Understanding Reward Ambiguity Through Optimal Transport Theory in Inverse Reinforcement Learning

Ali Baheri
Department of Mechanical Engineering
Rochester Institute of Technology
Rochester, NY 14623
akbeme@rit.edu

Abstract

In inverse reinforcement learning (IRL), the central objective is to infer underlying reward functions from observed expert behaviors in a way that not only explains the given data but also generalizes to unseen scenarios. This ensures robustness against reward ambiguity—where multiple reward functions can equally explain the same expert behaviors. While significant efforts have been made in addressing this issue, current methods often face challenges with high-dimensional problems and lack a geometric foundation. This paper harnesses the optimal transport (OT) theory to provide a fresh perspective on these challenges. By utilizing the Wasserstein distance from OT, we establish a geometric framework that allows for quantifying reward ambiguity and identifying a central representation or "centroid" of reward functions. These insights pave the way for robust IRL methodologies anchored in geometric interpretations, offering a structured approach to tackle reward ambiguity in high-dimensional settings.

1 Introduction

In the reinforcement learning (RL), the derivation of reward functions has always been a subject of intricate debates and investigations [1]. Inverse RL (IRL) emerges as a pivotal methodology in this dialogue, where the aim is to decipher the underlying reward structure from observed expert behavior [2]. However, one of the intrinsic challenges that plagues IRL is the issue of reward ambiguity. A fundamental realization in IRL is that multiple distinct reward functions can lead to the same or similar expert behaviors. This many-to-one relationship between reward functions and policies complicates the task of pinpointing a unique reward landscape from demonstrations [2, 3].

Attempts to navigate this conundrum have been diverse. The maximum entropy IRL framework, for instance, integrates a probabilistic lens into IRL [4]. By postulating that expert demonstrations align not solely with optimality but also with an exponential reward distribution, this approach employs entropy maximization to favor reward functions yielding more uniformly distributed trajectories, serving as an inherent countermeasure to reward ambiguity. Further advancements came from Bayesian IRL, where the uncertainty in reward functions is addressed directly by employing a Bayesian framework [5, 6, 7]. Instead of seeking a single reward function, Bayesian IRL estimates a posterior distribution over reward functions given the demonstrations. InfoGAIL, an extension of the GAIL framework, is another noteworthy mention [8, 9]. It incorporates mutual information into the objective, aiming to capture multiple modes of the expert's behavior. By doing so, InfoGAIL effectively disentangles different underlying reasons (latent variables) for the expert's actions, which can be seen as a means to shed light on the ambiguous regions of the reward landscape.

Optimal transport (OT) theory presents a fresh, promising avenue for reward ambiguity problem. The fundamental tenet of OT lies in its ability to measure the *distance* between different probability distributions, or in our context, between various reward functions. This is achieved using a concept

called the Wasserstein distance. By understanding how far apart or close together different reward functions are, one can gain clearer insights into the nebulous territory of reward ambiguity that IRL grapples with. The true potential of OT for IRL is its *geometrically-grounded* framework. Where traditional approaches might only offer probabilistic insights, the marriage of IRL and OT provides a more structured, spatial understanding of reward landscapes. Imagine mapping reward functions within a geometric space, where the proximity between points indicates similarity. OT provides the tools to construct, measure, and interpret this space, and in doing so, can potentially revolutionize our comprehension of reward ambiguity. In this paper, our primary contribution is the fusion of the analytical power of OT with the challenges of reward ambiguity inherent in IRL. We advocate for a synthesis, where the discerning capabilities of IRL meet the geometric precision of OT, aiming to deliver a more holistic solution to one of RL's most enduring challenges.

