Reinforcement Learning under State and Outcome Uncertainty: A Foundational Distributional Perspective

Larry Preuett[1], **Qiuyi Zhang**[2], **Muhammad Aurangzeb Ahmad**[3]
lpreuett@uw.edu, qiuyiz@google.com, maahmad@uw.edu

[1]**University of Washington, Tacoma, USA**
[2]**Google DeepMind, California, USA**
[3]**University of Washington, Bothell, USA**

Abstract

In many real-world planning tasks, agents must tackle uncertainty about the environment's state and variability in the outcomes of any chosen policy. We address both forms of uncertainty as a first step toward safer algorithms in partially observable settings. Specifically, we extend Distributional Reinforcement Learning (DistRL)—which models the entire return distribution for fully observable domains—to Partially Observable Markov Decision Processes (POMDPs), allowing an agent to learn the distribution of returns for each conditional plan. Concretely, we introduce new distributional Bellman operators for partial observability and prove their convergence under the supremum p-Wasserstein metric. We also propose a finite representation of these return distributions via ψ-vectors, generalizing the classical α-vectors in POMDP solvers. Building on this, we develop Distributional Point-Based Value Iteration (DPBVI), which integrates ψ-vectors into a standard point-based backup procedure—*bridging DistRL and POMDP planning*. By tracking return distributions, DPBVI lays the foundation for future risk-sensitive control in domains where rare, high-impact events must be carefully managed. We provide source code to foster further research in robust decision-making under partial observability.

1 Introduction

Reinforcement Learning (RL) traditionally maximizes expected returns, treating each future reward distribution as a scalar. However, growing evidence suggests that modeling the full return distribution, rather than just its mean, can yield robust policies, better exploration, and richer theoretical insights (Bellemare et al., 2023). This distributional perspective, referred to as Distributional Reinforcement Learning (DistRL), captures variability and higher-order statistics of returns, providing agents with more profound information about environment stochasticity.

While DistRL has been well-studied in fully observable Markov Decision Processes (MDPs), many real-world systems are partially observable (Chiam et al., 2024; Komorowski et al., 2018; Caballero-Martin et al., 2024), forcing the agent to infer the latent state from noisy or incomplete observations. Such problems are naturally framed as Partially Observable Markov Decision Processes (POMDPs), but the literature on distributional methods in these settings remains sparse.

Beyond maximizing expected returns, many real-world settings require *risk-sensitive* planning, where limiting the probability of adverse or catastrophic outcomes is crucial. In partially observable domains, the stakes are even higher because the agent's uncertainty about the state can exacerbate the

impact of rare but high-cost events. For instance, a slight misjudgment of the patient's condition or the vehicle's surroundings in healthcare or autonomous navigation respectively can lead to undesirable or dangerous consequences. Incorporating risk considerations into POMDP planning explicitly accounts for such worst-case scenarios, making policies more conservative when high losses are possible and balancing exploration against the need to avoid significant harm. Establishing a foundational framework for distributional methods in POMDPs, as we do here, is thus a necessary first step toward risk-sensitive control in partially observable domains. Although we do not formally address risk-sensitive objectives, modeling full return distributions under partial observability is a necessary prerequisite for future work in this direction.

In this work, we present a formal extension of DistRL to POMDPs, bridging the gap between distributional theory and partial observability. Concretely, we make three major contributions:

- **Formal Extension of DistRL to POMDPs.** We define the partially observable distributional evaluation operator $\widetilde{\mathcal{T}}_{PO}$ and its optimality operator $\widetilde{\mathcal{T}}^*_{PO}$, showing that both are γ-contraction under the supremum p-Wasserstein metric ($1 \leq p < \infty$). This result generalizes classical DistRL convergence results to the partially observable setting.
- **Finite Representation via ψ-Vectors.** We introduce ψ-vectors, a distributional analog to the well-known α-vectors in POMDP theory. In the risk-neutral regime, we show that a finite set of ψ-vectors suffices to represent the optimal distributional value function while preserving the piecewise linear and convex (PWLC) property in the Wasserstein space.
- **Distributional Point-Based Value Iteration (DPBVI).** We adapt Point-Based Value Iteration (PBVI) to the distributional setting, yielding DPBVI, which uses ψ-vectors instead of α-vectors. Under risk-neutral objectives, DPBVI converges to the same policy as PBVI yet lays the groundwork for risk-sensitive extensions by maintaining full return distributions.

Alongside these theoretical results, we release our source code[1] to facilitate further exploration of risk-aware approaches in partially observable domains. Thus, our work unifies partial observability and DistRL, enabling distributionally robust decision-making where state uncertainty and return variability are central considerations.

2 Setting

We consider a finite, discrete-time Partially Observable Markov Decision Process (POMDP) $\mathcal{P}$. POMDPs are defined as a tuple: $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega, b_0, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is a set of discrete states, $\mathcal{A}$ is a set of discrete actions, $\mathcal{O}$ is the discrete space of noisy and/or incomplete state information, $T(s, a, s') = P(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the state transition model, $\Omega(o', s', a) = P(o_{t+1} = o' \mid s_{t+1} = s', a_t = a)$ is the sensor model, $b_0 \in \Delta$ is the initial belief, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $0 \leq \gamma < 1$ is the discount factor.

