-DQN [Ind.] can solve smaller instances of deep sea, we find that it does not scale well to larger problem sizes. We hypothesize that individual bootstrapping combined with repeated applications of the same projection may be less stable than the mixture bootstraps of PE-DQN.

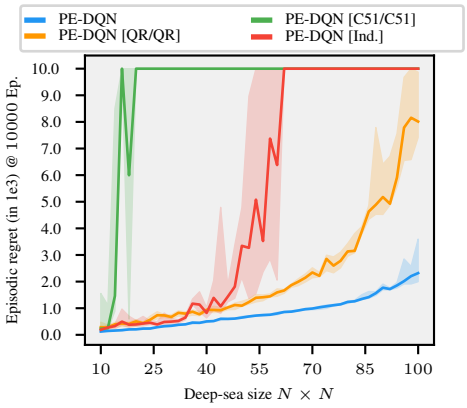

Figure 6: Ablation studies on extended deep sea environments. Median episodic regret over deep sea size. Lower is better. Shaded regions are the interquartile range of 10 seeds.

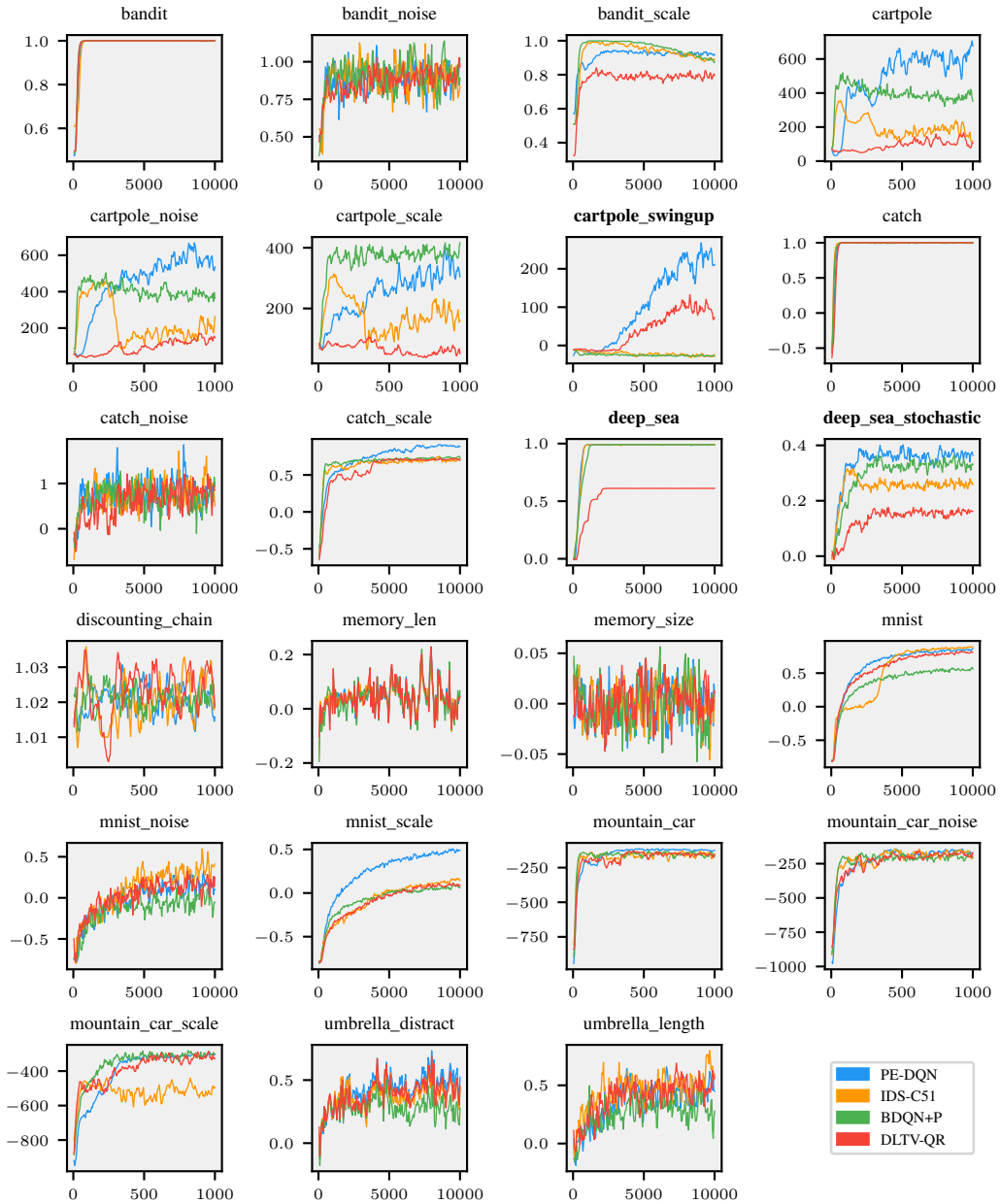

Figure 7: Averaged episodic return for all 23 bsuite tasks.