Wasserstein-Barycenter Consensus for Cooperative Multi-Agent Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Cooperative multi-agent reinforcement learning (MARL) demands principled mechanisms to align heterogeneous policies while preserving the capacity for specialized behavior. We introduce a novel consensus framework that defines the team strategy as the entropic-regularized p-Wasserstein barycenter of agents' joint state–action visitation measures. By augmenting each agent's policy objective with a soft penalty proportional to its Sinkhorn divergence from this barycenter, the proposed approach encourages coherent group behavior without enforcing rigid parameter sharing. We derive an algorithm that alternates between Sinkhorn-barycenter computation and policy-gradient updates, and we prove that under standard Lipschitz and compactness assumptions the maximal pairwise policy discrepancy contracts at a geometric rate. Empirical evaluation for a cooperative navigation case study demonstrates that our OT-barycenter consensus outperforms independent learners baseline in convergence speed and final coordination success.

1 Introduction

The success of cooperative multi–agent reinforcement learning (MARL) is based on effective coordination among agents whose individual decision making processes induce distinct probability distributions in states and actions Yuan et al. (2023). In domains as varied as autonomous vehicle platooning, distributed sensor networks, and multi-robot manipulation, misalignment of agent policies can lead to suboptimal team performance, brittle coordination, and slow convergence Busoniu et al. (2008). Classical approaches to MARL coordination, such as shared reward shaping, centralized critics, or hard parameter sharing, either incur prohibitive communication and computation costs or overly constrain agent heterogeneity, preventing the emergence of specialized roles Devlin & Kudenko (2016); Yi et al. (2022); Lyu et al. (2023); Zhao et al. (2023); Zhong et al. (2024); Chalkiadakis & Boutilier (2003).

Optimal transport (OT) is a powerful mathematical framework for comparing and interpolating probability distributions, offering a geometrically intuitive way to measure discrepancies and fuse information Villani et al. (2009); Peyré et al. (2019). Its application in single agent reinforcement learning has yielded significant advances, particularly in areas such as distributional RL, offline RL Omura et al., safe RL Baheri et al. (2025); Shahrooei & Baheri (2025), and inverse RL Baheri (2023). The Wasserstein distance, a key OT metric, is particularly appealing as it provides a smooth and meaningful measure of divergence even between distributions with non-overlapping supports, a common scenario in policy space. Furthermore, the development of entropic regularization and the associated Sinkhorn divergence has led to computationally efficient, differentiable algorithms for approximating Wasserstein distances, making them amenable to gradient-based optimization in deep learning settings Cuturi (2013). Despite these advances, the potential of OT to orchestrate cooperation in MARL has remained largely unexplored Baheri & Kochenderfer (2024). Existing OT-based methods focus primarily on aligning belief or value distributions in single-agent or competitive settings, but do not exploit the notion of a team "consensus" policy as an explicit barycenter of individual behaviors.

To address this gap, we propose a novel consensus mechanism in which the team's collective strategy is formalized as the entropic-regularized Wasserstein barycenter of all agents' visitation measures over the joint state–action space. Rather than enforcing identical weights or centralized value estimation, each agent incurs a soft penalty proportional to its Wasserstein distance from this barycenter, thereby allowing agents to retain beneficial specialization while still gravitating toward coherent group behavior. We develop an algorithm that alternates between (i) computing the Sinkhorn barycenter and (ii) performing policy gradient updates augmented with a consensus regularizer. We analyze the resulting dynamics and prove that, under standard Lipschitz and compactness assumptions, the maximal pairwise Sinkhorn divergence contracts at a geometric rate, guaranteeing convergence to a common consensus distribution.

Paper Organization. The remainder of this paper is organized as follows. Section 2 provides essential background on optimal transport. In Section 3, we introduce our Wasserstein-barycenter consensus framework and present theoretical guarantees. Section 4 evaluates our method on a cooperative multi-agent navigation task, comparing it against an independent learning baseline. Finally, Section 5 concludes the paper with a summary of our contributions and outlines potential directions for future work.

