When Distance Matters: Wasserstein Trust Regions for Multi-Agent Coordination

Chirayu Salgarkar
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester, NY 14623, USA
cms8111@rit.edu

Ali Baheri
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester, NY 14623, USA
akbeme@rit.edu

Abstract

This paper presents a new trust region optimization approach for cooperative multi-agent reinforcement learning through the incorporation of optimal transport. We replace traditional KL-divergence constraints with the Wasserstein-1 distance to define trust regions, using a dual formulation to transform the constrained optimization problem into a tractable problem over single non-negative dual variables per agent. We also introduce a coordination-aware adaptive trust-region (CAATR) mechanism, adjusting each agent's trust-region radius inversely proportional to teammate policy drift. The resulting Wasserstein multi-agent trust-region policy optimization (W-MATRPO) algorithm provides surrogate objective bounds through sequential optimization. Theoretical analysis establishes performance bounds for the multi-agent setting, and experimental analysis demonstrates improved exploration in an environment with local optima traps.

Introduction

Learning effective coordination strategies for multiple agents remains a central challenge in robotics, autonomous systems, and distributed control. MARL has become a valuable tool for modeling such systems with multiple decision-making entities each pursuing individual goals, with use-cases spanning autonomous traffic networks with vehicle coordination requirements (Zhang et al. 2024) or dynamic multi robot locomotion and planning (Orr and Dutta 2023). When multiple agents must work together to achieve a common goal, cooperative MARL provides a framework for learning behaviors to enable adaptive agent coordination based on experience and key objectives. However, cooperative MARL is not without its challenges. Factors hindering perfect coordination include the moving-target/non-stationarity problem (Papoudakis et al. 2019) (Hernandez-Leal et al. 2017), curse of dimensionality (Hao et al. 2022), and the credit-assignment problem (Zhou et al. 2020) (Nguyen, Kumar, and Lau 2018) in cooperative settings. For an overview of MARL, we refer readers to the surveys by (Busoniu, Babuska, and De Schutter 2008) (Oroojlooy and Hajinezhad 2023) (Canese et al. 2021).

Recently, trust-region methods have been used to attempt to rectify these aforementioned factors (Li et al. 2021) (Sun et al. 2022) (Li and He 2023). Trust-region methods constrain how much a policy can change per update step, in order to prevent overtly aggressive updates that could destabilize learning. One way to enforce such a method is through the use of Kullback-Leibler (KL) divergence (also known as relative entropy), as is done in trust region policy optimization (Schulman et al. 2015a). Designing trust-region policy optimization algorithms for MARL is difficult as it requires agents to coordinate their policy updates while maintaining monotonic improvement guarantees. Classical trust-region algorithms have been extended to the cooperative multi-agent setting by leveraging game-theoretic approaches (Wen et al. 2022) and by transforming the policy-update rule (Li and He 2023). Notably, the heterogeneous-agent trust region policy optimization (HATRPO) algorithm (Kuba et al. 2021) provided the first formal monotonic improvement guarantee in cooperative MARL through sequential updates bounded by KL divergence.

KL divergence has two very attractive benefits; (1) there exists a classical inequality that relates KL divergence to the total variation distance between distributions, known as Pinsker's inequality (Pinsker 1964) (Fedotov, Harremoës, and Topsoe 2003) and (2) there exists a closed-form solution for the KL divergence between two multivariate Gaussians (Li 2018). However, the use of KL divergence is not without its limitations. KL divergence is asymmetric (it is not a metric), implying that different update magnitudes could exist depending on the direction of comparison, which could skew the policy optimization techniques. Some research has also argued that KL divergence is not well-suited for knowledge distillation, as it cannot capture relationships between different categories and is poor at dealing with high-dimensional feature spaces (Lv, Yang, and Li 2024). Similarly, in unsupervised learning, when dealing with low dimensional manifolds, KL divergence likely fails as the model manifold and true distribution's support is often zero leading to an undefined KL divergence (Arjovsky, Chintala, and Bottou 2017). Both of these issues permeate MARL, where agents must learn distinct but coordinated policies (similar to these category relationships), often having been assigned various roles (Wang et al. 2020) (Makar, Mahadevan, and Ghavamzadeh 2001). These limitations motivate our exploration of alternative divergence metrics, namely from optimal transport theory.

Most generally, optimal transport refers to the mathematical problem of moving a distribution of mass from one location to another as efficiently as possible (Peyré, Cuturi, and others 2019). The Wasserstein distance is the most ubiquitous metric between probability distributions originally derived from the optimal transport problem. This metric, as well as the development of the Sinkhorn distance, a regularized approximation of the Wasserstein distance (Cuturi 2013) has been instrumental for providing a geometrically intuitive way to compare probability distributions, particularly useful in machine learning and reinforcement learning. (We refer the reader to the works of (Kolouri et al. 2017) (Peyré, Cuturi, and others 2019) (Villani and others 2008) for an overview of optimal transport and its applications in machine learning.) Recent literature has replaced KL divergence with Wasserstein distance for single-agent policy optimization settings (Pacchiano et al. 2020) (Terpin et al. 2022) (Song, Zhao, and He 2022); however, the integration of Wasserstein constraints in multi-agent policy optimization remains unexplored. It is difficult to maintain coordination guarantees while handling Wasserstein constraints among multiple interacting agents.

In this paper, we develop a Wasserstein-constrained multi-agent trust region algorithm by formulating a dual optimization framework that preserves multi-agent coordination while exploiting the geometric advantages of optimal transport. Our approach transforms the intractable primal problem into a tractable per-agent dual problem, where each agent solves for a single scalar dual variable that automatically balances exploration with coordination constraints. Additionally, we introduce a coordination-aware adaptive trust-region (CAATR) mechanism that dynamically adjusts each agent's trust-region radius based on teammate policy drift. The derived algorithm produces a multi-agent policy optimization scheme that: (1) provides a computationally tractable solution to Wasserstein-constrained multi-agent optimization through dual decomposition, (2) provides theoretical guarantees on surrogate objective improvement, and (3) demonstrates improved exploration and convergence to near-global optima through coordination-aware adaptation, avoiding local optima traps observed in fixed trust-region methods. Later in this paper, we discuss the success of our W-MATRPO + CAATR algorithm at escaping local optima traps in a differential game environment.

Contributions. Our main contributions are as follows:

- We derive a tractable dual formulation for Wasserstein-constrained policy optimization (W-MATRPO) in cooperative MARL.
- We introduce CAATR, a method to adaptively modulate trust region radius based on teammate stability.
- We provide theoretical analysis establishing surrogate objective bounds and demonstrate empirically that W-MATRPO+CAATR can escape local optima outperforming conventional methods.

Paper Organization. The remainder of this paper is structured as follows: Section 2 presents Related works. Section 3 presents the methodology, including preliminaries on Dec-POMDPs, the Wasserstein-constrained policy optimization dual formulation, and the W-MATRPO algorithm with CAATR. Section 4 establishes theoretical results including regularity conditions and surrogate objective bounds. Section 5 presents numerical experiments on differential games. Section 6 analyzes the results demonstrating W-MATRPO+CAATR's ability to escape local optima. Section 7 discusses limitations and future directions, and Section 8 concludes the paper.