In this setting, the agent is unable to directly observe the state of the environment and must rely on a developed belief state $b \in \Delta$, a probability distribution across $\mathcal{S}$. The belief state serves as sufficient information for the agent to behave optimally. The agent's belief is developed using the sequence of observations $b_t = P(s_t \mid b_0, a_0, o_1, \ldots, o_{t-1}, a_{t-1}, o_t)$ by

$$\begin{aligned} b_t(s_t) &= \tau(b_{t-1}, a, o) \\ &= \frac{\Omega(o_t, s_t, a_{t-1}) \sum_{s_{t-1} \in \mathcal{S}} T(s_{t-1}, a_{t-1}, s_t) b_{t-1}(s_{t-1})}{\sum_{s_t \in \mathcal{S}} \Omega(o_t, s_t, a_{t-1}) \sum_{s_{t-1} \in \mathcal{S}} T(s_{t-1}, a_{t-1}, s_t) b_{t-1}(s_{t-1})} \end{aligned} \tag{1}$$

In POMDPs, the objective is to learn an optimal, stationary policy $\pi^*(a \mid b)$ which maximizes the expected return $\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}(s_t, a_t)\right]$ for time horizon T. Policies in this setting may be viewed as conditional plans and are typically learned by learning the optimal value function

[1] Source code available at github.com/lpreuettUW/distributional_point_based_value_iteration.

$$V^*(b) = \max_{a \in \mathcal{A}} \left[\sum_{s \in \mathcal{S}} \mathcal{R}(s,a) b(s) + \gamma \sum_{o' \in \mathcal{O}} \sum_{s' \in \mathcal{S}} \Omega(o', s', a) \sum_{s \in \mathcal{S}} T(s, a, s') b(s) V^*(b') \right] \tag{2}$$

The n-th horizon value function is comprised of a set of α-vectors $V_n = \{\alpha_0, \alpha_1, \ldots, \alpha_m\}$ and is piecewise linear and convex (PWLC) (Sondik, 1978). Each α-vector is a $|\mathcal{S}|$-dimensional hyperplane and defines the value function for some bounded region of Δ (i.e., $\max_{\alpha \in V} \alpha \cdot b$). At each planning step, the next value function V_n may be computed from the previous value function V_{n-1} via the Bellman optimality operator $\mathcal{T}_{PO}$

$$\begin{aligned} V_n(b) &= \mathcal{T}_{PO} V_{n-1} \\ &= \max_{a \in \mathcal{A}} \left[\sum_{s \in \mathcal{S}} \mathcal{R}(s,a) b(s) + \gamma \sum_{o' \in \mathcal{O}} \max_{\alpha \in V_{n-1}} \sum_{s' \in \mathcal{S}} \Omega(o', s', a) \sum_{s \in \mathcal{S}} T(s, a, s') b(s) \alpha(s') \right] \end{aligned} \tag{3}$$

2.1 Point-Based Value Iteration

PBVI computes $V_t = \mathcal{T}_{PO} V_{t-1}$ in three steps. First, it creates projections for each action and observation

$$\begin{aligned} &\Gamma^{a,*} \leftarrow \alpha^{a,*}(s) = \mathcal{R}(s,a) \\ &\Gamma^{a,o'} \leftarrow \alpha_i^{a,o'}(s) = \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') \Omega(o', s', a) \alpha_i^{t-1}(s'), \forall \alpha_i^{t-1} \in V_{t-1} \end{aligned} \tag{4}$$

Next, the value of action a at belief b is computed by

$$\Gamma_b^a = \Gamma^{a,*} + \sum_{o' \in \mathcal{O}} \arg\max_{\alpha \in \Gamma^{a,o'}} (\alpha \cdot b) \tag{5}$$

Lastly, the best action at each belief is used to construct V_t

$$V_t \leftarrow \arg\max_{\Gamma_b^a, \forall a \in \mathcal{A}} (\Gamma_b^a \cdot b), \forall b \in \mathcal{B} \tag{6}$$

V_t is comprised of at most $|\mathcal{B}|$ α-vectors, thus requiring $|\mathcal{S}||\mathcal{A}||V_{t-1}||\mathcal{O}||\mathcal{B}|$ operations per backup.

2.2 Distributional Reinforcement Learning

Traditionally, the value function models the return expectation. DistRL, instead, aims to learn the compound distribution of the returns $Z^\pi(s)$ sourced to randomness in (1) reward R (2) transition P^π, and (3) distribution of the next-state value $Z(S')$. Let $\widetilde{\mathcal{T}}$ be the distributional Bellman operator such that

$$\widetilde{\mathcal{T}}^\pi Z(s) \overset{D}{=} R + \gamma P^\pi Z(s) \tag{7}$$

where $\overset{D}{=}$ denotes equality in distribution, R is the reward random variable, and $P^\pi Z(s) \overset{D}{=} Z(S')$ for $S' \sim T(s, a, \cdot)$.

The distributional Bellman operator has been shown to be a γ-contraction under the supremum p-Wasserstein metric space $\forall p \in [1, \infty]$ (Bellemare et al., 2023). The distributional Bellman optimality operator $\widetilde{\mathcal{T}}^*$, which selects actions that maximize the expected return, similarly converges to the fixed point under the supremum p-Wasserstein metric space $\forall p \in [1, \infty]$ (Bellemare et al., 2023) if there is a unique optimal policy and a mean-preserving distribution representation $\mathcal{F}$ is used (e.g., a categorical distribution representation). In the case of the existence of multiple optimal policies, however, convergence isn't guaranteed because differing distributions can be similar in expectation.

3 Distributional Reinforcement Learning Under Partial Observability

In the presence of partial observability we must account for additional sources of randomness; namely, randomness in (a) the unobservable state $\mathcal{S}$, (b) observations O and (c) sensing Ω^π. Therefore, the return distribution is learned with respect to belief instead of state (i.e., $Z(b)$).