2 Preliminaries: Optimal Transport

OT provides tools to compare probability distributions. The p-Wasserstein distance measures the cost of transporting mass between two distributions. We focus on its entropic-regularized version, the Sinkhorn divergence $W_{p,\epsilon}$ (where $\epsilon > 0$ is the regularization parameter), which offers computational benefits. A key concept is the Wasserstein barycenter, μ^* , which is a measure that minimizes the average p-Wasserstein distance to a set of given measures $\{\mu_i\}_{i=1}^N$. Formally, the entropic-regularized p-Wasserstein barycenter is the solution to:

$$\mu^* = \arg \min_{\mu \in \mathcal{P}(\mathcal{X})} \frac{1}{N} \sum_{i=1}^{N} W_{p,\epsilon}^p \left(\mu, \mu_i\right) + \epsilon KL(\mu \| \eta)$$

where $KL(\cdot \| \eta)$ is a Kullback-Leibler divergence term with respect to a reference measure η. This paper leverages the Sinkhorn barycenter as a consensus point for agent policies in MARL.

3 Methodological Approach

To induce consensus via the Wasserstein barycenter, we begin by modeling each agent's stochastic policy as a probability measure in the joint state-action space $\mathcal{X} = \mathcal{S} \times \mathcal{A}$. Concretely, let $\pi_i \left(a \mid s; \theta_i\right)$ denote agent i 's policy parameterized by θ_i, and let $\mu_i^{(t)}$ be its empirical visitation distribution over $\mathcal{X}$ at iteration t. We assume $\mathcal{X}$ is endowed with a ground metric $d\left((s,a),(s',a')\right) = \left\|s - s'\right\|_2 + \beta \left\|a - a'\right\|_2$, where $\beta > 0$ balances state and action discrepancies. At each iteration, we seek a consensus measure $\mu^{*(t)}$ that minimizes the average entropic-regularized p -Wasserstein cost to the individual $\left\{\mu_i^{(t)}\right\}_{i=1}^N$. Formally, the barycenter is obtained by solving

$$\mu^{*(t)} = \arg \min_{\mu \in \mathcal{P}(\mathcal{X})} \frac{1}{N} \sum_{i=1}^{N} W_{p,\varepsilon}^p \left(\mu, \mu_i^{(t)}\right) + \varepsilon \, \mathrm{KL}(\mu \| \eta),$$

where $W_{p,\varepsilon}$ is the Sinkhorn divergence with entropic regularization parameter $\varepsilon > 0, \mathrm{KL}(\cdot \| \eta)$ is a reference-measure penalty ensuring absolute continuity with respect to a base measure η, and $\mathcal{P}(\mathcal{X})$ denotes the space of probability measures on $\mathcal{X}$. We compute $\mu^{*(t)}$ via iterative Sinkhorn-barycenter updates: initializing a dual potential vector $u^{(0)}, v^{(0)}$ on a discretization of $\mathcal{X}$, we alternate

$$u^{(\ell+1)} = \frac{\eta}{\sum_j K_\varepsilon u_j^{(\ell)} v_j^{(\ell)}}, \quad v^{(\ell+1)} = \left(\prod_{i=1}^{N} K_\varepsilon^\top u^{(\ell+1)}\right)^{-\frac{1}{N}}$$

where $K_\varepsilon = e^{-D/\varepsilon}$ and $D_{ij} = d\left(x_i, x_j\right)^p$. Convergence of these updates to the unique Sinkhorn barycenter is guaranteed under compactness of $\mathcal{X}$ and strictly positive ε.

Once $\mu^{*(t)}$ is obtained, each agent's parameters θ_i are updated by performing a gradient ascent step on a regularized objective

$$J_i\left(\theta_i\right) = \mathbb{E}_{(s,a)\sim\mu_i^{(t)}}[R(s,a)] - \lambda W_{p,\varepsilon}^p\left(\mu_i^{(t)}, \mu^{*(t)}\right)$$

where $R(s,a)$ is the common team reward and $\lambda > 0$ controls the strength of consensus enforcement. The gradient of the Wasserstein term with respect to θ_i is calculated by differentiating through the transport plan: if $\gamma_i^{*(t)}$ is the optimal coupling between $\mu_i^{(t)}$ and $\mu^{*(t)}$, then