Related Work

We first compare our work with the most closely related papers that exist in the literature. (Terpin et al. 2022) developed single-agent TRPO with Wasserstein constraints using dual optimization. They developed a one-dimensional dual reformulation for the infinite-dimensional optimization problem in policy optimization, as well as an optimal policy update for the dual problem. While their work is instrumental for motivating our research, it is important to note that their approach has not been extended to the multi-agent setting, and we provide a different method for developing the dual formulation. (Song, Zhao, and He 2022) (Pacchiano et al. 2020) also study single-agent policy optimization through the incorporation of the Wasserstein distance. These setups are (1) single-agent, and are not necessarily generalizable to multi-agent coordination problems, and (2) do not study adaptive trust regions, only focusing on fixed trust regions. (Shek, Shi, and Tokekar 2025) considers the multi-agent trust-region policy optimization problem with adaptive trust regions through both a Karush-Kuhn-Tucker-based method and a greedy algorithm. This setup differs from our approach for two reasons: (1) this approach does not use the notions of optimal transport in the algorithm and (2) we use a different algorithm for adaptively generating the trust region, as we use a setup that considers prior policy updates while their approach allocates the KL divergence amongst agents. *On multi-agent trust region methods:* Trust region policy optimization (TRPO), first described in (Schulman et al. 2015a) is a method to optimize policies iteratively with guaranteed monotonic improvement through KL-constrained updates. Extending these guarantees in multi-agent settings is challenging due to the non-stationarity problem (Li and He 2023) (Matignon, Laurent, and Le Fort-Piat 2012) absent in the single-agent setting. One of the first papers to tackle this implementation was (Kuba et al. 2021), through the incorporation of sequential policy updates with agent-specific KL constraints. More specifically, (Kuba et al. 2021) (Zhong et al. 2024a) extended TRPO to multi-agent settings with sequential updates and monotonic improvement guarantees. This aforementioned algorithm is especially useful when agents exhibit heterogeneity, such as individual roles. Subsequently, this framework has been improved to include adaptivity (Shek, Shi, and Tokekar 2025) as well as the reformulation of safe MARL as a constrained Markov game, solved with multi-agent constrained policy optimization (MACPO) (Gu et al. 2022).

Other metrics have also been considered in the MARL problem. (Nasiri and Rezghi 2023) uses the Bregman divergence to implement heterogeneous-agent reinforcement learning through Mirror Descent Policy Optimization, and (Zawar, Sethi, and Roy 2024) extended the MACPO (Gu et al. 2022) approach to use Jensen-Shannon divergence rather than KL divergence. (Zawar, Sethi, and Roy) notes that they did not use the Wasserstein distance due to intractability, claiming Jensen-Shannon divergence as a "middle ground" (Zawar, Sethi, and Roy); we use a dual formulation to solve the MARL problem using the Wasserstein distance while maintaining computational tractability. *On optimal transport:* Optimal transport theory compares probability distributions by measuring minimum cost of mass transport from one distribution to another (Villani and others 2008) (Peyré, Cuturi, and others 2019). The Wasserstein distance (optimal transport metric) holds some advantages over KL divergence: (1) it is a true metric and is well-defined over non-overlapping supports, and (2) it incorporates the geometry of the underlying action space through the ground metric (Villani and others 2008). Cuturi's Sinkhorn algorithm (Cuturi 2013) used entropic regularization to make the Wasserstein distance tractable, greatly increasing its potential for machine learning research. Recent advances in dual and subdual methods have further improved scalability, yielding smooth variational problems amenable to numerical optimization (Cuturi and Peyré 2018) (Cuturi and Peyré 2016) (Khamis et al. 2024).

Optimal transport theory has found applications as a tool across many domains of reinforcement learning. (Baheri 2023) establishes a framework using optimal transport optimizing rewards while maintaining risk constraints, while (Baheri and Kochenderfer 2024) notes a potential synergy between optimal transport theory and the MARL, particularly in policy alignment and in addressing non-stationarity. Other methods by which optimal transport was used in RL include Wasserstein unsupervised reinforcement learning (He et al. 2022), distributional RL, curriculum learning, robustness, and imitation learning. In distributional RL, the Bellman operator is notably a contraction in the Wasserstein metric (Bellemare, Dabney, and Munos 2017). In curriculum learning, (Klink et al. 2022) framed curricula (the sequence of learning tasks) as interpolations between task distributions through a constrained optimal transport problem. In robustness, (Abdullah et al. 2019) formalized RL as a min-max game with a Wasserstein constraint, and (Hou et al. 2020) used the Wasserstein distance to measure disturbances of a reference transition kernel. In imitation learning, (Xiao et al. 2019) used Kantorovich potential as a reward function and found a connection between inverse RL and optimal transport. The most relevant papers at the intersection of optimal transport and trust region policy optimization include (Terpin et al. 2022) (Song, Zhao, and He 2022) (Pacchiano et al. 2020) described earlier. Our paper uses optimal transport theory in a multi-agent setting, while also incorporating a novel adaptive trust region setup, uniquely positioning this work in both the optimal transport and MARL domains.

Methodology

Preliminaries

We briefly summarize our notation. Cooperative multi-agent tasks are formally modeled as decentralized partially-observable Markov decision processes (Dec-POMDPs), defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \Omega, P, r, O, \gamma \rangle$ (Kaelbling, Littman, and Cassandra 1998)(Schulman et al. 2015a). We denote by $\mathcal{N} = \{1, \dots, N\}$ the set of N agents. The global state space is $\mathcal{S}$, with joint action space $\mathcal{A}$ and joint observation space Ω. Each agent i has its individual action space $\mathcal{A}_i$ and observation space Ω_i. The environment dynamics are characterized by the state transition probability function $P(s'|s, a)$, which determines the probability of transitioning to state s' given current state s and joint action a. The shared reward function $r(s, a)$ provides the immediate reward for all agents. The observation function $O(o|s', a)$ determines the probability of joint observation o given the new state s' and previous action a. Finally, $\gamma \in [0, 1]$ is the discount factor for future rewards.

At each timestep t, the environment is in global state $s_t \in \mathcal{S}$. The agents undergo a joint action $a_t = (a_1, ..., a_N) \in \mathcal{A}$, causing a transition to state s_{t+1} according to $P(s_{t+1}|s_t, a_t)$ yielding a shared reward $r_{t+1} = r(s_t, a_t)$. Agents do not observe s_t directly, but receive local observations $o_{i,t} \in \Omega_i$ according to the observation function $O(o_t|s_{t+1}, a_t)$. Each agent i operates using a local, parameterized policy $\pi_{\theta_i}(a_i|o_i)$. The collective goal is to find the parameters $\theta = \{\theta_1, ..., \theta_N\}$ for the joint policy $\pi_\theta(a|s) = \prod_{i=1}^{N} \pi_{\theta_i}(a_i|o_i)$ that maximizes the expected total discounted return:

$$J(\pi_\theta) = \mathbb{E}_{\substack{s_0 \sim p_0, a_t \sim \pi_\theta(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \tag{1}$$

This problem is addressed within the centralized training for decentralized execution (CTDE) paradigm, where a centralized learner uses global information during training to optimize decentralized policies (Amato 2024).

Wasserstein-Constrained Policy Optimization

Trust-region methods in single-agent reinforcement learning are valued for their guaranteed monotonic improvement (Schulman et al. 2015b), i.e. policy performance can be improved by maximizing a surrogate objective within its constrained policy neighborhood. We define this constrained policy neighborhood using the Wasserstein-1 distance, W_1. For two probability distributions μ and ν on the action space $\mathcal{A}$, the W_1 distance is defined as follows:

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{A} \times \mathcal{A}} c(a, a') \, d\gamma(a, a') \tag{2}$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions (couplings) with marginals μ and ν, and $c(a, a')$ is a transport cost function. Setting $c(a, a')$ to be a metric on the action space (e.g. Euclidean distance), the W_1 distance provides a geometrically meaningful measure of the difference between policies. This contrasts with KL divergence, which is a purely information-theoretic view on policy change, insensitive to the inherent metric of the action space.