3.1 The Partially Observable Distributional Bellman Operators

Let $\widetilde{\mathcal{T}}_{PO}$ be the partially observable distributional Bellman operator such that

$$\widetilde{\mathcal{T}}^\pi_{PO} Z(b) \stackrel{D}{=} R + \gamma \tau^\pi Z(b) \tag{8}$$

where $\tau^\pi Z(b) \stackrel{D}{=} Z(B')$ for $O \sim \Omega(\cdot, \mid S', A), S' \sim T(\cdot \mid S, A), A \sim \pi(\cdot \mid b)$. The partially observable distributional Bellman optimality operator $\widetilde{\mathcal{T}}^*_{PO}$ may be similarly defined, but where actions are selected to maximize the expected return. The distributional Bellman operators in the partially observable setting share convergence properties with those of the fully observable setting, because any POMDP can be rewritten as a belief MDP, where the state of the belief MDP is the belief of the POMDP.

We denote by $\mathfrak{P}_p(\mathbb{R})$ the set of Borel probability measures on $\mathbb{R}$ with finite p-th moment, ensuring that the p-Wasserstein distance is well-defined. The space $\mathfrak{P}_p(\mathbb{R})^\Delta$ then denotes the set of functions $\eta : \Delta \to \mathfrak{P}_p(\mathbb{R})$, where Δ is the belief space of the POMDP.

Theorem 1. *For any POMDP $\mathcal{M}$ with rewards bounded by $[R_{min}, R_{max}]$ the partially observable distributional Bellman operator is a contraction on $\mathfrak{P}_p(\mathbb{R})^\Delta$ in the supremum p-Wasserstein distance $\bar{w}_p, \forall p \in [1, \infty), \forall \gamma \in [0, 1)$. That is,*

$$\bar{w}_p(\widetilde{\mathcal{T}}^\pi_{PO}\eta, \widetilde{\mathcal{T}}^\pi_{PO}\eta') \leq \gamma \bar{w}_p(\eta, \eta'),$$

$\forall \eta, \eta' \in \mathfrak{P}_p(\mathbb{R})^\Delta$.

See Appendix A.1 for proof.

Theorem 1 follows from the standard γ-contraction properties of $\widetilde{\mathcal{T}}^\pi$. Because $\widetilde{\mathcal{T}}^\pi_{PO}$ is a contraction mapping, by the Banach Fixed-Point Theorem, there exists a unique $\eta^* \in \mathfrak{P}_p(\mathbb{R})^\Delta$ such that η^* is a fixed point.

Theorem 2. *Let $\widetilde{\mathcal{T}}^*_{PO}$ be the partially observable distributional Bellman optimality operator for a POMDP $\mathcal{M}$ with bounded rewards and let $p \in [1, \infty]$. Suppose there is a unique optimal policy π^* and, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution is $P_{\mathfrak{R}}(\cdot \mid s, a)$. For any initial return-distribution function $\eta_0 \in \mathfrak{P}_p(\mathbb{R})$, the sequence of iterates*

$$\eta_{k+1} \;=\; \widetilde{\mathcal{T}}^*_{PO}\, \eta_k$$

converges in the supremum p-Wasserstein distance $\overline{w}_p$ to η^{π^*}, the return distribution associated with the unique optimal policy.

Corollary 1. *Consider a mean-preserving projection $\Pi_\mathcal{F}$ for some representation $\mathcal{F}$. Suppose $\Pi_\mathcal{F}$ is a nonexpansion in the supremum p-Wasserstein distance, and let the finite domain $\mathfrak{P}_p(\mathbb{R})^\Delta$ be closed under $\Pi_\mathcal{F}$. If there is a unique optimal policy π^* then, under the conditions of Theorem 2, the sequence*

$$\eta_{k+1} = \Pi_\mathcal{F} \widetilde{\mathcal{T}}^*_{PO} \eta_k$$

converges to the fixed point η^{π^*}.

These results build on the convergence guarantees of the distributional Bellman optimality operator in Bellemare et al. (2023, Theorem 7.9), which assumes a unique optimal policy and analyzes the tabular MDP setting. Our proof adapts these arguments to the belief MDP by ensuring the required

continuity and compactness conditions hold, allowing the result to generalize to partially observable settings (see Appendix A.2). Bellemare et al. note that the unique optimal policy assumption is necessary for contraction guarantees but they empirically observe that algorithms like C51 often perform well even when this condition may not strictly hold Bellemare et al. (2023).

Theorem 2 states if there is a unique optimal policy, repeatedly applying $\widetilde{\mathcal{T}}^*_{PO}$ will converge to the optimal return distribution. Corollary 1 establishes that if we approximate the return distribution after each greedy update, the resulting approximate iteration still converges to the same fixed point, provided there is a unique optimal policy.

3.2 ψ-vectors

In classical POMDP value iteration, each conditional plan—a sequence of actions contingent on future observations—is captured by an α-vector, which stores said plan's expected return for each state. Thus, a belief-state's value is found by selecting the α-vector with the highest dot product $\alpha \cdot b$. While this approach gives the expected returns of each plan, it does not capture the uncertainty around possible outcomes.