2 Preliminaries

OT finds its roots in the study of moving and redistributing masses in the most efficient manner [10, 11]. Historically conceived by Gaspard Monge in the late 18th century, its motivations were primarily combinatorial. However, with the reformation of the theory by Kantorovich in the 20th century, it adopted a more functional analysis perspective, making it more broadly applicable [12, 13]. In the context of OT, we deal with two probability measures, μ and ν, defined on a common metric space. The foundational goal of OT is to "transport" μ to ν while minimizing a certain cost associated with moving each infinitesimal unit of mass. This cost is typically governed by a cost function c, which dictates the expense of moving a unit mass across a given distance. Mathematically, the p-Wasserstein distance between two probability measures μ and ν, defined with respect to a ground metric d on the space, is given by:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int d(x, y)^p d\gamma(x, y) \right)^{\frac{1}{p}} \tag{1}$$

In this representation, x and y are points within the metric space over which the probability measures μ and ν are defined. The function $d(x, y)$ computes the distance between these two points, rooted in the foundational ground metric d. Meanwhile, $\Gamma(\mu, \nu)$ signifies the set of all conceivable joint distributions (or transport plans) having μ and ν as marginals. The Wasserstein distance's merit is its efficacy to encapsulate the geometric discrepancies between probability measures. This capability has rendered it invaluable across various domains, from image processing to machine learning, economics, and computational biology [14, 15, 16, 17, 18, 19]. In the context of our study, the appeal of OT is derived from its capacity to numerically elucidate disparities between reward function distributions. This offers a unique perspective to examine the intricacies of reward ambiguity inherent in IRL.

3 Problem Formulation

IRL revolves around the fundamental task of inferring the most probable reward function that an expert is optimizing, given their behavior. A key challenge in IRL is the ambiguity of reward functions: multiple reward functions might give rise to the same, or nearly identical, expert policies. This set of reward functions is denoted as $R(\pi^*)$, where π^* is the expert's policy. A pertinent factor influencing the dynamics of this ambiguity is the dimensionality of the problem. In our context, the state space is represented as $\mathcal{S}$ and the action space as $\mathcal{A}$. The dimensionality, d, encapsulates the total number of state-action pairs and is given by:

$$d = |\mathcal{S}| \times |\mathcal{A}| \tag{2}$$

Higher the dimensionality, the more intricate and vast our reward space becomes, potentially complicating our understanding of $R(\pi^*)$. To navigate this complex space and to better understand the distribution and relationships between these reward functions, we turn to OT theory. This theory allows us to map each reward function in $R(\pi^*)$ to a representation in a space where distances (dissimilarities) between these functions can be meaningfully measured. We term this mapping as Φ, which translates a reward function, R, to its representative point, W, in the Wasserstein space. With this transformation, **our aims become twofold:**

1. Quantification. Measure the Wasserstein distance between different representations in the Wasserstein space, $\mathcal{W}(W_1, W_2)$. This measurement provides insight into the degree of ambiguity among reward functions.

2. Centroid Identification. Beyond just understanding the space, it is crucial to pinpoint a central, representative point in this space, a "centroid" W^*, that stands, on average, closest to all other reward function representations. Mathematically, this can be articulated as:

$$W^* = \text{argmin}_{W \in \Phi(R(\pi^*))} \sum_{W_i \in \Phi(R(\pi^*))} \mathcal{W}(W, W_i) \tag{3}$$

This centroid aids in identifying a "central" reward function representation, which can serve as a reference or benchmark in the Wasserstein space.

4 Theoretical Results

Upon formalizing the inherent challenges of reward ambiguity, particularly in the context of escalating dimensionality, the immediate imperative is to elucidate these complexities with rigorous theoretical underpinnings. The potential of OT, as postulated in the preceding sections, necessitates concrete validation. In this section, a suite of theorems is presented. These results not only affirm our foundational hypotheses but also offer a systematic framework to comprehend and navigate the intricacies of reward function ambiguity in IRL.[1]

Theorem 1: Convergence of Inferred Rewards in Wasserstein Space. Let R_{true} be the true underlying reward function of an expert, and R_n be the inferred reward function based on n expert trajectories. Then, under certain regularity conditions on the environment and expert behavior, as $n \to \infty$:

$$W_p(R_n, R_{\text{true}}) \to 0 \tag{4}$$

where W_p denotes the p-Wasserstein distance.