Applying the sequential update scheme of HATRPO (Zhong et al. 2024b), we formulate the per-agent policy improvement step as an optimization problem:

$$\begin{aligned} \max_{\pi_i^{\text{new}}} \quad & \mathbb{E}_{s,a_{-i}}[A^{\pi^{\text{old}}}(s,a_i,a_{-i})] \\ \text{s.t.} \quad & \mathbb{E}_{s\sim\rho}[W_1(\pi_i^{\text{old}}(\cdot|s),\pi_i^{\text{new}}(\cdot|s))] \leq \delta_i \end{aligned} \tag{3}$$

Here, the objective is to maximize the expected advantage of agent i's new policy, subject to the constraint that the average Wasserstein distance between the new and old policies remains within a trust-region of radius δ_i. The optimization problem (3) is intractable due to the infinite-dimensional policy space and the complexity of computing the Wasserstein constraint. We therefore develop a Lagrangian dual formulation that transforms this problem into an optimization over a single non-negative dual variable $\lambda_i \geq 0$ for each agent.

Theorem 1 (Dual Formulation)**.** *The Wasserstein constrained policy optimization problem admits the dual representation:*

$$\min_{\lambda_i\geq 0}\left\{\lambda_i\delta_i+\mathbb{E}_{s,a\sim\pi^{old}}\left[\Phi^{old}_{\lambda_i}(s,a_i)\right]\right\} \tag{4}$$

where

$$\Phi^{old}_{\lambda_i}(s,a_i) := \max_{a_i'\in\mathcal{A}_i}\left\{A^{\pi^{old}}(s,a_i',a_{-i})-\lambda_i c(a_i,a_i')\right\} \tag{5}$$

is the λ-regularized advantage function. Furthermore, strong duality holds between the primal and dual problems.

Proof. The primal problem is a concave maximization over a convex constraint set: the objective $\mathbb{E}_{s,a_{-i}}[A^{\pi^{\text{old}}}(s,a_i,a_{-i})]$ is linear in π_i^{new} (hence both convex and concave), and the Wasserstein constraint $\mathbb{E}_{s\sim\rho}[W_1(\pi_i^{\text{old}}(\cdot|s),\pi_i^{\text{new}}(\cdot|s))] \leq \delta_i$ defines a convex set. Setting $\pi_i^{\text{new}} = \pi_i^{\text{old}}$ yields $W_1(\pi_i^{\text{old}},\pi_i^{\text{old}}) = 0 < \delta_i$, satisfying Slater's condition for the constraint. By standard Lagrangian duality for constrained maximization problems, strong duality holds.

The Lagrangian for the maximization problem is

$$\begin{aligned}\mathcal{L}(\pi_i^{\text{new}},\lambda_i) =& \mathbb{E}_{s,a_{-i}}[A^{\pi^{\text{old}}}(s,a_i,a_{-i})] \\ & -\lambda_i\left(\mathbb{E}_{s\sim\rho}[W_1(\pi_i^{\text{old}}(\cdot|s),\pi_i^{\text{new}}(\cdot|s))]-\delta_i\right)\end{aligned} \tag{6}$$

The dual function is:

$$\begin{aligned} g(\lambda_i) &= \sup_{\pi_i^{\text{new}}}\mathcal{L}(\pi_i^{\text{new}},\lambda_i) \\ &= \lambda_i\delta_i + \sup_{\pi_i^{\text{new}}}\mathbb{E}_{s\sim\rho}\bigg[\mathbb{E}_{a_{-i}}\int_{\mathcal{A}_i}A^{\pi^{\text{old}}}(s,a_i,a_{-i})\,d\pi_i^{\text{new}}(a_i|s) \\ &\quad -\lambda_i W_1(\pi_i^{\text{old}}(\cdot|s),\pi_i^{\text{new}}(\cdot|s))\bigg] \end{aligned} \tag{7}$$

Since the optimization over $\pi_i^{\text{new}}(\cdot|s)$ can be performed independently for each state s (the conditional policies are separable), we can exchange the order of expectation and supremum. Define

$$\bar{A}_i(s,a_i) := \mathbb{E}_{a_{-i}\sim\pi_{-i}^{\text{old}}(\cdot|s)}[A^{\pi^{\text{old}}}(s,a_i,a_{-i})] \tag{8}$$

as the advantage function marginalized over other agents' actions. Since the expectation over a_{-i} does not depend on the choice of π_i^{new}, the inner supremum becomes

$$\begin{aligned}\sup_{\pi_i^{\text{new}}(\cdot|s)}\bigg\{&\int_{\mathcal{A}_i}\bar{A}_i(s,a_i)\,d\pi_i^{\text{new}}(a_i|s) \\ &-\lambda_i W_1(\pi_i^{\text{old}}(\cdot|s),\pi_i^{\text{new}}(\cdot|s))\bigg\}\end{aligned} \tag{9}$$

By Kantorovich duality for the W_1 Wasserstein distance (Kantorovich and Rubinshtein 1958), the supremum of a linear functional minus a transport cost admits the dual representation

$$\begin{aligned}&\sup_{\pi_i^{\text{new}}(\cdot|s)}\left\{\int_{\mathcal{A}_i}\bar{A}_i(s,a_i)\,d\pi_i^{\text{new}}(a_i|s)-\lambda_i W_1(\pi_i^{\text{old}}(\cdot|s),\pi_i^{\text{new}}(\cdot|s))\right\} \\ &\quad = \int_{\mathcal{A}_i}\max_{a_i'\in\mathcal{A}_i}\left[\bar{A}_i(s,a_i')-\lambda_i c(a_i,a_i')\right]d\pi_i^{\text{old}}(a_i|s)\end{aligned} \tag{10}$$

where the maximization is taken pointwise over the cost function c.

Substituting this back into the dual function and using the definition of $\bar{A}_i$, we have

$$\begin{aligned} g(\lambda_i) = \lambda_i\delta_i + \mathbb{E}_{s\sim\rho}\mathbb{E}_{a_i\sim\pi_i^{\text{old}}(\cdot|s)}\Bigg[&\max_{a_i'\in\mathcal{A}_i} \\ &\left[\mathbb{E}_{a_{-i}\sim\pi_{-i}^{\text{old}}(\cdot|s)}[A^{\pi^{\text{old}}}(s,a_i',a_{-i})]-\lambda_i c(a_i,a_i')\right]\Bigg]\end{aligned} \tag{11}$$

Exchanging the order of the maximum and the expectation over a_{-i} (valid since expectation preserves suprema), we obtain

$$g(\lambda_i) = \lambda_i\delta_i + \mathbb{E}_{s,a\sim\pi^{\text{old}}}\left[\Phi^{\text{old}}_{\lambda_i}(s,a_i)\right] \tag{12}$$

The dual problem $\min_{\lambda_i\geq 0} g(\lambda_i)$ yields the stated result. □

Remark 1. *Intuitively, the inner maximization process derives a new action a_i' that balances maximizing advantage against the transport cost to that action, scaled by λ_i. The outer minimization finds the optimal trade-off parameter λ_i that satisfies the trust-region constraint. The dual formulation avoids explicit computation of the Wasserstein distance integral during the policy optimization step, instead requiring only the solution to the inner maximization problem over a_i'. For continuous action spaces, the inner maximization $\max_{a_i'}\{A^{\pi^{old}}(s,a_i',a_{-i})-\lambda_i c(a_i,a_i')\}$ can be solved using: (1) gradient ascent when the advantage and cost functions are differentiable, (2) cross-entropy method (CEM) for non-differentiable cases, or (3) closed-form solutions for Gaussian policies with L_2 cost, where the optimal action is $a_i' = a_i + \nabla_{a_i}A/(2\lambda_i)$. For discrete action spaces, exhaustive search is tractable.*

W-MATRPO Algorithm

We solve the dual problem using an actor-critic architecture and a sequential update scheme. We use N decentralized actor networks, $\pi_{\theta_i}(a_i|o_i)$, which take only local observations as input. During training, a centralized critic, composed of a state-value network $V_\phi(s)$ and a state-action value network $Q_\phi(s,a)$, uses global information to compute a shared, low-variance estimate of the advantage function, $A^{\pi_{\text{old}}}(s,a) \approx Q_\phi(s,a) - V_\phi(s)$.

