Rationality Measurement and Theory for Reinforcement Learning Agents

Kejiang Qian [1] **Amos Storkey** [1] **Fengxiang He** [1]

Abstract

This paper proposes a suite of rationality measures and associated theory for reinforcement learning agents, a property increasingly critical yet rarely explored. We define an action in deployment to be perfectly rational if it maximises the hidden true value function in the steepest direction. The expected value discrepancy of a policy's actions against their rational counterparts, culminating over the trajectory in deployment, is defined to be expected rational risk; an empirical average version in training is also defined. Their difference, termed as rational risk gap, is decomposed into (1) an extrinsic component caused by environment shifts between training and deployment, and (2) an intrinsic one due to the algorithm's generalisability in a dynamic environment. They are upper bounded by, respectively, (1) the 1-Wasserstein distance between transition kernels and initial state distributions in training and deployment, and (2) the empirical Rademacher complexity of the value function class. Our theory suggests hypotheses on the benefits from regularisers (including layer normalisation, ℓ_2 regularisation, and weight normalisation) and domain randomisation, as well as the harm from environment shifts. Experiments are in full agreement with these hypotheses. The code is available at https://github.com/EVIEHub/Rationality.

1. Introduction

Reinforcement learning is rapidly advancing toward human-level capabilities in many domains, such as robotics (Nguyen & La, 2019), autonomous vehicles (Feng et al., 2023), finance (Liu et al., 2022b), and reasoning in large language models (LLMs) (Shao et al., 2024). They are increasingly embedded in real-world, high-stakes systems that directly impact human lives and social fabric. For example, we can expect to share public roads with autonomous vehicles in the near future; in financial markets, reinforcement learning already accounts for a substantial proportion of trading activities. The increasing penetration of reinforcement learning agents in society calls for an understanding of their behaviours through the economic lens. Rationality is fundamental to this end: it characterises the capability of agents to make decisions that maximise their utilities given accessible information, making it possible to economically study agent behaviour (von Neumann & Morgenstern, 1944; Dayan & Daw, 2008; Sen, 1994).

However, the rationality of reinforcement learning is rarely touched on in the literature. To address the issues, this paper proposes a suite of rationality measures for reinforcement learning agents and develops theory based on the measures.

We mathematically define an action to be *perfectly rational* if it maximises the actual value function (though it might be unknown) in the steepest direction. An agent can not be perfectly rational, i.e., of bounded rationality (Simon, 1990; Conlisk, 1996), leading to loss in action-value function, defined as *rational value loss*. This paper is particularly interested in the rationality in deployment (or "inference"). Cumulating the expected rational value loss over the trajectory in inference, we define an *expected rational value risk*. This measure is not directly accessible; we then define an estimator, *empirical rational value risk*, to be the empirical average version in training. Their difference, termed *rational risk gap*, measures the rationality of agents in deployment, given their observable behaviour in training, which is central in the theoretical development in this paper. To note, this suite of measures takes a "local and immediate" perspective: an action is defined to be rational if this *individual* move is optimal, given all information *available at that time*. This setting coincides with a large volume of literature in economics, such as Sen (2002); Gershman et al. (2015).

The rational risk gap is decomposed into two components: (1) an *extrinsic rational gap*, caused by the environment shifts between training and deployment, and (2) an *intrinsic rational gap*, determined by the algorithm itself. This decomposition provides a lens for understanding the sources of sub-rationality. We prove that the two components are upper bounded as follows. The extrinsic rational gap is bounded by $L_s H \cdot W_1(p_0^\dagger, p_0) + H^2 L_s (L_p + 1) \cdot W_1(p^\dagger, p)$, relying on the 1-Wasserstein distance $W_1(p_0^\dagger, p_0)$ between

[1]University of Edinburgh. Correspondence to: Fengxiang He <F.He@ed.ac.uk>.

Preprint. February 5, 2026.

initial state distributions $p_0^\dagger$ in inference and p_0 in training, 1-Wasserstein distance $W_1(p^\dagger, p)$ between transition kernels $p^\dagger$ in inference and p in training, Lipschitz constant L_s of the mapping from state to value function, Lipschitz constant L_p of the mapping from transition kernel to its induced state distributions, and horizon H of an episode. This term may help understand the *sim-to-real transfer* challenge (Da et al., 2025). The intrinsic rational gap has an upper bound $L_\Pi H\sqrt{\log|\mathcal{A}|} + 2\sum_{h=1}^{H}\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 3H^2\sqrt{\frac{\log(H/\delta)}{2T}}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, relying on the empirical Rademacher complexity of value-function class $\mathcal{Q}_\Pi$, Lipschitz constant L_Π of the mapping from policy π to its induced state distribution, the action space cardinality $|\mathcal{A}|$, and the training episode number T.

Our theory suggests empirically testable hypotheses: (1) regularisers, including layer normalisation (Ba et al., 2016), ℓ_2-regularisation, and weight normalisation (Salimans & Kingma, 2016) control the hypothesis complexity of value function class, contributing positively to rationality; (2) domain randomisation (Tobin et al., 2017) improves robustness across environments, also making benefits to rationality, and (3) environment shifts between training and deployments, are harmful to rationality. We conduct experiments to verify these hypotheses, employing Deep Q-Network (DQN) (Mnih et al., 2013) on the Taxi-v3 (Dietterich, 2000) and Cliff Walking environments (Sutton & Barto, 2018). The empirical results are in full agreement with the hypotheses.

To our best knowledge, this work is the first to develop a mathematical framework for measuring the rationality of reinforcement learning agents, indicating sources of irrationality. Our theory sheds light on understanding and improving the rationality of reinforcement learning, which is increasingly critical in this era, as we are inevitably and irreversibly marching into a human-AI co-existing society.

1.1. Related Works

Rationality of machine learning Efforts to study the rationality of machine learning are seen in the literature. Valiant (1995) proposes a philosophical definition: rationality is the ability to abstract and utilise available information to understand, predict, and control the environment, with a probably approximately correct (PAC) style criterion. Abel (2019) provides a formal characterisation for bounded rationality of reinforcement learning, showing that rational decisions depend on how agents represent environments, balancing simplicity and predictive accuracy. Analysing behavioural data from human participants, Evans et al. (2025) introduces the Wasserstein distance between the learned policy and prior as a constraint to model bounded rationality in reinforcement learning. Despite these conceptual formalisations and empirical works, a theoretical framework remains absent, summarised by Macmillan-Scott & Musolesi (2025).

Generalisability of reinforcement learning Generalisation in reinforcement learning is more subtle than in supervised learning because here data is generated by correlated trajectories, and learned policies influence observation. Existing papers establish within-environment guarantees in finite MDPs via PAC and regret analyses (Strehl et al., 2009; Jaksch et al., 2010; Azar et al., 2017); and approximate dynamic programming characterises how estimation and approximation errors propagate through Bellman backups (Munos & Szepesvári, 2008). More recent work replaces dependence on state space with structural complexity measures for rich observations, e.g., Eluder Dimension and Bellman-type ranks (Russo & Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Jin et al., 2021). For deep reinforcement learning, Liu et al. (2022a) casts temporal-difference error as a generalisation problem under neural function approximation; and Wang et al. (2019) analyses the generalisation gap in the reparameterisable settings.

2. Preliminaries

Episodic Markov decision process (EMDP) Suppose an agent, at state $s \in \mathcal{S}$, takes an action a from a finite space $\mathcal{A}$ that transits her to state s' sampled from transition kernel $p(\cdot \mid s, a) \in \Delta(\mathcal{S})$, and then receives an immediate reward r. The action is sampled from policy $\pi \in \Pi$, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, relying on state s. We assume the learning process is "episodic": agents, in every episode, start at initial states s_1 drawn from distribution $p_0(\cdot) \in \Delta(\mathcal{S})$, run for H time steps (i.e., the horizon), and yield returns $\sum_{h=1}^{H} r_h$ ($r_h \in [0,1]$). Intermediate policies $\pi_t = \{\pi_h^t\}_{h=1}^{H}$ are generated during the training of T episodes in total. Given a policy π and a transition kernel p, a trajectory $\mathbf{s}_{h:H}^t = (s_h^t, s_{h+1}^t, \cdots, s_H^t)$ is taken in episode t. This setting is termed EMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^{H}, \{p_h\}_{h=1}^{H}, p_0)$.

A policy can be evaluated by action-value function $Q_h^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{j=h}^{H} r_j \mid s_h = s, a_h = a\right]$ and value function $V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{j=h}^{H} r_j \mid s_h = s\right]$. A terminal condition $V_{H+1}^\pi(s) = 0$ indicates that no reward is gained beyond the horizon H. Value functions are recursively defined, governed by the Bellman equations: $V_h^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_h^\pi(s,a)]$, $Q_h^\pi(s,a) = r_h + \mathbb{E}_{s'\sim p_h(\cdot|s,a)}\left[V_{h+1}^\pi(s')\right]$.

Training-to-deployment shifts Suppose the training environment has transition kernels $p = \{p_h\}_{h=1}^{H}$ and an initial state distribution p_0. A policy π induces a state distribution $\mathcal{D}_h^\pi$ at time step h, termed *state distribution in training* (Cobbe et al., 2020; Wang et al., 2020). Similarly, the deployment environment has different transition kernels $p^\dagger = \{p_h^\dagger\}_{h=1}^{H}$ and initial distribution $p_0^\dagger$, under which a *state distribution in deployment* $\mathcal{D}_h^{\pi,\dagger}$ is induced.

Following Liu et al. (2022a); Wang et al. (2019), this paper assumes the *episode independence*, defined as below,

Assumption 1 (episode independence). For any $t = 1, \dots, T$, state s_h^t is sampled from a distribution $\mathcal{D}_h^{\pi_t}$, i.e., $s_h^t \sim \mathcal{D}_h^{\pi_t}$. The variables $\mathbf{s}_h^{1:T} = \{s_h^t\}_{t=1}^T$ are independent, but not necessarily identically distributed.

In this setting, the objective of a reinforcement learning algorithm is to find an optimal policy π^* that maximises the expected rewards over the trajectory in deployment:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{s_h \sim \mathcal{D}_h^{\pi,\dagger}} \left[V_h^{\pi,\dagger}(s_h) \right].$$

This paper employs Wasserstein distance (Kantorovich, 1960; Villani, 2008) to measure environment shifts.

Definition 1 (p-Wasserstein distance). Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a metric space equipped with a distance function d. μ and ν are two probability measures on $\mathcal{S}$. For any $p \geq 1$, the *p-Wasserstein distance* between μ and ν is defined to be

$$W_p(\mu, \nu) \triangleq \left(\inf_{\gamma} \int_{\mathcal{S} \times \mathcal{S}} d(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p},$$

where the infimum is taken over all joint distributions γ on $\mathcal{S} \times \mathcal{S}$ whose marginals coincide with μ and ν.

We employ the Total Variation (TV) distance to measure the distance between policies (Boucheron et al., 2013).

Definition 2 (Total Variation (TV) distance). The TV distance between two distributions μ, ν is defined as,

$$d_\Pi(\mu, \nu) \triangleq \sup_{x \in \mathcal{X}} \frac{1}{2} |\mu(x) - \nu(x)|.$$

The TV distance can be controlled by the Kullback-Leibler (KL) divergence (Pinsker, 1964).

Definition 3 (Kullback-Leibler (KL) divergence). The KL divergence between two distributions μ, ν is defined as,

$$\mathrm{KL}(\mu \| \nu) \triangleq \sum_{x \in \mathcal{X}} \mu(x) \log \frac{\mu(x)}{\nu(x)}.$$

Hypothesis complexity Let the class of value functions be $\mathcal{Q}_\Pi \triangleq \left\{ s \mapsto Q_h^*(s, a_h^\pi) : \pi \in \Pi,\ h \in [H] \right\}$. For brevity, we use $f(s) = Q_h^*(s, a_h^\pi) \triangleq Q_h^{\pi^*}(s, a_h^\pi)$. Rademacher complexity, and its empirical version (Bartlett & Mendelson, 2003), are employed to measure the hypothesis complexity. We define the empirical version here, and present Rademacher complexity in Appendix C.