To address this, we adopt a distributional viewpoint: rather than storing a scalar expectation, we store a distribution over returns for each plan. We call these ψ-vectors. Formally, a ψ-vector is

$$\Psi = \{\psi^s \mid s \in \mathcal{S}\}, \tag{9}$$

where ψ^s is a distribution over returns conditioned on being in state s. Given a set of candidate ψ-vectors Γ, the distributional value function at a belief b is then obtained by taking the maximum distributional inner product:

$$Z(b) \;=\; \max_{a \in \mathcal{A}} \max_{\Psi \in \Gamma_a} \big\langle \Psi, b \big\rangle, \tag{10}$$

where $\Gamma_a \subset \Gamma$ is the subset of ψ-vectors associated with action a. Here, $\langle \Psi, b \rangle$ denotes a mixture distribution of the ψ^s, weighted by the belief b. In the risk-neutral setting, we typically evaluate this mixture by taking its expectation, mirroring the $\alpha \cdot b$ calculation from classical POMDPs—but retaining the entire distribution allows us to capture richer information about potential outcomes.

Theorem 3. *Under the assumptions of Theorem 2 and Corollary 1, the optimal distributional value function $Z^*(b)$ can be represented as a finite set of ψ-vectors Ψ, $\forall a \in \mathcal{A}$, such that for any belief state $b \in \Delta$,*

$$Z^*(b) = \max_{a \in \mathcal{A}} \max_{\Psi \in \Psi^*_a} \langle \Psi, b \rangle,$$

and these ψ-vectors maintain piecewise linearity and convexity within the Wasserstein metric space.

See Appendix A.3 for proof.

By Theorem 3, Z^* inherits the PWLC structure of classical POMDPs, but within the Wasserstein metric space. Crucially, this representation remains finite in the risk-neutral case, ensuring computational tractability—just as a finite set of α-vectors suffices to represent the classical optimal value function V^*.

4 Distributional Point-Based Value Iteration

Building upon the foundations of PBVI and DistRL in the partially observable setting, we now introduce *Distributional Point-Based Value (DPBVI)*. Conceptually, DPBVI is PBVI with $\widetilde{\mathcal{T}}^*_{PO}$ in place of $\mathcal{T}^*_{PO}$. As in standard PBVI, we retain point-based belief expansion (Pineau et al., 2003), but our backup step updates a finite set of ψ-vectors rather than α-vectors.

Recall that in PBVI, we maintain a finite set of belief points $\mathcal{B} \subset \Delta$, where Δ is the probability simplex over states $\mathcal{S}$. Each backup step creates new α-vectors to approximate the PWLC value

function, then selects one α-vector per belief. DPBVI extends this idea to the distributional setting by replacing α-vectors with ψ-vectors, where each ψ-vector stores one distribution per state for a given action.

By Theorem 3, a finite set of these ψ-vectors suffices to represent the optimal distributional value function Z^* in the risk-neutral setting, just as α-vectors suffice in classical POMDPs. We represent return distributions using the categorical distribution parameterization, which is mean-preserving when its support spans all possible returns (Bellemare et al., 2023). This ensures that the expected returns align with the classical POMDP value function, preserving the key convergence guarantees from Corollary 1 while operating in the space of return distributions.

4.1 The DBPVI Backup

Given the previous set of ψ-vectors Γ^{t-1}, each DPBVI backup follows four steps: 1. action-observation projection, 2. categorical distribution projection, 3. belief-specific maximization, and 4. candidate ψ-vector selection.

We create candidate ψ-vectors by combining the immediate rewards $\widetilde{R}_a$ with the next-step distributions of returns, conditioned on actions a and observations o'. Formally,

$$\Gamma_a^{o',t} \leftarrow \left\{ \widetilde{R}_a + \gamma\, T(\cdot, a, s')\, \Omega(o', s', a)\, \psi^{s'} \;\middle|\; \Psi \in \Gamma^{t-1},\ s' \in \mathcal{S} \right\}. \tag{11}$$

This step yields $|\mathcal{A}| \times |\mathcal{O}| \times |\Gamma^{t-1}|$ projected distributions, denoted by $\Gamma^{\cdot,t}$. Because each distribution reflects one action-observation pair, it may be *subnormalized* until we later sum over observations.

The categorical distribution representation is not closed under $\widetilde{\mathcal{T}}^*_{PO}$ due to the γ-scaling and immediate reward addition operations on prior ψ-vectors. This means the subnormalized distributions in $\Gamma^{\cdot,t}$ may have heterogeneous supports. To ensure each ψ-vector has the same support, we apply the categorical projection operator Π_c

$$\Gamma^{\cdot,t} \leftarrow \{\Pi_c(\Gamma_a^{o',t}) \mid \Gamma_a^{o',t} \in \Gamma^{\cdot,t}\}.$$

For each belief $b \in \mathcal{B}$, we choose the combination of ψ-vectors across observations that yields the highest expected distribution (under the risk-neutral lens). Symbolically, for a single action a,

$$\Gamma_a^{b,t} = \sum_{o' \in \mathcal{O}} \operatorname*{arg\,max}_{\Psi \in \Gamma_a^{o',t}} \langle \Psi, b \rangle. \tag{12}$$

This step mirrors the cross-sum and maximize operation in PBVI, except each summand is a distributional mixture rather than a scalar $\alpha-$vector. Summing those best subnormalized distributions across all observations recovers a full next-step return distribution.

Finally, for each belief b, we pick the best action's distribution by

$$\Psi_b^t \leftarrow \operatorname*{arg\,max}_{\Gamma_a^{b,t},\, a \in \mathcal{A}} \langle \Gamma_a^{b,t}, b \rangle. \tag{13}$$

Collecting $\{\psi_b^t\}_{b \in \mathcal{B}}$ yields our updated set Γ^t. As in PBVI, we keep at most $|\mathcal{B}|$ vectors in Γ^t, thus maintaining a point-based approximation.