The W-MATRPO algorithm (Algorithm 1) begins by initializing the joint policy $\pi^{(0)}$ and policy drift history (line 1). At each iteration t (line 2), the algorithm collects trajectories by executing the current joint policy in the environment and computes the advantage function $A^{\pi^{(t)}}$ using the centralized critic (line 3). Next, it calculates adaptive trust-region radii $\{\delta_i^{(t+1)}\}$ for each agent using the CAATR mechanism detailed in Algorithm 2 (line 4). We then randomly order the agents by generating a random permutation σ of the agent indices $\{1,\dots,N\}$ (line 5) before initializing an empty set $\mathcal{U}_k$ to track updated agents (line 6). For each agent in this random order (lines 7-13), the algorithm retrieves the agent index i as the k-th element of the permutation, i.e., $i = \sigma(k)$ (line 8), then computes the importance-corrected advantage $M_i(s,a)$ using the set of previously updated agents (line 9). Next, the algorithm solves the dual optimization problem to find the optimal Lagrange multiplier λ_i^* that minimizes the dual objective (line 10). Using this optimal dual variable, the agent's policy is updated from $\pi_i^{(t)}$ to $\pi_i^{(t+1)}$ (line 11), and importance sampling correction is applied (line 12). The updated agent is then added to the set $\mathcal{U}_k$ (line 13). After all agents have been updated, the centralized critic parameters ϕ are trained using the collected trajectories (line 14). Finally, the algorithm stores the policy drift measurements $\text{Drift}_j^{(t)}$ for all agents, computed as the Wasserstein distance between consecutive policies (line 15).

Coordination-Aware Adaptive trust-region (CAATR) We introduce a feedback mechanism to adapt the trust-region radius δ_i based on teammate policy. The CAATR mechanism adjusts each agent's exploration constraints based on the collective behavior of its teammates. The radius for agent i at iteration t is set to be inversely proportional to the measured policy drift of its teammates:

$$\delta_i^{(t)} = \frac{C}{\sum_{j\neq i} W_1(\pi_j^{(t-1)}, \pi_j^{(t-2)}) + \epsilon} \tag{13}$$

where C is a positive hyperparameter and ϵ is adaptively set as

$$\epsilon = \max(\epsilon_{\text{base}}, \min(\epsilon_{\max}, D_{-i}/10)) \tag{14}$$

where D_{-i} is the teammate drift sum. Here, ϵ_{base} is the minimum regularization value preventing division by zero when all teammates' policies are perfectly stable, and $\epsilon_{\max}$ is the maximum regularization value that prevents the trust region from becoming too small even when teammates exhibit high drift.

The CAATR update procedure (Algorithm 2) takes as input the current and previous policies for all agents, along with hyperparameters C, $\varepsilon_{\text{base}}$, and $\varepsilon_{\max}$. For each agent j (lines 1-3), it computes the policy drift $\text{Drift}_j^{(t)}$ as the expected Wasserstein distance between the agent's current policy $\pi_j^{(t)}$ and previous policy $\pi_j^{(t-1)}$ (line 2). Then, for each agent i (lines 4-8), the algorithm computes the teammate drift sum D_{-i} by summing the drifts of all other agents except agent i (line 5). ϵ is adaptively set based on the teammate drift (line 6). We then set agent i's adaptive trust-region radius as $\delta_i^{(t+1)} = C/(D_{-i} + \varepsilon)$ (line 7). Finally, the algorithm returns the complete set of adaptive trust-region radii for all agents (line 9).

Algorithm 1 W-MATRPO with CAATR

```
Require: Initial policy $\pi^{(0)}$, parameters $C, \varepsilon_{\text{base}}, \varepsilon_{\max}$
 1: Initialize policy drift history
 2: for $t = 0, 1, 2, \dots$ do
 3:     Collect trajectories and compute joint advantage $A^{\pi^{(t)}}$
 4:     Compute trust-regions $\{\delta_i^{(t+1)}\}$ using Algorithm 2
 5:     Randomly order agents $\sigma$
 6:     Initialize $\mathcal{U}_k = \emptyset$ (set of updated agents)
 7:     for $k = 1$ to $N$ do
 8:         $i \leftarrow \sigma(k)$
 9:         Compute $M_i(s,a)$ using $\mathcal{U}_k$
10:         Solve dual problem for agent $i$: $\lambda_i^* \leftarrow \arg\min_{\lambda_i \geq 0} g(\lambda_i)$
11:         Update policy $\pi_i^{(t+1)}$ with $\lambda_i^*$
12:         Apply importance sampling correction
13:         $\mathcal{U}_k \leftarrow \mathcal{U}_k \cup \{i\}$
14:     end for
15:     Update centralized critic parameters $\phi$
16:     Store policy drift $\text{Drift}_j^{(t)}$ for all $j$
17: end for
```

Algorithm 2 CAATR trust-region update

```
Require: Current policies $\{\pi_j^{(t)}\}$, previous policies $\{\pi_j^{(t-1)}\}$, $C, \varepsilon_{\text{base}}, \varepsilon_{\max}$
Ensure: Adaptive trust-region radii $\{\delta_i^{(t+1)}\}$
 1: for $j = 1$ to $N$ do
 2:     Compute policy drift: $\text{Drift}_j^{(t)} \leftarrow \mathbb{E}[W_1(\pi_j^{(t)}, \pi_j^{(t-1)})]$
 3: end for
 4: for $i = 1$ to $N$ do
 5:     Compute teammate drift sum: $D_{-i} \leftarrow \sum_{j\neq i} \text{Drift}_j^{(t)}$
 6:     Set adaptive epsilon: $\varepsilon \leftarrow \max(\varepsilon_{\text{base}}, \min(\varepsilon_{\max}, D_{-i}/10))$
 7:     Set adaptive trust-region: $\delta_i^{(t+1)} \leftarrow \frac{C}{D_{-i}+\varepsilon}$
 8: end for
 9: return $\{\delta_1^{(t+1)}, \dots, \delta_N^{(t+1)}\}$
```