Definition 4 (empirical Rademacher complexity). Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{S}}$ be a function class and $\mathbf{s}^{1:n} = \{s^i\}_{i=1}^n$ be a sample set. Let $\boldsymbol{\sigma}^{1:n} = (\sigma^1, \dots, \sigma^n)$ be independent Rademacher random variables. The *empirical Rademacher complexity* of $\mathcal{F}$ on the sample set $\mathbf{s}^{1:n}$ is defined as

$$\hat{\mathfrak{R}}(\mathcal{F}, \mathbf{s}^{1:n}) \triangleq \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}^{1:n}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma^i f(s^i) \right].$$

3. Rationality Measures

This section defines a suite of rationality measures for reinforcement learning. To note, we are particularly interested in the rationality in deployment. Intuitively, training processes usually employ gradient descent, or its variants, that optimise objectives in the steepest direction. In light of this, the rationality in training can be characterised by the discrepancy between the hidden actual value function and the objectives in optimisation.

3.1. Rationality Measures

We first define *perfectly rational actions*. For brevity, we also call them *rational actions* if no ambiguity is caused.

Definition 5 (perfectly rational action). An action a_h° is called *perfectly rational*, if its policy maximises the true value function of the state s_h at the time step h:

$$a_h^\circ \sim \pi^\circ(\cdot \mid s_h); \quad \pi^\circ \triangleq \arg\max_{\pi \in \Pi} \mathbb{E}_{a_h \sim \pi} \left[Q_h^{*,\dagger}(s_h, a_h) \right].$$

Remark 1. As mentioned, the main goal is to study the rationality in deployment, which is not episodic.

A reinforcement learning agent may be of *bounded rationality*; i.e., the agent does not always take rational actions. This incurs *rational value loss*, defined as below.

Definition 6 (rational value loss). Let $p^\dagger, p_0^\dagger$ denote the transition kernel and initial state distribution of the inference environment. The rational value loss of action a_h^π drawn from policy $\pi \in \Pi$ at step h is defined as:

$$\mathcal{L}(a_h^\pi, s_h) \triangleq Q_h^{*,\dagger}(s_h, a_h^\circ) - Q_h^{*,\dagger}(s_h, a_h^\pi).$$

From the definitions, we directly prove Lemma 1.

Lemma 1. *If an action is perfectly rational, its rational value loss is zero.*

We then define *expected rational value loss* in deployment.

Definition 7 (expected rational value loss). Given state distribution in deployment $\mathcal{D}_h^{*,\dagger}$ induced by the optimal policy π^*, the expected rational value loss of any policy π at time step h is defined as:

$$\mathcal{R}_h(\pi) \triangleq \mathbb{E}_{s_h \sim \mathcal{D}_h^{*,\dagger}} \left[Q_h^{*,\dagger}(s_h, a_h^\circ) - Q_h^{*,\dagger}(s_h, a_h^\pi) \right].$$

The inference environment is supposed to be unknown; thus, the expected rational value loss is usually inaccessible. We then define an empirical version in training.

Definition 8 (empirical rational value loss)**.** Suppose an agent is trained by T episodes, taking a sequence of states $\{s_h^t\}_{t=1}^T$. Let p, p_0 denote the transition kernels and initial state distribution of the training environment. The empirical rational value loss of a policy π at time step h is defined as the average over the T episodes, as below,

$$\hat{\mathcal{R}}_h(\pi) \triangleq \frac{1}{T}\sum_{t=1}^{T}\left[Q_h^*(s_h^t, a_h^\circ) - Q_h^*(s_h^t, a_h^\pi)\right].$$

By cumulating the expected and empirical rational value loss over a trajectory, we define *expected rational value risk* and *empirical rational value risk* as follows.

Definition 9 (expected rational value risk)**.** Let $\mathcal{D}_h^{*,\dagger}$ denote the state distribution in deployment, at step h, induced by the optimal policy π^*, transition kernels $p^\dagger$, initial state distribution $p_0^\dagger$. The expected rational value risk of a policy π over a trajectory of horizon H is defined as:

$$\mathcal{R}(\pi) \triangleq \sum_{h=1}^{H}\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}\left[Q_h^{*,\dagger}(s_h, a_h^\circ) - Q_h^{*,\dagger}(s_h, a_h^\pi)\right].$$

Definition 10 (empirical rational value risk)**.** Suppose an agent is trained by T episodes, each of horizon H. Let p, p_0 denote the transition kernel and initial state distribution of the training environment. The empirical rational value risk of policy π is defined as the average over T episodes:

$$\hat{\mathcal{R}}(\pi) \triangleq \frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\left[Q_h^*(s_h^t, a_h^\circ) - Q_h^*(s_h^t, a_h^\pi)\right].$$

The gap between the expected and empirical rational value risks, termed *rational risk gap*, reflects how rational an agent is in deployment, given the behaviour in training.

We also define the *asymptotic rational risk gap* as below. It helps understand the asymptotic property of an agent in terms of rationality.

Definition 11 (asymptotic rational risk gap)**.** Under the same conditions of Definitions 9 and 10, we define the asymptotic rational risk gap of an agent to be $\lim_{T\to\infty}\left|\mathcal{R}(\pi) - \hat{\mathcal{R}}(\pi)\right|$.

"Local and immediate" perspective of rationality This paper takes a "local and immediate" perspective for defining rationality measures. For example, an action is defined to be rational if this *individual* move is optimal in terms of the value function, *at the time*. In other words, a rational agent is not expected to have the capabilities of overlooking the global landscape or anticipating the future, in a "global and long-term" view. We appreciate that such a more strategic perspective of rationality is also valuable, which is, however, out of the scope of this paper.

3.2. Decomposition of Rational Risk Gap

We now present a lemma on the decomposition of the rational risk gap, which indicates the sources of sub-rationality.

Lemma 2 (decomposition of rational risk gap)**.** *The rational risk gap* $\left|\mathcal{R}(\pi) - \hat{\mathcal{R}}(\pi)\right|$ *of policy* $\pi \in \Pi$ *over a trajectory of horizon* H *can be decomposed as follows,*

$$\begin{aligned}\left|\mathcal{R}(\pi) - \hat{\mathcal{R}}(\pi)\right| \leq\ & 2\underbrace{\sum_{h=1}^{H}\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h, a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h, a_h^\pi)\right|}_{\text{extrinsic rational gap}} \\ &+ 2\underbrace{\sum_{h=1}^{H}\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h, a_h^\pi) - \frac{1}{T}\sum_{t=1}^{T}Q_h^*(s_h^t, a_h^\pi)\right|}_{\text{intrinsic rational gap}},\end{aligned}$$

where $\mathcal{D}_h^*$ *is the state distribution in training induced by optimal policy* π^**, while* $\mathcal{D}_h^{*,\dagger}$ *is the state distribution in deployment induced by the same policy but under different transition kernels and initial state distribution.*

This lemma suggests that the rational risk gap can be decomposed into two components as follows:

Extrinsic rational gap: the distance between the true value in deployment, $\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h, a_h^\pi)$, and its counterpart in training, $\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h, a_h^\pi)$. Intuitively, it arises from the *training-to-deployment shifts*, closely linking to the more well-known *sim-to-real challenge* (Peng et al., 2018; Tobin et al., 2017; Andrychowicz et al., 2020). Specifically, changes of the transition kernel (p to $p^\dagger$) and of initial state distribution (p_0 to $p_0^\dagger$) induce different state distributions ($\mathcal{D}_h^*$ vs. $\mathcal{D}_h^{*,\dagger}$), and hence different optimal value functions ($Q_h^*(s_h, a_h^\pi)$ vs. $Q_h^{*,\dagger}(s_h, a_h^\pi)$).

Intrinsic rational gap: the difference between the expected value $\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h, a_h^\pi)$ and its empirical version $\frac{1}{T}\sum_{t=1}^{T}Q_h^*(s_h^t, a_h^\pi)$, both in training. This gap is determined by the joint effects of generalisability and the online setting of reinforcement learning, reflecting the capacity to learn the optimal policy in a dynamic environment.

4. Rationality Theory

This section develops theory for the rational risk gap. The theory relies on the following assumptions.

Assumption 2 (Lipschitz-continuous value)**.** L_s is a positive constant. We assume that the value function $V_h^\pi(\cdot)$ is L_s-Lipschitz under distance function d, for any policy $\pi \in \Pi$, $h \in \{1, \ldots, H\}$ and $s, \tilde{s} \in \mathcal{S}$,

$$|V_h^\pi(s) - V_h^\pi(\tilde{s})| \leq L_s d(s, \tilde{s}).$$

Assumption 3 (Lipschitz-continuous transition kernel)**.** L_p is a positive constant. We assume that the mapping from transition kernel to the induced state distribution is L_p-Lipschitz under the 1-Wasserstein distance for any $\pi \in \Pi$,

$$W_1\big(\mathcal{D}_h^{\pi,\dagger}, \mathcal{D}_h^{\pi}\big) \le L_p W_1(p^\dagger, p).$$

Assumption 4 (Lipschitz-continuous policy)**.** L_Π is a positive constant. We assume that the mapping $\pi \mapsto \mathcal{D}_h^\pi$ is L_Π-Lipschitz under the TV distance d_Π,

$$\sup_{f\in\mathcal{Q}_\Pi} \left|\mathbb{E}_{\mathcal{D}_h^\pi}[f] - \mathbb{E}_{\mathcal{D}_h^{\pi'}}[f]\right| \le L_\Pi\, d_\Pi(\pi,\pi'), \quad \forall \pi,\pi' \in \Pi.$$

Assumption 5 (Entropy-regularised policy)**.** We assume that the learned policy has the following KL bound:

$$\sup_{s\in\mathcal{S}} \mathrm{KL}(\pi_{t+1}(\cdot \mid s)\|\pi_t(\cdot \mid s)) \le \alpha.$$

Remark 2. These Assumptions are reasonably mild, following Bukharin et al. (2023); Gottesman et al. (2023); Wang et al. (2019); Schulman et al. (2018); Vieillard et al. (2020). They mean that (1) environments are smooth with respect to (w.r.t.) states, (2) the learned policy satisfies the smoothness condition, and (3) the learned policy does not go too far away.

4.1. Extrinsic Rational Gap Bound

We first study the extrinsic rational gap.

Theorem 1 (extrinsic rational gap bound)**.** *Let $\mathcal{D}_h^{*,\dagger}, \mathcal{D}_h^*$ denote the state distributions in inference and training, respectively. Under Assumptions 2–3, the extrinsic rational gap over a trajectory of horizon H is upper bounded by*

$$\sum_{h=1}^{H} \sup_{\pi\in\Pi} \left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h,a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h,a_h^\pi)\right|$$
$$\le L_s H \cdot W_1(p_0^\dagger, p_0) + H^2 L_s(L_p+1)\cdot W_1(p^\dagger, p).$$

This theorem shows that the extrinsic rational gap is determined by (1) $L_s H \cdot W_1(p_0^\dagger, p_0)$ that arises from the discrepancy between initial state distributions of $p_0^\dagger$ and p_0, and (2) $H^2 L_s(L_p+1)\cdot W_1(p^\dagger,p)$ caused by the difference between transition kernels of $p^\dagger$ and p.

A detailed proof is given in Appendix B.