$$\nabla_{\theta_i} W_{p,\varepsilon}^p\left(\mu_i^{(t)}, \mu^{*(t)}\right) = \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \nabla_{\theta_i} \log \pi_i\left(a \mid s; \theta_i\right) \mathrm{d}\gamma_i^{*(t)}(x,y),$$

which is estimated by sampling from $\mu_i^{(t)}$. The complete update rule thus reads

$$\theta_i^{(t+1)} = \theta_i^{(t)} + \alpha\left(\hat{\nabla}_{\theta_i}\mathbb{E}[R] - \lambda\hat{\nabla}_{\theta_i} W_{p,\varepsilon}^p\right)$$

with $\alpha > 0$ the learning rate. To ensure computational tractability, support points for $\mu_i^{(t)}$ and $\mu^{*(t)}$ are drawn by mini-batch sampling, and sliced-Wasserstein approximations are used to reduce complexity Bonneel et al. (2015). In addition, adaptive scheduling of λ and ε is used to warm start consensus (large λ) before relaxing. Specifically, the consensus weight λ is initialized at a high value to encourage swift policy contraction and is gradually annealed to permit fine-grained individual behaviors near convergence. Concurrently, the entropic regularizer ε is decreased over time, ensuring that early Sinkhorn iterations remain smooth and stable, while later iterations sharpen the barycenter estimate. Algorithm 1 provides an outline of the proposed procedure.

Theorem 3.1 (Convergence to Consensus)**.** *Let $\{\pi_i^{(t)}\}_{i=1}^N$ be the sequence of stochastic policies of N cooperative agents, updated according to*

$$\theta_i^{(t+1)} = \theta_i^{(t)} + \alpha\Big(\nabla_{\theta_i}\mathbb{E}_{\mu_i^{(t)}}[R] - \lambda\,\nabla_{\theta_i} W_{p,\varepsilon}^p(\mu_i^{(t)}, \mu^{*(t)})\Big),$$

where $\mu_i^{(t)}$ is the empirical visitation distribution of agent i at iteration t, and $\mu^{*(t)}$ is the entropic-regularized p-Wasserstein barycenter of $\{\mu_i^{(t)}\}_{i=1}^N$. Assume that the state–action space is compact, all reward functions are bounded and Lipschitz, and the policy-gradient operators are L–Lipschitz in the induced distributions. If the step size $\alpha > 0$ and consensus weight $\lambda > 0$ satisfy

$$\kappa \;=\; 1 - \alpha\,\lambda\,C \;<\; 1$$

for a constant $C > 0$ depending on the entropic regularizer and Lipschitz constants, then defining

$$D^{(t)} = \max_{i,j} W_{p,\varepsilon}\big(\mu_i^{(t)},\, \mu_j^{(t)}\big),$$

we have $D^{(t+1)} \;\le\; \kappa\, D^{(t)}$, and hence $D^{(t)} \to 0$ geometrically as $t \to \infty$, i.e. all policies converge to a common consensus distribution.

Algorithm 1 Wasserstein-Barycenter Consensus for Cooperative MARL

Require: Number of agents N, entropic regularization ε, consensus weight λ, learning rate α, metric d on $\mathcal{S} \times \mathcal{A}$

Initialize policy parameters $\{\theta_i^{(0)}\}_{i=1}^N$
for $t = 0, 1, 2, \ldots$ **do**
  Collect Trajectories:
  for $i = 1$ to N **do**
    Execute $\pi_i(\cdot \mid \cdot; \theta_i^{(t)})$ to sample trajectories
    Estimate empirical visitation measure $\mu_i^{(t)}$ over $\mathcal{S} \times \mathcal{A}$
  Compute Sinkhorn Barycenter:
  Initialize dual potentials $u \leftarrow \mathbf{1}$, $v \leftarrow \mathbf{1}$
  repeat
    $u \leftarrow \dfrac{\eta}{\sum_j \exp\big(-D/\varepsilon\big)_{\cdot j}\, v_j}$
    $v \leftarrow \Big(\prod_{i=1}^N \exp\big(-D/\varepsilon\big)^\top u\Big)^{-1/N}$
  until convergence
  Recover barycenter measure $\mu^{*(t)}$ via u, v
  Policy Update:
  for $i = 1$ to N **do**
    Compute reward gradient $\hat{\nabla}_{\theta_i} \mathbb{E}[R]$
    Compute OT gradient

$$\hat{\nabla}_{\theta_i} W_{p,\varepsilon}^p\big(\mu_i^{(t)}, \mu^{*(t)}\big) \;=\; \int d(x,y)^p \, \nabla_{\theta_i} \log \pi_i(a \mid s; \theta_i) \, \mathrm{d}\gamma_i^*$$