Because we manage distributions instead of scalars, each backup is slower by a factor of $|M|$, where $|M|$ is the number of bins in each categorical distribution. Concretely, one DPBVI backup runs in $\mathcal{O}(|\mathcal{S}|\ |\mathcal{A}|\ |\Gamma^{t-1}|\ |\mathcal{O}|\ |\mathcal{B}|\ |M|)$, compared to $\mathcal{O}(|\mathcal{S}|\ |\mathcal{A}|\ |V_{t-1}|\ |\mathcal{O}|\ |\mathcal{B}|)$ for PBVI. Nonetheless, DPBVI preserves the core PBVI structure—belief set, backup iteration, and finite policy representation—while enabling distributional modeling of returns under partial observability.

5 Experimental Results

We empirically validate that DPBVI recovers the same value function as PBVI under risk-neutral objectives, as predicted by theory. Our focus is not on performance or efficiency, but on whether distributional backup operations preserve correctness relative to standard scalar PBVI.

We evaluate DPBVI and PBVI in two small POMDP domains: (1) MiniGrid DoorKey (5×5), where the agent must reach a goal behind a locked door using a key, and (2) a Two-State Noisy-Sensor domain with binary state transitions and stochastic observations. See Appendix B for full environment and implementation details.

Table 1 reports runtime and iteration counts for both methods. As expected, DPBVI incurs greater computational cost due to distributional operations but converges in the same number of iterations as PBVI. To assess value function agreement, we compare the expected values of DPBVI's ψ-vectors with PBVI's scalar values across belief points. Figure 1 shows the maximum relative error per iteration. Errors remain very small, confirming that DPBVI tracks PBVI's value function closely.

We further test for compounding error by tightening the convergence threshold in the Two-State Noisy-Sensor environment (Figure 2). The relative error initially increases, then stabilizes and decreases, suggesting discrepancies stem from categorical projection, floating-point limits, and bootstrapping rather than theoretical flaws. These findings confirm that DPBVI is a correct extension of PBVI under risk-neutral assumptions and establish a foundation for future risk-sensitive variants.

6 Conclusion

In this paper, we presented the first comprehensive study of Distributional Reinforcement Learning in partially observable domains. Our main theoretical result showed that the partially observable distributional Bellman operator $\widetilde{\mathcal{T}}^{\pi}_{PO}$ and its optimality counterpart $\widetilde{\mathcal{T}}^{*}_{PO}$ are γ-contractions in the supremum p-Wasserstein metric ($1 \leq p < \infty$). These results extend the well-known contraction properties of distributional Bellman operators in fully observable settings to POMDPs.

As part of our analysis, we introduced ψ-vectors as the distributional analogs of α-vectors, proving that under risk-neutral control, the optimal distributional value function in a POMDP can be represented by a finite collection of these ψ-vectors while preserving PWLC in the Wasserstein metric space. This representation underlines how a distributional perspective naturally generalizes classical POMDP theory.

Building on these foundations, we described DPBVI, a risk-neutral algorithm that applies the distributional Bellman optimality operator to a finite set of belief points. This extension parallels the classical PBVI procedure yet replaces scalar α-vectors with distributional ψ-vectors, thereby capturing richer information about uncertainty in returns. Experiments (see Section 5 and Appendix B) show that DPBVI recovers the same solution as PBVI under risk-neutral control, confirming that the distributional approach does not compromise solution quality in the risk-neutral regime.

While our empirical and theoretical results focus solely on the risk-neutral setting, this work enables future exploration of risk-sensitive approaches by extending the distributional perspective to partially observable domains. Capturing full return distributions under uncertainty lays the groundwork for future work on risk-sensitive control. We hope these developments foster further exploration of DistRL in practical domains where both state uncertainty and variability in outcomes are central concerns.

That said, while our experiments validate theoretical correctness, they are limited to small synthetic environments and do not assess scalability or empirical performance in large-scale settings. Future work may explore practical extensions of DPBVI to high-dimensional problems and risk-sensitive criteria.

Acknowledgments

The authors thank Ankur Teredesai and Kevin Jamieson for their helpful feedback on early drafts of this work. Muhammad Aurangzeb Ahmad's work was supported by the National Institutes of Health, USA (NIH grant T32 GM 121290).

A Proofs

A.1 Proof of Theorem 1

Theorem 1. *For any POMDP $\mathcal{M}$ with rewards bounded by $[R_{min}, R_{max}]$ the partially observable distributional Bellman operator is a contraction on $\mathfrak{P}_p(\mathbb{R})^{\Delta}$ in the supremum p-Wasserstein distance $\bar{w}_p, \forall p \in [1, \infty), \forall \gamma \in [0, 1)$. That is,*

$$\bar{w}_p(\widetilde{\mathcal{T}}^{\pi}_{PO}\eta, \widetilde{\mathcal{T}}^{\pi}_{PO}\eta') \leq \gamma \bar{w}_p(\eta, \eta'),$$

$\forall \eta, \eta' \in \mathfrak{P}_p(\mathbb{R})^{\Delta}$.

Proof. We prove contraction of the partially observable distributional Bellman operator $\widetilde{\mathcal{T}}^{\pi}_{PO}$ by lifting the POMDP $\mathcal{M}$ to its equivalent fully observable belief MDP $\widetilde{\mathcal{M}}$, where states are beliefs $b \in \Delta$, actions are $a \in \mathcal{A}$, and transitions are governed by the belief update $\tau(b, a, o')$ with probability

$$P(o' \mid b, a) = \sum_{s,s'} b(s) T(s, a, s') \Omega(o' \mid s', a).$$

Assume the following:

- Δ is a Borel space (as it is a subset of a finite-dimensional simplex).
- $\mathcal{A}$ is finite.
- The reward function $R(b, a) := \mathbb{E}_{s\sim b}[R(s, a)]$ is bounded.
- The belief transition kernel $\tau(\cdot \mid b, a)$ is measurable.
- The policy $\pi : \Delta \to \mathcal{A}$ is fixed and measurable.