Proof sketch We first decompose the extrinsic rational gap (ERG) at time step $h \in [H]$ into two terms as follows,

$$\sup_{\pi\in\Pi} \left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h,a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h,a_h^\pi)\right|$$
$$= \underbrace{\sup_{\pi\in\Pi} \left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h,a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*,\dagger}(s_h,a_h^\pi)\right|}_{\mathrm{I}}$$
$$+ \underbrace{\sup_{\pi\in\Pi} \left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*,\dagger}(s_h,a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h,a_h^\pi)\right|}_{\mathrm{II}}.$$

Term I arises from the distance between the state distributions $\mathcal{D}_h^{*,\dagger}$ and $\mathcal{D}_h^*$. Under Assumption 3, this term admits the upper bound Term I $\le W_1(p_0^\dagger, p_0) + (h-1)L_p\, W_1(p^\dagger, p)$. Term II is shown in Lemma 3 that scales linearly with the Wasserstein distance between p and $p^\dagger$, with an additional dependence on the horizon.

Lemma 3. *Under Assumption 2, for any step $h \in [H]$, the optimal value discrepancy between the inference transition kernel $p^\dagger$ and training transition kernel p, under same training distribution $\mathcal{D}_h^*$, satisfies*

$$\sup_{\pi\in\Pi} \left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*,\dagger}(s_h,a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h,a_h^\pi)\right|$$
$$\le (H-h)\, L_s\, W_1(p^\dagger, p).$$

Combining these two terms over a trajectory of horizon H, we obtain an upper bound on the extrinsic rational gap in Theorem 1.

4.2. Intrinsic Rational Gap Bound

We then obtain the following high-probability upper bound for the intrinsic rational gap.

Theorem 2 (intrinsic rational gap bound)**.** *Under Assumptions 3, 4, and 5, let $\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi)$ denote the empirical Rademacher complexity of value function class $\mathcal{Q}_\Pi$ with a sequence of states $\mathbf{s}_h^{1:T} = \{s_h^t\}_{t=1}^T$ at time step $h \in [H]$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the upper bound on intrinsic rational gap is:*

$$\sum_{h=1}^{H} \sup_{\pi\in\Pi} \left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h,a_h^\pi) - \frac{1}{T}\sum_{t=1}^{T} Q_h^*(s_h^t, a_h^\pi)\right|$$
$$\le L_\Pi H \cdot \sqrt{\log|\mathcal{A}|} + 2\sum_{h=1}^{H} \hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 3H^2\sqrt{\frac{\log(H/\delta)}{2T}}.$$

This bound depends on the empirical Rademacher complexity $\sum_{h=1}^H \hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi)$, which measures the capacity of the value function class under finite-sample training. The term $L_\Pi\sqrt{\log|\mathcal{A}|}$ arises from policy shift between the initial uniform policy and the optimal policy π^*, which scales with the logarithm of the action space cardinality $|\mathcal{A}|$. The remaining term is a concentration term that decays at a rate $O(T^{-1/2})$, as the number of training episodes increases.

A detailed proof is provided in Appendix C.

Proof sketch We decompose the intrinsic rational gap into two terms:

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\frac{1}{T}\sum_{t=1}^{T}Q_h^*(s_h^t,a_h^\pi)\right| \\
&\leq \underbrace{\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right|}_{\text{I}} \\
&\quad+\underbrace{\sup_{\pi\in\Pi}\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)-Q_h^*(s_h^t,a_h^\pi)\right]\right|}_{\text{II}}.
\end{aligned}
$$

Term I can be bounded by the following Lemma 4.

Lemma 4 (policy drift bound)**.** *Under Assumptions 4 and 5, let $\mathcal{A}$ be a finite action space and $\pi\in\Pi$ be a policy. Set parameter $\alpha=4\log|\mathcal{A}|/T^2$. At time step $h\in[H]$ over T episodes, we have this policy drift bound,*

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right| \\
&\leq L_\Pi\sqrt{\log|\mathcal{A}|}.
\end{aligned}
$$

Then, we obtain the upper bound for Term II.

Lemma 5 (on-average generalisation bound)**.** *Let $\mathbf{s}_h^{1:T}=\{s_h^1,\ldots,s_h^T\}$ be independent random variables with $s_h^t\sim\mathcal{D}_h^{\pi_t}$ on a space $\mathcal{S}$. Define the averaged state distribution as $\bar{\mathcal{D}}_h\triangleq\frac{1}{T}\sum_{t=1}^{T}\mathcal{D}_h^{\pi_t}$. For any $\delta\in(0,1)$, with probability at least $1-\delta$, we have:*

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\bar{\mathcal{D}}_h}[Q_h^*(s_h,a_h^\pi)]-\frac{1}{T}\sum_{t=1}^{T}Q_h^*(s_h^t,a_h^\pi)\right| \\
&\leq 2\mathfrak{R}(\mathcal{Q}_\Pi)+\sqrt{\frac{H^2\log(1/\delta)}{2T}}.
\end{aligned}
$$

Combining the two lemmas over a trajectory of horizon H, we prove Theorem 2.

4.3. Main Result

We now obtain the main theorem on the rational risk gap bound directly from the two previous subsections.

Theorem 3 (rational risk gap bound)**.** *Under then same conditions of Theorems 1 and 2, for any $\delta\in(0,1)$, with probability at least $1-\delta$, the rational risk gap of policy $\pi\in\Pi$ over T episodes of horizon H can be bounded by:*

$$
\begin{aligned}
&\left|\mathcal{R}(\pi)-\hat{\mathcal{R}}(\pi)\right|\leq\beta_1\cdot W_1(p_0^\dagger,p_0)+\beta_2\cdot W_1(p^\dagger,p) \\
&+2L_\Pi H\cdot\sqrt{\log|\mathcal{A}|}+4\sum_{h=1}^{H}\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi)+6H^2\sqrt{\frac{\log(H/\delta)}{2T}},
\end{aligned}
$$

where $\beta_1=2L_sH$ and $\beta_2=2H^2L_s(L_p+1)$.

4.4. Expected Rational Value Risk Bound

We then obtain the following Corollary 4 from Theorem 3, which establishes an upper bound on the expected rational value risk, indicating the rationality of a learned policy.

Corollary 4 (expected rational value risk bound)**.** *Let a° be a rational action. Let $\pi\in\Pi$ be a learned policy. Under the same conditions of Theorem 1 and 2, for any $\delta\in(0,1)$, with probability at least $1-\delta$, we have*

$$
\begin{aligned}
&\sum_{h=1}^{H}\left[\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h,a_h^\circ)-\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h,a_h^\pi)\right] \\
&\leq\beta_1\cdot W_1(p_0^\dagger,p_0)+\beta_2\cdot W_1(p^\dagger,p)+2L_\Pi H\cdot\sqrt{\log|\mathcal{A}|} \\
&\quad+4\sum_{h=1}^{H}\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi)+6H^2\sqrt{\frac{\log(H/\delta)}{2T}},
\end{aligned}
$$

where $\beta_1=2L_sH$ and $\beta_2=2H^2L_s(L_p+1)$.

4.5. Sim-to-Real Transfer Challenge

Simulation is a common training ground for reinforcement learning, which often suffers from the *reality gap*: a policy that performs excellently in a simulator can fail spectacularly in the real world, which differs in subtle but consequential ways. The mismatch can be reflected in distribution shifts in both observations and transition dynamics, and the learned policy may overfit to simulator-specific quirks rather than robust principles. This challenge is also referred to as the *sim-to-real transfer* Challenge (Tobin et al., 2017; Peng et al., 2018; Andrychowicz et al., 2020).

Our theory provides a novel and powerful lens to study the sim-to-real transfer challenges. Specifically, the extrinsic rational gap partially characterises this challenge and sheds light on how to mitigate it in terms of rationality. In the following section for experiments, we empirically study how environment shifts would negatively influence rationality.

4.6. Asymptotic Rationality

We also directly obtain the following corollary on the asymptotic property of rationality.

Corollary 5 (asymptotic rational risk gap bound)**.** *Under the same conditions of Theorems 1 and 2, for any $\delta\in(0,1)$, with probability at least $1-\delta$, the rational risk gap of policy $\pi\in\Pi$ over T episodes of horizon H can be bounded by:*

$$
\begin{aligned}
&\lim_{T\to\infty}\left|\mathcal{R}(\pi)-\hat{\mathcal{R}}(\pi)\right|\leq\beta_1\cdot W_1(p_0^\dagger,p_0)+\beta_2\cdot W_1(p^\dagger,p) \\
&+2L_\Pi H\cdot\sqrt{\log|\mathcal{A}|}+4\sum_{h=1}^{H}\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi),
\end{aligned}
$$

where $\beta_1=2L_sH$ and $\beta_2=2H^2L_s(L_p+1)$.

4.7. Technical Novelties in Proofs

Our theory cannot be developed directly from the learning theory developed for supervised learning or static domain adaptation (Zhang et al., 2012; He et al., 2024). In our reinforcement learning settings, data is generated online, and the state distribution evolves along two coupled axes: (i) *environment shifts*, caused by mismatched transition dynamics between the training and inference environments, and (ii) *online policy shift within environments*, arising from continual policy updates in a dynamic environment.

As a consequence, the usual i.i.d. (independent and identically distributed) assumption is violated even within an environment that is fundamental in the mentioned literature. The central technical challenge of this work is therefore to define and control the rational risk gap of a reinforcement learning agent under non-i.i.d. conditions and environment shifts.

To address this challenge, our theory avoids assuming i.i.d. trajectories; instead, our theory only requires episodic independence, a weaker and more realistic assumption in reinforcement learning, as seen in Liu et al. (2022a); Wang et al. (2019). We further introduce the Wasserstein distance as a metric to quantify distribution shifts induced by environment shifts and the online setting in one environment, which enables a precise decomposition and control of the rational risk gap in reinforcement learning settings.

5. Experiments

We conduct experiments to empirically verify our measures and theoretical analysis.

5.1. Empirically Testable Hypotheses

A good theory can explain and suggest empirically testable hypotheses (Lakatos, 1968; Popper, 2005). Our theory leads to the following hypotheses.

H1: Benefits of regularisations Regularisers, such as layer normalisation (LN) (Ba et al., 2016), ℓ_2 regularisation (L2), and weight normalisation (WN) (Salimans & Kingma, 2016), can penalise hypothesis complexity. As suggested in Theorem 2, the reduced hypothesis complexity (measured here by the empirical Rademacher complexity), which indicates a smaller rational risk gap, corresponding to an improved rationality.

H2: Benefits of domain randomisation Domain randomisation (DR) is an augmentation technique that randomises parameters of the environment during training (Tobin et al., 2017). It is supposed to improve the robustness of reinforcement learning algorithms against distribution shifts across environments. As suggested in Theorem 1, this further improves the rationality.

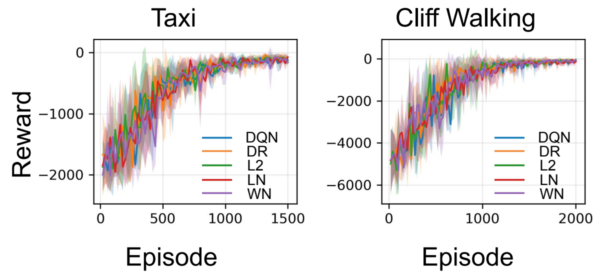


Figure 1. Reward curves of DQN under different regularisation and domain randomisation techniques in Taxi-v3 and Cliff Walking environments.

H3: Deficits of environment shifts Theorem 1 suggests that environment shifts enlarge the rational risk gap, as quantified by the 1-Wasserstein distance between transition kernels $W_1(p, p^\dagger)$ and initial state distributions $W_1(p_0, p_0^\dagger)$. Consequently, this means larger environment shifts lead to worse rationality.