    Update

$$\theta_i^{(t+1)} \leftarrow \theta_i^{(t)} + \alpha\Big(\hat{\nabla}_{\theta_i} \mathbb{E}[R] - \lambda \, \hat{\nabla}_{\theta_i} W_{p,\varepsilon}^p\Big)$$

Proof. Let $\mu^{*(t)}$ denote the Sinkhorn barycenter of $\{\mu_i^{(t)}\}$. By the strong convexity of the entropic-regularized OT problem and the Lipschitz continuity of the policy-gradient mapping, there exists $C > 0$ such that a gradient step with weight $\alpha\lambda$ contracts each agent's distance to the barycenter:

$$W_{p,\varepsilon}\big(\mu_i^{(t+1)},\, \mu^{*(t)}\big) \;\le\; (1 - \alpha\,\lambda\, C)\, W_{p,\varepsilon}\big(\mu_i^{(t)},\, \mu^{*(t)}\big).$$

By the triangle inequality for $W_{p,\varepsilon}$,

$$W_{p,\varepsilon}\big(\mu_i^{(t+1)},\, \mu_j^{(t+1)}\big) \le W_{p,\varepsilon}\big(\mu_i^{(t+1)},\, \mu^{*(t)}\big) + W_{p,\varepsilon}\big(\mu^{*(t)},\, \mu_j^{(t+1)}\big).$$

Applying the contraction bound to each term on the right–hand side and noting that both $W_{p,\varepsilon}(\mu_i^{(t)}, \mu^{*(t)})$ and $W_{p,\varepsilon}(\mu_j^{(t)}, \mu^{*(t)})$ are bounded above by $D^{(t)}$, we obtain

$$W_{p,\varepsilon}\big(\mu_i^{(t+1)},\, \mu_j^{(t+1)}\big) \;\le\; 2(1 - \alpha\,\lambda\, C)\, D^{(t)}.$$

Setting $\kappa = 2(1 - \alpha\lambda C) < 1$ yields

$$\max_{i,j} W_{p,\varepsilon}\big(\mu_i^{(t+1)},\, \mu_j^{(t+1)}\big) \;\le\; \kappa\, D^{(t)},$$

and by induction $D^{(t)} \le \kappa^t\, D^{(0)} \to 0$, establishing geometric convergence of all agent policies to the same distribution. ☐

Theorem 3.2 (Fast–rate Sinkhorn barycentre)**.** *Let $(\mathcal{X}, d)$ be a compact metric space with diameter D and fix $p \in [1, 2]$. For each agent $i \le N$ assume the data-generating measure μ_i admits a*