Sequential Updates and Loss Functions Agents are updated sequentially in random order to ensure that policy improvements account for the changing behavior of teammates within each iteration. An importance sampling correction is applied to the learning signal at each step to minimize biasing of the advantage estimates. Formally, let $\mathcal{U}_k$ denote the set of agents updated before agent i in the current iteration. The importance-corrected learning signal for agent i is:

$$M_i(s,a) = A^{\pi^{\text{old}}}(s,a) \prod_{k \in \mathcal{U}_k} \frac{\pi_k^{\text{new}}(a_k|o_k)}{\pi_k^{\text{old}}(a_k|o_k)} \tag{15}$$

where $A^{\pi^{\text{old}}}(s,a)$ is the original advantage estimate. The policy parameters θ_i and dual variable λ_i for each agent are updated using the following optimization scheme:

$$\begin{aligned}\mathcal{L}_i(\theta_i,\lambda_i) = &- \mathbb{E}_{s,a\sim\mathcal{D}}[M_i(s,a)] \\ &+ \lambda_i \left(\mathbb{E}_{s\sim\mathcal{D}}[W_1(\pi_i^{\text{old}}(\cdot|s), \pi_i^{\theta_i}(\cdot|s))] - \delta_i\right)\end{aligned} \tag{16}$$

We minimize $\mathcal{L}_i$ with respect to θ_i using gradient descent:

$$\theta_i \leftarrow \theta_i - \alpha_\theta \nabla_{\theta_i} \mathcal{L}_i(\theta_i, \lambda_i) \tag{17}$$

and maximize with respect to λ_i using gradient ascent (equivalent to minimizing the dual):

$$\lambda_i \leftarrow \lambda_i + \alpha_\lambda \nabla_{\lambda_i} \mathcal{L}_i(\theta_i, \lambda_i) \tag{18}$$

where α_θ and α_λ are the learning rates. This ensures λ_i acts as a Lagrange multiplier enforcing the trust-region constraint while being consistent with the dual problem $\min_{\lambda_i \geq 0} g(\lambda_i)$. The gradient ascent on λ_i automatically adjusts its value: if the policy violates the trust region constraint (W_1 distance exceeds δ_i), λ_i increases to penalize large updates more strongly; if the constraint is satisfied with margin, λ_i decreases to allow more aggressive policy improvements. (Note that while the dual formulation avoids Wasserstein distance computation in the policy gradient step through the inner maximization, the loss function still requires W_1 estimates for the constraint term and CAATR requires them for adaptation.) Critic parameters ϕ are updated by minimizing the standard mean-squared TD-error:

$$\mathcal{L}(\phi) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma V_{\phi_{\text{target}}}(s') - Q_\phi(s,a)\right)^2\right] \tag{19}$$

Theoretical Results

We now present further theoretical results for our approach. We begin by stating the regularity conditions required for our analysis. For convenience, we provide a table of notation, seen in Table 1:

Assumption 1 (Regularity of Spaces and Dynamics)**.** *The state space $\mathcal{S}$ is a compact subset of a Euclidean space, and each agent's action space $\mathcal{A}_i$ is a compact metric space equipped with a distance function that makes it complete and separable. The reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is continuous. For every continuous function $w : \mathcal{S} \to \mathbb{R}$, the mapping $(s,a) \mapsto \int_{\mathcal{S}} w(s')dP(s'|s,a)$ is continuous in both s and a.*

Assumption 2 (Continuity of Advantage and Cost)**.** *For every joint policy $\pi \in \Pi$, the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is continuous in both state and action. The transport cost function $c : \mathcal{A}_i \times \mathcal{A}_i \to \mathbb{R}_{\geq 0}$ is continuous and satisfies $c(a_i, a_i) = 0$ for all $a_i \in \mathcal{A}_i$.*

Table 1: Notation for Theoretical Analysis

Symbol	Description
λ_i	Dual variable for agent i
λ_i^*	Optimal dual variable for agent i
$\Phi_\lambda(s, a_i)$	λ-regularized advantage function
$\mathcal{D}_\lambda(s, a_i)$	Set of maximizers of regularized advantage
$c(a_i, a_i')$	Transport cost between actions a_i and a_i'
δ_i	Trust region radius (Wasserstein distance bound)
$\underline{T}_\lambda(s, a_i)$	Closest maximizer transport map
$\overline{T}_\lambda(s, a_i)$	Furthest maximizer transport map
$f_{\#}\mu$	Pushforward of measure μ through map f
t^*	Mixing parameter for optimal transport maps
$\tilde{\pi}_i$	Updated policy for agent i
W_1	Wasserstein-1 distance
Id	Identity map
γ_s	Transport plan at state s
ρ_π	State distribution under policy π

We use the standard c-transform identity from optimal transport theory: $\sup_\nu \{\langle f, \nu\rangle - \lambda W_c(\mu,\nu)\} = \int \sup_y [f(y) - \lambda c(x,y)] d\mu(x)$, which allows us to convert the supremum over probability measures to a pointwise maximization.

Remark 2. *Assumption 1 is standard in the MDP literature (Hernández-Lerma and Lasserre 2012) and guarantees existence of optimal stationary policies. Assumption 2 ensures that the regularized advantage function is well-behaved and that the Wasserstein distance is a proper metric.*

Having established the regularity conditions, we characterize the optimal policy update through the λ-regularized advantage function. For any $\lambda \geq 0$, define:

$$\Phi_\lambda(s, a_i) := \max_{a_i' \in \mathcal{A}_i} \{A^\pi(s, a_i', a_{-i}) - \lambda c(a_i, a_i')\}, \tag{20}$$

and its associated set of maximizers:

$$\mathcal{D}_\lambda(s, a_i) := \arg\max_{a_i' \in \mathcal{A}_i} \{A^\pi(s, a_i', a_{-i}) - \lambda c(a_i, a_i')\}. \tag{21}$$

The regularized advantage Φ_λ balances maximizing the advantage function against a transport cost penalty scaled by λ, and plays a central role in characterizing optimal transport maps.

Corollary 1 (Optimal Policy Characterization)**.** *Under Assumptions 1 and 2, let $\lambda_i^* \geq 0$ minimize the dual problem from Theorem 1. Then there exist measurable selection maps $\underline{T}_{\lambda_i^*}, \overline{T}_{\lambda_i^*} : \mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i$ satisfying:*

$$\underline{T}_{\lambda_i^*}(s, a_i) \in \arg\min_{a_i' \in \mathcal{D}_{\lambda_i^*}(s, a_i)} c(a_i, a_i'), \tag{22}$$

$$\overline{T}_{\lambda_i^*}(s, a_i) \in \arg\max_{a_i' \in \mathcal{D}_{\lambda_i^*}(s, a_i)} c(a_i, a_i'). \tag{23}$$

When $\lambda_i^* > 0$, there exists $t^* \in [0, 1]$ such that:

$$\begin{aligned} t^* &\mathbb{E}_{s,a_i\sim\pi}[c(a_i, \underline{T}_{\lambda_i^*}(s, a_i))] \\ &+ (1 - t^*)\mathbb{E}_{s,a_i\sim\pi}[c(a_i, \overline{T}_{\lambda_i^*}(s, a_i))] = \delta_i. \end{aligned} \tag{24}$$

An optimal policy for problem (P) is given by:

$$\begin{aligned}\tilde{\pi}_i(\cdot|s) = &t^* \underline{T}_{\lambda_i^*}(s, \cdot)_{\#}\pi_i(\cdot|s) \\ &+ (1 - t^*)\overline{T}_{\lambda_i^*}(s, \cdot)_{\#}\pi_i(\cdot|s),\end{aligned} \tag{25}$$

where $t^* = 0$ if $\lambda_i^* = 0$.