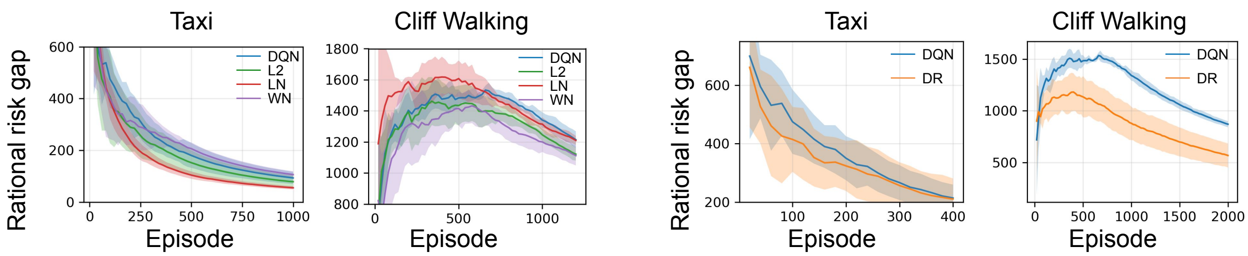
5.2. Implementation Details

We present major implementation details below. Full details are given in Appendix D.

Environment setups Two popular Gym environments are employed in our experiments: Taxi-v3 (Dietterich, 2000) and Cliff Walking (Sutton & Barto, 2018). We modify their environment dynamics to create two distinguished, training and inference settings. For the training environment, we choose the action randomisation rate from 0% to 70%, whereby the environment may override the agent's learned action with a random action. The inference environment takes the original environment without randomisation. Agents are trained under a non-zero probability of action randomisation, and then evaluated in the inference environment. In this way, we simulate distribution shifts between the training and inference environments.

Reinforcement learning algorithm We employ a typical reinforcement learning algorithm, Deep Q-Network (DQN) (Mnih et al., 2013) with softmax action selection, in our experiments.

Training implementations Agents are trained on a finite number of challenge levels, with the probabilities of executing a random action chosen from $\{0\%, 10\%, 30\%, 50\%, 70\%\}$. All results are averaged over five independent runs, with standard deviations reported as shaded regions.



(a) Regularisation *(b)* Domain randomisation

Figure 2. Rational risk gap of DQN under different regularisation and domain randomisation techniques in Taxi-v3 and Cliff Walking environments.

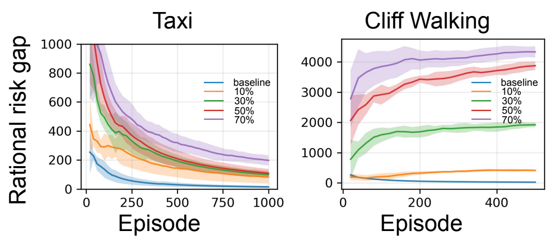


Figure 3. Rational risk gap of DQN across different environment levels in Taxi-v3 and Cliff Walking environments. We evaluate DQN under increasing challenge levels of training environments (0%, 10%, 30%, 50%, 70%), presenting the probability of action randomisation during training.

Experiment design For verifying Hypotheses H1 and H2, agents are trained in both environments with challenge level of 25% and evaluated in the original environments. We repeat experiments with different random seeds and compare rational risk gaps across methods. For Hypothesis H3, we fix the DQN's hyperparameters and vary challenge levels in {0%, 10%, 30%, 50%, 70%}, constructing different transition kernels. We repeat experiments and measure rational risk gaps in original environments.

Justification Our experiment settings ensure that the actual value functions are accessible, enabling rigorous empirical verification of our rationality measures and theory. A more detailed justification is in Appendix D.1.

Reproducibility The code is available at https://github.com/EVIEHub/Rationality.

5.3. Experimental Results

All setups run reasonably well in terms of reward, as shown in Figure 1. This ensures that our experiments are for rationality, controlling irrelevant variables.

H1: Regularisation Figure 2a illustrates the benefits of regularisation on DQN across the considered environments. In both environments, ℓ_2 regularisation consistently reduces rational risk gap; layer normalisation provides a stronger control in Taxi-v3 environment; and weight normalisation is more effective in the Cliff Walking environment compared to vanilla DQN.

H2: Domain randomisation Figure 2b illustrates the benefits of domain randomisation on the rationality. Compared to the DQN baseline, domain randomisation effectively reduces the rational risk gap in both environments, especially in the Cliff Walking environment.

H3: Environment shifts Figure 3 reports the rational risk gap of DQN under different challenge levels of training environments. Rational risk gap shows a clear, positive correlation with the challenge levels, which fully supports the hypothesis that environment shifts are harmful to rationality.

6. Conclusion

We introduce a rationality framework for reinforcement learning agents, an understudied but increasingly important lens for interpreting AI behaviour. We mathematically define perfectly rational actions, and quantify bounded rationality by a rational risk gap. The rational risk gap admits a clean decomposition into an extrinsic component , and an intrinsic one, each controlled by an upper bound. These bounds yield concrete practical implications: regularisation and domain randomisation can reduce intrinsic irrationality, while environment shift predictably worsens extrinsic irrationality. Comprehensive experiments support these predictions, collectively validating our theory.

Acknowledgements

K. Qian was supported in part by the UKRI Grant EP/Y03516X/1 for the UKRI Centre for Doctoral Training in Machine Learning Systems (https://mlsystems.uk/).

References

Abel, D. Concepts in bounded rationality: perspectives from reinforcement learning. *Brown University Master thesis*, 2019.

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Bukharin, A., Li, Y., Yu, Y., Zhang, Q., Chen, Z., Zuo, S., Zhang, C., Zhang, S., and Zhao, T. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68121–68133, 2023.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2048–2056, 2020.

Conlisk, J. Why bounded rationality? *Journal of economic literature*, 34(2):669–700, 1996.

Da, L., Turnau, J., Kutralingam, T. P., Velasquez, A., Shakarian, P., and Wei, H. A survey of sim-to-real methods in rl: Progress, prospects and challenges with foundation models. *arXiv preprint arXiv:2502.13187*, 2025.

Dayan, P. and Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, 2008.

Dietterich, T. G. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research*, 13(1):227–303, 2000.

Evans, B. P., Ardon, L., and Ganesh, S. Modelling bounded rational decision-making through wasserstein constraints. *arXiv preprint arXiv:2504.03743*, 2025.

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., and Liu, H. X. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953): 620–627, 2023. doi: 10.1038/s41586-023-05732-2.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349 (6245):273–278, 2015. doi: 10.1126/science.aac6076.

Gottesman, O., Asadi, K., Allen, C. S., Lobel, S., Konidaris, G., and Littman, M. Coarse-grained smoothness for reinforcement learning in metric spaces. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1390–1410, 2023.

He, Y., Wang, H., Li, B., and Zhao, H. Gradual domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 25(361):1–40, 2024.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 1704–1713, 2017.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: new rich classes of rl problems, and sample-efficient algorithms. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

Kantorovich, L. V. Mathematical methods of organizing and planning production. *Management Science*, 6:366–422, 1960.

Lakatos, I. Criticism and the methodology of scientific research programmes. In *Proceedings of the Aristotelian society*, volume 69, pp. 149–186. JSTOR, 1968.

Liu, F., Viano, L., and Cevher, V. Understanding deep neural function approximation in reinforcement learning via ϵ-greedy exploration. *Advances in Neural Information Processing Systems*, 35:5093–5108, 2022a.

Liu, X.-Y., Xia, Z., Rui, J., Gao, J., Yang, H., Zhu, M., Wang, C., Wang, Z., and Guo, J. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1835–1849, 2022b.

Macmillan-Scott, O. and Musolesi, M. (ir)rationality in ai: State of the art, research challenges and open questions. *arXiv preprint arXiv:2311.17165*, 2025.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

Nguyen, H. and La, H. Review of deep reinforcement learning for robot manipulation. In *2019 Third IEEE international conference on robotic computing (IRC)*, pp. 590–595. IEEE, 2019.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810, 2018.

Pinsker, M. S. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, 1964.

Popper, K. *The logic of scientific discovery*. Routledge, 2005.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv preprint arXiv:1602.07868*, 2016.

Schulman, J., Chen, X., and Abbeel, P. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2018.

Sen, A. The formulation of rational choice. *The American Economic Review*, 84(2):385–390, 1994.

Sen, A. *Rationality and freedom*. Harvard University Press, 2002.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Simon, H. A. Bounded rationality. *Utility and probability*, pp. 15–18, 1990.

Strehl, A. L., Li, L., and Littman, M. L. Pac model-free reinforcement learning. *Journal of Machine Learning Research*, 2009.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free methods. In *Conference on Learning Theory (COLT)*, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press Cambridge, 2018.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Valiant, L. G. Rationality. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 3–14, 1995.

Vieillard, N., Pietquin, O., and Geist, M. Munchausen reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4235–4246, 2020.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2008. doi: 10.1007/978-3-540-71050-9.

von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, 1944. doi: 10.1515/9781400829460.

Wang, H., Zheng, S., Xiong, C., and Socher, R. On the generalization gap in reparameterizable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 6648–6658, 2019.

Wang, K., Kang, B., Shao, J., and Feng, J. Improving generalization in reinforcement learning with mixture regularization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Zhang, C., Zhang, L., and Ye, J. Generalization bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

A. Notation

Table 1. Notation

Symbol	Description
$\mathcal{S}$	State space
$\mathcal{A}$	Finite action space
$\lvert\mathcal{A}\rvert$	Cardinality of action space
H	Horizon length
T	Number of training episodes
s_h^t	State at step h of episode t
π	Stochastic policy, mapping states to action distributions
π^*	optimal policy under transition kernel $p^\dagger$
π_t	Policy used in episode t
p	Transition kernel of the training environment
$p^\dagger$	Transition kernel of the inference environment
p_0	Initial state distribution of the training environment
$p_0^\dagger$	Initial state distribution of the inference environment
$\mathcal{D}_h^\pi$	State distribution at step h induced by policy π under p
$\mathcal{D}_h^{\pi,\dagger}$	State distribution at step h induced by policy π under $p^\dagger$
r_h	Reward function at step h
$V_h^\pi(s)$	Value function of policy π at step h under transition kernel p
$Q_h^\pi(s,a)$	Action value function of policy π at step h transition kernel p
$V_h^{\pi,\dagger}(s)$	Value function of policy π at step h under transition kernel $p^\dagger$
$Q_h^{\pi,\dagger}(s,a)$	Action value function of policy π at step h under transition kernel $p^\dagger$
$\mathcal{R}_h(\pi)$	expected rational value loss of policy π at step h
$\hat{\mathcal{R}}_h(\pi)$	empirical rational value loss of policy π at step h
$\mathcal{R}(\pi)$	expected rational value risk of policy π
$\hat{\mathcal{R}}(\pi)$	empirical rational value risk of policy π
$\mathcal{Q}_\Pi$	Class of value functions $\mathcal{Q}_\Pi : \mathcal{S} \to \mathbb{R}$
$\mathfrak{R}(\mathcal{Q}_\Pi)$	Rademacher complexity of $\mathcal{Q}_\Pi$
$\hat{\mathfrak{R}}(\mathcal{Q}_\Pi)$	Empirical Rademacher complexity of $\mathcal{Q}_\Pi$
$W_1(p^\dagger, p)$	1-Wasserstein distance between transition kernels
L_s	Lipschitz constant of value functions w.r.t. states
L_p	Lipschitz constant of induced state distributions w.r.t. transition kernels
L_Π	Lipschitz constant of induced state distributions w.r.t. policy
$\mathcal{O}(\cdot)$	Asymptotic complexity notation

B. Proof of Theorem 1

In this section, we prove the extrinsic rational risk bound in Theorem 1. We define the integral probability metric (IPM).

