Probabilistic Constrained Reinforcement Learning with Formal Interpretability

Yanran Wang [1] **Qiuchen Qian** [1] **David Boyle** [1]

Abstract

Reinforcement learning can provide effective reasoning for sequential decision-making problems with variable dynamics. Such reasoning in practical implementation, however, poses a persistent challenge in interpreting the reward function and the corresponding optimal policy. Consequently, representing sequential decision-making problems as probabilistic inference can have considerable value, as, in principle, the inference offers diverse and powerful mathematical tools to infer the stochastic dynamics whilst suggesting a probabilistic interpretation of policy optimization. In this study, we propose a novel Adaptive Wasserstein Variational Optimization, namely AWaVO, to tackle these interpretability challenges. Our approach uses formal methods to achieve the interpretability for convergence guarantee, training transparency, and intrinsic decision-interpretation. To demonstrate its practicality, we showcase guaranteed interpretability with a global convergence rate $\Theta(1/\sqrt{T})$ in simulation and in practical quadrotor tasks. In comparison with state-of-the-art benchmarks, including TRPO-IPO, PCPO, and CRPO, we empirically verify that AWaVO offers a reasonable trade-off between high performance and sufficient interpretability.

1. Introduction

Sequential decision-making problems can be represented using Reinforcement Learning (RL) or optimal control technologies to efficiently determine optimal policies or control strategies in the presence of uncertainties (Levine, 2018). Nevertheless, such reasoning poses an ongoing challenge to create a convincing interpretation of sequential decision-making and its corresponding optimal policies (Devidze et al., 2021; Levine, 2022). This challenge in comprehension poses a significant barrier to real-world RL's implementation in safety-critical domains, such as advanced manufacturing (Napoleone et al., 2020), autonomous navigation (Fernandez-Llorca & Gómez, 2023) and financial trading (McNamara, 2016).

Key Challenges. The challenges surrounding interpretability in the context of RL can be conceptualized through three distinct phases: **a. Guarantee of convergence** ensures that an RL framework converges towards an optimal policy, e.g., in an asymptotic manner. **b. Transparency in training convergence** emphasizes the identification of the underlying processes that an RL algorithm employs to reach convergence during its training. An instance is the convergence rate, where, based on a given number of training iterations, the rate enables the prediction of the expected level of convergence with a certain degree of confidence. **c. Interpretation of decisions** seeks to explain *why* these specific sequential decisions were made within a given state and environment. Specifically, this interpretation involves clarifying the quantitative impact of latent factors on these sequential decisions. Moreover, due to legal mandates in industries, this facet of interpretation is of even greater significance, particularly in ensuring the trustworthiness of self-driving vehicles (Fernández Llorca & Gómez, 2021; Fernandez-Llorca & Gómez, 2023), aerospace engineering (Brat, 2021; Torens et al., 2022), and high-frequency trading (McNamara, 2016).

One widely adopted approach to achieving model interpretability involves the use of post-hoc explanation methods. These methods provide retrospective rationales for model predictions, often through the creation of saliency maps or exemplars, as discussed in previous research (Lipton, 2018; Kenny et al., 2021). Despite their popularity, these approaches may produce incomplete or inaccurate explanations (Slack et al., 2020; Zhou et al., 2022). In response to these limitations, recent studies have shifted their focus towards intrinsic interpretability (Rudin, 2019; Kenny et al., 2022). These methods, however, face challenges in providing a transparent and comprehensive view in the decision-making process. While they present decision explanations that are user-friendly, as demonstrated in (Kenny et al., 2022), corresponding to **Key Challenge c**, there is no guaranteed transparency, as highlighted in **Key Challenges**

[1]Systems and Algorithms Laboratory, Imperial College London, South Kensington, London. Correspondence to: Yanran Wang <yanran.wang.20@imperial.ac.uk>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

a and **b**. Establishing such transparency is crucial and serves as a prerequisite for underpinning user trust and predicting the system's capabilities.

To our best knowledge, we present the first intrinsically interpretable constrained RL framework through the lens of probabilistic inference. Specifically, we reframe constrained RL as Wasserstein variational optimization, leveraging an enhanced foundational inference framework known as augmented Probabilistic Graphical Models (PGMs). This is illustrated in Figure 1. Our proposed **A**daptive Sliced **Wa**sserstein **V**ariational **O**ptimization (AWaVO), as elaborated in Figure 2, consists of two primary steps:

a. *Optimality-Rectified Policy Optimization using Distributional Representation (ORPO-DR)*: ORPO is conducted to dynamically adapt to uncertainties (Algorithm 1, Section 4.2). More importantly, Distributional Representation (DR) provides an entire distribution of the action-value function, contributing to heightened transparency in the convergence process (as outlined in **Theorem 5.4** and **Theorem 5.5**). Consequently, this efficiently tackles a significant portion of the deficiencies outlined in **Key Challenges a** and **b**;

b. *Wasserstein Variational Inference (WVI)*: as detailed in Section 4.1, WVI is subsequently performed to achieve the probabilistic interpretation of decisions, thereby tackling **Key Challenge c**.

Our contributions can be summarized as follows:

- **Adaptive Generalized Sliced Wasserstein Distance**, referred to as A-GSWD, incorporates the Sliced Wasserstein Distance (SWD) along with adaptive Radon transforms to handle dynamic uncertainties. Specifically, the proposed A-GSWD adaptively determines the hypersurfaces' slicing directions to enhance the precision of distribution distance computation;
- **Adaptive Sliced Wasserstein Variational Optimization**, abbreviated as AWaVO, employs inference to reformulate the problem of sequential decision-making. To tackle all Key Challenges, AWaVO leverages ORPO-DR to enhance the transparency of convergence under dynamic uncertainties. Additionally, WVI is employed to provide probabilistic decision-interpretation;
- **Formal methods for interpretation** are employed to demonstrate theoretical comprehension on the metric judgment of A-GSWD, transparency of training convergence, and probabilistic interpretation of decisions. The practical hardware implementation and additional demonstrations are showcased in a video [1].

[1] https://github.com/Alex-yanranwang/AWaVO

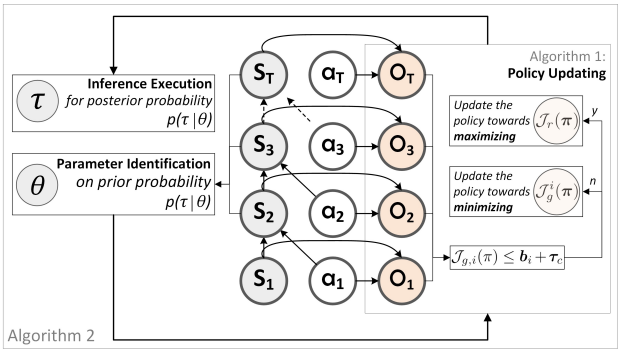


Figure 1. A new graphical model for constrained RL: refer to Algorithm 2 for a comprehensive overview of *(i) Parameter Identification*, *(ii) Policy Updating* and *(iii) Inference Execution*.

2. Related Work

Reinforcement Learning as Inference. The relationship between sequential decision-making and probabilistic inference has been explored extensively in recent years (Levine, 2018; Okada & Taniguchi, 2020; Liu et al., 2022). Despite variations in terminology, the core inference frameworks remain consistent, namely, PGMs (Koller & Friedman, 2009). While substantial research exists on learning and inference techniques within PGMs (Levine, 2018), the direct connection between RL (or control) and probabilistic inference is not immediately apparent. (Welch et al., 1995) establishes that control and inference are dual perspectives of the same problem. This connection offers novel insights and enhanced understanding within control problems by leveraging mathematical tools of inference (Toussaint & Storkey, 2006; Kappen et al., 2012). Moreover, the study on 'RL as inference' represents another prominent trend. Specifically, (Levine, 2018) demonstrates that RL is equivalent to probabilistic inference under dynamics. (Chua et al., 2018; Okada & Taniguchi, 2020) approach dynamics modeling by employing Bayesian inference optimization. Furthermore, (O'Donoghue et al., 2020) revisits the formalization of 'RL as inference' and demonstrates that with a slight algorithmic modification, this approximation can perform well even in problems where it initially performs poorly. In this study, we formalize constrained RL as Wasserstein variational optimization to achieve decision-interpretations.

Optimal Transport Theory. Forming effective metrics between two probability measures is a fundamental challenge in machine learning and statistics communities. The optimal transport theory, particularly the Wasserstein distance, has garnered significant attention across various domains (Solomon et al., 2014; Kolouri et al., 2017; Schmitz et al., 2018; Wang & Boyle, 2023) due to its accuracy, robustness, and stable optimization. Nevertheless, due to its computational demand on high-dimensional data, recent

advancements emphasize computational efficiency through differentiable optimization (Peyré et al., 2017). Among these methods, Sinkhorn distance (Cuturi, 2013; Altschuler et al., 2017) introduces entropy regularization to smoothen the convex regularization. Another notable approach involves slicing or linear projection (Ng, 2005), i.e., Sliced Wasserstein Distance (SWD) (Bonneel et al., 2015), which leverages the measures' Radon transform for efficient dimensionality reduction. Then, variants of SWD, such as Generalized SWD (GSWD) (Kolouri et al., 2019), improve projection efficiency. These advancements contribute to the efficiency in optimal-transport-based metrics. However, they suffer from reduced accuracy as SWD only slices distributions using linear hyperplanes, which may fail to capture the complex structures of data distributions. To overcome the accuracy limitation, Augmented SWD (ASWD) (Chen et al., 2021) projects onto flexible nonlinear hypersurfaces, enabling the capture of intricate data distribution structures. Building upon the ASWD framework, we introduce an adaptive variant called A-GSWD which leverages the projection onto nonlinear hypersurfaces and combines it with ORPO-DR to achieve adaptivity. This adaptive approach enhances the efficiency and accuracy of Wasserstein distance computation, improving upon the limitations of previous methods.

3. Problem Formulation and Preliminaries

Sequential Decision-making as Probabilistic Inference. A sequential decision-making problem, formalized as a standard RL or control problem, can be seen as an inference problem (Levine, 2018):

$$p(\tau|\mathcal{O}_{0:T-1}=1) \propto \int \underbrace{\prod_{t=0}^{T-1} p(\mathcal{O}_t=1|\boldsymbol{s}_t,\boldsymbol{a}_t)}_{:=p(\mathcal{O}|\tau)} \cdot \underbrace{p(\boldsymbol{s}_0)\left\{\prod_{t=0}^{T-1} p(\boldsymbol{a}_t|\boldsymbol{s}_t,\theta)p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)\right\}}_{\overset{\text{Markov property}}{:=} p(\tau|\theta)} \cdot \underbrace{p(\theta|D)}_{:=p_D(\theta)} \mathrm{d}\theta \tag{1}$$

where $\boldsymbol{s}_t, \boldsymbol{a}_t, \tau = \{(\boldsymbol{s}_t, \boldsymbol{a}_t)\}_{t=0}^{T-1}$ and $D = \{(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})\}$ are states, actions, a trajectory and observed training dataset. $\mathcal{O}_t = \{\mathcal{O}_{r,t}, \mathcal{O}_{g_i,t}\} \in \{0,1\}$ represents an additional binary variable of the optimality for $(\boldsymbol{s}_t, \boldsymbol{a}_t)$ in PGM (Levine, 2018; Okada & Taniguchi, 2020). $\mathcal{O}_{r,t}=1$ and $\mathcal{O}_{g_i,t}=1$ signify that the trajectory τ is optimized and compliant with the constraints, respectively.

In Equation 1, we can deconstruct the various components: **the probability** $p(\boldsymbol{a}_t|\boldsymbol{s}_t,\theta)$ signifies the stationary policy π which maps one state $\boldsymbol{s}_t$ to one action $\boldsymbol{a}_t$, where $\boldsymbol{a}_t \sim p(\cdot|\boldsymbol{s}_t,\theta) = \pi(\cdot|\boldsymbol{s}_t)$ at each time step t; **the transition probability** $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$ represents state transitions (also known as forward-dynamics models), where $\boldsymbol{s}_{t+1} \sim p(\cdot|\boldsymbol{s}_t,\boldsymbol{a}_t)$ (Chua et al., 2018) at each time step t; **the prior probability** $p_D(\theta)$ is derived from the posterior probability $p(\theta|D)$, where the parameter θ is inferred from the training dataset D; and lastly, **the optimality likelihood** $p(\mathcal{O}|\tau)$ is defined in relation to the expected reward and utility formulation of several trajectories, expressed as $\mathcal{F}_r \cdot p(\mathcal{O}_r|\tau) := \widetilde{r}(\tau)$ and $\mathcal{F}_g \cdot p(\mathcal{O}_{g_i}|\tau) := \widetilde{g}_i(\tau)$, where the operator family $\mathcal{F} = \{\mathcal{F}_r, \mathcal{F}_g\}$ and the optimality family $\mathcal{O} = \{\mathcal{O}_r, \mathcal{O}_{g_i}\}$ establish this relationship. In **Section 4.1** and **Section 5**, we offer theoretical understanding to illustrate how such specific definitions influence the RL's global convergence.

Constrained Reinforcement Learning as Probabilistic Graphical Models. Specifically, we consider a Constrained Markov Decision Process (CMDP) (Altman, 1999), a formal framework for constrained RL, which is formulated as a discounted Markov decision process with additional constrained objectives, i.e., a tuple $\langle S, A, P, R, G, \gamma\rangle$: S is a finite set of states $\{\boldsymbol{s}\}$; A is a finite set of actions $\{\boldsymbol{a}\}$; $P: S \times A \to S$ is a finite set of transition probabilities $\{p(\boldsymbol{s'}|\boldsymbol{s},\boldsymbol{a})\}$; $R: S \times A \times S \to \mathbb{R}$ is a finite set of bounded immediate rewards $\{r\}$; $G: S \times A \times S \to \mathbb{R}$ comprises a finite collection of unity functions $\{g\}$, where, upon satisfying the expected constraints g_i, the unity-optimality variable is specified as $\mathcal{O}_{g_i}=1$; and $\gamma \in [0,1]$ is the discount rate. A CMDP is presented as:

$$\max_{\pi} \mathcal{J}_r(\pi), \quad \text{s.t.} \quad \mathcal{J}_{g,i}(\pi) \le \boldsymbol{b}_i + \boldsymbol{\tau}_c, \quad i=1,...,n \tag{2}$$

where $\mathcal{J}_r(\pi) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t,\boldsymbol{a}_t)|\pi, \boldsymbol{s}_0 = s]$ and $\mathcal{J}_{g,i}(\pi) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g_i(\boldsymbol{s}_t,\boldsymbol{a}_t)|\pi, \boldsymbol{s}_0 = s]$ are the value function associated with the immediate reward r and the utility g, respectively; b_i is a fixed limit for the i-th constraint; and $\boldsymbol{\tau}_c$ is the tolerance. Figure 1 shows how constrained RL can be viewed as a novel variation of PGMs. A complete table of nomenclature is provided in **Appendix A** for convenient reference.