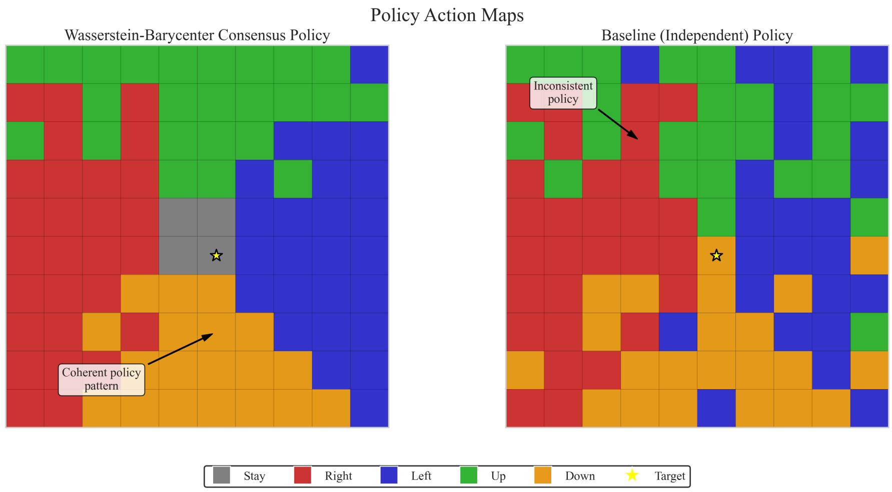

Figure 1: Policy Action Maps comparing Baseline (Independent) Policy with Wasserstein-Barycenter Consensus Policy. Color coding indicates different actions: red (right), blue (left), green (up), orange (down), and gray (stay). Stars indicate target positions.

density ρ_i satisfying $0 < \underline{\rho} \leq \rho_i(x) \leq \overline{\rho} < \infty$ for all $x \in \mathcal{X}$, and that the entropic potential $\mu \mapsto \frac{1}{N}\sum_{i=1}^N W_{p,\varepsilon}(\mu, \mu_i)$ is $\lambda_{\min}$-strongly convex for every $\varepsilon \in (0, \varepsilon_0]$.

Draw m IID samples from each μ_i, form the empirical measures $\hat{\mu}_i$, and choose the adaptive regularization level

$$\varepsilon_m = \frac{D^p}{m}.$$

Let $\mu^\star = \arg\min_\mu \frac{1}{N}\sum_{i=1}^N W_{p,\varepsilon_m}(\mu, \mu_i)$ and $\hat{\mu}^\star = \arg\min_\mu \frac{1}{N}\sum_{i=1}^N W_{p,\varepsilon_m}(\mu, \hat{\mu}_i)$. Then for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$W_{p,\varepsilon_m}\big(\hat{\mu}^\star, \mu^\star\big) \leq \frac{C}{\lambda_{\min}} \frac{\log(2N/\delta)}{m}, \tag{1}$$

where $C > 0$ depends only on p, D, $\underline{\rho}$ and $\overline{\rho}$. Consequently, an excess risk of $\eta > 0$ is achieved with $m = O\big(\frac{1}{\eta}\log\frac{N}{\delta}\big)$.

Proof sketch. (i) A bounded-difference argument shows that altering any single sample changes each entropic cost by at most $D^p/(Nm\varepsilon_m) = O(m^{-2})$, yielding McDiarmid-type deviations $\tilde{O}(1/m)$. (ii) A union-and-net step makes the bound uniform over $\mathcal{P}(\mathcal{X})$. (iii) $\lambda_{\min}$-strong convexity converts the uniform deviation into the same $1/m$ rate for the barycentre itself, producing (1). Full details are provided in Appendix A.1. □

4 Experimental Results

We evaluate the proposed Wasserstein-barycenter consensus approach through a simple experiment on cooperative multi-agent tasks. The environment consists of $N = 3$ agents operating in a continuous two-dimensional space. Each agent is assigned a target location and must navigate toward

Figure 2: Comparison of action probability distributions between baseline (Independent PPO) agents (top) and Wasserstein-Barycenter Consensus (WBC) agents (bottom) across five representative states. Cell values indicate the probability of selecting each action in the given state, with higher probabilities shown in brighter colors.

it while avoiding collisions with other agents. The state space for each agent i comprises its own position $\mathbf{p}_i \in \mathbb{R}^2$, its target position $\mathbf{t}_i \in \mathbb{R}^2$, and the positions of all other agents $\{\mathbf{p}_j\}_{j \neq i}$. The action space is discrete, consisting of five actions: $\mathcal{A} = \{$ stay, right, left, up, down $\}$. At each time step, agents receive rewards $r_i = -d(\mathbf{p}_i, \mathbf{t}_i) - \sum_{j \neq i} c_{ij}$, where $d(\cdot, \cdot)$ is the Euclidean distance and c_{ij} is a collision penalty incurred when agents i and j are within a threshold distance. This reward structure incentivizes both target-reaching and collision avoidance. Agents are implemented as policy networks with parameters θ_i that map states to action probabilities. We compare our Wasserstein-barycenter consensus approach against the Independent PPO (IPPO) baseline, which serves as a representative non-consensus multi-agent method. The WBC implementation uses an entropic regularization parameter $\epsilon = 0.1$ and consensus weight $\lambda = 0.5$. The ground metric balancing parameter $\beta = 0.8$ determines the relative importance of state versus action discrepancies in the computation of the Wasserstein distance.