Implications. The theorem explaining the convergence of inferred rewards in the Wasserstein space provides key guarantees for IRL. At its core, it confirms that, given a broad range of expert demonstrations, our estimated reward structures gradually approach the true underlying reward function driving the expert's decisions. This points to a consistent pattern in expert behaviors, addressing the well-known issue in IRL where multiple reward functions could seemingly explain the same expert actions. Furthermore, the theorem's foundation in OT enhances the reliability of IRL, suggesting that even with potential inaccuracies and variability in real-world demonstrations, the geometric and topological insight from the Wasserstein distance offers a way to identify and get closer to the true reward function. Therefore, this convergence theorem not only strengthens the theoretical basis of IRL but also highlights the crucial role of OT in improving the precision of reward predictions based on expert data.

Theorem 2: Existence and Uniqueness of Optimal Reward. Let $\mathcal{R}$ be a compact metric space representing the space of all reward functions, π^* be an expert's policy, $\mathcal{S} = \{R_1, R_2, \ldots, R_N\}$ be a set of inferred reward functions from $\mathcal{R}$, where each R_i induces a policy equivalent to π^*, and $W_p(\cdot, \cdot)$ be the p-Wasserstein distance induced by a ground metric on $\mathcal{R}$. Then:
1. Existence: There exists a reward function $R_{\text{centroid}} \in \mathcal{R}$ such that

$$R_{\text{centroid}} = \underset{R \in \mathcal{R}}{\text{argmin}} \sum_{i=1}^{N} W_p(R, R_i) \tag{5}$$

This R_{centroid} minimizes the total p-Wasserstein distance to all reward functions in $\mathcal{S}$.

2. Uniqueness: If the ground metric on $\mathcal{R}$ induces strict convexity, then R_{centroid} is unique.

Implications. The theorem elucidating the existence and uniqueness of the optimal reward representation within the Wasserstein space offers profound insights for the IRL. It posits that amidst the vast space of plausible reward functions that can induce a given expert policy, there emerges a central or "centroid" reward function that serves as the most representative embodiment with respect to the Wasserstein distance. This centrality implies a kind of "mean" or "average" behavior which, in the face of reward ambiguities, can provide a benchmark for the modeling of underlying motivations or preferences. Furthermore, the condition for uniqueness, contingent on strict convexity of the ground metric, reassures that this centroid reward is not only representative but singularly distinctive, thereby eliminating concerns of multiple competing central representations.

[1] Proofs for the presented theorems can be found in the supplementary material.

Theorem 3: Robustness of OT-Integrated IRL to Noisy Demonstrations. Let ϵ be a measure of the noise introduced in the demonstrations such that $\epsilon \geq 0$, with $\epsilon = 0$ representing noise-free demonstrations. Then, for every $\epsilon > 0$, there exists a constant $C(\epsilon)$ such that:

$$W_p\left(R^*\left(\pi^*\right), R^*\left(\pi^{*\prime}\right)\right) \leq C(\epsilon) \tag{6}$$

This implies that the distance between reward functions inferred from noisy and noise-free demonstrations remains bounded in the Wasserstein space, ensuring robustness of the OT-integrated IRL process to noisy expert demonstrations.

Implications. The theorem underscoring the robustness of the OT integration in IRL against noisy demonstrations bears significant ramifications for practical implementations. It suggests that even when expert demonstrations are marred by imperfections or inaccuracies – a common occurrence in real-world scenarios – the inferred motivations or reward structures remain steadfastly close to those derived from noise-free demonstrations. This bounded divergence in the Wasserstein space acts as a guarantee of stability, ensuring that the model's interpretation of an expert's intent is not drastically swayed by minor fluctuations or disturbances in the input data.

Theorem 4: Impact of Dimensionality on Reward Ambiguity in Wasserstein Space. Given the set S_d of reward functions for dimension d, let Δ_d be the average pairwise p-Wasserstein distance between reward functions in S_d defined as

$$\Delta_d = \frac{2}{|S_d|\left(|S_d|-1\right)} \sum_{R_i, R_j \in S_d, i<j} W_p\left(R_i, R_j\right) \tag{7}$$

Then, as d increases, Δ_d also increases. This indicates that as the dimensionality of the state-action space grows, the average pairwise Wasserstein distance between potential reward functions grows as well, suggesting a more dispersed reward landscape and thus intensified reward ambiguity.