Let $\eta, \eta' \in \mathfrak{P}_p(\mathbb{R})^{\Delta}$ be two distribution-valued value functions. Define the supremum p-Wasserstein metric as:

$$\bar{w}_p(\eta, \eta') := \sup_{b\in\Delta} W_p(\eta(b), \eta'(b)), \quad \forall p \in [1, \infty).$$

Fix $b \in \Delta$ and let (G_b, G'_b) be an optimal coupling of $\eta(b)$ and $\eta'(b)$. Define:

$$\tilde{G}_b := R(b, \pi(b)) + \gamma G_{b'}, \quad \tilde{G}'_b := R(b, \pi(b)) + \gamma G'_{b'},$$

where $b' \sim \tau(\cdot \mid b, \pi(b))$ is the next belief under the belief transition kernel.

Then:

$$W_p(\widetilde{\mathcal{T}}^{\pi}_{PO}\eta(b), \widetilde{\mathcal{T}}^{\pi}_{PO}\eta'(b)) \leq \mathbb{E}_{b'}\left[W_p(\eta(b'), \eta'(b'))\right] \leq \gamma \bar{w}_p(\eta, \eta').$$

Taking the supremum over $b \in \Delta$, we conclude:

$$\bar{w}_p\left(\widetilde{\mathcal{T}}^{\pi}_{PO}\eta, \widetilde{\mathcal{T}}^{\pi}_{PO}\eta'\right) \leq \gamma \bar{w}_p(\eta, \eta'),$$

as desired. □

A.2 Proof of Theorem 2

Theorem 2. *Let $\widetilde{\mathcal{T}}^*_{PO}$ be the partially observable distributional Bellman optimality operator for a POMDP $\mathcal{M}$ with bounded rewards and let $p \in [1, \infty]$. Suppose there is a unique optimal policy π^* and, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution is $P_{\mathfrak{R}}(\cdot \mid s, a)$. For any initial return-distribution function $\eta_0 \in \mathfrak{P}_p(\mathbb{R})$, the sequence of iterates*

$$\eta_{k+1} \;=\; \widetilde{\mathcal{T}}^*_{PO}\, \eta_k$$

converges in the supremum p-Wasserstein distance $\overline{w}_p$ to η^{π^*}, the return distribution associated with the unique optimal policy.

Proof. We lift the POMDP $\mathcal{M}$ to its equivalent belief MDP $\widetilde{\mathcal{M}}$, where each belief $b \in \Delta$ serves as a fully observable state. The transition dynamics $\tau(b, a, o')$ of the belief MDP are Markovian, and the reward function is defined as $\mathcal{R}(b, a) := \mathbb{E}_{s\sim b}[\mathcal{R}(s, a)]$, which is bounded by assumption.

From this point forward, we use $b \in \Delta$ to denote the state of the belief MDP. We assume the following conditions hold in the belief MDP:

- The belief state space Δ is compact and convex
- The reward function $\mathcal{R}(b, a)$ is continuous in $b\; \forall a \in \mathcal{A}$
- The belief update function $\tau(b, a, o')$ is continuous in b for all $a \in \mathcal{A}$ and $o' \in \mathcal{O}$
- The action space $\mathcal{A}$ is finite
- The policy space Π is fixed and measurable
- There exists a unique optimal policy π^*

Let η^{π^*} denote the return distribution induced by the optimal policy. The partially observable distributional Bellman optimality operator $\widetilde{\mathcal{T}}^*_{PO}$ can be interpreted as a greedy update rule over return distributions in the belief MDP.

Define $Q_\eta(b, a) := \mathbb{E}_{Z\sim\eta(b,a)}[Z]$ as the expected return under the return distribution η. The *action gap* at state $b \in \Delta$ is defined as

$$\text{GAP}(Q_\eta, b) := \min\{Q_\eta(b, a^*) - Q_\eta(b, a) : a^*, a \in \mathcal{A}, a^* \neq a, Q_\eta(b, a^*) := \max_{a'\in\mathcal{A}} Q_\eta(b, a')\}$$

The global action gap is then defined as

$$\text{GAP} := \inf_{b\in\Delta} \text{GAP}(Q_\eta, b)$$

We now justify that $Q_\eta(b, a)$ is continuous in b for all $a \in \mathcal{A}$. The return distribution $\eta(b, a)$ is defined recursively from the reward distribution and belief transitions. Under our assumptions that the reward function $\mathcal{R}(b, a)$ and belief update $\tau(b, a, o')$ are continuous in b, it follows that $Q_\eta(b, a) = \mathbb{E}_{Z\sim\eta(b,a)}[Z]$ is continuous in b as well. Since the action space is finite, the maximum and second maximum of $\{Q_\eta(b, a)\}_{a\in\mathcal{A}}$ are continuous in b, so $\text{GAP}(Q_\eta, b)$ is also continuous in b.

By the Extreme Value Theorem, since $\text{GAP}(Q_\eta, b)$ is continuous on the compact state space Δ, the infimum is attained and equal to the minimum. Since the optimal policy is unique, $\text{GAP}(Q_\eta, b) > 0\; \forall b \in \Delta$, thus $\text{GAP} > 0$.