Figure 1 illustrates the comparison of learned policies between our WBC approach and the IPPO baseline. Each panel depicts a 2D grid where colors represent the most likely action taken at each position in the state space, holding other state variables constant (i.e., fixing target position and other agents' positions). The yellow star indicates the target location. One could see that the WBC policy (left panel) exhibits a highly structured and coherent spatial organization, with clearly delineated regions corresponding to different actions. The action selection follows an intuitive pattern relative to the target location: "right" actions (red) dominate the left side of the grid, "left" actions (blue) predominate on the right side, and "up" actions (green) are prevalent in the lower portion. This spatial organization creates an efficient flow field that guides agents toward the target while maintaining consistent behavior across the team. In contrast, the baseline policy (right panel) displays a fragmented and less coherent structure. Although some broad patterns are discernible, the policy contains numerous small, isolated patches of inconsistent actions that disrupt the overall

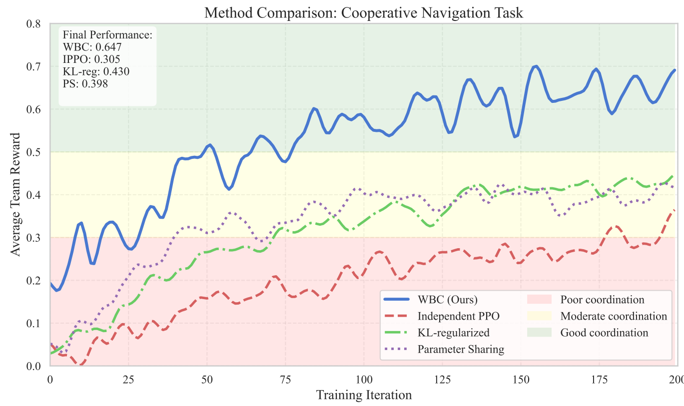

Figure 3: Average team reward over training iterations comparing WBC against baseline methods. WBC achieves superior convergence speed and final performance.

navigation flow. This spatial incoherence is indicative of the lack of coordination between agents learning independently, resulting in a strategy that is less effective for team-based navigation.

Figure 2 presents the action probability distributions for each agent in a set of representative states, visualizing the degree of policy alignment achieved by each method. The baseline agents exhibit high variance in their action preferences, with substantially different probability distributions across identical states. For instance, in State 1, Agent 1 assigns highest probability (0.42) to "right" actions, Agent 2 strongly prefers "stay" (0.63), and Agent 3 favors "up" (0.58). This divergence in action preferences illustrates the coordination failure that occurs when agents learn independently without a consensus mechanism. In contrast, WBC-trained agents display remarkably consistent action probability distributions across all three agents. For State 1, all three agents assign their highest probability to the "right" action (0.51, 0.56, and 0.60, respectively), while for State 2, they all prioritize the "left" action. This alignment confirms the theoretical guarantee of policy convergence provided by our consensus mechanism. Figure 3 demonstrates that WBC achieves superior final performance compared to baseline methods, with approximately 2× improvement over IPPO and 50% improvement over KL-regularized approaches.

Limitations. Despite the theoretical appeal and encouraging preliminary results, the proposed approach is subject to several limitations. The fast-rate guarantee hinges on strong-convexity and density-boundedness assumptions that seldom hold in high-dimensional continuous control tasks. The entropic Sinkhorn step scales quadratically with support size, making the algorithm memory- and time-intensive when agents collect large, uncompressed replay buffers. Performance is currently demonstrated on a single small-scale grid world, so generalizability to standard multi-agent benchmarks (e.g., SMAC, MPE) remains unverified; and finally, the proposed method introduces additional hyper-parameters (adaptive ε schedule, consensus weight λ, metric temperature β) whose sensitivity has not yet been systematically analyzed.