4. Method: Adaptive Sliced Wasserstein Variational Optimization (AWaVO)

In this section, we present AWaVO's two primary submodules: WVI and ORPO-DR. The detailed algorithm is outlined in Algorithm 2, and the overarching algorithmic structure is depicted in Figure 2.

4.1. WVI: Wasserstein Variational Inference

Variational Inference for Dynamic Uncertainties. Given uncertainties in a dynamics model, it is reasonable to assume that the optimal trajectories $\{\tau\}$ are uncertain. To infer optimal policies under uncertainties, let us consider a variational

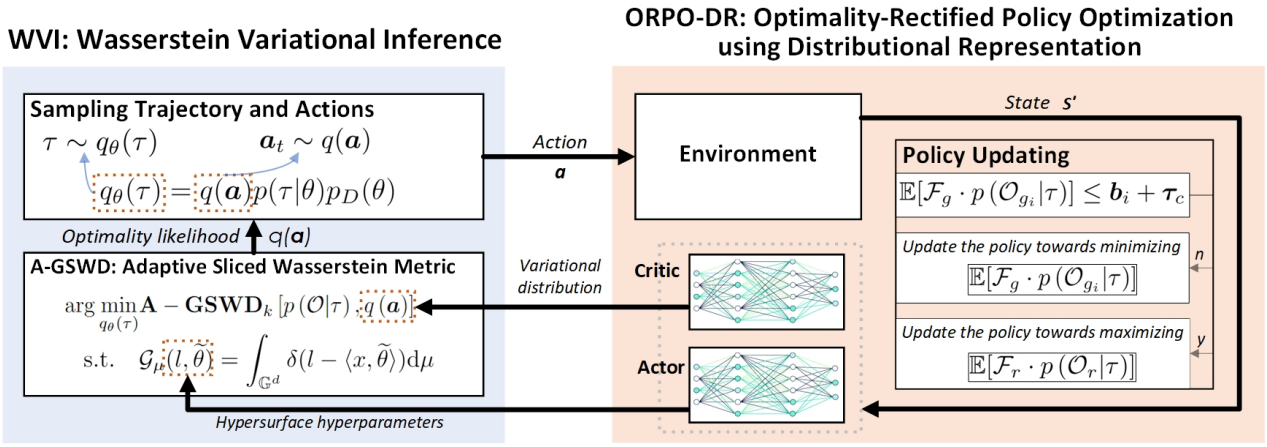


Figure 2. The algorithmic framework of AWaVO. We reform constrained RL as a Wasserstein variational optimization setup, consisting of two primary submodules: ORPO-DR and WVI (**Section 4**).

inference: $D(q_\theta(\tau)||p(\tau|\mathcal{O}))$, where, for simplicity, we use $p(\tau|\mathcal{O})$ to represent $p(\tau|\mathcal{O}_{t:T}=1)$; and $D(\cdot||\cdot)$ represents a distance metric between two probabilities. Building upon Equation 1, the variational distribution $q_\theta(\tau)$ is constructed as $q_\theta(\tau) = q(\boldsymbol{a})p(\tau|\theta)p_D(\theta)$. The construction suggests an assumption that the state transitions are controlled by $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$. According to Equation 1, we formulate the posterior as $p(\tau|\mathcal{O}) \propto p(\mathcal{O}|\tau)p(\tau|\theta)p_D(\theta)$ (see **Assumption 2** for our implementation details).

While Kullback-Leibler (KL) divergence is widely used in conventional variational inference, its application in certain practical implementations can be risky due to its limitations, including asymmetry and infinity, arising when there are unequal supports. In this section, we extend the Wasserstein distance into the variational inference, and present the derivation of how we transform the GSWD between the two posteriors to the optimality likelihood $p(\mathcal{O}|\tau)$ and its approximation $q(\boldsymbol{a})$.

Adaptive Generalized Sliced Wasserstein Distance. GSWD has exhibited high projection efficiency in previous studies (Kolouri et al., 2019; Chen et al., 2021) (please refer to **Appendix B** for a comprehensive background and definition of Wasserstein distance). However, the identification of the hypersurface hyperparameters, such as l and $\widetilde{\theta}$, remains to be a challenge. The selection of these parameters, specifying the hypersurface along with its slicing direction, is generally a task-specific problem and requires prior knowledge or domain expertise. We now present a new adaptive sliced Wasserstein distance, called A-GSWD, that integrates GSWD with ORPO-DR, an adaptive process for determining the parameters of a hypersurface.

Definition 4.1. Given SWD and GSWD (defined in **Appendix B.1** and **Appendix B.2**, respectively), we define A-GSWD by utilizing ORPO-DR for the adaptive slicing:

$$\mathbf{A}-\mathbf{GSWD}_\mathbf{k}(\mu,\nu) = \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_\mu\left(\cdot,\widetilde{\theta}\right), \mathcal{A}_\nu\left(\cdot,\widetilde{\theta}\right)\right)\mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}}$$

where $\mu, \nu \in P_k(\mathcal{X})$ are two measures of probability distributions over the space $\mathcal{X}$ (see **Appendix B.1** for details). $l \in \mathbb{R}$ and $\widetilde{\theta} \in \mathcal{R}_{\widetilde{\theta}}$ represent the parameters of hypersurfaces, both of which are the outputs from actor networks in ORPO-DR (see Figure 2 for details). The Adaptive Generalized Radon Transforms (AGRT) $\mathcal{A}$ is used as a push-forward operator $\mathcal{A}_\mu$, defined by $\mathcal{A}_\mu(l,\widetilde{\theta}) = \int_{\mathbb{G}^d} \delta(l - \alpha(x,\widetilde{\theta}))\mathrm{d}\mu$, where $\alpha(x,\widetilde{\theta})$ is *a defining function* satisfying the conditions **H.1-H.4** in (Kolouri et al., 2019). $\mathcal{R}_{\widetilde{\theta}} \subset \mathbb{R}^d$ is a compact set of all feasible parameters $\widetilde{\theta}$, where $\mathcal{R}_{\widetilde{\theta}} = \mathbb{S}^{d-1}$ for $\alpha(\cdot,\widetilde{\theta}) = \langle \cdot, \widetilde{\theta}\rangle$.

Although the proposed adaptive slicing method, i.e., A-GSWD, improves the efficiency and accuracy of the Wasserstein distance computation, its demonstration on a valid metric guarantee remains a problem (Kolouri et al., 2019). In **Section 5**, we prove that the proposed A-GSWD is a true metric that satisfies non-negativity, symmetry, the triangle inequality and $\mathbf{A}-\mathbf{GSWD}_k(\mu,\mu) = 0$, respectively.

We then employ A-GSWD to address the variational inference, i.e., minimizing the distance $D(q_\theta(\tau)||p(\tau|\mathcal{O})) = \mathbf{A}-\mathbf{GSWD}_k(q_\theta(\tau), p(\tau|\mathcal{O}))$ between the variational distribution $q_\theta(\tau)$ and the posterior distribution $p(\tau|\mathcal{O})$. Subsequently, the variational inference can be reformulated to the minimization problem, as shown in WVI of Figure 2: $\arg\min_{q_\theta(\tau)} \mathbf{A}-\mathbf{GSWD}_k\left(q\left(\boldsymbol{a}\right), p\left(\mathcal{O}|\tau\right)\right)$, where $p(\mathcal{O}|\tau)$ represents the optimality likelihood, and the detailed derivation is in **Appendix D.1**.

4.2. ORPO-DR: Optimality-Rectified Policy Optimization using Distributional Representation

The current policy optimization for constrained RL can be classified into two categories: primal-dual and primal approaches (Xu et al., 2021). The former, transforming the constrained problem into an unconstrained one, are most commonly used although sensitive to Lagrange multipliers and other hyperparameters, such as the learning rate. On the other hand, the latter (i.e., primal approaches) require less hyperparameter tuning but have received less attention in terms of convergence demonstration compared to the primal-dual approaches.

Policy Optimization combining Optimality Likelihood. Based on **Section 4.1**, a constrained RL problem, as outlined in Equation 2, can be iteratively substituted and resolved as:

$$\begin{cases} \arg\max_{q(\boldsymbol{a})} \mathbb{E}[\mathcal{F}_r \cdot p\left(\mathcal{O}_r|\tau\right)],\ \mathbb{E}[\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)] \leq \boldsymbol{b}_i + \boldsymbol{\tau}_c \\ \arg\min_{q(\boldsymbol{a})} \mathbb{E}[\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)], \quad otherwise \end{cases}$$

where we recall that $\{\mathcal{F}_r, \mathcal{F}_g\}$ are two operators defined as $\mathcal{F}_r \cdot p\left(\mathcal{O}_r|\tau\right) := \widetilde{r}(\tau)$ and $\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right) := \widetilde{g}_i(\tau)$, respectively. Furthermore, we can calculate the accumulated reward and utility function: $\widetilde{r}(\tau) = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t)]$ and $\widetilde{g}_i(\tau) = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t g_i(\boldsymbol{s}_t, \boldsymbol{a}_t)]$. Consequently, we obtain $\mathcal{J}_r(\pi) = \mathbb{E}[\widetilde{r}(\tau)]$ and $\mathcal{J}_{g,i}(\pi) = \mathbb{E}[\widetilde{g}_i(\tau)]$ if $T = \infty$.

If we only define $\mathcal{F}_r \propto \log[\cdot]$, it becomes equivalent to the formulation used in (Levine, 2018; Okada & Taniguchi, 2020; 2018). In this case, we can retrieve an optimization process that resembles Model Predictive Path Integral (MPPI) (Okada & Taniguchi, 2018). The design of reward functions in the traditional RL is typically based on task-specific heuristics which is often considered as much an art as science. We will present such interpretation in **Section 5** to show how the reward operator family $\mathcal{F}$ acts on convergence, as well as a more rigorous approach to ensure guaranteed global convergence rate during the training process. Additionally, in **Section 6**, we empirically verify these theoretical guarantees.

Policy Updating. As shown in Algorithm 1, we first update the policy towards either maximizing $\mathbb{E}[\mathcal{F}_r \cdot p\left(\mathcal{O}_r|\tau\right)]$ or minimizing $\mathbb{E}[\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)]$ by using the distributional representation (introduced in **Appendix C**), where the gradient of actor and critic network, denoted as δ_{θ^μ} and δ_{θ^Q}, are defined in Equation 8 in **Appendix C**. Then, as shown in ORPO-DR of Figure 2, the actor network outputs the parameters to dynamically determine the hypersurfaces and the corresponding slicing directions; and the critic network provides an entire state-action distribution, which is directly utilized as the variational distribution of the optimality likelihood $q(\boldsymbol{a})$ in A-GSWD, as shown in Figure 2.

Algorithm 1 ORPO-DR: Optimality-Rectified Policy Optimization using Distributional Representation

Input: $\boldsymbol{s}_k$, $\boldsymbol{s}_{k+1}$, $\boldsymbol{\tau}_c$, θ^μ, θ^Q
Output: updated θ^μ, θ^Q
Constraint Estimation:
$\mathcal{J}_{g,i}(\pi_\theta(\boldsymbol{s}, \boldsymbol{a})) = \mathbb{E}[\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)],\ \forall i \in [1, p]$
Policy Improvement:
if $\mathcal{J}_{g,i}(\boldsymbol{\pi}) \leq \boldsymbol{b}_i + \boldsymbol{\tau}_c, \forall i \in [1, p]$ **then**
  update the policy towards maximizing $\mathbb{E}[\mathcal{F}_r \cdot p\left(\mathcal{O}_r|\tau\right)]$:
  $\theta^\mu \leftarrow \theta^\mu + l_\mu \delta_{\theta^\mu}$, and $\theta^Q \leftarrow \theta^Q + l_\theta \delta_{\theta^Q}$
else
  update the policy towards minimizing $\mathbb{E}[\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)]$:
  $\theta^\mu \leftarrow \theta^\mu - l_\mu \nabla_{\theta^\mu} \widetilde{g}_i(\tau)$, and $\theta^Q \leftarrow \theta^Q - l_Q \nabla_{\theta^Q} \widetilde{g}_i(\tau)$
end if

Algorithm 2 AWaVO: Adaptive Sliced Wasserstein Variational Optimization

Input: $\boldsymbol{s}_k$, $\boldsymbol{s}_{k+1}$, θ^μ, θ^Q
Output: $\boldsymbol{a}_k$
Initialize:
$\boldsymbol{\theta} = [\theta^\mu, \theta^Q]$: the parameters of actor and critic network
repeat
  for $t = 0, 1, 2, ..., T-1$ **do**
    Parameter Identification: achieve $p_D(\theta)$ by doing inference of the posteriors $p(\theta|D)$ (Section 3)
    Policy Updating: $\{\theta^\mu, \theta^Q\} \leftarrow$ Exec. Algorithm 1 ($\boldsymbol{s}_k$, $\boldsymbol{s}_{k+1}$, $\boldsymbol{\tau}_c$, θ^μ, θ^Q)
    Inference Execution: do inference of the posterior probability, as described in Section 4.1:
    $p(\tau|\mathcal{O}_{t:T}) \leftarrow \arg\min_{q_\theta(\tau)} \mathbf{A} - \mathbf{GSWD}_k\left(q\left(\boldsymbol{a}\right), p\left(\mathcal{O}|\tau\right)\right)$
    sample actions $\boldsymbol{a}_k \leftarrow p(\tau|\mathcal{O}_{t:T})$, execute $\boldsymbol{a}_k$, and observe $\boldsymbol{s}_{k+1}$
  end for
until convergence