Definition 12 (IPM). Let $\mathcal{Q}_\Pi \subseteq \{f : \mathcal{S} \to \mathbb{R}\}$ be a class of bounded measurable functions. For any probability measures μ, ν on $\mathcal{S}$, the integral probability metric (IPM) induced by $\mathcal{Q}_\Pi$ is defined as

$$D_{\mathcal{Q}_\Pi}(\mu,\nu) \triangleq \sup_{f\in\mathcal{Q}_\Pi} \big|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\big|.$$

We now restate our Lemma 3.

Lemma 3. *Under the assumption 2, for any step $h \in [H]$, the optimal value difference between the inference transition kernel $p^\dagger$ and training transition kernel p under same training distribution $\mathcal{D}_h^*$ satisfies*

$$\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*} Q_h^{*,\dagger}(s_h, a_h^\pi) - \mathbb{E}_{s_h\sim\mathcal{D}_h^*} Q_h^*(s_h, a_h^\pi)\right| \le (H-h)\, L_s\, W_1(p^\dagger, p).$$

Proof. For any $h \in \{1, \dots, H\}$ and any $s \in \mathcal{S}$, the Bellman expectation equations give

$$Q_h^*(s, a_h^\pi) = r_h(s, a_h^\pi) + \int_{\mathcal{S}} V_{h+1}^*(s')\, p(\mathrm{d}s' \mid s, a_h^\pi),$$

and

$$Q_h^{*,\dagger}(s, a_h^\pi) = r_h(s, a_h^\pi) + \int_{\mathcal{S}} V_{h+1}^{*,\dagger}(s')\, p^\dagger(\mathrm{d}s' \mid s, a_h^\pi).$$

Subtracting the two equations yields

$$\begin{aligned} &Q_h^{*,\dagger}(s, a_h^\pi) - Q_h^*(s, a_h^\pi) \\ &= \int_{\mathcal{S}} V_{h+1}^{*,\dagger}(s')\, p^\dagger(\mathrm{d}s' \mid s, a_h^\pi) - \int_{\mathcal{S}} V_{h+1}^*(s')\, p(\mathrm{d}s' \mid s, a_h^\pi). \end{aligned}$$

Adding and subtracting $\int_{\mathcal{S}} V_{h+1}^*(s')\, p^\dagger(\mathrm{d}s' \mid s, a_h^\pi)$ inside the integrand gives

$$\begin{aligned} &Q_h^{*,\dagger}(s, a_h^\pi) - Q_h^*(s, a_h^\pi) \\ &= \int_{\mathcal{S}} \big(V_{h+1}^{*,\dagger}(s') - V_{h+1}^*(s')\big) p^\dagger(\mathrm{d}s' \mid s, a_h^\pi) + \int_{\mathcal{S}} V_{h+1}^*(s')\, \big(p^\dagger - p\big)(\mathrm{d}s' \mid s, a_h^\pi). \end{aligned}$$

Since $V_{h+1}^\pi(\cdot)$ is L_s-Lipschitz, by the Kantorovich–Rubinstein duality,

$$\sup_{\pi\in\Pi}\left|\int_{\mathcal{S}} V_{h+1}^*(s')\, \big(p^\dagger - p\big)(\mathrm{d}s' \mid s, a_h^\pi)\right| \le L_s\, W_1(p^\dagger, p).$$

Taking absolute values and using Jensen's inequality,

$$\begin{aligned} \big|Q_h^{*,\dagger}(s, a_h^\pi) - Q_h^*(s, a_h^\pi)\big| \le &\left|\int_{\mathcal{S}} V_{h+1}^{*,\dagger}(s') - V_{h+1}^*(s') p^\dagger(\mathrm{d}s' \mid s, a_h^\pi)\right| \\ &+ L_s\, W_1(p^\dagger, p). \end{aligned}$$

Hence, for any $h \in \{1, \dots, H\}$,

$$\begin{aligned} \sup_{\pi\in\Pi}\sup_{s\in\mathcal{S}}\big|Q_h^{*,\dagger}(s, a_h^\pi) - Q_h^*(s, a_h^\pi)\big| \le &\sup_{\pi\in\Pi}\sup_{s\in\mathcal{S}}\big|Q_{h+1}^{*,\dagger}(s, a_{h+1}^\pi) - Q_{h+1}^*(s, a_{h+1}^\pi)\big| \\ &+ L_s\, W_1(p^\dagger, p). \end{aligned} \tag{1}$$

We now prove by backward induction on h. For all $h \in \{1, \dots, H\}$, we first assume that

$$\sup_{\pi\in\Pi}\sup_{s\in\mathcal{S}}\big|Q_h^{*,\dagger}(s, a_h^\pi) - Q_h^*(s, a_h^\pi)\big| \le (H-h)\, L_s\, W_1(p^\dagger, p). \tag{2}$$

By the terminal condition $V_{H+1}^\pi(\cdot) \equiv V_{H+1}^{\pi,\dagger}(\cdot) \equiv 0$, we have

$$\begin{aligned} &\sup_{\pi\in\Pi}\sup_{s\in\mathcal{S}}\big|Q_H^{*,\dagger}(s, a_h^\pi) - Q_H^*(s, a_h^\pi)\big| \\ &= \sup_{\pi\in\Pi}\sup_{s\in\mathcal{S}}\Big|\int_{\mathcal{S}} V_{H+1}^{*,\dagger}(s')p^\dagger(ds'|s, a_h^\pi) - \int_{\mathcal{S}} V_{H+1}^*(s')p(ds'|s, a_h^\pi))\Big| = 0. \end{aligned}$$

Moreover, note that the RHS of (2) at $h = H$ equals

$$(H - H)\, L_s\, W_1(p^\dagger, p) = 0,$$

so (2) holds for $h = H + 1$.

Fix any $h \in \{1, \dots, H\}$. Assume that (2) holds at time $h + 1$, i.e.,

$$\sup_{\pi\in\Pi} \sup_{s\in\mathcal{S}} \left| Q^{*,\dagger}_{h+1}(s, a^\pi_h) - Q^*_{h+1}(s, a^\pi_h) \right| \le (H - h - 1)\, L_s\, W_1(p^\dagger, p). \tag{3}$$

Applying the recursion (1) and then substituting (3), we obtain

$$\begin{aligned}
\sup_{\pi\in\Pi} \sup_{s\in\mathcal{S}} \left| Q^{*,\dagger}_{h}(s, a) - Q^*_{h}(s, a^\pi_h) \right| &\le \sup_{\pi\in\Pi} \sup_{s\in\mathcal{S}} \left| Q^{*,\dagger}_{h+1}(s, a^\pi_h) - Q^*_{h+1}(s, a^\pi_h) \right| \\
&\quad + L_s\, W_1(p^\dagger, p) \\
&\le (H - h - 1)\, L_s\, W_1(p^\dagger, p) + L_s\, W_1(p^\dagger, p) \\
&= (H - h)\, L_s\, W_1(p^\dagger, p).
\end{aligned}$$

This proves that (2) holds at time h whenever it holds at time $h + 1$.

By backward induction from $h = H + 1$ down to $h = 1$, (2) holds for all $h \in \{1, \dots, H\}$.

The final claim follows since

$$\begin{aligned}
&\sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^{*,\dagger}_h(s_h, a^\pi_h) - \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^*_h(s_h, a^\pi_h) \right| \\
&\qquad \le \sup_{\pi\in\Pi} \sup_{s\in\mathcal{S}} \left| Q^{*,\dagger}_h(s, a^\pi_h) - Q^*_h(s, a^\pi_h) \right| \\
&\qquad \le (H - h)\, L_s\, W_1(p^\dagger, p).
\end{aligned}$$

□

We are ready to prove the upper bound on the extrinsic rational gap in Theorem 1.

Theorem 1 (extrinsic rational gap bound)**.** *Let $\mathcal{D}^{*,\dagger}_h, \mathcal{D}^*_h$ denote the state distributions in inference and training. Under Assumptions 2–3, the extrinsic rational gap over a trajectory of horizon H is upper bounded by*

$$\begin{aligned}
&\sum_{h=1}^{H} \sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}^{*,\dagger}_h} Q^{*,\dagger}_h(s_h, a^\pi_h) - \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^*_h(s_h, a^\pi_h) \right| \\
&\qquad \le L_s H \cdot W_1(p^\dagger_0, p_0) + H^2 L_s (L_p + 1) \cdot W_1(p^\dagger, p).
\end{aligned}$$

Proof. We simply add and subtract $\sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}^{*,\dagger}_h} Q^{*,\dagger}_h(s_h, a^\pi_h) - \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^*_h(s_h, a^\pi_h) \right|$ to these two terms.

$$\begin{aligned}
&\sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}^{*,\dagger}_h} Q^{*,\dagger}_h(s_h, a^\pi_h) - \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^*_h(s_h, a^\pi_h) \right| \\
&= \underbrace{\sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}^{*,\dagger}_h} Q^{*,\dagger}_h(s_h, a^\pi_h) - \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^{*,\dagger}_h(s_h, a^\pi_h) \right|}_{\mathrm{I}} \\
&\quad + \underbrace{\sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^{*,\dagger}_h(s_h, a^\pi_h) - \mathbb{E}_{s_h\sim\mathcal{D}^*_h} Q^*_h(s_h, a^\pi_h) \right|}_{\mathrm{II}}.
\end{aligned}$$

Term I. This term describes the discrepancy of distributions induced by the difference between two different transition kernels $p^\dagger$ and p as well as the initial state distributions $p^\dagger_0$ and p_0.

According to the definition of IPM, the first term $\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h,a_h^\pi)-\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^{*,\dagger}(s_h,a_h^\pi)\right|$ satisfies:

$$\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h,a_h^\pi)-\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^{*,\dagger}(s_h,a_h^\pi)\right|\leq D_{\mathcal{Q}_\Pi}(\mathcal{D}_h^{*,\dagger},\mathcal{D}_h^*).$$

To relate this to the difference between kernels $p^\dagger$ and p, we use Assumption 2, which assumes that every $f\in\mathcal{Q}_\Pi$ is L_s-Lipschitz. By the Kantorovich–Rubinstein duality, this obtains

$$D_{\mathcal{Q}_\Pi}(\mathcal{D}_h^{*,\dagger},\mathcal{D}_h^*)\leq L_s\,W_1(\mathcal{D}_h^{*,\dagger},\mathcal{D}_h^*).$$

According to Assumption 3, we can bound the distribution shift by the 1-Wasserstein distance of initial state distributions and transition kernels:

$$\begin{aligned}W_1(\mathcal{D}_h^{*,\dagger},\mathcal{D}_h^*)&\leq\sup_{s\in\mathcal{S}}W_1\big(p_0^\dagger(s),\,p_0(s)\big)+L_p\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}W_1\big(p^\dagger(\cdot\mid s,a),\,p(\cdot\mid s,a)\big)\\&=W_1(p_0^\dagger,p_0)+(h-1)L_p\,W_1(p^\dagger,p).\end{aligned}$$

Thus, the environment shift is bounded by:

$$\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h,a_h^\pi)-\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^{*,\dagger}(s_h,a_h^\pi)\right|\leq L_sW_1(p_0^\dagger,p_0)+(h-1)L_sL_pW_1(p^\dagger,p).\tag{4}$$

Term II. This term quantifies the shift introduced by the difference between the transition kernel $p^\dagger$ in inference and the transition kernel p in training. Based on the lemma 3, we have

$$\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^{*,\dagger}(s_h,a_h^\pi)-\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)\right|\leq(H-h)L_sW_1(p^\dagger,p).\tag{5}$$

Combining these two bounds of 4 and 5,

$$\begin{aligned}&\sum_{h=1}^H\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}}Q_h^{*,\dagger}(s_h,a_h^\pi)-\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)\right|\\&\leq\sum_{h=1}^H\Big[L_s\cdot W_1(p_0^\dagger,p_0)+(h-1)L_sL_p\cdot W_1(p^\dagger,p)+(H-h)L_s\cdot W_1(p^\dagger,p)\Big]\\&\leq L_sH\cdot W_1(p_0^\dagger,p_0)+H^2L_s(L_p+1)\cdot W_1(p^\dagger,p),\end{aligned}$$

which concludes the proof. □

C. Proof of Theorem 2

In this section, we prove the upper bound on the intrinsic rational gap in Theorem 2.