Fix $\epsilon = \frac{1}{2}\text{GAP}(Q^*)$. The standard Bellman optimality operator is known to be a γ-contraction under the L^∞ norm. Consequently, $\exists k \in \mathbb{N}$ such that

$$||Q_{\eta_k} - Q^*||_\infty < \epsilon \qquad \forall k \geq K$$

For any fixed b, let a^* be the optimal action in that state. Then for any $a \neq a^*$, we have that

$$\begin{aligned} Q_{\eta_k}(b, a^*) &\geq Q^*(b, a^*) - \epsilon \\ &\geq Q^*(b, a) + \text{GAP}(Q^*) - \epsilon \\ &> Q_{\eta_k}(b, a) + \text{GAP}(Q^*) - 2\epsilon \\ &= Q_{\eta_k}(b, a) \end{aligned}$$

Thus, any greedy selection rule applied to η_k will yield the optimal policy π^* $\forall k \geq K$. From this point onward, the Bellman updates correspond to evaluation under a fixed policy π^*, and we may invoke Theorem 1 with the initial return distributions $\eta_0 = \eta_k$ to conclude that $\eta_k \to \eta^{\pi^*}$. Hence, $\widetilde{\mathcal{T}}^*_{PO}$ is a γ-contraction and converges to the unique fixed point under the supremum p-Wasserstein metric.

□

A.3 Proof of Theorem 3

Theorem 3. *Under the assumptions of Theorem 2 and Corollary 1, the optimal distributional value function $Z^*(b)$ can be represented as a finite set of ψ-vectors Ψ, $\forall a \in \mathcal{A}$, such that for any belief state $b \in \Delta$,*

$$Z^*(b) = \max_{a \in \mathcal{A}} \max_{\Psi \in \Psi^*_a} \langle \Psi, b \rangle,$$

and these ψ-vectors maintain piecewise linearity and convexity within the Wasserstein metric space.

Proof. By corollary 1, there exists a unique optimal distributional value function Z^*. Let $\Gamma^* = \{\Psi^*_a \mid \forall a \in \mathcal{A}\}$. Then for any belief b,

$$\begin{aligned} Z^*(b) &= \max_{a \in \mathcal{A}} \max_{\Psi \in \Psi^*_a} \langle \Psi, b \rangle \\ &\overset{(a)}{=} \max_{\Psi \in \Gamma^*} \langle \Psi, b \rangle \\ &\overset{(b)}{=} \max_{\alpha \in \Gamma^*} \alpha \cdot b \\ &\overset{(c)}{=} V^*(b) \end{aligned}$$

(a) follows from the definition of Γ^* as the union of Ψ_a for all a, (b) because of risk-neutrality, $\max_{\Psi \in \Gamma^*} \langle \Psi, b \rangle$ is determined by the maximum expectation. That is, $\arg\max_\Psi \langle \Psi, b \rangle = \arg\max_\Psi \mathbb{E}[\langle \Psi, b \rangle]$, where $\alpha = \mathbb{E}[\Psi]$. (c) follows since $\alpha \in \Gamma^*$ precisely matches the expected returns that define the classical optimal value function V^* in a POMDP. Finally, by Sondik's result (Sondik, 1978), V^* is PWLC and has finite representation. Therefore, Z^* is PWLC and may be represented by a finite set of ψ-vectors. □

B Detailed Experimental Results

Our primary goal in these experiments is to verify that DPBVI converges to the same solution as PBVI under risk-neutral objectives, rather than to improve upon the performance or efficiency of PBVI. Indeed, DPBVI is slower due to the additional cost of handling distributions, but it offers a framework that can be extended to risk-sensitive criteria in future work.

B.1 Environments

We evaluate on two small POMDP environments, each with its own dynamics models:

Table 1: Results of DPBVI and PBVI: average runtime and iteration counts for convergence criterion $\epsilon = 1\text{e}{-3}$.

Environment	Algorithm	Avg. Runtime (s)	# Iterations
MiniGrid DoorKey	PBVI	0.014	11
	DPBVI	54.805	11
Two-State	PBVI	0.083	789
	DPBVI	0.397	789

MiniGrid DoorKey (5x5) This environment, adapted from the MiniGrid suite (Chevalier-Boisvert et al., 2023, MiniGrid-DoorKey-5x5-v0), places an agent in a 5x5 grid containing a locked door and a key. The agent's objective is to reach a goal state behind a locked door requiring a key. We assume both perfect sensor and transition models, a discount factor $\gamma = 0.9$, a reward of 1.0 upon reaching the goal state, and a known start state. Because the start state is known and the dynamics models are perfect, we set $\mathcal{B}$ to include complete confidence in being in one of the ten states along the optimal path (i.e., $|\mathcal{B}| = 10$).

Two-State Noisy-Sensor This is a simple domain with two states, $\{s_0, s_1\}$, and a noisy sensor adapted from the environment in Russell & Norvig (2020, Section 17.5). The two actions are *Go* and *Stay*. *Go* changes states with probability 0.9 and *Stay* stays in the same state with probability 0.9. The agent received a reward of 1.0 each timestep it is in stay s_1 and 0 otherwise. With probability 0.6 the sensor reports the correct state. We set the discount factor $\gamma = 0.99$ and $\mathcal{B}$ to include twenty evenly spaced beliefs between complete belief of being in either state (i.e., $|\mathcal{B}| = 20$).

B.2 Setup

We disable belief-point expansion in both environments and fix a finite set $\mathcal{B}$ of belief states as described in Section B.1. PBVI is implemented as in (Pineau et al., 2003), while DPBVI follows the backup procedure from Section 4. Return distributions in DPBVI use 51 categorical atoms, with each ψ-vector initialized to place all probability mass at zero.