5. Formal Methods for Interpretability

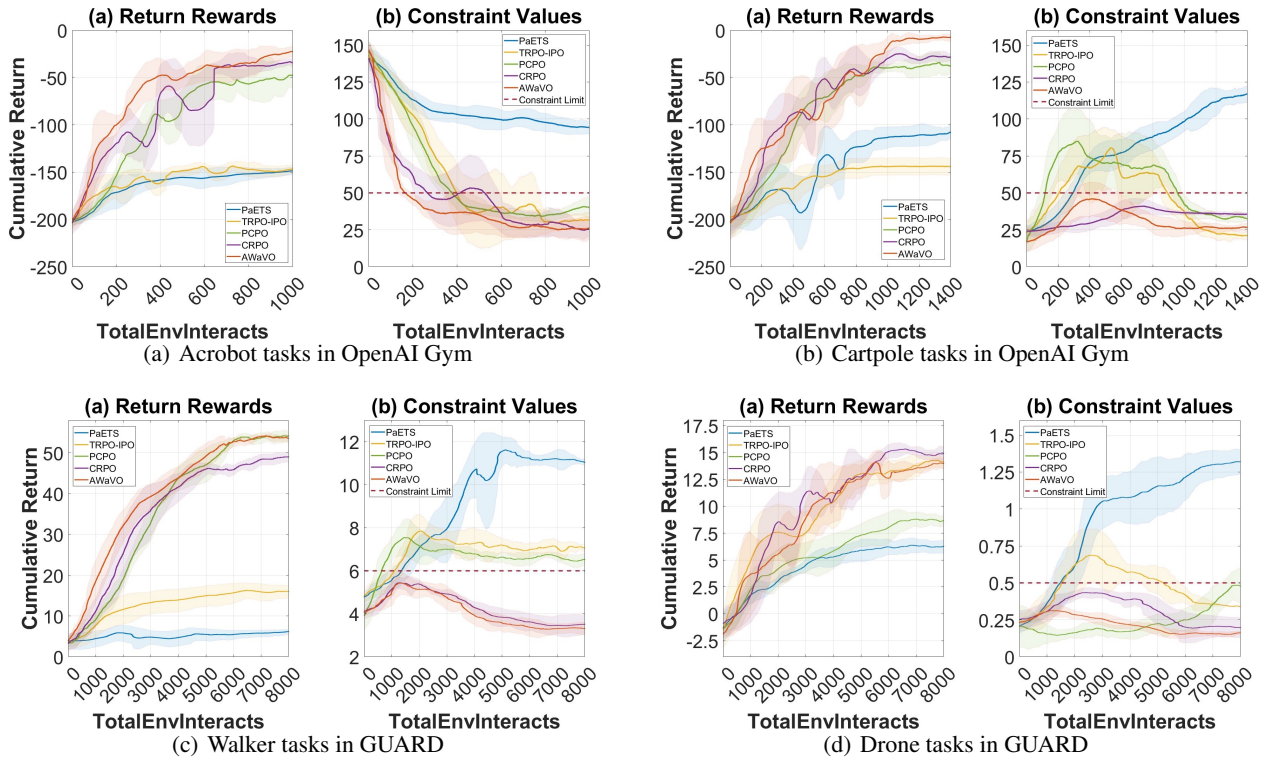
Proposition 5.1. *(Pseudo-metric): Given two probability measures $\mu, \nu \in P_k(\mathcal{X})$ and a mapping $\alpha : \mathcal{X} \to \mathcal{R}_{\widetilde{\theta}}$, the adaptive slicing A-GSWD, defined in **Definition 4.1**, with order k in the range $[1, \infty)$, is a pseudo-metric that satisfies non-negativity, symmetry, the triangle inequality and $\mathbf{A} - \mathbf{GSWD}_k(\mu, \mu) = 0$. See **Appendix D.3** for Proof.*

Remark 5.2. The adaptive slicing A-GSWD, with order $k \in [1, \infty)$, is a true metric *if and only if* the AGRT $\mathcal{A}$, defined in **Definition 4.1**, is an injective mapping.

We make the following three assumptions.

Assumption 1. *We define the function:* $p(\tau|\mathcal{O}) = p(\mathcal{O}|\tau)p(\tau|\theta)p_D(\theta)$.

Assumption 2. *Let $\Psi(s, a)$ be a feature vector, and χ_π be a stationary distribution in CMDP: $(s, a) \sim \chi_\pi$. There exists a constant $\hat{C}_0$ such that for any $\varrho \geq 0$, it holds that*



(a) Acrobot tasks in OpenAI Gym (b) Cartpole tasks in OpenAI Gym

(c) Walker tasks in GUARD (d) Drone tasks in GUARD

Figure 3. Performance comparison over 10 seeds. CRPO and AWaVO outperform PaETS, with a trade-off highlighted: although PaETS offers probabilistic interpretation with Bayesian networks, its convergence is generally unstable. Our proposed AWaVO achieves a better balance between high performance and interpretability. In contrast to two other constrained RL algorithms, i.e., TRPO-IPO and PCPO, we observe an interesting result: PCPO performs better in tasks like Acrobot, Cartpole, and Walker, while TRPO-IPO outperforms PCPO in the more complex drone tasks (Figure 3(d)). Further, in Figure 5, we will explore more complex real-world tasks using an aerial robot.

$p(\left|x^\intercal \Psi(s,a)\right| \leq \varrho) \leq \hat{C}_0 \cdot \varrho$, *where* $x \in \mathbb{R}^d$.

Assumption 3. *We define the family of functions:*

$$\mathcal{B}_{R,\infty} = f((s,a);\theta_q) = f((s,a);\theta_{q,0}) + \int \mathbf{1}({\theta_q}^\intercal \Psi(s,a) > 0) \cdot \omega(\theta_q)^\intercal \Psi(s,a) \mathrm{d}\varphi(\theta_q)$$

where $f((s,a);\theta_q)$ *is an* H*-layer neural network corresponding to the initial parameter* $\theta_{q,0}$*. The weighted function* $\omega(\theta_q) : \mathbb{R}^d \to \mathbb{R}^d$ *satisfies* $\|\omega(\cdot)\|_\infty \leq C_R/\sqrt{d}$*, where* $C_R \in \mathbb{R}$ *denotes an upper bounded value and* $d \geq 2$*.* $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}$ *represents the density of the weight distribution. We assume* $Q^\pi \in \mathcal{B}_{R,\infty}$*, for all* π*.*

Assumption 1 specifies that in our implementation, the term $p(\tau|\mathcal{O})$ is explicitly formulated as $p(\mathcal{O}|\tau)p(\tau|\theta)p_D(\theta)$. **Assumption 2** implies that the probability density of the distribution $\Psi(s,a)$ is uniformly upper-bounded over the unit sphere, a condition achievable in most ergodic Markov chains (Mitrophanov, 2005; Xu et al., 2021). **Assumption 3** implies a mild and broadly applicable regularity condition on Q^π, as $\mathcal{B}_{R,\infty}$ can be interpreted as a function class with infinite width neural networks, thus representing a sufficiently general set of functions.

To establish a link between the reward operator family $\mathcal{F}$ and the global convergence of ORPO-DR, here, we introduce **Conditions 5.3** and then present **Theorem 5.4** *(Global Convergence).*

Conditions 5.3. *The reward operator family* $\mathcal{F} = \{\mathcal{F}_r, \mathcal{F}_g\}$ *satisfies that: (i)* $\mathcal{F}_r$ *is monotonically increasing and continuously defined on* $(0,1]$*, and the range covers* $[r_{\min}, r_{\max}]$*; and (ii)* $\mathcal{F}_g$ *is monotonically decreasing and continuously defined on* $(0,1]$*, and the range covers* $[r_{\min}, r_{\max}]$*.*

Theorem 5.4. *(Global Convergence): Given the policy in the* i*-th policy improvement* $\boldsymbol{\pi^i}$*,* $\boldsymbol{\pi^i} \to \boldsymbol{\pi}^*$ *and* $i \to \infty$*, suppose Assumption 1 holds, there exists* $Q^{\boldsymbol{\pi}^*}(s,a) \geq Q^{\boldsymbol{\pi}^i}(s,a)$ *if and only if the reward operator family* $\mathcal{F}$ *satisfies the both* ***Conditions 5.3****. See* ***Appendix D.3*** *for Proof.*

Subsequently, we present a more rigorous comprehension of how $\mathcal{F}$ precisely influences the convergence rate. To our best knowledge, this study represents the first attempt to develop an intrinsic interpretation of how the reward function design influences convergence within the RL community.

Theorem 5.5. *(Global Convergence Rate): Let* m *and* H *be the width and layers of a neural network,* $K_{td} = (1 -$

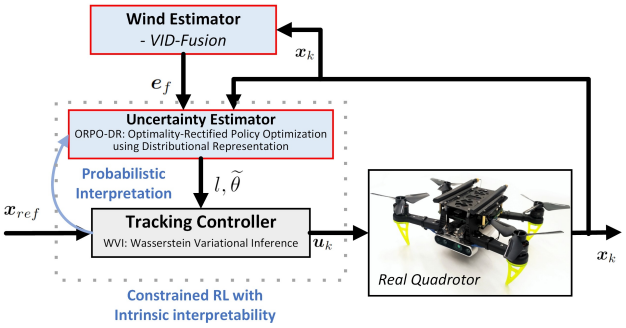


Figure 4. We use our AWaVO as the tracking controller for a quadrotor, where ORPO-DR is employed as the uncertainty estimator, and WVI using A-GSWD is leveraged as the controller.

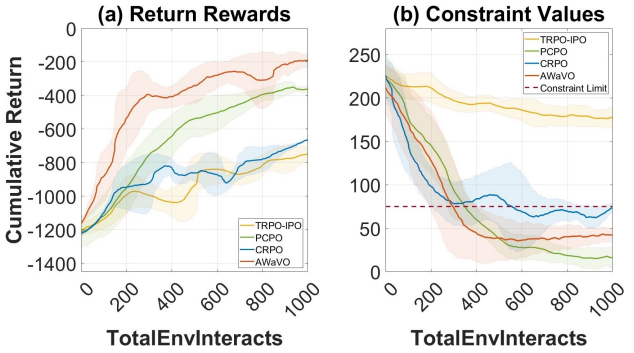


Figure 5. Performance comparison in a real quadrotor: our AWaVO slightly outperforms the constrained RL approach, i.e., PCPO, whilst achieving interpretability in Figure 7.

$\gamma)^{-\frac{3}{2}} m^{\frac{H}{2}}$ *be the iterations required for convergence of the distributional Temporal Difference (TD) learning (defined in Equation 10),* $l_Q = \frac{1}{\sqrt{T}}$ *be the policy update (in Line 4 of Algorithm 1) and* $\boldsymbol{\tau}_c = \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) + \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}})$ *be the tolerance (in Line 3 of Algorithm 1). Suppose Assumptions 1-3 hold. There exists a global convergence rate of* $\Theta(1/\sqrt{T})$*, and a sublinear rate of* $\Theta(1/\sqrt{T})$ *if the constraints are violated with an error of* $\Theta(1/m^{\frac{H}{4}})$*, with probability of at least* $1-\delta$*. This holds if and only if the reward operator family* $\mathcal{F}$ *satisfies* ***Conditions 5.3****. See* ***Appendix D.3*** *for Proof.*

Probabilistic interpretation on sequential decisions. We now quantitatively establish the relationships between latent factors, such as disturbances, that possibly influence decision-making and the sequential decisions, namely trajectories, by providing a probabilistic interpretation. Referring to the abbreviation presented in Equation 1, we reform it as: $p(\tau|D) = p(\mathcal{O}|\tau) \cdot p(\boldsymbol{s}, \boldsymbol{a}|\theta) \cdot p_D(\theta)$. Then, the latent factors are denoted by $L = \{L_i\}_{i=0}^{M-1}$, where M represents the total number of defined factors. By applying the chain rule to the posterior probability $p(\tau|D)$, we have $\{p(\tau|L_i)\}_{i=0}^{M} = \left\{\frac{p(\tau|D)}{p(L_i|D)}\right\}_{i=0}^{M}$, where the equation provides a decomposition of the joint posterior probability $p(\tau|D)$ into conditional probabilities that involve individual factors L_i. This decomposition is not only notable for its theoretical simplicity but facilitates a practical probabilistic understanding of how each factor influences policy in real-world safety-critical scenarios, such as robot autonomy. In **Section 6**, we showcase numerical examples to illustrate such probabilistic interpretation.

6. Experiments

In this section, we explain our empirical assessments of AWaVO's performance in simulated platforms and real-world robot tasks. Initially, we perform tasks with multiple constraints in OpenAI Gym framework (Brockman et al., 2016). Then we showcase AWaVO's practicality through real quadrotor Flight Tasks (FTs) to provide a more comprehensive assessment of its performance. These evaluations serve a dual purpose: to validate AWaVO's performance; and, critically, to empirically demonstrate its quantitative interpretability. This interpretability includes confirming properties such as the guaranteed convergence rate as demonstrated in **Theorem 5.5** and the probabilistic decision interpretation discussed in **Section 5** within the context of sequential decision-making tasks.

Comparative Performance in Simulated Tasks. We conduct tasks with multiple constraints in OpenAI Gym (Brockman et al., 2016) and GUARD (Zhao et al., 2023) (a constrained RL benchmark): Acrobot, Cartpole, Walker and Drone. We use four appropriate constrained RL benchmarks: PaETS (Okada & Taniguchi, 2020), i.e., a Bayesian RL combining with variational inference, TRPO-IPO (Liu et al., 2020), i.e., an enhanced variant of TRPO-Lagrangian (Bohez et al., 2019), PCPO (Yang et al., 2020), i.e., an advanced variant of CPO (Achiam et al., 2017) and CRPO (Xu et al., 2021), i.e., a primal constrained RL approach.

The AWaVO parameter settings given in Table 2 of **Appendix E.1** are based on selected benchmarks, i.e., CRPO (Xu et al., 2021) and GUARD (Zhao et al., 2023). According to our proposed **Proposition 5.1** and the **Proposition 1** presented in (Kolouri et al., 2019), the defining function $\alpha(\cdot, \widetilde{\theta})$ can be defined as homogeneous polynomials, i.e., $\alpha(\cdot, \widetilde{\theta}) = \sum_{|\kappa|=m} \widetilde{\theta}_\kappa x^\kappa$, where the defining function α is injective if the degree of the polynomial m is odd. Thus we set $m = 3$ based on (Kolouri et al., 2019). The comprehensive task descriptions are available in **Appendix E.2**.