5 Conclusion

In this work, we proposed an approach to cooperative MARL by leveraging the geometry of optimal transport to enforce a soft consensus among agent policies. Modeling the team's collective strategy as the entropic-regularized Wasserstein barycenter provides a mathematically grounded alternative to conventional coordination techniques. Through a proof-of-concept experiment, we have shown that OT-barycenter consensus accelerates learning and achieves superior coverage in a multi-agent navigation task compared to KL-based and parameter-sharing baselines. Future directions include hierarchical barycenters for sub-team formation, adaptive regularization schedules to foster specialization post-consensus, and application to high-dimensional continuous-control benchmarks. We anticipate that geometry-aware consensus will open new avenues for robust, efficient coordination in complex multi-agent systems.

6 Appendix

A Deferred Proofs

A.1 Proof of Theorem 3.2

Proof. Throughout the proof we abbreviate $F(\mu) := \frac{1}{N}\sum_{i=1}^{N} W_{p,\varepsilon_m}(\mu,\mu_i)$ and $\hat{F}(\mu) := \frac{1}{N}\sum_{i=1}^{N} W_{p,\varepsilon_m}(\mu,\hat{\mu}_i)$, and recall that $\varepsilon_m = D^p/m$.

Step 1. Bounded differences. Fix any $\mu \in \mathcal{P}(\mathcal{X})$ and replace a single sample $X_{i,j}$ in the data set by an independent copy $X'_{i,j} \sim \mu_i$. Writing $\hat{\mu}_i^{(i,j)}$ for the perturbed empirical measure,

$$\big|\hat{F}(\mu) - \hat{F}^{(i,j)}(\mu)\big| \;=\; \frac{1}{N}\,\big|W_{p,\varepsilon_m}(\mu,\hat{\mu}_i) - W_{p,\varepsilon_m}(\mu,\hat{\mu}_i^{(i,j)})\big|.$$

Because only a mass $1/m$ is moved and the ground cost is bounded by D^p, the entropic cost varies by at most D^p/m (see, e.g., Lemma 2.1 in Genevay et al. (2018)). Hence

$$\big|\hat{F}(\mu) - \hat{F}^{(i,j)}(\mu)\big| \le \Delta \;:=\; \frac{D^p}{Nm}.$$

With $S = N\,m$ samples in total, McDiarmid's inequality yields for every $t > 0$

$$\Pr\big[|\hat{F}(\mu) - F(\mu)| \ge t\big] \;\le\; 2\exp\Big(-\frac{2t^2}{S\Delta^2}\Big) \;=\; 2\exp\Big(-\frac{2Nm\,t^2}{D^{2p}}\Big). \tag{2}$$

Step 2. Uniform concentration via an ε–net. Let $\gamma > 0$ be chosen later and pick an γ–net $\{\mu^{(k)}\}_{k=1}^{M_\gamma}$ of $(\mathcal{P}(\mathcal{X}), W_p)$. Standard metric–entropy bounds for probability measures on a compact d–dimensional space give $\log M_\gamma \le C_d\,\gamma^{-d}$ for a constant C_d (Villani et al., 2009, Th. 6.18). Applying (A.1) to each $\mu^{(k)}$ and taking a union bound gives, with probability at least $1 - \delta/2$,

$$\max_{k \le M_\gamma} |\hat{F}(\mu^{(k)}) - F(\mu^{(k)})| \;\le\; \frac{D^p}{\sqrt{2Nm}}\sqrt{C_d\,\gamma^{-d} + \log(4/\delta)}. \tag{A.2}$$