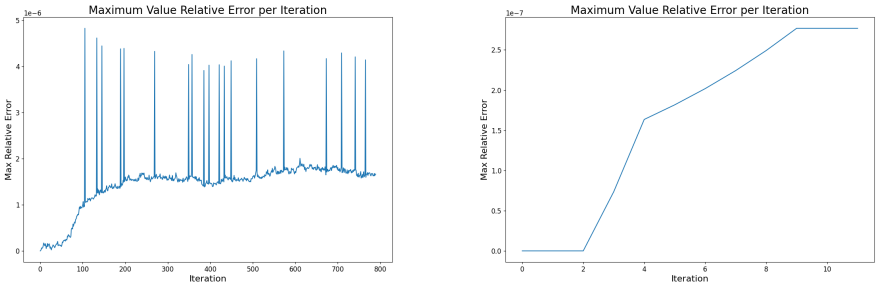
For the MiniGrid DoorKey environment, we set the distribution support to $[0, 5]$, reflecting the small range of cumulative rewards in a 5×5 grid. In the Two-State Noisy-Sensor environment, we adopt a larger support $[0, 100]$ to accommodate higher returns. Upon convergence, we compare the expected values of DPBVI's distributions against the scalar value function from PBVI, using identical discount factors, reward functions, and stopping criteria in both algorithms.

B.3 Results and Discussion

Table 1 compares the average runtimes and the number of iterations required for DPBVI and PBVI to converge on the MiniGrid DoorKey and Two-State environments. In both domains, DPBVI recovers the same optimal risk-neutral solution as PBVI, though with significantly higher runtime due to the additional cost of distributional operations.

To validate that DPBVI converges to the same value function as PBVI, we examine the maximum relative error between their value functions across belief points over successive backups. These results are shown in Figure 1 for both environments with a convergence criterion of $\epsilon = 1\text{e}{-3}$.

In both domains, we observe very small value differences throughout the run, confirming that DPBVI effectively recovers the same solution as PBVI. In the Two-State environment (Figure 1a), the relative error increases gradually and exhibits intermittent spikes. We hypothesize these spikes arise from categorical projection error, floating-point precision limits, and bootstrapping. In MiniGrid DoorKey (Figure 1b), the relative error increases slightly before plateauing around $1.2\text{e}{-7}$, again validating close alignment between both methods.



(a) Two-State Noisy-Sensor Environment (b) MiniGrid DoorKey Environment

Figure 1: Maximum relative error between DPBVI and PBVI value functions per iteration with convergence criterion $\epsilon = 1\text{e}{-3}$.

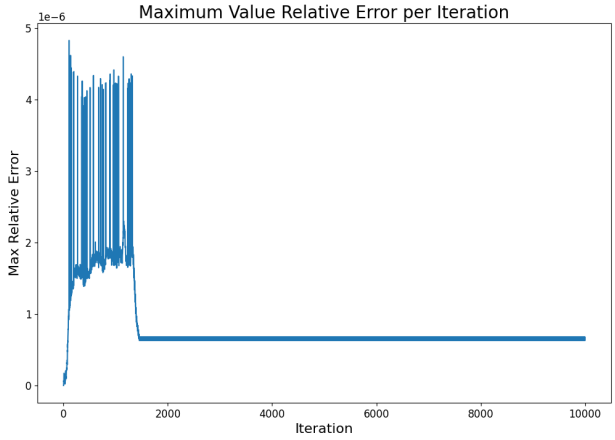


Figure 2: Maximum relative error between DPBVI and PBVI value functions with convergence threshold $\epsilon = 1\text{e}{-6}$ in the Two-State Noisy-Sensor environment.

To test whether the value function discrepancy is a function of the number of backups—suggesting compounding error or a design flaw in DPBVI—we tightened the convergence threshold to $\epsilon = 1\text{e}{-6}$ and set the iteration limit to $10,000$ in the Two-State environment. Results are shown in Figure 2.

As shown in Figure 2, the relative error increases over approximately the first $1{,}500$ backups, peaks just below $3\text{e}{-4}$, and then gradually decreases and plateaus near $5\text{e}{-5}$. This behavior supports our hypothesis that the discrepancy arises from bootstrapping, floating-point error, and categorical projection—rather than compounding error over time. Notably, neither PBVI nor DPBVI fully converged under this stricter threshold, yet the error remained bounded and decreased with additional backups. The MiniGrid DoorKey results remained unchanged under this more restrictive convergence criterion.

References

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Daniel Caballero-Martin, Jose Manuel Lopez-Guede, Julian Estevez, and Manuel Graña. Artificial intelligence applied to drone control: A state of the art. *Drones*, 8(7), 2024. ISSN 2504-446X.

DOI: 10.3390/drones8070296. URL https://www.mdpi.com/2504-446X/8/7/296.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Jodi Chiam, Aloysius Lim, and Ankur Teredesai. Nudgerank: Digital algorithmic nudging for personalized health. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 4873–4884, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. DOI: 10.1145/3637528.3671562. URL https://doi.org/10.1145/3637528.3671562.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, November 2018. ISSN 1078-8956, 1546-170X. DOI: 10.1038/s41591-018-0213-5. URL https://www.nature.com/articles/s41591-018-0213-5.

Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: an anytime algorithm for pomdps. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pp. 1025–1030, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020. ISBN 9781292401133. URL http://aima.cs.berkeley.edu/.

Edward J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978. DOI: 10.1287/opre.26.2.282. URL https://doi.org/10.1287/opre.26.2.282.