AWaVO's training depicted in Figure 3 initially corresponds to the benchmarks provided by CRPO (Xu et al., 2021) and GUARD (Zhao et al., 2023). The distinction in iterations lies

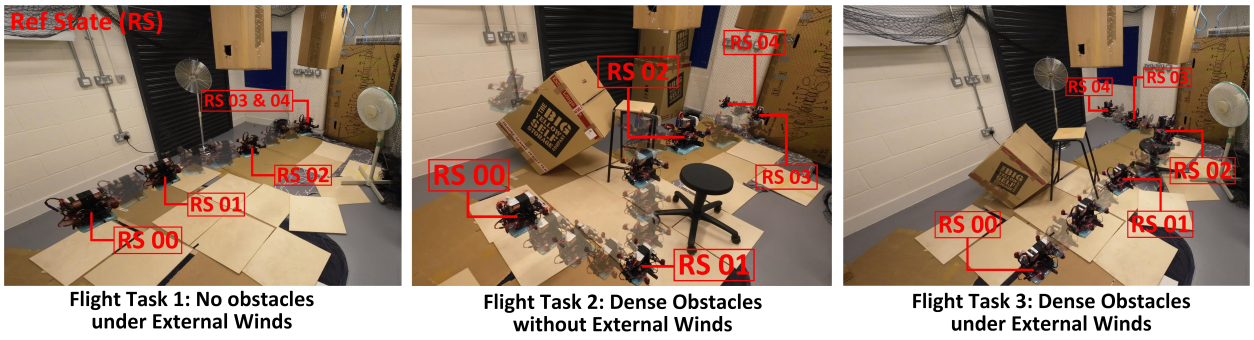


Figure 6. Real quadrotor Flight Tasks (FTs): *(i) FT 1* - tracking reference trajectories under external forces without obstacles; *(ii) FT 2* - tracking trajectories around dense obstacles; and *(iii) FT 3* - tracking trajectories under external forces around dense obstacles.

in showcasing the entire convergence process across four discrete tasks. We establish the constraint limit to facilitate a straightforward comparison of constraint convergence; see **Appendix E.3** for additional details. The tolerance is set as $\boldsymbol{\tau}_c = 0.5$, following CRPO (Xu et al., 2021). By analyzing the comparative training performances in Figure 3, we observe that the superiority of CRPO and AWaVO stems from their primal constrained RL nature, which involves training under constraints and ensuring global convergence rate. Although CRPO exhibits comparable or slightly better convergence performance than AWaVO, as evident in Figure 3(d), we place greater emphasis on two other aspects: training convergence under uncertainties and decision-making interpretation. In Figure 5 below, we provide comparative demonstrations in real robot tasks to showcase how AWaVO effectively balances a trade-off between performance and interpretability in a more complex sequential decision-making scenario.

Furthermore, we empirically verify the formal method **Theorem 5.5** *(Global Convergence Rate)* on the convergence rate, and conclude that, based on the average performance, the convergence rate of AWaVO is in the range of $\Theta(1/\sqrt{T}) < C_{rate} \leq \Theta(1/T^{1.2})$. According to the results shown in Figure 3(d), CRPO performs better than our AWaVO in the simulated drone task, with the absence of disturbances. Subsequently, in Figure 5 below, we further evaluate these approaches in a real-world physical environment characterized by varying uncertainties, leading to different outcomes. It is worth noting that our approach incorporates two optimizations for handling uncertainties: variational inference and policy updating. This combination reduces the frequency of policy updates whilst enhancing our ability to handle uncertainties. In the upcoming real robot task, we will introduce variable disturbances to demonstrate our capability to optimize policies under uncertain conditions.

Comparative Performance in Real-world Tasks. We demonstrate the effectiveness of AWaVO by practical implementation in real-world decision-making problems. The

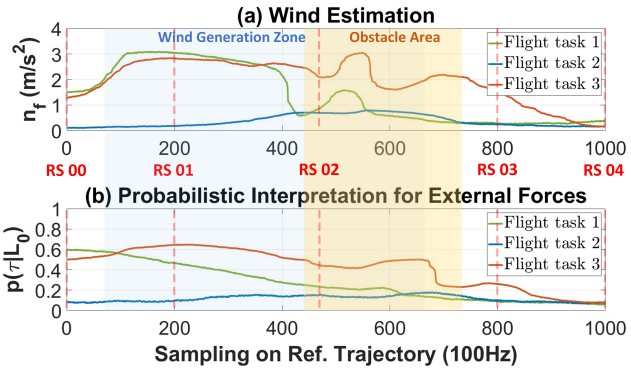


Figure 7. Probabilistic interpretation of decisions: the probability $p(\tau|L_0)$ reveals the degree to which the measurement of external forces (Ding et al., 2021), denoted as n_f, influences the decisions made by the quadrotor. For additional discussion on the case of 'RS 02', as an example instance, please refer to **Appendix E.4**.

quadrotor's tracking control system, shown in Figure 4, is an end-to-end learning-based framework. The technical specification of our quadrotor is shown in Table 3 of **Appendix E.1**. The training convergence is demonstrated in Figure 5. The aim of the real-world tasks (shown in Figure 6) is to track the reference effectively and accurately, where VID-Fusion (Ding et al., 2021) is used to measure external forces such as aerodynamic effects. To view the hardware experiments in action, please refer to the accompanying video demonstration at `https://github.com/Alex-yanranwang/AWaVO`.

Next, we illustrate the interpretation of sequential decisions, i.e., the actual control commands fed into the four motors. Leveraging the Intel RealSense D435i depth camera onboard, we can detect obstacles and estimate external forces. These latent factors, denoted as $L = L_0, L_1$, represent external forces and obstacles, respectively. The probability $p(\tau|L)$ reveals why the quadrotor makes these decisions and quantifies the extent to which factor L contributes to the sequential decisions, i.e., the real-time trajectory τ. Fig-

ure 7 presents a quantitative interpretation, i.e., $p(\tau|L_0)$, indicating the magnitude and evolution that the external force n_f impacts on the current control decisions. Please see **Appendix E.4** for further discussion on Figure 7.

Practically, this probabilistic interpretation represents significant progress in addressing a longstanding and challenging question: why do machine systems powered by Artificial Intelligence (AI) technologies make certain decisions, and what are the exact latent factors influencing those decisions? Such progress holds particular value for safety-critical industries like self-driving vehicles, aerospace engineering and high-frequency trading in financial services, particularly in cases where AI-based approaches exhibit erratic performance and thorough analysis is necessary.

7. Limitation

The primary limitation we encounter is ensuring the trustworthiness of the posterior probability generated by the critic network, which operates as a Bayesian network. Our ongoing efforts involve applying statistical methods to establish a specific confidence interval for the Bayesian network's outcomes. Additionally, we intend to study AWaVO's suitability for more real-world safety-critical applications. For scalability, please refer to **Appendix E.5**. This may serve as a starting point from which to formally analyze AWaVO's scalability and how the approach could extend to larger and more complex real-world domains. Another limitation is the lack of human-centered validation (although this would need to be taken on an *ad hoc* basis per deployment context), and considering policy or regulatory issues.

8. Conclusion

Enthusiasm towards the possible applications of constrained RL is growing worldwide. The insufficient ability to interpret agent actions and policy optimizations, however, poses a significant hurdle in deploying RL in safety-critical domains like advanced manufacturing and financial trading. Our primary motivation in introducing AWaVO, an intrinsically interpretable RL framework, is to tackle key challenges concerning convergence guarantees, optimization transparency, and sequential-decision interpretation. Empirical results demonstrate that the proposed AWaVO balances a reasonable trade-off between high performance and quantitative interpretability in both simulation and real quadrotor tasks.

Acknowledgments

This work was partially supported by NSF-UKRI [grant number NE/T011467/1]; and the Engineering and Physical Sciences Research Council [grant number EP/X040518/1].

Impact Statement

This paper presents work whose goal is to advance knowledge about interpretability and reproducibility in the RL community. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. Additionally, we provide the video and additional demonstration to the real-world quadrotor experiments.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Altman, E. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

Bahouri, H., Chemin, J.-Y., and Danchin, R. *Fourier analysis and nonlinear partial differential equations*, volume 343. Springer, 2011.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Beylkin, G. The inversion problem and applications of the generalized radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984.

Bohez, S., Abdolmaleki, A., Neunert, M., Buchli, J., Heess, N., and Hadsell, R. Value constrained model-free continuous control. *arXiv preprint arXiv:1902.04623*, 2019.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

Brat, G. Are we ready for the first easa guidance on the use of ml in aviation? In *SAE G34 Meeting*, 2021.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference and q-learning provably converge to global optima. *arXiv preprint arXiv:1905.10027*, 2019.

Chen, X., Yang, Y., and Li, Y. Augmented sliced wasserstein distances. In *International Conference on Learning Representations*, 2021.

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Clement, P. and Desch, W. An elementary proof of the triangle inequality for the wasserstein metric. *Proceedings of the American Mathematical Society*, 136(1):333–339, 2008.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

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

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Devidze, R., Radanovic, G., Kamalaruban, P., and Singla, A. Explicable reward design for reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:20118–20131, 2021.

Ding, Z., Yang, T., Zhang, K., Xu, C., and Gao, F. Vidfusion: Robust visual-inertial-dynamics odometry for accurate external force estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14469–14475. IEEE, 2021.

Ehrenpreis, L. *The universality of the Radon transform*. OUP Oxford, 2003.

Fernández Llorca, D. and Gómez, E. Trustworthy autonomous vehicles. *Publications Office of the European Union, Luxembourg,, EUR*, 30942, 2021.

Fernandez-Llorca, D. and Gómez, E. Trustworthy artificial intelligence requirements in the autonomous driving domain. *Computer*, 56(2):29–39, 2023.

Gannoun, A., Saracco, J., and Yu, K. Comparison of kernel estimators of conditional distribution function and quantile regression under censoring. *Statistical Modelling*, 7 (4):329–344, 2007.

Gorbach, N. S., Bauer, S., and Buhmann, J. M. Scalable variational inference for dynamical systems. *Advances in neural information processing systems*, 30, 2017.

Homan, A. and Zhou, H. Injectivity and stability for a generic class of generalized radon transforms. *The Journal of Geometric Analysis*, 27:1515–1529, 2017.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kappen, H. J., Gómez, V., and Opper, M. Optimal control as a graphical model inference problem. *Machine learning*, 87:159–182, 2012.

Kenny, E. M., Ford, C., Quinn, M., and Keane, M. T. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence*, 294: 103459, 2021.

Kenny, E. M., Tucker, M., and Shah, J. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2022.

Koenker, R. *Quantile regression*, volume 38. Cambridge university press, 2005.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Levine, S. Understanding the world through action. In *Conference on Robot Learning*, pp. 1752–1757. PMLR, 2022.

Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Liu, Y., Ding, J., and Liu, X. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4940–4947, 2020.

Liu, Z., Cen, Z., Isenbaev, V., Liu, W., Wu, S., Li, B., and Zhao, D. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pp. 13644–13668. PMLR, 2022.

Lorenz, E. N. and Emanuel, K. A. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, 55(3):399–414, 1998.

McNamara, S. The law and ethics of high-frequency trading. *Minn. JL Sci. & Tech.*, 17:71, 2016.

Mitrophanov, A. Y. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

Napoleone, A., Macchi, M., and Pozzetti, A. A review on the characteristics of cyber-physical systems for the future smart factories. *Journal of manufacturing systems*, 54:305–335, 2020.

Ng, R. Fourier slice photography. In *ACM Siggraph 2005 Papers*, pp. 735–744. 2005.

Nietert, S., Goldfeld, Z., Sadhu, R., and Kato, K. Statistical, robustness, and computational guarantees for sliced wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

O'Donoghue, B., Osband, I., and Ionescu, C. Making sense of reinforcement learning and probabilistic inference. *arXiv preprint arXiv:2001.00805*, 2020.

Okada, M. and Taniguchi, T. Acceleration of gradient-based path integral method for efficient optimal and inverse optimal control. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3013–3020. IEEE, 2018.

Okada, M. and Taniguchi, T. Variational inference mpc for bayesian model-based reinforcement learning. In *Conference on robot learning*, pp. 258–272. PMLR, 2020.

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86), 2017.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pp. 435–446. Springer, 2012.

Radon, J. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 5(21): 124, 2005.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pp. 306–314. PMLR, 2014.

Torens, C., Durak, U., and Dauer, J. C. Guidelines and regulatory framework for machine learning in aviation. In *AIAA Scitech 2022 Forum*, pp. 1132, 2022.

Toussaint, M. and Storkey, A. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on Machine learning*, pp. 945–952, 2006.

Villani, C. The wasserstein distances. In *Optimal transport*, pp. 93–111. Springer, 2009.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wang, Y. and Boyle, D. Trustworthy reinforcement learning for quadrotor uav tracking control systems. *arXiv preprint arXiv:2302.11694*, 2023.

Wang, Y., O'Keeffe, J., Qian, Q., and Boyle, D. Quadue-ccm: Interpretable distributional reinforcement learning using uncertain contraction metrics for precise quadrotor trajectory tracking. In *Conference on Robot Learning*, pp. 2306–2316. PMLR, 2023.

Welch, G., Bishop, G., et al. An introduction to the kalman filter. 1995.

Xu, T., Liang, Y., and Lan, G. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR, 2021.

Yang, T. Y., Rosca, J., Narasimhan, K., and Ramadge, P. J. Projection-based constrained policy optimization. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Zhao, W., Chen, R., Sun, Y., Liu, R., Wei, T., and Liu, C. Guard: A safe reinforcement learning benchmark. *arXiv preprint arXiv:2305.13681*, 2023.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9623–9633, 2022.

Appendix

A. Notation Table

Table 1. Main Notation Conventions

Parameters	Definition
$\boldsymbol{s}_t$	A state $\boldsymbol{s}$ sampled at a discrete timestamp t.
$\boldsymbol{a}_t$	An action $\boldsymbol{a}$ sampled at a discrete timestamp t.
τ	A trajectory.
D	Observed training dataset.
S	A finite set of states $\{\boldsymbol{s}\}$.
A	A finite set of actions $\{\boldsymbol{a}\}$.
P	A finite set of transition probabilities $\{p\}$.
R	A finite set of bounded immediate rewards $\{r\}$.
G	A finite collection of unity functions $\{g\}$.
$\mathcal{O}_t = \{\mathcal{O}_{r,t}, \mathcal{O}_{g_i,t}\}$	Binary variables of the optimality. $\mathcal{O}_{r,t} = 1$ and $\mathcal{O}_{g_i,t} = 1$ signify that the trajectory τ is optimized and compliant with the constraints.
$p_D(\theta)$	The posterior probability $p(\theta\|D)$.
$p(\mathcal{O}\|\tau)$	The optimality likelihood.
$\widetilde{r}(\tau)$	The accumulated reward $\widetilde{r}(\tau) = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t)]$.
$\widetilde{g}_i(\tau)$	The utility function $\widetilde{g}_i(\tau) = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t g_i(\boldsymbol{s}_t, \boldsymbol{a}_t)]$.
$\mathcal{F} = \{\mathcal{F}_r, \mathcal{F}_g\}$	$\mathcal{F}_r$ and $\mathcal{F}_g$ are the operators, with specifically defined as $\mathcal{F}_r \cdot p(\mathcal{O}_r\|\tau) := \widetilde{r}(\tau)$ and $\mathcal{F}_g \cdot p(\mathcal{O}_{g_i}\|\tau) := \widetilde{g}_i(\tau)$.
b_i	A fixed limit for the i-th constraint.
$\boldsymbol{\tau}_c$	The tolerance.
l	A curve on the hypersurface in the spatial Radon transform.
$\widetilde{\theta}$	An unit vector tangent to l. Both l and $\widetilde{\theta}$ represent the parameters of hypersurfaces.
$q_\theta(\tau)$	A variational distribution.
$q(\boldsymbol{a})$	An approximation of $p(\mathcal{O}\|\tau)$.