Lemma 7. *Under Assumption 5. Let $\mathcal{A}$ be a finite action space. Assume $\pi_1(\cdot\mid s)$ is uniform over $\mathcal{A}$ for all $s\in\mathcal{S}$, and for some $\alpha>0$, $\sup_{s\in\mathcal{S}}\mathrm{KL}\big(\pi_{t+1}(\cdot\mid s)\,\|\,\pi_t(\cdot\mid s)\big)\leq\alpha,\quad\forall\,t=1,\dots,T-1$. Then for all $t\geq1$,*

$$\begin{aligned}d_\Pi(\pi^*,\pi_t)&\leq d_\Pi(\pi^*,\pi_1)+\sum_{i=1}^{t-1}d_\Pi(\pi_{t+1},\pi_t)\\&\leq\sqrt{\frac{\log|\mathcal{A}|}{2}}+\sqrt{\frac{(t-1)^2\alpha}{2}}.\end{aligned}$$

Proof. By definition, for any $s\in\mathcal{S}$, the total variation distance $d_\Pi(\cdot,\cdot)$ is a metric on the probability simplex over $\mathcal{A}$, and satisfies the triangle inequality. Therefore, for any $s\in\mathcal{S}$,

$$d_\Pi\big(\pi^*(\cdot\mid s),\pi_t(\cdot\mid s)\big)\leq d_\Pi\big(\pi^*(\cdot\mid s),\pi_1(\cdot\mid s)\big)+\sum_{i=1}^{t-1}d_\Pi\big(\pi_{i+1}(\cdot\mid s),\pi_i(\cdot\mid s)\big).$$

Taking the supremum over $s \in \mathcal{S}$ on both sides obtains

$$d_\Pi(\pi^*, \pi_t) \le d_\Pi(\pi^*, \pi_1) + \sum_{i=1}^{t-1} d_\Pi(\pi_{i+1}, \pi_i). \tag{6}$$

Since $\pi_1(\cdot \mid s)$ is uniform over $\mathcal{A}$ for all $s \in \mathcal{S}$, we have for any s,

$$\begin{aligned}\mathrm{KL}\big(\pi^*(\cdot \mid s) \,\|\, \pi_1(\cdot \mid s)\big) &= \sum_{a \in \mathcal{A}} \pi^*(a \mid s) \log \frac{\pi^*(a \mid s)}{1/|\mathcal{A}|} \\ &= \log|\mathcal{A}| + \sum_{a \in \mathcal{A}} \pi^*(a \mid s) \log \pi^*(a \mid s)\big) \\ &\le \log|\mathcal{A}|,\end{aligned}$$

By Pinsker's inequality, for any $s \in \mathcal{S}$,

$$d_\Pi\big(\pi^*(\cdot \mid s), \pi_1(\cdot \mid s)\big) \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}\big(\pi^*(\cdot \mid s) \,\|\, \pi_1(\cdot \mid s)\big)} \le \sqrt{\frac{\log|\mathcal{A}|}{2}}.$$

Taking the supremum over s, we have

$$d_\Pi(\pi^*, \pi_1) \le \sqrt{\frac{\log|\mathcal{A}|}{2}}. \tag{7}$$

By assumption, for all $i = 1, \dots, t-1$,

$$\sup_{s \in \mathcal{S}} \mathrm{KL}\big(\pi_{i+1}(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)\big) \le \alpha.$$

Applying Pinsker's inequality again, we obtain for each i,

$$\begin{aligned}d_\Pi(\pi_{i+1}, \pi_i) &= \sup_{s \in \mathcal{S}} D_{\mathrm{TV}}\big(\pi_{i+1}(\cdot \mid s), \pi_i(\cdot \mid s)\big) \\ &\le \sup_{s \in \mathcal{S}} \sqrt{\tfrac{1}{2}\,\mathrm{KL}\big(\pi_{i+1}(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)\big)} \le \sqrt{\frac{\alpha}{2}}.\end{aligned}$$

Consequently,

$$\sum_{i=1}^{t-1} d_\Pi(\pi_{i+1}, \pi_i) \le (t-1)\sqrt{\frac{\alpha}{2}} = \sqrt{\frac{(t-1)^2\alpha}{2}}. \tag{8}$$

Combining the bounds 6, 7 and 8, we conclude that

$$d_\Pi(\pi^*, \pi_t) \le \sqrt{\frac{\log|\mathcal{A}|}{2}} + \sqrt{\frac{(t-1)^2\alpha}{2}},$$

which completes the proof. □

Then, we restate and prove the policy drift bound in Lemma 4.

Lemma 4 (policy drift bound)**.** *Under Assumptions 4 and 5, let $\mathcal{A}$ be a finite action space and $\pi \in \Pi$ be a policy. Set parameter $\alpha = 4\log|\mathcal{A}|/T^2$. At time step $h \in [H]$ over T episodes, we have this policy drift bound,*

$$\begin{aligned}&\sup_{\pi \in \Pi} \left| \mathbb{E}_{s_h \sim \mathcal{D}_h^*} Q_h^*(s_h, a_h^\pi) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{s_h^t \sim \mathcal{D}_h^{\pi^t}} Q_h^*(s_h^t, a_h^\pi) \right| \\ &\le L_\Pi \sqrt{\log|\mathcal{A}|}.\end{aligned}$$

Proof. This term, $\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right|$, measures the discrepancy between the state distribution $\mathcal{D}_h^*$ induced by the optimal policy π^* and the state distributions $\{\mathcal{D}_h^{\pi_t}\}_{t=1}^T$ induced by the learned policy π_t over T training episodes.

We apply the IPM definition:

$$\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right|\le D_\Pi(\mathcal{D}_h^*,\mathcal{D}_h^{\pi_t}),\quad \forall i=1,\cdots,t-1.$$

According to Assumption 4 and by the triangle inequality, we have

$$\begin{aligned}
&\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{s\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right|\\
&\qquad=\sup_{\pi\in\Pi}\left|\frac{1}{T}\sum_{t=1}^T\left(\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right)\right|\\
&\qquad\le\frac{1}{T}\sum_{t=1}^T\left(\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^*}Q_h^*(s_h,a_h^\pi)-\mathbb{E}_{s_h^t\sim\mathcal{D}_h^{\pi_t}}Q_h^*(s_h^t,a_h^\pi)\right|\right)\\
&\qquad\le\frac{1}{T}\sum_{t=1}^T D_{\mathcal{Q}_\Pi}(\mathcal{D}_h^*,\mathcal{D}_h^{\pi_t})\\
&\qquad\le\frac{L_\Pi}{T}\sum_{t=1}^T d_\Pi(\pi^*,\pi_t)\\
&\qquad\le L_\Pi\left(d_\Pi(\pi^*,\pi_1)+\frac{1}{T}\sum_{t=2}^T\sum_{i=1}^{t-1}d_\Pi(\pi_{i+1},\pi_i)\right).
\end{aligned}$$

We apply Lemma 7 to bound the $d_\Pi(\pi^*,\pi_t)$, which decomposes the distance to the optimal policy into two components: the discrepancy between initial policy π_1 and optimal policy π^*, and the cumulative step size of policy updates, each constrained by the KL divergence.

For some $\alpha>0$, $\sup_{s\in\mathcal{S}}\mathrm{KL}\big(\pi_{t+1}(\cdot\mid s)\,\|\,\pi_t(\cdot\mid s)\big)\le\alpha,\quad\forall\, t=1,\dots,T$. According to Lemma 7, we have:

$$\begin{aligned}
d_\Pi(\pi^*,\pi_t)&\le d_\Pi(\pi^*,\pi_1)+\sum_{i=1}^{t-1}d_\Pi(\pi_{i+1},\pi_i)\\
&\le\sqrt{\frac{\log|\mathcal{A}|}{2}}+\sqrt{\frac{(t-1)^2\alpha}{2}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{L_\Pi}{T}\sum_{t=1}^T d_\Pi(\pi^*,\pi_t)&\le L_\Pi\sqrt{\frac{\log|\mathcal{A}|}{2}}+\frac{L_\Pi}{T}\sum_{t=2}^T\sqrt{\frac{(t-1)^2\alpha}{2}}\\
&\le L_\Pi\sqrt{\frac{\log|\mathcal{A}|}{2}}+L_\Pi\sqrt{\frac{T^2\alpha}{8}}.
\end{aligned}$$

Then, we set the parameter $\alpha=4\log|\mathcal{A}|/T^2$ and obtain:

$$\frac{L_\Pi}{T}\sum_{t=1}^T d_\Pi(\pi^*,\pi_t)\le L_\Pi\sqrt{\log|\mathcal{A}|},$$

and completing the proof. □

We now define the Rademacher complexity on a function class $\mathcal{F}$.

Definition 13 (Rademacher complexity)**.** Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{S}}$ be a function class and $\mathbf{s}^{1:n} = (s^1, \ldots, s^n)$ be independent samples drawn from a distribution over $\mathcal{S}$. Let $\boldsymbol{\sigma}^{1:n} = (\sigma^1, \ldots, \sigma^n)$ be independent Rademacher random variables. The *Rademacher complexity* of $\mathcal{F}$ is defined as

$$\mathfrak{R}(\mathcal{F}) \triangleq \mathbb{E}_{\mathbf{s}^{1:n}} \mathbb{E}_{\boldsymbol{\sigma}^{1:n}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma^i f(s^i) \right].$$

We then restate the on-average generalisation bound in Lemma 5

Lemma 5 (on-average generalisation bound)**.** *Let $\mathbf{s}_h^{1:T} = \{s_h^1, \ldots, s_h^T\}$ be independent random variables with $s_h^t \sim \mathcal{D}_h^{\pi_t}$ on a space $\mathcal{S}$. Define the averaged state distribution $\bar{\mathcal{D}}_h \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathcal{D}_h^{\pi_t}$, and the Rademacher complexity $\mathfrak{R}(\mathcal{Q}_\Pi)$ of value function class $\mathcal{Q}_\Pi$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have:*

$$\sup_{\pi \in \Pi} \left| \mathbb{E}_{s_h \sim \bar{\mathcal{D}}_h} [Q_h^*(s_h, a_h^\pi)] - \frac{1}{T} \sum_{t=1}^{T} Q_h^*(s_h^t, a_h^\pi) \right| \leq 2\mathfrak{R}(\mathcal{Q}_\Pi) + \sqrt{\frac{H^2 \log(1/\delta)}{2T}}.$$

Proof. The proof follows the classical symmetrisation techniques. We define

$$\Phi(s_h^1, \ldots, s_h^T) \triangleq \sup_{f \in \mathcal{Q}_\Pi} \left\{ \mathbb{E}_{s_h \sim \bar{\mathcal{D}}_h} [f(s_h)] - \frac{1}{T} \sum_{t=1}^{T} f(s_h^t) \right\}.$$