Proof. By Assumptions 1 and 2, the advantage function and cost are continuous, so the correspondence $(s, a_i) \mapsto \mathcal{D}_\lambda(s, a_i)$ is closed-valued and measurable. Since the cost function $c(a_i, \cdot)$ is continuous on the compact set $\mathcal{D}_\lambda(s, a_i)$, it attains its minimum and maximum. The existence of measurable selections T_λ and $\overline{T}_\lambda$ follows from standard measurable selection theorems.

When $\lambda_i^* > 0$, the trust region constraint is active. Consider the transport plan:

$$\begin{aligned}\gamma_s := & t^*(\mathrm{Id}, T_{\lambda_i^*}(s, \cdot))_{\#}\pi_i(\cdot|s) \\ & + (1 - t^*)(\mathrm{Id}, \overline{T}_{\lambda_i^*}(s, \cdot))_{\#}\pi_i(\cdot|s),\end{aligned} \tag{26}$$

where Id denotes the identity map. This plan has marginals $\pi_i(\cdot|s)$ and $\tilde{\pi}_i(\cdot|s)$, yielding:

$$W_1(\pi_i(\cdot|s), \tilde{\pi}_i(\cdot|s)) \leq \int_{\mathcal{A}_i \times \mathcal{A}_i} c(a_i, a_i')\, d\gamma_s(a_i, a_i') \tag{27}$$

$$= t^* \mathbb{E}[c(a_i, T_{\lambda_i^*})] + (1 - t^*)\mathbb{E}[c(a_i, \overline{T}_{\lambda_i^*})]. \tag{28}$$

The parameter t^* is chosen to satisfy the trust region constraint with equality. When $\lambda_i^* = 0$, the constraint is not binding and setting $t^* = 0$ suffices. □

We now establish what guarantees the dual solution provides.

Proposition 1 (Surrogate Objective Bound)**.** *Let π^{old} denote the current joint policy and $\lambda_i^* \geq 0$ the optimal dual variable for agent i. When agent i updates to π_i^{new} via Corollary 1 while teammates remain at π_{-i}^{old}, the surrogate objective satisfies:*

$$\mathbb{E}_{s,a\sim\pi^{old}}\left[\frac{\pi_i^{new}(a_i|s)}{\pi_i^{old}(a_i|s)} A^{\pi^{old}}(s, a)\right] \geq \lambda_i^* \delta_i. \tag{29}$$

Proof. By strong duality, the primal optimal equals the dual optimal. The left side is the primal objective at π_i^{new}, which equals $\lambda_i^* \delta_i + \mathbb{E}[\Phi_{\lambda_i^*}(s, a_i)]$. Since $\Phi_\lambda(s, a_i) \geq A^{\pi^{\text{old}}}(s, a_i, a_{-i})$ and $\mathbb{E}_a[A^{\pi^{\text{old}}}(s, a)] = 0$, the result follows. □

Remark 3. *The dual variable λ_i^* represents the shadow price of the trust region constraint: each unit of Wasserstein distance δ_i contributes at least λ_i^* to surrogate improvement. However, actual performance $J(\pi^{new})$ depends on state distribution shift, which we cannot bound tightly without MDP-specific structure.*

NUMERICAL EXAMPLES

Multi-Agent Differential Game

We evaluate the proposed algorithm on an n-agent differential game where agents must coordinate to reach a global optimum while avoiding local optima. The joint reward function for $n \in \{2, 3, 5, 7, 9\}$ agents is given by

$$\begin{aligned} r(a_1, \ldots, a_n) = & \alpha_g \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(a_i - 5)^2}{\sigma_i^2}\right) \\ & + \alpha_l \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(a_i - 1)^2\right)\end{aligned} \tag{30}$$

where the global optimum is located at $(5, \ldots, 5)$ with coefficient $\alpha_g = 10/((2\pi)^{n/2}\prod_i \sigma_i)$, and a local optimum exists at $(1, \ldots, 1)$ with coefficient $\alpha_l = 6.5/(2\pi)^{n/2}$. The variance terms are set as $\sigma_1 = \sigma_3 = 1$ and $\sigma_2 = 3$ to introduce asymmetry among agents, creating different learning dynamics that test coordination under non-uniform gradients. Actions are constrained to $[0, 7]$ for all agents.

Each agent i maintains a Gaussian policy $\pi_i(\cdot) = \mathcal{N}(\mu_i, \sigma_i^2)$ initialized at $\mu_i = 1.5$, $\sigma_i = 0.5$, placing agents near the local optimum. The challenge is for agents to coordinate their policy updates to escape the local optimum basin and converge to the global optimum. We implement the W-MATRPO algorithm with CAATR. (Note that hyperparameters are shown in Table 2.) The 1-Wasserstein distance between Gaussian policies is computed as

$$W_1(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = |\mu_1 - \mu_2| + |\sigma_1 - \sigma_2|. \tag{31}$$

Table 2: Differential Game Hyperparameters

Parameter	**3 Agents**	**5 Agents**	**7 Agents**	**9 Agents**
Iterations	4000	4000	4000	4000
Batch size	30	30	30	30
CAATR constant (C)	0.02	0.02	0.10	0.15
Trust region (δ)	0.1	0.1	0.1	0.1
Initial μ_0	1.5	1.5	1.5	1.5
Initial σ_0	0.5	0.5	0.5	0.5
Critic LR	0.2	0.2	0.2	0.2

Results

We find that W-MATRPO with CAATR exhibits fundamentally unique behavior compared to KL–based TRPO methods in multi-agent settings. Most interestingly, the algorithm is successful at escaping local optima in differential game environments, while both adaptive and non-adaptive HATRPO variants seem to remain in local optima traps, thereby not exploring other locales. However, it is noteworthy that the W-MATRPO with CAATR algorithm did not end up finally converging to the global optima. This has the following implication: the inclusion of distance measures that respect the underlying structure of the action space may be beneficial in multi-agent coordination tasks, such as that required to escape local optima and approach a global optimum, as seen in Figure 1.

The CAATR mechanism has mixed success. In the differential game, combining CAATR with standard HATRPO seemed to provide minimal benefit when included with HATRPO. However, when paired with Wasserstein constraints, CAATR appears to facilitate coordinated exploration as seen in Figure 1. HATRPO includes a backtracking line search mechanism, which we hypothesize interferes with any benefit offered by the CAATR, as step size reductions and rejected updates artificially suppress the measured policy drift that CAATR uses to coordinate trust region adaptation. (This issue is not present in the W-MATRPO algorithm as the dual formulation does not require or use backtracking.) CAATR's value seems to emerge when the underlying trust region geometry already supports efficient exploration, serving to

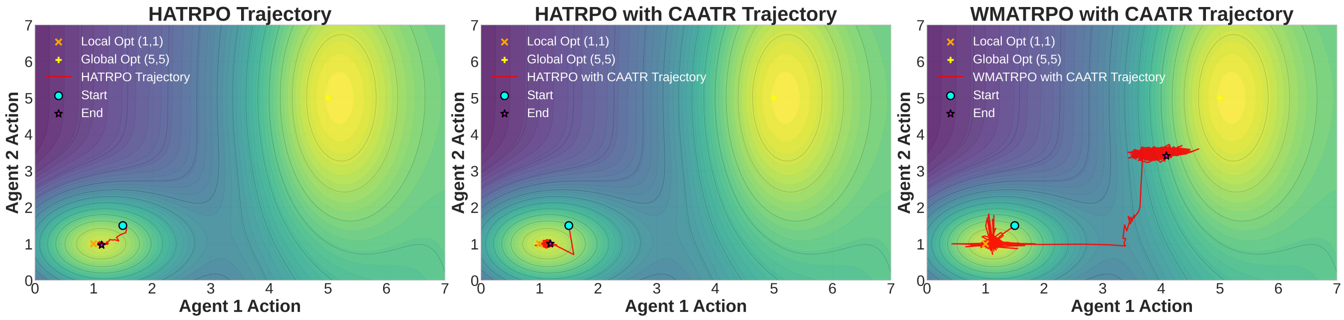

Figure 1: 2-agent differential game setup tested with three algorithms, HATRPO, HATRPO with CAATR and WMATRPO with CAATR. Only WMATRPO with CAATR successfully escapes local maximum and approaches global maximum.