B. Background on Wasserstein Distance

B.1. Sliced Wasserstein Distance

A fundamental challenge in both machine learning and statistics communities is to form effective metrics between pairs of probability distributions. Weaker notions, such as divergence measures, including KL divergence (Kullback & Leibler, 1951), have been proposed and widely used. However, such measures do not satisfy the two basic properties of a metric, namely symmetry and triangle inequality. To address this issue, interest has rapidly increased in optimal transport in recent years. In this subsection, we introduce the Wasserstein distance and its variants, including SWD (Rabin et al., 2012; Nietert et al., 2022) and GSWD (Kolouri et al., 2019), as metrics that conditionally satisfy the properties.

Let $\Gamma(\mu, \nu)$ be a set of all transportation plans $\gamma \in \Gamma(\mu, \nu)$, where γ is a joint distribution over the space $\mathcal{X} \times \mathcal{X}$, and $\mu, \nu \in P_k(\mathcal{X})$ are two measures of probability distributions over $\mathcal{X}$. $P_k(\mathcal{X})$ represents a set of Borel probability measures with finite k-th moment on a Polish metric space (Villani et al., 2009). $d(x, y)$ represents a distance function over $\mathcal{X}$. The Wasserstein distance of order $k \in [1, \infty)$ between two measures μ, ν is defined as (Villani et al., 2009): $W_k(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^k \mathrm{d}\gamma(x, y)\right)^{1/k}$. This definition, however, involves solving an optimization problem that is computationally expensive in practical implementation, particularly for high-dimensional distributions. Thus sliced k-Wasserstein distance (Rabin et al., 2012; Nietert et al., 2022), defined over spaces of hyperplanes in $\mathbb{R}^d$, is proposed as a

computationally efficient approximation:

$$\mathbf{SWD}_k(\mu,\nu)=\left(\int_{\mathbb{S}^{d-1}} W_k^k\left(\mathcal{R}_\mu\left(\cdot,\widetilde{\theta}\right),\mathcal{R}_\nu\left(\cdot,\widetilde{\theta}\right)\right)\mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} \tag{3}$$

where Radon transform $\mathcal{R}$ (Radon, 2005) is introduced in SWD to map a function $f(\cdot)$ to the hyperplanes $\left\{x\in\mathbb{R}^d|\langle x,\widetilde{\theta}\rangle=l\right\}$, i.e., $\mathcal{R}f(l,\widetilde{\theta})=\int_{\mathbb{R}^d} f(x)\delta(l-\langle x,\widetilde{\theta}\rangle)\mathrm{d}x$: $l\in\mathbb{R}$ and $\widetilde{\theta}\in\mathbb{S}^{d-1}\subset\mathbb{R}^d$ represent the parameters of these hyperplanes. In the definition of SWD, the Radon transform $\mathcal{R}_\mu$ is employed as the push-forward operators, defined by $\mathcal{R}_\mu(l,\widetilde{\theta})=\int_{\mathbb{R}^d}\delta(l-\langle x,\widetilde{\theta}\rangle)\mathrm{d}\mu$ (Kolouri et al., 2019).

B.2. Generalized Sliced Wasserstein Distance

While SWD offers a computationally efficient way to approximate the Wasserstein distance, the projections are limited to linear subspaces, such as hyperplanes $\{x\}$. Due to the nature of these linear projections, the resulting metrics typically have low projection efficiency in high-dimensional spaces (Kolouri et al., 2019; Deshpande et al., 2019). Thus various variants of SWD are proposed to enhance its projection effectiveness. Specifically, the GSWD (Kolouri et al., 2019), defined in Equation 4, is proposed by incorporating nonlinear projections. Its main novelty is that Generalized Radon Transforms (GRTs) $\mathcal{G}$ (Beylkin, 1984; Ehrenpreis, 2003; Homan & Zhou, 2017), i.e., $\mathcal{G}f(l,\widetilde{\theta})=\int_{\mathbb{R}^d} f(x)\delta(l-\beta(x,\widetilde{\theta}))\mathrm{d}x$, are used to define the nonlinear projections towards hypersurfaces rather than linear projections to the hyperplanes in SWD. Let $\beta(x,\widetilde{\theta})$ be *a defining function* when satisfying the conditions **H.1-H.4** in (Kolouri et al., 2019).

$$\mathbf{GSWD}_k(\mu,\nu)=\left(\int_{\mathcal{X}_{\widetilde{\theta}}} W_k^k\left(\mathcal{G}_\mu\left(\cdot,\widetilde{\theta}\right),\mathcal{G}_\nu\left(\cdot,\widetilde{\theta}\right)\right)\mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} \tag{4}$$

where $\widetilde{\theta}\in\mathcal{X}_{\widetilde{\theta}}$ and $\mathcal{X}_{\widetilde{\theta}}$ is a compact set of all feasible parameters $\widetilde{\theta}$ for $\beta(\cdot,\widetilde{\theta})$, e.g., $\mathcal{X}_{\widetilde{\theta}}=\mathbb{S}^{d-1}$ for $\beta(\cdot,\widetilde{\theta})=\langle\cdot,\widetilde{\theta}\rangle$. The GRT operator $\mathcal{G}_\mu$ is utilized as the push-forward operator, i.e., $\mathcal{G}_\mu(l,\widetilde{\theta})=\int_{\mathbb{G}^d}\delta(l-\beta(x,\widetilde{\theta}))\mathrm{d}\mu$. For the theoretical properties of a metric, SWD is a true metric that satisfies both symmetry and triangle inequality (Bonnotte, 2013), where the approximation error is obtained and analyzed in (Nadjahi et al., 2020). The GSWD defined by Equation 4 is a true metric if and only if $\beta(\cdot)$ in $\mathcal{G}$ is an injective mapping (Chen et al., 2021).

C. Background on Distributional Representation in Bellman Equation and Temporal Difference Learning

C.1. Reasoning behind Distributional Representation

The motivation for employing a distributional representation is twofold. Firstly, it provides more comprehensive and richer value-distribution information, thereby enhancing the stability of the learning process. This stability is particularly important for Bayesian learning processes, which often encounter challenges in achieving stable convergence. Secondly, the distributional representation contributes significantly to interpretability. As illustrated in Equation 16 of the proof, it uses quantiles derived from the distributional representation to formally establish the transparency of the convergence process outlined in **Theorem 5.5**.

C.2. Distributional Representation in Bellman Equation

Unlike traditional RL, where the primary objective is to maximize the expected action-value function Q, the distributional Bellman equation (Bellemare et al., 2017) was proposed to approximate and parameterize the entire distribution of future rewards. In the setting of policy evaluation, given a deterministic policy π, the *Bellman operator* $\mathcal{T}^\pi$ is defined as (Bellemare et al., 2017; Dabney et al., 2018):

$$\mathcal{T}^\pi Z(\boldsymbol{s}, \boldsymbol{a}) :\overset{D}{=} R(\boldsymbol{s}, \boldsymbol{a}) + \gamma Z(S', A') \tag{5}$$

where Z^π denotes the state-action distribution, and $R(\boldsymbol{s}, \boldsymbol{a})$ denotes the reward distribution. In control setting, a distributional *Bellman optimality operator* $\mathcal{T}$ with quantile approximation is proposed in (Dabney et al., 2018):

$$\mathcal{T} Z(\boldsymbol{s}, \boldsymbol{a}) :\overset{D}{=} R(\boldsymbol{s}, \boldsymbol{a}) + \gamma Z(\boldsymbol{s'}, \underset{a'}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{p}, R}[Z(\boldsymbol{s'}, \boldsymbol{a'})]) \tag{6}$$

where we let $Z_\theta := \frac{1}{N} \sum_{i=1}^{N} \delta_{q_i(\boldsymbol{s}, \boldsymbol{a})}$ be a quantile distribution mapping one state-action pair $(\boldsymbol{s}, \boldsymbol{a})$ to a uniform probability distribution supported on q_i. Based on Equation 5, a contraction is demonstrated (Dabney et al., 2018) over the Wasserstein metric: $\bar{d}_\infty(\Pi_{W_1} \mathcal{T}^\pi Z_1, \Pi_{W_1} \mathcal{T}^\pi Z_2) \leq \bar{d}_\infty(Z_1, Z_2)$, where $\bar{d}_k := \sup W_k(Z_1, Z_2)$ denotes the maximal form of the k-Wasserstein metrics. W_k, $k \in [1, \infty]$ denotes the k-Wasserstein distance. Π_{W_1} is a quantile approximation under the minimal 1-Wasserstein distance W_1.

C.3. Distributional Representation in Temporal Difference Learning

Building upon the aforementioned contraction guarantees, we utilize distributional TD learning to estimate the distribution of state-action value, denoted as Z. In each iteration, we have the following:

$$\begin{aligned} \zeta^i_{k+1}(\boldsymbol{s}, \boldsymbol{a}) &= \zeta^i_k(\boldsymbol{s}, \boldsymbol{a}) + l_{td} \boldsymbol{\Delta}^i_k \\ &= \zeta^i_k(\boldsymbol{s}, \boldsymbol{a}) + l_{td} \times \bar{d}_\infty(\Pi_{W_1} \mathcal{T}^\pi (h_i(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s'}) + \gamma \zeta^i_k(\boldsymbol{s'})), \Pi_{W_1} \mathcal{T}^\pi \zeta^i_k(\boldsymbol{s}, \boldsymbol{a})) \end{aligned} \tag{7}$$

where $\zeta^i_k \in S \times A$ represents the estimated distribution of the state-action distribution Z in the k-th TD-learning-iteration for all $i = 0, ..., p$. The TD learning rate is denoted as l_{td}. The function $h_i : S \times A \times S \to \mathbb{R}$ maps the triple $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s'})$ to a *real number*. Specifically, h_i is defined as $h_i = r$ when $i = 0$; and $h_i = g_i$ when $i \in [1, n]$. The distributional TD error $\boldsymbol{\Delta}^i_k$ in Equation 10 is calculated by $\bar{d}_\infty(\Pi_{W_1} \mathcal{T}^\pi (h_i(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s'}) + \gamma \zeta^i_k(\boldsymbol{s'})), \Pi_{W_1} \mathcal{T}^\pi \zeta^i_k(\boldsymbol{s}, \boldsymbol{a}))$.

In Algorithm 1, the gradient of actor and critic network, denoted as δ_{θ^μ} and δ_{θ^Q}, can be calculated as follows:

$$\begin{aligned} \delta_{\theta^\mu} &= (1/N) \sum \nabla_{\theta^\mu} \pi_{\theta^\mu}(\boldsymbol{s}_n) \mathbb{E}[\nabla_a Z_{\theta^Q}(\boldsymbol{s}_n, \boldsymbol{a})]_{a = \pi_{\theta^\mu}(\boldsymbol{s}_n)} \\ \delta_{\theta^Q} &= (1/N) \sum \nabla_{\theta^Q} \bar{d}_\infty(\Pi_{W_1} \mathcal{T}^\pi Z_{\theta^Q}(\boldsymbol{s}_n, \boldsymbol{a}_n), \Pi_{W_1} \mathcal{T}^\pi \widetilde{g}_i(\tau)) \end{aligned} \tag{8}$$

where Π_{W_1} represents a quantile approximation under the minimal 1-Wasserstein distance W_1.

D. Algorithm Details and Proofs

D.1. Detailed Derivation of the Objective Function

The aim of variational inference is to minimize the distance $D(q_\theta(\tau)||p(\tau|\mathcal{O})) = \mathbf{A} - \mathbf{GSWD}_k(q_\theta(\tau), p(\tau|\mathcal{O}))$ between the variational distribution $q_\theta(\tau)$ and the posterior distribution $p(\tau|\mathcal{O})$. Let $\mathcal{P}_{trans} = p(\boldsymbol{s}, \boldsymbol{a}|\theta)\, p_D(\theta)$, $\tilde{x} = x/\mathcal{P}_{trans}$ and $\tilde{y} = y/\mathcal{P}_{trans}$ and recall **Definition 4.1**, i.e., the definition of $\mathbf{A} - \mathbf{GSWD}_k$. Then, the variational inference can be reformulated to the minimization problem:

$$
\begin{aligned}
&\arg\min_{q_\theta(\tau)} \mathbf{A} - \mathbf{GSWD}_k\left(q_\theta(\tau), p(\tau|\mathcal{O})\right) = \arg\min_{q_\theta(\tau)} \mathbf{A} - \mathbf{GSWD}_k\left(q(\boldsymbol{a}) \cdot \mathcal{P}_{trans}, p(\mathcal{O}|\tau) \cdot \mathcal{P}_{trans}\right) \\
&= \arg\min_{q_\theta(\tau)} \left(\int_{\mathcal{X}_\theta} W_k^k \left(\mathcal{A}_{q(\boldsymbol{a})\cdot\mathcal{P}_{trans}}\left(\cdot, \widetilde{\theta}\right), \mathcal{A}_{p(\mathcal{O}|\tau)\cdot\mathcal{P}_{trans}}\left(\cdot, \widetilde{\theta}\right)\right) \mathrm{d}\widetilde{\theta} \right)^{\frac{1}{k}} \\
&\overset{(i)}{=} \arg\min_{q_\theta(\tau)} \left(\int_{\mathcal{X}_\theta} W_k^k \left(\mathcal{P}_{trans} \cdot \mathcal{A}_{q(\boldsymbol{a})}\left(\cdot, \widetilde{\theta}\right), \mathcal{P}_{trans} \cdot \mathcal{A}_{p(\mathcal{O}|\tau)}\left(\cdot, \widetilde{\theta}\right)\right) \mathrm{d}\widetilde{\theta} \right)^{\frac{1}{k}} \\
&= \arg\min_{q_\theta(\tau)} \inf_{\gamma \in \Gamma(\mathcal{P}_{trans}\cdot\mathcal{A}_{q(\boldsymbol{a})}, \mathcal{P}_{trans}\cdot\mathcal{A}_{p(\mathcal{O}|\tau)})} \left(\int_{\mathcal{X}_\theta} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^k \mathrm{d}\gamma(x,y) \mathrm{d}\widetilde{\theta} \right)^{\frac{1}{k}} \\
&= \arg\min_{q_\theta(\tau)} \inf_{\gamma \in \Gamma(\mathcal{A}_{q(\boldsymbol{a})}, \mathcal{A}_{p(\mathcal{O}|\tau)})} \left(\int \int d(\mathcal{P}_{trans}\tilde{x}, \mathcal{P}_{trans}\tilde{y})^k \mathrm{d}\gamma(\mathcal{P}_{trans}\tilde{x}, \mathcal{P}_{trans}\tilde{y}) \mathrm{d}\widetilde{\theta} \right)^{\frac{1}{k}} \\
&\overset{(ii)}{=} \arg\min_{q_\theta(\tau)} \inf_{\gamma \in \Gamma(\mathcal{A}_{q(\boldsymbol{a})}, \mathcal{A}_{p(\mathcal{O}|\tau)})} \left(\int \int \left(\mathcal{P}_{trans} \cdot d(\tilde{x}, \tilde{y})\right)^k \mathrm{d}\gamma(\mathcal{P}_{trans}\tilde{x}, \mathcal{P}_{trans}\tilde{y}) \mathrm{d}\widetilde{\theta} \right)^{\frac{1}{k}} \\
&\overset{(iii)}{=} \arg\min_{q_\theta(\tau)} \mathcal{P}_{trans} \cdot \inf_{\gamma \in \Gamma(\mathcal{A}_{q(\boldsymbol{a})}, \mathcal{A}_{p(\mathcal{O}|\tau)})} \left(\int \int d(\tilde{x}, \tilde{y})^k \mathrm{d}\gamma(\tilde{x}, \tilde{y}) \mathrm{d}\widetilde{\theta} \right)^{\frac{1}{k}} \\
&= \arg\min_{q_\theta(\tau)} \mathcal{P}_{trans} \cdot \mathbf{A} - \mathbf{GSWD}_k\left(q(\boldsymbol{a}), p(\mathcal{O}|\tau)\right)
\end{aligned}
\tag{9}
$$