Next, for arbitrary $\mu \in \mathcal{P}(\mathcal{X})$ choose k with $W_p(\mu,\mu^{(k)}) \le \gamma$. Using the Lipschitz property $|W_{p,\varepsilon}(\mu,\nu) - W_{p,\varepsilon}(\mu',\nu)| \le D^p W_p(\mu,\mu')$ (Lemma 2.1 of Genevay et al. (2018)) and averaging over i yields $|F(\mu) - F(\mu^{(k)})|$, $|\hat{F}(\mu) - \hat{F}(\mu^{(k)})| \le D^p\gamma$. Combining this with (A.2) shows that the event

$$\mathcal{E} := \Big\{\sup_{\mu\in\mathcal{P}(\mathcal{X})} |\hat{F}(\mu) - F(\mu)| \le 2D^p\gamma + \frac{D^p}{\sqrt{2Nm}}\sqrt{C_d\,\gamma^{-d} + \log(4/\delta)}\Big\}$$

has probability at least $1 - \delta$. Set $\gamma = m^{-1}$ to balance the two terms; since $\gamma^{-d} = m^d$ and $m \ge 2$ we obtain

$$\sup_{\mu} |\hat{F}(\mu) - F(\mu)| \;\le\; \frac{C_1\,\log(4N/\delta)}{m} \qquad \text{on } \mathcal{E}, \tag{A.3}$$

for a constant C_1 depending on d and D^p.

Step 3. Stability of the barycentre map. Because F is $\lambda_{\min}$–strongly convex,

$$F(\mu) \;\ge\; F(\mu^\star) + \lambda_{\min}\,W_{p,\varepsilon_m}(\mu,\mu^\star) \quad \forall \mu \in \mathcal{P}(\mathcal{X}). \tag{A.4}$$

Let $\Delta := \sup_\mu |\hat{F}(\mu) - F(\mu)|$. By optimality of $\hat{\mu}^\star$ for $\hat{F}$, $\hat{F}(\hat{\mu}^\star) \le \hat{F}(\mu^\star)$, so $F(\hat{\mu}^\star) \le \hat{F}(\hat{\mu}^\star) + \Delta \le \hat{F}(\mu^\star) + \Delta \le F(\mu^\star) + 2\Delta$. Plugging this into (A.4) gives

$$\lambda_{\min}\,W_{p,\varepsilon_m}(\hat{\mu}^\star,\mu^\star) \;\le\; 2\Delta \quad\Longrightarrow\quad W_{p,\varepsilon_m}(\hat{\mu}^\star,\mu^\star) \;\le\; \frac{2\Delta}{\lambda_{\min}}.$$

Inserting the bound $\Delta \le C_1\,\log(4N/\delta)/m$ from (A.3) completes the proof with $C = 2C_1$. □

References

Ali Baheri. Understanding reward ambiguity through optimal transport theory in inverse reinforcement learning. *arXiv preprint arXiv:2310.12055*, 2023.

Ali Baheri and Mykel J Kochenderfer. The synergy between optimal transport theory and multi-agent reinforcement learning. *arXiv preprint arXiv:2401.10949*, 2024.

Ali Baheri, Zahra Shahrooei, and Chirayu Salgarkar. Wasserstein adaptive value estimation for actor-critic reinforcement learning. *arXiv preprint arXiv:2501.10605*, 2025.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Georgios Chalkiadakis and Craig Boutilier. Coordination in multiagent reinforcement learning: A bayesian approach. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 709–716, 2003.

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

Sam Devlin and Daniel Kudenko. Plan-based reward shaping for multi-agent reinforcement learning. *The Knowledge Engineering Review*, 31(1):44–58, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023.

Motoki Omura, Yusuke Mukuta, Kazuki Ota, Takayuki Osa, and Tatsuya Harada. Offline reinforcement learning with wasserstein regularization via optimal transport maps. In *Reinforcement Learning Conference*.

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

Zahra Shahrooei and Ali Baheri. Risk-averse reinforcement learning: An optimal transport perspective on temporal difference learning. *arXiv preprint arXiv:2502.16328*, 2025.

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

Yuxuan Yi, Ge Li, Yaowei Wang, and Zongqing Lu. Learning to share in networked multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:15119–15131, 2022.

Lei Yuan, Ziqian Zhang, Lihe Li, Cong Guan, and Yang Yu. A survey of progress on cooperative multi-agent reinforcement learning in open environment. *arXiv preprint arXiv:2312.01058*, 2023.

Biao Zhao, Weiqiang Jin, Zhang Chen, and Yucheng Guo. A semi-independent policies training method with shared representation for heterogeneous multi-agents reinforcement learning. *Frontiers in Neuroscience*, 17:1201370, 2023.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32):1–67, 2024.