Implications. The theorem indicates that higher dimensionality in the state-action space corresponds to a greater average pairwise Wasserstein distance between possible reward functions. This, in turn, points to a more scattered distribution of reward functions in the geometric space. The spread of these reward functions implies that as we consider problems with more dimensions, the ambiguity in determining the "true" reward function (among many that can explain the expert behavior) becomes more pronounced.

5 Discussion

OT is effective at analyzing and comparing probability distributions by viewing them as comprehensive entities. However, this approach can pose challenges when dealing with reward functions that exhibit multiple modes. A unimodal representation in the Wasserstein space may not capture all the distinct peaks of a multimodal function. As a result, the overall representation might be oversimplified or inaccurate. To capture the diverse peaks of multimodal reward functions, it could be beneficial to represent them using multiple points or clusters in the Wasserstein space. Each cluster could represent a different peak or mode, allowing for a more detailed depiction of the reward function. Yet, this method introduces its own set of challenges, particularly in computation. Using multiple clusters necessitates specialized clustering algorithms suitable for the Wasserstein space, which could increase computational demands. Nonetheless, recent improvements in OT algorithms provide promising solutions to these challenges [20]. While employing multiple clusters enhances the representation's accuracy, it can make the decision-making process more complex. How should an agent prioritize between multiple modes of high reward? Incorporating these insights into actionable policies might require additional layers of analysis or even a hierarchical approach to decision-making.

6 Conclusion

IRL faces with issues such as reward ambiguity, where multiple reward functions can justify identical expert behaviors. In this paper, we presented an approach by incorporating optimal transport theory, specifically utilizing the Wasserstein distance. This method brings forth a geometric insight into reward ambiguity, paving the way for a more systematic understanding. While the theoretical underpinnings presented in this work shed light on several facets of the reward ambiguity challenge, a crucial avenue for future research is empirical validation. Rigorous testing and application of our theoretical insights to real-world IRL scenarios will be essential to fully gauge the efficacy and broad applicability of our proposed framework.

7 Acknowledgment

This research was supported in part by the National Science Foundation (NSF) under Award No. 2132060.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[3] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[4] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[5] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

[6] Jaedeug Choi and Kee-Eung Kim. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. *Advances in neural information processing systems*, 25, 2012.

[7] Jaedeug Choi and Kee-Eung Kim. Map inference for bayesian inverse reinforcement learning. *Advances in neural information processing systems*, 24, 2011.

[8] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. *Advances in neural information processing systems*, 30, 2017.

[9] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

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

[11] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[12] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006.

[13] Leonid Vitalevich Kantorovich. On a problem of monge. *Journal of Mathematical Sciences*, 133(4):1383–1383, 2006.

[14] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

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

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

[17] Yicheng Luo, Zhengyao Jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. *arXiv preprint arXiv:2303.13971*, 2023.

[18] Alfred Galichon and Bernard Salanié. Cupid's invisible hand: Social surplus and identification in matching models. *The Review of Economic Studies*, 89(5):2600–2629, 2022.

[19] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[20] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86), 2017.

8 Supplementary Material

Proof of Theorem 1.

Lemma 1. For every trajectory and any inferred reward function R, there exists a constant C such that: $|R(s,a)| \leq C$ for all state-action pairs (s,a).

Proof. The reward values should be bounded for practical tasks and environments; otherwise, they could lead to unbounded utility values making the problem ill-posed.

Lemma 2. *Consistency of IRL Inference.* Given the true reward function $R_{\text{true,}}$, as the number of expert trajectories $n \to \infty$, the expected reward under the inferred policy π_n based on R_n converges to the expected reward under the expert's policy π^* based on R_{true}.

Proof. Let's consider the set of all possible trajectories τ generated using the expert's policy π^*. Since the policy is fixed and deterministic, the distribution of trajectories is also fixed, denoted by $p(\tau \mid \pi^*)$. Given a sufficiently large number n, the likelihood of observing a particular trajectory τ using the inferred policy π_n should approach the likelihood of observing τ using the expert's policy π^*. Therefore, as $n \to \infty$:

$$\mathbb{E}_{\pi_n}\left[R_n(s,a)\right] \to \mathbb{E}_{\pi^*}\left[R_{\text{true}}(s,a)\right] \tag{8}$$

Proposition. *Compactness in Wasserstein Space.* Given a set of inferred reward functions $\{R_n\}$ based on increasing number of expert trajectories, the sequence is relatively compact in the Wasserstein space.