where (i) follows from the push-forward operator definition: $\mathcal{A}_\mu(l, \widetilde{\theta}) = \int_{\mathbb{G}^d} \delta(l - \alpha(x, \widetilde{\theta}))\mathrm{d}\mu$. (ii) follows from $d(c\tilde{x}, c\tilde{y}) = cd(\tilde{x}, \tilde{y}), c \in (0, 1)$ as d is a metric. (iii) follows from the fact that $\mathrm{d}\gamma(c\tilde{x}, c\tilde{y}) = \mathrm{d}\gamma(cx, cy) = \mathrm{d}\gamma(x, y) = \mathrm{d}\gamma(\tilde{x}, \tilde{y})$, since $\mathrm{d}\gamma(c\tilde{x}, c\tilde{y})$ is the measure of the subset of $\mathcal{X} \times \mathcal{X}$, which is just the re-scaled version of the subset $(\tilde{x}, \tilde{y})$ by the map $(x, y) \mapsto (x, y)$.

Equation 9 presents that the objective can be transformed to the minimization problem, i.e., $\arg\min_{q_\theta(\tau)} \mathbf{A} - \mathbf{GSWD}_k\left(q(\boldsymbol{a}), p(\mathcal{O}|\tau)\right)$, where $p(\mathcal{O}|\tau)$ represents the optimality likelihood.

D.2. Definition of Distributional Temporal Difference

We use distributional TD learning to estimate the distribution of state-action value, denoted as Z. In each iteration, we have the following:

$$
\begin{aligned}
\zeta_{k+1}^i(\boldsymbol{s}, \boldsymbol{a}) &= \zeta_k^i(\boldsymbol{s}, \boldsymbol{a}) + l_{td} \boldsymbol{\Delta}_k^i \\
&= \zeta_k^i(\boldsymbol{s}, \boldsymbol{a}) + l_{td} \times \bar{d}_\infty(\Pi_{W_1} \mathcal{T}^\pi(h_i(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + \gamma \zeta_k^i(\boldsymbol{s}')), \Pi_{W_1} \mathcal{T}^\pi \zeta_k^i(\boldsymbol{s}, \boldsymbol{a}))
\end{aligned}
\tag{10}
$$

where $\zeta_k^i \in S \times A$ represents the estimated distribution of the state-action distribution Z in the k-th TD-learning-iteration for all $i = 0, ..., p$. The TD learning rate is denoted as l_{td}. The function $h_i : S \times A \times S \to \mathbb{R}$ maps the triple $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$ to a *real number*. Specifically, h_i is defined as $h_i = r$ when $i = 0$; and $h_i = g^i$ when $i \in [1, p]$. The distributional TD error $\boldsymbol{\Delta}_k^i$ in Equation 10 is calculated by $\bar{d}_\infty(\Pi_{W_1} \mathcal{T}^\pi(h_i(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + \gamma \zeta_k^i(\boldsymbol{s}')), \Pi_{W_1} \mathcal{T}^\pi \zeta_k^i(\boldsymbol{s}, \boldsymbol{a}))$.

D.3. Proofs

Here we present the proofs of **Proposition 5.1** *(Pseudo-metric)*, **Theorem 5.4** *(Global Convergence)* and **Theorem 5.5** *(Global Convergence Rate)*, as outlined in **Section 5**. The associated propositions, namely **Proposition D.1** *(Policy Evaluation)* and **Proposition D.2** *(Policy Improvement)*, are explicated and clarified by their respective proofs below.

Proposition 5.1. *(Pseudo-metric): Given two probability measures $\mu, \nu \in P_k(\mathcal{X})$ and a mapping $\alpha : \mathcal{X} \to \mathcal{R}_{\widetilde{\theta}}$, the adaptive slicing A-GSWD, defined in **Definition 4.1**, with order k in the range $[1, \infty)$, is a pseudo-metric. This pseudo-metric satisfies non-negativity, symmetry, the triangle inequality, and $\mathbf{A-GSWD}_k(\mu, \mu) = 0$.*

Proof: The non-negativity property naturally arises from the fact that the Wasserstein distance W_k is a metric (Villani et al., 2009). To prove symmetry, since the k-Wasserstein distance is a metric (Villani et al., 2009):

$$W_k\left(\mathcal{A}_\mu(\cdot, \widetilde{\theta}; \alpha), \mathcal{A}_\nu(\cdot, \widetilde{\theta})\right) = W_k\left(\mathcal{A}_\nu(\cdot, \widetilde{\theta}), \mathcal{A}_\mu(\cdot, \widetilde{\theta})\right)$$

Thus, there exists (Chen et al., 2021):

$$\begin{aligned}\mathbf{A-GSWD_k}(\mu, \nu) &= \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_\mu(\cdot, \widetilde{\theta}), \mathcal{A}_\nu(\cdot, \widetilde{\theta})\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} \\ &= \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_\nu(\cdot, \widetilde{\theta}), \mathcal{A}_\mu(\cdot, \widetilde{\theta})\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} = \mathbf{A-GSWD_k}(\nu, \mu)\end{aligned}$$

Therefore, symmetry holds. Then, we prove the triangle inequality. Since the triangle inequality holds for the Wasserstein distance, we can obtain $W_k\left(\mathcal{A}_{\mu_1}, \mathcal{A}_{\mu_3}\right) \leq W_k\left(\mathcal{A}_{\mu_1}, \mathcal{A}_{\mu_2}\right) + W_k\left(\mathcal{A}_{\mu_2}, \mathcal{A}_{\mu_3}\right)$. Thus, there exists:

$$\begin{aligned}\mathbf{A-GSWD}_k(\mu_1, \mu_3) &= \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_{\mu_1}, \mathcal{A}_{\mu_3}\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} \\ &\leq \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_{\mu_1}, \mathcal{A}_{\mu_2}\right) + W_k^k\left(\mathcal{A}_{\mu_2}, \mathcal{A}_{\mu_3}\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} \\ &\leq \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_{\mu_1}, \mathcal{A}_{\mu_2}\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} + \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_{\mu_2}, \mathcal{A}_{\mu_3}\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}}\end{aligned} \tag{11}$$

where the derivation of Equation 11 is based on the Minkowski inequality (Bahouri et al., 2011), which establishes that $\mathbf{A-GSWD}_k$ satisfies the triangle inequality. Since $W_k(\mu, \mu) = 0$ for any μ (Villani et al., 2009), we have $\mathbf{A-GSWD}_k(\mu, \mu) = \left(\int_{\mathcal{R}_{\widetilde{\theta}}} W_k^k\left(\mathcal{A}_\mu, \mathcal{A}_\mu\right) \mathrm{d}\widetilde{\theta}\right)^{\frac{1}{k}} = 0$.

Therefore, A-GSWD is a pseudo-metric that satisfies non-negativity, symmetry, the triangle inequality, and $\mathbf{A-GSWD}_k(\mu, \mu) = 0$. ■

Remark 5.2. *The adaptive slicing A-GSWD, with order $k \in [1, \infty)$, is a true metric if and only if the AGRT $\mathcal{A}$, defined in **Definition 4.1**, is an injective mapping.*

Proof: Given the indiscernibility property for the Wasserstein distance W_k (Villani et al., 2009), it follows $W_k(\mu_1, \mu_2) = 0$ if and only if $\mu_1 = \mu_2$. Consequently, $\mathbf{A-GSWD}_k(\mu_1, \mu_2) = 0$ is equivalent to $\mathcal{A}_{\mu_1}(l, \widetilde{\theta}) = \mathcal{A}_{\mu_2}(l, \widetilde{\theta})$. The equality $\mathcal{A}_{\mu_1}(l, \widetilde{\theta}) = \mathcal{A}_{\mu_2}(l, \widetilde{\theta})$ implies $\mu_1 = \mu_2$ if and only if $\mathcal{A}$ is an injective mapping.

Therefore, A-GSWD is a metric if and only if $\mathcal{A}_{\mu_1}(l, \widetilde{\theta}) = \mathcal{A}_{\mu_2}(l, \widetilde{\theta})$ implies $\mu_1 = \mu_2$, i.e., the AGRT $\mathcal{A}$ is an injective mapping. ■

Proposition D.1. *(Policy Evaluation) (Dabney et al., 2018; Wang et al., 2023)): we consider a quantile approximation Π_{W_1} under the minimal 1-Wasserstein distance W_1, the Bellman operator $\mathcal{T}^\pi$ under a deterministic policy π and $Z_{k+1}(\boldsymbol{s}, \boldsymbol{a}) = \Pi_{W_1}\mathcal{T}^\pi Z_k(\boldsymbol{s}, \boldsymbol{a})$. The sequence $Z_k(\boldsymbol{s}, \boldsymbol{a})$ converges to a unique fixed point $\widetilde{Z_\pi}$ under the maximal form of ∞-Wasserstein metric $\bar{d}_\infty$.*

Proof: We recall a contraction proved in (Dabney et al., 2018) over the Wasserstein Metric:

$$\bar{d}_\infty(\Pi_{W_1}\mathcal{T}^\pi Z_1, \Pi_{W_1}\mathcal{T}^\pi Z_2) \leq \bar{d}_\infty(Z_1, Z_2) \tag{12}$$

where Equation 12 implies that the combined operator $\Pi_{W_1}\mathcal{T}^\pi$ is an ∞-contraction. Based on Banach's fixed point theorem, $\mathcal{T}^\pi$ has a unique fixed point, i.e., $\widetilde{Z}_\pi$. Furthermore, the definition of *Bellman optimality operator*, defined as Equation 6, implies that all moments of Z are bounded. Therefore, we conclude that the sequence $Z_k(\boldsymbol{s}, \boldsymbol{a})$ converges to $\widetilde{Z}_\pi$ in $\bar{d}_\infty$ for $p \in [1, \infty]$. ■

Proposition D.2. *(Policy Improvement): Given an old policy $\boldsymbol{\pi_{old}}$, a new policy $\boldsymbol{\pi_{new}}$ and $Q(s,a) = \mathbb{E}[Z(s,a)]$, suppose Assumption 1 holds, there exists $Q^{\boldsymbol{\pi_{new}}}(s,a) \geq Q^{\boldsymbol{\pi_{old}}}(s,a)$ when performing Algorithm 1, $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$ if and only if the reward operator family $\mathcal{F} = \{\mathcal{F}_r, \mathcal{F}_g\}$ satisfies the both* ***Conditions 5.3****.*

Proof: We recall that $\{\mathcal{F}_r, \mathcal{F}_g\}$ are two operators defined as $\widetilde{r}(\tau) := \mathcal{F}_r \cdot p\left(\mathcal{O}_r|\tau\right)$ and $\widetilde{g}^i(\tau) := \mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)$, respectively. Suppose Assumption 1 holds. Since the two optimization objectives in policy updating, i.e., $\max\mathbb{E}[\mathcal{F}_r \cdot p\left(\mathcal{O}_r|\tau\right)]$ and $\min\mathbb{E}[\mathcal{F}_g \cdot p\left(\mathcal{O}_{g_i}|\tau\right)]$ (see **Section 4.2**), and $p\left(\mathcal{O}|\tau\right)$ is defined on $(0,1]$, we can conclude the both **Conditions 5.3** that *(i)* $\mathcal{F}_r$ is monotonically increasing and continuously defined on $(0,1]$, and the range covers $[r_{\min}, r_{\max}]$; *(ii)* $\mathcal{F}_g$ is monotonically decreasing and continuously defined on $(0,1]$, and the range covers $[r_{\min}, r_{\max}]$.

Then based on Equation 6, there exists:

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}[Q(s_t, \pi(s_t))] \leq \max_{a' \in \mathcal{A}} \mathbb{E}[Q(s_t, a')] \\ &= \mathbb{E}[Q(s_t, \pi'(s_t))] \end{aligned} \tag{13}$$

where $\mathbb{E}_\pi[\cdot] = \sum_{a \in A} \boldsymbol{\pi}(a|s)[\cdot]$, and $V^\pi(s) = \mathbb{E}_\pi \mathbb{E}[Z_k(s,a)]$ is the value function. According to Equation 13 and Equation 6, it yields:

$$\begin{aligned} Q^{\boldsymbol{\pi_{old}}} &= Q^{\boldsymbol{\pi_{old}}}(s_t, \boldsymbol{\pi_{old}}(s_t)) \\ &= r_{t+1} + \gamma \mathbb{E}_{s_{t+1}} \mathbb{E}_{\boldsymbol{\pi_{old}}} Q^{\boldsymbol{\pi_{old}}}(s_{t+1}, \boldsymbol{\pi_{old}}(s_{t+1})) \\ &\overset{(i)}{\leq} r_{t+1} + \gamma \mathbb{E}_{s_{t+1}} \mathbb{E}_{\boldsymbol{\pi_{new}}} Q^{\boldsymbol{\pi_{old}}}(s_{t+1}, \boldsymbol{\pi_{new}}(s_{t+1})) \\ &\leq r_{t+1} + \mathbb{E}_{s_{t+1}} \mathbb{E}_{\boldsymbol{\pi_{new}}} [\gamma r_{t+2} \\ &+ \gamma^2 \mathbb{E}_{s_{t+2}} Q^{\boldsymbol{\pi_{old}}}(s_{t+2}, \boldsymbol{\pi_{new}}(s_{t+2}))|] \\ &\leq r_{t+1} + \mathbb{E}_{s_{t+1}} \mathbb{E}_{\boldsymbol{\pi_{new}}} [\gamma r_{t+2} + \gamma^2 r_{t+3} + ...] \\ &= r_{t+1} + \mathbb{E}_{s_{t+1}} V^{\boldsymbol{\pi_{new}}}(s_{t+1}) \\ &= Q^{\boldsymbol{\pi_{new}}} \end{aligned} \tag{14}$$

where (i) relies on Equation 13, and $\boldsymbol{\pi_{new}}$ corresponds to the maximum Q in the Bellman function. Therefore, we have $Q^{\boldsymbol{\pi_{new}}}(s,a) \geq Q^{\boldsymbol{\pi_{old}}}(s,a)$. ■

Then we provide **Lemma D.3** and the proof of **Theorem 5.4**.