Table 3: Performance Comparison of Multi-Agent Algorithms. Our proposed **W-MATRPO with CAATR** (highlighted in blue) consistently achieves the lowest distance to the global reward surface across all agent configurations, with an average improvement of 64% over baseline methods. The distance metric measures convergence quality, where lower values indicate superior performance.

Agents	Algorithm	Final Actions	Distance ↓	Improvement
3	Standard HATRPO	(0.995, 1.002, 1.013)	6.9227	baseline
	HATRPO with CAATR	(1.063, 1.040, 1.011)	6.8623	0.9%
	W-MATRPO with CAATR	**(3.829, 3.517, 3.466)**	**2.4335**	**64.8%**
5	Standard HATRPO	(1.004, 0.992, 0.999, 0.999, 1.017)	8.9393	baseline
	HATRPO with CAATR	(1.008, 0.963, 0.973, 1.030, 1.006)	8.9534	-0.2%
	W-MATRPO with CAATR	**(3.566, 3.571, 3.537, 3.509, 3.508)**	**3.2690**	**63.4%**
7	Standard HATRPO	(1.016, 1.027, 1.025, 1.035, 1.026, 1.034, 1.015)	10.5157	baseline
	HATRPO with CAATR	(0.958, 0.953, 1.024, 0.988, 0.977, 0.961, 0.998)	10.6361	-1.1%
	W-MATRPO with CAATR	**(3.536, 3.557, 3.554, 3.522, 3.481, 3.592, 3.506)**	**3.8760**	**63.1%**
9	Standard HATRPO	(1.014, 0.996, 0.952, 0.987, 0.978, 1.001, 0.964, 0.954, 0.976)	12.0590	baseline
	HATRPO with CAATR	(1.042, 0.981, 1.038, 0.990, 1.033, 1.035, 0.944, 1.029, 0.987)	11.9741	0.7%
	W-MATRPO with CAATR	**(3.540, 3.565, 3.569, 3.574, 3.577, 3.574, 3.492, 3.525, 3.587)**	**4.3330**	**64.1%**

↓ *Lower is better*

modulate the exploration rate based on team stability. The algorithm's performance across different multi-agent settings seems consistent (see Table 3), with the distance from global optimum scaling approximately linearly with the number of agents. This suggests that the coordination challenge intensifies with team size, but the fundamental escape mechanism remains effective for W-MATRPO as the number of agents increase.

Limitations and Future Directions

The most notable limitation is the apparent difficulty of reaching global maxima, as observed in the ablation study. The tradeoff of improved exploration may potentially cause paths to global maxima to be avoided in transit. Reduction in the basin gap was observed to have a positive effect on performance, indicating that reduction in search space improves the performance of this approach. Another important limitation to be noted is the computational overhead required in solving the dual optimization problem. This increases wall-clock training time for our approach as compared to similar approaches, which is inevitable due to the computational cost of the approach. Second, our theoretical guarantees assume accurate advantage estimation, which may not hold early in training when the critic is poorly calibrated.

The results also indicate that even if Wasserstein-based approaches are seemingly successful at avoiding local optima, superior tenets of KL-based approaches (lower computational overhead, overall performance) suggest that both methods seem to be useful in certain scenarios. Future work should investigate hybrid approaches that dynamically select between Wasserstein and KL constraints based on the observed optimization landscape. Additionally, the framework should be extended to competitive or mixed-agent settings to determine if geometric awareness provides similar benefits in scenarios where agents may not share a joint policy, or when objectives are not aligned. Finally, this approach needs further evaluation on larger-scale benchmarks such as SMACv2 (Ellis et al. 2023).

Conclusion

This work presents a tractable dual formulation for Wasserstein-constrained trust region policy optimization in cooperative MARL. By replacing KL-divergence constraints with Wasserstein-1 distance, we reduce the infinite-dimensional primal optimization to a one-dimensional convex problem over a single dual variable per agent, with explicit characterization of optimal policy updates and bounds on surrogate objective improvement. We introduce a coordination-aware adaptive trust region mechanism (CAATR) that modulates each agent's trust region inversely proportional to teammate policy drift. Empirical evaluation on differential games demonstrates that the resulting W-MATRPO algorithm achieves improved exploration compared to KL-based methods, successfully escaping local optima where HATRPO fails. The primary contribution is computational: the dual formulation provides a practical algorithm for Wasserstein trust regions without requiring explicit Wasserstein distance computation during optimization. Future work should investigate when geometric trust regions provide advantages over information-theoretic constraints, extend the framework to competitive settings, and evaluate on larger-scale benchmarks to determine the generality of the observed exploration benefits.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[Abdullah et al. 2019] Abdullah, M. A.; Ren, H.; Ammar, H. B.; Milenkovic, V.; Luo, R.; Zhang, M.; and Wang, J. 2019. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*.

[Amato 2024] Amato, C. 2024. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2409.03052*.

[Arjovsky, Chintala, and Bottou 2017] Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In Precup, D., and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 214–223. PMLR.

[Baheri and Kochenderfer 2024] Baheri, A., and Kochenderfer, M. J. 2024. The synergy between optimal transport theory and multi-agent reinforcement learning. *arXiv preprint arXiv:2401.10949*.

[Baheri 2023] Baheri, A. 2023. Risk-aware reinforcement learning through optimal transport theory. *arXiv preprint arXiv:2309.06239*.

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

[Busoniu, Babuska, and De Schutter 2008] Busoniu, L.; Babuska, R.; and De Schutter, B. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38(2):156–172.

[Canese et al. 2021] Canese, L.; Cardarilli, G. C.; Di Nunzio, L.; Fazzolari, R.; Giardino, D.; Re, M.; and Spanò, S. 2021. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences* 11(11):4948.

[Cuturi and Peyré 2016] Cuturi, M., and Peyré, G. 2016. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences* 9(1):320–343.

[Cuturi and Peyré 2018] Cuturi, M., and Peyré, G. 2018. Semidual regularized optimal transport. *SIAM Review* 60(4):941–965.

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

[Ellis et al. 2023] Ellis, B.; Cook, J.; Moalla, S.; Samvelyan, M.; Sun, M.; Mahajan, A.; Foerster, J. N.; and Whiteson, S. 2023. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[Fedotov, Harremoës, and Topsoe 2003] Fedotov, A. A.; Harremoës, P.; and Topsoe, F. 2003. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory* 49(6):1491–1498.

[Gu et al. 2022] Gu, S.; Kuba, J. G.; Wen, M.; Chen, R.; Wang, Z.; Tian, Z.; Wang, J.; Knoll, A.; and Yang, Y. 2022. Multi-agent constrained policy optimisation.

[Hao et al. 2022] Hao, X.; Mao, H.; Wang, W.; Yang, Y.; Li, D.; Zheng, Y.; Wang, Z.; and Hao, J. 2022. Breaking the curse of dimensionality in multiagent state space: A unified agent permutation framework. *arXiv preprint arXiv:2203.05285*.