Let $\tilde{s}_h^1, \ldots, \tilde{s}_h^T$ be an independent ghost sample with $\tilde{s}_h^t \sim \mathcal{D}_h^{\pi_t}$. Since

$$\mathbb{E}_{s_h \sim \bar{\mathcal{D}}_h} [f(s_h)] = \mathbb{E}_{\tilde{s}_h^1, \ldots, \tilde{s}_h^T} \left[\frac{1}{T} \sum_{t=1}^{T} f(\tilde{s}_h^t) \right],$$

we have

$$\begin{aligned}
\mathbb{E}[\Phi] &= \mathbb{E}_{s_h^1, \ldots, s_h^T} \left[\sup_{f \in \mathcal{Q}_\Pi} \left(\mathbb{E}_{s_h \sim \bar{\mathcal{D}}_h} [f(s_h)] - \frac{1}{T} \sum_{t=1}^{T} f(s_h^t) \right) \right] \\
&= \mathbb{E}_{s_h^1, \ldots, s_h^T} \left[\sup_{f \in \mathcal{Q}_\Pi} \left(\mathbb{E}_{\tilde{s}_h^1, \ldots, \tilde{s}_h^T} \left[\frac{1}{T} \sum_{t=1}^{T} f(\tilde{s}_h^t) \right] - \frac{1}{T} \sum_{t=1}^{T} f(s_h^t) \right) \right] \\
&= \mathbb{E}_{s_h^1, \ldots, s_h^T} \left[\sup_{f \in \mathcal{Q}_\Pi} \left(\mathbb{E}_{\tilde{s}_h^1, \ldots, \tilde{s}_h^T} \left[\frac{1}{T} \sum_{t=1}^{T} f(\tilde{s}_h^t) - \frac{1}{T} \sum_{t=1}^{T} f(s_h^t) \right] \right) \right],
\end{aligned}$$

Jensen's inequality gives

$$\mathbb{E}[\Phi] \leq \mathbb{E}_{s_h^1, \ldots, s_h^T} \left[\mathbb{E}_{\tilde{s}_h^1, \ldots, \tilde{s}_h^T} \left[\sup_{f \in \mathcal{Q}_\Pi} \frac{1}{T} \sum_{t=1}^{T} (f(s_h^t) - f(\tilde{s}_h^t)) \right] \right].$$

Introduce independent Rademacher variables $\sigma_1, \ldots, \sigma_t$. By symmetry,

$$\begin{aligned}
&\mathbb{E}_{s_h^1, \ldots, s_h^T} \left[\mathbb{E}_{\tilde{s}_h^1, \ldots, \tilde{s}_h^T} \left[\sup_{f \in \mathcal{Q}_\Pi} \frac{1}{T} \sum_{t=1}^{T} (f(s_h^t) - f(\tilde{s}_h^t)) \right] \right] \\
&\leq \mathbb{E}_{s_h^1, \ldots, s_h^T} \left[\mathbb{E}_{\tilde{s}_h^1, \ldots, \tilde{s}_h^T} \left[\mathbb{E}_{\boldsymbol{\sigma}^{1:T}} \left[\sup_{f \in \mathcal{Q}_\Pi} \left(\frac{1}{T} \sum_{t=1}^{T} \sigma^t (f(s_h^t) - f(\tilde{s}_h^t)) \right) \right] \right] \right] \\
&\leq 2\, \mathbb{E}_{\mathbf{s}_h^{1:T}} \mathbb{E}_{\boldsymbol{\sigma}^{1:T}} \left[\sup_{f \in \mathcal{Q}_\Pi} \frac{1}{T} \sum_{t=1}^{T} \sigma^t f(s_h^t) \right] \\
&= 2\mathfrak{R}(\mathcal{Q}_\Pi).
\end{aligned} \tag{9}$$

Then, we apply McDiarmid's inequality. In EMDP with bounded reward $0 \le r_h \le 1$, the value function satisfies that $0 \le |f(s) - f(s')| \le H$. Replacing a single sample s_h^t to $\tilde{s}_h^t$, we have

$$\frac{1}{T}\sum_{t=1}^{T} f(s_h^t) \quad \text{by at most} \quad \frac{|f(s_h^t) - f(\tilde{s}_h^t)|}{T} \le \frac{H}{T}.$$

Hence $\Phi(s_h^1, \dots, s_h^T)$ satisfies bounded differences with $c_t = H/T$. Therefore,

$$\Pr\big(\Phi - \mathbb{E}[\Phi] \ge \varepsilon\big) \le \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{T} c_t^2}\right) = \exp\left(-\frac{2T\varepsilon^2}{H^2}\right).$$

Setting the Right-Hand Side (RHS) equal to δ obtains

$$\varepsilon \le \sqrt{\frac{H^2 \log(1/\delta)}{2T}}.$$

Combining this with $\mathbb{E}[\Phi] \le 2\,\mathfrak{R}(\mathcal{Q}_\Pi)$ in equation (9) obtains the high-probability inequality, and we conclude the proof. ☐

We combine the upper bound on policy drift in Lemma 4 and the above bound in Lemma 5 to prove the restated Theorem 2.

Theorem 2 (intrinsic rational gap bound)**.** *Under Assumptions 3, 4 and 5, let $\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi)$ denote the empirical Rademacher complexity of value function class $\mathcal{Q}_\Pi$ with a sequence of states $\mathbf{s}_h^{1:T} = \{s_h^t\}_{t=1}^T$ at time step $h \in [H]$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the upper bound on intrinsic rational gap is:*

$$\sum_{h=1}^{H} \sup_{\pi\in\Pi} \left| \mathbb{E}_{s_h \sim \mathcal{D}_h^*} Q_h^*(s_h, a_h^\pi) - \frac{1}{T}\sum_{t=1}^{T} Q_h^*(s_h^t, a_h^\pi) \right| \\ \le L_\Pi H \cdot \sqrt{\log|\mathcal{A}|} + 2\sum_{h=1}^{H} \hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 3H^2\sqrt{\frac{\log(H/\delta)}{2T}}.$$

Proof. Taking the supremum over $f \in \mathcal{Q}_\Pi$ and using the triangle inequality, we obtain

$$\sup_{f\in\mathcal{Q}_\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}_h^*} f(s_h) - \frac{1}{T}\sum_{t=1}^{T} f(s_h^t) \right| \le \underbrace{\sup_{f\in\mathcal{Q}_\Pi} \left| \mathbb{E}_{s_h\sim\mathcal{D}_h^*} f(s_h) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{s_h\sim\mathcal{D}_h^{\pi_t}} f(s_h^t) \right|}_{\text{Term I}} \\ + \underbrace{\sup_{f\in\mathcal{Q}_\Pi} \left| \frac{1}{T}\sum_{t=1}^{T} \Big(\mathbb{E}_{s_h\sim\mathcal{D}_h^{\pi_t}} f(s_h^t) - f(s_h^t) \Big) \right|}_{\text{Term II}}. \tag{10}$$

Under Assumptions 3 and 4, we have

$$\text{Term I} \;\le\; L_\Pi \sqrt{\log|\mathcal{A}|}. \tag{11}$$

Applying Lemma 5 with the independent (non-identical) sample $s_h^1, \dots, s_h^t$ and the averaged source distribution $\bar{\mathcal{D}}_h$, we obtain that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\text{Term II} \;\le\; 2\,\mathfrak{R}(\mathcal{Q}_\Pi) + \sqrt{\frac{H^2\log(1/\delta)}{2T}}, \tag{12}$$

where the Rademacher complexity is

$$\mathfrak{R}(\mathcal{Q}_\Pi) \triangleq \mathbb{E}_{\mathbf{s}_h^{1:T}} \left[\mathbb{E}_{\boldsymbol{\sigma}^{1:T}} \left[\sup_{f\in\mathcal{Q}_\Pi} \frac{1}{T}\sum_{t=1}^{T} \sigma^t f(s_h^t) \right] \right].$$

We now replace $\mathfrak{R}(\mathcal{Q}_\Pi)$ to the empirical Rademacher complexity with a sequence of states $\mathbf{s}_h^{1:T} = \{s_h^t\}_{t=1}^T$.

$$\hat{\mathfrak{R}}(\mathcal{Q}_\Pi, \mathbf{s}_h^{1:T}) \triangleq \mathbb{E}_{\boldsymbol{\sigma}^{1:T}} \left[\sup_{f \in \mathcal{Q}_\Pi} \frac{1}{T} \sum_{t=1}^T \sigma^t f(s_h^t) \right].$$

By definition, $\mathfrak{R}(\mathcal{Q}_\Pi) = \mathbb{E}_{\mathbf{s}_h^{1:T}} \left[\hat{\mathfrak{R}}(\mathcal{Q}_\Pi, \mathbf{s}_h^{1:T}) \right]$.

Replacing a single sample s_h^t can change $\hat{\mathfrak{R}}(\mathcal{Q}_\Pi)$ by at most H/T, since $|f(s) - f(s')| \leq H$ for all $f \in \mathcal{Q}_\Pi$. Hence, by McDiarmid's inequality, for any $\varepsilon > 0$,

$$\Pr\left(\mathbb{E}_{s_h^1, \ldots, s_h^t} [\hat{\mathfrak{R}}(\mathcal{Q}_\Pi, \mathbf{s}_h^{1:T}] - \hat{\mathfrak{R}}(\mathcal{Q}_\Pi, \mathbf{s}_h^{1:T)} \geq \varepsilon \right) \leq \exp\left(-\frac{2T\varepsilon^2}{H^2} \right).$$

Setting $\varepsilon = H\sqrt{\frac{\log(1/\delta)}{2T}}$ and recalling that $\mathfrak{R}(\mathcal{Q}_\Pi) = \mathbb{E}_{\mathbf{s}_h^{1:T}} \left[\hat{\mathfrak{R}}(\mathcal{Q}_\Pi, \mathbf{s}_h^{1:T}) \right]$, we obtain that with probability at least $1 - \delta$,

$$\left| \mathfrak{R}(\mathcal{Q}_\Pi) - \hat{\mathfrak{R}}(\mathcal{Q}_\Pi, \mathbf{s}_h^{1:T}) \right| \leq 2\sqrt{\frac{H^2 \log(1/\delta)}{2T}}. \tag{13}$$

Combining (12) and (13), and using a union bound over the two probabilistic events, we obtain that with probability at least $1 - \delta$,

$$\text{Term II} \ \leq \ 2\,\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 3\sqrt{\frac{H^2 \log(1/\delta)}{2T}}. \tag{14}$$

Substituting (11) and (14) into the decomposition (10), we conclude that with probability at least $1 - \delta$, for all $h \in [H]$,

$$\begin{aligned}
\sum_{h=1}^H \sup_{f \in \mathcal{Q}_\Pi} \left| \mathbb{E}_{s_h^t \sim \mathcal{D}_h^*} f(s_h^t) - \frac{1}{T} \sum_{t=1}^T f(s_h^t) \right| &\leq \sum_{h=1}^H \left[L_\Pi \sqrt{\log |\mathcal{A}|} + 2\,\hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 3H\sqrt{\frac{\log(H/\delta)}{2T}} \right] \\
&\leq L_\Pi H \cdot \sqrt{\log |\mathcal{A}|} + 2\sum_{h=1}^H \hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 3H^2\sqrt{\frac{\log(H/\delta)}{2T}}.
\end{aligned}$$

We complete the proof.

□

We restate the Corollary 4, which can be directly derived from our Theorems 1 and 2.