Lemma D.3. *(Bellemare et al., 2017): The Bellman operator $\mathcal{T}^\pi$ is a p-contraction under the p-Wasserstein metric $\bar{d}_p$.*

Theorem 5.4. *(Global Convergence): Given the policy in the i-th policy improvement $\boldsymbol{\pi^i}$, $\boldsymbol{\pi^i} \to \boldsymbol{\pi}^*$ and $i \to \infty$, suppose Assumption 1 holds,* there exists $Q^{\boldsymbol{\pi}^*}(s,a) \geq Q^{\boldsymbol{\pi}^i}(s,a)$ *if and only if* the reward operator family $\mathcal{F}$ satisfies the both **Conditions 5.3**.

Proof: Since **Proposition D.2** suggests $Q^{\boldsymbol{\pi}_{i+1}}(s,a) \geq Q^{\boldsymbol{\pi}_i}(s,a)$, the sequence $Q^{\boldsymbol{\pi}_i}(s,a)$ is monotonically increasing *if and only if* the reward operator family $\mathcal{F}$ satisfies the both **Conditions 5.3**. Furthermore, **Lemma D.3** implies that the the state-action distribution Z over $\mathbb{R}$ has bounded p-th moment, so the first moment of Z, i.e., $Q^{\boldsymbol{\pi}_i}(s,a)$, is upper bounded. Therefore, the sequence $Q^{\boldsymbol{\pi}_i}(s,a)$ converges to an upper limit $Q^{\boldsymbol{\pi}_*}(s,a)$ with $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$. ■

To prove **Theorem 5.5**, we provide **Lemma D.4** and its proof below.

Lemma D.4. *(Convergence rate of neural TD learning): Let m be the width of the actor-critic networks, and $\bar{Z}_t = \frac{1}{N}\sum_{i=1}^{N} \delta_{q_i(\boldsymbol{s},\boldsymbol{a})}$ be an estimator of Z_t^i. Suppose Assumption 2 holds, In the TD learning, with probability at least $1-\delta$, there exists*

$$\left\| \Pi_{W_1} \bar{Z}_t - \Pi_{W_1} Z_t^* \right\| \leq \Theta(m^{-\frac{H}{4}}) + \Theta([(1-\gamma)K]^{-\frac{1}{2}}[1 + \log^{\frac{1}{2}} \delta^{-1}]) \tag{15}$$

Proof: Utilizing Gluing Lemma (Villani, 2009; Clement & Desch, 2008) for Wasserstein distance W_p, where we employ the 1-Wasserstein distance W_1 and the one-dimensional quantile $q_t^{i,*}$, we establish:

$$\begin{aligned}
&\left\|\Pi_{W_1}\bar{Z}_t - \Pi_{W_1}Z_t^*\right\| = \sum_{i=1}^{N}\left\|\bar{q_t^i} - q_t^{i,*}\right\| \\
&= \sum_{i=1}^{N}\left\|f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q_{K_{td}}) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q^*})\right\| \\
&\leq \sum_{i=1}^{N}\left\|f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q_{K_{td}}) - f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q})\right\| \\
&+ \sum_{i=1}^{N}\left\|f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q_{K_{td}}) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q^*})\right\| \\
&\overset{(i)}{\leq} \Theta(m^{-\frac{H}{4}}) + \sum_{i=1}^{N}\left\|f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q_{K_{td}}) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q^*})\right\| \\
&\overset{(ii)}{\leq} \Theta(m^{-\frac{H}{4}}) + \Theta([(1-\gamma)K_{td}]^{-\frac{1}{2}}[1+\log^{\frac{1}{2}}\delta^{-1}])
\end{aligned} \tag{16}$$

where H denotes the layers of the neural network. Suppose Assumption 2 holds. Then (i) follows from **Lemma 5.1** in (Cai et al., 2019), where each quantile represents as a form of local linearization (Koenker, 2005; Gannoun et al., 2007):

$$\begin{aligned}
&\sum_{i=1}^{N}\left\|f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q_{K_{td}}) - f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q})\right\|^2 \leq \frac{1}{m^H}\sum_{i=1}^{N} b_r \\
&\left|[(\mathbf{1}(W_i^{(h)}x_i^{(h-1)}>0) - \mathbf{1}(W_i^{(0)}x_i^{(h-1)}>0))\cdot W_i^{(h)}x_i^{(h-1)}]^2\right| \\
&\leq \frac{4C_0}{m^H}\sum_{i=1}^{N}[\sum_{r=1}^{m}\mathbf{1}(\left|W_{i,r}^{(0)}x_i^{(h-1)}\right| \leq \left\|W_{i,r}^{(h)} - W_{i,r}^{(0)}\right\|_2)] \\
&\leq \frac{4C_0}{m^H}(\sum_{r=1}^{m}\left\|W_{i,r}^{(h)} - W_{i,r}^{(0)}\right\|_2^2)^{\frac{1}{2}}(\sum_{r=1}^{m}\left\|\frac{1}{W_{i,r}^{(0)}}\right\|_2^2)^{\frac{1}{2}} \leq \frac{4C_0C_1}{m^{\frac{H}{2}}}
\end{aligned} \tag{17}$$

where the constant $C_0 > 0$ and $C_1 > 0$. Thus we upper bound $\sum_{i=1}^{N}\left\|f_i^{(H)} - f_{0,i}^{(H)}\right\| \leq \Theta(m^{-\frac{H}{4}})$, which holds (i) in Equation 16. Suppose Assumption 3 holds. Then (ii) follows from **Lemma 1** in (Rahimi & Recht, 2008), with probability at least $1-\delta$, there exists:

$$\begin{aligned}
&\sum_{i=1}^{N}\left\|f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q_{K_{td}}) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q^*})\right\| \\
&\leq \frac{1}{\sqrt{1-\gamma}}\sum_{i=1}^{N}\left\|f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q_\pi}_{K_{td}}) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^{Q^*})\right\| \\
&\leq \frac{C_3}{\sqrt{(1-\gamma)K_{td}}}(1+\sqrt{\log\frac{1}{\delta}})
\end{aligned} \tag{18}$$

where (iii) holds, and therefore Equation 16 holds. ■

Theorem 5.5. *(Global Convergence Rate): Let m and H be the width and the layer of the neural network, $K_{td} = (1-\gamma)^{-\frac{3}{2}}m^{\frac{H}{2}}$ be the iterations required for convergence of the distributional TD learning (defined in Equation 10), $l_Q = \frac{1}{\sqrt{T}}$ be the policy update (in Line 4 of Algorithm 1) and $\boldsymbol{\tau}_c = \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) + \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}})$ be the tolerance (in Line 3 of Algorithm 1). Suppose Assumptions 1-3 hold.* There exists a global convergence rate of $\Theta(1/\sqrt{T})$, and a sublinear rate of $\Theta(1/\sqrt{T})$ if the constraints are violated with an error of $\Theta(1/m^{\frac{H}{4}})$, with probability at least $1-\delta$. Importantly, this conclusion holds if and only if the reward operator family $\mathcal{F}$ satisfies both **Conditions 5.3**.

Proof: **Proposition D.2** suggests that the sequence $Q^{\boldsymbol{\pi}_i}(s,a)$ achieves global convergence *if and only if* the reward operator family $\mathcal{F}$ satisfies **Conditions 5.3**. Then we let $\triangle_{\theta^Q} = \theta^Q_{t+1} - \theta^Q_t$, and suppose the critic networks are H-layer neural networks. Based on **Lemma 6.1** in (Kakade & Langford, 2002), there exists

$$
\begin{aligned}
&(1-\gamma)[\mathcal{J}_r(\boldsymbol{\pi}^*) - \mathcal{J}_r(\boldsymbol{\pi}_t)] \\
&= \mathbb{E}[Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}) - \mathbb{E}Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}')] \\
&= \mathbb{E}[\nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}} - \mathbb{E}[\nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}'),\theta^Q)^{\mathrm{T}}]]\triangle_{\theta^Q} \\
&+ \mathbb{E}[Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}) - \nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}] \\
&+ \mathbb{E}[\nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}'),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q} - Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}')] \\
&= \frac{1}{l_Q}\big[l_Q\mathbb{E}[\nabla_\theta \log(\boldsymbol{\pi}_t(\boldsymbol{a}|\boldsymbol{s}))^{\mathrm{T}}]\triangle_{\theta^Q} - \frac{l_Q^2\mathcal{L}_f}{2}\|\triangle_{\theta^Q}\|_2^2\big] \\
&+ \mathbb{E}[Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}) - \nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}] + \frac{l_Q\mathcal{L}_f}{2}\|\triangle_{\theta^Q}\|_2^2 \\
&+ \mathbb{E}[\nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}'),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q} - Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}')] \\
&\overset{(i)}{\leq} \frac{1}{l_Q}\mathbb{E}[\log(\frac{\boldsymbol{\pi}_{t+1}(\boldsymbol{a}|\boldsymbol{s})}{\boldsymbol{\pi}_t(\boldsymbol{a}|\boldsymbol{s})})] + \frac{l_Q\mathcal{L}_f}{2}\|\triangle_{\theta^Q}\|_2^2 \\
&+ \sqrt{\mathbb{E}[Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}) - f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q})]^2} \\
&+ \sqrt{\mathbb{E}[f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - \nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}]^2} \\
&+ \sqrt{\mathbb{E}[\nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}'),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q} - f^{(H)}((\boldsymbol{s},\boldsymbol{a}'),\triangle_{\theta^Q})]^2} \\
&+ \sqrt{\mathbb{E}[f^{(H)}((\boldsymbol{s},\boldsymbol{a}'),\triangle_{\theta^Q}) - Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}')]^2} \\
&= \frac{1}{l_Q}\big[\mathbb{E}[\mathcal{D}_{KL}(\boldsymbol{\pi}^*||\boldsymbol{\pi}_t)] - \mathbb{E}[\mathcal{D}_{KL}(\boldsymbol{\pi}^*||\boldsymbol{\pi}_{t+1})]\big] \\
&+ 2\sqrt{\mathbb{E}[f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - \nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}]^2} \\
&+ 2\sqrt{\mathbb{E}[Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}) - f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q})]^2} + \frac{l_Q\mathcal{L}_f}{2}\|\triangle_{\theta^Q}\|_2^2
\end{aligned}
\tag{19}
$$

where (i) follows from the $\mathcal{L}_f$-Lipschitz property of $\log(\boldsymbol{\pi}_t(\boldsymbol{a}|\boldsymbol{s}))$. Suppose Assumption 2 holds. Next, we upper bound the term $\sqrt{\mathbb{E}[f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - \nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}]^2}$ as shown below.

$$
\begin{aligned}
&\sqrt{\mathbb{E}[f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - \nabla_\theta f^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}]^2} \\
&= \sum_{i=1}^N \Big\| f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - \nabla_\theta f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}\Big\| \\
&\leq \sum_{i=1}^N \big[\Big\| f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - \nabla_\theta f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}\Big\| \\
&+ \Big\|\nabla_\theta f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q} - \nabla_\theta f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\theta^Q)^{\mathrm{T}}\triangle_{\theta^Q}\Big\|\big] \\
&= 2\sum_{i=1}^N \Big\| f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q}) - f_{0,i}^{(H)}((\boldsymbol{s},\boldsymbol{a}),\triangle_{\theta^Q})\Big\| \\
&\overset{(ii)}{\leq} \frac{4\sqrt{C_0C_1}}{m^{\frac{H}{4}}}
\end{aligned}
\tag{20}
$$

where (ii) follows from Equation 17. Suppose Assumption 3 holds. Then, in order to upper bound

$\sqrt{\mathbb{E}[Q_{\pi_t}(\boldsymbol{s},\boldsymbol{a}) - f^{(H)}((\boldsymbol{s},\boldsymbol{a}), \triangle_{\theta^Q})]^2}$, taking expectation of Equation 19 from $t=0$ to $T-1$, yields

$$
\begin{aligned}
&(1-\gamma)\big[\mathcal{J}_r(\boldsymbol{\pi}^*) - \mathbb{E}[\mathcal{J}_r(\boldsymbol{\pi})]\big] \\
&= (1-\gamma)\frac{1}{T}\sum_{t=0}^{T-1}[\mathcal{J}_r(\boldsymbol{\pi}^*) - \mathcal{J}_r(\boldsymbol{\pi}_t)] \\
&\leq \frac{1}{T}\big[\frac{1}{l_Q}\mathbb{E}[\mathcal{D}_{KL}(\boldsymbol{\pi}^*||\boldsymbol{\pi}_t)] + \frac{8T\sqrt{C_0C_1}}{m^{\frac{H}{4}}} + \frac{Tl_Q\mathcal{L}_f}{2}d_\theta^2 \\
&+ 2\sum_{t=0}^{T-1}\sum_{i=1}^{N}\Big\|f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}), \theta_{t+1}^Q - \theta_t^Q) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}), \theta^{Q^*})\Big\|\big] \\
&= \frac{\mathbb{E}[\mathcal{D}_{KL}(\boldsymbol{\pi}^*||\boldsymbol{\pi}_t)]}{l_QT} + \frac{8\sqrt{C_0C_1}}{m^{\frac{H}{4}}} + \frac{l_Q\mathcal{L}_f}{2}d_\theta^2 \\
&+ \frac{2}{T}\sum_{i=1}^{N}\Big\|f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}), \theta_{K_{td},t}^Q) - f_i^{(H)}((\boldsymbol{s},\boldsymbol{a}), \theta^{Q^*})\Big\| \\
&\overset{(iii)}{\leq} \frac{\mathbb{E}[\mathcal{D}_{KL}(\boldsymbol{\pi}^*||\boldsymbol{\pi}_t)]}{l_QT} + \frac{8\sqrt{C_0C_1}}{m^{\frac{H}{4}}} + \frac{l_Q\mathcal{L}_f}{2}d_\theta^2 \\
&+ \frac{4\sqrt{C_0C_1}}{Tm^{\frac{H}{4}}} + \frac{2C_3}{T\sqrt{(1-\gamma)K_{td}}}(1+\sqrt{\log\frac{1}{\delta}})
\end{aligned}
\tag{21}
$$

where (iii) follows from **Lemma D.4** (Equation 16). Thus, substituting $K_{td} = (1-\gamma)^{-1}m^{\frac{H}{2}}$ and $l_Q = \Theta(1/\sqrt{T})$ into Equation 21, with probability at least $1-\delta$, yields:

$$
\begin{aligned}
&\mathcal{J}_r(\boldsymbol{\pi}^*) - \mathbb{E}[\mathcal{J}_r(\boldsymbol{\pi})] \leq C_5\frac{1}{(1-\gamma)\sqrt{T}} + C_6\frac{1}{(1-\gamma)m^{\frac{H}{4}}} \\
&+ C_7\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}} + 2C_3\frac{\sqrt{\log\frac{1}{\delta}}}{(1-\gamma)Tm^{\frac{H}{4}}} \\
&\leq \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) + \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}}\sqrt{\log\frac{1}{\delta}})
\end{aligned}
\tag{22}
$$

where $C_5 = \mathbb{E}[\mathcal{D}_{KL}(\boldsymbol{\pi}^*||\boldsymbol{\pi}_t)] + \frac{\mathcal{L}_f d_\theta^2}{2}$, $C_6 = 8\sqrt{C_0C_1}$ and $C_7 = 4\sqrt{C_0C_1} + 2C_3$. Therefore, there exists:

$$
\begin{aligned}
\mathcal{J}_r(\boldsymbol{\pi}^*) - \mathbb{E}[\mathcal{J}_r(\boldsymbol{\pi})] \leq \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) \\
+ \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}}\sqrt{\log\frac{1}{\delta}})
\end{aligned}
\tag{23}
$$

where Equation 23 suggests that there exists a global convergence rate of $\Theta(1/\sqrt{T})$, with probability at least $1-\delta$.