Proof of Theorem:

From Lemma 1, we know the rewards are bounded. This implies that the set of potential reward functions resides in a compact space. From Lemma 2, as $n \to \infty$, the difference in expected rewards under the inferred policy and the expert's policy diminishes. Given the compactness of the space of reward functions (from the proposition) and the convergence of expected rewards, it implies that the sequence $\{R_n\}$ has a convergent subsequence in the Wasserstein space. By the properties of Wasserstein convergence, this convergent subsequence must converge to the true reward function R_{true} in the Wasserstein metric. Thus, for any $\epsilon > 0$, there exists an N such that for all $n > N$:

$$W_p\left(R_n, R_{\text{true}}\right) < \epsilon \tag{9}$$

This implies that as $n \to \infty, W_p\left(R_n, R_{\text{true}}\right) \to 0$.

Proof of Theorem 2.

Lemma 1. The sum of p-Wasserstein distances, as a function on $\mathcal{R}$, is lower bounded and continuous.

Proof. *Lower Bounded*: By the non-negativity of the Wasserstein distance, the sum is clearly non-negative and thus lower bounded by 0. *Continuity*: The Wasserstein distance itself is continuous by its definition, and the sum of continuous functions remains continuous. Given that $\mathcal{R}$ is compact and the sum of p-Wasserstein distances is continuous on $\mathcal{R}$, by the Extreme Value Theorem, the function attains its minimum. Hence, a minimizer R_{centroid} exists.

Lemma 2. If the ground metric space is strictly convex, then the p-Wasserstein distance is also strictly convex.

Proof. Assume two distinct reward functions R_1 and R_2 in $\mathcal{R}$. Consider a t such that $0 < t < 1$ and look at the function:

$$f(t) = W_p\left(tR_1 + (1-t)R_2, R_i\right) \tag{10}$$

Given the properties of Wasserstein distances, $f(t)$ is a convex function. However, with strict convexity in the ground metric, $f(t)$ is strictly convex. This means: $f(t) < tW_p\left(R_1, R_i\right) + (1-t)W_p\left(R_2, R_i\right)$ for all i and for all t in $0 < t < 1$. When summing over all i, if R_1 and R_2 were both minimizers, this would lead to a contradiction, as a convex combination of them would yield a strictly smaller sum of Wasserstein distances. Therefore, only one of them can be the minimizer, establishing uniqueness.

Proof of Theorem:

Directly follows from Lemma 1, the sum of p-Wasserstein distances as a function over the compact space $\mathcal{R}$ is lower bounded and continuous. Hence, it attains a minimum. Thus, an R_{centroid} that minimizes this function exists. Directly follows from Lemma 2, if the ground metric is strictly convex, then the p-Wasserstein distance is strictly convex. This strict convexity ensures that the minimizer R_{centroid} is unique.

Proof of Theorem 3.

Let $R^*(\pi^*)$ be the true reward function obtained from the noise-free expert policy π^*. Given the noisy demonstration policy $\pi^{*\prime}$, the reward function inferred is $R^*(\pi^{*\prime})$. Define the noise function as $\eta(s,a)$ for state-action pair (s,a). This function produces a noise value that perturbs the action taken by the expert.

$$\pi^{*\prime}(a \mid s) = \pi^*(a \mid s) + \epsilon\eta(s,a) \tag{11}$$

Using the definition of the Wasserstein distance, we express the distance between these two reward functions:

$$W_p\left(R^*(\pi^*), R^*(\pi^{*\prime})\right) = \left(\int \left|R^*(\pi^*)(s,a) - R^*(\pi^{*\prime})(s,a)\right|^p dsda\right)^{\frac{1}{p}} \tag{12}$$