Corollary 4 (expected rational value risk bound)**.** *Let a° be the rational action, and let $\pi \in \Pi$ be any policy learned from the training environment. Under the setting of Theorems 1 and 2, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, we have*

$$\begin{aligned}
&\sum_{h=1}^H \left[\mathbb{E}_{s_h \sim \mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^\circ) - \mathbb{E}_{s_h \sim \mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^\pi) \right] \\
&\leq \beta_1 \cdot W_1(p_0^\dagger, p_0) + \beta_2 \cdot W_1(p^\dagger, p) + 2L_\Pi H \cdot \sqrt{\log |\mathcal{A}|} \\
&\quad + 4\sum_{h=1}^H \hat{\mathfrak{R}}_h(\mathcal{Q}_\Pi) + 6H^2\sqrt{\frac{\log(H/\delta)}{2T}},
\end{aligned}$$

where $\beta_1 = 2L_s H$ and $\beta_2 = 2H^2 L_s(L_p + 1)$.

Proof. We prove the result by bounding the expected rational value risk to the same form as the rational risk gap.

By optimality of π°, for any policy $\pi \in \Pi$ we have

$$\frac{1}{T} \sum_{t=1}^T \left[Q_h^*(s_h^t, a_h^\circ) - Q_h^*(s_h^t, a_h^\pi) \right] \geq 0.$$

Adding it to the expected rational value risk has

$$\begin{aligned}
&\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\circ}) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\pi}) \\
&\leq \mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\circ}) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\pi}) \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T} Q_h^{*}(s_h^t, a_h^{\circ}) - \frac{1}{T}\sum_{t=1}^{T} Q_h^{*}(s_h^t, a_h^{\pi}).
\end{aligned}$$

By regrouping terms and applying the triangle inequality, we obtain

$$\begin{aligned}
&\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\circ}) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\pi}) \\
&\leq 2\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\pi}) - \frac{1}{T}\sum_{t=1}^{T} Q_h^{*}(s_h^t, a_h^{\pi})\right|.
\end{aligned}$$

It has the same decomposition as the rational risk gap,

$$\begin{aligned}
&\sum_{h=1}^{H}\left[\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\circ}) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*}(s_h, a_h^{\pi})\right] \\
&\leq 2\sum_{h=1}^{H}\underbrace{\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*,\dagger}} Q_h^{*,\dagger}(s_h, a_h^{\pi}) - \mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h, a_h^{\pi})\right|}_{\text{environment shift}} \\
&\quad + 2\sum_{h=1}^{H}\underbrace{\sup_{\pi\in\Pi}\left|\mathbb{E}_{s_h\sim\mathcal{D}_h^{*}} Q_h^{*}(s_h, a_h^{\pi}) - \frac{1}{T}\sum_{t=1}^{T} Q_h^{*}(s_h^t, a_h^{\pi})\right|}_{\text{intrinsic rational gap}}.
\end{aligned}$$

Applying Theorem 1 to the environment shift term and Theorem 2 to the intrinsic rational gap term completes the proof. □

D. Additional Details of Environments

This appendix presents the tables describing the environments and their division into training and inference settings. Table 2 summarises the key differences across environments, including state and action space dimensions. Tables 3 and 4 show the state components for Taxi-v3 and Cliff Walking environments.

Table 2. Environment information

	State Space	Action Space	Reward Space
Taxi-v3	$\mathcal{S} \in \{0, \ldots, 499\}$	$\mathcal{A} \in \{0, \ldots, 5\}$	$R \subset \mathbb{R}$
CliffWalking-v0	$\mathcal{S} \in \{0, \ldots, 47\}$	$\mathcal{A} \in \{0, \ldots, 3\}$	$r \subset \mathbb{R}$

Table 3. State description for Taxi-v3.

Index	State Component	Description
0	Taxi row position	Discrete grid row index
1	Taxi column position	Discrete grid column index
2	Passenger location	One of four landmarks or in taxi
3	Destination	One of four landmarks

Environment shifts in training and inference. Since the original Taxi and Cliff Walking environments do not distinguish between training and inference settings, we introduce an *action randomisation* mechanism with probability $0 \leq \varepsilon \leq 1$ to

Table 4. State description for Cliff Walking.

Index	State Component	Description
0	Agent row position	Discrete grid row index
1	Agent column position	Discrete grid column index

Algorithm 1 Action randomisation for training and inference environment

Require: Base transition kernel $p^\dagger(\cdot \mid s, a)$, number of states n_S, number of actions n_A, slip probability $\varepsilon \in [0, 1]$
Ensure: Training transition kernel $p(\cdot \mid s, a)$

Compute averaged transition kernel
for $s = 0, 1, \ldots, n_S - 1$ **do**
  Initialise empty kernel $\bar{p}(\cdot \mid s)$
  for $a = 0, 1, \ldots, n_A - 1$ **do**
    for all $(\rho, s', r, \text{done}) \in p^\dagger(\cdot \mid s, a)$ **do**
      $\bar{p}(s', r, \text{done} \mid s) \leftarrow \bar{p}(s', r, \text{done} \mid s) + \frac{1}{n_A}\rho$
    end for
  end for
end for
Randomise actions
for $s = 0, 1, \ldots, n_S - 1$ **do**
  for $a = 0, 1, \ldots, n_A - 1$ **do**
    Initialise empty kernel $p(\cdot \mid s, a)$
    for all $(\rho, s', r, \text{done}) \in p^\dagger(\cdot \mid s, a)$ **do**
      $p(s', r, \text{done} \mid s, a) \leftarrow (1 - \varepsilon)\,\rho$
    end for
    for all $(q, s', r, \text{done}) \in \bar{p}(\cdot \mid s)$ **do**
      $p(s', r, \text{done} \mid s, a) \leftarrow p(s', r, \text{done} \mid s, a) + \varepsilon\, q$
    end for
  end for
end for
Output: p

construct two distinct transition kernels, p and $p^\dagger$. During training ($\varepsilon > 0$), the environment executes the agent's chosen action with probability $1 - \varepsilon$ and replaces it with a uniformly random action with probability ε, which leads to a perturbed transition kernel p. During inference ($\varepsilon = 0$), no action randomisation is applied, and the agent is evaluated under the original transition kernel $p^\dagger$. The pseudo-code for constructing the training transition kernel is provided in Algorithm 1.

D.1. Additional Details of Experimental Settings

Hyperparameters. We report the hyperparameters of DQN used in our experiments. All other hyperparameters are left at their default settings.

Computing the actual value functions $Q_h^{*,\dagger}(\cdot, \cdot)$ and $Q_h^*(\cdot, \cdot)$. The tabular setting allows us to compute the actual value functions under both the inference and training environments, which is generally intractable in non-tabular reinforcement learning settings. We compute $Q_h^{*,\dagger}$ and Q_h^* exactly by backward induction and derive the corresponding optimal policy π^*. To match our experimental setting, we further implement the optimal policy using a softmax parameterisation with a fixed temperature.

Justification on calculating expected rational value risk and rational risk gap The selected environments provide access to the state distributions in both training and deployment, which enables the calculation of the expected rational value risk, and further the rational risk gap.

Table 5. Hyperparameters for Deep-Q Network.

Hyperparameters	
Batch size	64
Replay buffer size	50,000
Softmax temperature τ	10^{-7}
Episodes	5,000
Warm-up steps	1,000
Learning rate	0.001
Target network update period	500
Optimiser	Adam
Hidden dimension	128
Initial exploration rate	1.0
Final exploration rate	0.05
Exploration decay episode	3,000

D.2. Additional Empirical Results

In this section, we present the reward curves and empirical rational value risk of DQN across five challenge levels of the Taxi-v3 and Cliff Walking environments in training, as shown in Figure 4 and Figure 5. They indicate that (1) the training in all cases is running well, and (2) rationality has a negative correlation with the environment shifts.

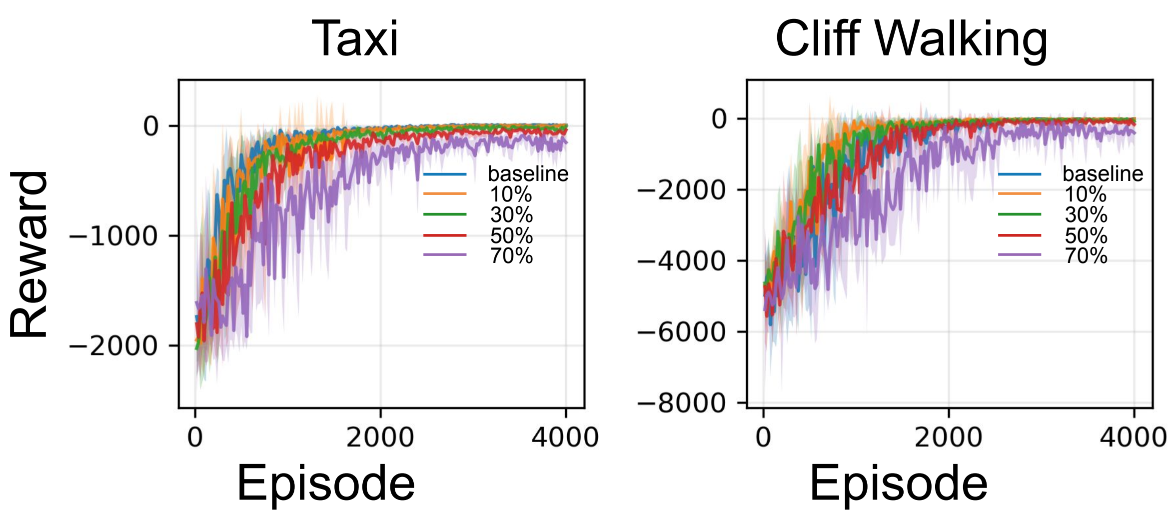


Figure 4. Reward curves of DQN across different environment levels in Taxi-v3 and Cliff Walking environments. We evaluate DQN under increasing challenge levels of training environments (0%, 10%, 30%, 50%, 70%), presenting the probability of action randomisation during training.

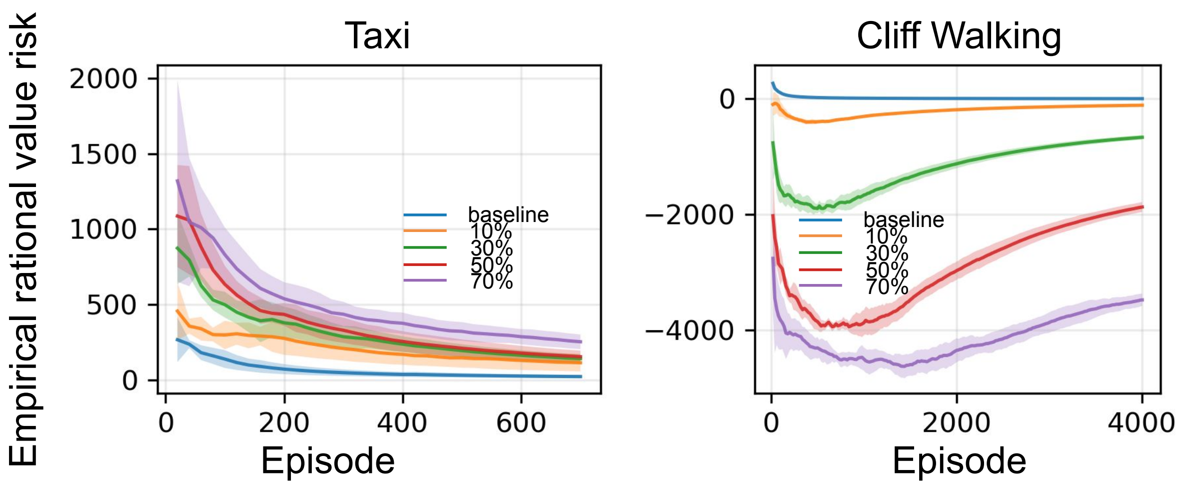


Figure 5. Empirical rational value risk of DQN across different environment levels in Taxi-v3 and Cliff Walking environments. We evaluate DQN under increasing challenge levels of training environments (0%, 10%, 30%, 50%, 70%), presenting the probability of action randomisation during training.