[He et al. 2022] He, S.; Jiang, Y.; Zhang, H.; Shao, J.; and Ji, X. 2022. Wasserstein unsupervised reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6884–6892.

[Hernandez-Leal et al. 2017] Hernandez-Leal, P.; Kaisers, M.; Baarslag, T.; and De Cote, E. M. 2017. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*.

[Hernández-Lerma and Lasserre 2012] Hernández-Lerma, O., and Lasserre, J. B. 2012. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media.

[Hou et al. 2020] Hou, L.; Pang, L.; Hong, X.; Lan, Y.; Ma, Z.; and Yin, D. 2020. Robust reinforcement learning with wasserstein constraint. *arXiv preprint arXiv:2006.00945*.

[Kaelbling, Littman, and Cassandra 1998] Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101(1-2):99–134.

[Kantorovich and Rubinshtein 1958] Kantorovich, L. V., and Rubinshtein, S. 1958. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics* 13(7):52–59.

[Khamis et al. 2024] Khamis, A.; Tsuchida, R.; Tarek, M.; Rolland, V.; and Petersson, L. 2024. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE transactions on pattern analysis and machine intelligence*.

[Klink et al. 2022] Klink, P.; Yang, H.; D'Eramo, C.; Peters, J.; and Pajarinen, J. 2022. Curriculum reinforcement learning via constrained optimal transport. In *International Conference on Machine Learning*, 11341–11358. PMLR.

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

[Kuba et al. 2021] Kuba, J. G.; Chen, R.; Wen, M.; Wen, Y.; Sun, F.; Wang, J.; and Yang, Y. 2021. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*.

[Li and He 2023] Li, H., and He, H. 2023. Multiagent trust region policy optimization. *IEEE Transactions on Neural Networks and Learning Systems* 35(9):12873–12887.

[Li et al. 2021] Li, W.; Wang, X.; Jin, B.; Sheng, J.; and Zha, H. 2021. Dealing with non-stationarity in marl via trust-region decomposition. *arXiv preprint arXiv:2102.10616*.

[Li 2018] Li, J. Z. 2018. *Principled approaches to robust machine learning and beyond*. Ph.D. Dissertation, Massachusetts Institute of Technology.

[Lv, Yang, and Li 2024] Lv, J.; Yang, H.; and Li, P. 2024. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. *Advances in Neural Information Processing Systems* 37:65445–65475.

[Makar, Mahadevan, and Ghavamzadeh 2001] Makar, R.; Mahadevan, S.; and Ghavamzadeh, M. 2001. Hierarchical multi-agent reinforcement learning. In *Proceedings of the fifth international conference on Autonomous agents*, 246–253.

[Matignon, Laurent, and Le Fort-Piat 2012] Matignon, L.; Laurent, G. J.; and Le Fort-Piat, N. 2012. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review* 27(1):1–31.

[Nasiri and Rezghi 2023] Nasiri, M. M., and Rezghi, M. 2023. Heterogeneous multi-agent reinforcement learning via mirror descent policy optimization. *arXiv preprint arXiv:2308.06741*.

[Nguyen, Kumar, and Lau 2018] Nguyen, D. T.; Kumar, A.; and Lau, H. C. 2018. Credit assignment for collective multiagent rl with global rewards. *Advances in neural information processing systems* 31.

[Oroojlooy and Hajinezhad 2023] Oroojlooy, A., and Hajinezhad, D. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence* 53(11):13677–13722.

[Orr and Dutta 2023] Orr, J., and Dutta, A. 2023. Multi-agent deep reinforcement learning for multi-robot applications: A survey. *Sensors* 23(7):3625.

[Pacchiano et al. 2020] Pacchiano, A.; Parker-Holder, J.; Tang, Y.; Choromanski, K.; Choromanska, A.; and Jordan, M. 2020. Learning to score behaviors for guided policy optimization. In *International Conference on Machine Learning*, 7445–7454. PMLR.

[Papoudakis et al. 2019] Papoudakis, G.; Christianos, F.; Rahman, A.; and Albrecht, S. V. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*.

[Peyré, Cuturi, and others 2019] Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11(5-6):355–607.

[Pinsker 1964] Pinsker, M. S. 1964. Information and information stability of random variables and processes. *Holden-Day*.

[Schulman et al. 2015a] Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015a. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

[Schulman et al. 2015b] Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015b. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

[Shek, Shi, and Tokekar 2025] Shek, C. L.; Shi, G.; and Tokekar, P. 2025. Multi-agent trust region policy optimisation: A joint constraint approach. *arXiv preprint arXiv:2508.10340*.

[Song, Zhao, and He 2022] Song, J.; Zhao, C.; and He, N. 2022. Efficient wasserstein and sinkhorn policy optimization.

[Sun et al. 2022] Sun, M.; Devlin, S.; Beck, J.; Hofmann, K.; and Whiteson, S. 2022. Trust region bounds for decentralized ppo under non-stationarity. *arXiv preprint arXiv:2202.00082*.

[Terpin et al. 2022] Terpin, A.; Lanzetti, N.; Yardim, B.; Dorfler, F.; and Ramponi, G. 2022. Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions. *Advances in Neural Information Processing Systems* 35:19786–19797.

[Villani and others 2008] Villani, C., et al. 2008. *Optimal transport: old and new*, volume 338. Springer.

[Wang et al. 2020] Wang, T.; Dong, H.; Lesser, V.; and Zhang, C. 2020. Roma: Multi-agent reinforcement learning with emergent roles. *arXiv preprint arXiv:2003.08039*.

[Wen et al. 2022] Wen, Y.; Chen, H.; Yang, Y.; Li, M.; Tian, Z.; Chen, X.; and Wang, J. 2022. A game-theoretic approach to multi-agent trust region optimization. In *International conference on distributed artificial intelligence*, 74–87. Springer.

[Xiao et al. 2019] Xiao, H.; Herman, M.; Wagner, J.; Ziesche, S.; Etesami, J.; and Linh, T. H. 2019. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*.

[Zawar, Sethi, and Roy] Zawar, R.; Sethi, P. S.; and Roy, R. Jensen-shannon divergence in safe multi-agent rl. In *The Second Tiny Papers Track at ICLR 2024*.

[Zawar, Sethi, and Roy 2024] Zawar, R.; Sethi, P. S.; and Roy, R. 2024. JENSEN-SHANNON DIVERGENCE IN SAFE MULTI- AGENT RL. In *The Second Tiny Papers Track at ICLR 2024*.

[Zhang et al. 2024] Zhang, R.; Hou, J.; Walter, F.; Gu, S.; Guan, J.; Röhrbein, F.; Du, Y.; Cai, P.; Chen, G.; and Knoll, A. 2024. Multi-agent reinforcement learning for autonomous driving: A survey. *arXiv preprint arXiv:2408.09675*.

[Zhong et al. 2024a] Zhong, Y.; Kuba, J. G.; Feng, X.; Hu, S.; Ji, J.; and Yang, Y. 2024a. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research* 25(32):1–67.

[Zhong et al. 2024b] Zhong, Y.; Kuba, J. G.; Feng, X.; Hu, S.; Ji, J.; and Yang, Y. 2024b. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research* 25(32):1–67.

[Zhou et al. 2020] Zhou, M.; Liu, Z.; Sui, P.; Li, Y.; and Chung, Y. Y. 2020. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. *Advances in neural information processing systems* 33:11853–11864.