Because the only difference between $R^*(\pi^*)$ and $R^*(\pi^{*\prime})$ is the noise, express this difference as a function of ϵ :

$$R^*(\pi^{*\prime})(s,a) = R^*(\pi^*)(s,a) + g(\epsilon,s,a) \tag{13}$$

where $g(\epsilon,s,a)$ is the change in reward due to the noise. Using the Triangle Inequality, we bound the difference:

$$\left|R^*(\pi^*)(s,a) - R^*(\pi^{*\prime})(s,a)\right| \leq \left|g(\epsilon,s,a)\right| \tag{14}$$

Plugging this into our expression for W_p, we get:

$$W_p\left(R^*(\pi^*), R^*(\pi^{*\prime})\right) \leq \left(\int |g(\epsilon,s,a)|^p dsda\right)^{\frac{1}{p}} \tag{15}$$

Assume $g(\epsilon,s,a)$ can be upper-bounded by some function $h(\epsilon)$ that only depends on ϵ. This means that for each $\epsilon > 0$, the change in the reward due to the noise is bounded by $h(\epsilon)$.

$$W_p\left(R^*(\pi^*), R^*(\pi^{*\prime})\right) \leq h(\epsilon) \tag{16}$$

Let $C(\epsilon) = h(\epsilon)$. Therefore, the Wasserstein distance between the reward functions obtained from the true and noisy demonstrations remains bounded by $C(\epsilon)$, showing the robustness of the approach to noisy demonstrations.

$$W_p\left(R^*(\pi^*), R^*(\pi^{*\prime})\right) \leq C(\epsilon) \tag{17}$$

Proof of Theorem 4.

We assume:

1. The reward functions in $\mathcal{R}_d$ are Lipschitz continuous.
2. The state-action space $\mathcal{S} \times \mathcal{A}$ is a compact subset of $\mathbb{R}^d$.
3. The set of reward functions S_d that induce the same policy is non-empty and has positive measure for each dimensionality d.

Lemma. For any fixed policy π^*, the variance of the reward functions in S_d that induce π^* grows with the dimensionality d.

Proof. Consider two reward functions $R_1, R_2 \in S_d$ that both lead to π^* and are sampled randomly (with respect to some fixed distribution on S_d). The variance for a specific state-action pair (s,a) can be considered as:

$$\operatorname{Var}(R(s,a)) = E\left[R^2(s,a)\right] - E^2[R(s,a)] \tag{18}$$

For the entire space, the average variance is:

$$V_d = \int_{\mathcal{S}\times\mathcal{A}} \operatorname{Var}(R(s,a)) dsda \tag{19}$$

Given our assumptions, as d increases, the state-action space $\mathcal{S} \times \mathcal{A}$ becomes more diverse. Consequently, V_d should also increase due to the Lipschitz continuity of the reward functions and the compactness of the state-action space.

Main Theorem:

The average pairwise p-Wasserstein distance, Δ_d, grows with dimensionality d.

Proof. Given the above lemma, we know that the variance of reward functions grows as d increases. By the properties of Wasserstein distances, when the variance of a distribution increases, the Wasserstein distance also typically grows. More formally, consider two reward functions R_1, R_2 in S_d. The p-Wasserstein distance between them is:

$$W_p\left(R_1, R_2\right) = \left(\int_{\mathcal{S}\times\mathcal{A}} \left|R_1(s,a) - R_2(s,a)\right|^p dsda\right)^{\frac{1}{p}} \tag{20}$$

With the growth of d, the Lipschitz continuity and our lemma suggest that the average difference between $R_1(s,a)$ and $R_2(s,a)$ for any random (s,a) pair will increase. Consequently, $W_p\left(R_1, R_2\right)$ will also grow, on average. Thus, integrating over all possible pairs R_1, R_2 in S_d, it follows that:[2]

$$\Delta_d = \frac{2}{\left|S_d\right|\left(\left|S_d\right| - 1\right)} \sum_{R_i, R_j \in S_d, i<j} W_p\left(R_i, R_j\right) \tag{21}$$

will increase as d grows, which completes our proof.

Remarks. The proof provides a high-level understanding grounded in the inherent properties of high-dimensional spaces and the complexity of policy definition.

[2]The term $\frac{2}{|S_d|(|S_d|-1)}$ arises from considering the number of unique pairwise combinations of reward functions in S_d.