Following *Line 6* in Algorithm 1 and recalling Equation 19, Equation 20 and Equation 21, the convergence process is similarly stated for the constraint approximation $\mathcal{J}_g^i(\boldsymbol{\pi}),\ \forall i \in [1,p]$ here

$$
\begin{aligned}
\mathbb{E}[\mathcal{J}_g^i(\boldsymbol{\pi})] - \mathcal{J}_g^i(\boldsymbol{\pi}^*) \leq \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) \\
+ \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}}\sqrt{\log\frac{1}{\delta}})
\end{aligned}
\tag{24}
$$

the constraint violation is then bounded below

$$\begin{aligned}\mathbb{E}[\mathcal{J}_g^i(\boldsymbol{\pi})] - \boldsymbol{b}_i &\leq \left[\mathcal{J}_g^i(\boldsymbol{\pi}^*) - \boldsymbol{b}_i\right] + \left[\mathbb{E}[\mathcal{J}_g^i(\boldsymbol{\pi})] - \mathcal{J}_g^i(\boldsymbol{\pi}^*)\right] \\ &\leq \boldsymbol{\tau}_c + \left[\mathbb{E}[\mathcal{J}_g^i(\boldsymbol{\pi})] - \mathcal{J}_g^i(\boldsymbol{\pi}^*)\right] \\ &\leq \boldsymbol{\tau}_c + \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) + \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}}\sqrt{\log\frac{1}{\delta}})\end{aligned} \tag{25}$$

where we have $\boldsymbol{\tau}_c = \Theta(\frac{1}{(1-\gamma)\sqrt{T}}) + \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}})$, therefore, we obtain:

$$\begin{aligned}\mathbb{E}[\mathcal{J}_g^i(\boldsymbol{\pi})] - \boldsymbol{b}_i \leq \Theta(&\frac{1}{(1-\gamma)\sqrt{T}}) \\ &+ \Theta(\frac{1}{(1-\gamma)Tm^{\frac{H}{4}}}\sqrt{\log\frac{1}{\delta}})\end{aligned} \tag{26}$$

where Equation 26 suggests that there exists a sublinear rate of $\Theta(1/\sqrt{T})$ if the constraints are violated with an error of $\Theta(1/m^{\frac{H}{4}})$, with probability at least $1-\delta$. ■

E. Experiment Supplementary

E.1. Experimental Setting

The parameter setting of AWaVO is shown in Table 2, and the technical specification of the quadrotor is shown in Table 3.

Table 2. Parameter Setting of AWaVO

Parameters	Definition	Values
$l_{\mu,cart}$	Learning rate of actor in Cartpole (Xu et al., 2021)	0.0005
$l_{\theta,cart}$	Learning rate of critic in Cartpole (Xu et al., 2021)	0.0005
$l_{\mu,acro}$	Learning rate of actor in Acrobot (Xu et al., 2021)	0.005
$l_{\theta,acro}$	Learning rate of critic in Acrobot (Xu et al., 2021)	0.005
$l_{\mu,guard}$	Learning rate of actor in Walker and Drone (Zhao et al., 2023)	0.001
$l_{\theta,guard}$	Learning rate of critic in Walker and Drone (Zhao et al., 2023)	0.001
μ	Actor neural network: fully connected with H hidden layers (m neurons per hidden layer)	-
θ	Critic neural network: fully connected with H hidden layers (m neurons per hidden layer)	-
D	Replay memory capacity	10^6
B	Batch size	128
γ	Discount rate	0.998
m	the width of neural network	128
H	the layer of neural network	2
T	Length in each episode	500
N	Time steps	20

Table 3. Technical Specification of Hardware

No.	Component	Specific Model
1	Frame	QAV250
2	Sensor - Depth Camera	Intel RealSense D435i
3	Sensor - Down-view Rangefinder	Holybro ST VL53L1X
4	Flight Controller	Pixhawk 4
5	Motors	T-Motor F60 Pro IV 1750KV
6	Electronic Speed Controller	BLHeli-32bit 45A 3-6s
7	On-board Companion Computer	DJI Manifold 2-c (CPU Model: Intel Core i7-8550U)
8	Mounts	3D Print for Sensors/ Computer/Controller/Battery

E.2. Task Descriptions in the Simulated Platforms

Acrobot and Cartpole tasks in OpenAI Gym. In Cartpole (Brockman et al., 2016), the pole movement is constrained within the range of $[-2.4, 2.4]$. Each episode has a maximum length of 200 steps and is terminated if the angle of the pole exceeds 12 degrees. During training, the agent receives a reward of $+1$ for each step taken. However, it incurs a penalty of $+1$ if (*i*) it enters the areas $[-2.4, -2.2]$, $[-1.3, -1.1]$, $[-0.1, 0.1]$, $[1.1, 1.3]$, or $[2.2, 2.4]$, or (*ii*) the angle of the pole exceeds 6 degrees.

In Acrobot (Brockman et al., 2016), the agent is rewarded for swinging the end-effector at a height of 0.5, where each episode has a maximum length of 500 steps. Conversely, it faces a penalty if (*i*) torque is applied to the joint when the first pendulum swings in an anticlockwise direction, or (*ii*) if the second pendulum swings in an anticlockwise direction with

respect to the first pendulum.

Walker and Drone tasks in GUARD. Walker (Zhao et al., 2023), a bipedal robot, comprises four primary components: a torso, two thighs, two legs, and two feet. Notably, unlike the knee and ankle joints, each hip joint possesses three hinges in the x, y, and z coordinates, enabling versatile turning. Maintaining a fixed torso height, Walker achieves mobility through the control of 10 joint torques.

Drone in GUARD (Zhao et al., 2023) is designed to emulate a quadrotor, simulating the interaction between the quadrotor and the air by applying four external forces to each of its propellers. These external forces are configured to counteract gravity when no control actions are applied. To maneuver in three-dimensional space, the Drone utilizes four additional control forces applied to its propellers.

E.3. Further Details on Constraint Limit Setting

In accordance with the benchmark (Xu et al., 2021), we established the constraint limit as 50 in Acrobot, as depicted in Figure 3(a). In the remaining scenarios, namely Cartpole in Figure 3(b), Walker in Figure 3(c), Drone in Figure 3(d), and the real quadrotor in Figure 5, the constraint limit serves as a lower boundary, indicating the level of tolerance the constraints can endure. The agent's stable performance for specific tasks occurs when it operates below this constraint limit. We hypothesize that there may be potential benefits in establishing a fixed limit, b_i, by decoupling the cumulative value into specific fixed limits. This is left for future work.

E.4. Further Discussion on Probabilistic Interpretation of Sequential Decisions

Curves in Figure 7 (a) and (b). The curves in Figure 7 (a), provided as a reference, give the estimated values of external aerodynamic forces (winds) in real-time. These estimates are derived from the signals collected from onboard sensors. In Figure 7 (b), the curves illustrate the quantitative impact of external forces L_0 on current sequential decisions, specifically, the planned trajectory τ. This impact is quantified as parts of pulse width modulation signals that are fed into the motors to either resist or cooperate with the measured (or identified) aerodynamic forces. Figure 7 (b) aims to quantitatively interpret and visually convey the decision-making process in response to external forces.

Figure 7 (b) becomes particularly important when the agent makes sub-optimal decisions leading to events like quadrotor crashes or collisions. These curves prove valuable for interpreting and performing quantitative analyses of distinct environmental factors, such as winds and obstacles, allowing for an understanding of their magnitudes of influence on the current decision-making process.

An instance to interpret Figure 7 (b). In the case of Reference State 02 (RS 02), located in an area with a combination of wind and obstacles, both aerodynamic effects (i.e., external forces) from winds and obstacles act simultaneously on the quadrotor. In Flight Task 3 (FT 3), represented by the red curve, we can observe the influence of external forces (i.e., aerodynamic effects from winds and obstacles) on the current trajectory planning decisions. The value is approximately 0.40 at RS 02, implying that the ongoing trajectory planning decisions have a $\sim 40\%$ probability of being influenced by the aerodynamic effects.

Comparing FT 1 and FT 2, where the values at RS 02 are approximately 0.20 (FT 1) and 0.18 (FT 2), respectively, we can decouple the aerodynamic effects generated by the wind on the body (FT 1) and obstacles (FT 2). Quantitatively, at RS 02, situated in an area with a mix of wind and obstacles, the red $p(\tau|L_0)_{FT3}$ is approximately equal to the sum of $p(\tau|L_0)_{FT1}$ (only wind) and $p(\tau|L_0)_{FT2}$ (only obstacles).

E.5. Scalability

To explore the full scalability of our methodology, we have implemented it on the Lorenz 96 system (Lorenz & Emanuel, 1998; Gorbach et al., 2017), characterized by the following equations: $dx_i/dt = (x_{i+1} - x_{i-2}) \cdot x_{i-1} - x_i + f_\theta$, where f_θ represents a scalar forcing parameter, and $x_{-1} = x_{I-1}, x_0 = x_I, x_{I+1} = x_1$, with I denoting the number of states in the deterministic system. This system serves as a simplified model for weather forecasting. We choose it as a versatile framework for expanding the number of states in our inference task. In our experiment, we range from 50 to 1000 states, with approximately one-third of the states randomly designated as unobserved.

We show the results in Figure 8 and Figure 9. Our methodology successfully infers a system with 1000 states in less than 400 seconds (shown in Figure 9). From visual inspection, we observe that the unobserved states are inferred, with

the approximation error remaining consistent regardless of the problem's dimensionality. Given that many real-world RL problems involve state spaces significantly larger than 1,000 states, our future work will explore such scalability concerns on larger models.

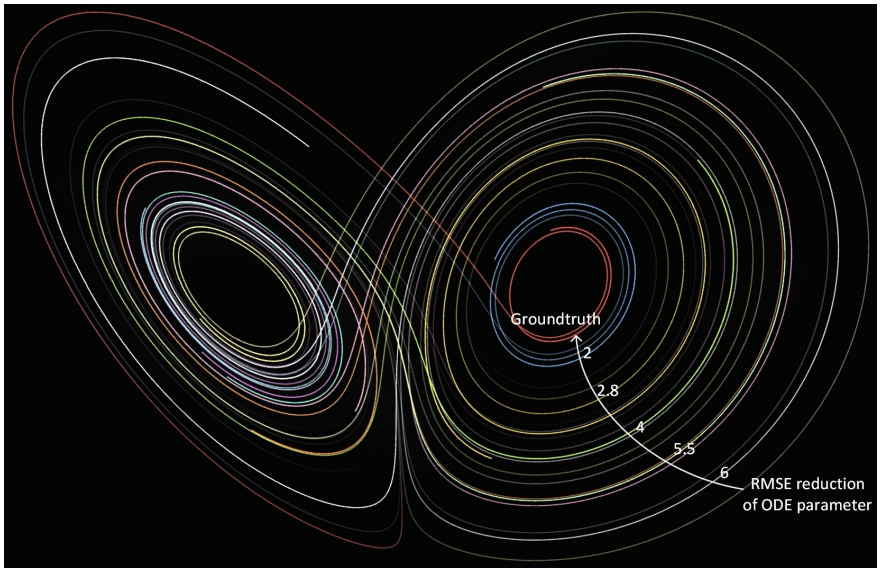


Figure 8. Scalability performance on Lorenz 96 system (Lorenz & Emanuel, 1998; Gorbach et al., 2017).

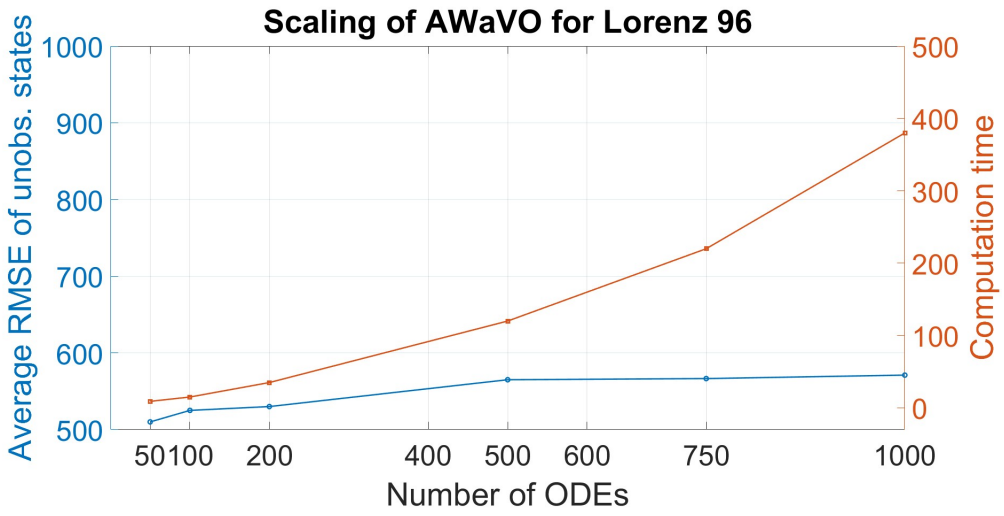


Figure 9. Average RMSE and computation time of the unobserved state for scalability.